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

DISEÑO DE LA INVESTIGACIÓN IMPLEMENTACIÓN DE UN PROTOTIPO PARA LA ACTUALIZACIÓN DE *DASHBOARDS* Y PREDICCIÓN DE PRODUCTOS POTENCIALES PARA LOS CLIENTES, EN UNA EMPRESA COMERCIALIZADORA DE PRODUCTOS DE GUATEMALA

Sergio Geovany Guoz Tubac
Asesorado por el M. sc. Ing. Johnatan Esaú Franco Clara

Guatemala, enero de 2024

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

DISEÑO DE LA INVESTIGACIÓN IMPLEMENTACIÓN DE UN PROTOTIPO PARA LA ACTUALIZACIÓN DE *DASHBOARDS* Y PREDICCIÓN DE PRODUCTOS POTENCIALES PARA LOS CLIENTES, EN UNA EMPRESA COMERCIALIZADORA DE PRODUCTOS DE GUATEMALA

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

SERGIO GEOVANY GUOZ TUBAC
ASESORADO POR M. SC. ING. JOHNATAN ESAÚ FRANCO CLARA

AL CONFERÍRSELE EL TÍTULO DE

INGENIERO EN CIENCIAS Y SISTEMAS

GUATEMALA, ENERO 2024

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANO	Ing. José Francisco Gómez Rivera (a.i.)
VOCAL II	Ing. Mario Renato Escobedo Martinez
VOCAL III	Ing. José Milton De León Bran
VOCAL IV	Ing. Kevin Vladimir Cruz Lorente
VOCAL V	Ing. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN GENERAL PRIVADO

DECANO	Ing. José Francisco Gómez Rivera (a.i.)
EXAMINADOR	Ing. Edgar Estuardo Santos Sutuj
EXAMINADOR	Ing. Juan Alvaro Díaz Ardavin
EXAMINADOR	Ing. Edgar René Ornelis Hoil
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

EEPFI-PP-0425-2023
Guatemala, 23 de abril de 2023

Director
Carlos Gustavo Alonzo
Escuela De Ingenieria En Sistemas
Presente.

Estimado Mtro. Alonzo

Reciba un cordial saludo de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería.

El propósito de la presente es para informarle que se ha revisado y aprobado el Diseño de Investigación titulado: **IMPLEMENTACIÓN DE UN PROTOTIPO PARA LA ACTUALIZACIÓN DE DASHBOARDS Y PREDICCIÓN DE PRODUCTOS POTENCIALES PARA LOS CLIENTES EN UNA EMPRESA COMERCIALIZADORA DE PRODUCTOS DE GUATEMALA**, el cual se enmarca en la línea de investigación: **Análisis de datos - Análisis de datos**, presentado por el estudiante **Sergio Geovany Guoz Tubac** con cui **2931956250409**, quien optó por la modalidad del "PROCESO DE GRADUACIÓN DE LOS ESTUDIANTES DE LA FACULTAD DE INGENIERÍA OPCIÓN ESTUDIOS DE POSTGRADO". Previo a culminar sus estudios en la Maestría en Artes en Ingeniería Para La Industria Con Especialidad En Ciencias De La Computación.

Y habiendo cumplido y aprobado con los requisitos establecidos en el normativo de este Proceso de Graduación en el Punto 6.2, aprobado por la Junta Directiva de la Facultad de Ingeniería en el Punto Décimo, Inciso 10.2 del Acta 28-2011 de fecha 19 de septiembre de 2011, firmo y sello la presente para el trámite correspondiente de graduación de Pregrado.

Atentamente,

"Id y Enseñad a Todos"

Mtro. Johnatan Esau Franco Clara
Asesor(a)

Mtro. Carlos Gustavo Alonzo
Coordinador(a) de Maestría

UNIVERSIDAD DE SAN CARLOS
ESCUELA DE POSTGRADO
FACULTAD DE INGENIERIA
DE GUATEMALA

Mtro. Edgar Dario Alvarez Coti
Director
Escuela de Estudios de Postgrado
Facultad de Ingeniería

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
FACULTAD DE INGENIERÍA
DIRECCIÓN
ESCUELA DE POSTGRADO

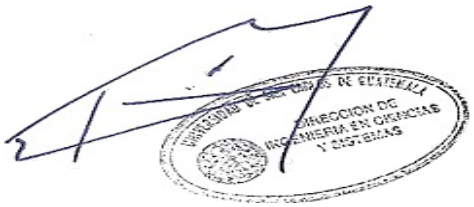
UNIVERSIDAD DE SAN CARLOS
DE GUATEMALA

FACULTAD DE INGENIERÍA

EEP-EICS-0424-2023

El Director de la Escuela De Ingenieria En Sistemas de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del Asesor, el visto bueno del Coordinador y Director de la Escuela de Estudios de Postgrado, del Diseño de Investigación en la modalidad Estudios de Pregrado y Postgrado titulado: **IMPLEMENTACIÓN DE UN PROTOTIPO PARA LA ACTUALIZACIÓN DE DASHBOARDS Y PREDICCIÓN DE PRODUCTOS POTENCIALES PARA LOS CLIENTES EN UNA EMPRESA COMERCIALIZADORA DE PRODUCTOS DE GUATEMALA**, presentado por el estudiante universitario **Sergio Geovany Guoz Tubac**, procedo con el Aval del mismo, ya que cumple con los requisitos normados por la Facultad de Ingeniería en esta modalidad.

ID Y ENSEÑAD A TODOS



Mtro. Carlos Gustavo Alonzo
Director
Escuela De Ingenieria En Sistemas

Guatemala, abril de 2023

Decanato
Facultad e Ingeniería

24189101- 24189102

LNG.DECANATO.OIE.65.2024

El Decano 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: **IMPLEMENTACIÓN DE UN PROTOTIPO PARA LA ACTUALIZACIÓN DE DASHBOARDS Y PREDICCIÓN DE PRODUCTOS POTENCIALES PARA LOS CLIENTES EN UNA EMPRESA COMERCIALIZADORA DE PRODUCTOS DE GUATEMALA**, presentado por: **Sergio Geovany Guoz Tubac** después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

Firmado electrónicamente por: José Francisco Goméz Rivera
Motivo: Orden de impresión
Fecha: 28/01/2024 11:15:08
Lugar: Facultad de Ingeniería, USAC.

Ing. José Francisco Gómez Rivera
Decano a.i.

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANO a.i.
Facultad de Ingeniería

Guatemala, enero de 2024

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:

DISEÑO DE LA INVESTIGACIÓN IMPLEMENTACIÓN DE UN PROTOTIPO PARA LA ACTUALIZACIÓN DE *DASHBOARDS* Y PREDICCIÓN DE PRODUCTOS POTENCIALES PARA LOS CLIENTES, EN UNA EMPRESA COMERCIALIZADORA DE PRODUCTOS DE GUATEMALA

Tema que me fuera asignado por la Dirección de la Escuela de Ciencias y Sistemas, con fecha 23 de abril de 2023.

Sergio Geovany Guoz Tubac

ACTO QUE DEDICO A:

Dios	Por ser el supremo creador del universo y por permitirme cumplir este sueño.
Mis padres	Héctor Sergio Guoz y Vilma Yolanda Tubac, por su apoyo incondicional, quienes son fuente de inspiración para alcanzar este logro.
Mis hermanos	Karen, Jorge y Jhonatan Guoz Tubac, por su constante apoyo y compañía durante mi vida.
Mis abuelos	María Ollej, Leonardo Tubac, Isidro Guoz y Rosa Esquit, por sus sabias enseñanzas y consejos durante toda mi vida.
Mis tíos	Carlos (q. d. e. p.) y Cristina Guoz, por el apoyo y ánimos brindados.

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	Por ser la *alma mater* que me permitió nutrirme de conocimientos.
Facultad de Ingeniería	Por proporcionarme los conocimientos y forjarme como profesional.
Mis amigos	Por los momentos compartidos durante la carrera.
Mi asesor	Msc. Ing. Johnatan Franco, por haberme guiado durante el trabajo de graduación.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
Q	Quetzal

GLOSARIO

API	Interfaz de programación de aplicaciones
Athena	Servicio de análisis interactivo
AWS	Amazon Web Services, colección de servicios de computación en la nube.
AMS	Servicios administrados por AWS
Bucket	Es un contenedor de objetos.
Cloud computing	Computación en la nube
Crawler	Herramienta para el manejo de catálogos de datos
Data mart	Almacén de datos con información específica
Data warehouse	Almacén de datos
ETL	Extracción, Transformación y Carga
Glue	Herramienta de integración de datos
Glue DataBrew	Herramienta visual de preparación de datos

IoT	Internet de las cosas
OLAP	Procesamiento analítico en línea
OLTP	Procesamiento de transacciones en línea
QuickSight	Herramienta utilizada para la generación de *dashboards*.
RDBMS	Sistema de administración de bases de datos relacionales.
S3	Simple Storage Service de AWS, almacenamiento en la nube
SageMaker	Herramienta de AWS utilizada para generar y exponer modelos de predictivos
Serverless	Se refiere a servicios sin servidor

RESUMEN

Los procesos de actualización de *dashboards* de forma manual generan mayor consumo de recursos, pero por medio de tecnologías en la nube se puede implementar la automatización de este proceso, para contar con información en la toma de decisiones. Los modelos predictivos también apoyan a que el negocio sea competitivo por medio de sistemas de recomendación.

Las tecnologías en la nube permiten realizar análisis y presentación de información de una forma ágil, por medio de servicios administrados por Amazon Web Services (AWS, siglas en inglés); con herramientas *serverless* se pueden generar *dashboards* y modelos predictivos que pueden ser utilizados en sistemas existentes. Los procesos de extracción, transformación y carga de información se pueden efectuar de manera eficiente y con recursos computacionales mayores de los que podría tenerse de forma local, lo que a su vez conlleva una reducción de costos.

El presente diseño de investigación busca la implementación de un prototipo para la mejora de un proceso de actualización de *dashboards* a través de herramientas y tecnologías en la nube, específicamente AWS, para automatizar el proceso de extracción, transformación y carga de información, con el fin de generar *dashboards* que aporten en la toma de decisiones. Con la información almacenada en la nube, se busca implementar un modelo para la predicción de productos potenciales para los clientes de una empresa de Guatemala.

El diseño de investigación plantea la implementación de un prototipo a través de cinco fases, midiendo el proceso actual, seguido por el proceso Extracción, Transformación y Carga (ETL, siglas en inglés), la generación de *dashboards*, generación del modelo predictivo y, por último, la evaluación del nuevo proceso.

1. INTRODUCCIÓN

Automatizar procesos e implementar modelos predictivos implican una ventaja competitiva para las empresas frente a la competencia. Se deben tener procesos eficientes de actualización de *dashboards* en la toma de decisiones y que este aporte en este proceso.

El siguiente trabajo de investigación consiste en una implementación de un prototipo de actualización de *dashboards*, que pretende sustituir un proceso manual de actualización que conlleva tiempo y genera lentitud en el procedimiento. Esto se efectuará mediante una reingeniería del proceso e implementación de tecnologías en la nube; así mismo, es importante analizar la información para que oriente en la toma de decisiones y que aporte en la competitividad de la empresa frente a su competencia. Para ello, se elabora un modelo predictivo con el fin de conocer productos potenciales para los clientes; esto se llevará a cabo en una empresa comercializadora de productos en Guatemala.

En el primer capítulo se definen los antecedentes de la investigación, se presentan trabajos relacionados con implementaciones de arquitecturas y servicios en la nube, orientados al análisis de información y también a la generación de modelos predictivos que aportan en la toma de decisiones organizacionales.

En el segundo capítulo se justifica la implementación del prototipo para la actualización de *dashboards* y la creación del modelo predictivo, enfocado en la reingeniería y optimización del proceso de generación de *dashboards*. También

se definen los aportes que el modelo podría tener para apoyar a la organización ante la competencia.

En el tercer capítulo se definen los alcances de la investigación, se delimita el enfoque y lo que abarca el prototipo, definiéndolo desde las perspectivas investigativa, técnica y resultados.

En el cuarto capítulo se presenta la teoría que sustenta la investigación, se describen temas que aportan y son base para la elaboración del prototipo acerca de arquitecturas en la nube, procesos de análisis de datos y algoritmos predictivos.

En el quinto capitulo se detallan los resultados obtenidos tras implementación del prototipo, los datos obtenidos del nuevo proceso de actualización de dashbords y la efectividad del modelo predictivo; también se muestran los modelos y técnicas utilizadas.

En el sexto capítulo se presenta el análisis y la interpretación de los resultados obtenidos, con implementación del nuevo proceso de actualización de *dashboards* comparado con el anterior, así como la precisión del modelo predictivo para identificar productos potenciales para los clientes.

Con la implementación del prototipo se pretende que el negocio tenga *dashboards* actualizados en periodos cortos, que sea un proceso automatizado y eficiente para aportar en la toma de decisiones. Por otro lado, con el análisis de información, por medio de un modelo predictivo, aportará también en la toma de decisiones y pretende identificar productos potenciales para los clientes.

2. ANTECEDENTES

La computación en la nube nos provee de varios modelos de servicios, tales como *Software as a Service*, *Platform as a Service* e *Infraestructure as a Service*, las cuales se pueden aprovechar para implementar distintos tipos de arquitecturas como las arquitecturas monolíticas, *serverless* y microservicios, entre otras.

En un trabajo de investigación de una arquitectura de Internet de las Cosas (IoT, siglas en inglés) escalable, basada en AWS, se implementó una arquitectura para la ganadería inteligente, la cual fue construida con servicios *serverless* y con varios grupos de servicios tales como almacenamiento, bases de datos, cómputo, análisis y seguridad, entre otros.

La organización de dicha arquitectura se forma en grupos como: reconocimiento de datos, transmisión de datos, repositorio de datos, almacenamiento y procesamiento de datos, aprendizaje automático, notificaciones y otros; esta organización permite cumplir con buenas prácticas y con un buen diseño de arquitectura. Esta arquitectura, en pruebas de rendimiento, tuvo la capacidad para manejar la cantidad requerida de 10,000 dispositivos IoT, resultando en una arquitectura eficaz. (Dineva & Atanasova, 2021)

Una arquitectura organizada es útil, fácil de mantener y tiene buen rendimiento para el objetivo de análisis y procesamiento de información.

Una arquitectura similar a la anterior es la arquitectura de servicios administrados por AWS, que puede ser útil en la generación de un proceso de extracción, transformación y carga de datos, con el fin de llenar un repositorio de información para mostrar datos en sistemas de *dashboard* y, también, para el uso de otros servicios que son útiles en el análisis de datos.

Al tener definida una arquitectura se procede con la implementación de búsqueda de patrones de compra entre productos, utilizando el algoritmo *a priori*, el cual funciona al aplicar las reglas de asociación: soporte, confianza e importancia. La regla de soporte se basa en el porcentaje de apariciones de un artículo, la regla de confianza se basa en porcentajes de probabilidad que mide la relación entre productos, y la regla de importancia valida si dos artículos han sido comprados con frecuencia; para indicar que hay una relación entre artículos, el valor de importancia debe ser mayor a 1, si es menor a 1 puede considerarse que la relación de los artículos no es frecuente. (Lee, 2022)

También existen algoritmos más complejos para pronosticar resultados. En el caso de estudio nos interesa el pronóstico de las ventas que un negocio puede tener; para ello existen los modelos de *machine learning*, tales como: Regresión Lineal, Árboles de Decisión, Redes Neuronales y XGBoost, entre otros, que tienen muchas aplicaciones para la resolución de problemas.

En un estudio comparativo de modelos de pronóstico sobre demanda para una empresa minorista multicanal, se realizó un modelo para el pronóstico de ventas de una empresa minorista, en el cual tomaron datos semanales de ventas para el análisis, realizando la comparación de varios algoritmos y además un modelo híbrido, que es la combinación de tres algoritmos. Para tratar de eliminar las deficiencias de los modelos Random Forest y XGBoost, proponen un modelo híbrido que combina Random Forest y XGBoost para

alimentar un algoritmo de regresión logística, lo que resulta en un modelo más veraz y con mejores resultados. (Mitra et al., 2022)

En otro estudio, donde crean un modelo para la predicción de ventas, utilizando una comparación de algoritmos, titulado *Modelos de aprendizaje automático para la previsión de series temporales de ventas*, concluye que se tienen mejores resultados para la predicción de ventas de un negocio un método con un enfoque de regresión, comparado con un método de series de tiempos, debido al supuesto de los métodos de regresión de que los patrones volverán a repetirse en el futuro. (Pavlyshenko, 2019)

Al tener un modelo entrenado de *machine learning*, la siguiente parte es hacer uso del modelo, por lo que en este paso entran las herramientas de *cloud computing*, tal como SageMaker de AWS, que provee muchas ventajas y facilidades para su implementación y puede ser utilizado en diversas aplicaciones para casos reales. La de nuestro interés es en la predicción de productos, sin embargo, se han realizado diversos trabajos con esta herramienta.

Por ejemplo, para la implementación de un modelo, utilizando la red neuronal Long Short-Term Memory para predicción de clima, después de haber entrenado el modelo, se puede utilizar de una manera muy rápida y sencilla sin que perdamos tiempo en configuraciones tediosas, debido a que SageMaker ofrece la facilidad de que el modelo pueda utilizarse a través de una Interfaz de Programación de Aplicaciones (API, siglas en inglés), accesible desde cualquier sistema que estemos desarrollando. (Hoang et al., 2020)

Podemos concluir que, al aplicar una arquitectura de servicios, generaríamos un proceso efectivo que implica bajos costos. Tendríamos un

proceso automatizado de actualización de *dashboards*, pero también al utilizar servicios de S3 podemos utilizar Amazon Glue y SageMaker, para presentación y análisis de los datos almacenados con anterioridad. De esta forma, se puede mejorar el sistema en que se aprovechan los datos para los *dashboards*.

3. PLANTEAMIENTO DEL PROBLEMA

La información en el entorno empresarial es importante para tener una ventaja ante la competencia y posicionarse en el mercado. La empresa productora y comercializadora de productos alimenticios en Guatemala tiene un sistema de visualización de información, en el cual se basan los gerentes para la toma de decisiones. Estos *dashboards* se actualizan en intervalos de tiempo que no son de provecho para la empresa, porque el proceso de actualización del repositorio de información, en donde se almacenan los datos, es poco eficiente; el tiempo que toma en realizarse este proceso es muy tardado, por lo que no puede hacerse en periodos muy cortos. La información que se obtiene al final del proceso no es la ideal; en la mayoría de casos no provee un análisis que ayude a predecir qué productos tendrán mayor demanda en distintas temporadas del año.

El tener una actualización lenta del *dashboard* y una mala predicción de productos demandados en el mercado, se debe a que en la empresa se maneja una gran cantidad de datos, los cuales se analizan de una forma poco eficiente. El modelo de datos no es el óptimo para generar información a partir de él; aparte de ello, se desaprovechan los datos históricos de la empresa porque solamente se toman periodos cortos para el análisis de datos.

El proceso de actualización de *dashboards*, que inicia desde la extracción hasta la presentación de los datos, se realiza manualmente en la mayoría del proceso.

La organización de los recursos computacionales a disposición no está especializada para buscar patrones de información entre gran cantidad de datos, por lo que la tarea de procesamiento y análisis tiene una presentación tardía.

Los problemas mencionados anteriormente no ayudan a tomar decisiones óptimas porque no beneficien al crecimiento del negocio, por no contar con toda la información analizada de forma efectiva.

Por aparte, también representa la pérdida de oportunidades de ventas de la empresa, que es un efecto negativo del proceso actual, que está relacionado con tener mayores pérdidas por materia prima no vendida. Además, los clientes no son estimulados para comprar los productos y tampoco se les ofrece productos en temporadas en las que aumenta la demanda. De esta forma, la empresa se ve afectada por la competencia directa y nuevos competidores que surgen en el mercado.

Con una mala arquitectura y herramientas no especializadas solo se pueden realizar análisis superficiales de datos y no un análisis profundo, en el que se puedan visualizar patrones de demanda de productos, así como información de utilidad para la toma de decisiones.

Por lo anterior, se plantean las siguientes preguntas de investigación:

3.1. Pregunta central

¿De qué forma se puede aumentar la efectividad del proceso de actualización de *dashboards* y predicción de productos más solicitados, en una empresa comercializadora de alimentos de Guatemala?

Partiendo de ello, se tienen las siguientes preguntas auxiliares que apoyarán en el desarrollo de esta investigación.

3.2. Preguntas auxiliares

- ¿Qué se necesita para la implementación del nuevo proceso de actualización de *dashboards* y la predicción de resultados, en una empresa comercializadora de alimentos en Guatemala?

- ¿Con qué herramientas se puede realizar la predicción de los productos más solicitados por los clientes de una empresa comercializadora de alimentos de Guatemala?

- ¿Cómo se puede identificar la efectividad del proceso de actualización de *dashboards* y predicción de productos más solicitados por los clientes en una empresa comercializadora de alimentos de Guatemala?

4. JUSTIFICACIÓN

El trabajo de investigación está bajo la línea de investigación de minería y análisis de datos. En los negocios, los procesos manuales de actualización de *dashboards* implican tiempo y esfuerzo de los colaboradores; no contar con información actualizada en ocasiones hace perder ventaja a la empresa ante la competencia, por no tomar las decisiones adecuadas para cada situación. En ese sentido, la implementación de un prototipo para aumentar la efectividad en el proceso de actualización de *dashboards* beneficiaría directamente a la empresa comercializadora de alimentos en Guatemala. Así mismo, la implementación de herramientas y tecnologías que provee AWS para realizar el análisis de información ayudaría a conocer de mejor forma el mercado.

Con el rediseño del proceso de actualización de *dashboards* y la implementación de herramientas de AWS, se logrará tener información actualizada en tiempos cortos, sin la necesidad de invertir mucho tiempo para realizar el proceso manual de procesamiento de información. Por otra parte, la evaluación del modelo de predicción de productos de la empresa contra los datos históricos de la empresa podrá darnos certeza del modelo, el cual puede ser utilizado para implementar estrategias en el negocio y, una vez implementado, la empresa podrá seguir afinando el modelo para lograr mejores resultados de predicción.

La aplicación de este prototipo ayudaría en la reducción de costos de mano de obra y tiempo, al realizar un proceso más efectivo que nos permita obtener resultados en menor tiempo y con menor esfuerzo. También será de

apoyo en la toma de decisiones para la predicción de productos que el cliente puede llegar a necesitar.

El desarrollo de este trabajo de investigación también puede servir de ejemplo para que otras empresas del sector implementen una infraestructura en la nube, así como tecnologías para análisis de datos de distintas fuentes y ser analizados para varios objetivos, con el fin de generar sistemas de información útiles para los negocios.

Por lo tanto, es importante realizar este trabajo de investigación para implementar un proceso eficiente de extracción, transformación y carga de datos para repositorios en la nube, y tener actualizados los sistemas de reportería, en una empresa comercializadora de alimentos de Guatemala, con el objetivo de mejorar su competitividad. El proceso puede aplicar a pequeñas o medianas empresas que busquen una solución similar, para manejar y actualizar su información.

5. OBJETIVOS

5.1. General

Implementar un prototipo para aumentar la efectividad del proceso de actualización de *dashboards* y predicción de los productos más solicitados por los clientes, en una empresa comercializadora de alimentos de Guatemala.

5.2. Específicos

- Implementar una arquitectura en la nube y efectuar una reingeniería y automatización del proceso de actualización de *dashboards*, en una empresa comercializadora de alimentos de Guatemala.

- Implementar herramientas que provee AWS, para el análisis de información y predicción de los productos más solicitados por los clientes de una empresa comercializadora de alimentos de Guatemala.

- Evaluar la efectividad del proceso de actualización de *dashboards* y predicción de productos más solicitados por los clientes, en una empresa comercializadora de alimentos de Guatemala.

6. ALCANCES

El enfoque de este trabajo de investigación es descriptivo y explicativo, abordando la optimización y reingeniería de un proceso de actualización de *dashboards* por medio de la extracción, transformación, carga y presentación de la información, utilizando los servicios de AWS. También se desarrollará un modelo predictivo de productos potenciales para los clientes.

6.1. Perspectiva de investigación

- Describir la reingeniería y automatización del proceso de actualización de *dashboards*, utilizando la nube como herramienta para la optimización y mejora del proceso.

- Descripción del desarrollo de un modelo de productos potenciales para los clientes, utilizando los servicios de AWS para el proceso que va desde el entrenamiento hasta la exposición del modelo.

- Comparación del uso de herramientas y tecnologías en la nube, para validar la efectividad del proceso de actualización de *dashboards* y del modelo diseñado.

6.2. Perspectiva técnica

- Diseño e implementación de un proceso eficiente y automatizado de extracción, transformación, carga y presentación de *dashboards* para el negocio, partiendo de sus bases de datos locales.

- Implementar una arquitectura orientada a servicios en la nube, utilizando AWS como proveedor para el proceso de ETL y análisis de información.

- Utilizar servicios de AWS para el análisis de información y generar un modelo de predicción de productos potenciales para los clientes.

6.3. Perspectiva de resultados

Prototipo para aumentar la efectividad del proceso de actualización de *dashboards* y análisis de información, identificando productos potenciales para los clientes en una empresa de alimentos de Guatemala. El proyecto está enfocado en ejecutar una reingeniería del proceso de actualización de *dashboards* por medio de la integración de tecnologías y servicios en la nube; también se desarrollará un modelo de predicción de productos potenciales para los clientes.

Para ello se tienen las funciones siguientes del prototipo:

- Reingeniería del proceso de actualización de *dashboards*.

- Proceso automatizado y efectivo para la actualización de *dashboards*.

- Uso de la arquitectura orientada a servicios para la implementación del proceso encargado del ETL y el análisis de información.

- Presentación de *dashboards* con Amazon QuickSight para el uso en usuarios finales.

- Uso de los servicios de AWS para la generación de un modelo de predicción de productos potenciales para el cliente.

- Validación del modelo de predicción de productos potenciales para los clientes, por medio de un conjunto de datos de prueba.

7. NECESIDADES A CUBRIR Y ESQUEMA DE SOLUCIÓN

La empresa de alimentos de Guatemala requiere información actualizada para la toma de decisiones, tener *dashboards* actualizados; se debe realizar este proceso de una forma automatizada para tener la información en el momento que sea necesario. Pero en la empresa se ejecuta un proceso poco efectivo de actualización de *dashboards*, porque en ocasiones los empleados deben generar archivos intermedios entre la base de datos y la herramienta con la que generan *dashboards*.

Para mitigar el problema causado por un proceso lento y manual de actualización de *dashboards* en la empresa comercializadora de productos alimenticios de Guatemala, se implementará un prototipo para la actualización de *dashboards* y predicción de productos potenciales para los clientes. Se utilizarán los servicios de AWS para la implementación del sistema.

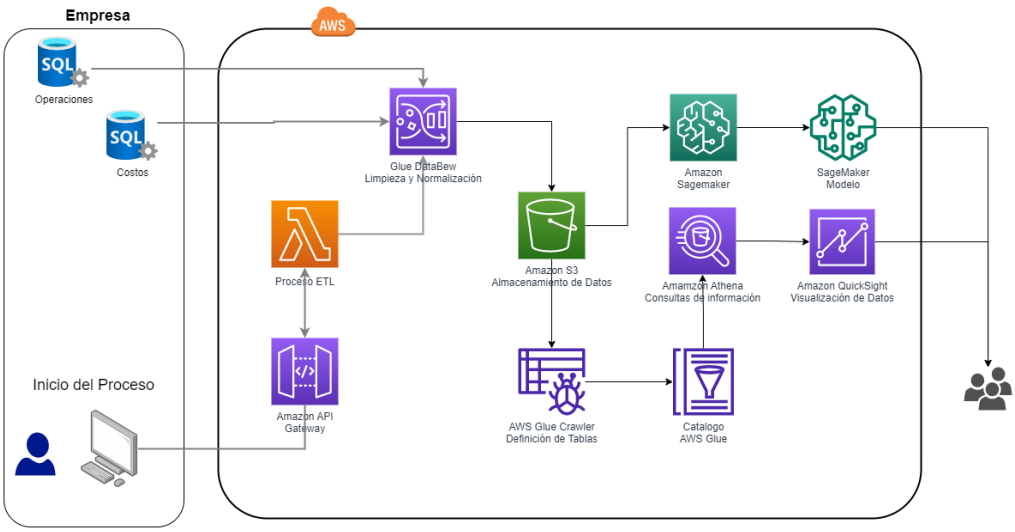
Además, se hará una reingeniería del proceso de actualización de *dashboards*, utilizando un proveedor en la nube para manejar todo el proceso y análisis de datos, evitando que las personas tengan que descargar la información del sistema de ventas. Se condensará la información para después utilizarla como base en los *dashboards* y otros análisis.

Se implementará una arquitectura en la nube, utilizando como proveedor AWS. Algunos de los servicios que se utilizarán son API Gateway, funciones Lambda, almacenamiento en *buckets* de S3, AWS Glue para el proceso de ETL y Amazon QuickSight para la presentación de datos como *dashboards*. Todo

estará conectado para compartir información, tanto como entrada para análisis o como información lista para usuarios finales.

Figura 1.

Arquitectura del proceso de actualización de dashboards



Nota. Flujo del proceso de actualización de datos, generación de *dashboards* y de un modelo de recomendación por medio de herramientas de AWS, utilizando servicios *serverless* para la arquitectura. Elaboración propia, realizado con draw.io.

El proceso utilizará principalmente una arquitectura orientada a servicios como se muestra en la figura 1; se hará uso de AWS Glue para todo el proceso de limpieza y transformación de datos y para aplicar catálogos de datos que usará Amazon Athena.

El proceso inicia cuando el usuario, a través de una interfaz en un sistema que ya posee la empresa, inicia la gestión por medio de la llamada a una función lambda.

El servicio de AWS Glue DataBrew se conectará directamente con las bases de datos de la empresa, para realizar la limpieza y normalización de los datos. Seguido de ello, se almacenará la información en el servicio de S3, tanto los datos de las operaciones de la empresa como los costos.

El servicio de AWS Glue Crawler es un rastreador que accederá a los archivos almacenados en S3, extraerá metadatos y creará definiciones de las tablas en AWS Glue Data Catalog.

Amazon Athena nos dará la capacidad de realizar consultas en lenguaje SQL desde los modelos de datos que se crearán con AWS Glue, por lo que podremos mostrar datos de importancia después del proceso de limpieza y normalización, listo para la última etapa, que consiste en la visualización de datos que se realizará con Amazon QuickSigth.

Con esto se pretende tener una arquitectura eficiente de costos bajos. Un factor importante es que la arquitectura está orientada a servicios, por lo cual solamente se pagará por el uso de cada servicio según el tipo de pago que impone AWS. Con esto se reducirá el tiempo invertido en procesos manuales y se lograrán los objetivos de poseer un proceso limpio y automatizado, para obtener *dashboards* actualizados que provean información útil en la toma de decisiones.

8. MARCO TEÓRICO

8.1. Arquitecturas *cloud*

A continuación, se presentan algunas arquitecturas que se pueden implementar con diversos proveedores de servicios en la nube, enfocándonos en arquitecturas sin servidor.

8.1.1. Arquitectura *serverless*

Las arquitecturas *serverless* se refieren a un sistema en la que siguen existiendo servidores; sin embargo, los desarrolladores ya no se preocupan por la administración de estos. La plataforma *serverless* se encarga de los temas referentes a la cantidad de servidores, así como a la capacidad, la cual se aprovisiona a medida que se requiera según la carga de trabajo. (Baldini et al., 2017)

En este tipo de arquitectura se encuentran las funciones lambda de AWS, la cual permite codificar funciones que pueden ser llamadas en cualquier momento; no se necesita configurar un servidor, solamente se codifica la función; dentro de ella permite llamar o conectarse a otros servicios de AWS, tales como DynamoDB o S3.

8.1.2. Arquitectura microservicios

Podemos definir como arquitectura de microservicios a aquella aplicación en donde todos sus módulos son microservicios, es decir, que está organizada

por componentes (Dragoni et al., 2017). Estas arquitecturas tienen un nivel bajo de acoplamiento. Cada componente se encarga de una tarea específica del sistema; estos pueden ser llamados por los demás servicios por medio de peticiones HTTP.

Una de las grandes ventajas de este tipo de arquitectura es que en caso de que algún microservicio falle, los demás siguen operando, lo cual robustece el sistema porque solamente afecta a los componentes asociados al microservicio que esté fallando.

Los microservicios se basan en el concepto de modularización, en el que cada microservicio es implementado y operado de forma independiente, que ofrece acceso a su lógica interna y datos por medio de una interfaz de red bien definida. (Jamshidi et al., 2018)

8.1.3. Arquitectura con servicios administrados por AWS

Según el sitio web de AWS (2023), AMS (Servicios Administrados por AWS) ayuda con la integración de los servicios de AWS a escala y a operar de una mejor forma y con una mayor seguridad. Ofrece seguridad, disponibilidad, eficiencia, resiliencia y conformidad. AMS está enfocado en permitir que los usuarios se centren directamente en el giro de negocio y en la innovación.

AWS provee diversas categorías de servicios que se complementan para construir arquitecturas complejas, que funcionan a través de la interconexión de servicios. De esta forma, podemos diseñar una arquitectura específica para las necesidades del negocio de manera más rápida.

Según Chung (2023), el beneficio que tiene AMS de AWS es que permite implementar una arquitectura en la que podemos conectarnos a bases de datos locales, ayuda al descubrimiento automático de datos con AWS DataBrew, transformación y catalogación de datos y análisis de datos sin licencias de terceros.

8.2. Manejo de datos

Para el manejo de datos hay diversas formas de procesar y de almacenar la información. Se puede optar por uno o varios tipos de procesos y almacenamientos, según sea el objetivo que se desee alcanzar; si deseamos utilizar los datos para inteligencia de negocios y también para ciencia de datos podríamos implementar un *data lake* y posteriormente un *data warehouse*.

8.2.1. Proceso de ciencia de datos

Según Kotu & Beshpande (2019), el proceso de la ciencia de datos se compone de varios pasos, en los que se puede realizar de forma iterativa. Se mencionan a continuación los primeros cuatro pasos que tienen mayor relevancia:

- Conocimiento previo: este paso contribuye a definir el problema que se quiere resolver y la información que se necesita para la resolución del problema. Se puede decir que el proceso de la ciencia de datos comienza con una necesidad, una pregunta o un objetivo de negocio. Se debe tener un planteamiento del problema bien definido, debemos conocer el problema que deseamos solucionar. Sin un panorama claro del problema al que se enfrenta, será muy complicado obtener un

conjunto de datos correctos y, por lo tanto, también será difícil elegir el algoritmo correcto para la solución del problema.

- Preparación de datos: este paso es en el que se trabaja con los datos. Es uno de los que consume mayor tiempo porque los datos generalmente no vienen de la forma que lo requieren los algoritmos de ciencias de datos, por lo que en este paso se debe estructurar la información según lo requiera el algoritmo que se vaya a utilizar. Si tenemos la información en distinta estructura debemos aplicar funciones, filtros, transformación de datos, entre otros métodos, con el fin de tener la información en el formato requerido.

- Modelado de datos: partiendo de que el modelo es una representación de los datos y las relaciones existentes entre ese mismo conjunto de datos, en este paso del proceso se construye el modelo por medio de datos de entrenamiento, después se evalúa con un conjunto de datos de prueba; en caso de que el modelo sea poco preciso, se puede iterar en el paso de construcción del modelo para obtener el modelo final.

- Aplicación: en este paso el modelo se libera a producción para que pueda ser utilizado en el negocio; en esta parte también se ven temas como la preparación para producción, la integración técnica, tiempo de respuesta, mantenimiento del modelo y asimilación.

Este proceso nos da una guía del orden de los pasos que debemos seguir para lograr implementar modelos, con el fin de resolver alguna necesidad o con el objetivo de descubrir patrones según sea el caso, basado en el giro del negocio.

Los procesos de ciencia de datos en general son iterativos. Estando en cierta etapa se puede retornar a las previas para afinar la precisión del modelo que se requiera.

8.2.2. Proceso de extracción, transformación y carga

El proceso de ETL sirve para la integración de datos, generalmente para realizar análisis y hacer reportes sobre los datos finales. Según Martínez et al. (2013), cada uno de los pasos se define de la siguiente manera:

- Extracción: es el proceso de selección de datos; pueden provenir de distintas fuentes, ya sean sistemas transaccionales, archivos de texto u hojas de cálculo. Todos los datos que se van seleccionando se cargan en memoria y se conocen como datos crudos, que serán utilizados para la siguiente fase.

- Transformación: la entrada a esta fase son los datos crudos; se realiza la limpieza, personalización, cálculos sobre datos y funciones de agregación. El fin de este proceso es generar información normalizada, datos formateados, estructurados y resumidos, según sea la finalidad y las necesidades del negocio.

- Carga: en esta etapa se toman como entrada los datos procesados en la fase anterior, con la finalidad de la inserción de los datos estructurados hacia un *data warehouse* o algún otro repositorio de datos.

8.2.3. Repositorios de datos

Los repositorios de datos son útiles para el almacenamiento, principalmente para la organización de la información; se crean con el fin de recuperar y trabajar con dicha información.

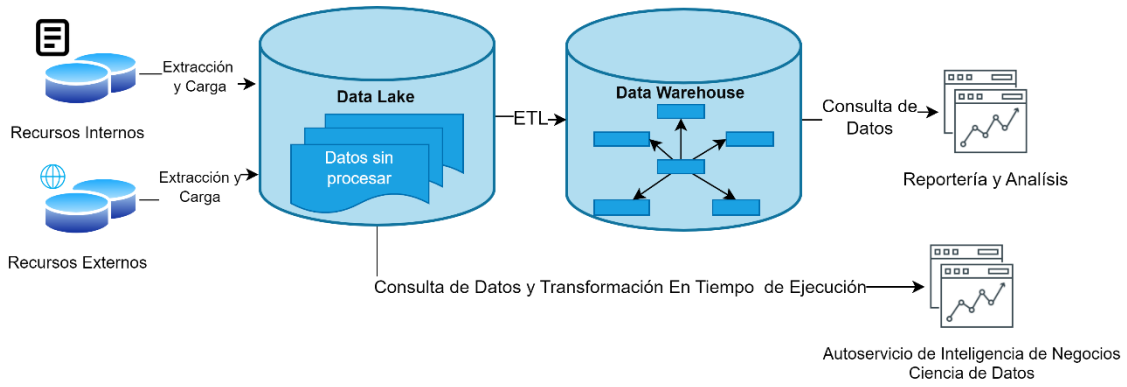
8.2.3.1. *Data lake*

Según Llave (2018), son varios los propósitos de un lago de datos. Pueden funcionar como un área de ensayo o como fuente para un *data warehouse*, además de funcionar como plataforma de experimentación y como fuente directa de inteligencia empresarial de autoservicio.

Un lago de datos puede contener gran cantidad de información estructurada o no estructurada, tal como se ve en la figura 2, proviene de una extracción y una carga de datos, pero no de una transformación de datos, esto debido a que los datos generalmente son almacenados en su estado original.

Figura 2.

Implementación de un data lake



Nota. Diagrama de implementación de un *date lake* en conjunto con un *data warehouse* para análisis y reportería. Adaptado de M. Llave. Data lakes in business intelligence: reporting from the trenches [Lagos de datos en inteligencia empresarial: informes desde las trincheras]. *Procedia Computer Science. 138* (1), p. 521. (https://doi.org/10.1016/j.procs.2018.10.071)

8.2.3.2. *Data mart*

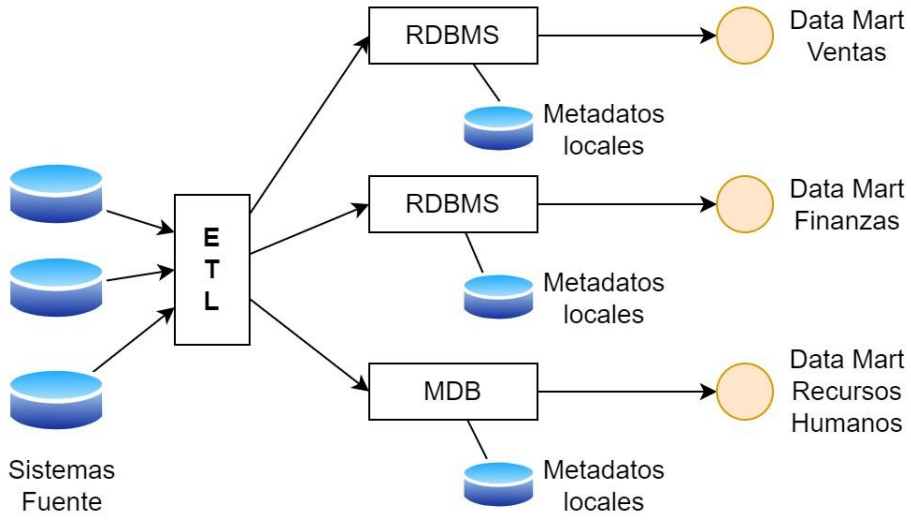
Un *data mart* es una colección de datos perteneciente a un área más específica dentro de una organización; apoya en la toma de decisiones según el área al que pertenece el *data mart* (Ranjan, 2009). Con esto podemos definir un *data mart* como un subconjunto de datos de toda la organización, que está especializado para una parte de la compañía. Podría ser que esté enfocado solamente en las ventas, en los costos de producción o por departamentos. Esta separación se realiza según sean las necesidades del negocio, en relación con lo que les interesa analizar o reportar.

Con un *data mart* puede ser más rápido realizar un análisis sobre cierta área del negocio, esto debido a que solo se toma el conjunto de datos de interés, por lo que estaría más optimizado para poder extraer la información.

En la figura 3 tenemos la arquitectura de un *data mart*. Según Sahama & Croll (2007), en donde inicialmente se tienen los sistemas y se almacena la información, seguido de ellos está el proceso de extracción, transformación y carga, que distribuye la información hacia Sistemas de Administración de Bases de Datos Relacionales (RDBMS, siglas en inglés), en los cuales ya solamente se encuentra la información relacionada a un área de la empresa. En este caso, hay *data marts* de ventas, finanzas y recursos humanos, lo que puede ser de beneficio en cuanto a optimización y solamente se tienen los datos que esa área de la empresa utilizará.

Figura 3.

Arquitectura de un data mart



Nota. Diagrama de la utilización de varios *data marts* para cada área de la organización. Adaptado de T. Sahama & P. Croll. A data warehouse architecture for clinical data warehousing [Una arquitectura de almacén de datos para el almacenamiento de datos clínicos]. *Conferences in Research and Practice in Information Technology Series. 68* (1). p. 230. (https://www.researchgate.net/publication/27473559_A_Data_Warehouse_Architecture_for_Clinical_Data_Warehousing)

8.2.3.3. *Data warehouse*

Un *data warehouse* es un gran repositorio de datos. Estos pueden provenir de una o varias fuentes con el fin de integrarse y almacenarse de una forma estructurada; todo esto con el fin de apoyar en la toma de decisiones (Nambiar & Mundra, 2022). Por lo tanto, para que los datos lleguen a almacenarse en un *data warehouse* pasan por un proceso de extracción, transformación y carga de la información; los datos se guardan filtrados y procesados, que sirven para cierto propósito trazado por el negocio.

Según Nambiar & Mundra (2022), un *data warehouse* contiene información histórica del negocio y se definen tres tipos de arquitecturas; son las siguientes:

- Arquitectura de un nivel: este tipo de arquitectura tiene solamente una capa, no separa el procesamiento analítico del transaccional, por lo que el almacenamiento es bajo debido a que no hay redundancia y los datos se almacenan una sola vez; sin embargo, el rendimiento es bajo por ser una misma capa en donde se opera con los datos.

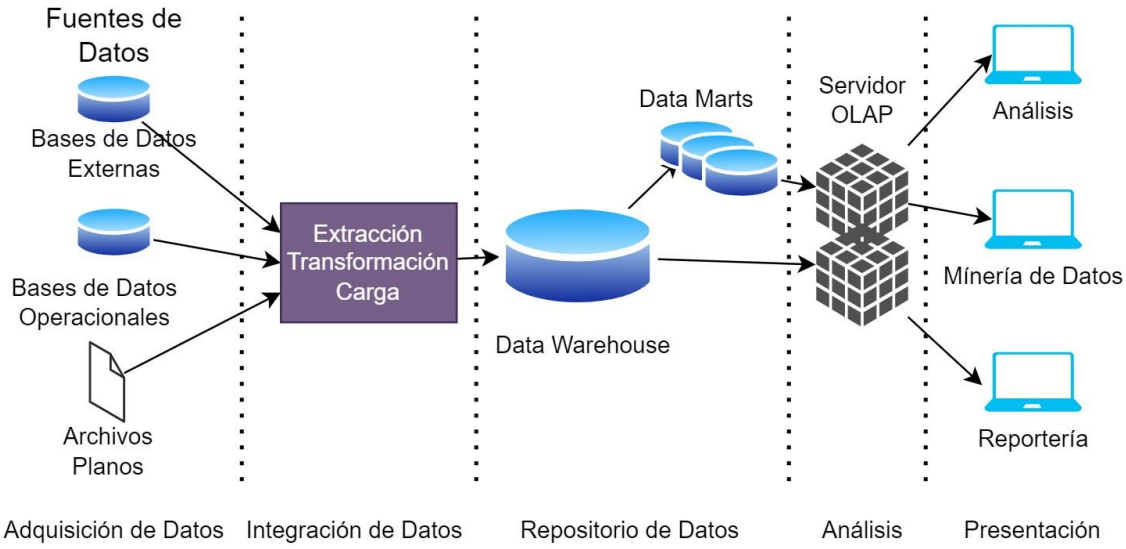
- Arquitectura de dos niveles: existen dos capas, una es para la preparación de los datos, en donde se procesan y se limpian para pasarlos a la siguiente capa. En la siguiente capa se encuentra el almacén de datos que serán accesibles para los usuarios finales. Este tipo de arquitectura puede llegar a tener problemas de conectividad según la calidad de la red, debido a la conexión entre las capas.

- Arquitectura de tres niveles: en estas arquitecturas hay tres niveles: el nivel inferior, medio y superior. En el nivel inferior se lleva a cabo el proceso de extracción, transformación y carga de información: los datos se limpian y procesan. En el nivel medio está el servidor de Procesamiento Analítico en Línea (OLAP, siglas en inglés), el cual es una representación de la base de datos que está más orientada a ser entendible por los usuarios finales. Por último, está el nivel superior, que es la capa del cliente; en esta capa están todas las herramientas con las que se puede acceder al repositorio de datos, para extraer la información y ser de utilidad para el análisis y toma de decisiones del negocio.

En la figura 4 podemos ver una arquitectura de un *data warehouse* con varias capas; se tiene la inicial de adquisición de datos, en donde están las fuentes de datos, después se realiza el proceso de ETL hacia un *data warehouse*. Lo interesante es que podemos generar *data marts* usando como fuente el *data warehouse*. Después va hacia un servidor OLAP, listo para ser utilizado por la última capa, que es la de presentación y para los usuarios finales.

Figura 4.

Arquitectura data warehouse de varias capas



Nota. Diagrama de capas de una arquitectura *data warehouse*, desde la adquisición hasta la presentación de la información, usando cubos OLAP. Adaptado de E. Saddad et al. Lake data warehouse architecture for big data solutions [Arquitectura de almacén de lago de datos para soluciones de big data]. *International Journal of Advanced Computer Science and Applications*. *11* (8). p. 418. (https://doi.org/10.14569/IJACSA.2020.0110854)

8.2.4. Almacenamiento y análisis de datos en AWS

Para el almacenamiento y análisis de datos existen varios servicios en AWS, algunos de ellos son Amazon S3 y Amazon Redshift.

8.2.4.1. Amazon Simple Storage S3

Según el sitio web de AWS (2022), es un servicio enfocado al almacenamiento de objetos; este servicio se adapta a cualquier sector, por lo que tanto pequeñas como grandes empresas pueden almacenar sus datos a un bajo costo. Los beneficios que ofrece este servicio son escalabilidad, disponibilidad, seguridad y rendimiento.

Uno de los usos que se le puede dar a este servicio es como de un lago de datos, debido a que se puede almacenar distintos tipos de datos, estructurados o no estructurados, para posteriormente utilizarlos para realizar análisis por medio de algoritmos de *machine learning* o para otros usos como la generación de informes históricos.

8.2.4.2. Amazon Redshift

Según el sitio web de AWS (2022), Amazon Redshift permite el análisis y creación de modelos de *machine learning*; se enfoca en convertir grandes cantidades de datos en información útil. Puede analizar varios tipos de datos, pueden ser estructurados o no estructurados, al igual que lagos de datos, a base de hardware y *machine learning*.

Un factor importante de Amazon Redshift es que también puede funcionar sin servidor, lo que indica que no nos preocupamos por la

administración detallada de dónde o cómo se almacenan los datos y solamente se cobra por el espacio ocupado.

Dependiendo del tamaño del negocio puede llegar a representar un costo elevado, en especial si es una pequeña o mediana empresa: soporta grandes cantidades de datos. Para grandes negocios puede llegar a ser una herramienta muy potente.

8.3. *Machine learning*

Existen varias categorías de *machine learning*, que son útiles para efectuar predicciones.

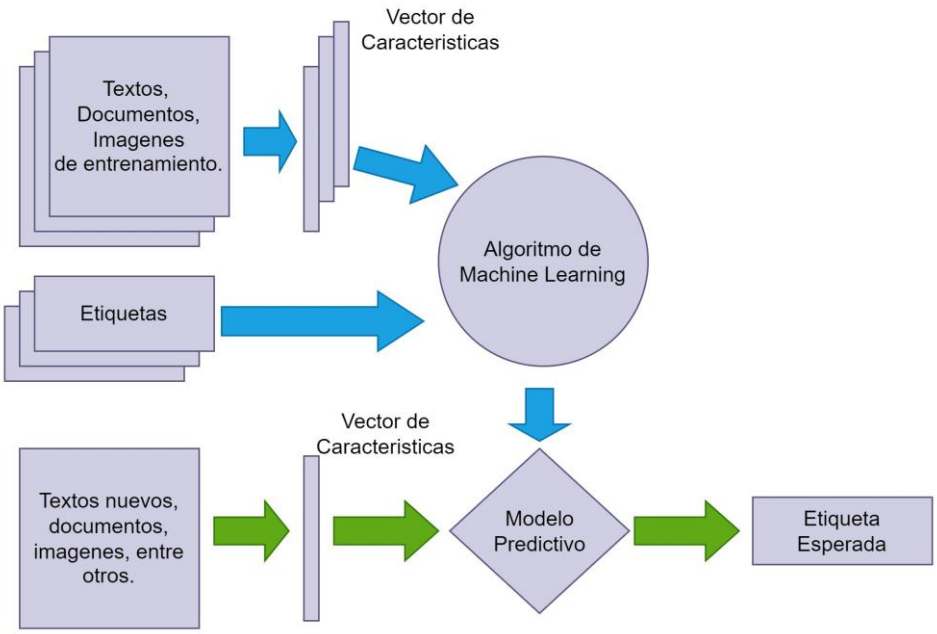
8.3.1. Aprendizaje supervisado

El aprendizaje supervisado se refiere a las variables de entrada y también las variables de salida; el objetivo es usar un algoritmo con el fin de mapear una función basado en las entradas y salidas. El propósito es que cuando se tenga un nuevo dato de entrada, la función pueda predecir el valor de salida. (Brownlee, 2016)

El objetivo de estos algoritmos es generar un estimador que pueda predecir la etiqueta de un objeto, basado en sus características (Nasteski, 2017). Por ello y con base en un conjunto de características, podemos obtener un resultado ya sea categórico o valor real, apegado al comportamiento descubierto por el algoritmo.

Figura 5.

Ejemplo del proceso completo del aprendizaje supervisado



Nota. Diagrama de aprendizaje supervisado, partiendo de datos y etiquetas, se implementa un algoritmo para generar un modelo predictivo. Adaptado de V. Nasteski. An overview of the supervised machine learning methods [Una descripción general de los métodos de aprendizaje automático supervisados]. *Horizons. 4.* p. 5. (https://doi.org/10.20544/HORIZONS.B.04.1.17.P05)

El proceso general de un algoritmo de aprendizaje supervisado se muestra en la figura 5. A partir de un conjunto de datos que pueden ser textos, documentos, imágenes u otros tipos de datos, se convierten en vectores de características, a cada uno de estos registros se le asigna su etiqueta correcta. Tanto los vectores de características como las etiquetas se ingresan al algoritmo seleccionado de aprendizaje supervisado, para generar el modelo predictivo.

Para probar el modelo se tiene otro conjunto de datos de entrada, el cual se convierte en vectores de características y se ingresa al modelo para obtener la etiqueta según los patrones aprendidos por el modelo.

El aprendizaje supervisado se puede dividir según el resultado, por lo que pueden ser de clasificación y de regresión.

8.3.1.1. Clasificación

Si el dato de salida toma un valor categórico es considerado un modelo de clasificación (Brownlee, 2016). Los modelos de aprendizaje supervisado correspondientes a este grupo pueden clasificar los elementos basados en sus características. Una de las implementaciones podría ser clasificar a un cliente basado en sus gustos, que en este caso las características del cliente serían la entrada y, según el modelo, lo clasificaría con una de las etiquetas previamente definidas

8.3.1.2. Regresión

Cuando el dato de salida toma un valor numérico real como el valor de una moneda, es considerado modelo de regresión (Brownlee, 2016). Estos modelos tienen muchas posibles aplicaciones en todo tipo de entornos: en un entorno empresarial puede ser de utilidad, por ejemplo, para predecir la cantidad de producto que se debe solicitar a los proveedores basado en datos históricos del negocio, esto debido a que el resultado proporcionado por el modelo es un dato numérico real. También se puede predecir la demanda o los productos potenciales para los clientes, que es uno de los objetivos de este trabajo de graduación.

8.3.2. Algoritmos de aprendizaje supervisado

A continuación, se describen varios algoritmos de aprendizaje supervisado.

8.3.2.1. Algoritmo Random Forest

Es un algoritmo clasificador que está integrado por una serie de estructuras tipo árbol. Son predictores de tipo árbol, en la que cada árbol depende de los valores de un vector aleatorio muestreado de forma independiente; cada árbol tiene la misma distribución de probabilidad. Es importante mencionar que el error de generalización de este algoritmo depende de la fortaleza y correlación entre los árboles. (Breiman, 2001)

Este tipo de algoritmos está basado en árboles de decisión; son útiles cuando necesitamos implementar soluciones en las que se requiera clasificar los datos por categorías u otro agrupador. También se puede implementar para problemas de regresión en los que necesitamos predecir valores, basándonos en datos históricos.

8.3.2.2. Algoritmo XGBoost

Extreme Gradient Boosting es un algoritmo de aprendizaje supervisado, basado en árboles de decisión, con la peculiaridad de que tiene un proceso llamado impulso, con el cual se logran obtener modelos que tengan mayor grado de precisión; el boosting se refiere a crear muchos modelos de forma secuencial y, en cada nuevo modelo, se corrigen deficiencias del modelo anterior. (Mitchell & Frank, 2017)

Este algoritmo representa una mejora con el concepto del impulso, por lo que es bueno utilizarlo en casos en los que se busca solucionar problemas de clasificación o regresión.

8.3.2.3. Algoritmo de regresión lineal

Este tipo de algoritmos hace una separación en clases, de los vectores de entrada al modelo para el entrenamiento. El objetivo de este algoritmo es agrupar elementos basado en sus características, esto se logra a través de clasificar los elementos basado en el valor que produce la combinación lineal de las características; se considera uno de los clasificadores más rápidos. (Osisanwo et al., 2017)

Es un algoritmo en el que se puede establecer una relación lineal entre una variable de entrada y una de salida. También es utilizado para la predicción de la demanda, basándose en características del producto o de los clientes, dependiendo de las necesidades del negocio sobre la relación que se quiere llegar a conocer.

8.3.2.4. Algoritmo de regresión logística

Este tipo de algoritmos sirve, mayormente, para casos en los que se quiere un resultado de dos opciones; la probabilidad de que el evento de interés ocurra es la razón de la probabilidad de que dividido por la probabilidad de que el evento no ocurra. (LaValley, 2008)

En este caso los algoritmos de regresión logística podrían ser útiles en casos cuando se requiera tener resultados en dos categorías, por ejemplo, para saber si un cliente compraría o no cierto producto que saldrá al mercado. De

esta forma se identifica si el producto será bien recibido o habrá que tomar acciones como aumentar el marketing o, en caso extremo, evitar lanzar ese producto porque no será bien recibido.

8.3.2.5. Media Móvil Integrada Autorregresiva

Son algoritmos de modelos lineales capaces de hacer una representación de series de tiempo estacionarias y no estacionarias; sirven para pronosticar series de tiempo y no asume que exista algún patrón en los datos históricos de la serie que se quiere pronosticar (Gahirwal, 2013). Este modelo es conocido como Autoregresive Integrated Moving Average (ARIMA, siglas en inglés).

Este tipo de algoritmos nos ayuda a predecir valores futuros, basados en datos que se han obtenido con anterioridad, cuyos datos no necesariamente deben tener una tendencia lineal; sin embargo, funciona mayormente con datos estacionarios. Los datos estacionarios indican que no tendrá mayor variación a través del tiempo, sino que se mantiene estable.

8.3.3. Aprendizaje no supervisado

En el aprendizaje supervisado se tienen etiquetas preestablecidas para el entrenamiento del modelo. La categoría de algoritmos de aprendizaje no supervisado es aquella que descubre patrones y relaciones subyacentes en los datos, sin necesidad de referencias o etiquetas. (Celecia et al., 2021)

En la categoría de aprendizaje no supervisado existen diversos algoritmos complejos para buscar patrones en un conjunto de datos, es útil cuando se tienen grandes cantidades de datos y se requiere conocer el patrón

que siguen los datos sin que se tengan los datos con etiquetas, sino que el algoritmo identifica los patrones.

8.3.4. Amazon SageMaker

Según el sitio web de AWS (2023), Amazon SageMaker es un servicio para crear modelos de *machine learning* con muchas ventajas, debido a que permite acceder a grandes cantidades de datos estructuras o no estructurados; permite automatizar los procesos de creación de modelos, y también se reducen los tiempos de entrenamiento con la infraestructura optimizada.

Esto hace que Amazon SageMaker sea una herramienta útil para la creación de modelos sin montar una infraestructura para el entrenamiento del modelo, sino que AWS provee la infraestructura y la opción de acceso a los datos, haciendo que haya un ahorro en costos de generar un modelo.

SageMaker provee una integración sencilla a las fuentes de datos, por lo que es una ventaja al momento del entrenamiento del modelo, dado que podemos tener un conjunto de datos en Amazon S3, que también es un servicio de bajo costo.

9. PROPUESTA DE ÍNDICE DE CONTENIDOS

10. METODOLOGÍA

10.1. Descripción del problema

El trabajo de investigación es de tipo cuantitativo, está orientado a evaluar la efectividad de un proceso de actualización de *dashboards* por medio del tiempo que toma la extracción, transformación y presentación de información. Por aparte, se genera y evalúa la precisión de un modelo de predicción de productos potenciales para los clientes, por medio de herramientas de generación de modelos.

10.2. Diseño

El diseño es experimental porque por medio de las variables de tiempo se medirá el cambio en el proceso de actualización de *dashboards*. Así mismo, la generación del modelo que identificará los productos potenciales para los clientes es experimental, ya que se evaluará la precisión del modelo, partiendo de darle datos de entrada al modelo y validar la correcta predicción.

10.3. Alcance

El enfoque de este trabajo de investigación es descriptivo y explicativo, por la optimización y reingeniería del proceso de actualización de *dashboards* y la generación del modelo para identificar productos potenciales para los clientes.

10.4. Variables

Las variables que se utilizarán en el trabajo de investigación serán las siguientes:

Tabla 1.

Variables de investigación

Variables	Definición	Subvariables	Indicadores
Tiempo de actualización	Es el tiempo que tardan los colaboradores en realizar el proceso de actualización de *dashboards*.	Tiempos por cada subproceso.	Medición del tiempo del nuevo proceso.
Tasa de éxito en el proceso de actualización.	Relación de la cantidad de actualizaciones exitosas por la cantidad total de actualizaciones.	Cantidad de éxitos. Cantidad total de iteraciones.	Conteo del número de procesos finalizados correctamente.
Tasa de error en el proceso de actualización.	Relación de la cantidad de actualizaciones fallidas por la cantidad total de actualizaciones.	Cantidad procesos fallidos. Cantidad total de iteraciones.	Conteo del número de procesos finalizados con error.
Precisión del modelo	Indica la precisión que tiene el modelo en cuanto a la predicción de datos.		El accuracy generado por el modelo.

Nota. Variables e indicadores para la implementación del prototipo, serán de ayuda para medir si se alcanzan los objetivos planteados. Elaboración propia, realizado con Excel.

10.5. Fases del estudio

El estudio tendrá cinco fases y por último se redactará el informe final.

10.5.1. Fase uno: medición del sistema manual y definición del nuevo proceso

En esta fase se medirá el tiempo que tardan los empleados en realizar el proceso de actualización de *dashboards*. Con el fin de conocer el tiempo actual y compararlo con el tiempo al implementar el nuevo proceso, se tomarán varias mediciones de tiempo.

En esta parte del proceso se aplicará:

- Toma de valores de tiempo del proceso actual.
- Media aritmética para obtener el tiempo de actualización representativo.
- También se elaborará el servicio con API Gateway y Lambda, para iniciar el procedimiento de actualización de *dashboards* y que se actualice de forma automática.

Los puntos relevantes de esta fase son:

- Tiempo medio de actualización de *dashboards* para posterior análisis.
- Creación de la función de actualización en caso de ser necesaria.
- Opción de llamar a la función desde el sistema del negocio.

10.5.2. Fase dos: extracción, transformación y carga de información

Para la implementación del nuevo proceso de actualización de *dashboards* se hará la preparación de la información. Esto se realizará con la obtención de los datos, principalmente se extraerán los datos importantes para el análisis de ventas que es uno de los enfoques del estudio.

Al tener la información preparada se conectará a los servicios de AWS, en donde se realizará la transformación de los registros, se hará la limpieza de datos en caso de ser necesario y se tendrán los datos preparados para la siguiente fase.

En esta fase también se almacenarán los datos en el servicio de S3 que provee AWS. Los datos se organizarán en archivos para facilidad de análisis y entrenamiento del modelo de identificación de productos potenciales para los clientes. Este repositorio de información servirá como datos de entrada para los servicios de Amazon SageMaker y Amazon Glue Crawler, que se encargará de convertir los archivos en tablas para que sea consultada y luego mostrada en los *dashboards*.

La técnica de recolección de información para esta fase es:

- Base de datos, extrayendo principalmente los datos esenciales para realizar la actualización de los *dashboards* con información de utilidad. La información obtenida también serán los datos para el entrenamiento y pruebas del modelo predictivo que se generará.

Las tareas que se deben ejecutar en esta fase son las siguientes:

- Extracción de la información desde la base de datos.
- Limpieza y normalización de los datos con servicios de AWS.
- Carga de información en los servicios de AWS.

Al tener una arquitectura orientada a servicios administrados por AWS no se implementarán servidores adicionales para este proceso, sino se utilizarán los servicios que nos provee AWS para cada paso del proceso.

10.5.3. Fase tres: generación de *dashboards*

En esta fase nos apoyaremos en el servicio de Amazon Athena para la consulta de información en las tablas temporales que se crearon con AWS Glue Crawler. Después se hará uso de AWS QuickSight para la generación de *dashboards*, de los cuales, el usuario final podrá acceder para aprovechar este nuevo proceso de actualización y presentación de *dashboards*.

Esta fase se elaborará tomando en cuenta la fase anterior en la que se almacena la información en AWS; para ello se trabajará con lo siguiente:

- Repositorio de datos almacenados en AWS en la fase de carga.
- Se elaborará un diagrama entidad relación para el modelado de datos.

Se desarrollarán las siguientes acciones para tener *dashboards* actualizados de forma eficiente:

- Implementación del servicio AWS Glue Crawler para definición del modelo de datos.
- Implementación de un catálogo de tablas con AWS Glue.
- Consulta de información con Amazon Athena.
- Visualización de datos con Amazon QuickSight.

10.5.4. Fase cuatro: generación del modelo

Al tener los datos almacenados en AWS S3, se procederá con el entrenamiento del modelo con Amazon SageMaker, el cual nos facilita la creación y exposición de modelos de *machine learning*. El modelo tendrá un conjunto de datos de entrenamiento y otro conjunto de datos de prueba con el

fin de validar la precisión del modelo. De esta forma se conocerá si los datos con los que se entrenó el modelo son suficientes y tendrá predicciones efectivas en la mayoría de los casos.

La información que se tomará en esta fase es la siguiente:

- Datos almacenados en AWS S3 provenientes de la base de datos inicial. Los datos se prepararán con el fin de que el algoritmo los trabaje de forma más eficiente.

Los puntos principales para tomar en cuenta en esta fase son:

- Preparación de un conjunto de datos para entrenamiento del modelo.
- Preparación de un conjunto de datos de prueba.
- Implementación del algoritmo de análisis predictivo.
- Validación del modelo
- Generación de matriz de confusión.
- Exposición de la API para consumo del modelo.

Se aplicará un algoritmo predictivo para la elaboración del modelo; el algoritmo será de la categoría de aprendizaje supervisado.

10.5.5. Fase cinco: evaluación del modelo

Se efectuará la comparación entre el tiempo que tardaban los empleados en realizar todo el proceso de actualización de *dashboards* comparado con el tiempo que les tomará el nuevo proceso. Esto se hará corriendo varias veces el proceso de actualización de *dashboards* para conocer también la tasa de éxito y error del nuevo proceso.

La forma de comparación será de la siguiente forma:

- Una tabla de valores antes y después del nuevo proceso.

Por otra parte, se mostrarán los datos generados por el modelo de *machine learning* para conocer si el entrenamiento fue efectivo. Esto se realizará por medio de los valores de precisión y la matriz de confusión, para identificar si es un modelo que logrará cumplir con el fin de identificar productos potenciales para los clientes.

La forma de mostrar la precisión del modelo será de la siguiente forma:

- Una tabla que muestre los valores generados por el modelo para identificar si es un modelo con un grado alto de precisión.

10.5.6. Redacción del informe final

Se elaborará el informe final contemplando todas las fases, iniciando con la recolección de datos para implementar el nuevo proceso de actualización de *dashboards* hasta la presentación de los mismos. Se elaborará de la siguiente forma:

- Proceso inicial del prototipo
  - Descripción del proceso actual
  - Toma de tiempo del proceso actual de actualización de *dashboards*
- Proceso de extracción, transformación y carga
  - Modelado de datos
  - Descripción de la extracción

    - Descripción de la transformación de datos
    - Diagrama de la carga de información
- Generación de *dashboards*
    - Descripción del proceso de actualización y generación de *dashboards*
    - Modelado de datos
    - Consulta de la información
    - Presentación de la información
- Generación del modelo de aprendizaje supervisado
    - Proceso de entrenamiento
    - Validación del modelo
    - Liberación para utilización del modelo
- Evaluación de resultados
- Conclusiones
- Recomendaciones

10.6. Técnicas de recolección de información

- Observación del tiempo de actualización: se efectuará la toma del tiempo que tardan los empleados en realizar el proceso de actualización de *dashboards* y se comparará con el tiempo que tardan con el nuevo proceso, cuando ya se haya implementado el prototipo, con el fin de identificar la variación del tiempo con el nuevo proceso.

- Base de datos: se extraerán principalmente los datos esenciales para realizar la actualización de los *dashboards* con información de utilidad. La información obtenida también serán los datos para el entrenamiento y pruebas del modelo predictivo que se generará.

11. TÉCNICAS DE ANÁLISIS DE LA INFORMACIÓN

Las técnicas que se utilizarán en el desarrollo del trabajo de investigación consisten en diagramas e implementaciones de estadística inferencial, por medio de programación con Python.

Las técnicas de análisis se aplicarán durante el desarrollo del trabajo de investigación y también en la fase de presentación de resultados, cuando ya se haya elaborado la arquitectura en la nube y se hayan implementado los servicios que servirán para la actualización de *dashboards* y para la generación del modelo de aprendizaje supervisado. Es decir, serán de apoyo para la validación de resultados. Las herramientas serán las siguientes:

- Diagrama de diseño de entidad relación para modelar la información de donde se extraerá la información para mostrarla en *dashboards*.

- Análisis descriptivo por medio de una tabla de tiempos, en el que los empleados tardan en actualizar información para los *dashboards* con el proceso normal y el tiempo que tardará el nuevo proceso en actualizar información.

- Algoritmo de predicción: el modelo se generará con algoritmos de predicción y servirá para la identificación de productos potenciales para los clientes; se hará con librerías de Python.

- Media aritmética para obtener el tiempo promedio que tardaba el proceso de actualización anterior contra la media aritmética del tiempo del nuevo proceso.

- Regresión lineal por medio del algoritmo XGBoost con el que se generará el modelo predictivo; esto se implementará con el servicio Amazon SageMaker. Se verificará la precisión del modelo para conocer qué tan efectivo será a la hora de realizar predicciones.

- Se obtendrá la matriz de confusión para conocer la cantidad de falsos positivos y falsos negativos que produce el modelo.

Los datos se presentarán en tablas, siendo las siguientes:

- Comparativa de tiempos de actualización antes y después.

- Precisión del modelo y cantidad de datos de entrenamiento y pruebas.

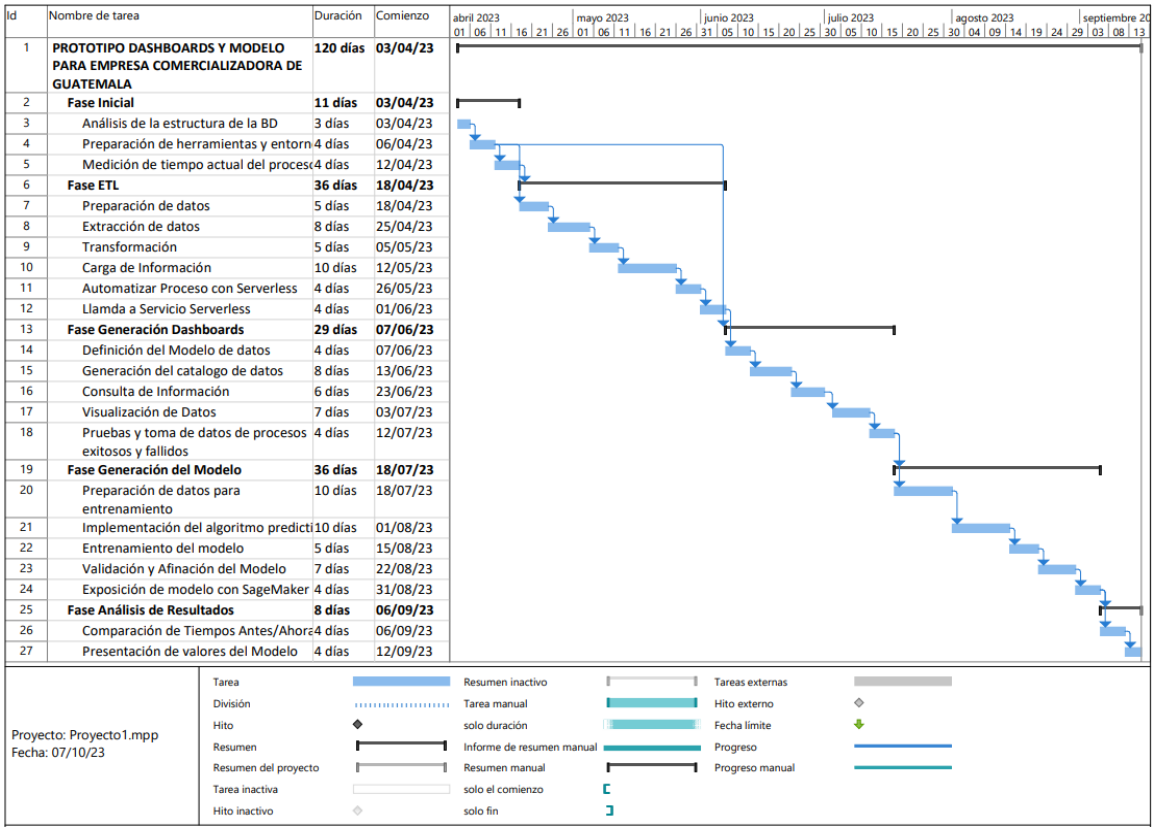
- Matriz de confusión

12. CRONOGRAMA

El cronograma de las actividades que se realizarán en el trabajo de investigación se presenta a continuación en la figura 6:

Figura 6.

Cronograma de actividades



Nota. Planificación de las actividades agrupadas por fases, que conlleva la implementación del prototipo. Elaboración propia, realizado con Microsoft Project.

13. FACTIBILIDAD DEL ESTUDIO

13.1. Factibilidad operativa

Para el desarrollo del proyecto de graduación no se dependerá de más recursos humanos, solamente del apoyo del asesor. Durante la elaboración del prototipo solamente se toman datos iniciales por medio de la observación del proceso actual de actualización de *dasboards*, para tener una referencia y compararla más adelante.

Se contará con infraestructura que se implementará por medio de una arquitectura de servicios administrados por AWS.

En factibilidad operativa contamos con lo necesario para la realización del proyecto.

13.2. Factibilidad técnica

Para este estudio se necesitan herramientas especializadas para el análisis de datos y la presentación de información; para ello, se utilizarán los recursos computacionales en la nube. Para el análisis de datos requiere de mayor capacidad de procesamiento para la aplicación de algoritmos predictivos; sin embargo, es un servicio que provee AWS, por lo que este estudio contará con las herramientas necesarias para llevarse a cabo.

En cuanto a conocimientos, se debe conocer de manera general cómo funcionan los servicios de AWS, y cómo pueden interactuar entre sí para lograr los objetivos.

En conclusión, el estudio cuenta con los recursos técnicos suficientes para el desarrollo del proyecto.

13.3. Factibilidad económica

El presente trabajo de investigación se ejecutará con recursos propios del estudiante de maestría. Siendo una investigación explicativa y descriptiva, se tendrán en cuenta los siguientes recursos:

Tabla 2.

Recursos necesarios para la investigación

Recurso	Costo
Almacenamiento	Q. 50.00
AWS Glue	Q. 200.00
AWS Athena	Q. 300.00
AWS QuickSight	Q. 300.00
AWS SageMaker	Q. 250.00
Asesor	Q. 2,500.00
TOTAL	**Q. 3,600.00**

Nota. Presupuesto necesario para la implementación del prototipo, se detallan los costos de los servicios de AWS que se utilizarán para elaborar el prototipo. Elaboración propia, realizado con Excel.

Siendo los recursos aportados suficientes para la investigación, se considera que es factible la ejecución del estudio.

Basado en la factibilidad operativa, técnica y económica, contamos con los recursos necesarios para elaborar y concluir de forma exitosa el proyecto de investigación, por lo cual se considera un proyecto factible.

REFERENCIAS

Amazon Web Services (s.f.). *AWS managed services* [Servicios administrados por AWS]. https://aws.amazon.com/es/managed-services/

Baldini, I., Castro, P., Chang, K., Cheng, P., Fink, S., Ishakian, V., Mitchell, N., Muthusamy, V., Rabbah, R., Slominski, A., & Suter, P. (2017). Serverless computing: current trends and open problems [Computación sin servidor: tendencias actuales y problemas abiertos]. *Springer, Singapore*. https://doi.org/10.48550/arXiv.1706.03178

Breiman, L. (2001). Random forests [Bosques aleatorios]. *Machine Learning, 45*, 5-32. https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2016). *Master machine learning algorithms discover how they work and implement them from scratch* [Domine los algoritmos de aprendizaje automático, descubra cómo funcionan e impleméntelos desde cero]. Machine Learning Mastery. https://datageneralist.files.wordpress.com/2018/03/master_machine_learning_algo_from_scratch.pdf

Bustamante, A., Galvis, E., y Gómez, L. (2013). Técnicas de modelado de procesos de ETL: una revisión de alternativas y su aplicación en un proyecto de desarrollo de una solución de BI. *Scientia Et Technica, 18*(1), 185-191. https://doi.org/10.22517/23447214.8727

Celecia, A., Figueiredo, K., Rodriguez, C., Vellasco, M., Maldonado, E., Silva, M., Rodrigues, A., Nascimento, R. & Ourofino, C. (2021). Unsupervised machine learning applied to seismic interpretation: towards an unsupervised automated interpretation tool [Aprendizaje automático no supervisado aplicado a la interpretación sísmica: hacia una herramienta de interpretación automatizada no supervisada]. *Sensors, 21*(19), 6347. https://doi.org/10.3390/s21196347

Chung, P., & Vartak, M. (11 de junio de 2021). *Building a cloud-based OLAP cube and ETL architecture with AWS Managed Services* [Creación de una arquitectura ETL y cubo OLAP basada en la nube con servicios administrados de AWS]. Amazon Web Services. https://aws.amazon.com/es/blogs/architecture/building-a-cloud-based-olap-cube-and-etl-architecture-with-aws-managed-services/

Dineva, K., & Atanasova, T. (2021). Design of scalable IoT architecture based on AWS for smart livestock [Diseño de arquitectura IoT escalable basada en AWS para ganadería inteligente]. *Animals, 11*(9), 2697. https://doi.org/10.3390/ani11092697

Dragoni, N., Giallorenzo, S., Lluch, A., Mazzara, M., Montesi, F., Mustafin, R., & Safina, L. (2017). Microservices: yesterday, today, and tomorrow [Microservicios: ayer, hoy y mañana]. *Present and Ulterior Software Engineering*, *4*, 195–216. https://doi.org/10.48550/arXiv.1606.04036

Gahirwal, M., & M., V. (2013). Inter time series sales forecasting [Pronóstico de ventas entre series temporales]. *Information Technology, Vivekanand Education Society's Institute of Technology*.

https://www.researchgate.net/publication/235761175_Inter_Time_Series_Sales_Forecasting

Hoang, D., Yang, P., Cuong, L., Trung, P., Tu, N., Truong, L., Hien, T. & Nha, V. (2020). Weather prediction based on LSTM model implemented AWS machine learning platform [Predicción del tiempo basada en el modelo LSTM implementado en la plataforma de aprendizaje automático de AWS]. *International Journal for Research in Applied Science & Engineering Technology (IJRASET), 8*, 283-290. https://doi.org/10.22214/ijraset.2020.5046

Huttunen, J., Jauhiainen, J., Lehti, L., Nylund, A., Martikainen, M., & Lehner, O. (2019). Big data, cloud computing and data science applications in finance and accounting [Aplicaciones de big data, computación en la nube y ciencia de datos en finanzas y contabilidad]. *ACRN Journal of Finance and Risk Perspectives 8*, 16-30. https://www.acrn-journals.eu/resources/SI08_2019b.pdf

Jain, N., & Srivastava, V. (2013). Data mining techniques: a survey paper [Técnicas de minería de datos: un estudio]. *International Journal of Research in Engineering and Technology*, *2*(11), 116-119. http://dx.doi.org/10.15623/ijret.2013.0211019

Jamshidi, P., Pahl, C., Mendonça, N., Lewis, J., & Tilkov, S. (2018). Microservices: the journey so far and challenges ahead [Microservicios: el viaje hasta ahora y los desafíos futuros]. *IEEE Internet Computing, 35*(3), 24-35. http://dx.doi.org/10.1109/MS.2018.2141039

Kotu, V., & Beshpande, B. (2019). *Data science concepts and practice* [Conceptos y práctica de la ciencia de datos]. Morgan Kaufman.

LaValley, M. (2008). Logistic regression [Regresión logística]. *Circulation, 117*(18), 2395 - 2399. https://doi.org/10.1161/circulationaha.106.682658

Lee, G. (2022). *Automatización de una pequeña empresa mediante un proceso de ciencia de datos.* [Tesis de máster, Universitat Politècnica de València]. Archivo digital. http://hdl.handle.net/10251/189021

Llave, M. (2018). Data lakes in business intelligence: reporting from the trenches [Lagos de datos en inteligencia empresarial: informes desde las trincheras]. *Procedia Computer Science*, *138*(1). 516-524. https://doi.org/10.1016/j.procs.2018.10.071

Mitchell, R., & Frank, E. (2017). Accelerating the XGBoost algorithm using GPU computing [Acelerando el algoritmo XGBoost usando computación GPU]. *PeerJ Computer Science*. https://doi.org/10.7717/peerj-cs.127

Mitra, A., Jain, A., Kishore, A., & Kumar, P. (2022). A comparative study of demand forecasting models for a multi-channel retail company: a novel hybrid machine learning approach [Un estudio comparativo de modelos de previsión de la demanda para una empresa minorista multicanal: un novedoso enfoque híbrido de aprendizaje automático]. *Operations Research Forum, 3*(4), 1-22. https://doi.org/10.1007/s43069-022-00166-4

Nambiar, A., & Mundra, D. (2022). An overview of data warehouse and data lake in modern enterprise data management [Una descripción general

del almacén de datos y el lago de datos en la gestión de datos empresariales moderna]. *Big Data Cognitive Computing, 6*(4), 132. https://doi.org/10.3390/bdcc6040132

Nasteski, V. (2017). An overview of the supervised machine learning [Una descripción general del aprendizaje automático supervisado]. *Horizons, 4*, 51-62. https://doi.org/10.20544/horizons.b.04.1.17.p05

Osisanwo, F., Akinsola, J., Awodele, O., Hinmikaiye, J., Olakanmi, O., & Akinjobi, J. (2017). Supervised machine learning algorithms: classification and comparison [Algoritmos de aprendizaje automático supervisados: clasificación y comparación]. *International Journal of Computer Trends and Technology, 48*(3), 128-138. http://dx.doi.org/10.14445/22312803/IJCTT-V48P126

Pavlyshenko, B. (2019). Machine-Learning models for sales time series forecasting [Modelos de aprendizaje automático para la previsión de series temporales de ventas]. *Data, 4*(1), 15. https://doi.org/10.3390/data4010015

Ranjan, J. (2009). Business intelligence: concepts, components, techniques and benefits [Inteligencia de negocios: conceptos, componentes, técnicas y beneficios]. *Journal of Theoretical and Applied Information Technology, 9*(1), 60-70. https://www.jatit.org/volumes/Vol9No1/9Vol9No1.pdf

Ravat, F., & Zhao, Y. (2019). Data lakes: trends and perspectives [Lagos de datos: tendencias y perspectivas]. *International Conference on Database and Expert*, *11706*(1), 304-313. https://doi.org/10.1007/978-3-030-27615-7_23

Saddad, E., Mokhtar, M., El-Bastawissy, A., & Hazman, M. (2020). Lake data warehouse architecture for big data [Arquitectura de lake data warehouse para big data]. *International Journal of Advanced Computer Science and Applications, 11*(8), 417-424. https://dx.doi.org/10.14569/IJACSA.2020.0110854

Sahama, T., & Croll, P. (2007). A data warehouse architecture for clinical data warehousing [Una arquitectura de almacén de datos para el almacenamiento de datos clínicos]. *Proceedings Australasian Workshop on Health Knowledge Management and Discovery, 68*(1), 227-232. https://www.researchgate.net/publication/27473559_A_Data_Warehouse_Architecture_for_Clinical_Data_Warehousing