TECNOLÓGICO NACIONAL DE MÉXICO

EDUCACIÓN

SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

PRIMERA EDICIÓN

CIENCIA DE LOS DATOS

Propuestas y casos de uso

Coordinadores:
Rubén Pizarro Gurrola
José Gabriel Rodríguez Rivas
Marco Antonio Rodríguez Zúñiga
Jeorgina Calzada Terrones

CIENCIA DE LOS DATOS

Universidad Pedagógica de Durango

1948-2020 72 ANIVERSARIO

EDUCACIÓN
SECRETARÍA DE EDUCACIÓN PÚBLICA

TECNOLÓGICO NACIONAL DE MÉXICO

CIENCIA DE LOS DATOS

Ciencia de los Datos

Propuestas y casos de uso

Coordinadores:

Rubén Pizarro Gurrola

José Gabriel Rodríguez Rivas

Marco Antonio Rodríguez Zúñiga

Jeorgina Calzada Terrones

Docentes del Tecnológico Nacional de México del Instituto Tecnológico de Durango

Septiembre 2020

9 786078 730100

Ciencia de los Datos.

Propuestas y casos de uso

Primera Edición: Septiembre 2020 ©

Rubén Pizarro Gurrola

José Gabriel Rodríguez Rivas

Marco Antonio Rodríguez Zúñiga

Jeorgina Calzada Terrones

Editado en Durango, Dgo, México

Cuidaron la edición:

Universidad Pedagógica de Durango

Instituto Tecnológico de Durango

Diseño Editorial

Luis Fernando Galindo Vargas

ISBN: 978-607-8730-10-0

Comité Científico Dictaminador

Dr. Omar David Almaraz Rodríguez. Universidad Pedagógica de Durango (UPD).

Dr. Isidro Amaro Rodríguez, Tecnológico Nacional de México. Instituto Tecnológico de Durango.

MES. Josefa del Carmen Hernández Ancona. Tecnológico Nacional de México. Instituto Tecnológico de Campeche

D.C.C. Blanca Cecilia López Ramírez. Tecnológico Nacional de México. Instituto Tecnológico de Roque

MC. Javier Nájera Frías. Tecnológico Nacional de México. Instituto Tecnológico de El Salto.

MC. Liliana Elena Olguín Gil. Tecnológico Nacional de México. Instituto Tecnológico de Tehuacán.

MGTI Eliceo Manuel Martín Pacheco Can. Tecnológico Nacional de México. Instituto Tecnológico de Campeche.

MGTI Elías Gabriel Fidel Pacheco Can. Tecnológico Nacional de México. Instituto Tecnológico de Campeche.

Dr. Rafael Moisés Rosas Sánchez. Tecnológico Nacional de México. Instituto Tecnológico de Tehuacán.

M.S.L. Luis Ramón Sánchez Rico. Tecnológico Nacional de México. Instituto Tecnológico de Roque.

MC. Francisco Vázquez Guzmán. Tecnológico Nacional de México. Instituto Tecnológico de Tehuacán.

MTI. Eduardo Vázquez Zayas. Tecnológico Nacional de México. Instituto Tecnológico de Tehuacán.

Tabla de contenido

Prólogo

Rubén Pizarro Gurrola

Tecnológico Nacional de México. Instituto Tecnológico de Durango

rpizarro@itdurango.edu.mx

El libro es el resultado de una integración y adaptación de diversos productos académicos desarrollados como consecuencia de cursos relacionados con “Ciencia de los Datos e Internet de las Cosas” ofrecidos en el Instituto Tecnológico de Durango durante los años 2018 y 2019.

Los lectores tendrán un panorama claro sobre ¿lo que es?, ¿para qué sirve? ¿qué incluye?, ¿qué herramientas existen?, ¿cómo? y ¿en dónde se puede usar “Ciencia de los Datos”. El Libro va dirigido a la comunidad de científicos de datos, académicos, investigadores, emprendedores, innovadores y público en general que deseen involucrarse con el paradigma Ciencia de los Datos.

El libro "no" es un tutorial para el aprendizaje de herramientas relacionadas con Ciencia de los Datos, "no" es un libro a detalle de temas tales como Bases de datos, *Big Data*, *Machine Learning*, Minería de Datos, entre otros, "no" es una metodología de trabajo; el documento es una integración de distintos temas en donde los autores asocian y relacionan de manera general características del concepto denominado Ciencia de los Datos.

¿Cómo debiera leerse este libro?; el libro contiene distintos tópicos asociados a “Ciencia de los Datos”, el lector "no" debe esperar que cada capítulo sea un antecedente del siguiente, el lector "no" debe pensar que tiene un libro con temas sucesivos en orden de importancia. "no" son capítulos en donde se hace necesario leer un capítulo anterior para comprender el siguiente, es decir, "no" va de temas sencillos a temas más complejos como algún libro de tipo académico. El libro es propiamente una percepción de aspectos que se asocian con “Ciencia de los Datos”.

Ahora bien, al identificar cada capítulo, los lectores que les interese algún tema en particular puede dirigirse y dar lectura al mismo sin necesidad de haber leído algún capítulo anterior, es decir son temas independientes pero relacionados con “Ciencia de los Datos”.

La nube de palabras de la figura 1, identifica las palabras principales de este prólogo.

Figura 1. Nube de palabras del prólogo. Fuente propia

Los datos y la información en las empresas son un activo necesario para la acertada toma de decisiones. Hoy en día y la gran cantidad de datos, la diversidad de formatos y la velocidad que se crean, requiere del uso de herramientas y tecnología para su rápido almacenamiento, procesamiento, análisis e interpretación veraz y oportuna, con ello obtener valor y conocimiento con la finalidad de realizar acciones que busquen la eficiencia y productividad en las organizaciones.

La información proviene de todos lados, sensores que recibe datos, publicaciones en las redes sociales, imágenes y vídeos digitales, registros de compra y transacciones, señales de GPS de los móviles, datos de las aplicaciones WEB, además de los tradicionales medios de comunicación como radio, televisión, prensa entre otros.

La Ciencia de los Datos como eje central de este libro, es un campo multidisciplinario que involucra los procesos y sistemas para extraer el conocimiento o un mejor entendimiento de grandes volúmenes de datos y sus diferentes formas estructurados y no estructurados. Pretende aprovechar aspectos tales como base de datos, tecnología big data, aprendizaje automático (machine learning) e, internet de las cosas (*internet of things*). (Pizarro, Amaro, López, & Galindo, 2018).

Estas tecnologías son parte de la Ciencia de los Datos que se pueden utilizar para transformar los datos en información con la finalidad de obtener razonamiento y comprensión para convertirla de manera inmediata en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en las organizaciones y en las personas. (Pizarro, Amaro, López, & Galindo, 2018).

A pesar de que el término Big Data se asocia principalmente con cantidades de datos exorbitantes, se debe dejar de lado esta percepción, Big Data no va dirigido solo a gran tamaño, sino que abarca tanto volumen como variedad de datos y velocidad de acceso y procesamiento; además la Tecnología de Información y Comunicaciones (TIC) propicia que una gran cantidad se datos, estos deben procesase, entenderse y transformarse en decisiones de valor, esto es el reto del big data (Pizarro, Amaro, López, & Galindo, 2018) .

El concepto de Machine Learning en la Ciencia de los Datos, es una área de la Inteligencia Artificial que engloba un conjunto de tareas, técnicas y algoritmos que hacen posible el aprendizaje automático a través del entrenamiento con grandes volúmenes de datos (Pizarro, Amaro, López, & Galindo, 2018).

Por otra parte, en un futuro cercano cualquier objeto cotidiano estará dotado de algún tipo de sensor que enviará información. El internet de las cosas (IoT) en la

Ciencia de los Datos, está generando volúmenes masivos de datos estructurados y no estructurados.

Las personas con las competencias adquiridas en el uso de éstas y otras herramientas y tecnologías, se dotan con ciertas características que dan lugar al surgimiento de un nuevo perfil profesional, el científico de datos (Data Scientist), que serían las personas capacitadas que deben saber de procesos, de tecnologías, del análisis e interpretación estadística, de comunicación de negocios, entre otros atributos.

El perfil del científico de datos es un ejemplo de la evolución de las profesiones que hacen uso de la información con el uso de la tecnología de información y comunicaciones (TIC), un híbrido entre un programador, analista, comunicador y consejero. Se trata de un profesional dedicado a analizar e interpretar grandes bases de datos. (Pizarro, Amaro, López, & Galindo, 2018).

Con lo anterior y con el ánimo de difundir y promover una comunidad de científicos de datos, en el Instituto Tecnológico de Durango (ITD), se tiene el cuerpo académico en formación denominado “Desarrollo de Tecnologías de Información basada en Ciencia de los Datos” con clave Prodep “ITDUR-CA-14” y cuyo propósito es el generar productos académicos de investigación, vinculación y difusión que sean innovadores y productivos para hacer frente a los nuevos retos de las empresas y organizaciones.

De igual forma en el ITD en los años 2018 y 2019, se impartieron por parte de docentes integrantes del cuerpo académico y colaboradores, dos cursos especiales de titulación relacionados con “Ciencia de los datos e Internet de las cosas” y cuyo objetivo fue entre otras cosas acercar a los exalumnos con necesidad de lograr cerrar su ciclo y culminar su proceso de titulación desarrollando productos académicos relacionados con Ciencia de los Datos.

Como resultado del trabajo desarrollado de los participantes en los cursos, los integrantes y colaboradores del cuerpo académico denominados coordinadores del libro, realizaron un proceso de clasificación, selección, evaluación, ajustes y

modificaciones de los productos académicos concluidos y presentan en este libro los resultados.

El libro que lleva por nombre "**Ciencia de los Datos. Propuestas y casos de uso**", se plantean temas, casos y propuestas de implementación de aspectos relacionados con Ciencia de los Datos, incluye títulos tales como: Bases de datos SQL y NoSQL. Comparativo SQL server & MongoDB; Comparación de herramientas para visualización de datos (Tableau - Power BI); Big Data y su impacto en la sociedad; R como herramienta de Ciencia de los Datos aplicada a la productividad; Big Data: Análisis de estrategias de marketing digital; Comparativo de herramientas para análisis y visualización de datos: Tableau y R; Análisis de datos masivos en el campo de la salud; Herramientas de Big Data; Ciencia de los Datos aplicado en las Pymes; Análisis de Datos Geoespaciales en Protección Civil utilizando R y Python; Machine Learning aplicado a la salud; Análisis comparativo y uso de R y Python enfocado al análisis descriptivo de datos de una entidad financiera.

Cada uno de los trabajos fueron editados por sus autores y coautores; adaptados, modificados y evaluados por los coordinadores de libro. En cada uno de ellos, se trata de manera particular los temas citados. Los trabajos se presentan en la modalidad de capítulos de tal forma que cada capítulo viene incluido la autoría del participante y en algunos casos en coautoría respetando de manera general la esencia, el pensamiento y la forma de redacción de cada autor.

En cada capítulo se presenta una estructura que identifica los elementos de introducción, marco de referencia, desarrollo, conclusiones y la referencia bibliográfica respectiva de cada capítulo.

En la introducción de cada capítulo, se presentan elementos de una justificación del título, el objetivo general y los específicos, el propósito, el impacto y beneficio entre otras cosas.

Con respecto al marco de referencia en cada capítulo, se da a conocer los conceptos que sustentan los aspectos relevantes del título del capítulo.

En el apartado de desarrollo en algunos capítulos, se enriquecen conceptos; en algunos casos se hace un análisis personal de los temas tratados en el marco de referencia; en otros capítulos, se habla de la relación que existe de los conceptos establecidos en el marco de referencia con una propuesta de implementación y caso de uso relacionados con la naturaleza de la Ciencia de los Datos y del tema del capítulo.

En el **capítulo 1**, se presenta el título “Bases de datos SQL y NoSQL. Comparativo SQL server & MongoDB” desarrollado por **Juan Daniel Carrillo Mercado y Gabriel Arturo Lugo Morales**. Los autores presentan aspectos teóricos y conceptuales de las bases de datos SQL y NoSQL; ponen de manifiesto el enfoque relacional y No relacional; identifican la relación de las bases de datos con el paradigma de la Ciencia de los Datos, Inteligencia de Negocios y Big Data; exponen un comparativo de las características principales de los enfoques SQL y NoSQL basado en los atributos de herramientas manejadoras de bases de datos SQL Server y MongoDB.

Al final del capítulo presentan un comparativo y resumen mediante modelo de fortalezas, oportunidades, debilidades y amenazas utilizando la herramienta FODA para destacar los aspectos principales de los dos gestores de bases de datos.

En el **capítulo 2** se presenta el tema “Big Data y su impacto en la sociedad”. desarrollado por “**Carolina Sosa Hernández**” en coautoría con “**José Gabriel Rodríguez Rivas**”. Los autores abordan los conceptos de Big Data, partiendo de los temas relacionados con bases de datos, herramientas y tecnologías de Big Data, entre otros.

En el apartado de desarrollo, la autora y su coautor presentan cinco ejemplos en donde queda de manifiesto el uso del Big Data como casos de éxito y dignos de tomarse en cuenta.

Los ejemplos que se muestran del uso de Big Data son: primer caso. El séptimo arte; segundo caso. ¿Saldrás de viaje?, los Destinos Turísticos Inteligentes

(DTI) son la mejor opción; tercer caso. Un nuevo perfil: El periodista de Datos; cuarto caso. Una campaña política exitosa: Barack Obama y finalmente el quinto caso. BBVA Bancomer y Secretaría de Turismo (SECTUR).

El **capítulo 3** llamado **"**Herramientas Big Data**"** desarrollado por **"Jaime Alonso Chacón Soto"** y **"Jeorgina Calzada Terrones"** tiene mucha relación con el capítulo 2, aquí el autor y su coautora describen algunas de las herramientas relacionadas con Big Data, desde aquellas que tienen propósito para la ingesta de datos, el almacenamiento, el procesamiento, el análisis hasta las que tienen que ver con la visualización de los datos.

Los autores hacer tablas comparaciones destacando características de las distintas herramientas dejando bien el claro las oportunidades de hacer uso de cada una de ellas.

Finalmente, presentan una propuesta en donde mencionan las aplicaciones del Big Data en el ámbito laboral, profesional, social, empresarial y gubernamental, argumentando además algunos casos relacionados con los Big Data, en donde cabe la posibilidad de implementarse. El capítulo termina con un análisis personal sobre lo que se propone y finaliza con sus correspondientes conclusiones y recomendaciones.

En el **capítulo 4** titulado "Big Data. Análisis de estrategias de Marketing Digital", desarrollado por "**María Dolores Concepción De Lara Gurrola", Jesús Eduardo Carrillo Morales**" y "**Luis Fernando Galindo Vargas**" se aborda de igual como en el capítulo 2 y 3 conceptos relacionados con el Big Data, los autores enriquecen con temas de Marketing, Marketing digital, *Inbound marketing* y *Outbound Marketing*.

Al final del capítulo los autores recomiendan estrategias elementales y uso de tecnología existente para el posicionamiento de marca e imagen de la empresa, del producto o servicio que vende y para mejorar la relación con el cliente. En sus análisis y conclusiones del documento, los autores hacen una reflexión de la

relación que existe entre Big Data y el Marketing Digital y las posibilidades que trae consigo el uso de estas tecnologías para las empresas.

El **capítulo 5** titulado “Análisis comparativo y uso de R y Python enfocado al análisis descriptivo de datos de una entidad financiera” desarrollado por “**Nohemí Garcia Hernádez, Claudia Elizabeth Serrato Bacio**” y “**Marco Antonio Rodríguez Zúñiga**”. Se presentan temas y conceptos asociados a estadística descriptiva; Ciencia de Datos, Minería de Datos; análisis de datos y análisis predictivo; Machine Learning así como algunos tópicos relacionados con el proceso de otorgamiento de créditos a clientes

Los autores en el apartado de desarrollo, presentan algunas herramientas que existen en el mercado para análisis predictivo, así mismo, hacen una comparativa entre R y Python y enriquecen el tema con la presentación de herramientas para análisis predictivo, de igual forma dan a conocer un panorama de un proceso de otorgamiento de créditos para una empresa financiera utilizando únicamente R como ejemplo del caso.

Con respecto a la empresa financiera que se cita en el capítulo, los autores omiten el nombre de la misma por cuestiones de confidencialidad y solo se le llamará “la empresa financiera”.

Se menciona sobre el modelo de *ScoreBoard* como una técnica para el otorgamiento de créditos a los clientes de una entidad financiera, se hace referencia de un ejemplo ficticio con datos de dominio público de un análisis de regresión logística para valorar el otorgamiento de crédito a un cliente.

El **capítulo 6** es titulado “Ciencia de los Datos aplicado en las Pymes” desarrollado por “**César Omar Domínguez Gurrola**”, “**Marco Antonio Rodríguez Zúñiga**” y “**Jeorgina Calzada Terrones**”. Los autores en su marco de referencia abordan los principales temas en orden de aparición tales como Business Intelligent o Inteligencia de negocios, Big data, Ciencia de los Datos y la importancia del científico de datos, Machine Learning, Minería de Datos.

Proponen en el contexto del estado de Durango y la importancia de que sea la Ciencia de los Datos un detonante para el desarrollo del Municipio y del Estado.

Por otra parte, en el apartado de desarrollo de éste capítulo los autores ponen de manifiesto un conjunto de casos que bien pudieran ser referentes para que se tomen de ejemplo y se impuse la tecnología relacionada con Ciencia de los Datos y con ello potenciar el desarrollo económico de las PyMES, de la ciudad y en el Estado de Durango.

El **capítulo 7** es titulado “Ciencia de los Datos con R como herramienta aplicada a la productividad” desarrollado por “**Beatriz Eneida Chavez Atayde**” y “**Rubén Pizarro Gurrola**”. La autora y su coautor presentan las características principales del lenguaje de programación R y el entorno de trabajo R Studio.

Realizan una asociación interesante de características del Ingeniero Industrial con el Ingeniero en Sistemas Computacionales de la siguiente manera:

Con respecto a los industriales hace mención en distintas técnicas y herramientas que son esenciales en el campo laboral tales como: Diagrama de Flujo de Procesos, Diagrama de Causa-Efecto, Diagrama de Pareto, Modelo ANOVA, Histograma, Gráfica de Corrida, Gráfica de control, Diagrama de Dispersión, Modelo de Regresión, entre otros. De todas ellas hace hincapié en las siete Herramientas de la Calidad y el modelo Sixsigma.

Con respecto a las características del Ingeniero en Sistemas, la autora y su coautor hacen referencia a que los Ingenieros Industriales pueden enriquecer sus habilidades computacionales y analíticas haciendo el uso del lenguaje R y el entorno R Studio para desarrollar y aplicar las técnicas de calidad mencionadas en el párrafo anterior.

Como propuesta, en el capítulo hacen mención que se utilice R y R Studio como lenguaje y herramienta para potenciar las técnicas analíticas que se mencionaron.

Ofrecen cuatro ejemplos prácticos del uso de R y R Studio mencionando cuatro herramientas: el diagrama de pareto, el modelo ANOVA, el modelo de

correlación y regresión lineal para predicciones y el modelo de causa efecto, todos ellos perfectamente emulabes en R y R Studio.

La autora y su coautor hacen mención del proceso de un *MRP Controller* de una empresa maquiladora de la localidad de Durango.

Terminan concluyendo y con buenos deseos de que sea posible implementar la propuesta de combinar herramientas y técnicas de la Ingeniería Industrial con la programación R, mencionan que se puede potencializar las ventajas por encima de las opciones actuales para el análisis de datos.

El **capítulo 8** lleva por nombre “Análisis de datos masivos en el campo de la salud” desarrollado fue escrito y desarrollado por “**Teresita De Jesús Camacho Cepeda**” y “**Marco Antonio Rodríguez Zúñiga**”.

La autora y su coautor hablan sobre los datos que se generan en el campo de la salud, mencionan lo siguiente: “estos crecen de manera exponencial y a gran velocidad, gracias a la implementación de técnicas y herramientas Big Data se puede capturar, almacenar, procesar y analizar esta ingesta cantidad de datos, acción que con las herramientas convencionales no se puede realizar eficientemente”.

El capítulo aborda los conceptos de análisis de datos, identifican el área de oportunidad que existe de adecuar técnicas y herramientas Big Data en el Sector Salud.

En el capítulo se mencionan aspectos tales como el análisis de datos masivos, tipos de datos, origen de los datos; las herramientas y técnicas más recomendadas para la aplicación de Big Data, la analítica predictiva y se mencionan algunos tipos de algoritmos; el impacto y beneficio de la aplicación de esta nueva tecnología.

Se plantea una propuesta de implementación utilizando Big Data, esta propuesta se trata de unificar todos los datos relacionados con el Sector Salud con el objetivo de brindar atención médica. La propuesta que se documenta, se denomina “Unificación de Contenidos Generales en el Sector Salud”.

El **capítulo 9** titulado “Machine Learning aplicado a la salud” fúe elaborado por su autora “**Sayra María Vargas Arroyo**” y su coautor **“Rubén Pizarro Gurrola**”

El Aprendizaje Automático (Machine Learning), es un área de la Inteligencia Artificial (IA), que engloba un conjunto de técnicas que hacen posible la extracción de conocimiento a través del entrenamiento y validación con grandes volúmenes de datos. Las Técnicas avanzadas de Machine Learning (ML) incluye algoritmos supervisados y no supervisados; tales como máquinas de soporte vectorial (SVM), árboles de decisión, regresión logística para predicciones; clasificación para identificar la pertenencia a un grupo o a una etiqueta y clustering para asociaciones y diferencias de datos en grupos, entre otros, pueden ser utilizadas en la actualidad en muchas áreas de trabajo y por supuesto en la ciencia médica.

La autora y su coautor, describen los conceptos relacionados al Machine Learning; la clasificación de los algoritmos, supervisados y no supervisados; ventajas y desventajas de los algoritmos: presentan ejemplos y casos de uso en el Sector Salud para los diferentes tipos de algoritmos, así como el proceso para la construcción de modelos; mencionan los principales lenguajes de programación y herramientas más usadas para análisis de grandes volúmenes de datos; documentan aplicaciones de Machine Learning es decir, resñonden a la pregunta ¿en dónde se está desarrollando y aplicando?.

El desarrollo del capítulo define una propuesta de trabajo en el Sector Salud, para la implementación de herramientas de Machine Learning, se analizan procesos para llevar a cabo las principales actividades para implementar la propuesta, se define el flujo de trabajo necesario para llevar a cabo un proceso de construcción de un modelo, utilizando estrategias para que el sistema de clasificación sea lo más óptimo posible.

Se presentan casos de uso de predicción del riesgo de cáncer y diagnóstico y del uso de las técnicas de Machine Learning en medicina cardiovascular. En ambos casos se muestran el poder de los algoritmos Machine Learning, observando cuál de los algoritmos es el clasificador perfecto y con la tasa de error más baja.

El **capítulo 10** tiene por nombre “Análisis de datos georreferenciados en protección civil usando R y Python” y fue desarrollado por “**Armando Urbina Retana**” en coautoría con “**Rubén Pizarro Gurrola**”.

El autor y coautor en este capítulo presentan una propuesta que incluye aspectos de análisis de datos geoespaciales mediante lenguajes de programación de uso libre R y Python para integrarlos dentro de un Sistema de Información Geográfico de uso libre (QGIS). La idea se centra en que un QGIS tenga características de visualizar mapas cartográficos que muestren zonas de riesgos en la ciudad de Durango, Dgo. México realizando cruces de datos de la unidad de Protección Civil del propio Municipio.

Presentan una importante justificación en el hecho de que es un área de oportunidad de que el ciudadano con aplicaciones adecuadas, se entere de manera pronta y con certeza sobre contingencias relacionadas con Protección Civil, es decir en donde hay desastre para tomar las debidas precauciones y reacciones al respecto.

En el marco de referencia, presentan los temas de Protección Civil, Sistemas de Información Geográfica, Sistema de Gestores de Base de Datos, Sistemas de Informaciém Geográfica con R, Sistemas de Información Geográfica con Python, herramientas Webmapping, Ciencia de Datos y datos espaciales.

En el apartado de desarrollo del tema, presentan una propuesta de trabajo como área de oportunidad, el proceso para llevar acabo las actividades, la descripción de cómo llevarlo a cabo, las tecnologías y recursos a utilizar.

Terminan haciendo una reflexión sobre la propuesta con un análisis a título personal.

El **capítulo 11** denominado “Comparativo de herramientas para visualización de datos: Tableau y Power BI” fue desarrollado por las autoras “**Ana Georgina Soledad Núñez Martínez”, “América Herrera Domínguez**” y “**José Gabriel Rodríguez Rivas**”.

Las autoras y su coautor dan a conocer en el capítulo conceptos acerca del término Business Intelligence (BI) y su relación con las herramientas de visualización de datos, hablan acerca de las definiciones, aplicaciones, características y cuales herramientas existen en el mercado de manera generalizada,

En su marco de referencia hacen mención de distintas herramientas de visualización de datos, entre ellas PowerBI y Tableau.

Posteriormente en el apartado de desarrollo, proporcionan datos específicos sobre Tableau y Power BI, sus ventajas bondades, restricciones, proceso de instalación, entre otras cosas que van llevando al lector a una comprensión sencilla de ¿qué son? y ¿para qué? sirven estas herramientas.

Así mismo, las autoras y su coautor mencionan casos recopilados de la literatura en donde algunos ejecutivos de empresas hablan y citan sobre las bondades tanto de las herramientas Tableau como de PowerBI respectivamente.

En su propuesta, dan a conocer la importancia de que la empresa Financiera Nacional de Desarrollo Agropecuario, Rural, Forestal y Pesquero (FND) utilice herramientas como Tableau y PowerBI.

Terminan en su propuesta, haciendo un comparativo mediante el modelo de Fortalezas, Oportunidades, Debilidades y Amenazas (FODA) haciendo mención de las dos herramientas.

Las autoras y su coautor hacen una reflexión a manera de análisis de la propuesta y finalizan con las conclusiones pertinentes de su trabajo.

El **capítulo 12** denominado **“**Herramientas para análisis y visualización de datos: Tableau y R**”** fue desarrollado por los autores(as) “**Nahibe Susana Orrante Vázquez” y “Jesus Raymundo Rodríguez Díaz**” y **José Gabriel Rodríguez Rivas.**

Los autores de este capítulo ofrecen una descripción acerca de R y Tableau como herramientas para análisis y visualización de datos.

En el contenido del capítulo abordan conceptos relacionados con Ciencia de los Datos y la relación existente con herramientas para el análisis y visualización de datos.

En el apartado de desarrollo, enfatizan dos propuestas:

La primera es acerca de utilizar R y Tableau para procesar, analizar y visualizar información relacionada con Área de Vigilancia Epidemiológica del Estado de Durango; específicamente tratar datos del sistema que se encuentra en el área de Epidemiología de la Dirección General de Epidemiología (DGEPI) denominado Sistema Único Automatizado para la Vigilancia Epidemiológica (SUAVE).

En la segunda propuesta mencionan el uso de R y Tableau para enriquecer el análisis y la visualización de datos en la Residencia General de Conservación de Carreteras dependiente de la Secretaria de Comunicaciones y Transportes; propiamente en información relacionada con informes correspondientes a todo el proceso de licitación, contratación, ejecución y terminación de las obras además del informe del programa nacional de conservación de carreteras, entre otros.

Los autores del capítulo terminan con un análisis sobre las posibilidades y los beneficios de utilizar herramientas poderosas y flexibles como R y Tableau en los procesos de análisis y visualización de datos, con ello mejorar procesos de toma de decisiones y estar en un contexto integral al mundo de la Ciencia de los Datos.

El libro en su totalidad ha sido riguorosamente cuidado, auto evaluado, co-evaluado por los coordinadores con el uso de rúbrica establecida para este propósito y que se encuentra en el enlace siguiente: https://docs.google.com/forms/d/e/1FAIpQLSf4EELt-KJSCHSOmb5kAGtehoMw98wpqg2qKghX8fFkLkxbSg/viewform?usp=sf_link

El libro fue dictaminado por un Comité Cientifico Dictaminador integrado y conformado por docentes académicos e investigadores que laboran en instituciones petenecientes al Tecnológico Nacional de México (TecNM) y a la Universidad Pedagógica de Durango (UPD).

Dentro de las instituciones que pertenecen al TecNM participaron docentes del Instituto Tecnológico de El Salto (ITES), Instituto Tecnológico de Roque (ITR), Instituto Tecnológico de Tehuacán (ITT), Instituto Tecnológico de Campeche (ITC) y por supuesto del Instituto Tecnológico de Durango.

El rol y funciones que realizaron los académicos del comité de este libro, fue el de observar y dictaminar recomendaciones a las cuales se les dio seguimiento por parte de los coordinadores del libro, se realizaron los ajustes pertinentes para obtener finalmente un producto valorado, cuidado y enriquecido.

La siguiente lista de personalidades fueron los académicos investigadores que formaron parte de este comité científico, el orden de aparición obedece a una cuestión de tipo alfabético:

- Dr. Omar David Almaraz Rodríguez de la Universidad Pedagógica de Durango (UPD).
- Dr. Isidro Amaro Rodríguez del Tecnológico Nacional de México. Instituto Tecnológico de Durango.
- MES. Josefa del Carmen Hernández Ancona del Tecnológico Nacional de México. Instituto Tecnológico de Campeche
- D.C.C. Blanca Cecilia López Ramírez del Tecnológico Nacional de México. Instituto Tecnológico de Roque
- MC. Javier Nájera Frías del Tecnológico Nacional de México. Instituto Tecnológico de El Salto.
- MC. Liliana Elena Olguín Gil del Tecnológico Nacional de México. Instituto Tecnológico de Tehuacán.
- MGTI Eliceo Manuel Martín Pacheco Can del Tecnológico Nacional de México. Instituto Tecnológico de Campeche.
- MGTI Elías Gabriel Fidel Pacheco Can del Tecnológico del Nacional de México. Instituto Tecnológico de Campeche.
- Dr. Rafael Moisés Rosas Sánchez del Tecnológico Nacional de México. Instituto Tecnológico de Tehuacán

- M.S.L. Luis Ramón Sánchez Rico del Tecnológico Nacional de México. Instituto Tecnológico de Roque.
- MC. Francisco Vázquez Guzmán del Tecnológico Nacional de México. Instituto Tecnológico de Tehuacán.
- MTI. Eduardo Vázquez Zayas del Tecnológico Nacional de México. Instituto Tecnológico de Tehuacán.

La función del comité editorial fue la de observar y realizar recomendaciones en cada uno de los capítulos con la finalidad de fortalecer el producto académico.

Con respecto a los resultados de la rúbrica se obtuvieron en una escala de calificación de 1 a 5, siendo la más baja 1 y las más alta 5 se distinguen algunos datos en lo general. Determinando resultados de la encuesta de la rúbrica aplicada y con valores primordialmente entre 4 y 5 se destacan las siguientes descripciones:

Con respecto al apartado de introducción de cada capítulo, se establece que un 92% definen los aspectos generales del título, con un 92.9% deteminan la claridad de la justificación, con 98% se establece con claridad el objetivo general y con un 92% definen el propósito del del capítulo.

En el apartado de marco de referencia de cada capítulo el comité científico dictaminador describe que un 99% hay un adecuado hilo conductor de ideas; un 85% hablan de homogeneidad de ideas conforme al título; un 85% indican la formalidad de citación y referencia; un 92% coinciden en párrafos no plagiados; un 84% deteminan que hay adecuada citación de figuras

Con respecto al desarrollo establecen que un 78% los capítulos describen la empresa o unidad de trabajo a donde se dirige la propuesta, así como la relación que se tiene con los conceptos incorporados en el marco de referencia, además, coinciden que un 95% se cumplen los objetivos específicos citados en introducción, un 85% en ideas claras y con un 90% los párrafos describen con ideas claras en relación a herramientas, técnicas que se utilizarían para llevar a cabo una propuesta relacionada con Ciencia de los Datos.

Con respecto a las conclusiones se describe que un 99% son ideas claras y concluyentes, un 93% de las ideas rescatan el cumplimiento de los objetivos general y específicos y un 93% las ideas son homogéneas conforme a los objetivos y el título del capítulo.

Con todo esto, es importante reconocer el trabajo de autores y coautores de capítulos, académicos e investigadores del comité científico dictaminador, coordinadores del libro y las Instituciones involucradas, mencionado que una variable para el éxito es la perseverancia y el esfuerzo individual pero el trabajo en equipo es el ingrediente secreto.

Finalmente, con un profundo agradecimiento a todos y cada uno de ellos, resulta grato recordar una frase: "*Mi padre solía decir que nunca es demasiado tarde para hacer algo que siempre quisiste hacer. Y decía que nunca sabes lo que puedes lograr hasta que lo intentas*". Michael Jordan en "A Humbled Jordan Learns New Truths" (1994) de The New York Times.

Referencias

Pizarro, R., Amaro, I., López, J. R., & Galindo, L. (2018). Formación de científicos de datos en el Instituto Tecnológico de Durango. En N. Bocanegra Vergara, *INNOVACIÓN Y TECNOLOGÍAS DE LA INFORMACIÓN Y LA COMUNICACIÓN EN EDUCACIÓN SUPERIOR* (págs. 47-59). Durango Dgo.: REDIE.

Capítulo 1

Bases de datos SQL y NoSQL. Comparativo SQL server & MongoDB

Juan Daniel Carrillo Mercado
Tecnológico Nacional de México. Instituto Tecnológico de Durango
03040123@itdurango.edu.mx

Gabriel Arturo Lugo Morales
Tecnológico Nacional de México. Instituto Tecnológico de Durango
alugo@itdurango.edu.mx

1.1. Introducción

Las bases de datos han sido importantes para el almacenamiento y análisis de datos. A través de los años han surgido herramientas que han ayudado a poder realizar esta tarea de manera satisfactoria.

En los últimos años ha crecido la necesidad de almacenar y analizar grandes volúmenes de información, por lo que los sistemas de bases de datos relacionales se han enfrentado a un gran reto.

En este documento presenta cómo los sistemas de gestión de bases de datos han evolucionado para poder satisfacer el gran crecimiento de almacenamiento y tratamiento de los datos. Se mencionará el papel que han tenido las bases de datos

relacionales, se mencionará el surgimiento de un nuevo paradigma para almacenar y tratar los datos, como lo son las bases de datos NoSQL.

El objetivo general de este capítulo es realizar un comparativo de las bases de datos SQL con respecto a las bases de datos NoSQL listando las principales características, ventajas, así el análisis de gestores respectivos.

De manera específica se pretende lo siguiente:

- Mostrar la evolución de los sistemas de gestión de bases de datos relacionales o SQL, describiendo sus funcionalidades.
- Describir las características principales del modelado de datos de los sistemas relacionales, así como su principal lenguaje de consulta SQL.
- Explicar la necesidad del surgimiento de las bases de datos NoSQL, sus características y los tipos de bases de datos NoSQL.
- Mostrar algunos de los principales sistemas de gestión de bases de datos.
- Mostar algunas de las características de SQL Server como uno de los sistemas de gestión de bases de datos relacionales más populares en el mercado.
- Mostrar como MongoDB está teniendo una gran aceptación entre las bases de datos NoSQL, así como el lenguaje de consulta que utiliza para gestionar los datos.

Dentro de los alcances de este apartado se pretende mostrar un panorama general de las bases de datos relacionales y como éstas han desempeñado un papel importante en el almacenamiento y trato de los datos, a la vez como van cobrando fuerza las bases de datos NoSQL; visualizar las ventajas y desventajas de las bases de datos SQL y las NoSQL; tener una visión general de cuáles son las similitudes en sus lenguajes de consulta entre SQL Server y MongoDB y cuáles son sus principales diferencias como sistemas de gestión de bases de datos.

El capítulo da a conocer cómo fueron surgiendo las bases de datos NoSQL como respuesta a nuevas necesidades en el tratamiento y procesamiento de los datos que las bases de datos relacionales no han podido satisfacer. De igual forma, menciona las características que son parte de las bases de datos NoSQL.

Una justificación para la elaboración de este capítulo es dar a conocer a la comunidad las nuevas herramientas de sistemas de gestión de bases de datos que están haciendo frente a las necesidades de la gestión y almacenamiento de grandes volúmenes de datos y así poder tomar una decisión de cual elegir de acuerdo con las necesidades de los sistemas a desarrollar.

1.2. Marco de referencia

Actualmente ha aumentado la demanda para almacenar, consultar y analizar datos. Este fenómeno se ha presentado en las empresas, organizaciones y en particulares. Y ahora con el uso del internet y los dispositivos móviles, la demanda de datos se ha disparado exponencialmente.

Ahora bien, surgen algunos cuestionamientos como los siguientes:

- ¿Qué aspectos habrá que abordar sobre las bases de datos?,
- ¿Cómo surge la necesidad de estructurar los datos en forma de bases de datos?,
- ¿Cuáles son las aplicaciones de las bases de datos?,
- ¿Qué sistemas de datos se utilizaban antes de las bases de datos relacionales?,
- ¿Cómo fueron surgiendo los sistemas de gestión de bases de datos?, ¿Qué características tienen los sistemas de gestión de bases de datos?,
- ¿Cuáles son tendencias y ventajas de los sistemas de gestión de bases de datos?,
- ¿Cómo surgen los sistemas de bases de datos NoSQL?

El origen de las bases de datos fue evolutivo, debido a necesidades de almacenamiento y procesamiento, los sistemas de bases de datos tuvieron que requerir mayor capacidad y funcionalidades, así se da el surgimiento de herramientas para interactuar con los datos y la información.

1.2.1. Bases de datos

Una base de datos en un sentido simple es un conjunto de información organizada y se requieren herramientas para su tratamiento y uso.

Las bases de datos son un elemento indispensable de las buenas prácticas en el desarrollo de software empresarial. Una base de datos es lo que le dará sentido al desarrollador para conectar los programas con los datos bajo un esquema de optimización y mejor rendimiento en las aplicaciones.

Debido al crecimiento de las tecnologías de información, sumado al uso de las tecnologías móviles con acceso a internet, el volumen de datos que manejan tanto las organizaciones como los particulares se ha incrementado notablemente, hoy en día es difícil de imaginar una aplicación que no utilice una base de datos.

Arias (2015) menciona, “las bases de datos son colecciones de información (datos) que se relacionan para crear un sentido y dar más eficiencia a una encuesta, un estudio organizado, o la estructura de datos de una empresa, son de vital importancia para las organizaciones y en las últimas décadas son parte principal de los sistemas de información” (pág. 13).

En las primeras décadas del siglo XX, el uso de las bases de datos fue creciendo en las empresas. En esos años solo algunas personas interactuaban directamente con los sistemas de bases de datos, otros sin darse cuenta manejaban datos en forma de base de datos como: informes impresos, expedientes, registro de transacciones, entre otros (Silberschatz, Korth, & Sudarshan, Fundamentos de Bases de Datos. Cuarta Edición, 2002).

Actualmente casi todas las empresas u organizaciones usan las bases de datos: en agencias de viajes en el sector turismo; escuelas y el giro educación; en la industria en el ámbito de producción, inventarios y calidad; en el comercio y tiendas departamentales y tiendas autoservicio; aeropuertos; bancos; áreas de recursos humanos y financieros en las empresas; gobierno y sus instituciones; ventas y marketing; medicina y sector salud; y en todas aquellas aplicaciones que tenga que ver con servicios web, móviles y sus transacciones; entre muchos otros.

1.2.2. Sistemas de gestión de bases de datos

Silberschatz, Korth, & Sudarshan (2020), establecen que “un sistema de bases de datos es una colección de archivos interrelacionados y un conjunto de programas que permitan a los usuarios acceder y modificar estos archivos” (pág. 3).

Un sistema de gestor de base de datos también nombrado sistema manejador de bases de datos (SMBD) es una herramienta computacional que permite a los usuarios establecer esquemas de definición de la base de datos, utilizar un lenguaje de manipulación para administrar, programar y realizar consultas y modificaciones sobre los datos (Silberschatz, Korth, & Sudarshan, Fundamentos de Bases de Datos. Cuarta Edición, 2002).

Las bases de datos contienen información relevante para la empresa, organización o particular. La función principal de los sistemas de gestión de bases de datos es almacenar y recuperar la información de una base de datos de manera rápida, práctica y eficiente.

Los sistemas de gestión de bases de datos están desarrollados para manejar grandes cantidades de datos. Además, debe de proporcionar la fiabilidad de la información almacenada, a pesar de las interrupciones en el sistema.

Los primeros sistemas de gestión de bases de datos se realizaron para ejecutar procesos administrativos y así reducir el papeleo, tener un mayor control sobre los datos y facilitar su consulta (Silberschatz, Korth, & Sudarshan, Fundamentos de Bases de Datos. Cuarta Edición, 2002).

En los inicios de los sistemas de gestión de bases de datos estaban ligados y dependientes a los sistemas operativos, así como al hardware. Los sistemas de gestión de bases de datos estaban diseñados para para facilitar el manejo de grandes volúmenes de datos con operaciones complejas.

Cuando el programador realizaba una aplicación, este tenía que conocer detalles del diseño físico de la base de datos, por lo que la programación de las aplicaciones se convertía en algo complicado.

En su artículo Aguilar Romero & Rodríguez García, (2016), mencionan que un sistema gestor de bases de datos pueden verse como una capa intermedia que integra el lenguaje de definición de datos (DDL) y el lenguaje de manipulación de datos (DML) para facilitar la manipulación de tablas, registros y realizar consultas y operaciones generalmente basadas en un estándar de un lenguaje para ejecución de consultas (SQL) (pág. 9), además, los sistemas de gestión de bases de datos proporcionan las siguientes tareas:

- Definición y creación de las bases de datos mediante lenguaje de definición de datos (DDL)
- Manipulación de los datos realizando consultas, inserciones, eliminaciones y actualizaciones con lenguaje de manipulación de datos (DML)
- Acceso controlado a los datos mediante mecanismos de seguridad de acceso a los usuarios. Esto implica que los usuarios que no tengan autorización a la base de datos no puedan tener acceso a ella.
- Mantener la integridad y consistencia de los datos.
- Controlar la concurrencia a la base de datos para compartición de recursos.
- Mecanismos de copias de respaldo y recuperación para restablecer la información en caso de fallos de sistema. (Aguilar Romero & Rodríguez García, 2016).

El sistema de gestión de bases de datos administra tanto los datos físicos y su almacenamiento, lo que lo convierte en una herramienta muy útil. Conscientes de esta situación, los sistemas de gestión de bases de datos utilizan una herramienta de vistas que facilitan a que cada usuario tenga su propia visión de la base de datos.

No todos los sistemas de gestión de bases tienen las mismas funcionalidades, estas dependen según el producto. Pero en general, los grandes sistemas de gestión de bases de datos proporcionan las funcionalidades que se citaron e incluso más. Pretendiendo así ofrecer un sistema de gestión de bases de datos que controle cualquier tipo de requisitos y que sea confiable a cualquier fallo.

Los sistemas de gestión de bases de datos evolucionan cada vez más rápido para satisfacer los requisitos de los usuarios. El gran aumento de aplicaciones, web, escritorio, móviles, aparatos electrónicos, hoy en día es necesario almacenar grandes cantidades de datos, imágenes, videos, sonido, entre otros. Respecto a esta gran demanda los sistemas de gestión de bases de datos nunca permanecerán estáticos.

Los sistemas de gestión de bases de datos se han adaptado a las tecnologías recientes: la multimedia, la de orientación a objetos e internet.

Los sistemas de gestión de bases de datos incorporan la posibilidad de tipos abstractos de datos. El éxito del desarrollo de software orientado a objetos ha hecho que esta práctica se extienda a todos los ámbitos de la informática, por lo que ha requerido que los sistemas de gestión de bases de datos relacionales incluyan el paradigma orientado a objetos que integra aspectos tales como objeto, identidad de objeto, objetos compuestos, métodos, encapsulación y herencia, entre otros. (Meza Quiñones, 2018).

El rápido crecimiento de sistemas de información por medio de la web hace que los sistemas de gestión de bases de datos agreguen recursos en su servidor web para que se puedan incluir instrucciones SQL y mostrar la información de las bases de datos a los usuarios.

Los sistemas de gestión de bases de datos tienen características y funcionalidades diversas, entre las que se destacan: optimización en la redundancia de datos; consistencia y mismo significado de los datos; compartición de datos por distintos usuarios; posibilidad de mantenimiento; integridad de los datos mediante reglas y restricciones; seguridad y accesibilidad, replicaciones; productividad para no preocupares por aspectos de bajo nivel; respaldos y restaurado; concurrencia para accesos al mismo tiempo y en paralelo de distintos usuarios y de manera remota; entre otros.

1.2.3. Big Data, Inteligencia de Negocios, Ciencia de los Datos y Bases de datos

Ahora bien, la pregunta es ¿qué relación existe entre las Bases de Datos con los conceptos de Inteligencia de Negocios, Big Data y Ciencia de los Datos?

Actualmente los sistemas de gestión de bases de datos se han adaptado a este tipo de almacenes de datos incorporando herramientas tales como: creación y mantenimiento de réplicas, diferentes orígenes de datos, estructuras físicas que permiten el análisis multidimensional, diversos tipos de datos, entre otros.

En vista de estos grandes volúmenes de datos y adaptándose a las necesidades de los usuarios, han surgido los sistemas de gestión de bases de datos no relacionales, los cuales ha ido creciendo su implementación y se han visto como una opción a escoger de acuerdo a las necesidades las empresas, organizaciones y particulares.

En los últimos años se ha tenido el auge del concepto de Data Warehouse (DW) o almacén de datos, entendiéndose como un repositorio de datos que contiene los datos históricos de la organización, extraídos de las bases de datos operacionales, archivos de trabajo y cualquier otra fuente de información importante para la toma de decisiones. (Formia & Estevez, 2019). Esto aumenta definitivamente el volumen de información.

A lo largo de los años las empresas, organizaciones y particulares han ido acumulando grandes cantidades de datos que han ido almacenando en Data Warehouse, por lo que ha requerido que los sistemas de gestión de bases de datos administren de manera confiable los datos para poder realizar consultas y análisis.

Como lo menciona Jones (2019), en su libro intitulado: La guía definitiva de análisis de Big Data para empresas, técnicas de minería de datos, recopilación de datos y conceptos de inteligencia empresarial, la inteligencia de negocios o empresarial “es un término amplio que cubre aspectos como el análisis de procesos, extracción de datos, análisis descriptivo y evaluación comparativa del rendimiento”. (pág. 192).

Con lo anterior, se hace imprescindible el uso del recurso de los datos que se encuentra inmerso y de manera natural en las bases de datos.

Las bases de datos, representan la información organizada de la empresa, la tecnología Big Data requiere de manera imprescindible tener acceso de las bases de datos internas y externas de la empresa y representa un modelo evolutivo que hace referencia al enorme volumen de datos de formato heterogéneo (estructurados, semiestructurados, y no estructurados) que se generan continuamente y cuyo análisis mediante las tecnologías habituales y tradicionales de los sistemas manejadores de bases de datos se ve dificultado (Hueso, y otros, 2019).

Como lo expresa Hueso, y otros (2019) “En la actualidad, el término Big Data se ha expandido para incorporar el análisis e interpretación de los datos, constituyendo una Ciencia de los Datos (data science)”. (pág. 2)

Jones (2019) menciona que “los datos modernos no están estructurados y son diferentes de los datos tradicionales. Por lo tanto, necesita tener métodos avanzados de análisis de datos”. (págs. 49-50). La figura 1, indica el tipo de datos requerido para el año 2020, más del 80% de los datos no están estructurados.

Se necesitan manejadores de bases de datos con mayores funcionalidades o con adaptaciones e integración con otras herramientas para obtener mayor potencia y rendimiento en el procesamiento del volumen de datos, variedad, variabilidad, a grandes velocidades de recuperación, con veracidad y valor, así como para poderlos visualizarlos de distintas formas.

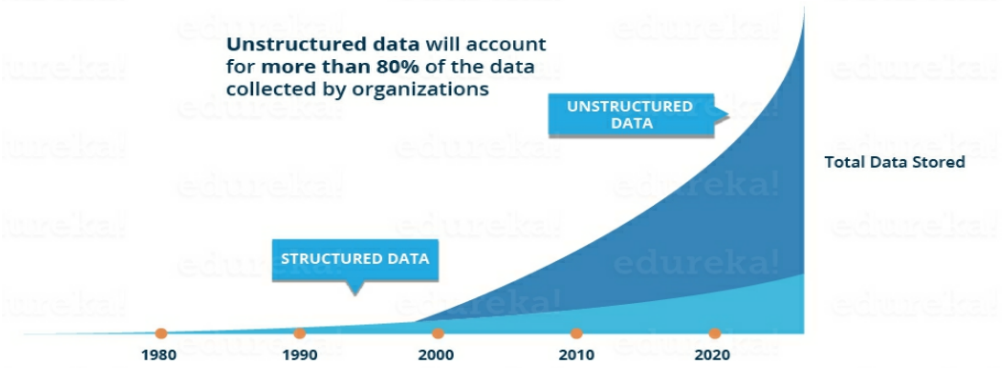


Figura 1. Datos estructurados y No estructurados para el año 2020. (Jones, Ciencia de los Datos. Lo que saben los mejores científicos de datos sobre el análisis de datos, minería de datos, estadísticas, aprendizaje automático y Big Data que usted desconoce, 2019)

1.2.4. Principales sistemas de gestión de bases de datos en la actualidad.

Actualmente existe una gran cantidad de sistemas de gestión de bases de datos, estos son usados de acuerdo a las necesidades y solvencia económica de las empresas, organizaciones o particulares. Se muestra en la figura 2, los diez principales sistemas de gestión de bases de datos

Rank			DBMS	Database Model	Score		
Jul 2020	Jun 2020	Jul 2019			Jul 2020	Jun 2020	Jul 2019
1.	1.	1.	Oracle	Relational, Multi-model	1340.26	-3.33	+19.00
2.	2.	2.	MySQL	Relational, Multi-model	1268.51	-9.38	+38.99
3.	3.	3.	Microsoft SQL Server	Relational, Multi-model	1059.72	-7.59	-31.11
4.	4.	4.	PostgreSQL	Relational, Multi-model	527.00	+4.02	+43.73
5.	5.	5.	MongoDB	Document, Multi-model	443.48	+6.40	+33.55
6.	6.	6.	IBM Db2	Relational, Multi-model	163.17	+1.36	-10.97
7.	7.	7.	Elasticsearch	Search engine, Multi-model	151.59	+1.90	+2.77
8.	8.	8.	Redis	Key-value, Multi-model	150.05	+4.40	+5.78
9.	9.	↑ 11.	SQLite	Relational	127.45	+2.64	+2.82
10.	10.	10.	Cassandra	Wide column	121.09	+2.08	-5.91

Figura 2. Principales sistemas de gestión de bases de datos. (DB-Engines, 2020)

1.2.5. Bases de datos relacionales

Las bases de datos relacionales son unas de las más utilizadas por las empresas, organizaciones y particulares para administrar sus datos, ya que estos en su mayoría tienen una estructura que cumple con las especificaciones del modelo de datos relacionales.

Una de las principales características que presentan las bases de datos relacionales actualmente es que manejan como un estándar el Lenguaje Estructurado de Consulta (SQL).

Una base de datos relacional es un conjunto de tablas estructuradas en registros y que se relacionan entre sí mediante claves.

El modelo de datos relacional continúa siendo hoy en día la forma más común de estructurar y almacenar la información en cualquier tipo de sistema de información.

Las bases de datos relacionales son aquellas que se apegan y cumplen con el modelo relacional, su estructura principal la componen tablas, atributos y relaciones que representan información (Osorio Rivera, 2008).

Los componentes básicos de una base de datos relacional son:

- Tablas. Son estructuras regulares conformadas por renglones y columnas Las tablas no pueden encontrarse duplicadas dentro de una misma base de datos.
- Columnas. Es una división lógica de la tabla, son los atributos que forman la estructura y caracterizan a cada tabla.
- Registros. También conocidas como filas, es decir los registros son los valores divididos en columnas de una fila.
- Relaciones. Son asociaciones entre tablas mediante valores comunes en ellas. Dichos valores se conocen como llaves o claves. Las claves pueden ser primarias o foráneas Clave primaria.

- Clave Primaria. Cada registro pose una clave única, estas claves identifican de manera única a un registro. Este registro en el valor de la clave primaria no puede repetirse en la tabla.
- Clave foránea. Estas claves son referencias colocadas en tablas secundarias de la relación de tablas. Estas contienen el mismo valor que la clave primaria del registro de la tabla principal (Osorio Rivera, 2008).

Otra de las características de las bases de datos relacionales es que es posible aplicar principios de álgebra relacional y de cálculo relacional para manipular los datos contenidos en la base de datos.

La integridad y seguridad de los datos son importante en las bases de datos relacional porque permiten proteger los datos de los daños provocados por los usuarios ya sea de manera intencional o no, así como algún fallo en el sistema.

Como parte de la integridad de los datos las bases de datos relacionales deben de asegurar que alguna actualización de los datos no viole las restricciones de integridad que se especifican en una base de datos.

1.2.6. Leguaje SQL

Una de las principales características de los sistemas de gestión de bases de datos relacional es contar con un leguaje de consulta que sea fácil y mas amigable con el usuario. (Osorio Rivera, 2008).

El principal lenguaje de consulta utilizado por los sistemas de gestión de bases de datos relacionales es el Lenguaje Estructurado de Consulta (SQL).

El lenguaje SQL está basado en el álgebra relacional (Silberschatz, Korth, & Sudarshan, Fundamentos de Bases de Datos. Cuarta Edición, 2002), no solo se pueden realizar consultas sino también modificación de los datos y establecer restricciones de integridad en estos.

1.2.6.1. Componentes del lenguaje SQL

El lenguaje SQL está compuesto por varios partes:

- Lenguaje de definición de datos (DDL). Mediante este lenguaje se proporciona órdenes para crea, modificar o eliminar un esquema de relaciones, crear, modificar o eliminar tablas, crear, modificar o eliminar índices.
- Lenguaje de manipulación de datos (DML). Permite realizar consultas a los datos utilizando el álgebra relacional sobre los registros. Incluye instrucciones para insertar, borrar y modificar registros en una base de datos.
- Definición de vistas. Mediante el lenguaje de definición de datos se pueden crear vistas de los datos
- Control de transacciones. Por medio de instrucciones SQL se puede especificar el comienzo y el final de una transacción.
- SQL incorporado y SQL dinámico. Se pueden definir instrucciones SQL en lenguajes de programación como C++, C#, Java, PHP, entre otros.
- Integridad. Mediante el lenguaje de definición de datos se puede especificar las restricciones de integridad que deben de contener los datos almacenados en la base de datos.
- Autorización. Incluye órdenes para dar acceso a los datos. (Silberschatz, Korth, & Sudarshan, Fundamentos de Bases de Datos. Cuarta Edición, 2002).

1.2.6.2. Estructura básica de una consulta SQL

Una instrucción de consulta SQL en su forma básica contiene tres clausulas: *select*, *from*, *where*. El conjunto de estas instrucciones son de gran utilidad a la hora de la extracción de datos de una base de datos relacional. (Osorio Rivera, 2008).

- *select*. Se utiliza para listar o proyectar los campos que se requieren de una consulta.
- *from*. Muestra las relaciones o tablas que se debe de analizar en una instrucción.
- *where*. Corresponde a la condición de álgebra relacional para mostrar o filtrar registros.

Una instrucción simple en SQL tiene la siguiente forma:

Select Atributo1, Atributo2,…AtributoN

From Relacion1, Relacion2, …RelacionN

Where Condición de **JOIN** y fitrado

1.2.6.2.1. Cláusula order by

SQL ofrece un control sobre el orden de los registros. La cláusula *order by* hace que los registros obtenidos de una consulta sean presentados en un orden específico. De manera predeterminada la cláusula *order by* muestra los registros de manera ascendente. Esta cláusula permite el que se pueda especificar que los datos se muestren de manera descendente solo con agregarle la condición desc. Ejemplo:

```
select nombre, sueldo, sueldo-(sueldo*iva) /100

From Empleado

Order by Sueldo desc
```

1.2.6.3. Operaciones sobre conjuntos

El lenguaje SQL ofrece varias operaciones sobre conjuntos, las más comunes son: *union*, *intersect* y *except* (Silberschatz, Korth, & Sudarshan, Fundamentos de Bases de Datos. Cuarta Edición, 2002). Las figuras 3 al 5 interactúan sobre relaciones y corresponden a operaciones del álgebra relacional de unión "∪", intersección "∩" y diferencia "-".

- **Union.** Devuelve la suma de dos o más conjuntos de datos.

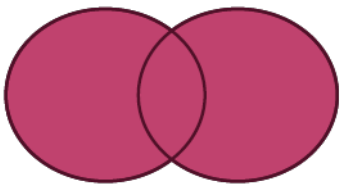


Figura 3. Representación de una unión. Elaboración propia

- **Intersec**. Devuelve la intersección de dos o más conjuntos de datos.

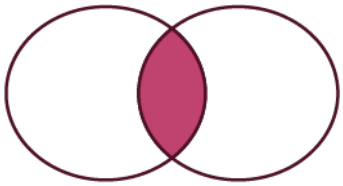


Figura 4. Representación de un *intersec*. Elaboración propia

- **Except.** Devuelve la diferencia de dos o más conjuntos de datos.

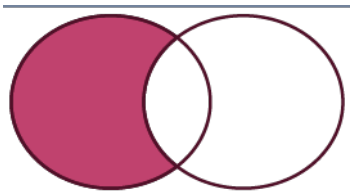


Figura 5. Representación de un *except*. Elaboración propia

1.2.6.4. Cláusulas para modificar datos

El lenguaje SQL también cuenta con instrucciones para insertar, modificar o eliminar datos.

1.2.6.4.1. Cláusula insert

Esta instrucción es utilizada para agregar registros a una tabla contenida en una base de datos. Los valores que se insertan deben de tener la misma estructura de los atributos de la tabla. Un ejemplo básico de esta instrucción sería.

```
insert into Personal values (1234, 'Pérez', 'Juan')
```

1.2.6.4.2. Cláusula update

Esta instrucción se utiliza para actualizar los valores de los registros contenidos en una tabla de una base de datos. La cláusula **where** de la instrucción **update** puede contener cualquier constructor legar en la cláusula **where** de una instrucción **select**. (Silberschatz, Korth, & Sudarshan, Fundamentos de Bases de Datos. Cuarta Edición, 2002). Un ejemplo básico es el siguiente:

```
update Personal set nombre = 'José'

where id=1234
```

1.2.6.4.3. Clausula delete

La instrucción *delete* permite borrar registros de una tabla contenida en una base de datos un borrado mediante SQL, se expresa de la siguiente manera:

```
delete from Tabla
where condición
```

Un ejemplo aplicando esta condición sería:

```
delete from Personal
where id=1234
```

1.2.6.5. Actualidad en bases de datos relacionales

Las bases de datos relacionales seguirán siendo el fuerte de las empresas, organizaciones y particulares para los próximos años, ya que los sistemas de gestión de bases de datos relacionales continúan evolucionando y adaptándose a las necesidades de los clientes que manejan grandes volúmenes de datos estructurados siguiendo el modelo de datos entidad relación.

Las bases de datos relacionales SQL son y seguirán siendo una parte importante en el manejo de los datos estructurados, pero debido a lo poco flexible en lo que respecta a la modificación de las estructuras de los datos, han salido al mercado sistemas de gestión de bases de datos con una mayor flexibilidad, como lo son las bases de datos NoSQL (Baldassari Valencia, 2019).

1.2.7. Bases de datos NoSQL

El término NoSQL significa *not only SQL* (No solo SQL). El movimiento Big Data es uno de los grandes impulsores de la tecnología NoSQL.

Como lo define Baldassari Valencia, (2019), “NoSQL puede ser definido como “*Not only SQL*” y es un sistema de gestión de bases de datos la cual podría ser la siguiente generación de tecnologías que es no relacional, distribuida, escalable horizontalmente, de código abierto y más rápida, puesto que no implementa las propiedades ACID las cuales aseguran la confiabilidad de las transacciones sobre las bases de datos.”. (pág. 3)

Empresas como Google, Amazon, Facebook, twitter, son algunas de las que encontraron limitaciones en los sistemas de gestión de bases de datos relacionales, así que iniciaron el desarrollo de Sistemas de gestión de bases de datos no relacionales.

Debido a la gran cantidad de datos que se genera actualmente en la nube, es necesario que los nuevos sistemas de gestión de bases de datos contengan disponibilidad total, tolerancia a fallos, grandes almacenamientos de penta bytes de datos que se encuentren distribuidos en miles de servidores, que contengan un excelente desempeño y que se pueda tener una escalabilidad horizontal de manera fácil, que puedan manejar grandes cantidades de datos estructurados y no estructurados, nuevos métodos de consulta, libertad para nuevos esquemas y tener un control total sobre los datos (Bender, Deco, González Sanabria, Hallo, & Ponce, 2014).

Aplicaciones como Facebook y Twitter no encontraban una solución en los sistemas de gestión de bases de datos relacionales, debido a los millones de transacciones que se manejaban por medio del lenguaje SQL los servidores quedan expuestos a saturación y disminución de rendimiento.

Esta problemática requirió de soluciones que soporten grandes cantidades de transacciones y esto se podía solucionar de dos maneras: seguir utilizando los sistemas de gestión de bases de datos relacionales, pero con máquinas más potentes, con mayor memoria, procesadores de última generación, discos de mayor capacidad, entre otros, (escalabilidad vertical), o desarrollando sistemas de gestión de bases de datos utilizando máquinas configuradas en clúster (escalabilidad horizontal).

Debido a esta gran necesidad, surgen las bases de datos altamente escalables, y ya que las bases de datos relacionales son complicadas de escalar horizontalmente, se empezaron a utilizar las bases de datos no relacionales o NoSQL. (Baldassari Valencia, 2019).

Las bases de datos no relacionales proporcionan un rendimiento en lectura y escritura de datos que los sistemas de gestión de bases de datos relacionales no pueden lograr, dadas estas diferentes necesidades y retos técnicos con los que se fueron encontrando los sitios web más utilizados, sobre todo en cuanto a disponibilidad y escalabilidad, es donde las bases de datos NoSQL tuvieron un gran impacto (García Cebreiros, 2016).

Es por eso que hablar de bases de datos NoSQL, es hablar de estructuras que permiten tratar datos en a las que las bases de datos relacionales generan problemas, como es el rendimiento cuando se generan millones de transacciones diarias. Además, las bases de datos NoSQL son sistemas de almacenamiento que no cumple con el esquema entidad relación, tampoco tiene una estructura en sus datos en forma de tabla si no que tiene otros tipos de estructurar la información.

1.2.7.1. Justificación de las de las bases de datos NoSQL

Los principales motivos con los que se cuenta para comenzar a valorar la posibilidad de dejar atrás las bases de datos relacionales y comenzar a fijarse en las bases de datos NoSQL son los siguientes:

- Evitar la complejidad innecesaria. Las bases de datos tradicionales tienen gran cantidad de funcionalidades y restricciones para mantener la consistencia de los datos, en ciertos casos, mucho más de lo necesario. Esto hace que, a nivel global, las operaciones en la base de datos tarden más tiempo lo cual el rendimiento se ve afectado.
- Alto rendimiento. Muchas bases de datos NoSQL proporcionan un rendimiento superior al que ofrecen las bases de datos relacionales.
- Escalabilidad horizontal y hardware de bajo costo. Al contrario que las bases de datos relacionales, los sistemas NoSQL han sido diseñados para escalar horizontalmente (Baldassari Valencia, 2019).
- Complejidad y costo de levantar un clúster de base de datos. Facilidad y sencillez con la que estos sistemas son capaces de añadir y quitar nodos del sistema.

- Comprometer la fiabilidad a cambio del rendimiento. La fiabilidad en los datos es un tema muy importante, pero existen ciertos momentos sin despreciar la confianza en los que se puede demandar un aumento del rendimiento a cambio de un nivel menos de fiabilidad.
- Movimientos en lenguajes de programación y Frameworks de desarrollo. Dejar las capas de acceso a bases de datos independientes del resto del código. Esto, que ya de por si es una buena práctica para bases de datos relacionales, adquiere un importante significado para el movimiento NoSQL, que se ha dado prisa en desarrollar los conectores para sus bases de datos para los distintos lenguajes de programación.
- Requisitos de Cloud Computing**.** La alta escalabilidad, especialmente horizontal, y que los tiempos de administración sean lo mínimo posible.

1.2.7.2. Tipos de bases de datos NoSQL

Hay varios tipos de bases de datos NoSQL, cada una con sus peculiaridades y ventajas. Entre las principales se encuentran: basadas en clave-valor, basadas en columnas, basadas en documentos y basadas en grafos.

Se identifican diversos tipos de bases de datos NoSql encontradas en el documento de Gracia del Busto & Yanes Enríquez, (2012) y en acenswhitepapers, (2014).

1.2.7.2.1. Basadas en clave-valor

Este tipo de base de datos consiste en un mapa o diccionario en el cual se pueden almacenar y obtener valores por medio de una clave; favorece la escalabilidad y omite las consultas complejas y permite que la recuperación de la información sea rápida.

Los sistemas de clave-valor prometen un rendimiento excelente para volúmenes de datos muy grandes, son las más simples de entender ya que guardan tuplas de orden clave-valor (Baldassari Valencia, 2019).

En un sistema relacional existen bases de datos y dentro de cada base de datos se tienen tablas formadas por filas y columnas. En un sistema clave-valor

existen contenedores, también se les llama *cabinets*, en cada contenedor se puede tener tantas parejas de clave-valor como se desee. En cada contenedor es posible tener datos de la misma naturaleza o totalmente diferentes, todo depende de los desarrolladores de la aplicación.

1.2.7.2.2. Basadas en columnas

Este tipo de base de datos almacenan la información por columnas, cada clave única apunta a un conjunto de subclaves, que son tratadas como columnas. Se utilizan con más frecuencia para la lectura de grandes volúmenes de datos que para la escritura. Añade cierta flexibilidad, ya que permite añadir columnas a las filas necesarias sin alterar el esquema completo. Este tipo de bases de datos también es conocido también como BigTable.

Este tipo de base de datos son en realidad lo que se podría suponer, tablas de datos donde las columnas de valores de datos representan el almacenamiento estructural. Los datos son almacenados como secciones de las columnas de datos en lugar de filas de datos, como en la mayoría de los gestores relacionales. Esto tiene ventajas para los almacenes de datos, sistemas de gestión de relaciones con clientes, catálogos de bibliotecas de tarjetas y otros sistemas de consulta donde los agregados se calculan a través de un gran número de elementos de datos similares.

1.2.7.2.3. Basadas en documentos

En este tipo de bases de datos los datos se almacenan en forma de documentos de tipo XML o JSON, encapsulando pares de clave-valor en documentos y utilizando etiquetas para los valores de las claves. Permite estructuras de datos complejas y realizar consultas muy potentes. Este tipo de base de datos proporcionan una gran versatilidad y se ha convertido en el favorito para muchos desarrolladores. (Baldassari Valencia, 2019).

Las bases de datos de documentales son consideradas por muchos como un escalón superior ante los simples gestores de llave-valor, puesto que permiten encapsular pares de llave-valor en estructuras más complejas denominadas documentos.

1.2.7.2.4. Basadas en grafos

Este tipo de base de datos representan los datos como nodos de un grafo y sus relaciones con las aristas del mismo grafo, en este tipo de base de datos se puede usar la teoría de grafos para recorrer la base de datos ya que esta puede describir los atributos de los nodos y las aristas.

Este tipo de base de datos debe de estar normalizado, esto es que cada tabla tendría una sola columna y cada relación dos, esto es así para que cualquier cambio en la estructura en la información tenga solo efecto solo en la estructura de la información de manera local (Gracia del Busto & Yanes Enríquez, 2012).

1.2.8. Diferencias con las bases de datos SQL

Existen diferencias principales que se pueden encontrar entre las bases de datos NoSQL y las SQL. (acenswhitepapers, 2014).

- La mayoría de las bases de datos NoSQL no utilizan el lenguaje SQL o solo lo utilizan como un lenguaje de apoyo.
- No utilizan estructuras fijas como tablas para el almacenamiento de los datos. Las bases de datos NoSQL permiten hacer uso de otros tipos de modelos de almacenamiento.
- No permite operaciones JOIN. Al tener volúmenes de datos tan extremadamente grandes se evitan las operaciones JOIN.
- Arquitectura distribuida. Los datos de las bases de datos NoSQL pueden encontrarse distribuidos en varias máquinas (acenswhitepapers, 2014).

1.2.9. Principales sistemas de gestión de bases de datos NoSQL

A continuación, se mencionan algunos sistemas de gestión de bases de datos NoSQL, así como el tipo de base de datos que maneja.

- Cassandra**.** Es una base de datos de tipo clave-valor creada por Apache. Utiliza un lenguaje de consulta CQL (Cassandra Query Language). Puede

correr en cualquier plataforma que cuente con Java (Apache Cassandra, 2016).

- HBase**.** Apache HBase es un almacén de big data distribuido y escalable de forma masiva del ecosistema de Apache Hadoop. Apache HBase es una base de datos abierta, distribuida, versionada y no relacional, modelada según el Bigtable de Google (APACHE HBASE, 2020).
- MongoDB**.** Es una base de datos basada en documentos con la escalabilidad y flexibilidad. Almacena datos en documentos flexibles, similares a JSON, lo que significa que los campos pueden variar de un documento a otro y la estructura de datos se puede cambiar con el tiempo. Es una base de datos distribuida en su núcleo, por lo que la alta disponibilidad, la escala horizontal y la distribución geográfica están integradas y son fáciles de usar (MongoDB, 2020).
- Neo4j. Es una base de datos basada en grafos. Esta base de datos está desarrollada en Java. Neo4j tienen mejor rendimiento que las bases de datos relacionales y las NoSQL. La clave es que, aunque las consultas de datos aumenten exponencialmente, el rendimiento de Neo4j no desciende, frente a lo que sí sucede con las bases de datos relacionales (neo4j, 2020).

1.2.10. Clasificación de bases de datos de acuerdo al teorema CAP

Las bases de datos relacionales están ampliamente extendidas y cuentan con grandes y potentes herramientas su manejo. Existe, además, una gran cantidad de documentación en línea y una gran comunidad de usuarios. Numerosas aplicaciones salen al mercado pensadas para integrarse fácilmente con este tipo de bases de datos. A diferencia, las bases de datos NoSQL son relativamente nuevas y disponen de pocas herramientas de gestión que no ofrecen muchas de las opciones que son encontradas en las herramientas para la gestión de las bases de datos relacionales (Gracia del Busto & Yanes Enríquez, 2012)

Es por esta razón, que al momento de elegir una base de datos NoSQL se tenga en cuenta el teorema CAP: Consistencia (C), Disponibilidad (A) y Tolerancia

a particiones (P). Según Eric Brewer, creador de este teorema, plantea que en los sistemas distribuidos solo se puede tener dos de las tres garantías (la C, la A o la P), y por lo tanto es preciso elegir la más importante. La figura 6, representa la clasificación de las bases de datos conforme al teorema CAP.

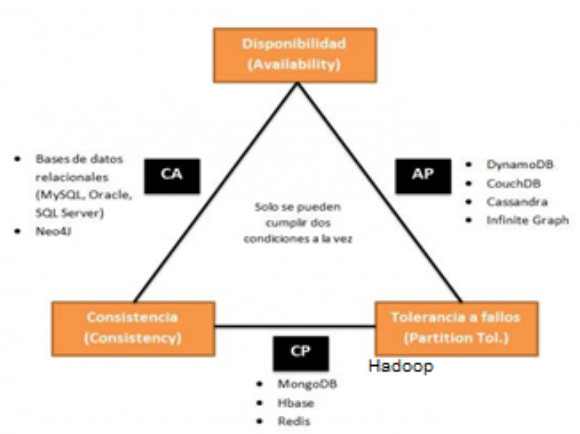


Figura 6. Clasificación de las bases de datos según el teorema CAP (GENBETA, 2014).

- Consistencia: Al realizar una consulta o inserción siempre se tiene que recibir la misma información, con independencia del nodo o servidor que procese la petición.
- Disponibilidad: Que todos los clientes puedan leer y escribir, aunque se haya caído uno de los nodos.
- Tolerancia a particiones: Implica, que el sistema tiene que seguir funcionando, aunque existan fallos o caídas parciales que dividan el sistema (GENBETA, 2014).

1.3. Desarrollo

En este apartado, se muestran las características de los sistemas manejadores de base de datos SQL Server y de MongoDB.

Al final del tema, se muestra un análisis FODA con las apreciaciones y comparaciones personales del autor sobre ambos manejadores.

1.3.1. SQL Server

SQL Server es un sistema de base de datos relacional de Microsoft, desarrollado para las pequeñas, medianas y grandes empresas.

SQL Server desde la versión del 2012 permite el desarrollo de aplicaciones diversas y proporciona una amplia inteligencia empresarial lo que le permite ser una base sólida sobre la que las empresas pueden construir su infraestructura de información (Montalvo, 2019).

SQL Server es un sistema de gestión de base de datos integrado en Windows, pero actualmente la versión más reciente también puede ser instalada en distribuciones de Linux.

1.3.1.1. Características de SQL Server

SQL Server tiene una gran capacidad de gestión de datos conservando su integridad y coherencia, realiza las siguientes funciones:

- Almacenar datos.
- Administrar las restricciones de integridad definidas.
- Garantizar la coherencia de los datos almacenados, a pesar de que existan errores en el sistema.

SQL Server se integra a Windows en los siguientes niveles:

- Eventos. Lleva un registro de los errores generados. Los errores son centralizados lo que facilita su diagnóstico.
- Rendimiento. Mediante el analizador de rendimiento es sencillo detectar los cuellos de botella y tomar las acciones necesarias para evitar estos problemas.
- Tratamientos en paralelo. Es capaz de aprovechar las bondades de la arquitectura de multiprocesos. Esto es que cada instancia se le asigna su propio proceso de ejecución para explotar al máximo los componentes del equipo.

- Seguridad. Puede utilizar la seguridad gestionada por Windows, así permite a los usuarios tener un único usuario y contraseña. Pero también puede gestionar su propio sistema de seguridad.
- Servicios de Windows para la ejecución de los componentes de software correspondientes al servidor. Mediante estos servicios es mucho más fácil gestionar el servidor y disfrutar de las funcionalidades de Windows.
- Directorio activo. Se registra de forma automática en el directorio activo. Esto facilita la búsqueda de instancias que estén funcionando dentro del dominio.
- Puede gestionar bases de datos con rendimiento en memoria y procesamiento de transacciones en línea, también como bases de datos con procesamiento analítico en línea.

De acuerdo a al sitio de Microsoft SQL Server Engine, (2019) las nuevas funcionalidades de la herramienta están inmersas en estos aspectos: Inteligencia en todos tus datos con clústeres de Big Data; capacidad de elegir el lenguaje y la plataforma; rendimiento líder del sector; la plataforma de datos más segura; alta disponibilidad incomparable; BI móvil integral; SQL Server en Azure (Microsoft SQL Server, 2019).

1.3.1.2. Cliente/Servidor

La arquitectura que utiliza SQL Server para la administración de los datos de las aplicaciones se basa en cliente/servidor. El software cliente habilita a que los equipos cliente se conecten a una instancia de Microsoft SQL Server en una red. Un "cliente" es una interfaz de tipo front-end que utiliza los servicios que proporciona un servidor de SQL Server. El servidor se encarga de la gestión de los datos y de la administración de los recursos del servidor entre las diferentes transacciones de los clientes. (Microsoft SQL Server, 2019). En la figura 7 se muestra un esquema en donde hay peticiones de clienetes y el motor de base de datos como servidor dando respuesta a esas peticiones de los clientes.

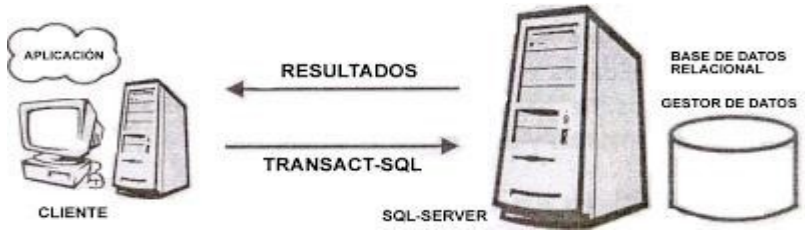


Figura 7. Funcionamiento cliente servidor. Fuente: (González, s.f.)

1.3.1.3. Motor de base de datos

El Motor de base de datos de SQL Server es el principal servicio para almacenar, procesar y proteger los datos. El Motor de base de datos proporciona un acceso controlado y un procesamiento de transacciones rápido para cumplir los requisitos de las aplicaciones consumidoras de datos más exigentes. Los servicios del Motor de base de SQL Server permiten crear aplicaciones de alto rendimiento de base de datos para el procesamiento de transacciones en línea y el soporte en línea de procesamiento analítico (Microsoft SQL Server Engine, 2019).

La parte principal y más importante de SQL Server es el motor de almacenamiento, ya que controla cómo se almacenan los datos en el disco. El Motor de base de datos proporciona acceso y procesamiento de transacciones. Por ejemplo, la definición de los tipos de datos de una tabla y columna, mejora el rendimiento de las consultas, crear y mantener índices. Puede dividir grandes tablas e índices a través de múltiples estructuras de almacenamiento, aprovechando el particionamiento (Montalvo, 2019).

1.3.1.4. Servicio de análisis de datos

Este componente de SQL Server incluye las herramientas para crear y administrar aplicaciones de procesamiento analítico en línea (OLAP) y de minería de datos (Montalvo, 2019).

1.3.1.5. Seguridad

De acuerdo con las especificaciones de Microsoft SQL Server Seguridad, (2017), SQL Server proporciona una arquitectura de seguridad creada para permitir a los administradores de bases de datos y desarrolladores desarrollar aplicaciones de base de datos seguras y contrarrestar las amenazas. El marco de seguridad de

SQL Server permite el acceso a los elementos de la base de datos mediante autenticación y autorización (Microsoft SQL Server Seguridad, 2017).

- La autenticación es el proceso de inicio de sesión para SQL Server por el cual un usuario solicita el acceso mediante el envío de credenciales que el servidor evalúa. La autenticación establece la identidad del usuario.
- La autorización es el proceso con el que se determinan los recursos susceptibles de protegerse a los que tiene acceso un usuario y las operaciones que les están permitidas. SQL Server admite dos modos de autenticación, el modo de autenticación de Windows y el modo mixto (Microsoft SQL Server Seguridad, 2017).

1.3.1.5.1. Seguridad basada en roles

SQL Server utiliza la seguridad basada en roles. Los roles se crean, se les establecen los permisos y posteriormente, se asignan a los usuarios. Existen roles fijos de servidor y base de datos que son un conjunto fijo de permisos asignados.

- Los roles fijos de servidor están pensados para su uso en la administración de SQL Server y los permisos asignados a estos roles se pueden modificar.
- Los roles fijos de base de datos integran un conjunto de permisos diseñados para administrar grupos de permisos.
- Para utilizar los objetos de base de datos, se deben asignar inicios de sesión a cuentas de usuario, se podrán agregar entonces a roles de base de datos a los usuarios y heredarán los conjuntos de permisos asociados con estos roles. (Microsoft SQL Server, 2017).

1.3.1.6. Interfaces de programación

A través de Transact-SQL (T-SQL), se tiene acceso a una programación amplia, simple y potente. También se puede ampliar las características del servidor mediante el aprovechamiento de las capacidades de los lenguajes de programación dentro del “common language runtime” (CLR) tales como Microsoft Visual Basic o Microsoft Visual C #.

SQL Server aprovecha las capacidades del lenguaje XML ya que se integra directamente en el motor, lo que permite almacenar y consultar datos XML, así como devolver conjuntos de resultados en una variedad de formatos XML.

1.3.1.7. Replicación

SQL Server tiene la capacidad de distribuir copias de datos, así como mantener todas las copias sincronizadas con el conjunto de datos maestros. A través de los años, las capacidades de distribución de SQL Server se han expandido desde el mantenimiento de múltiples copias de sólo lectura a ser capaz de hacer los cambios de datos en toda la red de bases de datos al tiempo que el motor de replicación sincroniza todos los cambios en el ambiente.

SQL Server incluye soporte para hacer periódicamente instantáneas de un conjunto de datos que se aplican a múltiples máquinas. Después de la aplicación de una instantánea inicial, la replicación transaccional transfiere los cambios incrementales de datos desde el publicador hacia cada suscriptor.

La replicación transaccional también tiene capacidades adicionales: permite que los cambios sean hechos a un suscriptor y sincronizado de nuevo a un publicador, y dejar que una arquitectura punto a punto sea implementada por medio de muchas instancias de SQL Server se pueden replicar entre sí como iguales.

La replicación de mezcla permite a los usuarios móviles, desconectados, puedan tomar conjuntos locales de datos, realizar cambios a nivel local, y luego sincronizar todos los cambios con el servidor.

1.3.1.8. Disponibilidad

SQL Server proporciona varias tecnologías para garantizar la disponibilidad de los datos: la conmutación por error, la copia de base de datos, el envío de registros, y la replicación.

El espejeo de la base de datos se basa en los procesos internos de gestión de registro en el motor de almacenamiento, para mantener una segunda copia de una base de datos con una latencia extremadamente baja. La base de datos

espejeada se puede ejecutar en un modo sincrónico, lo que garantiza que las transacciones no se pierden nunca si ocurre un fallo en la base de datos primaria.

El traspaso de registros se basa en la “copia de seguridad/restauración” del motor junto con el agente SQL Server, para programar aplicaciones automatizadas de copias de seguridad del registro de transacciones en un servidor secundario.

La replicación aprovecha las capacidades compatibles con el motor de replicación, descriptos anteriormente, que le permite mantener una copia de la totalidad o una parte de una base de datos.

En una instancia, se puede aplicar el espejado de base, el traspaso de registros, y la réplica para proporcionar copias redundantes de bases de datos completas o subconjuntos de bases de datos, que pueden ser utilizadas en el caso de una falla en la plataforma de base de datos primaria (Hotek, 2009).

1.3.1.9. Servicio Broker

El servicio Broker permite una gestión de consultas asíncronas. Permite que una aplicación cliente envié una gran cantidad de peticiones a SQL Server y este pueda tratar la peticiones una detrás de la otra. SQL Server permite administrar la cola de mensajes para dar una respuesta puntual a cada petición.

El servicio Broker ayuda a los desarrolladores a crear aplicaciones seguras y escalables. Proporciona una plataforma de comunicación donde los componentes de las aplicaciones independientes trabajen como un conjunto funcional. (Montalvo, 2019)

1.3.2. MongoDB

MongoDB es un sistema de gestión de base de datos no relacional de código abierto desarrollada por la compañía 10gen, el lenguaje en el que ha sido desarrollado es C++, este sistema de gestión de base de datos fue lanzado en el año 2009 y el tipo de base de datos que administra es basada en documentos.

Este sistema de gestión de base de datos surge como una nueva tendencia para las bases de datos NoSQL, bases de datos que no tienen un esquema fijo, que

son más rápidas en el acceso a datos y escalan mejor que las bases de datos relacionales.

Una base de datos en MongoDB consiste en un conjunto de colecciones las cuales equivaldrían a las tablas en los sistemas de gestión de bases de datos relacionales y cada colección puede tener diferentes tipos de documentos u objetos. Un documento es lo que equivaldría a un registro en una base de datos relacional.

Los documentos en MongoDB tienen una estructura JSON, pero es almacenado en un formato más rico llamado BSON el cual es una lista de pares clave-valor. Los valores pueden ser: Valor primitivo, un arreglo de documentos, o una nueva lista de clave valor o documentos (MongoDB, 2020).

1.3.2.1. Modelo enriquecido de datos

Al reemplazar el concepto de la fila con un modelo más flexible, al permitir documentos y arreglos embebidos, permite representar relaciones jerárquicas complejas en un registro. Esto ayuda a los desarrolladores de lenguajes orientados a objetos a tratar los datos como objetos.

MongoDB es de libre esquema, es decir que las claves de un documento no están predefinidas o fijadas de ninguna manera, esto facilita al momento de hacer migraciones de grandes cantidades de información ya que no es necesario modificar el esquema, así las claves nuevas o las que falten pueden ser tratadas a nivel de aplicación. Esto les da a los desarrolladores una gran flexibilidad al momento de trabajar con la evolución de los modelos de datos.

1.3.2.2. Escalabilidad

El tamaño de las bases de datos de las aplicaciones ha ido creciendo a un ritmo increíble. La escalabilidad implica brindar un beneficio para las bases de datos NoSQL ya que permite la disminución de tráfico en las transacciones y búsquedas de la información (Correa Leal, 2015).

Conforme ha ido creciendo la gran cantidad de datos que se requieren almacenar, los desarrolladores se han enfrentado a una difícil decisión, ¿cómo

escalar las bases de datos?, ¿conseguir máquinas más potentes o particionar datos a través de varios equipos de cómputo?

MongoDB fue desarrollado pensando en el principio de escalabilidad, ya que su modelo de datos basado en documentos permite que este divida automáticamente los datos a través de múltiples equipos. Esto se puede realizar al balancear los datos y la carga de un clúster, el cual distribuya los documentos automáticamente permitiendo a los desarrolladores en concentrarse solo en la programación de la aplicación y no en escalarla. Cuando sea necesaria más capacidad, solo se agrega una nueva máquina al clúster y dejan que MongoDB organice todo.

1.3.2.3. Datos como documentos

MongoDB almacena los datos como documentos en una representación binaria llamada BSON (Binary JSON). La estructura BSON enriquece la representación popular de JSON (JavaScript Object Notation) para poder incluir otros tipos de datos como int, long, y float. Los documentos BSON contienen uno o más campos, y cada campo tiene un valor de un tipo de dato específico, incluyendo arreglos, datos binarios y sub-documentos (MongoDB, 2020).

Los documentos que tienen una estructura similar se organizan como colecciones. La colección es la representación de una tabla de una base de datos relacional, los documentos a las filas y los campos a las columnas. La figura 8 muestra un ejemplo de código BSON que acepta MongoDB.

Figura 8. Estructura de un documento en MongoDB. Fuente: (MongoDB, 2020)

1.3.2.4. Modelo de consultas

MongoDB ofrece controladores para la mayoría de los lenguajes de programación para un desarrollo natural. incluyen Java, .NET, Ruby, PHP, JavaScript, Node.js, Python, Perl, PHP, Scala y otros (MongoDB, 2020).

Una diferencia fundamental en comparación con bases de datos relacionales es que el modelo de consulta MongoDB se implementa como métodos o funciones dentro de la API de un lenguaje de programación específico, en contraposición a un lenguaje completamente separado como SQL. Esto, junto con la afinidad entre el modelo de documento JSON de MongoDB y las estructuras de datos utilizadas en la programación orientada a objetos, hace simple la integración con las aplicaciones.

1.3.2.4.1. Tipos de consultas

A diferencia de otras bases NoSQL, MongoDB no se limitada solo a las operaciones clave valor, y así los desarrolladores pueden construir aplicaciones poderosas creando consultas complejas.

Una consulta puede devolver un documento, un subconjunto de campos específicos dentro del documento o agregaciones complejas de muchos documentos, por ejemplo:

- Las consultas de clave-valor devuelven resultados basados en cualquier campo del documento, generalmente la llave primaria.
- Las consultas de rango devuelven resultados basados en valores como las desigualdades (por ejemplo, mayor que, menor que, igual que, entre).
- Las consultas geoespaciales devuelven resultados basados en criterios de proximidad, intersección e inclusión que pueden ser especificados por un punto, línea, circulo o polígono.
- Las consultas de búsqueda de texto revuelven resultados en orden de relevancia basados en argumentos de texto usando operadores booleanos (AND, OR, NOT).
- Las consultas de agregación devuelven agregaciones (count, min, max, average y similares a GROUP BY de SQL).

- Las consultas Map Reduce ejecutan procesamiento complejo de datos que es expresado en Javascript (sitiobigdata.com, 2017).

1.3.2.5. BSON

Los documentos en MongoDB son un concepto abstracto, la representación concreta de un documento varía según el controlador del lenguaje que se utilice. Debido a que los documentos se utilizan ampliamente para las comunicaciones en MongoDB, también es necesario que haya una representación de los documentos que es compartido por todos los drivers, las herramientas y procesos del ambiente MongoDB y esa representación se llama Binario JSON (BSON). BSON es un formato binario capaz de representar e interpretar cualquier documento MongoDB como una cadena de bytes.

1.3.3. Comparativo SQL Server MongoDB

Algunos conceptos que son utilizados en SQLServer son semejantes en MongoDB. Se muestran algunos conceptos comunes en cada sistema en la tabla 1.

Tabla 1.

Conceptos comunes enre SQL Server y MongoDB. Elaboración propia

SQL Server	MongoDB
Base de datos	Base de datos
Tabla	Colección
Fila	Documento
Columna	Campo
Índex	Índex
Clave primaria	Clave primaria
Joins	Documentos embebidos
Agregación	Agregación *pipeline*

Tanto SQL Server como MongoDB tienen un lenguaje rico para realizar consultas.

SQL Server utiliza SQL como lenguaje para realizar sus operaciones el cual tiene una gran cantidad de instrucciones que se han ido enriqueciendo a través de los años.

MongoDB utiliza su propio lenguaje de consulta el cual le permite realizar potentes transacciones.

A continuación, en la tabla 2, se identifica un comparativo de ejemplos de algunas instrucciones más utilizadas tanto en SQL Server como MongoDB.

Tabla 2

Comparativa y ejemplo de instrucciones SQL Server y MongoDB

Instrucción	SQL Server	MongoDB
Seleccionar todos los documentos en una colección.	`SELECT * FROM Alumnos`	`db.Alumnos.find( {} )`
Seleccionar documentos de una colección de acuerdo a la condición.	`SELECT * FROM Alumnos WHERE IdAlumno = '007'`	`db.Alumnos.find( {IdAlumno:"007"})`
Seleccionar documentos de una colección de acuerdo con un operador de consulta.	`SELECT * FROM Alumnos WHERE Calificacion in (7, 10)`	`db.Alumnos.find({Calificacion: {$in:[7,10]}})`
Seleccionar documentos de una colección de acuerdo con el operador AND.	`SELECT * FROM Alumnos WHERE Grado = "A" AND Calificacion <7`	`db.Alumnos.find({Grado: "A", Calificacion:{$lt:7}})`

Seleccionar documentos de una colección de acuerdo con el operador OR	**SELECT** * **FROM** Alumnos **WHERE** Grado = "A" **OR** Calificacion **<**7	**db**.Alumnos.**find**({**$or**:[{Grado: "A"}, {Calificacion:{**$lt**:7}}]})
Insertar uno o más documentos	**INSERT INTO** Alumnos (IdAlumno, Nombre, Grado, Calificacion) **VALUES**('008','Juan Solis','B',7)	**db**.Alumnos.**insertOne**({IdAlumno:"008",Nombre:"Juan Solis",Grado:"B",Calificacion:7}) **db**.Alumnos.**insertMany**([{IdAlumno:"009",Nombre:"Jose Solis",Grado:"A",Calificacion:8},{ IdAlumno:"010",Nombre:"Maria Casio",Grado:"C",Calificacion:6}])
Actualizar uno o más documentos	**UPDATE** Alumnos **SET** Calificacion=10 **WHERE** IdAlumno='008'	**db**.Alumnos.**updateOne**({IdAlumno:"008"},{$set:{Calificacion:10}}) **db**.Alumnos.**updateMany**({"Calificacion":{$lt:8}},{$set:{Grado:"Z"}})
Borrar uno o más documentos	**DELET FROM** Alumnos **WHERE** IdAlumno='008'	**db**.Alumnos.**deleteOne**({IdAlumno:"008"}) **db**.Alumnos.**deleteMany**({IdAlumno:"008"})

La instrucción JOIN es una de las mejores instrucciones de SQL Server y de las bases de datos relacionales en general, ya que esta permite realizar consultas a varias tablas.

MongoDB no soporta la instrucción JOIN, pero lo realiza mediante datos embebidos como arreglos e incluso como otro documento. Esto sería un documento dentro de otro.

Para almacenar los datos, SQL Server requiere primero definir las tablas y las columnas, pero en MongoDB no se define un esquema, el propio sistema de gestión de base de datos acomoda la estructura según los requerimientos.

A continuación, se muestra un breve análisis de fortalezas, debilidades, oportunidades y amenazas para SQL Server y MongoDB, de acuerdo a lo expuesto en este documento.

Figura 9. FODA SQL Server. Fuente: Elaboración propia

Figura 10. FODA MongoDB Fuente: Elaboración propia

Conclusiones

Pensar en un cambio de bases de datos SQL a bases de datos NoSQL, es un gran desafío para los desarrolladores de sistemas que trabajan con bases de datos. Esto es, debido a que desde los inicios de los sistemas de gestión de bases de datos relacionales han sido parte fundamental en los sistemas de datos, además el lenguaje de consulta SQL es un lenguaje estándar para este tipo de base de datos.

Las bases de datos relacionales han desempeñado un papel sorprendente, hasta que se fueron creando sistemas de software en el cual requerían altos volúmenes de transacciones y datos de distintos tipos, esto provocó una afectación en la escalabilidad de los sistemas, ya que difícilmente se podían almacenar y modelar altos volúmenes de datos con este tipo bases de datos. Para resolver esta situación se aumentaron las capacidades de los equipos, pero esta solución generó un alto costo en infraestructura y sistemas demasiado complejos.

Es por eso que las bases de datos NoSQL han dado una solución a los desafíos que enfrentan las bases de datos relacionales, porque ofrecen esquemas dinámicos, modelado flexible, arquitectura escalable y almacenamiento de grandes volúmenes de datos.

Las bases de datos relacionales seguirán existiendo, y aun tienen un gran potencial en el mercado, ya que para ciertas necesidades específicas se seguirán utilizando, además en el mercado los sistemas de gestión de bases de datos relacionales se siguen actualizando y brindando una gama amplia de funcionalidades para los usuarios que adaptan sus necesidades a este tipo de bases de datos.

Las bases de datos NoSQL seguirán evolucionando debido al aumento en el tratamiento de grandes volúmenes de datos y estarán listas para resolver este tipo de necesidades actuales.

Tanto las bases de datos SQL como las NoSQL tienen ventajas y desventajas si se comparan entre sí. Al momento de tratarlas por separado, se descubre que cada una tiene sus bondades y características muy particulares para el ambiente en el que se desean implementar. Para las bases de datos SQL se distingue la consistencia y la seguridad; para las bases de datos NoSQL tener una escalabilidad horizontal, almacenar grandes volúmenes de datos, realizar transacciones en grandes cantidades de datos y flexibilidad en el modelado de datos.

Existe en el mercado un amplio abanico de alternativas de distinta y variada oferta para elegir qué tipo de gestor base de datos utilizar y resolver las necesidades requeridas a los sistemas y aplicaciones que se desarrollen.

Referencias

acenswhitepapers. (24 de 02 de 2014). *acens. The Cloud services company de telefonía*. Obtenido de Bases de datos NoSQL. Qué son y tipos que nos podemos encontrar: https://www.acens.com/wp-content/images/2014/02/bbdd-nosql-wp-acens.pdf

Aguilar Romero, M., & Rodríguez García, J. L. (2016). Comparación de opciones para inteligencia de negocios en los principales sistemas gestores de bases de datos del mercado. *Economía y Administración (E&A) , Vol.7 (1)*, 5-20.

Apache Cassandra. (01 de 01 de 2016). *Apache Cassandra*. Obtenido de Architecture and Overview: https://cassandra.apache.org/doc/latest/architecture/overview.html

APACHE HBASE. (14 de 07 de 2020). *APACHE HBASE*. Obtenido de Welcome to Apache HBase: https://hbase.apache.org/

Arias, Á. (2015). *Bases de Datos con MySql. Segunda Edición.*

Baldassari Valencia, H. D. (2019). Estudio comparativo de motores de bases de datos SQL y NoSQL para la gestión de información transaccional. *DISERTACIÓN PREVIA A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SISTEMAS Y COMPUTACIÓN.* . Quito, Ecuador, Ecuador: FACULTAD DE INGENIERÍA DE SISTEMAS Y COMPUTACIÓN. PONTIFICIA UNIVERSIDAD CATÓLICA DEL ECUADOR. .

Bender, C., Deco, C., González Sanabria, J., Hallo, M., & Ponce, J. (2014). *Tópicos avanzados de Bases de datos.* México, Perú, Brasil, Uruguay: 1a ed. - Iniciativa Latinoamericana de Libros de Texto Abiertos.

Correa Leal, L. G. (2015). Análisis comparativo entre la base de datos no relacional MongoDb con la base de datos Postgresql, sistema para la gestión de clientes y registro de pagos de la clínica odontológica Ortho Dent. *Trabajo de Gado de Ingeniero en Sistemas Computacionales*. Ibarra, Ecuador, Ecuador: Universidad Técnica del Norte. Facultad de Ingeniería en Ciencias Aplicadas. carrera de Ingeniería en Sistemas Computacionales.

DB-Engines. (11 de 07 de 2020). *DB-Engines*. Obtenido de DB-Engines Ranking: https://dbengines.com/en/ranking

Formia, S. A., & Estevez, E. (2019). Implementación y Maduración de un Data Warehouse –Caso de Estudio de la Agencia de Recaudación Tributaria de Río Negro (ARTRN). *11o. Simposio Argentino de Informática en el Estado (SIE) - JAIIO 46 (Córdoba, 2017)*, 90-100.

García Cebreiros, R. (01 de 07 de 2016). Análisis de bases de datos y tendencias tecnológicas. *Un caso de uso en Twitter para aplicaciones de análisis de sentimientos y opiniones*. Alicante, Alicante, España: Universidad de Alicante.

GENBETA. (28 de 01 de 2014). *GENBETA: Dev*. Obtenido de GENBETA: Dev: https://www.genbeta.com/desarrollo/nosql-clasificacion-de-las-bases-de-datos-segun-el-teorema-cap

González, R. (s.f. de s.f. de s.f.). *BASE DE DATOS - SQL SERVER*. Obtenido de Cliente Servidor: https://perusql20005.blogspot.com/2010/01/cliente-servidor.html

Gracia del Busto, H., & Yanes Enríquez, O. (2012). Bases de datos NoSQL. *Revista Telem@tica. Vol. 11. No. 3*, 21-33.

Hotek, M. (2009). *Microsoft SQL Server 2008.* Obtenido de Step by Step: http://docshare02.docshare.tips/files/17067/170672242.pdf

Hueso, M., Ibeas, J., Revuelta, I., Santos, F., Soler, M. J., & Buades, J. M. (2019). Big data y ciencia de los datos para una nefrología personalizada: ¿estamos preparados para una "nefrología inteligente"? *Sociedad Española de Nefrología. Servicios de edición de Elsevier España S.L.U.*, 10.

Jones, H. (2019). *Analítica de Datos. La guía definitiva de análisis de Big Data para empresas, técnicas de minería de datos, recopilación de datos y conceptos de inteligencia empresarial.* México: Independently published.

Jones, H. (2019). *Ciencia de los Datos. Lo que saben los mejores científicos de datos sobre el análisis de datos, minería de datos, estadísticas, aprendizaje automático y Big Data que usted desconoce.* México: Amazon Mexico Services, Inc.

Meza Quiñones, E. E. (2018). Introducción a los Sistemas de Datos. Modelo Relacional. Diseño de Bases de Datos Distribuidos. Modelo Orientado a Objeto. Aplicaciones Comerciales. Uso de Modeladores de Portales Web. Aplicaciones Web de Sistemas Educativos. *Para optar al Título de Segunda Especialidad Profesional Especialidad: Informática Educativa*. Lima, Perú, Perú: FACULTAD DE CIENCIAS. Escuela Profesional de Matemática e Informática.

Microsoft SQL Server. (30 de 03 de 2017). *Microsoft SQL Server Documentación*. Obtenido de Roles de servidor y base de datos en SQL Server: https://docs.microsoft.com/es-es/dotnet/framework/data/adonet/sql/server-and-database-roles-in-sql-server

Microsoft SQL Server. (01 de 10 de 2019). *SQL Server 2019*. Obtenido de Nuevas capacidades de SQL Server 2019: https://www.microsoft.com/es-es/sql-server/sql-server-2019-features

Microsoft SQL Server Engine. (26 de 07 de 2019). *Microsoft*. Obtenido de SQL SEver Engine: https://docs.microsoft.com/en-us/sql/database-engine/install-windows/install-sql-server-database-engine?view=sql-server-ver15

Microsoft SQL Server Seguridad. (30 de 03 de 2017). *Microsoft Documentación SQL Server y ADO.NET*. Obtenido de Información general sobre la seguridad de SQL Server: https://docs.microsoft.com/es-es/dotnet/framework/data/adonet/sql/overview-of-sql-server-security

Mike, H. (2009). *Microsfot SQL Server 2008.* Obtenido de Step by Setp: http://docshare02.docshare.tips/files/17067/170672242.pdf

MongoDB. (15 de 07 de 2020). *MongoDB*. Obtenido de The database for modern applications: https://www.mongodb.com/es

Montalvo, G. (01 de 07 de 2019). Análisis Comparativo de Migración de Motor de BDD de SQL Server a Oracle en una Empresa del Sector Automotríz. *Trabajo previo a la obtención del título de Ingeniero en Sistemas y Computación*. Quito, Quito, Ecuador: Facultad Ingeniería Carrera Sistemas y Computación. Pontificia Universidad Católica del Ecuador.

neo4j. (15 de 07 de 2020). *neo4j*. Obtenido de The Native Graph Database for Today's Connected Applications: https://neo4j.com/neo4j-graph-database/

Osorio Rivera, F. L. (2008). *Bases de Datos Relacionales. Teoría y Práctica.* Medellín Colombia: Instituto Tecnológico Metropolitano.

Silberschatz, A., Korth, H. F., & Sudarshan, S. (2002). *Fundamentos de Bases de Datos. Cuarta Edición.* MADRID, BUENOS AIRES, CARACAS, GUATEMALA, LISBOA, MÉXICO, entre otras: McGraw Hill.

sitiobigdata.com. (17 de 12 de 2017). *MongoDB Arquitectura y modelo de datos.* Obtenido de Tipos de Querys: https://sitiobigdata.com/2017/12/27/mongodb-arquitectura-y-modelo-de-datos/#

Capítulo 2

Big Data y su impacto en la sociedad

Carolina Sosa Hernández
Tecnológico Nacional de México. Instituto Tecnológico de Durango
02040145@itdurango.edu.mx

José Gabriel Rodríguez Rivas
Tecnológico Nacional de México. Instituto Tecnológico de Durango
gabriel.rodriguez@itdurango.edu.mx

2.1. Introducción

Desde el preludio de la civilización, el hombre creó e implementó procesos que le permitieron el uso óptimo de los recursos a su alcance para garantizar la satisfacción de sus necesidades y la prosperidad de los pueblos. De tal manera que, en la época antigua surgieron grandes imperios, todos con una sólida organización, visión de orden y un extraordinario talento para dirigir.

Ya en la edad media se observaba a las organizaciones formales como la Iglesia Católica, consolidar los conceptos de disciplina, unidad de mando y dirección, autoridad, jerarquía y responsabilidad; y a la Milicia aportar sus principios estratégicos y de organización sí como el uso de palabras restringidas, entre las que se encuentran: estrategia, tácticas, operaciones, reclutamiento y logística. (Hernández, 2002).

La Revolución Industrial caracterizada por la aparición y difusión masiva de la máquina de vapor junto con la introducción del carbón y del hierro facilitó el crecimiento acelerado de varias industrias como la textil, con varias consecuencias sociales.

Muchos y grandiosos inventos surgieron en los siglos XIX y XX dejando en claro que el avance tecnológico, industrial, humanístico y de comunicaciones, es una constante en el marco de la existencia humana. Como ejemplo de éstos inventos en el ámbito de la computación se encuentran (Sánchez, 2005).

- 1621 Reglas del cálculo
- 1671 Multiplicadora –Gottfried Leibniz
- 1805 Las cintas perforadoras en los telares –Joseph Marie Jacquard
- 1847 La algebra de la lógica o booleana - George Boole
- 1944 MARK I –IBM y Howard Aiken
- 1946 ENIAC- Primer ordenador electrónico
- 2ª generación de computadoras (1955-1964) transistor IBM 7090 y 7094
- 3ª generación de computadoras (1965-1970) circuito integrado, IBM 360, minicomputadore
- 4ª generación de computadoras (1971- adelante) microprocesador PC, estación de trabajo
- 1960 Inicio de los Sistemas de información
- 1969 Internet
- 1973 Teléfono móvil
- 1976 Supercomputadora- Van Tassel /Cray
- 1977 Ordenador personal –IBM
- 1979 Disco compacto
- 1990 Word Wide Web –CERN
- 1993 Global position system GPS- EE.UU
- 2004 Web 2.0 modificación del uso y/o interacción web de estático-lectura a dinámico lectura-escritura-modificación –O´reilly
- Internet de las cosas
- Computación en la nube
- Internet del Todo
- Big Data.

En este breve recorrido histórico, surge el Big Data sutilmente en el rico y extenso conocimiento guardado y administrado por los imperios, el clero y la milicia; en el despertar de la revolución industrial o en los inventos del siglo XX. De esa información generada (datos), se tiene constancia en archivos, libros, informes, revistas, historiales y con la aparición de la informática, ahora en un soporte digital

Una importante justificación para escribir sobre Big Data, es para los estudiosos de los fenómenos sociales, el hecho de que existan más de 3,000 millones de personas conectadas a Internet y más de 7,000 millones de suscripciones a telefonía celular en el mundo, les conduce a investigar la reciente y monumental avalancha de información. Por ende, la tarea de las Ciencias Sociales en lo que respecta a los grandes volúmenes de datos, es sin duda, entender qué fenómenos sociales y qué dilemas éticos trae consigo y cómo su análisis puede ayudar a entender, proyectar y resolver problemáticas sociales.

Si se realizara una encuesta donde se le preguntara a las personas que entienden por Big Data, probablemente muchos no sabrían dar una respuesta, otros tantos se atreverían a comentar que se trata de un término puramente técnico y solo unos cuantos atinarían a decir que se trata de la cantidad masiva de datos derivada de la creciente ola tecnológica que impacta directa o indirectamente nuestra manera de vivir.

Como consecuencia de lo anterior, ha surgido un genuino interés por conocer los resultados obtenidos de dicha labor, siendo el porqué de este trabajo de tipo académico, estableciéndose los siguientes objetivos.

Objetivo general:

Identificar las características del Big Data desde una de las perspectivas fundamentales de la humanidad: la social, a través de la recopilación de casos de éxito y el estudio de la literatura actual cuyos autores compartan una opinión imparcial.

Objetivos específicos:

- Definir el concepto de Big Data.

- Describir los elementos técnico prácticos del Big Data.
- Analizar el Big Data a la vida cotidiana mediante ejemplos reales de impacto social.
- Analizar sus posibles riesgos y presentar algunas conclusiones.

Es el marco teórico se define el concepto de Big Data, mencionando de forma general los diferentes tipos de bases de datos, así como las técnicas y tecnologías utilizadas actualmente para procesar y analizar los macrodatos.

En el desarrollo. Se describen cinco casos en los que Big Data ha tenido una injerencia positiva en la sociedad sin dejar de lado los riesgos que este cambio trae consigo. Se presenta además la opinión y el aprendizaje personal del autor; así como las ventajas, desventajas y problemática descubiertas sobre el objeto de la investigación.

2.2. Marco de referencia

2.2.1. Era digital y Big Data

El volumen de la información ha crecido a un ritmo sin precedentes haciéndose más compleja su administración, por lo que desde las grandes corporaciones hasta el ciudadano promedio han tenido que avanzar al mismo ritmo creando y utilizando métodos para su compresión.

De acuerdo con (Gartner, sf), define el Big Data como “un gran volumen, velocidad o variedad de información que demanda formas costeables e innovadoras de procesamiento de información que permitan ideas extendidas, toma de decisiones y automatización del proceso”. Sus dimensiones pueden describirse de la siguiente forma:

- **Volumen**: Se refieren a la cantidad de datos disponibles. Cada día, las empresas registran un aumento significativo de sus datos, estos son creados por personas (P&P), maquinas-personas (M&P) y máquinas-maquinas (M&M). A mediados del año 2012, en una encuesta de la empresa IBM, se preguntó a expertos el

volumen de datos requerido para ser considerado Big Data, a lo que respondieron que para ser considerada como tal la base de datos tiene que ser superior a 1,114 terabytes.

Para darse una idea de ¿qué tan grande es, BIG?, 16 millones de fotografías de Facebook pueden ser almacenadas en un terabyte; sin embargo, estas apreciaciones varían.

- **Variedad:** Existen diversas formas de representar los datos, pudiendo ser datos estructurados y no estructurados; estos últimos son los que se generan desde páginas web, redes sociales, foros, archivos de búsquedas, correos electrónicos o bien pueden originarse de sensores en diferentes actividades de las personas. Por ejemplo, Si se toma una base de datos de Twitter, en un twit se ve que además del texto hay ligas a una serie de recursos culturales diversos como videos, audios o fotografías, los cuales la mayoría de las veces se encuentran no estructurados.
- **Velocidad:** Se refiere a la velocidad con que se originan los datos, de las diversas fuentes de datos estructurados y datos no estructurados como pueden ser páginas Web, bases de datos, redes sociales, call centers, datos geoespaciales, datos semiestructurados (XML, RSS) provenientes de audio y video, los datos generados por los termómetros, el Internet de las cosas, las RFID, entre otras, en suma con las interacciones del hombre.

Basta pensar en la cantidad de individuos que pasan por los controles de seguridad en los aeropuertos del mundo, los registros de cámaras de vigilancia o las transacciones bancarias para apreciar el continuo e implacable flujo de generación de datos.

El volumen, variedad y velocidad son los atributos técnicos que pueden ayudar a definir Big Data, en la actual literatura hay cierto consenso al respecto, sin embargo, resultan insuficientes cuando existen otras propiedades necesarias para entenderlo, como la veracidad, objetividad, representatividad y los dilemas éticos asociados (Meneses, 2018).

2.2.2. Tecnologías Big Data

Entonces, ¿cómo procesar y analizar esos volúmenes de información? Para ello, es necesario conocer la evolución de las bases de datos y comprender la forma en la que se almacena y organiza la información.

Las bases de datos comenzaron a aparecer entre los años 1950 y 1960, mayormente impulsadas por el incremento de dos factores tecnológicos: la confiabilidad de los procesadores computacionales y la capacidad de almacenamiento en cintas y unidades de disco. (Castro, González y Callejas, 2012).

En consecuencia, las bases de datos se convirtieron en un gran soporte para organizar la información. Surgiendo varios modelos que se explican brevemente a continuación.

2.2.2.1. Tipos de Base de datos

2.2.2.1.1. Jerárquicas

En este tipo de base de datos se organiza la información con una jerarquía del tipo padre / hijo. Existe una serie de nodos que contendrán atributos y que se relacionarán con nodos hijos de tal forma que puede existir más de un hijo para el mismo padre, pero un hijo sólo tiene un padre.

Las entidades en este tipo de modelo se denominan segmentos y sus atributos se llaman campos. Las estructuras jerárquicas fueron usadas ampliamente en los primeros sistemas de gestión de datos de unidad central, como el Sistema IMS por IBM. Su desventaja consiste en que pierden simplicidad a medida que el "volumen de la información" va incrementándose.

2.2.2.1.2. En Red

Este tipo de base de datos organiza la información en **registros y enlaces**. En los registros se almacenan los datos utilizando atributos. Los enlaces realizan las relaciones entre los registros de la base de datos. Las bases de datos en red

son parecidas a las jerárquicas pero se diferencian en que en ellas puede haber más de un padre.

En este modelo se pueden representar perfectamente relaciones de muchos a muchos. Tienen la gran desventaja que son complejas y tienen gran dificultad de manejo ocasionando que se estén abandonando completamente este tipo.

2.2.2.1.3. Relacionales

Se basan en el uso de tablas (relaciones). Las tablas se representan gráficamente como una estructura rectangular formada por filas y columnas. Cada columna (atributo) almacena información sobre una propiedad determinada de la tabla. Cada fila (tupla) posee una ocurrencia o ejemplar de la instancia o relación representada por la tabla.

Uno de los objetivos de toda base de datos es producir un conjunto de respuestas a partir de una o varias solicitudes o consultas. En el caso del modelo relacional las consultas se especifican utilizando el SQL (por sus siglas en inglés Structured Query Language) o Lenguaje de Consulta Estructurado, un estándar que permite al usuario expresarse en forma declarativa sin ninguna instrucción detallada de programación. (Castro, González y Callejas, 2012).

En algunas ocasiones derivado de fallas en el diseño o a problemas indetectables se pueden producir los siguientes problemas:

- Redundancia. Cuando existen datos que se repiten continua e innecesariamente.
- Anomalías en operaciones de modificación de datos. Existe cuando al insertar un solo elemento, se debe de repetir registros en una tabla en el que varían unos pocos atributos, o que al eliminar un elemento suponga eliminar varios registros.

2.2.2.1.4. Orientadas a objetos

Desde el surgimiento de la programación orientada a objetos, se consideró en adaptar las bases de datos a estos lenguajes. Disponen de mayor expresividad y son adecuadas para almacenar muchos tipos de datos diferentes. Capacidad de manejar herencia entre tablas lo hace muy poderoso.Son las bases de datos de tercera generación (la primera fue las bases de datos en red y la segunda las relacionales). Su modelo conceptual se suele diseñar en UML.

2.2.2.1.5. Objeto Relacional

Toman lo mejor de dos de los modelos de bases de datos más usadas, resultando en un híbrido entre el modelo relacional y el modelo orientado a objetos. Se añade a las bases de datos de tipo objeto relacional la posibilidad de almacenar procedimientos de usuario, disparadores, tipos definidos por el usuario con sus propias propiedades y consultas recursivas.

2.2.2.1.6. Bases de datos No relacionales

A pesar de la importante contribución de las bases de datos relacionales al ser una de las más utilizadas en el mundo para todo tipo de proyectos, el surgimiento en los últimos años de una incontenible recopilación de datos no estructurados y datos semi-estructurados que surgen de diversas fuentes como redes sociales, blogs, Internet de las cosas, entre otros, surge la necesidad de un nuevo paradigma o modelo para hacer frente a este nuevo tipo de datos.

Para satisfacer las actuales necesidades informáticas surgen las bases de datos NoSQL. Es común pensar que el concepto NoSQL es la oposición directa al de SQL, sin embargo, esa “afirmación” difiere del planteamiento original ya que NoSQL se define como “No Solo SQL” termino que es usado para denominar a todas las bases de datos que no siguen los principios las bases de datos relacionales.

2.2.3. Herramientas para el Big Data

Big Data ha propiciado la aparición del Data Science o Ciencia de los Datos, término que hace referencia al conjunto de tecnologías y técnicas necesarias para el tratamiento de la información masiva desde los puntos de vista estadístico e informático, resolviendo al tiempo el problema de almacenamiento de los datos.

Derivado de ello, un nuevo perfil ha surgido, el “Data Scientist” –Científico de Datos. Las personas de este perfil, deben saber de herramientas computacionales, análisis e interpretación estadística.

Por lo tanto, podemos decir que para una excelente gestión de los datos es necesario contar con un experto en el campo y una adecuada infraestructura tecnológica (hardware y software) basada, entre otros aspectos, en técnicas que posibiliten un correcto almacenamiento y posterior análisis de los mismos. Lo anterior da la pauta para hablar de las tecnologías que iniciaron el ecosistema Big Data.

2.2.3.1. Hadoop.

Las tecnologías de Big Data se clasifican en las que dan soporte a la captura, la transformación, el procesamiento y el análisis de los datos, ya sean estructurados, semiestructurados o no estructurados.

Hadoop es una librería de Apache definida como un framework que permite el procesamiento de grandes volúmenes de datos a través de clústeres de computadoras que utilizan modelos de programación muy sencilla. Está diseñado pensando en la escalabilidad; desde un par de servidores hasta cientos de máquinas o nodos.

Cuenta con dos componentes principales, el primero es HDFS (Sistema de Archivos Distribuidos Hadoop) que permite distribuir los archivos en distintos equipos y tiene tres pilares principales: Namenode, Datanodes, Jobtracker. Sus principales características son (Hernández, Duque y Moreno, 2017):

- Tolerancia a fallos

- Acceso a datos en streaming
- Facilidad para el trabajo
- Modelo sencillo de coherencia
- Portabilidad de convivencia.

El segundo componente de Hadoop es MapReduce. MapReduce es un modelo de programación distribuida que permite el procesamiento masivo de datos a gran escala de manera paralela, desarrollado como alternativa escalable y tolerante a fallos. MapReduce divide el procesamiento en dos funciones: Map y Reduce.

Función Map(). Funcion caracterizada por trabajar con grandes volúmenes de datos. Estos datos son divididos en dos o más partes. Cada una de estas partes contiene colecciones de registros o líneas de texto. La finalidad de la función es calcular un conjunto de valores intermedios basados en el procesamiento de cada registro, agruparlos de acuerdo a la clave intermedia y posteriormente enviarlos a la función Reduce(), la cual se ejecuta para cada elemento de cada lista de valores intermedios que recibe. El resultado final se obtiene mediante la recopilación e interpretación de los resultados de todos los procesos que se ejecutaron.

Se puede decir entonces que MapReduce divide la informática en dos fases:

- **De asignación**, en la que los datos se dividen en bloques que se pueden procesar por subprocesos independientes, incluso ejecutándose en máquinas distintas;
- **De reducción**, que combina la salida de los asignadores múltiples en un resultado final.

Ambas fases en el ecosistema de Hadoop se han clasificado de alto rendimiento para el procesamiento de datos a gran escala demostrando la reducción de tiempo comparado con los sistemas de escritorio tradicionales.

2.3. Desarrollo

2.3.1. Big Data: Un enfoque optimista

En la actualidad los datos son fuente de valor económico, social y político, más si se trata de grandes volúmenes. Se les considera un bien público que beneficia la transparencia y mejora la toma de decisiones.

Por ejemplo, se tiene registro de los proyectos que impulsa la Organización de Naciones Unidas bajo el nombre Global Pulse, los cuales demuestran cómo el análisis de grandes volúmenes de datos puede ayudar a frenar una epidemia como el ébola al mapear el recorrido de los casos mediante teléfonos celulares, alertando a los servicios de salud de los diversos continentes.

En tanto que la práctica científica ha sido ya modificada con el análisis de cuantiosos datos en diversos campos disciplinarios como la física y la biología. En estudios económicos es posible realizar proyecciones bursátiles y detectar tasas de inflación en tiempo real rastreando los precios de millones de productos.

El periodismo inmerso en el colosal caudal de datos, ha dado paso al periodismo de datos. Los *Papeles de Panamá*, que develaron una compleja red de evasión de impuestos a escala global, son ejemplo del poder informático (Ascanio, 2016). Se analizaron 11.5 millones de archivos con más de un millón de imágenes, más de dos millones de archivos en pdf, tres millones de bases de datos y cinco millones de correos electrónicos en la más voluminosa filtración de la historia del periodismo.

En lo que a redes sociales se refiere, la energía social en Twitter por ejemplo, se ha analizado con el objeto de comprender fenómenos sociopolíticos (Meneses, 2018). Facebook, que reúne diariamente grandes cantidades de datos de sus usuarios (intereses, lugares a donde van, redes de amigos, horarios de conexión, instituciones a las que pertenecen y mucho más), crea perfiles que le permiten ubicar las publicidades de una manera selectiva: pañales para las madres de niños pequeños, whisky para los amantes de las bebidas, viajes a México para quienes visitaron la página de una agencia de viajes, entre otros.

Figura 1. Facebook y Twitter, los titanes de las redes sociales.
Fuente: (Agudo, 2014)

Por la razón anterior, puede ser considerada como una empresa de publicidad con sondeos y encuestas permanentes que le permiten elaborar mensajes individualizados. Siendo aún mejor si se considera que son los mismos clientes quienes producen los contenidos que mantendrán interesados a otros clientes y tomando en cuenta que no hay una persona decidiendo a quién ofrecerle pañales o whisky, sino que esa tarea la hará un algoritmo, es decir, un programa que al ser alimentado con *Big Data* «aprenderá» qué ofrecer a cada quien según sus intereses.

Estos algoritmos, además, son capaces de aprender por prueba y error para mejorar su participación, si a mujeres de cierta localidad, edad, nivel cultural, entre otros, les interesó tal producto, probablemente a otras con el mismo perfil también les interese.

El especialista Martin Hilbert resume el fenómeno de la siguiente manera: el acceso a Big Data convirtió a las ciencias sociales, de las que siempre se burlaron, en las ciencias más ricas en datos. Nosotros nunca tuvimos datos, y por eso nunca funcionaban las políticas públicas. Y de la noche a la mañana, 95% de los sujetos que estudiamos pasó a tener un sensor de sí mismo 24 horas al día (Hildbert, 2017).

Los biólogos siempre dijeron “eso no es ciencia, no tienen datos. Pero ellos no saben dónde están las ballenas en el mar. Hoy, nosotros sí sabemos dónde están las personas y también sabemos qué compran, qué comen, cuándo duermen, cuáles son sus amigos, sus ideas políticas, su vida social” (Hildbert, 2017). A través del Big Data, se pueden prever comportamientos y aprender de las experiencias anteriores.

La información viene y va en todas direcciones y con ello surgen nuevas formas de procesarla y, a partir de ahí, se obtiene un conocimiento ampliamente detallado de la población, desde los estados de ánimo hasta los consumos, pasando por los hábitos para moverse o quiénes son sus amigos. Quien accede a esa información y tiene la capacidad de procesarla posee una poderosa herramienta para influir sobre la población. (Magnani, 2017).

Los grandes datos tienen la ventaja de que las ciencias computacionales pueden capturar huellas, movimientos, opiniones y prácticas culturales de millones de personas, lo cual es imposible para el etnógrafo tradicional y para los estudios de la sociología. Se trata de una acumulación gigantesca de huellas e imaginaciones contenidas en el hardware y software que pueden develar hechos y situaciones únicas (Meneses, 2018).

Por todo lo anterior y bajo un enfoque optimista, se expone a detalle cinco casos de éxito que hacen referencia a lo expresado anteriormente.

2.3.2. Primer caso. El séptimo arte

Es difícil prever hasta dónde pueden llegar las operaciones Big Data. Uno de los ejemplos más llamativos es su introducción en el cine, en específico la película Amanecer parte 2. La película es un ejemplo de cómo interpretar los sentimientos a través del análisis de datos, analizando los tuits generados en la promoción de la película. La campaña de marketing fue a nivel mundial y estaba basada en las redes sociales. Los movimientos en Twitter se analizaron con social sentiment index de IBM.

Mediante la herramienta de IBM se pueden prever los comportamientos del mercado utilizando un análisis con grandes cantidades de datos aportados por las redes sociales. En el caso de la película Amanecer parte 2, se analizaron los tuits publicados once días antes de su estreno en los Estados Unidos y los compararon con los tuits de las películas competidoras en ese momento, para determinar los sentimientos que cada una de ellas generaba y de esta manera predecir el comportamiento en taquilla.

El estudio arrojo como resultado que el sentimiento positivo de un 90% generado en el pre lanzamiento, pasó a un 75% el día del estreno, Además el estudio que el cambio no fue por las críticas negativas, sino por los lamentos por finalizar la saga. (Ferrer-Sapena y Sánchez-Pérez, 2013).

Figura 2. Big Data y su influencia en el cine
Fuente: (merkactiva, 2017)

2.3.3. Segundo caso. ¿Saldrás de viaje?, los DTI son la mejor opción.

Por mucho tiempo se ha imaginado un espacio en el que la tecnología y la sociedad se integren de manera armónica. En la actualidad es frecuente encontrar a viajeros cada vez más exigentes. Buscan nuevas emociones y mayor número de viajes. Por tales motivos en España surgió un proyecto que fue recogido oficialmente en el Plan Nacional e Integral de Turismo, aprobado por el Consejo de Ministros en el mes de junio de 2012. El reto, transformar los destinos turísticos en Destinos Turísticos Inteligentes (DTI).

Para la Sociedad Mercantil Estatal para la Gestión de la Innovación y las Tecnologías Turísticas, S.A.M.P., SEGITTUR (2014) un Destino Turístico Inteligente es *"*un espacio turístico innovador, accesible a todos, consolidado sobre una infraestructura tecnológica de vanguardia que garantiza el desarrollo sostenible del territorio, facilita la interacción e integración del visitante con el entorno, e

incrementa la calidad de su experiencia en el destino y la calidad de vida de los residentes".

Figura 3. DTI, una nueva forma de viajar.
Fuente: (Aenor, sf)

Consecuentemente, todas las acciones en torno al proyecto son enfocadas al desarrollo sostenible de sus tres vertientes principales medio ambiental, económica y socio-cultural.

El proyecto tiene como antecedente a la "Ciudad Inteligente" CI. Considerada como una zona de límites geográficos y político administrativos totalmente definidos, la cual otorga primacía a las tecnologías de información y comunicaciones (TIC), con el objetivo diseñar una ciudad dotada de tecnología innovadora, que facilite el desarrollo urbano sostenible y mejore la calidad de vida de los ciudadanos.

En el caso de los DTI el territorio es el eje central en torno al cual se estructuran sus dos pilares básicos: las Nuevas Tecnologías de la Información y de la Comunicación (NTIC) que dotan al espacio de inteligencia permitiéndole proveer entre muchas otras cosas sistemas de movilidad al visitante, permitiéndole tener auténticas experiencias; y el desarrollo turístico sostenible y accesible que busca satisfacer las necesidades presentes de los residentes y turistas y mejorar las

futuras administrando los recursos económicos, sociales y estéticos, respetando la integridad cultural, los procesos ecológicos esenciales, la diversidad biológica y los sistemas de apoyo a la vida. En este punto es válido acentuar las **diferencias** entre una CI y un DTI (de Ávila y Sánchez, 2015).

- El Destino Turístico Inteligente es impulsado por el sector turístico público y privado.
- El público objetivo es el turista, no el ciudadano;
- Los límites geográficos pueden coincidir con los de un municipio o no (ejemplos: Costa del Sol, Camino de Santiago, entre otros.)
- La interacción va más allá de la propia estancia en la ciudad. En los DTI comienza antes de que el visitante llegue al destino, continúa durante su estancia y se prolonga hasta después de su marcha.
- Los DTI están ligados al aumento de competitividad del mismo y a la mejora de la experiencia del turista. Las ciudades inteligentes están orientadas a mejorar la gobernabilidad de la misma y a incrementar la calidad de vida de los ciudadanos.

2.3.3.1. El alma del sistema: la Información *(«Big Data»)*

Es innegable que el incremento de la comunicación digital, móvil y entre los objetos a través de sensores, la interacción del visitante con el destino a través de las redes sociales y mediante el uso intensivo de las tecnologías, genera un volumen de datos que es captado en tiempo real. Estos macrodatos son almacenados, analizados y gestionados para su óptimo aprovechamiento por las tecnologías Big Data. Algunos ejemplos de la implementación y puesta en marcha de las tecnologías son:

- La nueva versión del portal oficial de España (2013), Spain.info, se ha diseñado para tener un mejor posicionamiento semántico, incluir la oferta

privada y obtener amplia información sobre qué buscan los turistas y cómo piensan los potenciales turistas.

- Sistema de sensores. Necesario para captar de manera anónima la información sobre el comportamiento del turista para la futura toma de decisiones.
- App móviles. En lo relativo a las aplicaciones móviles, SEGITTUR ha desarrollado una plataforma de generación de aplicaciones para dispositivos móviles denominada «Spain in Apps». Ya ha lanzado una serie de 10 apps para monitorear de igual manera el comportamiento de los turistas.
- Escucha activa. El uso de redes como Twitter, Facebook, Instagram entre otros se está empleando para saber lo que los turistas dicen de los destinos turísticos inteligentes. Información valiosa para toda mejora.

En resumen, el conocimiento de lo que piensa y quiere el turista permite tomar medidas para corregir lo malo y potenciar lo bueno. Por lo que se puede decir, potencializar el turismo en cualquier país es una acertada decisión que trae consigo un cúmulo de beneficios para los gobiernos, los residentes y sobre todo para las personas que por placer recorren un lugar.

2.3.4. Tercer caso. Un nuevo perfil: El periodista de Datos

El Open Data y el Big Data, así como las redes sociales, transformaron el trabajo en los medios de comunicación en un mundo informativamente globalizado. Esta transformación permitió el surgimiento de nuevos perfiles que pluralizan las expectativas laborales y exploran nuevas fórmulas para contar historias. Es el caso del periodista de datos.

El periodista de datos, en términos generales es una persona capaz de tratar y analizar grandes bases de datos para generar historias útiles para los ciudadanos. El internet y las nuevas herramientas informáticas maximizan la cantidad de fuentes de información. El colosal volumen de datos (Big Data) almacenados y disponibles

en la red se convierten en su materia prima, dónde busca, selecciona, procesa, analiza y compara para finalmente hacer una publicación.

En una sociedad en la que aumenta la información y la necesidad de transparencia, es propio del periodista utilizar y explicar los datos. Al mismo tiempo, el componente informático es parte esencial de la producción de noticias. Este nuevo esquema de trabajo lleva al periodista de los datos a considerar lo siguiente cada vez que desee emprender una tarea:

- Recopilación de datos e información. Aunque parezca evidente no hay que olvidar que un texto de calidad debe aportar datos. Para ello se pueden emplear técnicas de *scraping* para obtener información de sitios web.
- Limpieza y limpiado de la información. Se puede utilizar Google Refine, que permite elaborar porcentajes o patrones.
- Contextualización y combinación. Implica la búsqueda de antecedentes y la información contextual.
- Sería el caso de la combinación de datos con diferentes formatos de geolocalización.
- Comunicación (Visualización de datos-Infografía). Con la irrupción del Big Data y Open Data se ha logrado que los gráficos se basen en grandes cantidades de datos ofreciendo información de calidad de forma atractiva e interactiva.
- Esto se puede hacer con todo tipo de herramientas y aplicaciones (Tableau, CartoDB, Google Fusion, Many Eyes, etc.).

En resumen, el volumen, variabilidad y velocidad de los datos afecta directamente la tarea de los gobiernos, las empresas y a los investigadores y ratifica la importancia de la recopilación, análisis y representación de la información, técnicas de cálculo y cuantificación, un contexto que facilite la comprensión y sobre todo un profesional que sea el enlace entre medios de comunicación, tecnología y sociedad.

En el ámbito del periodismo de datos figuran personajes como Adrián Holovaty, con su proyecto *Chicago Crime*, donde se extrae la importancia del uso de base de datos, así como su aplicación a la visualización de información para mejorar el acceso a la información periodística. García y Catalina (2018) recopilan los siguientes casos:

- Un equipo de periodistas del diario argentino La Nación, genero una base de datos sobre los subsidios a colectivos y compañías desde el año 2006. El trabajo consistió en la disposición pública de datos, la sistematización del tratamiento de la información, y la proyección con estadísticas oficiales, dando como resultado un trabajo de interés periodístico.
- En el caso de la publicación brasileña Gazeta do Povo, através de la aplicación “Retratos Paraná”, se hizo posible el acceso a estadísticas de las ciudades del Estado, donde se difundió información sobre el desarrollo del Estado a partir de indicadores sobre la sociedad, la economía, la política, el medio ambiente, la educación y cultura.
- En otro caso, el Texas Tribune, puso a disposición del público el salario de los empleados públicos del Estado, permitiendo consultar, a partir del nombre, la agencia empleadora o la función.
- Mientras que el Chicago Tribune, publicó el reporte titulado “Illinois School Report Cards”, que analizaba las escuelas públicas en Illinois, con un claro componente visual que permite conocer cada uno de los centros.
- El Programa de Periodismo de Investigación de la Universidad de Berkeley-California, la revista Mother Jones desarrollaron la propuesta “Terrorists for the FBI”, que intentaba profundizar el papel del FBI en las redes sociales, donde se pudieron encontrar patrones de comportamiento.

En años más recientes, Stone (2014), realizó una extraordinaria recopilación e aplicación del Big data en los medios y junto a los diferentes formatos de visualización, se empleó para diferentes escenarios:

- Segmentación de la audiencia. La personalización de los contenidos periodísticos, así como de los anuncios de acuerdo a los intereses de los lectores.
- Análisis de comentarios, análisis de la calidad de los productos y servicios, así como análisis de los intereses de la audiencia en base al comportamiento de navegación del usuario con la finalidad de mejorar el trabajo de los periodistas.
- Seguimiento de noticias en tiempo real. Mediante técnicas de minería de datos, y tomando en consideración los tuits, se puede generar un sistema de alerta temprana de noticias. Con el uso de algoritmos especializados se identifican, clasifican y alertan a los clientes de información clave en tiempo real. (García y Catalina, 2018).

Figura 4. Herramientas del Periodista de datos
Fuente: (IIPJM, 2016)

2.3.5. Cuarto caso. Una campaña política exitosa: Barack Obama

Vender un producto o idea implica encontrar la forma de darlo a conocer. El saber el cómo, dónde, cuándo y a quien comunicárselo es todo un arte que comúnmente se conoce como Mercadotecnia (Marketing) en donde se definen las

estrategias y procesos adecuados para lograr dicho objetivo. (Castro Martínez, 2012)

En lo que respecta a las campañas políticas todos los esfuerzos y estrategias de mercadeo están enfocados en las personas que aspiran a un cargo público. Se busca darlos a conocer y comunicar a los votantes una imagen de ellos. El objetivo, convencer al mayor número de personas para que otorguen su voto al candidato promocionado y con ello asegurar su nominación y elección.

Para tener mayores probabilidades de éxito en este tipo de campañas, se debe tomar en consideración las siguientes variables:

- **Posicionamiento**. Se busca dar una imagen propia y autentica al candidato para que se diferencie del resto.
- **Segmentación del electorado**. Una vez realizado el primer sondeo, se obtiene información valiosa acerca de los votantes. A partir de ello, se procede a la segmentación y la selección del "público objetivo" hacia el cual van dirigidas todas las estrategias del marketing.
- **Selección de medios**. Es de suma importancia conocer los medios que son más frecuentados por el público objetivo para poder llegar a ellos.
- **Mensaje**. Debe ser claro y conciso.
- **El candidato**. El candidato es el "producto que se desea vender", por lo tanto se debe estudiar detenidamente la imagen que se desea proyectar y empezar a construirla cuidadosamente. Dentro de estos aspectos, es importante su vestimenta, manera de hablar, lenguaje corporal, su pasado, entre otros. La más mínima falla puede provocar un impacto negativo que podría ser causa de fracaso (Ancin, 2018).

En la actualidad, las ventas y la publicidad tienen su auge a través de la "mobile communication" (comunicación móvil). Cada año la inversión en publicidad y medios de comunicación aumenta. La información generada a partir del uso de

los medios representa una ganancia o pérdida de adeptos según el criterio con el que sea juzgada la aparición.

Cumpliendo con la máxima “si no sales en los medios, no existes”, los candidatos hacen lo necesario para aparecer el mayor número de veces. Para ello, es inminente la necesidad de recursos por lo que una estrategia innovadora es la clave para hacerse de adeptos y por lo tanto de “donaciones”.

En el 2008 Estados Unidos eligió a su primer presidente afroamericano, el demócrata Barack Obama quien es sin duda, el mejor ejemplo quizá a nivel internacional de lo que el marketing, la Internet, el correo electrónico, los celulares, los blogs, entre otros, pueden hacer para lograr la empatía con la sociedad y lograr adeptos, concibiendo una nueva forma de hacer política.

Figura 5. Barack Hussein Obama II.
Fuente (Instagram, 2018)

Barack Hussein Obama II, es hijo de madre estadunidense y padre keniano, miembro del partido demócrata, fue senador por Illinois en 2004, convirtiéndose así en el quinto afroamericano en la historia de Estados Unidos en ocupar esta posición. Obama, se convirtió en candidato a la presidencia el 3 de junio de 2008 al vencer a la también senadora Hillary Diane Rodham Clinton.

Obama se supo “vender” como el candidato del cambio y como un icono de esperanza y progreso para la nación norteamericana. Supo reconocer la diversidad de los votantes y atendió estos rasgos como propios de la población afroamericana e hispana, no por sus orígenes sino por su condición de ciudadanos y fue así como

se abanderó con el principio de igualdad logrando finalmente el apoyo más notable por parte de este sector (Castro, 2012).

La estrategia de Obama a través de Internet se centró en sus sitios https://barackobama.com/, perfiles de Facebook, Twitter, MySpace, YouTube, Wikipedia, entre otros.

Figura 6. MySpace: MyBarackObama.com.
Fuente: (My Space, 2018)

Chris Hughes 25 años, cofundador de Facebook, se convirtió en director de organización interna y uno de los jugadores clave detrás de MyBO (my.barack obama.com). David Plouffe asesor de campañas electorales del Partido Demócrata de los Estados Unidos, mencionó que “La tecnología siempre se ha utilizado como una red para capturar personas en una campaña por una causa, pero no para organizarla” (Hughes y otros, 2010).

Un ejemplo de la eficacia de MyBO fue el programa vecino a vecino, este fue lanzado en Septiembre de 2008 y permitió a usuarios registrados, ver una lista de votantes indecisos, los cuales era necesario ser llamados o visitarlos de puerta en puerta. Los voluntarios fueron emparejados con ellos de acuerdo a ciertas características o perfiles como profesión, edad, idioma, entre otros. Los voluntarios contaban con acceso a un script personalizado e interfaces fáciles para informar los resultados de sus esfuerzos a la campaña. Los voluntarios usaron la herramienta

para hacer 8 millones de llamadas en lugar de ir de puerta en puerta para registrar a los nuevos votantes (Hughes y otros, 2010).

Otra emotiva estrategia fue la “Cena con Barack”. Comúnmente, las cenas de recaudación de fondos permiten a los donantes de cantidades altas comprar el acceso. Pero la campaña de Obama hizo lo contrario, seleccionó a cuatro donantes de cualquier cantidad que habían compartido sus historias para encontrarse con Obama en un entorno de cena y discutir sus problemas. Los eventos fueron transmitidos en YouTube y los sitios web de la campaña. Fueron las poderosas historias de la gente común las que lograron la cuantiosa recaudación.

Stephen Geer se unió lanzando el programa de correo electrónico en mayo del 2007. El equipo de correo electrónico tenía tres objetivos: mensaje, movilización y dinero. En términos de movilización, su palabra triple mantra era: respetar, empoderar e incluir. El equipo de correo electrónico trabajó para identificar y aprovechar los seguidores existentes que se habían registrado en MyBO y organizaron eventos y fiestas. Adicionalmente la campaña desarrolló más de 7,000 correos electrónicos personalizados adaptados a cada persona objetivo (Hughes y otros, 2010).

También se lanzó el servicio de mensajes cortos (SMS) de texto de la campaña en mayo de 2007. El programa de mensajes de texto comenzó con un servicio básico de SMS, que permitía a los usuarios de teléfonos móviles enviarse mensajes cortos de texto. Con el tiempo, el programa de mensajes de texto desarrolló tonos de llamada y fondos de pantalla, así como aplicaciones de iPhone y video. Los partidarios enviaban un mensaje de texto con la palabra "HOPE" para suscribirse a los mensajes de texto de la campaña. En general, la campaña envió entre 5 y 20 mensajes de texto dirigidos por mes (Hughes y otros, 2010).

Se planeó una manifestación en Nueva York siguiendo el éxito de una en Texas durante el primer mes de la campaña. Voluntarios ayudaron a organizar la reunión. El equipo envió correos electrónicos a los seguidores de todo el país. "Queríamos que todos se hicieran cargo del mitin y se sintieran responsables por su éxito ", dijo Rospars.

Durante los cuatro días previos a la elección, el equipo trabajó en varias cosas para apoyar a la campaña, Por ejemplo, si un partidario había dado un código postal u otra información, y visitó los sitios web justo antes la elección, las páginas principales enumeraban una línea de eventos cerca de ellos. También podían “adivinar” dónde vivió en función de su dirección IP, y de esta forma mostrarle un evento cercano a su ubicación. Otras acciones como asegurarse de que los votantes conocieran la ubicación de sus centros de votación era prioridad. Cualquiera que haya dado una dirección de correo electrónico recibió un recordatorio para votar junto con el sondeo de dirección y horario (Hughes y otros, 2010).

Además, en estados de batalla, el sitio web haría una lista de cinco personas que tenían el mismo lugar de votación y alentó al partidario para llamar o tocar sus puertas y "*llevarlos consigo*... Eso fue parte de nuestra estrategia de nunca dejar que la gente se sienta como si no hubiera algo más que ellos podrían hacer para ayudar ", dijo Rospars. El día de las elecciones, Twitter se utilizó para publicar números gratuitos y cadenas de mensajes de texto para encontrar lugares de votación, así como oportunidades de voluntariado. Después, Obama ganó, el millón de personas que habían estado recibiendo actualizaciones de texto y anuncios recibieron un mensaje final: "Todo esto sucedió por ti. Gracias, Barack".

En otras de las acciones que causaron gran impacto se encuentra el video convertido en himno “Yes we can”, canción compuesta por el rapero Will.I.Am de los Black Eyed Peas. Además de ello, se hizo partícipes a otros cantantes, actores, políticos, activistas e influencers, por citar sólo algunos: Leonardo Di Caprio, Jennifer Aniston, Eva Longoria, Pearl Jam, Bruce Springsteen, la familia Kennedy, Oprah Winfrey, el ex presidente Jimmy Carter, entre otros, que le redituaron en la legitimación de su campaña y su proyecto.

La campaña fue dinámica (Castro, 2012), elaborada de tal forma que se adaptara a los momentos cambiantes de la elección entre las que se destacan un conjunto de acciones que se listan a continuación:

- 750 millones de dólares recaudados en poco menos de dos años.

- Recopilaron 13 millones de correos electrónicos de personas interesadas en recibir información directamente del equipo Obama.
- 4 millones de personas se inscribieron para recibir mensajes SMS de la campaña.
- 2 millones de personas crearon su propia web a través de la plataforma de participación on-line MyBarackObama.com,
- 5 millones de “amigos” en otras páginas de redes sociales en la web (por ejemplo, Facebook, MySpace, entre otros.).

La comunicación a través del correo electrónico permitió a Barack Obama informar a sus seguidores de sus actividades, debates, cruzadas para recaudar fondos, sobre todo porque el marketing por e-mail tiene múltiples ventajas: el costo es bajo, es instantáneo, interactivo y se abarca a un gran número de contactos. Facebook y Twitter le permitieron compartir más, interactuar con los cibernautas además de enviar información siempre novedosa y actualizada.

Es importante mencionar y no dejar fuera del radar que más allá de la estrategia en medios, la campaña que diseñaron para el demócrata Barack Obama no fue únicamente un fondo político o social, sino que retomaron la crisis económica, la guerra, la desigualdad, el racismo, la migración, la seguridad social y el papel de la mujer. En este punto, Michelle Obama con su feminidad y fortaleza fue una pieza clave. Su presencia fue concebida como el cúmulo de valores familiares y de pareja, dio el empuje a su esposo, no al caminar detrás, sino al lado (Castro, 2012).

La unificación del candidato y su mensaje; una estrategia simple y disciplinada, la integración de las nuevas tecnologías al proyecto político, la participación masiva del segmento joven de la sociedad y la personalidad del candidato dieron la presidencia de los Estados Unidos de Norteamérica a Barack Obama en 2008.

Figura 7. Triunfo político de Barack Obama- 2008.
Fuente**:** (El universal, 2012)

Obama, no fue el primero en utilizar la tecnología, pero sí fue el primero en utilizarla como tela de fondo de su campaña. Supo recopilar, procesar, canalizar y aprovechar toda la información generada, impulsando una nueva generación de activistas tecnológicos y una nueva forma de hacer campaña. De esta forma Obama se convierte en un revolucionario en el uso de las nuevas tecnologías aplicadas a la política. El marketing político nunca volverá a ser igual; tal vez ahora el uso de la tecnología móvil se convierta en una obligación más que en una necesidad cuando de electores se trata (Castro, 2012), (Aaker y Chang, 2010).

2.3.6. Quinto caso. BBVA Bancomer y SECTUR

El fenómeno del Big Data ha llamado la atención de las más prestigiadas marcas en todas las esferas de la sociedad. Tal es el caso del corporativo bancario BBVA Bancomer quien desarrollo en el año 2016 “Big Data y Turismo” en colaboración con la Secretaría de Turismo de México (SECTUR). El objetivo de este convenio es compartir la información sobre el comportamiento comercial durante un año de los turistas nacionales y extranjeros, en 12 análisis estadísticos,

que comprenden los 111 Pueblos Mágicos y los principales corredores turísticos (BBVA, 2016).

Una de las tareas de la tecnología es hacer que convenios como "Big Data y Turismo" permitan a las dependencias públicas tener información que les sea útil para generar propuestas de valor integral, ofertas acordes con las necesidades de los turistas nacionales y extranjeros, e impulsar el crecimiento de esta actividad relevante en el país. El proyecto "Big Data y Turismo" permitió conocer el comportamiento comercial de 86 millones de usuarios de tarjetas bancarias nacionales y extranjeras.

En un año se registraron 1,5 mil millones de transacciones, a través de los medios de pago de BBVA de clientes y no clientes de la institución, con un monto de 813 mil millones de pesos. Con el proyecto "Big Data y Turismo" se identificaron cuáles son las compras habituales por el mercado local y las que se hacen como visitante, a partir de una innovadora definición de entorno habitual de los usuarios de tarjeta.

Con el estudio realizado a través de la colaboración de BBVA Data&Analytics y SECTUR para extraer y analizar grandes cantidades de datos con el objetivo de describir la actividad de un territorio permitió conocer, por ejemplo, que en Cozumel el gasto principal realizado por mexicanos se realizó en restaurantes y viajes. Algunos otros datos que arrojo el análisis incluyen:

- Los turistas nacionales realizan el pago con tarjeta, principalmente para viajes o excursiones, mientras que los turistas internacionales principalmente en actividades de entretenimiento.
- En la mayoría de los Pueblos Mágicos analizados, los turistas realizaron el pago con tarjeta en restaurantes, alimentos, entretenimiento y viajes.
- En cuanto al origen de los turistas internacionales resalto que el mayor porcentaje de gasto es realizado por turistas estadounidenses, mientras que Argentina ocupa el segundo lugar

Además de profundizar en los patrones del gasto de los turistas nacionales y extranjeros, el análisis identifica las zonas y servicios de interés para los visitantes, siempre en completo apego a las normas de confidencialidad del sector financiero (BBVA, 2016).

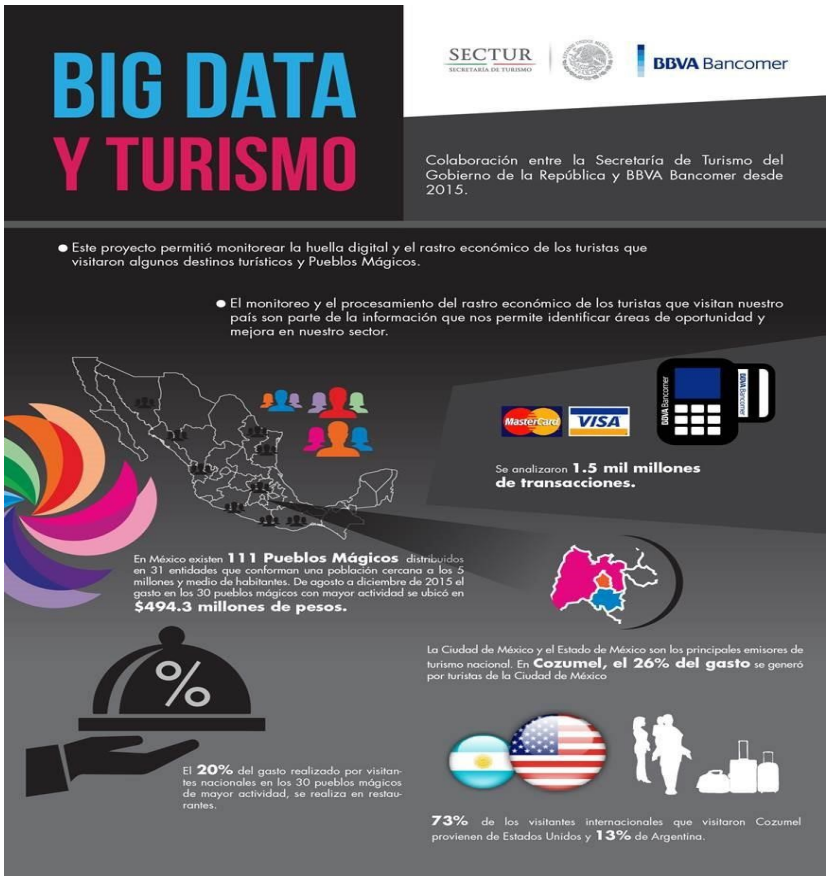


Figura 8. Big Data y Turismo.
Fuente: (BBVA, 2016)

2.3.7. Riesgos del Big Data

A pesar de la creciente popularidad y aceptación de Big Data es necesario mantener una postura abierta y analizar la contraparte de este fenómeno. Para tal efecto, se analizan tres de los riesgos más significativos que lo rodean.

2.3.7.1. Privacidad de las personas

Para el derecho, los macrodatos representan una tarea por demás importante ya que deben ser objeto de protección mediante regulaciones ***ex profeso*** al relacionarse con el derecho fundamental a la privacidad.

Figura 9. Big Data y el derecho fundamental a la privacidad.
Fuente: (OH Strategy, 2017)

Big Data permiten tratar cantidades masivas de datos provenientes de fuentes dispares, con el objetivo de poder otorgarles una utilidad. Sin embargo, también va acompañada de riesgos. Quizás uno de los más relevantes sea que este análisis masivo de datos descansa sobre la privacidad de las personas.

Los algoritmos con que se indagan los datos no siempre son neutrales. En 2016, *Pro Publica,* el medio de investigación periodística estadounidense, analizó los algoritmos utilizados por el sistema judicial para predecir los casos de reincidencia delictiva. Encontró que los algoritmos habían sido creados con una tendencia racista.

La recolección de datos que realizan empresas como Google, Facebook, Amazon, Apple, así como las compañías de telecomunicaciones, no se llevan a cabo mediante contratos transparentes sobre su utilización y protección. Esta situación se observa cuando los usuarios aceptan contratos de uso sin leer las condiciones que esto implica.

Las redes sociales son empresas que almacenan datos. Facebook, por ejemplo, reduce las expresiones a sus siete iconos; Twitter, por su lado, permite frases de hasta 140 caracteres, por lo que se puede decir que son decisiones preestablecidas para cuantificar la sociabilidad. Así mismo, estas empresas no

comparten con los gobiernos toda la información que guardan por su supuesto compromiso con la libertad de expresión, y mucho menos con la comunidad científica (Meneses, 2018). Evidencias como éstas ponen en claro que no se puede aceptar a ojos cerrados y sin cuestionamientos la propuesta de que los grandes datos nos acercan a un universo donde el comportamiento humano pueda construir una “mejor sociedad”.

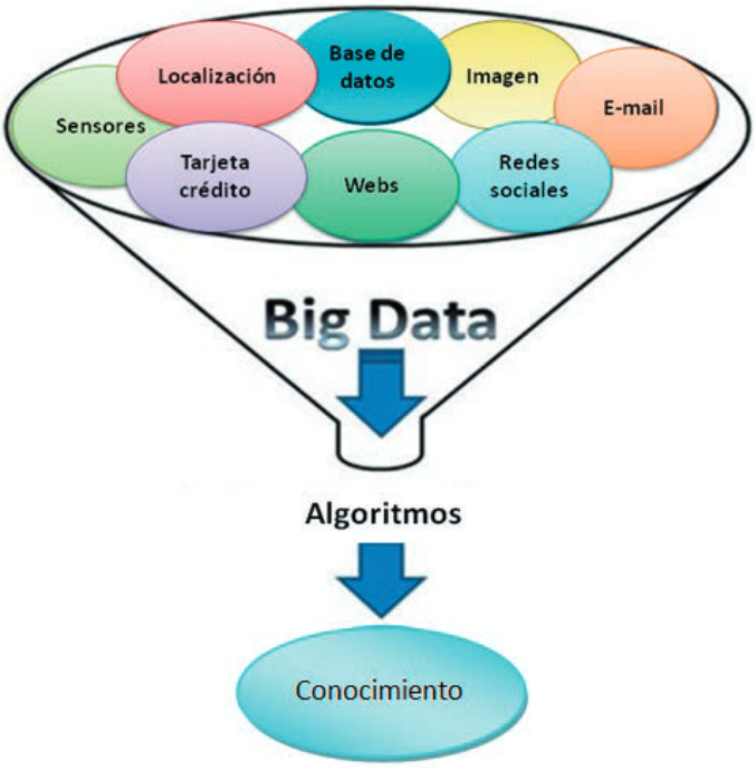


Figura 10. Big Data y el derecho fundamental a la privacidad.
Fuente: (Gil, 2016)

Otro ejemplo de la falta de parámetros éticos lo ofrece el reciente caso protagonizado por Facebook y la Universidad de Cornell, quienes manipularon el algoritmo de la red social para observar las reacciones psicológicas de los usuarios ante informaciones negativas. Los resultados revelaron que las expresiones de otros influyen en nuestras emociones. Para obtener esta evidencia Facebook introdujo en el perfil de 700,000 personas determinados contenidos, lo que causó una gran controversia (Meneses, 2018).

Para analizar Facebook, Instagram, YouTube o Twitter, se debe partir reconociendo no sólo el impacto social de dichas redes, sino también los limitantes

para expresar el sentir de los usuarios al postear y compartir una información; así como los atributos de tipo geopolítico, económico y cultural (Meneses, 2018). Además del riesgo de atentar contra la privacidad de las personas, Big Data trae consigo otros de igual importancia que merecen ser puestos sobre la mesa.

2.3.7.2. Conclusiones erróneas que nadie revisa: errores por azar y por confusión

Uno de los propósitos principales del Big Data es descubrir patrones para realizar predicciones futuras. Para ello, es importante encontrar una verdadera relación entre las variables analizadas diferenciando entre la causalidad (causa-efecto) y la casualidad (azar o confusión).

En la estadística, la correlación es el grado de relación que existe entre dos variables, es decir, se dice que dos variables están correlacionadas cuando el aumento o disminución de una provoca un cambio en la otra. Cuando dos variables se correlacionan, es posible que también presenten una relación de **causalidad,** de igual forma existe correlaciones donde no hay causalidad sino casualidad. Por lo que habrá que estudiar si existe un patrón verdadero entre ambas variables o si por el contrario se trata de una mera coincidencia. Las correlaciones por casualidad son falsas y se les conoce como espuria. Las hay de dos tipos: error por azar y error por confusión. Se Analizan cada una de ellas para tener una visión más amplia del por qué es importante tenerlo en consideración al relacionarlo con Big Data (Gil, 2016).

En primer lugar, **indagar en el error por azar**. El estadístico Stanley Young ha denominado «la tragedia de los grandes conjuntos de datos» a lo siguiente: entre más variables se estudian en un gran conjunto de datos, más probabilidades habrá de encontrar relaciones falsas sin ningún significado real, aunque ambas presenten una fuerte relación estadística. Esto implica que, de interpretarse de forma errónea, el analista puede terminar siendo engañado por los datos.

Se toma como ejemplo el caso de Ricardo Galli, doctor en informática y activista del software libre, que llevó a cabo un experimento donde se tienen los

siguientes datos de evolución de cinco variables económicas de los últimos años, (Figura 11) donde se puede ver claramente que esta pequeña cantidad de datos no muestra ninguna correlación entre variables (Gil, 2016).

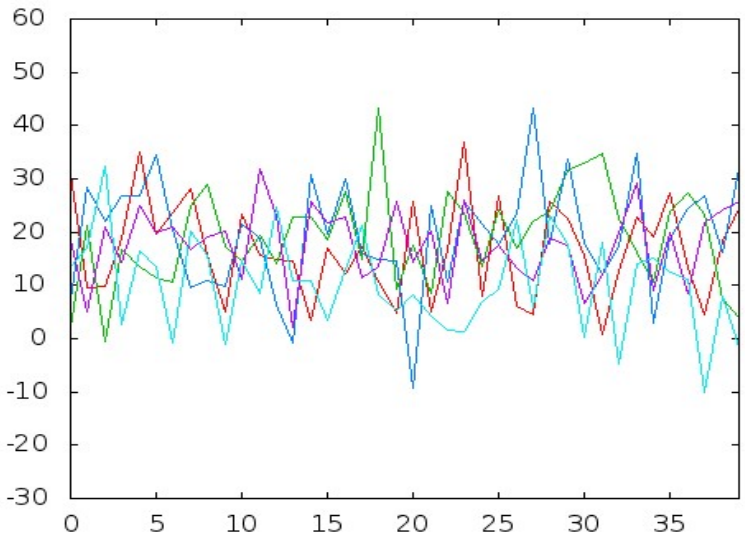


Figura 11. Experimento de Ricardo Galli- variables 1 a la 5.
Fuente: (Gil, 2016)

Pero ahora se supone que, en lugar de cinco variables, se puede analizar mil variables (algo similar a lo que ocurriría con el Big Data). El gráfico sería algo similar a lo mostrado en la figura 12.

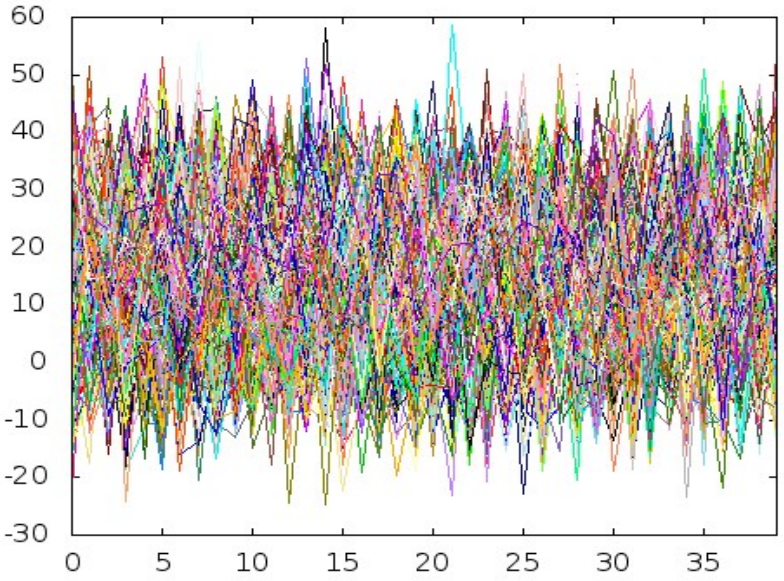


Figura 12. Experimento de Ricardo Galli- con 1,000 variables.
Fuente: (Gil, 2016)

Se debe imaginar las consecuencias que una mala interpretación de variables económicas puede tener en la toma de decisiones de un país. Así pues, entre más variables analicemos estamos expuestos a encontrar más relaciones espurias, que si no son cuestionadas, nos harán caer en conclusiones erróneas, mismas que por la confianza ciega que parece haber por los datos, pueden ser aceptadas para tomar decisiones sobre las personas, empresas o países.

Enseguida, se analiza el llamado **error por confusión** con el siguiente ejemplo: En 1952 el matemático polaco J. Neyman observó que en diversas regiones rurales existía una relación directa positiva entre el número de cigüeñas que habitaban en los pueblos y el número de nacimientos en esos pueblos. Es decir, la evolución de la población de las cigüeñas y de los habitantes señalaba que aquellas regiones que tenían una mayor población de cigüeñas, también presentaban una mayor tasa de natalidad (Gil, 2016).

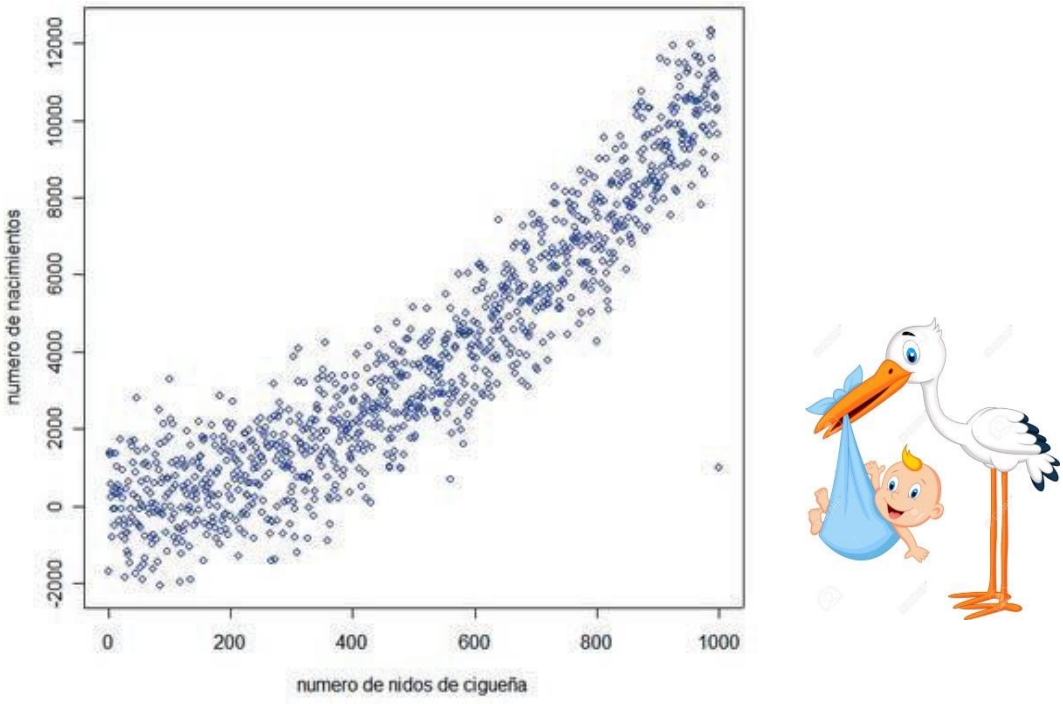


Figura 13. Tasa de natalidad y población de cigüeñas.
Fuente: (Gil, 2016)

En base al resultado anterior, se podría deducir lo siguiente: ¿Será ésta la demostración de que a los niños los trae la cigüeña? Probablemente esto podría

demostrar la prueba de ello, pero es bien sabido que realmente no es así. De esta forma, ambas variables no tienen una relación de causa-efecto. La razón de esta relación es que en ambos casos dependen de una tercera variable “la calidad de las cosechas”. Por lo tanto, en los años, con más sol, lluvia y alimentos, las cigüeñas criaban más, al igual que los habitantes de dichas regiones.

Lo que sucede en esta supuesta correlación, es lo que se conoce **como error por confusión**, donde, dos eventos que son independientes, pero que parecen estar relacionados sin que la relación sea real; en este caso la población de cigüeñas y la tasa de natalidad (Gil, 2016). De esta forma se observa la relación falsa entre las 2 variables, y el factor desconocido con relaciones ocultas pero reales (Figura 14).

Figura 14. Error por confusión.
Fuente: (Gil, 2016)

En conclusión, la correlación no implica causalidad, puede ser simple azar o una tercera variable que esté influyendo en la relación, creando una falsa apariencia de causa-efecto. Siendo de mayor escrutinio en el contexto del Big Data, ya que este hecho se agudiza, porque el número de correlaciones que podemos encontrar aumenta drásticamente. Debido a esto, es necesario ser críticos con los resultados, buscar la causa real de la relación entre variables y exigir rigor científico en los resultados que arrojan los datos.

2.3.7.3. La toma de decisiones automatizadas.

La evolución de la ciencia de datos permite hacer el siguiente cuestionamiento, ¿cuándo hace falta la intervención de una persona para supervisar los resultados y conclusiones obtenidos de forma automática antes de una toma de decisiones?.

Existen decisiones automatizadas sin la intervención del ser humano como pueden ser el otorgamiento de un préstamo o en un diagnóstico médico. Se debe considerar que existe participación humana cuando se crean los algoritmos que analizarán los datos para la toma de decisiones, pero en muchas ocasiones no se vuelve a realizar un control humano para comprobar esa decisión. La investigación básica y la intuición están siendo reemplazadas por fórmulas algorítmicas. Es decir La investigación básica y la intuición están siendo reemplazadas por algoritmos.

Antes este escenario, surge la interrogante: Se debe de confiar ciegamente en los algoritmos que hacen que en muchas ocasiones las empresas tomen decisiones sobre nosotros sin que podamos saber por qué las han tomado. Muchos consideran que el sector del marketing es idóneo para probar y corregir las nuevas herramientas matemáticas y tecnológicas, pues tiene pocos riesgos y muchos beneficios. Consideremos que un error en marketing puede provocar que un consumidor vea un anuncio erróneo, y un acierto lograr el aumento de las ventas (Gil, 2016).

Sin embargo, si está técnica es llevada a otros sectores como el bancario, el asegurador, y sobre todo el sanitario, donde la situación seria preocupante por la magnitud del riego. Así pues, la duda de que tan necesaria es la intervención humana está latente con mayor intensidad en estos sectores.

A la fecha, las organizaciones que basan sus decisiones en la Ciencia de los Datos, pocas veces revisan los resultados, esto debido a que partidarios de la automatización afirman que, es precisamente la promesa de un resultado basado en los datos y libre del sesgo humano, una de las principales virtudes de los algoritmos. De este modo el objetivo a lograr por los desarrolladores será mejorar

la calidad de los algoritmos para obtener una mejor clasificación del individuo. En Europa, conscientes de esta problemática, introdujeron una disposición la cual prohíbe tomar decisiones trascendentales para una persona sobre la única base del análisis automático de datos (Gil, 2016).

2.3.8. Análisis personal

Desde la utilización del oro, el metal precioso, como base de las economías del mundo, se observó que su impacto en la vida histórica, uso y consecuencias estaban determinados por las particularidades de cada país. Mientras que a Europa la monetización le sirvió para su expansión comercial y crecimiento económico, en la India, su acumulación era necesaria para cancelación de impuestos, crecimiento fiscal del Estado y recaudación de bienes en caso de guerra.

Expresando una opinión personal, aventuraré a decir que, en el actual siglo XXI, la **información** se postula como el nuevo ORO, valorado, resguardado y disputado constantemente para luego ser usado en distintos menesteres de la vida humana.

Siguiendo esa misma línea de ideas, es de vital importancia que se tome conciencia de qué el uso de todo “metal precioso” demanda cualidades como responsabilidad, ética, profesionalismo, objetividad, aspiraciones al bien común e interés por defender la integridad de las personas involucradas en el proceso.

La ciencia, en particular las ciencias sociales, ha mostrado un verídico interés sobre la contribución de la tecnología y la información en la evolución de la sociedad, percibiendo un nuevo tipo, “la sociedad tecnológica”.

La comunidad científica y profesional al igual que las potencias mundiales, tiene en sus manos la responsabilidad de encausar todo conocimiento hacia el bien común. No obstante, existen como en todos los acontecimientos humanos, infinidad de opiniones muchas a favor, otras en contra y el resto neutrales que entorpecen las buenas acciones.

Se citan algunas ventajas:

- La correcta utilización de la técnica y la tecnología de Big Data abren el panorama conceptual de cualquier situación y permite obtener información de calidad que sirve como soporte ante la toma de decisiones.
- Big Data, se suma a los ya existentes métodos de análisis científicos aportando entre otros aspectos rapidez, eficacia y exactitud.
- Ofrece la oportunidad de formar grupos de trabajo multidisciplinarios logrando que los resultados enriquezcan aún más a los involucrados.

Se citan algunas desventajas:

- Se encuentra en plena construcción conceptual, lo que impide ver con claridad sus características, alcances y limitaciones.
- Aunque están implícitos en las actividades diarias, las organizaciones públicas y privadas, en su gran mayoría, carecen de la teoría que les proporcione una guía, del hardware y software adecuados, pero, sobre todo de profesionales capacitados para tratar los macrodatos.
- Otro de los puntos menos favorables es el que trata sobre la privacidad y protección de las personas que dan origen a los datos. A la fecha, no existen políticas efectivas orientadas a la seguridad de la información y del individuo.

Quizá el problema más significativo al que Big Data puede enfrentarse es la diferencia cultural, política, social y económica de los países. Situación que puede impedir su ideal desarrollo.

2.4. Conclusiones

La tecnología evoluciona. Desde sus orígenes que datan de finales de la Segunda Guerra Mundial (1935-1945) hasta hoy, ha ido cuesta arriba. Siempre proponiendo e innovando. Su auge, ha desencadenado el Big Data nombrado también macrodatos o información masiva. Un porcentaje de estos se encuentran accesibles y disponibles; son diversos y de rápida creación, sin embargo, para poder administrarlos y analizarlos es necesario valerse de técnicas y tecnologías específicas, ya que las herramientas y procesos tradicionales están limitados para esa tarea.

Hadoop, MapReduce, Casandra, Hbase entre otros, son tecnologías que procesan y analizan la información masiva. Existen también, tecnologías que ayudan a la fácil comprensión de los resultados, estas son catalogadas como herramientas de visualización.

En la mayoría de la literatura encontrada se hace referencia no solo al aspecto tecnológico del Big Data que lo define como - un gran volumen de datos con características como la variabilidad y velocidad, los cuales pueden ser almacenados, procesados, analizados y visualizados con la finalidad de tomar decisiones inteligentes. Sino también a su naturaleza social, política y cultural.

Las opiniones se dividen, mientras unos piensan que es el inicio de un crecimiento exponencial en los sectores social y económico, otros lo ven como un medio para el control y dominio de las masas. Es indudable que Big Data está detrás de muchos sucesos sociales, políticos, económicos y culturales de los últimos tiempos.

Pese a eso, el autor se suma a los científicos, empresarios, gobernantes y ciudadanía en general que ve al Big Data con ojos optimistas. Con el convencimiento de que este, al igual que los otros inventos creados por el hombre, lo llevara a su óptimo desarrollo siempre y cuando se tenga presente lo mencionado con antelación, el respeto a la privacidad, el derecho a la libre expresión y la autonomía al tomar decisiones y al llevar a cabo alguna actividad.

En lo referente a la escasez de profesionales en el área, es un aliciente pensar que, a través del presente trabajo, los jóvenes se inclinen hacia el estudio de la Ciencia de los Datos.

Referencias

Aaker, J. y Chang, V. (2010). Obama and the power of social media and technology. *The European Business Review*, 16-21.

Aenor. (s.f). Destinos Turísticos Inteligentes. Aenor [Figura] Recuperado de: https://www.aenor.com/certificacion/administracion-publica/destino-turistico-inteligente

Ancin, A. I. (2018). Análisis de los distintos tipos de campañas publicitarias y sus diferentes aplicaciones para lograr el top of mind de las marcas. Revista Caribeña de Ciencias Sociales.

Agudo, S. I. (2014) ¿Qué pasó en Facebook y Twitter durante 2013?. Principiantes en Social Media. [Figura]. Recuperado de: https://principiantesensocialmedia.com/2014/01/10/cambios-en-redes-sociales-durante-2013/

Ascanio, M. B. (2016). Los papeles de Panamá y sus implicaciones periodísticas y sociales. *Comunicación: estudios venezolanos de comunicación*, (174), 61-71.

BBVA (2016). BBVA muestra cómo el 'big data' puede potenciar el turismo en México. Recuperado de: https://www.bbva.com/es/bbva-muestra-big-data-puede-potenciar-turismo-mexico/

Castro, M. L. (2012). El marketing político en Estados Unidos: el caso Obama. *Norteamérica, 7*(1), 209-222.

Castro, R. A., González, S. J. S., y Callejas, C. M. (2012). Utilidad y funcionamiento de las bases de datos NoSQL. *Facultad de Ingeniería, 21*(33), 21-32.

de Avila, M. A. L., y Sánchez, S. G. (2015). Destinos turísticos inteligentes. *Economía industrial*, (395), 61-69.

El Universal (2012). Gana Obama. Periodico online El Universal. Recuperado de: http://archivo.eluniversal.com.mx/primera/40704.html

Ferrer-Sapena, A., y Sánchez-Pérez, E. (2013). Datos abiertos, big data: ¿Hacia dónde nos dirigimos? *Anuario ThinkEPI, 7*, 150-156.

García, J. A., y Catalina, G. B. (2018). Una perspectiva documental y bibliotecológica sobre el big data y el periodismo de datos. *Investigación Bibliotecológica: archivonomía, bibliotecología e información, 32*(74), 77-99.

Gartner (s,f) Big Data. *Gartner* Research and Advisory Company. Recuperado de: https://www.gartner.com/en/information-technology/glossary/big-data

Gil, E. (2016). Big data, privacidad y protección de datos. *Madrid: Agencia Estatal Boletín Oficial del Estado*.

Hernández, L. E., Duque, M. N., y Moreno, C. J. (2017). Big data: una exploración de investigaciones, tecnologías, y casos de aplicación. Tecnologías, 20(39).

Hernández, S. (2002). *Administración: Pensamiento, proceso, estrategia y vanguardia.* México: McGraw-Hill.

Hilbert, M. (2017). CV & BIO. MartinHilbert.net. Recuperado de: https://www.martinhilbert.net/tag/big-data/

Hughes, S. G. F., Allbright-Hannah, K., Goodstein, S., Grove, S., Zuckerberg, R., Sladden, C., & Bohnet, B. (2010). Obama and the power of social media and technology. *The European Business Review (May-June 2010)*, 16-21.

Magnani, E. (2017). Big data y política. El poder de los algoritmos. *Nueva Sociedad*(269).

IIPJM. (2015). Curso de Postgrado Internacional de Periodismo de Datos. Instituto Internacional de Periodismo José Martí. [Figura]. Recuperado de: https://periodismojosemarti.wordpress.com/2015/12/11/4877/

IIPJM. (2016). ¿Qué es el periodismo de datos?. Instituto Internacional de Periodismo José Martí. [Figura]. Recuperado de: https://periodismojosemarti.wordpress.com/2016/11/03/que-es-el-periodismo-de-datos/

Meneses, R. M. E. (2018). Grandes datos, grandes desafíos para las ciencias sociales. *Revista Mexicana de Sociología, 80*(2).

Merkactiva (2017). El cine, el mejor influencer. Merkactiva. Blog de Mercadotecnia. [Figura] Recuperado de: http://www.merkactiva.com/blog/cine-mejor-influencer/

Sánchez, M. L. A. (2005). *Informática*. Pearson Educación. ISBN, 9789702605393

SEGITTUR (2014). Destino Turístico Inteligente. Sociedad Mercantil Estatal para la Gestión de la Innovación y las Tecnologías Turísticas. Recuperado de: https://www.segittur.es/es/DTI/

Stone, M. L. 2014. “Big data for Media”. Reuters Institute for the Study of Journalism, 1-31. UK: University of Oxford. http://www.bigdatamedia.org/wp-content/uploads/2017/03/Big-Data-For-Media_2014-Stone.pdf

Capítulo 3

Herramientas Big Data

Jaime Alonso Chacón Soto
Tecnológico Nacional de México. Instituto Tecnológico de Durango
95040538@itdurango.edu.mx

Jeorgina Calzada Terrones
Tecnológico Nacional de México. Instituto Tecnológico de Durango
jcalzada@itdurango.edu.mx

3.1. Introducción

Con el paso del tiempo la tecnología ha ido creciendo y avanzando en todos los sectores conocidos, tanto ha sido su crecimiento que hoy en día existen un mayor número de dispositivos en el mundo que seres humanos. Dándose por consecuencia un incremento de datos generados por dichos dispositivos.

El término Big Data se refiere a la gran cantidad de datos generados a cada momento en todo el planeta por los dispositivos que se utilizan y la cual da un enfoque en el entendimiento y toma de decisiones al analizarlos de una manera correcta.

Para el buen uso de esta gran cantidad de datos Big Data ha determinado 5 características que deben poseer, ellos los denominan como las 5 V de Big Data esto para obtener un mejor entendimiento de los datos a analizar donde comprenden el Volumen (cantidad de datos), Velocidad (a la que se procesan ser

requiere su análisis), Variedad (datos estructurados, semi-estructurados y no estructurados, Veracidad (grado de confianza que se establece sobre los datos) y Valor (conocimiento e información útil analizado) (Universidad de Alcalá, s.f.)

Por estos motivos, es necesario conocer las herramientas existentes para el uso adecuado de los datos desde su recolección, almacenamiento, procesamiento y análisis. Así mismo, es necesario saber los diferentes tipos de fuentes de datos que existen; las estructura que los integran; los tiempos que son requeridos para ser procesados, si es en el momento mismo o a un tiempo posterior y la forma en que se presentan, descriptiva o gráfica.

Los datos provienen de distintas fuentes como redes sociales, visitas de páginas web, reproducción de videos, reproducción de música, generación de comentarios, correos electrónicos, aplicaciones, entre otros, también hay datos generados por máquinas a través de sensores, equipos biométricos, GPS entre algunos otros. Actualmente los dispositivos conectados a internet superan en número a la población mundial y cada día aumenta este número de dispositivos.

Big Data y Ciencia de los Datos, se están convirtiendo en una de las ramas más importantes de la computación hoy en día y una de las más rentables refiriéndose en la parte económica, esto a nivel mundial, ya que se utiliza en cualquier sector, desde las ciencias, diversión, cultura, finanzas, salud hasta ganar puestos políticos de países de primer mundo como lo fue en el caso de Estados Unidos de Norte América.

Por esto se pretende documentar las herramientas adecuadas para la recolección o ingesta, almacenamiento, procesamiento, análisis y visualización de datos. Se abordará acerca de dichas herramientas y se realizará la evaluación de algunas de ellas.

En la actualidad los datos se producen por diferentes fuentes como páginas web, reproducciones de música videos, redes sociales, aplicaciones de celulares, transacciones comerciales, así como las que se generan por máquinas ya sea sensores de humedad, temperatura, infrarrojos, entre otros, esto da lugar a que

exista un universo de datos que se pueden utilizar para conocer tendencias, hacer pronósticos y predicciones de una manera más efectiva y real.

Por estos motivos es necesario saber, aprender y comprender cuales son las mejores herramientas que existen para el manejo de Big Data. Se estima que, en el año 2020 estén más de 30 mil millones de dispositivos conectados a internet, generando una enorme cantidad de datos segundo a segundo la cual puede ser analizada con las herramientas adecuadas. (OBS Business School, s.f.).

En Internet se generan 4,1 millones de búsquedas en Google, se escriben 347.000 twists, se comparten 3,3 millones de actualizaciones en Facebook, se suben 38.000 fotos a Instagram, se visualizan 10 millones de anuncios, se suben más de 100 horas de vídeo a YouTube, se escuchan 32.000 horas de música en streaming, se envían 34,7 millones de mensajes instantáneos por Internet o se descargan 194.000 apps (OBS Business School, s.f.).

Las herramientas que se documentan, comprenden desde la ingesta hasta el análisis y visualización de datos, existen herramientas comerciales y gratuitas (las cuales son las más utilizadas).

El objetivo general, es dar a conocer las herramientas en el contexto Big Data y elaborar una propuesta asociada a esta tecnología.

De manera específica se busca:

- Realizar un análisis y propuesta de herramientas Big Data para su implementación que son utilizadas normalmente en la ingesta, almacenamiento, procesamiento y análisis.
- Identificar las herramientas de análisis gráfico de Big Data.
- Realizar la comparación de algunas de estas herramientas para conocer sus ventajas y desventajas.
- Conocer los pasos a seguir para la instalación de algunas herramientas.

Cada día más empresas, gobiernos, comercios y ciudades se están involucrando en el análisis de datos con más fuerza para determinar tendencias,

patrones, evitar crímenes, prevenir enfermedades, crear ciudades inteligentes, ser más productivos y mejorar la vida misma de las personas.

El análisis de datos cada día tiene mayor valor para las personas dedicadas al manejo de ellos. Por este motivo la revista Harvard Business Review en su artículo de Davenport & Patil, (2012) citan como la profesión de "data scientist" como "la más sexy del siglo XXI" (Davenport & Patil, 2012).

El mundo va hacia una evolución del Big Data. Su crecimiento impulsa la creación de nuevos empleos, servicios, estrategias e incluso infraestructura especializada capaz de gestionar mayores volúmenes de datos y simplificar su procesamiento y su análisis (Portafolio, 2019).

El impacto y beneficio de este capítulo, es comprender aspectos de Big Data, adentrarse a un tema el cual cada día empieza a ser más común para todos, se puede decir que es el siguiente paso que se está empezando a dar para el uso y manejo del internet ya que es con los avances de las herramientas, programación y de los dispositivos que se utilizan para el manejo de dichas herramientas se obtienen beneficios al instante y ayudan a comprender circunstancias o mejorar situaciones que hace años no se imaginaba.

3.2. Marco de referencia

El concepto de Big Data se aplica en los casos en los que la información recolectada no se captura, gestiona, analiza o procesar por las herramientas habituales y tradicionales. (Pérez Marqués, 2015). Se está hablando de una cantidad enorme de información con términos de medidas de almacenamiento como petabytes, exabytes, zettabytes, yettabytes, entre otros.

La figura 1, identifica los valores numéricos de unidades de medida en bytes y cuáles son considerados datos normales y cuáles Big Data y mucho más allá.

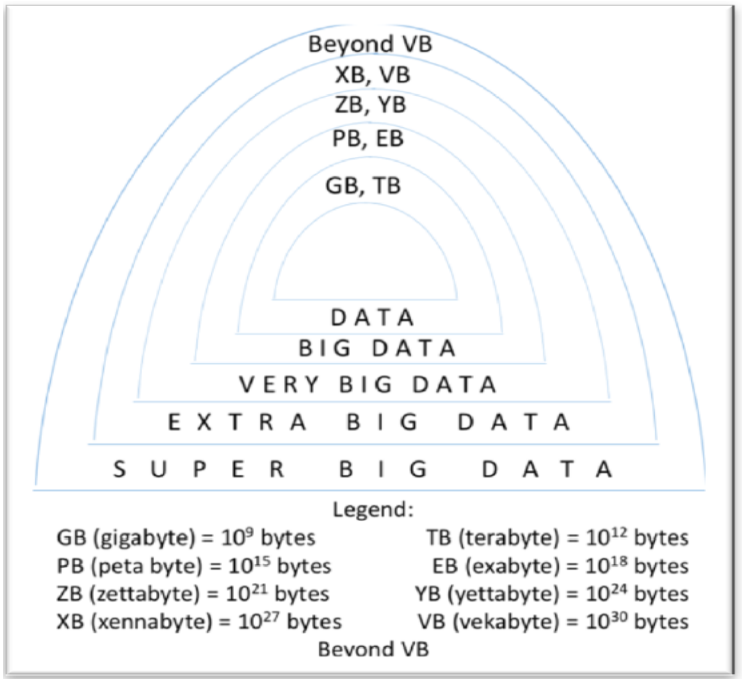


Figura 1. Big Data conforme al volumen de datos. (Sánchez Villaseñor, Herraientas, retos, oportunidades, seguridad y tendencias del Big Data, 2019)

Algunas características relacionadas al Big Data serían las siguientes:

La cantidad (volumen), en Big Data se habla de una ingesta de datos; la generación de datos (velocidad) en Big Data tiene que ver con recolectar al momento en lugar de al final del día; en cuanto a la estructura de datos (variedad) se identifica semiestructurados y no estructurados en lugar de solo estructurados; con lo referente a la fuente, Big Data es totalmente distribuido en lugar de centralizado; la integración en Big Data es más compleja que lo tradicional; con respecto al almacenamiento en Big Data se encuentra el paradigma NoSQL y sistemas de archivos distribuidos (HDFS) en lugar de únicamente SQL y bases de datos relacionales tradicionales y finalmente con respecto al acceso en Big Data se tiene que darse en tiempo real, es decir más rápido y eficiente que la usanza tradicional. (Sánchez Villaseñor, Herraientas, retos, oportunidades, seguridad y tendencias del Big Data, 2019).

A lo anterior hay que agregar además de que Big Data se debe entender perfectamente que los datos son cambiantes y muy dinámicos (variabilidad) y

deben ser confiables (veracidad), útiles (valor) y que puedan presentarse de maneras diversas para su mejor comprensión y entendimiento (visualización).

Sánchez Villaseñor, (2019) en su tesina (como se cita en *Salazar, 2016)* menciona que “Big Data hace referencia al tratamiento y análisis de enormes repositorios de datos cuyo tamaño está más allá de capturar, almacenar, administrar, analizar y procesar con herramientas de bases de datos o analíticas convencionales” (Salazar Argonza, 2016).

Gil, (2016) menciona que “el Big Data es el conjunto de tecnologías que permiten tratar cantidades masivas de datos provenientes de fuentes dispares, con el objetivo de poder otorgarles una utilidad que proporcione valor” (pág. 15).

¿Qué sucede en un minuto en términos de procesamiento de datos en internet?, es decir, ¿qué pasa en 60 segundos en el mundo online. La figura 2, identifica algunos valores cuantificables de cantidades de búsquedas, fotos subidas, horas de video, documentos consultados, mensajes, correos, entre muchas otras cosas. De aquí se desprende que se necesita y requiere distintas herramientas para el tratamiento de todos estos tipos de datos.

Figura 2. 60 segundos en el universo online (Twitter, 2018).

En términos de Big Data resulta importante hacer mención de las herramientas diseñadas para tratar la gran cantidad de datos que comprenden en

una arquitectura como es la recolección o ingesta de datos, almacenamiento, procesamiento y análisis de estos datos.

El ciclo de vida del Big Data puede comprenderse bajo ciertas fases principales generación y adquisición, almacenamiento, procesamiento y análisis de datos, en las cuales se tienen ciertos tipos de herramientas. La figura 3, identifica estas cuatro fases incluyendo herramientas de visualización de datos. En los siguientes apartados se describen algunas herramientas asociadas a cada fase.

Figura 3. Fases de Big Data incluyendo visualización (Sánchez Villaseñor, Herraientas, retos, oportunidades, seguridad y tendencias del Big Data, 2019).

3.2.1. Herramientas de ingesta

Las herramientas de ingestión en proyectos Big Data tienen que ver con las características para extraer datos de distintas fuentes, es decir los datos deben ser “ingestados”, lo que significa que son transferidos para procesamiento (Rayón, 2016).

3.2.1.1. Kafka

Es un proyecto de código abierto que se comenzó a desarrollar en el año 2009 por la empresa LinkedIn. Y en el 2011 fue donado por Apache, la cual estuvo en desarrollo hasta Octubre de 2012 donde paso a formar parte de los proyectos de alto nivel. Esta desarrollado en Java y Scala puede manejar distintos orígenes de datos como son las redes sociales o sensores y trabaja en procesamiento de tiempo real.

En la figura 4, se identifica el logotipo de Kafka como una plataforma de almacenamiento distribuido y con posibilidades de réplica. Muy rápido y eficiente en lecturas y escrituras (Rayón, 2016).

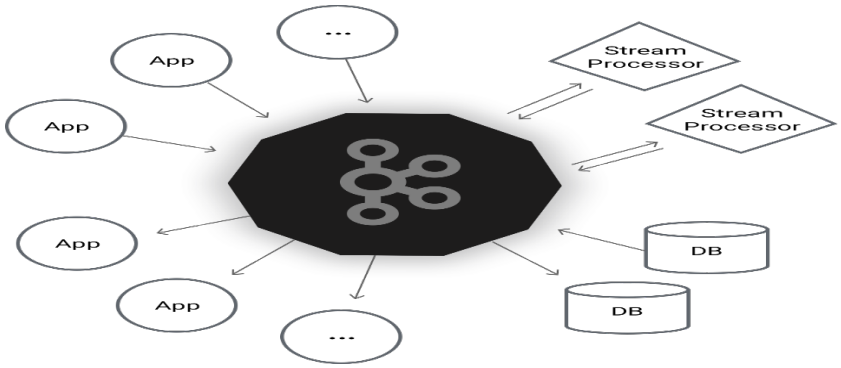


Figura 4. Logotipo Kafka (Apache Kafka, s.f.).

3.2.1.2. Sqoop

Es una herramienta de alto nivel, está diseñada para transferir grandes cantidades de datos a Hadoop provenientes de bases de datos estructurados como lo es MySql, Oracle, Postgress o un data warehouse.

La trasferencia es realizada al leer fila por fila de cada tabla y las importa a sistema de archivos distribuido (HDFS) de Hadoop para obtener de salida una gran variedad de archivos, los cuales pueden ser de formato .CSV, Avro (sistema de compresión diseñado por Apache para el proyecto Hadoop (Big Data Dummy, 2017)), de secuencia o binarios. El logotipo que lo distingue se presenta en la siguiente figura.

Figura 5. Logotipo Sqoop (Apache Sqoop, 2019).

3.2.1.3. Flume

Es una herramienta de la fundación Apache, es un sistema distribuido, gratuito y muy eficiente para la recopilar, agregar y mover grandes cantidades de datos en logs los cuales pueden ser diferentes tipos de orígenes desde diferentes servidores, web o de algún otro tipo de servidor que pueda brindar información relevante. Este sistema permite realizar la ingesta de datos semiestructurados como no estructurados.

Utiliza una arquitectura sencilla y flexible que está basada en el flujo de datos tipo streaming, permite la creación de bastantes aplicaciones analíticas en línea. La figura 6 muestra un flujo de trabajo en donde se utiliza flume

Figura 6. Transportar datos en flume e imagen de flume. (Big Data Dummy Analytics, 2017)

3.2.2. Herramientas de Almacenamiento

En los siguientes párrafos se describen herramientas relacionadas con la característica de almacenar grandes cantidades de datos, relacionándose con la característica de volumen y velocidad en términos de Big Data.

3.2.2.1. Hadoop

Apache Hadoop es una infraestructura que permite el procesamiento distribuido de grandes conjuntos de datos en grupos de computadoras utilizando modelos de programación simples y eficientes. Está diseñado para crecer desde servidores individuales hasta miles de máquinas, cada una de las cuales ofrece computación y almacenamiento local. Hadoop está pensado para detectar y manejar fallas en la capa de aplicación, por lo que ofrece un servicio de alta disponibilidad en la parte superior de un grupo de computadoras, cada una de las cuales puede ser propensa a fallas (Apache Hadoop, s.f.).

MapReduce es el motor de procesamiento de Hadoop utilizando nodos en varios servidores y su sistema de almacenamiento es Hadoop Distributed File System (HDFS) permite a las aplicaciones ejecutarse en varios cluster, de esta manera su procesamiento es más rápido.

Hadoop es una herramienta de alto nivel ya que es la gran referencia al hablar de Big Data, ya que es usado y diseñado por una gran cantidad de

contribuyentes utilizando lenguaje Java. La figura 7 identifica el logo de Apache hadoop.

Figura 7. Apache hadoop. (Apache Hadoop, s.f.)

3.2.2.2. MongoDB

MongoDB es una base de datos orientada a documentos muy similares a los de tipo JSON, es una base de datos No SQL y por consiguiente es muy flexible. Su lenguaje es de gran alcance es muy extenso y comprensible esto proporciona un apoyo extraordinario para filtrar y clasificar cualquiera que sea el campo deseado. Las consultas son de tipo JSON y por esta razón fácilmente compuestas. (MongoDB, 2020). La figura 8 identifica el logo de MongoDB

Figura 8. mongoDB (MongoDB, 2020)

3.2.2.3. Cassandra

Esta base de datos está diseñada para cuando es necesario obtener escalabilidad y se requiere alta disponibilidad sin afectar el rendimiento. Tiene gran tolerancia a fallas en el hardware básico o la infraestructura en la nube. El soporte de Cassandra para la replicación en múltiples centros de datos es el mejor de su clase, por consiguiente genera confiabilidad para el almacenamiento de los datos. Es una base de datos NoSQL que almacena los datos a través de una agrupación de llave – valor con un identificador que permite obtener la información de una manera mucho más rápida (Apache Cassandra, 2016).

El Lenguaje de Consulta es CQL ofrece un modelo cercano a SQL, se desarrolló inicialmente por Facebook con el fin de mejorar la búsqueda en la bandeja de entrada, el 2008 se lanza como un proyecto *source code* de Google siendo en el 2010 cuando alcanza el nivel de proyecto de alto nivel. Una

particularidad de esta herramienta es que no necesita de Hadoop para que poder trabajar con ella. La figura 9 identifica el logotipo de cassandra.

Figura 9. Logotipo de Cassandra (Apache Cassandra, 2016)

3.2.2.4. HBase

Es un almacén de Big Data distribuido y escalable, es utilizada cuando se requiere un acceso aleatorio y en tiempo real de lectura y escritura en Big Data. Esta herramienta almacena tablas demasiado grandes que pueden tener miles de millones de filas, así como miles de millones de columnas sobre hardware sencillo. Está modelada a partir de Bigtable de Google el cual es un sistema de almacenamiento distribuido para datos estructurados. Generalmente esta base de datos se utiliza para operaciones en línea la cual permite trabajar con las operaciones mucho más rápido que otras herramientas. (Apache HBase, 2020)

Esta herramienta fue un proyecto de la empresa Powerset para poder trabajar y procesar las enormes cantidades de datos utilizando un lenguaje natural. Actualmente es un proyecto de la fundación Apache de nivel superior. La figura 10 identifica el logo actual de HBASE.

Figura 10. Apache HBASE (Apache HBase, 2020)

3.2.3. Herramientas de Procesamiento

Cuando se encuentran los datos disponibles en el sistema ya es posible iniciar con el procesamiento de datos, existen diferentes alternativas y herramientas para atender los problemas de procesamiento de Big Data.

3.2.3.1. Spark

Es un sistema de computación tipo clúster de propósito general y está orientado a la velocidad. Proporciona APIs en Java, Scala, Python y R. Cuenta con un motor optimizado la cual soporta la ejecución de gráficos. Soporta un conjunto extenso de herramientas de alto nivel y en las cuales incluyen Spark SQL el cual está diseñado para el procesamiento de datos estructurados como son del tipo SQL, también existe MLlib en el que se puede implementar machine learning, cuanta con Spark Graphs que permite trabajar con bases de datos basados en grafos y Spark Streaming que habilita procesar datos que provienen de una fuente de tipo Streaming. El logotipo que lo distingue lo presenta la figura 11:

Figura 11. Spark. (Apache Spark, s.f.)

3.2.3.2. Apache Hive

Este software de almacenamiento facilita a los usuarios la lectura, escritura y administración de los datos, se pueden hacer consultas utilizando la sintaxis de SQL. Para la implementación es necesario contar con una versión superior de Java 1.7 y trabajar con Hadoop 2x, ya que no es recomendada la implementación de versiones anteriores a estas.

Es un proyecto de código abierto por parte de la Fundación Apache que inicialmente empezó como un sub proyecto de Hadoop, y con el paso del tiempo se convirtió en un proyectó de alto nivel gracias a lo amigable que es para los usuarios. El logotipo que lo distingue se identifica en la figura 12.

Figura 12. Logotipo de Apache Hive (Apache Tive TM, 2014)

3.2.3.3. Cloudera Impala

Esta herramienta proporciona consultas SQL las cuales son rápidas e interactivas directamente con los datos de Hadoop que son almacenados en HDFS o HBase. Utiliza los mismos metadatos, para las consultas utiliza la sintaxis SQL y el controlador ODBC. Esto otorga una plataforma familiar y unificada para poder realizar consultas en tiempo real o por lotes. Cuenta con un sistema único para el procesamiento y el análisis de grandes cantidades de datos, esto puedan evitar el modelado costoso y los trabajos de Extraer, Transformar y Cargar (ETL) utilizados solo para análisis.

Figura 13. Logotipo Cloudera (Cloudera, s.f.)

3.2.3.4. Apache Pig

Esta herramienta está diseñada para analizar grandes conjuntos de datos y a través de un lenguaje de alto nivel permite el análisis. Una de las cualidades más destacadas de los programas de Pig se debe a su estructura para poder manejar conjuntos de datos muy grandes. Pig consiste en un compilador el cual produce secuencias de programas MapReduce. El lenguaje de Pig actualmente consiste en

el lenguaje textual Pig Latin, que cuenta con las siguientes cualidades que lo hace una herramienta poderosa:

- Facilidad de programación: se codifican explícitamente como secuencias de flujo de datos, lo que facilita su escritura, comprensión y mantenimiento.
- Oportunidades de Optimización: permite optimizar al sistema su ejecución automáticamente lo que permite al usuario solo preocuparse en la semántica y no en la eficiencia.
- Extensibilidad: permite a los usuarios crear sus propias funciones para procesos especiales. El logotipo que lo distingue e visualiza en la figura 14:

Figura 14. Apache Pig (Apache Pig, 2018)

3.2.3.5. Apache Storm

Es un sistema de computación distribuida en tiempo real, gratuito y de código abierto. Esta herramienta facilita el procesamiento de flujos de datos ilimitados, provocando que se pueda realizar el procesamiento en tiempo real, es muy fácil de utilizar y tiene la factibilidad de poderse usar casi con cualquier lenguaje de programación.

Puede ser utilizado en tiempo real, en procesos de aprendizaje automático en línea, mejoras continuas y muchas más. Es muy rápido ya que ya que puede procesar más de un millón de registros por segundo por nodo. Escalable, tolera los fallos, fácil de configurar y garantiza que los datos serán procesados. Se presenta en la figura 15 el logotipo de Apache Storm

Figura 15. Apache Storm (Apache Storm, 2019)

3.2.4. Herramientas de Análisis

El análisis de datos es el proceso que busca obtener respuestas a interrogantes y patrones ocultos en la información, es la etapa más importante en el Big Data ya que las respuestas servirán de apoyo para tomar decisiones dentro de la empresa o institución. Se citan algunas herramientas que permiten análisis de datos en ecosistema Big Data.

3.2.4.1. Jupiter Lab

Este proyecto tiene su origen en la aplicación web Ipython Notebook la cual registra sesiones de trabajo con Python para trabajar en cuadernos electrónicos.

Un cuaderno de trabajo puede contener código fuente, textos explicativos, ecuaciones, gráficos, controles interactivos. Es fácil de instalar, el cual nos puede dar resultados en tiempo real, es muy amigable para los usuarios.

El logotipo que lo distingue es el que se muestra en la figura 16.

Figura 16. Apache Storm (Jupyter, 2020).

3.2.4.2. Python

Es un lenguaje de programación muy poderoso, es muy fácil de aprender, cuenta con licencia de código abierto y permite al usuario el análisis de gran

cantidad de datos de una manera rápida y sencilla, cuenta con un Shell poderoso, fácil de instalar en equipos de cómputo que no necesita especificaciones de hardware especial, es multiplataforma.

Es administrado por la fundación Python Software, es compatible con la licencia pública general de GNU a partir de la versión 2.1.1. El logotipo que lo distingue es el siguiente:

Figura 17. Python (Python, s.f.)

3.2.4.3. Lenguaje R

Software de licencia libre para computación estadística y gráficos. Se compila y se ejecuta en una amplia variedad de plataformas UNIX, Windows y MacOS. R tiene una gran variedad de técnicas estadísticas en el modelado lineal y no lineal, pruebas estadísticas clásicas, clasificación, agrupación, y análisis de tiempo, técnicas gráficas, y es altamente extensible. Permite generar gráficos de excelente calidad, incluyendo símbolos matemáticos y fórmulas. Cuenta con funciones naturales para el análisis, es un lenguaje de programación bien desarrollado, fácil de aprender y muy efectivo que incluye condicionales, bucles, funciones recursivas definidas por el usuario y de entrada y salida, entre otras. La figura 18 muestra el logotipo de R.

Figura 18. Lenguaje de programación R. (R-project.org, 2019)

3.2.4.4. Scala

Es un lenguaje directamente orientado a objetos ya que todo es un objeto, cuenta con una sintaxis ligera para definir funciones anónimas, soporta funciones

de primer orden, permite que las funciones sean anidadas, y soporta currying. Las construcciones incorporadas al lenguaje para reconocimiento de patrones modelan tipos algebraicos los cuales son usados en muchos lenguajes de programación funcionales.

El reconocimiento de patrones de Scala se puede extender al procesamiento de datos XML con la ayuda de patrones de expresiones regulares. Sus características hacen a Scala una excelente herramienta para el desarrollo de aplicaciones como Web Services. Se presenta en la figura 19 el logotipo de Scala.

Figura 19. Scala (Scala, s.f.)

3.2.5. Herramientas de visualización de datos

Son herramientas que permiten visualizar la información de manera ejecutiva y resumida con distintas y poderosas funcionalidades.

3.2.5.1. IBM Spss

Este lenguaje ofrece el análisis estadístico avanzado, cuenta con una gran cantidad de algoritmos de machine learning, puede analizar texto, extensibilidad de código abierto, se puede integrar al big data y cuenta con una implementación continua en sus aplicaciones. Su facilidad de uso, flexibilidad y escalabilidad permiten sea accesible para los usuarios con cualquier nivel de conocimientos y para los proyectos conjuntos de todos los tamaños y complejidad, es el software estadístico líder mundial utilizado para resolver problemas empresariales y de investigación mediante el análisis.

3.2.5.2. Tableau

Es una herramienta de análisis gráfica fácil de usar y muy potente, puede convertir los datos de múltiples fuentes de información, combina varias fuentes de datos en una sola pantalla, trabaja con cualquier hoja de cálculo y base de datos

de cualquier tamaño y combina varias vistas en un cuadro de mando. La figura 20 identifica el logotipo de Tableau.

Figura 20. Tableau (Tableau Software, LLC, s.f.)

3.2.5.3. RapidMiner

Es un programa informático utilizado para el análisis y la minería de datos, permite realizar procesos de análisis de datos utilizando el encadenamiento de operadores todo esto mediante un entorno gráfico. Es utilizado en las ramas de investigación, educación, capacitación, y en la creación de prototipos y también es usado en aplicaciones empresariales.

Su desarrollo fue realizado en Java, se puede integrar a lenguajes como Python y R, tiene la característica de poder usarse de manera gráfica, en línea de comandos, tipo batch y en otros programas a través de sus bibliotecas. Este incluye herramientas de visualización de datos y gráficos. La figura 21 identifica el logotipo de RapiMiner.

Figura 21. RapidMiner (RapidMiner, 2020)

3.2.5.4. SAS Studio

Esta herramienta se encuentra conectada a un server SAS con la finalidad de poder procesar comandos y eficientar el proceso, este servidor SAS puede estar alojado en un entorno de nube, un servidor local o una copia de SAS en la máquina local. Esto provoca que una vez procesado el código los resultados son enviados a SAS Studio. Una de sus características está en que es compatible con los

navegadores web mas utilizados como lo es Internet Explorer, Google Chrome, Safari y Mozilla Firefox.

Este lenguaje trabaja principalmente sobre tablas de datos y nos ofrece la facilidad de poder leerlas, transformarlas, combinarlas, resumirlas, crear informes a partir de ellas de una manera rápida y sencilla.

3.2.6. Aplicaciones de Big Data

Hoy en día el proceso adecuado para el análisis de los datos generados es de vital importancia ya que esto permite llegar al objetivo deseado no dependiendo de la fuente donde se generan estos datos. En la actualidad estos datos se utilizan para lograr un objetivo ya sea en el ámbito empresarial, científico, entretenimiento, deportivo, comercial, salud, transporte, ambiental, entre otros. Se describen algunos de ellos:

- Comercial dirigido a clientes: las empresas hoy en día necesitan tener conocimiento de lo que los clientes necesitan, por lo tanto necesitan crear modelos predictivos. Por eso es necesario tener las herramientas adecuadas así como la mejor estrategia para desarrollar un plan de trabajo y saber cuál es la mejor opción para estos casos ya que en la actualidad los datos llegan por fuentes como son redes sociales, logs, de navegación, envíos de mensajes de texto, videos reproducidos entre otros. Un ejemplo muy claro es en la industria automotriz de la cual pueden predecir las tendencias de los clientes así como mejorar sus productos al saber la forma en que los clientes conducen.
- En el deporte: Con el uso de sensores cada vez más capaces de registrar la fuerza, el alcancé, velocidad, y otras variables del deporte, se pueden realizar grandes análisis para obtener un mejor rendimiento en los deportistas en la actualidad. El registro de los datos ayudan a los atletas a ser más competitivos, un ejemplo de esto es en mundo del tenis en el cual llevan un registro de unos 41 millones de data points para determinar patrones y estilos de los jugadores, manejan un programa llamado *Slam Tracker* basado en la tecnología IBM SPSS el cual realiza un análisis predictivo.

- En la ciencia: La gran cantidad de datos que se generan mediante sensores o registros obtenidos de alguna otra fuente es necesario contar con las mejores herramientas de ingesta, almacenamiento y procesamiento para el análisis de dichos datos, un dato de esto es en el CERM que se encuentra en Suiza y que se encarga del gran colisionador de hadrones que genera una enorme cantidad de datos.
- En máquinas y dispositivos: Las herramientas de análisis ayudan a ser la maquinaria de mejor calidad y más inteligente, un ejemplo es al utilizar las herramientas de análisis para mejorar las redes de energía a partir de los medidores, así como también vehículos inteligentes gracias a la gran cantidad de sensores que hoy en día utilizan con los datos generados vuelve más óptima el manejarlos así como la seguridad del conductor y ahorros de consumibles.
- En las ciudades: Con los datos que se generan día a día en las ciudades se hacen análisis para optimizar los servicios de los ayuntamientos. Es posible usar datos en el manejo de los recursos, para conocer la calidad del aire, monitorear fugas de agua. Con las herramientas adecuadas se puede determinar patrones y comportamientos de flujo de vehículos en avenidas principales y así buscar soluciones reales.
- En la salud: Con la capacidad de procesamiento de plataformas de análisis de Big Data se pretende tener un mejor control en la salud pública, un ejemplo al poder analizar la búsqueda de medicamento desde la red en determinada área geográfica para determinar si existe el riesgo de una pandemia como ocurrió hace años con el virus AH1N1 el cual google lo determinó tiempo antes de ser publicado, al analizar la búsqueda de medicamentos y síntomas de gripe. Así también la capacidad de procesamiento de plataformas permite ya decodificar cadenas enteras de ADN en minutos esto permite encontrar un mejor tratamiento y conocer sus patrones de propagación. Un ejemplo es en el hospital de Toronto en la unidad de neonatos al analizar y grabar los latidos del corazón así como los patrones de respiración de cada bebe pueden predecir infecciones 24 horas antes de que aparezcan.

- Trading Financiero: El sector financiero y de servicios es un entorno que necesita obtener resultados en segundos y con la gran cantidad de datos que se generan a cada momento es necesario tener la mejor estrategia como las mejores herramientas para obtener análisis y resultados en tiempo real.

Estos fueron algunos de los sectores en los cuales se pueden utilizar las herramientas y lenguajes que se utilizan en el Big Data, ya que los datos obtenidos son de diferentes fuentes y de tipo estructurado, no estructurado o semiestructurado. Y en la mayoría de ellos es necesario obtener resultados en tiempo real.

3.3. Desarrollo

En este apartado se hacen comparaciones de distintas herramientas Big Data, con la ayuda de tablas comparativas se reflejan las principales características de las herramientas de ingesta, de almacenamiento, de procesamiento, de análisis, y de visualización de gráficos.

3.3.1. Herramientas de Ingesta de datos

Comparación entre Herramientas más comunes de Big Data para la ingesta de datos usados comúnmente en el ámbito de análisis de los datos comparando a Sqoop y Flume, la tabla 1 identifica las comparaciones.

Tabla 1.

Comparativo de Sqoop y Flume. Elaboración propia

Comparativo	Sqoop	Flume
Naturaleza Básica	Funciona bien con cualquier Sistema Manejador de Base de datos Relacional (RDBMS).	Funciona bien para la fuente de datos de Streaming que se generan continuamente.
Flujo de datos	Es utilizado específicamente para la transferencia de datos en paralelo. Por esta razón, la salida podría estar en múltiples archivos.	Se utiliza para recopilar y agregar datos debido a su naturaleza distribuida.
Arquitectura	Está basada en conectores, lo que provoca que los conectores sepan cómo conectarse a una fuente de datos diferente.	Basada en agentes, donde el código escrito en ella se conoce como un agente que es responsable de obtener datos.
Se Utiliza	Para copiar datos más rápido y luego usarlos para poder generar resultados analíticos.	Para extraer los datos cuando se desea analizar patrones, causas raíz o análisis de sentimientos en las redes sociales.
Actuación	Reduce el almacenamiento excesivo y sus cargas de procesamiento las trasfiere a otros sistemas.	Es tolerante a las fallas, robusto y su mecanismo de confiabilidad es sostenible para recuperación y recuperación de fallas.

3.3.2. Herramientas de Almacenamiento

En 1999 se desarrolló el teorema CAP, este afirma que en sistemas distribuidos es imposible garantizar a la vez: consistencia, disponibilidad y tolerancia a particiones (Manjarrez Antaño, Martínez Castro, & Cuevas Valencia, 2014). Solo dos de tres de estos aspectos se pueden alcanzar en un ambiente distribuido.

- Consistencia: al realizar una consulta o inserción siempre se tiene que recibir la misma información, con independencia del nodo o servidor que procese la petición

- Disponibilidad: que todos los clientes puedan leer y escribir, aunque se haya caído uno de los nodos.
- Tolerancia a particiones: Los sistemas distribuidos pueden estar divididos en particiones (generalmente de forma geográfica). Así que esta condición implica, que el sistema tiene que seguir funcionando aunque existan fallos o caídas parciales que dividan el sistema. (Manjarrez Antaño, Martínez Castro, & Cuevas Valencia, 2014)

En la figura 22 se muestra el modelo del teorema CAP que muestra algunas de las bases de datos según las condiciones que cumplen del teorema CAP

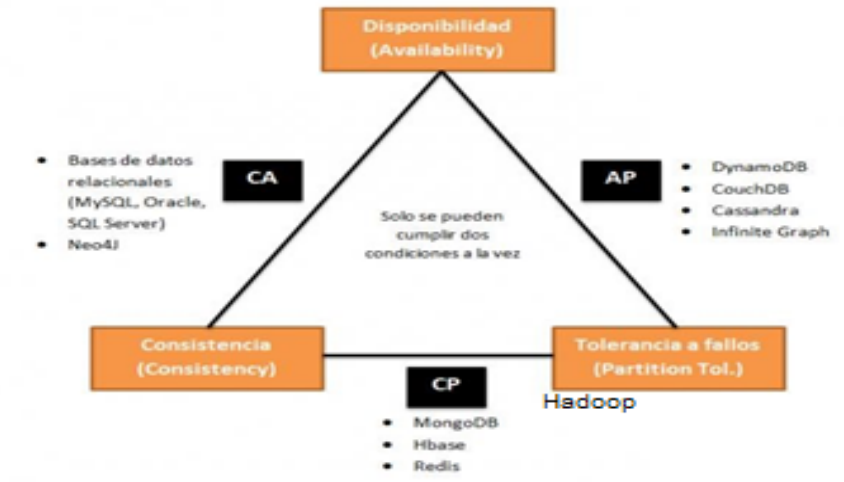


Figura 22. Teorema CAP. (Genbeta, 2014)

La tabla 2 hace un comparativo e MongoDB y Hadoop

Tabla 2

Comparativo MongoDB y Hadoop. Elaboración propia

Comparativo	MongoDB	Hadoop
En sistemas RDMBS	Puede mejorar RDMBS o en algunos casos reemplazarlo, puede utilizarse en diversos casos de uso.	Hadoop no remplaza el sistema RDMBS, ya que es un complemento que ayuda al almacenamiento de datos.
Entorno	Es una base de datos la cual está diseñada en lenguaje C++.	Es una colección de diferentes software, está diseñado en Java.
Formato de datos	Deben provenir de un formato .CSV o en JSON.	Maneja diferentes formatos como sea posible, estructurados como no estructurados.
Diseño	Permite procesar y analizar grandes cantidades de datos.	Para el almacenamiento y la recuperación de los datos.
Costos en Hardware	Si es muy rentable para su implementación.	Al ser una colección de software si se maneja un costo mayor.
El manejo en la memoria	Es muy eficiente en el manejo de la memoria.	Puede optimizar el espacio en memoria.
Marco de referencia	Este almacena los datos en colecciones en formato JSON o BSON.	Utiliza un sistema de archivos distribuidos de Hadoop (HDFS) y MapReduce.
Debilidad	Poca tolerancia a fallas, lo cual ocasiona pérdida de datos.	Depende de 'NameNode', que es el único punto de falla.
Fuerza	Solución más robusta, es más flexible que Hadoop. Y este sistema puede reemplazar a un RDBMS existente.	Su gran fortaleza es que está diseñado para manejar Big Data. Excelente sistema de almacenamiento que permite manejar procesos por lotes y trabajos ETL de larga duración.

3.3.3. Herramientas de Procesamiento

Existe gran variedad en Herramientas de procesamientos de Big Data en las cuales se puede realizar la comparación de 2 de ellas, siendo algunas de las más comunes y pertenecen a la fundación Apache pero no por este motivo son muy similares ya que las 2 herramientas son diseñadas para diferentes tipos de procesamiento y existen diferencias entre una y otra, las cuales son Apache Pig y Apache Hive. El comparativo se indica en la tabla 3.

Tabla 3

Apache Pig vs Apache Hive. Elaboración propia.

Comparativo	Apache Pig	Apache Hive
Definición	Sistema de flujo de datos de alto nivel y de código abierto con lenguaje simple de aprender conocido (Pig Latin).	De código abierto y similar a SQL el cual es utilizado para consultas analíticas. usa un lenguaje declarativo llamado HiveQL.
Procesamiento de datos	Lenguaje de procedimientos considerado de alto nivel, es amigable.	En lotes, procesamiento analítico en línea (OLAP) y para grandes bases de datos empresariales
Velocidad de procesamiento	Latencia más alta, trabaja con MapReduce en segundo plano y es mas rápido comparado con Hive	Trabaja el MapReduce en segundo plano por esta razón también tiene una latencia alta
Compatible Hadoop	Se ejecuta sobre MapReduce	Se ejecuta en MapReduce
Esquema Utilizado	No tiene un concepto de esquema y puede almacenar datos en un alias	El esquema utilizado es para insertar datos en tablas
Se utiliza la interfaz Web	No puede ser usados en la interfaz web ya que no es compatible	Si es compatible con esta interfaz
Tipos de datos	Estructurados y semiestructurados.	Solo con estructurados
Formato AVRO	Admite AVRO y estructuras de datos complejas y formato binario.	No es compatible con este tipo de archivos.
Uso	Los usuarios en general son programadores e investigadores.	Es utilizado para el análisis de datos.

3.3.4. Herramientas de Análisis

Entre las herramientas de análisis de datos masivos, destaca 2 de ellos por su poder en el análisis, ser amigables para programadores y obtener resultados en el análisis deseado. Estos son R y Python los cuales tienen bastante aceptación.

Cuentan con características en común:

- Open Source la cual permite una libre descarga en comparación a otras herramientas comerciales como es SAS y SPSS.
- Son lenguajes de programación con paquetes avanzados relacionados con estadística.
- Comunidades en línea las cuales son un apoyo para los usuarios
- En lo económico según un estudio realizado por O´Relly los científicos de datos que utilizaron herramientas de código abierto obtuvieron un mejor salario en comparación a los científicos que utilizan herramientas patentadas (US$130,000 código abierto vs. US$90,000 patentados).

Si bien ambos lenguajes de programación están ganando protagonismo en la comunidad de análisis de datos, están luchando por convertirse en el lenguaje de elección de los científicos de datos. La comparación se muestra en la tabla 3.

Tabla 3

Comparativo R y Python. Elaboración propia

Comparativo	**R**	**Python**
Bases	Manejados por el grupo R-Core y R Project, está escrito principalmente en lenguaje C, Fortran	Python Software Foundation(PSF), este lenguaje está inspirado en C, Modula-3 y esencialmente por ABC.
Usuarios	Investigadores de Big Data, pero está tomando fuerza en el ámbito del mercado empresarial	Programadores, analistas de datos, desarrolladores y científicos de datos.
Propósito	Uso amigable para el análisis de datos, estadística y utiliza diversidad en modelo de graficas	Este lenguaje está dirigido a la productividad así como legibilidad de código
Comunidad	Usuarios Stackoverflow (preguntas y respuestas), gran comunidad y documentación	Stackoverflow , los usuarios contribuyen con código y documentación
Flexibilidad	Para la realización de modelos estadísticos, fácil usar en fórmulas complejas.	Fácil sintaxis que ayuda en la depuración y codificación. Usado para realizar scripts en sitios web o en alguna otra aplicación.
Algunas tareas	Trabajo exploratorio de datos, es más fácil para usuarios principiantes. Sus modelos estadísticos se pueden realizar al escribir un código de pocas líneas	Lenguaje de programación hecho y derecho, una herramienta excelente para la implementación de algoritmos utilizados en la producción.
Uso	Análisis de datos computación independiente o realizar el análisis en un servidor individual.	Cuando la tarea de análisis de datos debe integrarse con alguna aplicación web o si el código debe incorporarse a una base de datos en producción.
Desventajas	Lento en curva de aprendizaje, necesario descargar paquetes.	No cuenta con la mismas cantidad de biblioteca en comparación con R
Ventajas	Gráficas excelentes, cuenta con un enorme catálogo para el uso de análisis de uso de datos, interfaz con GitHub y disponibilidad de RMarkdown.	Jupyter para compartir datos, sus cálculos matemáticos son fáciles y rápidos., bastante legibilidad de código , gran velocidad, excelentes funciones en Python.

3.3.5. Herramientas de visualización de datos

El análisis grafico es una forma de consulta en la que los datos se muestran de forma gráfica. Con esto se busca tener el propósito de una comprensión, razonamiento y toma de decisiones efectivos. Esto se usa principalmente en conjuntos de datos muy grandes y complejos.

Los objetivo de estas herramientas son:

- Trata de proporcionar una forma transparente en el proceso de los datos e información para un mejor análisis de los datos.
- Integra disciplinas científicas para mejorar la división del trabajo entre humanos y máquinas.
- Obtener un conocimiento más profundo de grandes datos, dinámicos y complejos.
- Proporciona informes más comprensibles, oportunos y defendibles.
- Comunicar de una mejor manera el informe para una mejor toma de decisiones.

Existen lenguajes de análisis visual que son muy usados en la comunidad de análisis de datos, como lo es SAS y Tableau. Se realiza una comparación que se muestra en la tabla 5.

Tabla 5.

Comparativo SAS Vs Tableau. Elaboración propia.

Comparativo	SAS	Tableau
Su Clasificación	Por debajo de las 10 primeras de herramientas usadas en el análisis de datos gráficos	Es la herramienta más usada en el ámbito de análisis y visualización de datos.
Clientes	Staples, Scotia Bank, Instituto Australiano de Salud y Bienestar	Accenture, Coca Cola, Bank of America, PayPal, Google, Skype, Citigroup, Dell, Walmart, The New York Times, US Army, Air Force.
Volumen de datos	Gran volumen de datos sin ningún problema	Se utiliza también en pequeñas y medianas empresas
Instalación	Complejo de instalar	Fácil de instalar y administrar
Soporte de su Funcionalidad	Existe un sistema de "Proceso almacenado" que es un riesgo alto si deja de funcionar, se puede programar para hacer literalmente cualquier cosa	Literalmente es una herramienta de arrastrar y soltar y si alguna funcionalidad no se ejecuta, hasta ahí llega su desempeño
Capacidad	Para escalar, es necesario un costo extra.	Viene con todas las funcionales cargadas, no puede crecer
Visualización	Deficiente en visualización de datos comparado con Tableau.	Gran potencial para visualización de datos
Características	El panel de análisis visual SAS tiene múltiples pestañas, con el panel interactivo y explorador de datos visuales	Tablero de Tableau puede contar una historia, cuenta con parámetros y mapas
Ventajas	Es excelente para modelar y exploración de datos	Cuenta con conectividad con Google Analytics y fuentes de datos dispersas
Apoyo	Excelente soporte de servicios al cliente, así como de expertos en el análisis.	No es necesaria la asistencia, se actualiza con frecuencia y los problemas se resuelven rápidamente.
Costo y Licencias	Su licencia es anual y su costo es único	El costo se difiere para cada una de sus características.

3.3.6. Propuesta de trabajo.

Las herramientas que se han analizado tienen el propósito de facilitar la captación o ingesta, el análisis, el almacenamiento, el procesamiento y la visualización de los datos. Los datos pueden provenir de diferentes orígenes o fuentes.

Se han analizado datos desde ya hace tiempo por parte de científicos e investigadores con el fin de comprender patrones o determinar causales de algún problema, lo que ahora representa el gran reto es la escala en la que estos son generados.

Esta explosión de "grandes datos" está transformando la manera en que se conduce una investigación y por este motivo es necesario conocer y entender las herramientas que se pueden utilizar para lograr analizar esta gran cantidad de datos relacionados con el descubrimiento científico, investigación ambiental y biomédica, educación, salud, seguridad nacional, empresarial entre otros. De entre los proyectos que se pueden mencionar donde se ha llevado a cabo el uso de una solución de Big Data se encuentran:

- El *Language, Interaction and Computation Laboratory (CLIC)* en conjunto con la Universidad de Trento en Italia, son un grupo de investigadores cuyo interés es el estudio de la comunicación verbal y no verbal tanto con métodos computacionales como cognitivos.
- Lineberger Comprehensive Cancer Center - Bioinformatics Group utiliza Hadoop y HBasepara para analizar datos producidos por los investigadores de *The Cancer Genome Atlas(TCGA)* para soportar las investigaciones relacionadas con el cáncer.
- El PSG College of Technology, India, analiza múltiples secuencias de proteínas para determinar los enlaces evolutivos y predecir estructuras moleculares. La naturaleza del algoritmo y el paralelismo computacional de Hadoop mejora la velocidad y exactitud de estas secuencias.

- La *Universidad Distrital Francisco Jose de Caldas* utiliza Hadoop para apoyar su proyecto de investigación relacionado con el sistema de inteligencia territorial de la ciudad de Bogotá.
- La *Universidad de Maryland* es una de las seis universidades que colaboran en la iniciativa académica de cómputo en la nube de IBM/Google. Sus investigaciones incluyen proyectos en la lingüística computacional (machine translation), modelado del lenguaje, bioinformática, análisis de correo electrónico y procesamiento de imágenes.

A nivel general estas herramientas se han utilizado en diferentes ámbitos de la sociedad y que hoy se siguen usando como lo es en:

- En la prevención de crimen. Gracias a la implementación de las aplicaciones Smart Steps, que coteja toda la información que es censada en los barrios de Massachusetts en Estados Unidos, ya es posible predecir en un 68% si ocurrirá un crimen en un área determinada, lo que ha permitido desarrollar mapas en los que se identifican zonas de alto riesgo. El uso de estas herramientas permiten a la policía analizar las incidencias delictivas de las zonas, la movilidad demográfica, la ocupación laboral de los pobladores y el perfil psicológico de los 6.5 millones de habitantes del estado en menos de un mes, lo que permite generar acciones de contención criminal antes de que un acto delictivo ocurra.
- En el uso de la movilidad. Al cierre de 2016, Uber contaba con 39 mil conductores registrados en México y poco más de un millón 200 mil usuarios inscritos en el país, otorgando el 3er lugar a nivel mundial utilizando esta plataforma. Gracias al uso de herramientas, el sistema Uber gestiona los millones de billones de destinos, esto le permite tener una base de datos de los diferentes puntos de inicio del trayecto a su destino y de esta manera les permite obtener tarifas adecuadas.
- En la salud. Los científicos de datos pueden conocer, en tiempo real, la actividad de las ambulancias de una ciudad, de esta manera pueden obtener información de los riesgos o de accidentes lo cual les permite realiza modelos para el despliegue de las ambulancias. Con estos datos analizados en segundos podría

facilitar la tarea de distribuir adecuadamente la red de hospitales de un estado, e incluso, a planificar los dispositivos de seguridad y salud frente a catástrofes naturales de grandes dimensiones.

- En el área de recursos humanos. Mediante el análisis de los datos de gran volumen de información, el departamento de recursos humanos puede generar mejores análisis para determinar porqué emigran los empleados o porqué se quedan a laborar en la empresa llevando a cabo registros de empleados anteriores y aplicaciones a vacantes, entre otros. Según un estudio realizado por la financiera y compañía de capital humano Workday, un 96% de los directores de atracción de talento considera que el Big Data permitirá avanzar para "predecir tendencias y lo que pasará en el futuro", con estas herramientas los gestores de capital humano tienen un amplio conocimiento de las necesidades de las empresas así como de las cualidades o defectos del sector empresarial.
- En las comunicaciones. En la actualidad las compañías de telefonía móvil e internet tienen un número impresionante de datos sobre sus clientes: cantidad de llamadas que realizan, tiempo de duración, horario en el que se ejecutan, cobertura de redes y cuentan con reportes de interrupciones en los servicios. Una de estas empresas a nivel mundial fue la empresa T-Mobile, al filtrar las conversaciones en redes sociales de sus clientes, redujo a la mitad su número de quejas, reconociendo cada eventualidad por sector y área. De esta manera lograron promocionar paquetes especiales para sus usuarios, ofertándoles un servicio específico antes de que abandonaran la compañía.
- En las ventas. La tienda Macy's es el perfecto ejemplo de cómo el uso de las herramientas adecuadas para el análisis puede ser más que útil a la hora de fidelizar con el usuario por medio del feedback en sus 900 almacenes y boutiques distribuidos a lo largo de Estados Unidos y su tienda online (a la que diariamente acceden 14 millones de personas). Gracias a este análisis de datos que generan los consumidores, los propietarios de "la tienda más grande del mundo" se han ahorrado 557 mil dólares anuales por concepto de "análisis y envíos de email-marketing", empezaron a generar estrategias de descuento en tiempo real de acuerdo a lo que se comentaba en redes sociales.

- En el deporte. Las decisiones en el mundo de los deportes se basan en criterios usualmente subjetivos y monetarios, en el año 2002 cuando Billy Beane, gerente deportivo del equipo de béisbol de los *Oakland Athletics*, contrató a un grupo de jugadores infravalorados pero los cuales se les puede realizar un contrato económico muy por debajo de lo habitual y que los convertía en jugadores rentables, utilizando un criterio de selección estadístico basado en el uso de herramientas de análisis. Se hizo una selección idónea en una base de 1918 jugadores, cotejando varios porcentajes como por ejemplo el número de veces que se pone en base o las veces que realizaba un hits y dejando en el pasando o en segundo plano el número de veces que realizaba un cuadrangular. Esta historia fue llevada al cine en la cual la película *Moneyball: El Juego de la Fortuna*, Beane llevó a enfrentar a los atléticos de Oakland, valuado en 45 millones de dólares, en contra de los Yankees de Nueva York New cuyos jugadores se cotizan en 125 millones.
- En la política. El expresidente de Estados Unidos, Barack Obama, decidió utilizar un análisis de datos para su reelección en 2012. Tras un primer análisis, los esfuerzos de la campaña se enfocaron en tres aspectos: registro (recoger datos de los votantes convencidos), persuasión (dirigirse a los dudosos de una forma eficaz) y voto del electorado (asegurarse de que los partidarios fueran a ejercer el voto sí o sí). Para validar la información de 235 millones 248 mil posibles votantes utilizó herramientas de análisis Big Data, el demócrata y su comisión recogieron datos a pie de campo y realizaron un *feedback* muy rápido vía notificaciones de *email,* con esto detectó en qué estados o zonas funcionaría mejor la publicidad, ayudándolo a optimizar el gasto en comerciales de televisión y enfocándolos a redes sociales.

3.3.7. Enlaces de instalación y configuración de herramientas

En este apartado se indican algunos los enlaces electrónicos (URL) que se recopilaron al momento de la escritura de este documento , son guía para los procesos de instalación de las herramientas citadas.

3.3.7.1. Instalación de Hadoop

Se realiza la descarga en el siguiente link: https://hadoop.apache.org/releases.html

3.3.7.2. Instalacion de Mongodb

Para iniciar con la instalación es necesario ingresar a la siguiente página para realizar la descarga https://www.mongodb.com/es una vez dentro de la página seleccionar el apartado para obtener la prueba de MongoDB.

3.3.7.3. Instalación de Tableau

Para la instalación de esta herramienta es necesario ingresar a la siguiente página para realizar la descarga del software: https://www.tableau.com/es-es/trial/tableau-software#form

3.3.7.4. Instalación del lenguaje R

Para la instalación de lenguaje de R es necesario ingresar a la página: https://www.r-project.org/; también se sugiere que se instale la herramienta RStudio desde el enlace https://rstudio.com/products/rstudio/

3.3.7.5. Instalación del lenguaje Python

La instalación de Python se debe descargar del siguiente link: https://www.python.org/

3.3.8. Análisis de propuesta

Para utilizar de cualquier herramienta de Big Data es necesario conocer el origen en done se va a obtener los datos, estimar su volumen, su manejo en tiempos de respuesta en los que se requiere el resultado y sobre todo conocer que es lo que se está buscando en el análisis para poder determinar que herramientas se necesitan para llevar a cabo estas acciones.

Mundialmente el análisis de datos es la especialidad de moda o una de las que se tiene pensado mayor crecimiento por el número importante de empresas de todo tipo desde venta de productos hasta el entretenimiento utilizan dichos datos

generados, así como gobiernos, políticos, científicos y ciudades han empezado a utilizar el análisis de datos.

La fundación de Software Apache es la punta de lanza para el desarrollo de estas herramientas ya que tienen una de las cualidades más fuerte al ser *Open Source* (código abierto) y con la gran comunidad que tienen no solo en el desarrollo, sino también en el soporte llevan una delantera muy importante en la Big Data.

Se mencionó que existen diferentes herramientas para trabajar con Big Data ya sea de tipo no comercial y comercial; que son de distinta naturaleza y de diferente propósito; además, que algunas de ellas son diseñadas únicamente para procesos especiales.

Para el uso del análisis de datos con las herramientas de Big Data existen grandes ventajas que otorgan estas herramientas, como lo es el procesamiento de grandes volúmenes de datos, las velocidades para trabajar con ellos, muchas de estas herramientas son de código abierto, comunidades que existen para apoyar en el uso de estas, la facilidad de obtener resultados de diferentes tipos de fuentes, los resultados obtenidos los cuales permite interpretar información de conducta, patrones a seguir o prevenir eventos en tiempo real si así se requiere, algunos de los lenguajes que se utilizan son muy potentes y fáciles de aprender, los resultados presentados son de manera muy legible y entendible para la toma de decisiones.

Algunas de las desventajas que se encuentran es que algunas de estas herramientas es que no son compatible con fuentes complejas de datos, el costo alto de varias de ellas, tal vez, se requiere hardware especial para ejecutar procesos eficientemente, algunos de los lenguajes son complicados de aprender. Sin embargo, siempre habrá que evaluar el costo beneficio de enfrentar el reto.

3.3.8.1. Herramientas de Ingesta (Sqoop vs Flume)

Entre las dos herramientas de las cuales se hicieron comparaciones para la captación de datos que se utilizan en Big Data la conclusión es que si necesita realizar dicha ingesta de datos de registro textuales en Hadoop, la herramienta Flume es la opción correcta gracias a su naturaleza y a que es una herramienta

especial para trabajar con una gran cantidad de datos no estructurados, esta herramienta tiene una gran flexibilidad en la ingesta y Sqoop no está diseñada para trabajar con datos no estructurados.

Sqoop no es la mejor opción para el manejo de datos basado en eventos y es usado para bases de datos estructurados únicamente, Flume es una herramienta más poderosa para la ingesta de datos.

3.3.8.2. Herramientas de Almacenamiento (MongoDB vs Hadoop)

En la comparación de MongoDB y Hadoop para el almacenamiento de datos como la solución de Big Data es necesario realizar una considerable investigación acerca de lo que se necesita obtener o cual es la mejor de ellas para este almacenamiento de datos masivos.

Si se busca el procesamiento de datos en tiempo real pero con baja latencia o se necesita reemplazar el RDBMS o se va iniciar un sistema transaccional nuevo en su totalidad es recomendable MongoDB.

Si se necesita un análisis de datos de larga ejecución y que se consulten datos en el momento que se necesiten, entonces Hadoop sería la opción. Se debe considerar la velocidad y el volumen de los datos a almacenar donde Hadoop es la mejor opción ya que está diseñado para manejar información a mayor escala y es uno de los temas que se debe tener en cuenta para una mejor escalabilidad o expansión del proyecto.

La que sea de las dos opciones pueden ser excelentes para una solución escalable que tenga que procesar grandes cantidades de datos complejos. Mucha gente que utiliza MongoDB recomienda que se utilicen ambos sistemas en conjunto ya que MongoDB delega tareas en tiempo real y el procesamiento de datos se lo deja a Hadoop la cual es una ventaja de las dos herramientas al trabajar en conjunto y hacer más eficiente las tareas. Una de las desventajas en trabajar en conjunto es que una mala configuración sería desastroso para la confiabilidad y sostenibilidad de alguna plataforma.

Hadoop tiene la gran ventaja de ser tolerante a fallos algo que MongoDB por su estructura y lo deja en una desventaja considerable ante Hadoop.

3.3.8.3. Herramientas de Procesamiento (Apache Pig vs Apache Hive)

Estas dos herramientas de procesamiento de datos son utilizados en los cluster de Hadoop y son consideradas herramientas poderosas para el análisis de datos y la Extracción, Transformación y Carga de datos (ETL siglas en ingles).

El usuario debe elegir los tipos de datos de entrada y lo que requiere a la salida ya que como se mencionó anteriormente trabajan en el cluster de Hadoop y comparten características muy similares en su procesamiento, velocidad y compatibilidad con Hadoop.

Apache Pig tiene una gran ventaja al poder ser usada para la interfaz WEB ya que Hive no es compatible con esta interfaz, así como Pig admite datos estructurados como semiestructurados y admite el formato AVRO, y Hive tiene estas limitantes.

La herramienta Pig es más poderosa que Hive, esta herramienta es utilizada en su mayoría por científicos de datos y en su defecto Hive solo es utilizada para el análisis de datos únicamente.

Un usuario debe seleccionar una herramienta basada en los tipos de datos y la salida esperada. Ambas herramientas proporcionan una forma única de analizar Big Data en el clúster de Hadoop. Según la discusión anterior, el usuario puede elegir entre Apache Pig y Apache Hive para sus requisitos.

3.3.8.4. Herramientas de Análisis (R vs Python)

Son dos de los lenguajes que se utilizan con mayor frecuencia en el análisis de datos ya que cuentan con gran aceptación por parte de programadores por ser lenguajes sencillos de aprender en comparación a otros existentes, como ya se ha mencionado, cuentan con una gran comunidad de desarrolladores independientes que los vuelve más poderosos en el análisis de datos, aunque hay que mencionar que muchas veces estas mejoras son solucionadas por R y después por Python, es más fácil de aprender el código Python que el del lenguaje R, pero R cuenta con

RStudio el cual facilita el entendimiento de los datos y variables para lograr un análisis adecuado con pocas líneas de código, R tiene mayor biblioteca de funciones que Python, RStudio tiene mayor variedad en gráficas para el uso del analista de datos.

Una de las cualidades importante de Python es que además de realizar análisis se puede integrar en aplicaciones Web o a una base de datos en producción mientras que R solo se puede utilizar en procesos independientes o individuales.

En pocas palabras estos 2 lenguajes analizados son excelentes solo que cada uno de ellos es para un propósito en específico y para el uso de análisis de datos que se requiere.

3.3.8.5. Herramientas de Análisis Graficas (SAS VS Tableau)

En la comparativa de herramientas de visualización de Big Data solo cabe mencionar que SAS se encuentra entre las primeras 10 herramientas más usadas mientras que Tableau es la número uno, esto debido a que es una herramienta diseñada para el uso de pequeñas, medianas y grandes empresas mientras que SAS está diseñada para grandes empresas, SAS maneja costos más elevados un ejemplo es que maneja una anualidad y para adquirir otros paquetes de esta aplicación llevan un costo extra y en Tableau todo viene incluido desde que se adquiere. Tableau es una herramienta más completa y más funcional que cualquier otra en ambiente gráfico; sin embargo las versiones comerciales están fuera del alcance de presupuestos académicos, pero si al alcance de presupuestos empresariales.

El análisis de datos existe en contextos y ámbitos como la ciencia, astrología, salud, empresarial, financiero, agricultura, minería, comercial, gubernamental, deportes, entre otros, ha cambiado la forma de ver sucesos que cotidianamente se presentan. Obteniendo mejorías en la vida diaria previniendo enfermedades, desastres naturales, bajando costos de producción, analizando causantes de problemas, y muchos otros beneficios.

Este tipo de análisis ha llegado a cambiar ciudades enteras a nivel mundial, un ejemplo muy claro es la ciudad de Barcelona considerada como una ciudad inteligente la cual con análisis de datos a llegado a obtener beneficios en el transporte, consumo de agua, fallos de electricidad, reutilización de residuos, emergencias hospitalarias, entre otros. En general al utilizar las herramientas adecuadas para el análisis se pueden obtener un sin fin de mejoras en todos los ámbitos.

Conclusiones y Recomendaciones

Ante el crecimiento constante de datos y de dispositivos que generan dichos datos se debe estar preparado para lograr un entendimiento de cómo se puede llevar un correcto proceso de análisis para lograr resultados sorprendentes, las herramientas, los términos usados y los beneficios que se pueden obtener de éstos análisis en cualquier sector son demasiados con la apropiada tecnología y conocimiento de ella.

Este documento ofrece un panorama general de herramienta Big Data, menciona la existencia de herramientas para ingesta o recolección datos, almacenamiento, procesamiento, análisis de datos y visualización de datos.

La mayoría de las herramientas analizadas en este documento trabajan en un ecosistema de Hadoop el cual permite trabajar con miles de nodos que lo vuelve más eficiente en el análisis de datos.

Se documentó también de los beneficios que se obtienen con estas herramientas, se podría decir que en todo tipo de sectores públicos, privados y comerciales pudieran estar presentes. En contextos como salud, alimentación, entretenimiento, cultural, ambiental, política, empresarial y muchos más, el Big Data puede dar material para analizar varios aspectos.

Se puede medir comportamientos de los usuarios, de temperatura, de humedad, de salud, de lo que sea necesario solo con tener algún sensor que emita algún dato.

Ha cambiado la forma de hacer negocios con Big Data y sus herramientas disminuyendo costos, reorganizando procesos, innovando y mejorando decisiones. En general Big Data y sus herramientas son un beneficio a nivel mundial y lo que falta por analizar es bastante ya que cada día la cantidad de datos será mayor.

De estas herramientas se puede decir que se ha llegado a comprender mejor el tema de análisis de grandes datos así como tener una idea de que es el próximo paso en lo que se refiere a la información, y que en poco tiempo en México llegará a ser una de las especialidades más rentables, como lo es en gran parte del mundo.

Estas herramientas no solo pueden ser útiles para grandes empresas refiriéndonos en lo comercial, sino también en pequeñas y medianas empresas. Además de los beneficios que pueden tener en otros ámbitos y ser necesario gente capacitada para poder realizar este tipo de análisis.

Finalmente se concluye que la mayoría de las herramientas descritas son de fácil acceso y de una instalación no muy complicada, así como muy amigables en lo referente a la integración con el usuario, varias de ellas son de fácil comprensión para usuarios de nivel básico en conocimientos de cómputo.

Referencias

Apache Cassandra. (01 de 01 de 2016). *Apache Software Foundation*. Obtenido de Apache Cassandra: https://cassandra.apache.org/

Apache Hadoop. (s.f. de s.f. de s.f.). *Apache Hadoop*. Obtenido de Apache Hadoop: http://hadoop.apache.org/

Apache HBase. (14 de 07 de 2020). *Apache HBase*. Obtenido de Welcome to Apache HBase: https://hbase.apache.org/

Apache Kafka. (s.f. de s.f. de s.f.). *Apache Kafka.* Obtenido de A distributed streaming platform: http://kafka.apache.org/

Apache Pig. (01 de 01 de 2018). *Apache Pig*. Obtenido de Welcome to Apache Pig!: http://pig.apache.org/

Apache Spark. (s.f. de s.f. de s.f.). *Apache Spark*. Obtenido de Apache Spark™ is a unified analytics engine for large-scale data processing.: http://spark.apache.org/

Apache Sqoop. (18 de 01 de 2019). *The Apache Software Fundation*. Obtenido de Apache Sqoop: http://sqoop.apache.org/

Apache Storm. (01 de 01 de 2019). *Apache Storm*. Obtenido de Apache Storm: http://storm.apache.org/

Apache Tive TM. (01 de 01 de 2014). *Apache Tive TM*. Obtenido de Apache Tive TM: https://hive.apache.org/

Big Data Dummy. (10 de 01 de 2017). *Big Data Dummy. Analytics, NoSQL and Microservices*. Obtenido de Formatos de Fichero: https://bigdatadummy.com/2017/01/10/apache-avro/#avro

Big Data Dummy Analytics. (17 de 02 de 2017). *Big Data Dummy Analytics, NoSQL and Microservices*. Obtenido de Apache Flume Ingesta: https://bigdatadummy.com/2017/02/07/apache-flume/

Cloudera. (s.f. de s.f. de s.f.). *Cloudera Documentation*. Obtenido de Cloudera Documentation: https://docs.cloudera.com/documentation/enterprise/5-3-x/topics/impala_intro.html

Davenport, T., & Patil, D. (01 de 10 de 2012). *Data. Data Scientist: The Sexiest Job of the 21st Century*. Obtenido de Data. Data Scientist: The Sexiest Job of the 21st Century: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Genbeta. (28 de 01 de 2014). *Genbeta: Dev*. Obtenido de Genbeta: Dev: https://www.genbeta.com/desarrollo/nosql-clasificacion-de-las-bases-de-datos-segun-el-teorema-cap

Gil, E. (2016). *Big data, privacidad y protección de datos.* Madrid: Publisher: Agencia Española de Protección de Datos y Boletín Oficial del Estado ISBN: 9788434023093.

Jupyter. (01 de 01 de 2020). *Jupyter*. Obtenido de jupyter: https://jupyter.org/

Manjarrez Antaño, A. C., Martínez Castro, J. M., & Cuevas Valencia, R. E. (2014). Migración de Bases de Datos SQL a NoSQL . *Tlamati, Número Especial 3 CICOM* , 144-148.

MongoDB. (15 de 07 de 2020). *MongoDB*. Obtenido de The database for modern applications: https://www.mongodb.com/es

OBS Business School. (s.f. de s.f. de s.f.). *OBS Business School*. Obtenido de Noticias. En 2020, más de 30 mil millones de dispositivos estarán conectados a Internet:

https://obsbusiness.school/es/noticias/estudio-obs/en-2020-mas-de-30-mil-millones-de-dispositivos-estaran-conectados-internet

Pérez Marqués, M. (2015). *Bid Data. Técnicas, herramientas y aplicaciones.* México: AlfaOmega Grupo Editor S.A de C.V.

Portafolio. (04 de 12 de 2019). *Portafolio. Cambios y desafíos para el Big Data en 2020*. Obtenido de El empleo de esta tecnología disruptiva gana cada vez más terreno en distintos ámbitos.: https://www.portafolio.co/innovacion/cambios-y-desafios-para-el-big-data-en-2020-536199

Python. (s.f. de s.f. de s.f.). *Python*. Obtenido de Python: https://www.python.org/

RapidMiner. (01 de 01 de 2020). *RapidMiner*. Obtenido de RapidMiner: https://rapidminer.com/

Rayón, Á. (18 de 12 de 2016). *Deusto Data.* Obtenido de TECNOLOGÍAS DE INGESTA DE DATOS EN PROYECTOS «BIG DATA» EN TIEMPO REAL: https://blogs.deusto.es/bigdata/tecnologias-de-ingesta-de-datos-en-proyectos-big-data/

R-project.org. (01 de 01 de 2019). *R-project.org*. Obtenido de R-project.org: https://www.r-project.org/

Salazar Argonza, J. (01 de 01 de 2016). Big Data en la educación. *Revista Digital Universitaria ISSN: 1607 - 6079. Publicación mensual*, 16. Obtenido de Big Data en la educación: http://www.revista.unam.mx/vol.17/num1/art06/#

Sánchez Villaseñor, O. (01 de 04 de 2019). Herraientas, retos, oportunidades, seguridad y tendencias del Big Data. *Tesina para obtener el título de Ingeniero en Computación*. Toluca, Estadode México, México: UNAM. Facultad de Ingeniería.

Scala. (s.f. de s.f. de s.f.). *Scala*. Obtenido de Scala: https://docs.scala-lang.org/getting-started/index.html

Tableau Software, LLC. (s.f. de s.f. de s.f.). *Tableau*. Obtenido de Que es Tableau: https://www.tableau.com/es-mx

Twitter. (25 de 06 de 2018). *Twitter.* Obtenido de Cloud Business: https://twitter.com/cloudbusinesspe/status/1011361824946651136

Universidad de Alcalá. (s.f. de s.f. de s.f.). *Universidad de Alcalá*. Obtenido de Las cinco Vs. que sirven para explicar el Big Data: https://www.master-bigdata.com/big-data-5-v/

Capítulo 4

Big Data. Análisis de Estrategias de Marketing Digital

María Dolores Concepción De Lara Gurrola
Tecnológico Nacional de México. Instituto Tecnológico de Durango

03041017@itdurango.edu.mx

Jesús Eduardo Carrillo Morales
Tecnológico Nacional de México. Instituto Tecnológico de Durango

03040123@itdurango.edu.mx

Luis Fernando Galindo Vargas
Tecnológico Nacional de México. Instituto Tecnológico de Durango

lgalindo@itdurango.edu.mx

4.1. Introducción

Actualmente se está utilizando un nuevo esquema de investigación informático, útil para muchas empresas, que proporciona información que anteriormente no se tenían en cuenta.

Este documento, aporta una guía que lo llevarán a desarrollar las habilidades para el mejor desempeño en Marketing Digital además de servir de base para identificar más oportunidades de crecer si se exploran nuevos campos de investigación con el Big Data.

El análisis de Big Data ayuda a las organizaciones a aprovechar sus datos y utilizarlos para identificar nuevas oportunidades dentro de su especialidad. Es un paso más allá de la estadística, se puede procesar con el uso de herramientas como programas de lealtad, reportes de ventas, análisis web y bases de datos de los clientes.

La economía obliga a las empresas de nivel global a competir por la información, centralizar esta misma controlando aspectos de negocio basados en grandes volúmenes de datos. El marketing digital puede aportar estrategias que se pueden tomar en cuenta al querer incursionar en el comercio electrónico de cualquier producto o servicio dirigido a los consumidores finales creando nuevos horizontes que apuntalarán a nuevos esquemas de negocios de todo tipo de mercado.

En estos tiempos se cuenta con tecnología que permite tratar grandes cantidades de datos en un tiempo muy limitado, de manera que su análisis casi puede ser utilizado en tiempo real para lograr una toma de decisiones en base de los resultados que se obtengan, ya sean comerciales como de cualquier otro tipo en cualquier campo que se deseen aplicar.

La recopilación de grandes cantidades de datos y la búsqueda de tendencias dentro de los mismos, permiten que las empresas vean un gran crecimiento sin problemas y de manera eficiente gracias al marketing digital en gran medida a que se adapta para la creación de los perfiles de búsqueda del usuario.

Personas con o sin conocimientos tecnológicos, tienen la incertidumbre de cómo se almacena toda la información generada en el mundo: en Facebook, Twitter, o como el buscador tan conocido que es Google es capaz de manejar todas las búsquedas que se hacen a diario.

El objetivo de este documento es hacer un análisis de diferentes formas que el marketing digital puede apoyar a las empresas y en las cuales se implementa el Big Data para obtener máximo beneficio en cuanto la presencia que proyecta el marketing digital.

De manera específica se busca lo siguiente:

- Identificar aspectos generales de los paradigmas Business Intelligence, Data Mining y Machine Learning.
- Determinar los aspectos y características del paradigma Big Data y la relación con otros temas.
- Identificar las herramientas que conformas el Big Data
- Identificar las características del Marketing Digital
- Determinar aspectos de implementación de Marketing Digital y Big Data en empresas de manera general. (posible propuesta)

Al integrar patrones de búsquedas, conductas y tecnologías a fin de explotar este conocimiento como nueva tendencia dentro de las proyecciones de ventas o posicionamiento de marcas.

Este documento muestra una forma, de cómo, desde un email, publicaciones, redes sociales y publicidad en videos proyectan el marketing especializado a áreas en las que muestra interés cada usuario, señaliza o maneja algún tipo de perfil en base a los comportamientos que el Big Data y otras herramientas de análisis pueden ofrecer.

El propósito de este capítulo es mostrar algunos esquemas de análisis donde las empresas puedan recurrir al Big Data y sus herramientas en el proceso de toma de decisiones, concientizar a dichas empresas de las diferentes estrategias de posicionamiento las cuales pueden proyectarse a mayor número de usuarios específicos a fin de obtener mayor beneficio y ganancias en su ramo.

El impacto que se tiene al utilizar el Big Data en el marketing digital permite conocer mejor a los clientes de una empresa o futuros clientes potenciales, por lo cual tendría un gran impacto en el incremento de ventas ya que se puede ofertar al consumidor un producto de acuerdo con las necesidades que requiera.

Un beneficio de utilizar el Big data y sus herramientas adecuadas, es que se pueden procesar los datos tomando los criterios adecuados para llegar a lograr una estrategia de marketing digital exitosa.

4.2. Marco de referencia

Para comprender más lo que es el Big Data, se debe de conocer su historia y como interactúa con otros factores como Business Intelligence, Data Mining y Machine Learning los cuales serán la base para optar por la mejor opción dentro del Marketing Digital.

En 1958 Hans-Peter Luhn escribe un artículo donde habla por primera vez del *Business Intelligence* (BI) en el cual menciona “negocios es una colección de actividades llevadas a cabo con cualquier propósito, ya sea ciencia, tecnología, comercio, industria, derecho, gobierno, defensa, entre otros.

La noción de inteligencia también se define aquí, en un sentido más general, como la "capacidad de comprender las relaciones entre los hechos presentados de tal manera que guíe la acción hacia una meta deseada.” (Luhn, 1958).

Por otra parte, es importante mencionar que el conocimiento es generado a partir del buen uso de la información, la información es el significado de los datos y es derivado a partir de ellos. Existe una asociación importante entre datos, información y conocimiento. Si la información genera conocimiento, entonces ha ocurrido proceso de inteligencia en un ser inteligente. (Ahumada Tello & Perusquia Velasco, 2016).

En los siguientes apartados, se describen los conceptos de Inteligencia de Negocios; Big Data y sus herramientas; Minería de Datos y sus herramientas; Marketing Digital y sus características y acepciones; entre otros conceptos.

4.2.1. ¿Qué es el Business Intelligence?

La inteligencia de negocio (BI) es un conjunto de estrategias, acciones y herramientas que apoyan a la administración y creación de conocimiento mediante el análisis de variables y datos históricos en una empresa. (Ahumada Tello & Perusquia Velasco, 2016)

El término Inteligencia de Negocios integra una amplia variedad de tecnologías, herramientas, acciones, estrategias, plataformas de software, aplicaciones y procesos. (Peña Ayala, 2006)

La finalidad de la Inteligencia de Negocios es apoyar en el proceso de toma de decisiones que enriquecen el conocimiento de la empresa mejorando procesos productivos que permite obtener ventajas competitivas en las empresas.

La Inteligencia de negocio en relación con el Big Data, obtiene beneficio del procesamiento y almacenamiento de grandes cantidades de datos que son generados por las aplicaciones y sistemas de información; al tener los datos e información el BI y el Big Data los convierten en valor y conocimiento para la empresa sustentando así decisiones de mejor calidad. (García Merino & García Merino, 2017)

4.2.2. ¿Qué es el Big Data?

Por su significado en español grandes datos. El Big Data en la informática ha tomado gran significado al recolectar la información de varias fuentes estructuradas o no estructuradas, con facilidad de encontrar, analizar y dicha información.

Las organizaciones han encontrado en el Big Data una gran técnica para la toma de decisiones, con un conjunto de herramientas informáticas que permiten la manipulación, gestión y análisis de la información tomando en cuenta ciertos parones de cómo se comporta la información obtenida.

En la figura 1 se presenta al centro el concepto de Big Data y alrededor distintas tecnologías, disciplinas y conceptos con las que se relaciona.

La figura 1 muestra que Big Data se relaciona eminentemente con la disciplina de estadística y distintas herramientas para análisis de datos; se relaciona con el paradigma NoSQL para el tratamiento de datos no solo estructurados; evidentemente se hace necesario el uso de herramientas para el almacenamiento y procesamiento de información por eso la razón de tener relación con el cómputo en la nube; así mismo, el uso de cómputo móvil para la visualización y comunicación de datos a usuarios finales, entre otras cosas.

Figura 1. Big Data, tecnologías, disciplinas y conceptos con que se relaciona. (SEGITTUR, 2019).

Big Data se relaciona con una gran ingesta de información, el tamaño y el número no son las únicas variables grandes que están implicadas. Los principales conceptos agrupados que han constituido y definido este nombre han sido las denominadas tres V's: volumen, variabilidad y velocidad.

Las nuevas fuentes de datos que alimentan al Big Data tienen propiedades no solo por su volumen, sino también por otras características distintivas frente a las fuentes de datos tradicionales. Big Data generalmente recurre al esquema de las 3 vs (volumen, velocidad y variedad), si bien a veces se añaden algunos atributos para alcanzar las 5 v's (veracidad y valor) o incluso las 7 v's (variabilidad y visualización) (Gutiérrez Puebla, 2018).

Big Data puede ser visto como una tendencia y un presente en el avance de la tecnología que identifica un área de oportunidad hacia un nuevo enfoque para mejorar la toma de decisiones, se utiliza para procesar, analizar y describir enormes cantidades de datos (estructurados, no estructurados y semi- estructurados) y viene a ser una evolución del mundo de las bases de datos relacionales. El concepto de Big Data se utiliza en las organizaciones con fuentes de información interna y externa que no puede ser procesada o analizada utilizando herramientas y procesos convencionales (Puyol Moreno, 2014).

En resumen, la tecnología de los grandes datos ó Big Data es todo aquello que tiene que ver con grandes volúmenes de información que se mueven o analizan

a alta velocidad y que pueden presentar una compleja variabilidad en cuanto a la estructura de su composición.

La figura 2 identifica un esquema de las 7 v's, el significado y traducción de cada 'v' se presentan a continuación:

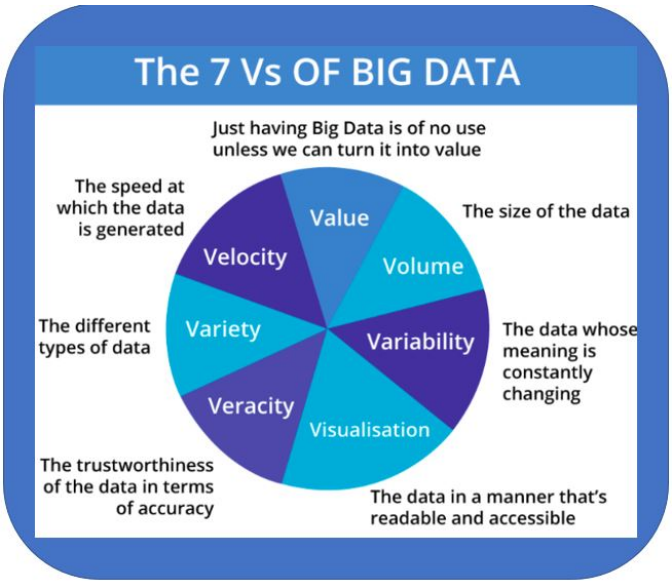


Figura 2. Las 7 v's del Big Data

Fuente: (Kudupu, 2018)

4.2.3. Características de Big Data

- Volumen de datos.
- Variedad de Información.
- Velocidad de la información.
- Veracidad
- Valor
- Variabilidad
- Visualización

4.2.3.1. Volumen de datos

Para Big Data, considerar la información dentro de este mismo debe de sobrepasar el terabyte de información, realmente no se conoce una cantidad específica, se podría mencionar claramente el volumen de información entre petabyte o zettabyte.

El volumen de los datos para conformar parte del Big Data, se puede tomar de cualquier medio conectado posible, ya sea sensores de movimiento, de temperatura, atmosféricos, celulares, relojes inteligentes entre otros.

Cada día, hora, segundo los sensores, tabletas, teléfonos y sistemas inteligentes generan muchos datos que aumentan de manera exponencial. En la época actual, el almacenamiento de los datos en las nuevas tecnologías no tienen más de dos años. La mayoría de estos datos no se procesan ni se analizan por los sistemas y aplicaciones tradicionales, dado que no tienen la funcionalidad ni la potencia de procesarlos. (Puyol Moreno, 2014).

4.2.3.2. Variedad de Información.

La variedad de datos puede ser obtenidos de diversas maneras en todo el mundo; por ejemplo, móviles, video, sistemas de localización (GPS), sensores digitales, automóviles, medidores eléctricos, anemómetros, entre otros, los cuales pueden medir y comunicar el posicionamiento, movimiento, vibración, temperatura, humedad, y hasta los cambios químicos que sufre el aire.

La variedad hace referencia a la diversidad de tipos, formatos y fuentes de datos, pueden ser estructurados como lo que se conoce en forma de tablas de las bases de datos relacionales; semiestructurados como lo que se observa en los archivos HTML o JSON; y no estructurados como archivos de texto, correos electrónicos, imágenes o vídeos, entre otros. Los datos no estructurados que pueden ser tratados más adecuadamente con sistemas NoSQL. (Gutiérrez Puebla, 2018)

4.2.3.3. Velocidad de la información.

Relacionado con el volumen está la rapidez a la que los datos son generados y procesados. Se generan de forma continua, de manera que es posible seguir procesos en secuencia y hacer análisis en tiempo casi real (Gutiérrez Puebla, 2018).

Debido a la gran variedad y volumen de la información susceptible de ser procesada bajo herramientas Big Data, las aplicaciones que analizan datos masivos requieren que la velocidad de respuesta sea lo suficientemente rápida como para lograr obtener la información correcta en el momento justo (Barranco Fragoso, 2012).

4.2.3.4. Veracidad

La veracidad hace referencia a la confiabilidad en relación directa a que existe una gran cantidad de datos. Actualmente ya se puede trabajar con poblaciones en lugar de con muestras y de que deriven de acciones de la gente. Por ejemplo, al hacer una encuesta se registra lo que la gente dice que hace; ahora por medio del monitoreo y seguimiento de acciones de las personas, se registra lo que la gente hace y cómo se comporta; estos son elementos a favor de la veracidad de los datos (Gutiérrez Puebla, 2018).

4.2.3.5. Valor

En la actualidad disponer de datos cobra un valor imprescindible y creciente, se compara a los datos con el petróleo de la cuarta revolución industrial. El dato independiente no tiene valor; lo que le da valor económico es darle significado y transformarlo en información y la vez en conocimiento útil para la tomar decisiones y ejecutar acciones (Gutiérrez Puebla, 2018).

4.2.3.6. Variabilidad

Los datos con el tiempo cambian, además, debido a la variabilidad de los procesos en las empresas, los datos cambian constantemente.

La variabilidad se refiere a los datos cuyo significado está continuamente cambiando. Este es, cuando la recopilación de los mismos se basa en el

procesamiento del lenguaje, ya que las palabras no tienen definiciones estáticas, y su significado puede variar mucho dependiendo del entorno (Maroto, 2018).

4.2.3.7. Visualización

Una aplicación para procesar, analizar y visualizar datos estructurados tiene sentido en sistemas tradicionales, ahora, con la gran cantidad de datos masivos, es necesario disponer de una manera de representarla la información de un modo más accesible y fácil de leer, y aquí es donde entra en juego el concepto de la visualización (Maroto, 2018).

Se necesita una forma de presentar y visualizar la gran cantidad de información más allá del tradicional formato de informe y gráficos con ejes ‘x’ y ‘y’, esto es un gran reto para el Big Data. (Maroto, 2018).

4.2.3.8. ¿Para qué el Big Data?

El Big Data, da a las organizaciones públicas y privadas la enorme oportunidad de obtener una mayor riqueza de decisión. Tradicionalmente las empresas para tomar decisiones tenían que apoyarse más en la intuición y habilidades directivas, Big Data está provocando, sin duda, un cambio de paradigma, un gran impacto en la cadena de valor del negocio. El reto está en aprovechar toda la infraestructura de este recurso, las herramientas, junto con una nueva forma de pensar, para innovar en aspectos que hasta ahora eran poco probables (Maroto, 2018).

El Big Data puede apoyar en procesos de análisis avanzados en algunas de estas cuestiones:

- Análisis descriptivo. Explica lo que los datos dicen que aconteció o está ocurriendo. Como ejemplo el caso de una tienda de autoservicio identificar y describir con datos el aumento repentino de las preferencias de productos por parte de los clientes o por el contrario identificar el que ya un producto se no vende.
- Análisis predictivo: Se anticipa un probable resultado; se busca determinar por qué ha sucedido en otras ocasiones un determinado evento y tratar así

de predecir, de acuerdo a los datos, lo que va a suceder. Aquí vale la pena la experimentación y la simulación. El caso de las ventas experimentar que pasaría si se ofrecen estrategias de ventas para los productos que no se venden o que pasaría si se lanzan estrategias de marketing para retener más a los clientes. Además, se puede aplicar algoritmos de predicción para pronosticar los comportamientos de los clientes y prepararse a los acontecimientos.

- Análisis prescriptivo: identifica cómo hacer que algo ocurra, es decir, qué se debería hacer y qué opción es la que se debería emplear. Para el caso de las ventas la prescripción sería encontrar la mejor estrategia para atraer a los clientes y aumentar ventas o, por el contrario, disminuir costos. (Antón Carranza, 2017).

El análisis en general se realiza de acuerdo al objetivo de encontrar patrones, relaciones y tendencias para poder localizar la información necesaria para la toma de decisiones de la venta de un servicio o producto.

4.2.4. Inconvenientes Big Data

Existen ciertos inconvenientes para el uso de Big Data principalmente el rechazo natural y la adopción del HW y SW, se citan algunos inconvenientes:

- Un largo proceso para su adopción
- Se requiere un gran esfuerzo
- Rechazo natural del personal
- Gasto económico
- Problemas de información no actualizada, entre otros (Ruiz García, 2016).

Big data transformará los negocios y las empresas modificarán a la sociedad. Se debe confiar en que las ventajas pesen más que las desventajas, pero hay que transitar razonadamente, puesto que la sociedad no parece todavía estar lista para la gestión de los datos que ya tiene capacidad para recoger (Cukier, 2014).

4.2.5. Elección de la fuente de datos

La elección de los datos en Big Data es muy importante, dicha información se tomará de varias fuentes, de redes sociales, de búsquedas en la WEB, de sensores, de bases de datos internar y externas, entre otros.

- Big data puede utilizar estos tipos de datos que son:
- Estructurados
- No estructurados
- Semi-Estructurados

4.2.5.1. Datos Estructurados

Los datos estructurados es información almacena de una manera organizada, relacionada y de fácil acceso, la organización de estos suele estar almacenados en tablas o campos de tipo fecha, numéricos, datos de caracteres y otros. Por ser un arreglo de datos la información se puede tomar linealmente.

Estos datos pueden ser obtenidos de diferentes lados como un punto de venta el cual maneja una base de datos SQL que registra los movimientos del día almacenando en sus registros semanas, meses y años de informaciones de venta.

4.2.5.2. Datos No Estructurados

Por lo contrario, los datos No estructurados carecen de organización no están definidos, estos son generados de manera masiva, ya sea, por un escrito, algún archivo en formato pdf o publicaciones en redes sociales (Facebook, twitter y muchos más).

Como se puede ver en la figura 3, se observa que los datos además de provenir de fuentes estructuradas (abajo izquierda), también puede provenir de otras fuentes en otros formatos, redes sociales, streaming o video de aplicaciones móviles, sensores múltiples y diversos dispositivos, estando en formatos semiestructurados y tal vez no estructurados.

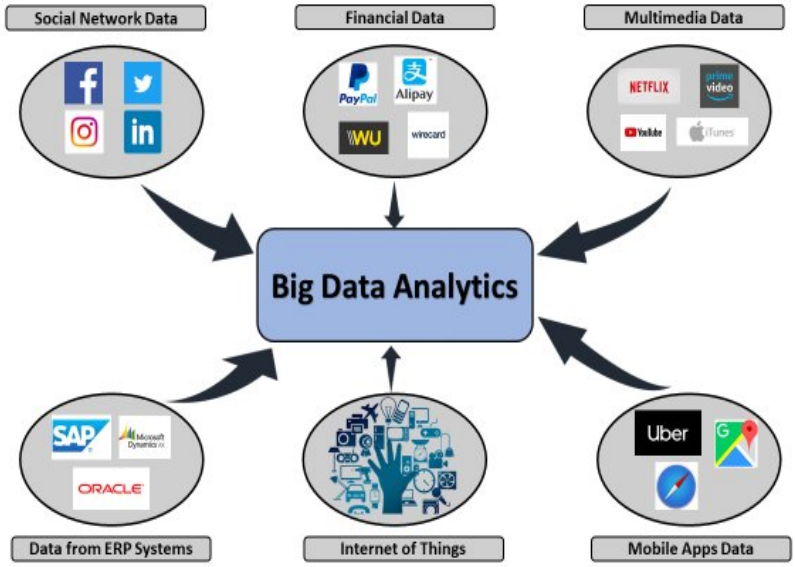


Figura 3. Fuentes de Datos Estructurados y No estructurados.

Fuente: (Lenz, 2019)

4.2.6. Herramientas para manejo del Big data

En los siguientes puntos, se hace mención de conceptos y distintas herramientas localizadas en la literatura consultada y que se relacionan con Big Data, son conceptos y herramientas para el procesamiento, análisis y almacenamiento de información principalmente.

4.2.6.1. Hadoop

Es un software diseñado para el manejo masivo de información estructurada, no estructurada o semiestructurada. Su desarrollo fue inspirado por Google's MapReduce y Google File System, desarrollado originalmente en Yahoo! y actualmente administrado como proyecto de Apache Software Foundation. Hadoop tiene una estructura de software de código abierto para el procesamiento de grandes bases de datos en sistemas distribuidos.

De manera específica Hadoop, es un software de desarrollo creada en código libre y abierto bajo licencia Apache. Básicamente, permite programar tareas intensivas de computación masiva, haciendo piezas pequeñas y distribuyendo el trabajo en un conjunto de varios procesadores y/o procesos. (Rodríguez, 2016)

Hadoop está disponible bajo la licencia Apache 2.0, es una biblioteca de software que permite el procesamiento distribuido de grandes conjuntos de datos a través de grupos de ordenadores que utilizan modelos sencillos de programación.

Hadoop fué diseñado para cambiar de trabajar con equipos individuales a utilizar varios servidores ejecutando y procesando datos al mismo tiempo, cada computadora local procesa y almacena de manera independiente para luego a través de procesos Map-Reduce juntar los resultados. Hadoop es un ambiente de código abierto, el cual permite escribir y ejecutar aplicaciones distribuidas que procesar grandes cantidades de datos.

4.2.6.2. Cassandra.

Cassandra es un sistema de gestión de base de datos diseñado para administrar y procesar una gran cantidad de datos distribuidos y almacenados por distintos servidores, proporciona un servicio de alta disponibilidad, robustez y confiabilidad además de una importante característica de alta tolerancia a fallas. (Castillo, Garcés, & Navas, 2017)

4.2.6.3. MongoDB

MongoDB es una base de datos no relacional (NoSql) orientada al manejo de documentos, tiene características de versatilidad, potencia y facilidad de uso, al igual que en su capacidad para manejar tanto grandes como pequeños volúmenes de datos. Es una base de datos que no tiene concepto del esquema relacional y uso de tablas, esquemas, SQL, columnas o filas. No cumple con las características ACID, que por sus siglas en inglés significan *Atomicity, Consistency, Isolation and Durability* (Atomicidad, Consistencia, Aislamiento y Durabilidad, en español).

MongoDB permite las operaciones CRUD, siendo éstas las siglas de *Create, Read, Update and Delete* (Crear, Recuperar, Actualizar y Eliminar); para almacenar y recuperar los datos hace uso del formato JSON, pero utiliza BSON, que es una forma binaria del formato JSON, el cual ocupa menos espacio al almacenar los datos. BSON es más rápida y eficiente para convertir a un formato de datos de un lenguaje de programación.

MongoDB es una base de datos para manejo de documentos de código abierto y encabeza el liderazgo en bases de datos NoSQL. Está desarrollado en el lenguaje de programación C++. Ofrece una alta disponibilidad, escalabilidad y particionamiento a cambio de consistencia y soporte transaccional. En términos prácticos, esto tiene que ver qué en lugar de tablas y filas, MongoDB utiliza documentos para hacerla rápida, flexible, escalable (Castillo, Garcés, & Navas, 2017).

4.2.6.4. Nube de cómputo (o cloud computing)

Cómputo en la nube o cloud computing es un concepto que tiene que ver con utilizar SW y HW que no está físicamente y precisamente localizado en casa o en el lugar de trabajo.

Cómputo en la nube proporciona una manera de acceder a servidores de almacenamiento, bases de datos y una gran variedad de servicios y aplicaciones por medio de Internet.

Los proveedores de servicios y aplicaciones en la nube normalmente cobran por disponer dichos servicios, aunque algunos ofrecen disponibilidad gratuita y de manera temporal.

La intención del cómputo en la nube es que a través de una inversión económica por parte de la empresa, ésta reemplace parcial o totalmente su infraestructura de cómputo, de tal manera que el entorno de cómputo pueda ser administrada por el proveedor externo, aumentando en capacidad de almacenamiento, administración de redes, seguridad, análisis de datos, desarrollo de aplicaciones, entre otros servicios (Guevara, 2018).

4.2.6.5. Extract, transform, and load.

Herramientas de software utilizadas para extraer datos de fuentes externas, transformarlos para satisfacer las necesidades operativas y cargarlos en una base de datos.

4.2.6.6. HBase

Base de datos abierta, distribuida y no relacionada, creada en Google's Big Table. Inicialmente desarrollada por Powerset, en la actualidad es gestionada por Apache Software Foundation como parte de Hadoop.

4.2.6.7. MapReduce

Modelo de programación ideado por Google para procesamiento de grandes bases de datos en ambientes distribuidos. MapReduce también fue desarrollado por Hadoop.

4.2.6.8. SQL (structured query language o lenguaje de consulta estructurado).

Lenguaje de computación diseñado para administrar base de datos relacionales, permitiendo especificar distintos tipos de operaciones tales como las que se mencionan en CRUD. (*Create, Read, Update, Delete*).

4.2.6.9. Visualización de datos

Son las herramientas de software que permiten la visualización de datos de manera gráfica que facilitan tanto el análisis y la representación de las mismas, tal es el caso de herramientas tales como *Tableau, SAP, RapidMiner, PowerBI, IBM Watson Studio*, entre otros.

4.2.7. Machine Learning

Como parte de la inteligencia artificial, *Machine Learning* utiliza un conjunto de algoritmos supervisados y No supervisados para realzar tareas relacionadas con el aprendizaje en los sistemas y aplicaciones, este aprendizaje es el proceso de proceso describir, analizar y aprender como si lo hiciera una persona.

Algunos algoritmos de aprendizaje automático son aquellos que permite predecir con base en la historia de los datos, regresión lineal, múltiple, logística, polinomial, support vector machine, árboles de regresión , entre otros.

Hay algoritmos de clasificación dentro del aprendizaje automático que permiten identificar una distinción conforme una etiqueta proporcionada al conjunto de datos, con ello el algoritmo clasifica la pertenencia del dato y determinar en

donde queda su clasificación, algunos algoritmos pueden ser, árboles de clasificación, *support vector machine*, *k means*, entre otros.

Hay tareas que permiten agrupar datos en grupos (cluster) no definidos de manera inicial, sino que la propia similitud y patrones semejantes encontrados en los datos se detecta un nuevo conocimiento y nuevos grupos que identifica a un conjunto de datos, estos algoritmos caen dentro del grupo de algoritmos No supervisados.

La figura 4, identifica una clasificación de estos algoritmos entre supervisados y No supervisados además de aprendizaje por reforzamiento y aprendizaje profundo que tienen que ver con el razonamiento y el uso de redes neuronales como para reconocer imágenes y rostros, entre otras cosas.

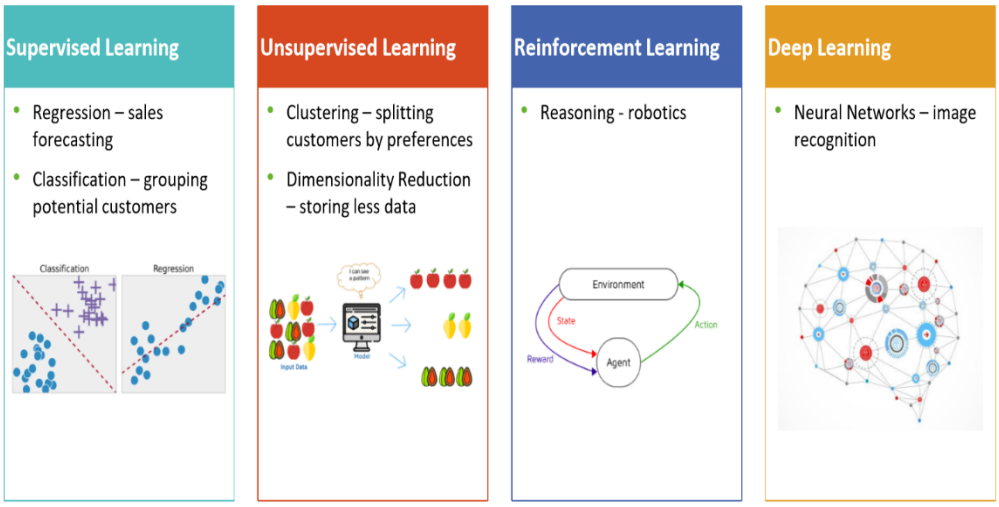


Figura 4. Clasificación de algoritmos de Machine Learning.
Fuente: (Clariba, s.f.)

La inteligencia artificial es la forma en la que se describe los objetos que simulan la inteligencia y *Machine Learning* es el motor que servirá para hacerlos funcionar. *Machine Learning* identifica patrones complejos a partir de enormes cantidades de información, procesándolos y determinar su comportamiento.

Los algoritmos de Machine Learning tienen la capacidad para mejorar sin algún tipo de ayuda externa les permite desarrollar modelos para descubrir tendencias, que es lo que necesita el consumidor y en qué mercado lo requiere.

4.2.8. Data Mining

Data Mining, se refiere al proceso de extraer conocimiento de bases de datos. Su objetivo es descubrir situaciones anómalas y/o interesantes, tendencias, patrones y secuencias en los datos (Molina, 2000).

Data Mining o Minería de Datos es el proceso completo de la obtención del conocimiento, intenta obtener patrones y modelos de la recolección de datos. Estos datos o modelos obtenidos recurrentemente suelen ser útiles o no requerir una valoración subjetiva por parte del usuario.

Algunos algoritmos de Data Mining cotidianamente comprenden de tres componentes:

- El modelo, que contiene parámetros que han de fijarse a partir de los datos de entrada.
- El de preferencia, que sirve para comparar modelos alternativos.
- El algoritmo de búsqueda, que viene a ser como cualquier otro programa de inteligencia artificial (IA). El criterio de preferencia suele ser algún tipo de heurística y los algoritmos de búsqueda empleados suelen ser los mismos que en otros programas de inteligencia artificial.

Las principales diferencias entre los algoritmos de Data Mining se hallan en el modelo de representación elegido y la función del mismo, es decir según la definición del objetivo y tarea.

4.2.8.1. Herramientas de Data Mining

Las herramientas de Data Mining empleados en el proceso de descubrir conocimiento en las bases de datos (KDD) se pueden clasificar en dos grandes grupos

- Técnicas de verificación, en las que el sistema permite evaluar hipótesis proporcionadas por el usuario.

- Métodos de descubrimiento, en los que se han de identificar patrones que puedan ser de interés y de forma automática, incluyendo en este grupo todas las técnicas de predicción (Valcárcel Asencios, 2004)

El resultado generado con la aplicación de algoritmos de minería de datos que pertenecen a los métodos de descubrimiento, puede ser de tipo descriptivo o predictivo. Las predicciones sirven para identificar pronósticos y tendencias, es decir aspectos futuros de los datos, en tanto que una descripción puede ayudar a su comprensión y entendimiento. (Valcárcel Asencios, 2004).

4.2.9. e-businnes

Se puede interpretar como el proceso de compra y venta por internet, la interacción con clientes, productos y proyectos, a través de tecnologías, comunicaciones múltiples y canales de distribución que se proyectan en el mercado en red. Este modelo se diseñó para aumentar la satisfacción del cliente, aumentar los ingresos y la productividad en la empresa.

Incorpora el uso estratégico de las tecnologías de la información y la comunicación para interaccionar con clientes, proyectos y socios a través de la comunicación múltiple y los canales de distribución (Siebel, 2001).

4.2.10. ¿Qué es Marketing?

El marketing es una disciplina que se dedica al análisis de comportamientos de clientes, productos, precios, ofertas, demandas por medio del cual individuos y empresas satisfacen sus necesidades de innovar, crear, modificar el proceso de mercadear y vender algún bien o servicio.

4.2.10.1. ¿Qué es Marketing Digital?

Es la proyección digital de una marca, producto o servicio y que está asociado al marketing de internet, existen compradores en todo el mundo, para todo lo que se pueda vender. Es aquí, donde se analiza que personas usan el internet, las necesidades que tienen, sus gustos y lo que recurrentemente están acostumbrados a comprar.

En esta la era del internet las diferentes variables o comportamientos online pueden ser medidos como:

- Tiempo activo en cada sitio.
- Accesos a páginas y dominios.
- Búsquedas rápidas.

Dentro de este mismo esquema se pueden catalogar algunos tipos de perfiles de internautas. Los cuales McKinsey y Media Metrix mencionan como: (Rivera, 2015)

- Simplificados.
- Surferos
- Conectores
- Rutinarios
- Buenos negociantes
- Amantes del deporte

Por otra parte, la función de marketing digital nace del aspecto tradicional innovándose con nuevas tecnologías, permite que las empresas se posicionen cercanos a los consumidores, mejoren las relaciones con éstos y hagan de sus clientes unos fans (aficionados y gustos por lo que venden) utilizando para ello diversas plataformas que hay en el mundo digital (Mariscal Suárez, 2018).

La llegada de las redes sociales ha permitido que el cliente sea una parte importante para la organización y que, mediante las mismas, el usuario esté en constante comunicación con la organización, generando oportunidades para la empresa en tiempo real. Es aquí en donde el paradigma Big Data tiene relevancia en apoyar en analizar la información y descubrir conocimiento viable para innovar en los procesos de ventas y mercadotecnia.

El marketing digital se ha posicionado en nuevos aspectos, actualmente es una de las herramientas necesarias para las empresas. A través de Internet, plataformas de comunicación digital, redes sociales y otras tecnologías, las empresas se están cada vez adentrando a la digitalización y apoyar en ello a las funciones de venta y mercadotecnia de los productos y servicios que ofrecen. (Mariscal Suárez, 2018).

4.2.10.2. Estrategias del marketing

Al emprender cualquier negocio siempre es necesario tener presente a quien se va a ofrecer el producto o servicio.

Segmentando, el mercado de consumo en diversas características comunes. Se seleccionan los segmentos del mercado meta y aquellos que se convertirán en mercados secundarios.

Si la competencia en el mercado primario al cual se enfoca tiene mucha competencia o en él se encuentra alguien que lo acapara, es mejor que tomar uno de los mercados secundarios para preparar la competencia explotando los puntos débiles que se obtengan del líder o los competidores.

Se identifican algunas estrategias del marketing:

- Estrategias de mercado meta
- Estrategia de producto
- Estrategia de precios
- Estrategia de distribución
- Estrategia de ventas
- Estrategia de promoción
- Estrategia de publicidad.

4.2.10.3. Tácticas para crear marketing digital.

Dentro de las tácticas de mercado figuran algunas que pueden ser útiles para crear esquemas de ventas o servicios capaces de llegar a cualquier usuario, pero no solo es que el consumista pruebe algunos productos sino el tratar de conservar la fijación de la marca producto servicio promovido, y quiera conocer más de ellos.

Estas tácticas tienen la intención de mantener un esquema de venta activo y renovando intereses para atraer cada vez más la atención de los mercados. Se pueden clasificar estas tácticas como:

- Atraer gente nueva: consiste en generar el interés a los productos y servicios a ofrecer.

- Retener gente: esta parte es la más difícil para cualquier mercado, ya que, no por realizar una sola venta o colocado algún servicio quiere decir que estos mismos sean de interés para el consumidor, por lo cual se debe de gestionar interés constante hacia el consumidor.
- Hacer venta: teniendo en cuenta que el producto o servicio cumple la necesidad del cliente se debe de consolidar la relación y hacer que el consumidor obtenga el servicio o producto.
- Atraer gente de vuelta. Ya consciente que el producto o servicio cumple con las medidas necesarias y para su venta, debe de tomar acciones para que el cliente siga con la marca ofrecer más oportunidades donde sienta que la posición de la empresa es estable y vuelva a consumir este producto o servicio.

El marketing digital puede apoyar en identificar segmentos clave de clientes dentro de la base de datos, así como comportamientos especiales de compra en los segmentos. Esto permite realizar estrategias y mensajes a diferentes audiencias de mercado (Jones, 2019).

4.2.10.4. Inbound Marketing

El Inbound marketing “es una estrategia se basa en atraer clientes con contenido útil y relevante, agregando valor en cada una de las etapas del recorrido del comprador”. Con el inbound marketing, los clientes potenciales encuentran la empresa a través de distintos canales como blogs, motores de búsqueda y redes sociales.

A diferencia del marketing tradicional, el inbound no necesita esforzarse por llamar la atención de los clientes potenciales, ya que, al crear contenido diseñado para abordar los problemas y las necesidades de los clientes ideales, atraerá prospectos calificados y generarás confianza y credibilidad para la empresa. (HubSpot, 2017).

Las formas del Inbound marketing son blogs de actualidad y noticias de ventas y ofertas; campañas en social media y correos electrónicos; Search Engine

Optimization (SEO), videos virales, seminarios basados en Web, entre otros. (Patrutiu-Baltes, 2016).

Redes sociales como lo son Facebook, Instragam, Twitter, WhatsApp, entre otras se han convertido en herramientas accesibles en el mercado en donde existen vendedores y compradores. Hoy en día el internet es un mecanismo de enlace, una fuente de información y un buscador de datos, en donde las organizaciones deben aprovechar las condiciones para generar información estratégica que logre captar la atención del consumidor, con ello aumentar las ventas y generar utilidades. (Mariscal Suárez, 2018).

4.2.10.5. Outbound Marketing

El marketing tradicional está asociado con el outbound marketing, lo que significa que la estrategia de marketing es llevar los productos a los clientes, mientras que el marketing digital es sinónimo del término de inbound marketing, cuyo objetivo principal es ganar el interés objetivo por parte de los clientes. (Patrutiu-Baltes, 2016).

Finalmente, el outbound marketing (marketing de salida) tiene como objetivo promover los productos y servicios para el audiencia directamente, mientras que el inbound marketing ayuda a resaltar bienes y servicios de manera indirecta a través de tecnología citada que permite dar a conocer los productos y servicios, en el contexto en el que los consumidores se identifican como objetivo de la empresa. (Kudupu, 2018).

4.2.11. ¿Qué es el ADN Digital?

El ADN digital es la marca que tienen los usuarios, su propio código de conducta que los mueve como se manejan dentro de la red, los movimientos que varían de usuario a usuario.

Esto genera un rastro digital de los aparatos electrónicos que suelen utilizar, desde pagos realizados con tarjetas de bancos, las conexiones o visualizaciones de los usuarios en equipos de cómputo y teléfonos inteligentes, todo esto enriquece al internet de las cosas y se concentra en el Big Data.

La figura 5 identifica que como personas o empresas en el mundo digital al realizar cualquier acontecimiento en la red o internet, llámese operación, transacción, consulta, búsqueda, compra, petición, solicitud, registro, mail, descarga, carga, entro otras acciones, existe una evidencia y una historia que deja huella. Cada uno de los usuarios de la WEB deja un registro de las acciones que realiza. Es ahí donde esa información y características de uso en la WEB e internet puede ser convertida en conocimiento y ser aprovechado por las empresas para promover productos y servicios dado las preferencias y gustos de los clientes y usuarios. A esto rastro es lo que se le conoce como ADN digital.

Figura 5. ADN Digital y ADN de la empresa
Fuente: (Mejía Llano, 2019)

4.2.11.1. Empresas y disrupción digital

En un estudio plasmado en el artículo por Falcó Rojas (2019) menciona que las empresas están comenzando a evolucionar digitalmente. Esto sugiere que las organizaciones están comenzando a tomar la disrupción digital más en serio (Falcó Rojas, 2019).

Entendiéndose por disrupción digital los cambios y adopciones que las empresas tienen para adaptarse a las nuevas tecnologías y modificar sus procesos para enfrentar los retos actuales del mercado.

En el artículo Falcó Rojas (2019) en sus resultados de medir y evaluar la madurez digital de las empresas, se afirma que ya no hay que convencer a nadie de la importancia del cambio hacia la digitalización. El 85% de los encuestados está de acuerdo con que ser un negocio digital es imprescindible para el éxito de la empresa. (Falcó Rojas, 2019).

4.3. Desarrollo

Se analizan algunos ejemplos en donde se han desarrollado estrategias donde el Big Data ayuda a la toma de decisiones, se muestran algunos resultados, tomando en cuenta estos casos. Se analiza el cómo la conducta de usuarios es una fuente de información relevante.

4.3.1. ¿Cómo utiliza Amazon el Big Data?

Empresas como Google, Facebook, Amazon, han basado su estrategia empresarial sobre el concepto de Big Data (Revuelta Bayod, 2018).

Debido al uso del Big Data, la empresa Amazon ha aumentado de manera importante sus ventas, además, ha logrado reestructurar completamente su estrategia de negocios.

Mediante esta tecnología, el proceso de estar en contacto con los clientes de para empresa se ha innovado y evolucionado de manera constante, hasta poder indicar que el conocimiento mismo de los consumidores es la clave del éxito de Amazon.

Gracias a distintos algoritmos complejos, Amazon ha personalizado completamente el proceso de compra, habilitando a cada usuario vea contenidos específicos y adaptados a sus gustos cuando utiliza el sitio. La frase de "los clientes interesados en este producto también compraron esto" es, en efecto, un recurso bien aprovechado por herramientas de Big Data.

¿Ha pasado que las empresas digitales recomienden un producto en el que hace tiempo se estaba pensando adquirir? ¿o uno que vagamente se había considerado adquirir? Estas acciones son producto del análisis recurrente de la información brindada por cada cliente que, combinada con el uso de Inteligencia Artificial puede determinar preferencias y distinciones para los usuarios.

En un mundo que se transforma digitalmente en el camino de disponer de tecnología que permita hacer las cosas fáciles para la vida, la empresa Amazon ha aprovechado las herramientas para mejorar las relaciones con los clientes y acercarse de manera efectiva a sus consumidores.

4.3.2. Netflix: Las claves del éxito basado en Big Data

Netflix extrae datos de los patrones de visualización de películas para saber cuál es el interés del usuario y lo utiliza para tomar decisiones.

Netflix recopila y analiza datos del consumo que realizan sus clientes; lo que buscan, lo que adquieren, lo que ven, sus etiquetas; dónde, cuándo y cómo consumen cada contenido. Sin duda, lo importante está en cómo se benefician de esta analítica de datos para mejorar sus servicios en varios aspectos:

- Segmentar al cliente.
- Realizar recomendaciones individualizadas de contenido conforme gustos.
- Ofrecer una plataforma de contenidos simple, usable y personalizada para cada consumidor, de acuerdo a su experiencia.
- Pronosticar claves de éxito y tendencias

En Netflix todo ese conocimiento, esa inteligencia, hace que la definición de cada producto, cada programa, deje de ser un arte y se convierta en una ciencia.

La información que se ha recopilado durante mucho tiempo es lo que les soporta decidir qué películas y series incluir en sus ofertas. Y de esta forma la empresa Norte Americana cubre todas las necesidades de los países en donde se utiliza. En cada uno de ellos, ofrece una programación acorde a los datos que ha recopilado, almacenado y analizado de los espectadores de esa región.

La tecnología Big Data, ha sido utilizada por Netflix, de tal forma que ha pasado de ser un distribuidor de contenido a transformarse en una de las productoras de mayor éxito.

Netflix es capaz de determinar cuántas horas debe un suscriptor utilizar el servicio para evitar que entre en situación de cliente poco habitual. De tal forma que, en el momento en que detecta de que la cuota media de uso es menor al promedio que ha identificado, toma medidas para incrementarla mediante ofertas y promociones personalizadas de acuerdo a el análisis de los datos.

4.3.3. Analítica descriptiva, predictiva y prescriptiva en el Marketing Digital

Un nuevo enfoque se está dando al realizar análisis en los datos en las organizaciones, involucra aspectos de Big Data y de Marketing Digital, estos aspectos tienen que ver con analítica descriptiva, predictiva y prescriptiva.

La analítica descriptiva, intenta responder a preguntas sobre lo que sucedió en el pasado, es el obtener conocimiento de la historia de los datos, por lo regular, todo esto lo que trata es de informar. Se intenta responder a inquietudes tales como: ¿cuál es el cliente o producto más rentable? ¿qué cantidad de ingresos por ventas fueron generados en el primer trimestre del año?, entre otras.

El análisis predictivo tiene que ver con pronosticar y anteponerse a situaciones futuras, conocer que va a suceder en términos de probabilidad, regresión, utilizando para ello la historia y el conocimiento adquirido. Implica determinar tendencias y patrones del futuro. Algunas inquietudes a responder con este tipo de análisis son las siguientes: ¿cuál es el número aproximado de un centro de llamadas en el próximo trimestre? ¿cuál es la siguiente mejor oferta para este cliente?, ¿qué clientes es probable que se alejen?, ¿qué cliente puede reaccionar emotivamente para comprar o caso contrario para no comprar?, entre otras.

El análisis prescriptivo por otra parte, responde a preguntas asociadas con ¿cómo puedo manejar esto?, ¿cuál es la mejor alternativa? Este es el momento en el que el análisis se vuelve operativo. Es totalmente dependiente del negocio y del caso. Algunos de los ejemplos utilizados para demostrar el punto incluyen: Saber

que una persona tiene una mayor probabilidad de emocionarse a favor, por lo que se le presenta una oferta o promoción. Determina el historial de visualización de un cliente en la Web, y luego sugiere productos que el consumidor pueda ver y leer (Jones, 2019).

4.3.4. Propuesta de implementación Big Data

La propuesta es un bosquejo en lo general de las estrategias que debieran emprender las empresas del giro comercial, consiste en tomar en cuenta estrategias de marketing digital como son:

- Inbound Marketing
- Outbound Marketing
- Crear identidad digital
- Trabajar la Ley de enfoque
- Posicionarse en buscadores

Teniendo en cuenta ya algunos aspectos de estrategias de Marketing, se toma en cuenta algunos casos de éxito, donde el posicionamiento del Marketing pudiera ser muy asertivo, demostrando que se pueden implementar nuevas estrategias para los servicios utilizados en la actualidad. Al surgir nuevas necesidades se crean nuevas estrategias y así sucesivamente.

Para toda empresa comercial, se propone, que en lugar de utilizar marketing tradicional, hay una nueva tendencia donde el posicionamiento de la marca es la principal meta, optimizando los recursos que la red puede ofrecer como: redes sociales, inbound marketing, outbound marketing, anuncios en páginas en general no siendo contenido basura (spam), los cuales serán capaz de atraer la atención del consumidor y serán enfocados directamente a las preferencias del cliente y a las necesidades de la empresa.

Dentro de las tecnologías que se pueden utilizar para posiciones marca y negocio están las siguientes:

- Anuncios en Youtube.

- Facebooks ads.
- Google Adworks
- Twitter.
- SEM.
- Display Ads.

Estas tecnologías son adecuadas para mostrar al mayor porcentaje de usuarios un producto o servicio, dichas tecnologías forman parte de la información que recolecta el Big Data para enriquecer las estrategias de marketing digital.

Una ventaja de utilizar en las empresas comerciales el Big Data y el Marketing digital es que permite ampliar los sectores de posibles ventas por lo cual se incrementa la producción de un bien o servicio, genera más ganancias, produce más campos de estudio ya que genera una cantidad ilimitada de información, los mensajes claros y objetivos producirán una relación más clara con los clientes.

Las desventajas que se pudieran presentar son que la información no sea clara y objetiva, el impacto no sea el proyectado, el manejo de la información no sea el adecuado, no sea bien visto por la comunidad, nocivo a la vista, pudiera tener resistencia al cambio por parte de los usuarios. Al optar por un proyecto de proyección este podría saturar la imagen de la empresa, producto e imagen de la misma obteniendo un resultado negativo, el cual podría considerarse como perdidas en cualquier circunstancia.

El impacto que se tendría al aplicar el Big Data al marketing digital sería el incremento de las ganancias económicas tanto real y potencialmente, debido a que se amplía el campo de posibles clientes al ofrecer un producto o servicio de acuerdo a las necesidades de cada usuario, utilizando las ventajas que ofrece las redes, utilizando como beneficio a favor lo que es las búsquedas que realicen mediante el navegador, redes sociales, entre otros.

Tal impacto tiene un beneficio para el sector privado y el sector público ya que se aumenta la productividad, la calidad de vida de los ciudadanos, así como ser más competitivos.

Conclusiones y Recomendaciones

A medida que la información avanza y los medios de comunicación se amplían el Big Data y el Marketing Digital surgen como una fase de comunicación e información que emerge a partir de la revolución tecnológica y como esta iniciada en la década del setenta. En este sentido, si bien ofrece oportunidades para alcanzar tanto mejoras en competitividad y productividad como en opciones para la mejora en la calidad de vida, con potenciales beneficios para los países en desarrollo, se requieren importantes desarrollos en el área de Tecnología de Información y Comunicaciones (TIC).

En realidad, Big Data abre oportunidades, pero el análisis de datos masivos para la toma de decisiones inteligentes pone de manifiesto la necesidad de generar capacidades para superar la brecha digital. Por lo tanto, la materialización de los potenciales beneficios para los países en desarrollo requiere la elaboración de políticas activas y específicas que tengan en cuenta la generación y apropiación de rentas informacionales, el manejo de la privacidad en cuanto a los datos personales, el desarrollo de capacidades para la creación de valor y la difusión de la información y el conocimiento para contribuir a la reducción de desigualdades socioeconómicas.

Las recomendaciones para tener en cuenta sería no resistirse al cambio, explorar los nuevos horizontes que ofrece el Big Data y ampliar así la proyección que se desea obtener, aprovechar las nuevas plataformas donde el marketing digital deja de ser funcional para optar por otras alternativas.

En la actualidad el *inbound marketing* marca una nueva tendencia. Los *AdWords* de *Google* son la premisa en elección para mostrar las características de los productos, así como, los *displays ads* los cuales al generar *banners* para la invasión visual pasiva en sitios de mayor trámite de usuarios son verdaderamente convenientes hoy en día.

Referencias

Ahumada Tello, A., & Perusquia Velasco, J. M. (2016). Inteligencia de negocios: estrategia para el desarrollo de competitividad en empresas de base tecnológica. *Contaduría y Administración. Revista Internacional. UNAM*, 127-158.

Antón Carranza, M. (2017, 07 14). Identificación del talento en la Organización: El Big Data aplicado al fútbol. *Trabajo de Fin de Grado. Grado en Administración y Dirección de Empresas*. Valladolid, Valladolid, España: Facultad de Ciencias Económicas y Empresariales. Universidad de Valladolid, España.

Barranco Fragoso, R. (2012, 06 18). *IBM.* Retrieved from IBM Developer: https://www.ibm.com/developerworks/ssa/local/im/que-es-big-data/index.html

Castillo, J. N., Garcés, J. R., & Navas, M. P. (2017). Base de Datos NoSQL: MongoDB vs. Cassandra en operaciones CRUD (Create, Read, Update, Delete). *Revista Publicando. 4 No 11. (1). 2017, 79-107. ISSN 1390-9304*, 79-107.

Clariba. (s.f., s.f. s.f.). *Clariba website*. Retrieved from Machine Learning Empresarial ¿Qué es la Inteligencia Artificial?: https://es.clariba.com/machine-learning-for-business

Cukier, K. (2014). Los big data y el futuro de los negocios. In F. González, *Reinventar la empresa en la era digital* (p. 451). Madrid, España: BBVA OpenMind.

Falcó Rojas, F. R. (2019). Análisis empírico de la transformación digital en las organizaciones. *International Journal of Information Systems and Software Engineering for Big Companies (IJISEBC)*, 35-52.

García Merino, E. M., & García Merino, M. J. (2017). Análisis de los Modelos de Inteligencia de Negocios basados en Big Data en las Pymes del Ecuador. *Revista Científica. Ciencia Tecnología*, 1-12.

Guevara, R. E. (2018, 09 01). Servicios de Cómputo en la Nube. Cloud Computing. *Trabajo monográfico para obtener el grado de Ingeniero en Redes*. Chetumal, Quintana Roo, México: Universidad de Quintana Roo. División de Ciencias e Ingeniería.

Gutiérrez Puebla, J. (2018). Big Data y nuevas geografías: la huella digital de las actividades humanas. *Documents d'Anàlisi Geogràfica. eISSN: 2014-4512*, 195-217.

Jones, H. (2019). *Ciencia de los Datos. Lo que saben los mejores científicos de datos sobre el análisis de datos, minería de datos, estadísticas, aprendizaje automático y Big Data que usted desconoce.* México: Amazon Mexico Services, Inc.

Kudupu, P. (2018, 02 12). *Web Snippets.* Retrieved from Web Snippets: http://www.prathapkudupublog.com/2018/01/7-vs-of-big-data.html

Lenz, R. (2019). Big Data: Ethics and Law. *SSRN Electronic Journal. DOI: 10.2139/ssrn.3459004*, 37.

Luhn, H. P. (1958). A Business Intelligence System . *IBM Journal*, 314-319.

Mariscal Suárez, T. E. (2018, 01 19). Aproximación teórica del Big Data sobre el marketing Digital. *Componente práctico del examen complexivo previo a la obtención del grado de Ingeniería en Marketing*. Guayaquil, Guayaquil, Ecuador: Universidad Católica de Santigo de Guayaquil. FACULTAD DE ESPECIALIDADES EMPRESARIALES CARRERA DE MARKETING.

Maroto, C. (2018). Big Data y su impacto en el sector público. *Harvard Deusto Business Review*, 16-25.

Mejía Llano, J. C. (2019, 10 29). *Marketing Digital, Social Media y Transformación Digital*. Retrieved from TRANSFORMACIÓN DIGITAL: INCORPORE EL MUNDO DIGITAL EN EL ADN DE SU EMPRESA + VIDEO: https://www.juancmejia.com/juan-carlos-en-los-medios/transformacion-digital-incorpore-el-mundo-digital-en-el-adn-de-su-empresa-video/

Molina, L. (2000). *Torturando los Datos hasta que Confiesen. Departamento de Lenguajes y Sistemas Informáticos.* Catalunya: Departamento de Lenguajes y Sistemas Informáticos.

Patrutiu-Baltes, L. (2016). Inbound Marketing - the most important digital marketing strategy. *Bulletin of the Transilvania University of Braşov. Series V: Economic Sciences*, 61-68.

Peña Ayala, A. (2006). *Inteligencia de Negocios: Una Propuesta para su Desarrollo en las Organziaciones.* México: Instituto Politécnico Nacional. Dirección de Publicaciones. Retrieved from https://es.calameo.com/read/0009834562d4384832b9e

Puyol Moreno, J. (2014). UNA APROXIMACIÓN A BIG DATA. *Revista de Derecho UNED , núm. 14*, 471-505.

Revuelta Bayod, M. J. (2018). Big Data: crisis y nuevos planteamientos en los flujos de comunicación de la cuarta revolución industrial. *Área Abierta. Revista de comunicación audiovisual y publicitaria. ISSN: 1578-8393 / ISSNe: 1578-8393*, 309-324.

Rivera, M. (2015, 07 15). *Pymempresario.* Retrieved from Estrategia de marketing online: https://www.pymempresario.com/2015/07/estrategia-de-marketing-online/

Rodríguez, M. F. (2016). Distribución de un Analizador de Contenido de Twitter utilizando el Framework Hadoop Map-Reduce. (C. d. Aires, Ed.) *PLADEMA, Universidad Nacional del Centro, Tandil, Buenos Aires, Argentina*.

Ruiz García, E. (2016, 18 01). Estudio y Evalaución de Sistemas Big Data de tratamiento de información. *Trabajo de Fin de Grado. Escuela Técnica Superior de Ingenieros de Telecomunicación*. Madrid, Madrid, España: Universidad Politécnica de Madrid.

SEGITTUR. (2019, 06 01). SEGITTUR turismo e innovación. Madrid, Madrid, España. Retrieved from SEGITTUR: https://www.segittur.es/opencms/export/sites/segitur/.content/galerias/descargas/documentos/Presentacin-Destinos-Tursticos-Inteligentes.pdf

Siebel, T. M. (2001). *Principios del e-Business.* Barcelona, Buenos Aires, Edo de México, Santiago de Chile: Ediciones Granica.

Valcárcel Asencios, V. (2004). Data Miningyel Descubrimiento del Conocimiento. *Revista de la Facultad de Ingeniería Industrial. Vol. (7) 2. UNMSM*, 83-86.

Capítulo 5

Análisis comparativo a través del uso de R y Python enfocado al análisis descriptivo de datos de una entidad financiera

Nohemí García Hernández
Tecnológico Nacional de México. Instituto Tecnológico de Durango
00041233@itdurango.edu.mx

Claudia Elizabeth Serrato Bacio
Tecnológico Nacional de México. Instituto Tecnológico de Durango
00041233@itdurango.edu.mx

Marco Antonio Rodríguez Zúñiga
Tecnológico Nacional de México. Instituto Tecnológico de Durango
mrodriguez@itdurango.edu.mx

5.1. Introducción

Las instituciones financieras en la búsqueda de rendimientos de sus activos cuentan con diferentes productos dirigidos a sus clientes. En particular, el otorgamiento de créditos grupales a personas físicas y morales, es uno de sus principales servicios financieros.

Las instituciones buscan que los ingresos y los egresos por algún producto sean tal, que los primeros sean iguales o mayores a los segundos; sin embargo, de

manera natural al otorgar un crédito grupal las instituciones financieras se encuentran ante un grupo que pudiera o no cubrir sus obligaciones adquiridas.

Otorgar un crédito a clientes minoristas como no minoristas, es el negocio principal de una institución financiera. Al hacerlo, las instituciones requieren tener sistemas adecuados para decidir a quién se le puede otorgar un crédito. La puntuación de crédito es clave en la evaluación de riesgos para analizar y cuantificar el riesgo crediticio de un posible deudor. Fundamentalmente, la puntuación de crédito tiene como objetivo cuantificar la probabilidad de que un cliente pague la deuda. El resultado del ejercicio de puntuación de crédito es una puntuación que refleja la solvencia de un cliente.

En los últimos períodos, las instituciones financieras han recolectado una gran cantidad de información que describe el comportamiento predeterminado de sus clientes. Algunos ejemplos son información histórica sobre la fecha de nacimiento, el sexo, los ingresos, el estado laboral, entre otros, de un cliente. Todos estos datos se han almacenado muy bien en enormes bases de datos (por ejemplo, relacionales) o almacenes de datos.

Al mismo tiempo, las instituciones financieras han acumulado considerable experiencia empresarial sobre sus productos de crédito. Actualmente el objetivo de la puntuación de crédito es analizar a los clientes de alto y bajo riesgo, con más detalle y elaborar un modelo de decisión basado en la estadística, que permita determinar futuras solicitudes de crédito y, en última instancia, decidir cuáles aceptar y cuáles rechazar.

Para los clientes históricos, se sabe cuáles resultaron ser buenos y malos pagadores. Este estado bueno/malo es ahora la variable de destino binario y que será de gran utilidad en el momento de calcular la puntuación sobre los clientes. El objetivo de esta puntuación de crédito es cuantificar esta relación de la forma más precisa posible para ayudar en las decisiones del otorgamiento del crédito, al momento del monitoreo y de la gestión. Las instituciones financieras obtienen una puntuación de los patrones, en la solicitud de préstamo, (generalmente préstamos, compromisos de préstamo y garantías).

Una vez que se tenga el modelo de puntuación de crédito construido, se podrá utilizar para decidir si la solicitud de crédito debe ser aceptada o rechazada, o para derivar la probabilidad de un incumplimiento futuro. En resumen, la puntuación de crédito es una herramienta clave de gestión de riesgos para que una institución financiera administre, comprenda y modele de manera óptima el riesgo crediticio al que está expuesto.

Una justificación de este trabajo es que el uso de los lenguajes de programación R y Python, son herramientas que permiten realizar análisis de datos con aspectos descriptivos e inferenciales entre otras cosas.

Un estudio comparativo entre estas dos herramientas permitirá tener una base para promover propuestas de uso en empresas financieras de cualquier tipo en relación a los procesos de análisis de datos.

Ahora bien, con respecto a las entidades financieras el uso de R y Python se convertirá en dos herramientas que pudieran dar soporte complementario a los análisis de riesgos en el otorgamiento de crédito, así como conocer la cartera en riesgo y con ellos lograr disminuir eventualidades que no beneficien a la entidad financiera, en base a sus procesos.

Como objetivo general se busca realizar un estudio que permita comparar las características principales entre las herramientas de R y Python para análisis descriptivos de datos a gran escala. Por medio de estas herramientas desarrollar una propuesta viable para la aplicación en aspectos de evaluación de riesgos en el otorgamiento de créditos.

De manera específica se quiere lograr lo siguiente:

- Identificar los datos estadísticos de una empresa financiera a evaluar.
- Identificar diferencias entre los lenguajes de programación R y Python
- Identificar elementos de comparación para aspectos descriptivos entre R y Python.
- Determinar en base a la comparación de los lenguajes de programación R y Python cuál es el más viable para realizar el análisis de datos.

- Identificar aspectos o metodologías que existen para evaluar riesgos crediticios en los bancos.
- Establecer elementos para el análisis descriptivo en el conocimiento del cliente.
- Determinar qué datos se requieren para realizar la evaluación de riesgos crediticios.

Se pretende lograr que en un futuro la empresa financiera adopte el uso de algún lenguaje de programación planteados en este proyecto para el análisis de datos para la evaluación de riesgos en el otorgamiento de créditos a sus clientes.

Un aspecto importante que se espera lograr en este documento, es la difusión de los lenguajes de R y Python para que un futuro la empresa financiera adopte el uso de algún lenguaje de programación planteados en este proyecto para el análisis de datos para la evaluación de riesgos en el otorgamiento de créditos a sus clientes. Con ellos se logrará disminuir la cartera en riesgo y evitar pérdidas para la empresa financiera.

Además, para la empresa financiera, se visualiza reducción de costos con utilización de alguno de estos lenguajes, ya que no será necesario adquirir o dar mantenimiento a algún paquete estadístico con licencia, ya que R y Python son gratuitos.

5.2. Marco de referencia

Los datos son un elemento y recurso importante para las empresas y organizaciones. Si no se tienen las herramientas adecuadas para procesarlos y signifiquen algo importante, no tienen valor alguno. La Ciencia de los Datos es un campo multidisciplinario que aplica técnicas matemáticas, estadísticas y computacionales y se puede aplicar a cualquier área o giro económico-laboral tales como sociología, psicología, salud, industria, comercio, investigación, educación, turismo, deporte, gobierno, política, entre muchos otros.

Los datos vienen en diferentes formas, en un nivel avanzado, existen tres categorías principales; estructurados, semiestructurados y no estructurados. Los científicos de los datos son los responsables de recopilar, analizar e interpretar grandes cantidades de datos para ayudar a empresas y organizaciones en la toma de decisiones.

La Ciencia de los Datos es un paradigma interdisciplinario, se manejan varias herramientas, principios de aprendizaje automático y algoritmos cuyo propósito es predecir patrones de grandes volúmenes de información.

¿Qué tan diferente es de la estadística? La figura 1 tiene la respuesta. Se observa que el analista puede afrontar retos de administración de análisis y exploración de datos, por el contrario, el científico de datos involucra además de exploración de datos, aspectos de aprendizaje automático e ingeniería de datos sin descuidar por supuesto los aspectos del negocio.

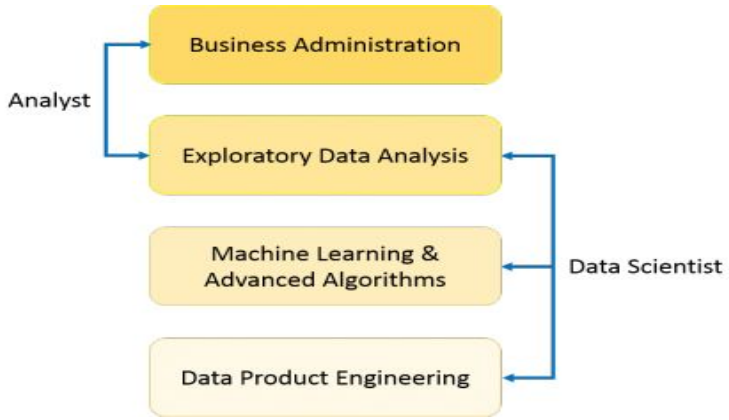


Figura 2. Analista de Datos vs. Científico de datos (Jones, 2019).

Un analista de Datos realiza el análisis de datos exploratorios a nivel de la administración del negocio, lo que está sucediendo al procesar el historial de los datos y un científico de los datos, explicara cómo extraer información, además que utilizara diferentes algoritmos de aprendizaje automático para resaltar la ocurrencia de un evento en el futuro (Jones, 2019).

En la vida diaria, es común que se recolecte información que sirven para tomar decisiones, va desde escuchar el reporte meteorológico y decidir que ropa

usar, así como decidir qué camino tomar para llegar a determinado lugar, según el tráfico vial.

El análisis de los datos es el proceso de examinar sistemáticamente los datos con el propósito de destacar información útil, ya que es la base de la investigación científica. Lehkyi, (2020) menciona que el análisis de datos se puede dar en cuatro pasos: entender el problema, coleccionar datos, procesar datos y tomar decisiones (Lehkyi, 2020).

De la misma manera, los administradores de empresas deben tomar decisiones todos los días. Como el administrador de negocios no se deben tomar decisiones basadas en la intuición o discernimientos basados en la experiencia. Para tomar una buena decisión se pueden aprender procedimientos y métodos que ayudaran a tomar decisiones basadas en hechos concretos. Estos procedimientos y métodos implican recolección, presentación, elaboración de resúmenes de un conjunto de datos y obtener conclusiones acerca de tales datos, entonces estará haciendo Estadística.

Son muchas las situaciones en las que tomar decisiones es un papel muy importante. Para ello se tiene que conocer la situación concreta que se está analizando por lo que se deben manejar datos, analizarlos y presentarlos. Para todo esto existe la Estadística.

5.2.1. Estadística

Sin duda alguna, la estadística ocupa uno de los lugares más importantes dentro de las investigaciones científicas, ya que por medio de esta se realizan evaluaciones cuantitativas sobre las hipótesis de investigación, que posteriormente desarrollan modelos predictivos, se llegan a estimar algunos parámetros y se pueden analizar experimentos.

La estadística es la rama de las matemáticas que examina las formas de procesar y analizar datos. Ofrece procedimientos para recolectar y transformar los datos de manera que sean de utilidad para quienes toman decisiones en los negocios. (Levine, Krehbiel, & Berenson, 2006)

Los datos son información específica de hechos, que permite realizar estudios, analizarlos y conocer más sobre ellos. La información es el conjunto de los datos que tienen un significado.

Los datos al inicio son imperfectos, en el sentido que no brinda información de utilidad. Es necesario aprender métodos que permitan obtener información a partir de datos observados y analizados para comprender mejor la situación que los mismos presentan. Existen muchos métodos estadísticos cuyo propósito es ayudar a exponer las características sobresalientes e interesantes de los datos que pueden ser usados en casi todas las áreas del conocimiento.

La estadística es un conjunto de técnicas y métodos científicos que permiten al investigador interpretar la información numérica, elegir muestras representativas para realizar inferencias, contrastar hipótesis, estimar y predecir mediante relaciones causa- efecto y tomar decisiones. (Saenz, 2005).

El uso de algún programa de computación dirigido a la estadística, es importante para la ciencia básica como para la aplicada; ya que por medio de este existen muchas posibilidades de automatización de los cálculos estadísticos que son muy complejos en lo que se refiere al análisis de datos

Existen aplicaciones en la investigación científica con altos costos en sus licencias (SPSS, Minitab, Statgraphic, MS Excel, entre otros). Sin embargo, en los últimos años han surgido algunas aplicaciones que son de software libre, tal es el caso de R y Python.

5.2.1.1. Estadística Descriptiva

Es el conjunto de técnicas para analizar, describir e interpretar los datos recolectados sobre un fenómeno de interés, con el fin de tomar decisiones, obtener conclusiones o plantear hipótesis (Jorge Andrés Alvarado Valencia, 2008).

Es un conjunto de técnicas numéricas y graficas con las que se intenta descubrir la estructura de un conjunto de datos. (Saenz, 2005)

Para analizar datos, es posible servirse de un conjunto de herramientas y de técnicas conocidas de la Estadística Descriptiva. Antes de ello, resulta necesario clasificar las variables que caracterizan a los datos en dos tipos: escalares y categóricas:

- **Variables escalares:** son variables susceptibles de medición cuantitativa, también denominadas variables métricas. Es decir, son representaciones numéricas reales de una característica de interés como el tiempo de producción, el número de defectos de una pieza o su longitud. Estas variables, a su vez se clasifican en:
    - **Variables escalares discretas:** son aquellas que pueden tomar solo ciertos valores dentro de un intervalo determinado de los números reales. Ejemplo: número de hijos por familia, números de trabajadores por empresa, entre otros.
    - **Variables escalares continuas:** pueden tomar todos los valores posibles dentro de un intervalo determinado de los números reales. Ejemplo: ingreso de un hogar, peso en kilogramos de una persona, entre otros.
- **Variables categóricas:** son aquellas variables cuyos posibles valores no son susceptibles de medición cuantitativa directa pero sí pueden ser clasificados. También conocidas como variables no métricas. Ejemplo: el género de una persona, su estado civil o su estado socioeconómico. Las variables categóricas se dividen en dos grupos: nominales y ordinales.
    - **Nominales:** sus valores se identifican con una clase o grupo de elementos de acuerdo con una característica. Ejemplo: color de un carro, estado civil de una persona.
    - **Ordinales:** variables cuyos posibles valores, una jerarquía u ordenación. Sus valores representan una propiedad de orden de un conjunto o una posición relativa pero no son una representación numérica real de la variable, por ejemplo: nivel educativo de una persona, nivel de preferencia por un producto, en escala de 1 a 5.

5.2.2. Ciencia de los Datos

Es un campo interdisciplinario que involucra métodos científicos, procesos y sistemas para extraer conocimiento o un mejor entendimiento de datos en sus diferentes formas, ya sea estructuradas o no estructuradas. (Alex, 2015)

5.2.3. Análisis de datos.

5.2.3.1. Análisis

El análisis: implica categorización, ordenamiento, manipulación y resumen de datos para responder a ciertas preguntas de investigación de un tema en específico. (Saenz, 2005).

La distinción y separación de las partes de un todo hasta llegar a conocer sus principios o elementos.

5.2.3.2. Análisis de datos

Es un proceso en cual permite inspeccionar, limpiar y transformar datos con el objetivo de obtener información de utilidad lo que sugiere conclusiones y apoya en la toma de decisiones.

5.2.3.3. Análisis comparativo

Proceso para identificar mediante un análisis, diferentes aspectos que se relacionan o no entre dos o varios objetos, semejanzas y/o similitudes. Implica examinar dos o más cosas, es una expresión de las semejanzas de dos o más cosas.

5.2.4. Minería de datos

También denominada exploración de datos, es un campo de la estadística y la ciencia de la computación referida al proceso que intenta descubrir patrones en grandes volúmenes de conjuntos de datos. Utiliza los métodos de la inteligencia artificial, aprendizaje automático, estadística y sistemas de bases de datos. El objetivo principal es extraer información de un conjunto de datos y transformarla en una estructura comprensible para su uso posterior. (Maimon & Rokach, 2010)

5.2.5. Machine Learning

Es una disciplina científica de la inteligencia artificial que crea sistemas que aprenden automáticamente. Aprender en este contexto significa, identificar patrones complejos en millones de datos. De acuerdo a Arthur Samuel en 1959, le da a las computadoras la habilidad de aprender sin ser explícitamente programadas.

Las técnicas de aprendizaje automático (Machine Learning) están en el pleno desarrollo de transformación y se puede definir como, un conjunto de métodos capaces de detectar automáticamente patrones en los datos para realizar predicciones, o para tomar otros tipos de decisiones en entorno de incertidumbre. (Management Solutions, 2018).

Los componentes principales del aprendizaje automático se pueden clasificar en cuatro grupos:

- Las fuentes de información, que pueden aportar datos tantos estructurados como no estructurados, y que son la base del reto de componentes.
- Las técnicas y algoritmos para el tratamiento de información no estructurada como pueden ser: texto, voz, video, entre otros; y para la obtención de patrones a partir de datos.
- La capacidad de autoaprendizaje, que permite que el algoritmo se adapte a los cambios en los datos.
- El uso de sistemas y software como vehículo para la visualización de la información y la programación.

Machine Learning utiliza algoritmos para convertir un conjunto de datos en un modelo predictivo. El tipo de algoritmo que funciona mejor (supervisado, no supervisado, clasificación, regresión, entre otros.) depende del tipo de problema que está resolviendo, los recursos informáticos y la naturaleza de los datos. (Robertson, 2019)

Las técnicas de aprendizaje automático se emplean para localizar en los datos y crear modelos que pronostiquen los resultados futuros. Hay disponible una amplia gama de algoritmos de aprendizaje automático, incluidas regresiones

lineales y no lineales, algoritmos de clasificación, de clustering, redes neuronales, máquinas de vectores de soporte, árboles de decisión, entre otros.

Hay dos categorías principales de problemas que el aprendizaje automático a menudo resuelve: regresión y clasificación. La regresión es para datos numéricos (ejemplo, ¿cuál es el ingreso probable para alguien con una dirección y profesión determinadas) Y la clasificación es para datos no numéricos (ejemplo, ¿el solicitante incumplirá con este préstamo?).

5.2.5.1. Análisis predictivo

Implica una variedad de técnicas estadísticas de modelización, aprendizaje automático y minería de datos que analiza datos actuales e históricos para hacer predicciones acerca del futuro o acontecimientos no conocidos.

Los problemas de predicción, son un subconjunto de problemas de regresión para datos de series temporales. Los problemas de clasificación a veces se dividen en binarios (sí o no) y problemas de categorías múltiples (animales, vegetales o minerales).

El análisis predictivo emplea datos históricos para predecir eventos futuros. Por lo general, los datos históricos se utilizan para crear un modelo matemático que capture las tendencias importantes. Este modelo predictivo se usa entonces con los datos actuales para predecir lo que pasará a continuación, o bien para sugerir acciones que llevar a cabo con el fin de obtener resultados óptimos. (Mathworks, s.f.)

El análisis predictivo ayuda a los equipos de sectores tan diversos entre los que se encuentran: financiero, sanidad, farmacéutico, automoción, aeroespacial y fabricación.

El análisis predictivo es el proceso de utilizar el análisis de datos para realizar predicciones basadas en los datos. En este proceso se hace uso de los datos junto con técnicas analíticas, estadísticas y de aprendizaje automático a fin de crear un modelo predictivo para predecir eventos futuros.

El término “análisis predictivo” describe la aplicación de una técnica estadística o de aprendizaje automático para crear una predicción cuantitativa sobre el futuro.

Hay disponibles grandes cantidades de datos y mediante el análisis predictivo, los operadores pueden convertir esta información en conocimientos que permite pasar a la acción. Ver figura 2.

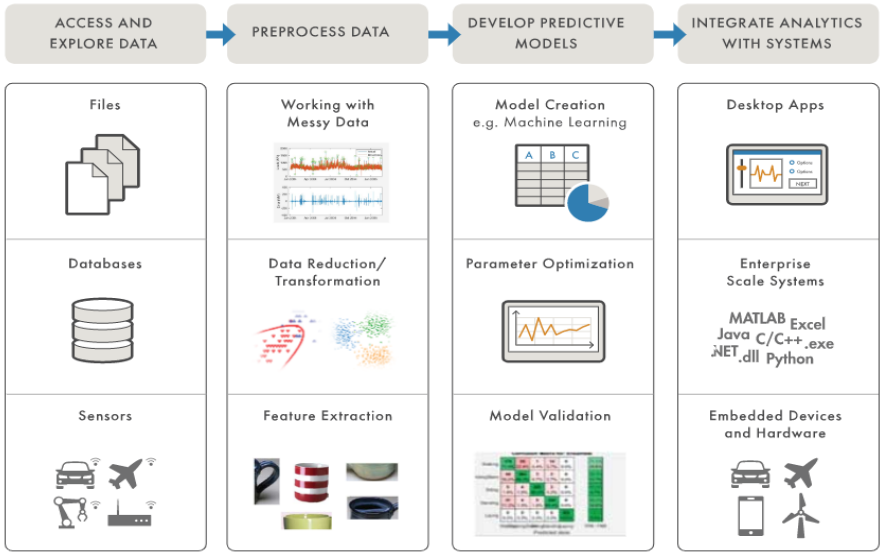


Figura 2. Flujo de trabajo de Análisis predictivo. (Mathworks, s.f.)

5.2.5.2. Algoritmos de regresión lineal

Los algoritmos de regresión lineal se ajustan a una línea recta u otra función que es lineal en sus parámetros, como un polinomio, a los datos numéricos, generalmente realiza inversiones de matriz para minimizar el error al cuadrado entre la línea y los datos. El error al cuadrado se usa como la métrica porque no le importa si la línea de regresión está por encima o por debajo de los puntos de datos; solo le importa la distancia entre la línea y los puntos.

5.2.5.3. Algoritmos de regresión NO lineal

Los algoritmos de regresión no lineal, ajustan curvas que no son lineales en sus parámetros a los datos (polinomial), son un poco más complicados porque, a diferencia de los problemas de regresión lineal, no se pueden resolver con un método determinista. En cambio, los algoritmos de regresión no lineal implementan

algún tipo de proceso de minimización iterativo, a menudo algún ajuste en el método de descenso más pronunciado.

5.2.6. Lenguajes de programación R y Python para análisis de datos

Se describen los dos lenguajes de programación más populares en el ámbito de Ciencia de los Datos y del análisis de dato, R y Python.

5.2.6.1. R

Es un lenguaje de programación para efectuar análisis de datos estadísticos y visualizar gráficas de los mismos datos. Además, es un software libre, gratuito, accesible y siempre a la vanguardia. (r-project.org, 2019)

5.2.6.2. Python

Es un lenguaje de programación y/o software de libre uso, ejecución, distribución y modificación, además de que no tiene costo para quien lo usa. Una de sus características es que es de alto nivel. Programación orientada a objetos. Dispone de un rico ecosistema compuesto de librearías open source para matemáticas, estadísticas, machine learning y ciencia en general. (python, s.f.)

5.2.7. Otorgamiento de créditos de una entidad financiera

Se conceptualizan algunos elementos que forman parte del proceso de otorgamiento de créditos de una entidad financiera y que resultan necesario conocer de ellos para una mejor comprensión de este trabajo.

- **Cartera de préstamos:** Es un conjunto de créditos y financiamientos que los bancos, compañías de inversión o incluso agencias del gobierno poseen o manejan. (Condusef, 2019)
- **Cartera en riesgo:** Mide la porción de la cartera de créditos “contaminada” por deudas atrasadas y en riesgo de no ser pagadas como porcentaje de la cartera total. (Condusef, 2019)
- **Cliente:** Es una persona física o jurídica que recibe un servicio o adquiere un bien a cambio de un dinero u otro tipo de retribución. (Revuelta, 2012).
- **Crédito:** Es una operación financiera en donde una persona llamada “Acreedor” (Entidad financiera), presta una cierta cifra monetaria a otra

persona llamada “Deudor”, quién a partir de ese momento, garantiza al acreedor que retomara esta cantidad solicitada en el tiempo previamente estipulado más una cantidad adicional, llamada “Intereses**”.** (Revuelta, 2012)**.**

- **Entidad Financiera:** Entidad o agrupación que tiene como objetivo y fin ofrecer servicios de carácter financiero y que van desde la simple intermediación y asesoramiento al mercado de los seguros o créditos bancarios. (SanJuán, 2019).
- **Entidades de crédito:** bancos o cajas de ahorro (financieras). (SanJuán, 2019).
- **Institución financiera:** compañía que presta servicios financieros a los agentes económicos de la sociedad. (SanJuán, 2019).
- **Morosidad:** Situación en la que un deudor se ha retrasado tres meses en el pago de los intereses y/o el principal de su deuda. Se trata de una situación de alto riesgo pero que aún no ha caído en la categoría de crédito fallido o irrecuperable. (Revuelta, 2012).
- **Morosidad bancaria:** Es un indicador del nivel de riesgo de que los deudores de los bancos privados (generalmente nos referimos a las personas que piden crédito), no cumplen con sus obligaciones de pago. (Condusef, 2019)
- **Tipos de instituciones financieras:** Se clasifican en dos tipos: **Instituciones Bancarias,** este tipo de entidades, puede captar fondos del público en forma de dinero o recursos financieros de distinto tipo. Su principal actividad es la de captar fondos de agentes con excedentes de capital, para prestarlo a agentes con déficit. Además, estas pueden también conceder garantías y avales, emitir dinero electrónico o realizar transferencias bancarias entre otras actividades. **Instituciones NO Bancarias,** la principal diferencia de estas con las anteriores no pueden captar depósitos del público (SanJuán, 2019).

5.3. Desarrollo

Este apartado de este capítulo muestra algunas herramientas adecuadas para el análisis predictivo y un comparativo entre los lenguajes de programación R y Python como elementos esenciales de este producto académico. Posteriormente, da a conocer herramientas para análisis predictivo relacionadas con el sector financiero. Al final del apartado se muestra una propuesta de utilizar regresiones logísticas para el otorgamiento de crédito de una empresa financiera. Se muestra un ejemplo basado en lenguaje de programación R.

5.3.1. Big Data en el análisis predictivo.

"Gastar ahora y pagar después", es una oferta que en la actualidad muchas empresas financieras hacen a sus clientes para aumentar su cartera de préstamos. Sin embargo, se debe tener la conciencia de que implica un riesgo al tomar ese tipo de decisiones.

Por ello es importante que tanto la empresa financiera como el cliente puedan cumplir con sus obligaciones crediticias, pagar dentro del plazo establecido. Para ello la empresa financiera deberá evaluar el riesgo de incumplimiento de cada cliente y así poder decidir mejor a quien si debe otorgar la oferta del crédito.

5.3.1.1. Evaluación de riesgo de otorgamiento crédito

Los avances tecnológicos han permitido a las empresas financieras reducir riesgo del crédito otorgado, al utilizar la gran variedad de los datos sobre los clientes. Mediante técnicas estadísticas y de aprendizaje automáticos, se analizan y se reduce a un solo valor conocido como puntaje de crédito, que representa el riesgo del crédito, el cual puede servir de guía en el proceso de decisión. A mayor puntaje mayor solvencia de pago del cliente.

La calificación crediticia es una forma de inteligencia artificial basada en modelos predictivos que evalúan la probabilidad de que un cliente no cumpla con la obligación crediticia, se vuelva un delincuente o persona insolvente. El modelo predictivo aprende al utilizar los datos históricos de un cliente junto con los datos del grupo de pares y otros datos para predecir la probabilidad de que ese cliente muestre un comportamiento definido en el futuro.

El mayor beneficio de la calificación crediticia es la capacidad de ayudar a tomar decisiones de manera rápida y eficiente, como aceptar o rechazar a un cliente o aumentar o disminuir el valor del préstamo, la tasa de interés o el plazo. La velocidad y la precisión resultantes de tomar tales decisiones han hecho que la calificación crediticia sea la piedra angular en la gestión de riesgos en todos los sectores, incluidos la banca, las telecomunicaciones, los seguros y el comercio minorista.

5.3.1.2. Puntaje de crédito

El puntaje crediticio arroja una calificación, la cual se puede utilizar durante toda la experiencia del cliente en la relación entre la organización financiera y el cliente. A pesar de que se desarrolló para los departamentos de riesgo crediticio, también lo están utilizando los departamentos de marketing ya que de la misma manera ellos se han beneficiado de las técnicas de la calificación crediticia en sus campañas de marketing. Ver figura 3.

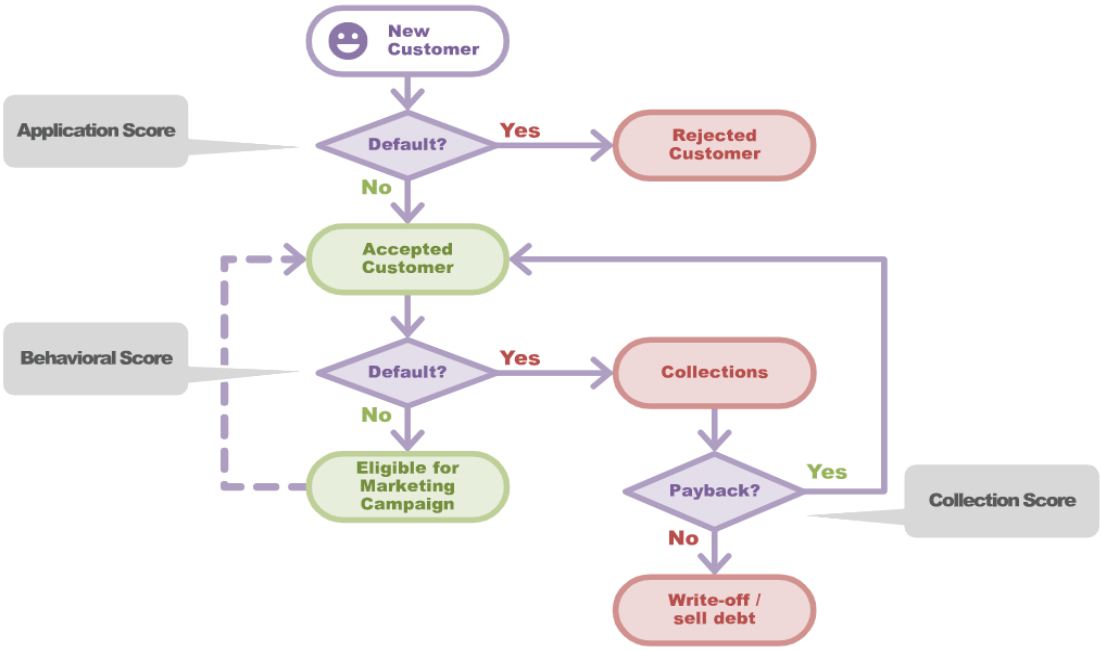


Figura 3. Puntajes de crédito de un cliente. (Mashanovich, 2017).

Como se muestra en la Figura 3, se utilizan diferentes puntajes de crédito en diferentes etapas del recorrido del cliente:

- El puntaje de la solicitud, se refiere a la evaluación del riesgo en el incumplimiento de los nuevos solicitantes al tomar decisiones sobre la aceptación o rechazo del solicitante.
- El puntaje de comportamiento, se refiere a la evaluación del riesgo de incumplimiento asociado con un cliente existente al tomar decisiones relacionadas con la administración de cuentas, como el límite de crédito, la administración por exceso de límite, los nuevos productos y similares.
- El puntaje de cobro, se utiliza en estrategias de cobro para evaluar la probabilidad de que los clientes en cobro paguen la deuda.

5.3.2. Herramientas para análisis predictivo

El análisis predictivo es un cambio en el juego de los negocios, las compañías tienen a su alcance soluciones avanzadas y sencillas del análisis de datos y con ello conocer que es lo que puede suceder en un lapso corto de tiempo.

Un científico de datos debe tener conocimientos en ciencia aplicada con una larga experiencia en su industria y formación en materia científica (aprendizaje supervisado y no supervisado), que le permita llegar a soluciones creativas y con criterio.

Un científico de datos, como la persona que es mejor en estadística que cualquier ingeniero de software y mejor en ingeniería de software que cualquier estadístico. (Kozyrkov, 2018)

En un mundo en donde la especialización es un valor indispensable, este perfil profesional se ha convertido en una especie aplicado a los datos.

Grandes empresas como IBM, Facebook, HP, Oracle, Amazon, o LinkedIn, se mueven en el mundo del Big Data para obtener mejores ventajas competitivas.

La clave de todo ese proceso es la Ciencia de los Datos, la mejora de algoritmos que permita reducir costos, perfeccionar procesos, controlar niveles de riesgos entre otros.

El científico de datos (Data Scientist) va mucho más de lo que está acostumbrando en términos de innovación, su responsabilidad empieza diseñando un prototipo con las tecnologías que mejor se adapten al problema en cuestión, algunas de ellas son: Hadoop, MongoDB, Spark, Python, R, entre otros.

Con el desarrollo de la analítica avanzada es una facilidad con la que los ejecutivos de las empresas, al margen del trabajo que ya realizan otros perfiles como lo son el científico de datos y el analista de datos, pueden hacer predicciones y entender el futuro de sus equipos.

Precisamente la analítica avanzada presenta dos puntos clave:

El uso de API's: en los sectores como el financiero el retail o el energético, usan este tipo de interfaces de programación de aplicaciones para construir modelos predictivos y extraer valor de los datos para sacar conclusiones y tomar decisiones, predecir el comportamiento de los clientes para modificar la oferta y los precios, así como conocer la opinión acerca de los productos y servicios, también conocer cómo se puede aumentar la productividad y rendimiento, y por ultimo prevenir y detectar algún tipo de fraude.

La introducción del PMML (lenguaje de marcado del modelo predictivo): es un lenguaje de texto XML desarrollado por Data Mining Group (DMG), es un lenguaje estándar que se emplea para representar los modelos predictivos, lo que permite que una misma solución pueda compartir diferentes aplicaciones compatibles con PMML.

Este tipo de lenguaje ofrece verdadera interoperabilidad a los integrantes que realizan el análisis predictivo, como los son IBM, SAS, SAP, Oracle, Microsoft, Alteryx o KNIME, entre otros.

A continuación, se enlistan algunas soluciones para el mercado del análisis predictivo:

5.3.2.1. IBM

IBM es uno de los grandes del análisis predictivo ya que cuenta con varias soluciones como los son:

- La analítica de clientes: sirve para conocer los grados de satisfacción del cliente, si poder ofertar productos personalizados, prevé que clientes están a punto de cambiar de proveedor, detecta tendencias de mercado en redes sociales y técnicas de venta cruzada.
- Analítica operacional: el pensamiento real es que las empresas tengan herramientas para realizar la evaluación de costos operativos, velocidad, flexibilidad y calidad a través de la recolección, almacenamiento y análisis de datos así como buscar el valor en tiempo real.
- Analítica predictiva para Big Data: en la mayoría de los casos las empresas disponen de una gran cantidad de datos, el problema es que son datos no estructurados o semiestructurados por ello la toma de decisiones en ese escenario es imposible. La clave es contar con herramientas que organicen los datos, extraigan relaciones y hagan proyecciones sin tener conocimientos técnicos elevados. Esto se consigue con herramientas que combinen los datos no estructurados de forma sencilla, elabore resúmenes visuales y de lenguaje plano para que el analista de negocio (no de datos) entienda el valor de la información, pueda establecer proyecciones de demanda y perfiles personales de clientes. Las soluciones predictivas de IBM son abiertas y soportan la tecnología Hadoop.
- Análisis de amenazas y fraude: los modelos predictivos ayudan a anticiparse a cualquier amenaza o fraude en las compañías. Estas herramientas detectan patrones inusuales en la información con técnicas de minería y análisis de datos.

5.3.2.2. SAS Visual Statistics

Es una plataforma para analizar de forma sencilla una gran cantidad de datos almacenados en Hadoop. La interfaz es muy amigable que permite al usuario deslizar y soltar variables para obtener cuadros de mando.

La herramienta de SAS tiene algunas características interesantes en analítica:

- SAS Analytics Visual Explorer: crea modelos de forma interactiva a partir de variables múltiples, para poder visualizar la información obtenida con esas variables sólo es necesario jalar y soltar: gráficos de barras, histogramas, diagramas de caja, mapas de calor, mapas geográficos, burbujas, diagramas de dispersión, la herramienta en si es muy visual.
- Técnicas interactivas de modelado descriptivo.
- Construcción de modelos predictivos con varias técnicas como lo son, la regresión lineal, modelos lineales generalizados, regresión logística y árboles de clasificación.
- Comparación de modelos con la creación de resúmenes: gráficos de elevación, gráficos ROC, estadísticas de concordancia y tablas de clasificación errónea.

A continuación se muestra una figura con algunos ejemplos de gráficas que permite visualizar el resultado de los análisis de los datos. Ver figura 4.

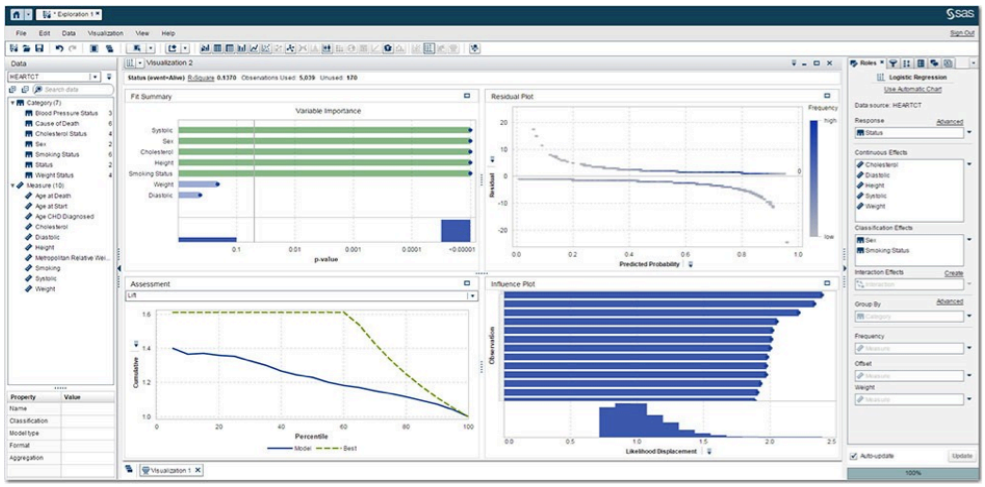


Figura 4. SAS Visual Statistics. (sas Visual Statics, 2019)

5.3.2.3. SAP

Esta empresa dispone de una de las soluciones más conocida que es SAP HANA, que unifica las capacidades de una base de datos y una plataforma de aplicaciones es una herramienta ideal para el análisis predictivo, esta plataforma proporciona bibliotecas de procesamiento de texto y procesamiento espacial, la ventaja que tiene es que es capaz de procesar una gran cantidad de datos en tiempo real sin retardos.

Escanea 319000 millones de símbolos por segundo, estableció un nuevo record como mayor almacén de datos del mundo equivalente a 12.1 petabytes de datos. En resumen, se podría pensar que esta aplicación solo serviría para almacenar y procesar grandes volúmenes de datos.

5.3.2.4. Oracle Advanced Analytics

Es una herramienta que integra en una sola plataforma la base de datos y el análisis avanzado de datos de Oracle R Enterprise y Oracle Data Mining; brinda un análisis de datos en tiempo real en cuestión de predicción, emite recomendaciones y proporciona alertas tempranas de posibles fraudes.

- Oracle R Enterprise, cuenta con bibliotecas de R, que es un lenguaje de programación utilizado en la estadística para analizar datos de cualquier base de datos.
- Oracle Data Mining, cuenta con algoritmos de minería de datos de gran alcance que sirve para construir, evaluar, compartir y extender modelos de análisis predictivo (Oracle Database, s.f.).

5.3.2.5. RapidMiner Studio

Rapid Miner Studio es una plataforma de análisis predictivo de aprendizaje automático, minería de datos y análisis de negocio; permite la carga, transformación y modelado de grandes volúmenes de datos a partir de fuentes como lo son Excel, Access, Oracle, IBM, DB2, Microsoft SQL, SAP Sybase, Ingres, MySQL, Pstgres, SPSS, Dbase, y cualquier fuente de texto plano. Cualquier empresa puede integrar en esta plataforma su propio algoritmo relacionado con datos a través de aplicaciones abiertas (RapidMiner, 2020).

5.3.2.6. Alteryx 7.1

Es una plataforma que permite el acceso, la gestión y el análisis predictivo de los datos en la misma herramienta. Es una solución que integra todas las funcionalidades del lenguaje de programación R para el estudio estadístico, pero sin que el usuario necesite tener unos conocimientos avanzados en estadística. Gracias a esta aplicación, las empresas pueden manejar grandes volúmenes de datos, ser capaz de interpretarlos al igual que otras herramientas del análisis

predictivo, los usuarios pueden jalar y soltar variables para crear sus propios modelos.

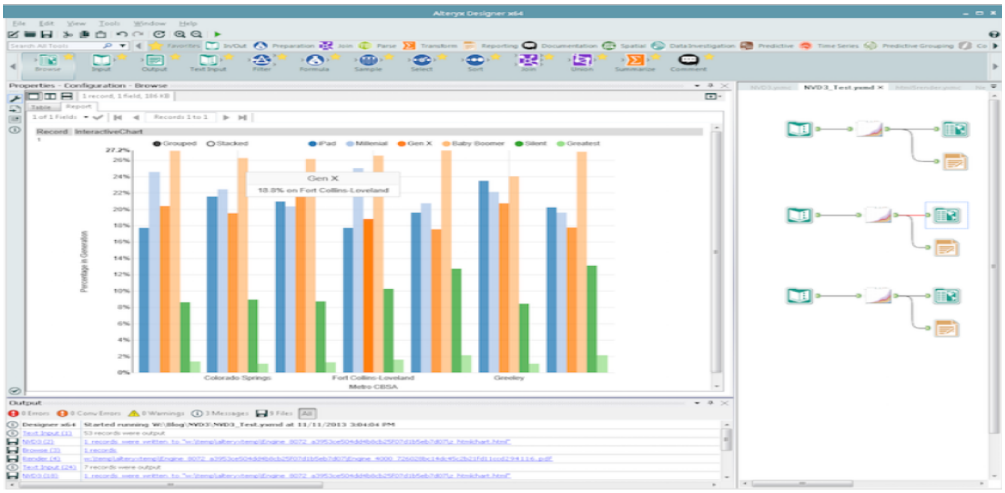


Figura 5. Interfaz Alteryx 7.1. (alterix, s.f.).

5.3.2.7. Microsoft soluciones de SQL Server

Esta plataforma de bases de datos, ofrece a los clientes la obtención de información de valor del análisis predictivo de los datos mediante la integración de tecnología en memoria y de alto; ya que utiliza un conjunto de herramientas para implementar y administrar bases de datos en la nube y en entornos locales y permite el análisis predictivo en los siguientes dos sentidos:

- El analista puede usar técnicas de minería de datos con herramientas cono Excel para extraer información, hacer visualizaciones y graficas con los datos y con ello realizar resúmenes visuales.
- Los desarrolladores pueden usar SQL Server para crear soluciones de minería y análisis de datos.

5.3.2.8. KMINE Analytics Platform

KMINE es una plataforma de minería de datos para crear modelos predictivos visuales. Está escrita en lenguaje Java es una solución en código abierto, con esta herramienta se puede hacer por ejemplo:

- Visualizar datos en histogramas, mapas etc.
- Crear modelos estadísticos, arboles de decisión, regresiones, entre otros.

- Realizar informes personalizados
- Se pueden incorporar funcionalidades escritas en R y Python.

5.3.2.9. Tableau, SAP Predictive Analytics & Fiserv

Otras herramientas no menos importantes deben ser sin lugar a duda las siguientes:

Tableau: Inteligencia empresarial y análisis: ayuda a las personas y organizaciones a ser más motivadas por los datos como el líder confiable en análisis.

SAP Predictive Analytics: Software de análisis predictivo: brinda información predictiva a los usuarios comerciales y científicos de datos para una mejor toma de decisiones.

Fiserv: Gestión del riesgo de delitos financieros: para transformar la forma en que los bancos, las cooperativas de crédito y las instituciones de ahorro y crédito. Posibilita a las instituciones respondan rápidamente a los nuevos ataques de delitos financieros. Este software es comercial, tiene costo y existen múltiples aplicaciones para los análisis predictivos para un sinfín de problemas. (GetApp, 2019).

El análisis de los datos, es una parte central de cualquier proyecto. Para obtener nuevos patrones, tendencias, estructuras, entre otros. Se pueden realizar con diferentes herramientas y lenguajes de programación.

5.3.3. Comparativo entre R y Python

R fue escrito por Robert Gentleman y Ross Ihaka del Departamento de Estadística de la Universidad de Auckland en Nueva Zelanda y actualmente R se ha enriquecido con las aportaciones de personas de todo el mundo, quienes pueden realizar modificaciones en el código fuente. R tiene más de 4038 paquetes, cada uno para desarrollar algo en particular. (Julio Fernando Suárez Cifuentes, 2013)

Algunas ventajas que ofrece R en comparación de otras herramientas se citan de la siguiente manera:

- Posibilita la realización de gráficos excelentes.

- Es flexible y tiene muchas librerías.
- Se pueden programar procedimientos y aplicaciones propias.
- R está disponible en varias formas: el código fuente escrito principalmente en C (y algunas rutinas en Fortran), esencialmente para máquinas Unix y Linux, o como archivos binarios pre compilados para Windows y Linux.
- Tiene funciones para análisis estadísticos y gráficos; estos últimos se pueden ver de inmediato en su propia ventada y ser guardados en varios formatos (jpg, png, bmp,ps, pdf, emf, pictex, xfig).
- Es un lenguaje interpretado y no compilados.
- Es un lenguaje orientado a objetos.
- Es gratuito.

Python es un lenguaje de programación creado por Guido van Rossum a principias de los años 90. Es un lenguaje similar a Perl, pero con una sintaxis muy limpia y que favorece un código legible. Es un lenguaje interpretado o de script, con tipado dinámico, fuertemente tipado, multiplataforma y orientado a objetos. Se muestran a continuación una tabla comparativa entre R y Python:

Tabla 1.

Comparativo del lenguaje de programación R y Python.

Lenguaje de Programación	Python	R
Creado por:	Guido Van Rossum en 1991.	Ross Ihaka y Robert Gentleman en 1995
Propósito:	Enfatiza la productividad y la legibilidad del código.	Se centra en un mejor análisis de datos fácil de usar, estadísticas y gráficos.
Utilizado por:	Es utilizado por los programadores que quieren profundizar en estadística de los datos o aplicar técnicas estadísticas, los desarrolladores recurren a ciencia de los datos.	Se usa principalmente en lo académico y la investigación. R se está expandiendo en el mercado empresarial.
	"Cuanto más cerca estés de trabajar en un entorno de ingeniería, más preferirás Python"	"Cuanto más cerca estés de las estadísticas, la investigación y la ciencia de los datos, más preferirás R"
Algunas Librerías y Paquetes:	• **Pandas** para manipular datos. • **SciPy /Numy** para los científicos. • **Sckikit-learn** técnicas de aprendizaje automático. • **Matplotlib** para hacer gráficos. • **Statsmodels** para explorar datos, estimar modelos estadísticos y realizar pruebas estadísticas y pruebas unitarias.	• **Tydiverse** varios paquetes • **readr** para leer datos • **Dplyr, Plyr y Data.table** para manipular datos. • **ggplot2** de gráficos • **Stringr** para manejar cadena de caracteres.
Uso de Datos y Tareas	Es generalmente usado cuando el análisis de los datos ser integrado en aplicaciones Web o si el código de estadísticas debe incorporarse a una base de datos de producción. Como lenguaje de programación completa, es una excelente herramienta para implementar algoritmos para el uso de producción.	• Utilizado para cuestiones estadísticas • Excelente en algoritmos d machine learning

5.3.4. Caso BBVA Research

En la actualidad con el desarrollo tecnológico una de las consecuencias importantes ha sido la generación de grandes volúmenes de información y datos en distintos ámbitos.

La recopilación ingesta de datos, el análisis y su utilización en la toma de decisiones es, actualmente, uno de las principales fuentes de diferenciación en un mundo globalizado y en el cual la información y las tecnologías están al alcance de todos a un bajo costo.

Una empresa capaz de almacenar y analizar de forma adecuada la información, obtendrá ventajas competitivas que le permitirán diferenciarse de sus competidores y ocupar una posición líder en la industria que participa.

El uso de Big Data ha entrado con mucha fuerza en el marketing, utilizándose como herramienta para promocionar productos o servicios a un cliente, sin conocerlo, solo con la información obtenida de sus búsquedas o preferencias en la Web (huella digital). Esto, está permitiendo el desarrollo de nuevos productos y servicios que sacan provecho de la información, recopilándola y usándola para generar soluciones, mejorar la experiencia de servicio y aumentar la fidelidad y retención de los clientes. (BBVA Research, 2017).

Big Data que se posiciona en varios sectores: en primera instancia en el sector de Telecomunicaciones como herramienta para predecir y prevenir la pérdida de clientes y en la realización de campañas. Poco a poco, se han ido sumando la banca y procesos de ventas en tiendas departamentales, autoservicio y comercio, en la búsqueda de sacar el máximo provecho a la información que generan sus clientes actuales y potenciales en sus interacciones en las redes sociales. (BBVA Research, 2017).

En las empresas de servicios bancarios, la revolución digital está iniciando y los próximos pasos en esta industria apuntan a una experiencia digital totalmente adherida para el cliente. Los bancos digitales ofrecen y ofrecerán servicios de forma digital, tomando la oportunidad de adelantarse a las necesidades de las personas y

empresas y tendrán estrategias de inversión y ahorro que mejor se adaptan a las necesidades de cada cliente, basándose solamente en los datos disponibles. (BBVA Research, 2017)

No todos los países y no todas las compañías avanzan al mismo ritmo en el uso y aplicación de prestaciones digitales. El nivel de desarrollo y penetración de las tecnologías determinan la velocidad a la cual los países y las empresas se incorporan a la economía digital. (BBVA Research, 2017).

Big Data permite analizar la información económica en tiempo real, sin esperar hasta que las cifras se publiquen, lo que ocurre con rezagos que habitualmente van desde una semana hasta más de un mes.

La tecnología Big Data permite recopilar y procesar en tiempo real las decisiones de personas, empresas y del gobierno y, con ello, hacer un seguimiento en línea de la evolución de la economía. Este es un enorme y revolucionario avance en materia de análisis, permite realizar mejores estimaciones y analizar a detalle las decisiones de los agentes económicos, así como entender mejor el mecanismo de transmisión de las políticas económicas. El desarrollo de mejores modelos de predicción usando estos datos es todavía un tema pendiente.

El estudio de la huella digital está siendo usado por algunas empresas para calcular índices de consumo y ha sido usado en apoyo a las autoridades mexicanas para fortalecer el diseño de políticas públicas para impulsar el crecimiento ordenado de la actividad turística del país y para ofrecer un mejor servicio por parte de las empresas de bienes y servicios turísticos.

Con lo anterior basado en el caso BBVA, se establece una propuesta de uso de análisis de datos de cartera de clientes una empresa financiera para lograr conocer las percepciones de las personas.

5.3.5. Propuesta de análisis predictivo para la empresa financiera

En una empresa financiera se generan grandes cantidades de información en relación a los clientes a los cuales se les tiene otorgado un crédito, la cantidad de cartera en riesgo es muy elevada aunado a estos factores se ha disminuido

constantemente el número de clientes y por consiguiente la cartera activa hablando en términos monetarios.

Para “la empresa financiera” tomada como referencia y caso de uso, el proceso para otorgar un crédito al cliente el proceso de evaluación del crédito es demasiado lento porque los procesos de análisis y evaluaciones de riesgos de clientes tarda mucho tiempo.

Dentro de la empresa financiera, el riesgo de crédito se evalúa de manera subjetiva, es decir, la información del cliente se obtiene a través de relaciones personales entre los clientes y los asesores de la financiera.

Los préstamos son a menudo un proceso de juicio de una persona que evalúa las solicitudes en función de ciertos criterios, por ejemplo: ¿Es el solicitante o cualquiera de su familia conocido por la organización?; ¿Cuál es el monto del préstamo solicitado?; ¿Cuál es la capacidad de pago del solicitante?; ¿En que invertirá el préstamo?

Para ello la financiera debe tener la capacidad de poder evaluar el riesgo de incumplimiento de cada cliente para poder formar el grupo solidario, y en base a esta evaluación la empresa como tal, pueda decidir si es viable otorgar el crédito o no, sin embargo, esto no debiera ser de manera subjetiva.

Con el tiempo, han evolucionado las técnicas de modelado sobre el riego de crédito. Recientemente se incluyen enfoques sofisticados que utilizan cientos o miles de modelos diferentes, diversos marcos de validación y técnicas de conjunto con múltiples algoritmos de aprendizaje para obtener una mayor precisión.

Por lo anterior, como propuesta se recomienda el uso de R y Python para el análisis predictivo sobre las necesidades de los clientes y así poder generar estrategias más específicas en cuanto el otorgamiento de los créditos, conocer el motivo principal de la deserción de clientes, disminuir la cartera en riesgo y así generar mejores cifras positivas para la entidad financiera.

Al estar obteniendo la información en tiempo real se podrán trazar estrategias para generar nuevos productos y servicios al sector de clientes con los cuales se

trabaja, sin necesidad de pagar a empresas externas para realizar encuestas ya que de esta manera se pierde tiempo, dinero y clientes.

En base a lo anterior se propone realizar un análisis para determinar la siguiente información:

- ¿Qué colaboradores de la empresa han enviado más cantidad de clientes a cartera vencida y cuál es el motivo específico?
- ¿Cómo se podrá contrarrestar esta pérdida para la empresa financiera?
- ¿Cuál ha sido el motivo principal de la pérdida de clientes en la empresa?
- ¿Qué estrategia se puede utilizar para regresar a todos esos clientes perdidos?

Se propone un modelo que permita realizar el análisis de los datos de manera eficiente, rápida y oportuna. Se pretende que en base a la información obtenida se puedan lograr tomar mejores decisiones dentro de la organización financiera.

Se sugiere utilizar el modelo que se base en la regresión logística denominado "*Credit Scorecard*" generalmente conocido como cuadro de mando estándar; porque se basa en la regresión logística como el modelo subyacente.

El resultado del modelo *Credit Scorecard* consiste en un conjunto de atributos (características del cliente) que se muestran en forma de tabla. Dentro de un atributo, los puntos ponderados se asignan a cada valor de atributo en el rango y la suma de esos puntos es igual a la calificación crediticia final.

Para el modelo se debe documentar lo siguiente:

- La unidad de análisis (como nivel de cliente o producto),
- Marco poblacional (por ejemplo, solicitantes de préstamos a través de la puerta) y tamaño de la muestra.
- Definiciones operativas ("malo" o "bueno") y supuestos de modelo (por ejemplo, excluyendo clientes fraudulentos),
- Horizonte temporal de observación (como el historial de pagos de los clientes en los últimos dos años) y ventanas de rendimiento, ese es el marco de tiempo para el que se aplica la definición de "mala",

- Fuentes de datos y métodos de recopilación de datos.

5.3.5.1. Preparación de datos

La preparación de los datos es pieza clave para el desarrollo de un cuadro de mando crediticio. Implica la recopilación de datos, la combinación de múltiples fuentes de datos, agregaciones, transformaciones, limpieza de datos, y la observación de la amplitud y profundidad de los datos para obtener una comprensión clara y transformar la cantidad de datos en calidad de los datos para que se pueda preparar con confianza la siguiente fase.

El primer paso en el análisis exploratorio es la lectura de los datos y luego explorar las variables y casos que hay, los tipos de datos de las variables y el rango de los valores que toman. Algunas variables pueden ser:

- Id de préstamo: un identificador único para la información del préstamo.
- Id de cliente: identificador único del cliente. Los clientes pueden tener más de un préstamo.
- Monto actual del préstamo: es el monto del préstamo que fue completamente pagado o la cantidad que se incumplió.
- Plazo: una variable categórica que indica si se trata de un préstamo a corto o largo plazo.
- Puntuación de crédito: un valor entre 0 y 800 que indica el riesgo que se puede tener al otorgar el crédito, es el historial crediticio.

El lenguaje que se sugiere para iniciar con el análisis predictivo es R. De esta manera y con los datos históricos se podrá predecir posibles acontecimientos negativos y así contrarrestar perdidas de clientes y cartera activa. También esta herramienta será de mucha utilidad ya que se podrá determinar la causa principal de la deserción de los mismos clientes y con ellos generar estrategias para incrementar y retener a la mayor cantidad de clientes.

5.3.5.2. Proceso para evaluación del riesgo

Hay básicamente dos enfoques principales para evaluar el riesgo crediticio: Ambos dependen de la información histórica, pero el tipo de información que utilizan es diferente los enfoques son los siguientes

El enfoque de juicio es un enfoque cualitativo, basado en la experiencia empresarial y el sentido común, el experto en crédito o comité de crédito, que es un grupo de expertos en crédito, tomará una decisión sobre el riesgo crediticio. Por lo general, esto se hace sobre la base de la inspección de los cinco Cs del solicitante y el préstamo (Chi Dung, 2018).

- El carácter mide el carácter y la integridad del prestatario (por ejemplo, reputación, honestidad, entre otros).
- El capital mide la diferencia entre los activos del prestatario (por ejemplo, coche, casa, etc.) y pasivos (por ejemplo, gastos de alquiler, entre otros).
- La garantía mide la garantía proporcionada en caso de que se produzcan problemas de pago (por ejemplo, casa, coche, entre otros).
- La capacidad mide la capacidad de pago del prestatario (por ejemplo, la situación del trabajo, los ingresos, entre otros).
- La condición mide las circunstancias de la institución que proporciona el crédito (por ejemplo, condiciones de mercado, presión competitiva, carácter estacional, entre otros).

Al analizar esta información, se realiza una evaluación cualitativa o subjetiva del riesgo crediticio. Aunque el enfoque de juicio puede parecer subjetivo y, por lo tanto, poco sofisticado a primera vista, todavía es muy frecuentemente utilizado por los bancos para carteras de crédito muy específicas.

Con el surgimiento de técnicas de clasificación estadística a principios de la década de 1980, las instituciones financieras se interesaron cada vez más en abandonar el enfoque de los juicios y optar por un enfoque estadístico más formal basado en datos.

El enfoque estadístico se basa en el análisis estadístico de los datos históricos para encontrar la relación multivariada óptima entre las características de

un cliente y la variable binaria de destino bueno/malo. Es menos subjetivo que el enfoque de juicio, ya que no está vinculado al conocimiento y la experiencia de un experto en crédito en particular.

Tiene como objetivo construir cuadros de mando, que se basan en correlaciones multivariadas entre insumos (como la edad, el estado civil, los ingresos, la cantidad de ahorro) y una variable denominada objetivo que refleja el riesgo de incumplimiento. En otras palabras, un cuadro de mandos asignará puntuaciones a cada una de esas entradas.

En esta propuesta, las puntuaciones se asignarán a la edad, el estado civil, los ingresos y la cantidad de ahorro. Todas esas puntuaciones se sumarán y se compararán con el umbral crítico que especifica el nivel mínimo de calidad de crédito requerida. Si la puntuación agregada supera el umbral, se concederá crédito. Si cae por debajo del umbral, se retendrá el crédito.

En la experiencia, se pueden aplicar enfoques híbridos. En un primer paso, una institución financiera puede generar valores informativos por puntuación judicial. Un ejemplo puede ser una opinión experta de un analista de crédito sobre la ética de pago de un prestatario (por ejemplo, como un número discreto entre 1 y 5). En un segundo paso, la institución puede agregar esta puntuación de juicio y otra información dura en una puntuación estadística.

En términos generales, el enfoque estadístico de la puntuación de crédito tiene muchas ventajas en comparación con el enfoque de juicio. En primer lugar, es mejor en términos de velocidad y precisión. Se puede tomar decisiones más rápidas de lo que se pudo hacer con el enfoque de los juicios. Esto es especialmente relevante cuando se trabaja en un entorno en línea donde las decisiones de crédito deben tomarse rápidamente. Debido a que un cuadro de mandos de crédito es esencialmente una fórmula matemática, se puede programar y evaluar fácilmente de una manera automatizada y rápida.

Otra ventaja de tener modelos estadísticos de puntuación de crédito es la coherencia. Ya no se basa en la confianza de la experiencia, la intuición o el sentido

común. Hoy en día sólo es una fórmula matemática, y la fórmula siempre se evaluará exactamente de la misma manera si se le da el mismo conjunto de insumos, como la edad, el estado civil, los ingresos, entre otros.

Por último, los modelos estadísticos de puntuación de crédito normalmente también serán más potentes que los modelos de juicio. Este aumento del rendimiento permitirá una reducción de la pérdida de deuda y los costos operativos, y como resultado también perfeccionará la gestión de la cartera.

5.3.5.3. Regresión logística para desarrollar un modelo de cuadro de mandos

La regresión logística es una técnica de clasificación de puntuación de crédito muy popular debido a su simplicidad y buen rendimiento. Al igual que con la regresión lineal, una vez que se han estimado los parámetros, la regresión se puede evaluar de una manera sencilla, contribuyendo a su eficiencia operativa. Desde un punto de vista de interpretabilidad, se puede transmitir fácilmente en un cuadro de mandos de crédito ejecutable, fácil de usar y basado en puntos.

Un aspecto técnico importante del desarrollo de la regresión logística es la selección de variables. La regresión logística tiene un procedimiento integrado para realizar la selección de variables. Se basa en una prueba de hipótesis estadística para verificar si el coeficiente de una variable contenida en el modelo es significativamente diferente de cero.

En la puntuación de crédito, es muy importante tener en cuenta que la importancia de la estadística es sólo un criterio de evaluación que se tiene al momento de realizar la selección de una variable.

En la regresión logística, esto se puede evaluar fácilmente examinando el código del coeficiente de regresión. Es preferente que un coeficiente contenga el mismo código que pronostica el experto en crédito; ya que de lo contrario pueden ser renuentes al hacer uso del modelo.

Antes de crear un modelo de clasificación binaria (por ejemplo, Regresión logística), un paso común es realizar análisis de datos exploratorios y del filtrado de

variables. Este es el paso en el que se dan a conocer los datos y se eliminan variables que están mal condicionadas o no contienen información que aporte en la predicción referente a la acción de interés. El propósito de este paso no debe confundirse con el de las técnicas de selección de variables, como la regresión escalonada, donde se seleccionan las variables que entran en el modelo final. En su lugar, este es un paso precursor diseñado para garantizar que los enfoques implementados durante las fases de modelado final se configuran para el éxito.

El peso de la evidencia (WOE) y el valor de la información (IV) proporcionan un gran cuadro para el análisis exploratorio y el filtrado de las variables para que a su vez se clasifiquen en binarios. WOE y IV se han utilizado ampliamente en el mundo del riesgo crediticio durante varias décadas, y la teoría subyacente se remonta a la década de 1950. WOE y IV son técnicas simples, pero potentes para realizar la transformación y selección variable.

Estos conceptos tienen una enorme conexión con la técnica de modelado de regresión logística. Es ampliamente utilizado en la puntuación de crédito para medir la separación de buenos clientes vs malos clientes. Al mismo tiempo, las ventajas de la transformación WOE son que permiten controlar los valores que faltan y manejar valores atípicos.

La transformación se basa en el valor logarítmico de las distribuciones. Esto se alinea con la función de salida de regresión logística No hay necesidad de variables ficticias.

Mediante el uso de una técnica de binning adecuada, puede establecer una relación auto reconstruyente (aumento o disminución) entre la variable independiente y la variable dependiente.

En resumen, el análisis WOE y IV permite:

- Considerar la contribución independiente de cada variable al resultado.
- Detectar relaciones lineales y no lineales.
- Clasificar las variables en términos de fuerza predictiva "univariada".

- Visualizar las correlaciones entre las variables predictivas y el resultado binario.
- Comparar sin problemas la fuerza de las variables continuas y categóricas sin crear variables ficticias.
- Manejar sin problemas los valores que faltan sin censura.
- Evaluar el poder predictivo de los valores que faltan.

Por convención, los valores estadísticos generados en el modelo de regresión logística P-Value para la selección variable se pueden utilizar de la siguiente manera:

- Menos de 0,02: el predictor no es útil para el modelado (separar las mercancías de los malos).
- De 0,02 a 0,1: el predictor solo tiene una relación débil con la relación de probabilidades De bienes/malas.
- De 0,1 a 0,3: el predictor tiene una relación de resistencia media con la relación de probabilidades De bienes/malas.
- De 0,3 a 0,5: el predictor tiene una fuerte relación con la relación entre bienes y malos.
- Más de 0,5: se debe comprobar claramente, al seleccionar las variables que tienen IV superior a 0,5 debido a una relación sospechosa.

5.3.5.4. Características clave de un modelo útil de cuadro de mandos

Al adoptar un cuadro de mandos para ponerlo en producción, se debe realizar una evaluación a fondo; dependiendo de la configuración exacta y el uso del modelo, es posible que sea necesario evaluar diferentes aspectos durante la evaluación para garantizar que el modelo sea aceptable para su implementación.

Las características clave del modelo de cuadro de mandos exitoso son:

- Interpretabilidad: un cuadro de mandos debe ser interpretable, es decir, que se requiere una comprensión más profunda del comportamiento predeterminado detectado.

- Los modelos que admiten al usuario comprender las razones subyacentes relacionadas al mismo, indican la predeterminación de un cliente por medio de modelos de caja blanca; mientras que los modelos matemáticos complejos, incomprensibles se conocen a menudo como modelos de caja negra.
- Precisión estadística: esta se refiere a la eficacia de detección y la corrección del cuadro de mandos en el etiquetado de los clientes como predeterminados. Existen varios criterios de evaluación estadística que se pueden aplicar para evaluar este aspecto, como la tasa de aciertos, las curvas de elevación, el área debajo de la curva (AUC), entre otros. También puede referirse al significado estadístico, lo que quiere decir, que los patrones que se han encontrado en los datos tienen que ser válidos y no ser la consecuencia del ruido.
- Costo económico: Desarrollar e implementar un cuadro de mandos implica un costo significativo para una organización. El costo total incluye los costos de recopilar, pre procesar y analizar los datos, y los costos para poner los cuadros de mando resultantes en producción. Además, los costos de software, recursos humanos e informáticos también se deben tener en cuenta. Posiblemente también se deben comprar datos externos (por ejemplo, de la oficina de crédito) para enriquecer los datos internos disponibles.
- Cumplimiento normativo: Un cuadro de mandos debe estar bien alineado y cumplir con todas las regulaciones y leyes aplicables. Se debe respetar otras regulaciones como por ejemplo, con respecto a la privacidad y/o la discriminación.

El uso más importante de las puntuaciones de las solicitudes es decidir sobre la aprobación del préstamo. Las puntuaciones también se pueden utilizar con fines de fijación de precios. Los precios basados en el riesgo (a veces también denominados precios ajustados al riesgo) establecen el precio u otras características (por ejemplo, plazo del préstamo, garantía) del préstamo en función

del riesgo percibido medido por la puntuación de la solicitud. Una puntuación más baja implicará una tasa de interés más alta y viceversa.

5.3.5.5. Regresión logística en R

Se utiliza el ejemplo simulado con los datos contenidos en la dirección http://www.creditriskanalytics.net/datasets-private2.html descargando el archivo y cargando los datos en R y R Studio. (Baesens, 2016).

El conjunto de datos HMEQ (Home Equity Loans) informa características e información de morosidad para 5960 préstamos con garantía hipotecaria. Un préstamo con garantía hipotecaria es un préstamo en el que el deudor utiliza el capital de su casa como garantía subyacente.

Se muestran los datos de 30 registros de datos limpios y transformados generados por la recreación en el enlace https://rpubs.com/rpizarrog/601901. (Pizarro, 2020).

Tabla 2.

Datos limpios y transformados

BAD	LOAN	MORTDUE	VALUE	REASON	JOB	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
1	1100	25860	39025	HomeImp	Other	10.50	0.00	0.00	94.37	1.00	9.00	33.78
1	1300	70053	68400	HomeImp	Other	7.00	0.00	2.00	121.83	0.00	14.00	33.78
1	1500	13500	16700	HomeImp	Other	4.00	0.00	0.00	149.47	1.00	10.00	33.78
1	1500	73760.82	101776	DebtCon	Other	8.92	0.25	0.45	179.77	1.19	21.30	33.78
0	1700	97800	112000	HomeImp	Office	3.00	0.00	0.00	93.33	0.00	14.00	33.78
1	1700	30548	40320	HomeImp	Other	9.00	0.00	0.00	101.47	1.00	8.00	37.11
1	1800	48649	57037	HomeImp	Other	5.00	3.00	2.00	77.10	1.00	17.00	33.78
1	1800	28502	43034	HomeImp	Other	11.00	0.00	0.00	88.77	0.00	8.00	36.88
1	2000	32700	46740	HomeImp	Other	3.00	0.00	2.00	216.93	1.00	12.00	33.78
1	2000	73760.82	62250	HomeImp	Sales	16.00	0.00	0.00	115.80	0.00	13.00	33.78
1	2000	22608	101776	DebtCon	Other	18.00	0.25	0.45	179.77	1.19	21.30	33.78
1	2000	20627	29800	HomeImp	Office	11.00	0.00	1.00	122.53	1.00	9.00	33.78
1	2000	45000	55000	HomeImp	Other	3.00	0.00	0.00	86.07	2.00	25.00	33.78
0	2000	64536	87400	DebtCon	Mgr	2.50	0.00	0.00	147.13	0.00	24.00	33.78
1	2100	71000	83850	HomeImp	Other	8.00	0.00	1.00	123.00	0.00	16.00	33.78
1	2200	24280	34687	HomeImp	Other	8.92	0.00	1.00	300.87	0.00	8.00	33.78
1	2200	90957	102600	HomeImp	Mgr	7.00	2.00	6.00	122.90	1.00	22.00	33.78
1	2200	23030	101776	DebtCon	Other	19.00	0.25	0.45	179.77	1.19	21.30	3.71
1	2300	28192	40150	HomeImp	Other	4.50	0.00	0.00	54.60	1.00	16.00	33.78
0	2300	102370	120953	HomeImp	Office	2.00	0.00	0.00	90.99	0.00	13.00	31.59
1	2300	37626	46200	HomeImp	Other	3.00	0.00	1.00	122.27	1.00	14.00	33.78
1	2400	50000	73395	HomeImp	ProfExe	5.00	1.00	0.00	179.77	1.00	0.00	33.78
1	2400	28000	40800	HomeImp	Mgr	12.00	0.00	0.00	67.20	2.00	22.00	33.78
1	2400	18000	101776	HomeImp	Mgr	22.00	0.25	2.00	121.73	0.00	10.00	33.78
1	2400	73760.82	17180	HomeImp	Other	8.92	0.00	0.00	14.57	3.00	4.00	33.78
1	2400	34863	47471	HomeImp	Mgr	12.00	0.00	0.00	70.49	1.00	21.00	38.26
0	2400	98449	117195	HomeImp	Office	4.00	0.00	0.00	93.81	0.00	13.00	29.68
1	2500	15000	20200	HomeImp	Other	18.00	0.00	0.00	136.07	1.00	19.00	33.78
1	2500	25116	36350	HomeImp	Other	10.00	1.00	2.00	276.97	0.00	9.00	33.78
0	2500	7229	44516	HomeImp	Self	8.92	0.00	0.00	208.00	0.00	12.00	33.78

Las variables de los datos

- Los datos tienen las siguientes características:
- Son trece variables del conjunto de datos
- BAD - MALO: 1 = solicitante incumplido en préstamo o gravemente moroso; 0 = préstamo pagado por el solicitante

- LOAN - PRÉSTAMO: monto de la solicitud de préstamo
- MORTDUE: Monto adeudado por hipoteca existente
- VALUE - VALOR: valor de la propiedad actual
- REASON - RAZÓN: DebtCon = consolidación de deuda; HomeImp = mejoras para el hogar
- JOB - TRABAJO: categorías ocupacionales
- YOJ: Años en el trabajo actual
- DEROG: número de informes despectivos importantes
- DELINQ: Número de líneas de crédito morosas
- CLAGE: Edad de la línea de crédito más antigua en meses.
- NINQ: número de consultas de crédito recientes
- CLNO: número de líneas de crédito
- DEBTINC - DEUDA: relación deuda-ingreso

5.3.5.6. Resumen del modelo

Siguiendo un enlace https://rpubs.com/rpizarrog/601901 de Pizarro, (2020), puede apreciarse la recreación de un modelo de regresión logística para la predicción otorgamiento o No de crédito.

La variable dependiente o de respuesta del modelo es BAD y el resto son las variables independientes.

```
modelo <- glm(data = datos.Entrena, formula = BAD ~ LOAN + VALUE + REASON
+ JOB + DEROG + DELINQ + CLAGE + NINQ + CLNO + DEBTINC, family = binomial)

# modelo

summary(modelo)
```

En la tabla 3 se resume el modelo de regresión logística mostrando los valores de los coeficientes así como los valores significativos de Pr(>|t|) que representa la relación estadística entre las variables independiente y la variable de respuesta.

En la tabla se observa que la gran mayoría de las variables independientes tienen relación significativa con la variable dependiente. La variable llamada JOB en

su categoría Office y ProfExe respectivamente representan poca relación estadísticamente significativa.

Tabla 3.

Coeficientes del modelo de regresión logística

Coefficients:	Estimate Std.	Error	t value	Pr(>|t|)	Siginif.
(Intercept)	4.013e-02	3.277e-02	1.225	0.220822	
LOAN	-2.824e-06	5.256e-07	-5.372	8.20e-08	***
VALUE	3.189e-07	1.076e-07	2.965	0.003048	**
REASONHomeImp	4.623e-02	1.229e-02	3.761	0.000171	***
JOBOffice	-3.055e-02	2.011e-02	-1.519	0.128731	
JOBOther	4.375e-02	1.731e-02	2.527	0.011534	*
JOBProfExe	1.182e-02	1.904e-02	0.621	0.534604	
JOBSales	1.726e-01	4.164e-02	4.146	3.45e-05	***
JOBSelf	9.100e-02	3.350e-02	2.716	0.006633	**
DEROG	8.769e-02	7.001e-03	12.525	< 2e-16	***
DELINQ	1.132e-01	5.113e-03	22.132	< 2e-16	***
CLAGE	-6.289e-04	6.962e-05	-9.032	< 2e-16	***
NINQ	2.788e-02	3.458e-03	8.062	9.76e-16	***
CLNO	-2.674e-03	6.061e-04	-4.412	1.05e-05	***

5.3.5.7. Análisis predictivo

Bajo el mismo ejemplo, imaginar que llegan cuatro clientes a solicitar un crédito con las siguientes características:

Los primeros dos solicitan un préstamo de 2000 con valores diferentes en las otras variables independientes; los últimos dos solicitan préstamos de 2500 como se muestra en la tabla siguiente:

Tabla 4.

Nuevos clientes para análisis predictivo

LOAN	MORTDUE	VALUE	REASON	JOB	YOJ	DEROG	DELINQ	CLAGE	NINQ	CLNO	DEBTINC
2000	43000	53000	HomeImp	Other	3	0	0	86.067	2	25	33.7799
2000	62536	85400	DebtCon	Mgr	2.5	0	0	147.13	0	24	33.7799
2500	27116	34350	HomeImp	Other	10	1	2	276.97	0	9	33.7799
2500	9229	42516	HomeImp	Self	8.92	0	0	208	0	12	33.7799

Emulando la predicción en R y siguiendo la simulación en el enlace citado https://rpubs.com/rpizarrog/601901 de (Pizarro, 2020).

```
predicciones <- predict(modelo, nuevas_peticiones, se.fit = TRUE)
prediccion_prob <- exp(predicciones$fit) / (1 + exp(predicciones$fit))
prediccion_prob <- as.data.frame(prediccion_prob)

names(prediccion_prob) <- c("Prob.Predic")
kable(prediccion_prob)
```

y

```
prediccion_prob <- prediccion_prob %>%
    mutate(Préstamo = if_else(Prob.Predic > 0.5, "NO", "SI"))
kable(prediccion_prob)
```

Generando valores ajustados siguientes y de manera respectiva 0.2803811, 0.1092540, 0.4540155 y 0.2252095 y los valores probabilísticos finales conforme a la fórmula:

Prob=Log(fit/(1+Log(fit))) y se tiene el siguiente resultado:

Tabla 5

Resultado de la predicción

Prob.Predic	Préstamo
0.3	SI
0.13	SI
0.51	**NO**
0.24	SI

A todos los clientes se les presta dinero excepto al cliente número 3. Porque su probabilidad de No Pago es por encima del 50%.

Con lo anterior y para fines práctico y de claridad, sólo se describe el modelo regresión logística, dejando de lado todo el modelo ScoreBoard que se encuentra

en la dirección origen de referencia. http://www.creditriskanalytics.net/datasets-private2.html (Baesens, 2016).

5.3.5.8. Análisis de la propuesta

Actualmente R ofrece un amplio conjunto de librerías para el desarrollo de software de riesgo, puede integrarse con casi cualquier base de datos del mercado hablando tanto de las públicas como de las privadas, dejando construir aplicaciones en donde se pueda controlar en todo momento la gestión de los datos, mientras que el entorno R-Studio es un marco completo y muy amigable de desarrollo de código R.

El lenguaje R, es una herramienta que, una vez realizado el esfuerzo inicial imprescindible para adquirir la competitividad mínima, presenta un formidable potencial de estudio a casi cualquier necesidad o problema de análisis de datos que a una empresa se le pueda presentar, siendo un aliado muy interesante y recomendable para aplicar en el ámbito de la Ciencia de los Datos.

Después de analizar la información anteriormente descrita, se puede concluir que hay muchísimas herramientas de software que tratan la problemática, pero todas cuentan con ciertas desventajas.

Aunque la mayoría de las instituciones financieras están implementando y utilizando sistemas de calificación crediticia hoy en día, se enfrentan a una serie de limitaciones. Una primera limitación se refiere a los datos que se utilizan para estimar los modelos de puntuación de crédito. Dado que los datos son el ingrediente principal, y en la mayoría de los casos el único, para construir estos modelos, su calidad y capacidad predictiva es clave para el éxito de los modelos.

La calidad de los datos se refiere, por ejemplo, al número de valores y valores atípicos que faltan, la rectitud y representatividad de los datos. Los problemas de calidad de datos pueden ser difíciles de detectar sin conocimientos específicos del dominio, pero tienen un impacto importante en el desarrollo del cuadro de mandos y las medidas de riesgo resultantes. La disponibilidad de datos de alta calidad es un requisito previo muy importante para crear buenos modelos de puntuación de

crédito. Sin embargo, los datos no sólo deben ser de alta calidad, sino que también deben ser predictivos, en el sentido de que las características capturadas están relacionadas con la probabilidad de incumplimiento del cliente.

Además, antes de construir un modelo de cuadro de mandos, se requiere de realizar una reflexión a fondo sobre por qué un cliente incumple y qué características podrían estar potencialmente relacionadas con esto. Los clientes pueden incumplir debido a razones desconocidas o información no disponible para la institución financiera, lo que plantea otra limitación al desempeño de los modelos de calificación crediticia.

Las técnicas estadísticas utilizadas en el desarrollo de modelos de puntuación de crédito suelen suponer un conjunto de datos de tamaño suficiente que contiene suficientes valores predeterminados. Esto puede no ser siempre el caso para tipos específicos de carteras donde sólo se dispone de datos limitados, o sólo se observa un número bajo de incumplimientos. Para este tipo de carteras, es posible que se tenga que recurrir a métodos alternativos de evaluación del riesgo utilizando, por ejemplo, el juicio de expertos basado en los cinco Cs, como se ha discutido anteriormente.

Conclusiones y Recomendaciones

Los resultados que se pretenden lograr con la propuesta descrita, es que la institución financiera adopte un lenguaje de programación, para realizar el análisis estadístico de datos, obtenidas de las bases de datos con las cuales trabaja; asimismo al adoptar la nueva forma de trabajar

Con el análisis descriptivo de la información se pretende lograr tomar mejores decisiones en cuanto al otorgamiento de créditos, así como también el cuidado de la cartera en riesgo, y en conjunto lograr una cartera de clientes más sana en la empresa financiera.

Se pretende lograr que en un futuro la empresa financiera adopte el uso de algún lenguaje de programación planteados en este proyecto para el análisis de

datos para la evaluación de riesgos en el otorgamiento de créditos a sus clientes. Con ellos se lograría disminuir la cartera en riesgo y evitar pérdidas para la empresa financiera.

Apoyar en la difusión de los lenguajes de R y Python para que un futuro la empresa financiera adopte el uso de algún lenguaje de programación planteado en este proyecto para el análisis de datos para la evaluación de riesgos al otorgar créditos a sus clientes. Con ellos se lograría disminuir la cartera en riesgo y evitar pérdidas para la empresa financiera.

Además, se reducirán costos en la utilización de alguno de estos lenguajes, ya que no será necesario adquirir o dar mantenimiento a algún paquete estadístico con licencia, ya que R y Python son gratuitos.

Se efectuó un análisis comparativo de los lenguajes R y Python, relacionado a una entidad financiera, para realizar análisis de datos estadísticos además se describen distintas técnicas que se tienen en la actualidad para el análisis y predicción de datos estadístico en una entidad financiera.

Se presentó una propuesta de trabajo para elegir la herramienta más viable hablando de lenguajes de programación para el análisis estadístico como los son R y Python.

Se sugirió el lenguaje R ya que por sus características está más enfocado al análisis y predicción de eventos en una entidad financiera.

Se mostró un ejemplo de código R en el análisis, predicción y modelado de datos de una entidad financiera. Ya que es una herramienta que, una vez realizado el esfuerzo inicial imprescindible para adquirir la competitividad mínima, presenta un formidable potencial de estudio a casi cualquier necesidad o problema de análisis de datos que a una empresa se le pueda presentar, siendo un aliado muy interesante y recomendable en un contexto de la Ciencia de los Datos.

El modelo *Credit Scorecard* es un modelo que consiste en un conjunto de atributos o características de un cliente. Dentro de un atributo, los puntos

ponderados se asignan a cada valor del atributo en el rango y la suma de esos puntos es igual a la calificación crediticia final.

El análisis propuesto en la entidad financiera es para evaluar el riesgo de otorgamiento de un crédito a los clientes que solicitan algún tipo de préstamo en un banco.

Con toda esta información recabada, se pretende que en la empresa se tomen mejores decisiones, lograr tener una cartera en riesgo más sana y sobre todo clientes satisfechos con los productos proporcionados por la institución.

Finalmente, el concepto de Big Data ya que permite analizar la información económica en tiempo real, sin necesidad de esperar hasta que las cifras sean publicadas, lo que ocurre con rezagos que regularmente van desde una semana hasta más de un mes.

La estadística ocupa uno de los lugares más importantes dentro de las investigaciones científicas, ya que por medio de esta se realizan evaluaciones cuantitativas sobre las hipótesis de investigación, que después desarrollan modelos predictivos, se llegan a estimar algunos parámetros y se pueden analizar experimentos.

Referencias

Alex, L. (2015). *Data Sciencie and Data Scientist in Global Association for Research Methods and Data.*

alterix. (s.f., s.f. s.f.). *AUTOMATION THAT LETS DATA SPEAK AND PEOPLE THINK.* Retrieved from AUTOMATION THAT LETS DATA SPEAK AND PEOPLE THINK: https://www.alteryx.com/

BBVA Research. (2017, 06 05). *BBVA Research. Artículo de Prensa.* Retrieved from Chile. El potencial de Big Data como herramienta: https://www.bbvaresearch.com/wp-content/uploads/pdf/67403_173947.pdf

Condusef. (2019). Buro de Entidades Financieras. *Condusef, 412*, 412. Retrieved 08 24, 2019, from https://www.condusef.gob.mx/Revista/index.php/usuario-inteligente/servicios-financieros/412-buro-de-entidades-financieras

GetApp. (2019, 07 14). *GetApp. Software de análisis predictivo.* Retrieved from Software de análisis predictivo: https://www.getapp.com.mx/directory/628/predictive-analytics/software

Jones, H. (2019). *Ciencia de los Datos. Lo que saben los mejores científicos de datos sobre el análisis de datos, minería de datos, estadísticas, aprendizaje automático y Big Data que usted desconoce.* México: Amazon Mexico Services, Inc.

Jorge Andrés Alvarado Valencia, J. J. (2008). *FUNDAMENTOS DE INFERENCIA ESTADÍSTICA.* Colombia: Pontificia Universidad Javeriana.

Kozyrkov, C. (2018, 12 22). *Ciencia & Datos.* Retrieved from ¿Qué diablos es Ciencia de Datos?. En la búsueda de una definición inútil.: https://medium.com/datos-y-ciencia/qu%C3%A9-diablos-es-ciencia-de-datos-f1c8c7add107

Lehkyi, S. (2020, 07 10). *What Is Data Analysis?* Retrieved from It is not that complex actually.: https://towardsdatascience.com/what-is-data-analysis-7bb27b5f0d4d

Levine, D. M., Krehbiel, T. C., & Berenson, M. L. (2006). *ESTADÍSTICA PARA ADMINISTRACIÓN. Cuarta edición.* México, Argentina, Brasil: Pearson Educación.

Maimon, O., & Rokach, L. (2010). *Data Mining and Knowledege Discovery Handbook. Secodn Edition.* New York: Springer.

Management Solutions. (2018, 01 01). *MS Management Solutions. Making things happend.* Retrieved from Machine Learning. Una pieza clave en la transformción de los modelos de negocios: https://www.managementsolutions.com/sites/default/files/publicaciones/esp/machine-learning.pdf

Mashanovich, N. (2017, 09 14). *Credit Scoring: The Development Process from End to End.* Retrieved from Credit Scoring: The Development Process from End to End: https://www.worldprogramming.com/blog/datascience/credit_scoring_development_pt1/

Mathworks. (s.f., s.f. s.f.). *Mathworks.* Retrieved from Predictive Analytics: https://es.mathworks.com/discovery/predictive-analytics.html

Oracle Database. (s.f., s.f. s.f.). *Database Machine Learning.* Retrieved from Oracle Machine Learning: https://www.oracle.com/mx/database/technologies/datawarehouse-bigdata/machine-learning.html

Pizarro, R. (2020, 04 20). *RPubs by RSTUDIO.* Retrieved from Evaluar Clientes ScoreBoard: https://rpubs.com/rpizarrog/601901

python. (s.f., s.f. s.f.). *python*. Retrieved from python: https://www.python.org/

RapidMiner. (2020, 01 01). *RapidMiner*. Retrieved from RapidMiner: https://rapidminer.com/

Revuelta, M. (2012). *https://www.google.com/search?q=Revuelta,+M+(2012),+Diccionario+de+finanzas+Bolet%C3%ADn+de+Estudios+Econ%C3B3micos,+67(206)+43.* Retrieved 07 28, 2019, from https://www.google.com/search?q=Revuelta,+M+(2012),+Diccionario+de+finanzas+Bolet%C3%ADn+de+Estudios+Econ%C3B3micos,+67(206)+43

Robertson, S. (2019, 05 01). *Morioh.* Retrieved from Machine Learning algorithms explained: https://morioh.com/p/3d8e92fbbb61/machine-learning-algorithms-explained

r-project.org. (2019, 01 01). *r-project.org*. Retrieved from r-project.org: https://www.r-project.org/

Saenz, J. M. (2005). *ebooks.* Retrieved 08 03, 2019, from https://www.abebooks.com/book-search/author/sarabia-alegria-jose-maria/

SanJuán, F. (2019). *https://economipedia.com/definiciones/institucion-financiera.html.* Retrieved 08 02, 2019, from https://economipedia.com/definiciones/institucion-financiera.html

sas Visual Statics. (2019, 07 14). *SAS Visual Statics.* Retrieved from SAS Visual Statics: https://www.sas.com/es_mx/software/visual-statistics.html

Capítulo 6.

Ciencia de los Datos aplicada en las PyMES en Durango

César Omar Domínguez Gurrola
Tecnológico Nacional de México. Instituto Tecnológico de Durango
03040131@itdurango.edu.mx

Marco Antonio Rodríguez Zúñiga
Tecnológico Nacional de México. Instituto Tecnológico de Durango
mrodriguez@itdurango.edu.mx

Jeorgina Calzada Terrones
Tecnológico Nacional de México. Instituto Tecnológico de Durango
jcalzada@itdurango.edu.mx

6.1. Introducción

En este capítulo se busca dar a conocer las tecnologías relacionadas con el concepto de Ciencia de los Datos más actuales que pudieran aplicar las Pequeñas y medianas empresas (PyMES) del Estado de Victoria de Durango, México.

Se busca que las PyMES al conocer estos aspectos, puedan adoptar y tener ventajas competitivas tales como definir la entrada a algún nuevo mercado, optimizar y/o disminuir costos, planificar la producción de los siguientes meses,

analizar la rentabilidad por producto, sugerir promociones u ofertas por perfil de clientes, identificar claramente el público objetivo, mejorar la eficiencia operativa, estandarizar procesos, definir necesidades específicas de los clientes, agilizar procesos de contratación de personal, generar campañas de mercadotecnia personalizadas, encontrar las mejores rutas y transportes para entregas, entre otras.

La Ciencia de los Datos, Business Intelligence (inteligencia de negocios), Datos Masivos, Machine Learning (aprendizaje automático), entre otras tecnologías de hoy en día, ya han ayudado a muchas empresas y negocios de diferentes partes del mundo incluido México, cada vez se utilizan más y se hacen más populares, el darlas a conocer debe despertar la curiosidad entre las PyMES de Durango.

Los casos reales que se documentan en este capítulo dan un claro ejemplo de los alcances que puede tener el implementar estas tecnologías en las PyMES, además de las nuevas oportunidades de negocios que se generan, sin dejar de lado los nuevos empleos que surgen a partir de estas tecnologías.

El estado de Durango cuenta con 39 Municipios, además de tener 1,759,848 habitantes según el último censo del INEGI levantado en el año 2015. Durango se encuentra en el lugar número 28 de 32 Estados en cuanto al crecimiento neto de PyMES, muy por debajo de la media nacional la cual está en 6.2, Durango aparece con un -1.8. (INEGI. Censos Económicos, 2019)

En la imagen 1 se muestra un mapa del territorio nacional que indica una clasificación de número de establecimientos de comercios al por menor y en donde está ubicado el Estado de Durango con 23845 según datos de INEGI. Censos Económicos 2019, (2019) unidades comerciales comparado con los demás estados está en los de color verde de menor cantidad de comercios al por menor (INEGI. Censos Económicos, 2019).

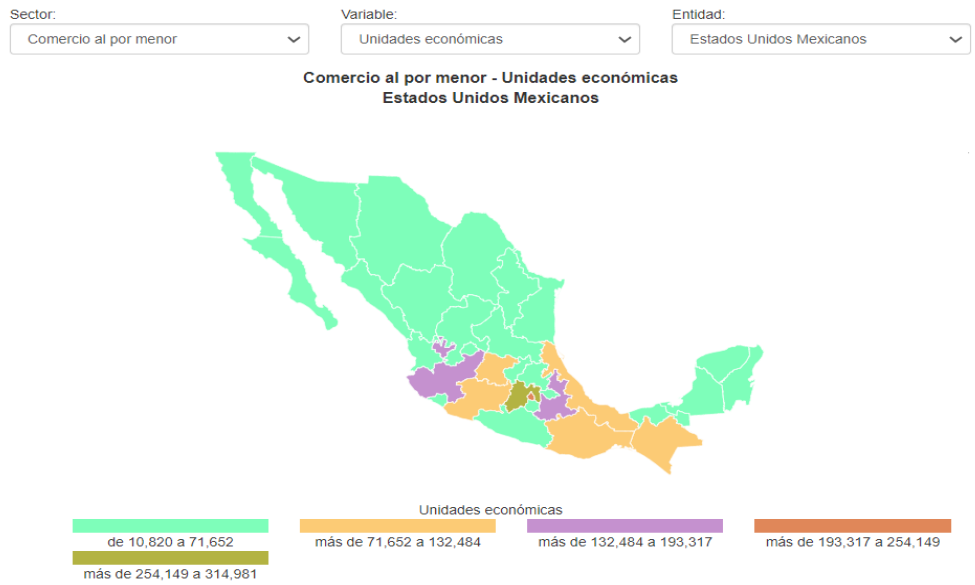


Figura 1. Clasificación de comercios al por menor en México. (INEGI. Censos Económicos, 2019)

En los países avanzados, con altos indicadores de bienestar social y económico, denotan que su progreso es proporcional a la inversión que han hecho en educación, ciencia, tecnología e innovación. Por consecuencia, la lección es clara: si se quiere progresar social y económicamente, se tiene que promover el desarrollo educativo, científico y tecnológico (DURANGO AL DIA, 2018).

Entre las herramientas más usadas destacan los Datos Masivos, que se puede resumir como un conjunto de tecnologías y herramientas capaces de obtener, almacenar, procesar grandes cantidades de datos e información y transformarlas mediante sistemas computacionales y estadísticos, en información útil para toma de decisiones.

La Ciencia de los Datos extrae, manipula y estudia grandes volúmenes de información ya sean datos estructurados o no estructurados y los convierte en el recurso más valioso para las empresas, con el cual se generan estrategias empresariales.

Los científicos de datos tienen mucha importancia en las organizaciones que los contratan debido a que son los encargados de concentrar los datos generados por el negocio, y por medio de sus habilidades en diferentes campos como

estadística avanzada, experiencia en algoritmos, Machine Learning, minería de datos, análisis de la información, logran darle sentido a la información y usarla en favor de las empresas para mejorar o cambiar aspectos específicos.

Son muy pocas las PyMES que entienden que el poder real de la información no está en los ceros y en los unos, si no en su análisis. Obtener las conclusiones correctas y convertir los datos en información valiosa permite revolucionar el mundo de los negocios, a todos los niveles. Son contadas las compañías que se han dado cuenta rápidamente del potencial de los Datos Masivos.

En el comercio al por menor, un ticket de compra proporciona más información de lo que se ve a simple vista, es una de las entradas de datos más importantes de las que dispone un negocio, nos da la oportunidad de hacer recomendaciones personalizadas a los clientes, basadas en compras anteriores, consultas del cliente o de clientes que hagan o hayan hecho compras similares, también puede ayudar a desarrollar un asistente personal de compras, que ayude a las personas a tener una extraordinaria experiencia en el comercio on line.

Este capítulo tiene un sentido de difusión y divulgación, al leerla y razonarla pueda despertar en el lector ya sea en el empresario o cualquier persona la curiosidad y el sentido emprendedor con un enfoque tecnológico que hoy en día toda empresa debe tener.

Las universidades en otros países redoblan la apuesta con programas de análisis y Ciencia de los Datos. Producto de la demanda insatisfecha de expertos en análisis y administración de datos, esto más que futuro es el presente en varios países del mundo. En Universidades, Escuelas de Educación Superior e Institutos Tecnológicos del País, se tiene que empezar a pensar en la necesidad de programas relacionados con las exigencias empresariales en el ámbito de Ciencia de los Datos.

Al dar a conocer estas tecnologías a las PyMES, se pretende que conozcan el impacto que pueden llegar a tener al aplicarlas y con ellos detonar el desarrollo

económico en el Estado de Durango, ya que en la entidad hay 2 principales fuentes de empleo el Gobierno y las pequeñas y medianas empresas.

Una justificación para el desarrollo de este documento es contribuir con la difusión de nuevas Tecnologías basadas en Ciencia de los Datos para disminuir el rezago que existe en Durango en materia de desarrollo PyMES ya que representan una gran parte de la economía Duranguense.

Con este documento académico se intenta proponer las bases, para poder comenzar a integrar a esas PyMES al uso de la Ciencia de los datos, para que se puedan dar cuenta de algunas herramientas que pueden utilizar y poder ser más competitivos en este mundo cada vez más conectado al internet.

También se dan ejemplos de casos reales ya aplicados en otras ciudades de otros países para que se puedan dar cuenta que todo esto ya está pasando, que no son cuentos futuristas y que tarde o temprano llegará al País y a los estados y que como la gran mayoría de las oportunidades laborales, profesionales y comerciales quien esté preparado es quien tiene más posibilidades de continuar.

El objetivo general que se persigue es dar a conocer a las PyMES en Durango las tecnologías inherentes a la Ciencia de los Datos para ayudar a sus negocios a mejorar procesos de diferentes formas.

Estas diferentes formas en que se pueden beneficiar las PyMES pueden ser: identificar claramente el público objetivo, mejorar la eficiencia operativa, estandarizar procesos, definir necesidades específicas de los clientes, agilizar procesos de contratación de personal, generar campañas de mercadotecnia personalizadas, encontrar las mejores rutas y transportes para entregas, entre otras.

Con lo anterior, de manera específica se persigue dar a conocer a las PyMES como la Ciencia de los Datos puede dotar de herramientas que les permitan dar ventajas competitivas y que ayuden a que los negocios crezcan y por ende que tengan un periodo de vida mayor y no desaparezcan.

Se persigue también, ayudar a las PyMES a comprender que la Ciencia de los Datos es una evolución natural de la tecnología y que al invertir en ella más que

una pérdida de tiempo y dinero, como se suele ver en las empresas, es un detonante que ayuda a ser más competitivos

Se pretende también, mostrar a los negocios que en estados y ciudades de otros países se está aplicando la Ciencia de los Datos en beneficio de las PyMES y que no son un concepto de moda como muchas veces se menciona debido a la desinformación de las personas muchas veces causada por el atraso científico y tecnológico que existe en el Estado.

Por medio de la aplicación de la Ciencia de los Datos se favorece el empoderamiento y facilita a las personas dueños o encargados de las PyMES la toma de decisiones.

Las PyMES que cuentan con el servicio de Científicos de Datos pueden tomar decisiones basados en pruebas cuantificables, estas decisiones basadas en información previamente analizada pueden llevar a las empresas a una mayor rentabilidad que es lo que buscan todos los negocios, además de una mayor eficiencia operativa y una estandarización de procesos.

En los negocios que están orientados a los clientes, la información producida por la Ciencia de los Datos ayuda a identificar y definir el público objetivo, así como identificar sus necesidades específicas, ayudando a la empresa en toma de decisiones básicas como de que productos ofrecer y tan complicadas como el *stock* (existencia) a manejar o si se necesita o no rentar un local más grande para hacer crecer el negocio.

En negocios que están orientados al reclutamiento de personas o en cualquier tipo de agencia donde se reclute personal, la Ciencia de los Datos ayuda en el procesamiento de las pruebas de aptitudes y vocacionales basadas en los datos del perfil solicitado para encontrar al mejor candidato para cubrir el puesto ayudando a descartar a aquellas personas que no cumplen los perfiles y ahorrando así a las PyMES pérdida de tiempo y dinero, por ejemplo en capacitación de una persona que no cumple el perfil y que se le tendrían que enseñar habilidades con las que no cuenta, además de agilizar los procesos de contratación.

Las empresas orientadas a las ventas y mercadotecnia o con esos departamentos dentro de la empresa pueden usar la Ciencia de los Datos para mejorar la relación con los clientes, así poder verificar que productos compran con qué frecuencia, que productos de los que se venden no ha comprado, que productos que se tienen se han buscado o comprado en otras empresas y generar campañas de mercadotecnia personalizadas para poder aprovechar todas las ventanas de oportunidad de ventas que se puedan generar.

Las empresas que están enfocadas a productos financieros pueden tomar el ejemplo de los bancos los cuales están usando la minería de los datos para ayudarlos a la detección de fraudes y con esto darles mayor confianza a sus clientes.

Las PyMES orientadas a envíos o que tienen servicios de entrega a domicilio utilizan la Ciencia de los Datos para encontrar las mejores rutas para poder hacer sus entregas así como en que horario es más común o solicitado este servicio, incluso se puede llegar a determinar el transporte más indicado para llevar acabo la entrega como puede ser un bicicleta, una motocicleta, un automóvil o algo de mayor tamaño dependiendo de los resultados de las variables que se evalúen como puede ser distancia, tamaño del objeto a entregar, tiempo de entrega, trafico, condiciones climatológicas, entre otros.

6.2. Marco de referencia

En este apartado, se exponen algunos conceptos relacionados tales como Datos Masivos y sus características; Business Intelligence o inteligencia de negocios con sus aspectos de cuadro de mando integral y sistemas de soporte a la decisión y ejecutivos; Ciencia de los Datos sus características y el papel del científico de datos, Machine Learning o aprendizaje automático; al final de este marco de referencia se describe el contexto PyMES en el estado de Durango, México.

6.2.1. Big Data (Datos Masivos)

La expresión Datos Masivos tiene su nacimiento académicamente por el profesor Francis Diebold, de la Universidad de Pensilvania que utiliza el término en su artículo publicado en 2003 “Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting” . Diebold, (2003) cita lo siguiente:

> “Recientemente, mucha buena ciencia, ya sea física, biológica o social, se ha visto obligada a confrontar, y a menudo se ha beneficiado, del fenómeno del " Datos Masivos ". El concepto de Datos Masivos Datos Masivos se refiere a la explosión en la cantidad (y a veces, calidad) de datos disponibles y potencialmente relevantes, en gran parte el resultado de avances recientes y sin precedentes en la tecnología de almacenamiento y registro de datos.” (pág. 2).

Los Datos Masivos son una recopilación de datos tanto de fuentes tradicionales como digitales dentro y fuera de las empresas que representan una fuente para el descubrimiento y el análisis continuo ya sea de los productos o servicios ofrecidos, así como del comportamiento de los consumidores.

El concepto Big Data o Datos Masivos también es conocido como un conjunto de tecnologías y herramientas capaces de obtener, almacenar y procesar grandes cantidades de datos e información, siempre y cuando estén dentro del tiempo y coste asumibles para una organización.

El registro de las interacciones con la web, con las redes sociales, así como las transacciones al adquirir un producto, los registros financieros, uso de tarjetas y todo lo que deje registro electrónico en internet pueden ser considerados parte de los Datos Masivos.

La capacidad que tienen los Datos Masivos, para transformar los datos disponibles gracias a sistemas computacionales y estadísticos, en información útil para generar una ventaja competitiva para las PyMES y beneficios para los clientes es una de sus características más sobresalientes.

Las diferentes herramientas que se utilizan en el análisis de los Datos Masivos ayudan a realizar predicciones para que las PyMES mejoren la toma de decisiones, siempre de la mano del análisis de los clientes ya que al predecir su comportamiento se les puede dar una atención personalizada para cumplir sus requerimientos incluso antes de que ellos mismos sepan cuáles son.

Los diferentes autores no se ponen de acuerdo en las características de los Datos Masivos, pero las más citadas son las famosas 5V's que son: volumen, velocidad, variedad, veracidad, valor. Algunos otros autores incluyen otras tres características como lo es la variabilidad, la volatilidad y la visualización.

6.2.1.1. Volumen

El volumen habla sobre la cantidad de información y/o datos que están almacenados y se utilizan de las bases de datos. La computación en la Nube (cloud computing) es de las nuevas tecnologías que son paralelas a este desarrollo.

El internet de las cosas (IoT) ayuda a que cada vez se genere una creciente cantidad de datos relacionados al comportamiento de las personas, lo que permite analizar sus diferentes comportamientos para generar servicios hasta ahora impensables para las personas.

Otras tecnologías necesarias y que se han popularizado dado que permiten un análisis de forma económicamente viable para los negocios son las herramientas y lenguajes de programación como Hadoop, R, Python, entre otros, que son empleados para estos fines.

El volumen tiene un carácter masivo, las unidades de medida para estas ingentes cantidades de datos son los terabytes (un billón de bytes), petabytes (mil billones de bytes), exabytes (un millón de billones de bytes). (Gutiérrez Puebla, 2018).

6.2.1.2. Velocidad

La característica de velocidad en los Datos Masivos, es una parte muy importante ya que los datos se están generando de manera muy rápida, para que

los resultados arrojados de estos análisis puedan ser relevantes se necesita que se realicen en tiempo real.

A los clientes les encanta disponer de los efectos de los Datos Masivos para obtener de manera rápida los resultados de los mismos y así facilitarles más sus vidas, las empresas pueden obtener a través de la oferta de más servicios o productos una mayor rentabilidad de forma rápida, al momento que un cliente hace una compra se le puede proponer una segunda compra o accesorios para lo que adquirió.

La velocidad tiene que ver con la rapidez para la adquisición de datos, ya que estos se pueden obtener de diferentes maneras y a ritmos diferentes así como la velocidad para almacenarse, procesarse y administrarse. (Sánchez Villaseñor, 2019).

6.2.1.3. Variedad

La variedad, se origina de la manera en que se obtienen los datos derivados del Internet de las cosas y todos los dispositivos conectados a internet.

Las bases de datos estructuradas son aquellas bases que se crean de manera tradicional con formatos claros que llevan una estructura de Bases de Datos aquellos datos que se conforman tradicionalmente con formatos claros que contienen una estructura de datos.

Las no estructuradas, al contrario de las anteriores no tienen formatos transparentes se tienen que vincular con otros datos para poder obtener mejores resultados, pudieran ser videos, audios comentarios en Facebook, Twitter, Instagram, entre otros.

La variedad hace referencia a la diversidad de tipos, formatos y fuentes de datos, desde datos estructurados y no estructurados. Aquí aparece el tradicional enfoque de datos con herramientas y manejadores con enfoque SQL y además las del tipo NoSQL (Not Only SQL) que almacenan y procesan datos en formato JSON HTML, entre otros (Gutiérrez Puebla, 2018).

6.2.1.4. Veracidad

Los datos necesarios para realizar cualquier análisis deben de ser creíbles ya que estos datos son demasiado importantes ya que basados en ellos se generan decisiones y una mala información podría afectar tanto a las empresas como a sus clientes. Hay que tomar en cuenta que mientras más datos se tengan es más difícil poder definir cuáles son de utilidad para poder separarlos de los datos falsos y que no sirven para nada.

Con la aplicación de los Datos Masivos en los negocios cambia radicalmente la forma de toma de decisiones, ya que antes se tomaban linealmente y ahora con el análisis de los grandes volúmenes de información se plantean muchísimas combinaciones que pueden determinar la dirección de la empresa, permitiendo ajustar la información que se tiene a las necesidades de la empresa, pasando a lo que se denomina Decisiones Dirigidas por Datos (DDD).

La veracidad hace referencia al nivel de confianza y de calidad de los datos. Obtener datos de alta calidad se ha convertido en todo un arte e imprescindible para las organizaciones, sobre todo cuando se trata de datos no estructurados (Gil, 2016).

6.2.1.5. Valor

El objetivo final de los procesos de Datos Masivos es crear valor, ya sea entendido como oportunidades económicas o como innovación. Sin él, los esfuerzos dejan de tener sentido, esto significa que si los datos que arrojan los proceso de análisis masivo de datos (Big Data) no se utilizan y no tienen significado entonces de nada sirve (Gil, 2016).

El valor de los datos significa rentabilidad como resultado de la gestión de los datos (BBVA. BIG DATA, 2017).

6.2.1.6. Visualización

Poder visualizar los datos es necesario para comprenderlos y entenderlos con la finalidad de tomar decisiones bajo una correcta interpretación de los mismos. (Gil, 2016); es permitir y lograr que la gran cantidad de datos recolectados,

procesados y analizados sean comprensibles y sencillos de leer mediante técnicas y herramientas adecuadas para la visualización de datos (BBVA. BIG DATA, 2017).

6.2.1.7. Variabilidad

La variabilidad es el entendimiento que se tienen de los datos dependiendo del contexto de los mismos, en ocasiones se pueden obtener diferentes conclusiones dependiendo de la interpretación que se le haya dado y del entorno en donde se mueven.

La variabilidad tiene que ver las variaciones que aparecen en función, por ejemplo, de los dobles sentidos que puede tener un tipo de expresión, la ironía en las frases, los modismos, las costumbres, las expresiones coloquiales. En consecuencia, es necesario comprender el contexto y el significado real de la información. (incubicÓn By Structuralia, 2019).

Lo anterior tiene sentido por ejemplo cuando un algoritmo y analizador de texto libre encuentra frases en redes sociales, documentos, reportes, entre otros que las tiene que asociar a un entorno económico, de gobierno familiar, deportivo, industrial, dependiendo del contexto.

6.2.1.8. Volatilidad

Esta característica mencionada en el documento de Sánchez Villaseñor, (2019) implica al “tiempo de almacenamiento de los datos después de procesarlos, ya que la volatilidad tiene impacto directo en los macro datos, como el volumen y la veracidad, por ello en las organizaciones existen políticas de almacenamiento de datos para que la información no tenga interferencias ni daño” (pág. 21).

Algunos artículos proponen agregar **vulnerabilidad** que tiene que ver con la seguridad de los datos y la **validez** que tiene que ver con la limpieza y que tan precisos son los datos. (datahack, 2020).

La figura 2 resume e identifica las características de las 10 Vs de los Datos Masivos mencionadas anteriormente:

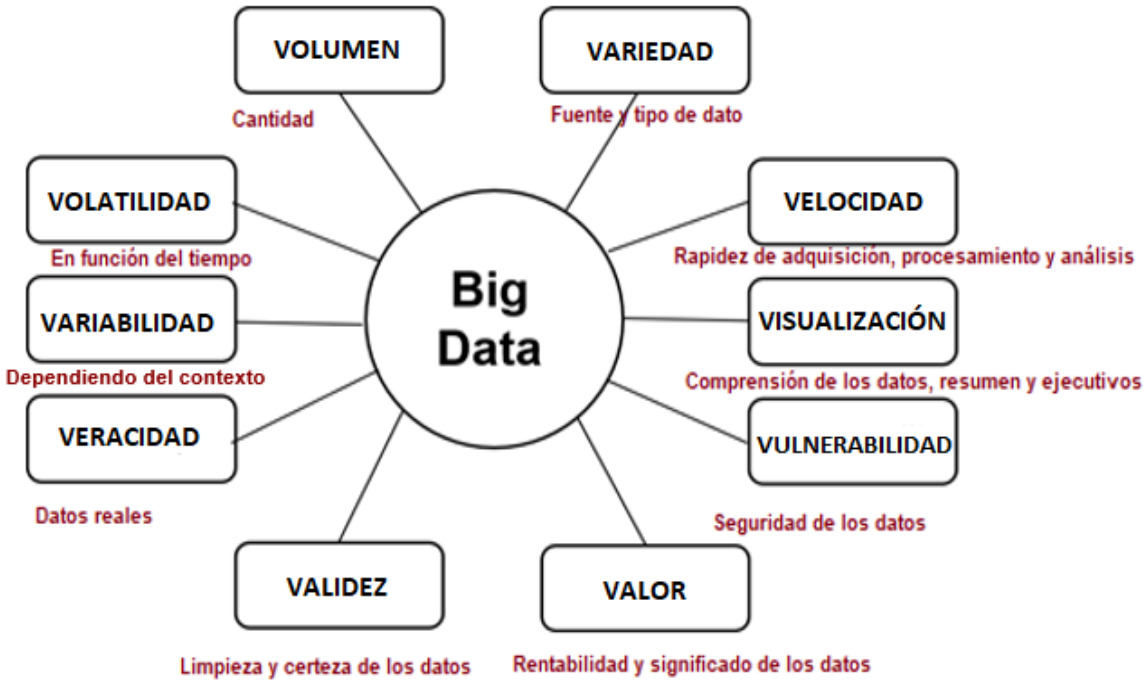


Figura 2. 10 Vs del Big Data. (infogoal.com Goal oriented solutions, s.f.) y elaboración propia

6.2.2. Business Intelligence

Tradicionalmente se ha asociado al Business Intelligence (BI) o inteligencia de negocios como el proceso para procesar datos, convertir esos datos en información y consecuentemente en conocimiento.

Como lo mencionan en su línea de investigación Díaz, Osorio, Amadeo, & Romero, (2013) “es la habilidad para transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios” (pág. 226)

El BI es un conjunto de metodologías, prácticas y capacidades dirigidas al tratamiento de información que permite tomar mejores decisiones a las empresas. La práctica se logra al desarrollar a través de sistemas de TIC un conocimiento profundo de los procesos de interés de la empresa (Silva Solano, 2017).

Puede decirse que el BI es el antecedente inmediato de los Datos Masivos dado que el BI trata con información interna y externa de la empresa, estructurada y no estructurada, la extrae, la procesa, la analiza y la convierte en información homogénea y comprensible para obtener conocimiento que permite una mejor toma de decisiones.

Con el avance tecnológico y recordando las características inherentes a los Datos Masivos de la figura 2 (volumen, variedad, velocidad, valor, veracidad, variabilidad, volatilidad, validez, vulnerabilidad y visualización), es necesario que se adecúen técnicas y herramientas para poder hacer frente a los nuevos retos con que se necesita en las organizaciones para el tratamiento, procesamiento y análisis de los datos; esto es Datos Masivos es una evolución del BI o BI está inmerso en Datos Masivos.

La inteligencia de negocios proporciona información privilegiada para dar respuesta a uno o varios problemas en los negocios actuando como un factor estratégico generando una ventaja competitiva en diferentes ámbitos según el giro o sector al que pertenezca el negocio por mencionar algunas serían: definir la entrada a algún nuevo mercado; control en las finanzas; optimizar y disminuir costos; planificar la producción de los siguientes meses ; analizar la rentabilidad por cada producto; sugerir promociones u ofertas por perfil de clientes; definir un nuevo producto o servicio según necesidades de sus clientes, entre otros.

Finalmente como lo menciona en su trabajo de tesis Sánchez Carrillo & Patnoll Gonzales (2019) “El Business Intelligence actúa como un factor estratégico para una empresa, generando una ventaja sobre la competencia, que no es otra cosa que proporcionar información precisa y fiable para tomar decisiones” (pág. 24)

Los principales productos de Business Intelligence (BI) que existen hoy en día son:

- Cuadros de Mando Integrales (CMI)
- Sistemas de Soporte a la Decisión (DSS)

- Sistemas de Información Ejecutiva (EIS) (Sinergia e Inteligencia de Negocio S.L., s.f.)

6.2.2.1. El Cuadro de Mando Integral (CMI)

El CMI es un modelo de administración que tiene objetivos relacionados entre sí, medidos a través de indicadores aplicando planes de acción que permitan alinear el comportamiento de los miembros de la organización con la estrategia de la empresa, su función primordial es la implantación y comunicación de la estrategia a toda la empresa (cmigestión, 2019).

Los cuadros de mando generalmente están basados en indicadores económico-financieros, el modelo de CMI diseñado por Kaplan y Norton (Balanced Scorecard), el cual describe un balance del aspecto estratégico de la empresa con respecto a otras 3 áreas que son; clientes, procesos internos y el recurso humano como se ilustra en la figura 3.

Figura 3. Modelo CMI de Kaplan Norton. (cmigestión, 2019)

6.2.2.2. Sistemas de Soporte a la Decisión (DSS)

Los sistemas de soporte a la decisión, conocidos por sus siglas en ingles DSS (Decision support system) son sistemas informáticos que procesan grandes cantidades de datos con el fin de proporcionar información y soporte al usuario para la toma de decisiones oportunas. El objetivo principal de los sistemas de soporte a la decisión es proporcionar la mayor cantidad de información significativa en el

menor tiempo posible, con el fin de ayudar a las empresas en la toma de decisiones para generar ventajas competitivas.

De acuerdo al trabajo de investigación de Sánchez Carrillo & Patnoll Gonzales (2019), quien a su vez rescata el concepto el concepto de Cohen & Asín (2004), quien define a un sistema de soporte para la toma de decisiones "como un conjunto de programas y herramientas que permiten obtener de manera oportuna la información que se requiere durante el proceso de la toma de decisiones que se desarrolla en un ambiente de incertidumbre" (pág. 16).

Los sistemas de soporte a la decisión son en la mayoría de los casos de procesamiento analítico en línea mejor conocido como minería de datos esto significa que extraen información importante de manera implícita en los datos.

6.2.2.3. Sistemas de Información Ejecutiva (EIS)

Los sistemas de información ejecutiva pueden definirse como un sistema de información diseñados especialmente para que los altos ejecutivos o dueños de negocios puedan tener acceso a la información interna y externa de sus negocios proporcionándoles todas las herramientas necesarias que les ayuden en quehacer diario (Guevara Vega, 2016).

Las principales características de los Sistemas de información ejecutiva son: extraen, filtran, organizan, consolidan y distribuyen datos; distribuyen datos internos y externos; identifican y analizan tendencias; se integran con otros sistemas de información; personalizan la información para cada ejecutivo; tienen capacidades "drill down" que significa profundizar en detalles mientras se observando un informe o gráfica.

6.2.2.4. Data Warehouse

Los sistemas y componentes del Business Intelligence se diferencian de los sistemas operacionales en que los datos están desnormalizados para apoyar consultas de alto rendimiento además que estos datos se nutren de diferentes sistemas operacionales normalizados traduciéndolos e integrándolos, esto también es conocido como Datawarehouse, mientras que en los sistemas operacionales los datos están normalizados para apoyar operaciones de inserción, modificación y

borrado de datos. Lo anterior se muestra en la figura 4 que refleja el proceso de la extracción de los datos para entrar a un proceso de transformación y carga para poderse presentar en informes y reportes ejecutivos utilizando Data Warehouse como repositorio de información.

Figura 4. DataWarehouse Fuente: (areatecnologia.com, s.f.)

6.2.3. Ciencia de los Datos

La ciencia de los datos se puede definir como el estudio de la información, así como la extracción y manipulación de grandes volúmenes de datos tanto estructurados como no estructurados para poder convertirlos en el recurso más valioso de un negocio.

Cuando se trata de Datos Masivos y el procesamiento de los datos es imprescindible adherir la Ciencia de los Datos, como un arte para el análisis de estos grandes datos. Un enfoque diferente del Business Intelligent (BI) y la Ciencia de los Datos es precisamente el tratamiento a la gran cantidad de datos no estructurados que se generan diariamente, a diferencia de los datos estructurados tradicionalmente utilizados en BI. (Jones, 2019).

Ciencia de datos pueden ayudar a las empresas u organizaciones en muchos aspectos tales como:

- Aumenta la eficiencia tanto de la venta como de los procesos
- Ayuda a controlar los costos

- Reconociendo nuevas oportunidades de mercado
- Aumentar las ventajas competitivas
- Ayuda a tomar decisiones más informadas, entre muchos otros.

Las disciplinas informáticas, de estadística y matemáticas son las más empleadas en la Ciencia de los Datos, así como algunas técnicas más específicas como pueden ser, el análisis de conjuntos de datos, Machine Learning o aprendizaje automático, minería de datos, herramientas de graficación y visualización.

La Ciencia de Datos utiliza técnicas de programación para analizar datos, para esto se requiere el desarrollo de cuatro habilidades:

Programación. La programación se puede definir como el proceso mediante el cual se le dan instrucciones a una computadora para poder generar un resultado mediante la habilidad computacional de poder reducir una tarea compleja en una serie de códigos interpretados. Cabe mencionar que no todos los problemas pueden ser solucionados por una computadora, pero pueden ayudar a resolver por lo menos alguna de las partes.

Estadística. La estadística es muchas cosas menos aburrida por lo menos para los científicos de datos ya que es increíble todo lo que puede lograrse tan solo obteniendo de los datos, la información más sencilla como pueden ser, media, mediana desviación estándar y los cuartiles, cuando se tiene conocimiento y por qué no decirlo algo de suerte la información obtenida puede ser reveladora solo hay que profundizar en ella paso a paso.

Comunicación. Esta habilidad es una de las más importantes ya que sin la capacidad de poder comunicar los resultados que se obtienen, no se podrían aplicar, algunas de estas formas de comunicación que se deben de llevar acabo para poder colaborar con grupos interdisciplinarios pueden ser, encontrar maneras de visualizar los datos de formas que permitan a otras personas interpretarlos y poder obtener soluciones o conclusiones basados en ellos, buscar la manera de explicar procesos complejos, interpretar un modelo estadístico de forma que tenga sentido para el público en general. Esta habilidad requiere empatizar con los demás

para encontrar la forma de discutir los datos usados y los resultados obtenidos con interlocutores muy diversos: público en general, especialistas de diferentes disciplinas, funcionarios públicos, colegas, dueños o directivos de negocios, entre otros.

Conocimiento de dominio. El conocimiento de dominio es la experiencia acumulada en un campo en particular de toda actividad humana como puede ser ganadería, mercadotecnia, física, matemáticas, relaciones públicas, aprendizaje en los niños. Esta habilidad es un complemento básico de las habilidades estadístico-analíticas ya que por medio del conocimiento de dominio se puede plantear que preguntas se deben responder en un complejo problema, además se puede dar cuenta si los resultados obtenidos mediante complejos análisis estadísticos tienen sentido.

Las cuatro habilidades están presentes en cada tarea que involucra la Ciencia de Datos, en mayor o menor medida de acuerdo a la complejidad de cada tarea.

6.2.4. Científico de datos

El científico de datos es una profesión muy atractiva entre las demás profesiones ya que en estos tiempos está llamando la atención en todas las empresas no importando si son privadas, publicas, de servicios, comercio, industriales, agropecuarias de extracción, entre otros. La figura 5 esquematiza la imagen de un científico de datos.

Figura 5. Científico de datos. Fuente: http://cioal.com/wp-content/uploads/sites/2/2018/11/Data-Analyst.jpg

Los científicos de datos adquieren importancia en las organizaciones que los contratan debido a que son los encargados de concentrar los datos generados por él o los negocios, en la mayoría de los casos son datos no estructurados y por medio de sus habilidades en diferentes campos como estadística avanzada, experiencia en algoritmos y codificación de los mismos, Machine Learning, minería de datos, análisis de la información.

Para poder tener los mejores resultados los Científicos de datos deben poseer inteligencia emocional además de ser expertos en el análisis de datos y por si fuera poco, ser capases de transformar esa información no estructurada en modelos que puedan ser interpretados por otros como pueden ser, ejecutivos de nivel directivos o dueños de las empresas para que ellos entiendan de manera clara y precisa, y puedan darle la importancia a esa información para que les ayude en la toma de decisiones y con estas puedan aumentar sus ventajas competitivas, puedan controlar sus costos, hacer más eficiente los procesos y todo lo que se necesite implementar basado en los resultados del análisis de datos.

La información con la que trabaja el científico de los datos puede provenir de cualquier fuente digital, las cuales están en constante crecimiento, algunas de las

fuentes más comunes son, redes sociales como Facebook, teléfonos inteligentes, dispositivos conectados en la nueva era que es el internet de las cosas o el internet del todo, así como cosas tan simples que en ocasiones no se puede percatar como pueden ser, simples encuetas, compras por internet, búsquedas y visualización de videos. Con toda esta información, el científico de datos identifica la manera de resolver problemas mediante la identificación de patrones, a este proceso se le conoce como minería de datos.

6.2.5. Machine learning (Aprendizaje Automático)

El inicio del Machine Learning parte de la teoría de que las computadoras pueden aprender de los datos sin ser programados para tareas específicas, cada vez que los modelos son expuestos a nuevos datos estos se adaptan de diferentes formas, aprenden de los análisis anteriores para producir nuevas decisiones confiables y repetibles.

Muchos de los algoritmos que se están aplicando a Datos Masivos tienen mucho tiempo de haber sido creados pero al aplicarlos a los grandes volúmenes de información, están dando resultados cada vez más interesantes dada la velocidad con que se está creando y analizando la información hoy en día. La figura 6 obtenido como banner de https://www.freepik.es/vector-premium/machine-learning-banner-web-icon-set-mineria-datos-algoritmo-red-neuronal_2436829.htm representa el Machine Learning que identifica diversas aserciones: utiliza minería de datos, algoritmos diversos de regresión, clasificación clustering; existe aprendizaje con los datos, puede emplear utiliza redes neuronales y aprendizaje profundo; es parte de la inteligencia artificial entre otras cosas.

La figura 6, identifica aspectos de Machine Learning.

Figura 6. Machine Learning: Fuente: (freepick, s.f.).

Una de las principales aplicaciones del Machine Learning es en la mercadotecnia y las ventas, se utiliza para analizar el historial de compras que ha tenido una persona para en base a eso promocionar distintos artículos que podrían gustarle a la persona y busca la manera de mostrárselos sistemáticamente para poder generar una nueva venta.

El futuro muy cercano del comercio físico y electrónico estará basado en la capacidad que puedan tener las empresas para generar, capturar, almacenar y analizar los datos, para poder dar a cada cliente una experiencia de compra personalizada para poder tener mayor probabilidad de éxito en una venta y así poder tener mayor crecimiento en un tiempo menor y evitar desaparecer en los primeros años de vida de un negocio.

Los conceptos de inteligencia artificial y Machine Learning, muchas veces son utilizados indistintamente, pero no son exactamente lo mismo; El Machine Learning deriva de la inteligencia artificial, está basado en que las máquinas pueden aprender de manera autónoma al darles acceso a los datos y la inteligencia artificial es un concepto mucho más amplio, en resumen explica que las máquinas son capaces de realizar tareas de una forma tan desarrollada que las personas las consideran inteligentes.

6.2.6. Contexto PyMES y Tecnología en el Estado de Durango

El lugar donde se propone implementar la Ciencia de los Datos enfocada a las PyMES es el estado de Durango el cual cuenta con 39 Municipios, además de tener 1,759,848 habitantes según el último censo del INEGI levantado en el año 2015.

Según el "Análisis de la demografía de los establecimientos 2012" cuyo objetivo es: Generar información al número de establecimientos micro, pequeños y medianos, ubicados en todo el territorio nacional, para los sectores de industrias manufactureras, comercio y servicios privados no financieros, con un personal ocupado de 0 a 100 personas.

Durango se encuentra en el lugar número 28 de 32 Estados en cuanto al crecimiento neto de PyMES, muy por debajo de la media nacional la cual está en 6.2, Durango aparece con un -1.8 como se puede apreciar en la figura 7.

Figura. 7 Crecimieto neto PyMES Fuente: (INEGI. Censos Económicos, 2019)

Según el (**DENUE**) que es el Directorio Estadístico Nacional de Unidades Económicas, en él se ofrecen los datos de identificación, ubicación, actividad

económica y tamaño de los negocios activos en el territorio nacional. En su última actualización realizada en abril del 2019 Durango cuenta con 63,004 empresas de las cuales 62,509 cuentan con 100 empleados o menos, las cuales son consideradas como PyMES.

Se tiene el dato de que el 99.21% de las empresas registradas en Durango son PyMES, destacando el comercio al por menor con 24,424 representando el 39.07 % de las PyMES.

6.3. Desarrollo

Son muy pocas las PyMES que entienden que el poder real de la información no está en los ceros y en los unos, si no en su análisis. Obtener las conclusiones correctas y convertir los datos en información valiosa permite revolucionar el mundo de los negocios, a todos los niveles. Son contadas las compañías que se han dado cuenta rápidamente del potencial de los Datos Masivos.

Los Datos Masivos se utilizan en una gran variedad de casos de diversas áreas, independientemente del tamaño de las empresas, los datos de los dispositivos móviles son los más analizados, seguidos de la información de geolocalización que es, de momento, de las más cotizadas, enseguida se tienen las redes sociales e Internet, además, en las grandes compañías, también analizan la información proveniente de sensores y demás dispositivos conectados para ganar en eficiencia y efectividad.

6.3.1. Ciencia de los Datos en el sector Retail

En el seminario ”Data Science en el Sector Retail” , organizado por la Cátedra de Ciencia de datos y aprendizaje automático UAM-IIC, el 1 de diciembre del 2017 en la Escuela Politécnica Superior (EPS) de la Universidad Autónoma de Madrid

(UAM); se analizó los avances tecnológicos específicos en Data Science para la industria del comercio al por menor (Retail). (iic. Instituto de Tecnología del Conocimiento, 2017).

Este sector en específico, está enfocado en el aumento de las ventas y la retención de clientes, actualmente genera grandes volúmenes de datos constantemente. (iic. Instituto de Tecnología del Conocimiento, 2017)

Implantar proyectos de Datos Masivos y Data Science supone un impulso estratégico y diferencial para afrontar el futuro del comercio al por menor.

En el comercio al por menor, un ticket de compra proporciona más información de lo que se ve a simple vista, es una de las entradas de datos más importantes de las que dispone un negocio, además de indicar del número de clientes y el volumen de ventas, también proporciona información valiosa de la experiencia de usuario; para el fabricante representa la dinámica de venta del distribuidor al usuario final de sus productos.

Figura 8. Datos en Ticket de compra Fuente: (Aldea, 2017)

Al analizar los datos de un ticket de compra y/o factura simplificada ofrece la oportunidad de hacer recomendaciones personalizadas a los clientes, basadas en compras anteriores, consultas del cliente o de clientes que hagan o hayan hecho compras similares en diferentes lapsos de tiempo, incluso en diferentes sucursales si se cuenta con ellas, conocer el tipo de detalles que incluyen los tickets de compra

también puede ayudar a desarrollar un asistente personal de compras, que permita a las personas a tener una extraordinaria experiencia en el comercio en línea.

El ticket, como una solución centrada en el cliente, claramente es un acto de compra, pero mediante un análisis se pueden generar cupones personalizados para incentivar la venta de ciertos productos basados en los consumos del cliente, se puede sugerir productos similares o dar rotación a productos con bajas ventas con promociones adecuadas, además se puede incentivar la frecuencia de visita del establecimiento al incluir fechas límite para obtener los beneficios de los cupones.

Como un extra a toda esta variedad de beneficios que se le dan al cliente para incentivar las ventas, se pueden alimentar algoritmos que detecte anomalías en caja, así como generar análisis que ayuden a controlar el stock (existencia) en tiempo real, entre otros.

Uno de los retos más grandes en el comercio al por menor es la optimización de procesos, los cuales se pueden llevar acabo por medio de algunas de las aplicaciones prácticas de Datos Masivos y Machine Learning como son, la creación de modelos de predicción de ventas, segmentación personalizada de clientes, predicción del cambio de preferencias en los clientes y detección de fraude, entre otras.

Cuando un negocio maneja millones de datos de diferentes productos (de gran consumo, de baja rotación, de alta rotación, entre muchos otros) segmentados por secciones, lo mejor es usarlos para mejorar el servicio mediante la optimización de los procesos, los cuales pueden usarse para detectar algún fraude interno.

Se pueden cruzar con algunos otros datos, por ejemplo, con datos de meteorología para mejorar la predicción de ventas dependiendo del clima como cuando empezar a vender impermeables y botas plásticas en los días lluviosos, se pueden cruzar con datos escolares para planificar la demanda de uniformes y calzado escolar o lo más común que es hacer un perfilado de clientes para poder ofrecer las mejores opciones dependiendo de los gustos y el historial de compras.

Al aplicar el análisis de datos y algunas técnicas se pueden ayudar al comercio al por menor a identificar cuánto se puede llegar a vender de un producto, el cual tiene baja rotación si se le aplica un descuento o una promoción de 2X1 o 3 X2, entre otros.

En época de cambios es muy importante saber cómo cambian las cosas, cuando se crea una nueva tendencia y hacia dónde se dirige y sobre todo saber que está en el gusto de los clientes, últimamente se están aplicando técnicas de geolocalización y Datos Masivos aplicados al comercio al por menor.

Al usar las adecuadas técnicas de geolocalización puede ayudar a los negocios a localizar cual es el mejor punto para poder optimizar sus ventas dependiendo del análisis de los clientes así como determinar cuál sería el horario de apertura idóneo dependiendo del tráfico tanto de autos como de personas que están en determinada zona, incluso se puede determinar cómo se mueve el cliente dentro de la tienda para poder colocar los productos de manera estratégica para tener una mayor impacto de determinados productos y así incentivar la venta.

Los Datos Masivos está generando un mundo nuevo lleno de oportunidades, las empresas que están en condiciones de aprovecharse de ello son las primeras que llegan al mercado con productos y servicios que cubren y dan solución a las necesidades y deseos de los clientes.

La mejora de la experiencia del cliente y la eficiencia de los procesos del negocio deben ser las prioridades al realizar un análisis de Datos Masivos.

Con los datos que provienen de los smartphones o teléfonos inteligentes, de las aplicaciones móviles, del comercio electrónico, de los diferentes sistemas de punto de venta, entre otros, las empresas están en una posición inigualable con la capacidad de recopilación de esos grandes volúmenes de información y el correcto análisis pueden saber que productos o servicios son los que tienen mayor demanda así como cuales no la tienen.

Se puede hacer un análisis más profundo donde se le incluyeran diferentes variables para obtener mejores resultados, por ejemplo, el clima, determinar cuál es

el mejor método para vender en días lluviosos, tal vez pudiera dar como resultado el que sea factible meter un servicio de entrega a domicilio ya que en días lluviosos la gente tiende a salir menos de sus hogares.

Las PyMES que son más ágiles en estos temas están ajustando sus estrategias de negocio para aumentar sus ventas y fidelizar a sus clientes para poco a poco irle ganando más mercado a sus competidores, cuando se ejecutan bien, las mejoras en la experiencia del cliente pueden ayudar a fidelizarlo además de incrementar los ingresos de manera exponencial.

6.3.2. Big Data en Kroger

Ejemplos de uso de Big Data o Datos Masivos, se pueden encontrar en todo el mundo, se cita el caso concreto en Estados Unidos de América, para ser más específicos en un supermercado de nombre Kroger.

Primeramente, empezar con unos datos duros provenientes de la Asociación de mercadotecnia Directo de EEUU la cual indica que la tasa de retorno de promociones o cupones de correo directo es en promedio 3.7 %, el supermercado Kroger, realizando ligeras modificaciones en las estrategias de mercadotecnia y ventas provocadas por el uso de Datos Masivos, pudo obtener una tasa de uso de cupones enviados por correo directo del 70% en un plazo de seis semanas.

Lo más interesante es como lo logró porque parece fácil decirlo, pero requiere de un excelente análisis, su éxito se basó en personalizar su software de correo directo basándose en el historial de compra del cliente individual, además de que Kroger tiene un programa de tarjeta de fidelidad que está clasificada en el número uno de la industria de la alimentación en EEUU (Softwarecamp. Capacitación para el futuro. , 2016)

Más del 90% de sus clientes utilizan la tarjeta cuando compran lo que les ayuda a reunir información de sus clientes con la cual se determina cual es la mejor opción de cupón a enviar en el correo personalizado para darle al cliente justo lo

que necesita y de esta manera incentivar la venta, además todos sabemos y hablo por experiencia personal que vas por una cosa al supermercado o a cualquier tienda y sales de ahí con 2 o 3 productos más mínimo.

En la figura 9, se muestra una tienda de supermercado de la cadena Kroger

Figura 9. Tienda Kroger. Fuente: (CASANARE POSITIVO HEMP, s.f.).

Hablando de Business Intelligence, Datos Masivos y/o Ciencia de los Datos, cabe hacer una reflexión, si las empresas quieren multiplicar sus ventas de manera exponencial y aprovechar las nuevas oportunidades de negocios que se están creando por las nuevas tecnologías tienen que hacerse algunas peguntas tales como:

- ¿Qué estrategias están desarrollando para la aplicación de tecnologías basadas en herramientas de Business Intelligence, Datos Masivos y/o Ciencia de los Datos? (EL COMERCIO, 2018)
- ¿Cuáles de estas tendencias deberán tener en cuenta para mantenerse competitivas y en el mercado? (EL COMERCIO, 2018)
- ¿Se conocen qué estrategias en Business Intelligence, Datos Masivos y/o Ciencia de los Datos está utilizando la competencia? (EL COMERCIO, 2018)

6.3.3. CASO BBVA

Otro ejemplo sería el lanzamiento por BBVA de su herramienta Commerce360 en España, que permite conocer datos que hasta ahora solo eran accesibles por las grandes empresas, y que les proporciona una ventaja competitiva a las PyMES.

Es una herramienta web que pone el Business Intelligence avanzado a disposición de todos los comercios que tengan o contraten el TPV (Terminal Punto de Venta) BBVA.

Commerce360 se tienen gráficas, análisis y sugerencias que permiten disponer de información estadística del negocio y el entorno más cercano. El acceso a la herramienta web se realiza con claves personalizadas manteniendo la seguridad de los datos de las empresas, el seguimiento y actualización de los datos se hace de manera mensual (BBVA. BIG DATA, 2017)

Con Commerce360 se pone las tecnologías de datos disponibles únicamente para las grandes empresas a disposición de las PyMES. Gracias a esta herramienta los comerciantes pueden conocer mejor su sector y a sus clientes, lo que les permite mejorar su toma de decisiones (BBVA. BIG DATA, 2017).

BBVA lleva trabajando desde el 2014 con BBVA Data & Analytics, esta herramienta inicio con un lanzamiento parcial en más de 300 comercios en Andalucía España. En el año 2016 abre su contratación a todos sus clientes de BBVA en España que dispongan de un TPV BBVA. Los establecimientos que han utilizado esta herramienta destacan que aporta a su negocio nuevas oportunidades de mejora (BBVA. BIG DATA, 2017)

El Grupo BBVA es pionero en el uso de tecnologías de Datos Masivos. Ahora pone toda su experiencia y conocimiento al servicio de los clientes a través de esta nueva herramienta. Además, pone al servicio de la sociedad los beneficios de su transformación digital, que inició hace ya una década.

El análisis realizado por BBVA Data & Analytics es especialmente útil para las PyMES, pero también tiene su reflejo en la economía del país, ya que al

incrementar las ventas y la actividad de los comercios y en los servicios de hotelería y turismo, además de muchos más, pueden contribuir a generar más empleos y a incrementar el PIB. (BBVA. BIG DATA, 2017).

6.3.4. Caso Confectionary Holding S.L

Otro caso muy concreto en donde tal vez nadie hubiera pensado que se podrían utilizar los Datos Masivos y todas sus innovaciones hoy en día es el de Confectionary Holding S.L, en este momento es un grupo empresarial, referente en el sector del Dulce a nivel nacional (España) e Internacional. Como proveedor global de Dulce (Turrones, Dulces y especialidades, Chocolates y Bombones, grajeados, entre otros.). Experiencia avalada por casi 300 años de historia (CONFECTIONARY HOLDING, s.f.).

Figura 10. Empresa Confectionary Holding S.L (CONFECTIONARY HOLDING, s.f.).

En esta empresa en la cual la receta ha permanecido casi invariable desde 1725. Almendra, miel y azúcar, no significa que no puedan apostar por la innovación. Los Datos Masivos han tomado las plantas de producción de Jijona (Alicante) y Alcaudete (Jaén), dando como resultado la mejora de procesos, el aumento de la competitividad y la agilización de la toma de decisiones.

El objetivo principal de Confectionary Holdings, al aplicar Datos Masivos como lo menciona Samaniego (2017), “es convertir los datos, la información en bruto, en smart data, primero, y en good data, después. No solo se trata de transformar datos en información entendible, sino en información que sea comprensible para la persona que tiene que tomar decisiones”. (Samaniego, 2017).

Desde el año 2015, la maquinaria de las dos plantas de producción se ha ido conectando de forma gradual y se le ha dotado de diferentes sensores en la cadena de suministros y proveedores. Todos estos sensores arrojan datos los cuales se analizan de forma continua y los resultados se ponen a disposición de toda la compañía divididos en tres niveles principales: (Samaniego, 2017)

- Para la directiva,
- Para los mandos medios
- Para cada puesto de trabajo específico.

De igual forma como se describe en Samaniego (2017) la verdadera potencia, en dotar de información a este segundo y tercer nivel (a los mandos medios y a cada persona y su trabajo específico); proveer de información al resto de la organización ha permitido descentralizar la toma de decisiones, dotar de autonomía a los trabajadores y ganar agilidad de respuesta en un mercado que demanda rapidez (Samaniego, 2017).

De esta manera la intuición y la experiencia que tanto significado tenían antes ahora están siendo sustituidas por las decisiones tomadas con datos, de esta manera se ha implicado a una parte importante de la empresa en la toma de decisiones. Esto ha contribuido a mejorar la productividad y el desempeño de las fábricas, así como a enriquecer y simplificar la organización en un entorno cada vez más complejo.

Los Datos Masivos parecen muy lejanos y poco entendibles para muchas empresas que prefieren invertir en algo más tangible y en aspectos más concretos, como, por ejemplo, los puntos de venta. Sin embargo, muy pronto, el mercado estará conduciendo hacia situaciones mucho más complejas que se van a hacer

muy difíciles de gestionar y es en donde el uso de la tecnología pudiera ayudar. (Samaniego, 2017).

6.3.5. Smart Cities

La Ciencia de los Datos, Business Intelligence, Datos Masivos, Machine Learning, Internet de las Cosas (IoT) entre otras, están siendo aprovechadas en muchos países principalmente en Estados Unidos y Europa con el auge de las ciudades inteligentes o Smart Cities.

El estudio Cities In Motion Index (CIMI) 2017, realizado por el trabajo en conjunto entre el Centro para la Globalización y la Estrategia y el Departamento de Estrategia de la IESE Business School de la Universidad de Navarra (España). En él se contemplan 10 indicadores con soportes estadísticos rigurosos y sus conceptos más relevantes para definir cuáles son las ciudades a nivel global y regional que se identifican como más innovadoras, globalizadas y ambientalmente sostenibles.

Entre los indicadores que destaca el informe de Berrone & Enric (2017) se incluyen, por ejemplo,

- Capital humano que incluye indicadores tales como alta educación, número de universidades, movilidad estudiantil, museos, galerías y gastos en eventos de recreación, entre otros.
- Cohesión social. agrupa aspectos tales como relación decesos por número de habitantes, indicadores de criminalidad e inseguridad, violencia, paz social, aspectos de salud, condiciones sociales, indicadores de precios de inmuebles, entre otros.
- Economía: integra algunos indicadores como índice de precios, número de empresas, empleos, salario mínimo, generación de producto interno bruto (PIB), tiempo para iniciar un negocio, tiempo para consolidar un negocio, porcentaje de empresarios y emprendedores comparado con la población, condiciones industriales y comerciales, entre otros.

- Administración pública relaciona aspectos como tazas de impuestos, , reservas totales, reservas per cápita, embajadas, puertos, personajes como activistas, líderes, hombres de negocios, entre otros.
- Gobernanza. Es un término utilizado para medir la eficiencia en procesos administrativos gubernamentales tales como fortalea en las leyes, indicadores de corrupción, eficiencia, porcentajes de sitios WEB para eficiencia en procesos, indicadores de calidad en el servicio al público, datos abiertos, entre otros.
- Calidad del medio ambiente que reflejas aspectos principalmente de contaminación tales como emisiones de contaminantes e indicadores de contaminación, limpieza agua, metales, generación de energías, entre otros.
- Movilidad y transporte tiene que ver con indicadores de tráfico, calidad en los servicios de rutas urbanas, conectividad de aeropuertos, puertos, centrales de autobuses, medios de trasportes, indicadores de accidentes, flujos vehiculares, entre otros.
- Planeación urbana agrupa aspectos como saber y determinar proyectos de crecimiento futuro de una ciudad que contenga aspectos de sustentabilidad y calidad de vida en los ciudadanos.
- Indicadores excepcionales con reconocimiento internacional que denota aspectos tales como el número de turistas, visitantes, pasajeros que arriban a la ciudad, hoteles, tiendas de autoservicios, eventos relacionados con conferencias de talla internacional, eventos turísticos, regionales, entre algunos otros.
- Aspectos y uso de tecnología que tiene que ver con indicadores de usuarios WEB, con telefonía móvil, direcciones IP, anchos de banda, calidad en servicios de internet, cantidad estantes de internet libre, cantidad de suscriptores en servicios digitales (CABLE) entre otros. (Berrone & Enric, 2017).

La figura 11 muestra las primeras 10 ciudades inteligentes dentro de un ranking a nivel mundial de entre 151 ciudades de acuerdo a Berrone & Enric (2017).

Ranking	City	Performance	CIMI
1	New York City-USA	A	100,00
2	London-United Kingdom	A	98,71
3	Paris-France	A	91,97
4	Boston-USA	RA	88,90
5	San Francisco-USA	RA	88,46
6	Washington, D.C.-USA	RA	86,10
7	Seoul-South Korea	RA	84,91
8	Tokyo-Japan	RA	84,85
9	Berlin-Germany	RA	83,40
10	Amsterdam-Netherlands	RA	82,86

Figura 11. Ranking de Smart Citys (Berrone & Enric, 2017)

En Latinoamérica hay varios países como Argentina, Brasil, Chile, Colombia y México entre otros, los cuales cuentan con ciudades inteligentes, enseguida se muestra el ranking regional de ciudades en América Latina.

Algunas de las características que menciona en su artículo digital (Ospina, s.f.) son que una ciudad inteligente tiene que ver con que tenga indicadores relacionados con la sustentabilidad, la inclusión, el transporte, la generación de economía y precisamente que su desarrollo esté pensado en la calidad de vida de sus ciudadanos. (Ospina, s.f.).

CIUDAD	POSICIÓN REGIONAL	POSICIÓN GLOBAL 2014	POSICIÓN GLOBAL 2015	POSICIÓN GLOBAL 2016
Buenos Aires (Argentina)	1	82	80	83
Santiago (Chile)	2	88	89	85
Ciudad de México (México)	3	100	90	87
Medellín (Colombia)	4	102	100	96
Montevideo (Uruguay)	5	94	101	99

Figura 12. Ranking de Smart Citys en América Latina

Generalmente las primeras ciudades en comenzar a implementar estas tecnologías son las capitales de los Países sin importar lo grande o difícil que parezca implementar estas tecnologías en las grandes urbes.

El hablar de ciudades inteligentes en este capítulo de Ciencia de los Datos aplicados a las PyMES es que precisamente a las empresas se les debe impulsar para que promuevan estrategias y proyectos relacionados con el adecuado uso de los datos, de su organización y del exterior a través de lo que existe y provea la ciudad en términos de datos; con ello puedan obtener un beneficio económico, es decir aumentar utilidades a través de las ventas y la adecuada relación con sus cliente; eficiencia en procesos internos y externos de su negocio.

Conclusiones y Recomendaciones

La Ciencia de los Datos, Business Intelligence, Datos Masivos, Machine Learning, entre otras tecnologías de hoy en día, han ayudado a muchas empresas y negocios de diferentes partes del mundo incluido México, cada vez se utilizan más y se hacen más populares.

Este documento debe despertar la curiosidad entre las PyMES de Durango y poco a poco comenzar a implementar las tecnologías mencionadas para poder tener ventajas competitivas pero sobre todo para ponerse al día por que en otros lugares ya se están utilizando y esto los pone en desventaja

Los casos reales de los que se habló en este capítulo ofrecen un claro ejemplo de los alcances que puede tener el implementar estas tecnologías en las PyMES, además de las nuevas oportunidades de negocios que se generan, sin dejar de lado los nuevos empleos que surgen a partir de estas tecnologías

La intención con este capítulo del libro es aportar un granito de arena en materia tecnológica para con ello ayudar al desarrollo de las PyMES.

Al dar a conocer estas tecnologías a las PyMES, se pretende que conozcan el impacto que pueden llegar a tener al aplicarlas y con ellos detonar el desarrollo económico ya que en Durango hay 2 principales fuentes de empleo el Gobierno y las PyMES, al darle crecimiento a las empresas ellas podrán generar más empleos y mejor pagados.

También se espera que una vez que se apliquen estas áreas de conocimiento por las PyMES, las grandes empresas puedan voltear a ver como una opción para instalarse en el Estado al ver que se cuenta con todo lo que necesitan, además de que hay mano de obra calificada con esto ayudar a que el Estado de Durango tenga un desarrollo de todos los sentidos.

Referencias

Aldea, V. (19 de 01 de 2017). *La factura simplificada y el ticket son lo mismo*. Obtenido de La factura simplificada y el ticket son lo mismo: https://anfix.com/blog/ticket-gastos-y-facturas-simplificadas/

areatecnologia.com. (s.f. de s.f. de s.f.). *TECNOLOGÍA*. Obtenido de DATA WAREHOUSE: https://www.areatecnologia.com/informatica/data-warehouse.html

BBVA. (27 de 06 de 2017). *BIG DATA*. Obtenido de Llega Commerce360, una herramienta de tecnología 'big data', para ayudar al desarrollo de las pymes: https://www.bbva.com/es/bbva-presenta-commerce360-herramienta-que-usa-tecnologia-big-data-para-ofrecer-a-los-clientes-mayor-conocimiento-sobre-sus-negocios-y-les-ayuda-a-crecer/

BBVA. BIG DATA. (26 de 02 de 2017). *BIG DATA. Las siete 'V' del Big Data*. Obtenido de BIG DATA. Las siete 'V' del Big Data: https://bbvaopen4u.com/es/actualidad/las-siete-v-del-big-data

Berrone, P., & Enric, R. J. (2017). *IESE Cities in Motion.* Navarra: IESE Business School. University of Navarra.

CASANARE POSITIVO HEMP. (s.f. de s.f. de s.f.). *Kroger, la cadena de supermercados más grande de los EE. UU, está presionando sobre el cáñamo*. Obtenido de Kroger, la cadena de supermercados más grande de los EE. UU, está presionando sobre el cáñamo: http://casanarepositivoparahemp.com/2020/02/12/kroger-la-cadena-de-supermercados-mas-grande-de-los-ee-uu-esta-presionando-sobre-el-canamo/

cmigestión. (s.f. de s.f. de 2019). *Cuadro de Mando Integral*. Obtenido de Cuadro de Mando Integral: https://cmigestion.es/cuadro-de-mando-integral/

Cohen, D., & Asín, E. (2004). *Sistemas de Información para los Negocios un enfoque de toma de decisiones.* México: McGraw-Hill.

CONFECTIONARY HOLDING. (s.f. de s.f. de s.f.). *Confectionary Holding*. Obtenido de Las webs de nuestras marcas: http://www.confectionaryholding.com/index.php/es/?option=com_content&view=article&id=67

datahack. (27 de 01 de 2020). *datahack. BIG DATA FAMILY*. Obtenido de LAS 10 V'S DEL BIG DATA: https://www.datahack.es/10-vs-del-big-data/

Díaz, F. J., Osorio, M. A., Amadeo, A. P., & Romero, D. L. (2013). *Aplicando estrategias y tecnologías de Inteligencia de Negocio en sistemas de gestión académica.* Paraná, Entre Rios: XV WORKSHOP DE INVESTIGADORES EN CIENCIAS DE LA COMPUTACIÓN.

Diebold, F. (2003). “Big data” dynamic factor models for macroeconomic measurement and forecasting. *Cambridge: Cambridge University Press*, 115-122.

DURANGO AL DIA. (08 de 02 de 2018). *DURANGO AL DIA. La fuerza de la verdad*. Obtenido de Necesita Durango más investigadores: http://www.durangoaldia.com/necesita-durango-mas-investigadores/2018/02/

EL COMERCIO. (17 de 05 de 2018). *Zona Ejecutiva Tendencias*. Obtenido de Business Intelligence: 10 temas que serán tendencia este 2018: https://archivo.elcomercio.pe/especial/zona-ejecutiva/tendencias/business-intelligence-10-temas-que-seran-tendencia-este-2018-noticia-1993317

freepick. (s.f. de s.f. de s.f.). *freepick*. Obtenido de freepick: https://www.freepik.es/vector-premium/machine-learning-banner-web-icon-set-mineria-datos-algoritmo-red-neuronal_2436829.htm

Gil, E. (2016). *Big data, privacidad y protección de datos.* Madrid: Publisher: Agencia Española de Protección de Datos y Boletín Oficial del Estado ISBN: 9788434023093.

Guevara Vega, C. P. (12 de 08 de 2016). *Evaluando Software.com*. Obtenido de Qué es un EIS: sistema de Información Ejecutiva: https://www.evaluandosoftware.com/eis-sistema-informacion-ejecutiva/

Gutiérrez Puebla, J. (2018). Big Data y nuevas geografías: la huella digital de las actividades humanas. *Documents d’Anàlisi Geogràfica. eISSN: 2014-4512*, 195-217.

iic. Intituto de Tecnología del Conocimiento. (10 de 11 de 2017). *Seminario: Data Science en el Sector Retail*. Obtenido de Seminario: Data Science en el Sector Retail: https://www.iic.uam.es/noticias/data-science-sector-retail-seminario/

incubicÓn By Structuralia. (23 de 07 de 2019). *Incubicon*. Obtenido de Del volumen a la visualización, descubre las 7 V´s del Big Data: https://blog.incubicon.com/del-volumen-a-la-visualizacion-descubre-las-7-v-del-big-data

INEGI. Censos Económicos. (16 de 07 de 2019). *INEGI*. Obtenido de Censos económicos 2019: https://www.inegi.org.mx/programas/ce/2019/

infogoal.com Goal oriented solutions. (s.f. de s.f. de s.f.). *GOAL DIRECTED LEARNING*. Obtenido de Big Data Tutorial: http://infogoal.com/datawarehousing/big_data_tutorial.htm

Jones, H. (2019). *Ciencia de los Datos. Lo que saben los mejores científicos de datos sobre el análisis de datos, minería de datos, estadísticas, aprendizaje automático y Big Data que usted desconoce.* México: Amazon Mexico Services, Inc.

Ospina, J. P. (s.f. de s.f. de s.f.). *Ciudades Intelgentes en Amética Latina*. Obtenido de Ciudades Intelgentes en Amética Latina: http://conexionintal.iadb.org/2018/11/27/267_e_ideas6/

Samaniego, A. (13 de 12 de 2017). *Hablemos de Empresas*. Obtenido de Cuando Big Data es igual a 'real money': tres casos de éxito de Big Data orientado a negocio: https://hablemosdeempresas.com/grandes-empresas/casos-de-exito-big-data-orientado-a-negocio/

Sánchez Carrillo, J. A., & Patnoll Gonzales, L. J. (2019). Desarrollo de un DataMart para el Soporte de la Toma de Decisiones en el áea de ventas de la empresa. *Tesis. PRESENTADA PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS*. Chiclayo , Chiclayo , Perú: Universidad de Lambayeque. Facultad de Ciencias de Ingeniería.

Sánchez Villaseñor, O. (01 de 04 de 2019). HERRAMIENTAS, RETOS, OPORTUNIDADES, SEGURIDAD Y TENDENCIAS DEL BIG DATA. *Tesina para obtener el título de INGENIERO EN COMPUTACIÓN*. Toluca , Estadode México, México: UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO. FACULTAD DE INGENIERÍA.

Silva Solano, L. (2017). Business Intelligence: un balance para su implementación. *InnovaG*, 27-36.

Sinergia e Inteligencia de Negocio S.L. (s.f. de s.f. de s.f.). *Business Intelligence*. Obtenido de ¿Qué es Business Intelligence?: https://www.sinnexus.com/business_intelligence/

Softwarecamp. Capacitación para el futuro. . (12 de 07 de 2016). *BigData y el caso de éxito con efectividad del 70%*. Obtenido de Ventas inteligentes con BigData: https://softwarecamp.mx/articulos/big-data-y-el-caso-del-mailing-con-efectividad-del-70/

Capítulo 7

Ciencia de los Datos con R como herramienta aplicada a la productividad

Beatriz Eneida Chávez Atayde
Tecnológico Nacional de México. Instituto Tecnológico de Durango
02040824@itdurango.edu.mx

Rubén Pizarro Gurrola
Tecnológico Nacional de México. Instituto Tecnológico de Durango
rpizarro@itdurango.edu.mx

7.1. Introducción

La segunda década del siglo XXI ha mostrado avances exponenciales en cuanto a tecnología se refiere, hoy a diferencia de otras épocas, es posible extraer la inmensa cantidad de información que se produce en todos ámbitos, se puede analizar con herramientas computacionales y con el uso de lenguajes de programación (Alcalde San Miguel, 2009, p. 245).

Este documento trata sobre una propuesta de aplicar aspectos de Ciencia de los Datos en procesos de calidad y producción en las empresas de arneses. Los procesos de producción abarcan no solamente el aspecto tecnológico, sino que también están inmersas las relaciones humanas y los datos dentro de su composición, representando una verdadera área de oportunidad la implementación

de análisis detallados que fusionen por un lado las filosofías de la calidad, las iniciativas de mejora continua y el análisis de datos.

El capítulo relaciona las disciplinas de Ingeniería en Sistemas y la Ingeniería Industrial. De manera general en las empresas cada una de estas áreas de estudio mantiene una relación alejada, sin embargo, pueden ser complementarias y aumentar su simbiosis, la productividad empresarial con un enfoque en la calidad. Todo esto mediante análisis de datos utilizando lenguaje de programación R.

La Ciencia de Datos contempla conocimientos de una o más disciplinas: finanzas, medicina, geología, matemáticas, computación, estadística, ingeniería, entre otros. Considera aspectos de investigación tales como prueba, hipótesis y la variación de los resultados. Los resultados obtenidos deben ser confiables e involucrar más aspectos analíticos que los enfoques anteriores, sin olvidar que debe poseer un lenguaje automatizado (García Nocetti, 2017).

La Ciencia de Datos surge ante la necesidad del análisis de grandes volúmenes de información. Dicha disciplina impacta en el grado de fundamentación de las decisiones que se toman en las organizaciones a partir de la información brindada por los sistemas de información utilizados (Rodríguez, 2015).

Una importante justificación para este trabajo académico es que existe la necesidad de fusionar los conocimientos de mejora continua desde una óptica más analítica considerando el crecimiento exponencial en la información recopilada por el sector automotriz que envuelve a todas las esferas de su competencia, tales como recursos humanos, producción, ingeniería de métodos, entre otros.

Como todo sistema, existen deficiencias en los procesos; de acuerdo a la experiencia en el ámbito, se tienen antecedentes y se han observado, fallas en el sistema actual, lo cual ha llevado a que se tengan que tomar medidas adicionales para la validación y verificación de la información que el sistema actual arroja, desembocando es disminución del aprovechamiento de las capacidades de un sistema de planificación de los materiales (*MRP controller*), hecho por el cual se invierte tiempo en hacer verificaciones manuales, lo cual dificulta el desarrollo del

resto de las actividades, pues el tener que contar un inventario de un solo componente manualmente sin tener confianza en los datos que arroja el sistema, por el precedente de las inconsistencias en los resultados.

El lenguaje R es una herramienta esencial en la Ciencia de los Datos, y su enfoque es estadístico, al utilizar R, permitiría desarrollar una metodología más robusta en cuanto al análisis de datos que sea acorde con las necesidades y las habilidades de un *MRP controller* independientemente de su área de especialización y sea a la par funcional como área de oportunidad para la empresa APTIV.

Como objetivo general se pretende presentar una propuesta de implementación del uso del lenguaje de programación R, como herramienta para el análisis de datos de un *MRP Controller* en la empresa APTIV.

De manera específica se busca lo siguiente:

- Presentar un marco de referencia de las herramientas de calidad
- Identificar las técnicas de análisis de datos relacionadas con la filosofía Six Sigma.
- Identificar el lenguaje de programación R y su entorno R Studio como propuesta en el ámbito de la Ciencia de los Datos.
- Relacionar los conceptos de las herramientas estadísticas de calidad con el lenguaje R para análisis de datos.
- Generar una propuesta de implementación para un *MRP Controller*

Como propósito se tiene que, mediante la adecuada implantación de la propuesta, se pueda establecer un mejor control de calidad y productividad en las actividades, garantizando cumplir con las exigencias de los procesos mediante análisis de datos utilizando lenguaje de programación R y R Studio.

El documento cuenta con ejemplos prácticos para implementar análisis de datos relacionados con el enfoque de las herramientas de calidad.

7.2. Marco de referencia

Las herramientas de la Calidad fueron establecidas en 1968 por el Ingeniero japonés Kaoru Ishikawa. Constituyen por si mismas un conjunto de técnicas estadísticas que no demandan de un conocimiento experto para implementarse en los procesos de equipo. Con ellas es posible resolver aproximadamente el 95% de los problemas que presenta una organización; sobre todo en las áreas de calidad y productividad.

Estas herramientas con el transcurso del tiempo se conocieron como “las siete herramientas básicas de la calidad” y pueden ser definidas generalmente como métodos para la mejora continua y la solución de problemas.

Enumerando las siete herramientas básicas de la calidad:

i. Diagrama Causa – Efecto (Diagrama de Ishikawa)
ii. Diagrama de flujo
iii. Hoja de Comprobación (Hojas de Verificación).
iv. Gráficos de Control.
v. Histograma.
vi. Diagrama de Pareto.
vii. Diagrama de Dispersión.

El éxito de estas técnicas puede explicarse debido a la capacidad que han demostrado para implementarse dentro de un amplio rango de problemas, desde el control de calidad hasta las áreas de producción, marketing y administración entre otras. Las organizaciones de servicios públicos representan excelentes campos para su aplicación independientemente que su origen sea el área industrial.

Una de las ventajas generales de dichas herramientas es su funcionalidad gráfica en mandos medios y altos de la organización, facilitan la planificación, el establecimiento de objetivos y la solución de problemas.

Las herramientas permiten analizar y organizar la relación entre datos cualitativos, clarificar interrelaciones, establecer prioridades y planificar tareas complejas para alcanzar una meta. Son de suma importancia para mejorar

procesos, productos y sistemas. En datos cuya naturaleza es la cualitativa, puede decirse que ayuda a la resolución de problemas y a la gestión de ideas innovadoras.

El hecho de contar con una alternativa de análisis robusto como R y R Studio para la elaboración de diagnósticos y reportes puede fortalecer por un lado el estatus laboral del interesado y por otro las opciones de implementación en metodologías de mejora continua abarcarían no sólo a los egresados de ingenierías afines a los temas de calidad, sino que también podrían contemplar con mayor facilidad a los egresados de ingenierías en sistemas e ingenieros informáticos.

7.2.1. Herramientas de calidad y estadísticas

Se presentan algunas herramientas útiles para el monitoreo de calidad y relacionadas con las siete herramientas de calidad.

7.2.1.1. Diagrama de Ishikawa

El diagrama de Ishikawa o diagrama de causa-efecto consiste en una representación gráfica sencilla en la que puede verse de manera relacional una especie de espina central, que es una línea en el plano horizontal, representando el problema a analizar, que se escribe a su derecha (Garza, 2003, p. 72).

Esta herramienta creada a lo largo del siglo XX en el contexto industrial y posteriormente en el de los servicios, permite facilitar el análisis de problemas y sus soluciones en esferas como lo pueden ser; calidad de los procesos, los productos y servicios. En la figura 1 se muestra un ejemplo del diagrama causa efecto

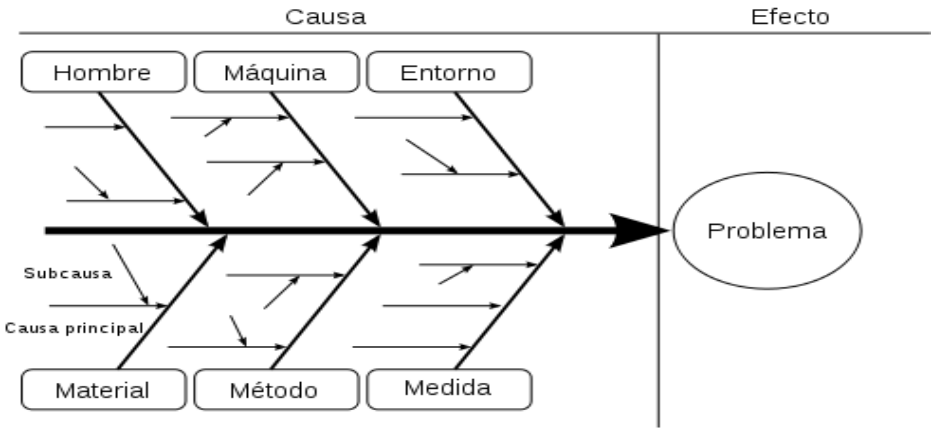


Figura 1 Diagrama de Ishikawa (lean, 2016)

7.2.1.2. Diagrama de Flujo

Un diagrama de flujo es una imagen gráfica de los pasos que se tienen que dar para realizar un proceso; iniciando con un símbolo de entrada, después por medio de otros símbolos representar una serie de acciones y procesos, finalizando con un símbolo de salida o terminación. Representa el orden en que las tareas se llevan a cabo en la organización, también denota la relación lógica entre todas las tareas que componen el diagrama. (Fernández y Fernández & Quintanar Morales, 2015).

Algunas características del diagrama de flujo es que presenta claridad en la información; está formado por símbolos que representan inicio, proceso o acciones, decisiones y finalización; cada símbolo representa una acción específica y las flechas entre los símbolos representan el orden de realización de las acciones. (Fernández y Fernández & Quintanar Morales, 2015).

7.2.1.3. Hoja de comprobación

Las hojas de comprobación son igualmente nombradas hojas de control o de verificación o chequeo, son formatos estructurados en forma de tabla cuya finalidad es registrar la ocurrencia o frecuencia de ciertos sucesos, mediante un método sencillo.

Son formatos utilizados para realizar actividades repetitivas, ayudan a controlar que se cumplan una lista de requisitos y recopilar datos de manera ordenada de forma sistemática. Se emplean para hacer comprobaciones de cumplimiento de actividades o productos (Gómez, 2017). La figura 2, muestra un ejemplo genérico de una hoja de verificación de avisos de incidencia en la vía pública:

AVISOS DE INCIDENCIAS EN LA VÍA PÚBLICA							
		DISTRITOS					TOTAL
		A	B	C	D	E	
INCIDENCIAS INFORMADAS	Alcantarillado	120	30	72	101	25	348
	Calzada	84	18	128	32	64	326
	Aceras	224	54	59	49	16	402
	Señales de tráfico	48	37	97	29	14	225
	Limpieza	173	46	131	42	35	427
	Alumbrado	56	19	53	102	27	257
	Semáforos	31	8	27	16	8	90
	Parques y Jardines	30	12	24	29	16	111
	Arboles	27	28	13	45	25	138
	Total	793	252	604	445	230	2324

Periodo registrado: 1-10-2015 / 30-10-2015
Verificador/a: ********** **********
Método: Análisis de las incidencias informadas por ciudadanos
Periodicidad: 1/mes

Figura 2. Hoja de comprobación. (Aiteco Consultores, Aiteco Consultores S.L., s.f)

7.2.1.4. Gráfico de control

Los gráficos de control se usan para controlar las incidencias sobre los procesos, bajo un enfoque de identificar posibles anomalías e inconsistencias en la ejecución del proceso.

Se pretende que un gráfico de control identifique visualmente, el correcto funcionamiento de los procesos. La gráfica es una serie de punto unidos que deben estar dentro de los límites, de tal forma que si así se encuentran el proceso está controlado, los picos hacia arriba y hacia abajo significa anomalías o cosas que se salen de control (González González & Jimeno Bernal, 2012).

La figura 3, muestra un ejemplo genérico de gráfico de control con un pico en el número de muestra 16:

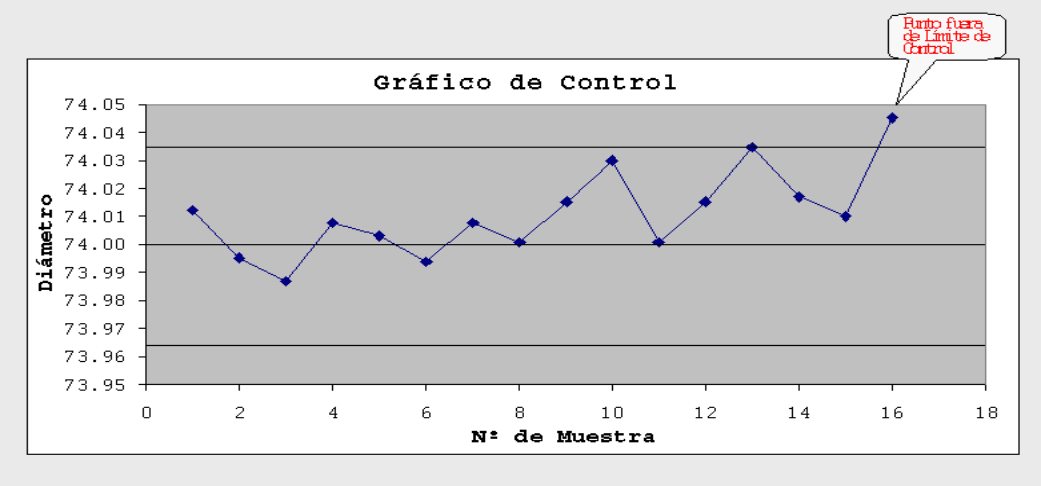


Figura 3. Diagrama de control. (González González & Jimeno Bernal, 2012)

7.2.1.5. Histograma de Frecuencias

Es una representación gráfica para datos cuantitativos, se elabora con datos anticipadamente resumidos con una distribución de frecuencia absoluta, relativa o porcentual. Se construye colocando la variable de interés en el eje horizontal y la frecuencia en el eje vertical. El valor de cada clase se indica dibujando una barra o rectángulo desde la base con valor de cero a una altura igual a la frecuencia correspondiente (Anderson, Sweeney & Williams, 2008).

Se puede evidenciar comportamientos, observar el grado de homogeneidad, acuerdo o concisión entre los valores de todas las partes que componen la población o la muestra, o, en contraposición, poder observar el grado de variabilidad, y por ende, la dispersión de todos los valores que toman las partes, también es posible no evidenciar ninguna tendencia y obtener que cada miembro de la población toma por su lado y adquiere un valor de la característica aleatoriamente sin mostrar ninguna preferencia o tendencia (Garza, 2003, p. 155).

7.2.1.6. Diagrama de Pareto

El diagrama de Pareto se parece a un histograma en el que se ordenan cada una de las clases o columnas por orden de mayor a menor frecuencia. Algunas ocasiones sobre este diagrama se incorpora un diagrama de frecuencias acumuladas.

El beneficio de utilizar el diagrama es para poder establecer un orden de prioridades en la toma de decisiones dentro de una organización. (Garza, 2003, p. 160).

En la figura 4, obtenida de Lind, Marchal, & Wathen (2015) es una representación del consumo de agua de manera genérica, al ver la imagen las actividades de riego, baño personal y albercas representan 82.1% del consumo de agua; se puede pensar en optimizar el consumo de agua en estas tres áreas.

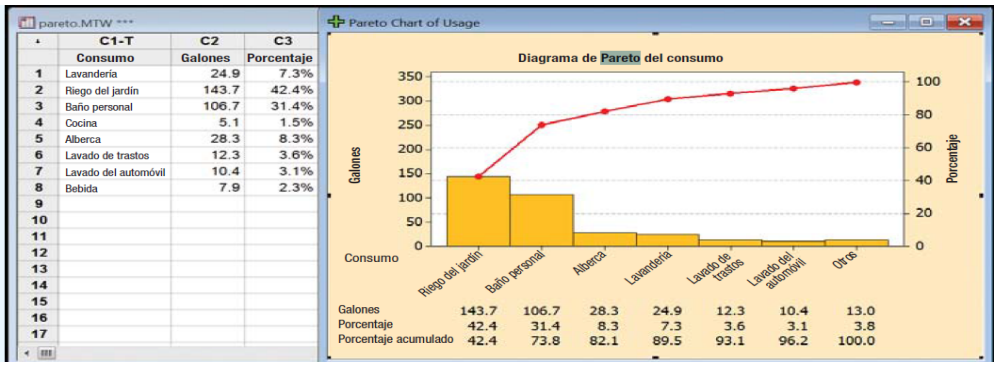


Figura 4. Diagrama de Pareto. (Lind, Marchal, & Wathen, 2015)

7.2.1.7. Diagrama de dispersión

Una herramienta gráfica que permite mostrar la relación entre variables es el diagrama de dispersión**.** Para realizar un diagrama de dispersión se necesitan dos variables, una variable independiente X y una dependiente Y. Una de las variables se escala sobre el eje horizontal (eje *X*) de una gráfica y la otra variable, a lo largo del eje vertical (eje *Y*). Por lo general, una de las variables depende hasta cierto grado de la otra (Lind, Marchal, & Wathen, 2015).

En la figura 5, se muestra una imagen genérica de un diagrama de dispersión en donde en el eje de las X está la variable edad y en el eje de las Y está la ganancia, la imagen representa que a mayor edad la gente tiene más dinero para comprar vehículos más caros de tal forma que la ganancia para el vendedor es mayor.

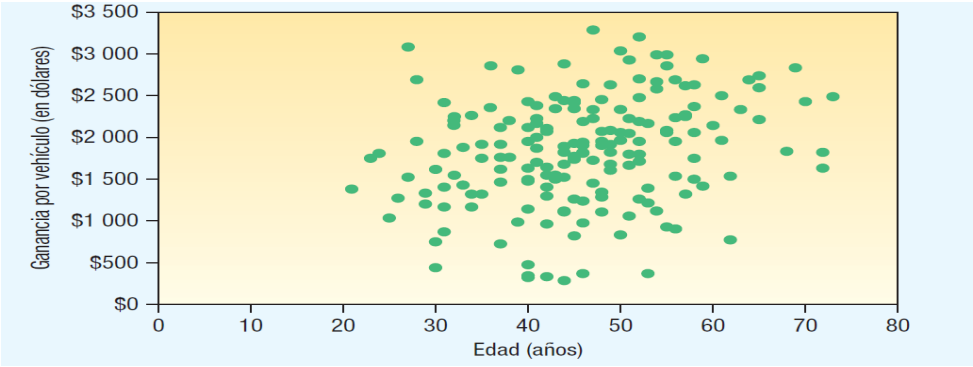


Figura 5. Diagrama de dispersión. (Lind, Marchal, & Wathen, 2015)

Existen otras herramientas que ayudan a procesos de calidad y que de igual forma están dentro de la disciplina de la estadística. Se citan algunas de estas

herramientas: análisis de varianza, correlación lineal, regresión, diagrama de árboles, entre otros.

7.2.1.8. Análisis de varianza

En estadística, el ANOVA ó Análisis de la Varianza, es una colección de modelos estadísticos. Esta sirve como la distribución del estadístico de prueba en varias situaciones; con ella se pone a prueba si dos muestras provienen de poblaciones que tienen varianzas iguales, y también se aplica cuando se desea comparar varias medias poblacionales en forma simultánea (Lind, Marchal, & Wathen, 2015).

En ambas situaciones, las poblaciones deben seguir una distribución normal, y los datos deben ser al menos de escala de intervalos (Spiegel, Schiller, & Srinivasan, 2007, p. 336).

7.2.1.9. Correlación lineal

En probabilidad y estadística, la correlación indica la fuerza y la dirección de una relación lineal y proporcionalidad entre dos atributos estadísticos. Se determina que las dos variables están correlacionadas cuando los valores de alguna de ellas, se comporta de manera homogénea y sistemáticamente igual con respecto a los valores de la otra, existe correlación entre ellas si al disminuir los valores de una lo hacen también los de la otra variable y viceversa.

La correlación como un complemento a la interpretación del diagrama de dispersión identifica un coeficiente de relación, la cual es un estadístico cuantitativo de la fuerza de la relación entre dos variables en escala de intervalo o de razón (Lind, Marchal, & Wathen, 2015).

7.2.1.10. Regresión lineal

La regresión lineal simple se basa en analizar como los cambios en una variable no aleatoria, inciden sobre otra variable aleatoria, en el caso de existir una relación funcional entre ambas variables que puede ser establecida por una expresión lineal representada por una línea recta se obtiene una regresión lineal simple (Alicia Vila, 2018).

La regresión permite y se utiliza para establecer predicciones, estimaciones y pronósticos con la historia de los datos; a la variable que se va a predecir se le llama **variable dependiente** y a la variable o variables que se usan para predecir el valor de la variable dependiente se les llama **variables independientes o predictoras**. (Mendenhall, Beaver, & Beaver, 2010).

7.2.2. La empresa Aptiv

Se menciona la empresa Aptiv dado que la propuesta se encamina a promover eficiencia en los procesos que tiene que ver con un *MRP Controller* que ahí se utiliza.

La Empresa: Aptiv Services SA de CV es una organización del giro de Maquiladora Automotriz (Export) y su actividad principal es el desarrollo y manufactura de autopartes, es decir, para el sector automotriz. Es una empresa de alta tecnología que integra soluciones más seguras, más verdes y más conectadas para el sector de la automoción.

Aptiv Services (antes Delphi) es una empresa como lo menciona Deplhi (2018) tiene la capacidad para asegurar una nueva y exitosa empresa para todos está fundamentada en la probada trayectoria de constante innovación además de tener un compromiso con el planeta, con la responsabilidad de fomentar un negocio fuerte y sostenible que haga del mundo un lugar mejor (Delphi, 2018).

7.2.3. Metodologia Six Sigma

Six Sigma (6σ) es un conjunto modelo de calidad que apoya a los procesos en la búsqueda de los cero defectos.

Six Sigma es una filosofía con base estadística compuesta por varias etapas: definir, medir, analizar, mejorar y controlar. Identifica el número el número de desviaciones estándar determinadas al resultado de un proceso. Su objetivo es promover la calidad en los mismos (Navarro Albert, 2017).

En los procesos industriales se presenta el costo de baja calidad, ocasionado por: fallas internas, de los productos defectuosos; retrabajo y problemas en el control de materiales; fallas externas, de productos regresados; garantías y penalizaciones; evaluaciones del producto, debido a inspección del proceso y producto; utilización, mantenimiento y calibración de equipos de medición de los procesos y productos; auditorias de calidad y soporte de laboratorios; y prevención de fallas, debido al diseño del producto, pruebas de campo, capacitación a trabajadores y mejora de la calidad.

La aplicación del análisis Six Sigma en los procesos industriales permite que se puedan detectar problemas en el proceso de producción tales como: cuellos de botella, productos con algún defecto, pérdidas de tiempo, etapas críticas, entre otras.

La metodología Six Sigma, utiliza herramientas estadísticas que se citaron en puntos anteriores: diagrama de flujo o de procesos, diagrama de causa-efecto, diagrama de Pareto, histograma, gráfica de corrida, gráfica de control, diagrama de dispersión y el modelo de regresión, entre otros.

7.2.4. Lenguaje de programación R

Es un freeware (software de libre licencia), es un lenguaje y entorno para computación estadística y gráficos, es un proyecto GNU (R-project.org, 2019).

R proporciona una amplia variedad de técnicas estadísticas (modelos lineales y no lineales, pruebas estadísticas clásicas, análisis de series temporales, clasificación, agrupamiento, ente otros) además de potente para la generación y visualización de gráficas y altamente extensible (R-project.org, 2019).

Una de las fortalezas de R es la facilidad con la que se pueden producir y visualizar análisis de datos bien diseñados con calidad de publicación, incluidos símbolos y fórmulas matemáticas (R-project.org, 2019).

Algunas características genéricas de R son: freeware, liviano, manipulación de datos, automatización de código, cálculos potentes, y precisos, versatilidad, organización de proyectos, tratamiento de grandes volúmenes de datos,

replicabilidad, detección y corrección de errores, estadística en su máximo nivel, despliegue de gráficos y multiplataforma, entre otros.

7.2.5. R Studio

RStudio es un entorno de desarrollo integrado (IDE) para R. Incluye una consola, editor de resaltado de sintaxis que admite la ejecución directa de código, así como herramientas para el trazado, el historial, la depuración y la gestión del espacio de trabajo (Studio, 2019).

R Studio tiene características que es importante resaltar: integración, ejecución de código, resaltado de la sintaxis, ayuda, completado de comandos, atajos de teclado, navegador de objetos, gestión del historial de comandos, navegación del código, visualización e importación de datos, integración de gráficos, gestor de proyectos, control de versiones, generación de documentos, entre otros.

7.2.6. Ciencia de los Datos

A diario en el mundo se generan una gran cantidad de datos e información, su fuente es muy diversa: sitios WEB, sensores que reciben información climática y de otro tipo, de las redes sociales, imágenes y vídeos digitales, registros de compra y transacciones y señales de GPS de los móviles, entre otros. Toda esto se integra en el paradigma Big Data y es en donde se produce el inminente nacimiento de la disciplina que conozca y trate con esta información, la Ciencia de los Datos.

Debido a las grandes cantidades de información, la Ciencia de los Datos requiere de programas robustos para el manejo y procesamiento de datos, además de la capacidad analítica y visionaria de quien estructura la información para facilitar la toma las decisiones.

El reto actual consiste en la posibilidad de analizar científicamente toda esta información, empleando los conocimientos y herramientas existentes en los campos de la estadística, inteligencia artificial, matemáticas, bases de datos y programación, negocios y comunicación, entre otros.

En el presente trabajo se expone cómo el empleo de la Ciencia de los Datos, con R y R Studio como herramienta como propuesta de trabajo en la empresa Aptiv.

7.3. Desarrollo

Se presenta una propuesta del uso de R aplicado a datos relacionados con el trabajo de un MRP Controller.

7.3.1. MRP Controller

Un *MRP Controller* y quien lo maneja, es un sistema que requiere un análisis detallado que permita determinar acciones preventivas y/o correctivas en los procesos, además de identificar situaciones que influyan de forma determinante en el esquema de trabajo que el puesto requiere.

Las funciones de un *MRP Controller* están relacionadas con la administración de los recursos de transportación; validar requerimientos semanales para soportar necesidades del cliente; identificación y manejo de riesgos de impacto a líneas; control de excesos y obsoletos; cumplir y controlar niveles de inventario de materia prima; apoyo al cumplimiento para clarificar el camino (*glade path)*, cumplimiento del objetivo; control interno.

Algunas habilidades requeridas para el puesto de *MRP Controller*: habilidades de comunicación hablada y escrita, de negociación, matemáticas y habilidad para manejo de grandes cantidades de datos, actitud de servicio, administración del tiempo o disponibilidad de horario, administración de recursos, toma de decisiones, trabajo bajo presión y proactivo, entre otras.

7.3.2. La propuesta

Toda vez identificada la importancia de que un MRP Controller utiliza gran cantidad de datos y variables, se propone que mediante el entorno R Studio y el lenguaje de programación R, se pueden manipular y utilizar las herramientas de calidad.

Las herramientas de calidad para un MRP Controller, permiten analizar datos como resultado de la ejecución de procesos normales y procesos que presentan mayor conflicto, facilitan la visión global de oportunidades y problemas.

Algunos de los análisis que un MRP Controller pudiera realizar con las herramientas mencionadas y a través de R y R Studio, serían las que se enlistan a continuación:

- **Aplicación del análisis de Pareto en la variable Defecto.** A través de este análisis se podrá identificar el veinte por ciento de los elementos que generan el ochenta por ciento de los números de partes defectuosas, y/o con problemas de inventarios (excesos o cortos), para de esta forma identificarlos y tener retroalimentación con los proveedores para evitar material que impacte negativamente en la construcción de arneses.
- **Análisis de varianza en las variables MRP y StdPack.** Con la herramienta mencionada es posible conocer si existen variaciones significativas entre la cantidad de componentes destinados y el lugar de origen de los mismos. El parámetro F de Fisher-Snedecor determinará a través de funciones como la cantidad de grados de libertad y la significatividad asintótica si debe rechazarse o aceptarse la hipótesis nula de no variación en los promedios.
- **Análisis de regresión y correlación.** Permite estimar si la cantidad de inventario disponible al día actual coincide con el inventario existente en almacén. La pendiente de la ecuación de regresión estimará el crecimiento o decrecimiento de la variable independiente sobre la dependiente y el coeficiente de determinación mostrará el nivel de asociación entre ambas variables.
- **Histograma.** Permite estratificar la cantidad de material perdido en almacén considerando elementos tales como la regla de Sturges para determinar la cantidad de intervalos del histograma.
- **Diagrama de Ishikawa en variables Defectos.** ¿Cuáles son las características que determinan los defectos en los componentes analizados? ¿material? ¿mano de obra? ¿el método de ensamble, no es el adecuado? Con esta herramienta se podrá desglosar a detalle las características que incidan en las fallas.

Esta propuesta establece que R como herramienta de análisis estadístico es conveniente dado que se adapta para ofrecer una solución óptima, así la propuesta ofrece flexibilidad de que los códigos generados en R son adaptables a otros

procesos y dependerá de las necesidades específicas de cada *MRP Controller* y los procesos que este desee mejorar con esta propuesta.

Es importante puntualizar que cada *MRP Controller* maneja materiales, proveedores (locales, nacionales, internacionales continentales e internacionales intercontinentales) clientes, procesos, stock, días de inventario, procesos de calidad, muy diferentes entre sí, a su vez que todas y cada una de las necesidades y adaptaciones de cada *MRP Controller* serán de utilidad para el *MRP Manager* o *Procurement Supervisor*.

Un *Procurement Supervisor* es el encargado de presentar un concentrado cada determinado tiempo de la situación del departamento de *Ordering* y cuando se requiere una escalación para la resolución de algún inconveniente que está fuera del nivel del *MRP Controller*, de lo anterior, por mencionar algún ejemplo, hay un *MRP* que maneja exclusivamente cable, otro que maneja proveedores Internacionales americanos, uno más encargado de aquellos proveedores que son europeos y el traslado de la materia prima es marítimo así como uno que maneja proveedores mexicanos únicamente.

Esto es la razón por la cual, un proceso estándar para el manejo de todas las hojas de trabajo de cada *MRP Controller* no ofrece solución a las necesidades específicas de cada proveedor, pues no es igual el stock que hay que tener para un proveedor cuyo tránsito es de 4 semanas , al que se debe tener con un proveedor que tiene un tiempo de tránsito de 15 minutos, las variables y las variantes son muchas y ello desemboca en la necesidad de crear una propuesta que optimice y se adapte a las necesidades individuales de la Hoja de trabajo.

Se presentan en la propuesta, algunos códigos en lenguaje R para ciertas herramientas citadas con datos extraídos de la literatura relacionada con R y estadística. Se presentan todos excepto el histograma dada la facilidad de generar el mismo con la instrucción *hist()* de R.

Los ejemplos que se muestran en los siguientes puntos, son prácticas desarrolladas de forma *R NoteBook y MarkDown* con datos diferentes a los

procesos de producción de la empresa citada, sin embargo, pueden servir de referencia, con lo cual, perfectamente se puede recrear en el lenguaje de programación R y su entorno R Studio.

Los documentos *NoteBook* y *Markdown* son archivos que contienen combinaciones de código y resultados de datos de manera tabular y gráfica; para los siguientes ejemplos los archivos *NoteBook* y *Markdown* contienen instrucciones R y su ejecución.

Primero se muestra un ejemplo de análisis de Pareto de productos elaborados con algunos defectos; segundo, se presenta el ejemplo de análisis ANOVA de una variable “one way” con el caso de personas sin y con el consumo de alcohol y el grado de atracción o selectividad hacia otra persona; tercero, se presenta un ejemplo en R de análisis de regresión con el caso de *auto-mpg.csv*; cuarto se presenta caso de diagrama de Ishikawa con datos de causa y efecto.

En las prácticas que se presentan en formato *NoteBook,* se sigue una secuencia*,* los códigos se muestran sombreados, se muestran las gráficas realizadas en R propias del código y los resultados se observan tal y como se generan en R Studio.

7.3.2.1. Análisis de Pareto

Se cita el ejemplo desarrollado en *markdown* del siguiente enlace: https://rpubs.com/rpizarro/624226, se muestra su desarrollo a manera de referencia en el siguiente R Notebook:

R Notebook

7.3.2.1.1. Las librerías

Se utiliza la librería qcc para diagramas de pareto.

```
library(knitr) # Para ver tablas más amigables en formato html markdown)
library(qcc) # Para diagrama de Pareto
```

La siguiente práctica en R Booknote genera dos diagramas de Pareto con dos conjuntos de datos diferentes. Datos y datos2.

7.3.2.1.2. Los datos

Simulando un conjunto de desperfectos de una línea de producción de costura.

- Se simula que se tienen datos resumidos de defecto de una línea de producción.
- Se generan dos vectores defectos y cantidades
- Se integra en un data frame a partir de los dos vectores.

```
defectos <- c("Piel Arrugada", "Costura con fallas ", "Reventado de piel"
, "Mal montada")
cantidades <-c( 99, 135, 369, 135 )
datos <- data.frame(defectos, cantidades)
kable(datos)
```

defectos	cantidades
Piel Arrugada	99
Costura con fallas	135
Reventado de piel	369
Mal montada	135

7.3.2.1.3. Generando el diagrama de Pareto

A partir de la variable cantidades del conjunto de datos. Se muestran en las figuras 6 y 7 los diagramas de pareto.

```
pareto.chart(datos$cantidades)
```

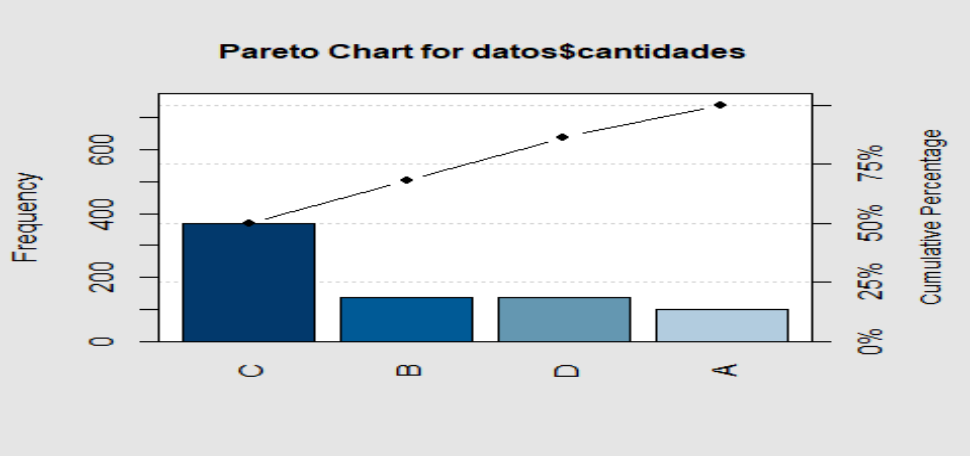


Figura 6. Diagrama de Pareto frecuencia cantidades. (Pizarro, Diagrama de Pareto, 2020)

```
##
## Pareto chart analysis for datos$cantidades
##     Frequency Cum.Freq. Percentage Cum.Percent.
##   C 369.00000 369.00000   50.00000     50.00000
##   B 135.00000 504.00000   18.29268     68.29268
##   D 135.00000 639.00000   18.29268     86.58537
##   A  99.00000 738.00000   13.41463    100.00000
```

7.3.2.1.4. Generando datos aleatorios

De un conjunto de 100 datos simulados llamados muestras generados de manera aleatoria, se simula que hay una cantidad de defectos 1, 2 o 3 defectos por cada muestra.

Se visualizan los primeros y últimos seis registros de los 100 registros

```
muestras <- 1:100
cant.defectos <- sample(1:3, 100, replace = TRUE)

datos2 <- data.frame(muestras, cant.defectos)



kable(head(datos2))
```

muestras	cant.defectos
1	1
2	2
3	2
4	1
5	2
6	3

```
kable(tail(datos2))
```

	muestras	cant.defectos
95	95	3
96	96	1
97	97	3
98	98	1
99	99	3
100	100	2

7.3.2.1.5. Generar frecuencia de los datos

```
tabla.defectos <- table(datos2$cant.defectos)

tabla.defectos

##
##  1  2  3
## 34 31 35
```

7.3.2.1.6. Generar diagrama de Pareto para datos

```
pareto.chart(tabla.defectos)
```

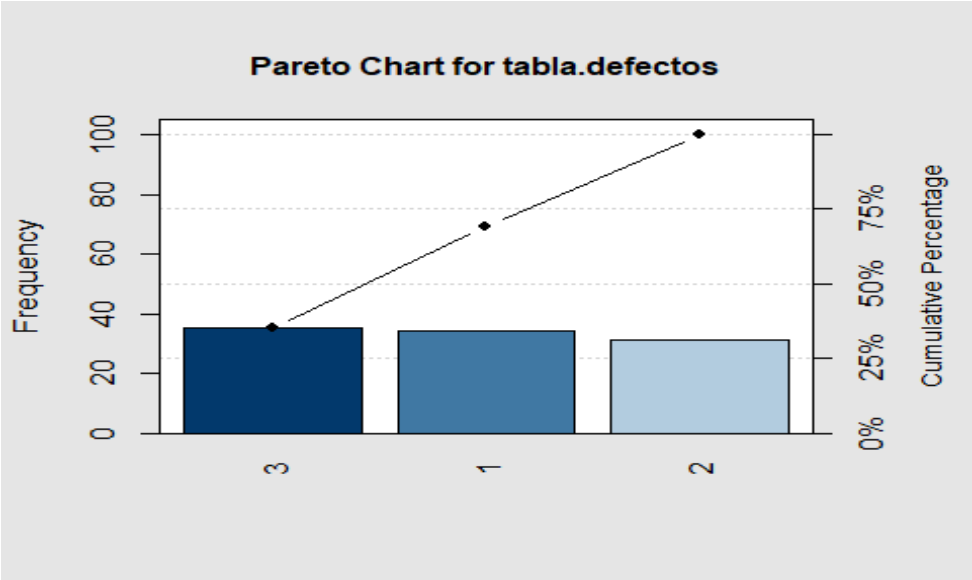


Figura 7. Diagrama de Pareto frecuencia defectos. (Pizarro, Diagrama de Pareto, 2020)

```
##
## Pareto chart analysis for tabla.defectos
##      Frequency Cum.Freq. Percentage Cum.Percent.
##   3         35        35         35           35
##   1         34        69         34           69
##   2         31       100         31          100
```

7.3.2.2. Análisis de varianza (one way)

Se cita el ejemplo desarrollado en *markdown* del siguiente enlace: https://rpubs.com/rpizarro/624230, se muestra su desarrollo a manera de referencia en el siguiente R Notebook:

R Notebook

Aplicar ANOVA de una vía sobre un conjunto de datos relacionados con la atracción que tienen las personas hacia otras personas cuando SI consumen alcohol o NO consumen.

7.3.2.2.1. Las librerías

La librería WRS2 permite tener acceso a los datos (*data(goggles)*) que se utilizan en la práctica.

```
# install.packages("WRS2")
# install.packages("psych")
# install.packages("car")
library(WRS2) # Para disponer de los datos goggles; data(goggles)
library(psych) # Para variables descriptivas
library(car) # Para prueba de normalidad y homocestacidad
```

7.3.2.2.2. Los datos

- Los datos corresponden a 48 participantes (24 hombres y 24 mujeres) que se han dividido en 3 grupos de 8 participantes cada uno.
- Cada grupo asistió a un club nocturno, a un grupo no se le dio alcohol, otro tomó 2 bebidas y el último 4 bebidas de alcohol.
- Al final de la noche el investigador tomó una fotografía a la pareja elegida por el participante y un grupo de jueces independientes evaluó el poder de atracción de dicha persona.
- La función *data(goggles)* accede directamente a los datos.
- Se muestran los primeros seis y últimos seis datos con la función *head()* y *tail()* respectivamente.
- Se muestran también 10 registros aleatorios con reemplazo utilizando la función *sample().*

```
data(goggles)
datos <- goggles
head(datos)
```

```
##   gender alcohol attractiveness
## 1 Female    None             65
## 2 Female    None             70
## 3 Female    None             60
## 4 Female    None             60
```

```
## 5 Female    None             60
## 6 Female    None             55

tail(datos)

##    gender alcohol attractiveness
## 43   Male 4 Pints             30
## 44   Male 4 Pints             55
## 45   Male 4 Pints             35
## 46   Male 4 Pints             20
## 47   Male 4 Pints             45
## 48   Male 4 Pints             40

datos[sample(1:nrow(datos), 10, replace = TRUE),]

##    gender alcohol attractiveness
## 14 Female 2 Pints             60
## 27   Male    None             80
## 6  Female    None             55
## 44   Male 4 Pints             55
## 23 Female 4 Pints             50
## 2  Female    None             70
## 30   Male    None             75
## 25   Male    None             50
## 24 Female 4 Pints             50
## 16 Female 2 Pints             50
```

La base de datos tiene 3 variables:

- el sexo (variable *gender*: hombre o mujer),
- el alcohol consumido (variable alcohol: nada, 2 bebidas o 4 bebidas), Para los datos la variable viene con valor de Pintas que significa que equivale a una bebida de, aproximadamente, 568 mililitros para los Ingleses. La pinta americana equivale a unos 473 mililitros. A su vez, también se denominan pintas a los vasos que se utilizan para beber este tipo de cervezas.
- el nivel de atracción física de la pareja encontrada (variable *attractiveness*: puntaje de 0 a 100 dado por los jueces).

7.3.2.2.3. Explorando datos

- Hay 48 observaciones y 3 variables.
- gender es el género y es de tipo factor con etiquetas de ‘Female’ de Femenino y ‘Male’ de Masculino.

- alcohol de tipo factor, ‘None’ no consumió alcohol, ‘2 Pints’ consumió 2 bebidas y ‘4 Pints’ consumió 4 bebidas
- attractiveness de tipo numérico entero con valores entre 0 y 100 significando el puntaje dado por los jueces.

```
str(datos)
```

```
## 'data.frame':    48 obs. of  3 variables:
##  $ gender        : Factor w/ 2 levels "Female","Male": 1 1 1 1 1 1 1 1
1 1 ...
##  $ alcohol       : Factor w/ 3 levels "None","2 Pints",..: 1 1 1 1 1 1
1 1 2 2 ...
##  $ attractiveness: int  65 70 60 60 60 55 60 55 70 65 ...
```

```
summary(datos)
```

```
##     gender      alcohol   attractiveness
##  Female:24   None   :16   Min.   :20.00
##  Male  :24   2 Pints:16   1st Qu.:53.75
##              4 Pints:16   Median :60.00
##                           Mean   :58.33
##                           3rd Qu.:66.25
##                           Max.   :85.00
```

7.3.2.2.4. Estadísticos descriptivos por grupo utilizando *describeBy*()

Se calcula los estadísticos descriptivos por grupo utilizando la función *describeBy()* de la librería *psych*. Los principales estadísticos que genera la función *describeBy()* son los siguientes:

- item name. Nombre del grupo, X1, X2, X3…
- item number. Número de variables del grupo, la salida es 1
- number of valid cases. Número de casos válidos, número de observaciones ‘n’ del grupo
- mean. La media del grupo
- standard deviation. La desviación estándar
- median. La mediana
- mad: median absolute deviation (from the median). La desviación media absoluta de la mediana
- minimum. El valor mínimo
- maximum. El valor máximo

- *skew*. La curtosis
- *standard error*. El error estándar

```
describeBy(datos$attractiveness, datos$alcohol)

##
##  Descriptive statistics by group
## group: None
##    vars  n  mean   sd median trimmed   mad min max range skew kurtosis
se
## X1    1 16 63.75 8.47   62.5   63.57 11.12  50  80    30 0.29    -1.07
2.12
## ------------------------------------------------------------
## group: 2 Pints
##    vars  n  mean   sd median trimmed  mad min max range skew kurtosis
se
## X1    1 16 64.69 9.91     65   64.64 7.41  45  85    40 0.08    -0.23
2.48
## ------------------------------------------------------------
## group: 4 Pints
##    vars  n  mean    sd median trimmed   mad min max range  skew kurtos
is   se
## X1    1 16 46.56 14.34     50   46.79 14.83  20  70    50 -0.22    -1.
21 3.59
```

Utilizando summary(por grupo)

Para cada grupo se puedo haber utilizado *summary(),* con resultados iguales pero con menores estadísticos que la función *describeBy()*

```
summary (datos[datos$alcohol=="None",])

##     gender     alcohol   attractiveness
##  Female:8   None   :16   Min.   :50.00
##  Male  :8   2 Pints: 0   1st Qu.:58.75
##             4 Pints: 0   Median :62.50
##                          Mean   :63.75
##                          3rd Qu.:70.00
##                          Max.   :80.00

summary (datos[datos$alcohol=="2 Pints",])

##     gender     alcohol   attractiveness
##  Female:8   None   : 0   Min.   :45.00
##  Male  :8   2 Pints:16   1st Qu.:60.00
##             4 Pints: 0   Median :65.00
##                          Mean   :64.69
```

```
##                          3rd Qu.:70.00
##                          Max.   :85.00
```

```
summary (datos[datos$alcohol=="4 Pints",])
```

```
##     gender     alcohol   attractiveness
##  Female:8   None   : 0   Min.   :20.00
##  Male  :8   2 Pints: 0   1st Qu.:33.75
##             4 Pints:16   Median :50.00
##                          Mean   :46.56
##                          3rd Qu.:55.00
##                          Max.   :70.00
```

Los dos primeros grupos presentaron valores similares en el atractivo de las parejas que encontraron, sin embargo, estos valores disminuyeron en el grupo que consumió más alcohol:

- No consume alcohol, 63.75±8.47
- 2 bebidas de alcohol, 64.69±9.91
- 4 bebidas de alcohol, 46.56±14.34

Los valores se presentan como media±sd (desviación estándar).

7.3.2.2.5. Visualizando datos con boxplot()

La función *boxplot()* permite visualizar e identificar datos atípicos, para este ejemplo no existe datos extraños. El caracter tilde o virulilla ´~´ significa en esta función que se generan los diagramas de cajas de la variable numérica *attractiveness* en función de la variable factor llamada alcohol que significa cada grupo. La figura 8 muestra el diagrama de caja.

```
boxplot(datos$attractiveness~datos$alcohol, xlab = "Consumo", ylab = "Sel
ectividad")
```

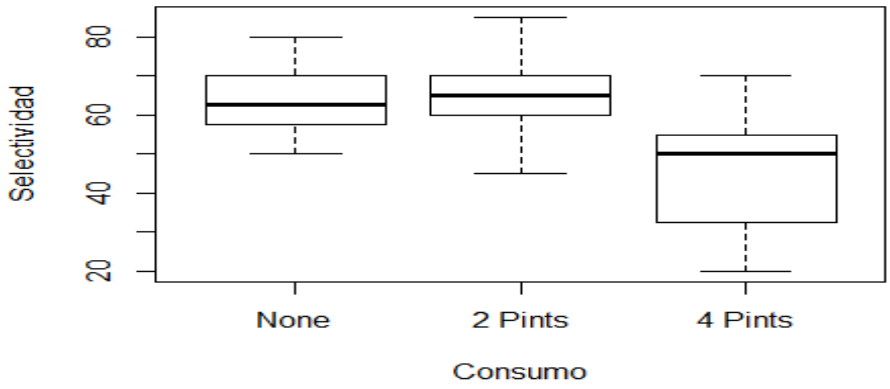


Figura 8. Gráfica de cajas consumo de alcohol. (Pizarro, ANOVA una via. Caso consumo alcohol, 2020).

7.3.2.2.6. Prueba de Normalidad

Para este conjunto de datos, hay dos tipos de variables: variables independientes: Gender y alcohol y variable dependiente: attractiveness

Se utiliza la prueba de normalidad de Shapiro-Wilks que funciona bien para conjuntos de datos pequeños.

Para esta prueba la hipótesis nula implica que los datos siguen una distribución normal, y la hipótesis alternativa indica lo contrario.

Por tanto, si el *p-valor* de la prueba es inferior a *0.05* (el nivel alfa de significación que se toma por defecto) se rechaza la hipótesis nulas y se dice que la respuesta no sigue una distribución normal en cada grupo de estudio.

En caso contrario, si el *p-valor* es mayor a *0.05*, no se estaría incumpliendo el supuesto de normalidad.

```
by(datos$attractiveness, datos$alcohol, shapiro.test)
```

```
## datos$alcohol: None
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.95498, p-value = 0.5725
##
```

```
## ------------------------------------------------------------
## datos$alcohol: 2 Pints
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.94489, p-value = 0.4132
##
## ------------------------------------------------------------
## datos$alcohol: 4 Pints
##
##  Shapiro-Wilk normality test
##
## data:  dd[x, ]
## W = 0.952, p-value = 0.522
```

De lo anterior, en los 3 grupos la variable *attractiveness* tiene distribución normal, aclarando que ¡solo se tienen 16 observaciones por grupo!

- Primer grupo: p-value = 0.5725
- Segundo grupo 2 bebidad: p-value = 0.4132
- Gercer grupo 4 bebidas: p-value = 0.522

7.3.2.2.7. Homogeneidad de varianza

Se utilizan distintos tipos de pruebas para evaluar la homocedasticidad (homogeneidad de varianza), es decir, para contrastar si varios grupos o muestras de la misma población son homocedásticas (tiene la misma varianza).

Las diferencias entre las pruebas se deben a su sensibilidad al supuesto de normalidad visto anteriormente.

Para esta prueba la hipótesis nula implica que los datos presentan homogeneidad de varianza entre los grupos, por lo cual si el *p-valor* es inferior a *0.05* se estaría incumpliendo este supuesto.

Si *p-valor* es mayor a *0.05* existe homogeneidad de

- Prueba de Bartlett: *p-value = 0.1092*

```
bartlett.test(datos$attractiveness, datos$alcohol)

##
##  Bartlett test of homogeneity of variances
```

```
##
## data:  datos$attractiveness and datos$alcohol
## Bartlett's K-squared = 4.4295, df = 2, p-value = 0.1092
```

- Prueba de Levene: *Pr(>F) = 0.1095*

```
leveneTest(datos$attractiveness, datos$alcohol)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##       Df F value Pr(>F)
## group  2  2.3238 0.1095
##       45
```

- Prueba de Flinger: p-value = 0.1115

```
fligner.test(datos$attractiveness, datos$alcohol)
```

```
##
##  Fligner-Killeen test of homogeneity of variances
##
## data:  datos$attractiveness and datos$alcohol
## Fligner-Killeen:med chi-squared = 4.3876, df = 2, p-value = 0.1115
```

De lo anterior, en todos los casos se obtiene *p>0.05*, es decir, no se encuentran problemas de heterocedasticidad (falta de homogeneidad de varianza).

Las pruebas de normalidad y homogeneidad de varianza no indican problemas con los supuestos estadísticos clásicos.

7.3.2.2.8. Prueba ANOVA

El procedimiento ANOVA de un factor genera un análisis de varianza de un factor para una variable dependiente cuantitativa respecto a una única variable de factor (la variable independiente). El análisis de varianza se utiliza para contrastar la hipótesis de que varias medias son iguales.

ANOVA es una prueba estadística para analizar si más de dos grupos difieren entre sí de manera significativa en sus medias y varianzas (Hernández, Fernández, & Baptista, 2014).

¿Existen diferencias entre los grupos de personas conforme y de acuerdo a la variable consumo de alcohol y el nivel de selectividad en cuanto a la atracción hacia otras personas valorado en la variable attractiveness?

¿al consumir alcohol los sujetos se vuelven menos selectivos a la hora de elegir pareja?

¿en qué momento se vuelven más selectivos?

¿con 2 bebidas o con 4 bebidas de alcohol?

Para el caso del ANOVA de un factor entre grupos, las hipótesis que corresponden son:

- **H0**: Para los grupos de personas que no consumen alcohol, para el grupo que consume 2 bebidas y para el grupo que consume 4 bebidas; el atractivo físico de las parejas de sujetos con distinto consumo de alcohol es similar (los sujetos, independientemente del nivel de consumo de alcohol que tengan encima, son igual de selectivos a la hora de encontrar pareja).
- **H1**: alguno es distinto (existen diferencias entre al menos alguno de los 3 grupos de consumo de alcohol, alguno de ellos es más selectivo pero no se indica cuál, hasta las pruebas post hoc)
- Se utiliza el caracter virulilla o tilde '~' para indicar que se hace la prueba ANOVA de la variable *attractiveness* en función de la variable alcohol (representado como fórmula en R attractiveness ~ alcohol).
- Se utiliza la variable llamada análisis (analisis sin acento) para determinar los resultados de la prueba.

```
analisis <- aov(data = datos, attractiveness ~ alcohol)

summary(analisis)
```

```
##             Df Sum Sq Mean Sq F value   Pr(>F)
## alcohol      2   3332  1666.1   13.31 2.88e-05 ***
## Residuals   45   5634   125.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Existen diferencias estadísticamente significativas entre los grupos de consumo de alcohol ($F(2,45)=13.31$, $p<0.05$). *Pr > F = 2.88e-05 (tiene tres asteriscos)*

7.3.2.2.9. Pruebas post hoc

Como se ha detectado diferencias entre los grupos, la pregunta específica es ¿en qué grupos son significativamente distintos?

Para ello se utiliza la prueba de comparación múltiple post hoc de Tukey (también llamada prueba HSD).

La prueba de Tukey es una prueba estadística para comparaciones posteriores (post hoc) en el ANOVA unidireccional o de un factor. (Hernández, Fernández, & Baptista, 2014)

```
TukeyHSD(analisis)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = attractiveness ~ alcohol, data = datos)
##
## $alcohol
##                       diff        lwr       upr     p adj
## 2 Pints-None        0.9375  -8.650654 10.525654 0.9695381
## 4 Pints-None      -17.1875 -26.775654 -7.599346 0.0002283
## 4 Pints-2 Pints   -18.1250 -27.713154 -8.536846 0.0001067
```

Se identifica que existen diferencias entre 4 Pints-None y 4 Pints-2 Pints ($p<0.001$), pero no entre 2 Pints-None.

Observando los gráficos y estadísticos descriptivos que han calculado anteriormente, se puede afirmar que solo para 4 bebidas (pintas) de alcohol disminuye estadística y significativamente el nivel de atracción (selección) de las parejas.

Dicho de otro modo, se detectan diferencias significativas en el atractivo de las parejas de sujetos con distinto consumo de alcohol, en particular se observa que los sujetos se volvieron menos selectivos en la búsqueda de pareja a partir de 4 bebidas de alcohol.

Moraleja, *un poco de alcohol nos hace menos selectivos en la búsqueda de pareja pero ¿si se abusa del alcohol?* … … …

7.3.2.3. Análisis de regresión y correlación

Se presenta el caso de correlación y regresión lineal de los datos de autos que se encuentra en el archivo markdown del servicio rpubs del enlace siguiente: https://rpubs.com/rpizarro/578026. El ejemplo es un caso que se trata en temas de aprendizaje automático pero que queda muy adecuado para justificar el concepto de correlación y regresión lineal como herramienta predictora.

Regresión lineal múltiple de autos

7.3.2.3.1. Las librerías

```
library(readr)
# install.packages("corrplot") # Nuevo
library(corrplot) # Para correlación
library(caret)  # Para dividir conjunto de datos

# install.packages("MASS") # NUEVO
library(MASS)
```

7.3.2.3.2. Los datos

```
datos <- read.csv("https://raw.githubusercontent.com/rpizarrog/Curso-Titulacion-Data-Science-/master/2020/datos/auto-mpg.csv")


head(datos) # Los primeros seis
```

```
##   No mpg cylinders displacement horsepower weight acceleration model_year
## 1  1  28         4          140         90   2264         15.5         71
## 2  2  19         3           70         97   2330         13.5         72
## 3  3  36         4          107         75   2205         14.5         82
## 4  4  28         4           97         92   2288         17.0         72
## 5  5  21         6          199         90   2648         15.0         70
## 6  6  23         4          115         95   2694         15.0         75
##              car_name
## 1 chevrolet vega 2300
## 2     mazda rx2 coupe
## 3        honda accord
## 4     datsun 510 (sw)
## 5         amc gremlin
## 6          audi 100ls
```

```
tail(datos) # Los últimos seis
```

```
##      No  mpg cylinders displacement horsepower weight acceleration mod
el_year
## 393 393 13.0         8          350        155   4502         13.5
72
## 394 394 16.5         6          168        120   3820         16.7
76
## 395 395 34.5         4          105         70   2150         14.9
79
## 396 396 38.1         4           89         60   1968         18.8
80
## 397 397 30.5         4           98         63   2051         17.0
77
## 398 398 19.0         6          232        100   2634         13.0
71
##                  car_name
## 393  buick lesabre custom
## 394    mercedes-benz 280s
## 395  plymouth horizon tc3
## 396 toyota corolla tercel
## 397    chevrolet chevette
## 398           amc gremlin
```

```
summary(datos) # Antes de categorizar o factor
```

```
##        No              mpg          cylinders      displacement
##  Min.   :  1.0   Min.   : 9.00   Min.   :3.000   Min.   : 68.0
##  1st Qu.:100.2   1st Qu.:17.50   1st Qu.:4.000   1st Qu.:104.2
##  Median :199.5   Median :23.00   Median :4.000   Median :148.5
##  Mean   :199.5   Mean   :23.51   Mean   :5.455   Mean   :193.4
##  3rd Qu.:298.8   3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:262.0
##  Max.   :398.0   Max.   :46.60   Max.   :8.000   Max.   :455.0
##
##    horsepower        weight      acceleration     model_year
##  Min.   : 46.0   Min.   :1613   Min.   : 8.00   Min.   :70.00
##  1st Qu.: 76.0   1st Qu.:2224   1st Qu.:13.82   1st Qu.:73.00
##  Median : 92.0   Median :2804   Median :15.50   Median :76.00
##  Mean   :104.1   Mean   :2970   Mean   :15.57   Mean   :76.01
##  3rd Qu.:125.0   3rd Qu.:3608   3rd Qu.:17.18   3rd Qu.:79.00
##  Max.   :230.0   Max.   :5140   Max.   :24.80   Max.   :82.00
##
##            car_name
##  ford pinto    :  6
##  amc matador   :  5
##  ford maverick :  5
##  toyota corolla:  5
##  amc gremlin   :  4
##  amc hornet    :  4
##  (Other)       :369
```

7.3.2.3.3. Convertir la variable categórica de Cilindros a factor

Para conocer su frecuencia con *summary()*

```
datos$cylinders <- factor(datos$cylinders, levels = c(3,4,5,6,8),labels = c('3c', '4c','5c','6c','8c'))

summary(datos) # Después de categorizar o factor
```

```
##        No              mpg        cylinders  displacement     horsepower
##  Min.   :  1.0   Min.   : 9.00   3c:  4    Min.   : 68.0   Min.   : 46.0
##  1st Qu.:100.2   1st Qu.:17.50   4c:204    1st Qu.:104.2   1st Qu.: 76.0
##  Median :199.5   Median :23.00   5c:  3    Median :148.5   Median : 92.0
##  Mean   :199.5   Mean   :23.51   6c: 84    Mean   :193.4   Mean   :104.1
##  3rd Qu.:298.8   3rd Qu.:29.00   8c:103    3rd Qu.:262.0   3rd Qu.:125.0
##  Max.   :398.0   Max.   :46.60             Max.   :455.0   Max.   :230.0
##
##      weight      acceleration     model_year              car_name
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   ford pinto    :  6
##  1st Qu.:2224   1st Qu.:13.82   1st Qu.:73.00   amc matador   :  5
##  Median :2804   Median :15.50   Median :76.00   ford maverick :  5
##  Mean   :2970   Mean   :15.57   Mean   :76.01   toyota corolla:  5
##  3rd Qu.:3608   3rd Qu.:17.18   3rd Qu.:79.00   amc gremlin   :  4
##  Max.   :5140   Max.   :24.80   Max.   :82.00   amc hornet    :  4
##                                                 (Other)       :369
```

7.3.2.3.4. Tabla de correlación

Es una prueba estadística para analizar la relación entre dos variables medidas en un nivel por intervalos. (Hernández, Fernández, & Baptista, 2014).

- *Hipótesis a probar*: correlacional, del tipo de "a mayor *X*, mayor *Y*", "a menor *X*, menor *Y*", "altos valores en *X* están asociados con altos valores en *Y*", "bajos valores en *X* se asocian con bajos valores de *Y* ". La hipótesis de investigación señala que la correlación es significativa.
- *Variables*: dos, la prueba en sí no considera a una como independiente y a otra como dependiente, ya que no evalúa la causalidad. La noción de causa-efecto

(independiente-dependiente) es posible establecerla teóricamente, pero la prueba no asume dicha causalidad.

- El significado de las correlaciones son las siguientes:
  - −0.90 = Correlación negativa muy fuerte.
  - −0.75 = Correlación negativa considerable.
  - −0.50 = Correlación negativa media.
  - −0.25 = Correlación negativa débil.
  - −0.10 = Correlación negativa muy débil.
  - = No existe correlación alguna entre las variables.
  - +0.10 = Correlación positiva muy débil.
  - +0.25 = Correlación positiva débil.
  - +0.50 = Correlación positiva media.
  - +0.75 = Correlación positiva considerable.
  - +0.90 = Correlación positiva muy fuerte.
  - +1.00 = *Correlación positiva perfecta* ("A mayor *X*, mayor *Y*" o "a menor *X*, menor *Y*", de manera proporcional. Cada vez que *X* aumenta, *Y* aumenta siempre una cantidad constante).
  - El *signo indica la dirección de la correlación* (positiva o negativa); y *el valor numérico, la magnitud de la correlación.* (Hernández, Fernández, & Baptista, 2014).

```
cor(x=datos[,-c(1,3, 8,9)], method = "pearson")
```

```
##                     mpg displacement horsepower     weight acceleratio
n
## mpg           1.0000000   -0.8042028 -0.7756163 -0.8317409    0.420288
9
## displacement -0.8042028    1.0000000  0.8975294  0.9328241   -0.543684
1
## horsepower   -0.7756163    0.8975294  1.0000000  0.8636062   -0.688055
0
## weight       -0.8317409    0.9328241  0.8636062  1.0000000   -0.417457
3
## acceleration  0.4202889   -0.5436841 -0.6880550 -0.4174573    1.000000
0
```

7.3.2.3.5. Mostrando correlaciones con pairs()

```
pairs(x=datos[,-c(1,3,8,9)], lower.panel = NULL)
```

La figura 9 muestra las correlaciones generadas por *pairs()* y la figura 10 la misma correlación pero con la función *corroplot().* Se muestran las figuras 11 y 12 como complementos del modelo de regresión.

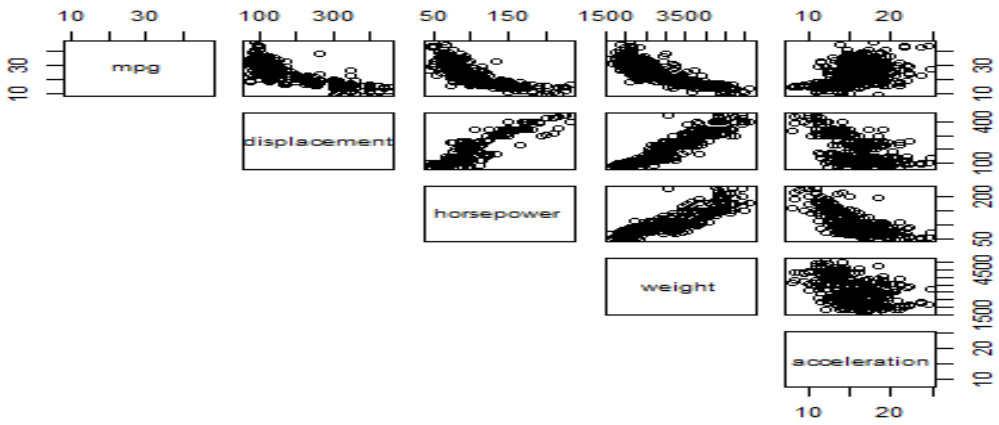


Figura 9. Correlaciones entre variables de los datos usando pairs() (Pizarro, Regresion lineal de autos, 2020).

7.3.2.3.6. Correlación con corrplot()

```
corrplot(corr = cor(x=datos[,-c(1,3,8,9)], method = "pearson"), method =
"number")
```

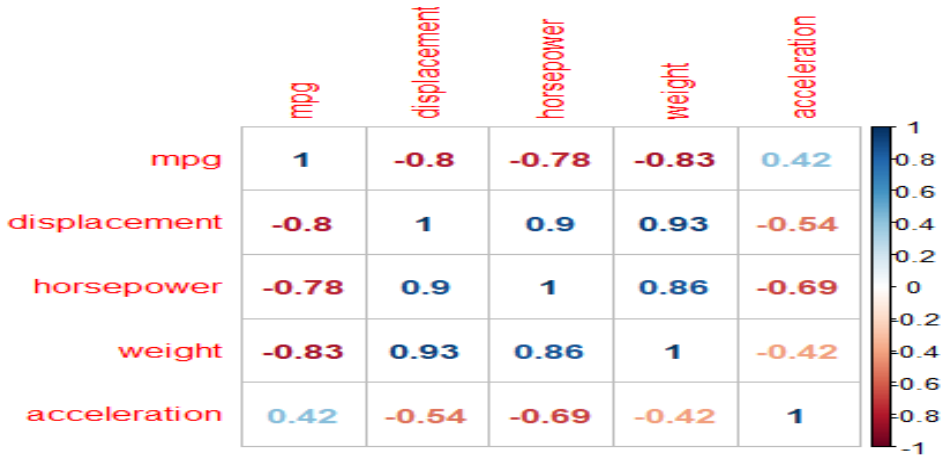


Figura 10. Correlaciones entre variables de los datos usando pairs() (Pizarro, Regresion lineal de autos, 2020).

7.3.2.3.7. Dividir conjunto de entrenamiento y datos de validación o prueba

```
set.seed(2018)
entrena <- createDataPartition(datos$mpg, p=0.7, list = FALSE)
# entrena

nrow(entrena) # Cuantos datos de entrenamiento

## [1] 280
```

7.3.2.3.8. Generando el modelo lineal con los datos de entrenamiento

- Los datos de entrenamiento son los registros de datos que están en esas posiciones que fueron generados por createDataPartition
- Los datos de validación son los que no son del entrenamiento
- No interesan las columnas 1, 8 y 9 de los datos, no resultan significativas para la regresión.

```
datosentrenamiento <- datos[entrena, -c(1,8,9)]

datosvalidacion <- datos[-entrena, -c(1,8,9)]

# Se observa los datos, datos de entrenamiento y datos de validación
head(datos)

##   No mpg cylinders displacement horsepower weight acceleration model_y
ear
## 1  1  28        4c          140         90   2264         15.5
71
## 2  2  19        3c           70         97   2330         13.5
72
## 3  3  36        4c          107         75   2205         14.5
82
## 4  4  28        4c           97         92   2288         17.0
72
## 5  5  21        6c          199         90   2648         15.0
70
## 6  6  23        4c          115         95   2694         15.0
75
##              car_name
## 1 chevrolet vega 2300
## 2     mazda rx2 coupe
## 3        honda accord
## 4     datsun 510 (sw)
## 5         amc gremlin
## 6          audi 100ls

head(datosentrenamiento)
```

```
##    mpg cylinders displacement horsepower weight acceleration
## 3 36.0        4c          107         75   2205         14.5
## 4 28.0        4c           97         92   2288         17.0
## 5 21.0        6c          199         90   2648         15.0
## 6 23.0        4c          115         95   2694         15.0
## 7 15.5        8c          304        120   3962         13.9
## 8 32.9        4c          119        100   2615         14.8
```

```
head(datosvalidacion)
```

```
##     mpg cylinders displacement horsepower weight acceleration
## 1  28.0        4c          140         90   2264         15.5
## 2  19.0        3c           70         97   2330         13.5
## 11 12.0        8c          429        198   4952         11.5
## 13 13.0        8c          302        130   3870         15.0
## 14 27.9        4c          156        105   2800         14.4
## 18 14.0        8c          400        175   4385         12.0
```

```
modelo <- lm(mpg ~ ., data = datosentrenamiento)

modelo
```

```
##
## Call:
## lm(formula = mpg ~ ., data = datosentrenamiento)
##
## Coefficients:
##  (Intercept)   cylinders4c   cylinders5c   cylinders6c   cylinders8c
##    37.284202      6.231475      8.248195      2.131026      4.568171
## displacement    horsepower        weight  acceleration
##     0.002245     -0.057543     -0.004665      0.050745
```

- Prediccion de mpg está dada por la fórmula:
- mpg = 37.284202 + 6.231475 * 4c + 8.248195 * 5c + 2.131026 * 6c + 4.568171 * 8c + 0.002245 * displacement - 0.057543 * horsepower -0.004665 * weight + 0.050745 * acceleration

7.3.2.3.9. Summary(modelo)

- La pregunta es: ¿Éste modelo predice bien o predice mal?
- No todas las variables son importantes, hay algunas que tienen mayor presencia que el resto.
- summary(), para ver otras variables

```
summary(modelo)
```

```
##
## Call:
## lm(formula = mpg ~ ., data = datosentrenamiento)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.0606  -2.4686  -0.4435   1.9821  16.0907
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  37.2842024  3.6497412  10.216  < 2e-16 ***
## cylinders4c   6.2314753  2.4926855   2.500  0.01301 *
## cylinders5c   8.2481946  3.8091396   2.165  0.03123 *
## cylinders6c   2.1310256  2.7759570   0.768  0.44335
## cylinders8c   4.5681710  3.2054454   1.425  0.15527
## displacement  0.0022449  0.0108924   0.206  0.83687
## horsepower   -0.0575428  0.0202773  -2.838  0.00489 **
## weight       -0.0046652  0.0009999  -4.665 4.84e-06 ***
## acceleration  0.0507454  0.1443575   0.352  0.72547
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.092 on 271 degrees of freedom
## Multiple R-squared:  0.7304, Adjusted R-squared:  0.7224
## F-statistic: 91.75 on 8 and 271 DF,  p-value: < 2.2e-16
```

- Los asteriscos significan que las variables aportan mayor significación a la variable independiente “mpg”
- El valor de residuos es como se distribuyen los errores en la estimación
- Los errores que se cometen en la regresión se identifican con el *boxplot* (diagrama de caja), identifican que tanto los datos están cerca de la mediana

```
boxplot(modelo$residuals)
```

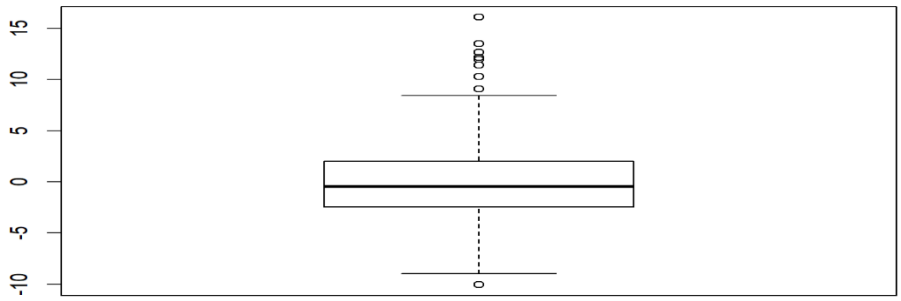


Figura 11. Diagrama de caja boxplot del caso autos. (Pizarro, Regresion lineal de autos, 2020)

7.3.2.3.10. Error cuadrático medio del modelo en los datos de entrenamiento

```
sqrt(mean((modelo$fitted.values - datosentrenamiento$mpg) ^ 2))
```

```
## [1] 4.026021
```

Evaluar con los datos de validación que no serán los de entrenamiento

```
mpg_prediccion <- predict(modelo, datosvalidacion)
```

Error cuadrático medio del modelo en los datos de validación

```
sqrt(mean((mpg_prediccion - datosvalidacion$mpg) ^ 2))
```

```
## [1] 3.894627
```

7.3.2.3.11. Predecir dos nuevos autos con ciertas características

- 4c y 5c
- dsplacement = 80 y 100
- horsepow = 70, 90
- weight 2000, 2100
- acelera = 14, 16

```
cyl = c(4,5)
dis = c(80,100)
hp = c(70,90)
wei = c(2000, 2100)
ace = c(14, 16)

nuevosdatos <- data.frame(cylinders=cyl, displacement=dis,horsepower=hp,
weight=wei, acceleration = ace)
nuevosdatos
```

```
##   cylinders displacement horsepower weight acceleration
## 1         4           80         70   2000           14
## 2         5          100         90   2100           16
```

7.3.2.3.12. Predecir por la Fórmula

- Se toma solo un coeficiente de cilindraje porque o es
- de 4 columna 2
- de 5 columna 3
- de 6 columna 4
- de 8 columna 6
- modelo$coefficients

- 38.607311983 + 7.212652193 + 0.006877506 * 80 + -0.072208661 *70* + -*0.005155968* 2000 + 0.024851517 * 14

```
# Directo
38.607311983 + 7.212652193 + 0.006877506 * 80 + -0.072208661 *70 + -0.005
155968 * 2000 + 0.024851517 * 14
```

```
## [1] 31.35154
```

```
# Predecir para cuatro cylndros
modelo$coefficients
```

```
##  (Intercept)  cylinders4c  cylinders5c  cylinders6c  cylinders8c displ
acement
## 37.284202446  6.231475322  8.248194569  2.131025643  4.568170956  0.00
2244864
##   horsepower       weight acceleration
## -0.057542812 -0.004665209  0.050745431
```

```
mpg.predict4 = modelo$coefficients[1] +
              modelo$coefficients[2] * 1 +
              modelo$coefficients[6] * dis[1] +
              modelo$coefficients[7] * hp[1] +
              modelo$coefficients[8] * wei[1] +
              modelo$coefficients[9] * ace[1]
mpg.predict4
```

```
## (Intercept)
##    31.04729
```

```
# Predecir para cinco cilindros
modelo$coefficients
```

```
##  (Intercept)  cylinders4c  cylinders5c  cylinders6c  cylinders8c displ
acement
## 37.284202446  6.231475322  8.248194569  2.131025643  4.568170956  0.00
2244864
##   horsepower       weight acceleration
## -0.057542812 -0.004665209  0.050745431
```

```
mpg.predict5 = modelo$coefficients[1] +
              modelo$coefficients[3] * 1 +
              modelo$coefficients[6] * dis[2] +
              modelo$coefficients[7] * hp[2] +
              modelo$coefficients[8] * wei[2] +
              modelo$coefficients[9] * ace[2]
mpg.predict5
```

```
## (Intercept)
##    31.59302
```

Predecir por la función predict()

```
# .....
nuevosdatos$cylinders <- factor(nuevosdatos$cylinders, levels = c(4,5),la
bels = c('4c','5c'))
mpg.predict <- predict(modelo, nuevosdatos)
mpg.predict
```

```
##        1        2
## 31.04729 31.59302
```

Gráficas varias del modelo

```
par(mfrow=c(2,2))
plot(modelo)
```

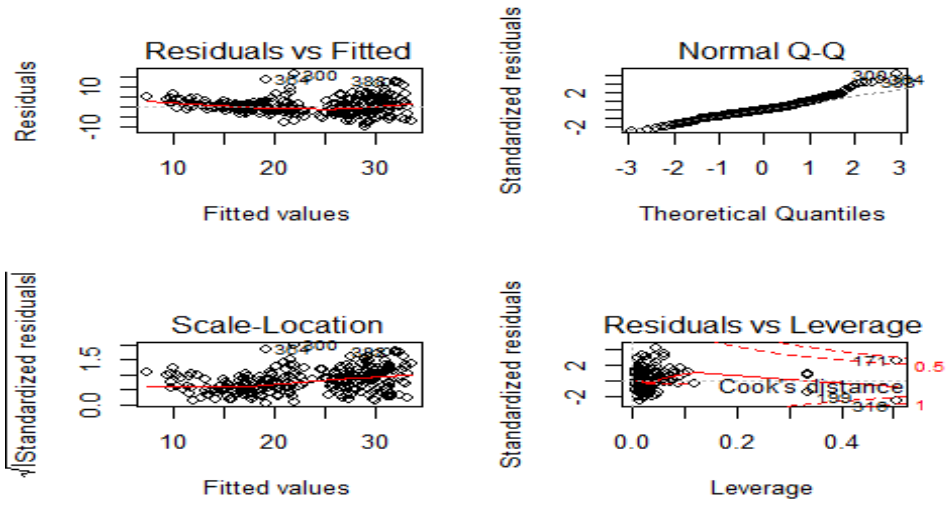


Figura 12. Varias gráficas del modelo de regresión. (Pizarro, Regresion lineal de autos, 2020)

7.3.2.4. Diagrama de Ishikawa

Se presenta un ejemplo de diagrama de Ishikawa que se obtienen del enlace siguiente: https://rpubs.com/ctellez_gdl/58735. El diagrama de Ishikawa se presenta en el siguiente *NoteBook* de R y R Studio

R Notebook

7.3.2.4.1. Las librerías

```
# Se instala la librería qcc
#install.packages("qcc")
# Se activa librería qcc
library(qcc)
```

7.3.2.4.2. Los datos

Se crea una lista con varias columnas llamada causas, se visualiza la lista

```
causas <- list(Measurements=c("Micrometers", "Microscopes","Inspectors"),
               Materials=c("Alloys", "Lubricants", "Suppliers"),
               Personnel=c("Shofts", "Supervisors", "Training","Operators
"),
               Environment=c("Condensation", "Moisture"), Methods=c("Brak
e", "Engager", "Angle"),
               Machines=c("Speed", "Lathes", "Bits", "Sockets"),
                   effect="Surface Flaws")

causas
```

```
## $Measurements
## [1] "Micrometers" "Microscopes" "Inspectors"
##
## $Materials
## [1] "Alloys"     "Lubricants" "Suppliers"
##
## $Personnel
## [1] "Shofts"      "Supervisors" "Training"    "Operators"
##
## $Environment
## [1] "Condensation" "Moisture"
##
## $Methods
## [1] "Brake"   "Engager" "Angle"
##
## $Machines
## [1] "Speed"   "Lathes"  "Bits"    "Sockets"
##
## $effect
## [1] "Surface Flaws"
```

7.3.2.4.3. Diagrama causa y efecto

Se utiliza la función cause.and.effect de la librería qcc

```
cause.and.effect(causas, effect="Surface Flaws")
```

Se muestra en la figura 13 y 14 ejemplo de este código del diagrama causa efecto.

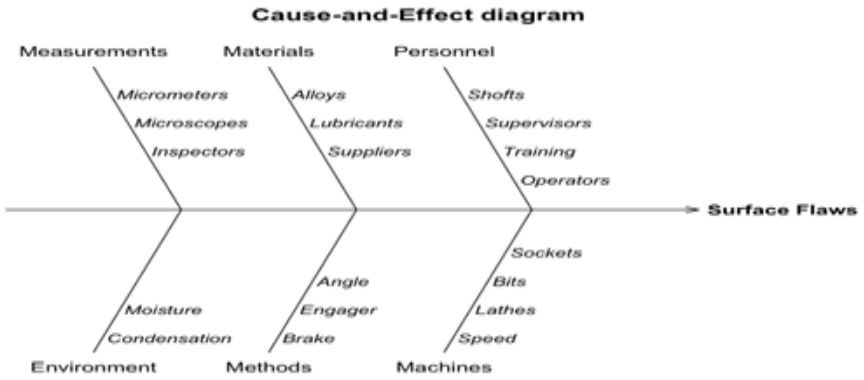


Figura 13. Diagrama de Ishikawa. (Telles Martínez, 2017)

7.3.2.5. Diagrama Sixsigma causa y efecto

Se construyen los datos

```
# Se elabora el diagrama
effect <- "Flight Time"
causes.gr <- c("Operator", "Environment", "Tools", "Design",
               "Raw.Material", "Measure.Tool")
causes <- vector(mode = "list", length = length(causes.gr))
causes[1] <- list(c("operator #1", "operator #2", "operator #3"))
causes[2] <- list(c("height", "cleaning"))
causes[3] <- list(c("scissors", "tape"))
causes[4] <- list(c("rotor.length", "rotor.width2", "paperclip"))
causes[5] <- list(c("thickness", "marks"))
causes[6] <- list(c("calibrate", "model"))
```

Se utiliza la función ss.ceDiag() de la librería sixsigma.

```
ss.ceDiag(effect, causes.gr, causes, sub = "Paper Helicopter Project")
```

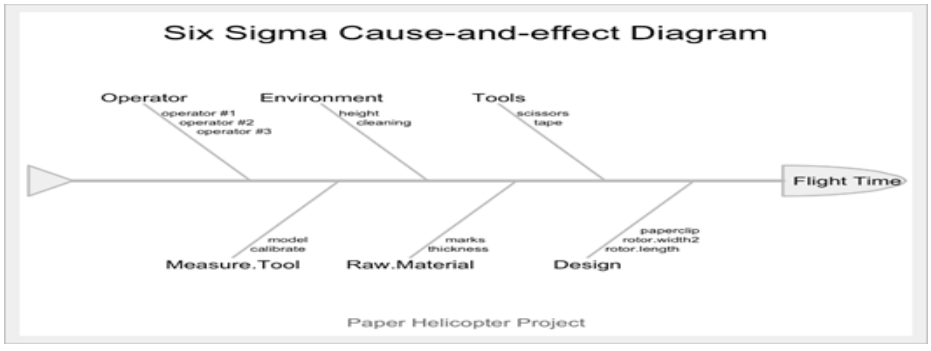


Figura 14. Diagrama de Ishikawa. (Telles Martínez, 2017)

Conclusiones

En este capítulo se explica la importancia de seguir y conservar los modelos de trabajo que han sido pilar de la industria desde sus inicios hasta el presente, al mismo tiempo se destaca que las bases históricas y aquellos modelos que son base para procesos industriales y de producción, ahora van de la mano con la tecnología.

La propuesta implica combinar herramientas y técnicas para la Ingeniería Industrial mediante programación R; dicha propuesta de implementación se pretende que sea posible, pues se tienen ventajas considerables por encima de las opciones actuales y las anteriormente utilizadas para un *MRP Controller*.

El planteamiento presenta una oportunidad de dar a conocer, R y R Studio, como herramientas para empezar a divulgar el tema de Ciencia de los Datos y a los Científicos de Datos en empresas del estado de Durango como una necesidad más que como una inversión, si bien el beneficio será principalmente a nivel personal y laboral, de llegar a ser posible esta implementación se estará atravesando una barrera cultural y educativa en la cual muchos estados ya están adelantados, pues el empezar a crear una necesidad y "plantar la semilla" de tal modo que se da a conocer que en el Instituto Tecnológico de Durango se está optando y apostando por generar tener egresados con capacitación a la par de cualquier cuidad y con el nivel de instituciones privadas.

Esta opción de implementación, de ser aceptada, tendrá un impacto laboral importante pues de acuerdo a los análisis realizados , representará una optimización de tiempo en las actividades que se realizan, los datos que se obtengan serán más exactos y por ende confiables, la cadena de beneficios obtenidos a partir de la optimización del manejo de la hoja critica (hoja de trabajo) , tendrá un impacto importante ya que el mantener y tener inventarios más sanos, es decir, control de exceso, y de obsoletos, órdenes a proveedor en tiempo y cantidad más adecuada, tiempos de tránsito mejor analizados a futuro.

La Ciencia de los Datos no es solamente un tema novedoso y relacionado a colectar información, es un tema visionario y que requiere y exige una ejecución,

que puede ayudar a las empresas a desarrollar e implementar estrategias mucho más robustas y personalizadas.

Se puede usar Ciencia de los Datos en distintas organizaciones desde niveles altos hasta niveles más bajos llegando a todos los procesos.

Bibliografía

Aiteco Consultores, S. (s.f). *Aiteco Consultores S.L.* Obtenido de Hojas de Comprobación, de Control o Verificación: Aiteco Consultores S.L.

Alcalde San Miguel, P. (2009). Calidad. (1a. Edición). En P. Alcalde San Miguel, *Calidad. (1a. Edición).* Paraninfo.

Alicia Vila, M. S. (16 de agosto de 2018). *Universidad Oberta de Catalunya*. Obtenido de www.uoc.edu: http://estudios.uoc.edu/es/buscador-resultados?searchWords=correlacion+lineal

Anderson, D., Sweeney, D., & Williams, T. (2008). *Estadística para administración y economía Estadística para administración y economía. 10a. Edición.* México, D.F: Cengage Learning Editores,S.A. de C.V.

Delphi. (01 de 01 de 2018). *Delphi Technologies*. Obtenido de What is Delphi Technologies: https://www.delphi.com/company

Fernández y Fernández, C. A., & Quintanar Morales, J. A. (2015). Reducciones temporales para convertir la sintaxis abstracta del diagrama de flujo de tareas no estructurado al álgebra de tareas. *ReCIBE. Revista electrónica de Computación, Informática, Biomédica y Electrónica*, 1-35.

García Nocetti, F. (01 de 05 de 2017). *Ciencia de datos y big data.* Obtenido de Nexos: https://www.nexos.com.mx/?p=31892

Garza, E. G. (2003). *Administracion de la Calidad Total.* México: Pax.

Gómez, G. (01 de 12 de 2017). USO DE HERRAMIENTAS DE CALIDAD EN INDUSTRIA TEXTIL: CASO CONFECCIONES WINTER S.A. *TRABAJO FINAL DE GRADUACION. FACULTAD DE CIENCIAS ECONOMICAS*. Buenos Aires, Argentina, Argentina: UNIVERSIDAD NACIONAL DEL CENTRO DE LA PROVINCIA DE BUENOS AIRES.

González González, R., & Jimeno Bernal, J. (01 de 01 de 2012). *PDCA Home*. Obtenido de Diagramas de control: Gráficos para controlar procesos: https://www.pdcahome.com/diagramas-de-control/

Hernández, R., Fernández, C., & Baptista, M. d. (2014). *Metodología de la Investigación.* México: McGRAW-HILL / INTERAMERICANA EDITORES, S.A. DE C.V.

lean, p. (16 de 09 de 2016). *progressa lean*. Obtenido de Diagrama Causa-Efecto (Diagrama Ishikawa): https://www.progressalean.com/diagrama-causa-efecto-diagrama-ishikawa/

Lind, D. A., Marchal, W. G., & Wathen, S. A. (2015). *Estadística aplicada a los Negocios y a la Economía.* México, D.F: McGrawn Hill Education. McGRAW-HILL/INTERAMERICANA EDITORES, S.A. DE C.V.

Mendenhall, W., Beaver, R. J., & Beaver, B. M. (2010). *Introducción a la probabilidad y estadística.* México, D.F.: Cengage Learning Editores, S.A. de C.V.

Navarro Albert, E. G. (2017). Metodología e implementación de Six Sigma. 3C Empresa: investigación y pensamiento crítico. *3C Empresa (Edición Especial). Área de Innovación y Desarrollo, S.L.*, 73-80.

Pizarro, R. (05 de 29 de 2020). *ANOVA una via. Caso consumo alcohol.* Obtenido de ANOVA una via. Caso consumo alcohol: ANOVA una via. Caso consumo alcohol

Pizarro, R. (29 de 05 de 2020). *Diagrama de Pareto.* Obtenido de Diagrama de PAreto: https://rpubs.com/rpizarro/624226

Pizarro, R. (24 de 02 de 2020). *Regresion lineal de autos.* Obtenido de Regresion lineal de autos: https://rpubs.com/rpizarro/578026

Rodríguez, G. d. (2015). LA CIENCIA DE LOS DATOS Y SU IMPACTO. *Revista Científica ECOCIENCIA*, 15.

R-project.org. (01 de 01 de 2019). *R-project.org*. Obtenido de R-project.org: https://www.r-project.org/

Spiegel, M., Schiller, J., & Srinivasan, R. A. (2007). *Probabilidad y Estadística .* Mexico D.F: McGraw-Hill.

Studio, R. (01 de 01 de 2019). *R Studio*. Obtenido de R Studio: https://rstudio.com/products/rstudio/

Telles Martínez, C. (07 de 02 de 2017). *Ishikawa.R.* Obtenido de Ishikawa.R: https://rpubs.com/ctellez_gdl/58735

Capítulo 8

Análisis de datos masivos en el campo de la salud

Teresita De Jesús Camacho Cepeda
Tecnológico Nacional de México. Instituto Tecnológico de Durango
03040118@itdurango.edu.mx

Marco Antonio Rodríguez Zúñiga
Tecnológico Nacional de México. Instituto Tecnológico de Durango
mrodriguez@itdurango.edu.mx

8.1. Introducción

Este capítulo tiene la finalidad de identificar conceptos básicos sobre el análisis de datos masivos en el campo de la salud, conocer técnicas y herramientas para el análisis de estos datos además de conocer la aplicación de Big Data en el mismo campo.

Se da a conocer el impacto y beneficio que tiene el uso del análisis masivo en el campo de la salud, así como la importancia de su implementación.

En el marco contextual se mencionan los diferentes tipos de datos que maneja Big Data al igual que sus características, se hace mención de las fuentes de datos masivos y la descripción de cada una de estas fuentes.

Así mismo se informa sobre las herramientas Big Data más conocidas y se da una breve descripción de cada una de ellas.

Se muestra la analítica predictiva que es la de objeto de interés para la aplicación de la propuesta, describiendo algunos de los algoritmos genéricos y se muestra una tabla con las características de algunos de estos algoritmos, con el objetivo de identificar su aplicación en la propuesta del sistema de unificación de contenidos generales del sector salud.

Se menciona la importancia y los beneficios de aplicar Big Data y la influencia que tiene en el campo o sector salud. Se da a conocer la importación de Big Data en el sector y los beneficios que brinda en diferentes áreas.

Se describen los beneficios más notorios en el ámbito de la salud y el impacto de Big Data en la Biomedicina, con ello se enlistan algunos ejemplos.

Se menciona la gran responsabilidad que carga el análisis de datos masivos y se tipifica un decálogo con puntos que se deben considerar para la protección de datos sensibles en el campo de la salud.

Se plantea una propuesta para unificar contenidos generales del sector salud con el objetivo de integrar todos los datos generados por diversas fuentes con la finalidad de tener una sola base de datos para poder realizar consultas y tener a disposición y poder realizar predicciones y prevención de enfermedades mejorando el sector y la calidad de vida de las personas gracias a una atención médica personalizada utilizando técnicas y herramientas para el análisis masivo de datos, dando solución a muchas de las problemáticas que actualmente se presentan en las instituciones de salud.

El presente trabajo académico se justifica dado que la enorme cantidad de datos generados del sector salud va creciendo de manera exponencial a tal grado que pareciera imposible acumular toda esa información, gracias a Big Data esta tarea es posible y aplicado de manera correcta y eficiente se puede anticipar y prevenir enfermedades.

En el sector salud existen resultados de laboratorio, radiografías, historiales clínicos y muchos más que implican un gran volumen de estos datos que se generan

a gran velocidad y provienen de fuentes muy variadas. Esto justifica plenamente el desarrollo de este producto académico.

Big Data es cada vez más útil en el sector salud sumado a la introducción de avances tecnológicos como por ejemplo los relojes inteligentes donde toda esa información recopilada puede ser utilizada para diagnosticar y tratar a un paciente, es como tener un registro actualizado y a la mano que se puede estar alimentando constantemente para que esa información sea utilizada por los profesionales de la salud.

Big Data permite analizar detalladamente los datos de un lugar del mundo para identificar tendencias específicas que permitirá predecir y prevenir enfermedades.

Como objetivo general se pretende describir el análisis del manejo y procesamiento de datos masivos con el fin de conocer su aplicación en áreas concretas del campo de la salud mediante el uso de técnicas y herramientas Big Data.

De manera específica se busca lo siguiente: identificar conceptos del análisis de datos masivos; mencionar técnicas y herramientas; determinar aplicaciones de Big Data en el campo de la salud y definir una propuesta de implementación del análisis de datos en el campo de la salud.

Este documento tiene como propósito realizar una indagación sobre el análisis de datos masivos en el campo de la salud y dar a conocer cuáles son las técnicas y herramientas que se utilizan actualmente para el uso masivo de la información, así como adopción e integración de Big Data en este sector, aunado a conocer los beneficios que trae consigo el uso de esta tendencia en el sector.

El uso de datos en el ámbito de la salud puede dar mucha información experiencias previas para tomar medidas y encontrar soluciones exitosas en relación al bienestar del paciente (Martínez Velasco, 2019).

8.2. Marco de referencia

En primer lugar, es importante estar familiarizados con el significado de Big Data, también conocido como Análisis de Datos Masivos.

8.2.1. Datos masivos y Big Data

El manejo de datos masivos actualmente es una tendencia tecnológica conocida como Big Data que maneja grandes y complejos conjuntos de datos e información difíciles de procesar utilizando herramientas convencionales, la dificultad está en cómo acceder, distribuir y utilizar la gran cantidad de datos “No Estructurados” para que ayuden a tomar mejores decisiones (Joyanes Aguilar, 2013).

La situación que presenta el sector salud es que los pacientes y hospitales generan y almacenan grandes cantidades de datos en formatos de papel o electrónicos que impiden utilizar la información de manera eficiente.

Big Data se convierte en una gran oportunidad de acceder a grandes cantidades de información, es decir, obtener más información contextualizada, diagnósticos concretos y correctos, y atención personalizada. El almacenamiento de grandes cantidades de datos hace posible un avance significativo en la prevención y diagnóstico con una constante conexión con diferentes sectores de salud a nivel mundial facilitando un diagnóstico con parámetros más amplios a los que estábamos acostumbrados antes de tener esta tecnología.

Dentro del análisis masivo de datos se encuentra la analítica predictiva que no es más que el almacenamiento de grandes cantidades de información tanto del sector público o privado en el sector salud para detectar patrones de enfermedades y disminuir su afectación en los pacientes.

Varias prácticas de la medicina se ven favorecidas por la actual tecnología ya que tiene gran impacto en la sociedad pues mejora el diagnóstico y tratamiento, ayuda a que la intervención sea menos invasiva en la atención del paciente.

La atención médica del futuro será más eficaz si se logra aprovechar el potencial de la gran cantidad de datos que se tendrán. Mucho se habla de la relación

de Big Data con la Medicina de las “4P”, personalizada, predictiva, preventiva y participativa, la cual significa un cambio de paradigma en el modelo de atención médica, que hace posible tratar al paciente en etapas iniciales de su enfermedad. (Vidal Ledo, Morales Suárez, Menéndez Bravo, González Cárdena, & Portuondo Sao, 2020)

- Personalizada: Gracias a Big Data y a los avances de la aplicación del genoma humano en la medicina se podrá ofrecer al paciente la terapia más adecuada.
- Predictiva: El Big Data y el análisis de datos procedentes de sensores que miden frecuencia cardiaca, frecuencia respiratoria, presión sanguínea, entre otros, en conjunto con análisis clínico o radiografías se puede detectar alguna infección antes de padecer síntomas.
- Preventiva: El impacto del Big Data en la medicina preventiva con gran repercusión fue un ejemplo en el caso de la actriz Angelina Jolie ya que se realizó una prueba de secuenciación de ADN y reveló que tenía una mutación en el gen BRCA1 que aumentaba hasta en un 87% la probabilidad de cáncer de mama, gracias al diagnóstico temprano y a la doble mastectomía se previno la enfermedad.
- Participativa: Big Data aprovechará todos los datos de los wearables (vestimenta dotada de sensores), como el pulso, glucosa, temperatura para mejorar la atención de los pacientes y estén enterados de su salud actual.

Para poder utilizar el máximo de Big Data lo mejor sería capturar, almacenar y analizar todo dato existente sobre análisis clínicos, historiales médicos, secuenciación de ADN de pacientes, informaciones de redes sociales para conformar una base de datos compartida entre el sector salud.

8.2.2. Big Data

El Big Data tiene que ver con la recopilación de datos, estos pueden ser de fuentes tradicionales o de fuentes digitales tanto dentro como fuera de una empresa.

El concepto Big Data también hace referencia a un grupo de tecnologías y herramientas que son capaces de capturar, almacenar y procesar enormes cantidades de datos en poco tiempo y con un costo aceptable para una empresa.

Big Data se conoce como una nueva generación de tecnologías, arquitecturas y estrategias diseñadas especialmente para la captura y el análisis de grandes volúmenes de datos que son provenientes de diferentes fuentes a una muy alta velocidad con la finalidad de obtener valor económico de ellos.

El Big Data se puede encontrar en datos digitales como por ejemplo el comportamiento de la web o las interacciones de las redes sociales.

Es importante una vez que se menciona el significado Big Data conocer sus Vs. En total son 7 Vs:

Figura 1. Las 7 Vs del Big Data. (Montealegre Gallo, 2017)

- Volumen: Esta es una característica que se refiere a las grandes cantidades masivas de datos que tienen que ser procesadas y analizadas por la Tecnología Big Data.
- Variedad: En esta característica tiene que ver con los diferentes tipos de fuentes de datos que Big Data puede admitir y gestionar. Estos tipos de datos pueden ser estructurados (es decir bien definidos o los datawarehouse) y también no estructurados (como las radiografías o resonancias magnéticas). Este punto significa que los datos son diferentes en función de sus características dependiendo de su origen, forma, tipo, estructurados o no estructurados.
- Velocidad: Esta característica se refiere a datos en movimiento, es decir la velocidad del flujo de datos, desde que se originan hasta que se presentan como información útil para la toma de decisiones; para determinados propósitos en el sector salud resulta crítico que la rapidez con la que se realizan estos procesos sea la más cercana a la realidad ya que se podría salvar una vida.

- Veracidad: Esta característica se refiere a la fiabilidad de los datos, que la información sea de calidad y sobre todo veraz, pero por mucho que se dedique a la limpieza de los datos para obtener mejor calidad, en algunos casos no se puede eliminar la imprevisibilidad ni la incertidumbre, aun así, estos datos masivos deben tenerse en cuenta.
- Variabilidad: ¿El flujo de datos es regular, o varía? ¿Se puede contar con estos datos incluso en condiciones impredecibles? Esta V se refiere a la necesidad de obtener datos relevantes considerando todas las circunstancias posibles.
- Visualización: Es la posibilidad de identificar patrones y claves útiles, poder ser visualizado de manera ágil, con herramientas dinámicas que permitan su representación para la constante búsqueda de variables o tendencias que ayuden a los procesos de negocio.
- Valor: Big Data es una combinación de las anteriores. Cada empresa podrá adoptar esta tecnología bajo diferentes enfoques, pero con un objetivo común: mejorar el rendimiento y la toma de decisiones. (Sánchez Villaseñor, 2019).

8.2.3. Tipos de Datos

Actualmente un hospital promedio arroja 665 Terabytes de datos al año, pero en su mayoría no son útiles ya que por lo menos el 80% de estos datos son “No estructurados”, lograr que toda esta información sea útil sería de gran ayuda en el sector salud.

El manejo de Datos Masivos se distingue por manejar datos estructurados, datos semiestructurados y datos no estructurados.

Datos Estructurados: Son los datos que ya tienen un esquema definido, que pueden ser ingresados a un campo específico como fecha, numero o nombre y almacenarlos en tablas.

Datos Semiestructurados: Son los datos que no tienen un esquema o campo definido, pero tienen muchas etiquetas o marcadores para distinguir los diferentes elementos del dato, por ejemplo, los datos clínicos.

Datos no estructurados: Son los datos que no tienen esquema o campo determinado y no es posible almacenarlos en tablas, son tomados en cuenta como objetos o documentos (Joyanes Aguilar, 2013).

La figura 2, identifica de otra manera el concepto de los tipos de datos.

Figura 2. Tipos de Datos Big Data. (Facultad de Estudios Estadísticos. , s.f.)

8.2.4. La fuente (Captura) de Datos Masivos

Existen cinco categorías de fuentes de datos, cada una de ellas comprende diferente tipo de información, los que se mencionan se refieren al sector salud.

- Wen and Social Median: Obtiene datos de clicks, Twitter, Facebook, blogs y contenidos web, Big Data puede recabar datos abundantes para el sector salud por medio de redes sociales específicas para profesionales de la medicina.
- Machine- to- Machine (M2M): Es la tecnología que permite la conexión entre dispositivos como los sensores que registran datos específicos como la temperatura, humedad, velocidad, presión también pueden ser las lecturas de medidores inteligentes, RFID, y señales GPS. En el caso del sector salud los que recaban los datos pueden ser los datos y sensores en dispositivos weareables, también de Smartphone con pacientes monitorizados.
- Big Transaction Data: Estos datos transaccionales pueden ser datos estructurados o datos no estructurados, provenientes de telecomunicaciones o registros de llamadas.
- Biometrics: Es toda información Biométrica como huellas digitales, el escaneo de retina, el reconocimiento facial, en el caso del sector salud la genética (ADN).

- Human Generated: Son datos que generan las personas, por ejemplo, los datos que se guardan en los Call Center en las llamadas telefónicas, notas de voz, documentos, correos electrónicos y en caso del sector salud estudios médicos, registros electrónicos o recetas médicas.

Figura 3. Fuente de Datos Masivos Big Data. (Rayo, 2017)

8.2.5. Herramientas Big Data.

El manejo de Datos Masivos necesita utilizar herramientas que puedan capturar datos de diferentes fuentes, la plataforma Apache Hadoop es la que se ha utilizado desde los inicios en distintos proyectos Big Data; Oracle, IBM y Microsoft aparecen como los principales proveedores de herramientas Big Data.

Con respecto a Hadoop, este tiene tres componentes fundamentales:

- ***Hadoop Distributed File System (HDFS)***: Son datos que están en el clúster de Hadoop que se segmentan en partes pequeñas a las que se les llama bloques y que estos a su vez se distribuyen por el clúster y así la función *map* y *reduce* son ejecutadas en subconjuntos pequeños, procesan grandes volúmenes de datos.
- ***Hadoop MapReduce***: Es el componente principal de Hadoop, *MapReduce* tiene que ver con que Hadoop realice dos procesos por separado: *map* selecciona un conjunto de datos y lo convierte en otro donde son divididos en *tuplas* (pares clave/valor); *reduce* obtiene la salida de *map* como datos de entrada y mezcla las tuplas en un conjunto más pequeño de las mismas.
- ***Hadoop Common*:** Son librerías que soporta Hadoop.

La tabla 1 identifica otros aspectos de Big Data relacionados con Hadoop.

Tabla 1

Herramientas Big Data

Plataforma	Descripción
Hadoop	Maneja grandes volúmenes de datos y distribuye cargas de procesamiento, compuesto por Map Reduce y archivos distribuidos
PIG	Lenguaje de alto nivel para flujo de datos, paraleliza grandes volúmenes de tipo MapReduce que pueden ser interpretados por Hadoop
MAHAOUT	Biblioteca escalable para hacer minería de datos y aprendizaje automático, se integra con otras herramientas Hadoop, entorno de trabajo similar R
Hbase	Administrador de base de datos para ambiente distribuidos
Spark	Motor de base de datos que usa Hadoop, con soporte para aprendizaje automático. Compatible con Python R y SQL
Cassandra	Gestor de Base de datos a gran escala para datos de misión crítica, escalable, tolerante a fallas y alta replicación

8.2.6. Analítica Predictiva

La analítica predictiva existe desde hace décadas antes de que se diera a conocer su gran importancia gracias a las opciones que ofrece Big Data como la gran cantidad de datos que se capturan de personas como transacciones en línea o redes sociales, sensores como dispositivos GPS, así como el poder de procesamiento que se hizo costeable tanto en la nube como en Hadoop.

La analítica predictiva se refiere a una tecnología que aprende con base en los datos (experiencia) para predecir un comportamiento para mejores decisiones. La esencia de los datos masivos es que permite hacer predicciones, al aplicar matemáticas a grandes cantidades de datos se puede hacer probabilidades.

Por ejemplo en el ámbito de la salud en términos predictivos y como lo menciona Gutiérrez Martínez & Febles Estrada, (2019) “La medicina del futuro será mucho más eficaz si se logra aprovechar mejor el potencial que representarán la gran cantidad de datos que se tendrán” (Gutiérrez Martínez & Febles Estrada, 2019)

La analítica predictiva se materializa por la creación de modelos predictivos, un modelo predictivo es como un mecanismo que va a predecir el comportamiento de un individuo por ejemplo con un click o una compra, las características del individuo se toman como datos entrantes y de salida genera una puntuación predictiva, entre más sea su puntuación, las probabilidades son mayores de que el individuo muestre un comportamiento predictivo.

En los modelos predictivos se aplican funciones matemáticas y algoritmos para determinar una correlación entre un conjunto de datos de entrada, los algoritmos son parte de los métodos y técnicas de la minería de datos.

Existe un grupo de algoritmos genéricos que apoyan a la analítica predictiva:

- **Máquinas de vectores de soporte (SVM):** Son algoritmos de aprendizaje supervisado para dar solución a problemas de clasificación y regresión, este construye hiperplanos en un espacio de dimensión mayor al encontrado calculando al que proporcione mayor separación entre los dos subconjuntos será el hiperplano óptimo, esto les dará etiqueta de clase y función de regresión que otorga el valor predictivo.
- **Redes Neuronales (NN)**: Aquí se presenta una estructura de aprendizaje automatizado basado en el sistema nervioso de animales, está constituido por un conjunto de nodos conectados entre sí, establece una correlación entre los nodos de entrada y destino.
- **Árboles de Decisión**: Con base en un conjunto de datos se construye un diagrama de construcción lógica, este modelo es más fácil de usar y entender, se parece al sistema de predicción basado en reglas, sirve para presentar una serie de condiciones sucesivas para solucionar un problema.
- **De agregación o clustering**: Es un conjunto de técnicas para clasificar un conjunto de individuos en grupos homogéneos. De los más usados el algoritmo *k-means* trata de obtener una partición de un conjunto con n observaciones en k número de grupos, en donde cada observación es del grupo más cercano a la media, estos elementos son los centroides, el conjunto inicial elige al azar al

centroide (partición), calcula la media a cada grupo y repite el proceso hasta que el centroide es el mismo.

- **Reglas de asociación**: Esta aplica si la importancia está en las asociaciones de los elementos de entrada tomando en cuenta que la variable destino no es importante, un claro ejemplo es que hay de común en las personas que compran leche y pañales y además cerveza, este es un análisis de cesta de compra utilizado en decisiones de marketing.

8.2.7. Herramientas del Análisis Predictivo

Las organizaciones a nivel mundial están dándole mucho valor al análisis predictivo. Así mismo, están poniendo mucho énfasis en adquirir herramientas que les permitan hacer este tipo de análisis.

Algunos fabricantes de herramientas para análisis predictivo sería los siguientes:

- **IBM**: Tiene un conjunto importante de prestaciones y pone la predicción en el centro. Tiene una opción para cada organización que empieza un análisis predictivo. Su software ofrece realizar modelos, análisis y aplicaciones predictivas ya sea en la empresa físicamente como en la nube.
- **SAS**: Es una potencia en la analítica, ofrece casi toda característica de un científico de datos. Se actualiza con base en las necesidades de usuarios. SAS Visual Analytics ofrece solución “Todo- en- uno”, herramientas de visualización y de análisis predictivo, estas soluciones se añaden a las de código abierto de R, Python y Hadoop**.**
- **Actores Fuertes (Strong Performers)**: Pertenece al grupo de Oracle, Dell, Alpine Data Labs, Alteryx, RapidMiner entre otros, todos tienen atractivo en sus productos lo cual los hacen buenas opciones para las empresas, los que tienen mejor puntaje en la estrategia son Oracle, Alteryx y RapidMiner solo por mencionar algunos.
- **Microsoft y Predixion software**: Son también de esta categoría, Microsoft se centra en los servicios de la nube al igual que Predixion Software, pero este le da capacidades de análisis predictivo a Excel.

- **Código Abierto**: El software de código abierto es un motor de análisis predictivo, el lenguaje de programación de código abierto R para estadísticas y análisis predictivo siempre ha estado presente y es de los más apoyados. Los desarrolladores de aplicaciones cuentan con bibliotecas API así reparan datos y construyen modelos predictivos utilizando Python, Java y Scala, en Python se puede utilizar NumPy y Scipy para preparar datos y construir modelos predictivos.

8.2.8. Big Data en Sector Salud

De acuerdo a la necesidad de los datos en el campo de la dalus y como lo menciona Biedma Ferrer & Bourret, (2019) “en la próxima década la medicina estará caracterizada por la utilización intensiva e inteligente de los datos” (Biedma Ferrer & Bourret, 2019).

Big Data aporta un gran beneficio al sector salud tanto para la prevención, detección y tratamiento de padecimientos, se enlistan algunos de los beneficios:

- Mejora el diagnóstico por parte de los médicos debido a que se les brinda información ordenada como punto de referencia.
- Existe un mayor control con la asistencia y comunicación con los pacientes.
- Se puede predecir alguna enfermedad con base en el historial médico del paciente.
- Se agilizan los sistemas de gestión y de pago a proveedores.
- Ayuda a los investigadores con búsqueda de tratamientos.

Cabe mencionar que para poder aprovechar todos estos beneficios y que todo este desarrollo tecnológico tenga impacto en la calidad y eficiencia en el sector salud o en cualquier otro, será de suma importancia crear una cultura tecnológica que pueda garantizar un buen manejo de esta, ligado a la ética, valores y responsabilidad del uso de la misma.

Para sacar el máximo de Big Data es necesario capturar, almacenar y analizar todos los datos sobre ensayos clínicos, historiales médicos, secuenciación de ADN, información de redes sociales, entre otros, con la finalidad de formar una

gran base de datos donde se involucre toda institución relacionada con el sector. La figura 4, identifica un ecosistema de fuentes de datos que se pueden aprovechar en el sector salud.

Figura 4. Integración de Fuentes Datos en el sector salud. (IDC Salud, 2013)

Las técnicas de Big Data se enfrentan a algunos retos ocasionados por el avance de mecanismos de almacenamiento y gestión de datos, computación en la nube y por los resultados que permiten obtener, almacenar y procesar datos de todo tipo, se enlistan algunos:

- Necesidad de integrar la información.
- Documentar la información de manera digital sin mayor esfuerzo por los profesionales de la salud.
- Analizar datos no estructurados.
- Si los de datos (stocks de información) cuya integración se tendrá que asimilar.
- La parte técnica y legal que brinde seguridad al compartir e intercambiar datos.
- Medios para asegurar la calidad.

8.2.8.1. Importancia de aplicar Big Data en el Sector Salud

A continuación, se tipifica la importancia de aplicar Big Data el el Sector Salud:

- **Costo vs Efectividad**: Con la aplicación de Big Data se mejora el análisis de costo vs efectividad de los tratamientos para poblaciones específicas. Actualmente se sufre un cambio demográfico y epidemiológico constante en

todo el planeta aunado al avance tecnológico; todo el sector salud enfrenta presión por el gasto, esto obliga a los gobiernos a determinar cuáles son los tratamientos que no ponen en riesgo la sustentabilidad financiera del sector. Big Data brinda la oportunidad de que se cambie la estrategia de detectar a los pacientes con alto costo de atención por la atención del impacto de diferentes tratamientos para obtener el más costeable.

- **La Mortalidad**: Big Data previene la mortalidad por diferentes enfermedades, la implementación de infinidad de sensores como en los celulares, ropa entre otros ayudan a tener un monitoreo constante y masivo que si se analizan de manera adecuada se puede enviar una alerta al prestador de salud más cercano para que alguna persona con falla cardiaca por ejemplo sea atendida de manera inmediata.
- **Cuidados Intensivos**: En el sector salud hay servicios que presentan mayor riesgo para los pacientes que otros como es el caso de cuidados intensivos ya que se interactúa con muchas variantes constantemente, el análisis Big Data crea mejores estrategias para minimizar riegos como, por ejemplo: infecciones por la aplicación de catéteres, sangrados abundantes o medicaciones incorrectas.
- **Epidemias**: El uso de Big Data es útil para determinar por medio de dispositivos móviles o redes sociales puntos de infección o dispersión de un modo más rápido y preciso, ya que en las recientes pandemias para ser más específicos del ébola y chikungunya se mostraron deficiencias en el sistema de vigilancia de nuestra región dejando ver la intervención tardía con impacto notable en el aumento de mortalidad por esta pandemia. Actualmente el panorama del COVID19.
- **Análisis de Factores de Riesgo**: La alimentación, la salubridad, el medio ambiente, la educación y el empleo son factores de los que depende el sector salud, de manera que si Big Data realiza cualquier mejora en los sectores mencionados ocasiona un gran impacto en la salud de la población.

8.2.8.2. Aplicaciones Big Data en Sector Salud

En el sector salud se registran petabytes de datos (petabyte (PB) equivale 10^{15} bytes = 1 000 000 000 000 000 de bytes) en su mayoría esta información no es empleada para la práctica médica, descubrir métodos o conocimientos, o bien para hacer pruebas, ya que actualmente la atención médica sigue siendo personalizada, el diagnostico se determina con base en la interpretación de los médicos y teniendo en cuenta ese diagnóstico se aplica tratamiento a la persona atendida, así crea la posibilidad de hacer un registro de cada una de las citas del paciente, esto facilita el acceso a un formato electrónico, que pueda personalizar las recomendaciones al paciente.

La integración de un sistema inteligente que se alimente de datos con base en la experiencia diaria y que se documente en bases de datos clínicos y que estas a su vez permitan extraer y dar a conocer nueva información de manera constante para mejorar la calidad del sector salud.

Los mayores beneficios se darían en hospitales, clínicas públicas y privadas, las farmacéuticas y claro que los pacientes. Las consultoras Garther y Forrester mencionan que el ahorro del sector salud de los Estado Unidos al aplicar Big Data es de alrededor de 300 millones de dólares. En la figura 5 se presentan algunas industrias beneficiadas con el uso del Big Data.

Figura 5. Industrias que usan Big Data. (QuestionPro, 2019)

A continuación, se mencionan algunos de los beneficios más notables en el sector salud:

8.2.8.2.1. Gestión Hospitalaria

La detallada información de los tratamientos de pacientes puede definir cuáles son los tratamientos más eficientes ya sea de manera individual o grupal. El análisis de datos ayudaría a la buena toma de decisiones médicas y así facilitar el detectar algún error en los tratamientos médicos.

Conocer cuáles son las instituciones de salud con mayor rendimiento, el desempeño profesional y si los procedimientos son los adecuados.

Con base en los perfiles de pacientes se pueden obtener modelos predictivos.

En la parte financiera y contable se podría implementar un sistema de pago automatizado que lleve un control de gastos de la institución.

8.2.8.2.2. Industria Farmacéutica

Sin duda alguna, la industria farmacéutica es una de las más importantes empresas a nivel mundial por lo que estas aportan. En este campo los datos masivos también pueden ser de mucha utilidad. Por ejemplo:

- Las técnicas actuales de análisis de datos masivos pueden mejorar cualquier análisis.
- Brinda los pormenores de los resultados de ensayos clínicos.
- Aporta una mejora en la planificación, diseño y selección de ensayos clínicos en potenciales pacientes.
- Avances en el estudio de ADN.
- Análisis en los patrones de enfermedades.
- Reducción de costos para desarrollo en fármacos.

En general la aplicación de Big Data en el sector salud mejoraría de manera considerable la calidad de la asistencia médica, aumentaría la satisfacción del paciente ya se tendrían nuevas formas de atención y la eficacia de la industria farmacéutica aumentaría.

En otros casos de aplicación de Big Data en la salud seguirán surgiendo modelos nuevos de negocios basados en la virtualización, aplicaciones móviles,

seguimiento a pacientes y protocolos cambiarán totalmente el sector de como lo conocemos actualmente. El informe que fue publicado por la Universidad Internacional de Valencia señala que las búsquedas que se realizan en internet ayudan a detectar pandemias, predicen casos de gripe y realizan estimaciones de salud pública.

8.2.8.3. Big Data en la Biomedicina

La aplicación de Big Data en el sector salud es un claro ejemplo de cómo el análisis inteligente puede mejorar nuestra vida diaria.

En los Estados Unidos y Canadá ya son conocidos los beneficios de utilizar Big Data en la Biomedicina, ahora si se reflexiona sobre la cantidad de información relacionada a un paciente solo por mencionar algunos como: grupo sanguíneo, presión sanguínea, nivel de glucosa, alérgenos, patologías y resonancias magnéticas, si combinamos esta información con la edad, género, hábitos alimenticios, actividad física se puede diseñar política de atención sanitaria, protocolos y tratamientos que se aproximan a la biomédica centrada en un paciente.

El sector salud ha evolucionado al grado de utilizar Big Data y la inteligencia artificial para ámbitos administrativos y de asistencias como es el caso de la cirugía asistida por robot o en la enfermería asistentes virtuales.

En la actualidad ya no es de sorprenderse el utilizar la implantación de dispositivos subcutáneos que recopilan información de nuestra salud para mejorarla, gracias al uso de análisis inteligentes que predicen determinados escenarios que Big Data hace posible, tampoco desconocemos la integración de dispositivos en prendas de vestir que monitorizan los signos vitales las 24 horas del día o el desarrollo de análisis inteligente en robótica quirúrgica.

Big Data se retroalimenta debido a que el análisis inteligente hace posible la creación de mejores dispositivos biotecnológicos que recolectan más información y de mayor calidad permitiendo un avance mayor en este campo.

El análisis de secuenciación de ADN es gracias al aumento de capacidad de procesamiento y almacenamiento de datos aunado a la evolución de

algoritmos. Hace 10 años obtener la secuenciación del genoma humano tenía un costo de 1 billón de dólares actualmente.

Ufe Technologies presentó una herramienta llamada Ion Proton, esta herramienta es capaz de secuenciar por completo el genoma humano en un día por un monto de 1000 dólares los precios irán bajando hasta que se pueda obtener esta información por algunos cientos de dólares, este descenso se debe a la gran evolución en la medicina, la importancia de este evento es obtener el ADN de millones de personas y cruzar todos esos datos.

El genoma de una persona es de aproximadamente 100 gigabytes que es lo equivalente a 102 400 fotografías, es decir un millón de genomas serían cientos de petabytes de datos, el beneficio se encuentra en combinar el perfil genético con los datos del día a día más el entorno en el que se encuentra para saber a la perfección los riesgos de padecer cáncer, diabetes o alguna otra enfermedad.

Esta inteligencia lleva a la medicina personalizada donde se conocerá el tratamiento correcto para el paciente en el momento correcto, el beneficio no solo es en el diagnóstico como anteriormente ya se mencionó, sino también se aplica a la farmacéutica con la creación de nuevos medicamentos más eficaces y por ende la reducción de costes.

8.2.8.4. Big Data Revolucionando el Sector Salud

Big Data está revolucionando el sector, se considera que es el sector que menos utiliza la información de manera correcta para adaptarse a las necesidades del paciente probablemente sea por la crisis económica que tiene en cintura al sector, así que si queremos hacer sostenible el sector estamos obligados a aplicar la medicina de precisión, es decir, darle a cada quien lo que necesita, se tiene que ser efectivo y adelantarse a las necesidades de los pacientes para la prevención de enfermedades.

La medicina de precisión se basa en tres pilares:

- El incremento de la capacidad de lectura del ADN, incluida la técnica de secuenciación masiva y de ensamblaje para una lectura coherente.

- Consolidación de Big Data y el análisis de datos en la biomedicina.
- Utilizar los datos en la práctica clínica: ensayos clínicos a la medida, farmacogenómica, terapia digital y fármacos personalizados.

La propuesta de ofrecer más calidad, mejor atención y obtener la reducción de costes en el sector salud, solo se va a lograr manejando la información de manera correcta.

Big Data en pocas palabras es la disponibilidad del volumen masivo de datos de una gran variedad que permiten su análisis para poder identificar tendencias y patrones de comportamiento, el concepto está ligado al de reutilización y liberación de datos por lo que estos son transformados en información,

Con lo anterior se crea la necesidad de tener bases de datos abiertas para que la comunidad científica pueda utilizarlos bajo un marco regulatorio con base en la confidencialidad y a la privacidad, esta información debe ser protegida y es obligatorio evitar que el uso de datos de la salud identifique a las personas a menos de que sea bajo su consentimiento, es importante mencionar que actualmente existen las herramientas necesarias para el manejo correcto de los datos.

Actualmente la mayoría de las organizaciones sanitarias clasifican a los pacientes según el riesgo, centrando la atención en los grupos más vulnerables, además los avances en relación a la genética y en la medicina personalizada permite usar biobancos de ADN y así poder hacer predicciones sobre enfermedades futuras de una persona, para poder lograrlo se utilizan complejos sistemas de análisis que estudian la relación entre los genes y las patologías como el cáncer y además analiza el ADN para definir si una personas es propensa de padecer alguna enfermedad específica.

Se enuncian algunos ejemplos para revolucionar el sector con Big Data:

- Predicción de hospitalización con base en patologías por factores ambientales.
- Identificación de pacientes de alto riesgo.
- Toma de decisiones en consulta.

- Análisis de salud de toda la población.
- Seguimiento de tendencias.
- Ensayos clínicos, Big Data encontrará al paciente y realizará el proceso casi automático.
- Efectividad de medicamentos y seguimiento de efectos secundarios.
- Evaluación del sector salud.
- Vigilancia epidemiológica.

8.2.8.5. Responsabilidad del Big Data

Big Data trae innumerables beneficios en el sector salud, pero también riesgos. La ética es una disciplina que interactúa con lo jurídico-político que brinda un tipo de guía para realizar acciones, si se toma en cuenta sólo a manera de reflexión se estaría cayendo en un gran error.

Los profesionales de la salud están incorporando sus prácticas en nuevos sistemas de registro y almacenamiento de datos para utilizarlos en lo que se llama el gobierno electrónico y gracias a esto se tiene una sensación de pérdida de privacidad en este caso el ejercicio de la salud está olvidando el concepto de secreto profesional.

Es cierto que gracias a las tecnologías de la información se mejora la atención médica, facilita el acceso y mejora el sistema, sin embargo, esta información personal requiere cuidados especiales.

Durante el intercambio y comparación de bases de datos surgen riesgos y problemas éticos debido a que la naturaleza del Big Data es cambiante no se distingue la práctica académica y comercial.

Actualmente se tiene la propuesta de un decálogo que es importante analizar, tomar en cuenta lo jurídico-político y tener claro que se puede y debe modificar en caso de riesgos y nuevos retos en el manejo masivo de la información:

- **i**. Reconocer que datos son de personas y pueden hacer daño si se utilizan de manera incorrecta: La identificación personal y grupal sumada a la estigmatización brinda la posibilidad de la identificación personal y la

geolocalización, siempre se debe tener presente que los datos provienen de personas y que aunque los datos no tienen dignidad, estos datos son de una persona que cuenta con ella.

- **ii** La privacidad es más que un valor: Si la intención fuera tomar en cuenta solo lo “Publico” o “Privado” se realizaría una reducción de un espectro muy complejo. Privado va más allá de decir que no es público y público no quiere decir que es cualquier dato, por eso se tiene que tomar en cuenta diferentes espectros de datos: ¿Cuáles?, ¿Quiénes?, ¿Para qué?, ¿Por qué? etc.
- **iii** Protegerse contra la re-identificación de datos: Es uno de los puntos más complejos desde el punto de vista técnico, se sabe que los metadatos ya son una realidad cotidiana como las redes sociales, correos entre otros, dan aviso de que son usados para brindar un mejor servicio, de igual manera existe la posibilidad de que se disgregan y se encripten para la protección de los mismos.
- **iv** Intercambio Ético de datos: El acceso democrático sería un buen paso a seguir, obtener los consentimientos, no vender o transferir las bases de datos entre particulares, este sería un buen inicio para establecer una base.
- **v** Fortalezas y limitaciones de Datos: Big no es sinónimo de automáticamente será mejor, los datos deben ser completos para el fin deseado, además correctos para hacer asociaciones y predicciones correctas.
- **vi** Debatir las decisiones de ética difíciles: Hay muchas decisiones que deben ser totalmente plurales y democráticas, esto implica a todos al igual que el Big Data, ¿Cómo será la intervención del Big Data a nivel global?, ¿Qué papel tendrían los comités de bioética?, ¿Cuál y cómo será el marco jurídico?, ¿Cómo manejar y explicar conflictos de interés con base en el manejo de datos? etc.
- **vii** Crear un código ya sea de conducta, de industria o de comunidad de investigación: Los códigos o reglas no son mágicos, pero si todos estamos dispuestos a seguirlos tienen más fuerza que si se imponen por alguna autoridad.
- **viii** Crear diseños de datos y sistemas que sean fáciles de auditar: La transparencia de la ética y lo legal deberá ser lo primero que empecemos a construir ante el Big Data, la transparencia no solo debe ser en los datos que se

estén manejando, sino que también en las personas que los manejan, esto debe ser durante todo el proceso no solo en el consentimiento, todo de modo paralelo con la farmacéutica, aseguradoras, gobiernos entre otros.

- **ix** Comprometerse con las grandes consecuencias de los datos y las prácticas de análisis: Una persona a la que se le ha dado seguimiento desde la ecografía prenatal, el preescolar, primaria, pubertad, adolescencia, redes sociales y más, la cuestión será ¿en qué momento se le preguntará por su asentimiento?, o si sus padres son los que darán el consentimiento y por qué o si no están de acuerdo ¿se tendría que destruir la información?.
- **x** Saber cuándo romper reglas: Aunque parece contradictorio, se debe estar abiertos a todas las posibilidades, a cambiar alguna de las propuestas del decálogo, a eliminar o incluir alguna, finalmente el objetivo es no perder de vista que la intención es la protección de datos personales.

Se debe estar consciente de todos los riesgos que la tecnología implica en la intimidad y confidencialidad de datos personales puesto que actualmente no se tiene un marco normativo que ampare.

8.2.9. Herramientas para el manejo de Datos Masivos

Los proyectos que aplican el análisis masivo de información exigen exorbitantes capacidades para almacenar, procesar, distribuir y visualizar información, por esta cuestión se propone el proyecto de unificación de contenidos generales en sector salud, puesto que el crecimiento de datos en estos campos es exponencial y como es bien sabido el sector salud ya cuenta con una base de datos, pero esta no está relacionada con otras fuentes externas.

Por tal motivo se necesita elegir las herramientas a implementar según las necesidades del sector, en este caso la taxonomía Big Data existen herramientas de recolección, de almacenamiento, de procesamiento y de visualización.

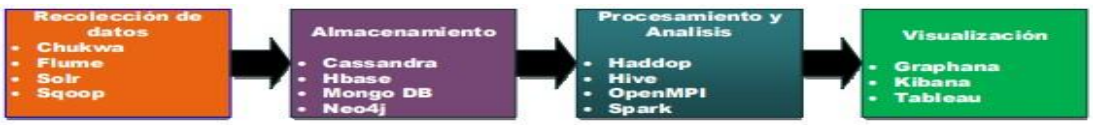


Figura 6. Taxonomía Big Data.

Fuente: Elaboración propia

8.2.9.1. Herramientas de Recolección

Herramienta de Recolección como **Chukwa** está diseñada para la recolección y análisis a gran escala de “logs” (grabación secuencial de base de datos de todos aquellos sucesos que afectan directamente a un proceso particular) de grandes sistemas distribuidos. Además de contar con las herramientas para mostrar los resultados del análisis y monitoreo, esta herramienta es guiada por (HDFS), este es el sistema de archivos distribuidos de Hadoop, heredando su escalabilidad y robustez, cabe mencionar que Apache Chukwa también ofrece herramientas de monitoreo y análisis de resultados. Incluye y se divide en agentes que son ejecutados en cada máquina para la recolección de datos; colectores que reciben los datos de los agentes y los almacena; procesos ETL (*Extract, Transform, Load*) que realizan la extracción, transformación y carga de información de los colectores y los HICC que permiten la visualización y monitoreo de toda la información es decir panel central.

Flume es una Herramienta Distribuida para recolección, agregación y transmisión de ingentes volúmenes de datos de diferentes fuentes, tiene arquitectura basada en la transmisión de datos altamente flexibles y configurable que permite adaptarse a cualquier tipo de situación: monitoreo de logs, descarga de datos de redes sociales o de correos electrónicos entre otros, además el destino de los datos también se puede configurar es decir no limitarlos a ser utilizados solo por Hadoop; su arquitectura se compone de:

- Evento: cantidad de datos que Flume transporta.
- Flujo: flujo de datos desde el origen al destino.
- Cliente: se ejecuta en punto de origen y entrega eventos del origen al agente Flume.
- Agenten: reenvía eventos.
- Source: absorbe los eventos entregados y decide a que canal enviarlos.
- Channel: almacena temporalmente los eventos para garantizar la durabilidad de flujos.

- Sink: elimina eventos del canal, los transmite al siguiente agente o los envía a la finalización del evento. En la figura 6 se identifican los elementos citados de Flume.

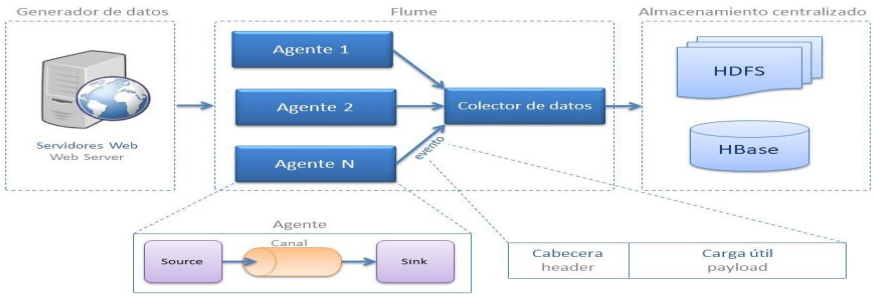


Figura 6. Apache Flume. (Calvo, 2018)

8.2.9.2. Herramientas de Almacenamiento de Datos Masivos

Cassandra: inicialmente planeado para el proyecto de Facebook, creado a través de apache es un BD no relacional distribuida (NoSQL) y basada en modelos de almacenamiento, tiene una gran disponibilidad de los datos además de ser tolerante a fallos, no exige equipos costosos (hardware) para manipularlos y guardarlos en la nube.

Cassandra está desarrollado en Java, es de código abierto, funciona de forma dinámica con implementaciones de código cerrado, su arquitectura se basa en peer to peer (P2P), es decir, todos los nodos tienen la misma importancia jerárquica, ningún nodo tienen roles diferentes y son distribuidos de la misma manera para evitar fallos, así se garantiza la ubicación y el estado de la información, esta información está disponibles en cualquier momento y garantiza un alto nivel de sincronización. La imagen 7 resume las características de Cassandra como manejador de base de datos del tipo NoSQL.

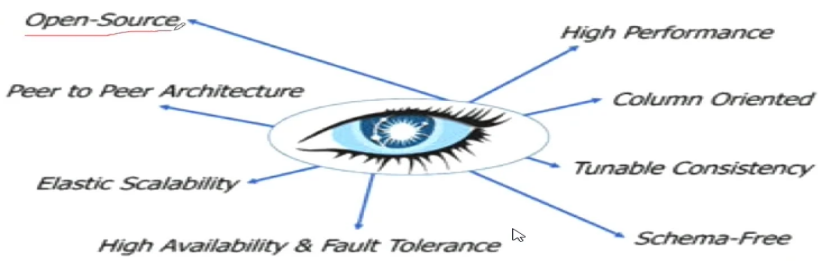


Figura 7.Características de Cassandra (Requena Mesa, 2019).

Apache HBase: base de datos no SQL de tipo columna es ejecutado en HDFS, es un sistema de gestión de bases de datos de correlación ordenada, distribuido y vacío que se ejecuta en un sistema de archivos distribuidos y está compuesto por:

- Servidor Maestro: administra las regiones y equilibra las cargas.
- Servidor de Región: realizan el trabajo y establecen la comunicación con el cliente, quien consulta la información.
- HFiles: representación física de datos en HBase, una característica es que no se habla de tipos de datos todo el almacenamiento es en Bytes.

A continuación, se muestra una tabla con las diferentes características de Cassandra, Big Table. En la imagen de la figura 9 se muestra un resumen del comparativo entre las tres herramientas mencionadas.

Características	*Cassandra*	*Big Table*	*HBase*
Tipo de licencia	Apache Open Source	Propietaria	Apache Open Source
Lenguaje de implementación	Java	C C++	Java
Ejemplos de uso	Banca, las finanzas	Gmail, Búsqueda de libros de Google	Mensajería, Sistemas de Monitoreo
Mejores usos	Escribir con frecuencia, leer menos	Diseñado para escalar a través de cientos o miles de máquinas	Lectura aleatoria al escribir en bases de datos grandes
Control de concurrencia	Multiversión de concurrencia y control	Logueado	Logueado
Transacciones	Local	Local	Local
Almacenamiento de datos	Disco	GFS	Hadoop
Gestión de Series de tiempo	Es un excelente complemento para los datos de series de tiempo, y es ampliamente utilizado para almacenar las métricas de rendimiento, seguimiento de flota, datos de sensores, registros, datos financieros.		Las consultas sobre esta serie de tiempo podrían recuperar un intervalo de versiones. Varias aplicaciones de monitoreo se han desarrollado sobre esta plataforma como OpenTSDB.
Disponibilidad	Alta disponibilidad	Alta disponibilidad	La disponibilidad es aceptable, han priorizado elevar los niveles de consistencia.
Tolerancia a Fallas	Elevada	Elevada	Elevada
Consistencia	Consistencia aceptable, priorizaron elevar la disponibilidad.	Alta consistencia	Alta consistencia.
Complejidad	Fácil instalación y configuración. Estructura familiar.	De compleja instalación, configuración y administración	De compleja instalación, configuración y administración. Debe dominarse los sistemas Hadoop y HDFS.
Documentación y Comunidad	Muy buena documentación y comunidad de desarrollo activa	Excelente documentación.	Excelente documentación y comunidad de desarrollo.

Figura 8. Comparativo Cassandra vs Big Table vs HBase

Fuente: Elaboración propia

Mongo BD: brinda escalabilidad, rendimiento tanto en escritura como en lectura y elevada disponibilidad, se basa en colecciones y documentos, tiene un conjunto de archivos existentes en un solo servidor. Las colecciones son un conjunto de documentos de base de datos, esta colección está dentro de una única base de datos y dentro de la colección hay datos variados, aunque con el mismo objetivo. Los documentos tienen designado clave –valor, está relacionado con esquemas dinámicos y permite tener diferentes estructuras en una sola colección y estos a su vez pueden tener diferentes tipos de datos.

Neo4j: es una base de datos orientada a grafos, especialmente utilizada para los datos no estructurados, esta base de datos es de las más utilizadas en Big Data, actualmente es utilizada para detectar fraudes en el sector bancario. Utiliza grafos para la representación de datos y la relación de ellos, para poder llevarlo a cabo hace uso de distintos tipos de grafos: grafos no dirigidos: que sus nodos y relaciones son intercambiables; grafos dirigidos: sus nodos y relaciones no son bidireccionales; grafos con peso en donde a las relaciones entre nodos se le asigna valor y después realiza operaciones; grafos etiquetados, definen los vértices con base en sus etiquetas asignadas; grafos con propiedad: tiene peso y etiquetas en que se asignan propiedades. La siguiente tabla resume las características y Beneficios de Neo4j

Tabla 2

Características y Beneficios Neo4j.

Característica	Beneficio
Open Source	Utilizada por miles de organizaciones
Interfaz Amigable	Es la mayor comunidad y más activa
Modelamiento de datos fácil	Brinda mayor rendimiento de lectura y escritura, ofrece alto rendimiento y velocidad sin desproteger la integridad de los datos
Consultas legibles	Alto rendimiento en procesamiento y almacenamiento, incluso con el crecimiento de datos.
Alto rendimiento	Fácil de aprender y dominar

Hive: inicialmente creada para hacer Hadoop más fácil de operar y administrar ingentes cantidades de datos almacenados en un ambiente distribuido, tiene un lenguaje similar al de SQL llamado Query Language (HQL), estas sentencias HQL son separadas por un servicio de Hive y enviadas a un proceso MapReduce ejecutados por el cluster de Hadoop, esta infraestructura funciona en la parte superior a Hadoop y está optimizado para realizar lecturas de bases de datos, la siguiente figura 10 muestra su arquitectura.

Jaql: fue donado por IBM a la comunidad de software libre, tiene como objetivo que el desarrollador de aplicaciones Hadoop se centre en que quiere obtener y no como obtenerlo, analiza la lógica y la distribuye en mappers y reducers solo cuando es y donde sea necesario, combina la facilidad de uso de un lenguaje de alto nivel con la capacidad de paralelismo y rendimiento de Hadoop.

Jaql tiene una infraestructura flexible para administrar y analizar datos semiestructurados como por ejemplo XML y CSV entre otros.

Oozie: permite ejecutar un conjunto de aplicaciones de Hadoop, en secuencia específica comúnmente conocido como flujo de trabajo, es de código abierto, simplifica el flujo de trabajo y la coordinación entre cada proceso, permite que el usuario defina acciones y las dependencias entre las mismas.

Redis: es de código abierto, es compatible con estructuras de datos como hilos, hash, conjuntos ordenados y de consultas de rango, mapas de bits y de índices geoespaciales con consultas de radio, es un motor de bases de datos en memoria con base en el almacenamiento en tablas de hashes pero si así es decidido puede usarcé como un base de datos persistente y durable.

8.2.9.3. Herramientas de Procesamiento

Pig: permite a usuarios Hadoop analizar conjuntos de datos y enfocarse menos a crear programas MapReduce, el lenguaje PigLatin maneja cualquier tipo de dato y Pig es el ambiente donde se ejecutan los programas, proceso parecido entre la máquina virtual java y una aplicación java.

Spark: permite realizar trabajos en memoria para optimizar tiempo de procesamiento proporcionando ingentes cantidades de datos en poco tiempo, utiliza algoritmos iterativos, nos proporciona API para Python y R.

Zookeeper: es de código abierto, asegura que toda operación de actualización respete el orden en el que fueron enviadas, los resultados son fallas o éxitos, el cliente ve el mismo servicio independiente de que servidor los contenga, al realizar una actualización la modificación se mantiene hasta que el cliente la sobre escriba, el cliente tiene garantizado que los servicios están actualizados.

Hadoop: está inspirado en la programación de MapReduce, consiste en separar en dos tareas mapper y reducer, para manipular datos distribuidos a nodos de un clúster y lograr paralelismo en el procesamiento, Hadoop está compuesto por tres partes:

Hadoop Distributed File System (HDFS) los datos del clúster de Hadoop son divididos en bloques y distribuidas por medio del clúster, así las funciones map y reduce se ejecutan en pequeños subconjuntos para brindar escalabilidad de ingentes volúmenes de datos, los bloques de datos hacia HDFS se almacenan en más de una ocasión y se almacenan en diferente rack para lograr la redundancia.

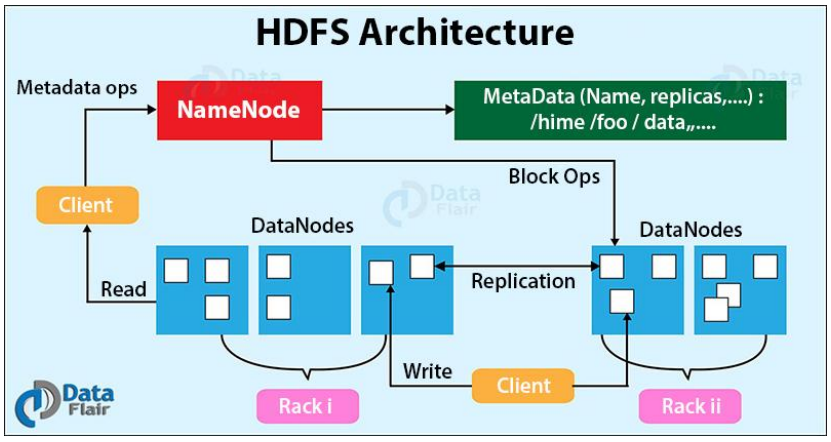


Figura 9. Arquitectura Haddop HDFS. (Data FLair, 2020)

Hadoop MapReduce: núcleo de Hadoop, son dos procesos, map toma un conjunto de datos y lo transforma en otro conjunto y los elementos individuales

se dividen en pares de llave – valor; reduce toma la salida de map como datos de entrada y combina los pares en grupos más pequeños.

En la parte intermedia Shuffle se obtienen los pares del proceso map para elegir el nodo que procesara esos datos dirigiendo la salida a un reduce en específico.

Es un Framework que brinda un sistema de procesamiento de datos de manar paralela y distribuida, tiene como objetivo resolver problemas utilizando grandes volúmenes de datos de manera paralela utilizando archivos HDFS, el framework cuenta con una arquitectura que cuenta con un servidor maestro y varios esclavos, es un servidor esclavo para cada nodo del clúster, el framework se dio por la necesidad de mejorar el procesamiento de datos entre las maquinas locales y exteriores en una red con la finalidad de reducir el tiempo de procesamiento de una tarea para la red, es un framework crea objetivos específicos es uno de los más utilizados en el mundo del Big Data.

Hadoop Common Components: son las librerías de soporte para proyectos Hadoop, contiene los archivos .jar y todos los scripts para poder ejecutar Hadoop, proporciona código fuente también, documentación y una muestra de proyectos de la comunidad Hadoop.

8.2.9.4. Herramientas de Visualización

Tableau: es una herramienta gratuita, es un modelo de análisis visual que permite hacer consultas a bases de datos de manera muy fácil además de intuitiva, su principal beneficio es que la interfaz de usuario es muy amigable con el cliente, la velocidad y calidad de los grafos son excelentes tanto en computadoras como en dispositivos móviles además facilita informes sencillos y objetivos. La figura 11 identifica una salida de información visual de Tableau.

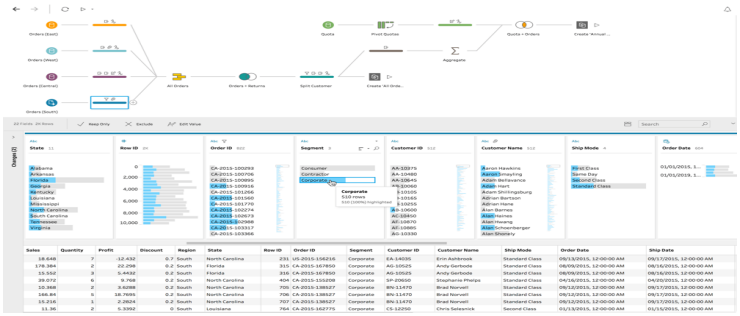


Figura 11. Tableau como herramienta para visualziicón de datos (Smart Solutions International, 2018)

8.3. Desarrollo

En este apartado, se describe la propuesta que se hace con nombre Unificación de Contenidos Generales en el Sector Salud, con la intención de dar solución a muchas de las necesidades de las personas que acuden a atenderse en diferentes instituciones de Salud, y así facilitar el trabajo de los prestadores del servicio para mejorar la calidad en la atención al paciente.

8.3.1. Unificación de contenidos generales en Sector Salud

En la actualidad ya existen países donde todos los datos médicos que se capturan desde cualquier institución de salud ya sean públicas, privadas o desde dispositivos móviles u otros dispositivos registrados están unificados, es decir, que se almacenan en una sola base de datos y pueden ser requeridos desde cualquier institución con la seguridad de que los datos están actualizados y completos independientemente de donde se realizó la actualización, en este caso de donde se atendió a la persona.

Un claro ejemplo es el proyecto MIDAS, este crea aplicaciones Big Data en Europa en conjunto con seis sistemas sanitarios: Euskadi, Inglaterra, Irlanda, Irlanda del Norte, Finlandia y E.U.A.

Gracias a este proyecto se realizan previsiones para abordar temas epidemiológicos como lo es la obesidad o el alcoholismo entre otros, además otro de sus beneficios es que podrá ser utilizado por otros sistemas de salud.

MIDAS será el primero de los proyectos que utiliza tecnologías de extracción de información desde diversas fuentes para después realizar un análisis y que las Instituciones de Salud tomen las mejores decisiones, siempre cumpliendo con los más altos estándares en la protección de datos y ética.

Actualmente existen muchas deficiencias en el Sector Salud siendo así de los sectores que mayores beneficios obtiene de Big Data, lamentablemente en nuestro país falta mucho para obtener los beneficios de los que anteriormente ya se han mencionado.

Dada las deficiencias que existen en las aplicaciones e interfaces para la captura y almacenamiento de datos en el sector salud, la propuesta consiste en que todos los datos del sector salud sean capturados, almacenados y accesibles desde un solo sistema de información de datos masivos.

Que cualquier institución relacionada con la salud tenga la obligación de dar de alta el sistema de análisis de datos masivos

Los procesos de atención en el sector salud actuales son deficientes, gracias al análisis de datos masivos, se pudiera contar con las herramientas necesarias para gestionar, almacenar, presentar y administrar los datos médicos de manera personalizada y a gran velocidad.

La perspectiva es que toda la información esté integrada y sea accesible desde cualquier punto, que los todos los documentos al igual que el historial clínico se encuentren en formatos digitales, que todo esté en la misma plataforma y así lograr la comunicación entre todas las instituciones relacionadas con la salud.

Para poder implementar la propuesta es necesario tomar en cuenta cuatro puntos importantes:

- Integración de los datos Médicos
- Gestión de Flujos de trabajo
- Visor Médico
- Archivo Médico

En el primer punto la integración de los datos se trata de armonizar los datos independientemente de la fuente o si son estructurados o no estructurados, el sistema de gestión de datos que se muestra en la imagen 14, lo que permite es que bioseñales, videos, imágenes los convierte en un mismo formato, esta es la base la unificación de contenidos generales en el sector salud.

La integración es tanto clínica como técnica, puesto que solo una base de datos necesita ser administrada.

En el segundo punto gestión de flujo de trabajo, una vez que los datos ya fueron almacenados en el sistema de análisis masivo, la siguiente tarea es obtener el máximo valor de esos datos, este punto garantiza la unificación de los datos.

El tercer punto visor médico, se necesita un visor multiformato para visualizar los datos que están en los más diversos formatos y que puedan ser convertidos a un formato estándar, considera también la visualización simultánea de imágenes radiológicas, resonancias entre otras.

Archivo médico es el punto número cuatro, los profesionales de la salud son los más beneficiados con estos archivos, gracias a este sistema de análisis se reduce la estructura de datos administrativos y los costes para el mantenimiento, otro beneficio es la liberación de espacio por el manejo de formatos estándar.

Tanto el sistema como la arquitectura deben diseñarse para ser escalables ya sea para el volumen de información ya almacenada como en consultas y actualizaciones. Se recomienda utilizar una infraestructura híbrida puesto que se tendrían servidores virtuales en entorno de cloud computing, cada institución podrá determinar cómo consume los recursos y cuantas máquinas virtuales necesita.

En el sistema de unificación de contenidos generales en el sector salud se tiene siempre presente que los datos digitales son el principal activo, estos datos deben de estar consolidados y relacionados con otros datos además de almacenados y a disposición de cualquier persona que lo requiere y desde el lugar que se requiera.

Para satisfacer esta necesidad se tiene que armonizar los datos médicos de las diferentes Instituciones con diferentes formatos con la finalidad de que se puedan visualizar por un único sistema (Unificación de Contenidos Generales en el Sector Salud), una vez que se realice lo anterior los datos estarán disponibles de manera interna por alguna institución específica y en la nube de manera general, además de tener toda esa información a la mano para su tratamiento continuo, así se lograría un ahorro considerable en tiempo y dinero y al mismo tiempo se facilitaría la relación y colaboración entre instituciones y profesionales de la salud, se mejoraría la calidad de la atención al cliente, no se duplicarían resultados o en caso de ya tener resultados de algún análisis no sería necesario realizarlo nuevamente si la personas es atendida en otra institución o por otro profesional.

8.3.2. Implementación de un Sistema para unificar datos generales en Sector Salud Nacional

La propuesta de trabajo se centra en unificar las datos en el sector salud, para brindar atención de calidad a las personas en tiempo mínimo y facilitar tratamientos a la medida, tomando en cuenta que los datos no estructurados son los más difíciles de unificar, esta sería la área de oportunidad a mejorar para poder continuar el proceso de unificación de datos del sector salud, actualmente hay tres áreas de investigación donde se aplican los conceptos de Big Data en datos no estructurados como lo son:

- Procesamiento de imágenes médicas: Estas imágenes son muy importantes para el diagnóstico, evaluación y planificación médica, como lo son las tomografías, resonancias, radiografías, mamografías, ultrasonidos y tiene como objetivo hacer su interpretación más fácil y rápida.

  La integración del análisis de datos masivos tiene el potencial para mejorar el diagnóstico y reduce considerablemente el tiempo para poder proporcionarlo.

- Señales médicas: Son muy parecidas a las imágenes médicas, el análisis de señales médicas fisiológicas, tiene un grado más significativo especialmente a la hora de adquirirlos y almacenarlos en alta resolución desde una infinidad de

monitores conectados al paciente, debido a esto deben integrarse constantemente a sistemas y análisis predictivos para asegurar su eficiencia.

Los dispositivos de monitoreo fisiológico generan y almacenan infinidad de datos de manera continua, pero por un breve periodo, con el sistema de unificación de contenidos generales en el sector salud, se almacenarían para su análisis constante y se relacionaría con los registros médicos existentes para un mejor diagnóstico y atención.

- Genómica: Hoy en día se tiene se ejecutan iniciativas para integrar datos genómicos, con el objetivo de brindar atención personalizada, para esto es necesario una respuesta rápida del análisis Big Data además de que los datos deben ser fiables.

8.3.3. Pasos a seguir para aplicar el sistema de unificación

Lo primero que se tiene que hacer para poder implementar la propuesta planteada, se debe tomar en cuenta que el recurso más difícil con el que se va a interactuar es el recurso humano, las personas en su mayoría son renuentes a cambios, así que se debe de brindar la asesoría y capacitación necesaria a todas las personas relacionadas directa o indirectamente en el sector, a continuación se enlistan las actividades a realizar para la implementación de la propuesta:

i. Reconocer que el Sector Salud actualmente cuenta con muchas deficiencias y es necesario implementar sistemas de análisis de datos masivos debido al crecimiento exponencial de datos, con el objetivo de unificar los contenidos generales del sector salud, y así agilizar los procesos de atención que actualmente son muy lentos además de poder implementar la medicina personalizada.
ii. Definir cuáles son las instituciones de salud que serán contempladas para la implementación del proyecto, estas instituciones ya sean públicas o privadas aportarán datos para su almacenamiento y posterior análisis, datos que posteriormente se podrán consultar desde otra institución.
iii. Establecer cuáles son los dispositivos electrónicos o sensores que serán considerados para aportar información y que esta sea almacenada en el sistema

de unificación, en la imagen 6 de dispositivos que se pudieran utilizar en el Sector Salud

iv. Definir cuáles herramientas de análisis de datos masivos de información se van a utilizar, esto aplica para el desarrollo del sistema de unificación de contenidos generales en el sector salud.

8.3.4. Descripción de los pasos a seguir para la aplicación del sistema de unificación

En el sector salud actualmente ya se cuenta con una base de datos de los pacientes, cabe mencionar que su infraestructura es deficiente, para que una persona sea atendida tiene que esperar un largo periodo de tiempo, una vez que llega con su médico a través el sondeo se realiza un diagnóstico y de ser necesario se realiza una solicitud de análisis clínicos que tardan a un más tiempo y posteriormente ya con los resultados necesita otra consulta para que le prescriban un tratamiento.

Si esta persona acude con otro médico a otra institución por una segunda opinión, se tendrá que realizar el proceso de atención nuevamente, sondeo, análisis y tratamiento, cuáles serían los resultados si los datos de este paciente ya estuvieran unificados en una sola plataforma y que gracias al análisis de datos masivo se obtiene un análisis predictivo para poder prescribir un tratamiento a la medida según el padecimiento, la calidad de atención en el sector salud aumentaría además de los tiempos de espera se reducen considerablemente y el ahorro económico seria también considerable, de aquí la necesidad y la importancia de la unificación de datos a través del uso del análisis masivo de información.

Para poder determinar cuáles instituciones están relacionadas son la salud, se tiene que tener un acercamiento con sectores de gobierno que faciliten el dato, cuántas y cuáles son las instituciones o empresas relacionadas con el rubro para tomarlas en cuenta en el momento de la implementación del sistema de unificación y así de esta manera los datos ya existentes y los nuevos datos sean almacenados en la misma base de datos y así poder ser consultados a través de la misma

plataforma con la seguridad de tener información actualizada y confiable sin importar el lugar donde requieren esa información y en tiempo reducido.

Es necesario llevar un registro de todos los dispositivos que arrojan datos médicos y que estén enviando datos constantemente de manera remota para alimentar el sistema de integración, en caso de que el portador necesite atención médica los datos ya estén disponibles para un diagnóstico casi inmediato y se le pueda otorgar tratamiento a la medida.

Para elegir las herramientas de análisis de datos masivos es necesario realizar una investigación y análisis de las herramientas ya existentes y cuáles son las adecuadas para el proyecto de unificación y diseñar el modelo que se pueda implementar, como se van a obtener los datos, como se van a almacenar, procesar y visualizar.

También será necesario la elaboración de manuales y capacitación del personal por medio de manuales impresos y digitales cursos presenciales antes de la implementación del sistema de unificación.

8.3.5. Análisis de la propuesta

Se debe tener siempre claro qué es lo que se quiere lograr con la implementación de tecnologías o herramientas de análisis de datos en una institución.

Debido al creciente número de población, a la saturación en salas de emergencia y hospitalización, en muchos de los casos, la atención médica de los pacientes se brinda por medio de evaluaciones superficiales, lo que puede ocasionar un diagnóstico erróneo y que posteriormente esa persona regrese con un problema de salud mayor, esto se puede volver repetitivo.

En la consulta médica se puede archivar el expediente del paciente, pero no sirve de mucho esa información para el análisis de los datos pues es un método obsoleto que no ayuda a analizar los datos de los pacientes además no permite realizar búsquedas dentro de esa información para poder brindar atención personalizada.

Con la información recabada y a los conocimientos adquiridos se puede decir que al analizar la propuesta del sistema de Unificación de contenidos generales el autor se percata que el primer paso para obtener una atención médica personalizada es la unificación de datos médicos y que se debe resolver la situación que actualmente se presenta.

Hoy en día el sector salud, aunque es de los más beneficiados al aplicar las tecnologías de Big Data también es donde menos apoyo económico se da para la implementación de la misma, es de suma importancia que se realicen grandes inversiones para poder contar con la arquitectura y la tecnología necesaria para que la aplicación de Big Data sea funcional.

Para poder aprovechar las ventajas del análisis masivo de datos, otro de los factores a los que se le debe atención es el personal con el que se debe de contar en este sector, se debe de contar con profesionales del sector salud, Ingenieros en Sistemas Computacionales, Licenciados en Informática, Licenciados en Ciencias de las Matemáticas y Estadísticas, profesionales que cuenten con los conocimientos y las habilidades necesarias para manejar las herramientas relacionadas con Big Data.

Se debe encontrar las herramientas necesarias para obtener, almacenar y analizar datos posteriormente tener ya identificadas las instituciones que serán partícipes de la implementación de herramientas y tecnologías Big Data además de estar siempre en contacto y de manera muy cercana con el sector.

No cabe duda de que si se implementan las técnicas y herramientas Big Data de manera correcta con base en las necesidades del Sector Salud se lograra llegar al objetivo fijado, en este caso, poder brindar atención médica personalizada aunado a la reducción de costes y la mejora en la toma de decisiones.

Se espera que dentro de pocos años Big Data reemplace totalmente las tecnologías aplicadas actualmente en el sector, y a sabiendas de que la aplicación del Big Data hoy en día es de manera progresiva, todavía falta mucho por hacer.

8.3.6. Resultados esperados

Al realizar este capítulo sobre el Análisis de Datos masivos en el campo de la Salud se obtienen resultados genéricos. Se destacan beneficios unificando contenidos generales en el sector salud y lo que esto conlleva a lo siguiente: minimizar el tiempo de espera para recibir atención médica, reducir el tiempo para obtener resultados de análisis clínicos, aumentar la veracidad de estos, tener un diagnóstico acertado y tratamiento personalizado asegurando la mejoría, en la salud de la persona atendida reducir la mortalidad, obtener un expediente clínico completo actualizado y disponible en cualquier momento y lugar donde se requiera para la atención médica, se prevendrán enfermedades antes de los síntomas y los costos generados en el sector y para la población en general disminuirán considerablemente, también se mejorará la relación y la cooperación entre las instituciones del sector para obtener un bien común, así como la relación médico – paciente, y la relación Sector Salud – Entidades de Gobierno.

En las instituciones de salud, las aglomeraciones de personas para recibir atención se reducen considerablemente debido a todo lo anteriormente mencionado aunado a contar con una base de datos actualizada los procesos administrativos y de atención son más ágiles pues los datos están siempre a la mano.

Se maximiza el uso de datos de manera estratégica para mejorar la toma de decisiones, ya sea en atención médica, reducción de costes o en minimizar el tiempo de respuesta en cualquier área del campo.

Gracias a la unificación de información se eliminarán los malos entendidos y se ahorrará tiempo y dinero.

Gracias a la proporción de datos emitidos por dispositivos electrónicos y sensores médicos y a las herramientas Big Data para procesarlos tenemos como resultado un seguimiento remoto a pacientes con enfermedades crónicas además del monitoreo de signos vitales de personas internadas en alguna institución médica o bien en personas de edad avanzada desde su domicilio, se envía una alerta de emergencia aumentando la eficiencia de los hospitales.

Big Data ayuda a fomentar el autocuidado obteniendo como resultado pacientes colaboradores ayudados de los dispositivos móviles enviando información de la salud actual de cada persona.

Con la implementación del sistema de Unificación de contenidos generales en el campo de la salud se obtiene una plataforma completa y flexible adaptable a otros sistemas de salud con la intención de que en tiempos no muy lejanos se pueda internacionalizar.

Crear un sistema tan eficiente que, sin importar el crecimiento exponencial de datos a analizar, procesar y visualizar, no reduzca la eficiencia ni afecte los resultados a obtener.

Se puede contar con una medicina de calidad por medio de la medicina de las 4P: predictiva, preventiva, personalizada y participativa.

Resolver la saturación de personas en las instituciones de salud por diversas situaciones relacionadas con la salud ya sean trámites administrativos o de atención médica o farmacéutica ya que la aplicación de Big Data brinda resultados mejorando todas las áreas relacionadas con el campo de la salud.

El sector salud brindará atención de excelente calidad, se adecuarán las instalaciones para la implementación y uso de herramientas Big Data necesarias.

La atención será personalizada e integral mejorando notoriamente el servicio que actualmente se brinda.

El profesional de la salud tendrá un panorama completo del expediente médico de cada paciente facilitando así el brindar un diagnóstico acertado aunado a prescribir medicamento exclusivo para cada paciente.

Gracias a la aplicación de Big Data en el campo de la salud, la formación podrá saber cuándo surtir un medicamento o cuando lo tiene que surtir puesto que ya sabrá con anticipación las necesidades del Sector Salud.

Se evitará la propagación de epidemias o en caso de que exista algún brote de enfermedades altamente infecciosas, se reducirá el tiempo de control de la misma.

Es sabido que entre más instituciones sean consideradas para la implementación del sistema de unificación de contenidos generales en el campo de la salud, la inversión será mayor, solo basta recordar que el resultado será proporcional o mayor.

Gracias al ahorro considerable por las reducciones de costes en cirugías innecesarias, medicamentos incorrectos, entre otros, los beneficios son generales para la población y el gobierno sin dejar de mencionar al sector salud.

Ser ejemplo para otras entidades y generar confianza en la aplicación del análisis de datos masivos para maximizar los beneficios de la mejor toma de decisiones.

Gracias al uso de herramientas Big Data es posible administrar y optimizar los recursos del sector y crear un plan estratégico para la mejora continua.

Los procesos que se llevan a cabo dentro de las instituciones se realizarán de manera prácticas y sencillas, la extracción de información será rápida y efectiva por medio de interfaces muy amigables con las personas que manejen la información.

La perspectiva de la gente que necesita atención médica será diferente a la actual, puesto que se generará confianza y se creará satisfacción por el servicio y la atención que brinda las instituciones del Sector Salud.

No será necesario acudir a otro estado o país para recibir atención o para obtener medicamento inexistente en la institución médica local.

El resultado de la propuesta sobre la unificación de información permite también la unificación de instituciones, en la creación de una institución homogénea, única en todo su ámbito.

Con la implementación de la propuesta, se piensa que serían pioneros en la implementación de Big Data en el campo de salud, contando con una institución que brinda servicio y atención de calidad.

Todos los datos almacenados serán manejados con total privacidad, se deberán cumplir reglas y ciertas restricciones para el manejo de información y serán protegidos contra cualquier amenaza.

El sector salud será sostenible nuevamente gracias a la aplicación de técnicas y herramientas Big Data.

El modelo de negocio del sector salud será transformado ya que actualmente se basa en el servicio que presta, entre más atención brinda mayor es su financiación, con Big Data se implementaría de forma contraria pues la remuneración a los profesionales de la salud y la financiación del sector salud se basará en el estado de salud de los pacientes o de la comunidad en general, la productividad se medirá con base en lo que no se ha gastado respecto a los recursos del sector.

Como resultado del análisis masivo de datos en el campo de la salud la atención médica no tendrá retrasos es decir: los medicamentos prescritos se brindarán en el momento, no se tendrá desabasto, el sistema para agendar citas médicas será más eficiente, no se tendrá sobresaturación de personas en las instituciones, las cirugías que sean requeridas no se podrán suspender por falta de personal médico, los resultados de análisis clínicos se darán en tiempos relativamente cortos, dependerá del tipo de análisis y número de estos que fue requerido, las vacunas recomendadas por el sector salud siempre estarán disponibles.

Se prevendrá y solucionará el problema en las personas alérgicas a componentes ambientales o a cualquier otro factor, mejorando la calidad de vida de los pacientes, que es el resultado más esperado.

Una vez aplicado el Sistema de Unificación de Contenidos Genéricos, la atención de las mujeres embarazadas será de mayor calidad, será posible prevenir

la mayoría de las complicaciones, facilitando terminar el periodo gestacional, y en cuanto nazca él bebe ya pertenecerá a una nueva generación de personas que reciben atención médica generando un historial médico desde su gestación; con los datos gestacionales también se podrá tener los medicamentos necesarios a la mano para las enfermedades a las que el bebé sea susceptible, en casos extraordinarios como ya ocurre en la actualidad realizar procedimientos quirúrgicos o prescripción de medicamentos al gestante y obtener como resultado la reduciendo el índice de mortandad en neonatos y sus madres.

El uso de dispositivos de imágenes digitales aumentará en todas las instituciones de salud ya que gracias a Big Data el análisis y la interpretación de las imágenes como radiografías, ultrasonidos, resonancias entre otros, será fácil y práctica, obteniendo más información de la que se puedes obtener con la interpretación a través del ojo humano, brindará facilidad para diagnosticar al profesional de la salud, dato que estarán almacenados para posteriores consultas.

Con el estudio del genoma humano ayudado de herramientas Big Data se logrará comprender como y porque no enfermamos, como resultado de esto podremos erradicar enfermedades alargando nuestra plaza de vida y aumentando la calidad de la misma.

Un resultado más, es el conocimiento adquirido al desarrollar este trabajo académico y la propuesta de implementación de un Sistema de Unificación de contenidos Generales en el campo de la salud, así como el análisis y las comparaciones realizadas para elegir las mejores técnicas y herramientas a aplicar para el desarrollo de las mismas, además de conocer las características y los beneficios de cada una de estas, al igual que su implementación.

Con Big Data se resuelve la necesidad creciente de las empresas para obtener herramientas capaces de procesar ingentes cantidades de datos de manera eficiente aunada a él enorme volumen de datos de muy diversas fuentes.

Como resultado de la propuesta y la implementación del sistema de Unificación de Contenidos en campo de la salud, se pretende lograr y obtener una

plataforma única y fácil de manejar por el profesional de la salud y amable con las personas que interactúan con el mismo sistema, lograr obtener un sistema que brinda accesibilidad y disponibilidad de la información y de los datos médicos en cualquier momento y lugar, generar satisfacción de los pacientes ganando la confianza de los personas en general, al brindarles las características de la medicina 4P, se piensa que el apoyo gubernamental hace falta para para que el sector se siga actualizando en técnica y herramientas Big Data.

Conclusiones y Recomendaciones

Se concluye que las técnicas y las herramientas Big Data no solo son para analizar, procesar y visualizar grandes cantidades de datos de fuentes variadas, con la finalidad de transformar datos en información y a su vez la información en conocimientos.

Es posible afirmar que la implementación del análisis masivo de datos surge por la necesidad de organizaciones para contar con herramientas y técnicas necesarias para manejar un volumen de datos que crece de manera exponencial, que con las herramientas actuales no se puede realizar eficientemente.

Big Data es una tecnología que se está aplicando poco a poco, aun así, los beneficios que esta ofrece es el uso de la analítica predictiva y la analítica avanzada para la gestión de datos masivos en tiempos reducidos utilizando mínimos recursos, en especial los datos no estructurados y así poder contar con modelos predictivos para mejorar la toma de decisiones.

Toda aquella empresa que aplique el análisis masivo de datos obtiene una gran ventaja competitiva por todos los beneficios que conlleva su aplicación, no es para sorprenderse que Big Data dentro de poco tiempo sustituya la tecnología actual de la mano con Business Intelligence para el desarrollo de soluciones complementarias, Big Data se centra en el procesamiento de datos masivos de volumen exponencial , incluidos los datos no estructurados y Business Intelligence el análisis avanzado de la información para crear una interface amigable con el usuario atractiva a la vista.

La aplicación del análisis masivo de datos es una muy buena opción para que el campo de la salud sea sustentable, ya que actualmente se les brinda presupuestos ajustados, un bajo porcentaje de dicho presupuesto es destinado a la implementación de nuevas tecnologías incluyendo al software y hardware necesario para su aplicación.

Hay que ser conscientes de que es necesario contar con un gran equipo de trabajo para poder implementar un proyecto Big Data, profesionales en el campo de la salud, que conozcan las áreas y disciplinas de informática, sistemas, matemáticas y estadística entre otros.

Big Data tiene el papel más importante en la transformación en el Campo de Salud hacia la medicina del futuro mejor conocida como la medicina de las 4P: predictiva, preventiva, personalizada y participativa; además de fomentar el autocuidado por medio de la participación de los pacientes y demás personas, recibiendo los datos que son recolectados de los dispositivos electrónicos y sensores, dándoles a conocer para que las personas estén enteradas de su estado de salud actual y puedan actuar con tiempo antes de cualquier complicación en su salud.

La base del Big Data es la digitalización de todos los datos existentes de manera física como las recetas, radiografías, resultados de estudios clínicos etc., datos estructurados y no estructurados también deben de estar armonizados.

Gracias a la digitalización de datos no estructurados los profesionales de la salud pueden brindar un diagnóstico atinado, anteriormente se brindaba un diagnóstico con base en una revisión superficial y a lo que el paciente transmitía, aumentando el porcentaje de fracaso en los tratamientos.

Es de vital importancia adaptar las leyes referentes al uso de la tecnología especialmente para la protección de los datos en conjunto con el sector salud.

Para que exista un verdadero ecosistema Big Data es necesario que las instituciones de salud pública o privada se integren con tecnología, política, infraestructura y la cultura.

La percepción de las personas es muy importante para la aceptación de la implementación del sistema de unificación de contenidos generales en el campo de la salud.

Big Data cuenta con herramientas que permiten optimizar recursos, así como herramientas que ayudan a disminuir el impacto al momento de implementar la infraestructura de Big Data.

Los costes del campo de la salud se reducen considerablemente gracias a la medicina 4P.

En este capítulo, se dio a conocer conceptos relacionados con el análisis de datos masivos, los diferentes tipos de datos, de donde provienen esos datos; las herramientas y técnicas más recomendadas para la aplicación de Big Data, la analítica predictiva y se mencionaron algunos tipos de algoritmos; el impacto y beneficio de la aplicación de esta nueva tecnología y se plantea una propuesta de implementación utilizando Big Data, esta propuesta se trata de unificar todos los datos relacionados con el sector salud con el objetivo de brindar atención médica de calidad.

Como conclusión el beneficio para la comunidad en general es muy basto ya que el buen estado de salud de la ciudadanía demuestra que hay un alto grado de calidad de vida y por consecuencia aumenta el entorno productivo aunado a los incrementos de ingresos económicos por ende se da la reducción de las desigualdades sociales, se reduce la urgencia de aumento del presupuesto al sector salud por la disminución de la demanda de atención médica, con base en esto obtenemos un sistema de salud eficaz que conlleva al crecimiento económico y a la creación de una sociedad satisfecha con respecto al campo de la salud aumentando la calidad de vida en general.

En el caso de los beneficios brindados al Instituto Tecnológico de Durango, es principalmente la aportación de ideas para solucionar problemas reales apoyados de los conocimientos y experiencias adquiridas en esta institución,

logrando el reconocimiento de la misma por las aportaciones de mejora en la comunidad.

Con respecto a la propuesta se recomienda lo siguiente:

- Se debe tener claro qué tipo de documento se quiere realizar y los resultados que se desean obtener.
- Tener la idea clara de la propuesta a desarrollar.
- Adentrarse del tema a tratar y de la tecnología a utilizar para su desarrollo.
- Contar con el conocimiento básico como mínimo de las técnicas y herramientas a implementar.
- Conocer alcances y limitaciones que se puedan presentar al momento del desarrollo e implementación de la propuesta ya anteriormente mencionada.
- Conocer cuales con los recursos económicos con los que se cuenta, la inversión que se va a realizar y tener presente que la remuneración no es inmediata, pero sí es segura.
- Tener en mente el equipo de trabajo con el que se va a contar para liderar el proyecto.
- Conocer cuáles son los bienes tecnológicos con los que cuenta.
- Tener en cuenta cual es la relación de la institución con las personas que van a recibir el beneficio de la aplicación del análisis masivo de datos.
- Saber cuál es la perspectiva del sector de las personas que van a acudir a atención médica.
- Contemplar los tiempos de análisis, desarrollo e implementación del proyecto.

Referencias

Biedma Ferrer, J. M., & Bourret, C. (2019). La Potencialidad del Big Data en el ámbito sanitario. Especial referencia al caso Español. *Revista de Economía & Administración, Vol. 16 No. 2. Julio - Diciembre de 2019*, 93-109.

Calvo, D. (2018, 07 02). *Apache Flume.* Retrieved from Apache Flume: https://www.diegocalvo.es/flume/

Facultad de Estudios Estadísticos. . (s.f., s.f. s.f.). *Universidad Complutense de Madrid*. Retrieved from ¿QUÉ ES BIG DATA?: https://www.masterbigdataucm.com/que-es-big-data/

Gutiérrez Martínez, J. A., & Febles Estrada, A. (2019). Hacia la medicina del 2030. *UCE Ciencia. Revista de postgrado. Vol. 7(1), 2019*, 1-9.

IDC Salud. (2013, 12 05). *Big Data y el sector salud.* Retrieved from El Ecosistema y distintas fuentes de información que se pueden integrar y aprovechar: https://es.slideshare.net/BEEVA_es/big-data-y-el-sector-salud

Joyanes Aguilar, L. (2013). *Big Data, Análisis de grandes volúmenes de datos en organizaciones.* México: Alfaomega Grupo Editor.

Martínez Velasco, V. (2019, 06 01). Analítica predictiva en la toma de decisiones de la ingeniería. Ejemplos de aplicaciones. *Analítica predictiva en la toma de decisiones de la ingenieria. ejemplos de aplicaciones.* Cataluña, Barcelona, España: Universidad Politécnica de Cataluña.

Montealegre Gallo, A. C. (2017, 09 01). *IMPORTANCIA DE LA SOLUCION BIG DATA EN LA APLICACIÓN DE MOVILIDAD UBER MOVEMENT*. Retrieved from Universidad Libre Colombia: https://repository.unilibre.edu.co/bitstream/handle/10901/11203/MONOGRAFIA%20DIPLOMADO%20BIG%20DATA%20Camilo%20Montealegre.pdf?sequence=1&isAllowed=y

QuestionPro. (2019, 06 02). *Qué e BigData.* Retrieved from Uso de Big Data en diferentes industrias: https://www.questionpro.com/es/que-es-big-data.html

Rayo, Á. M. (2017, 05 15). *Computer Training by Netmind*. Retrieved from Tipos de datos en Big Data: clasificación por categoría y por origen: https://www.bit.es/knowledge-center/tipos-de-datos-en-big-data/

Requena Mesa, A. (2019, 06 17). *Apache Cassandra.* Retrieved from Qué es Apache Cassandra: https://openwebinars.net/blog/que-es-apache-cassandra/

Sánchez Villaseñor, O. (2019, 04 01). HERRAMIENTAS, RETOS, OPORTUNIDADES, SEGURIDAD Y TENDENCIAS DEL BIG DATA. *Tesina para obtener el título de INGENIERO EN COMPUTACIÓN*. Toluca , Estadode México, México: UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MÉXICO. FACULTAD DE INGENIERÍA.

Vidal Ledo, M. J., Morales Suárez, I. d., Menéndez Bravo, J. A., González Cárdena, L. T., & Portuondo Sao, M. (2020). Medicina de precisión personalizada. *Revista Cubana Educación Médica Superior. 2020*, 15.

Capítulo 9

Machine Learning aplicado a la salud

Sayra María Vargas Arroyo
Tecnológico Nacional de México. Instituto Tecnológico de Durango
03040118@itdurango.edu.mx

Rubén Pizarro Gurrola
Tecnológico Nacional de México. Instituto Tecnológico de Durango
rpizarro@itdurango.edu.mx

9.1. Introducción

En el empeño de lograr resultados coherentes en el campo de la salud, ha resultado muy útil el empleo del Aprendizaje Automático (Machine Learning), enfrentándose con el desafío de construir programas computacionales aprendan con la experiencia y de los conocimientos previos. El sistema de aprendizaje inmerso en el paradigma Machine Learning, crea descripciones generales de conceptos, a partir de grandes cantidades de datos.

Este capítulo da a conocer conceptos de Machine Learning, sus técnicas, algoritmos, el proceso que debe realizar para la construcción de modelos, generalidades del entorno de la programación en R y Python y las herramientas de código abierto que incluyen librerías, conociendo su estado de arte y aplicaciones en diversos sectores, se presentan modelos para lograr procesar cantidades de datos con el fin de tener una idea clara de esta tecnología.

Se dará a conocer en qué consisten estas técnicas y su grado de aplicación, haciendo hincapié su uso en el Sector Salud.

La finalidad de la investigación es describir diversos modelos basados en el uso de Aprendizaje Automático, con el objetivo de ver la actuación de estos modelos sobre casos de uso.

Una justificación importante es la oportunidad de impulsar modelos y algoritmos de Machine Learning en el campo de la medicina. El Sector Salud se convierte en un campo ideal para el empleo de las técnica inmersas en el Aprendizaje Automático que pueden permitir mejorar la capacidad de investigación clínica y dirigir de manera más precisa los diagnósticos, predecir comportamientos y toma de decisiones, aplicando para ello la integración de diferentes recursos tecnológicos.

El objetivo general es realizar un diagnóstico y propuesta general de implementación de técnicas y algoritmos de Machine Learning dentro del Sector Salud

Como objetivos específicos se pretende:

- Contextualizar aspectos de Machine Learning.
- Citar los principales algoritmos supervisados de Machine Learning.
- Identificar los principales algoritmos no supervisados de Machine Learning .
- Describir sobre los principales lenguajes de programación para implementar modelos de Machine Learning.
- Realizar propuesta en lo general del uso de Machine Learning en el Sector Salud.
- Identificar casos de uso en el sector salud

Se va a desarrollar una propuesta del uso de técnicas y algoritmos de Machine Learning, para análisis de datos con la finalidad de que se genere una idea y visión de los beneficios en el Sector Salud.

La idea de la propuesta tendrá un significado, convertir datos en conocimiento que permita obtener de aprendizaje y conocimiento para actuar desde las primeras etapas, permitiendo analizar, interpretar y predecir futuros escenarios que ayuden a prevenir y reducir la carga de enfermedades crónicas y su impacto en la sociedad y por consecuencia mejorar la salud general de la población.

El impacto y beneficio de una propuesta de esta naturaleza están en razón de mejorar diagnósticos médicos, pronósticos o estimaciones de riesgos, según sea el caso, que sirva de ayuda en la toma de decisiones acerca de los pacientes permitiendo un tratamiento más personalizado y eficaz.

9.2. Marco de referencia

En este apartado de marco de referencia se abordan conceptos y aserciones relacionados con Machine Learning, las técnicas y algoritmos supervisados y no supervisados, se presenta tablas comparativas de fortalezas y debilidades de los algoritmos, se da a conocer los lenguajes R y Python como recomendaciones para aplicar Machine Learning entre otros aspectos.

A lo largo del documento se entenderá como sinónimos la expresión de aprendizaje supervisado con Machine Learning (ML).

9.2.1. Machine Learning (ML)

La cantidad y dimensión de datos que se manejan en las ciencias médicas, hablando de atención de pacientes en hospitales e investigación hace que la identificación de relaciones y patrones sea difícil y compleja para las personas. Debido a ello, resulta necesario recurrir a computadoras para la aplicación de cálculos matemáticos complejos que permitan visualizar y entender lo que los datos reflejan de manera que se puedan extraer conclusiones fehacientes. (González Vilanova, 2019).

“Machine Learning es una disciplina dentro del campo de la inteligencia artificial centrada en el entrenamiento de modelos matemáticos a través de datos con el fin de que algoritmos que incorporen estos modelos sean capaces de realizar

tareas inteligentes, generalmente basadas en el reconocimiento de patrones multiparamétricos" (González Vilanova, 2019).

En su libro Lantz, (2013) menciona que "el campo de estudio interesado en el desarrollo de algoritmos informáticos para transformar datos en acciones inteligentes se conoce como aprendizaje automático. Este campo se originó en un entorno donde los datos disponibles, los métodos estadísticos y la potencia de cálculo evolucionaron rápida y simultáneamente." (pág. 7).

Machine Learning es la ciencia y el arte de la programación de computadoras, para que se pueda aprender de los datos (Revuelta Briz, 2018).

Es una disciplina científica que nació a partir de las investigaciones en Inteligencia Artificial (IA), confiando en los patrones y la inferencia; que se encarga de generar algoritmos que tienen la capacidad de aprender y no tener que programarlos de manera explícita, de esta manera se construyen modelos en base a ejemplos y son utilizados para tomar decisiones o hacer predicciones por medio de las instrucciones programadas. (Revuelta Briz, 2018).

Los algoritmos se alimentan con un volumen de datos para que aprenda y sepa qué hacer en cada uno de estos casos, siendo afín a los estudios en estadísticas computacionales, enfocadas en hacer predicciones. Una de las ventajas importantes del Machine Learning es la gran capacidad de adaptación en entornos cambiantes (Revuelta Briz, 2018).

9.2.2. Clasificación de algoritmos de Machine Learning

Los algoritmos de aprendizaje automático se dividen en algoritmos supervisados y algoritmos no supervisados, la siguiente tabla muestra en forma general la clasificación de los algoritmos Machine Learning.

Tabla 1.

Algoritmos de Machine Learning. Elaboración propia.

Modelos	Tareas
Algoritmos de Aprendizaje Supervisado	
• Vecino más cercano (K-Nearest Neighbors)	• Clasificación
• *Bayesianos*	• Clasificación
• Árboles de decisión	• Clasificación
• Regresión lineal	• Predicción numérica
• Árboles de regresión	• Predicción numérica
• Árboles modelo	• Predicción numérica
• Redes neuronales	• Uso dual
• Máquinas de vectores de soporte (SVM)	• Uso dual
Algoritmos de Aprendizaje No Supervisado	
• Análisis de componentes principales	• Detección de patrones
• Algoritmos de Clustering (Agrupación)	• Agrupamiento

9.2.3. Algoritmos supervisados

El aprendizaje supervisado es una de las dos ramas principales del ML. En cierto modo, es similar a cómo los humanos aprenden una nueva habilidad, alguien más nos muestra qué hacer y luego podemos aprender siguiendo su ejemplo. En el caso de los algoritmos de aprendizaje supervisado, generalmente se necesitan muchos datos que proporcionan la entrada a nuestro algoritmo y cuál debería ser la salida esperada. El algoritmo aprenderá de estos datos y luego podrá predecir la salida en función de las nuevas entradas que no haya visto antes (Bironneau & Coleman, 2019). La figura 1 visualiza que la entrada que puede ser texto, datos, imágenes como datos de entrenamiento; ingresan en un modelo basado en un algoritmo con una etiqueta de salida como posible resultado, con una nueva entrada diferente a la fuente de entrenamiento y con el modelo preparado, se genera una salida predictiva o de clasificación.

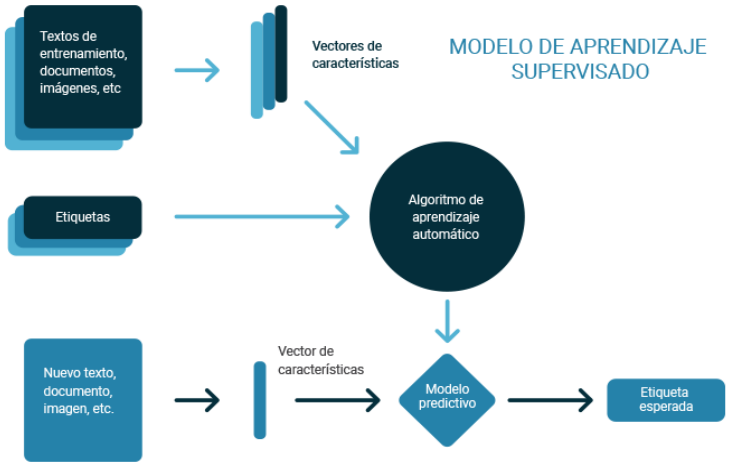


Figura 1. Modelo de aprendizaje supervisado (Luna González, 2018)

Los algoritmos más utilizados para resolver problemas con técnicas de aprendizaje supervisado son:

9.2.3.1. K-vecino más cercanos (K-Nearest Neighbors)

Este algoritmo es el más simple de todos los algoritmos de aprendizaje automático de clasificación, ofrece resultados razonablemente buenos en muchos casos en los que incluso, algoritmos más complejos se encuentran con problemas. Para construir este modelo sólo hace falta almacenar los datos del conjunto de entrenamiento y elegir cuántos vecinos se considerarán en su vecindario.

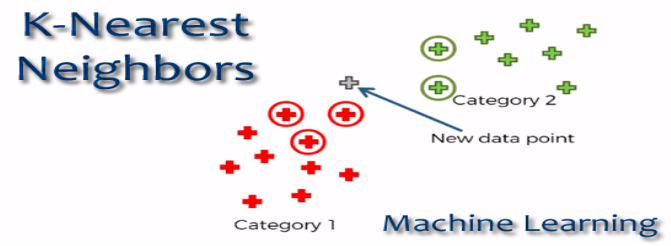


Figura 2. K-Nearest Neighbors (Avila Camacho, 2020)

A pesar de su simplicidad, este algoritmo tiene un grave inconveniente, y es que requiere que el conjunto de entrenamiento esté almacenado en todo momento, y no proporciona una solución explícita (un modelo), que pueda ser reutilizado posteriormente. En este sentido, es un representante de aproximación no paramétrica del ML.

Además, en principio es necesario escanear el conjunto entero cada vez que se clasifica un nuevo dato con el objetivo de encontrar los vecinos más cercanos, aunque para mitigar este problema se puede hacer una búsqueda aproximada, limitando el radio de búsqueda. Las fortalezas y debilidades de este algoritmo se describen en la tabla 2:

Tabla 2.

Fortalezas y Debilidades K-Nearest Neighbors. Elaboración propia.

Fortalezas	Debilidades
• Sencillo y efectivo. • No hace suposiciones sobre la distribución de datos subyacente. • Entrenamiento rápido	• Limita la capacidad de encontrar nuevas perspectivas en las relaciones entre las características. • Fase de clasificación lenta. • Requiere una gran cantidad de memoria. • Nominal características y datos faltantes requieren un procesamiento adicional.

9.2.3.2. Regresión lineal

Es un algoritmo de Machine Learning empleado para obtener una tendencia y un resultado numérico. Este algoritmo establece una relación lineal entre una o más variables independientes y una variable de respuesta o variable dependiente. Usualmente se utiliza el método de mínimos cuadrados para obtener esta relación. Se utiliza para realizar predicciones. En la figura 3, se muestra una línea de tendencia en rojo para datos que se grafican en el diagrama de dispersión en (color azul) en coordenadas que están dadas por coordenadas de ‘x’ y de ‘y’. Los valores a predecir están dados por la ecuación $y = b0 + b1x$

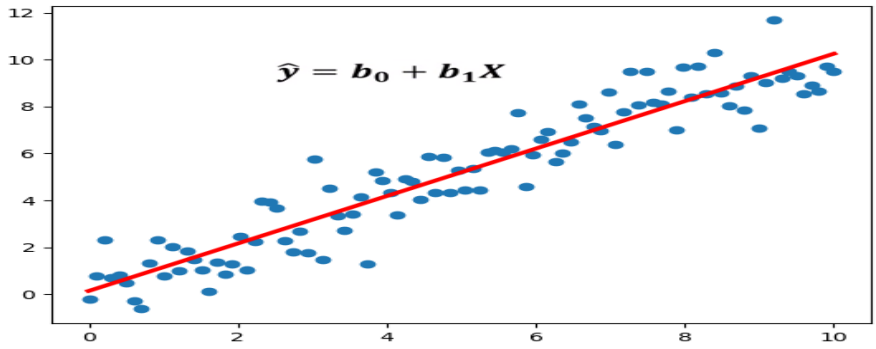


Figura 3. Regresión lineal (Didactalia classes, 2018).

Las fortalezas y debilidades de la regresión lineal se muestran en la siguiente tabla 3:

Tabla 3.

Fortalezas y Debilidades Regresión lineal simple. Elaboración propia

Fortalezas	Debilidades
• El enfoque más común para el modelado numérico de datos. • Se puede adaptar para modelar casi cualquier dato. • Proporciona estimaciones de la fuerza y el tamaño de las relaciones entre las características y el resultado.	• Hace suposiciones fuertes de los datos. • La forma del modelo debe ser especificada previamente por el usuario. • No le va bien con los datos faltantes. • Solo funciona con funciones numéricas, por lo que los datos categóricos requieren un procedimiento adicional • Requiere algunos conocimientos • de estadística para entender el modelo.

9.2.3.3. Regresión logística

El objetivo principal de la regresión logística, es el de modelar la influencia de las variables regresoras en la probabilidad de ocurrencia de un suceso en particular.

Sistemáticamente tiene dos objetivos: el primero es investigar cómo influye la probabilidad de ocurrencia de un suceso, la presencia de diversos factores, o no, y el valor o nivel de estos, el segundo es determinar el modelo más ajustado que describa la relación entre la variable respuesta y un conjunto de variables regresoras.

Los modelos de regresión logística son modelos estadísticos, por medio de los cuales se desea conocer la relación entre una variable dependiente cualitativa, (regresión logística binaria o binomial) o con más de dos valores (regresión logística multinomial).

Se utiliza una o más variables explicativas independientes, ya sean cualitativas o cuantitativas, la ecuación inicial del modelo es de tipo exponencial, se hace la transformación logarítmica (logit) permitiendo su uso como una función lineal.

Si la curva tiende infinito positivo la predicción se transforma a 1, y si la curva pasa el infinito negativo, la predicción será 0. Si la salida de la función Sigmoide es mayor que 0.5, se puede clasificar el resultado como 1 o SI, y si es menor que 0.5 se clasifica como 0 o No. Ejemplo: si el resultado es 0.75, se puede decir en términos de probabilidad, que existe un 75% de probabilidades de que el paciente sufra cáncer (González, Regresión Logística – Teoría, 2019).

La figura 4, muestra la gráfica de la función Sigmoide ‘S’ que refleja la predicción para resultados binarios, o es ‘Si’ o ‘1’ o es ‘No’ o ‘0’.

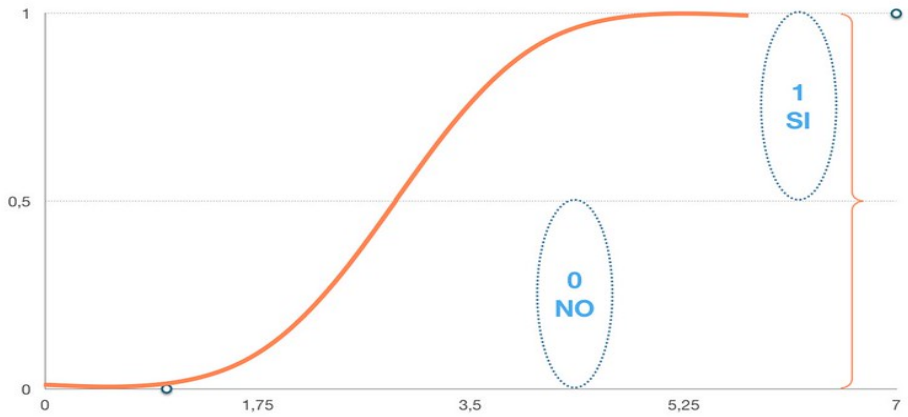


Figura 4. Regresión logística (González, Regresión Logística – Teoría, 2019)

9.2.3.4. Máquinas de vectores de soportes (SVM)

Las máquinas de soporte vectorial, son algoritmos de decisión que permiten resolver los problemas de regresión y clasificación de una manera muy eficiente debido a su sistema de aprendizaje automático. Estas se basan en la teoría del aprendizaje estadístico para la resolución de problemas de clasificación y regresión. (Parra, 2019).

El éxito de las máquinas de soporte vectorial radica principalmente en tres ventajas fundamentales: la primera hace referencia a una sólida fundamentación matemática, la segunda se refiere al concepto de minimizar el riesgo estructural; esto se significa que la minimización de la probabilidad de una clasificación errónea sobre nuevos ejemplos, este caso normalmente se presenta cuando hay pocos datos de entrenamiento, y la tercera ventaja radica en que disponen de potentes herramientas y algoritmos para hallar la solución de manera rápida y eficiente.

Los datos son mapeados por medio de un Kernel Gaussiano u otro tipo de Kernel a un espacio de características en un espacio dimensional más alto, en el que se busca la máxima separación entre clases. Esta función de frontera, cuando es traída de regreso al espacio de entrada, puede separar los datos en todas las clases distintas, formando cada una un agrupamiento.

Las fortalezas y debilidades del algoritmo SVM se describen en la tabla 4:

Tabla 4.

Fortalezas y Debilidades Support Vector Machine (SPM). Elaboración propia.

Fortalezas	Debilidades
• Se puede utilizar para problemas de clasificación o predicción numérica. • No está excesivamente influenciada por datos ruidosos y no muy propensos a sobrealimentación. • Puede ser más fácil de usar que las redes neuronales, especialmente debido a la existencia de varios algoritmos SVM bien soportados. • Ganando popularidad debido a su alta precisión y victorias de alto perfil en competiciones de minería de datos.	• Encontrar el mejor modelo requiere probar varias combinaciones de núcleos y parámetros de modelo. • Pude ser lento para entrenar, especialmente si el conjunto de datos de entrada tiene una gran cantidad de funciones o ejemplos • Resultados en un modelo complejo de caja negra que es difícil, si no imposible de interpretar.

9.2.3.5. Árboles de clasificación

Los árboles de clasificación o decisión separa las observaciones en ramas para construir un árbol con el propósito de mejorar la precisión de la predicción. Son un modelo surgido en el ámbito del ML y de la Inteligencia Artificial que partiendo de una base de datos, crea diagramas de construcciones lógicas que ayudan a resolver problemas de predicción y clasificación. (Parra, 2019).

De esta forma, se utilizan algoritmos matemáticos para identificar una variable y el inicio correspondiente de la variable que divide la entrada en dos o más subgrupos.

El objetivo del algoritmo es encontrar un par de variables de inicio que maximizan la homogeneidad de los resultantes en dos grupos de muestras o más.

En la figura 5 se muestra un ejemplo de árbol de decisión. El nodo raíz se encuentra arriba del árbol. Los nodos de decisión se visualizan por ramas sobre atributos particulares. Los arcos que identifican a un nodo corresponden a los posibles valores del atributo considerado en ese nodo. Cada línea conduce a otro nodo de decisión o a una hoja. Los nodos hoja representan la predicción de los datos. Para clasificar se recorre el árbol de arriba hacia abajo de acuerdo a los valores de los atributos probados en cada nodo y cuando se llega a una hoja, la instancia se clasifica de acuerdo a la clase que tiene la hoja (Contreras Morales, Ferreira Correa, & Valle, 2017).

El árbol en su forma gráfica se representa por un conjunto de nodos, hojas y ramas.

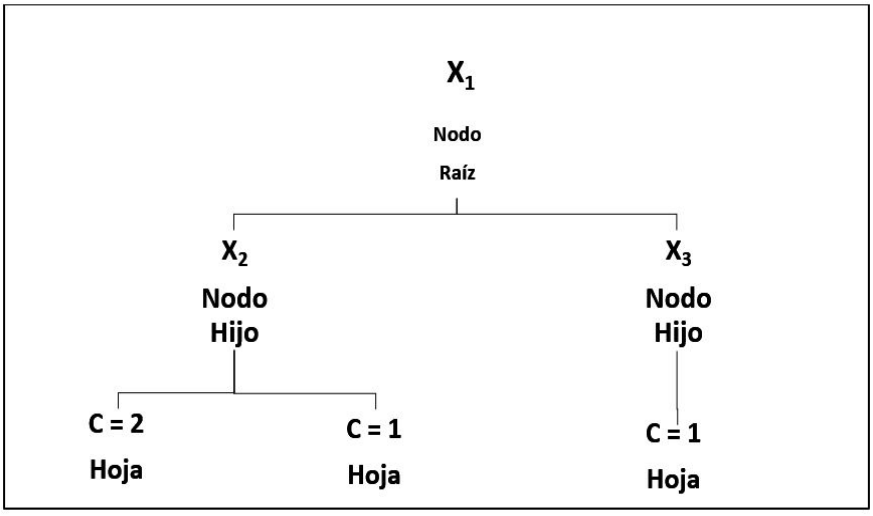


Figura 5. Árbol de decisión (Contreras Morales, Ferreira Correa, & Valle, 2017)

La tabla 5, muestra fortalezas y debilidades de los árboles de decisión:

Tabla 5.

Árboles de decisiones. Elaboración propia

Fortalezas	Debilidades
• Un clasificador de uso múltiple que funciona bien en la mayoría de los problemas. • El proceso de aprendizaje altamente automático puede manejar características numéricas o nominales. • Utiliza solo las características más importantes • Puede usarse en datos que se puede interpretar sin un fondo matemático (para árboles relativamente pequeños). • Más eficientes que otros modelos complejos • Puede usarse con pocos datos • o un número muy grande.	• Los modelos de árbol de decisión a menudo están sesgados hacia divisiones en características que tienen un gran número de niveles. • Es fácil disfrazar o vestir al modelo. • Puede tener problemas para modelar algunas relaciones debido a la dependencia de divisiones de ejes paralelos. • Pequeños cambios en los datos de entrenamiento pueden resultar en grandes cambios en la lógica de decisión • Los árboles grandes pueden ser difíciles de interpretar y las decisiones • que toman pueden parecer contradictorias • Pocos ejemplos

9.2.3.6. Bosques aleatorios

Para obtener un bosque aleatorio, primero hay que decidir de cuántos árboles se van a componer. Si se dispone de n árboles, entonces se crearán n nuevos conjuntos a partir del conjunto de entrenamiento a base de extraer elementos con reemplazo de dicho conjunto. Cada nuevo conjunto debe tener el tamaño del conjunto de entrenamiento.

Los bosques aleatorios es un conjunto de árboles, y de tantos árboles el analista selecciona el de mejor rendimiento. Se entrena un árbol con cada conjunto, obteniendo así n árboles diferentes.

Para realizar una predicción con el bosque aleatorio, lo que se hace es evaluar de todos los árboles sobre datos de entrenamiento el que tenga mejor rendimiento de predicción.

La predicción se basa en la evaluación de todo el conjunto de árboles. Con la configuración pertinente, es uno de los métodos que, con una menor complejidad, proporciona mejores resultados en una gran variedad de problemas de Machine Learning (Montes Núñez, 2017).

La figura 6 muestra un bosquejo del algoritmo de bosques aleatorios, de un conjunto de árboles evaluados, se elige al de mejor rendimiento.

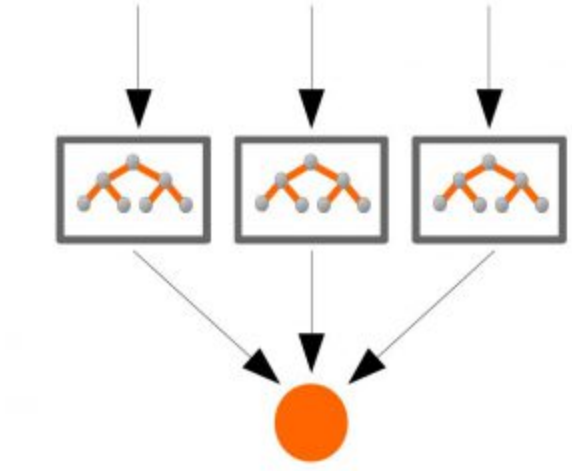


Figura 6. Bosques aleatorios. (Montes Núñez, 2017)

9.2.3.7. Redes neuronales

Las redes neuronales pretenden imitar el modelo biológico del cerebro, a través de modelos matemáticos organizados en redes interconectadas en paralelo y con una organización jerárquica. Intentan interactuar con los objetos del mundo real del mismo modo que el sistema nervioso biológico.

Una Red Neuronal o Neural Network está integrada por un conjunto de capas interconectadas en las que los valores de entrada (inputs) dan lugar a los valores de salida (outputs) a través de una serie de nodos con sus ponderaciones correspondientes. Dichos pesos se obtienen durante el proceso de entrenamiento de la red. Entre las capas de entrada y salida puede haber una o varias capas ocultas (hidden). Las fronteras entre clases que puede definir una Red Neuronal pueden ser complejas e irregulares (Montes Núñez, 2017).

Las capas son las unidades estructurales de la red neuronal, dentro de una capa las neuronas suelen ser del mismo tipo. La capa de entrada captura los datos o señales del entorno, la capa oculta realiza las operaciones matemáticas, procesa los datos y los dirige a la capa de salida la cual entrega la respuesta de la red a los datos capturados por la capa de entrada.

En la figura 7 se observa un esquema de red neuronal artificial interconectada la cual se compone de capas.

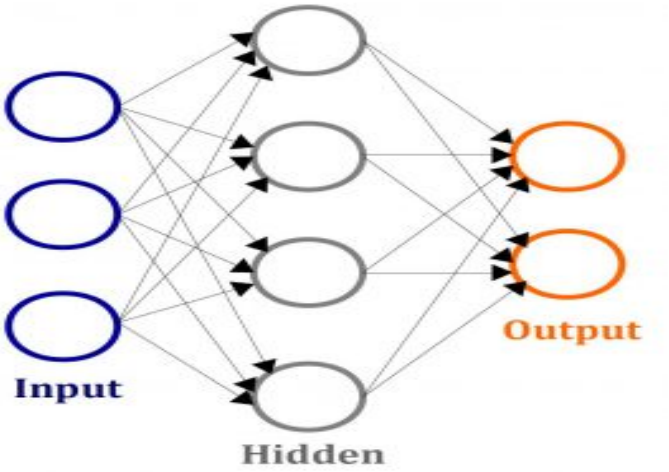


Figura 7. Esquema de Red Neuronal (Montes Núñez, 2017)

Las redes neuronales ofrecen las siguientes fortalezas y debilidades indicadas en la tabla 6:

Tabla 6.

Fortalezas y Debilidades de redes neuronales. Elaboración propia

Fortalezas	Debilidades
• Se puede adaptar a la clasificación o predicción numérica. • Entre los enfoques de modelado más precisos. • Hace algunas suposiciones sobre las relaciones subyacentes de los datos.	• Reputación de ser computacional intensivo y lento para entrenar, particularmente si la topología de la red es compleja. • Datos de entrenamiento fácil o excesivo. • Resultados en un modelo complejo de caja negra que es difícil, si no imposible, de interpretar.

9.2.3.8. Redes Bayesianas

Las Redes Bayesianas como herramienta describen el conocimiento probabilístico de las relaciones entre variables que afectan a un determinado fenómeno no determinista (Delgado & Tibau, 2015).

Tienen una gran aceptación en los procedimientos de toma de decisión en un gran número de campos, y su uso para la evaluación de riesgos es de gran popularidad en áreas tan diversas como la economía, la medicina, el riesgo medioambiental, los desastres ecológicos, ámbito forense, en tribunales de justicia, entre otros. (Delgado & Tibau, 2015).

Las principales fortalezas y debilidades de este algoritmo se observan en la tabla 7:

Tabla 7

Fortalezas y Debilidades de redes bayesianas. Elaboración propia

Fortalezas	Debilidades
• Fácil de construir y entender. • Las inducciones de estos clasificadores son extremadamente rápidas, requiriendo solo un paso para hacerlo. • Es muy robusto considerando atributos irrelevantes. • Toma evidencia de muchos atributos para realizar la predicción final	• Aumenta la complejidad (y el tiempo) tanto para aprender el modelo como para clasificación. • Tratar de mantener la misma estructura sencilla, pero considerando las dependencias entre atributos.

9.2.3.9. Aplicaciones de algoritmos supervisados

En la tabla 8, muestra casos de algoritmos supervisados aplicados a la salud.

Tabla 8

Algoritmos supervisados y sus aplicaciones. Elaboración propia

Algoritmos	Aplicacióno o casos
• Vecino más cercano	• Determinar con precisión un diagnóstico.
• Bayesianos	• Predicción de la Diabetes.
• Árboles de decisión	• Diagnóstico y tratamiento de pacientes con glaucoma.
• Regresión lineal	• Aplicación de la regresión lineal en un problema de nutrición.
• Árboles de regresión	• Diagnóstico de apendicitis aguda en niños.
• Redes neuronales	• Apoyo en el diagnóstico y tratamiento del paciente.
• Máquinas de vectores de soporte (SVM)	• Diagnóstico clínico de la enfermedad de Párkinson

9.2.4. Algoritmos no supervisados

Los algoritmos de aprendizajes supervisados se utilizan para construir modelos predictivos y los algoritmos no supervisados se utilizan para construir modelos descriptivos. El tipo de agoritmo que necesita usar depende de la tarea de aprendizaje que espera lograr. (Lantz, 2013).

El aprendizaje no supervisado utiliza datos de entrenamiento no etiquetados, hace predicciones para todos los puntos invisibles. No hay una variable dependiente del aprendizaje. La agrupación (clustering) y la reducción de dimensionalidad son ejemplos de problemas de aprendizaje no supervisados (Mohri, Afshin, & Ameet, 2018).

La Figura 8 muestra el funcionamiento de los algoritmos no supervisados en ML; a diferencia del supervisado, los datos de entrada no están clasificados ni etiquetados, y no son necesarias estas características para entrenar el modelo.

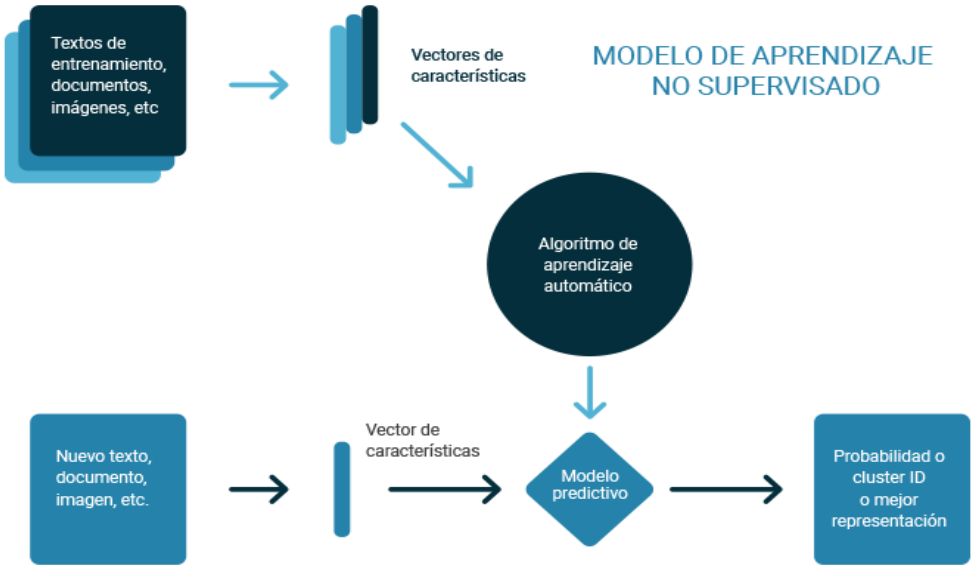


Figura 8. Modelo de aprendizaje no supervisado (Luna González, 2018).

9.2.4.1. Algoritmos de Clustering (agrupación)

Los algoritmos de Clustering clasifican los datos en grupos o clúster en función de las similitudes de sus atributos. Al mismo tiempo se busca que los datos agrupados en clúster diferentes sean claramente distintos.

El algoritmo de clustering jerárquico tiene una variedad de objetivos relacionados con agrupar una colección de objetos, es decir, observaciones, individuos, casos o filas de datos, en grupos o clusters, de modo que los datos que están dentro de cada grupo están más estrechamente relacionados con unos a otros que los objetos asignados a diferentes grupos (González, González, Ligdi, 2018).

En la figura 9, se identifican dos gráficas: del lado izquierdo dos grupos de datos y con la gráfica de la derecha un dendograma que presenta de manera visual clusters o grupos a partir de variable de entrada.

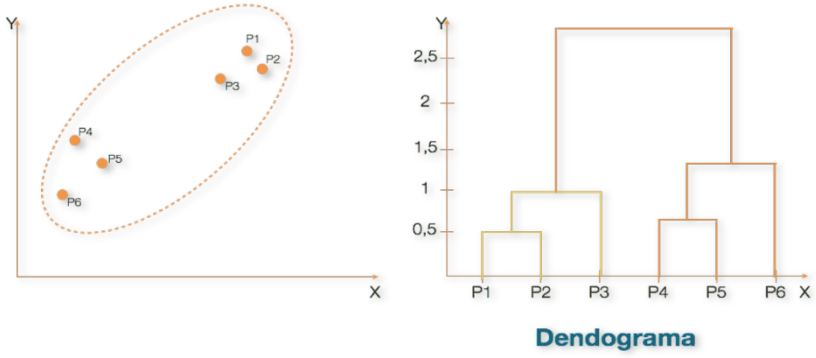


Figura 9. Algoritmo de Agrupamiento (González, González, Ligdi, 2018)

En la tabla 9, se describen las fortalezas y debilidades de este algoritmo.

Tabla 9

Fortalezas y debilidades del algoritmo de clustering. Elaboración propia

Fortalezas	Debilidades
• Usa simple principios para la identificación de grupos que pueden explicarse en términos no estadísticos. • Es altamente flexible y puede adaptarse para abordar casi todas sus deficiencias con ajustes simples.	• Debido a que utilizan un elemento de posibilidad aleatorio, no se garantiza que encuentre el conjunto óptimo de clústeres.

9.2.4.2. Análisis de componentes principales (PCA)

Este algoritmo realiza una reducción de atributos iniciales, de tal forma, que consigue representar la máxima información con el mínimo de atributos posibles. Es decir, convierte un conjunto de valores de características similares, en un conjunto de valores de características tal vez sin similitud.

El análisis de componentes principales es un método no supervisado que busca reducción de dimensionalidad, transforma una serie de variables posiblemente correlacionadas en una combinación lineal de variables no correlacionadas llamadas componentes principales. La idea detrás de ACP es que en lugar de tener varias variables se reducen pocas de ellas sin perder significado y relevancia en los datos. (Rassiga, 2017).

Por consecuencia, la gran fortaleza y ventaja del algoritmo es que permite preservar de mejor manera la variación original, es decir, se sacrifica la menor variabilidad original posible, condensando muchas variables en una sola. En lugar de tener muchas variables se reducen a unas cuantas sin perder relevancia de la información (Rassiga, 2017).

9.2.4.3. Deep Learning

Deep learning, también conocido en español como aprendizaje profundo, son aquellas técnicas de ML que hacen uso de arquitecturas de redes neuronales. El concepto “profundo” viene referido al número de capas que poseen estas técnicas.

Las redes neuronales convencionales posen una capa, las redes neuronales profundas contienen varias capas. Aprendizaje profundo es un aspecto de la inteligencia artificial que se ocupa de emular el enfoque de aprendizaje que utilizan las personas para obtener ciertos tipos de conocimiento. Puede considerarse una forma de automatizar el análisis predictivo (Sáez de la Pascua, 2019).

9.2.5. Tipos de problemas de algoritmos supervisados

En los siguientes puntos, se mencionan los problemas de regresión y clasificación respectivamente.

9.2.5.1. Problemas de regresión

Los problemas de regresión se caracterizan, en que la variable de respuesta Y es cuantitativa. Esto significa que la solución al problema es representada por una variable continua que puede ser flexiblemente determinada por las entradas X del modelo.

Los problemas de regresión que tienen entradas con dependencia temporal son también llamados problemas de predicción de series temporales o forecasting.

Los modelos de regresión predicen el valor de Y dados valores conocidos de variables X. Las predicciones dentro del rango de valores del conjunto de datos que se usa para ajustar el modelo reciben el nombre de interpolaciones. Por el contrario, aquella que está fuera del rango de los datos usado para ajustar o entrenar el modelo recibe el nombre de extrapolaciones, se basa fuertemente en supuestos, cuanto más lejos esté la extrapolación de los datos, más espacio hay para fallos debido a las diferencias entre las suposiciones y la realidad.

9.2.5.2. Problemas de clasificación

Los problemas de clasificación se caracterizan por tener una variable cualitativa Y como respuesta. Muchas veces las variables cualitativas también reciben el nombre de variables categóricas.

Predecir una respuesta cualitativa se le denomina clasificar. Algunas ocasiones los métodos encargados de clasificar, lo que hacen, es predecir la probabilidad de una observación de pertenecer a cada una de las categorías. De manera semejante, se comportan también como algoritmos de regresión y estos pueden ser:

- **Binaria**: {Sí, No}, {Azul, Rojo}, {Fuga, No Fuga}
- **Múltiple**: Comprará {Producto1, Producto2…}
- **Ordenada**: Riesgo {Bajo, Medio, Alto}

9.2.6. Proceso de Machine Learning (ML)

En su libro Lantz, (2013) establece una serie de pasos para implementar ML menciona que cualquier tarea puede llevarse a cabo por medio de estos pasos: (pág. 17)

- **Recolectar datos (dataset):** Dada la problemática a resolver, se debe investigar y obtener datos para alimentar el modelo. Es importante considerar la calidad y cantidad de información, ya que impactará directamente en el funcionamiento del modelo.
- **Explorar y preparar datos:** Se realiza un análisis exploratorio para detectar si hay correlaciones entre las distintas características seleccionadas. La selección de características, impactará directamente en los tiempos de ejecución y en los resultados. Se balancean los datos y se separan en dos grupos: uno para entrenamiento y otro para evaluación del modelo. En esta etapa se realizan otras tareas como: normalización, eliminar duplicados y corrección de errores. Otro aspecto importante es pensar siempre en la calidad y limpieza de los datos.
- **Elección del modelo:** Se elige el modelo a aplicar en base al problema.
- **Entrenamiento:** Se emplea los datos de entrenamiento para ejecutar el modelo. Los “pesos” se inician aleatoriamente, los pesos son los valores que multiplican o afectan a las relaciones entre las entradas y las salidas, se irán ajustando automáticamente por el algoritmo con cada entrenamiento.
- **El test o evaluación:** Se realiza con entradas que el modelo desconoce para verificar la precisión del modelo ya entrenado.
- **Ajustar parámetros:** Se aplica en el caso de no obtener resultados satisfactorios en el test, se realiza nuevamente el entrenamiento modificando los parámetros del modelo. Cada algoritmo tiene sus propios parámetros a ajustar. (Lantz, 2013).

9.2.7. Lenguajes de programación para Machine Learning

Actualmente y en los últimos tiempos, la popularidad y la capacidad de implementación de los lenguajes de programación para Machine Learning han

evolucionado y crecido grandemente, la razón es incluir múltiples dominios de problemas y oportunidades.

Por tal razón, son cada vez más las personas que desean aprender sobre Machine Learning, pero una pregunta hay que hacer ¿qué lenguaje de programación es apropiado aprender para Machine Learning? (González, Ligdi González. Inteligencia Artificial, 2018).

9.2.7.1. Lenguaje Python

Python es líder en desarrollos para machine learning, simple y de fácil aprendizaje (González, Ligdi González. Inteligencia Artificial, 2018), se puede observar una gran tendencia al uso de Python en grandes centros de investigación como el CERN (Organización Europea para la Investigación Nuclear).

Los paquetes NumPy (Python Numérico) y SciPy (Python Científico) son base para el trabajo científico. Es actualmente uno de los lenguajes más utilizados en inteligencia artificial y Ciencia de los Datos; por científicos de datos y expertos en Machine Learning, se caracteriza por hacer hincapié en una sintaxis limpia, que favorece un código legible (Challenger Pérez, Díaz Ricardo, & Becerra García, 2014).

9.2.7.2. Lenguaje R

R es un entorno de trabajo para la ejecución de análisis estadísticos y la creación de gráficos. La interfaz gráfica del programa es una consola de comandos, es decir, que para interactuar con el mismo hay que escribir líneas de código y ejecutarlas. (Ruiz Ruano & Puga, 2016).

R fue uno de los primeros lenguajes de programación desarrollados para la computación estadística, el análisis de datos con un buen soporte para la visualización, incluyendo modelado lineal y no lineal, pruebas clásicas estadísticas, análisis de series temporales, clasificación, clustering, entre otros.

R surgió como una opción indudable de lenguaje de programación entre muchos profesionales de la ciencia de datos. Dado que R era de código abierto y extremadamente poderoso en la construcción de modelos estadísticos sofisticados,

rápidamente se adoptó tanto en la industria como en el mundo académico. R tiene uno de los conjuntos de paquetes más populares (un conjunto de funciones y métodos para llevar a cabo un procedimiento complicado, que de lo contrario requiere mucho tiempo y esfuerzo para implementarlo), R es un proyecto de GNU. (Ramasubramanian & Moolayil, 2019).

9.2.7.3. Matlab

MATLAB proviene de Matrix Laboratory, es una herramienta de software matemático que contiene un entorno de desarrollo integrado con un lenguaje de programación muy particular. Es multiplataforma para Windows, Mac, Linux, Unix. Principalmente para operaciones con matrices.

9.2.8. Herramientas para aplicar modelos de Machine Learning

El aumento de software de código abierto hace que Machine Learning sea cada vez menos difícil de implementar. Las siguientes herramientas de código abierto incluyen librerías para Python y R:

9.2.8.1. NumPy

Es una biblioteca de manejo de datos, particularmente una que permite manejar grandes matrices multidimensionales junto con una gran colección de operaciones matemáticas.

Ventajas

- Capacidades de manipulación de matrices (y matrices multidimensionales) como transposición, remodelación, entre otros.
- Estructuras de datos altamente eficientes que mejoran el rendimiento y manejan la recolección de basura con una brisa.
- Capacidad para vectorizar la operación, de nuevo mejora el rendimiento y las capacidades de paralelización.

Desventajas

- Su alto rendimiento tiene un costo.

- Los tipos de datos son nativos del hardware y no de Python, por lo que incurren en una sobrecarga cuando los objetos Numpy se deben transformar de nuevo en equivalentes de Python y viceversa.

9.2.8.2. Pandas

Es una biblioteca de Python que proporciona estructuras de datos flexibles y expresivas (como marcos de datos y series), para la manipulación de datos, construido sobre Numpy, los pandas son tan rápidos y fáciles de usar.

Ventajas

- Extremadamente fácil de usar y con una pequeña curva de aprendizaje para manejar datos tabulares.
- Increíble conjunto de utilidades para cargar, transformar y escribir datos en múltiples formatos.
- Compatible con los objetos Numpy subyacentes y elije la mayoría de las bibliotecas de aprendizaje automático, como Scikit-learn.
- Capacidad para preparar gráficos / visualización.

Desventajas

- La facilidad de uso tiene el costo de una mayor utilización de la memoria. Pandas crea demasiados objetos adicionales para proporcionar acceso rápido y facilidad de manipulación.
- Incapacidad para utilizar infraestructura distribuida. Aunque los pandas pueden trabajar con formatos como archivos HDFS, no pueden utilizar la arquitectura de sistemas distribuidos para mejorar el rendimiento.

9.2.8.3. Scipy

Scipy es una biblioteca de computación científica para Python. También se construye sobre Numpy y es parte de Scipy Stack. Proporciona módulos, algoritmos para álgebra lineal, integración, procesamiento de imágenes, optimizaciones, agrupación, manipulación de matrices dispersas y muchos más.

9.2.8.4. Matplotlib

Es esencialmente una biblioteca de visualización, funciona a la perfección con objetos Numpy, proporciona un entorno de trazado como MATLAB para preparar gráficos, gráficos de alta calidad para publicaciones, cuadernos y aplicaciones web.

Ventajas

- Sintaxis extremadamente expresiva y precisa para generar gráficos altamente personalizables.
- Se puede usar fácilmente en línea con los portátiles Jupyter.

Desventajas

- Enorme curva de aprendizaje, requiere un poco de comprensión y práctica para usar Matplotlib.
- Gran dependencia de Numpy y otras bibliotecas de pila Scipy.

9.2.8.5. Scikit-Learn

Diseñado como una extensión de la biblioteca Scipy, Scikit-learn se ha convertido en el estándar, para muchas de las tareas de aprendizaje automático. Proporciona un paradigma de predicción y transformación simple pero potente para aprender de los datos, transformar los datos y finalmente predecir. Usando esta interfaz, proporciona capacidades para preparar modelos de clasificación, regresión, agrupamiento y conjunto. También proporciona una multitud de utilidades para pre-procesamiento, métricas y técnicas de evaluación de modelos.

Ventajas

- El paquete para llevar, que lo tiene todo, para los algoritmos clásicos de aprendizaje automático.
- Interfaz de ajuste y transformación consistente y fácil de entender.
- La capacidad para preparar tuberías no solo ayuda con la creación rápida de prototipos, sino también con implementaciones rápidas y confiables.

Desventajas

- Incapacidad para utilizar datos categóricos para algoritmos listos para usar que admiten dichos tipos de datos (los paquetes en R tienen tales capacidades).

9.2.8.6. Statsmodels

Esta biblioteca agrega herramientas y algoritmos estadísticos en forma de clases y funciones al mundo de Python. Construido sobre Numpy y Scipy, Stastmodels proporciona una extensa lista de capacidades en forma de modelos de regresión, análisis de series de tiempo, autor regresión.

Ventajas

- Conecta la brecha para la regresión y los algoritmos de series de tiempo para el ecosistema de Python.
- Análogo a ciertos paquetes R, por lo tanto, una curva de aprendizaje más pequeña.
- Enorme lista de algoritmos y utilidades para manejar casos de uso de regresión y series de tiempo.

Desventajas

- No tan bien documentado con ejemplos como Scikit-Learn.
- Ciertos algoritmos tienen errores con poca o ninguna explicación de los parámetros.

9.2.8.7. XgBoost

Más utilizada en diversas competiciones de ciencia de datos y casos de uso en el mundo real, XgBoost es probablemente una de las variantes más conocidas, permite una ejecución paralela y, por lo tanto, proporciona una mejora inmensa del rendimiento en comparación con los árboles potenciados por gradiente. Proporciona capacidades para ejecutar sobre marcos distribuidos como Hadoop fácilmente.

9.2.8.8. LigthGBM

LigthGBM (máquinas de aumento de gradiente), similar a XgBoost en la mayoría de los aspectos, a excepción de unos pocos sobre el manejo de variables categóricas y el proceso de muestreo para identificar la división de nodos.

LigthGBM utiliza un método novedoso llamado GOSS (muestreo de un lado basado en gradiente) para identificar la división de nodos. También tiene la capacidad de utilizar GPU para mejorar el rendimiento. Se informa durante algunas competiciones que LigthGBM es más eficiente en memoria que XgBoost.

9.2.8.9. Eli5

ELI5 es una de esas bibliotecas que proporciona las capacidades para depurar clasificadores y proporcionar una explicación sobre las predicciones.

9.2.8.10. TensorFlow

TensorFlow es una biblioteca matemática simbólica, que permite una programación diferenciable, un concepto central para muchas tareas de aprendizaje automático. Los tensores son el concepto central de esta biblioteca, que son objetos matemáticos genéricos para representar vectores, escaladores y matrices multidimensionales.

Ventajas

- Paquete de grado industrial que tiene un gran soporte comunitario con correcciones de errores frecuentes y mejoras a intervalos regulares.
- Capacidad para trabajar con un conjunto diverso de hardware, como plataformas móviles, web, CPU y GPU.
- Escalabilidad para manejar grandes cargas de trabajo y trabajos fuera de la caja.

Desventajas

- La interfaz de bajo nivel hace que sea difícil comenzar, enorme curva de aprendizaje
- No es fácil acostumbrarse a los gráficos de computación.

9.2.8.11. Theano

Fue una de las primeras bibliotecas en proporcionar capacidades para manipular arreglos multidimensionales, tiene capacidad para utilizar GPU de forma transparente. Está estrechamente integrado con Numpy, proporciona una sintaxis de diferenciación simbólica junto con varias optimizaciones para manejar números grandes y pequeños. Antes de la llegada de las nuevas bibliotecas, Theano era el bloque de construcción de facto para trabajar con redes neuronales.

Ventajas

- Facilidad de comprensión debido a su acoplamiento apretado con Numpy.
- Capacidad para utilizar GPU de forma transparente.
- Siendo una de las primeras bibliotecas de aprendizaje profundo, tiene una gran comunidad para ayudar y apoyar problemas.

Desventajas

- Es un proyecto obsoleto que no se desarrollará aún más.
- Sus API de bajo nivel a menudo presentaban una curva de aprendizaje empinada.

9.2.8.12. PyTorch

Es el resultado de la investigación y el desarrollo en el grupo de inteligencia artificial de Facebook. Admite gráficos dinámicos y una ejecución impecable (fue el único hasta TensorFlow 2.0).

Ventajas

- Uno de los marcos de aprendizaje profundo más rápidos.
- Capacidad para manejar gráficos dinámicos en oposición a los estáticos utilizados por la mayoría de las contrapartes.
- La implementación de Python ayuda a una integración perfecta con los objetos de Python y la sintaxis similar a Numpy.

Desventajas

- Aún ganando terreno y apoyo, por lo tanto, se retrasa en términos de material (tutoriales, ejemplos, etc.), para aprender.
- Capacidades limitadas como visualizaciones y depuración.

9.2.8.13. Keras

Keras es un marco de aprendizaje profundo de alto nivel que ha facilitado la forma en que desarrollamos y trabajamos con redes neuronales profundas. Desarrollado principalmente en Python, para Keras, el bloque de construcción básico es una capa. Dado que, la mayoría de las redes neuronales son diferentes configuraciones de capas, trabajar de tal manera facilita enormemente el flujo de trabajo.

9.2.9. Áreas de aplicación de Machine Learning

Actualmente el ML se está desarrollando y aplicando en estas áreas (Moscadó Pérez, 2018).

- **Seguridad informática:** Detección de ataques de intrusos, ayuda a la detección de fraudes online y con tarjeta de crédito. Antivirus detección de spam y detención de anomalías.
- **Reconocimiento de imágenes:** Por ejemplo: reconocimiento facial, dactilar, de objetos en un espacio, de voz y/o de escritura.
- **Conducción autónoma:** Reconocer imágenes en tiempo real, detectar obstáculos, señales de tráfico y predecir velocidades.
- **Salud:** Realizan un diagnóstico médico, analizan grandes volúmenes de imágenes en busca de algún tipo de enfermedad o predecir la posibilidad de tener una enfermedad dependiendo de una serie de características del paciente.
- **Análisis del mercado de valores:** Puede ayudar a los analistas financieros a predecir el precio de determinadas acciones.
- **Motores de recomendación:** Catalogan a un usuario según unos determinados parámetros y se les hace recomendaciones vinculadas con su catalogación.
- **Lingüística computacional:** Ayudan a las máquinas a mejorar la interactividad y la comunicación natural con los humanos, por ejemplo: traducción automática, redacción automática de textos y noticias.

- **Reconocimiento del habla:** Los algoritmos de ML, en especial los Deep Learning (DL), pueden trabajar paralelamente, sus principales ventajas son: mejoran la calidad de audio eliminando ruidos y permite detectar sonidos o canciones.
- **Motores de búsqueda:** Google utiliza esta tecnología para poder buscar en su banco de imágenes automáticas catalogadas o también puede buscar patrones o imágenes parecidas (Moscadó Pérez, 2018).

9.3. Desarrollo

La propuesta en lo general tiene que ver con aplicar técnicas y algoritmos de ML en el Sector Salud aprovechando las bondades de los lenguajes de programación R y Python principalmente

La gran cantidad de datos en hospitales y procesos de atenciones médicas y al gran aumento de números de parámetros y de la importancia de los mismos, se propone una solución que reside en la utilización de nuevas técnicas para el desarrollo de procesos de trasformación de los datos, con funcionalidad de selección, integridad, calidad y seguridad para monitoreo continuamente de los mismos.

Sin llegar a concretar un propuesta específica y tratando de dejar en el lector una idea clara de cómo sacar beneficio del ML en Sector Salud, la pregunta es, ¿en dónde se puede aplicar?. De la gran cantidad de aplicaciones que tiene el ML en el ámbito sanitario, se describen tres propuestas genéricas además de las citadas en la tabla 8 del punto 9.2.3.9.

Primero: con la gran cantidad de información en los hospitales, sería posible identificar un proyecto de ML para identificar relaciones, patrones en los datos. poder diferenciar pacientes similares por medio de un análisis y en consecuencia advertir a los médicos acerca de una probabilidad de enfermedad del paciente. (Luque Sucasaire, 2020). Se puede aplicar técnicas de regresión y clasificación dentro de los algoritmos supervisados de ML. De igual forma para encontrar

aspectos descriptivos se pueden usar algoritmos no supervisados como clustering para detectar patrones similares en grupos de pacientes con características similares y diferentes, con ello atacar problemas de manera oportuna.

Segundo: Utilizando algoritmos avanzados de Deep Learning DL para detectar enfermedades y clasificar pacientes a partir de imágenes que pueden ayudar a detectar tumores malignos en pulmón a partir de tomografías; dar soporte a análisis de resonancias magnéticas con el objetivo de detectar a pacientes con Alzheimer encontrando diferencias en patrones del cerebro e intensidad de la imagen, entre un cerebro con Alzheimer y uno sano; utilizar diagnóstico por imágenes DPI para aumentar certeza en reconocimiento de fracturas de extremidades y finalmente el uso de DL en resonancias magnéticas para la detección y clasificación de patrones en una afección cardíaca. (Leivi, 2019).

Tercero. Pensando en una escalabilidad global y propuesta futurista es integrar toda la información del Sector Salud de todo el Municipio, Estado y País con la finalidad de tener un gran abanico de alternativas para aplicar ML, DL y por supuesto aplicar la la fortaleza y potencial de la Inteligencia Artificial.

9.3.1. Proceso para llevar a cabo

Como parte de una propuesta, se debe pensar en una serie de actividades se deben de realizar para llevar a cabo un proceso de ML, se citan algunas de ellas:

- Integración de equipo de trabajo.
- Capacitación en R y Python.
- Reconocimiento de procesos de datos y de información (hospitales).
- Identificación de datos.
- Análisis de datos como casos específicos.
- Integración del análisis y resultados.
- Desarrollo e implementación de modelos.
- Implementación de pruebas.
- Comunicación de resultados a los directivos y médicos.

Así mismo, se sugiere realizar un conjunto de actividades extras que enriquecen y refuerzan las anteriores:

- Obtener la información elegida y procesarla para obtener características que puedan ayudar a mejorar la predicción.
- Determinar qué características y atributos se van a utilizar para entrenar el modelo. Estas características serán las necesarias para generar un modelo predictivo preciso. Es importante evitar que haya un número elevado de características para que el modelo no sea muy complejo, pero a la vez tiene que haber suficientes características para poder generar una predicción.
- Seleccionar el algoritmo de aprendizaje a utilizar.
- Mejorar el algoritmo de aprendizaje seleccionado, ajustando sus parámetros.
- Una vez ajustando los parámetros, se implementa y se evalúa su exactitud.

9.3.2. Casos de ML en el Sector Salud

En este apartado, se presentan dos casos extraídos de la literatura que abordan la predicción del cáncer y diagnóstico temprano así como el caso de la detección enfermedades cardiovasculares

9.3.2.1. Caso Wisconsin Breat Cáncer. Predicción del riesgo de cáncer y diagnóstico.

La clasificación es una de las tareas más importantes y esenciales en el aprendizaje automático, se han realizado diversas investigaciones para aplicar el aprendizaje automático en diferentes conjuntos de datos médicos y clasificar por ejemplo el cáncer de mama.

El objetivo de este caso, es evaluar la eficiencia y la eficacia de los algoritmos de clasificación, Máquinas de Vectores de Soporte (SVM), Bayer Naive (NB), Árboles de decisión (C4.5) y K-Nearest Neighbors (K-NN). En términos de precisión, sensibilidad, especificidad y precisión.

Los conjuntos de datos de Wisconsin Breat Cáncer, del repositorio de Machine Learning de UCI se utilizan en este estudio. Wisconsin tiene 699 casos

(benigno: 458, maligno: 241), 2 clases (65.5% maligno y 34.5% benigno), y 11 atributos de valores enteros.

La efectividad de todos los clasificadores en términos de tiempo para construir el modelo correctamente, instancias clasificadas, instancias incorrectas clasificadas y exactitud, se muestran en la tabla 10 y en la figura 10.

Tabla 10

Rendimientos de los algoritmos de clasificación .Caso Wisconsin Breat Cáncer (Asri, Mousannif, Moatassime, & Noeld, 2016)

Criterios de evaluación	**Clasificadores**			
	C4.5	**SVM**	**NB**	**K-NN**
Tiempo para construir un modelo (s)	0.06	0.07	0.05	0.01
Correctamente clasificado instancias	665	678	671	666
Incorrectamente clasificado instancias	34	21	28	33
Precisión	97.13	97.13	95.99	95.27

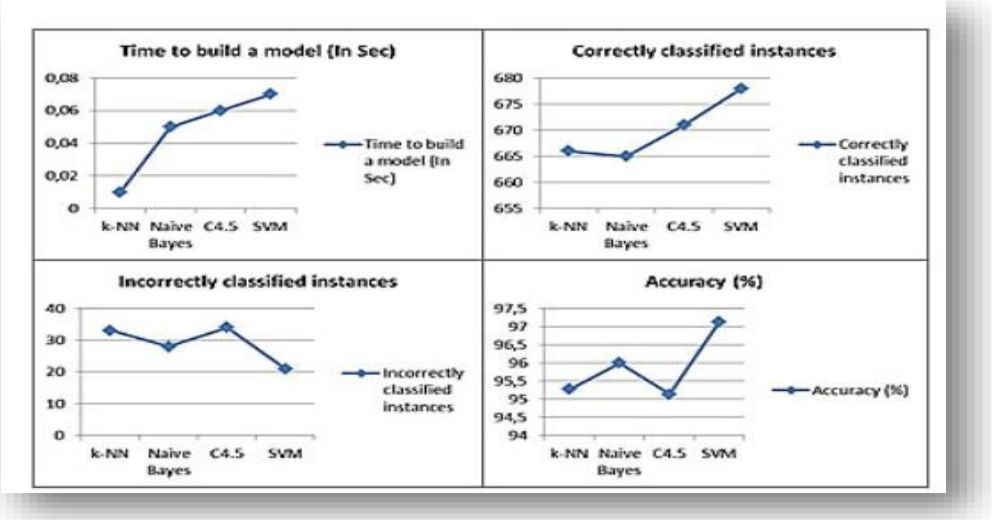


Figura 10. Gráfico comparativo de diferentes clasificadores. (Asri, Mousannif, Moatassime, & Noeld, 2016)

Para medir mejor el rendimiento de los clasificadores, el error de simulación también se considera en este estudio. Se evalúa la efectividad del clasificador en término de los criterios que se observan en la tabla 11:

- La estadística de Kappa (KS) como una medida de acuerdo al azar entre las clasificaciones y las clases verdaderas.
- Error absoluto promedio (MAE) como pronóstico cercano o predicciones de los resultados finales.
- Error cuadrático medio (RMSE).
- Error absoluto relativo (RAE).
- Error cuadrático relativo (RRSE).

En la tabla 11 se muestran algunos resultados de cada algoritmo y en la figura 11 se identifica un comparativo.

Tabla 11.

Error de entrenamiento y simulación. Caso WisconsinBreat Cáncer (Asri, Mousannif, Moatassime, & Noeld, 2016)

Criterios de evaluación	**Clasificadores**			
	C4.5	**SVM**	**NB**	**K-NN**
Estadística Kappa (KS)	0.89	0.93	0.91	0.89
Error absoluto promedio (MAE)	0.06	0.02	0.03	0.04
Error cuadrático medio (RMSE)	0.21	0.16	0.19	0.21
Error absoluto relativo (RAE) %	14	6.33	8.59	10.46
Error cuadrático relativo (RRSE) %	45	35.58	40.95	44.77

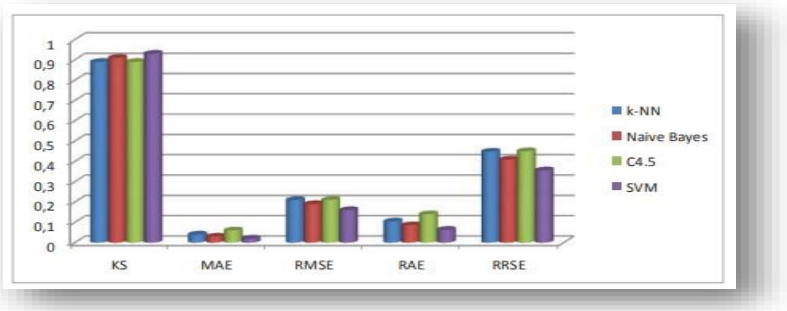


Figura 11. Diagrama comparativo: KS, MAE, RMSE, RAE Y RRSE (Asri, Mousannif, Moatassime, & Noeld, 2016)

Una vez que se construye el modelo predictivo, se puede verificar qué tan eficiente es. Para eso se comparan las medidas de precisión. Tasa verdaderos positivos (TP), Tasa de falso positivo (FP), Precisión, Sensibilidad, Media y Clase. Ver la tabla 12.

Tabla 12.

Comparación de medidas de precisión para C4.5, SVM, NB y K-NN. Caso Wisconsin Breat Cáncer (Asri, Mousannif, Moatassime, & Noeld, 2016)

	(TP)	(FP)	Precisión	Sensibilidad	Media	Clase
C4.5	0.95	0.05	0.96	0.95	0.963	Benigno
	0.94	0.04	0.91	0.94	0.93	Maligno
SVM	0.97	0.03	0.98	0.97	0.97	Benigno
	0.96	0.02	0.95	0.96	0.95	Maligno
NB	0.95	0.02	0.98	0.95	0.96	Benigno
	0.97	0.04	0.91	0.97	0.94	Maligno
K-NN	0.97	0.08	0.95	0.97	0.96	Benigno
	0.91	0.02	0.94	0.91	03	Maligno

Para comprender la eficiencia, la figura 12 presenta la curva ROC de los clasificadores que muestran mejor precisión de cada clasificador. La curva proporciona un gráfico que muestra el rendimiento de diferentes clasificadores.

Se pueden seleccionar modelos óptimos y descartar otros para obtener la mejor clasificación.

Las matrices de confusión representan una forma para evaluar el clasificador, cada fila representa las tasas en una clase real mientras que cada una de las columnas las predicciones, ver tabla 13.

Tabla 13.

Matrices de confusión. Caso Wisconsin Breat Cáncer. (Asri, Mousannif, Moatassime, & Noeld, 2016)

Algoritmo	Benigno	Malingno
C4.5	438	20
	14	227
SVM	446	12
	9	232
NB	436	22
	6	235
K-NN	445	13
	20	221

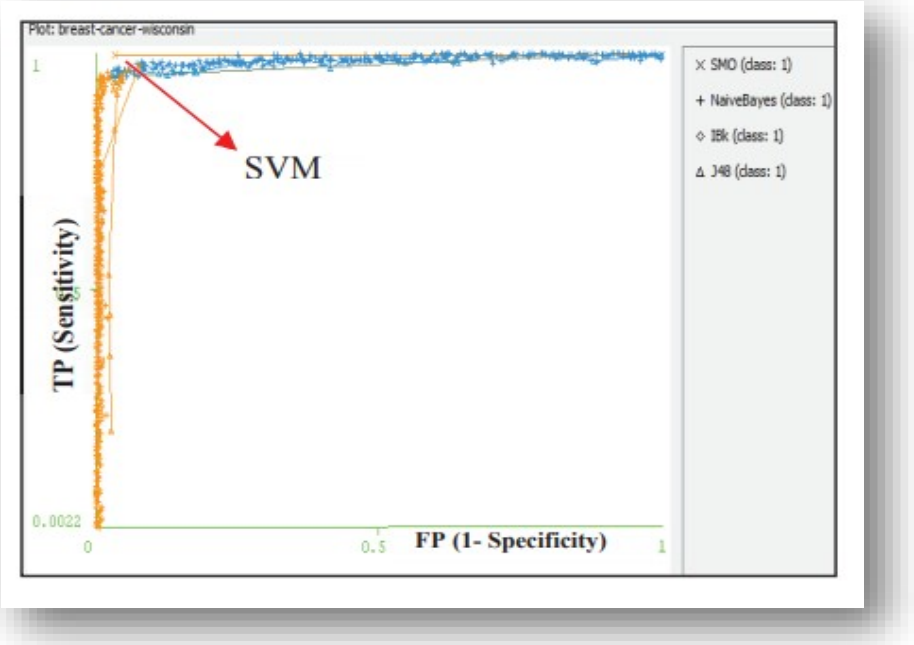


Figura 12. Curva ROC. (Asri, Mousannif, Moatassime, & Noeld, 2016)

Se observa que SVM es el clasificador perfecto y con la tasa de error más baja, enseguida le siguen los algoritmos NB, C4.5 y K-NN.

9.3.2.2. Técnicas de Machine Learning en medicina cardiovascular

Para este estudio se comparan los métodos de árboles de decisión C4.5, regresión logística y máquinas de soporte vectorial (SVM), se utiliza un conjunto de datos de Cleveland que hacen referencia a enfermedades cardiovasculares a través del repositorio de Machine Learning UCI. El conjunto de datos cuenta con 14 atributos y 303 registros (De la Hoz Manotas, Martínez Palacio, & Enrique, Técnicas de ML en medicina cardiovascular, 2013). Los atributos se muestran en la tabla 14.

De la distribución de los datos, se utiliza 60% para datos de entrenamiento, 20% para validación y 20% para la prueba. El caso identifica resultados como la matriz de confusión y la exactitud (número de aciertos entre número total de ejemplos de prueba). (De la Hoz Manotas, Martínez Palacio, & Enrique, Técnicas de ML en medicina cardiovascular, 2013). El caso busca hacer una comparación entre varios algoritmos con la finalidad de elegir el que mejor resultados ofrece.

Tabla 14.

Atributos. Caso medicina cardiovascular (De la Hoz Manotas, Martínez Palacio, & Enrique, Técnicas de ML en medicina cardiovascular, 2013)

No	Atributos	Valores
1	Edad	Edad en años
2	Sexo	0: Masculino, 1: Femenino
3	Tipo dolor de pecho	Valor 1: Angina típica Valor 2: Angina atípica Valor 3: Otro dolor Valor 4: Asintomático
4	Presión arterial en reposo	En mm Hg en la admisión del hospital
5	Colesterol	mg/dl
6	Nivel de azúcar >120 ml/dl	0: Falso, 1: Verdadero
7	Resultado electrocardiograma	Valor 0: Normal Valor 1: Anomalías Valor 2: Hipertrofia ventricular
8	Frecuencia cardiaca máxima alcanzada	
9	Ejercicio de inducción de angina	0: No, 1: Si
10	Depresión inducida	
11	Pendiente curva máxima del ejercicio	
12	Numero de vasos mayores (0-3)	
13	Thal	3: Normal 6: Defecto fijo 7: Defecto reversible
14	Diagnóstico de enfermedad cardiaca	0: Menor 50% 1: Mayor 50%

El caso recopilado indica que se aplicaron tres técnicas de clasificación para realizar un comparativo entre estos datos y se observó quien proporcionaba mayor precisión y menor error en las predicciones relacionadas con enfermedades cardiovasculares. Los métodos comparados fueron: máquinas de soporte vectorial, regresión logística y árboles de decisión. Se obtuvo la matriz de confusión de cada algoritmo que permitió evaluar sensibilidad, especificidad y precisión.

Las fórmulas utilizadas para el cálculo de la sensibilidad, especificidad y precisión son las siguientes:

- Sensibilidad = TP / (TP+FN)
- Especificidad = TN / (TN+FP)
- Precisión = (TP+TN) / TP+FP+TN+FN)

Donde:

- TP: Número de muestras clasificadas verdaderas (actualmente verdaderas)
- FP: Número de muestras clasificadas verdaderas (actualmente falsas)
- FN: Número de muestras clasificadas como falsas (actualmente verdaderas)
- TN: Número de muestras clasificadas falsas (muestras mal clasificadas)

Tabla 15.

Matriz de confusión en general. Elaboración propia

Precisión	**Ausencia**	**Presencia**
Ausencia	TP	FN
Presencia	FP	TN

Se observan las matrices de confusión para cada algoritmo, de las tablas 16 a la 18.

Tabla 16.

Matriz de confusión algoritmo SVM. Caso enfermedades cardiovasculares (De la Hoz Manotas, Martínez Palacio, & Enrique, Técnicas de ML en medicina cardiovascular, 2013)

Número de muestras	Ausencia	Presencia
Ausencia	139	25
Presencia	29	110

Tabla 17.

Matriz de confusión algoritmo Regresión logística. Caso enfermedades cardiovasculares (De la Hoz Manotas, Martínez Palacio, & Enrique, Técnicas de ML en medicina cardiovascular, 2013)

Número de muestras	Ausencia	Presencia
Ausencia	145	19
Presencia	29	110

Tabla 18.

Matriz de confusión algoritmo árboles de decisión. Caso enfermedades cardiovasculares (De la Hoz Manotas, Martínez Palacio, & Enrique, Técnicas de ML en medicina cardiovascular, 2013)

Número de muestras	Ausencia	Presencia
Ausencia	132	23
Presencia	39	100

Se muestran algunos datos comparativos de los métodos elegidos en la tabla 19 siguiente:

Tabla 19

Comparaciones de sensibilidad, especificidad y recisión. Caso enfermedades cardiovasculares (De la Hoz Manotas, Martínez Palacio, & Enrique, Técnicas de ML en medicina cardiovascular, 2013)

	Sensibilidad	Especificidad	Precisión
Máquina de soporte vectorial	84.8%	79.13%	82.17%
Regresión logística	88.41%	79.13%	84.15%
Árboles de decisión	80.48%	71.94%	76.56%

En la figura 13, se muestra de igual forma el comparativo

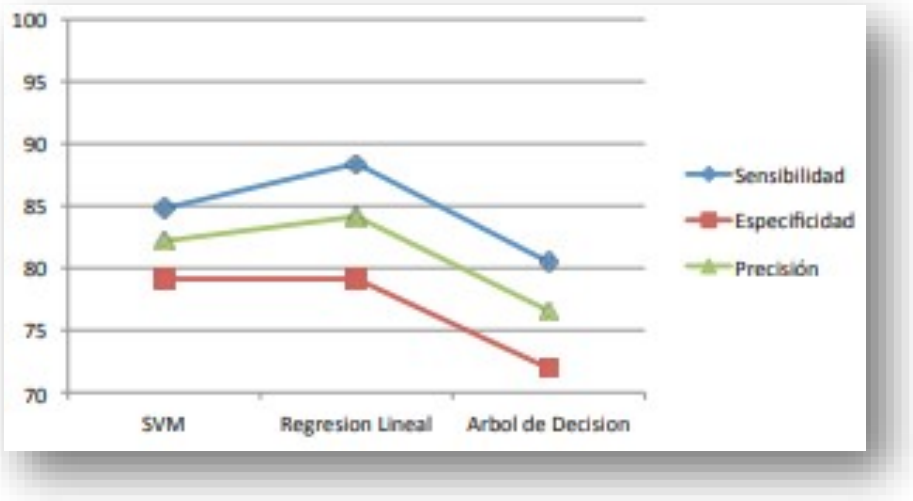


Figura 13. Gráfica comparativa (De la Hoz Manotas, Martínez Palacio, & Enrique, Técnicas de ML en medicina cardiovascular, 2013)

Se evidencia claramente que la regresión logística es, de los métodos comparados, el que mayor precisión brinda a los expertos en cuidados de la salud sobre la posibilidad de predecir enfermedades cardiacas con un porcentaje del 84.15%.

Las técnicas de Machine Learning permiten el análisis de datos a gran escala para lograr un mayor acierto relacionado con la detección o predicción de enfermedades cardiovasculares, tras someter a pruebas técnicas como son los árboles de decisión, las máquinas de soporte vectorial y la regresión logística.

9.3.3. Tecnologías y recursos

Para todo el análisis y tratamiento de los datos, así como la creación de los modelos se recomienda el uso de los lenguajes de programación y sus entornos de trabajo R y Python.

Conclusiones

Al recapitular el objetivo general del presente trabajo que textualmente establece “realizar un diagnóstico y propuesta general de implementación de técnicas y algoritmos de Machine Learning dentro del Sector Salud”, se cumple con una propuesta en lo general, sin llegar a establecer un diagnóstico real del estado del Sector Salud y a partir de ahí determinar una propuesta específica.

A lo largo de los apartados de marco de referencia y el desarrollo del capítulo se pude decir que se cumplen con los objetivos específicos citados.

Por otra parte, las instituciones de salud tienen un gran potencial de explotación de los datos, es necesario que evolucionen y den un paso en su transformación digital, se podrían ver ampliamente beneficiados gracias a la aplicación de disciplinas y herramientas tecnológicas, para la construcción de soluciones que permitan apalancar la creación de nuevas técnicas, una aportación ideal a este Sector Salud sería enfocarse a las Políticas de Salud Pública de manera sistemática, teniendo la oportunidad de hacer cambios y mejoras, para que permitan escalar diversos entornos y satisfacer las cambiantes necesidades en el sector salud en tiempo real y servir entonces para apoyar la labor médica.

La cantidad de datos médicos recopilados en historiales clínico existe y va a aumentar de forma exponencial, esto abre una puerta al conocimiento, al uso de Machine Learning para aplicar analítica avanzada.

La forma de diagnosticar y prevenir enfermedades evolucionará radicalmente. Debido a las nuevas tecnologías de computación, se tiene la habilidad de aplicar automáticamente cálculos matemáticos complejos a grandes cantidades

de información de manera repetida y cada vez más rápida, siendo su factor principal, la construcción de modelos automáticos en tiempo real.

Para posibles asociaciones sobre los datos se pueden rehusar los modelos obtenidos e integrarlos para obtener patrones globales, los casos de uso muestran que se obtuvieron modelos precisos.

Lo que hace rápido el proceso es que, durante el análisis, se centra en la tarea principal y no en la recolección de datos. Sin duda, el Machine Learning es uno de los protagonistas de este universo digital, en donde el procesamiento de los datos y el resguardo de la información son objetivos elementales.

Una restricción en relación a los datos, es en torno a la cantidad, velocidad y tipos de datos, la mayoría de las veces no están automatizados, presentándose como solución el manejo de las herramientas tecnológicas, que están diseñadas para el procesamiento de la información proveniente de diferentes sentidos, combinarlos con conocimientos y experiencias previas y llegar a conclusiones.

Predecir los resultados basados en el conocimiento de la observación de los datos, se ha convertido en trabajo muy productivo, en una alternativa en áreas de la investigación tecnológica, por lo cual el uso de esta herramienta que utilicen técnicas y algoritmos de ML traería consigo el ahorro de recursos para el Sector Salud.

Referencias

Asri, H., Mousannif, H., Moatassime, A. H., & Noeld, T. (2016). Using Machine Learning Algorithms for Breast Cancer Risk Prediction and Diagnosis. *Procedia Procedia Computer Science. ElSevier*, 1064-1069.

Avila Camacho, J. (01 de 05 de 2020). *JacobSoft*. Obtenido de K-Nearest Neighbors: https://www.jacobsoft.com.mx/es_mx/k-nearest-neighbors/

Bironneau, M., & Coleman, T. (01 de 05 de 2019). *Packt*. Obtenido de Types of ML algorithms: https://subscription.packtpub.com/book/big_data_and_business_intelligence /9781838550356

Challenger Pérez, I., Díaz Ricardo, Y., & Becerra García, R. A. (2014). El lenguaje de programación Python. *Ciencias Holguín. Centro de Información y Gestión Tecnológica. Santiago de Cuba*, 13.

Contreras Morales, E. F., Ferreira Correa, M., & Valle, M. A. (2017). Diseño de un modelo predictivo de fuga de clientes utilizando árboles de decisión. *Revista Ingeniería Industrial-Año 16 Nº1:*, 7-23.

De la Hoz Manotas, A. K., Martínez Palacio, U. J., & Enrique, M. P. (2013). Técnicas de ML en medicina cardiovascular. *Memoria Desarrollo Humano 11/20*, 41-46.

Delgado, R., & Tibau, X. A. (2015). Las Redes Bayesianas como herramienta para la evaluación del riesgo de reincidencia: Un estudio sobre agresores sexuales. *Revista Española de Investigación Criminológica*, 25.

Didactalia classes. (19 de 11 de 2018). *Didactalia classes*. Obtenido de Regesión Lineal Simple: https://didactalia.net/en/community/materialeducativo/resource/calculadora-de-regresion-lineal-simple---recta-de/a354dc49-3a49-47c4-92ba-3b720337ee11

González Vilanova, A. (2019). Métodos de machine learning en estudios biomédicos. *Trabajo final de grado en Biotecnología. Universitat Politécnica de València – Escola Tècnica Superior*. Valencia, Valencia, España: Universitat Politécnica de València – Escola Tècnica Superior.

González, L. (22 de 03 de 2018). *González, Ligdi*. Obtenido de ¿Por qué Machine Learning o aprendizaje automático es importante?: https://ligdigonzalez.com/introduccion-a-machine-learning/

González, L. (09 de 07 de 2018). *Ligdi González. Inteligencia Artificial*. Obtenido de Lenguajes de programación para Machine Learning: https://ligdigonzalez.com/lenguajes-de-programacion-para-machine-learning/

González, L. (28 de 06 de 2019). *Regresión Logística – Teoría*. Obtenido de Regresión Logística – Teoría: https://ligdigonzalez.com/regresion-logistica-multiple-machine-learning-teoria/

Lantz, B. (2013). *Machine Learning with R*. Lead Technical Editor;.

Leivi, A. E. (22 de 07 de 2019). Análisis de la implementación de Machine Learning en el diagnóstico por imágenes. *Trabajo de Grado de Maestría en Gestión de Servicios Tecnológicos y de Telecomunicaciones*. Victoria, Buenos Aires, Argentina: Universidad de San Andrés Escuela de Negocios.

Luna González, J. (08 de 02 de 2018). *SoldAI*. Obtenido de Tipos de aprendizaje automático: https://medium.com/soldai/tipos-de-aprendizaje-autom%C3%A1tico-6413e3c615e2

Luque Sucasaire, N. L. (01 de 2020). Análisis de sistemas para registros médicos electrónicos en clínicas y su enfoque al Machine Learning. *Grado Académico de Bachiller en Ingeniería Industrial*. San Pablo, Arequipa, Perú: Facultad de Ingeniería y Computación. Escuela Profesional de Ingeniería Industrial. Universidad Católica San Pablo.

Mohri, M., Afshin, R., & Ameet, T. (2018). *Foundations of Machine Learning, Second Edition.* London: Francis Bach, Editor. The MIT Press. Massachusetts Institute of Technology.

Montes Núñez, B. R. (11 de 05 de 2017). *Gfi Blog. New Challenges, New Ideas*. Obtenido de Algoritmos de entrenamiento en Machine Learning: https://blog.gfi.es/algoritmos-entrenamiento-machine-learning/

Moscadó Pérez, J. C. (01 de 12 de 2018). *Aprendizaje supervisado para la detección de amenzas WEB. Universitat Oberta de Catalunya.* Obtenido de Aprendizaje supervisado para la detección de amenzas WEB: http://openaccess.uoc.edu/webapps/o2/bitstream/10609/91066/6/jmoscardoTFM0119memoria.pdf

Parra, F. (25 de 01 de 2019). *Estadística y Machine Learning con R.* Obtenido de https://bookdown.org/content/2274/portada.html

Ramasubramanian, K., & Moolayil, J. (01 de 04 de 2019). *Packt*. Obtenido de Applied Supervised Learning with R: https://subscription.packtpub.com/book/big_data_and_business_intelligence_/9781838556334

Rassiga, P. (2017). Indice financiero multidimensional: una aplicación de componentes principales. Buenos Aires , Buenos Aires, Argentina: Universidad de San Andrés. Departamento Académico de Economía.

Revuelta Briz, R. (01 de 09 de 2018). Trabajo de fin de grado en INGENIERÍA INFORMÁTICA. *Aplicación de Técnicas de Machine Learning, Un caso práctico*. Cantabria, Cantabria, España: Universidad de Catabria. Facultad de Ciencias.

Ruiz Ruano, A. M., & Puga, J. L. (2016). R COMO ENTORNO PARA EL ANÁLISIS ESTADÍSTICO EN EVALUACIÓN PSICOLÓGICA. *Sección Monográfica. Papeles del Psicólogo*, 74-79.

Sáez de la Pascua, A. (31 de 01 de 2019). Deep learning para el reconocimiento facial de emociones básicas. *Grado en Ingeniería de Sistemas de Telecomunicaciones*. Catalunia, Catalunia, España: Universidad Politécnica de Catalunia.

Capítulo 10.

Análisis de Datos Geoespaciales en Protección Civil utilizando R y Python

Armando Urbina Retana
Tecnológico Nacional de México. Instituto Tecnológico de Durango
93040929@itdurango.edu.mx

Rubén Pizarro Gurrola
Tecnológico Nacional de México. Instituto Tecnológico de Durango
rpizarro@itdurango.edu.mx

10.1. Introducción

Este capítulo describe a la Ciencia de Datos en relación a un Sistema de Información Geográfica (GIS) para Protección Civil del Municipio de Durango con base al análisis de datos geoespaciales utilizando R y Python integrados para establecer un proceso de actualización de información cartográfica.

Constantemente se están generando datos, por lo tanto, el gobierno tiene que estar a la vanguardia. Se tiene que pensar en los datos como un conjunto de información sobre los cuales se obtienen para toma de decisiones o conclusiones.

Esto no es fácil para la información que se utiliza en la Dirección Municipal de Protección Civil, dado que la información que se tiene para el análisis de datos para el sistema de información geográfica no siempre son el total de los que desea estudiar, o son demasiado grandes para utilizar métodos comunes en estadística.

Otro problema es que los fenómenos que se pretenden estudiar no siempre corresponden a una distribución normal, por lo que métodos estadísticos comunes no son aplicables; es necesario hacer notar que existe una amplia gama de análisis que pueden ser aplicados a los datos.

Por lo anterior, la Ciencia de Datos y el Big Data se han convertido en un poderoso instrumento que intersecta estadística y los sistemas de información geográfica; por lo tanto, son conceptos importantes para el objetivo de éste producto académico.

Los fenómenos naturales de carácter destructivo siempre han aparecido de una forma recurrente e inevitable. En el territorio del Municipio de Durango, los fenómenos naturales han provocado variaciones al paisaje y algunos casos pérdidas económicas por daños a la infraestructura, pérdida de materiales como pérdidas humanas. Sin embargo, en los últimos años, estos fenómenos aunados con procesos de expansión urbana, pueden incrementar la magnitud de devastación.

La Protección Civil es un servicio público cuyo objetivo es prevenir las situaciones de grave riesgo colectivo o catástrofes, proteger y socorrer a las personas, los bienes y el medio ambiente, cuando dichas situaciones se producen, así como contribuir a la rehabilitación y reconstrucción de las áreas afectadas.

Actualmente en Protección Civil Municipal de Durango aún no posee un proceso para la publicación y actualización de la información cartográfica, lo que con lleva a que se realice la integración de un sistema de información basado en análisis de datos georreferenciados sobre el riesgo de desastres detallado a nivel municipal.

Agregado a esto, los desarrollos de estos sistemas de información cartográfica son realizados por software privativo por ejemplo ArtGis y Oracle, el cual representa un costo muy elevado.

Un Sistema de Información Geográfica se puede definir que está formado por cinco componentes: personal capacitado, datos geoespaciales y descriptivos, métodos analíticos, hardware y software.

El presente capítulo consiste en realizar un análisis de los datos geoespaciales utilizando software libre integrando R y Python con Sistema de Información Georreferenciado de software libre (QGIS) proponiendo que el sistema incluya base de datos, sistemas de tratamientos de información geográfica, tecnologías de desarrollo web, servidores web, que en conjunto permitan contar con un proceso de actualización de la información cartográfica; reduciendo el tiempo de actualización y publicación.

Como objetivo general se propone un Sistema de Información Geográfica que incluya la funcionalidad de análisis de datos geoespaciales utilizando R y Python para visualizar mapas cartográficos que muestren zonas de riesgos conforme a los cruces de datos.

De manera específica se busca lo siguiente:

- Describir las funciones básicas de Protección Civil.
- Investigar la herramienta de visualización de información geográfica (QGIS).
- Presentar casos de uso de sistemas geoespaciales y algunas funciones
- Describir las tecnologías de desarrollo a utilizar: R y Python en relación a QGIS
- Presentar y analizar propuesta

En la actualidad hay más facilidades de acceso a la información pública; es decir, se pueden consultar a través de una conexión de internet; por lo tanto, un sistema integral de información permite establecer bases de datos y realizar el análisis del peligro, de la vulnerabilidad y del riesgo antes desastres a escala nacional, regional, estatal y municipal, con objeto de generar mapas y sistemas geográficos de información. Con ello se estará en posibilidad de simular escenarios de desastres, emitir recomendaciones para la oportuna toma de decisiones y establecer efectivas medidas de prevención y mitigación.

Además, al desarrollarse e implementarse con software libres tendría costos bajos en relación a costos de licenciamiento, debido a que existen herramientas dentro de esta clasificación de software que proporciona las mismas, e incluso mejores funcionalidades, necesarias en comparación con el software propietario.

10.2. Marco de referencia

En este apartado, se mencionan aspectos y conceptos de Ciencia de los Datos y Big Data, los Sistemas de Información geográfica (GIS), de la unidad de Protección Civil y la información que manejan, herramientas para el desarrollo de proyectos (GIS), el caso de España y el caso del Municipio de Durango y sus distintos fenómenos naturales que le afectan.

10.2.1. Ciencia de datos y Big Data.

Al hablar de Ciencia de los Datos es necesario considerar varios elementos que conforman su arquitectura, la figura 1, identifica algunos de ellos: las fuentes u orígenes que producen los datos, distintas disciplinas en procesos de análisis, transformación y modelado de datos para finalmente generar ideas y generar información y conocimiento para tomar decisiones.

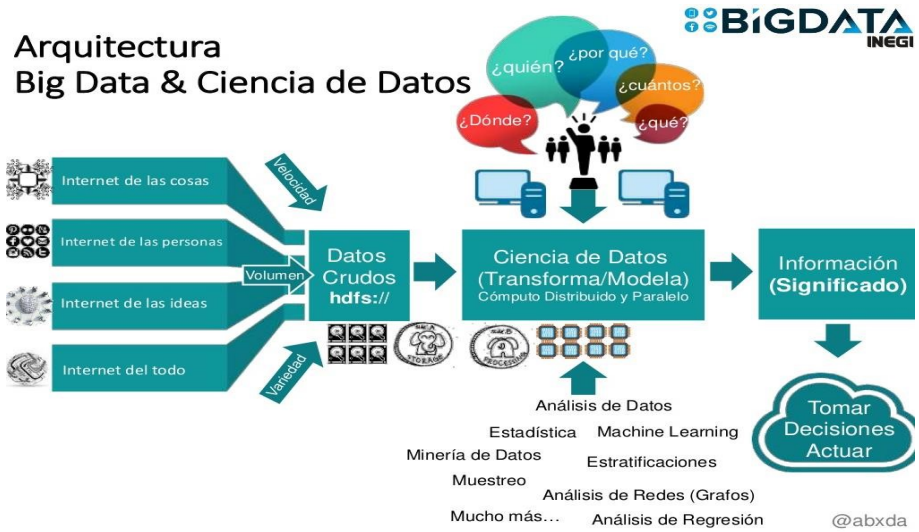


Figura 1. Ciencia de los Datos y Big Data (Coronado Iruegas, 2016)

Ciencia de Datos es un área de trabajo multidisciplinar que incluye procesos para obtener, preparar, limpiar, transformar, analizar, visualizar y modelar datos que generan conocimiento para entender problemas y oportunidades con la finalidad de apoyar el proceso de toma de decisiones. Los datos pueden ser estructurados, no estructurados además de heterogéneos. (Centro Mediterraneo. Univesidad de Granada, s.f.).

En ocasiones, se trata de grandes volúmenes de datos que por su complejidad se necesita de algoritmos y técnicas contemporáneas para obtener el conocimiento. La ciencia de Datos es un campo innovador con una alta aplicabilidad en ciencias de la salud, industria, servicios, educación, marketing, negocios, mercados financieros, transporte, comunicaciones, redes sociales, investigación, entre muchos otros. (Centro Mediterraneo. Univesidad de Granada, s.f.)

Big Data es un término ligado a la ciencia de datos para referirse a la información o grupo de datos que por su elevado volumen, diversidad y complejidad no pueden ser almacenados ni visualizados con herramientas tradicionales o, desde otra perspectiva, este anglicismo se emplea para describir el conjunto de procesos, tecnologías y modelos de negocio que están basados en la generación y recolección de datos y en capturar el valor que los propios datos encierran.

10.2.2. Panorama Tecnológico de la Ciencia de los Datos.

En la figura 2, se identifican tecnologías que permiten el tratamiento y procesamiento de los datos; en la parte izquierda las fuentes de los datos que son de diversa índole; al centro la infraestructura de cómputo, bases de datos, procesamiento distribuido y paralelo e identifica abajo el término de Big Data; a la derecha disciplinas y herramientas tales como estadística, análisis multivariado, Machine Learning. Se identifican a la derecha también, dos lenguajes de programación ideales para la Ciencia de los Datos, R y Python.

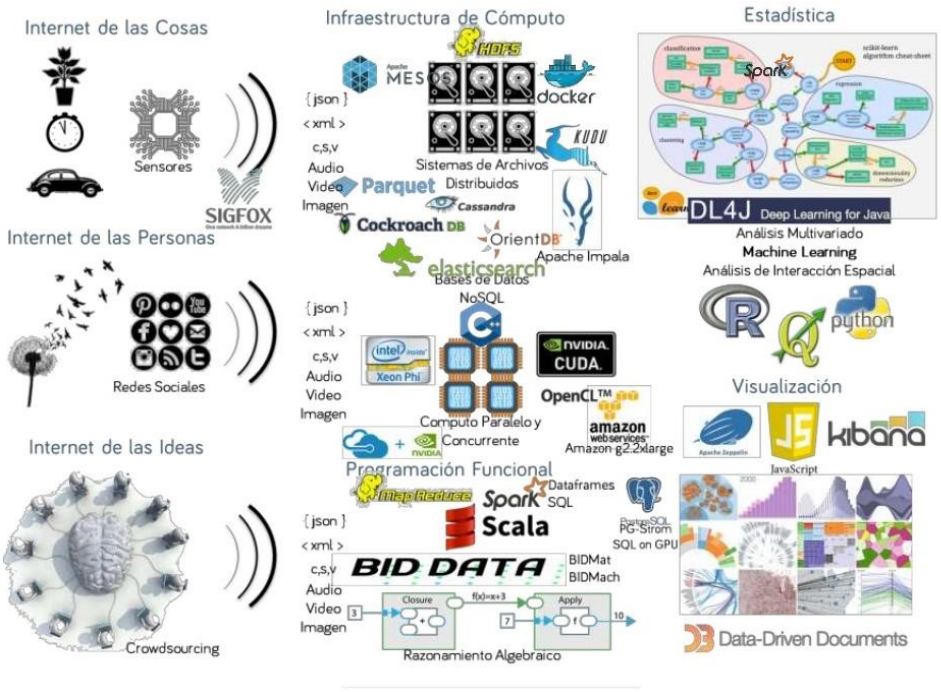


Figura 2. Panorama Tecnológico de la Ciencia de los Datos (Coronado Iruegas, 2016)

La Ciencia de los Datos representa una gran revolución para la innovación, la competitividad y la productividad, así como para la mejora de los servicios públicos, y gobiernos, empresas e instituciones deben estar diseñadas y operadas para aprovechar los datos digitales en la optimización, transformación y mejora de sus productos y servicios.

Big Data como herramienta y parte de la Ciencia de los Datos supone algunos aspectos:

- Apoyar el proceso de toma de decisiones a las organizaciones: almacenando, procesando y analizados datos para generar ideas innovadoras.
- Mejorar los procesos de retroalimentación de manera más oportuna y eficiente obteniendo valor y veracidad. Por ejemplo, obtener información a tiempo real del lanzamiento de un producto o anticipar el resultado de una nueva estrategia
- Conocimiento del mercado. No solo se trata de datos internos sino que al integrar datos externos ayuda a conocer a los clientes y adelantarse a sus necesidades, tendencias de consumo o deficiencias en la atención al cliente.

- Tecnología del presente y del futuro. La Ciencia de los Datos es un elemento diferenciador y de éxito entre las organizaciones. La tecnología del Big Data está en constante cambio y se visualiza que tendrá un rol todavía más importante en la toma de decisiones venideras (Universidad de Alcalá, s.f.).

10.2.3. Impacto y beneficios de aplicar la Ciencia de Datos

En el mundo actual de la tecnología que rodea a todos, la computación en la nube se va haciendo parte del quehacer diario, los servicios de Google, Facebook, Twitter, Dropbox, o , entre otros son cada vez más utilizados. Existe una mayor cantidad de dispositivos que están las 24 horas del día conectadas a internet. desde teléfonos, tabletas y TVs hasta automóviles incorporando en las actividades el paradigma de Internet de las cosas (IoT). En el mundo, se están generando datos constantemente, aquí entra la filosofía del Big Data; se está haciendo cada vez más necesario un nuevo perfil de profesionales de la información que puedan aplicar las técnicas, logaritmos y herramientas de la Ciencia de Datos.

Como lo menciona Jones (2019), la ciencia de datos es un arte. Los científicos de datos utilizan diferentes herramientas para lograr su tarea. A pesar de que todas estas herramientas son conocidas por la computadora, el científico de datos tiene la función de encontrar una manera en la que él o ella puedan usar todas las herramientas o las más adecuadas e integrarlas a los datos para desarrollar la respuesta correcta a una pregunta (Jones, 2019).

10.2.4. Sistema de Información Geográfico (GIS)

Un Sistema de Información Geoespacial contiene una recopilación y una generalidad de los conocimientos científicos contemporáneos en el campo de la geografía física, económica, cultural y política del área considerada.

Sirve como herramienta de consulta al añadir un valor esencial para el gestor público y al proporcionar el conocimiento de los diferentes aspectos que caracterizan el territorio en las actividades empresariales, al mismo tiempo constituye un punto de referencia para el público en general.

Un GIS es una aplicación computacional que integra datos geográficos con información descriptiva. Se usan para analizar y visualizar información espacio-físico-temporal; relaciona información en un contexto espacial y obtener aspectos de sus relaciones para la toma de decisiones (Sánchez Fleitas, Comas Rodríguez, & García Lorenzo, 2019).

Un Sistema de Información Geográfica o SIG (GIS en su acrónimo inglés, Geographic Information Systems) es un conjunto de componentes específicos que permiten a los usuarios crear consultas, integrar, analizar y representar de una forma eficiente cualquier tipo de información geográfica referenciada asociada a un territorio.

Por otra parte, el uso SIG facilita la visualización de los datos obtenidos en un mapa con el fin de reflejar y relacionar fenómenos geográficos de cualquier tipo. Además, permite realizar las consultas y representar los resultados en entornos web y dispositivos móviles en un modo ágil e intuitivo, con el fin de resolver problemas complejos de planificación y gestión, conformándose como un valioso apoyo en la toma de decisiones.

Los SIG conectados a la web es un desarrollo que permite explorar las ventajas existentes en la consulta de información geo referenciada. A través de los Web Map Service (WMS) estándar que permite publicar mapas según especificaciones OGC y Web Feature Service (WFS) que ofrece una interfaz de comunicación que permite interactuar en forma de consulta con los mapas mostrados por el estándar WMS.

10.2.5. Bases de Datos para GIS

Una base de datos para un sistema de información geográfico contiene dos principales componentes que son los datos espaciales y datos no espaciales (atributos). Se enlistan conceptos importantes para su uso dentro de las Bases de Datos Geográfica:

- Datos espaciales: se refiere a las particularidades geográficas de la ubicación descrita, es decir, los puntos geo localizados que conforman el perímetro de una

localidad o población, que estos a su vez se almacenan en cierto tipo de archivo que son interpretados por las aplicaciones geográficas.

- Datos no Espaciales: también llamados atributos, son los datos descriptivos que se almacenan en tablas y se administran por un manejador de base de datos.
- Capa Geográfica: son el mecanismo que se utiliza para visualizar un conjunto de datos geográficos en determinado software que sea capaz de visualizar este tipo de información.

Para tener una mejor comprensión en la información geográfica se debe de basar en tres expresiones fundamentales que son:

- ***Entidades:*** son representaciones de cosas ubicadas en la superficie de la Tierra o cerca de ella; se pueden presentar de forma natural (ríos y vegetación), pueden ser construcciones (carreteras, canalizaciones, pozos y edificios) o subdivisiones de tierra (condados, divisiones políticas, parcelas de terreno). Las entidades a su vez también se representan de tres formas:

i. *Puntos:* definen ubicaciones de entidades geográficas demasiado pequeñas por ejemplo ubicaciones de pozos, postes de teléfono, estaciones hidrométricas, direcciones, coordenadas GPS, picos de montañas, entre otros.

ii. Líneas*:* representan la forma y la ubicación de objetos geográficos demasiado estrechos para mostrarse como áreas, como por ejemplo centro de calle, arroyos, curvas de nivel y límites administrativos.

iii. Polígonos*:* son áreas cerradas, suelen ser figuras de muchos lados, que representan la forma y la ubicación de entidades como estados, condados, parcelas, tipos de suelo y zonas de uso del suelo.

- ***Atributos:*** son la información descriptiva que se manifiesta a través de símbolos de mapa, colores y etiquetas, por ejemplo, las carreteras que se muestran en función de su clase (si son autopistas, calles principales, caminos de terracería, pistas y calles residenciales)

- ***Imágenes:*** se hace referencia a una serie de tipos fuentes de datos basadas en celdas o en píxeles para satélites, fotografía aérea, modelos digitales de elevación, conjunto de datos raster (imágenes, fotografías), entre otros.

10.2.6. SIG con R

R es un entorno de software libre y lenguaje de programación que ofrece varias opciones, herramientas y bibliotecas para hacer análisis estadístico fácil y eficaz. Ha crecido en los últimos años gracias a su código abierto y es uno de los lenguajes de programación más utilizados en investigación por la comunidad estadística en un entorno de ejecución con gráficos, un depurador, acceso a ciertas funciones del sistema y la capacidad de ejecutar programas almacenados en archivos de script.

Las tecnologías utilizadas para la ciencia de los datos han mostrado un crecimiento particularmente rápido en los últimos años. En la figura 3, se indica el crecimiento de R con respecto a ser un lenguaje de programación para SIG.

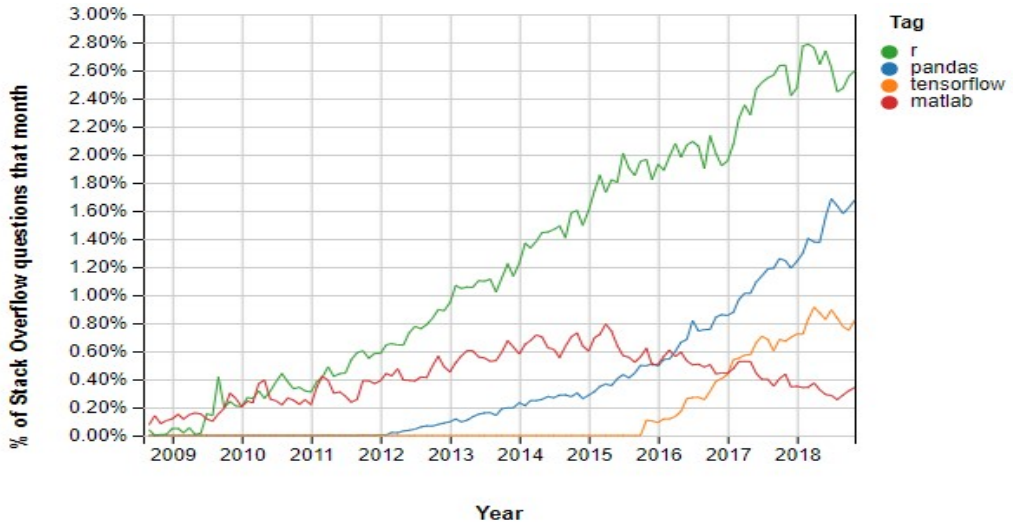


Figura 3. Gráfica comparativa de la creciente popularidad de R contra otros softwares. (Morales, 2018)

Como se puede ver en la figura 3, la popularidad de R ha crecido en los últimos años gracias a su código abierto. Es un lenguaje basado en la comunidad que proporciona poderosas herramientas para procesamiento, manipulación, visualización y publicación.

La interfaz del núcleo R es una ventana de línea de comandos que proporciona una excelente flexibilidad y control, que puede parecer poco amigable a usuarios acostumbrados a utilizar programas computacionales con menú y opciones seleccionadas con el ratón.

Los scripts son un archivo de texto con una serie de instrucciones; estos usualmente utilizados a través del IDE (Integrated Develpment Eviroment, entorno de desarrollo integrado). R Studio es un IDE para programar en R.

La relación entre el lenguaje de programación R y los Sistemas de Información Geográfica es natural. R se complementa con un SIG y especialmente útil para explotar grandes conjuntos de datos. Algunas características de R aplicado a SIG son:

- Clases para datos espaciales.
- Lectura y escritura de datos espaciales.
- Análisis de patrones de puntos.
- Geoestadística.
- Regresión espacial.
- Análisis ecológico.
- Algoritmos de procesamiento ráster.
- Detección remota.
- Teledetección y LiDAR (light detection and ranging).

10.2.6.1. Paquetes de R para trabajar con datos espaciales

Se han desarrollado paquetes modulares (actualmente son más de 13,500), que son complementos para temas específicos para la comunidad de desarrolladores.

En R existen cientos de paquetes que se pueden emplear para manejar información geográfica, se destacan algunos:

- Datos vectoriales: mediante el paquete *sf* y el paquete *sp.*
- *Datos ráster*: mediante el paquete ráster y stars.

- *rgdal o maptools*: se trata de un conjunto de herramientas para gestionar datos geográficos útiles, por ejemplo, importar archivos shapefile.
- *maps*: para visualizar mapas.
- *whitheboxR*: Este repositorio incluye 400 herramientas para realizar análisis geoespacial. Es la interfaz de R del programa de línea de comandos WhiteboxTools. Incluye herramientas para trabajar con teledetección y LiDAR.

El paquete sf (*Simple Features for R)* representa los objetos geográficos como objetos nativos de R utilizando estructuras de datos simples (listas, matrices, vectores…).

- Todas las funciones y métodos del paquete sf que emplean datos espaciales tienen el prefijo st_, que se refiere a espacial y temporal (al igual que en PostGIS).
- Los objetos geográficos son *data.frames* o *tibbles* con una columna de geometrías.
- Representa de forma nativa en R los 17 tipos de objetos geográficos simples para todas las dimensiones (XY, XYZ, XYM, XYZM)
- Interfaces a GEOS para soportar el modelo topológico DE9-IM.
- Se conecta a GDAL, es compatible con todas las opciones de controlador, columnas de fecha (*Date*) y fecha y hora (*DateTime*) (POSIXct) y transformaciones del sistema de coordenadas de referencia a través de PROJ.
- Utiliza serializaciones WKB (well-know-binary conocidas escritas en C++/Rcpp para fast I/O con GDAL y GEOS.
- Lee y escribe directamente en bases de datos espaciales como PostGIS usando DBI.

10.2.7. SIG con Python

Python es un lenguaje extremadamente útil para aprender en términos de SIG, ya que muchos (o la mayoría) de los diferentes paquetes de software de SIG (como ArcGIS, QGIS, PostGIS, entre otros, proporcionan una interfaz para hacer

análisis utilizando secuencias de comandos de Python. La figura 4 muestra el comparativo Python vs Java script en términos de SIG.

Figura 4. Gráfica comparativa de la creciente popularidad de Python contra JavaScript (Morales, 2018)

Como se puede ver en la figura anterior la popularidad de Python en los últimos años ha ido en aumento y cada vez son más los usuarios de estos programas, que, de modo sencillo, aprenden Python y se benefician de su utilización.

10.2.7.1. Paquetes de Python para análisis de datos y visualización

Este apartado cita algunos paquetes y librerías utilizadas en Python que permiten el análisis y la visualización de datos:

- Numpy: Paquete fundamental para computación científica con Python.
- Pandas: Estructuras de datos y herramientas de análisis de datos de alto rendimiento y fáciles de usar.
- Scipy: Una colección de algoritmos numéricos y cajas de herramientas específicas de dominio, que incluyen procesamiento de señales, optimización y estadísticas.

- Matplotlib: Biblioteca básica de trazado para Python. ○ Bokeh: Visualizaciones iterativas para la web (también mapas). ○ Plotly: Visualizaciones interactivas (también mapas) para la web.

10.2.7.2. Paquetes para SIG

Se citan de igual forma paquetes que utiliza Python que permite la integración con Sistemas de Información Geográficos y que soportan el formato de datos cartográficos y geoespaciales.

- GDAL: Paquete fundamental para procesar formatos de datos vectoriales y ráster (varios de los módulos dependen de esto). Utilizado para el procesamiento de trama.
- Geopandas: Trabajar con datos geoespaciales en Python de manera más fácil, combina las capacidades de los pandas y las formas.
- Shapely: Paquete Python para manipulación y análisis de objetos geométricos planos (basado en GEOS).
- Fiona: Lectura y escritura de datos espaciales (alternativa para geopandas).
- Pyproj: Realiza transformaciones cartográficas y cálculos geodésicos (basado en PROJ.4) o Pysal: Biblioteca de funciones de análisis espacial escritas en Python.
- Geopy: Biblioteca de geocodificación coordenadas a la dirección <-> dirección a las coordenadas.
- GeoViews: Mapas interactivos para la web.
- Geoplot: Biblioteca de visualización de datos geoespaciales de alto nivel para Python.
- Dash: es un marco de Python para crear aplicaciones web analíticas.
- OSMnx: Python para redes de calles. Recupere, construya, analice y visualice redes de calles de OpenStreetMap.
- Nettworkx: Análisis de red y enrutamiento de Python, por ejemplo, algoritmos Dijkstra y A*(A estrella es un algoritmo de búsqueda de primer orden).
- Cartopy: Haga que los mapas de dibujo para el análisis y visualización de datos sea lo más fácil posible.

- Scipy.spatial: Algoritmos espaciales y estructuras de datos. o Rtree: Indexación espacial para Python para búsquedas espaciales rápidas. o Rasterio: E / S raster limpio y rápico y geoespacial para Python. o RSGISLib: Teledetección y biblioteca de software SIG para Python.

10.2.8. Herramientas Webmapping

Web mapping es un concepto que en español se traduce como "cartografía en la web", este se refiere al proceso de diseñar, aplicar, generar, visualizar u ofrecer datos geoespaciales a través de la red de internet.

En este apartado se puntualizarán, de manera breve, algunas de las herramientas que aportan las actividades del *web mapping*.

- Mapbender es un software de mapeo web implementado en PHP y
- JavaScript, su configuración se basa principalmente en un modelo de datos almacenado en una base de datos PostgreSQL PostGIS o MySQL. Está desarrollado con código abierto y con licencia de la GNU GPL como software libre.
- OpenLayers es una biblioteca de JavaScript de código abierto bajo una derivación de la licencia BSD para mostrar mapas interactivos en los navegadores web.

10.2.9. Ciencia de los Datos y los datos geoespaciales.

Ciencia de los Datos es una combinación de varias herramientas, algoritmos y principios de aprendizaje automático con el objetivo de descubrir patrones ocultos a partir de los datos.

Cada vez más, los conjuntos de datos con localización son de un tamaño, variedad y tasa de actualización que excede la capacidad de las tecnologías de computación enfocada a los procesos geoespaciales, estos datos se llaman *GeoSpatial Big Data* (GSBD por sus siglas en inglés) o *Datos Masivos Geoespaciales* (DMG).

La gran diversidad de fuentes de datos masivos geoespaciales aumenta sustancialmente la diversidad de métodos de solución. Los nuevos algoritmos

pueden surgir a medida que se encuentren disponibles los conjuntos de datos masivos geoespaciales y de esta manera se crea la necesidad de una arquitectura flexible para integrar rápidamente nuevos conjuntos de datos y algoritmos asociados (Vázquez Pulido & Morales Bautista, 2019).

En la figura 5 Carranza (2016), menciona: muchas organizaciones orientadas a datos, han enfatizado la necesidad de incluir y potenciar en este proceso a las personas partiendo de sus necesidades y problemas. Esto exigen una revolución de datos, siempre que implique la transformación de paradigmas para la recopilación de datos demográficos, su organización y visualización para los gobiernos, la academia y el sector civil en todos los países. (Carranza Tresoldi, 2016)

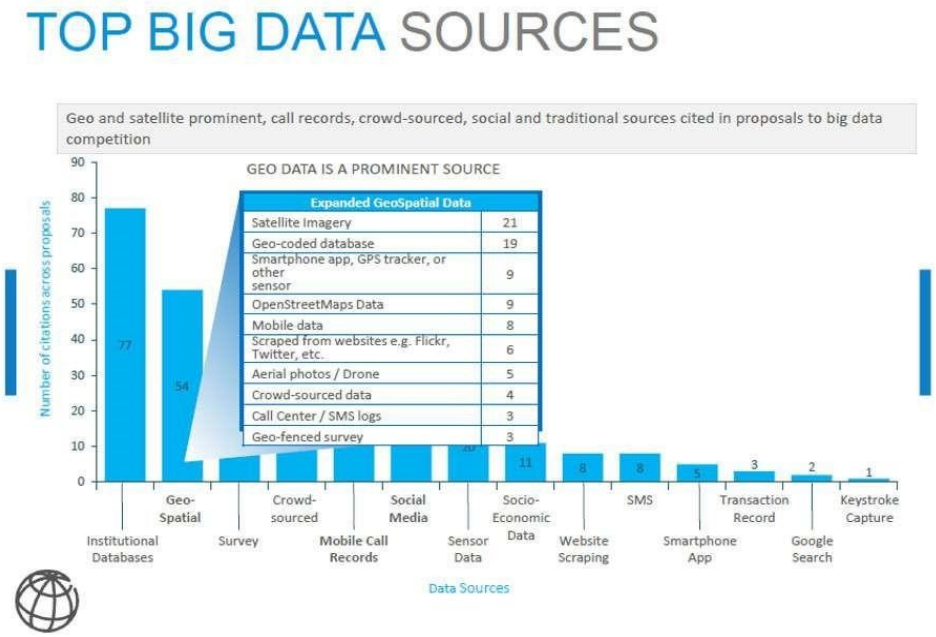


Figura 5 Top Big Data Source (Carranza Tresoldi, 2016).

Los datos geoespaciales describen objetos y cosas con relación al espacio geográfico, con coordenadas de ubicación en un sistema de referencia espacial. Estos datos tradicionalmente se recopilan mediante la localización geográfica (coordenadas x,y,z), la topografía terrestre, la fotogrametría, la teledetección y más recientemente, mediante el escaneo láser, la cartografía móvil, los contenidos geo-

etiquetados, la información geográfica participativa o colaborativa, los sistemas mundiales de navegación por satélite y sensores geoposicionados.

Se consideran como origen de los datos masivos geoespaciales, los siguientes:

- Imágenes de satélite.
- Datos existentes (BD, texto, video, fotos).
- Dispositivos móviles.
- GPS (*Global Positioning System*).
- Sensores (llantas, motores, RFID, tanques, válvulas).
- Internet de las cosas.
- Plataformas específicas (*waze, twitter, openstreetmaps, google maps*).
- Radar.
- LiDAR.Foto y video (ambos con geoposición) que proviene de sensores colocados sobre drones.
- Crowsourcing.

10.2.10. Protección Civil.

La historia de México ha mostrado las bondades culturales y sociales de un país tan diverso como complejo. Se ha visto el crecimiento y el desarrollo de este país a partir de la adaptación a los cambios que el transcurrir del tiempo exige; sin embargo, este crecimiento ha provocado también que la población viva entre riesgos que pasan inadvertidos, no sólo por la ausencia de información, sino también por la falta de medidas preventivas que ayudan a aminorar los riesgos a los que el ciudadano está expuesto.

Dada la ubicación geográfica de México en el mundo, se enfrenta al impacto de fenómenos naturales y humanos que han dejado a su paso importantes pérdidas materiales y humanas.

Experiencias como los sismos de 1985, la erupción del volcán Chichonal en 1982, frecuentes inundaciones en el sureste del país o la muerte de aficionados en

el Estadio Olímpico Universitario en 1985, entre otros dan muestran las vulnerabilidades.

El concepto de riesgo se encuentra ligado directamente a tres factores: peligro, exposición y vulnerabilidad, por lo que su conjunción depende de estos, ya que, si alguno no existe, el riesgo sería inexistente.

El peligro se relaciona con la probabilidad de ocurrencia de un fenómeno perturbador. La vulnerabilidad es la propensión al daño de un sistema expuesto, sea este topo físico (como la infraestructura) o social. Por su parte, la exposición se relaciona directamente con el valor que se asigne a la población, bienes y entorno que estén expuesto a un peligro o fenómeno perturbador.

Llevar el estudio de riesgo a un nivel muy detallado implica: integrar, múltiples y complejas fuentes de información, contar con gran capacidad de almacenamiento y procesamiento, generar mecanismos que garanticen su seguridad, difundir los resultados al público en general y tomar decisiones de manera oportuna.

Las herramientas para el diagnóstico de riesgo se enfocan principalmente a las tecnologías de la información, y en específico con los sistemas de información geográfica, la percepción remota, los sistemas de geoposicionamiento global además del desarrollo de aplicaciones específicas para la generación de escenarios de riesgo.

Por otra parte, es necesario mencionar que existen una taxonomía para los riesgos presentes establecidos en la Ley de Protección Civil (Cámara de Diputados del H Congreso de la Unión, 2019), los cuales son:

a. Fenómeno Antropogénico: Agente perturbador producido por la actividad humana.
b. Fenómeno Astronómico: Eventos, procesos o propiedades a los que están sometidos los objetos del espacio exterior incluidos estrellas, planetas, cometas y meteoros. Algunos de estos fenómenos interactúan con la tierra, ocasionándole situaciones que generan perturbaciones que pueden ser

destructivas tanto en la atmósfera como en la superficie terrestre, entre ellas se cuentan las tormentas magnéticas y el impacto de meteoritos.

c. Fenómeno Natural Perturbador: Agente perturbador producido por la naturaleza.
d. Fenómeno Geológico: Agente perturbador que tiene como causa directa las acciones y movimientos de la corteza terrestre. A esta categoría pertenecen los sismos, las erupciones volcánicas, los tsunamis, la inestabilidad de laderas, los flujos, los caídos o derrumbes, los hundimientos, la subsidencia y los agrietamientos.
e. Fenómeno Hidrometeorológico: Agente perturbador que se genera por la acción de los agentes atmosféricos, tales como: ciclones tropicales, lluvias extremas, inundaciones pluviales, fluviales, costeras y lacustres; tormentas de nieve, granizo, polvo y electricidad; heladas; sequías; ondas cálidas y gélidas; y tornados.
f. Fenómeno Químico-Tecnológico: Agente perturbador que se genera por la acción violenta de diferentes sustancias derivadas de su interacción molecular o nuclear. Comprende fenómenos destructivos tales como: incendios de todo tipo, explosiones, fugas tóxicas, radiaciones y derrames.
g. Fenómeno Sanitario-Ecológico: Agente perturbador que se genera por la acción patógena de agentes biológicos que afectan a la población, a los animales y a las cosechas, causando su muerte o la alteración de su salud. Las epidemias o plagas constituyen un desastre sanitario en el sentido estricto del término. En esta clasificación también se ubica la contaminación del aire, agua, suelo y alimentos.
h. Fenómeno Socio-Organizativo: Agente perturbador que se genera con motivo de errores humanos o por acciones premeditadas, que se dan en el marco de grandes concentraciones o movimientos masivos de población, tales como: demostraciones de inconformidad social, concentración masiva de población, terrorismo, sabotaje, vandalismo, accidentes aéreos, marítimos o terrestres, e interrupción o afectación de los servicios básicos o de infraestructura estratégica.

Es por esto que surge la necesidad de adoptar medidas que permitan actuar de manera consciente y preventiva ante fenómenos potencialmente destructivos de origen natural y humano (CENAPRED, 2014).

10.2.11. Caso de estudio: Sistema Nacional de Cartografía de Zonas Inundables de España.

El Ministerio de España ha puesto en marcha el Sistema Nacional de Cartografía de Zonas Inundables (SNCZI), un instrumento de apoyo a la gestión del espacio fluvial, la prevención de riesgos, la planificación territorial y la transparencia administrativa.

El eje central del es el visor cartográfico de zonas inundables, que permite a todos los interesados visualizar los estudios de delimitación del Dominio Público Hidráulico (DPH) y los estudios de cartografía de zonas inundables, elaborados por el Ministerio y aquellos que han aportado las Comunidades Autónomas.

El visor sirve de ayuda a los organismos de cuenca en la emisión de informes sobre autorizaciones en el DPH y zona de policía, en la gestión de avenidas en conexión con el S.A.I.H. (Sistema Automático de Información Hidrológica) y en la planificación de las actuaciones de defensa frente a inundaciones; agiliza la planificación y gestión de inundaciones por los servicios de Protección Civil; facilita la transmisión de información sobre zonas inundables a las administraciones competentes en planificación territorial y empresas promotoras; y permite a los ciudadanos conocer la peligrosidad de una zona determinada.

Los usuarios pueden consultar la información del Dominio Público y Zonas Inundables en URL: http://sig.mapama.es/snczi/visor.html?herramienta=DPHZI.

Se hace la recomendación para un funcionamiento óptimo y completo del visor, es necesario que disponga de una versión actualizada de cualquier navegador estándar, igualmente se recomienda para una completa visualización de la aplicación una resolución del monitor como mínimo de 1024 x 768 píxeles. (Ministerio para la Transición Ecológica, 2020).

En la figura 6, a ingresando a al enlace Visores Geográficos del área de actividad de Calidad y Evaluación Ambiental https://www.miteco.gob.es/es/cartografia-y-sig/visores/visores-calidad-evaluacion-ambiental.aspx se puede conocer los indicadores contaminantes notando un símbolo verde bajo, verde fuerte, amarillo, anaranjado rojo y más rojo los niveles de contaminación en lugares específicos de España, así como la fecha y hora del día.

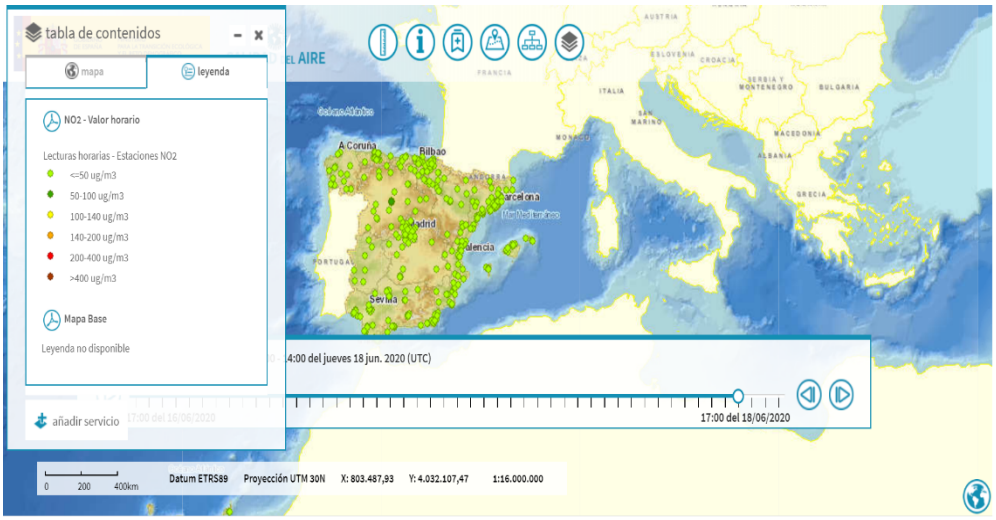


Figura 6. Mapa de España. Monitoreo de la calidad del aire. (Ministerio para la Transición Ecológica, 2020)

10.2.12. Caso Simca. Sistema de monitoreo de la calidad del aire del Estado de Durango.

En su portal (SIMCA, 2020), se muestra en la figura 7, un ejemplo semejante al caso del monitores de la calidad del aire en España, de punto anterior.

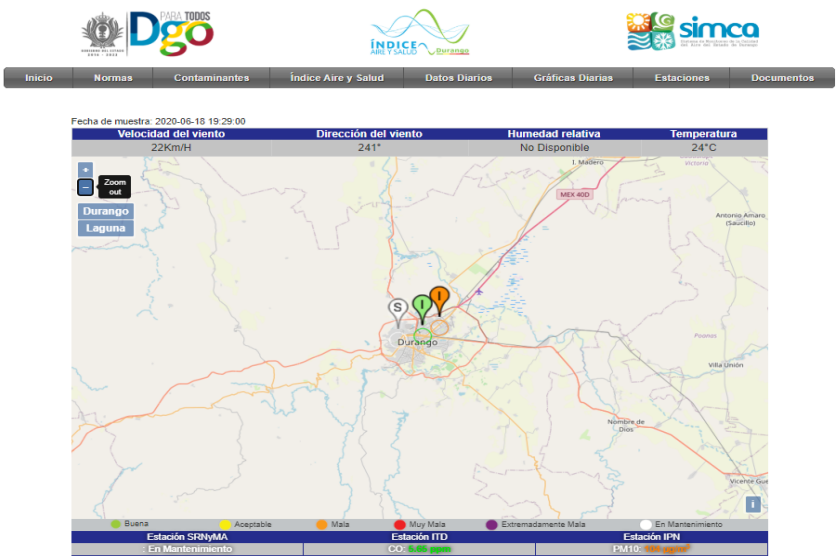


Figura. 7. Sistema de la Calidad del aire del Estado de Durngo simica. (SIMCA, 2020)

10.2.13. Caso Municipio de Durango

El municipio de Durango, no ha estado libre de fenómenos perturbadores de origen natural, los cuales frecuentemente se presentan e inciden de manera directa en perjuicio de sus habitantes.

Debido a esto, se han llevado registro de los fenómenos perturbadores que han ocurrido en la historia reciente del municipio por parte de la Dirección Municipal de Protección Civil.

10.2.13.1. Fenómenos geológicos

Con base en dicho registro, y respecto a los temas de carácter geológico, se han identificado de manera puntual las zonas que potencialmente pueden presentar deslizamientos de laderas, tales sitios son los márgenes y colonia aledañas al Cerro del Mercado, es decir, las Colonias Luz y Esperanza, Morga, Lázaro Cárdenas, Sergio Méndez Arceo, Rosas del Tepeyac, Cerro del Mercado y Guadalupe.

Basado en lo que se define en el atlas de riesgo del municipio de Victoria de Durango, se zonificó el peligro por deslizamientos en el Municipio de Durango; dicho peligro se ubica principalmente en la zona serrana, ya que como se mencionó

anteriormente, la pendiente del terreno es una condición para la existencia de este fenómeno.

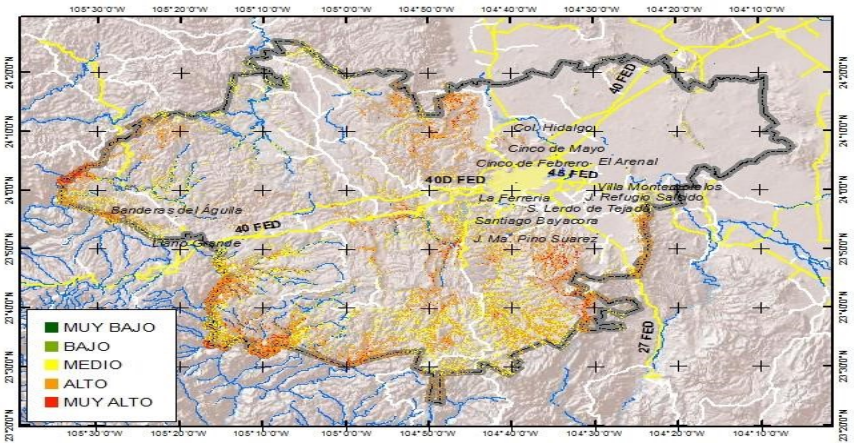


Figura 8. Zonificación del Peligro por deslizamientos en el Municipio de Durango

Fuente: Atlas de Riesgo del Municipio de Durango

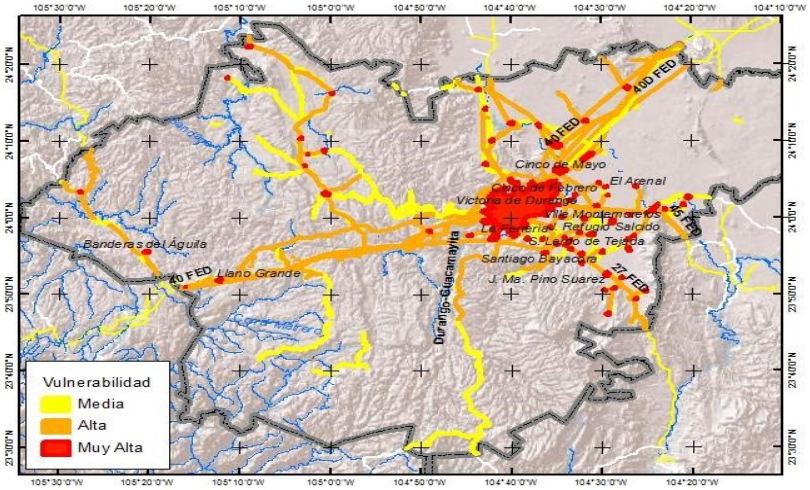


Figura 9. Mapa de vulnerabilidad general ante Procesos de Remoción en Masa en el Municipio de Durango. Fuente: Atlas de Riesgo del Municipio de Durango.

Existen en el Municipio de Durango, zonas con vulnerabilidad Media, Alta y Muy Alta, siendo la zona urbana la que se clasifica como vulnerabilidad Muy Alta.

10.2.13.2. Fenómenos hidrometeorológicos

El municipio de Durango es afectado por varios tipos de fenómenos hidrometeorológicos que pueden provocar daños materiales de importancia: principalmente está expuesto a sequias e inundaciones.

Acontecimientos como las sequías, las escazas precipitaciones pluviales han provocado que el Estado de Durango se encuentre entre los cinco estados con sequía extrema en el territorio nacional; paradójicamente, mientras en ciertas zonas las lluvias son virtualmente nulas, en otras, causan daños como las inundaciones, y por otro lado, las heladas y fríos producen afectaciones en las zonas de cultivo, y puede ser causa de enfermedades en los sectores de la población de corta o avanzada edad.

El conocimiento de los principales aspectos de los fenómenos hidrometeorológicos, la difusión de la cultura de Protección Civil en la población y la aplicación de las medidas de prevención de desastres pueden contribuir de manera importante en la reducción de los daños ante esta clase de fenómenos.

Existen diversos fenómenos hidrometeorológicos que se presentan en el municipio: ciclones (huracanes y ondas tropicales), tormentas eléctricas, sequías, temperaturas máximas extremas, vientos fuertes, inundaciones, masas de aire (heladas y temperaturas mínimas).

Con respecto a los fenómenos de carácter hidrometeorológico que han causado afectaciones en la región, se han registrado inundaciones en las inmediaciones del Río El Tunal y en los afluentes de la presa Gral. Guadalupe Victoria afectando a las localidades de El Pueblito, El Nayar, El Conejo, Francisco Villa Nuevo y Viejo, El Arenal, Tierra y Libertad, Libertad y Democracia, así como zonas bajas del Puente Dalila.

Se tiene registrado que en las Colonias Frac. Villas del Guadiana del I al VII, San Marcos, Las Nubes, Jardines de Cancún, Frac. San Juan, El Alacrán, Frac. Viva Reforma, Nuevo Durango I y II, Col. Isabel Almanza y Frac. Benito Juárez, existe la posibilidad de inundación por lluvias extremas, como de hecho ocurrió en

el año 2007 (inundación con hasta 150cm de tirante, y en el año 2010, cuando hubo 400 viviendas dañadas en la ciudad. Al respecto, además de registros municipales, existe evidencia hidrometeorográfica en los periódicos locales.

El registro sistemático de fenómenos perturbadores elaborado por la Dirección Municipal de Protección Civil representa un gran avance en la identificación de sitios que presentan diversos fenómenos que pudieran poner en riesgo a la población (H. Ayuntamiento del Municipio de Durango, 2012).

10.2.13.3. Peligro por lluvias extraordinarias

Las lluvias extraordinarias pueden afectar al Municipio de Durango de varias maneras. Puede ser un cúmulo de eventos a lo largo de varios días, incluso semanas, que como resultado sobrepasen el promedio de precipitación para el mes en el que ocurren. Pero también se pueden presentar como un solo evento o varios distribuidos en un máximo de 24 horas.

Las lluvias extraordinarias, para considerarse como tales deben superar los valores promedio mensuales de precipitación más una desviación estándar para cada una de las cuatro principales estaciones meteorológicas de la zona.

10.2.13.4. Sequías

La sequía meteorológica es una anomalía atmosférica transitoria en la que la disponibilidad de agua se sitúa por debajo de las necesidades de las plantas, los animales y la sociedad. La causa principal es una disminución significativa en la precipitación pluvial promedio de una zona dada. Si este fenómeno perdura por varias temporadas, deriva en una sequía hidrológica caracterizada por la desigualdad entre la disponibilidad natural de agua y las demandas naturales de agua.

La sequía afecta a todo el territorio municipal, aunque la mayor vulnerabilidad se encuentra en la zona rural del mismo, donde se ubican las actividades agropecuarias. Adicionalmente, en las Colonias Las Palmas, Héctor Mayagoitia, José María Morelos Norte, Paseo Las Mangas, Valentín Gómez Farías, Niños Héroes Norte, Arcoíris, Nuevo Amanecer, 1 de mayo, Arroyo Seco, Ejidal y Valle

Verde, las temperaturas mínimas extremas han causado daños a la salud de sus habitantes, sin que existan cuantificaciones exactas de los daños en este rubro.

En casos extremos se puede llegar a la aridez. Las consecuencias inmediatas de la sequía meteorológica son pérdida de cosechas, perdida de cabezas de ganado vacuno, ovino y caprino y en casos agudos, insuficiencia de agua para uso doméstico e industrial.

La humedad es un elemento central para la clasificación de la sequía agrícola e hidrológica, por lo que su cálculo es necesario para su interpretación y estudio. El Índice de Aridez de M. E. Hernández es una valoración del grado de humedad que existe en el ambiente mediante una sencilla ecuación que divide la precipitación promedio de un periodo de tiempo determinado, sobre la evaporación en el mismo periodo. El cálculo se realizó mediante la aplicación de la ecuación:

$$IA = P/E$$

Donde: IA: índice de aridez, P: precipitación anual (mm), E: evaporación anual (mm)

Tabla 1.

Índices de Aridez

Índice de Aridez	**Grado de Aridez**
<0.25	Árido
0.25-.50	Semiárido
0.50-2.0	Subhúmedo
>2.0	Húmedo

Con datos de otra estación, se presentan la imagen y se visualiza el índice de aridez a 50 años, se muestra en la figura 10.

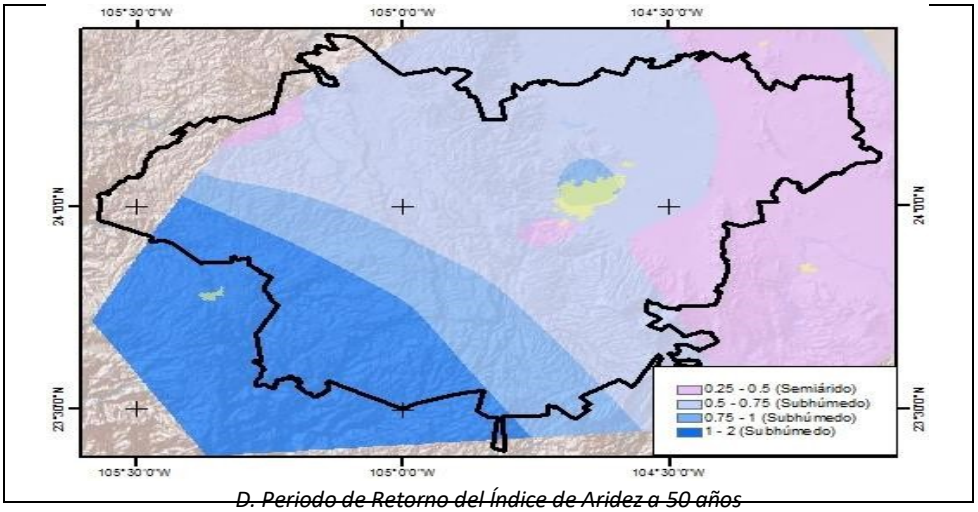


Figura 10. Periodo de Retorno del Índice de Aridez a 5, 10, 25 y 50 años. Fuente: Atlas de Riesgos del Municipio de Durango.

Otra información importante que existe en el Atlas de Riesgos del Municipio de Durango es la Clasificación de la Intensidad de la Sequía mensuales registrados por el Monitor de Sequía de América del Norte (NADM), con un valor de pixel de 1000m. La cartografía se generó mediante un sistema de información geográfica, a través de la media aritmética de los valores de 16 meses.

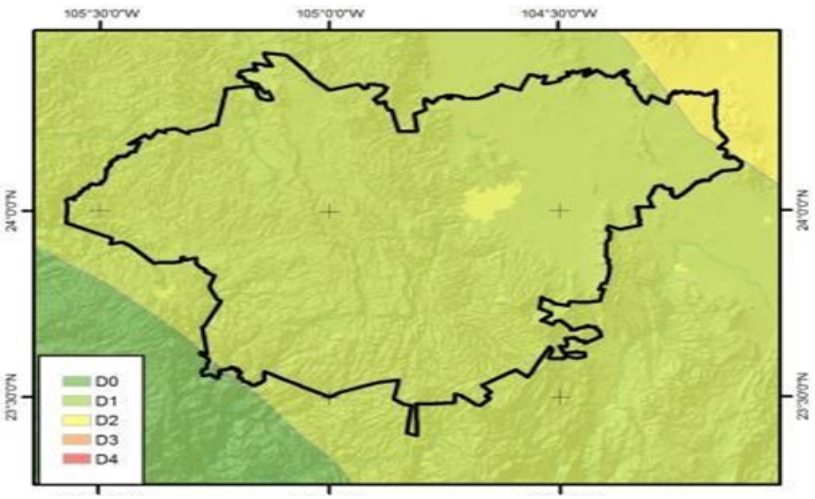


Figura 11. Clasificación de la Intensidad de la Sequía para un periodo determinado. Fuente: Atlas de Riesgos del Municipio de Durango.

En el caso de las áreas forestales, la vulnerabilidad se estima baja, debido a que en general la vegetación de esas zonas es más resistente a estos

eventos, y en todo caso, la repercusión económica no es tan álgida como en el caso de las zonas agrícolas.

Con base en lo anterior y debido a que hay sequía en el municipio desde hace más de un año, además de que la probabilidad de sequía a futuro es muy alta, se estima que, para el Municipio de Durango, el riesgo de sequía es MUY ALTO.

10.2.13.5. Inundaciones

Las inundaciones son un fenómeno en el cual se anega (inunda o llena) de agua un área determinada que generalmente está libre de ésta. El agua proviene del desbordamiento de ríos, represas, o escurrimientos de partes altas y se asocia a lluvias intensas, en el área o incluso en otras lejanas. A pesar de considerarse un fenómeno natural, tiene una alta influencia de los procesos de ocupación del territorio y construcción de infraestructura, ya que a menudo el riesgo existe cuando se establecen viviendas en zonas inundables y se crean embudos artificiales que impiden el libre tránsito de las avenidas de agua.

Las inundaciones son uno de los peligros más comunes en el Municipio de Durango, a menudo las inundaciones se desarrollan lentamente, pero las más dañinas son repentinas e incluso finalizan en sólo unas horas, sin señales visibles de lluvia en la zona inundada. Las inundaciones repentinas consisten en una avenida de agua con gran fuerza de arrastre y con una carga de escombros que encuentra en su paso. Las inundaciones ocurren sobre los márgenes de un río, canal o arroyo definido, pero también pueden generarse por la confluencia de aguas en zonas bajas.

Debido a la particular configuración del municipio, el riesgo de inundación es muy alto en varias localidades de la zona del Valle del Guadiana, incluyendo partes de la cabecera municipal, debido a que las aguas de la sierra bajan por arroyos cuyos márgenes están ocupados por viviendas, además de que algunos embalses naturales también han sido aprovechados para la construcción de casas.

Las inundaciones en las zonas urbanas se deben a depresiones en el terreno, que aprovecha el agua para embalsarse, o bien, zonas donde la pendiente y la orientación de las calles las convierte en cauces.

En las figura 12 y 13 se muestran ejemplos de mapa de riesgo de inundación en algunas zonas de la ciudad de Durango.

En el oriente de la Ciudad de Victoria de Durango, se construyó sobre una planicie de inundación antiguamente utilizada como tierras de cultivo (H. Ayuntamiento del Municipio de Durango, 2012).

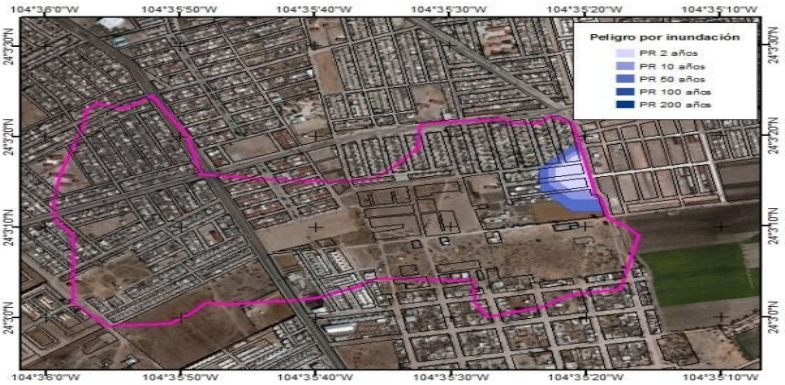


Figura 12. Peligro por inundación según diferentes periodos de retorno en la microcuenca San Luis. Fuente: Atlas de Riesgo del Municipio de Durango

Al poniente de la Ciudad de Victoria de Durango, se construyeron a los márgenes de un arroyo intermitente

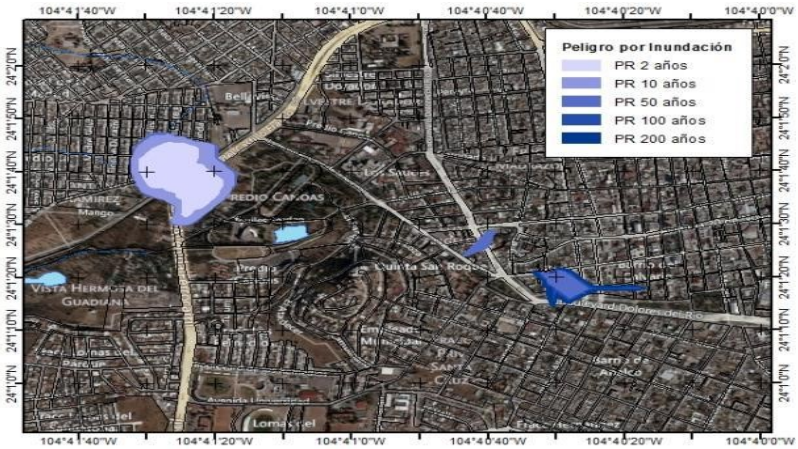


Figura 13. Peligro por inundación según diferentes periodos de retorno en la microcuenca Arroyo La Virgen. Fuente: Atlas de Riesgo del Municipio de Durango

10.2.13.6. Vientos Fuertes

El aire que circula sobre la Tierra se denomina viento, pero existen vientos de superficie y “vientos planetarios de altura”; estos últimos forman parte de la circulación general del aire en lo alto de la troposfera. La distribución desigual de la presión es lo que causa el movimiento del aire, ya que éste se desplaza desde las áreas de alta presión hacia áreas de baja presión, en un intento por lograr un equilibrio.

Los vientos de mayor intensidad pueden ser peligrosos ya que dañan a la infraestructura, produciendo ello a su vez, daños a las personas y a sus bienes. El fenómeno de los huracanes, se mide, de hecho, en función de los vientos, toda vez que son ellos los que causan los mayores perjuicios a la sociedad.

Para fines de protección civil, se presenta un mapa que muestra regiones con valores similares de intensidades máximas de viento, el cual se presenta en el anexo cartográfico en el mapa de la figura 14, en él se divide el municipio de Durango en zonas que representan bandas homogéneas de velocidad máxima de viento.

Periodo de retorno: 10 años *Periodo de retorno: 50 años* *Periodo de retorno: 200 años*

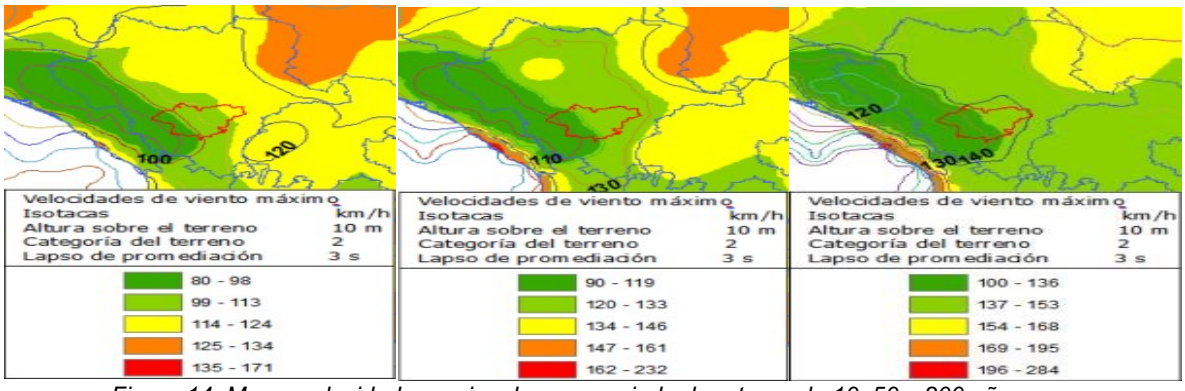


Figura 14. Mapa velocidades regionales con periodo de retorno de 10, 50 y 200 años.

Fuente: Atlas de Riesgo del Municipio de Durango

10.2.13.7. Heladas y temperaturas bajas

La helada es un fenómeno atmosférico que consiste en un descenso de la temperatura ambiente a niveles inferiores al punto de congelación del agua (0°C) y hace que el agua o el vapor que está en el aire se congele

depositándose en forma de hielo en las superficies, el cual se presenta en las primeras horas del día (de las 3 a las 6 horas).

El Municipio de Durango se caracteriza por una diversidad de condiciones de temperatura y humedad. Por su ubicación geográfica se encuentra entre dos regiones climáticas, la subhúmeda al poniente y la semiárida, al oriente. Debido a la forma del relieve, la altitud, extensión territorial y su localización relativamente cercana al océano, se producen diversos fenómenos atmosféricos, según la época del año; que ocasionan bruscos descensos de temperatura. Estos descensos de temperatura son más evidentes en las zonas de la Sierra, sin embargo, también ocurren en la región del Valle del Guadiana.

De acuerdo con registros históricos del Servicio Meteorológico Nacional, se pueden ubicar regiones donde es más común la incidencia de las bajas temperaturas expresadas como el fenómeno de las heladas. En general, se observa que, en el municipio de Durango, existe una estrecha relación entre las zonas más frías y los sistemas orográficos. (H. Ayuntamiento del Municipio de Durango, 2012).

El mapa de la figura15, indica la frecuencia de heladas registradas en el Municipio de Durango.

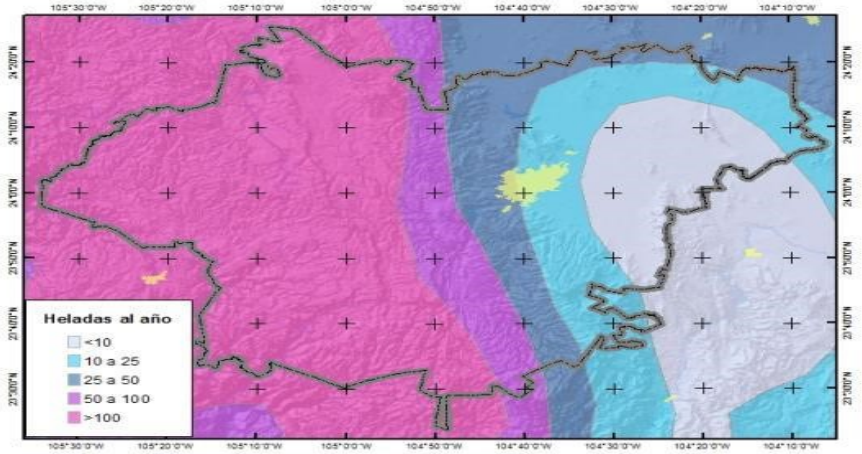


Figura 15. Mapa de frecuencia de heladas en el Municipio de Durango.
Fuente: UNAM, Atlas Nacional de México

Según los criterios establecidos por el Atlas Nacional de Riesgos (CENAPRED), **el peligro por heladas es muy alto** si hay más de 100 días al año con presencia de este fenómeno; alto si hay de 50 a 100 días; medio si hay de 25 a 50; bajo si hay de 10 a 25; y muy bajo si hay menos de 10 evento por año.

Los daños que producen las diferentes intensidades de heladas se refieren en la siguiente tabla.

Tabla 2.

Daños por diferentes intensidades de heladas. Fuente: Atlas de Riesgo del Municipio de Durango

Temperatura	Intensidad	Daños
0 a -3.5°	Ligera	El agua comienza a congelarse. Daños pequeños a las hojas y tallos de la vegetación. Si hay humedad el ambiente se torna blanco por la escarcha.
-3.6 a -6.4	Moderada	Los pastos, las hierbas y hojas de plantas se marchitan y aparece un color café o negruzco en su follaje. Aparecen los problemas de enfermedades en los humanos de sus vías respiratorias. Se comienza a utilizar la calefacción.
-6.5 a -11.5	Severa	Los daños son fuertes en las hojas y frutos de los árboles frutales. Se rompen algunas tuberías de agua por aumento de volumen del hielo. Se incrementan las enfermedades respiratorias. Existen algunos decesos por hipotermia.
< 11.5	Muy severa	Muchas plantas pierden todos sus órganos. Algunos frutos no protegidos se dañan totalmente. Los daños elevados son en las zonas tropicales.

En la figura 16 se presenta un mapa de teperaturas temperaturas mínimas absolutas en el Municipio de Durango:

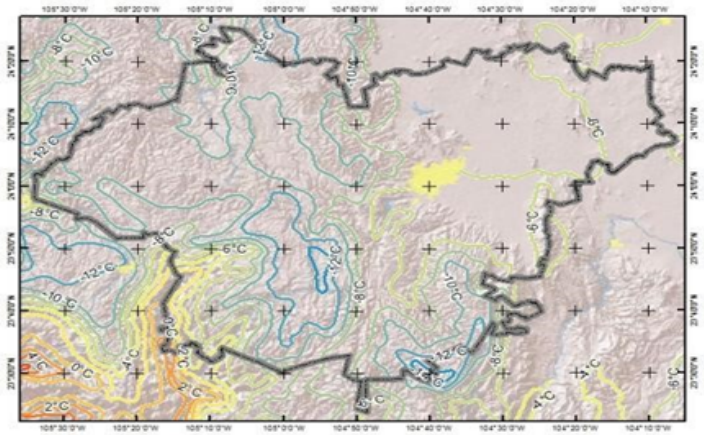


Figura 16. Mapa de temperaturas mínimas absolutas en el Municipio de Durango. Fuente: CONABIO.

Debido a la situación geográfica del Municipio de Durango son pocas las áreas que presentan nevadas. Este fenómeno ocurre principalmente en las regiones altas de la Sierra Madre Occidental al interior del Estado, y rara vez se presentan en el Valle del Guadiana.

En la figura 17, un mapa frecuencia de nevadas y en la figura 18 un indicador de peligro de nevadas

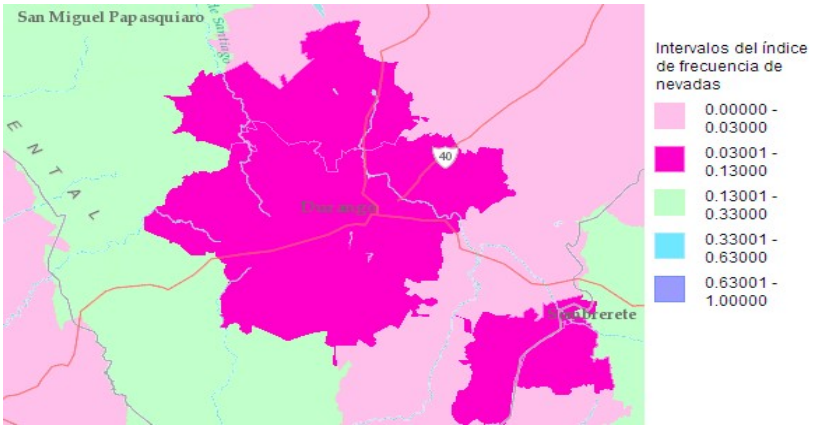


Figura 17. Índice de peligro por el mismo fenómeno en el Municipio de Durango. Fuente: Atlas de Riesgo del Municipio de Durango.

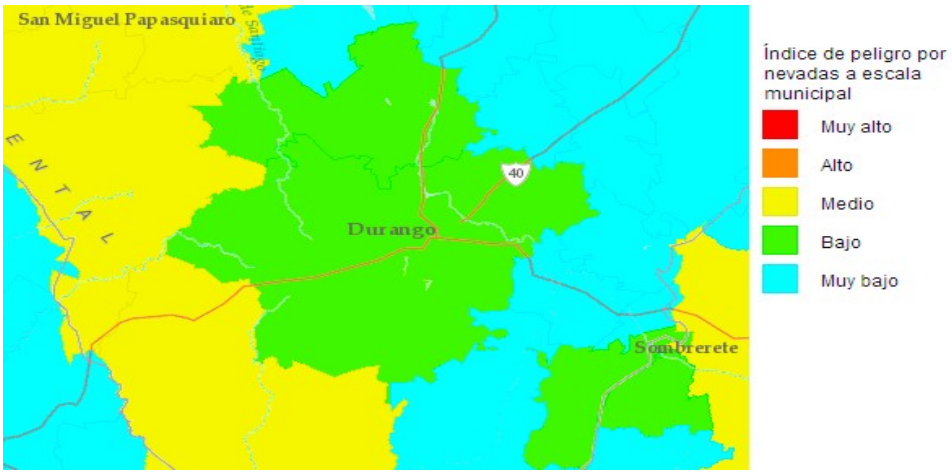


Figura 18. Índice de peligro por el mismo fenómeno en el Municipio de Durango. Fuente: Atlas de Riesgo del Municipio de Durango.

Debido a los importantes antecedentes del registro sistemático de fenómenos perturbadores en el Municipio de Durango, se presenta una propuesta de análisis de datos utilizando R y Python para visualizarlos en un SIG la cual tiene la finalidad de seguir una metodología específica para determinar, medir y evaluar el nivel de peligro y/o riesgo en el cual se encuentra la población.

Las propuestas de acciones y obras estarán enfocadas a la reducción y mitigación de riesgos; basadas en el desarrollo de bases de datos homologadas para cada uno de los fenómenos naturales perturbadores para su detección y localización de zonas de riesgo o peligro los cuales serán visualizados y ubicadas en la cartografía utilizando QGIS.

Por lo anterior, se sugiere proporcionar a la Dirección Municipal de Protección Civil una herramienta tecnológica que permita analizar los datos utilizando R y Python, la cual podrá identificar por medio de mapas cartográficos las principales zonas afectadas por agentes perturbadores para crear un conjunto de medidas que permitan, de manera oportuna, prevenir los riesgos. Así mismo, se tiene considerado que en un futuro esta herramienta sea una aplicación instalada en un dispositivo móvil para mantener informado al ciudadano en tiempo y forma.

10.3. Desarrollo

La propuesta de análisis de datos de Protección Civil utilizando R y Python en QGIS se fundamenta en las bases teóricas y sistémicas establecidas en la Guía para la elaboración de Atlas de Riesgos; ya que su estructura está diseñada como una plataforma informática apoyada en sistemas de información geográfica y bases de datos, la cual a su vez se conformó de acuerdo con los criterios de clasificación y los términos de referencia establecidos por el CENAPRED que integra información sobre: mapas de peligros por fenómenos perturbadores, mapas de susceptibilidad, inventario de bienes expuestos , inventario de vulnerabilidades, mapas de riesgos, escenarios de riesgos

Por otra parte, las herramientas y tecnologías utilizadas en el análisis de datos están evolucionando rápidamente, y la ciencia de datos necesita una metodología fundamental que sirva como estrategia de guía para resolver problemas. La metodología a utilizar es independiente de tecnologías o herramientas particulares, debe proporcionar un marco para continuar con los métodos y procesos que se utilizarán para obtener respuestas y resultados.

Una de estas metodologías corresponde a la Metodología Fundamental para la Ciencia de Datos, que consta de 10 etapas que representan un proceso iterativo que va desde la concepción de la solución hasta el despliegue, retroalimentación y refinamiento de la solución.

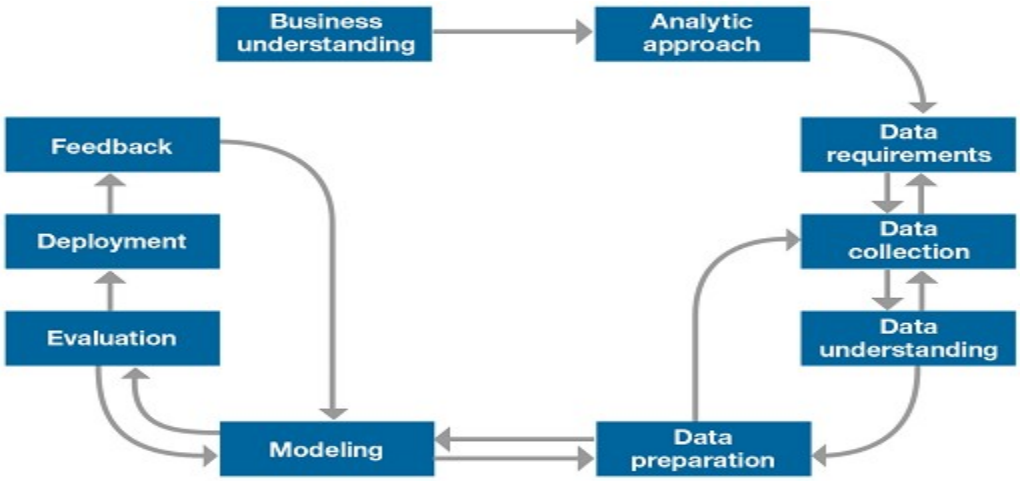


Figura 19. Esquema de Metodología de la Ciencia de los Datos (Patel, 2019)

10.3.1. Metodología Fundamental para la Ciencia de Datos

Se presenta la descripción de la metodología para Ciencia de los Datos de acuerdo a Patel (2019):

- Entendimiento del negocio (Business understanding): Cada proyecto, independientemente de su tamaño, comienza con el entendimiento del negocio, que sienta las bases para una resolución exitosa del problema del negocio.
- Enfoque analítico (Analytic approach): Después de indicar claramente un problema comercial, el científico de datos puede definir el enfoque analítico para resolverlo. Hacerlo implica expresar el problema en el contexto de las técnicas estadísticas y de aprendizaje automático para que el científico de datos pueda identificar las técnicas adecuadas para lograr el resultado deseado.
- Requerimientos de datos (Data requirements): La elección del enfoque analítico determina los requisitos de datos, para que los métodos analíticos que se utilicen requieran contenidos, formatos y representaciones de datos particulares, guiados por el conocimiento del dominio.
- Recopilación de datos (Data collection): El científico de datos identifica y recopila recursos de datos (estructurados, no estructurados y semiestructurados) que son relevantes para el dominio del problema. Al encontrar vacíos en la recopilación de datos, el científico de datos podría necesitar revisar los requisitos de los datos y recopilar más datos.
- Comprensión de los datos (Data understanding): Las estadísticas descriptivas y las técnicas de visualización pueden ayudar a un científico de datos a comprender el contenido de los datos, evaluar la calidad de los datos y descubrir información inicial sobre los datos.
- Preparación de datos (Data preparation): Comprende todas las actividades utilizadas para construir el conjunto de datos que se utilizarán en la etapa de modelado. Estos incluyen la limpieza de datos, la combinación de datos de múltiples fuentes y la transformación de datos en variables más útiles. Además, la ingeniería de características y el análisis de texto se pueden usar para derivar nuevas variables estructuradas, enriqueciendo el conjunto de predictores y

mejorando la precisión del modelo. Es la etapa que consume más tiempo, entre un 50% y un 90% del tiempo total del proyecto.

- Modelado (Modeling): A partir de la primera versión del conjunto de datos preparado, los científicos de datos utilizan un conjunto de datos históricos (en los que se conoce el resultado de interés) para desarrollar modelos predictivos o descriptivos utilizando el enfoque analítico ya descrito. El proceso de modelado es altamente iterativo.
- Evaluación (Evaluation): El científico de datos evalúa la calidad del modelo y verifica si aborda el problema comercial de manera completa y adecuada. Hacerlo requiere la computación de varias medidas de diagnóstico, así como otras salidas, como tablas y gráficos, utilizando un conjunto de pruebas para un modelo predictivo.
- Implementación (Deployment): Una vez que se ha desarrollado un modelo satisfactorio que ha sido aprobado por los patrocinadores comerciales, se implementa en el entorno de producción o en un entorno de prueba comparable. Tal despliegue a menudo se limita inicialmente para permitir la evaluación de su desempeño.
- Retroalimentación (Feedback): Al recopilar los resultados del modelo implementado, la organización obtiene comentarios sobre el rendimiento del modelo y observa cómo afecta su entorno de implementación. Al analizar esta información, el científico de datos puede refinar el modelo, aumentando su precisión y, por lo tanto, su utilidad.

El flujo de esta metodología enseña la naturaleza iterativa del proceso para la resolución de problemas. Los modelos no deben crearse una vez, luego implementarse y dejarse en su lugar sin cambios. En preciso, que, a través de la retroalimentación, el refinamiento y la redistribución, un modelo debe adaptarse continuamente a las condiciones, permitiendo que tanto el modelo como el trabajo detrás de éste proporcionen valor a la organización durante el tiempo que sea necesaria la solución.

10.3.2. Propuesta de solución

Las bases teóricas y sistémicas para la elaboración de un SIG en el tema de Protección Civil se deben de tomar en cuenta lo establecido en la Guía del CENAPRED para la prevención de riesgos y/o peligros.

El territorio municipal de Durango se encuentra sujeto a gran variedad de fenómenos que pueden causar desastres.

Por su cercanía a la Sierra Madre Occidental y su configuración de cuenca de sedimentación, el municipio está expuesto a inundaciones y flujos, que se resienten principalmente en las zonas bajas del oriente, en donde también se ubica la cabecera municipal y las principales localidades; las lluvias intensas además pueden causar deslaves y derrumbes en el interior del territorio.

La propuesta de Análisis de Datos de Protección Civil, corresponde al conjunto de tareas que tienden a la reducción de los impactos de los desastres más frecuentes a nivel municipal. Un requisito esencial para la puesta en práctica de las acciones de protección civil es contar con diagnósticos de riesgos, o sea, conocer las características de los eventos que pueden tener consecuencias desastrosas y determinar la forma en que estos eventos inciden en los asentamientos humanos, en la infraestructura y en el entorno. El proceso de diagnóstico implica la determinación de los escenarios o eventos más desfavorables que pueden ocurrir, así como de la probabilidad asociada a su ocurrencia.

Es importante explicar algunos conceptos generales sobre la medición del riesgo. El riesgo se calcula en función de una formulación probabilística, que en su planteamiento más general se expresa de la manera que se describe a continuación: $Riesgo = Peligro\ X\ Vulnerbilidad$.

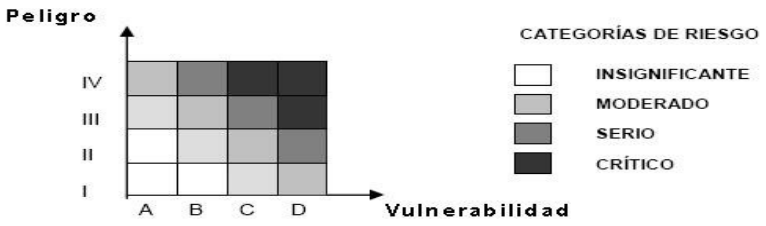


Figura 20. Representación gráfica de la medición del Riesgo en función del peligro y vulnerabilidad
Fuente: Atlas de Riesgo del Municipio de Durango

Se llama peligro P, a la probabilidad de que se presente un evento de cierta intensidad, tal que pueda ocasionar daños en un sitio dado. Se llama vulnerabilidad V, a la propensión de estos sistemas a ser afectados por el evento; la vulnerabilidad se expresa como una probabilidad de daño. Finalmente, el riesgo es el resultado de los dos factores. En este esquema, el riesgo se expresa como un resultado posible de un evento; ya que P y V son dos probabilidades.

Utilizando la Metodología Fundamental para la Ciencia de Datos enfocada al análisis de datos geoespaciales en un Sistema de Información Geoespacial (SIG) se suponen tres conjuntos de operaciones, que son los siguientes:

- Gestión de datos: involucra hacer operaciones necesarias para resolver un propósito, por lo que se incluyen en este planteamiento todos aquellos procesos para manejar y moldear adecuadamente la información. Las informaciones necesarias para el SIG serán generalmente de diversas fuentes y con distintos formatos, lo que implica que el SIG deberá homogeneizar y adaptar previamente los datos. En este proceso se deberá corregir ciertos problemas como las transformaciones y reproyecciones cartográficas necesarios, la introducción de los atributos temáticos que sean exactos, generar la topología de los elementos espaciales que se requieran, así como todos los ajustes y procesos que admitan que los datos sean completamente útiles para la obtención de la información, y a partir de ellos sean aplicados a las operaciones de análisis y representación en el SIG.
- Análisis de datos: conlleva al procesamiento de datos para la aplicación de diferentes operaciones y estudios en las diferentes características de los datos geográficos presentes, para así obtener nueva información que permita alcanzar soluciones a los problemas espaciales.

  El análisis de datos es el aspecto más característico de un dentro del SIG, ya que se puede considerar como el conjunto de operaciones más importante, ya que ayuda a obtener mayor conocimiento sobre un hecho, circunstancia o fenómeno presente o propuesto en un territorio.

- Presentación de datos: supone mostrar los datos mediante diferentes medios visuales, y la información generada después del análisis y gestión de los mismos. El SIG admite la representación de los datos de salida de distintas maneras, mediante tablas, gráficos, y primordialmente mediante mapas que se visualizarán en la interfaz del SIG, ya sea de forma analógico (para su uso físico), en diversos formatos para su uso en diferentes dispositivos informáticos, o mediante su publicación en la web a través de Internet.

10.3.3. Proceso para llevar a cabo actividades

Existe diferentes aspectos conectados entre sí que permiten un correcto funcionamiento de un SIG, estos se pueden dividir en cuatro componentes principales:

i. Componente informativo es el afín con los datos a emplear en el SIG y la visualización de los mismos. Se debe considerar en el dato sus tres aspectos fundamentales:
   - El temático corresponde con cualquier rasgo cuantitativo o cualitativo que identifique y defina al elemento o suceso a partir de partir de caracteres alfanuméricos que se está considerando
   - El espacial se relaciona con la forma geométrica y la localización geográfica del dato, la cual será factor fundamental para la información manejada en el SIG, ya que su importancia radica en la esencia de los datos, como su potencialidad para establecer las relaciones entre ellos.
   - El temporal de los datos tendrá incidencia en los elementos territoriales con un dinamismo elevado, por lo que es primordial la actualización continua para poder hacer análisis reales de una determinada situación.

ii. Componente instrumental hace referencia a las herramientas, instrumentos y aplicaciones que dan soporte informático al SIG mediante dispositivos, software y hardware.

iii. Componente metodológico establece los procesos y métodos encaminados al análisis de la información a partir de los avances y desarrollos científicos para el

estudio, la evaluación y la comprensión de los fenómenos naturales y sociales que se tiene lugar sobre el planeta.

iv. Componente organizativo comprende a las personas, los ámbitos de trabajo y sus procedimientos de gestión y organización que determinarán en gran medida la validez y operatividad del sistema.

10.3.4. Tecnologías y recursos

En la propuesta del análisis de datos de un SIG, se abarcan recursos tecnológicos para la solución de dicha idea, por tal motivo se enfocará principalmente a la explicación del software que se podrán utilizar para dicha propuesta.

El mundo de la programación es complejo y dinámico a la vez, debido a su constante cambio, adaptación y evolución para dar respuesta a las necesidades de múltiples sectores profesionales.

En el sector geoespacial esto no es la excepción, ya que, debido a su polifacética aproximación, existen una gran multitud de soluciones, aplicaciones y funcionalidades que pueden encuadrarse dentro del desarrollo o programación GIS.

De los softwares existentes para el análisis de datos en un SIG se tomarán en cuenta los siguientes:

- **Python:** es el principal lenguaje de programación y además el de mayor crecimiento. Python es un lenguaje de scripts, orientado a objetos y de alto nivel, con este leguaje podemos automatizar las tareas de geo procesamiento.
- **R:** es el lenguaje de programación con mayor capacidad para el análisis, el procesado y la visualización de datos desde un punto de vista estadístico.
- **MySQL:** se utiliza como acceso a bases de datos y lenguaje de control, se considera como el corazón de muchas operaciones de SIG debido a que forma parte integral e indispensable en el control y consulta de la información.

- **QGIS:** es un Sistema de Información Geográfica (SIG) de código libre. Permite manejar formatos raster y vectoriales, así como bases de datos. Algunas de sus características son:
- Soporte para la extensión espacial.,
- Manejo de archivos vectoriales,
- Soporte para un importante número de tipos de archivos raster.

10.3.5. Análisis de la propuesta

La Ciencia de los Datos está evolucionando rápidamente para convertirse en uno de los campos más populares en la industria de la tecnología. Con los rápidos avances en el rendimiento computacional que ahora permiten el análisis de conjuntos de datos masivos, podemos descubrir patrones e ideas sobre el comportamiento del usuario y las tendencias mundiales en una medida sin precedentes.

Aplicar la Ciencia de los Datos es ofrecer información útil a las personas adecuadas en el momento adecuado, por lo consiguiente el producto final estará orientado específicamente hacia la propuesta de análisis de la base de datos que es la parte medular del SIG, así como los lineamientos para su posterior expansión que se espera obtener como resultado de esta investigación es un documento que contenga una serie de recomendaciones para el diseño e implementación de un SIG para la interpretación de indicadores de fenómenos perturbadores en el Municipio de Durango. Además, los requerimientos para su uso y manejo.

Se espera contribuir en un diseño metodológicamente planeado como apoyo a la realización de un SIG cuyas aplicaciones estén dentro del marco de desarrollo sustentable, y al mismo tiempo demostrar su eficiencia para este tipo de aplicaciones.

Asimismo, se desea fomentar el apoyo al desarrollo, uso y administración de herramientas de SIG para la toma de decisiones. Esperando con esto no sólo el desarrollo de SIG para aplicaciones dentro del municipio de Victoria de Durango, sino también el desarrollo de herramientas de información que apoyen y fomenten un desarrollo sustentable del estado de Durango.

Conclusiones

Con respecto información relacionada con Protección Civil, la mayor parte de los datos que maneja está georreferenciada. Es decir, que se trata de información a la cual puede asignarse una posición geográfica, y es por tanto información que viene acompañada de otra información adicional relativa a su localización.

En los últimos años ha crecido la inquietud, investigación e interés en los sistemas de información geográfica. En parte debido a que los costos del hardware cada vez son más bajos y a que la capacidad de cómputo y almacenamiento son cada vez mayores.

La idea central consiste en realizar un análisis de los datos geoespaciales utilizando software libre integrando R y Python con QGIS proponiendo un sistema de información geográfica que incluya base de datos, sistemas de tratamientos de información geográfica, tecnologías de desarrollo web, servidores web, que en conjunto permitan contar con un proceso de actualización de la información cartográfica; reduciendo el tiempo de actualización y publicación.

Actualmente hay más facilidades de acceso a la información pública; es decir, se puede consultar a través de una conexión de internet; por lo tanto, un sistema integral de información permite establecer bases de datos y realizar el análisis del peligro, de la vulnerabilidad y del riesgo antes desastres a escala nacional, regional, estatal y municipal, con objeto de generar mapas y sistemas geográficos de información. Con ello se estará en posibilidad de simular escenarios de desastres, emitir recomendaciones para la oportuna toma de decisiones y establecer efectivas medidas de prevención y mitigación.

Además, al desarrollarse con software libres no tendrá costo de licenciamiento, debido a que existen herramientas dentro de esta clasificación de software que proporcionará las mismas, e incluso mejores, funcionalidades necesarias en comparación con el software propietario.

Partiendo de la idea central descrita anteriormente este documento menciona sobre los fenómenos perturbadores de carácter destructivo en el Municipio Victoria

de Durango, por tanto, no está exento a los desastres naturales por su ubicación geográfica y consecuente dinámica geológica y climática.

El objetivo de Protección Civil es prevenir las situaciones de grave riesgo colectivo o catástrofes, proteger y socorrer a las personas, los bienes y el medio ambiente, cuando dichas situaciones se producen, así como contribuir a la rehabilitación y reconstrucción de las áreas afectadas.

Para soportar la propuesta se presenta en este trabajo de investigación un análisis de los datos geoespaciales utilizando software libre integrando R y Python en QGIS proponiendo un sistema de información geográfica que incluya base de datos, sistemas de tratamientos de información geográfica, tecnologías de desarrollo web, servidores web, que en conjunto permitan contar con un proceso de actualización de la información cartográfica; reduciendo el tiempo de actualización y publicación.

Los beneficios para que tendrá Municipio Victoria de Durango se consideran dos puntos clave uno que con base a esta propuesta se tendrá un proceso de actualización de la información cartográfica; previendo los destrates por causa de los fenómenos naturales.

Otro beneficio que se busca con esta propuesta es que en un futuro se desarrolle y se implemente utilizando software libre (R, Python y QGIS); evitando el gasto por licenciamiento, debido a que herramientas dentro de esta clasificación de software que proporcionará las mismas, e incluso mejores, funcionalidades necesarias en comparación con el software propietario.

Por otro lado, se considera más que un beneficio es una propuesta para Instituto Tecnológico de Durango el tema de Protección Civil utilizando ciencias de datos con el objetivo de estimular el interés de la comunidad docente y estudiantil en este tema de tanta relevancia actual creando científicos de datos que son sencillamente un profesional dedicado a analizar e interpretar grandes bases de datos.

En el ámbito personal el realizar este trabajo con el tema de ciencia de datos enfocado al análisis de datos de Protección Civil utilizando herramientas de software libre llevó a conocer y actualizarse en este mundo de R y Python; como estos lenguajes son tan nobles que se pueden realizar un mundo de posibilidades para análisis de datos masivos con el fin de tomar decisiones para mejorar las actividades y mejoras de la vida actual.

El conocimiento para un profesional en la informática y tecnologías de la información va en proporción de la actualización de avances tecnológicos.

Para entender cómo funciona la Ciencia de los Datos no se requiere ser un científico de datos experto, se pueden conocer aspectos del enfoque de la Ciencia de los Datos en el manejo de información e integrar los modelos al desarrollo de aplicaciones realizando un análisis de datos para realizar las acciones necesarias y llegar a la toma de decisiones más apta para el problema que se presente.

Esta propuesta enfocada a ciencia de datos para el análisis de datos de Protección Civil utilizando R y Python en QGIS espera contribuir en un diseño metodológicamente planeado como apoyo a la realización de un SIG cuyas aplicaciones estén dentro del marco de desarrollo sustentable y al mismo tiempo demostrar su eficiencia para este tipo de aplicaciones.

Referencias

Cámara de Diputados del H Congreso de la Unión. (19 de 01 de 2019). *Ley General de Protección Civil*. Obtenido de Ley General de Protección Civil 2012. Ultima reforma 2018: http://www.diputados.gob.mx/LeyesBiblio/pdf/LGPC_190118.pdf

Carranza Tresoldi, J. (18 de 07 de 2016). *Geo Awesomenes*. Obtenido de The right human scale to measure Sustainable Development Goals: https://geoawesomeness.com/the-right-human-scale-to-measure-sustainable-development-goals/

CENAPRED. (2014). *Cenapred*. Obtenido de Centro Nacional de Prevención de Desastres: https://www.gob.mx/cenapred

Centro Mediterraneo. Univesidad de Granada. (s.f. de s.f. de s.f.). *Centro Mediterraneo. Univesidad de Granada*. Obtenido de Ciencia de Datos: Un Enfoque Práctico en la Era del Big Data (VI ed.): https://cemed.ugr.es/curso/20gr10/

Coronado Iruegas, A. A. (16 de 11 de 2016). *Slide Share.net. Discover. Share. Learn.* Obtenido de Big data taller INEGI sedesol: https://www.slideshare.net/acoronadoiruegas/big-data-taller-inegi-sedesol

Jones, H. (2019). *Ciencia de los Datos. Lo que saben los mejores científicos de datos sobre el análisis de datos, minería de datos, estadísticas, aprendizaje automático y Big Data que usted desconoce.* México: Amazon Mexico Services, Inc.

Ministerio para la Transición Ecológica. (18 de 06 de 2020). *Gobierno de España*. Obtenido de Ministerio para la Transición Ecológica: https://sig.mapama.gob.es/calidad-aire/

Morales, A. (01 de 01 de 2018). *MappingGIS. Formación que impulsa tu perfil GIS*. Obtenido de Lenguajes de programación para GIS y sus tendencias de crecimiento: https://mappinggis.com/2012/11/lenguajes-de-programacion-gis/

Patel, A. (18 de 08 de 2019). *ML Research Lab*. Obtenido de Data Science Methodology — How to design your data science project: https://medium.com/ml-research-lab/data-science-methodology-101-2fa9b7cf2ffe

Sánchez Fleitas, N., Comas Rodríguez, R., & García Lorenzo, M. M. (2019). Sistema Inteligente de Información Geográfica para las empresas eléctricas cubanas. *Ingeniare. Revista chilena de ingeniería, vol. 27* , 197-209.

SIMCA. (06 de 18 de 2020). *Sistema de Monitoreo de la Calidad del aire del Estado de Durango*. Obtenido de Sistema de Monitoreo de la Calidad del aire del Estado de Durango: http://calidadaire.durango.gob.mx/

Universidad de Alcalá. (s.f.). *Universidad de Alcalá. Master en Business Intelligence and Data Science*. Obtenido de Ventajas y Desventajas del uso del Big Data: https://www.master-bigdata.com/

Vázquez Pulido, J. C., & Morales Bautista, E. M. (01 de 01 de 2019). *Instituto Mexicano del Transporte*. Obtenido de Datos masivos geoespaciales para identificación de patrones de riesgo en la RNC: https://imt.mx/archivos/Publicaciones/PublicacionTecnica/pt540.pdf

Capítulo 11

Comparativo de herramientas para visualización de datos: Tableau y Power BI

Ana Georgina Soledad Núñez Martínez
Tecnológico Nacional de México. Instituto Tecnológico de Durango

00041260@itdurango.edu.mx

América Herrera Domínguez
Tecnológico Nacional de México. Instituto Tecnológico de Durango

03040966 @itdurango.edu.mx

José Gabriel Rodríguez Rivas
Tecnológico Nacional de México. Instituto Tecnológico de Durango

gabriel.rodriguez@itdurango.edu.mx

11.1. Introducción

Considerando que existen en el mercado diversas opciones de herramientas de visualización de datos, las cuales están enfocadas en extraer al máximo toda la información de la enorme cantidad de datos que se generan diariamente; y que cada vez con mayor frecuencia las empresas aplican la Inteligencia de Negocios (BI), para poder tomar la mejor decisión, es preciso ofrecer un comparativo de dos de las herramientas elegidas por los usuarios, Tableau y Power BI.

Por increíble que parezca, todo lo que existe alrededor en término de datos es apto de ser analizado, procesado y con ello obtener una ventaja competitiva ante el adversario en los negocios, sin importar el giro comercial de la empresa. Por tal motivo, es importante conocer las herramientas de visualización de datos y su relación con la inteligencia de negocios que hoy en día es aplicada con mayor frecuencia.

Este capítulo presenta aspectos de las herramientas de visualización de datos Tableau y Power BI, desde su instalación, funcionalidad, interfaz, fortalezas, debilidades, los productos y/o servicios que ofrece, además de cómo se pueden adecuar a las necesidades del usuario. Adicionalmente se ejemplifican algunos casos de éxito donde se han aplicado las 2 herramientas, con el fin de realizar un comparativo sobre ambas herramientas.

Los datos se encuentran prácticamente en todos lados, desde redes sociales, big data, dispositivos móviles, sistemas de predicción del clima y sensores entre otros, en términos generales todo lo que se puede clasificar como Internet de las Cosas a medida que estos datos van creciendo para los usuarios e incluso para algunos sistemas informáticos resulta difícil su manejo.

El almacenamiento y ordenamiento no es suficiente, cuando se puede obtener información relevante y precisa, la cual permita ayudar en la toma de decisiones para el ámbito en el que se encuentre. Es de suma importancia que se lleve a cabo en las empresas la integración de inteligencia en los procesos organizativos y de toma de decisiones, lo que conlleva a unir soluciones de visualización y análisis adaptables a las necesidades de negocio.

En una empresa Financiera de Desarrollo Agropecuario, Rural Forestal y Pesquero (FND), existen numerosos reportes y gráficas en Excel que diariamente se envían por correo electrónico a los titulares de las Agencias Estatales para monitorear sus resultados, pero que no pasan de ser simples reportes que incluso habrá usuarios que no le den el significado correcto.

Gracias a las nuevas herramientas que hay en el mercado, se puede facilitar este proceso y además llevarlo a su máxima expresión. Con la ayuda de Tableau y Power BI, la información enviada a cada Agencia Estatal podrá ser más sencilla de presentar, visualizar y sobre todo la posibilidad de detectar áreas de oportunidad para cualquier Estado. Incluso cada Agencia podrá elaborar sus propios reportes de cualquier variable que cada una requiera.

Como objetivo general se pretende realizar un comparativo de las herramientas de visualización de datos Tableau y Power BI, presentar casos de éxito y en base a la investigación obtenida, realizar un análisis de cuál es la mejor opción para el caso presentado.

De manera específica se pretende**:**

- Indagar el estado del arte acerca de las herramientas Tableau y Power BI.
- Analizar ventajas y desventajas de las herramientas para realizar un comparativo.
- Exponer casos exitosos en los cuales se aplicaron ambas herramientas de visualización de datos.
- Dar opinión acerca de cuál es la mejor opción y la que mejor se ajusta de acuerdo a las necesidades del usuario.

El propósito es comparar el uso y funcionamiento de las herramientas de visualización de datos Tableau y Power BI para sugerir la aplicación de alguna de estas en el área de trabajo. Lo anterior se realizará identificando fortalezas y debilidades, ajustando a las necesidades del usuario, conforme al tipo de información que se maneja y los reportes que se puedan generar.

Al investigar y conocer las características que poseen las herramientas de visualización de datos, Tableau y Power BI, se podrá sugerir el uso de alguna de ellas en el área de trabajo, lo cual permitirá facilitar el uso de los datos, poderlos convertir en la mayor información posible y que esta información pueda ser

presentada de tal manera que simplifiquen su entendimiento y se puedan detectar áreas de oportunidad.

11.2. Marco de referencia

11.2.1. Inteligencia de Negocios (Business Intelligence)

Hoy en día, ante los inminentes volúmenes de datos que se manejan diariamente, todas las compañías requieren de habilidades para tomar decisiones de forma correcta y rápida y así poder llegar a los resultados deseados. Para lograrlo existen múltiples opciones de solución empresariales, tecnologías y metodologías que pueden realizar la conversión de información a conocimiento.

La Inteligencia de Negocios (BI, *Business Intelligence*) por sus siglas en inglés, definida por Gómez y Bautista (2010), como una herramienta que cualquier organización, sin importar el giro al que se dedique, pueden soportar la toma de decisiones basadas en información precisa y oportuna; esta herramienta, que no son más que estrategias, productos, tecnologías y técnicas enfocadas a administrar y crear conocimiento.

A través del análisis de datos, BI va a garantizar que el conocimiento generado es el necesario para que permita elegir una alternativa que en ese momento sea la más conveniente para el éxito de la empresa, es decir mediante la aplicación de alguna herramienta de visualización de datos, como lo son Tableau o Power BI, será más rápida identificar aquellas áreas de oportunidad o incluso el punto al que se quiere llegar.

Cada compañía en el mundo sin duda, cuenta con su propio sistema de información diseñado específicamente para el sector de negocios al que atiende, dicho sistema será tan sencillo o robusto dependiendo de las necesidades de la misma, todos los datos generados, acabaran siendo volúmenes de datos obsoletos y solo ocuparan espacio, sin considerar que con el transcurso del tiempo estos datos son una fuente de conocimiento, básicamente la historia de la organización y que

en muchas ocasiones pueden ser tan poderosos para argumentar cualquier decisión.

"El poder competitivo de una empresa se basa en la calidad y cantidad de la información que sea capaz de usar en la toma de decisiones. Esta información aporta una diferencia o ventaja sobre el mercado, lo que se convertirá en conocimiento. Con la implementación de Inteligencia de Negocios se proporcionan las herramientas para aprovechar los datos almacenados en las bases de datos para utilizar la información como respaldo a las decisiones, reduciendo así, el efecto negativo que puede traer consigo una mala determinación" (Gómez y Bautista, 2010).

Gracias a que los datos son tratados, ordenados y reorganizados de tal manera que se pueda obtener sabiduría de estos, pasan de ser simples datos a información y esta información por consiguiente se trasforma en conocimiento, a esto se le llama ciclo del conocimiento. (Rodríguez, 2017).

Figura 1. Ciclo del conocimiento

Fuente: (Rodríguez, 2017).

La inteligencia de negocios definida como la habilidad corporativa para tomar decisiones, mediante el uso de metodologías, aplicaciones y tecnologías que permitan la reunión, depuración y transformación de los datos, para aplicar en ellos técnicas analíticas de extracción de conocimiento (Rud, 2000).

No habrá mayor precisión en la información que arroja un sistema de Business Intelligence, el cual responda las interrogantes que todo gremio quiere saber a cerca de su negocio; el ¿Qué pasó?, ¿Qué pasa ahora?, ¿Por qué pasó?, ¿Qué pasará? (Caralt, 2010).

Business Intelligence, Minería de Datos u Open Data, guardan una estrecha relación con Big Data: la cual es usada para hacer una descripción de la rápida integración y análisis de la información a escalas muy grandes, y como lo menciona Tascón (2013), “Frente a las denominadas “3V”, las cuales constituyen su esencia: Volumen, Variabilidad y Velocidad, además sugiere el surgimiento de una cuarta, la V de Visualización”.

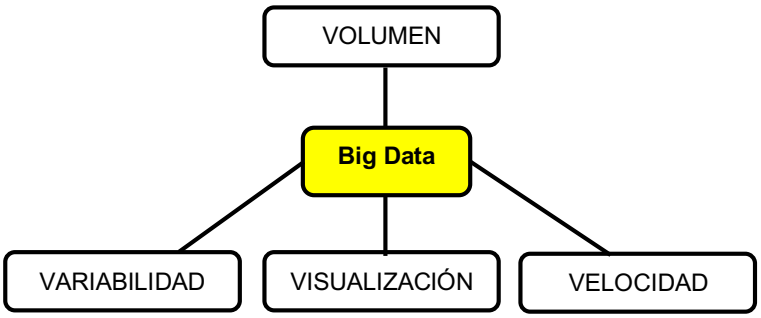


Figura 2. La Big Data y las 4V´s

Fuente: Táscón (2013)

La **variabilidad**, se refiere a los formatos y fuentes en los que se encuentran los datos, es la práctica de incorporar datos recopilados originalmente para propósitos dispares, análisis combinados, como por ejemplo registros médicos electrónicos combinados con las compras, historiales o actualizaciones de perfiles de redes sociales. En el caso de “FND” como ejemplo puede ser tipos de créditos que se ofrecen combinados con los índices de cartera vencida asociados además al género de las personas que son morosas para pagar.

El **Volumen** se refiere a la cantidad que se genera de datos en un espacio de tiempo determinado, en el caso de “FND” se relaciona con el número de créditos, tipos de crédito, monto total de créditos, índices de cartera vencida, índice de recuperación de cartera, entre otros, que se generaron en un periodo de tiempo específico.

La **Velocidad**, se puede referir al proceso de generación de datos, los cuales se compilan y analizan en tiempo real o casi en tiempo real, mediante algoritmos

que son operados sin la intervención humana, la rapidez de entrada y salida de los datos los cuales fluyen por los distintos canales, y en el caso de “FND” que tan rápido se puede extraer dicha información del sistema que el área de trabajo tiene.

11.2.2. Herramientas de visualización de datos

El análisis de datos y de negocios son disciplinas que han evolucionado en organizaciones y empresas prácticamente en todos los campos de saber, lo cual permite tener una decisión más eficaz y eficiente, se habla de tener un resultado además de acertado, rápido. Debido a esto, las herramientas de inteligencia de negocios han ido recogiendo las tecnologías como lo son OLAP (Procesamiento Analítico en Línea) por sus siglas en inglés, de informes y consultas (Reporting and Query), de visualización y de minería de datos tanto web, texto, medios sociales e incluso han llegado a técnicas de análisis de minería de sentimientos y minería de opinión. (Aguilar, 2016).

De acuerdo al estudio realizado por Turner y otros (2014), revela que, por increíble que esto parezca, no se ha llegado ni al 1% del análisis de los datos que se generan en el mundo, situación que es alarmante, dado que se pierde la enorme oportunidad de poder obtener información tan valiosa y necesaria para todos. La visualización de los datos se encargará de captar de manera inmediata la atención de quien tendrá que tomar decisiones importantes en una empresa, mediante un diseño que permita comprender la información presentada.

Llegado a este punto, surge la pregunta ¿qué son las herramientas de visualización de datos?

Cuando se habla de herramientas de visualización de datos, se refiere a diseñar o transformar los mismos, en gráficos que faciliten su comprensión, el objetivo será obtener información y dar soporte a la toma de decisiones, por tanto,

la información deberá comunicarse de una manera clara apoyándose de instrumentos que faciliten su comprensión.

La visualización de datos consiste en convertir la información en una imagen que ayude a identificar el significado de tantos datos almacenados. Es decir, del mundo de información que se maneja, se puede identificar fácilmente los puntos de oportunidad o bien si se llega o no al punto esperado.

Cuando se tiene a la mano toda esta información sintetizada es mucho más fácil identificar sobre todo tendencias, patrones, estilos que van a llevar muy seguramente al éxito. Una imagen vale más que mil palabras dice un refrán y acertadamente, todos conocen que es mucho más fácil analizar una imagen que cientos de registros o información organizada en tablas o gráficas básicas.

La visualización de los datos se encargará de “diseñar la comprensión de la información”.

Tabares y Hernández (2014), mencionan que el primer desafío que tiene la visualización de datos es el poder representar el conocimiento de tal manera que las personas puedan entenderlo, esto gracias al uso de los mecanismos o herramientas tan sofisticadas que hoy existen, aunque dichas herramientas estarán en constante enfrentamiento a inconvenientes por la gama tan grande de tamaños y diversidades de datos.

Como se ha mencionado, las organizaciones no alcanzan a aprovechar las ventajas de las múltiples oportunidades que tienen con toda la información que generan, el reto que tienen es el de poder separar, lo que es relevante y lo que no, todo siempre medido por tiempos y entre menor sea el tiempo de respuesta es mucho mejor.

Microsoft ha desarrollado y renovado herramientas como Power BI, pero además, existen en el mercado diversas opciones de software como Tableau, diseñadas para el tratamiento de los datos y su clasificación va de acuerdo al uso que se le da a la herramienta, o bien para el fin que fue desarrollada. Algunas están orientadas en la utilización de gráficos, diagramas y/o infografías, o incluso de su

uso gratuito. A continuación, se muestran algunas herramientas orientadas a la visualización de datos.

11.2.2.1. Tableau

Tableau es una de las herramientas de visualización de datos más populares y completas disponibles en el mercado. Su interfaz sencilla e intuitiva permite generar visualizaciones sobre grandes volúmenes de datos. Cuenta con un funcionamiento sencillo, además de la posibilidad de personalizar las informaciones. Tableau ofrece una versión gratuita.

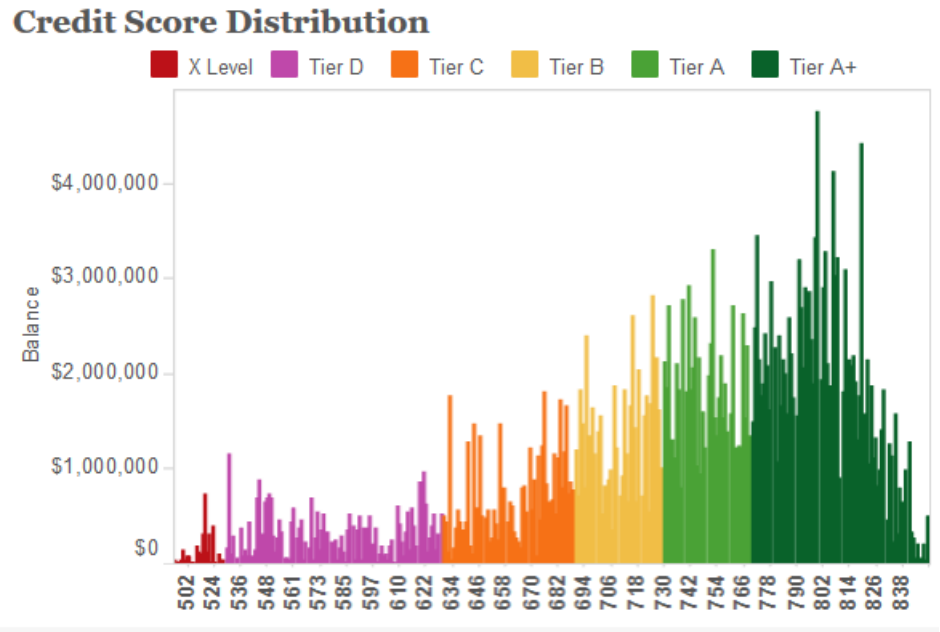


Figura 3. Ejemplo de gráfico de Tableau

Fuente: Tableau (2018a)

11.2.2.2. Qlik

La empresa Qlik ofrece dos productos para el análisis de datos: QlikSense y QlikView. QlikSense es una plataforma de análisis de datos en la nube. En él se pueden combinar datos de diferentes fuentes, sin importar lo extensa o complejas

que puedan ser. Cuenta con capacidades de Inteligencia Artificial (IA), búsquedas en lenguaje natural y soluciones interactivas.

QlikView por su parte, permite producir rápidamente aplicaciones y dashboards (tableros de mando) de analítica guiada interactiva, permitiendo a los usuarios tomar decisiones basadas en datos. Es una herramienta de Bussines Intelligence (Qlik, 2018)

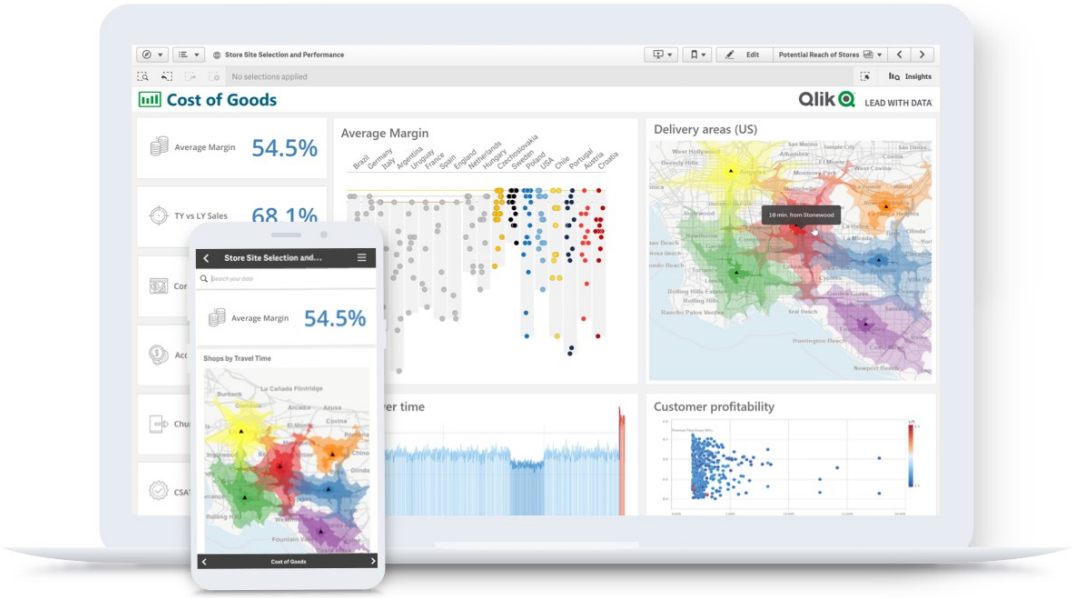


Figura 4. Qlink
Fuente: (Qlik, 2018)

11.2.2.3. Plotly

Plotly es una herramienta Web diseñada para el análisis y visualización de datos. Costa de una librería de gráficos científicos para Python, R, MATLAB, Perl, Julia, Arduino y REST. Con Plotly se pueden crear diferentes variedades de gráficos a partir de diversas fuentes de datos (Plotly, 2018). Esta herramienta es usada para generar desde gráficos sencillos hasta gráficos para Inteligencia Artificial, Machine Learning y Ciencia de los Datos. Cuenta con la capacidad de generar tableros interactivos (dashboards) para personalizar el análisis y visualización de los datos. En la siguiente figura se muestra un ejemplo de un dashboard realizado con plotly.

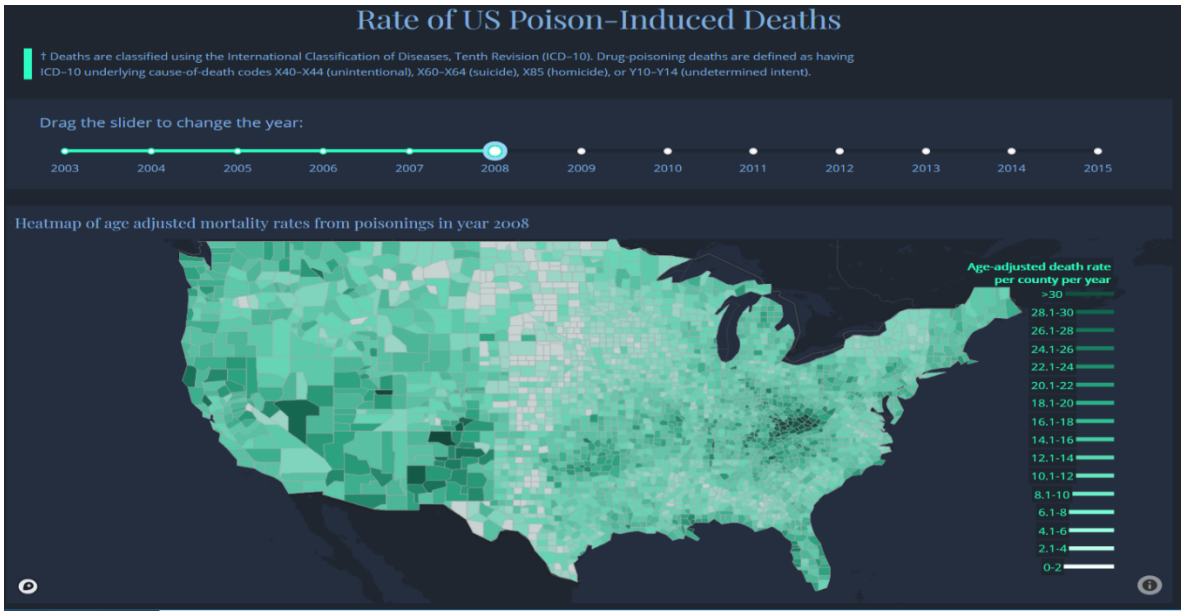


Figura 5. Dashboard Plotly
Fuente: (Plotly, 2018)

11.2.2.4. Carto

Carto es una herramienta para el análisis espacial. La plataforma de inteligencia de ubicación de CARTO permite a las organizaciones almacenar, enriquecer, analizar y visualizar los datos para la toma de decisiones espacialmente conscientes (Carto, 2018). Puede ser utilizado para analizar territorios de ventas o diseñar cadenas de suministro.

Figura 6. Análisis espacial Carto
Fuente: (Carto, 2018)

11.2.2.5. DataWrapper

Al igual que las otras herramientas, su interfaz es sencilla, intuitiva y clara. Esta herramienta ofrece la creación de gráficas, mapas y tablas. Es fácil de usar, responsiva, sin necesidad de instalar o actualizar (trabaja desde la nube), no requiere de códigos de programación (DataWrapper, 2018).

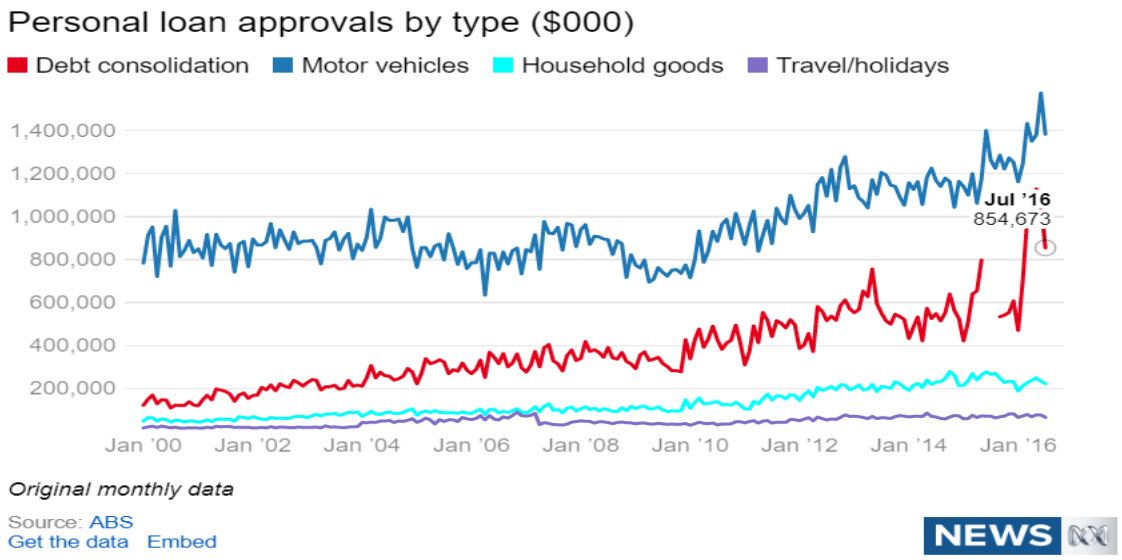


Figura 7. DataWrapper Chart

Fuente: (DataWrapper, 2018)

11.2.2.6. PowerBI

Power BI es un servicio de análisis de negocio basado en la nube que proporciona una vista única de los datos más críticos. Es una herramienta desarrollada por Microsoft y permite cientos de visualizaciones de datos, funcionalidades de inteligencia artificial integradas, integración perfecta de Excel y conectores para diferentes fuentes de datos (Power BI, 2018).

Figura 8. Power BI

Fuente: (Power BI, 2018)

En los siguientes puntos se detalla más características de las herramientas Tableau y PowerBI.

11.2.3. Herramienta Tableau

Tableau es una compañía de software fundada por Chris Stolte, Christian Chabot and Pat Hanrahan, en 2003 en Seattle, Estados Unidos, desarrolla productos interactivos de visualización de datos orientados en la Inteligencia de Negocios (Business Intelligence).

Sus fundadores realizaron una búsqueda de técnicas de visualización para llevar a cabo la exploración y análisis de datos relacionales y aunado a eso la combinación de un idioma estructurado de búsqueda por bases de datos y un lenguaje descriptivo que permitiera representar gráficos los cuales resultaron en la herramienta que hoy podemos utilizar de una forma muy práctica y sencilla.

Uno de los múltiples propósitos que Tableau ofrece, es el de ayudar a que las personas vean y comprendan sus propios datos, que puedan responderse sus propias preguntas, mediante un análisis mucho más rápido, útil, fácil y además

agradable a la vista. Todo esto representa una de las misiones que tiene el siglo XXI. Esta compañía representa el claro ejemplo de usar sus propios productos que desarrollan y ofrecen al mercado.

Dispone de diferentes servicios como análisis de encuestas, sitios webs, redes sociales, series temporales, hasta herramientas Big Data o DashBoards de negocios. Esta herramienta es tan potente que, puede hacer que una marca comercial tenga índices de popularidad más elevados, mediante el análisis de patrones en las redes sociales o un análisis estadístico de enfermedades gastrointestinales de un sector de la sociedad en un periodo de tiempo, por mencionar algunos (Ruíz, 2018).

Tableau como otras herramientas de visualización, ofrece además otros recursos adicionales para dar soluciones a las necesidades específicas de cada usuario, como lo son:

- **Tableau DeskTop (Tableau Escritorio):** Ediciones profesional y personal. Esta aplicación va a permitir analizar datos estructurados cualquiera que fuera la fuente, mediante la cual se podrá generar todo tipo de gráficos, informes interactivos, paneles de control, todo en un tiempo muy corto. La interfaz es realmente muy amigable, por lo que cualquier persona fácilmente puede familiarizarse con la herramienta.
- **Tableau Server (Tableau Servidor):** Es la solución de inteligencia comercial que ofrece análisis visual desde el navegador. Esta aplicación lo que va a permitir es el compartir o ver las publicaciones de otros usuarios que se encuentren en el servidor.
- **Tableau Online (Tableau en Línea):** es una extensión de Tableau Server la cual centraliza los datos en la nube y permite hacerlos públicos cuando los compartes. Publica orígenes de datos de Tableau Desktop, define conexiones de datos, agrega conexiones de datos en la nube y actualiza los datos de Salesforce y Google Analytic. BigQuery y Amazon Redshift.
- **Tableau Reader (Tableau Lector):** Permite la visualización de manera gratuita prácticamente a cualquier persona, para interactuar con los libros de

trabajo elaborados en Tableau Desktop, es preciso indicar que este producto es gratuito.

- **Tableau Public (Tableau Público):** es una herramienta de visualización de datos de acceso libre y gratuito, en donde cualquier persona en cuestión de minutos puede publicar datos, crear visualizaciones, ubicarlos y compartirlos directamente en sitios web, sin necesidad de escribir código.
- Existen las **Tableau Community**, las cuales permiten a los usuarios interactuar por medio de foros en línea y reafirmar lo aprendiendo o consultar dudas o experiencias con otros usuarios en tiempo real.

11.2.4. Herramienta Power BI

Power BI es una colección de servicios de software, aplicaciones y conectores que funcionan conjuntamente para convertir orígenes de datos sin relación entre sí en información coherente, interactiva y atractiva visualmente. Esto puede realizarse partiendo desde una sencilla hoja de cálculo de Excel como de una colección de almacenes de datos híbridos locales o basados en la nube, Power BI le permite conectar fácilmente los orígenes de datos, visualizar lo más importante y compartirlo al instante si así se desea.

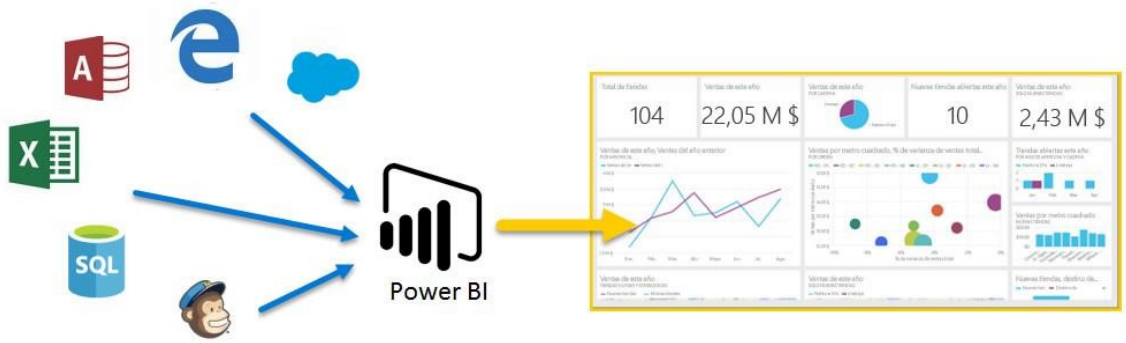


Figura 9. Power Bi entrada

Fuente: (Power Bi, 2018)

Por sus características representa una herramienta sencilla y rápida; con la gran capacidad de crear información rápida a partir de una hoja de cálculo de Excel o una base de datos local. Esta herramienta permite convertirse en un instrumento

personalizado de creación de informes y visualización, así como actuar como el motor de análisis y de decisión que impulsa proyectos en grupo, divisiones o empresas enteras. (Power Bi, 2018)

- **Power BI Desktop:** es el software donde se procesa y maquila toda la información que se importó desde cualquier fuente de datos. Se conecta con varias extensiones de datos, permite realizar consultas, crear y diseñar reportes e informes de inteligencia empresarial.
- **Power BI Service:** Es un servicio basado en la nube que proporciona una vista única de los datos más críticos del negocio, ya que permite crear y publicar informes para supervisar el estado de la empresa.
- **Power BI Mobile:** Es una aplicación que se conecta a cualquier dispositivo móvil (teléfonos y Tablet) desde cualquier lugar, permite acceder a la información empresarial mediante acceso directo e interactivo para ver los informes. Disponible para Android como IOS.

Figura 10. Las partes de Power BI

Fuente: (Power Bi, 2018)

11.2.5. La empresa

En este apartado se documenta acerca de la empresa a dónde se dirige la propuesta de uso de herramientas de visualización de datos.

Financiera Nacional de Desarrollo Agropecuario, Rural, Forestal y Pesquero (FND) es un organismo descentralizado de la Administración Pública Federal,

sectorizado en la Secretaría de Hacienda y Crédito Público (Banca de Desarrollo), que impulsa el desarrollo del medio rural a través de créditos accesibles para pequeños productores y MIPYMES.

Actualmente existen 5 Coordinaciones Regionales (Centro-Occidente, Noroeste, Norte, Sur y Sureste), que de acuerdo a su ubicación geográfica se clasifican cada uno de los estados del país. Cada estado es representado por una Agencia Estatal de Crédito, que también pueden subdividirse en Agencias de Crédito y Módulos de Atención, según sea el caso.

Cabe mencionar que el Estado de Durango pertenece a la Coordinación Regional Norte, cuenta con una Agencia Estatal de Crédito en el municipio de Durango, una Agencia de Crédito en Guadalupe Victoria y 2 Módulos de Atención en Santa María del Oro y Tamazula.

Figura 11. Distribución geográfica de las Coordinaciones Regionales de FND en el País

Fuente: (FND, 2018)

Misión:

Impulsar el desarrollo del medio rural y de las actividades del sector primario, a través del crédito y servicios financieros accesibles a productores, intermediarios financieros rurales y otros agentes económicos, con la finalidad de elevar la productividad y contribuir a mejorar el nivel de vida de la población.

Visión:

Ser la mejor opción de financiamiento para el medio rural y actividades vinculadas con el sector primario del país, reconocidos por la alta calidad en el servicio y comprometidos con el éxito de cada proyecto, que resulta en la generación de valor en la comunidad y la sustentabilidad de la institución.

Cada una de las Agencias Estatales, Agencias de Crédito y Módulos de atención es evaluada y monitoreada diariamente conforme a la meta establecida anualmente y mediante una serie de indicadores establecidos en los reportes que se dan a conocer vía correo electrónico. Los reportes enviados obedecen a los logros obtenidos del 1 de enero de cada ejercicio fiscal al día anterior del reporte y las variables que se consideran, entre otras, son: avance en la colocación de créditos, índices de cartera vencida, colocación acumulada al mismo periodo del ejercicio anterior, porcentajes de cumplimiento de la meta establecida anualmente y recuperación de cartera. Dichos reportes se realizan en una hoja de cálculo de Excel, presentadas en una forma plana, es decir, sin gráficos.

En la institución, cada Agente de Crédito está familiarizado con el formato, los indicadores y en cierta forma el renglón - columna donde va el porcentaje que les falta para cumplir la meta, o el porcentaje de cartera vencida para considerarse en la media.

Es preciso proponer una fuente de información más dinámica que les permita en una forma casi instantánea ver el comportamiento de cada Agencia y/o Módulo de atención e incluso los logros obtenidos por cada empleado con funciones de Ejecutivo de Financiamiento. Como propuesta, es en primera instancia, poder identificar mediante un comparativo cuál de las dos herramientas de visualización de datos puede ajustarse mejor a las necesidades del usuario.

Se propone implementar la herramienta de visualización de datos en la Agencia Estatal Durango a modo de prueba para *“evaluar los logros obtenidos de cada Ejecutivo de Financiamiento*”, quienes son la parte fundamental para el logro de las metas, debido a que son los ejecutores de dicha información.

La propuesta consiste en generar reportes que proporcionen información valiosa para el Agente Estatal y poder detectar áreas de oportunidad para la Agencia. Luego proponer la implementación de la herramienta a escalas mayores e incluso su uso a nivel Corporativo.

11.3. Desarrollo

Existen en el mercado diversas y muy completas herramientas de visualización. En este trabajo se tomaron en consideración Tableau y Power BI.

Estas herramientas van a permitir crear reportes de visualizaciones de datos dinámicos y sencillos, además de ofrecer servicios que son gratuitos y de los cuales se puede obtener muchos beneficios, además de que su interfaz es sumamente amigable lo cual permite que el usuario pueda manejarlas sin problema.

A lo largo del desarrollo de esta propuesta se detallarán más aspectos acerca de Tableau y PowerBI.

11.3.1. ¿Por qué Tableau?

La mayoría de los autores o publicaciones de artículos en internet hablan de Tableau, por su facilidad y rapidez para generar reportes o informes de análisis de datos. Es de las más utilizadas para la Inteligencia de Negocios lo cual es idónea para la propuesta de implementación en “FND”, ya que se requiere tener al día la respuesta a las preguntas ¿Cómo vamos?, ¿Cuánto nos falta?, entre otras.

Con esta herramienta resulta muy sencillo elaborar diagramas de interacción dinámicos, además que los gráficos que utiliza y los colores que se pueden usar realmente captan la atención deseada. Esto puede parecer una característica sin trascendencia para muchos, pero hay estudios relacionados sobre el uso de los colores y formas en que debe presentarse la información, de eso depende que el usuario capte o no su atención. Es más fácil y práctico identificar la información a partir de una imagen.

11.3.1.1. Instalación de la herramienta Tableau Desktop

La aplicación se descarga desde la URL https://public.tableau.com/s/. Por ser un producto internacionalizado, son compatibles con Unicode y con los datos almacenados en cualquier idioma. La interfaz y la documentación cuenta con 7 idiomas incluyendo el español.

Es necesario ingresar una cuenta de correo electrónico para realizar la descarga. Se captura la cuenta de correo en la caja de descripción y enseguida presionar botón de descargar la aplicación, se elige la opción correspondiente al sistema operativo con el que cuenta el equipo de cómputo (Windows o Mac).

Al terminar la instalación, la herramienta ofrece un libro digital para su aprendizaje. En Tableau Public se muestran las primeras opciones que tiene la aplicación, como la fuente de datos con la que se desea trabajar. Ciertamente en muchas empresas las principales fuentes de datos son archivos de Excel, aunque existen múltiples opciones con las que se puede trabajar en Tableau Public: Archivos de texto, Access, JSON, archivos de estadística, entre otros.

11.3.1.2. Requerimientos del Sistema

- Sistema Operativo: Windows 7 o posterior, para Mac OSX 10.11 o posterior.
- Navegadores web: Chrome en Windows, Mac y Android 4.4 o posterior, Microsoft Edge e Internet Explorer 11 en Windows, Mozilla Firefox y Firefox ESR en Windows y Mac, Apple Safari en Mac y iOS 8.x o posterior
- Fuentes de Datos: Microsoft Excel, Archivos de texto: archivos de valores separados por comas (.csv), archivos JSON, PDF, Archivos espaciales (ESRI Shapefiles, KML, GeoJSON y MapInfo), Archivos estadísticos; SAS (*.sas7bdat), SPSS (*.sav) y R (*.rdata, *.rda), Odata, Hojas de Google, Conectores de datos web.

11.3.1.3. ¿Cómo funciona Tableau?

Una vez terminado el proceso de instalación se ejecuta el acceso directo para abrir una hoja de Tableau Public, Se selecciona el archivo de datos con el que se desea conectar. Ver la figura 12.

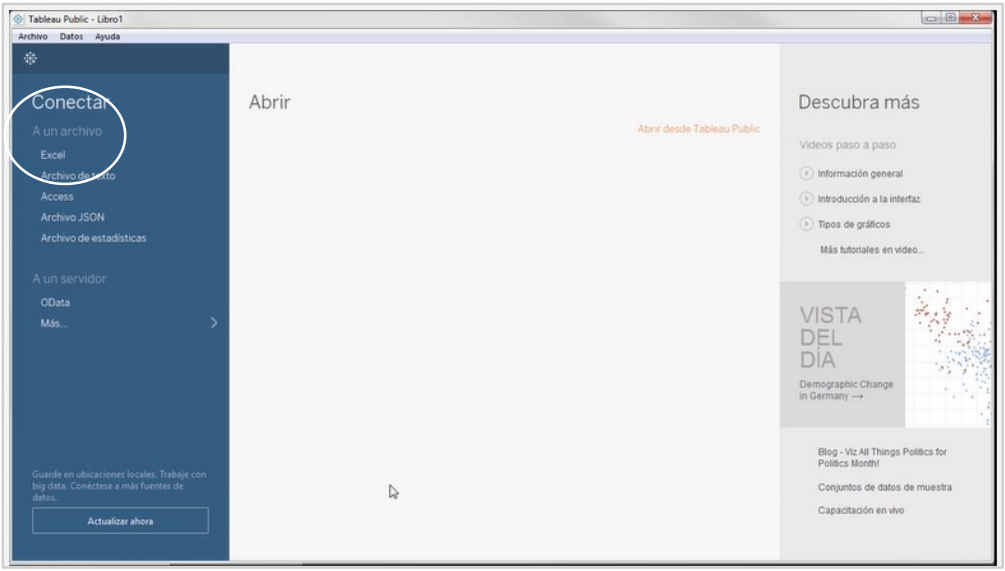


Figura 12. Conectar archivo de datos

Fuente: (Tableau, 2018b)

Para este ejemplo se eligió un archivo de Excel, que al seleccionarlo despliega una pantalla en la cual se deberá de indicar la ruta donde se encuentre ubicado el archivo. Ver la figura 13.

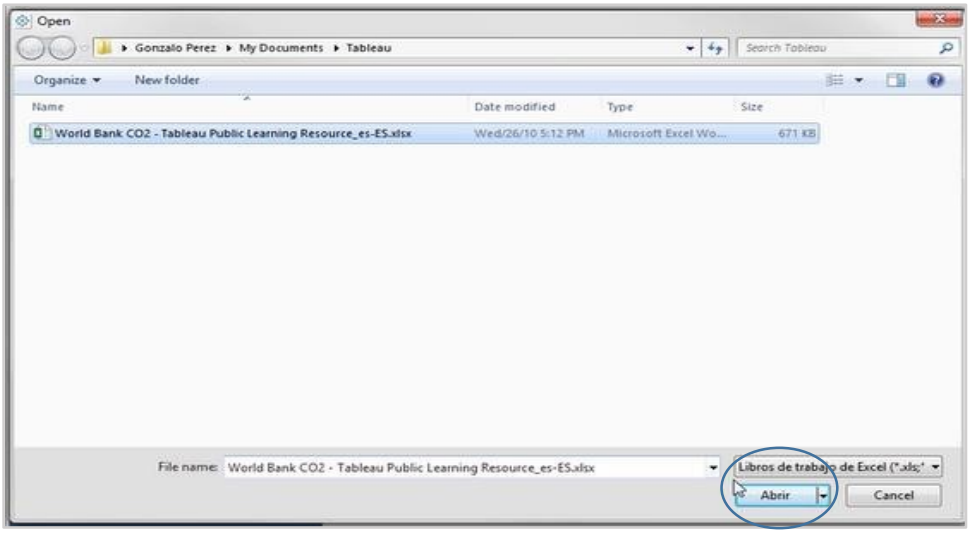


Figura 13. Selección de la fuente de datos

Fuente: (Tableau, 2018b)

En este punto Tableau Public se encuentra vinculado con el archivo seleccionado, se puede visualizar arrastrando el archivo que aparece en el panel izquierdo de la pantalla y soltarlo en el área que indica Arrastrar hojas aquí. Ver la figura. 14.

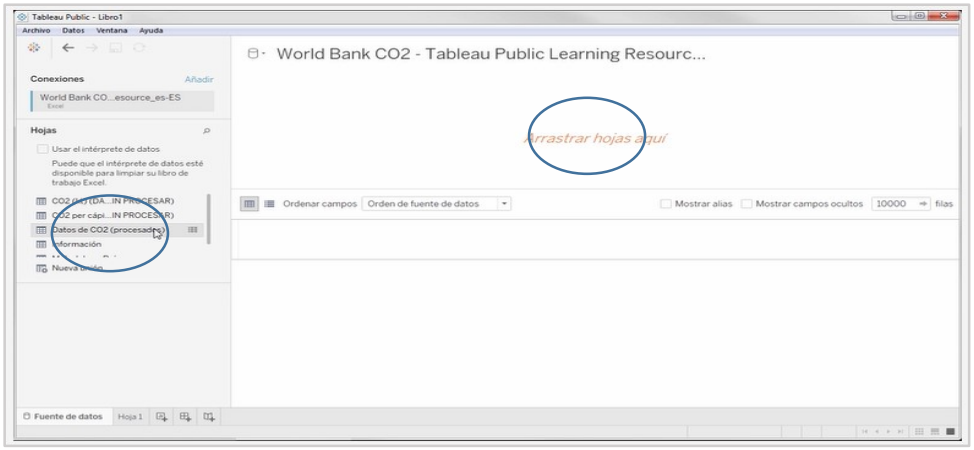


Figura 14 Vinculación de datos de Excel

Fuente: (Tableau, 2018b)

Se comenzará a tener una pequeña visualización en forma automática del conjunto de datos que contiene el archivo, los cuales se pueden identificar mediante el nombre del encabezado que aparecen en cada columna, también se podrá modificar tanto el nombre de la columna como la estructura del conjunto de datos si así se requiere. Ver la figura 15.

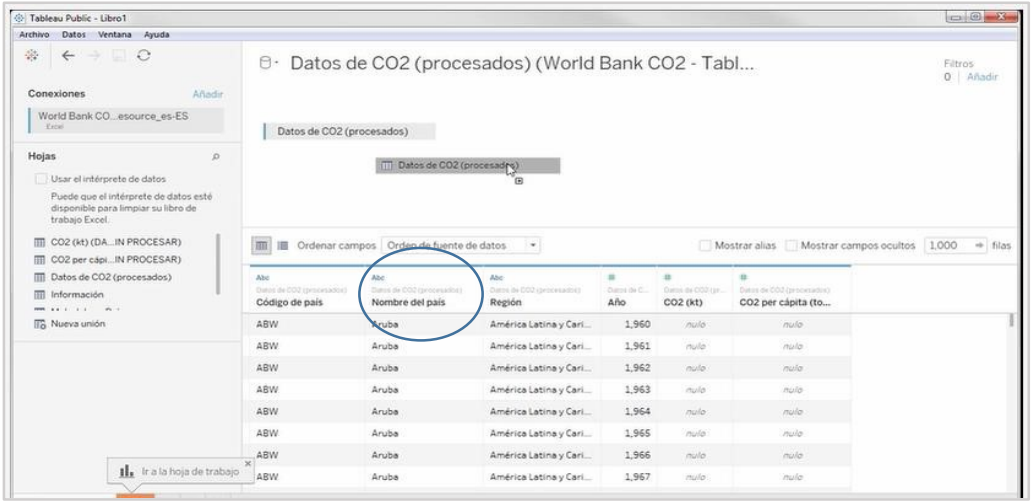


Figura 15. Visualización de los conjuntos de datos
Fuente: (Tableau, 2018b)

Para comenzar a realizar análisis de los datos se requiere abrir una hoja de trabajo, esta opción se encuentra ubicada en la parte inferior izquierda de Tableau Public, tal y como se ilustra en la figura 16.

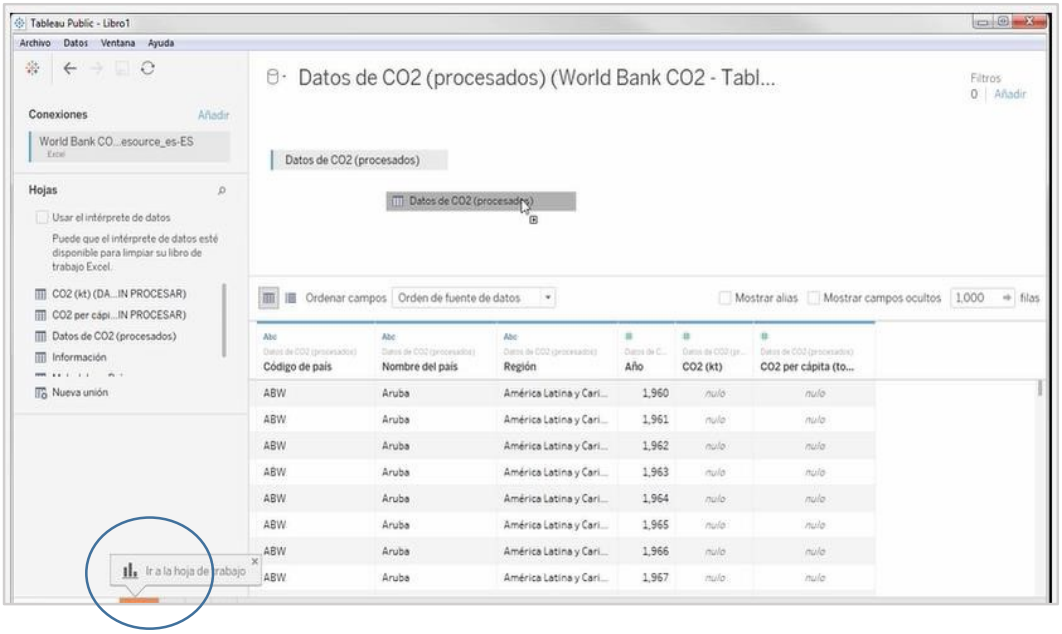


Figura 16. Ir a la hoja de trabajo
Fuente: (Tableau, 2018b)

Enseguida se selecciona el campo a graficar y se arrastra el campo al lienzo llamado Arrastrar un campo aquí, como se muestra en la figura 17. Esta representa una de las formas más sencillas de crear gráficos.

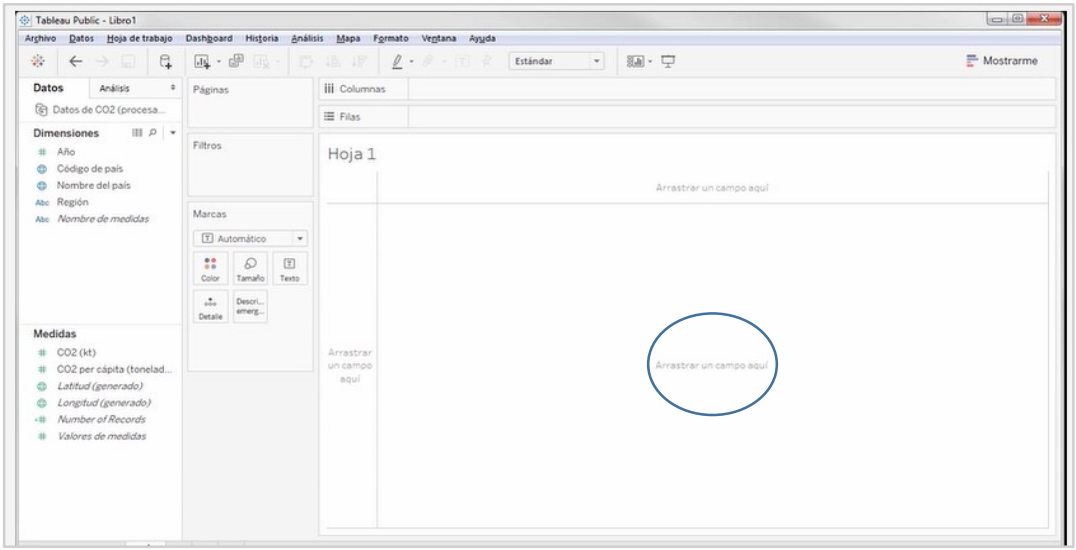


Figura 17. Lienzo para crear gráficos
Fuente: (Tableau, 2018b)

Esto permite visualizar los datos en forma casi instantánea y conforme se va agregando datos al lienzo la gráfica, ésta va realizando cambios que permite a primera vista identificar cual es el que más o cual es el que menos, conforme a las variables que se estén evaluando.

En la Figura 18, se puede observar que la primera gráfica solo se señala que países se evalúan, pero en la siguiente gráfica, cuando se le agrega otra variable al lienzo de trabajo, la gráfica cambia para mostrar cada país del mundo y que cantidad de CO2 son emitidas cada año.

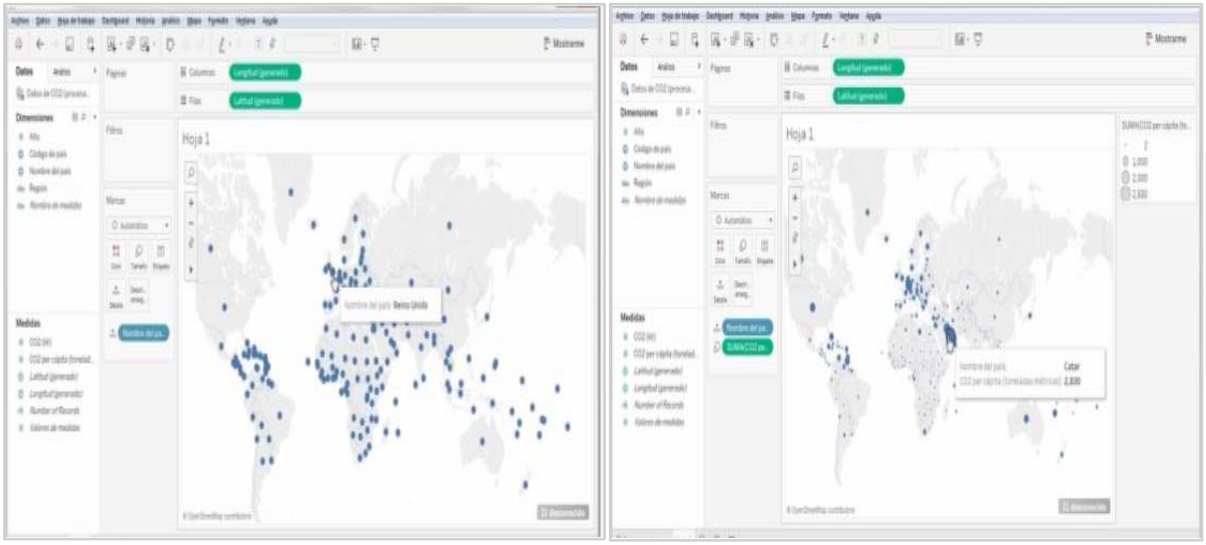


Figura 18. Visualización de datos
Fuente: (Tableau, 2018b)

11.3.1.4. Recursos adicionales que ofrece el portal de Tableau

- Videos paso a paso
- Datos de muestra
- Capacitación en vivo
- Además de brindar una certificación comprobando conocimientos técnicos con un examen de Qualified Associate en la siguiente dirección url https://www.tableau.com/es-es/support/certification.

11.3.2. Casos exitosos en la aplicación de la herramienta Tableau

A continuación, se presentarán algunos casos sobre empresas que experimentan el uso de la herramienta Tableau en sus procesos de trabajo.

11.3.2.1. Nacional Financiera – Análisis Financiero

A manera de reflexión esto es lo que menciona Guillermo Martínez Ceballos, Gerente de Información Financiera de Nacional Financiera:

"*Si no tuviésemos a Tableau, lo que se estaría haciendo ahora es esperar a que las áreas operacionales realicen compilaciones de informes, lo que tomaría de 10 a 15 días, o hasta un mes*". (Martínez, sf)

Hoy en día, los ejecutivos de Nacional Financiera tienen acceso a importantes puntos de datos con 24 horas de antelación. Al ajustar su estrategia de planificación sobre la marcha a lo largo de sus diferentes áreas de negocio (marketing, promociones y contabilidad), se encuentran en una mejor posición para alcanzar los objetivos anuales.

Con Tableau, el equipo trabaja desde una sola fuente, lo que conlleva una mayor colaboración y precisión. El volumen de datos es de 30 millones de registros debido a que tienen 15 años de historia. Cuentan con una data warehouse, el cual les ha permitido almacenar tal cantidad de información. Al ser una institución financiera, se debe dar seguimiento al cumplimiento de las metas, así como los indicadores que se tienen proyectados anual y mensualmente e ir anticipando incidencias.

Gracias a los tableros se puede comparar la información que se tiene con la información programada para ir viendo si se va cumpliendo o no el objetivo. Tableau permitió contar con la información 24 horas antes, lo que permite que tanto los informes ejecutivos como los informes de análisis que se realizan se vayan viendo con anticipación para ver incidencias, para los cierres de mes.

11.3.2.2. GNP centraliza sus datos y gana agilidad en la toma de decisiones

Con Tableau “se deja de ser un área que generaba reportes a ser un área que genera estrategias para la venta” (Cisneros, s.f.). “Con un par de clics se puede identificar quien es el que debe más y tomar acción para recuperar ese dinero” (Cisneros, s.f.).

Con una facturación anual de 56 mil millones de pesos mexicanos, operando en más de 40 oficinas, con 3,000 empleados y 4,250 agentes de ventas, el Grupo Nacional Provincial (GNP), es una de las compañías de seguros más grandes de México. Para gestionar sus datos y transformar el potencial de la información en oportunidades de rentabilidad del negocio, GNP eligió Tableau.

Para llevar a cabo la implementación de Tableau empezó con una prueba gratuita y en la actualidad se extiende por toda la compañía incluyendo el comité ejecutivo. Con Tableau, GNP logró centralizar todas sus bases de información en un único reporte y utilizar el poder de análisis de gerencia de inteligencia comercial para el día a día de la empresa.

11.3.3. ¿Porqué Power BI?

Microsoft Power BI Desktop se creó para los analistas y en el cual se pueden combinar visualizaciones interactivas con consultas y modelado de datos. Se pueden crear y publicar informes en Power BI. Ayuda a facilitar a otras personas información fundamental puntual, en cualquier momento y desde cualquier lugar.

Esta herramienta de visualización permite trabajar con orígenes de datos muy complejos, con lenguaje de programación y opciones encaminadas a expertos en tecnologías de la información. Pero, aun así cualquier usuario con conocimientos medios o avanzados de Excel, será capaz de crear informes mediante la importación de datos de Excel, para darle un giro significativo a la forma en que se presentan los datos a la empresa.

Power BI se puede conectar a Excel, la cual sigue representando aun y cuando existe la diversidad de opciones para el manejo de datos, la más utilizada en el mercado laboral. En Power BI se pueden crear paneles de datos de manera personalizados, lo cual nos aporta mayor flexibilidad de uso, ya que ubicamos el lugar donde se encuentra cada instrumento para la creación de un reporte. (Wong, 2018).

11.3.3.1. Instalación de la herramienta Power BI

Nuevamente se usará el apoyo de secuencias de figuras para mostrar la instalación de la herramienta Power BI Desktop. Se ingresará en el explorador de internet la siguiente URL https://powerbi.microsoft.com/es-es/desktop/. Al igual que Tableau, también ofrece múltiples opciones de idioma incluyendo español, lo cual facilita su instalación. Una vez ubicada la URL, seleccionamos la opción Comenzar gratis como se muestra en la siguiente figura.

Fig. 19. Instalación Power BI, comenzar gratis

Fuente: (Power BI, 2018)

En seguida se abrirá una ventana de diálogo que ofrece 2 opciones: descargar Power BI Desktop para Windows y registrarse en Power BI online. Se selecciona la opción *descargar*. Enseguida ir a opciones avanzadas de descarga para el cambio de idioma.

Figura 20. Power BI Desktop, Descargar

Fuente: (Power BI, 2018)

Microsoft Power BI, puede enviar actualizaciones y trucos para usar la herramienta, mediante un registro de correo electrónico que solicita en el proceso de instalación del software. Elegir la opción *Comenzar* para iniciar descarga.

Es importante comentar que este instrumento cuenta con enlaces para conectar con perfiles sociales de Facebook, de web, fuentes de datos para ejemplos y canal de YouTube con cursos para el aprendizaje de la misma.

Finalmente se indica *Siguiente*, seguido de validar condiciones y términos de licencia, se personaliza la descarga (accesos directos, ruta de la descarga, entre otros) para terminar con la opción Instalar y Finalizar. En la siguiente figura se muestra la hoja de trabajo que maneja Power BI Desktop.

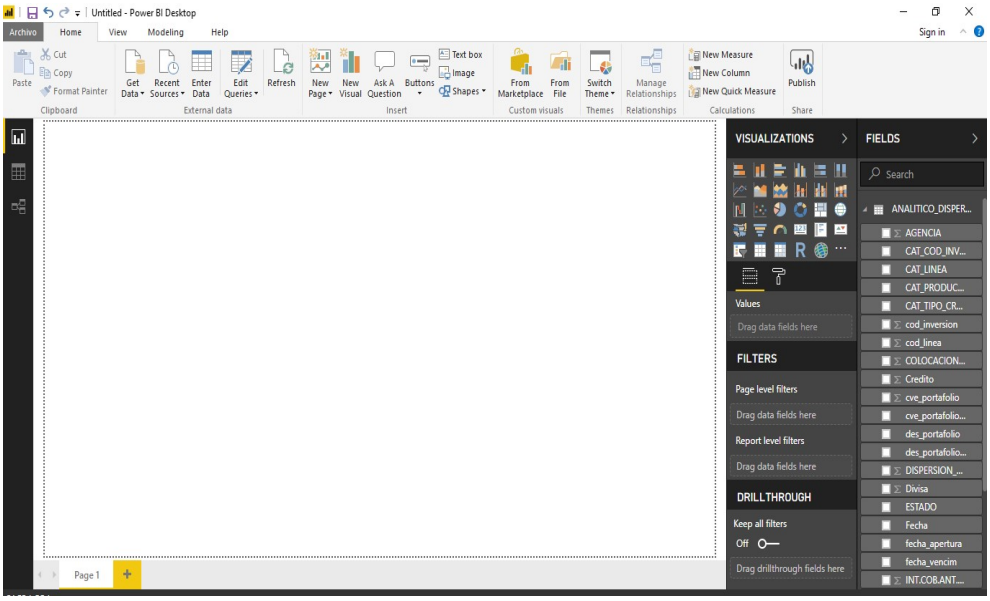


Figura 21 Power BI Desktop, hoja de trabajo

Fuente: (Power BI, 2018)

11.3.3.2. Requerimientos del Sistema

- **Sistema operativo:** Windows 10; Windows 7; Windows 8; Windows 8.1; Windows
- Server 2008 R2; Windows Server 2012; Windows Server 2012 R2
- **Navegadores web:** Internet Explorer 10 o superior.
- **Plataforma:** Disponible de 32 bits (x86) y 64 bits (x64).
- **Fuentes de Datos:** Microsoft Excel, Archivos de texto: archivos de valores separados por comas .CSV, JSON, PDF, XML, SQL, Internet Explorer, Access database,

11.3.3.3. ¿Cómo funciona Power BI?

Power BI consta de 3 partes importantes: **Power BI Desktop, Power BI Service y Power BI Mobile**. Ahora bien, Power BI Desktop es el primer paso en el flujo de trabajo ya que es donde se importan los datos de la fuente que necesite analizar.

Dentro de las particularidades que tiene esta herramienta es que cada usuario puede manejarla de distinta manera a otro usuario lo cual representa una ventaja por la facilidad de adecuarla al rol que juega cada usuario. Entendiendo que uno de ellos puede estar en un área diferente y aunque los datos sean los mismos se puede extraer reportes diferentes o incluso el mismo, pero visualizarlo en un informe Desktop o Power BI Server y algún otro desde su dispositivo móvil.

Después de la correcta instalación de la herramienta, se ejecuta el acceso directo que habilito el asistente de instalación. Se abrirá una pantalla como se muestra en la siguiente figura.

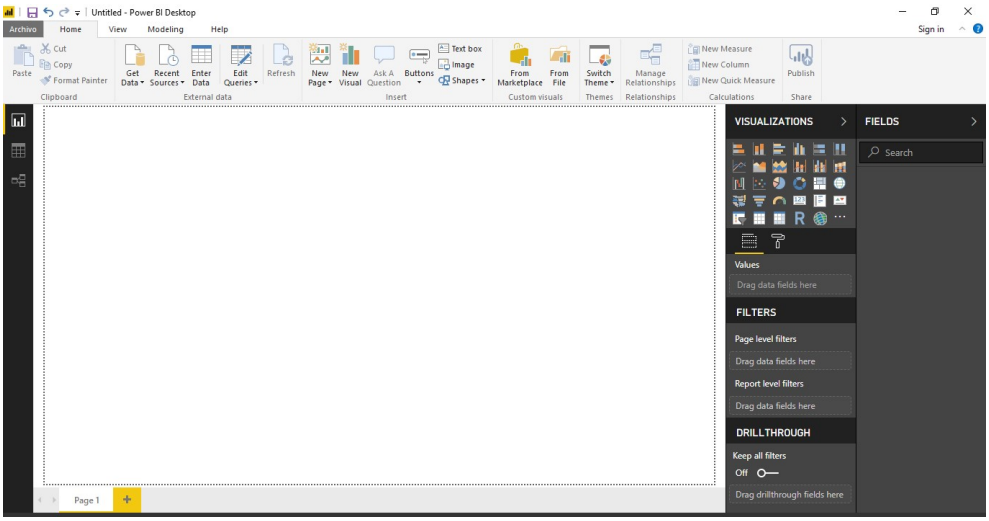


Figura. 22. Power BI Desktop, hoja de trabajo

Fuente: (Power BI, 2018)

Se debe conectar a la fuente de datos con la que se desea trabajar, en la siguiente figura se muestran las opciones que ofrece. Por ser Excel la fuente de datos que mayormente se utiliza, se ejemplificara con una base de datos.

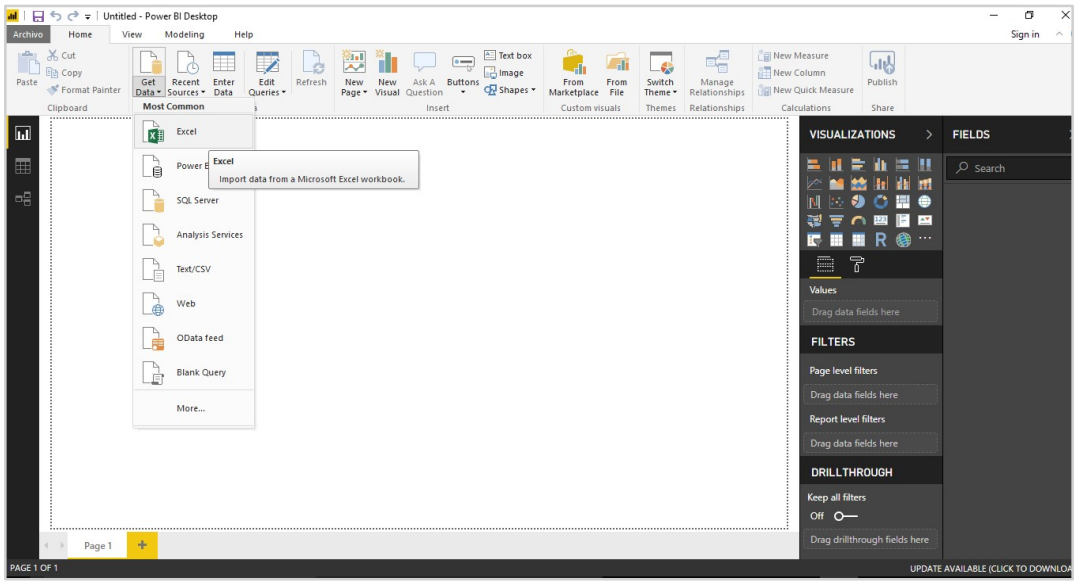


Figura 23. Opción de selección de fuente de datos
Fuente: (Power BI, 2018)

A continuación, se busca el archivo a utilizar para el análisis de datos. Se puede desplazar por los directorios y unidades de discos de almacenamiento para ubicar el archivo a utilizar.

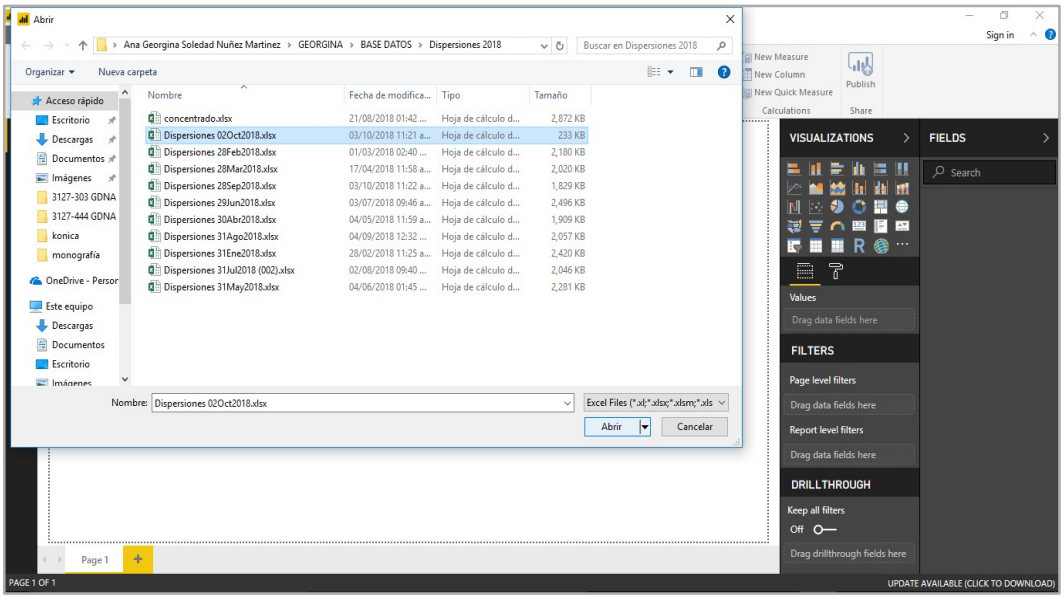


Figura 24. Búsqueda de fuente de datos
Fuente: (Power BI, 2018)

Al seleccionar un archivo, en el lado derecho se muestra una vista preliminar del archivo, tal y como se muestra en la siguiente figura.

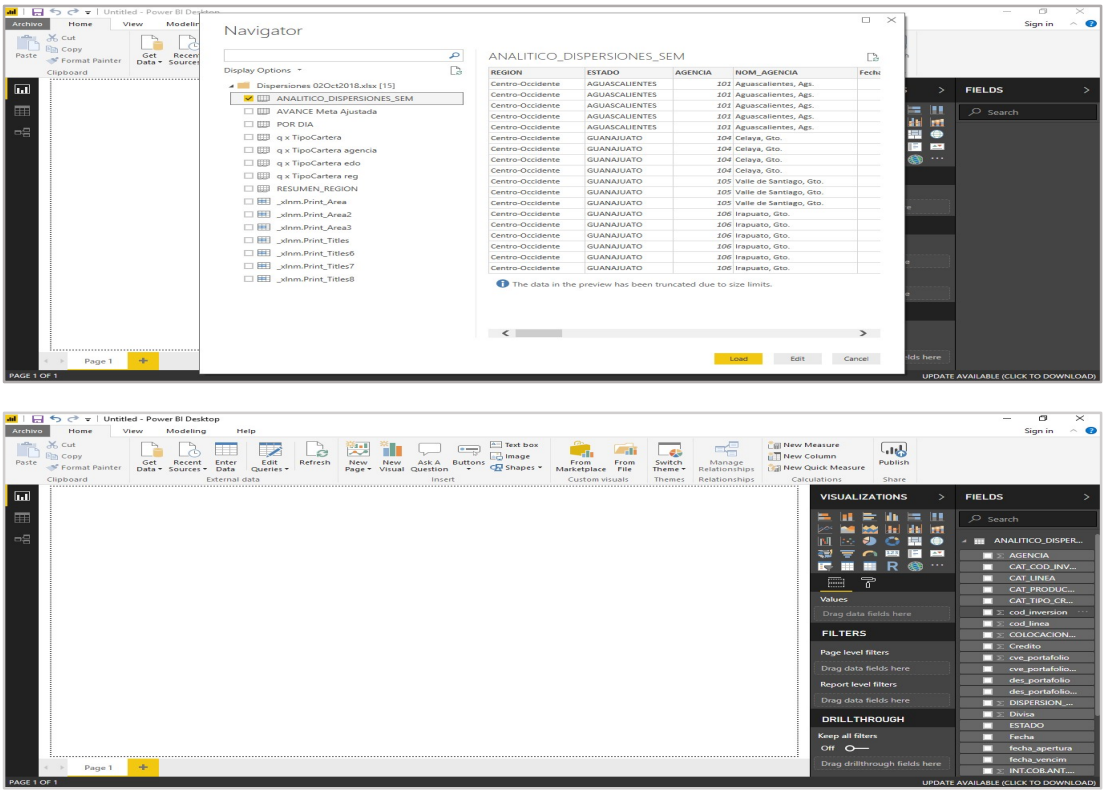


Figura 25. Vista preliminar de la fuente de datos
Fuente: (Power BI, 2018)

11.3.4. Casos exitosos en la aplicación de la herramienta Power BI

Se presentan algunos casos de empresas y usuarios que utilizan Power BI.

11.3.4.1. Policía Municipal de Nezahualcóyotl

La Dirección General de Seguridad Ciudadana Nezahualcóyotl, fue reconocida por la adopción de un programa de seguridad ciudadana de clase mundial. De todos es sabido que hay lugares como Nezahualcóyotl en donde existen mayores índices de actividad de bandas criminales.

Ahora, es reconocido por la implementación de un Modelo de Policía de Proximidad, el cual recibió premios nacionales e internacionales, gracias a las tecnologías de Microsoft aplicadas para maximizar y optimizar el uso de la información. Se recopilar en forma automatizada la información de las 1,800 cámaras del “C4” (Centro de Comando, Control, Comunicación y Cómputo) y un

helicóptero “el coyote”, se desarrolló una solución de inteligencia policial, con la capacidad analítica y predictiva lo cual conlleva a obtener inteligencia e información que facilita toma de decisiones tácticas y estratégicas.

La solución fue construida completamente en la nube de Azure. Utiliza Azure Data Factory para automatizar el flujo de datos y procesar la información, Microsoft Machine Learning para ejecutar los modelos matemáticos de predicción y Power BI para visualizar la información por medio de tableros.

Ahora se puede predecir el tipo de delito que se puede cometer en un cuadrante o incluso el modus operandi de los delincuentes. Luego por medio de tableros de Power BI se visualizan las tendencias de los delitos con distintas mezclas de datos (zona, horario, geografía, y detalles de los actos delictivos y con esto poder determinar en cierto tipo de delito cuando y a que se puede cometer. El punto es llegar a que cada jefe de sector tenga en la palma de su mano información fundamental para combatir la delincuencia mediante las aplicaciones móviles.

“*La información es un punto central de la operación de la policía. Con base en ella, se plantean operativos más eficientes, dirigidos a los delitos específicos que se están cometiendo. Nos permite una planeación estratégica a corto, mediano y largo plazo*.” (Gavela, 2018).

11.3.5. Propuesta de trabajo

Para cada una de las áreas de Financiera Nacional de Desarrollo Agropecuario, Rural, Forestal y Pesquero (FND) resulta muy importante el poder tener a la mano información que sea fácil de interpretar, donde se identifique claramente las áreas de oportunidad para mejorar el servicio que se ofrece.

Aunado a lo anterior, implementar acciones estratégicas que ayuden a revertir o modificar el cómo se hacen las cosas. Financiera Nacional de Desarrollo Agropecuario, Rural, Forestal y Pesquero (FND) es un organismo descentralizado de la Administración Pública Federal, sectorizado en la Secretaría de Hacienda y Crédito Público (Banca de Desarrollo), que impulsa el desarrollo del medio rural a través de créditos accesibles para pequeños productores y MIPYMES. (FND, s.f.).

La institución anualmente se le asigna a una meta de colocación de créditos a nivel nacional, por consiguiente, sus Coordinaciones Regionales, Agencias Estatales, Agencias de Crédito y Módulos de Atención en cada uno de los estados del País, son participes en el cumplimiento de la misma, además de ser sujetas a evaluaciones constantes sobre sus logros. Sin embargo, este análisis para el cumplimiento de la meta se hace diariamente y con proyecciones mensuales tomadas del ejercicio anterior en el mismo periodo de tiempo.

Por tal motivo es indispensable que además de las herramientas con las que cuenta para la medición o evaluación del cumplimiento de la colocación (bases de datos en Excel), se pueda tener otra alternativa que proporcione nuevas y mejores opciones de visualización y análisis de los datos que diariamente se generan.

No solo es importante medir el ¿Cuánto falta para la meta?, sino que se requiere información con la cual se pueda detectar variables donde se pueda proponer acciones o estrategias para el cumplimiento de su objetivo. La gestión de la implantación de la herramienta Power BI se genera en principio en Agencia Estatal Durango, que si bien es donde se concentra la información de todo el estado se comenzara con la que se genera en dicha oficina. En la figura 27 se muestra el organigrama de FND Agencia Estatal Durango.

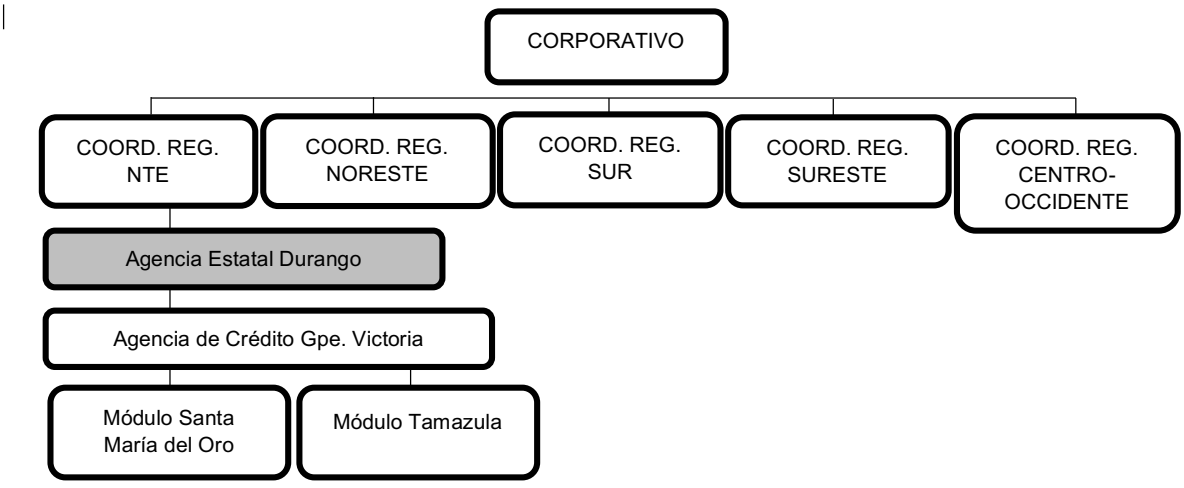


Figura 26. Organigrama FND – Agencia Estatal Durango

Fuente: Creación propia

La fuente de información que existe en la agencia es a través de los datos que arroja el sistema Terminal Financiera “TERFIN”, el cual comienza desde el registro del solicitante como cliente y como persona hasta la ministración del crédito.

Enseguida desde el área de TI quienes son los únicos con accesos para la migración de datos, emiten un comunicado donde se envía por medio del correo electrónico, la fuente de datos y reportes diariamente para su evaluación. Es aquí donde entraría la herramienta de visualización Power BI, ya que ellos envían tanto la base de datos como el reporte graficado con los indicadores más relevantes.

La propuesta consiste en generar la misma evaluación de los datos junto con sus gráficos para analizar los resultados en Excel y Power BI, comparar cual resulta más rápida y como se puede manipular la información de manera interactiva fácilmente, además de presentar los gráficos visualmente más ejecutivos.

11.3.6. Comparativo de herramientas de visualización de datos Tableau & Power BI

De acuerdo a la investigación realizada de ambas herramientas es difícil precisar cuál opción es mejor que otra, en realidad ninguna es mejor que la otra pues la diferencia radica en la utilidad que se le va a dar, el tipo de datos que se va a manejar incluso el análisis que se va a desarrollar. En base a eso se puede elegir la que mejor se adapte a las necesidades del usuario, sin embargo, no es que la otra herramienta no cumpla con las necesidades, sino que una se adapte mejor que la otra.

A simple vista parecieran similares, pero existen características las cuales las diferencian, en base a eso se puede determinar que Power BI se adapta mejor a la propuesta de trabajo ya que se integra perfectamente con bases de datos en Excel, y estas son las fuentes de datos que se generar en el área de trabajo.

Si lo que se está buscando es utilizar una herramienta de visualización la cual permita obtener información fácil de entender y poder mostrar, que además sea

a un costo razonable y que su fuente de datos principal sea el uso de bases de Excel, la opción es Power BI, por el contrario, si se requiere de una herramienta especializada en el análisis de datos más complejos que brinde opciones de gráficas más elaboradas, en donde no importe que su licencia de uso sea hasta 7 veces más cara la opción es Tableau por las siguiente razones:

- Mayor tiempo en el mercado, Tableau cuenta con 14 años de experiencia, lo cual la coloca en las favoritas por los usuarios.
- Amplia gama de fuentes de datos.
- Facilidad para crear DashBoards visuales, esto debido a que agrupa datos de distintas fuentes, carga descripciones emergentes con datos adicionales, excluye los valores que son atípicos y llega a visualizar los datos en distintas dimensiones geográficas al mismo tiempo.
- Gráficos fuera de lo tradicional, sobre todo para el uso de datos complejos que puede llegar a manejar.
- Mayor número de visualizaciones incluso que el mismo Power BI.
- Versatilidad cuando se refiere a las fuentes de datos que pueden ser procesados por la herramienta.
- Su valor comercial es mucho más caro y es necesario tener un poco más de conocimientos informáticos e incluso de análisis de datos.

A continuación, se presenta un análisis FODA de la herramienta Tableau.

Figura 27. Análisis FODA Tableau
Fuente: Elaboración propia

Se sugiere PowerBi por las siguientes razones:

- Fácil integración con otras herramientas de Microsoft: como lo son Excel, Azure, SQL Server y Cortana.
- Desde la aplicación ofrece una herramienta Training Videos, donde nos envía a tutoriales en su canal de YouTube.
- Versatilidad cuando se refiere a las fuentes de datos que pueden ser procesados por la herramienta más no se limita al ambiente Microsoft.
- Está diseñada para usuarios que no son expertos en análisis de datos.
- Interfaz amigable, por lo que la hace sencilla de manejar.
- Costo de la licencia profesional es de aproximadamente $191.00 por usuario y mes, lo cual es una gran ventaja por sobre otras opciones.

Enseguida, se presenta el análisis FODA de la herramienta Power BI.

Figura 28. Análisis FODA Power BI

Fuente: Elaboración propia

11.3.7. Análisis de la propuesta

Como profesionales de las TIC, se construye una pieza importante en el tratamiento automatizado de la información, de tal manera que debe existir el compromiso para promover y gestionar el uso o actualización de las herramientas que permitan brindar información en tiempo y forma.

Power BI es una excelente opción en lo que al análisis de datos se refiere. Por medio de la implementación de esta herramienta en el área de trabajo, se podrá obtener rápidamente un informe con las características que el usuario solicite, tan simple o tan sencillo como se requiera.

Para muchos, es tanta la familiaridad con herramientas como Excel que, muy probablemente exista cierta resistencia al cambio, pero es importante mencionar que con el ritmo de trabajo y la facilidad con la que se generan datos, se debe optar

por cambiar las herramientas y los procedimientos para que los datos puedan fluir tan rápido como son generados y poderlos transformar en información.

Vale la pena gestionar el uso de cualquiera de las herramientas Tableau Desktop o Power BI Desktop, a partir de que son de uso gratuito y que su interfaz realmente es muy amigable y fácil de usar para cualquier persona que maneje datos, pero sobre todo la inmensa utilidad que se le puede obtener de ambas herramientas.

Tableau y Power BI, sin duda dos de las herramientas más usadas en el mundo de la visualización de datos, cada una tendrá sus defensores, pero también sus detractores dependiendo cual se ajuste a las necesidades de información y visualización que requieran. Ambas pueden generar reportes con información relevante para la toma de decisiones, que finalmente es lo que el usuario requiere, sin embargo, Tableau se enfoca más a datos especializados, trabaja con fuentes de datos con mayor complejidad, como se ha descrito los usuarios que la usan deben tener un nivel de conocimiento más alto, lo cual no es limitante, pues su interfaz ayuda a aprender su modo de uso.

Power BI por su parte, es la opción que se adecua a cualquier usuario promedio, pues generalmente las fuentes de datos con las cuales se trabaja en las empresas son del tipo .xls. Como se ha mencionado es el tipo de gráficos que se encuentra en una y otra herramienta lo que llamaría la atención, pero finalmente es más el uso que se le va a dar, lo que pudiera ser la razón entre cual elegir e implementar. Tableau puede estar en ventaja en este sentido, pero si está en función la complejidad de la herramienta se elegiría Power BI.

Mediante la implementación de Power BI como herramienta de visualización de datos, se puede incentivar el uso de la misma por los encargados de generar reportes en la Agencia Estatal Durango. A través del tiempo observando los beneficios que conlleva, se puede sugerir la implementación a un nivel superior.

Con la herramienta, se pretende tener acceso a información más precisa y sencilla de comprender con la cual se puede apoyar al Agente Estatal de Crédito en

la toma de decisiones, sobre todo en implementar estrategias si así se requirieran para cumplir con la meta proyectada.

Se podrá identificar las áreas de oportunidad o de atención prioritaria donde nos indiquen los reportes obtenidos. Realizar cada vez mejor los reportes que son emitidos en la Agencia, los cuales brinden información relevante. Minimizar el tiempo de espera para generar un reporte, ya que en ocasiones la complejidad con el uso de grandes volúmenes de datos, los equipos tienden a saturarse.

Con la característica de ser un software de interfaz amigable, con una capacitación al personal, pudieran estar generando sus propios reportes, esto permitiría un análisis de sus récords personales. Por ser una herramienta ligera, no representa ningún inconveniente en el buen funcionamiento de los equipos de cómputo.

11.3.8. Impacto y beneficios de aplicar

Para la Agencia Estatal Durango, es de suma importancia tener de manera inmediata la información que nos indique diariamente, los niveles de cumplimiento de meta. Una buena medida de control usando indicadores interactivos permitirá estar informado sobre qué Agencia de Crédito o Módulo de atención e incluso que Ejecutivo de Financiamiento no está cumpliendo con su meta. Incluso se puede conocer qué tipo de financiamiento es el que mayor demanda tiene y por consiguiente que región es la más beneficiada con estos programas.

Claro está que con este tipo de herramientas es posible obtener información tan precisa y ser presentada de manera visualmente más interesante. Poder generar un histórico de colocación y evaluar en que periodos los créditos son más buscados aunado a esto, crear estrategias para los meses que con los datos históricos se conoce que no hay tanta demanda.

La información que puede resultar de unas tablas de Excel, que a simple vista no se encuentra relación entre sí, no tiene límites y el poder contar con estas herramientas que facilitan estos procesos y que son gratuitas, no hay lugar a pensar el por qué no usarlas.

Conclusiones

Quizás en el Estado de Durango existan numerosas empresas que no se han podido beneficiar con soluciones de Inteligencia de Negocios que brinda el basto mundo de las herramientas de visualización de datos tales como Tableau Desktop y Power BI Desktop.

En principio una razón es el poco interés en el uso de tecnologías e innovaciones, además del desconocimiento que existe de este tipo de herramientas, cuando en otros estados y en el mundo, están siendo parte de la toma de decisiones para que los negocios crezcan, sin distinción de giro al que se dediquen.

Gracias a este estudio, se pudo investigar de una manera generalizada a que le llamamos Inteligencia de Negocios y cuál es la relación con las herramientas de visualización de datos, esto conlleva a elegir dos de ellas para realizar un comparativo con respecto a sus características, opciones y facilidad de uso.

Ahora bien, tanto Tableau Desktop como Power BI Desktop poseen características que las hacen idóneas para aplicarlas en cualquier lugar de trabajo donde se manejen bases de datos, incluso sin necesidad de comprar una licencia pues de las grandes ventajas con las cuales cuentan son sus versiones gratuitas.

Ambas herramientas trabajan con fuentes de datos muy similares, cualquiera de las dos nos va a permitir generar reportes de una manera rápida, sencilla y fácil de interpretar.

Una ventaja importante de la implementación de estas herramientas es que no será necesario esperar que el área especializada en el tratamiento de las bases de datos genere los reportes especiales o necesarios para las áreas de trabajo, sino que cualquier usuario al que se capacite en el uso de la herramienta, será capaz de generar sus propios reportes para conocer un resultado, una tendencia, una moda, entre otros. Esto gracias a la facilidad de uso y su interfaz amigable, lo cual las hace atractivas, ya que prácticamente cualquier persona con conocimientos básicos de computación puede realizarlos.

Sin duda existe la justificación suficiente para implementar cualquier herramienta de visualización de datos en el área de trabajo, que en este caso es Power BI, visualmente parece más sencilla de aprender, existe mucho material audiovisual de tipo tutoriales para el aprendizaje, su instalación es muy sencilla e incluso al momento del registro para su descarga contactan al usuario para ofrecerle asistencia técnica mediante videoconferencias en Webmex, lo cual representa un plus a la hora de comparar y elegir la mejor opción. A final de cuentas quien determine cuál es la mejor opción, será el usuario final el cual probablemente lo determinará a partir de cual le resulte más práctica y fácil de usar.

Bibliografía

Aguilar, L. J. (2016). Big Data, Análisis de grandes volúmenes de datos en organizaciones. Alfaomega Grupo Editor.

Caralt, J. C. (2010). Introducción al Business Inteligence. Barcelona: UOC.

Carto (2018). Unlock the power of spatial analysis. Carto. Recuperado de: https://carto.com/

Cisneros, A. R. (s. f.). GNP centraliza sus datos y gana agilidad en la toma de decisiones. Soluciones Tableau. Recuperado de: https://www.tableau.com/es-es/solutions/customer/GNP-centraliza-sus-datos-y-gana-agilidad-en-la-toma-de-decisiones

Datawrapper. (2018). Why DataWrapper. DataWrapper Solutions. Recuperado de: https://www.datawrapper.de/why-datawrapper/

FND. (2018.). Acciones y programas. Financiera Nacional de Desarrollo Agropecuario, Rural, Forestal y Pesquero. Recuperado de https://www.gob.mx/fnd

Gavela, R. (2018). Combatiendo el crimen en México con solución de AI y modelos predictivos. Historias de clientes, Power BI, Microsoft. Recuperado de: https://customers.microsoft.com/es-es/story/seguridadneza-azure-powerbi-sql-es-mexico

Gómez, A. A. R., y Bautista, D. W. R. (2010). Inteligencia de negocios: Estado del arte. Scientia et technica, 1(44), 321-326.

Macías, R. G. (2016). *Business intelligence y logística empresarial*. Obtenido de https://www.gestiopolis.com/business-intelligence-logistica-empresarial/

Martínez, C. G. (s.f). Nacional Financiera alcanza sus objetivos financieros con el análisis de datos de autoservicio. Soluciones Tableau. Recuperado de: https://www.tableau.com/es-es/solutions/customer/nacional-financeria-hits-financial-goals-self-service-analytics

Plotly (2018). Plotly Dash App Gallery. Plotly. Recuperado de https://dash-gallery.plotly.host/Portal/

Power BI. (2018). Características Power BI. Power BI, Microsoft. Recuperado de: https://powerbi.microsoft.com/es-es/features/

Qlik. (2018). Business Intelligence de Qlik: Analítica e integración de datos. Qlik. Recuperado de: https://www.qlik.com/es-es

Rodríguez, C. I. (2017). Tendencias en Business Intelligence del Big Data al Social Intelligence. *Revista Tecnológica*, no. 10.

Rud, O. P. (2000). *Data Mining Cookbook Modeling Data for Marketing, Risk, and Customer Relationship*. John Wiley & Sons, 2001

Ruiz, R. Á. (2018). Minería de datos en redes sociales para pymes. *Universidad de Jaén*. Recuperado de https://hdl.handle.net/10953.1/7836

Sinnexus (2018). ¿Qué es Business Intelligence?. Sinnexus. Recuperado de https://www.sinnexus.com/business_intelligence/index.aspx

Tabares, L. F. y Hernández, J. F. (2014). Big Data Analytics: Oportunidades, Retos y Tendencias. Universidad de San Buenaventura, 20.

Tableau. (2018a). Tableau Galería Visual. Tableau. *Recuperado de https://www.tableau.com/solutions/gallery*

Tableau. (2018b). Tableau Resources. Tableau. *Recuperado de* https://public.tableau.com/es-es/s/resources

Tascón, M. (2013). Introducción: Big data. Pasado, presente y futuro. Telos: Cuadernos de comunicación e innovación, (95), 47-50.

Turner, V., Gantz, J. F., Reinsel, D., y Minton, S. (2014). The digital universe of opportunities: Rich data and the increasing value of the internet of things. IDC Analyze the Future, 16

Wong, V. (2018). *Compare 6 Types and 14 Data Visualitations Tools*. Recuperado de: https://it.toolbox.com/blogs/vincentwong/compare-6-types-and14-data-visualization-tools-091618

Capítulo 12

Herramientas para análisis y visualización de datos: Tableau y R

Nahibe Susana Orrante Vázquez
Tecnológico Nacional de México. Instituto Tecnológico de Durango
02040921@itdurango.edu.mx

Jesús Raymundo Rodríguez Díaz
Tecnológico Nacional de México. Instituto Tecnológico de Durango
03041057@itdurango.edu.mx

José Gabriel Rodríguez Rivas
Tecnológico Nacional de México. Instituto Tecnológico de Durango
gabriel.rodriguez@itdurango.edu.mx

12.1. Introducción

Se dice que una imagen vale más que mil palabras, los seres humanos han sido capaces de inventar diferentes herramientas de visualización a partir de la idea de que una imagen puede transmitir una mayor cantidad de información que un párrafo o una tabla.

“La visualización consiste en transformar información en imágenes que faciliten la extracción de significado” (Alcalde Perea, 2015), el objetivo es permitir

identificar patrones y tendencias que serían invisibles si esos mismos datos se presentaran de forma convencional”

Una imagen transmite una cantidad mayor de información que un complejo párrafo. Es por ello que existen herramientas, técnicas y disciplinas de conocimiento que hoy se integran en la “visualización de datos”, la cual consiste en transformar información en imágenes, a veces la información es cuantitativa y la visualización creada, muchas veces interactivas, puede ser llamada “de datos”; en ese caso, el objetivo del gráfico es permitir identificar patrones y tendencias que serían invisibles si esos mismos datos se presentaran en una tabla numérica.

Por otra parte, el análisis de datos es aplicado en las organizaciones y compañías para identificar información valiosa de una manera analítica y confiable, contribuyendo en la toma de decisiones en las empresas de una forma veraz, de igual forma cumple con la función de determinar el cumplimiento de metas.

Un primer caso que se menciona en este capítulo es el relacionado al Sector Salud. En este sentido, el análisis de la información es un factor determinante que facilita conocer cuáles son los problemas de salud pública de manera concreta, propositiva, oportuna y participativa para beneficio de los sectores sociales.

Para mejorar los resultados del análisis de información que se utilizan actualmente en los Servicios de Salud de Durango, se tomarán en cuenta las herramientas Tableau y R, los cuales tienen un enfoque de análisis estadístico, se pretende comparar y conocer sus ventajas y desventajas, con ello, determinar la factibilidad de la implementación para el mejoramiento del análisis de los datos.

En el Área de Vigilancia Epidemiológica se concentra la información de los casos nuevos de enfermedades, la cual es recopilada directamente por personal con preparación en el área médica, por lo cual se busca determinar cuál es la herramienta viable para que se trabaje en conjunto con el Área de Informática y así presentar el análisis estadístico con una interfaz gráfica (dimensiones, colores, etiquetas, gráficas, tablas).

Este capítulo en su primer contexto busca analizar y comparar las herramientas Tableau y R, para el análisis y visualización de la información que se utiliza en el Área de Vigilancia Epidemiológica, ya que no se cuenta con un sistema especializado para esa función.

Como justificación importante, el análisis de datos facilita que la organización mejore la información que tiene disponible para ayudar a tomar decisiones. Convertir los datos de la aplicación en representaciones visuales ayuda a los usuarios a describir conceptos, descubrir oportunidades, explorar opciones y llegar a tomar decisiones más óptimas, todo llevado a cabo por un medio muy persuasivo.

La importancia de este documento en su primer aspecto, busca ofrecer una propuesta para el uso de alguna herramienta para análisis y visualización de datos en el Área de Vigilancia Epidemiológica. Se busca que esté bien sustentada, dirigida y enfocada a mejorar el proceso de toma de decisiones en el área de trabajo citada.

Se presenta una primera propuesta al departamento de Vigilancia Epidemiológica, “la aplicación de la herramienta de análisis de información con base en la investigación documental realizada, buscando cubrir las necesidades del área.

Un segundo contexto aborda el caso relacionado con la Residencia General de Conservación de Carreteras del Centro de la Secretaría de Comunicaciones y Transporte (SCT) Durango.

En esta dependencia, se generan informes periódicamente que incluye presentaciones, gráficas y fichas técnicas a las diferentes áreas para dar a conocer los avances físicos y financieros de los trabajos que se ejecutan en el ejercicio fiscal vigente a los directivos de la institución, tales como el Residente General de Conservación de Carreteras, la Subdirección de Obras, la Dirección General del propio Centro SCT, así como a la Dirección General de Conservación de Carreteras.

Por tal motivo como segunda propuesta se pretende promover el uso de la herramienta de visualización de datos Tableau para darle un giro de 360 grados a la manera en que se ha presentado hasta hoy esta información a las distintas áreas.

Los objetivos principales de las propuestas son:

- Identificar, analizar y determinar una herramienta de análisis y visualización de datos, para la implementación de una propuesta de su uso en el Área de Vigilancia Epidemiológica.
- Realizar una propuesta para la implementación de Tableau como herramienta de visualización de datos que permita a los diferentes niveles de mando de la Residencia General de Conservación de Carreteras, Subdirección de Obras y Dirección General del Centro SCT Durango.

De manera específica se busca lo siguiente:

- Identificar aspectos de la Ciencia de los Datos, Bussines Intelligence (BI), herramientas para visualización y análisis de datos.
- Comparar y evaluar las herramientas en el entorno (BI): Tableau y R.
- Identificar los procesos de trabajo del Área de Vigilancia Epidemiológica.
- Identificar las características que se adecúan a las necesidades del Área de Vigilancia Epidemiológica y que contemple las nuevas tendencias de análisis de información.
- Identificar una propuesta del uso de Tableau o R en el Área de Vigilancia
- Presentar características y funciones de la herramienta Tableau como alternativa para la visualización de resultados.
- Presentar propuesta para presentar reportes con toda la información necesaria mediante la herramienta Tableau para visualizar de manera clara los resultados esperados para la toma de decisiones en la Residencia General de Conservación de Carreteras, Subdirección de Obras y Dirección General del Centro SCT Durango.

En el campo de la Ciencia de los Datos, existe una amplia gama de herramientas para visualización de los datos, por lo que resulta necesario el

análisis de dichas soluciones para optar por la que mejor se ajuste a las necesidades de la organización y al perfil de los usuarios.

En este sentido, en este documento se hace especial énfasis en las características y ventajas de las herramientas Tableau y R para el análisis y visualización de los datos.

Se pretende presentar dos propuestas basadas en aspectos de análisis y visualización de datos, con lo cual se buscar transformar datos en información, la información en conocimiento y el conocimiento en acción.

Así mismo se intenta con las propuestas, disminuir el tiempo de análisis, realizando cruces de datos para beneficio del área.

El impacto esperado en la Residencia General al implementar la herramienta Tableau, es la presentación de la información de forma visual y amigable facilitando la comprensión de la misma a los diferentes niveles de mando y así poder extraer el mayor conocimiento posible.

En consecuencia, el mayor beneficio que se tendrá es el poder identificar las obras con problemática a partir de los avances físicos y financieros que periódicamente se reportan a la Residencia General.

12.2. Marco de referencia

En los siguientes puntos, se identifican conceptos teóricos relacionados con el título de este capítulo: Ciencia de los Datos, Bussines Intelligence (BI), Tableau, lenguaje de programación R. De igual forma se abordan aspectos contextuales del sector salud y del contexto de la Residencia General de Conservación de Carreteras, Subdirección de Obras y Dirección General del Centro SCT Durango.

En el punto 12.2.1., se identifica un conjunto de términos necesarios para una adecuada comprensión de este capítulo, más adelante, en párrafos siguientes, se describen más ampliamente.

12.2.1. Conceptos

- **Datos:** Un dato es valor numérico discreto o continuo que representa algo de la realidad. En el ámbito empresarial un dato es una transacción de compra, venta, transferencia, depósito, retiro, entre otras cosas. Un dato por sí sólo no indica nada si no se le asocia dentro de un contexto para tenga significado y propósito. (Carrión, Diferencia entre Dato, Información y Conocimiento, s.f.)
- **Información:** Es aquello absolutamente esencial para comunicar algo de la forma más clara y objetiva posible, es un conjunto organizado de datos capaces de cambiar el estado de conocimiento en el sentido de las consignas transmitidas. La información tiene una estructura interna y puede ser calificada según varias características. (Carrión, Diferencia entre Dato, Información y Conocimiento, s.f.)
- **Conocimiento:** El conocimiento es algo más complexo, más grande, más profundo y más enriquecedor que los datos y la información. El conocimiento es una combinación de experiencias, valores, información y saber hacer que se utiliza como fuente la nuevas experiencias e información, y es trascendente para la toma de decisiones. Su origen se aplica en la mente de las personas. (Carrión, Diferencia entre Dato, Información y Conocimiento, s.f.)
- **Ciencia de datos:** Una acepción acerca de este concepto, es el arte de combinar y utilizar varias herramientas, principios de aprendizaje automático y algoritmos cuyo propósito es descubrir patrones ocultos y conocimiento a partir de datos. (Jones, 2019).
- **Big Data:** Se describe como un conjunto de datos que tienen un tamaño que supera la función normal de las herramientas de software de base de datos, como el almacenamiento, la captura, el procesamiento y el análisis. Se refiere a una colección de conjuntos de datos que son muy grandes y complejos, además de estructurados trata con datos no estructurados de manera que no se puede procesar utilizando herramientas de administración de bases de datos tradicionales. (Jones, 2019)
- **Visualización de datos:** Consiste en la aplicación de técnicas para seleccionar, procesar y poner a disposición de una audiencia una cantidad de datos,

dándoles significado para que, mediante su visualización, se conozcan sus relaciones de causa o dependencia, con el objetivo de señalar, denunciar o establecer conocimiento sobre un procedimiento, un fenómeno, una acción (Sedeño Valdellós, La visualización de datos como recurso social: posibilidades educativas y de activismo, 2016).

- **Herramienta de visualización:** Instrumentos para convertir la información en conocimiento, que combina técnicas provenientes de la estadística o la programación con el diseño artístico. (Sedeño Valdellós, La visualización de datos como recurso social: posibilidades educativas y de activismo, 2016)
- **Ofimática:** Son aplicaciones o programas que puede ser utilizados en tareas correspondientes a las oficinas, trabajos académicos y similares. Dichas herramientas permiten crear, cambiar, organizar, imprimir y transferir documentos de todo tipo, las aplicaciones ofimáticas pueden adquirirse por separado o en un conjunto de aplicaciones, llamado suite ofimática. Un ejemplo de suite ofimática es el la suite Micrososft Office que cuenta con múltiples aplicaciones tales como Excel, Word, Power Point, entre otros (Olivo Cabezas, 2016).
- **Tableau:** Es una herramienta vista como un modelo para el análisis y visualización de datos. (Tableau Softwate, LLC, s.f.)

12.2.2. Ciencia de los Datos

La ciencia de datos es un conjunto de principios fundamentales que apoyan y guían la extracción de información basada en principios y el conocimiento de los datos. Posiblemente el más cercano concepto relacionado con la ciencia de datos es la minería de datos, la real extracción de conocimiento a partir de datos a través de tecnologías que incorporan estos principios.

Una gran parte de lo que tradicionalmente se ha estudiado en el campo de las estadísticas es fundamental para la Ciencia de los Datos. Los métodos y metodología para visualizar datos son vitales. La ciencia de datos es compatible con

la toma de decisiones impulsada por datos, y en ocasiones permite tomar decisiones de forma automática a escala masiva, y depende de las tecnologías para el almacenamiento e ingeniería de Big Data.

El estadístico John Tukey definió el *análisis de datos* en 1961 en su obra "*The future of data análisis*" de la siguiente manera: procedimientos para analizar datos, técnicas para interpretar los resultados de dichos procedimientos, formas de planear la recolecta de datos para hacer el análisis más fácil, más preciso o más exacto (Martínez, s.f.)

El proceso de análisis de los datos se divide en tres partes principales: captura, procesamiento y consulta. La captura de los datos puede proceder de diferentes fuentes, el procesamiento consiste en acumular y manipular elementos de los datos capturados para poder producir información analizable, para la consulta necesariamente debe tener un formato concreto, para visualizarlo o publicarlo.

En cuanto a las herramientas para funciones estadísticas disponibles en el proceso de análisis y visualización de datos se encuentran algunas herramientas y aplicaciones tales como: Minitab, SPSS y Watson Studio de la marca IBM, PowerBi, KNime Analytics Platform, Tableau, entre otras.

Específicamente, Tableau es una herramienta de visualización de datos utilizada en el área de la inteligencia de negocios.

En la clasificación de herramientas no comerciales se encuentra, opciones como RapidMiner, Python y por supuesto R, estos dos últimos también etiquetados como Lenguajes de programación.

12.2.3. Bussines Intelligence (BI)

La inteligencia empresarial o Inteligencia de negocios (BI) es la infraestructura y conjunto de procedimientos que almacena, procesa y analiza los datos de la empresa después de recopilarlos. Inteligencia empresarial es un término amplio que cubre aspectos como el análisis de procesos, extracción de datos, análisis descriptivo y evaluación comparativa de indicadores del negocio. Está destinada a extraer todos los datos que genera una empresa y presentar

medidas de rendimiento fáciles de asimilar, y también tendencias que pueden influir en la toma de decisiones (Jones, 2019).

Los Software de BI y otras tecnologías para capturar, almacenar, analizar y generar información o conocimiento contribuyen para alcanzar el punto óptimo de la decisión y tiene como principales características:

- El reconocimiento de la experiencia
- El análisis de los datos contextualizados
- La capacidad de extraer e integrar datos de múltiples fuentes
- El procesamiento de los registros obtenidos en información útil para el conocimiento de la organización
- La búsqueda de relaciones de causa y efecto, trabajando con hipótesis y desarrollando estrategias y acciones competitivas (Puerta Gálvez, 2016).

Algunas de las herramientas líderes en el mercado según las estadísticas mostradas en el cuadro mágico de Gartner de febrero de 2018 son: Tableau, Microsoft BI y Qlik. Se visualiza la figura 1.

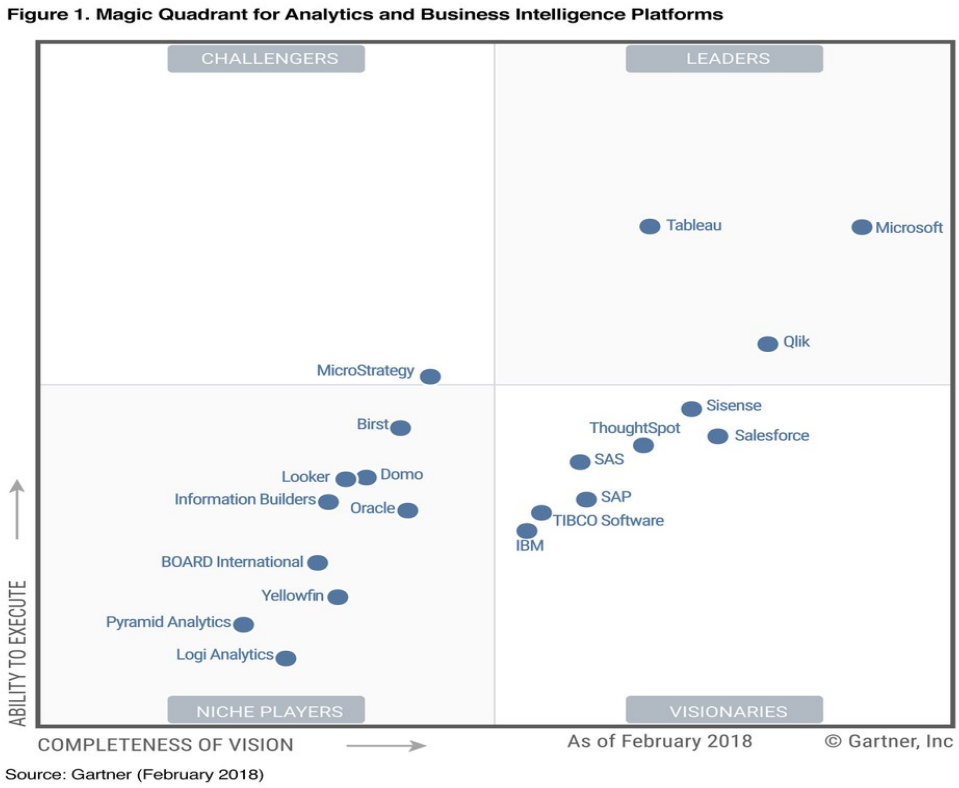


Figura 1. Cuadro mágico de Gartner. (De Juana, 2019)

12.2.4. Tableau

Tableau es una herramienta de visualización de datos, es decir, el usuario tiene la posibilidad de interactuar con los datos: comparar, filtrar, conectar unas variables con otras; además, la plataforma y los paneles que se pueden crear con la herramienta son muy visuales; facilita la comprensión rápida de los datos; tiene algunas ventajas interesantes con bases de datos; acepta formatos con Excel, Access y texto; se puede acceder a muchas bases de datos comunes como Microsoft SQL Server, MySQL, Oracle, entre otros (Flores Avendaño & Villacís Vera, 2017).

Tableau Desktop es una aplicación de software que permite a cualquier persona analizar cualquier tipo de datos de forma rápida y fácil. Con Tableau, se trabaja directamente con los datos, cambiando entre vistas fácilmente y sin soporte de Tecnologías de la Información.

Tableau presenta una función de arrastrar y soltar para crear rápidamente visualizaciones desde datos formados. Por lo tanto, utilizando datos recolectados previamente y arrastrándolos y soltándolos en el software Tableau ahorra tiempo. Además, las visualizaciones de Tableau se cambian fácilmente y analizado o especificado de acuerdo con el usuario.

En lugar de tener que ingresar manualmente la información, el usuario simplemente selecciona la parte de la visualización que necesita más exploración y puede crear un panel para analizar los datos.

Algunas características que presenta Tableau son las siguientes:

- Capacidad de modelado y análisis de datos con acceso a una variedad de fuentes de datos.
- Interfaz de usuario fácil de usar, intuitiva y con alta aceptación por parte de los usuarios.
- Análisis visual con inteligencia integrada para usuarios empresariales.
- No requiere que los usuarios posean amplios conocimientos de programación.
- Permite la publicación de informes en la web corporativa o cualquier tipo de sistema de la empresa que funcione sobre arquitectura web.
- Facilita la integración analítica de información procedente de diversas fuentes. (Ayala, Ortiz, & Maya, 2018).

Algunas de sus desventajas de Tableau son:

- La integración y modelización de datos tiene lugar en el cliente de escritorio.
- Falta de soporte de navegación guiada en paneles y aplicaciones de BI.
- Se utiliza principalmente en despliegues de un solo usuario y departamentales, causando un riesgo de “explosión” de informe cuando se utiliza Tableau como una solución de BI empresarial.

Tableau es una herramienta de visualización capaz de entregar visualizaciones interactivas en poco tiempo por su naturaleza de arrastrar y soltar, aumenta la eficacia mediante el poder de los datos, ofrece algo único combinando

un enfoque totalmente centrado en cómo las personas ven y comprenden los datos con una plataforma robusta y escalable, lo cual le ha permitido volverse efectiva incluso para las organizaciones más grandes del mundo.

Tableau puede usarse para definir y calcular nuevas variables y realizar manipulaciones de datos simples con el uso de fórmulas matemáticas como Excel, pero es mucho más potente y puede manejar millones de registros donde Excel falla. Además, cuenta con la capacidad para integrarse con varias plataformas, incluidas la Big Data Hadoop, la herramienta de análisis estadístico R y su soporte para Google BigQuery API.

Para acceder a la información almacenada en muchas bases de datos, los usuarios requieren conectores y cierta experiencia, la cual se puede adquirir mediante la práctica y el uso constante de esta herramienta de visualización de datos.

En la figura 2, se muestra un ejemplo de visualización de datos con Tableau, extraído de Tableau Softwate, LLC, (s.f.).

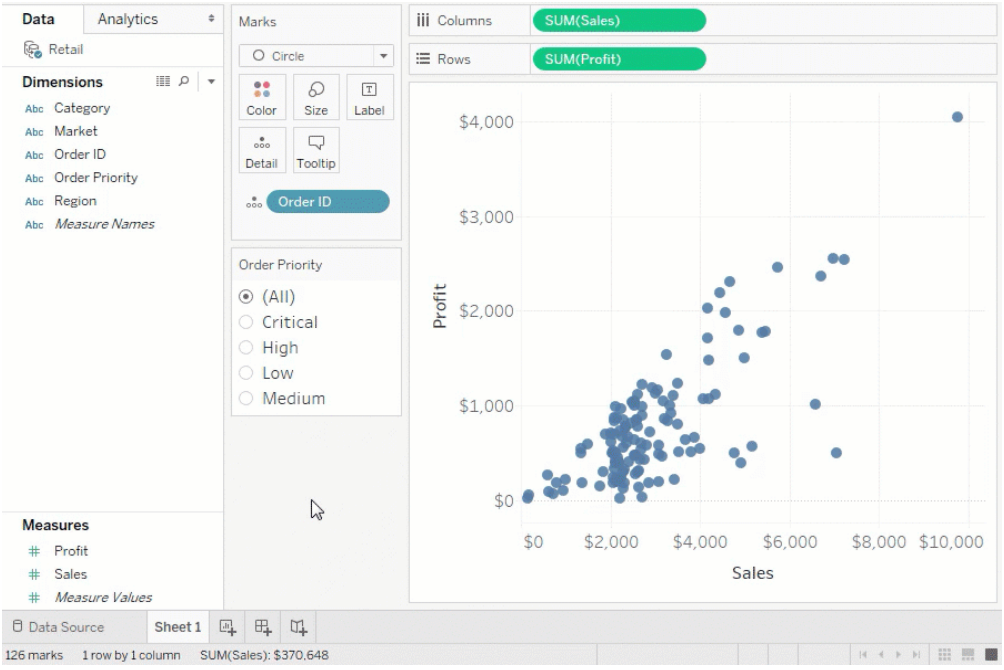


Figura 2. Visualización de datos en Tableau (Tableau Softwate, LLC, s.f.).

12.2.5. Lenguaje de programación R

R es un lenguaje de programación y un ambiente para análisis estadístico y la visualización de gráficos. Debido a su naturaleza se puede utilizar en una gran cantidad de tareas.

R abarca una amplia gama de técnicas estadísticas que van desde los modelos lineales a las más modernas técnicas de clasificación pasando por los test clásicos y análisis de series temporales.

Cuenta con la licencia GNU GPL o software libre, es decir, se puede instalar sin licencia o de uso comercial, respeta la libertad de los usuarios, paquetes descargables elaborados por una comunidad de programadores y software libre publicado por otras personas.

Algunas características de R son las que se citan a continuación:

- Posee una instalación eficaz de manejo y almacenamiento de datos.
- Contiene un conjunto de operadores para cálculos en matrices.
- Una colección grande, coherente e integrada de herramientas (paquetes) intermedias para el análisis de datos.
- Facilidades gráficas para el análisis y visualización de datos en pantalla o en papel.
- Un lenguaje de programación bien desarrollado, simple y efectivo que incluye condicionales, bucles, funciones recursivas definidas por el usuario e instalaciones de entrada y salida (r-project.org, 2019).

12.2.6. Los Servicios de Salud de Durango (SSD)

En este apartado, se describe las generalidades de la unidad de trabajo Servicios de Salud Durago (SSD) ya que es la unidad de trabajo en donde se pretende impulsar la propuesta de uso de herramientas para análisis y visualización de datos.

Los Servicios de Salud de Durango (SSD) son los encargados de establecer y conducir la política nacional en materia de asistencia social, servicios médicos y salubridad general, así como organizar y vigilar las instituciones de beneficencia

privada, en los términos de las leyes relativas, respetando la voluntad de los fundadores.

Así mismo, los SSD propician y coordinan la participación de los sectores social y privado en el sistema nacional de salud y determinan las políticas y acciones de inducción y concertación correspondientes, además de poner en práctica las medidas tendientes a conservar la salud y la vida de los trabajadores del campo y de la ciudad y la higiene industrial, con excepción de lo que se relaciona con la previsión social den el trabajo.

12.2.6.1. Antecedentes de los Servicios de Salud de Durango

En 1927 se iniciaron los trabajos de la primera delegación federal de salubridad instalada en el estado de Durango por la actual Secretaria de Salud.

Durante las primeras décadas del siglo XX aún sin especificación de documentos formales, la determinación de responsabilidades tenía cierta definición, correspondiendo a la federación el control de enfermedades transmisibles, la vigilancia, promoción del saneamiento ambiental, el control de los alimentos, bebidas y medicamentos, la educación para la salud; así como las acciones solo realizables mediante una organización nacional, tales como la vigilancia de los puertos y fronteras para efectos de control de enfermedades.

La responsabilidad de la organización de los servicios asistenciales se estaba a cargo de gobiernos estatales o municipales, incluyendo consultorios y hospitales para la atención médica.

El 21 de enero de 1985 se modificó la denominación de la dependencia a “Secretaría de Salud” conservando las siglas y el logotipo establecidos con anterioridad. En el año de 1998 se concluyó con 145 centros de salud; para 1999 se amplió a 159 unidades de primer nivel y para el año 2000 la infraestructura aumentó a 175; es decir, 30 Centros de Salud más en comparación a 1998.

Para el año 2000 es de relevancia señalar que en base a la sólida infraestructura lograda se llegó a la *certificación de cobertura universal del os servicios de salud en el estado de Durango*, por un organismo no gubernamental,

de carácter internacional como es la Organización Panamericana de la Salud (OPS). Dicha certificación fue otorgada luego de una exhaustiva verificación en base a indicadores internacionales de calidad, aplicados en las áreas del campo de trabajo en las cuatro jurisdicciones sanitarias (Durango, Gómez Palacio, Santiago Papasquiaro y Rodeo) incluyendo las zonas indígena, de la montaña, del semidesierto, las quebradas y de los valles.

Teniendo como objetivo fundamental en los Servicios de Salud de Durango mejorar y conservar las condiciones de salud de la población a la que se atiende, y cumpliendo una de las principales acciones en lo relativo a la ampliación de la cobertura de los servicios de salud, existía en la entidad para el año 2002 una infraestructura física de 179 unidades de primer nivel con 233 consultorios; 51 centros de salud se ubican en área urbana y 128 en localidades rurales.

En febrero de ese año 2005, se incorpora Durango al Sistema de Protección Social en Salud (Seguro Popular), con la finalidad de otorgar protección financiera a las familias sin seguridad social y con esto limitar el gasto catastrófico en salud realizado por las familias de menor capacidad económica.

Se crearon ese mismo año los institutos de salud preventiva y el de salud mental, los cuales tienen como finalidad impulsar la investigación científica de la medicina preventiva y los padecimientos mentales, así como coadyuvar en la atención de calidad.

Es de relevancia mencionar que a fines del año 2006 se inició la construcción del Hospital Regional de Alta Especialidad de Durango (HRAE); así, la transición epidemiológica observada en la base de la estructura poblacional, da origen a este hospital de tercer nivel de atención médica, con el objetivo de aumentar la capacidad de resolución en materia de salud, elevando la calidad de atención de los Servicios de Salud de Durango.

Durante el año 2007, se construyeron y se pusieron en operación el Hospital Integral de la Comunidad de Mapimí, de ocho camas; el Centro de Salud de Arnulfo

R. Gómez y la Unidad de Investigación Oncológica del Centro Estatal de Cancerología en la capital del Estado.

Para 2008, se entregaron varias Unidades de Especialidades Médicas (UNEME) como son: 2 Centros de Atención Primaria de Adicciones (CAPA), uno en Durango y otro en Gómez Palacio, UNEME-EC (Enfermedades Crónicas) en Durango y UNEME Salud Mental en Gómez Palacio. Asimismo, se pone en funcionamiento el Centro de Salud de Torreón de Cañas, el Hospital Integral de la Comunidad de Nuevo Ideal y el de Las Nieves, ambos con 8 camas.

A principio de 2009, se inician operaciones de las siguientes Unidades de Salud nuevas: UNEME-CAPA Sur en Durango, Dgo., UNEME-CAPA Santiago Papasquiaro, Centro de Salud Bosques del Valle en Durango, Dgo., Centro de Salud Col. La Virgen, en Durango, Dgo., Hospital Integral de Tamazula, Hospital Integral de Huazamota, Centro de Salud de Santa María de las Flores, en El mezquital, Dgo., Centro de Salud de Tierras Coloradas, en Pueblo Nuevo, Dgo. Y la UNEME-EC en Gómez Palacio, Dgo.

Durante el año 2010, se entregan varios proyectos de salud, entre los cuales están: el Laboratorio Estatal de Salud Pública, el Hospital General de Lerdo, la UNEME-CISAME en Durango, Dgo., el Centro de Salud de la Col. 5 de Mayo en Durango, Dgo., el Hospital Integral de Villa Unión, en Poanas, Dgo. Y la UNEME-EC de Santiago Papasquiaro.

Para el año 2011, se ponen en funcionamiento varias áreas del Hospital General “Victoria de Durango”, entre las cuales están: consulta externa, laboratorio e imagenología. También en este año inicia actividades el Hospital Integral de Nazas.

En el año 2012 se inaugura y entra en funciones la UNEME-EC de Bermejillo, en Mapimí, Dgo. Durante el año 2013, se pusieron en marcha el Centro de Salud con Servicios Ampliados (CESSA) de Durango, el Centro de Salud de la Col. Luz del Carmen y la UNEME-DEDICAM (Detección y Diagnóstico de Cáncer de Mama) de Gómez Palacio, Dgo.

Para el año 2014, se abrieron las siguientes Unidades Médicas: el Centro de Salud con Servicios Ampliados (CESSA) de Tepehuanes, Dgo., el día 20 de Enero; el día 1° de Abril de 2014, se ha puesto en operaciones el UNEME-EC de Lerdo, Dgo., así como el Área de hospitalización del Hospital General 450, en la Ciudad de Durango, Dgo., y el día 24 de Noviembre, se inauguró el Centro de Salud con Servicios Ampliados (CESSA) de Cd. Guadalupe Victoria, Dgo.

12.2.6.2. Servicios, funciones y objetivo que ofrecen los Servicios de Salud de Durango (SSD)

- Establecer y conducir la política nacional en materia de asistencia social, servicios médicos y salubridad general, con excepción de lo relativo al saneamiento del ambiente; y coordinar los programas de servicios a la salud de la administración pública federal, así como los agrupamientos por funciones y programas afines que, en su caso, se determinen.
- Crear y administrar establecimientos de salubridad, de asistencia pública y de terapia social.
- Aplicar a la asistencia pública los fondos que le proporcionen la lotería nacional y los pronósticos deportivos para la asistencia pública; y administrar el patrimonio de la beneficiencia pública en el Distrito Federal, en los términos de las disposiciones legales aplicables, a fin de apoyar los programas de servicios de salud.
- Organizar y vigilar las instituciones de beneficiencia privada, en los términos de las leyes relativas, e integrar sus patronatos, respetando la voluntad de los fundadores.
- Planear, normar, coordinar y evaluar el sistema nacional de salud y proveer a la adecuada participación de las dependencias y entidades públicas que presten servicios de salud, a fin de asegurar el cumplimiento del derecho a la protección de la salud.
- Asimismo, propiciará y coordinará la participación de los sectores social y privado en dicho sistema nacional de salud y determinará las políticas y acciones de inducción y concertación correspondientes.

- Dictar las normas técnicas a que quedará sujeta la prestación de servicios de salud en las materias de salubridad general, incluyendo las de asistencia social, por parte de los sectores público, social y privado, verificar su cumplimiento.
- Ejecutar el control sobre preparación, posesión, uso, suministro, importación, exportación de los de uso veterinario que no estén comprendidos en la convención de Ginebra.
- Poner en práctica las medidas tendientes a conservar la salud y la vida de los trabajadores del campo y de la ciudad y la higiene industrial, con excepción de lo que se relaciona con la previsión social en el trabajo.

12.2.6.3. Residencia General de Conservación de Carreteras

La Residencia General de Conservación de Carreteras forma parte de la Secretaria de Comunicaciones y Transportes la cual es una dependencia federal que tiene su origen funcional en la Secretaria de Estado y Derecho de Relaciones Exteriores e Interiores establecida el 8 de noviembre de 1821: Posteriormente, debido a las modificaciones efectuadas en el aparato de gobierno, las funciones relativas al ramo de comunicaciones y transportes se diseminaron entre varios organismos.

Los Centros SCT son las representaciones de la Secretaría en cada uno de los Estados que integran la Federación. Su misión es la de contribuir al desarrollo de los sistemas de comunicaciones y transportes en la entidad federativa, ejecutando y promoviendo los programas institucionales con seguridad, eficiencia y calidad, para el bienestar económico, social y cultural, con respecto al medio ambiente, al marco legal y ético.

La Residencia General cuenta con residencias de obras cuya función es llevar la supervisión y la correcta ejecución de los trabajos que se realizan en cada ejercicio fiscal, a cada una de estas residencias le corresponden diferentes tramos que conforman toda la Red Libre de Peaje del Estado de Durango.

12.3. Desarrollo

En la actualidad la Ciencia de los Datos permite apoyar los sistemas y procesos tecnológicos empleando las estadísticas recopilando una gran cantidad de información, garantizando una rapidez, precisión y eficacia, impulsando a la correcta toma de decisiones.

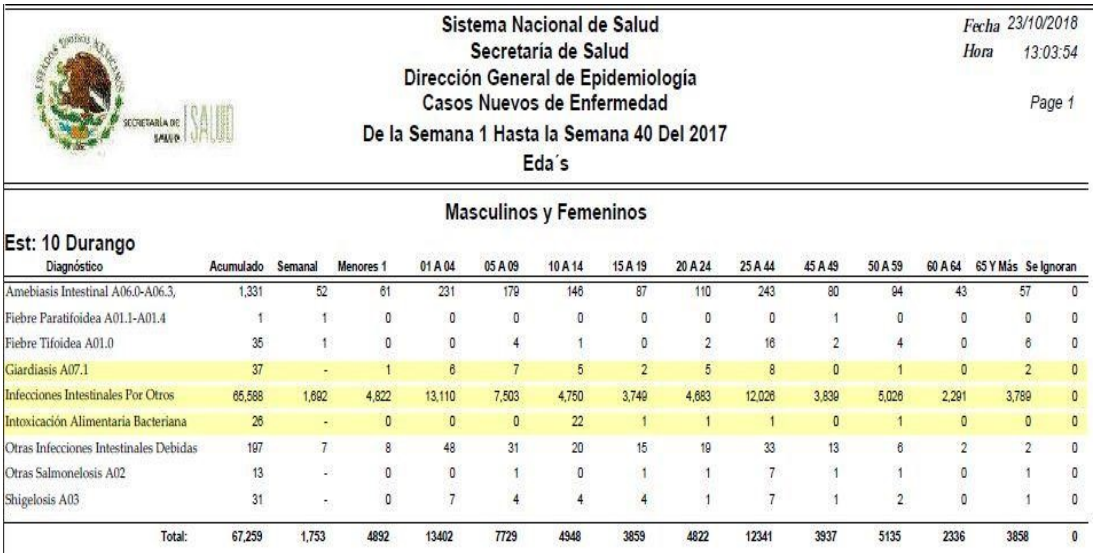
Hoy en día, en los Servicios de Salud de Durango se utiliza un sistema desarrollado en el año 2004 por el área de Epidemiología de la Dirección General de Epidemiología (DGEPI), el cual es el Sistema Único Automatizado para la Vigilancia Epidemiológica (SUAVE) para la captura y análisis de información de los casos nuevos de enfermedades registrados en el Estado, y como órgano rector en salud se concentra y analiza dicha información de todas las instituciones del sector salud (Secretaria de Salud, IMSS, ISSSTE, Prospera, DIF y SEDENA).

El SUAVE genera solamente reportes con información muy generalizada y no permite el análisis de la información de manera gráfica ni comparativa con años anteriores lo cual es fundamental para determinar el comportamiento de los diagnósticos de interés fundamental para la toma de decisiones para el área médica.

Algunos de los reportes del sistema SUAVE y que representan diagnósticos de interés Epidemiológico se citan a continuación:

- Enfermedades Diarreicas Agudas (EDAs).
- Infecciones Respiratorias Agudas (IRAs).
- Síndrome de Inmunodeficiencia Adquirida (SIDA).
- Virus de la Inmunodeficiencia Humana (VIH) .
- Conjuntivitis.
- Hepatitis A.
- Chicungunya.
- Enfermedades Febriles Exantemáticas (EFEs).
- Síndrome Coqueluchoide.
- Influenza.

En las figuras 3 y 4, se presenta un ejemplo de los reportes generados por el sistema separados por año 2017 y 2018 respectivamente de algunos padecimientos monitoreados constantemente.



Sistema Nacional de Salud
Secretaría de Salud
Dirección General de Epidemiología
Casos Nuevos de Enfermedad
De la Semana 1 Hasta la Semana 40 Del 2017
Eda´s

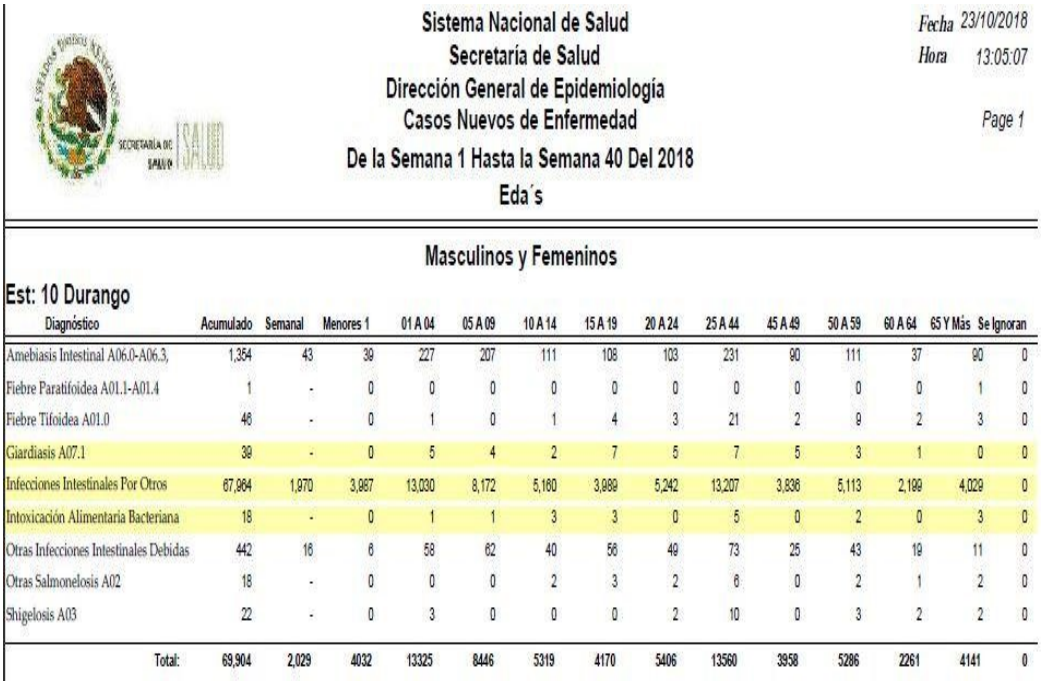
Fecha 23/10/2018
Hora 13:03:54
Page 1

Masculinos y Femeninos

Est: 10 Durango

Diagnóstico	Acumulado	Semanal	Menores 1	01 A 04	05 A 09	10 A 14	15 A 19	20 A 24	25 A 44	45 A 49	50 A 59	60 A 64	65 Y Más	Se Ignoran
Amebiasis Intestinal A06.0-A06.3,	1,331	52	61	231	179	146	87	110	243	80	94	43	57	0
Fiebre Paratifoidea A01.1-A01.4	1	1	0	0	0	0	0	0	0	1	0	0	0	0
Fiebre Tifoidea A01.0	35	1	0	0	4	1	0	2	16	2	4	0	6	0
Giardiasis A07.1	37	-	1	6	7	5	2	5	8	0	1	0	2	0
Infecciones Intestinales Por Otros	65,588	1,692	4,822	13,110	7,503	4,750	3,749	4,683	12,026	3,839	5,026	2,291	3,789	0
Intoxicación Alimentaria Bacteriana	26	-	0	0	0	22	1	1	1	0	1	0	0	0
Otras Infecciones Intestinales Debidas	197	7	8	48	31	20	15	19	33	13	6	2	2	0
Otras Salmonelosis A02	13	-	0	0	1	0	1	1	7	1	1	0	1	0
Shigelosis A03	31	-	0	7	4	4	4	1	7	1	2	0	1	0
Total:	67,259	1,753	4892	13402	7729	4948	3859	4822	12341	3937	5135	2336	3858	0

Figura 3. Reporte de Enfermedades Diarreicas Agudas (EDAs) 2017. SUAVE. Elaboración propia



Sistema Nacional de Salud
Secretaría de Salud
Dirección General de Epidemiología
Casos Nuevos de Enfermedad
De la Semana 1 Hasta la Semana 40 Del 2018
Eda´s

Fecha 23/10/2018
Hora 13:05:07
Page 1

Masculinos y Femeninos

Est: 10 Durango

Diagnóstico	Acumulado	Semanal	Menores 1	01 A 04	05 A 09	10 A 14	15 A 19	20 A 24	25 A 44	45 A 49	50 A 59	60 A 64	65 Y Más	Se Ignoran
Amebiasis Intestinal A06.0-A06.3,	1,354	43	39	227	207	111	108	103	231	90	111	37	90	0
Fiebre Paratifoidea A01.1-A01.4	1	-	0	0	0	0	0	0	0	0	0	0	1	0
Fiebre Tifoidea A01.0	46	-	0	1	0	1	4	3	21	2	9	2	3	0
Giardiasis A07.1	39	-	0	5	4	2	7	5	7	5	3	1	0	0
Infecciones Intestinales Por Otros	67,964	1,970	3,987	13,030	8,172	5,160	3,989	5,242	13,207	3,836	5,113	2,199	4,029	0
Intoxicación Alimentaria Bacteriana	18	-	0	1	1	3	3	0	5	0	2	0	3	0
Otras Infecciones Intestinales Debidas	442	16	6	58	62	40	56	49	73	25	43	19	11	0
Otras Salmonelosis A02	18	-	0	0	0	2	3	2	6	0	2	1	2	0
Shigelosis A03	22	-	0	3	0	0	0	2	10	0	3	2	2	0
Total:	69,904	2,029	4032	13325	8446	5319	4170	5406	13560	3958	5286	2261	4141	0

Figura 4. Reporte de Enfermedades Diarreicas Agudas (EDAs) 2018. SUAVE. Elaboración propia

En la figura 5, se presenta de manera general un archivo de la base de datos que utiliza el sistema SUAVE con las variables que cuenta para la generación de reportes.

CVE_ESTAD	CVE_JURISD	CVE_MUNI	CVE_UNIDA	NO_SEMAN	CVE_DIAGN	CVE_INSTIT	MENORE	DE0_1_A	DE0_5_A	DE1_0_A	DE1_5_A	DE2_0_A	DE2_5_A	DE4_5_A	DE5_0_A	DE6_0_A	DE6_5_Y	SE_IGNO	MENORE	DE0_1_A	DE0_5_A	DE1_0_A	DE1_5_A	DE2_0_A	DE2_5_A	DE4_5_A	DE5_0_A	DE6_0_A	DE6_5_Y	SE_IGNO	NO_TOT
10	01	034	5146	29	08	5	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
10	01	005	4003	26	49	4	0	0	0	0	0	0	0	0	0	1	2	0	0	0	0	0	0	0	0	0	0	1	1	0	5
10	01	005	4003	26	109	4	0	0	0	1	3	3	2	1	4	2	2	0	0	0	0	2	2	1	3	2	3	3	3	0	37
10	01	005	4001	26	179	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	1	1	0	0	6
10	04	018	3039	29	119	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
10	03	001	3027	29	08	3	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
10	03	001	3027	29	16	3	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	3
10	03	032	5134	26	109	5	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
10	03	037	5133	26	08	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	3
10	01	005	5024	29	110	5	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2
10	01	016	3004	29	109	3	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
10	04	028	3041	29	16	3	1	1	0	1	0	0	0	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	1	0	7
10	01	005	3002	29	109	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2
10	04	017	5054	29	08	5	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	1	0	0	3
10	02	012	3018	29	173	3	0	1	3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	7
10	04	028	1014	29	16	1	0	0	1	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	3
10	02	004	5117	29	110	5	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	4
10	01	023	5066	29	128	5	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
10	04	017	5015	29	16	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1

Figura 5. Variables y datos del sistema SUAVE. Elaboración propia

Algunos significados de las variables que muestra la figura 5, son claves de estados de la República Mexicana (estado 10 para Durango), claves de jurisdicciones, claves de municipios, claves de unidades médicas, número de semana, clave de diagnósticos, claves de pacientes atendido por edad masculinos y femeninos, entre otros.

En cada área se llevan a cabo tareas de índole médico y administrativo respectivamente, las cuales dan paso a la elaboración de reportes que proporcionan un control de las actividades realizadas. Toda esta información es capturada en cada área en formatos impresos y se ha venido haciendo el cambio en archivos

elaborados en computadora (Excel), con la finalidad de ofrecer el acceso inmediato de la información de un paciente.

12.3.1. Propuesta para solución para Área de Epidemiología del Sector Salud

Se proponen las herramientas de visualización R y Tableau ya que la información que genera el sistema SUAVE además de ser compatible con dicho software, no se requiere de una inversión económicamente hablando, ya que de otra manera se tendría que hacer una propuesta y esperar a que se apruebe el presupuesto de nivel federal.

Se busca que los informes generados por las dos propuestas permitan visualizar la información requerida por el o las distintas áreas involucradas de manera desglosada y comparativa por fecha para así observar el comportamiento de los diagnósticos y permita evitar o prever epidemias, brotes, aumentos inadecuados de enfermedades entre otras de interés social y epidemiológico

De manera adicional se tiene que tomar en cuenta, cuántos usuarios (empleados) usan actualmente las hojas de cálculo para realizar los análisis requeridos, si cuentan con conocimientos informáticos o conocimientos básicos en programación, cuántos tipos de reportes realizan, si la información se recibe de manera semanal, mensual o anual y que tipo de reporte son los de mayor interés.

12.3.2. Análisis de la propuesta para Área de Epidemiología del Sector Salud

Una de las fortalezas de R es la facilidad con la que se pueden producir gráficos de calidad de publicación bien diseñados, incluyendo símbolos matemáticos y fórmulas donde sea necesario.

R ofrece una poderosa manera de realizar análisis estadísticos en conjuntos de datos grandes. También es gratuito, lo que se convierte en un factor atractivo para su crecimiento. Debido a que tiene código abierto, se crean funciones y paquetes nuevos todo el tiempo. R se puede extender (fácilmente) a través de paquetes.

Tableau permite consumir sus datos de manera segura en un navegador, en el escritorio, con un dispositivo móvil o incorporados en una aplicación (Tableau, s.f.). Puede diseñar un dashboard interactivo con menús desplegables, controles deslizantes y otros elementos visuales en cuestión de minutos.

En las figuras 6 y 7, se muestran ejemplos visuales tanto de Tableau como de R. En éstas representaciones visuales se da a conocer comparativos EDAs de los años 2017 y 2018, con información real del sistema SUAVE, realizadas con base a los requerimientos de los usuarios.

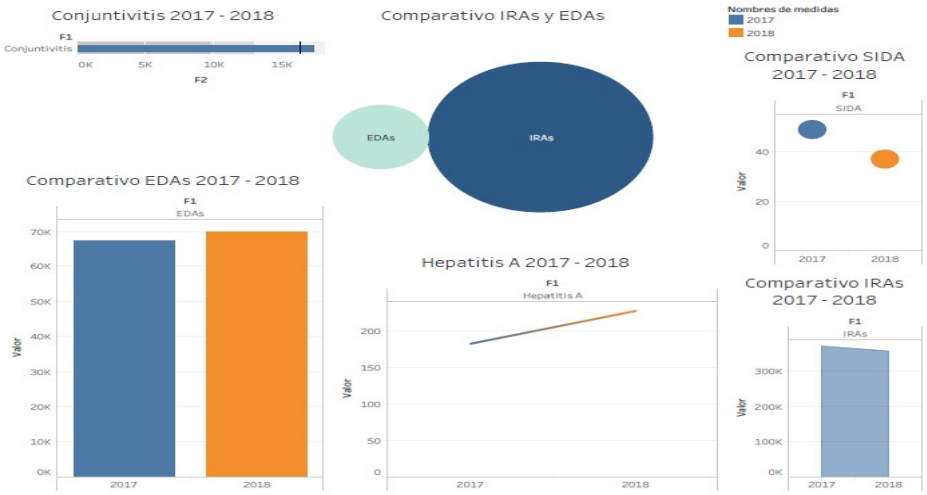


Figura 6. Dashborard de Tableau. Comparativos EDA's 2017-2018. Elaboración propia

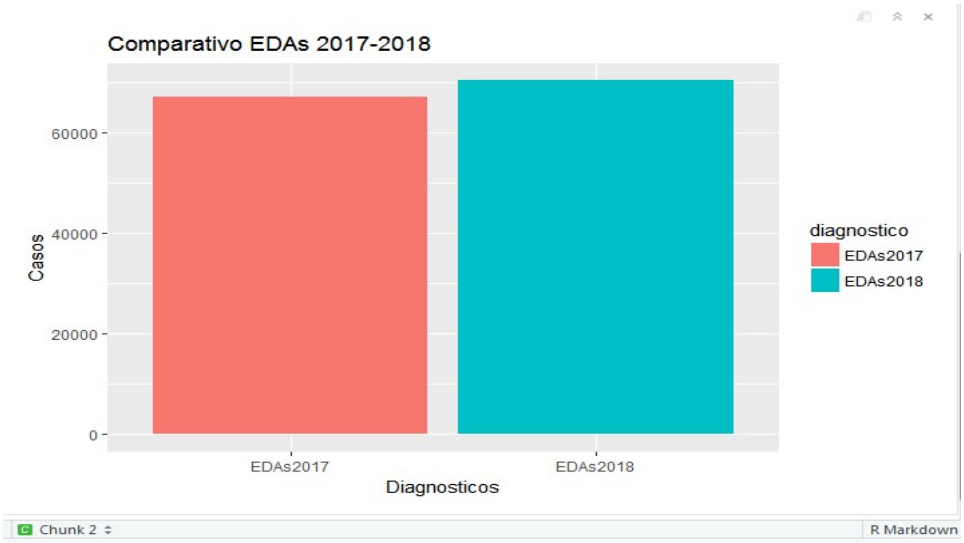


Figura 7. Comparativo generado en R. EDA's 2017-2018. Elaboración propia

Como parte de la propuesta, se sugiere una revisión a toda la información posible del Área de Vigilancia Epidemiológica, investigar documentos de estrategia interna, los reportes generados, entrevistas con los usuarios, explicaciones y capacitación de las ventajas de las herramientas de visualización de datos y de los distintos análisis de aplicar alguna de las herramientas R o Tableau.

Ahora bien, hablando de cuestiones sanitarias y al momento que se está viviendo con respecto a la pandemia denominada COVID 19 y que definitivamente tiene que ver con aspectos del Sector Salud, la figura 8, muestra una visualización de casos confirmados de la enfermedad en el todo el País con fecha de corte al 2 de agosto del año 2020.

Arriba en la imagen de la figura 8, se muestra el total de casos confirmados, defunciones, negativos, sospechosos y tasa de letalidad; del lado izquierdo con el mapa de la República Mexicana se pueden visualizar los casos por Estado, de lado derecho la cantidad de casos confirmados por día y la curva de acumulados y por debajo la distribución de casos por edad.

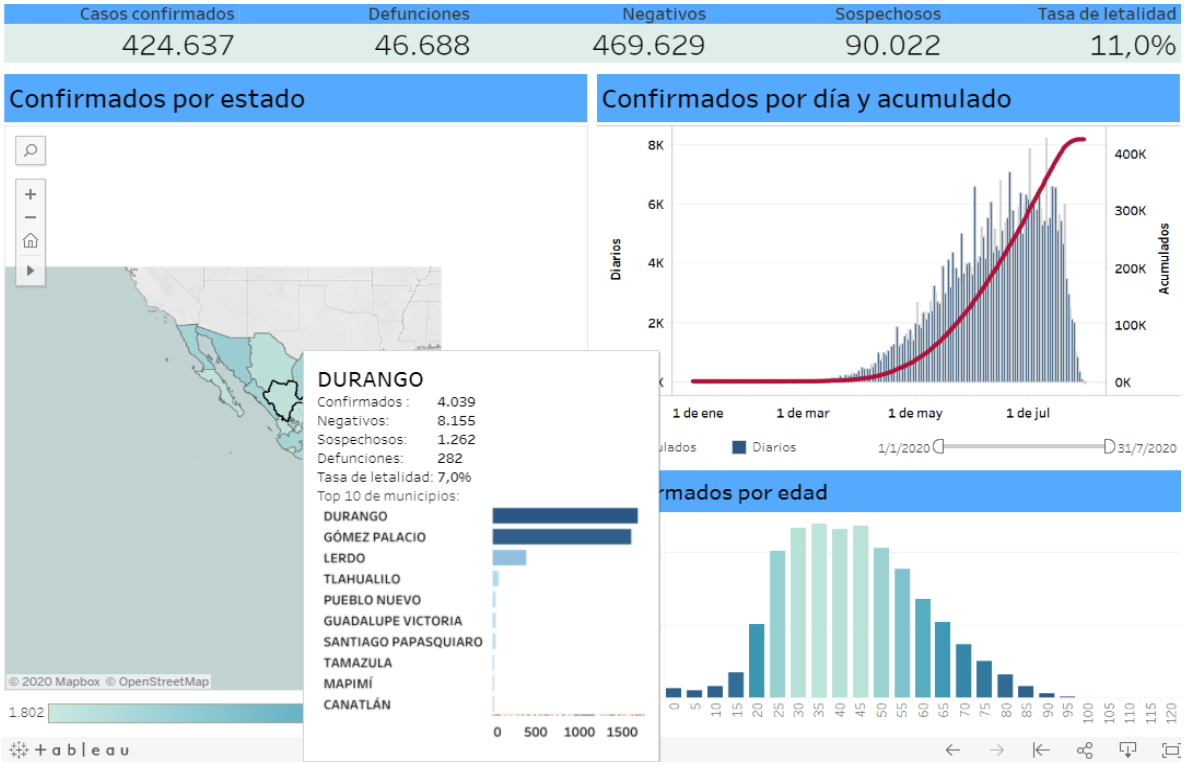


Figura 8. Datos Confirmado de COVID19, generados en Tableau. (Veláquez Luna, 2020).

Con la imagen anterior de la figura 8, del caso COVI19 extraído de la literatura se intenta reforzar la propuesta del uso de herramientas de análisis y visualización de datos como Tableau y el lenguaje de programación R para fortalecer procesos de interpretación de datos y toma de decisiones en las organizaciones en este caso en el Sector Salud.

12.3.3. Propuesta para solución para la Residencia General de Conservación de Carreteras de la S.C.T

Existe diversos informes de gran importancia dentro de las funciones que día con día se realizan en la Residencia General de Conservación de Carreteras, éstos dan se dan conocer periódicamente a los distintos niveles de mando tanto en el mismo Centro SCT como a Nivel Central en la Ciudad de México. En este apartado se citan tres ejemplos de informes.

Los informes, se elaboran en hojas de cálculo de Excel y contienen bastante información que implica que para visualizarlos sea necesario abrir el archivo con la herramienta correspondiente, o si se imprime forzosamente sea en un plotter, lo cual hace difícil de manipular al momento de analizar la información.

Un primer informe que se elabora quincenal y mensualmente, contiene la información correspondiente a todo el proceso de licitación, contratación, ejecución y terminación de las obras que se ejecutan en el Ejercicio Fiscal vigente.

Al principio de cada año personal del Departamento de Adecuaciones y Control del Programa de Obras que está adscrito a la Dirección General de Conservación de Carreteras turna a cada uno de los Centros SCT el formato en Excel con el Presupuesto de Egresos de la Federación (PEF) autorizado para el Ejercicio Fiscal en curso, a partir de las obras y recursos autorizados, periódicamente y conforme se realicen los procesos de Licitación los campos se van alimentando en el informe y se envía en las fechas correspondientes solo de manera digital por medio de correo electrónico.

Por mencionar algunos de los campos que concentra este informe es Estado, Programa, Tipo de Trabajos, Tramo, Ubicación, Meta, Asignación, Número de

Licitación, Fechas del procedimiento, Modalidad, Tipo de Contrato, Fecha de Firma del Contrato, Número de Contrato, Oficio de Autorización, Datos de la Empresa y así otros más hasta llegar a los avances físicos y financieros.

En la figura 9, se muestra con poca claridad este informe elaborado en Excel. La idea de la figura 9 no es tratar de que el lector adivine los datos de las columnas, sino que se genere una percepción de que la propuesta vertida en este documento es que se puede utilizar de herramientas con mayores funcionalidades para el análisis y la visualización de datos que ayuden a la comprensión y la toma de decisiones. Ahora bien, tampoco se trata de recomendar en la propuesta eliminar la herramienta Excel, se sugiere siga siendo usada como herramienta para almacenar datos y sea utilizada como fuente de datos para Tableau y R.

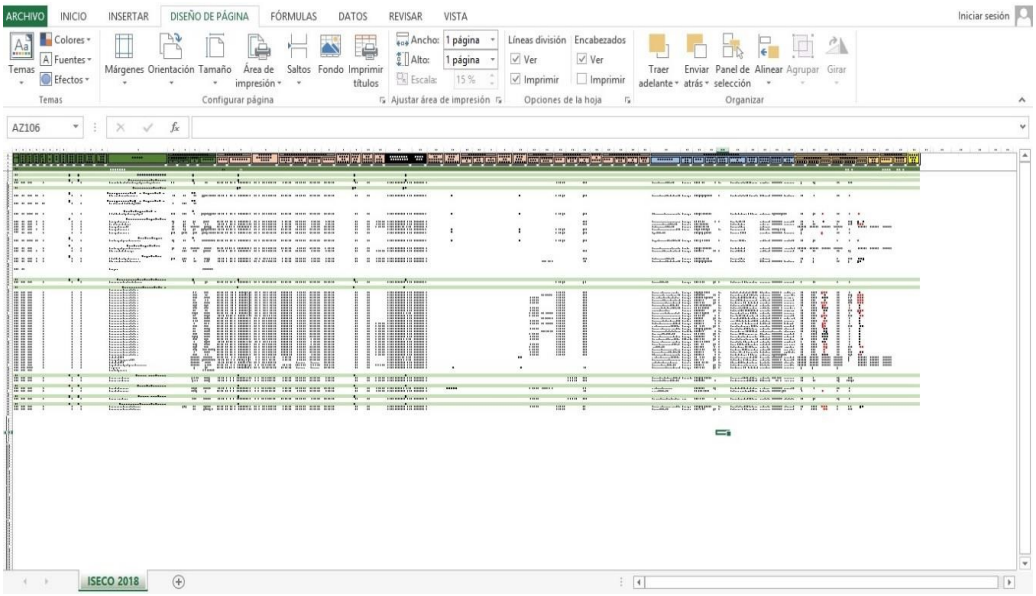


Figura 9. Informe del proceso de licitación, contratación, ejecución y terminación de las obras. Elaboración propia

Un segundo informe, existe el relacionado con el programa nacional de conservación de carreteras, este reporte se elabora y envía mensualmente en la Residencia General, a diferencia del primero este informe se envía en la fecha correspondiente de manera electrónica y documental.

Debido a que antes de enviarlo a la Dirección de Planeación y Evaluación, departamento adscrito a la Dirección General de Conservación de Carreteras; es

revisado, validado y rubricado por el Residente General, el Subdirector de Obras y finalmente por el Director General del Centro SCT.

Algunos datos que contiene el reporte son Tipo de Red, Partida Presupuestal, Número SAOP, Número de Contrato, Nombre de la Empresa, RFC, Tipo de Trabajo, Nombre de la Obra, Ubicación, Meta, Asignación, Avances Físicos y Financieros, fechas diversas, entre otros.

En la figura 10, se presente un ejemplo de un informe de conservación de carreteras; de igual forma no se percibe con claridad ni variables ni sus valores, sin embargo y de manera similar la propuesta es el uso de Tableau y R para diversos análisis de datos.

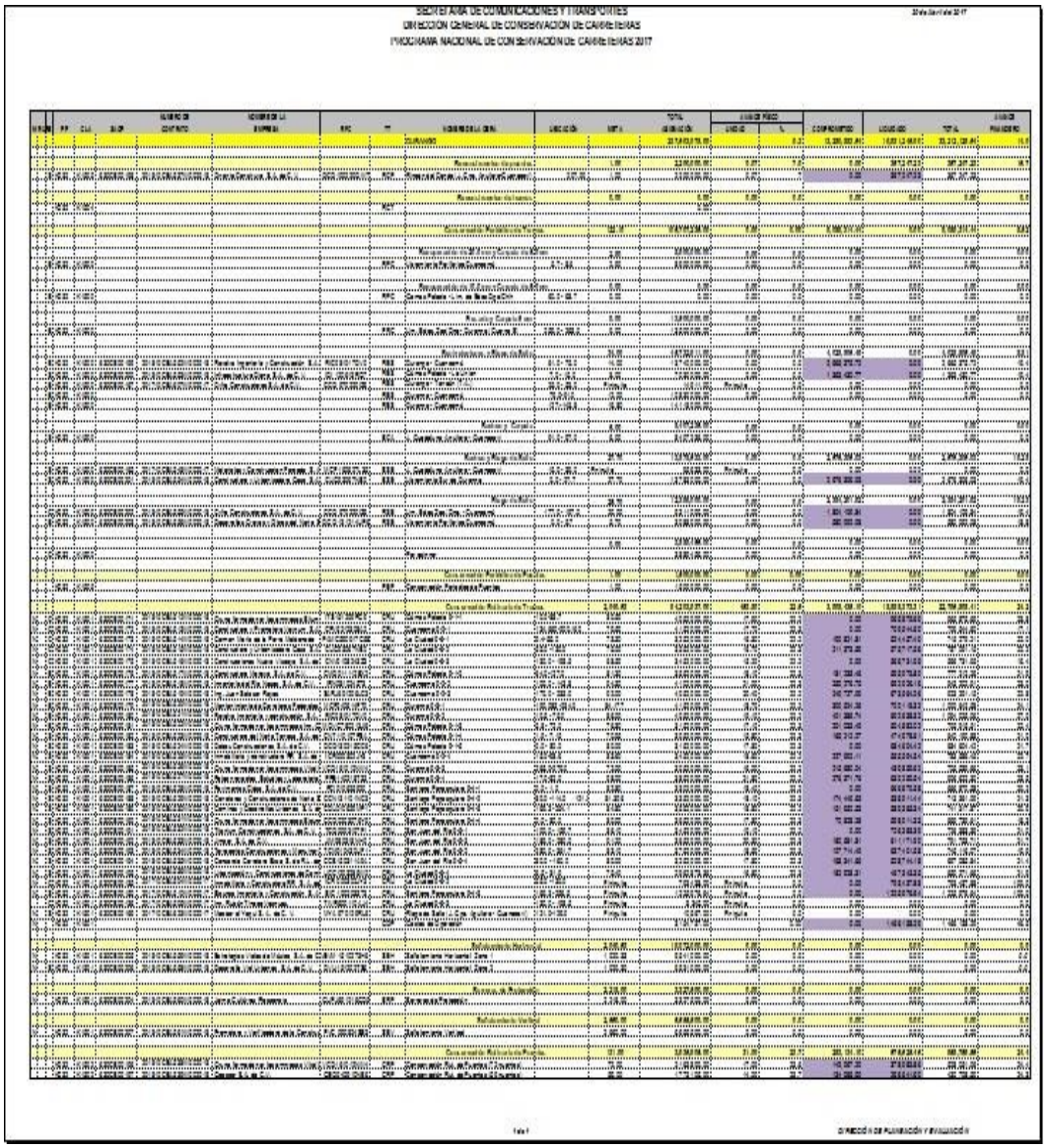


Figura 10. Informe Programa Nacional de Conservación de Carreteras. Elaboración propia

Por otra parte en la figura 11 (primera parte) y figura 12 (segunda parte), se presenta un reporte de dominio público extraído del portal de la SCT del enlace: http://www.sct.gob.mx/fileadmin/DireccionesGrales/DGCC/PDF/DGO_aff_julio2020.pdf que muestra el avance de obra para el Estado de Durango con fecha Julio 2020 y que de manera más clara se puede observar datos relacionados con este tipo de informe.

COMUNICACIONES
SECRETARÍA DE COMUNICACIONES Y TRANSPORTES

SUBSECRETARIA DE INFRAESTRUCTURA
DIRECCION GENERAL DE CONSERVACION DE CARRETERAS
Programa Nacional de Conservación de Carreteras
Avance Fisico - Financiero

JULIO 2020

No.	NOMBRE DE LA OBRA	UBICACIÓN		META	ASIGNACION	AVANCE FISICO		AVANCE FINANCIERO %
		km inicial	km final			UNIDAD	%	
	DURANGO				**318,733,293.00**		**86.31**	**69.00**
	RECONSTRUCCIÓN				**0.00**			
	Reconstrucción de Tramos				**0.00**			
	Recursos a Distribuir				0.00			
	Conservación Periódica			**112.10**	**174,619,902.85**	**111.88**	**99.57**	**85.16**
	Recuperación, Base Estabilizada y Riego de Sello			**21.00**	**57,487,293.00**	**20.78**	**98.70**	**64.99**
1	Cuencamé - Torreón	159.00	170.00	11.00	37,487,293.00	10.78	98.00	46.31
2	Cuencamé - Torreón	180.00	190.00	10.00	20,000,000.00	10.00	100.00	100.00
	Renivelación y Carpeta de 5.0 cm			**12.00**	**35,533,124.99**	**12.00**	**100.00**	**97.71**
3	Lím. de Edos. Zac./Dgo. - Durango (Cpo. A)	267.00	273.00	6.00	17,925,044.07	6.00	100.00	100.00
4	Lím. de Edos. Zac./Dgo. - Durango (Cpo. A)	273.00	279.00	6.00	17,608,080.92	6.00	100.00	95.38
	Carpeta de 5.0 cm			**22.00**	**27,901,469.13**	**22.00**	**100.00**	**97.65**
5	Durango - Cuencamé (16 KM EN TRAMOS AISLADOS)	25.00	55.00	16.00	19,947,163.53	16.00	100.00	97.50
6	Durango - Cuencamé	55.00	61.00	6.00	7,954,305.60	6.00	100.00	98.02
	Renivelación y Riego de Sello			**47.00**	**50,761,968.86**	**47.00**	**100.00**	**91.56**
7	J. Guadalupe Aguilera - Guanaceví	70.00	80.00	10.00	10,112,760.83	10.00	100.00	96.40
8	J. Guadalupe Aguilera - Guanaceví	80.00	91.00	11.00	12,550,000.00	11.00	100.00	91.34
9	Durango - Lím. de Edos. Dgo./Chih.	164.00	180.00	16.00	15,407,953.12	16.00	100.00	92.97
10	Durango - Lím. de Edos. Dgo./Sin.	90.00	100.00	10.00	12,691,254.91	10.00	100.00	86.21

Figura 11 Avance físico financiero de obra en Durango (Primera parte) de Julio 2020 (SCT. Subecretaría de Infraestructura. Dirección General de Conservación de Carreteras, 2020).

	Riego de Sello			**10.10**	**2,936,046.87**	**10.10**	**100.00**	**98.55**
11	Ramal Gómez Palacio - La Unión	1.90	7.00	5.10	1,495,439.97	5.10	100.00	100.00
12	Ramal Gómez Palacio - La Unión	35.00	40.00	5.00	1,440,606.90	5.00	100.00	97.05
	Conservación Rutinaria de Tramos			**1,968.23**	**101,074,656.00**	**1,357.69**	**70.83**	**48.44**
	Conservación Rutinaria de Tramos			153.70	8,656,000.00	153.70	100.00	45.61
	Conservación Rutinaria de Tramos			132.90	5,956,000.00	132.90	100.00	57.34
	Conservación Rutinaria de Tramos			80.00	3,456,000.00	80.00	100.00	50.49
	Conservación Rutinaria de Tramos			90.75	4,956,000.00	82.29	90.68	45.22
	Conservación Rutinaria de Tramos			99.25	5,556,000.00	58.96	59.41	49.01
	Conservación Rutinaria de Tramos			103.70	4,456,000.00	52.89	51.00	48.69
	Conservación Rutinaria de Tramos			129.80	6,556,000.00	68.80	53.00	47.20
	Conservación Rutinaria de Tramos			92.18	5,556,000.00	67.31	73.02	57.77
	Conservación Rutinaria de Tramos			119.80	5,656,000.00	73.98	61.75	52.59
	Conservación Rutinaria de Tramos			119.40	3,049,610.68	87.24	73.07	100.00
	Conservación Rutinaria de Tramos			120.61	1,756,000.00	86.84	72.00	100.00
	Conservación Rutinaria de Tramos			80.00	18,418,605.25	50.40	63.00	13.04

Figura 12 Avance físico financiero de obra en Durango (Segunda parte) de Julio 2020 (SCT. Subecretaría de Infraestructura. Dirección General de Conservación de Carreteras, 2020).

12.3.4. Análisis de la propuesta para para la Residencia General de Conservación de Carreteras de la SCT

En la actualidad en la Residencia General de Conservación de Carreteras se realizan periódicamente reuniones donde intervienen las Residencias de Obras responsables de la ejecución de las obras, personal de la Oficina Técnica encargada del trámite y pago de las estimaciones, así como de elaborar la presentación de los informes de Avances Físicos y Financieros, y finalmente el Residente General y el Subdirector de Obras, altos mandos que interpretan la información presentada para la toma de las decisiones que tengan lugar.

La forma en que se han realizado la entrega de resultados es mediante presentaciones elaboradas con Microsoft Power Point con gráficas simples generadas mediante Microsoft Excel. Como ejemplo de ello la figura 13.

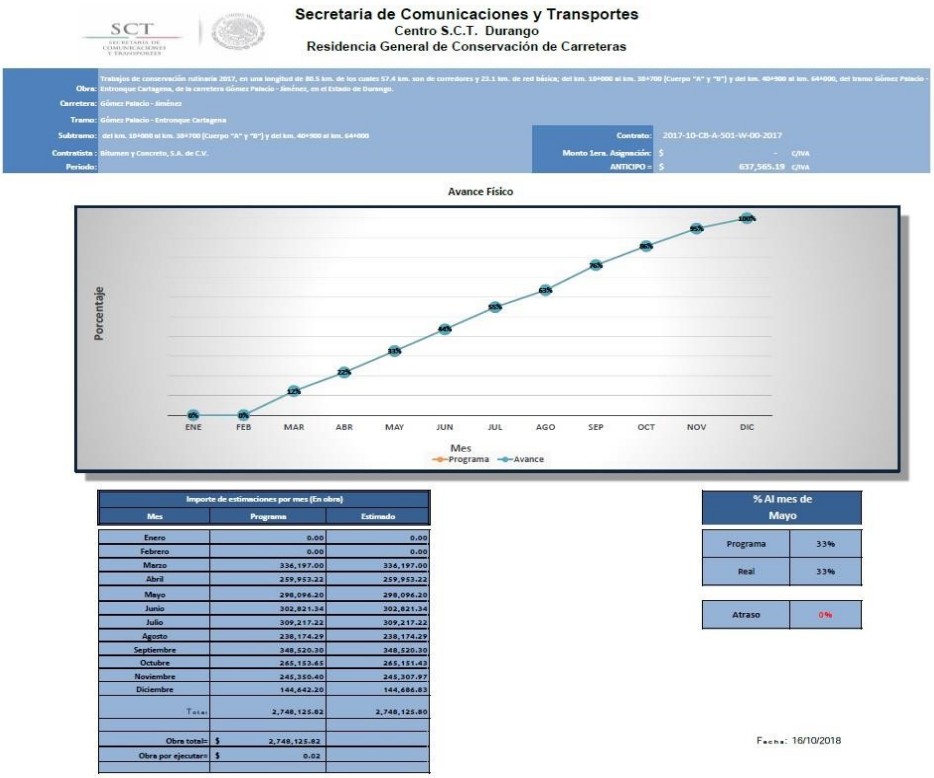


Figura 13. Visualización de la información de Avances Físicos y Financieros actualmente. Elaboración propia

Dentro de la gran cantidad de características que proporciona Tableau Desktop se pueden generar diferentes formas de visualización de la información la cual permite interactuar con la misma y así ampliar el panorama del conocimiento que se puede obtener para la toma de decisiones, si se cumple el objetivo de implementar dicha herramienta para representar los avances físicos y financieros en la Residencia General de Conservación de Carreteras, se generarían *dashboards* o reportes ejecutivos como lo que se muestra en la figura 14.

Figura 14. Dashboard de Avances Físicos y Financieros de contratos a cargo de la RGCC. Elaboración propia

Así mismo se recomienda aprovechar la funcionalidad de generación de mapas que integra la herramienta Tableau para de una manera visual representar la ubicación de las obras, la figura 15, muestra un mapa mundial con indicadores que representan a cada País a manera de ejemplo de visualización de mapas que se pueden hacer en Tableau.

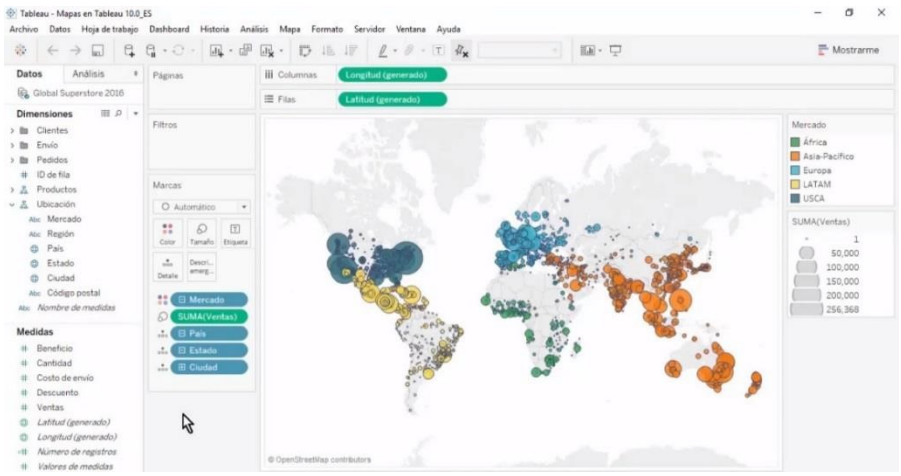


Figura 15. Un ejemplo de mapas en Tableau. Elaboración propia

En la figura 16 se gráficos con datos de INEGI como una muestra de aspectos visuales de población dispersa en estados de la República Mexicana haciendo una semejanza y una recomendación de realizar informes y reportes de la Residencia General de Conservación de Carreteras de la S.C.T utilizando herramientas relacionadas con Ciencia de los Datos, en este caso la herramienta visual de Tableau.

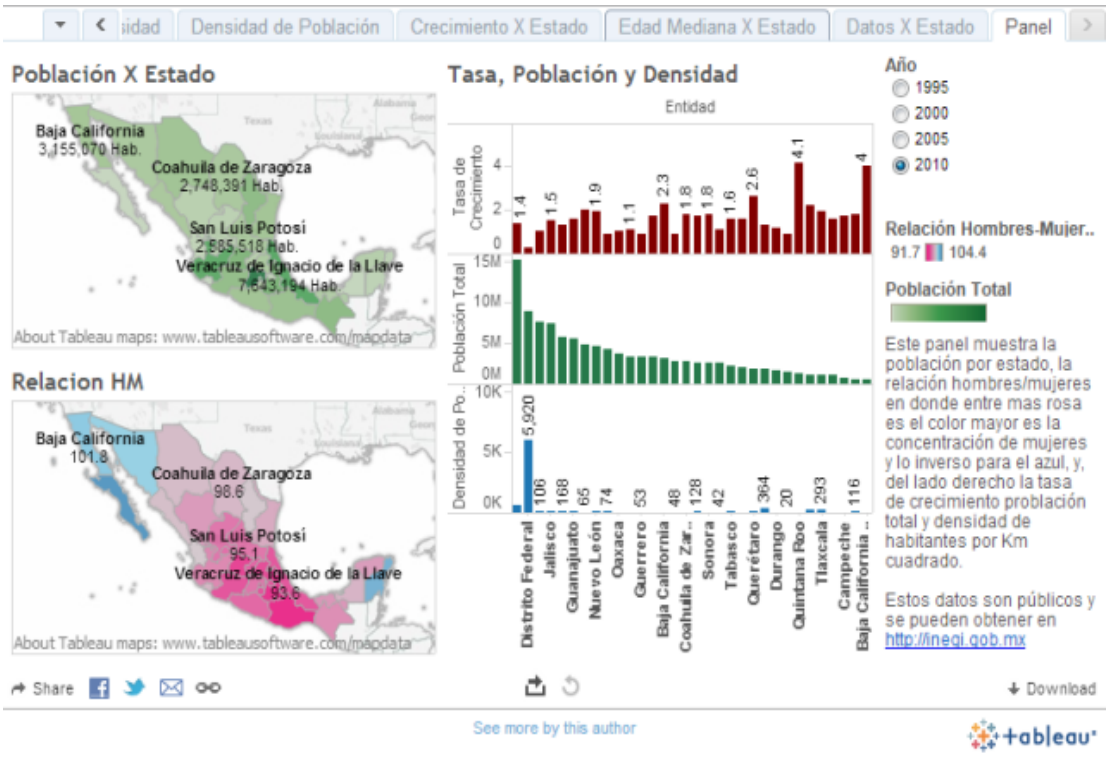


Figura 16. Ejemplo de visualizaciones en Tableau con de población de INEGI. (Sada, 2013).

Conclusiones

Debido a los avances de la tecnología producidos sobre todo a lo largo del siglo XX, la humanidad se ha transformado en una sociedad de la información con el uso de las nuevas tecnologías.

Está claro que las computadoras y las telecomunicaciones han reportado numerosos beneficios a las organizaciones, de hecho, gracias a la informática se permite estar comunicado independientemente del sector en donde se encuentre científico, industrial, sanitario, artístico, económico, de salud, entre otros.

El gran volumen y variedad de datos tienen un impacto en el desarrollo de tecnologías de análisis de datos empresariales, estos datos requieren mejoras más amplias en las tecnologías de recolección, almacenamiento, procesamiento y análisis.

Las herramientas de análisis y visualización de datos son intuitivas, fáciles de usar, ofrecen capacidades de autoservicio, procesamiento de memoria, visualización interactiva y flexibilidad. Con estas características se logra que las aplicaciones sean más rápidas, flexibles y que los usuarios puedan acceder directamente a los datos y a capacidades de análisis de datos con menos dependencia de intermediarios.

Después de conocer y ejecutar algunas de las funciones de la herramienta Tableau y ejecutar algunos comandos y funciones del lenguaje de programación R además basado en los requerimientos de los usuarios del área de Vigilancia epidemiológica de Los Servicios de salud de Durango, se determinó que ambos sistemas cumplen con la finalidad de analizar y presentar la información requerida por el área.

Sin embargo, se puede resaltar que a pesar de cumplir con el resultado requerido existen diferencias que fueron determinantes para la selección de uno de ellos.

Se observó que existen algunas limitaciones naturales desde el punto de vista usuarios, para el caso de utilizar R y poder aprovechar el poder de análisis y la capacidad para visualización que tiene lenguaje, se necesitan habilidades de programación.

La herramienta que se sugiere más viable para implementar en el área, es el sistema Tableau, ya que la curva de aprendizaje es más corta y por ser un sistema más intuitivo, no se requiere de conocimientos de programación para el análisis de la información.

A pesar de contar con personal informático en el área de Vigilancia Epidemiológica se requiere que también el personal médico pueda realizar análisis de información inclusive en otras áreas como las unidades u hospitales de las cabeceras municipales.

Otra de las ventajas observadas con el sistema Tableau es la capacidad de mostrar tableros (dashboard) de visualización del análisis de la información, lo cual

permite observar de manera más fácil y rápida y así poder detectar algo anormal en el comportamiento de la información y poder tomar decisiones y realizar acciones que beneficien a la población.

Se puede destacar que el uso de las tecnologías de la información y utilizando los sistemas adecuados a nuestras necesidades sin duda mejora los procesos y asegura un análisis, manejo y presentación de la información viable y por lo tanto en resultados de beneficio para todos.

Con respecto a la propuesta de Residencia General de Conservación de Carreteras, se plantea en base a una necesidad en el área de trabajo derivada de la forma en que se visualiza la información referente a los Avances Físicos y Financieros de las Obras que se ejecutan en la Residencia General.

Se puede destacar que el uso de las tecnologías de la información y utilizando los sistemas adecuados a las necesidades sin duda mejora los procesos y asegura un análisis, manejo y presentación, mediante los *dashboards* y aquellas opciones para crear mapas permita tener un panorama más amplio de aquellos detalles que tal vez con una presentación tradicional en Power Point o una tabla generada en Excel pueda representar. De esta forma los diferentes niveles de mando pueden extraer el mayor conocimiento posible y estén en la posibilidad de tomar las decisiones correctas en el momento oportuno.

Este trabajo es una ventana a las diferentes herramientas de visualización que están tomando fuerza, en específico Tableau Desktop, para aquel usuario que dentro de sus funciones esté la de presentar los resultados de las actividades o procesos que periódicamente se ejecuten en su lugar de trabajo, dejando atrás el uso de herramientas tradicionales como Power Point y Excel, que si bien son funcionales no cuentan con la capacidad de visualización que integra Tableau Desktop o cualquiera de sus versiones.

Para la Residencia General de Conservación de Carreteras esta propuesta de implementación deja una opción para cambiar el modo en que habitualmente se viene presentando la información a los altos mandos.

Implementar Tableau Desktop en la Residencia General y ver reflejados los resultados, puede ser el inicio de su uso no sólo en el área de trabajo objeto de éste producto académico, sino en los diferentes departamentos que conforman el Centro SCT Durango, por mencionar algunos como el Departamento de Recursos Humanos, el Departamento de Recursos Materiales, el Departamento de Autotransporte Federal, entre otros.

Referencias

Alcalde Perea, I. (2015). *Visualización de la información: De los datos al conocimiento.* Barcelona: UOC.

Ayala, J., Ortiz, J. G., & Maya, E. (2018). Herramientas de Business Intelligence (BI) modernas, basadas en memoria y con lógica asociativa . *REVISTA PUCE. ISSN: 2528-8156. NÚM.106*, 20.

Carrión, J. (s.f. de s.f. de s.f.). *Diferencia entre Dato, Información y Conocimiento.* Obtenido de Diferencia entre Dato, Información y Conocimiento: http://biblioteca.udgvirtual.udg.mx:8080/jspui/bitstream/123456789/869/3/Diferencia_entre_dato_informaci%c3%b3n.pdf

De Juana, R. (18 de 02 de 2019). *MCPRO*. Obtenido de El Cuadrante Mágico de Gartner: casi todo lo que tienes que saber: https://www.muycomputerpro.com/2019/02/18/el-cuadrado-magico-de-gartner-casi-todo-lo-que-tienes-que-saber

Flores Avendaño, P. A., & Villacís Vera, A. E. (01 de 09 de 2017). Análisis Comparativo de las Herramientas de Big Data. *En la Facultad de Ingeniría de la Pontificia Universidad Católica del Ecador*. Quito, Quito, Ecuador: Pontifica Universidad Católica del Ecuador. Facultad de Ingeniería de Sistemas y Computación.

Jones, H. (2019). *Analítica de Datos. La guía definitiva de análisis de Big Data para empresas, técnicas de minería de datos, recopilación de datos y conceptos de inteligencia empresarial.* México: Independently published.

Jones, H. (2019). *Ciencia de los Datos. Lo que saben los mejores científicos de datos sobre el análisis de datos, minería de datos, estadísticas, aprendizaje automático y Big Data que usted desconoce.* México: Amazon Mexico Services, Inc.

Martínez, R. (s.f. de s.f. de s.f.). *Empoderamiento de los ciudadnos en el análisis de los datos abiertos.* Obtenido de Empoderamiento de los ciudadnos en el análisis de los datos abiertos:

http://openaccess.uoc.edu/webapps/o2/bitstream/10609/72847/6/xyulexTFM0118memoria.pdf

Olivo Cabezas, R. C. (01 de 08 de 2016). Uso de la Herramienta Ofimática Word y su Influencia en el Proceso de Enseñanza–Aprendizaje a Estudiantes de la Escuela Educación Básica "Juan E. VEerdeSoto" Cantón Babahoyo, Provincia los Ríos. *Trabajo de Grado. Licenciatura en Ciencias de la Educación*. Babahoyo, Babahoyo, Ecuador: Universidad Técnica de Babahoyo.

Puerta Gálvez, A. (2016). *Business Intelligence Y La Tecnología de la Información.*

r-project.org. (01 de 01 de 2019). *r-project.org*. Obtenido de r-project.org: https://www.r-project.org/

Sada, R. (04 de 09 de 2013). *Tableau*. Obtenido de Poblacion MX INEGI: https://public.tableau.com/profile/ricardo.sada#!/vizhome/PoblacionMXINEGI/Panel

SCT. Subecretaría de Infraestructura. Dirección General de Conservación de Carreteras. (2020). *Programa Nacional de Conservación de Carreteras. Avance Avance Fìsico - Financiero.* SCT.

Sedeño Valdellós, A. (2016). La visualización de datos como recurso social: posibilidades educativas y de activismo. *Razón y palabra. Primera Revista Electrónica en Iberoamérica Especializada en Comunicación. Vol. 20, Núm 1*, 14.

Tableau. (s.f.). *Tableau para equipos y organizaciones*. Obtenido de https://www.tableau.com/es-es/products/teams-organizations

Tableau Softwate, LLC. (s.f.). *Tableau*. Obtenido de Que es Tableau: https://www.tableau.com/es-mx

Veláquez Luna, J. C. (02 de 08 de 2020). *Tableau Public*. Obtenido de COVID-19 en México: https://public.tableau.com/profile/jos.carlos.vel.zquez.luna#!/vizhome/proyecto_15964084802330/Dashboard1

CIENCIA DE LOS DATOS

Universidad Pedagógica de Durango

1948-2020 ANIVERSARIO