Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

PROPUESTA DE METODOLOGÍA PARA LA IMPLEMENTACIÓN DE CIENCIA DE DATOS EN LA TOMA DE DECISIONES GERENCIALES EN GUATEMALA

Diego Ernesto Sebastian Osorio García
Asesorado por el Ing. Oscar Alejandro Paz Campos

Guatemala, noviembre de 2021

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

PROPUESTA DE METODOLOGÍA PARA LA IMPLEMENTACIÓN DE CIENCIA DE DATOS EN LA TOMA DE DECISIONES GERENCIALES EN GUATEMALA

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

DIEGO ERNESTO SEBASTIAN OSORIO GARCÍA
ASESORADO POR EL ING. OSCAR ALEJANDRO PAZ CAMPOS

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO EN CIENCIAS Y SISTEMAS

GUATEMALA, NOVIEMBRE DE 2021

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANO	Inga. Aurelia Anabela Córdova Estrada
VOCAL I	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Br. Kevin Vladimir Armando Cruz Lorente
VOCAL V	Br. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	Inga. Aurelia Anabela Córdova Estrada
EXAMINADOR	Ing. César Augusto Fernández Cáceres
EXAMINADOR	Ing. César Rolando Batz Saquimux
EXAMINADOR	Ing. Sergio Arnaldo Méndez Aguilar
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

PROPUESTA DE METODOLOGÍA PARA LA IMPLEMENTACIÓN DE CIENCIA DE DATOS EN LA TOMA DE DECISIONES GERENCIALES EN GUATEMALA

Tema que me fuera asignado por la Dirección de la Escuela de Ingeniería en Ciencias y Sistemas, con fecha febrero de 2021.

Diego Ernesto Sebastian Osorio García

Guatemala, 28 de octubre de 2021

Ingeniero
Carlos Alfredo Azurdia
Coordinador de Privados y Trabajos de Tesis
Escuela de Ingeniería en Ciencias y Sistemas
Facultad de Ingeniería - USAC

Respetable Ingeniero Azurdia:

Por este medio hago de su conocimiento que en mi rol de asesor del trabajo de investigación realizado por el estudiante **DIEGO ERNESTO SEBASTIAN OSORIO GARCÍA** con carné **201503692 y CUI 3476 16437 0101** titulado **"PROPUESTA DE METODOLOGÍA PARA LA IMPLEMENTACIÓN DE CIENCIA DE DATOS EN LA TOMA DE DECISIONES GERENCIALES EN GUATEMALA.",** lo he revisado y luego de corroborar que el mismo se encuentra concluido y que cumple con los objetivos propuestos en el respectivo protocolo, procedo a la aprobación respectiva.

Al agradecer su atención a la presente, aprovecho la oportunidad para suscribirme,

Atentamente,

Ing. Oscar Alejandro Paz Campos
Colegiado No. 6430

Universidad San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Ingeniería en Ciencias y Sistemas

Guatemala 9 de noviembre de 2021

Ingeniero
Carlos Gustavo Alonzo
Director de la Escuela de Ingeniería
En Ciencias y Sistemas

Respetable Ingeniero Alonzo:

Por este medio hago de su conocimiento que he revisado el trabajo de graduación del estudiante **DIEGO ERNESTO SEBASTIAN OSORIO GARCÍA** con carné **201503692 y** CUI **3476 16437 0101** titulado **"PROPUESTA DE METODOLOGÍA PARA LA IMPLEMENTACIÓN DE CIENCIA DE DATOS EN LA TOMA DE DECISIONES GERENCIALES EN GUATEMALA"** y a mi criterio el mismo cumple con los objetivos propuestos para su desarrollo, según el protocolo aprobado.

Al agradecer su atención a la presente, aprovecho la oportunidad para suscribirme,

Atentamente,

Ing. Carlos Alfredo Azurdia
Coordinador de Privados
y Revisión de Trabajos de Graduación

ESCUELA DE INGENIERÍA EN CIENCIAS Y SISTEMAS

UNIVERSIDAD DE SAN CARLOS
DE GUATEMALA

FACULTAD DE INGENIERÌA
ESCUELA DE INGENIERÍA EN
CIENCIAS Y SISTEMAS

El Director de la Escuela de Ingeniería en Ciencias y Sistemas de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor con el visto bueno del revisor y del Licenciado en Letras, del trabajo de graduación ***"PROPUESTA DE METODOLOGÍA PARA LA IMPLEMENTACIÓN DE CIENCIA DE DATOS EN LA TOMA DE DECISIONES GERENCIALES EN GUATEMALA",*** *realizado por el estudiante, DIEGO ERNESTO SEBASTIAN OSORIO GARCÍA aprueba el presente trabajo y solicita la autorización del mismo.*

"ID Y ENSEÑAD A TODOS"

Msc. Carlos Gustavo Alonzo
Director
Escuela de Ingeniería en Ciencias y Sistemas

Guatemala, 30 de noviembre de 2021

USAC
TRICENTENARIA
Universidad de San Carlos de Guatemala

Decanato
Facultad de Ingeniería
24189101- 24189102
secretariadecanato@ingenieria.usac.edu.gt

DTG. 737.2021

La Decana de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Ingeniería en Ciencias y Sistemas, al Trabajo de Graduación titulado: **PROPUESTA DE METODOLOGÍA PARA LA IMPLEMENTACIÓN DE CIENCIA DE DATOS EN LA TOMA DE DECISIONES GERENCIALES EN GUATEMALA,** presentado por el estudiante universitario: **Diego Ernesto Sebastian Osorio García,** y después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

Inga. Anabela Cordova Estrada
Decana

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANO
FACULTAD DE INGENIERIA

Guatemala, noviembre de 2021

AACE/cc

ACTO QUE DEDICO A:

Dios	Por guiarme durante todo el camino, por ser mi fortaleza en los momentos difíciles, por la vida y darme la oportunidad de alcanzar esta meta profesional.
Mis padres	Alma García y Eliseo Osorio. Por apoyarme en todo momento, por los valores que me han inculcado y sobre todo por ser parte esencial en mi formación como persona y profesional.
Alanise Espinoza	Por su apoyo incondicional, paciencia y cariño.
Mis amigos	Por ser parte fundamental en mi vida y contribuir en este logro, y por acompañarme en esta etapa de mi vida.

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	Por ser mi *Alma Máter* de estudios y darme la oportunidad de forjar mi educación superior
Facultad de Ingeniería	Por ser una importante influencia en mi carrera, por haberme brindado tantos conocimientos y experiencias.
Mis amigos de la Facultad	Yimmi Pernillo, Roberto Cux, Jorge Delgado, Oscar Chacón, Kevin Anaya y Omar Lantán. Por compartir momentos inolvidables y el apoyo a lo largo de nuestra formación profesional.
Ing. Oscar Paz	Por su tiempo y ayuda en la etapa final de mi carrera.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
%	Porcentaje

GLOSARIO

Algoritmo Según la RAE es Conjunto ordenado y finito de operaciones que permite hallar la solución a un problema.

Cloud computing Según el diccionario panhispánico del español jurídico es un servicio digital que hace posible el acceso a un conjunto modulable y elástico de recursos informáticos que se pueden compartir.

Clustering Proceso que tiene el objetivo de lograr el agrupamiento de conjuntos de datos no etiquetados para construir subconjuntos de información conocida.

Data Lake Repositorio de datos centralizado que contiene datos de varias fuentes sin procesar.

GB Unidad de almacenamiento de información con capacidad de mil millones de bytes.

IDE Entorno de desarrollo integrado que sirve para el desarrollo y diseño de aplicaciones combinando herramientas de desarrollador en una interfaz gráfica.

IoT El Internet de las Cosas (Internet of Things en inglés), es la acción de interconectar a través de

	Internet objetos cotidianos que rodean a una persona.
KPI	Son indicadores clave de rendimiento utilizados para la evaluación de las acciones que contribuyen en el éxito de los objetivos.
Proceso *batch*	el procesamiento *batch* o por lotes es el proceso en el cual una computadora realiza lotes de trabajo de manera simultánea y en orden.
Python	Python es un lenguaje de programación interpretado, soporta la programación orientada a objetos, imperativa y funcional, es un lenguaje dinámico y multiplataforma.
R	Lenguaje de programación interpretado orientado a la computación estadística y gráficos.
Red Social	Según el diccionario de Oxford, red social es una página web en la que los internautas intercambian información personal y contenidos multimedia de modo que crean una comunidad de amigos virtual e interactiva.
Software	Según la RAE, es un conjunto de programas, instrucciones y reglas informáticas que permiten ejecutar ciertas tareas en una computadora.

TB	Unidad de almacenamiento de información con capacidad de un billón de bytes.

RESUMEN

Con el aumento de la aplicación de sistemas tecnológicos de información en las empresas se han incrementado los datos generados y almacenados.

Es de vital importancia su análisis para obtener información de valor que ayude a mejorar procesos o productos que ofrezca la organización, por lo que la alta gerencia debe tener una adecuada comprensión del uso que se le puede dar a los datos para que sean un activo más en la organización, esto es posible mediante la aplicación de la ciencia de datos.

La ciencia de datos es un concepto muy complejo que ha ganado popularidad en los últimos años, por lo que para su compresión se ofrece descripciones de los diferentes factores que se deben de tomar en cuenta para su implementación.

Reconociendo el alto impacto que tiene aplicar ciencia de datos dentro de las empresas, se propone una metodología de implementación de ciencia de datos para que las organizaciones guatemaltecas puedan obtener ventajas competitivas en el mercado, ofreciendo un plan estratégico de cinco pasos y una propuesta de arquitectura para su implementación, aprovechando la potencia y la versatilidad que tienen hoy en día las tecnologías basadas en la nube.

OBJETIVOS

General

Proporcionar una metodología de implementación de la ciencia de datos que sirva de guía técnica a las empresas guatemaltecas para utilizar los beneficios del análisis de datos en sus decisiones gerenciales.

Específicos

1. Dar a conocer los beneficios que da el análisis de datos dentro de las decisiones empresariales en Guatemala.

2. Proponer una metodología de implementación de la ciencia de datos para la toma de decisiones gerenciales.

3. Establecer un plan estratégico para aprovechar el valor de los datos generados por una empresa.

INTRODUCCIÓN

La cantidad de datos generada hoy en día por las empresas ha aumentado de manera considerable en los últimos años, con la llegada de la era digital nunca fue tan sencillo almacenarlos y poder procesarlos para beneficio del negocio.

La ciencia de datos se encarga del estudio multidisciplinar de los datos generados agregando valor a toda la información almacenada y así poder crear modelos de datos que ayuden a la toma de decisiones ejecutivas para mejorar los procesos y servicios ofrecidos por una organización.

En la aplicación de la ciencia de datos se debe de tener en cuenta que el punto de partida es la comprensión del negocio y como los ejecutivos perciben el análisis de datos. El segundo punto importante es la digitalización de los datos provenientes de cualquier fuente y saber la cantidad de datos que genera la organización. El último punto en tener en cuenta sobre la ciencia de datos es el uso de técnicas estadísticas para el modelado y análisis de datos además de la utilización de algoritmos computacionales tal como los de la inteligencia artificial capaces de procesar y analizar datos.

En Guatemala las empresas tienen una gran oportunidad para obtener ventajas competitivas a partir de decisiones basadas en datos, pero esto es posible dependiendo de su nivel de madurez digital, antes de comenzar con la aplicación de la ciencia de datos se debe de tener un nivel consciente de madurez digital para que la metodología que se propone sea fácil de implementar.

La propuesta de metodología de implementación de ciencia de datos se basa en un plan cíclico de cinco pasos en donde los datos de salida del modelo pueden ser utilizados como insumo de entrada para obtener una mejora continua del análisis y así la alta gerencia pueda mejorar sus estrategias para alcanzar sus objetivos.

Luego de establecer la metodología de ciencia de datos dentro de la empresa es recomendable tener una arquitectura basada en servicios de computación en la nube para aprovechar el escalamiento y la baja inversión de implementación lo cual significa una ventaja dentro del mercado guatemalteco.

1. CIENCIA DE DATOS EN LAS EMPRESAS

En los últimos años se ha tenido un crecimiento exponencial de los datos debido a que su costo de almacenamiento se ha reducido y que ya no son difíciles de conseguir gracias a la tecnología móvil, redes sociales y IoT. Pero no basta solo con tener los datos hace falta hacer buenas preguntas a los datos y de allí nace el concepto de Ciencia de datos (*Data Science*). “La ciencia de datos es el estudio de dónde proviene la información, qué representa y cómo se puede convertir en un recurso valioso para la creación de estrategias empresariales y de Tecnologías de la Información”[1]. La ciencia de datos emplea matemáticas, estadística, informática y administración de negocios para poder encontrar valor a los datos.

La importancia del uso de las técnicas de análisis de datos está en cómo dar valor a la múltiple cantidad de información almacenada, para que así una organización pueda tomar mejores decisiones a futuro basándose en datos del pasado.

Las empresas tecnológicas nativas tal como: Google, Facebook, Netflix, entre otros. Utilizan los datos generados por sus usuarios para analizarlos y así tomar decisiones personalizadas para cada una de las personas que utilizan sus servicios garantizando su satisfacción con el producto.

[1] ROUSE, Margaret. *¿Qué es ciencia de datos?* https://searchdatacenter.techtarget.com/es/definicion/Ciencia-de-datos. Consulta: 28 de febrero de 2021.

Acciones como las de estas empresas nativas digitales se pueden llevar a cabo en cualquier empresa en donde la mayoría de los gerentes no sabe que su organización genera múltiples datos provenientes de algún sistema informático, redes sociales o bien sus propios clientes. Esta información puede ayudar a tomar mejores decisiones gerenciales y llegar a un nivel más competitivo dentro del mercado.

1.1. Los datos en los negocios

"Con las grandes cantidades de datos disponibles, las empresas de casi todas las industrias se centran en explotar los datos para obtener una ventaja competitiva"[2]. Los datos están en todos lados y pueden provenir de diversas fuentes en las empresas desde una crítica de un cliente hasta las ventas de un producto bajo demanda.

Las fuentes de datos en un negocio suelen ser recursos internos tal como información transaccional, correos electrónicos e historiales de negociaciones, pero en los últimos años se han generado muchos datos de valor a través de fuentes externas como lo son las redes sociales en las cuales se puede saber lo que el cliente quiere y hacer del producto algo más atractivo al consumidor.

1.1.1. Percepción del análisis de datos por los directivos

El poder analizar el mercado y predecir el comportamiento de un producto o servicio en el futuro es clave para cualquier directivo, es por ello por lo que hoy en día se dice que si una empresa no aplica la ciencia de datos puede desaparecer. Entonces para los directivos el análisis de datos representa una

[2] PROVOST, Foster. y FAWCETT, Tom. *Data science for business. Sebastopol, CA: O'Reilly & Associates*, p. 1.

oportunidad para mejorar el negocio mediante la toma de decisiones cada vez más acertadas en el ámbito gerencial.

Aunque el análisis de datos es una herramienta muy poderosa en los negocios el concepto de la ciencia de datos no tiene una comprensión total, por lo que es necesario realizar una revisión completa del tema y su implementación en las empresas.

1.1.2. ¿Cómo los datos pueden mejorar a las empresas?

Los datos han ayudado a las empresas desde siempre, lo que ha cambiado hoy en día es la forma de almacenarlos y analizarlos para poder mejorar la toma de decisiones, que en general es el interés primordial de un negocio. Si bien existen muchos factores que pueden influir en la toma de decisiones gerenciales, no existe nada tan persuasivo como datos que demuestren que la organización va en la dirección correcta aprovechando las oportunidades que se presentan.

Las empresas pueden utilizar los datos para evaluar y mejorar ámbitos del negocio como: comprensión del rendimiento de un producto, servicio al cliente, búsqueda de nuevos clientes, campañas de publicidad, entre otros.

1.1.3. Implementación de la ciencia de datos

La implementación de la ciencia de datos ha sido utilizada por múltiples organizaciones en el mundo, obteniendo resultados positivos y negativos en la implementación. En todos los casos se observa que la ciencia de datos es un agente clave en los negocios, por lo que es significativo examinar experiencias desde diversos ámbitos partiendo de lo general a lo específico.

1.1.3.1. Ámbito mundial

Alrededor del mundo muchas compañías utilizan la ciencia de datos para hacer más sencillos sus procesos. El campo del análisis de datos ha crecido de tal manera que ya existen universidades en Alemania, Australia, Nueva Zelanda y Canadá que ofrecen grados y maestrías en ciencia de datos. Al ser relativamente nuevas las carreras en el campo del análisis de datos existe una gran brecha entre las personas que están capacitadas y las que no están capacitadas. Esta ampliación de la brecha de habilidades es diferente en las geografías.

> "El 70 % de las empresas estadounidenses reconocieron la brecha de habilidades, mientras que en India la brecha es del 64 %, en el Reino Unido es del 57 %, en Alemania es del 55 % y en Francia es del 52 %"[3].

Los países que tienen la capacidad de identificar la diferencia de habilidades entre las personas capacitadas y las que no lo están en el ámbito de la ciencia de datos, son los países que tienen más compañías que emplean sus datos de manera estratégica en la toma de decisiones o para optimizar algún proceso.

Un ejemplo de una compañía que a nivel mundial ha implementado la ciencia de datos para obtener mejor rendimiento en el mercado es Walmart, esta cadena de supermercados utiliza la información que obtiene de sus clientes para abastecer sus tiendas según eventos que vayan a ocurrir en una zona específica.

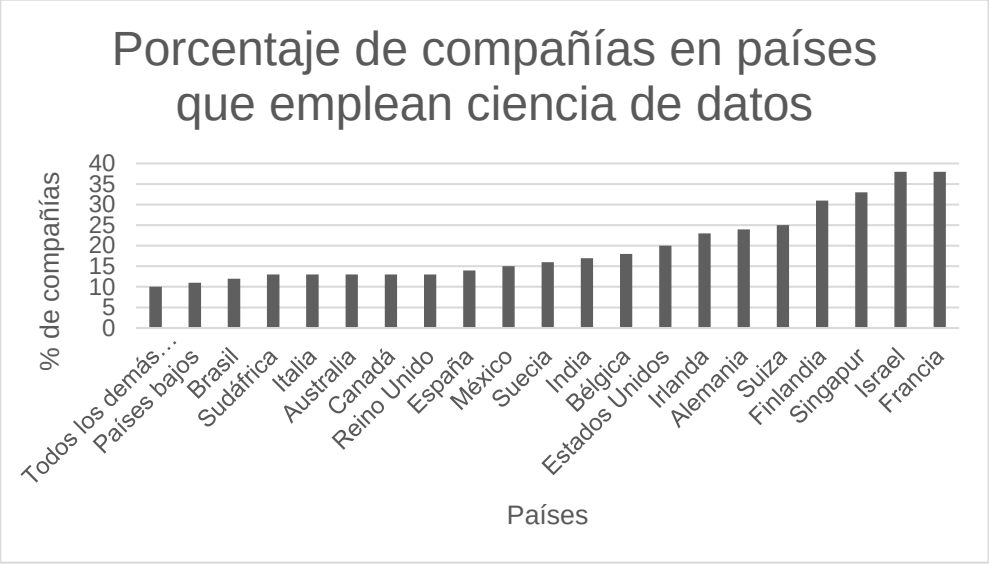
[3] SOME, Kamalika. *Analytics insight*. https://www.analyticsinsight.net/countries-which-hold-the-greatest-opportunities-for-data-scientists. Consulta: 3 de marzo de 2021.

En 2004 pudo predecir los productos que más se comprarían en la costa de Florida antes de que el huracán francés tocara tierra.

El ejemplo anterior demuestra la ventaja competitiva que da sobre otras empresas el hecho de poder predecir y adelantarse a lo que va a acontecer.

La ciencia de datos no solo se puede aplicar para problemas de empresas, sino también para resolver problemas de todo tipo tal es el caso de la implementación de un buen análisis de datos del departamento de policía de Santa Cruz en California que mediante inteligencia artificial calculan la probabilidad de que determinada zona tenga un crimen en un día concreto basándose en datos de delitos pasados y otros datos acontecimientos de relevancia en el lugar.

Figura 1. **Porcentaje de compañías en países que emplean ciencia de datos ciencia de datos**



Fuente: elaboración propia.

1.1.3.2. Ámbito latinoamericano

En Latinoamérica es menor el uso de la ciencia de datos lo que significa un gran campo de oportunidad de implementación, sin embargo, países como Brasil y México llevan una gran ventaja en cuanto la utilización del análisis de datos, pero existen registros de que Colombia, Chile, Perú y Argentina están empezando a utilizar el análisis de datos como herramienta funcional dentro del marco de toma de decisiones gerenciales especialmente en el área de publicidad y economía.

En la región latinoamericana son pocas las personas que cuentan con un grado o maestría en ciencia de datos lo cual amplía el panorama para inversión y así descubrir nuevos talentos y mejorar la competitividad del mercado mediante la tecnología. Algunas instituciones como la Sociedad de científicos de datos de México (SoCieDat), promueven el desarrollo de la ciencia de datos en México y son un referente para otras organizaciones en América Latina.

1.1.3.3. Ámbito guatemalteco

“Guatemala ocupa el lugar 107 en el índice Global de Innovación publicado en el año 2019, los niveles de competitividad son bajos y por lo tanto la productividad es baja”[4]. En consecuencia, surge la importancia de evaluar si la ciencia de datos y el análisis de la información puede ser una herramienta que ayude a las empresas a ser más productivas.

[4] ORANTES KESTLER, Alejandro. *La inteligencia artificial y las oportunidades para la empresa en Guatemala. Revista Ciencia Multidisciplinaria. CUNORI.* p. 141–146. https://doi.org/10.36314/cunori.v4i2.138. Consulta: 6 de marzo de 2021.

Las empresas guatemaltecas tienen una gran oportunidad de hacer de sus datos una herramienta muy poderosa para agregar valor a sus negocios. Instituciones en Guatemala como Agexport reconocen que el uso de técnicas de ciencia de datos puede llegar a tener un alto impacto en las organizaciones si se aprovecha la información para optimizar procesos o bien predecir basándose en la estadística, comportamientos futuros.

En Guatemala se genera mucha información no solo del sector empresarial, sino que del sector gobierno también, el reto está en poder aprovechar esa información y resolver problemas que se presenten según las necesidades de la organización o país.

1.2. Factores que tomar en cuenta

La ciencia de datos involucra principios, técnicas y procesos que facilitan la toma de decisiones en diferentes ámbitos de una empresa. Es necesario que los gerentes tengan el conocimiento necesario de todas las herramientas tecnológicas y estadísticas que pueden representar un punto clave en el análisis de datos que está generando su organización. A continuación, las más significativas.

1.2.1. Digitalización de los datos

Para poder hacer un buen análisis de datos utilizando las herramientas tecnológicas disponibles se necesitan digitalizar los datos. Hoy en día se puede transformar casi cualquier tipo de documento que utilicen las empresas a un documento digital para almacenarlo y luego obtener datos de valor. La digitalización de los datos reduce gastos, aumenta la productividad de los empleados y hace a la organización más eficiente.

Otros datos que se han digitalizado son los datos externos al negocio que son generados y capturados en todo momento como: datos demográficos de los clientes que interactúan en redes sociales, videoconferencias, cámaras de vigilancia, entre otros.

1.2.2. Big Data

Para el procesamiento de datos existen muchas herramientas y técnicas que no son ciencia de datos. La ingeniería y el procesamiento de datos son parte fundamental para que la ciencia de datos exista, sin embargo, no son parte de ella.

A menudo se suele confundir la ciencia de datos con las herramientas tecnológicas de procesamiento de datos, por ello es importante saber la diferencia entre ciencia de datos y procesamiento de datos. Las tecnologías de procesamiento de datos son útiles para tareas orientadas a datos que no requieren extraer conocimientos de la información sino más bien la gestión eficiente de transacciones de datos y administración de procesamiento en sistemas web.

A partir de que se requiere procesar de grandes volúmenes de datos surge Big Data. Según la definición de Gartner: “Big Data son activos de información de gran volumen, alta velocidad y/o gran variedad que exigen formas rentables e innovadoras de procesamiento de información que permitan una mejor compresión, toma de decisiones y automatización de procesos”[5].

[5] Big Data. *Gartner glossary.* https://www.gartner.com/en/information-technology/glossary/big-data. Consulta: 7 de marzo de 2021.

Por lo general Big Data se utiliza para implementar técnicas de minería de datos, sin embargo, su uso habitual en las empresas es del procesamiento de datos en apoyo a las técnicas de minería de datos o de ciencia de datos.

1.2.3. Inteligencia artificial

"Inteligencia artificial (IA), es la capacidad de una computadora o robot controlado por computadora para realizar tareas asociadas con seres inteligentes"[6]. La ciencia de datos hace uso de la inteligencia artificial como una herramienta dentro de sus operaciones de análisis de datos, especialmente en la parte del análisis predictivo en donde se utilizan algoritmos específicos que se entrenan para poder obtener un mejor resultado basándose en los datos.

La inteligencia artificial tiene varias ramas muy útiles dentro de la analítica de datos como Machine Learning que se basa en el entrenamiento de algoritmos mediante datos anteriores para que estos aprendan, después de un tiempo los algoritmos mejoran su desempeño logrando el objetivo de automatizar procesos.

En muchas ocasiones se utilizan redes neuronales como modelo computacional especializado en predecir respuestas basadas en datos históricos.

Si el resultado de la predicción no es el esperado se repite el proceso hasta entrenar completamente el modelo y disminuir los niveles de error.

[6] COPELAND, Jack. *artificial Intelligence / Definition, examples, and applications. Encyclopedia Britannica*. https://www.britannica.com/technology/artificial-intelligence. Consulta: 7 de marzo de 2021.

1.2.4. Estadística

La estadística es una herramienta muy útil dentro de la ciencia de datos, hay conceptos estadísticos que se deben de tener claros antes de poder hacer un análisis predictivo con recursos tecnológicos. Toda la base estadística es necesaria para poder interpretar los resultados que da un modelo predictivo, y se definirán conceptos básicos que se requieren dentro de la ciencia de datos.

1.2.4.1. Estadísticos

"Un estadístico es cualquier función real medible de la muestra de una variable aleatoria"[7]. Los estadísticos son medidas descriptivas que ayudan a tener cierta idea de cómo se distribuyen los datos que se están analizando. Existen cuatro tipos de estadísticos, estos son:

- Estadísticos de posición: Separan un conjunto ordenado de datos en grupos con la misma cantidad de unidades. Los más utilizados son: Deciles, cuartiles, percentiles y cuantiles.

- Estadísticos de centralización: Muestran valores representativos sobre como parece que se sitúan los datos. Estos son: Media, mediana y moda.

- Estadísticos de dispersión: Indican mayor o menor variabilidad de los valores respecto a las medidas de centralización. Entre ellos están: Varianza, desviación típica, coeficiente de variación, rango muestral.

[7] LÓPEZ, José. *Estadístico. Economipedia*. https://economipedia.com/definiciones/estadistico.html. Consulta: 9 de marzo de 2021.

- Estadísticos de forma: Aproximan una idea de cómo los datos se distribuyen. Estos son: Asimetría, medida de apuntamiento o curtosis.

1.2.4.2. Muestreo aleatorio

Cuando se tiene un conjunto de datos muy grande, normalmente denominado población, resulta conveniente tomar una muestra aleatoria de ese conjunto creando así un subconjunto más fácil de analizar, a esto se le conoce como muestreo aleatorio. Es importante definir la muestra con el menor sesgo posible para obtener resultados más certeros sobre el conjunto de datos inicial. Con ayuda de los estadísticos se puede llegar a inferir lo que se quiere estudiar del conjunto principal de datos a pesar de que dichas medidas descriptivas hayan sido aplicadas al subconjunto de la población. El procedimiento de seleccionar una muestra debe cumplir dos propiedades esenciales:

- Todas las muestras del mismo tamaño son equiprobables.

- Todos los elementos de la población tienen la misma probabilidad de ser elegidos.

1.2.4.3. Función de distribución de probabilidad

"Una distribución de probabilidad es una función estadística que describe todos los posibles valores y probabilidades que puede tomar una variable aleatoria dentro de un rango determinado"[8]. La función de probabilidad depende

[8] HAYES Adam. *Probability distribution. https://www.investopedia.com/terms/p/probabilitydistribution.asp*. Consulta: 18 de marzo de 2021.

de un numero de factores determinados que puede ser cualquier estadístico como: La media, desviación estándar, curtosis y asimetría.

Dentro de las funciones de probabilidad la más común es la función normal, aunque el proceso de determinar que función de probabilidad se ajuste más a los datos lo dictará el fenómeno que se estudie. Las funciones de probabilidad se pueden usar en cualquier ámbito en donde se necesite determinar el patrón que siguen ciertos datos que se estén analizando.

1.2.4.4. Contrastes de hipótesis

El mayor problema de la inferencia estadística es la toma de decisiones. El contraste de hipótesis es una prueba estadística que indica el proceso de decidir si una proposición de una población es aceptada o no, a esta proposición se le conoce como hipótesis estadística. Entonces una hipótesis estadística, es una proposición sobre la función de probabilidad de una variable aleatoria. Esta proposición puede ser acerca de la forma de la distribución de probabilidad o de los valores de los parámetros que definan la función.

El contraste de hipótesis en estadística tiene su fundamento en la información brindada por la muestra. Esto hace que si se rechaza la hipótesis se indica que los datos de la muestra determinan que es falsa y si se acepta únicamente significa que no se rechaza.

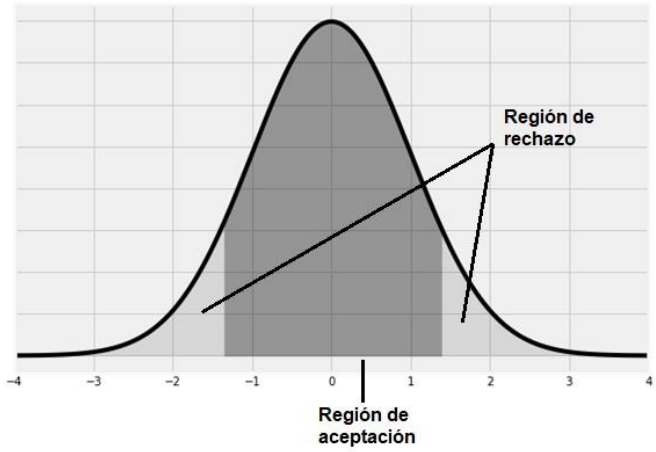
El contraste de hipótesis consiste en estudiar dos tipos de hipótesis: Hipótesis nula (H_0), que es la suposición que se hace acerca de los datos y a al caso contrario, a la hipótesis que no está contemplada dentro de la hipótesis nula se le conoce como hipótesis alternativa (H_1). De esta manera se pueden

dividir los resultados de la muestra en dos regiones; una región de aceptación y una región de rechazo.

Cuando se platea una hipótesis, esta puede ser simple o compuesta. La hipótesis es simple si se especifica el valor del parámetro, en cambio, una hipótesis es compuesta si tiene dos o más valores del parámetro.

Al momento de plantear los dos tipos de hipótesis es común hacerlo de manera que ambas sean complementarias, dependiendo del tipo de estadístico de contraste que se necesite.

Figura 2. **Regiones de aceptación y rechazo de hipótesis en una curva normal**



Fuente: elaboración propia, empleando Python 3.

Al momento de hacer un contraste de hipótesis se debe de elegir un estadístico de contraste que se utilizará para tomar una decisión. El estadístico es una variable aleatoria con una distribución determinada que dará las

probabilidades asociadas a un valor o a un conjunto de valores del estadístico de contraste.

Al realizar un contraste de hipótesis, hay que tener en cuenta que cuando se acepta o rechaza una proposición puede que se cometa un error. Cuando se rechaza la hipótesis nula, pero esta es correcta se conoce como error tipo 1. Cuando se comete el error de decisión cuando se acepta la hipótesis nula cuando esta es falsa se conoce como error tipo 2.

1.2.4.5. Test de chi cuadrado

"El estadístico ji-cuadrado (o chi cuadrado), que tiene distribución de probabilidad del mismo nombre, sirve para someter a prueba hipótesis referidas a distribuciones de frecuencias"[9]. Usualmente la prueba de chi cuadrado se utiliza para tres cosas distintas: una es para utilizarlo como elemento de bondad de ajuste observando si los datos que se están analizando tienen cierta tendencia hacia un extremo comparados con los datos esperados; el siguiente uso de la prueba de chi cuadrado es la prueba de homogeneidad y la prueba de independencia.

Para realizar la prueba de chi cuadrado se dispone de una tabla de frecuencias con datos. Para cada uno de los valores o intervalo de estos se indica la frecuencia absoluta observada (O_i) y también se calcula la frecuencia absoluta esperada (E_i). La prueba se basa en la sumatoria de las diferencias entre O_i y E_i.

[9] QUEVEDO RICARDI, Fernando. *The chi-square*. *Medwave*. https://www.medwave.cl/link.cgi/Medwave/Series/MBE04/5266. Consulta: 12 de marzo de 2021.

Si se llega a una concordancia perfecta entre las frecuencias esperadas y observadas la prueba tomará un valor de cero; en caso contrario, si existe una gran discrepancia entre las dos frecuencias el estadístico tomará un valor muy grande y en consecuencia se rechazará la hipótesis nula.

2. CONCEPTOS TÉCNICOS EN LA CIENCIA DE DATOS

El concepto de ciencia de datos es bastante nuevo, por ello no es tan conocido dentro del sector empresarial, sin embargo, desde 2015 se ha visto un crecimiento importante en cuanto a los empleos en los que se solicitan expertos en ciencia de datos.

La ventaja competitiva que otorga a las empresas el analizar los datos de manera adecuada y resolver problemas complejos con los resultados de los análisis es una de las razones por lo que las organizaciones contratan a los científicos de datos. La tarea de resolver un problema de análisis de datos y predicción con base en los datos no es sencilla, se necesita tener conocimiento de varias herramientas que ayudarán a generar ideas, tomar decisiones y realizar estrategias de negocio viables.

2.1. Definición de ciencia de datos

"Ciencia de datos es el descubrimiento del conocimiento profundo que puede ser obtenido a través de exploración e inferencia de datos"[10]. El concepto de ciencia de datos es bastante complejo debido a que es un campo interdisciplinar que junta la matemática y estadística, las ciencias de la computación y el análisis de negocios.

[10] ZHANG, Arthur. *Data analytics practical guide to leveraging the power of algorithms, data science, data mining, statistics, big data, and predictive analysis to improve business, work, and life*. Consulta: 12 de marzo de 2021.

En donde la parte de la computación se centra en utilizar una serie de algoritmos y lenguajes de programación, la parte de matemática se enfoca en el uso de técnicas estadísticas y por último la parte del análisis de negocios es la parte fundamental que le da valor a los datos dentro del contexto empresarial.

2.1.1. Comprensión del negocio

Para que la ciencia de datos funcione dentro del contexto empresarial se debe saber para qué fin se están analizando los datos. En el análisis predictivo de datos para poder obtener resultados que aporten valor se necesita generar preguntas importantes acerca de los datos.

La comprensión del negocio es un buen punto de partida para poder saber qué rumbo debe de llevar el análisis y que técnicas y herramientas poder utilizar. Un científico de datos se convierte dentro de la organización como un consultor de estrategias con habilidades para poder manejar problemas empresariales y algorítmicos que da un valor importante al negocio.

2.1.2. Comprensión de los datos

Una aplicación correcta de la ciencia de datos incluye una buena administración de los datos con los que se va a trabajar y tratarlos como el recurso valioso que son. Para poder comprender los datos se debe saber de dónde vienen los datos.

Muchas organizaciones obtienen los datos a través de fuentes internas como lo son transacciones, datos de registros históricos, correos institucionales, entre otros. Pero existen otras fuentes de datos externas como lo son las redes sociales.

La forma en que se obtienen los datos también influye en el proceso de conocimiento de los datos, por ejemplo, hay datos que se recopilan de forma pasiva por medio del método de observación sin intentar controlar variables involucradas, tal es el caso de los datos que vienen a través de los comentarios de los clientes. De hecho, la mayoría de los datos que obtienen las empresas se realiza mediante métodos de observación. Existe otra técnica de recopilación que se basa en diseñar y realizar un experimento en el que se controlan variables y se estudian otras, esta técnica se conoce como recopilación de datos experimental.

En las empresas farmacéuticas se suele frecuentar este último tipo de técnica de recopilación. Dentro de la comprensión de los datos se debe de considerar:

- Naturaleza de los datos: Se debe de considerar de dónde vienen los datos y cuál es su forma, muchas veces al recolectar datos este proceso puede ser sesgado o disperso y estos factores pueden afectar al resultado final. Es importante reconocer la naturaleza de los datos para corregirlos antes de empezar el análisis.

- Tiempo requerido: Es una regla general que la mayor parte del tiempo se pasará en entender los datos y hacer arreglos para que estos tengan el valor necesario para la organización.

- Costo de adquisición: Un componente muy importante es el costo que requiere obtener los datos, especialmente para las empresas en donde necesitan que su inversión retorne con algún valor agregado para la institución.

2.1.3. Preparación de los datos

Una vez se ha comprendido lo que se quiere conseguir con los datos y que datos se necesitan tener para poder realizar un análisis, se debe realizar una fase previa a la elaboración de un modelo para poder predecir y mejorar procesos de una empresa. En la etapa de preparación de los datos se suele pasar mucho tiempo depurando para que en procesos posteriores no existan problemas.

2.1.3.1. *Dataset*

"*Dataset* es la integración de datos heterogéneos en diversos formatos y de diversas comunidades…"[11]. Los *dataset* se suelen realizar al coleccionar una serie de datos con contenido relevante, separados por algún delimitador, que se utilizarán para un propósito específico. Dentro de la ciencia de datos un conjunto de datos es la fuente de información brindada por alguna organización o recolectada de manera que se agrupan los datos para su posterior análisis, normalmente bridado por medio de un archivo digital.

2.1.3.2. *Data frame*

Dentro del análisis de datos un *data frame* es una estructura de datos que se asemeja a una tabla con columnas y filas que contienen un conjunto de valores para cada una de las columnas. Normalmente un *data frame* se utiliza para el análisis estadístico dentro de los lenguajes de programación como R y Python. Las características comunes de un *data frame* son: cada columna debe

[11] RENEAR, Allen., SACCHI, Simone. & WICKETT, Karen. *Definitions of dataset in the scientific and technical literature. Proceedings of the American Society for Information Science and Technology*. https://doi.org/10.1002/meet.14504701240. Consulta: 14 de marzo de 2021.

de contener el mismo número de elementos y los nombres de las columnas no pueden ser vacíos.

2.1.3.3. Carga de datos

Una de las etapas más importantes dentro de la preparación de los datos es la carga. En este proceso se hace accesible la información al usuario después de ser recolectada y almacenada en algún *dataset*. Para realizar una carga de datos al sistema existen varios métodos, estos métodos pueden ser: basados en la nube, en donde se utiliza una herramienta que utiliza la velocidad y escalabilidad de la nube para procesar la información de manera rápida y procesos *batch*, que cargan grandes volúmenes de datos grandes por lo general de manera periódica para ahorrar tiempo de procesamiento.

2.1.3.4. Limpieza de datos

Un concepto clave para un buen análisis de datos es la limpieza de datos (*data cleaning)*. Según la naturaleza de los datos estos pueden venir de distintas fuentes y estar almacenados en distintos formatos. Esto da lugar a que la estructura de los datos sea diferente en cada caso.

El proceso de limpieza consiste en empezar a tratar los datos y cargarlos al IDE de trabajo para poder obtener información básica como: el número de columnas que tiene el archivo, un resumen de los estadísticos básicos, verificar si existen espacios en blanco, comprobar si se pueden eliminar o bien inferir los espacios vacíos o el tipo de dato que maneje cada columna.

El resultado de un buen proceso de limpieza es obtener un conjunto de datos fácil de manejar en las etapas posteriores, sencillo de entender y fácil de visualizar mediante gráficos.

2.1.3.5. Variables indicadoras

Una variable indicadora o comúnmente conocida como variable *dummy.* Es un tipo de variable artificial que se crea representando subgrupos de una categórica de manera numérica. Normalmente se utiliza este tipo de variables para el análisis de regresión debido a que es muy útil poder escribir una única ecuación que represente múltiples grupos en lugar de tener que escribir ecuaciones separadas para cada subgrupo. Dentro del contexto de una data *frame* se puede definir una variable indicadora como una columna más que determina a la categoría que se quiere dividir.

2.1.4. Operaciones con los datos

Existen múltiples operaciones que permiten manejar los datos antes de realizar un modelo que ayude al análisis de datos. Estas operaciones permiten realizar un análisis exploratorio preliminar de la información. Representa una parte crítica del análisis de datos en la que se prepara toda la información para ser procesada posteriormente.

2.1.4.1. Semilla de generación aleatoria

Cuando se está trabajando con análisis de datos especialmente con la generación de números pseudoaleatorios, es necesario que la creación de estos números no sea aleatoria. Detrás del uso de algoritmos que generan números pseudoaleatorios existe una buena semilla determinada para que

genera la misma secuencia de números cada vez que se ejecute el algoritmo, esto da como resultado un experimento totalmente repetible.

Dentro de la ciencia de datos una semilla de generación aleatoria es útil al momento de estar analizando los datos aleatorios y este análisis genera un respuesta adecuada a lo que se esperaba, entonces se necesita poder reproducir ese experimento varias veces para poder obtener siempre los mismos valores, para ello se determina con mucho cuidado una semilla de generación aleatoria que mediante el uso de algoritmos de generación de números pseudoaleatorios y herramientas tecnológicas producirá siempre los mismos resultados.

2.1.4.2. *Data frames* indicadores

Es una estructura de datos de números aleatorios o variables indicadoras que pueden conformar un posible *dataset*. Los *data frames* indicadores o *dummies data frame* son muy útiles dentro del análisis de datos especialmente para probar que el análisis o el modelo a crear sea correcto.

A estas estructuras de datos indicadores también se les conoce como *data frames* sintéticos y generalmente dan solución a problemas que se pueden encontrar en los datos reales como: valores faltantes, correlación incorrecta entre ciertas columnas o bien predicción incorrecta entre columnas de la estructura de datos original.

2.1.4.3. Agrupación de datos por categorías

Una característica importante de la ciencia de datos y de quien la implementa es la capacidad de poder agrupar los datos de manera que estos

permitan crear resúmenes de información que pueda ser desplegada en gráficos o bien analizada mediante algún modelo. Los datos deben de ser agrupados mediante algún criterio, ya sea alguna función de probabilidad que dicte la forma en que tienen que ser agrupados según un estadístico o bien mediante el uso de categorías importantes que se defina según el negocio.

2.1.4.4. Conjunto de entrenamiento

Antes de empezar a crear un modelo se suele dividir los datos en un conjunto de entrenamiento y también un conjunto de prueba. El conjunto de entrenamiento es la parte de los datos con la cual se construye el modelo, de los datos del conjunto se calculan los parámetros y se modelizan las ecuaciones para el análisis.

La importancia de dividir el conjunto de datos inicial y crear un conjunto de entrenamiento está en que los datos que se encuentren dentro de ese conjunto serán usados para ajustar el modelo, es decir, el modelo aprenderá de esos datos y generará resultados que concuerden con la información dada.

2.1.4.5. Conjunto de prueba

La contraparte del conjunto de entrenamiento es un conjunto que sirva de validación al modelo que se creó a partir de datos anteriores, normalmente los modelos se suelen ajustar mucho a los datos que se le proveen en el momento de su creación lo cual arroja datos incorrectos en un principio. La razón de ser de un conjunto de prueba es evaluar que tan correcta es la respuesta del modelo mediante datos con los cuales no fue creado y así comprobar la eficacia del análisis que realice.

2.2. Modelado

La fase de modelado en un análisis de datos es el lugar principal en donde se aplican técnicas estadísticas y algoritmos en conjunto con herramientas tecnológicas para obtener un patrón o modelo que obtiene todas las regularidades en los datos.

2.2.1. Regresión lineal

Una de las técnicas estadísticas más utilizada dentro del mundo del análisis de datos cuando se está creando el modelo de la solución al problema es la regresión lineal. Si se tiene un conjunto de datos históricos y se quiere predecir qué pasará en el futuro con base a los datos anteriores es una muy buena opción utilizar esta técnica.

El objetivo de un modelo de regresión lineal es el análisis de la relación entre dos variables, es decir, el efecto que puede causar una sobre otra o bien para predecir los valores de una variable a partir de otra.

“Esta relación se expresa como una ecuación que predice la variable de respuesta como una función lineal de los parámetros.”[12].

Dentro de la estadística hay muchas categorías para un modelo de regresión lineal, a continuación, se definen algunos tipos fundamentales, partiendo de la base en que se necesita poder determinar la relación entre una o más variables respecto de otra.

[12] ESPINO, Carlos. *Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo*. p. 18.

2.2.1.1. Regresión lineal simple

Una regresión lineal simple se define como el modelo en el que se tiene una variable dependiente e independiente de carácter cuantitativo con la siguiente estructura: $Y=\beta_0+\beta_1 X+\epsilon$. En la expresión anterior se admiten todos los tipos de factores que influyen a la variable dependiente, donde β_0 es el corte con la variable dependiente y β_1 es la pendiente.

Por último, existen una serie de valores no controlados, ϵ, que se agrupan bajo el nombre de error aleatorio o perturbación.

2.2.1.2. Regresión lineal múltiple

Este tipo de modelo es una extensión del modelo de regresión lineal simple que consiste en considerar más de una variable independiente. La forma generalizada del modelo de regresión múltiple es: $Y=\beta_0+\beta_1 x_1+ \ldots + \beta_n x_n+ \epsilon$. Para que se pueda encontrar la función que logre mejorar la distancia entre los valores observados y los esperados para cuando se tienen n variables explicativas.

Esta tarea se resume a encontrar los coeficientes de regresión x.

2.2.1.3. Suma de los cuadrados totales

La suma de cuadrados totales (SST), por sus siglas en inglés, es la sumatoria de las diferencias al cuadrado entre la variable dependiente y su media, expresada por la siguiente expresión: $SST = \sum(y_i-\bar{y})^2$. Se puede pensar en la suma de cuadrados totales como la dispersión alrededor de la media de las variables observadas.

2.2.1.4. Suma de los cuadrados de regresión

La suma de los cuadrados de regresión (SSR), por sus siglas en inglés, es la suma de las diferencias entre el valor que se ha predicho y la media de la variable dependiente y está dada por la siguiente expresión: $\sum(\hat{y}_i - \bar{y})^2$. En donde $\hat{y}_i$ es el valor predicho por el modelo y $\bar{y}$ es la media de la muestra.

2.2.1.5. Coeficiente de determinación

El coeficiente de determinación normalmente conocido como R^2, es una medida relativa que toma valores dentro del rango de 0 a 1. Este coeficiente representa la variabilidad explicada por la regresión lineal, dividida por la variabilidad total. Está definido por la siguiente expresión: SSR/SST.

2.2.2. Regresión logística

Cuando una variable dependiente es cualitativa binaria, el análisis de regresión logística es apropiado. La regresión logística se utiliza para describir datos y explicar la relación entre una variable binaria dependiente y una o más variables independientes nominales, ordinales, de intervalo o de razón. Maalouf explica que algunas de las ventajas de la regresión logística (LR) son que, puede utilizarse lógicamente para probabilidades y extenderse a problemas de clasificación de clases múltiples.

2.2.2.1. Razón de probabilidad

Una razón de probabilidad (ODD), por sus siglas en inglés, es el cociente entre la probabilidad de un evento verdadero y un evento falso p/q. Esto quiere decir que el rango de valores permitido para una razón de probabilidad siempre

será dentro de los números reales positivos. Normalmente se interpreta como ventaja competitiva.

2.2.3. Evaluación

La fase de evaluación tiene como propósito comprobar que todos los resultados que arrojen los modelos creados sean confiables y cumplan con el objetivo de brindar valor a los datos.

Normalmente se pueden obtener resultados inmediatamente después de la fase de modelado, pero siempre es recomendable realizar pruebas sobre los modelos con conjuntos de datos de pruebas para que se pueda tener la certeza de que las afirmaciones que se obtienen sean totalmente verdaderas, garantizando los objetivos originales.

Cabe destacar que en esta fase se destaca el propósito primordial de la ciencia de datos en las empresas que es respaldar la toma de decisiones empresariales enfocado en los resultados obtenidos del análisis y construcción de modelos que resuelven el problema inicial.

2.3. Selección de la herramienta de análisis

La ciencia de datos necesita realizar procesos de extracción, manipulación, análisis y predicción respecto a los datos que se ingresen, debido a esto se necesita una herramienta lo suficientemente versátil que pueda cumplir con los requerimientos de la ciencia de datos.

Existen múltiples opciones dentro del mercado computacional algunas de estas son: Python, R, JavaScript, entre otras.

Tabla I. **Resultados de la Encuesta de ciencia de datos y aprendizaje automático 2020**

Herramienta	% de encuestados que utilizan la herramienta
Python	86,7
SQL	42,1
R	23,9
C++	21,4
Java	18,8
C	18,5
JavaScript	16,7
MATLAB	12,4
Otro	10,9
Bash	9,9

Fuente: Dilmegani. *Top data science tools: The ultimate guide*. https://research.aimultiple.com/data-science-tools/. Consulta: 25 de marzo de 2021.

2.3.1. Python

Python es un lenguaje de programación interpretado creado en 1991 que se ha utilizado para crear sitios web mediante sus numerosos marcos de trabajo como Django. En los lenguajes interpretados predomina la estrategia de escribir rápidamente pequeños programas que serán ejecutados.

Python ha desarrollado una gran y activa comunidad de análisis de datos e informática científica. En los últimos 10 años, según Mackinney Python ha pasado de ser un lenguaje informático científico de vanguardia a uno de los más importantes para la ciencia de datos, Machine Learning y desarrollo de software en general.

Una de las grandes ventajas de Python es la gran cantidad de personas que se dedican a mantener bibliotecas científicas y de análisis de datos, lo que

hace que el lenguaje se vuelva más robusto y versátil. Dentro de las bibliotecas que se utilizan en ciencia de datos están: Pandas, NumPy, Matplotlib, Scikit-learn, entre otras.

2.3.2. Pandas

Biblioteca de Python que proporciona estructuras de datos de alto nivel y funciones diseñadas para que el trabajo con datos estructurados o tabulares sea rápido, fácil y expresivo, como describe Mackinney desde su creación en 2010, ha ayudado a hacer de Python un entorno de análisis de datos potente y productivo.

2.3.3. NumPy

Biblioteca de Python que proporciona un entorno numérico muy completo desde estructuras de datos hasta algoritmos necesarios para las aplicaciones científicas y numéricas. NumPy es una abreviatura de *Numerical* Python.

2.3.4. Matplotlib

Biblioteca de Python que proporciona un marco de trabajo para producir gráficas con base en datos bidimensionales. Aunque existen una gran cantidad de librerías para generar gráficos Matplotlib es la más utilizada por la gran integración con el resto de las bibliotecas de Python.

2.3.5. Scikit-learn

Biblioteca de Python que comenzó como un proyecto de David Cournapeau en el Google Summer of Code de 2010. Esta biblioteca se ha

convertido en la herramienta primordial para Machine Learning, albergando algoritmos de clasificación, regresión, aprendizaje supervisado y no supervisado, entre otros. Scikit-learn permite a Python convertirse en un lenguaje robusto para la ciencia de datos.

2.3.6. SciPy

SciPy es un ecosistema basado en una colección de paquetes de Python para matemáticas, ingeniería y ciencia en general. SciPy junto con NumPy representan un ambiente completo y maduro para la ciencia de datos en Python.

3. MACHINE LEARNING

El aprendizaje automático es una rama de la inteligencia artificial (IA), centrada en la creación de aplicaciones que aprenden de los datos y mejoran su precisión con el tiempo sin estar programadas para hacerlo. En ciencia de datos, un algoritmo es una secuencia de pasos de procesamiento estadístico. En el aprendizaje automático, los algoritmos están entrenados para encontrar patrones y características en cantidades masivas de datos con el fin de tomar decisiones y predicciones basadas en datos nuevos.

Dentro de la ciencia de datos, específicamente en la parte del modelado de datos el entrenar un agente inteligente, mediante el uso de aprendizaje automático, para que mejore el conocimiento o el desempeño del modelo creado resulta una gran ventaja al momento de analizar los datos y hacer predicciones que servirán para la toma de decisiones del negocio.

3.1. Agrupación y clasificación

La clasificación y agrupación son dos métodos de detección de patrones muy comunes utilizados en la inteligencia artificial, específicamente en el aprendizaje automático. La clasificación se basa en clases predefinidas a las que se asignan objetos. La clasificación identifica similitudes que puedan aparecer entre objetos y los agrupa según esas cualidades en común dentro de grupos distintos de objetos, a estos grupos se les conoce como clústeres.

Dentro de la ciencia de datos el uso de algoritmos de clasificación y agrupación permite realizar una segmentación supervisada de los datos para

encontrar grupos de objetos que difieren del resto o bien clasificarlos en algún grupo ya creado con anterioridad por la organización para el análisis posterior. En un modelo bien entrenado esto permite identificar patrones para predecir valores específicos basados en los datos de una variable objetivo.

3.1.1. Matriz de distancias

Dentro de la clasificación de objetos y grupos en el aprendizaje automático existe una serie de métodos para calcular la distancia entre cada una de las observaciones. Al resultado de este cálculo se le conoce como matriz de distancias.

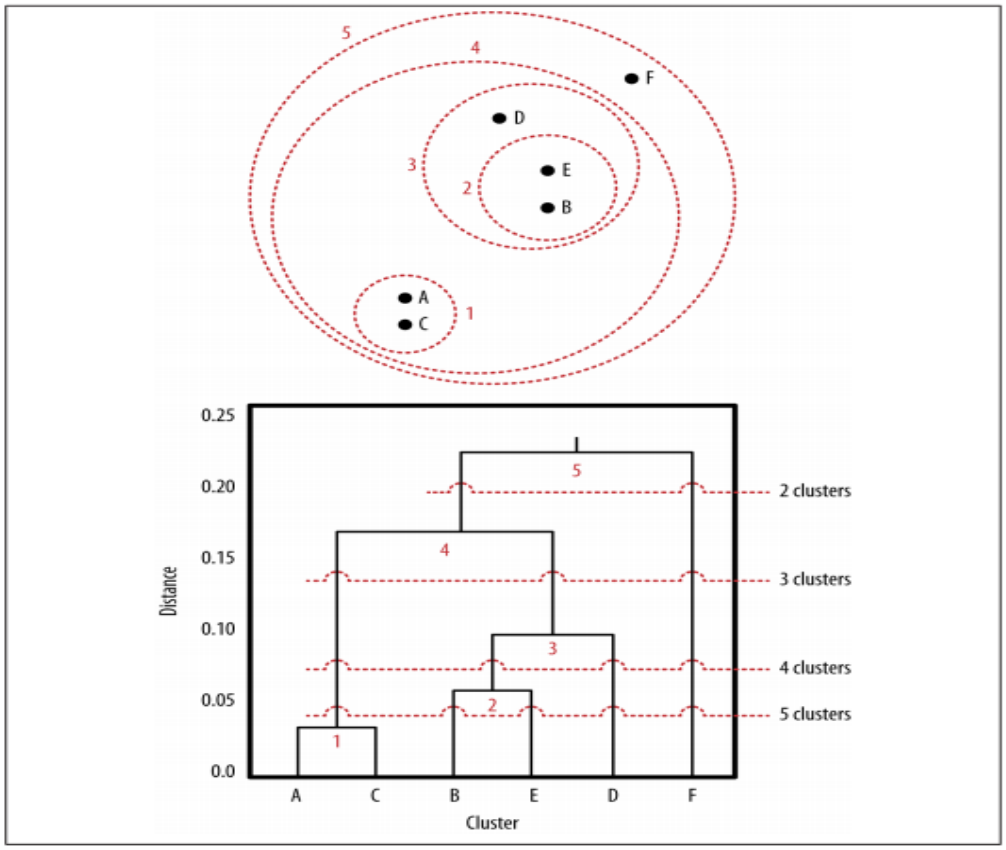
Los métodos que comúnmente se utilizan para medir las distancias entre dos grupos son: Distancia Euclidiana, distancia Manhattan y distancia de correlación de Pearson. La selección de un buen método para calcular la distancia es importante porque es muy influyente en los resultados de la agrupación. La matriz de distancias es la forma de representar y agrupar objetos según sus distancias entre sí. Por lo general se estandarizan las variables que se desean analizar mediante la matriz de distancias para hacerlas comparables en el caso que se mida en diferentes escalas.

3.1.2. Agrupación jerárquica

La agrupación jerárquica (*Hierarchical Clustering* en inglés), es un método que utiliza como base un algoritmo que agrupa los datos basándose en la distancia que existe entre cada uno de ellos, buscando que los datos que están dentro de un grupo sean los más similares entre sí.

La ventaja de una agrupación jerárquica es que da al científico de datos una perspectiva general de los datos similares agrupados antes de decidir qué datos extraer. Los agrupamientos jerárquicos generalmente se crean con cada nodo como un grupo, a medida que se va avanzando de forma iterativa en el algoritmo se van uniendo hasta lograr un único conjunto de datos. La forma en que se crean las agrupaciones es mediante la función de distancia que se elija. Muchas veces para representar los grupos es útil utilizar un dendograma que es un tipo de gráfica en forma de árbol que agrupa los datos en categorías.

Figura 3. **Agrupaciones basadas en distancias y dendograma correspondiente**



Fuente: PROVOST Foster. y FAWCETT, Tom. *Data science for business.* p. 165.

3.1.3. Agrupación completa

La agrupación completa es en donde la distancia se mide entre el par de datos más lejano entre dos grupos. Esto es equivalente a elegir el par de clústeres que tienen el diámetro más pequeño de distancia. El resultado son grupos con datos más compactos y con distancia pequeña entre los datos, pero los datos atípicos pueden ocasionar problemas en la formación de los grupos.

3.1.4. Método K-means

El método K-means se basa en un algoritmo iterativo que divide un conjunto de datos en k subconjuntos basándose en sus características. El algoritmo intenta que los datos dentro del clúster sean lo más parecido posible tratando así de separar los subconjuntos de elementos más diferentes. La forma en que el algoritmo realiza el agrupamiento de datos es minimizando la suma de las distancias que existe entre cada uno de los elementos y el centro del grupo.

El enfoque del método K-means es resolver problemas de optimización mediante la minimización de distancias entre los datos. El algoritmo permite separar los datos en grupos para visualizar la tendencia que siguen y así seleccionar un grupo para el análisis.

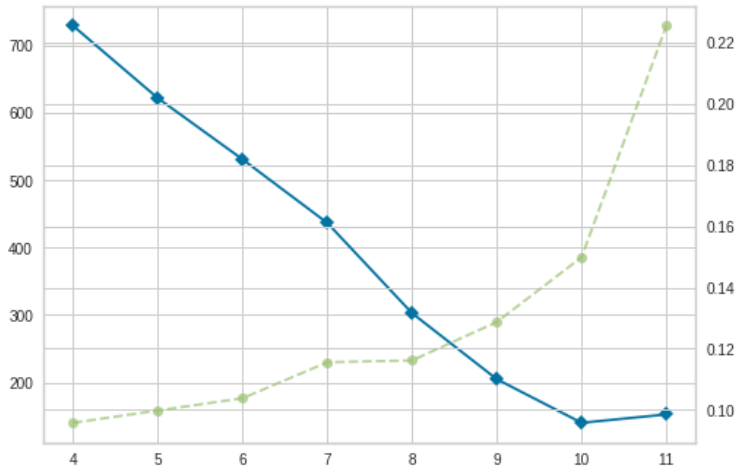
La desventaja que tiene el método es que se tiene que decir el número de subgrupos que se quiere obtener antes de empezar, esto significa un problema debido a que si se elige un k número de subgrupos muy pequeño el centroide no estará en el medio de los grupos y estos no serán representativos de los datos que se ingresen, en caso contrario si el número k es muy grande algunos

clústeres se dividirán y dos repartiendo así la información en subgrupos que probablemente no sean relevantes.

3.1.5. Método del codo

El método del codo dentro de la ciencia de datos ayuda a seleccionar el número óptimo de clústeres ajustando el modelo con un rango de valores k. La forma en que el método funciona es mediante la función de la suma de los cuadrados entre los puntos dados y los centroides. El valor k se elige justo donde la función de la suma de los cuadrados se aplana y empieza a formar un codo.

Figura 4. **Gráfica del método del codo**



Fuente: elaboración propia, empleando Python 3.

4. PROPUESTA DE METODOLOGÍA PARA LA IMPLEMENTACIÓN DE CIENCIA DE DATOS EN LA TOMA DE DECISIONES EMPRESARIALES

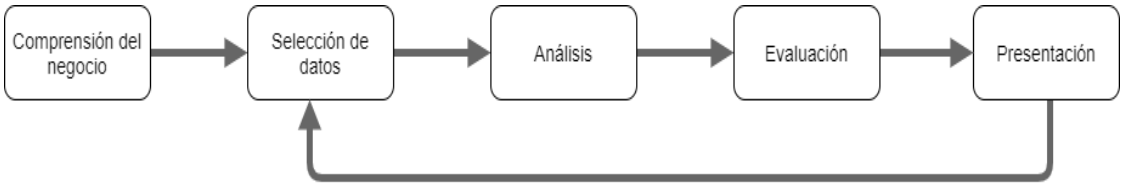
Según la encuesta nacional de empleo e ingresos realizada por el Instituto nacional de estadística en 2019 el 27,6 % de la población económicamente activa trabaja en el sector privado representado en su mayoría por pequeñas y medianas empresas.

Debido a la importancia que tienen las pymes dentro del sector económico de Guatemala, se brinda la oportunidad a los gerentes empresariales de tener una propuesta de metodología de implementación de ciencia de datos para mejorar la toma de decisiones empresariales con base en los datos generados en sus actividades, además de tomar en cuenta los recursos y capacidades tecnológicas con los que puedan contar.

La propuesta de metodología se conforma de un proceso cíclico conformado de 5 etapas que son:

- Comprensión del negocio
- Selección de datos
- Análisis
- Evaluación
- Presentación

Figura 5. **Proceso de implementación de la propuesta de metodología**



Fuente: elaboración propia.

4.1. Comprensión del negocio

El punto de partida de cualquier análisis es tener claro hacia dónde va dirigido para así seleccionar la estrategia correcta y obtener mejores resultados. Dentro del enfoque empresarial es importante definir el alcance y objetivos del análisis para poder realizar las preguntas adecuadas a los datos transformándolos en activos de valor para la organización. La comprensión del negocio es diferente para cada tipo de organización, pero, hay factores claves en las empresas ayudan a determinar por dónde empezar, estos son:

- Dirección: Dentro del sector empresarial los gerentes, directores o personas a cargo de la organización son los responsables de qué datos son relevantes como se almacenan. Normalmente en la dirección se dan las directrices que seguir durante el análisis de datos.

- Cultura organizacional: Dentro de las organizaciones existen hábitos, actitudes y tradiciones que pueden afectar a la calidad de los datos según la política que se tenga para recolectarlos y seleccionarlos. Estas políticas marcan el enfoque empresarial que lleva una organización.

4.1.1. Objetivos de la empresa

Los objetivos corporativos suelen ser la mejor herramienta para saber si se tiene o no una comprensión clara del negocio, normalmente marcan el camino de la empresa y sus aspiraciones. Para poder seleccionar los datos adecuados y generar un buen análisis siempre es recomendable revisar su definición y su estado en la organización.

La ciencia de datos utiliza los objetivos claros de la empresa para proporcionar una base semántica para la toma de decisiones creando valor añadido a los datos de la organización. Estos objetivos juegan un papel clave dentro de la estrategia a utilizar, para adquirir los datos necesarios a un costo considerable es necesario ver los datos como un objetivo más de la empresa para generarlos de calidad y producir un análisis de valor que ayude a la toma de decisiones posteriores dentro del negocio.

4.1.2. Nivel de madurez digital

La evaluación del nivel de madurez digital que tiene una empresa brinda información acerca de que tanto se aprovechan las oportunidades tecnológicas y sus beneficios dentro de la organización. Según el nivel de madurez resultante se puede determinar por dónde empezar el análisis o si realmente es viable para el negocio.

4.1.2.1. Nivel 1: En riesgo

Depende si la empresa opera sin sistemas de gestión integrados y no se tiene una estrategia de digitalización dentro de la organización. En este nivel no

se suele tener un canal digital de comunicación ni una web básica para el negocio.

4.1.2.2. Nivel 2: Consciente

La organización tiene un sistema de gestión de producción como un ERP o un CRM, también se cuenta con una estrategia de digitalización no implementada y una web de comercio electrónico.

4.1.2.3. Nivel 3: Competente

La empresa opera con una plataforma de IoT integrada con sistemas de gestión y producción, se obtienen pedidos por medio de la digitalización de la gestión y se controla la calidad por medio de dispositivos electrónicos además de que la gran mayoría de procesos productivos cuentan con procesos digitales integrados y KPIs en tiempo real.

4.1.2.4. Nivel 4: Experto

Organizaciones con modelos de negocio muy innovadores con optimización de procesos de mantenimiento con base en modelos predictivos, utilizan tecnologías de simulación, modelos virtuales y Big Data para sus datos, aparte de que se integran fácilmente con sistemas de terceros.

4.2. Selección de datos

Una vez se ha comprendido el negocio el siguiente paso es determinar qué datos seleccionar de las múltiples fuentes que tiene la organización. Esto depende del nivel de madurez digital que se tenga, pero también se determina

por hacer la pregunta correcta para que los datos tomen valor, este valor debe compensar el costo. En este punto de la metodología se obtiene la materia prima con la cual se construirá un modelo posteriormente.

4.2.1. Fuentes de datos disponibles

En las organizaciones hoy en día se puede llegar a recopilar una gran cantidad de datos de diversas fuentes, dependiendo del nivel de madurez digital estos pueden ser en documentos físicos o bien pueden ser capturados a través de medios digitales.

Para propósitos de ciencia de datos y debido al incremento de digitalización de documentos en esta metodología se consideran tres fuentes de datos importantes que se describen a continuación.

4.2.1.1. Datos transaccionales

La mayoría de las empresas tienen como principal fuente de datos transaccionales o también conocidos como datos estructurados, las transacciones realizadas por un negocio, por ejemplo: las ventas de sus productos o información de sus clientes.

Este tipo de información se suele encontrar dentro de bases de datos conectadas con alguna aplicación que la empresa utilice de manera interna.

La característica más importante de esta fuente de datos es que describen de manera precisa los intereses de la organización en las diferentes áreas que pueda cubrir.

4.2.1.2. Datos no estructurados

Los datos de clasificados en este tipo son datos cualitativos que normalmente se encuentran en imágenes, audio, vídeo, entre otros. Estos datos pueden ser importantes en el desarrollo de análisis de empresas de tráfico, seguridad o bien de digitalización que necesitan obtener respuestas a partir de datos de este tipo.

4.2.1.3. Datos externos

A pesar de que los datos externos se comprenden como toda aquella información que no está comprendida dentro de la organización, muchas veces cuesta definir qué datos son realmente externos, aunque existen algunas categorías como las redes sociales, valores de mercado, datos abiertos de estadística que pueden ser fácilmente accedidos o bien datos de paga que dan una información más precisa de un mercado.

4.2.2. Clasificación de datos de valor

El siguiente paso una vez se han definido las fuentes de datos y que toda la información que se necesite para hacer el análisis esté disponible se debe de clasificar los datos según la importancia que tengan en el problema. La forma que se recomienda en esta metodología es realizando las preguntas correctas a los datos según la comprensión del negocio y el problema que se esté resolviendo. A continuación, se definen cuatro categorías de preguntas que determinarán la clasificación de los datos.

4.2.2.1. Predicción

Uno de los principales objetivos de los negocios es poder evaluar riesgos y explorar oportunidades de mercado sin tener que arriesgar mucho, para ello se necesitan tener datos históricos del negocio y saber qué es lo que la gerencia necesita que suceda para poder seleccionar los datos correctos y generar una predicción a futuro.

4.2.2.2. Diagnóstico

A menudo las organizaciones necesitan saber cómo se encuentra el negocio, para poder seleccionar los datos correctos se necesita realizar preguntas relacionadas con la tendencia de la empresa para determinar por qué están pasando determinados eventos dentro de la organización.

4.2.2.3. Causa y efecto

Dentro de esta categoría se pueden encontrar datos que pueden marcar la diferencia en un análisis de un problema determinado, para encontrarlos se deben realizar preguntas que ayuden a comprender los factores clave y las consecuencias de la situación a analizar.

4.2.2.4. Recomendación

En el análisis de datos una de las tareas más comunes es realizar la recomendación de ciertas acciones, esto sucede cuando el análisis engloba muchas ideas generales de la organización, para poder dar una recomendación adecuada sobre una organización se debe de seleccionar los datos contestando a la pregunta: ¿Qué es lo que se necesita hacer?

4.2.3. Operaciones con los datos

Antes de empezar el análisis de datos, una vez identificados y clasificados es necesario realizar operaciones como la limpieza de datos y la carga al entorno de trabajo para que el proceso posterior sea mucho más sencillo y sea totalmente enfocado en generar modelos que permitan resolver el problema.

4.2.3.1. Carga de datos

La primera operación que se debe realizar con los datos es la carga al entorno de trabajo, para fines de esta metodología se utilizará como lenguaje de programación Python. Como se ha definido anteriormente existen múltiples fuentes de datos disponibles, una vez identificada la fuente se suele obtener los datos mediante archivos csv o bien archivos de texto plano. La información posteriormente se carga al lenguaje de programación por medio de la biblioteca pandas, almacenándolos en una estructura de datos llamada *data frame*.

Figura 6. **Código de carga de datos en Python**

```python
import pandas as pd
path =
 "https://raw.githubusercontent.com/desog/PropuestaMetodologia/main/Recursos/DataSets/incd.csv"
dataframe = pd.read_csv(path, encoding = 'latin1')
print(dataframe.head())
```

Fuente: elaboración propia, empleando Carbon.

4.2.3.2. Limpieza de datos

Una vez cargados los datos al entorno de trabajo se debe de realizar una limpieza de la información para determinar si está completa o bien si todos los

datos serán útiles para el análisis posterior. Normalmente se pueden encontrar elementos que falten debido a dos razones fundamentales: La extracción de los datos ha hecho que falten valores, o bien en el proceso de recolección no se pudieron obtener esos valores. Es importante poder determinar lo relevante que sean los valores faltantes para poder tomar la decisión de borrarlos o bien de agregarlos con técnicas estadísticas.

Figura 7. **Código de limpieza de datos en Python**

```
import pandas as pd
path =
"https://raw.githubusercontent.com/desog/PropuestaMetodologia/main/Recursos/DataSets/titanic.csv
"
dataframe = pd.read_csv(path)
pd.isnull(dataframe["body"]) #Verificación de valores faltantes
#Borrado de valores faltantes
dataframe.dropna(axis=0, how="all")
```

Fuente: elaboración propia, empleando Carbon.

4.3. Análisis

El análisis es el paso fundamental para poder llegar a tomar una buena decisión dentro de la organización, los datos y la ciencia de datos ayudarán a seleccionar la mejor opción para resolver el problema. Existen múltiples enfoques para procesar y analizar la información que ayudan a clasificar los problemas para elegir la mejor estrategia de análisis.

4.3.1. Enfoques de análisis

Cada problema es diferente y es importante poder identificar cuál es la mejor técnica para poder resolverlo, en esta metodología se definen 3 enfoques

para resolver diversos tipos de problemas que se encuentran comúnmente dentro de las empresas.

4.3.1.1. Enfoque estadístico

El enfoque estadístico se debe de utilizar cuando se está tratando de modelar un problema en el que se tiene una variable dependiente y se necesita determinar qué tan importante es para la variable independiente y la relación entre ellas. Existen múltiples métodos estadísticos que ayudan a tener mejor una mejor solución al problema, tal es el caso de la regresión lineal y polinómica que determinan mediante datos la relación entre dos o más variables. Un ejemplo de esto es cuando en las organizaciones se trata de determinar si la edad del consumidor influye en las ventas de un producto.

Otra técnica que es muy útil desde el enfoque estadístico es el contraste de hipótesis, utilizado cuando se desea comparar dos grupos y averiguar si difieren. En las empresas este tipo de análisis es muy útil cuando se desea saber si lanzar un producto a un mercado específico será beneficioso para la organización, se obtiene datos de dos grupos diferentes y se comparan para saber si tendrá éxito o no. Las pruebas de hipótesis están fuertemente relacionadas con el modelado estadístico, y muchas veces, se puede realizar una combinación de los dos métodos para realizar un análisis más robusto.

Una vez se ha determinado que el enfoque estadístico es el mejor para resolver el problema, la implementación en Python consiste en 4 pasos: La carga de datos al entorno de trabajo por medio de la librería pandas, La elección de las variables predictoras mediante de la librería Sklearn haciendo uso del método RFE que se encuentra dentro del módulo *feature_selection*, el tercer paso es el modelado del problema usando también la librería Sklearn y

los módulos para la regresión lineal y polinomial, se utiliza el módulo *metrics* de la misma librería para generar el coeficiente de correlación y RMSE. El último paso consiste en la predicción de valores futuros usando el modelo previamente construido.

Figura 8. **Regresión polinomial en código Python**

```python
import numpy as np
import pandas as pd
from sklearn.feature_selection import RFE
from sklearn.svm import SVR
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn.metrics import mean_squared_error,r2_score
path =
"https://raw.githubusercontent.com/desog/PropuestaMetodologia/main/Recursos/DataSets/space-
shuttle/o-ring-erosion-only.csv"
dataframe = pd.read_csv(path)
rasgos_prediccion = ["Experiencing thermal distress","Launch temperature (degrees F)","Leak-
check pressure (psi)"]
x = dataframe[rasgos_prediccion]
y = dataframe["Experiencing thermal distress"]
estimator = SVR(kernel = "linear")
selector = RFE(estimator,n_features_to_select=2,step=1)
selector = selector.fit(x,y)
x_pred = x[["Launch temperature (degrees F)"]]
poly_degree =2
polynomial_features = PolynomialFeatures(degree = poly_degree)
x_transform = polynomial_features.fit_transform(x_pred)
model = LinearRegression().fit(x_transform, y)
y_new = model.predict(x_transform)
#  rmse y r2
rmse = np.sqrt(mean_squared_error(y, y_new))
r2 = r2_score(y, y_new)
print('RMSE: ', rmse)
print('R2: ', r2)
x_new_min = 0.0
x_new_max = 70
x_new = np.linspace(x_new_min, x_new_max, 70)
x_new = x_new[:,np.newaxis]
x_new_transform = polynomial_features.fit_transform(x_new)
y_new = model.predict(x_new_transform)
```

Fuente: elaboración propia, empleando Carbon.

4.3.1.2. Enfoque de aprendizaje supervisado

Aunque el enfoque estadístico es útil en la mayoría de los casos existen situaciones más complejas en donde es necesario un análisis más robusto. En

inteligencia artificial, concretamente en Machine Learning hay técnicas de regresión y clasificación que se basan en el aprendizaje supervisado. El uso del aprendizaje supervisado es conveniente cuando se necesita que mediante información de entrada se predigan valores de salida.

El uso de algoritmos de aprendizaje supervisado en la ciencia de datos es muy común en el caso de que se quiera realizar un análisis predictivo. Un ejemplo claro del uso de esta técnica es en el riesgo de dar un crédito a alguna persona basado en su historial crediticio o cómo será la demanda de un producto en el futuro.

Los algoritmos más usados en este enfoque son los árboles de decisión que se enfocan en el cálculo de la entropía para determinar las opciones hasta lograr una predicción. Bayes y redes neuronales que determina una relación de causa y efecto basado en probabilidades para llegar a un resultado. La característica más importante de estos algoritmos es que se tiene un conjunto de datos de entrenamiento que son los que ajustan el modelo y una salida esperada que ayuda a determinar la exactitud de la predicción.

La implementación del enfoque de aprendizaje supervisado en Python se compone de 4 o 5 pasos, dependerá si las variables a analizar son categóricas o numéricas, el primer paso es la carga y división de datos en el conjunto de entrenamiento y de pruebas para analizar, el segundo paso dependerá sí las variables son categóricas se tendrán que transformar a variables numéricas para que la biblioteca sea capaz de analizarla, el siguiente paso es generar el modelo mediante la librería Sklearn, si el problema es de decisión los módulos DecisionTreeClassifier y Plot_tree son necesarios para obtener el modelo, en caso de que el problema sea de clasificación se puede utilizar Bayes mediante el módulo Naive_bayes con el método GaussianNB o redes neuronales con el

módulo Neural_network y el método MLPClassifier, por último se define el conjunto de test y se hace la predicción de la clasificación con el modelo creado.

Figura 9. **Árbol de decisión en código Python**

```python
import numpy as np
import matplotlib.pyplot as plt
from sklearn.tree import DecisionTreeClassifier, plot_tree
from sklearn.preprocessing import LabelEncoder
# Carga de datos:
Outlook =
['sunny','sunny','overcast','rain','rain','rain','overcast','sunny','sunny','rain','sunny','over
cast','overcast','rain']
Temperature =
['hot','hot','hot','mild','cool','cool','cool','mild','cool','mild','mild','mild','hot','mild']
Humidity =
['high','high','high','high','normal','normal','normal','high','normal','normal','normal','high'
,'normal','high']
Windy =
['false','true','false','false','false','true','true','false','false','false','true','true','fal
se','true']
Class = ['N','N','P','P','P','N','P','N','P','P','P','P','P','N']
le = LabelEncoder()
OutlookNumeros = le.fit_transform(Outlook)
TemperatureNumeros = le.fit_transform(Temperature)
HumidityNumeros = le.fit_transform(Humidity)
WindyNumeros = le.fit_transform(Windy)
ClassNumeros = le.fit_transform(Class)
datosCombinados = list(zip(OutlookNumeros,TemperatureNumeros,HumidityNumeros,WindyNumeros))
print(datosCombinados)
plt.figure()
clf = DecisionTreeClassifier().fit(datosCombinados,ClassNumeros)
plot_tree(clf,filled=True)
plt.show()
```

Fuente: elaboración propia, empleando Carbon.

Figura 10. **Redes neuronales en código Python**

```python
from sklearn import preprocessing
import pandas as pd
from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier
path =
"https://raw.githubusercontent.com/desog/PropuestaMetodologia/main/Recursos/DataSets/Car%20Evaluation/car.csv"
dataframe = pd.read_csv(path)
print(dataframe.head())
train, test = train_test_split(dataframe, test_size = 0.2)
#Obteniendo variables
buying = train["buying"]
maint = train["maint"]
doors = train["doors"]
persons = train["persons"]
lug_boot = train["lug_boot"]
safety = train["safety"]
clases = train["class"]
# Transformando variables
le = preprocessing.LabelEncoder()
buying_encoded = le.fit_transform(buying)
maint_encoded = le.fit_transform(maint)
doors_encoded = le.fit_transform(doors)
persons_encoded = le.fit_transform(persons)
lug_boot_encoded = le.fit_transform(lug_boot)
safety_encoded = le.fit_transform(safety)
clases_encoded = le.fit_transform(clases)
combinacion =
list(zip(buying_encoded,maint_encoded,doors_encoded,persons_encoded,lug_boot_encoded,safety_encoded))
print(clases)
print(clases_encoded)
#Modelado
model = MLPClassifier(activation='logistic',max_iter=700,hidden_layer_sizes=
(4,),alpha=0.001,solver='lbfgs')
model.fit(combinacion,clases_encoded)
#Test
buyingtest = test["buying"]
mainttest = test["maint"]
doorstest = test["doors"]
personstest = test["persons"]
lug_boottest = test["lug_boot"]
safetytest = test["safety"]
clasestest = test["class"]
le = preprocessing.LabelEncoder()
buyingtest_encoded = le.fit_transform(buyingtest)
mainttest_encoded = le.fit_transform(mainttest)
doorstest_encoded = le.fit_transform(doorstest)
personstest_encoded = le.fit_transform(personstest)
lug_boottest_encoded = le.fit_transform(lug_bottest)
safetytest_encoded = le.fit_transform(safetytest)
clasestest_encoded = le.fit_transform(clasestest)
combinaciontest =
list(zip(buyingtest_encoded,mainttest_encoded,doorstest_encoded,personstest_encoded,lug_boottest_encoded,safetytest_encoded))
print(clasestest)
print(clasestest_encoded)
#Predicción
print('Predicciones: ',model.predict(combinaciontest))
```

Fuente: elaboración propia, empleando Carbon.

4.3.1.3. Enfoque de aprendizaje no supervisado

Este enfoque es útil cuando lo que se necesita es clasificar la información para obtener un resultado, el ejemplo más claro es cuando se está realizando una segmentación de mercado para determinado producto. Otro caso de uso de los algoritmos de aprendizaje no supervisado es la búsqueda de patrones en un problema específico, especialmente útil cuando se tienen múltiples variables y es difícil determinar el comportamiento que tiene cada una de ellas.

Los algoritmos más utilizados en este tipo de enfoque son K-Means que es el método que utiliza las características de los datos agrupándolos en k grupos de manera iterativa hasta obtener una convergencia del cien por ciento entre los grupos creados, y el agrupamiento jerárquico que a diferencia de su homólogo no se tiene un número definido de grupos eso quiere decir que en cualquier momento de la iteración se puede detener, este último algoritmo es decisivo para clasificar datos aglomerados en forma divisiva.

Para implementar el enfoque del aprendizaje no supervisado en Python se hace uso del método KMeans que se encuentra dentro módulo clúster de la librería Sklearn. La implementación comienza con la carga de datos mediante la biblioteca pandas, el segundo paso es seleccionar los datos a analizar para luego empaquetarlos en un arreglo usando la librería Numpy.

El siguiente paso es seleccionar el número de grupos en los cuales se van a clasificar los datos, ajustar el modelo y predecir con los datos cargados en el sistema. Por último, se suele generar un gráfico de dispersión para identificar los grupos creados por el modelo.

Figura 11. **Algoritmo K-Means en código Python**

```python
import pandas as pd
import numpy as np
from sklearn import preprocessing
from sklearn.cluster import KMeans
path =
"https://raw.githubusercontent.com/desog/PropuestaMetodologia/main/Recursos/DataSets/incd.csv"
df= pd.read_csv(path)
df['Electricidad_Total']= df['Electricity - production(kWh)'] - df['Electiricty -
consumption(kWh)']
df['Gas_Total'] = df['Natural gas - production(cu m)'] - df['Natural gas - consumption(cu m)']
x = df['Electricidad_Total']
y = df['Gas_Total']
nombre = df['Country']
X = np.array(list(zip(x,y)))
kmeans = KMeans(n_clusters=3)
kmeans.fit(X)
y_kmeans = kmeans.predict(X)
print("Centroides: ", kmeans.cluster_centers_)
plt.scatter(X[:, 0], X[:, 1], c=y_kmeans, s=30, cmap='rainbow')
plt.xlabel("Electricidad")
plt.ylabel("Gas")
plt.show()
```

Fuente: elaboración propia, empleando Carbon.

4.3.2. División de datos

Una vez seleccionado el enfoque del análisis siempre es necesario poder dividir el conjunto de datos inicial en dos partes: Conjunto de entrenamiento y conjunto de pruebas. Cada uno de los conjuntos debe de tener los datos suficientes como para poder validar que el modelo es construido correctamente. La forma más sencilla de realizar la división es utilizando el principio de Pareto. El 80 % de los datos deberán de ser de entrenamiento y el 20 % serán de prueba.

4.3.2.1. Construcción del conjunto de entrenamiento

El conjunto de entrenamiento es la base de un modelo de cualquier tipo, representa los datos con los que se hacen los cálculos y se modelizan las ecuaciones. Normalmente estos datos son de entrada y ayudan a ajustar el modelo que se creará para que posteriormente. No hay que abusar del conjunto de entrenamiento porque muchas veces pasa que el modelo se sobre ajusta a los datos de entrada y no funciona para predicciones futuras.

4.3.2.2. Construcción del conjunto de prueba

El conjunto de prueba ayuda a validar el modelo, aunque el conjunto de prueba tiene menos datos que el conjunto de entrenamiento suele ser la forma más simple de probar que el modelo se ha ajustado a los datos de entrada que se han proporcionado por parte de la organización.

Figura 12. **División de datos código Python**

```python
import pandas as pd
from sklearn.model_selection import train_test_split
path =
 "https://raw.githubusercontent.com/desog/PropuestaMetodologia/main/Recursos/DataSets/incd.csv"
dataframe = pd.read_csv(path, encoding = 'latin1')
print(dataframe.head())
train, test = train_test_split(dataframe, test_size = 0.2, random_state = 42)
print(len(dataframe))
print(len(train))
print(len(test))
```

Fuente: elaboración propia, empleando Carbon.

4.3.3. Otras consideraciones

Existen algunas otras cosas que se deben de considerar al momento del análisis de datos tal como: la calidad de los datos, el volumen de variables y la selección de la semilla de generación aleatoria para el análisis. El análisis puede fallar por estos factores, por ello, es bueno tener siempre una estrategia para poder adoptar el enfoque adecuado sabiendo que los datos tendrán el valor necesario para poder ejecutar correctamente el análisis y ajustar un modelo que sea capaz de solucionar el problema.

4.3.3.1. Calidad de los datos

La calidad de los datos puede afectar al desarrollo del análisis con el enfoque correspondiente, es posible que los datos no tengan las variables correctas a analizar que permitan resolver el problema. Aunque el problema de la calidad no es correspondiente a la ciencia de datos si puede ser un factor que afecte a todo el análisis llegando a provocar que no se puede resolver el problema sin los datos correctos.

4.4. Evaluación

El proceso de evaluación es la cumbre del análisis, en esta fase se corrobora que el modelo creado y ajustado con los datos de entrada sea lo suficientemente bueno como para resolver el problema inicial, en este caso ayudar a la gerencia en la toma de decisiones. En esta parte el científico de datos deberá garantizar la calidad del análisis mediante medidas de diagnóstico y otros métodos de evaluación.

4.4.1. Métodos de evaluación

Poder medir el desempeño del análisis es fundamental para encontrar errores y lograr que los objetivos principales de la organización se cumplan. Los métodos de evaluación que se utilizan comúnmente son:

4.4.1.1. Simulación

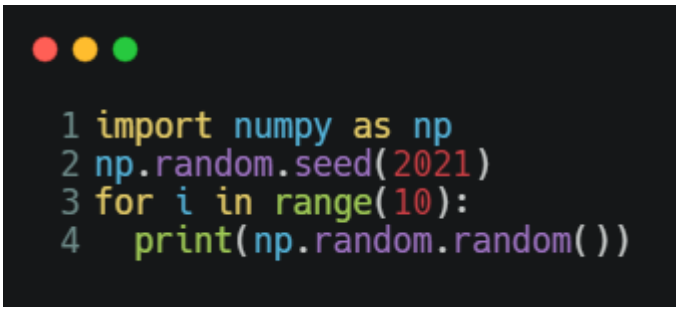
Simular datos de entrada para poder probar un modelo es siempre una buena estrategia para evaluar el análisis. Poder generar datos de prueba de manera aleatoria ayuda a comprobar el ajuste del modelo y su efectividad ante información nueva.

El hecho de realizar un experimento en un ambiente simulado da opción a evitar errores en la siguiente fase de la metodología. Lo más importante dentro de este método de evaluación es poder reproducirlo varias veces para su estudio, para ello, se utiliza una semilla de generación aleatoria que se define al inicio del experimento.

4.4.1.2. Selección de la semilla de generación aleatoria

El poder reproducir varias veces el resultado un modelo es importante, pero siempre es necesario mantener un grado de aleatoriedad especialmente en problemas que se resuelven con un enfoque estadístico. Obtener control sobre el determinismo de un proceso de experimentación es fundamental para poder garantizar el éxito de un modelo, la semilla aleatoria logra garantizar la reproductibilidad en un experimento.

Figura 13. **Generación de la semilla aleatoria código Python**



```
1 import numpy as np
2 np.random.seed(2021)
3 for i in range(10):
4   print(np.random.random())
```

Fuente: elaboración propia, empleando Carbon.

4.4.1.3. Gráficos y tablas

Durante el desarrollo de la evaluación para poder garantizar que el modelo aborda correctamente el problema de la organización. Se utilizan gráficos y tablas para interpretar la calidad del análisis y realizar un diagnóstico de una manera más sencilla. Dependiendo del enfoque seleccionado para resolver el problema se tendrá que elegir un tipo de gráfico o tabla diferente, los tipos son los siguientes:

- Gráfico de dispersión: Este es de los gráficos más utilizados en la ciencia de datos para describir el comportamiento de los datos. El uso de este tipo de gráfica es principalmente para el enfoque del aprendizaje no supervisado con el algoritmo de *clustering*.

Figura 14. **Creación de gráfico de dispersión código Python**

```
import matplotlib.pyplot as plt
import numpy as np
dataSet = np.array([[3,2],[2,2],[2,1],[1,2],[9,5],[9,4],[7,4],[8,3],[4,3],[8,4],[1,1],[1,3],
[3,3],[8,4.5],[8,5]])
plt.scatter(dataSet[:,0],dataSet[:,1], label = 'True Position')
plt.show()
```

Fuente: elaboración propia, empleando Carbon.

- Gráfico de línea: Este gráfico se utiliza para describir de manera simple el desarrollo de una o más variables a lo largo del tiempo. Esta gráfica es utilizada para el enfoque estadístico, especialmente en la regresión lineal o polinomial.

Figura 15. **Creación de gráfico de línea código Python**

```
import matplotlib.pyplot as plt
import numpy as np
x = np.array([2,3,6,8,10,12,14,16,20,26,28,30,34,36,38,44])
y = range(16)
plt.plot(x,y, color='blue', linewidth=3)
plt.title("Gráfico de lineas ", fontsize=10)
plt.xlabel('x')
plt.ylabel('y')
plt.show()
```

Fuente: elaboración propia, empleando Carbon.

- Tabla de clasificación: Las tablas de clasificación se utilizan en el enfoque de aprendizaje supervisado, ayudan a determinar la clase a la

que pertenecen los datos y las posibles predicciones que pueden realizar los algoritmos por el análisis de los datos.

4.5. Presentación

En esta etapa de la metodología se presentan los resultados del análisis ya evaluado a la gerencia. Aunque esta fase es la menos técnica es la que determina si el modelo tuvo éxito. Para que la presentación de los datos sea satisfactoria se debe de tener en cuenta que los resultados se entregan en informes de rendimiento que suelen contener: tablas para una comprensión sencilla de los datos, los gráficos utilizados durante la evaluación del modelo son una buena herramienta para mostrar el comportamiento de los resultados y por último se debe de mostrar las fuentes de datos que sirvieron en el análisis.

El fin de la ciencia de datos es dar valor a los datos almacenados por la organización y así ayudar a la toma de decisiones, se debe de destacar como el modelo creado durante la fase de análisis aporta ese valor y resuelve el problema planteado por la gerencia al principio del análisis.

5. PROPUESTA DE ARQUITECTURA MÍNIMA PARA CIENCIA DE DATOS

Independientemente de la organización la propuesta de arquitectura mínima busca tener la infraestructura necesaria para que todo el proceso de la metodología pueda realizarse con eficiencia. La arquitectura por utilizar se basa en la nube.

Hace uso de la nube pública Amazon Web Services (AWS), y sus herramientas para obtener mayor escalabilidad y flexibilidad para que pueda ser implementada en cualquier entorno.

5.1. Repositorio de datos

Las fuentes de datos empresariales pueden ser bases de datos, documentos en formato csv o simplemente archivos de texto plano. Todos estos recursos suelen ser generados a diario dentro de las organizaciones y describe muy bien sus alcances.

5.2. Fuentes de datos externas

Este tipo de recursos son todos aquellos datos que vienen de entidades externas a las empresas, Analíticas de Google para mejorar la publicidad, redes sociales u objetos de IoT que generen datos de interés para mejorar ámbitos de la organización.

5.3. Data Lake

Se recomienda utilizar un repositorio de datos de tipo Data Lake en donde se almacenen los datos de todas las fuentes de la organización para luego darles el formato que se necesite para el análisis. Para fines de la metodología se recomienda utilizar el servicio de Amazon S3 que ofrece almacenamiento de objetos con o sin estructurar.

La ventaja que ofrece el almacenamiento en la nube es el poco tiempo de implementación y los bajos costos de operación para grandes cantidades de datos, se aconseja tener el paquete de S3 estándar con acceso poco frecuente con un estimado de 1Tb de datos generados por mes.

Para llevar los datos sin estructurar al proceso de ciencia de datos se recomienda utilizar el servicio Amazon Kinesis para la transmisión de grandes cantidades de datos de manera sencilla con un bajo coste.

5.4. Funciones como servicio

La propuesta para que los datos alojados en el Data Lake sean transformados y puedan ser utilizados para generar el análisis, entrenamiento, evaluación y presentación correspondiente, es utilizar Lambda que es la tecnología de funciones como servicio que ofrece Aws.

Bajo este concepto se logra minimizar la cantidad de infraestructura que se debe implementar además de poder tener una alta respuesta totalmente escalable por parte de la ejecución de las funciones de esta etapa de la metodología.

5.5. Repositorio de datos transformados

Para el almacenamiento de los datos transformados durante la etapa de limpieza de la metodología que luego servirán de insumo para la etapa de modelado se propone utilizar el almacenamiento Amazon S3 bajo el estándar de acceso poco frecuente con un estimado de 1Tb de datos generados por mes.

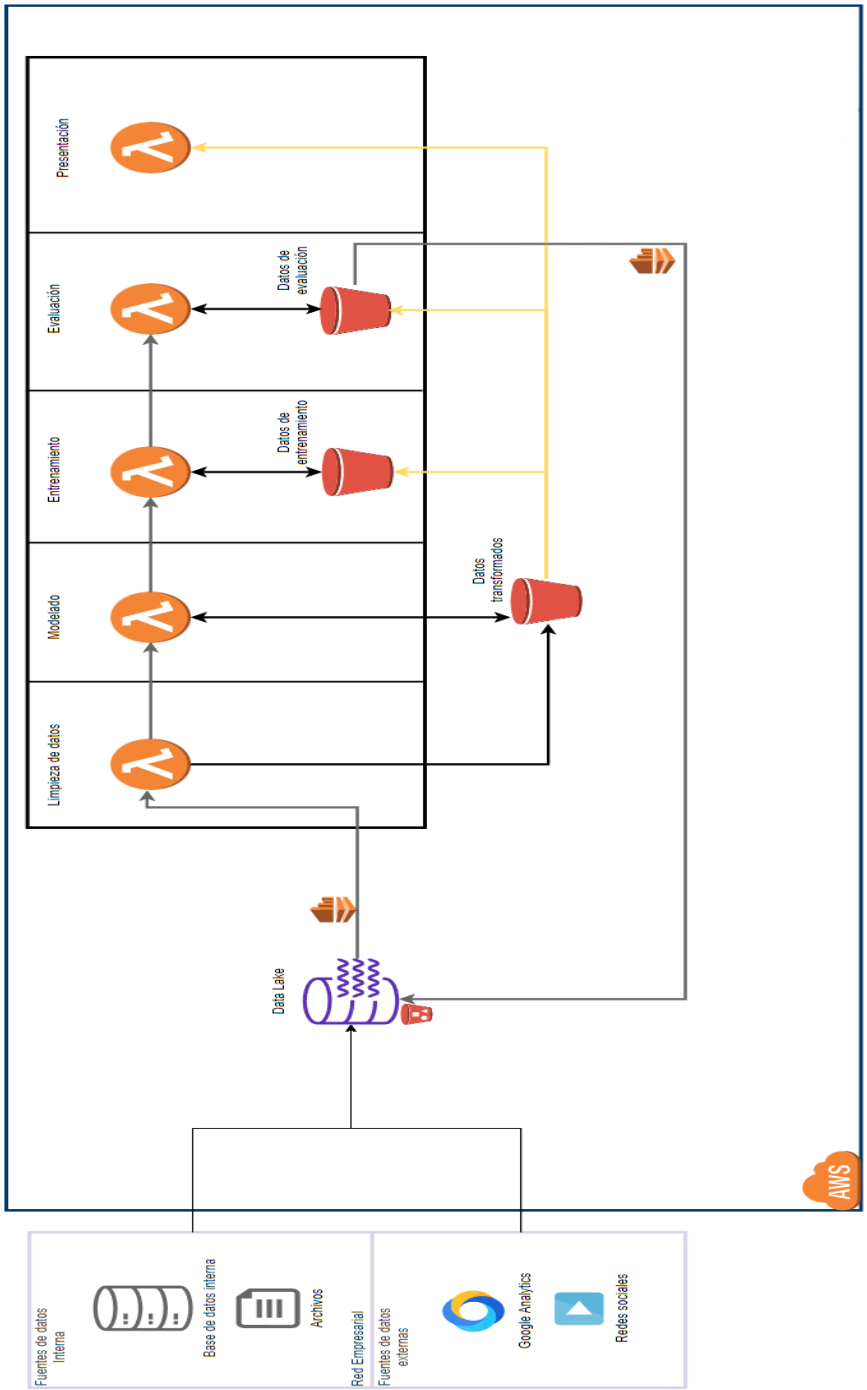
5.6. Repositorio de datos de entrenamiento

Para el almacenamiento de los datos que servirán de entrenamiento al modelo se propone utilizar el almacenamiento Amazon S3 bajo el estándar de acceso poco frecuente con un estimado de 500Gb de datos generados por mes.

5.7. Repositorio de datos de evaluación

De igual manera que en el repositorio de datos de entrenamiento se recomienda que para el almacenamiento de los datos que servirán para validar se utilice el almacenamiento Amazon S3 bajo el estándar de acceso poco frecuente con un estimado de 500Gb de datos generados por mes, aparte de utilizar el servicio de transmisión de datos Amazon Kinsesis para poder almacenarlos dentro del Data Lake y puedan ser reutilizados.

Figura 16. **Propuesta de arquitectura para ciencia de datos**



Fuente: elaboración propia, empleando Terrastruct.

CONCLUSIONES

1. La toma de decisiones basadas en datos, la mejora de los servicios ofrecidos y el análisis predictivo son beneficios que otorga el análisis de datos a las empresas.

2. Se propuso una metodología de implementación de ciencia de datos que considera la implementación desde cero para que sea utilizada por los ejecutivos en Guatemala en la toma de decisiones empresariales.

3. Se construyó un plan cíclico de cinco pasos que ayuda a sacar provecho de los datos generados por las empresas y que el resultado puede ser utilizado como insumo del mismo plan.

RECOMENDACIONES

1. Tomar en cuenta que los datos generados por las empresas guatemaltecas tienen valor, y mediante el análisis correcto pueden generar ventajas competitivas y aumentar la satisfacción de sus clientes.

2. Digitalizar lo más rápido posible la mayor cantidad de procesos dentro de las organizaciones, para aprovechar al máximo los beneficios del análisis de datos.

3. Analizar el comportamiento de los datos durante la etapa de análisis de la metodología, permitirá elegir la mejor técnica de ajuste al modelo y así aprovechar mejor el valor de la información.

REFERENCIAS

1. Big Data. *Gartner glossary.* [en línea]. <https://www.gartner.com/en/information-technology/glossary/big-data>. [Consulta: 7 de marzo de 2021].

2. CADY, Field. *Handbook for data scientists*. Somerset: John Wiley & Sons, Incorporated. New Jersey, Estados Unidos. 2017. 416 p.

3. COPELAND, Jack. *Artificial intelligence*. [en línea]. <https://www.britannica.com/technology/artificial-intelligence>. [Consulta: 7 de marzo de 2021].

4. DILMEGANI, Cem. *The Ultimate Guide to The Top 20 Data Science Tools*. [en línea]. < https://research.aimultiple.com/data-science-tools/>. [Consulta: 25 de marzo de 2021].

5. ESPINO, Carlos. *Análisis predictivo: técnicas y modelos utilizados y aplicaciones del mismo - herramientas Open Source que permiten su uso*. Trabajo de fin de grado de Ing. Informática. Universitat Oberta de Catalunya, España, 2017. 65 p.

6. HAYES, Adam. *Probability distribution.* [en línea]. <https://www.investopedia.com/terms/p/probabilitydistribution.asp>. [Consulta: 18 de marzo de 2021].

7. KAMPAKIS, Stylianos. *The decision maker's handbook to data science: A guide for non-technical executives, managers, and founders*. Londres: Apress, 2019. 156 p.

8. LÓPEZ, José. *Estadístico*. [en línea]. <https://economipedia.com/definiciones/estadistico.html>. [Consulta: 9 de marzo de 2021].

9. MAALOUF, Maher. *Logistic regression in data analysis: an overview. International journal of data analysis techniques and strategies*. doi: 10.1504/ijdats.2011.041335, 2011. 299 p.

10. Machine Learning. *IBM cloud education.* [en línea] <https://www.ibm.com/cloud/learn/machine-learning>. [Consulta: 25 de marzo de 2021].

11. MACKINNEY, Wes. *Python for data analysis* sebastopol: USA. O'Reilly Media, Inc, 2012. 505 p.

12. ORANTES KESTLER, Alejandro. *La inteligencia artificial y las oportunidades para la empresa en Guatemala*. Revista Ciencia Multidisciplinaria. CUNORI. 146 p.

13. PROVOST, Foster, & FAWCETT, Tom. *Data science for business*. Sebastopol, California: O'Reilly. 2013. 367 p.

14. QUEVEDO RICARDI, Fernando. *The chi-square. Medwave*. doi: 10.5867/medwave.2011.12.5266, 2011. 5 p.

15. Redacción APD. *Big data vs data science: Principales diferencias* [en línea]. <https://www.apd.es/big-data-vs-data-science/>. [Consulta: 27 de marzo de 2021].

16. RENEAR, Allen., SACCHI, Simone. & WICKETT, Karen. *Definitions of dataset in the scientific and technical literature. Proceedings of The American Society for Information Science and Technology*. doi: 10.1002/meet.14504701240, 2010. 4 p.

17. ROUSE, Margaret. *Novedades en la ciencia de los datos*. [en línea] <https://www.computerweekly.com/es/ehandbook/Novedades-en-la-Ciencia-de-los-Datos>. [Consulta: 27 de marzo de 2021].

18. SANDERS, Nathan. *A balanced perspective on prediction and inference for data science in industry*. *Issue 1*. doi: 10.1162/99608f92.644ef4a4, 2019. 20 p.

19. SOME, Kamalika. *Countries which hold the greatest opportunities for data scientists* [en línea]. <https://www.analyticsinsight.net/countries-which-hold-the-greatest-opportunities-for-data-scientists/>. [Consulta: 27 de marzo de 2021].

20. STEDMAN, Craig. *¿Qué es ciencia de datos?* [en línea]. <https://searchdatacenter.techtarget.com/es/definicion/Ciencia-de-datos>. [Consulta: 27 de marzo de 2021].

21. VANDERPLAS, Jake. *Python data science handbook*. Sebastopol, California: O'Reilly, 2017. 517 p.

22. ZHANG, Arthur. *Data analytics*. CreateSpace Independent Publishing Platform, 2017. 279 p.