LAGUNA

Santiago García Cifuentes

Nota de Aceptación

Certificamos que el presente Trabajo de Grado Satisface, en alcances y calidad, todos los requisitos que demanda un Trabajo de Grado de Maestría.

David Arango Londoño

_____
David Arango Londoño
Director Trabajo de Grado

_____
Gerardo M. Sarria Montemiranda
Jurado

_____
Hernán Darío Vargas Cardona
Jurado

Aprobado en cumplimiento de los requisitos exigidos por la Pontificia Universidad Javeriana Cali, para optar el título de Magister en Ciencia de Datos.

Camilo Rocha

_____
HERNÁN CAMILO ROCHA NIÑO Ph. D.
Decano Facultad de Ingeniería y Ciencias

_____
JUAN CARLOS MARTÍNEZ ARIAS
Director Posgrados de Ingeniería y Ciencias

Cali 27 Feb 2024

Acta de Correcciones al Documento de Trabajo de Grado

Santiago de Cali, 27 Feb 2024

Autor: Santiago García Cifuentes

Título del Trabajo de Grado: "LAGUNA"

Director: David Arango Londoño

Como indica el artículo 2.13 de las Directrices para Trabajo de Grado de Maestría, he verificado que el estudiante indicado arriba ha implementado todas las correcciones que los Jurados del Proyecto de Trabajo de Grado definieron que se efectuaran, como consta en el Acta de Evaluación correspondiente.

David Arango Londoño

_____

David Arango Londoño

C.C. 1130586950 de Cali

Santiago de Cali, 27 de Feb del 2024

Doctor
Diego Luis Linares Ospina
Director Maestría en Ciencia de Datos
Facultad de Ingeniería y Ciencias
Pontificia Universidad Javeriana de Cali

Cordial Saludo,

Con el fin de cumplir con los requisitos exigidos por la Universidad para optar por el título de Magíster en Ciencia de Datos, nos permitimos presentar a su consideración el proyecto denominado "LAGUNA", el cual fue realizado por el estudiante Santiago García Cifuentes con código 0065062 perteneciente a la Maestría en Ciencia de Datos, bajo la dirección de David Arango Londoño.

El suscrito director del Proyecto Aplicado autoriza para que se proceda a hacer la evaluación de este proyecto, toda vez que ha revisado cuidadosamente el documento y avala que ya se encuentra listo para ser presentado y sustentado oficialmente.

Atentamente,

Santiago García Cifuentes
C.C. 1112776252 de Cartago Valle

David Arango Londoño
C.C. 1130586950 de Cali

Documentación anexa:

Resumen del Proyecto Aplicado en formato digital (máximo 1 página).
Una copia digital (PDF) del documento del proyecto aplicado

FICHA RESUMEN

PROYECTO APLICADO – MAESTRÍA EN CIENCIA DE DATOS

LAGUNA:

1. ÁREA DE TRABAJO: Comercial.
2. TIPO DE PROYECTO: Aplicado.
3. ESTUDIANTE: Santiago García Cifuentes
4. CORREO ELECTRÓNICO: santiagogarciacifuentes@gmail.com
5. DIRECCIÓN Y TELÉFONO: Calle 71 Norte # 2B- 133 / 317 346 8962
6. DIRECTOR: David Arango Londoño
7. CORREO ELECTRÓNICO DEL DIRECTOR: david.arango@javerianacali.edu.co
8. PALABRAS CLAVE: Analítica de datos, TuChat, Las Ricuras de Sebastian, Análisis exploratorio, Limpieza de datos, Geocodificación, Análisis de sentimiento, Pronóstico ARIMA, Google Cloud, Looker Studio, Tablero de control.
9. FECHA DE INICIO: 22 Mayo 2022
10. DURACIÓN ESTIMADA: 12 meses
11. RESUMEN:
    Los chatbot son un modelo de interacción persona-computadora, es decir, es un programa informático diseñado para simular una conversación con usuarios humanos y esto es lo que TuChat ofrece a sus clientes. Sin embargo, sus clientes necesitan utilizar la información que es recolectada por medio de los chatbots para perfeccionar la estrategia comercial. De esta manera, este proyecto propuesto explora con uno de los clientes de TuChat que se llama Las Ricuras de Sebastian y aprovechando la información obtenida a través de los chatbots, logra organizar el flujo de los datos de la organización donde realizando un análisis exploratorio, limpieza y consolidación de las bases de datos, se logra obtener información más precisa y coherente. La identificación de horarios de alta y baja demanda a lo largo del día se realiza para optimizar la oferta de productos y promociones. Además, se implementan modelos avanzados, incluyendo geocodificación para ubicar geográficamente a los clientes, análisis de sentimiento para evaluar la satisfacción del cliente y un modelo ARIMA para pronosticar las ventas futuras. La visualización y acceso a los datos se facilita mediante Google Cloud y Looker Studio, culminando en la presentación de un tablero de control integral. Esta herramienta proporciona a Las Ricuras de Sebastian una representación visual clara de diversos indicadores clave, incluyendo resultados generales, análisis temporal y georeferenciación, permitiendo la toma de decisiones informadas y la mejora continua de las operaciones comerciales.

LAGUNA

Nombre del estudiante
0065062 - Santiago García Cifuentes

Proyecto Aplicado para optar al título de
Magister en Ciencia de Datos

Director
David Arango Londoño

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, ENERO 22 DE 2024

RESUMEN

Los chatbot son un modelo de interacción persona-computadora, es decir, es un programa informático diseñado para simular una conversación con usuarios humanos y esto es lo que TuChat ofrece a sus clientes. Sin embargo, sus clientes necesitan utilizar la información que es recolectada por medio de los chatbots para perfeccionar la estrategia comercial. De esta manera, este proyecto propuesto explora con uno de los clientes de TuChat que se llama Las Ricuras de Sebastian y aprovechando la información obtenida a través de los chatbots, logra organizar el flujo de los datos de la organización donde realizando un análisis exploratorio, limpieza y consolidación de las bases de datos, se logra obtener información más precisa y coherente. La identificación de horarios de alta y baja demanda a lo largo del día se realiza para optimizar la oferta de productos y promociones. Además, se implementan modelos avanzados, incluyendo geocodificación para ubicar geográficamente a los clientes, análisis de sentimiento para evaluar la satisfacción del cliente y un modelo ARIMA para pronosticar las ventas futuras. La visualización y acceso a los datos se facilita mediante Google Cloud y Looker Studio, culminando en la presentación de un tablero de control integral. Esta herramienta proporciona a Las Ricuras de Sebastian una representación visual clara de diversos indicadores clave, incluyendo resultados generales, análisis temporal y georeferenciación, permitiendo la toma de decisiones informadas y la mejora continua de las operaciones comerciales.

Palabras clave: Analítica de datos, TuChat, Las Ricuras de Sebastian, Análisis exploratorio, Limpieza de datos, Geocodificación, Análisis de sentimiento, Pronóstico ARIMA, Google Cloud, Looker Studio, Tablero de control.

ABSTRACT

Chatbots are a model of human-computer interaction, meaning it is a computer program designed to simulate a conversation with human users, and this is what TuChat offers to its clients. However, clients need to use the information collected through chatbots to refine their business strategy. In this way, this proposed project explores one of TuChat's clients called Las Ricuras de Sebastian, and leveraging the information obtained through chatbots, it manages to organize the flow of the organization's data. By conducting exploratory analysis, cleaning, and consolidation of databases, more precise and coherent information is obtained. Identification of high and low demand hours throughout the day is carried out to optimize product offerings and promotions. Additionally, advanced models are implemented, including geocoding to locate clients geographically, sentiment analysis to evaluate customer satisfaction, and an ARIMA model to forecast future sales. Data visualization and access are facilitated through Google Cloud and Looker Studio, culminating in the presentation of a comprehensive dashboard. This tool provides Las Ricuras de Sebastian with a clear visual representation of various key indicators, including overall results, temporal and georeferencing analysis, enabling informed decision-making and continuous improvement of business operations.

Keywords: Data Analytics, TuChat, Las Ricuras de Sebastian, Exploratory Analysis, Data Cleansing, Geocoding, Sentiment Analysis, ARIMA Forecast, Google Cloud, Looker Studio, Dashboard.

TABLA DE CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

1. INTRODUCCIÓN

TuChat, una empresa colombiana establecida en 2020 con sede en Cali Valle Colombia, se especializa en brindar soluciones de automatización a empresas que buscan optimizar procesos operativos, especialmente aquellas en el sector de alimentos y restaurantes. Centrándose en establecimientos con un promedio de 30 pedidos a domicilio diarios. TuChat ofrece una herramienta de chatbot que no solo simplifica la interacción con los clientes, sino que también se convierte en una fuente valiosa de datos sin procesar.

El enfoque de TuChat en restaurantes medianos, como “Las Ricuras de Sebastián”, responde a la necesidad de agilizar procesos en un sector que tradicionalmente dependía en gran medida de la atención telefónica. Las limitaciones operativas de atender pedidos únicamente por teléfono generaban costos innecesarios en nómina y afectaban la eficiencia en la gestión de pedidos. Aquí es donde surge la oportunidad para TuChat: implementar un chatbot en WhatsApp como un canal adicional de ventas, permitiendo a los clientes realizar pedidos de forma autónoma y eficiente, reduciendo la carga de trabajo asociada con las llamadas telefónicas.

La empresa cliente en este proyecto, Las Ricuras de Sebastián, un restaurante de comidas rápidas con 20 años de trayectoria, experimentó desafíos en la gestión de pedidos debido a la dependencia del teléfono fijo como único canal de atención. Con un promedio de 90 pedidos diarios, la necesidad de optimización se volvía evidente. La implementación del chatbot no solo liberó al restaurante de la necesidad de destinar personal exclusivo para atender el teléfono, sino que también abrió un nuevo canal de ventas más rápido y eficiente para los clientes, mejorando significativamente la experiencia general de hacer un pedido a domicilio.

En el contexto de este proyecto, denominado LAGUNA, se aborda la siguiente pregunta fundamental: ¿Cómo organizar y aprovechar los datos generados a través de las interacciones en el chatbot para convertirlos en información estratégica? La respuesta implica una transformación integral de los datos sin procesar almacenados en Google Cloud, en información visualmente accesible, brindando a los clientes de TuChat una herramienta analítica personalizada que facilita su toma de decisiones para el bien del negocio.

A lo largo de este proyecto, se exploró el enfoque metodológico, los desafíos encontrados y los resultados obtenidos en la implementación de esta solución, que no solo mejoró la eficiencia operativa sino que también proporcionó a los clientes una visión estratégica basada en datos para la toma de decisiones informadas.

2. DEFINICIÓN DEL PROBLEMA

2.1. PLANTEAMIENTO DEL PROBLEMA

En la era del Comercio Conversacional como lo menciona en su artículo E. Doudchitzky [1], donde la atención al cliente se ha expandido a través de diversos canales, la empresa TuChat, una empresa especializada en la implementación de chatbots, se enfrenta al desafío de aprovechar al máximo los datos generados por estos sistemas. Aunque ha logrado implementar con éxito chatbots en diferentes canales para optimizar la atención al cliente, la falta de un enfoque analítico integral impide la extracción de conocimientos estratégicos de los datos almacenados.

TuChat se encuentra en la encrucijada de organizar, procesar y visualizar eficientemente la información contenida en las conversaciones de los chatbots. Aunque la recolección y almacenamiento de datos se realizan de manera efectiva en Google Cloud, la carencia de un proceso analítico estructurado limita la capacidad de TuChat para proporcionar a sus clientes, como es el caso del restaurante “Las Ricuras de Sebastián”, de tener una visión completa y procesada de los datos relevantes para la toma de decisiones.

Esto implica reconocer la importancia crucial de un flujo de datos organizado como base fundamental en la estructura de una organización. En un entorno donde la información es un activo estratégico, la organización eficiente de los datos del chatbot se convierte en el punto de partida para construir herramientas avanzadas, como modelos de machine learning. La ingeniería de datos emerge como un componente esencial, facilitando la construcción y optimización de modelos predictivos y analíticos.

La importancia de esta transformación no solo radica en la mejora de la eficiencia operativa de TuChat sino también en las ventajas competitivas que brinda a empresas asociadas, como a “Las Ricuras de Sebastián”. La capacidad de procesar datos de manera eficiente permite una toma de decisiones más ágil y fundamentada, contribuyendo así a la mejora continua de las operaciones comerciales y a la adaptación proactiva a las demandas del mercado.

2.2 FORMULACIÓN DEL PROBLEMA

Ante esta situación, surge el principal interrogante: ¿Cómo estructurar y organizar un modelo de analítica de datos con un flujo de datos organizado como base fundamental en la estructura de TuChat que permita a sus clientes, como el restaurante “Las Ricuras de Sebastián”, aprovechar plenamente la información generada por los chatbots?.

Los datos que están actualmente dispersos en varias bases de datos almacenadas en Google Cloud, presentan un desafío en términos de desestructuración y aislamiento, dificultando su análisis y transformación en conocimiento estratégico.

Los problemas específicos a abordar incluyen la necesidad de consolidar la información de múltiples sedes de un cliente en una visualización única y comprensible, dando a conocer las métricas relevantes para un restaurante como:

- ¿Cuál es el total de las ventas de acuerdo a un rango de fechas?,
- ¿Cuál es la cantidad de pedidos?,
- ¿Cuál es el ticket promedio de las ventas?,
- ¿Cuál es la hora más frecuente de los pedidos?,
- ¿Cuál es el comportamiento de ventas de acuerdo a un rango de fechas?,
- ¿Qué tan frecuente es un cliente?,
- ¿Qué opinan y cómo se sienten los clientes con la experiencia del restaurante y su canal de venta que es el chatbot?
- ¿Cuál es el barrio donde más pedidos realizan los clientes por medio del chatbot al restaurante?

Por otro lado, para lograr obtener las métricas internas de TuChat como:

- ¿Cuál es el total de ventas?,
- ¿Cuál es la cantidad de pedidos en total y por cliente?,
- ¿Qué tan probable es que un cliente recomiende el restaurante, NPS?,

2.3 JUSTIFICACIÓN

Existe una necesidad de involucrar la tecnología y la digitalización para los negocios, como se evidenció durante la pandemia del COVID-19, donde las operaciones tradicionalmente físicas se transformaron digitalmente para continuar operando. Los sectores público y privado requirieron utilizar la tecnología para realizar actividades virtuales y conservar su productividad [27].

En este contexto, se plantea el reto de TuChat en este proyecto, el cuál es superar las limitaciones actuales en la estructuración y procesamiento de datos para ofrecer un valor agregado a las empresas asociadas y mejorar la eficiencia operativa interna de TuChat.

La importancia de abordar este problema radica en la creación de una base sólida para la toma de decisiones estratégicas. La capacidad de visualizar y comprender rápidamente los datos generados por los chatbots no solo mejorará la eficiencia operativa de TuChat, sino que también permitirá a las empresas asociadas, como “Las Ricuras de Sebastian”, obtener insights (descubrimientos) valiosos para optimizar sus procesos y mejorar la experiencia del cliente.

Un flujo de datos organizados como base fundamental en la estructura de una organización, se convierte en el punto de partida para construir herramientas avanzadas, como modelos de machine learning y la ingeniería de datos emerge como un componente esencial, facilitando la construcción y optimización de modelos predictivos y analíticos.

Este planteamiento busca no solo resolver la problemática actual sino también sentar las bases para un análisis de datos continuo y efectivo, proporcionando a TuChat y a sus clientes una herramienta integral para la toma de decisiones informadas.

3. OBJETIVOS DEL PROYECTO

3.1 OBJETIVO GENERAL

Desarrollar un modelo de analítica con un flujo de datos organizado como base fundamental, que potencie tanto para TuChat como para las empresas asociadas, cada uno de los datos que son almacenados actualmente y lograr convertirlos en información estratégica para el negocio.

3.2 OBJETIVOS ESPECÍFICOS

1. Realizar un análisis exploratorio de los datos para obtener las métricas e indicadores que generen valor a ambas partes del negocio.

2. Organizar y estructurar un repositorio que reciba y permita extraer todos los datos recopilados en el proceso del negocio y se puedan discriminar los datos de manera general como particular por cada negocio.

3. Elaborar un tablero de control que permita medir el comportamiento de las operaciones, mapas de georeferenciación de acuerdo a la ubicación de sus clientes, productos más vendidos, cálculo de métricas, calificación del nivel del servicio, sentimiento del cliente de acuerdo a su experiencia.

4. MARCO TEÓRICO

4.1. Empresa TuChat

TuChat es una empresa creada recientemente (2020), la cual se encuentra en etapa de crecimiento y evolución. Es un negocio B2B (Business to Business) donde este realiza la conexión con los clientes de sus empresas aliadas por medio de mensajería multicanal con robots hechos a la medida, logrando que la empresa potencie sus propios canales de comunicación con sus clientes como WhatsApp, Facebook, Instagram, Telegram, entre otros.

4.2. Comercio conversacional

Como lo menciona E. Doudchitzky [1] este concepto hace referencia a la aplicación de la comunicación por medio de chat en doble vía, es decir, el cliente contacta a la empresa o la empresa al cliente por medio de las apps de mensajería instantánea y el chatbot en tu e-commerce. Otra definición muy similar es la expuesta en el trabajo de grado de A.B. Molina donde indica que "es la incorporación de chats, mensajería u otras interfaces de lenguaje natural en el contexto de la comunicación bidireccional, para posibilitar la interacción entre personas, marcas o servicios y robots"[6].

En el siguiente gráfico se evidencia la dimensión masiva de las aplicaciones de mensajería en el mundo y su usabilidad en su momento.

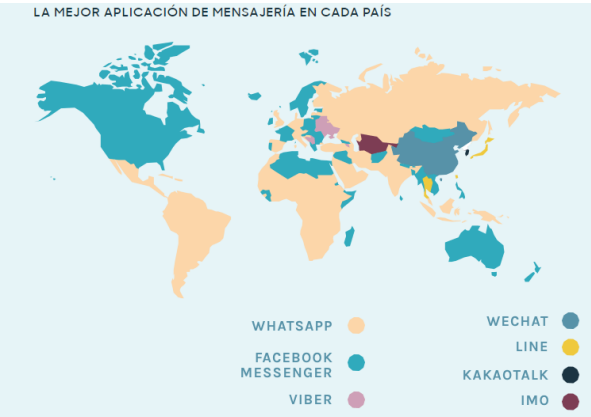


Figura 1: La mejor app de mensajería en cada país (2020) [7].

4.3. Ingeniería de Datos

La ingeniería de datos es una disciplina que se centra en el diseño, desarrollo y gestión de sistemas y arquitecturas para la recopilación, almacenamiento, procesamiento y análisis de datos. Su objetivo principal es asegurar que los datos estén disponibles, accesibles, confiables y preparados para ser utilizados en análisis y toma de decisiones. Este campo abarca diversas actividades, incluyendo la integración de datos, limpieza, transformación, carga (ETL), modelado de datos, y la implementación de sistemas de almacenamiento y gestión de bases de datos. La ingeniería de datos es esencial para garantizar la calidad y utilidad de los datos en entornos empresariales y científicos, especialmente en el contexto de la ciencia de datos y el análisis predictivo [25].

4.4. Bodega de datos (Data Warehouse)

Este es actualmente uno de los elementos más importantes para realizar inteligencia de negocios. Es aquí donde se almacena toda la información que una empresa u organización obtiene de sus diferentes fuentes [8].

W.H. Inmon, considerado el padre de las bodegas de datos en el año 1992, define los Data Warehouse como: "Un sistema orientado al usuario final, integrado, con variaciones de tiempo y sobre todo una colección de datos como soporte al proceso de toma de decisiones". Por otra parte, Ralph Kimball, considerado como uno de los más importantes precursores y padre del concepto Data Warehouse, lo define como: "una copia de los datos de la transacción estructurados específicamente para preguntar y divulgar" [9].

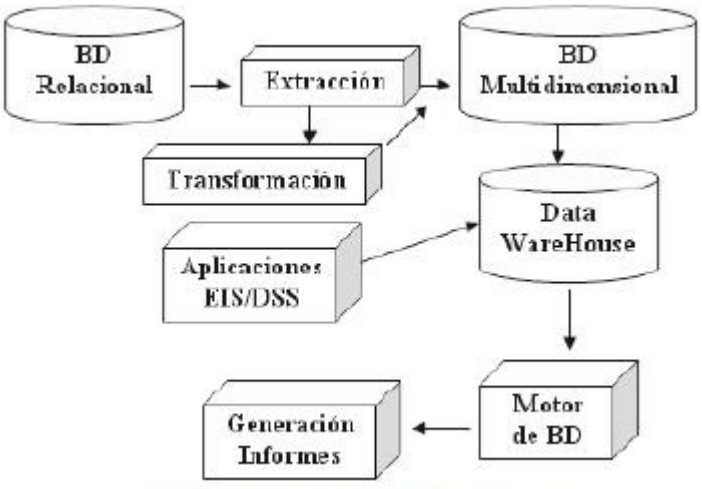


Figura 2: Proceso Data WareHouse [9].

4.5. Minería de textos

La minería de textos es un campo de la minería de datos que se centra en extraer patrones y conocimientos significativos a partir de datos no estructurados en forma de texto. Este proceso implica el análisis de grandes cantidades de documentos de texto para descubrir información oculta, relaciones, tendencias y patrones que pueden ser útiles para la toma de decisiones y la generación de conocimiento.

Existen varias categorías de minerías de textos: Extracción de información, recuperación de información, procesamiento de lenguaje natural, procesamiento de datos [4]. De acuerdo a lo mencionado, dentro de la minería de textos se tiene una categoría de análisis que se llama: **procesamiento del lenguaje natural** y dentro de este procesamiento se encuentran los modelos supervisados, no supervisados y de aprendizaje profundo [10]. En el desarrollo del proyecto se implementará uno de los modelos no supervisados basado en léxicones llamado Afinn para el desarrollo del análisis de sentimiento.

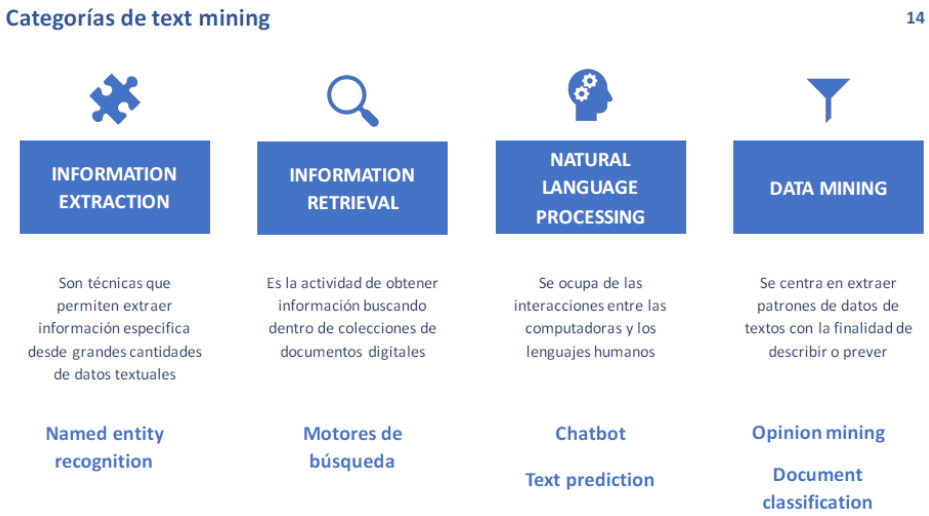


Figura 3: Categorías de la minería de textos [4].

4.5.1. Lenguaje Natural

Los datos textuales son datos no estructurados, pero generalmente pertenecen a un lenguaje específico que sigue una sintaxis y una semántica específica. Todos los datos de texto, como una palabra simple, una oración o un documento, se relacionan con lenguaje natural, la filosofía del lenguaje, la adquisición del lenguaje y el uso del lenguaje.

Para comprender el análisis de texto y el procesamiento del lenguaje natural, debemos comprender qué hace que un lenguaje sea “natural”. En términos simples, un lenguaje natural es un lenguaje desarrollado por humanos a través del uso y comunicación natural en lugar de construir y crear lenguaje artificial, como un lenguaje de programación de computadoras.[26].

4.5.2. Análisis de los sentimientos

El análisis de los sentimientos es quizás una de las aplicaciones más populares de procesamiento de lenguaje natural y análisis de texto, con una gran cantidad de sitios web, libros y tutoriales sobre este tema y parece funcionar mejor en textos subjetivos, donde las personas expresan opiniones, sentimientos y su estado de ánimo. Desde el punto de vista de la industria del mundo real, el análisis de los sentimientos se usa ampliamente para analizar encuestas corporativas, encuestas de retroalimentación, datos de redes sociales y reseñas de películas, lugares y productos básicos. La idea es analizar las reacciones de las sobre una entidad específica y tomar acciones perspicaces basadas en sus sentimientos.

Un corpus de texto consta de varios documentos de texto y cada documentos puede ser tan simple como una sola oración o tan complejo como un documentos completo con varios párrafos. Los datos textuales, a pesar de estar muy desestructurados, se pueden clasificar en dos grandes tipos de documentos. Los documentos fácticos suelen representar alguna forma de declaraciones o hechos sin sentimientos o emociones especificas asociadas a ellos. También se conocen como documentos objetivos. Los documentos subjetivos, por otro lado, expresan sentimientos, estados de ánimo, emociones y opiniones. Hay diferentes técnicas para analizar el sentimiento: Modelos basados en léxico no supervisados, modelos tradicionales de aprendizaje automático supervisado, modelos de aprendizaje profundo supervisado más nuevos y modelos avanzados de aprendizaje profundo supervisado [26].

4.6. Chatbot

Un chatbot es un programa de inteligencia artificial y un modelo de interacción persona-computadora (HCI), es decir, es un programa informático diseñado para simular una conversación con usuarios humanos, especialmente a través de Internet [3].

Se menciona que la historia de los chatbots comienza en la década de los años sesenta, como lo relata el artículo “De Eliza a Siri” [2] sobre la historia de los chatbots que comienza con Eliza, uno de los primeros programas de procesamiento del lenguaje natural, desarrollado en la década de 1960. Eliza tenía la capacidad de mantener conversaciones rudimentarias con usuarios, simulando ser un terapeuta conversacional.

Con el tiempo, la tecnología de asistentes virtuales avanzó gradualmente. Aparecieron diversos sistemas y plataformas que mejoraron la capacidad de comprensión del lenguaje natural y la interacción con los usuarios.

En 2011, Apple introdujo Siri, un asistente virtual personal integrado en los dispositivos iOS. Siri marcó un hito significativo al ofrecer una interfaz de usuario más avanzada y capacidades mejoradas de procesamiento de lenguaje natural.

El desarrollo de asistentes virtuales continuó evolucionando, con empresas como Google, Amazon y Microsoft lanzando sus propias versiones, como Google Assistant, Amazon Alexa y Cortana, respectivamente.

Hasta 2015, la tecnología de asistentes virtuales estaba presente en todas partes, integrándose en teléfonos inteligentes, altavoces inteligentes y otros dispositivos conectados. Estos sistemas eran capaces de realizar tareas variadas, desde responder preguntas hasta realizar acciones específicas mediante comandos de voz. Todo este auge trajo consigo una nueva era llamada la era del comercio conversacional y no hay una fecha exacta que marque el inicio concreto de la era del comercio conversacional, ya que es más un concepto evolutivo que ha ido ganando relevancia con el tiempo. Sin embargo, se puede decir que el término comenzó a ganar popularidad en la última década, especialmente a medida que las empresas empezaron a adoptar tecnologías de chatbot y mensajería para interactuar con los clientes.

El auge del comercio conversacional ha sido impulsado por avances tecnológicos en procesamiento de lenguaje natural, inteligencia artificial y la creciente preferencia de los consumidores por la mensajería instantánea como medio de comunicación. A medida que las empresas buscaban formas más efectivas de interactuar con sus clientes en línea, surgieron soluciones de comercio conversacional para facilitar la comunicación a través de diversos canales, como chats en sitios web, aplicaciones de mensajería y redes sociales.

Algunos ejemplos de chatbots exitosos en diferentes industrias son [5]:

Industria hotelera: Marriott International

Marriott International utiliza un chatbot en su página web y en su aplicación móvil para ayudar a los clientes a reservar habitaciones, hacer preguntas sobre el hotel y obtener información sobre las instalaciones.

Este chatbot es capaz de entender preguntas en lenguaje natural y proporcionar respuestas precisas y relevantes, lo que ha mejorado significativamente la experiencia del cliente y ha reducido el tiempo que los clientes necesitan para obtener información.

Industria de la moda: H&M

La popular marca de moda H&M utiliza un chatbot para ayudar a los clientes a encontrar ropa que se adapte a su estilo y preferencias.

Los clientes pueden interactuar con Kik a través de mensajes de texto y recibir recomendaciones de ropa que se ajusten a su presupuesto y estilo personal.

Industria Gastronómica

Empresas Colombianas como Frisby y Crepes & Waffles usan un chatbot en whatsapp para la toma de los pedidos, logrando llevar al cliente de inicio a fin con la selección de platos, bebidas, opciones personalizadas, ofrecer recomendaciones basadas en las preferencias del usuario o en los productos más populares. También, recopilan comentarios y opiniones de los clientes para mejorar la calidad del servicio.

Industria bancaria: Bank of America

Bank of America utiliza un chatbot llamado "Erica" que ayuda a los clientes con una amplia variedad de tareas bancarias, incluyendo el pago de facturas, la transferencia de dinero y la verificación de saldos de cuenta.

Gracias a Erica, los clientes pueden obtener ayuda con sus tareas bancarias en cualquier momento, desde cualquier lugar.

Industria de la salud:

Your.MD es un chatbot que utiliza inteligencia artificial para proporcionar información médica personalizada a los usuarios.

Los usuarios pueden interactuar con el chatbot a través de mensajes de texto y recibir recomendaciones de tratamiento y consejos de salud basados en sus síntomas y antecedentes médicos.

Your.MD ha demostrado ser especialmente útil para las personas que no tienen acceso a un médico o para aquellos que tienen preguntas de salud fuera de las horas de atención médica regulares.

4.7. Dashboards

Los dashboards o tableros de control, son herramientas visuales que permiten monitorear y analizar datos de manera comprensible para identificar rápidamente información valiosa para la toma de decisiones en las empresas. En el contexto de TuChat, se utilizarán dashboards para presentar métricas clave a los clientes, como Las Ricuras de Sebastián, y para el análisis interno de TuChat. Estos tableros proporcionarán visualizaciones intuitivas y personalizadas, facilitando la interpretación de datos complejos.

De acuerdo a lo mencionado en el artículo de la página cyberclick [11], las características sugeridas de un dashboard o tablero de control son las siguientes:

Características de un dashboard

Un dashboard es una especie de "resumen" que recopila datos de diferentes fuentes en un solo sitio y los presenta de manera digerible para que lo más importante salte a la vista. Estas son algunas de las características que debe tener este tablero de control:

Que sea personalizado: Un dashboard debe contener únicamente los datos relevantes para el negocio que se está evaluando.

Que sea visualmente agradable: La idea de un dashboard es que podamos obtener la información que buscamos a primera vista. Por ello, los datos se presentan en forma de gráficos y debemos contar con indicadores rápidos a través de claves de color, flechas hacia arriba o abajo logrando captar la atención del observador a primera vista.

Que sea práctico. La función principal de un dashboard siempre debe ser orientar las acciones, por ello, debe facilitar la información necesaria para que se pueda saber cuáles son los siguientes pasos a seguir para mejorar los resultados.

Hay varios programas de Business Intelligence (BI) ampliamente utilizados en el mercado, sin embargo, de acuerdo al artículo “Tamaño del mercado y análisis de acciones” de la página mordorintelligence en el 2023 los que mayor cuota de mercado tienen son: Power BI, Tableau y Looker [12].

Power BI

Es una colección de servicios de software, aplicaciones y conectores que funcionan conjuntamente para convertir orígenes de datos sin relación entre sí en información coherente, interactiva y atractiva visualmente. Sus datos podrían ser una hoja de cálculo de Excel o una colección de almacenes de datos híbridos locales y basados en la nube. Power BI permite conectarse con facilidad a los orígenes de datos, visualizar y descubrir qué es importante y compartirlo con cualquiera o con todos los usuarios que desee [13].

Tableau

Tableau es una plataforma de análisis visual que transforma la manera en que usamos los datos para resolver problemas. Además, permite a las personas y las organizaciones sacar el máximo partido de los datos [14].

Looker Studio

Looker es una plataforma de Business Intelligence que te permite visualizar grandes volúmenes de datos. Es la herramienta de Google Cloud Platform que se conecta con nuestras bases de datos y que permite visualizar estos datos en dashboards. Para una visión más global [15].

4.8. Flujo de datos en una organización

El flujo de datos en una organización generalmente es un proceso que involucra la captura, el almacenamiento, procesamiento, análisis, visualización y distribución de la información. A continuación, se presenta una descripción general del flujo de datos en una organización:

1. Captura de Datos:

El proceso comienza con la captura de datos desde diversas fuentes. Estas fuentes pueden incluir transacciones comerciales, interacciones con clientes, registros de empleados, sensores, redes sociales, encuestas, y más.
Los datos pueden ser tanto estructurados (como datos en una base de datos relacional) como no estructurados (como texto libre, imágenes, videos).

2. Almacenamiento de Datos:

Los datos capturados se almacenan en sistemas de almacenamiento de datos. Esto puede incluir bases de datos, data warehouses, data lakes u otros sistemas de almacenamiento según la naturaleza y el volumen de los datos.

3. Procesamiento de Datos:

Una vez que los datos están almacenados, pueden someterse a procesos de limpieza, transformación y enriquecimiento. Esto asegura que los datos sean coherentes, precisos y estén listos para el análisis.

Se aplican técnicas de integración para combinar datos de diferentes fuentes y asegurar la coherencia.

4. Análisis de Datos:

Los datos procesados se utilizan para realizar análisis. Esto puede incluir análisis descriptivos para entender patrones pasados, análisis predictivos para prever tendencias futuras, y análisis prescriptivos para sugerir acciones basadas en los resultados.

5. Visualización de Datos:

Los resultados del análisis se presentan a menudo a través de herramientas de visualización de datos como gráficos, tableros de control y reportes. Esto facilita la comprensión de los patrones y tendencias por parte de los usuarios no técnicos.

6. Toma de Decisiones:

Basándose en la información analizada y visualizada, los responsables de la toma de decisiones utilizan los datos para informar y respaldar decisiones estratégicas y operativas.

7. Distribución de Resultados:

Los resultados de los análisis se distribuyen a los diferentes departamentos y niveles de la organización para su uso. Esto puede implicar la automatización de informes, el acceso a tableros de control en tiempo real o la integración de resultados en sistemas específicos.

8. Retroalimentación y Mejora Continua:

Se recopilan comentarios sobre el proceso de análisis y las decisiones tomadas. Estos comentarios se utilizan para ajustar y mejorar continuamente el flujo de datos y los procesos analíticos de la organización.

Es importante señalar que el flujo de datos puede variar según la industria, el tamaño de la organización y los objetivos específicos. Además, en la era actual, donde la transformación digital es cada vez más relevante, la implementación de tecnologías emergentes como la inteligencia artificial y el aprendizaje automático también puede influir en la forma en que las organizaciones gestionan y utilizan sus datos [16].

En la siguiente figura se observa la estructura del flujo de datos en una organización

Figura 4: Notas de webinar de Data Camp. Flujo de datos en una organización.

4.9. Patrones Espaciales Georreferenciados

Este es un estudio de las coordenadas de los sitios en donde estas fueron tomadas y cuando el área de estudio es considerablemente grande se usa un geoposicionador para establecer dichas coordenadas (latitud y longitud). Un esquema general de datos georreferenciados es el siguiente:

Sitio	Latitud	Longitud	X_1	X_2	·	·	·	X_p
1	–	–	x_{11}	x_{12}	·	·	·	x_{1p}
2	–	–	x_{21}	x_{22}	·	·	·	x_{2p}
3	–	–	x_{31}	x_{32}	·	·	·	x_{3p}
·	–	–	·	·	·	·	·	·
·	–	–	·	·	·	·	·	·
·	–	–	·	·	·	·	·	·
n	–	–	x_{n1}	x_{n2}	·	·	·	x_{np}

Figura 4: Matriz datos georreferenciados [17].

4.10. Función Importrange

Es una de las funciones más importantes de Google Sheets, la cual responde a una de las preguntas más frecuentes entre los nuevos usuarios de Sheets: ¿Cómo traigo información de otras hojas en Google Sheets?. Su objetivo es generar un arreglo de datos. Éste es una tabla (que consta de filas y columnas) de valores.

Ejemplo de uso
IMPORTRANGE("https://docs.google.com/spreadsheets/d/abcd123abcd123", "hoja1!A1:C10")

IMPORTRANGE(A2,"B2")

Sintaxis
IMPORTRANGE(url_hoja_cálculo, string_rango)

url_hoja_cálculo: La URL de la hoja de cálculo desde la que se van a importar los datos.

El valor de url_hoja_cálculo debe estar entre comillas o hacer referencia a una celda que contenga la URL de una hoja de cálculo.
string_rango: String con el formato "[nombre_hoja!]Rango" (p. ej., "Hoja1!A2:B6" o "A2:B6") que indica el rango que se debe importar.

Cualquier actualización del documento fuente de IMPORTRANGE hará que todos los documentos receptores abiertos se actualicen y muestren una barra de carga verde. IMPORTRANGE también espera a que se completen los cálculos en el documento fuente antes de mostrar resultados en el documento receptor, incluso si no se debe realizar ningún cálculo en el rango fuente. [21].

4.11. Función QUERY

La función QUERY en Google Sheets no es solo una función; es un portal hacia el poder del lenguaje SQL. QUERY es una función exclusiva de Google Sheets que filtra y resume la información de alguna (o varias) hojas de cálculo.

Ejecuta una consulta sobre los datos con el lenguaje de consultas de la API de visualización de Google.

Ejemplo de uso
QUERY(A2:E6,"select avg(A) pivot B")

QUERY(A2:E6,F2,FALSO)

Sintaxis
QUERY(datos, consulta, [encabezados])

Datos: Rango de celdas en el que se hará la consulta.

Cada columna de datos sólo puede contener valores booleanos, numéricos (incluidos los tipos de fecha y hora) o de string.

En el caso de que una sola columna contenga datos mezclados, el tipo de dato mayoritario determina el tipo de datos de la columna para la consulta. Los tipos de datos de los valores que estén en minoría se consideran valores nulos [22].

4.12. Metodología CRISP-DM

Es una de las más empleadas actualmente para el desarrollo de proyectos de minería de datos. Sus siglas que significan Cross-Industry Standard Process for Data Mining, consta de seis etapas de desarrollo iterativas y éstas son:

1. Entendimiento del negocio.
2. Comprensión de datos.
3. Preparación de datos.

4. Modelado.
5. Evaluación.
6. Implementación.

La siguiente gráfica explica la dinámica del método CRISP-DM

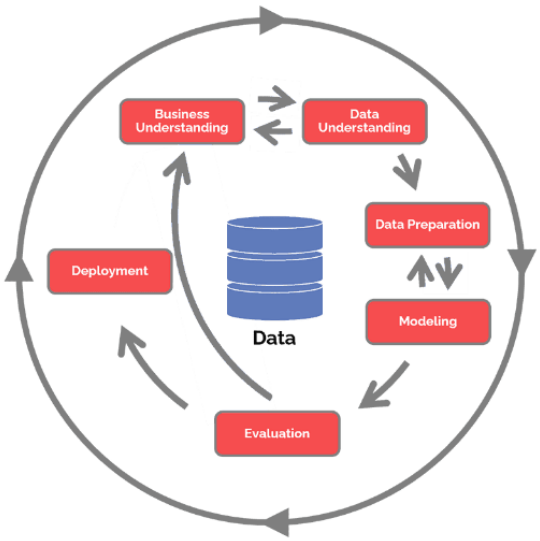


Figura 5: Ciclo de vida de CRISP-DM de la minería de datos [14].

Cada una de las etapas contiene unas actividades que orientan a cumplir el objetivo de cada una de ellas:

4.12.1. Entendimiento del negocio.

Esta es la etapa más importante, ya que si no se comprende ¿qué necesita el negocio? no servirán para nada las siguientes etapas. Las actividades dentro de esta etapa son:

1. **Identificación del problema:** Consiste en comprender la necesidad, restricciones, requisitos, beneficios entre otros.
2. **Determinación de objetivos:** Se deben especificar las metas a lograr, para poder proponer una solución desde minería de datos.
3. **Evaluación de la situación actual:** Especificar el estado antes de implementar la solución, para tener un punto de comparación que permita medir el éxito de lo implementado.

4.12.2. Comprensión de datos.

Las actividades principales de esta etapa son:

1. **Recolección de los datos:** Consiste en identificar las fuentes, si se debe utilizar técnicas de

recolección y obtener los datos a utilizar.

2. **Descripción de los datos:** Identifica el tipo, formato, volumetría y significado de cada dato.
3. **Exploración de los datos:** Consiste en realizar una exploración y conocer las características del conjunto de datos que se tiene para conocerlos lo mejor posible.

4.12.3. Preparación de los datos.

Normalmente es la etapa que consume mayor tiempo y es donde de acuerdo a las etapas anteriores se seleccionan los datos a utilizar y se hacen procesos de transformación para el uso en las siguientes etapas. Las actividades principales son:

1. **Limpieza de los datos:** Aplicación de técnicas para eliminar caracteres especiales, convertir en minúscula, eliminar espacios innecesarios, eliminar palabras que no agregan valor (stop words), normalizar valores, lematizar, stokenización, etc.
2. **Construir datos:** Obtener un dato que es efecto de la combinación de otros.
3. **Transformación de los datos:** Cambiar el formato o estructura de los datos sin modificar su significado, con el fin de aplicarles alguna técnica particular en la etapa de modelado.

4.12.4. Modelado:

Las actividades principales de esta etapa son:

1. **Selección de técnica de modelado:** Elegir la técnica apopiada para resolver el problema, los datos apropiados, las herramientas adecuadas.
2. **Selección de datos de prueba:** En algunos proyectos se requiere hacer la diferencia entre datos de prueba y datos de test
3. **Construcción del modelo:** Generar el mejor modelo, mediante un proceso iterativo y ajuste de parámetros.

4.12.5. Evaluación:

Esta etapa se centra en la evaluación de todos los aspectos de acuerdo a lo realizado en las etapas anteriores. Las principales actividades son:

1. **Evaluar resultados:** ¿Los modelos cumplen con los criterios de negocio? ¿Cuál(es) debemos aprobar para el negocio?
2. **Determinar continuar o regresar alguna etapa:** De acuerdo con los resultados de esta etapa, es aquí donde se determina seguir con la última fase de la metodología o regresar a alguna de las etapas anteriores o incluso partir de cero con un nuevo proyecto

4.12.6. Implementación:

El objetivo de esta etapa es que el cliente pueda acceder a los resultados y que estos agreguen valor al negocio y den respuesta a las necesidades de la primera etapa. Las principales actividades son:

1. **Planeación de la implementación:** Realice el plan para que todos los implicados estén sincronizados para la implementación del modelo.
2. **Documentar lo implementado:** Documentar los resultados de manera clara para el usuario final y asegurar de que todas las etapas de la metodología se documenten debidamente para obtener lecciones aprendidas durante el proceso.
3. **Realizar una retrospectiva de lo implementado:** Realice una retrospectiva del proyecto sobre lo que salió bien, lo que podría haber sido mejor y cómo mejorar en el futuro.

4.13. Metodología Ágil

Las metodologías ágiles es un concepto de gestión de proyectos que implica dividir el proyecto en fases y hace hincapié en la colaboración y la mejora continua. Estas metodologías surgieron como respuesta a los enfoques más tradicionales y rígidos, como el modelo en cascada. Los equipos siguen un ciclo de planificación, ejecución y evaluación.

Mientras que el enfoque en cascada tradicional consiste en que una disciplina participa en el proyecto y, a continuación, le pasa la responsabilidad al siguiente participante, la metodología ágil requiere equipos interdisciplinarios colaborativos.

Algunas de las metodologías ágiles más conocidas incluyen Scrum, Kanban, Extreme Programming (XP) y Lean, entre otras. Aunque tienen diferencias en sus prácticas específicas, comparten algunos principios fundamentales, como la entrega incremental, la colaboración estrecha entre los equipos multidisciplinarios, la capacidad de respuesta a los cambios y la entrega de valor continuo al cliente.

Las metodologías ágiles buscan adaptarse a la volatilidad del entorno empresarial y permitir la entrega de productos de alta calidad de manera más rápida y eficiente. Están ampliamente adoptadas no sólo en el desarrollo de software, sino también en diversas áreas como la gestión de proyectos, marketing y desarrollo de productos [23].

4.14. Metodología Lean Six Sigma

Lean six sigma es una filosofía y metodología que combina la manufactura con seis sigma, y establece cómo mejorar los procesos en una forma que involucra los costos de la mala calidad,

procesos fuera de control, el desperdicio y los factores críticos de los requerimientos de los clientes.

El objetivo principal de Six Sigma es mejorar la calidad y la eficiencia del proceso mediante la identificación y eliminación de variaciones que pueden causar defectos o fallos en la producción. La metodología se basa en un enfoque sistemático y cuantitativo para mejorar la calidad y reducir la variabilidad en los procesos.

Six Sigma utiliza un conjunto de herramientas y técnicas estadísticas para medir y analizar la capacidad de un proceso. El término "Six Sigma" hace referencia a la meta de alcanzar un nivel extremadamente bajo de defectos, equivalente a no más de 3.4 defectos por millón de oportunidades. La metodología se aplica en diferentes sectores y tipos de organizaciones, desde la fabricación hasta los servicios.

El enfoque Six Sigma sigue un ciclo de mejora continua conocido como DMAIC, que son las iniciales de las cinco fases del proceso:

- **Definir:** Identificar el problema, el alcance del proyecto y los objetivos de mejora.
- **Medir:** Recopilar datos relevantes y medir el rendimiento actual del proceso.
- **Analizar:** Utilizar herramientas estadísticas para analizar los datos y determinar las causas raíces de los problemas.
- **Mejorar:** Implementar soluciones para corregir y mejorar el proceso.
- **Controlar:** Establecer controles y monitoreo continuo para asegurar que el proceso se mantenga en el nuevo nivel de rendimiento.

El enfoque Six Sigma ha demostrado ser efectivo para mejorar la calidad y la eficiencia en diversas organizaciones, contribuyendo a la reducción de costos y la mejora de la satisfacción del cliente [24].

5. ANTECEDENTES

5.1. Evolución de los chatbots en el sector empresarial

Los chatbots han experimentado un crecimiento significativo en su adopción por parte de las empresas en los últimos años. Desde su introducción inicial como herramientas de atención al cliente básicas, los chatbots han evolucionado para convertirse en componentes integrales de las estrategias de automatización y servicio al cliente de las organizaciones. Estudios como el de "El auge de los bots" [28] han demostrado el impacto positivo de los chatbots en la eficiencia operativa y la experiencia del cliente en diversas industrias.

Se menciona que la historia de los chatbots comienza en la década de los años sesenta, como lo relata el artículo "De Eliza a Siri" [2] sobre la historia de los chatbots que comienza con Eliza, uno de los primeros programas de procesamiento del lenguaje natural, desarrollado en la década de 1960. Eliza tenía la capacidad de mantener conversaciones rudimentarias con usuarios, simulando ser un terapeuta conversacional.

5.2. Aplicaciones de los chatbots en la industria de alimentos y restaurantes

En el sector de alimentos y restaurantes, los chatbots han demostrado ser especialmente útiles para mejorar la experiencia del cliente y optimizar los procesos de pedidos y entrega. Investigaciones como la de la revista science direct [29], donde han destacado cómo los chatbots han permitido a los restaurantes aumentar la velocidad de servicio, reducir los errores en los pedidos y mejorar la satisfacción del cliente.

Sin embargo, en el mercado existen mitos acerca de los chatbots y su alto costo de implementación. Artículos como los de Chaty Blog [30] nos demuestra que cuando se dirige una pequeña empresa, cada interacción es importante. Y nada brilla más que una atención al cliente excepcional cuando se intenta impresionar a los clientes y fidelizarlos. En este artículo se exploraron cinco opciones para ayudar a mantener a los clientes contentos, informados y fieles.

5.3. Análisis de datos en el contexto de chatbots

El análisis de datos generado por las interacciones de los chatbots ofrece una oportunidad única para obtener información valiosa sobre las preferencias y comportamientos de los clientes. Estudios como el de "Aprovechar el modelado predictivo, la personalización del aprendizaje automático, la atención al cliente de PNL y los chatbots de IA para aumentar la lealtad del cliente" [31], han explorado diversas técnicas de análisis de datos, como el análisis de sentimientos y la predicción de la demanda, aplicadas a datos de chatbots para mejorar la toma de decisiones empresariales.

5.4. Casos de estudio y mejores prácticas

Se han documentado varios casos de estudio y mejores prácticas en la implementación y análisis de datos de chatbots en el sector empresarial. Ejemplos como el caso de estudio como el de "Crear bots de servicio al cliente que la gente no odie" [32] y el caso de estudio sobre cómo los "Factores organizacionales que afectan la implementación exitosa de chatbots para el servicio al cliente [33], ilustran cómo las empresas han utilizado eficazmente los chatbots y el análisis de datos para mejorar la eficiencia operativa, la experiencia del cliente y cómo grandes corporaciones tecnológicas como Microsoft, Google y Amazon predijeron que el comercio conversacional sería la próxima gran novedad.

6. METODOLOGÍA

Para el desarrollo de este proyecto se consideraron tres metodologías principales:

1. Metodología Ágil
2. Metodología Six Sigma
3. CRISP-DM.

Las tres metodologías exploradas presentan sus propias ventajas y desafíos. Las metodologías ágiles y Six Sigma ofrecían enfoques convencionales y probados en la industria, mientras que la metodología CRISP-DM (Cross-Industry Standard Process for Data Mining), se destacó por su enfoque sistemático y estructurado para abordar problemas de minería de datos. CRISP-DM se compone de seis etapas claramente definidas, desde la comprensión del negocio hasta la implementación, proporcionando una guía integral que facilita la gestión y ejecución efectiva de proyectos de este tipo.

Para resolver el objetivo específico número uno (1) se implementó la metodología CRISP-DM y el flujo de datos que debe tener una organización de la siguiente manera:

6.1 Entendimiento del negocio.

La situación actual y la necesidad de la empresa es la siguiente: La empresa TuChat por medio del chatbot del restaurante de estudio mencionado anteriormente, está capturando 18 mil datos al mes de las conversaciones con el cliente y se necesita sacar provecho de estos datos tanto para la empresa TuChat como para el restaurante cliente.

Se realizaron 4 sesiones de entendimiento de las necesidades con cada una de las partes: TuChat y el restaurante cliente Las Ricuras de Sebastián. Las 4 sesiones con TuChat fueron con el CEO de la empresa y las 4 sesiones con Las Ricuras de Sebastián fueron con el administrador general del restaurante.

6.2 Recolección de datos

Después de conocer y entender la necesidad que tiene la empresa con los datos, se continuó con la comprensión de estos teniendo en cuenta el flujo de datos en una organización. Todo inicia comprendiendo de donde la empresa recolecta los datos, sí de diferentes fuentes o de una sola fuente.

6.3 Almacenamiento de los datos

Una vez se tuvieron las bases de datos objetivo a trabajar, se empezó a validar la estructura y las variables que se van a utilizar, se realizó un listado de todas las variables que se tienen hoy en día para construir información de valor.

6.4 Procesamiento de los datos

Después de tener todas las bases de datos centralizadas en Google Cloud se debía empezar a procesar la información para resolver cada una de las necesidades del negocio. El primer reto del restaurante en estudio de este proyecto llamado Ricuras de Sebastián era unificar las bases de datos que son independiente por cada sede, es decir, cada sede tiene las siguientes bases de datos estructuradas: DB_Pedidos, DB_usuarios y DB_Traking y una base no estructurada llamada DB_Comentarios. Después se debía realizar un análisis exploratorio de los datos y una limpieza de los datos para ir respondiendo a las siguientes preguntas del negocio:

¿Cuál es la cantidad y el monto vendido?
¿Cantidad de clientes frecuentes?
¿Clientes de mayor consumo?

6.5 Modelado

De acuerdo al flujo de datos de la organización se continúa con el siguiente paso del método CRISP-DM: **El modelado.**

En esta etapa las principales actividades a realizar son: selección de técnicas, selección de los datos y construir el modelo que responde a la necesidad del negocio.

Se realizaron las técnicas de geocodificación, análisis de sentimiento y pronóstico Arima para responder a las siguientes preguntas del negocio:

¿Dónde se encuentran la mayoría de mis clientes?
¿Qué están comentando nuestros clientes?
¿Cuál es el sentimiento de nuestros clientes con el servicio del restaurante?
¿Cuánto aproximadamente se va vender el próximo mes?

6.6 Acceso a los datos y descubrimientos relevantes

Con esta sección de la metodología CRISP DM se logró resolver el objetivo específico número dos (2) de la siguiente manera:

En la etapa de evaluación e implementación y sincronizándolo con el flujo de los datos de la organización, continuamos con el acceso a los datos por medio de GoogleCloud y la interacción del cliente con los descubrimientos relevantes a través de la herramienta Looker Studio de Google; sin embargo, para conocer la relación que hay entre las bases de datos y cómo al momento de explorar se pueden realizar multifiltros entre todas, se creó un diagrama entidad - relación.

Después de este análisis se estructura un modelo conceptual de repositorio integrado teniendo en cuenta los siguientes componentes: la seguridad, las fuentes de datos, los procesos de ingesta de datos, el procesamiento, análisis de los datos y la visualización de estos.

Para dar cumplimiento al objetivo específico número tres (3), se logró construir un tablero integrador donde empresas podrán comprender la información recolectada y convertir los datos en información estratégica para la toma de sus decisiones.

Finalmente, cumpliendo con el objetivo general del proyecto, se logró desarrollar un modelo de analítica de datos que potencie, tanto para TuChat como para las empresas clientes, cada uno de los datos que son almacenados actualmente y convertirlos en información estratégica para el negocio.

7. RESULTADOS

7.1 Entendimiento del negocio.

De las sesiones realizadas con el CEO de la empresa TuChat y con el administrador general del restaurante, se identificaron las preguntas a resolver por parte de la solución propuesta a la necesidad. Estas preguntas clave a resolver se relacionan en la siguiente tabla:

Empresa TuChat	Restaurante Cliente
¿Cuál es el Ticket promedio? ¿Qué tan probable es que un cliente nos recomiende? NPS ¿Cuánto es el Costo de adquisición de clientes CAC? ¿Cuál es la Retención / Deserción? ¿Cuánto es el Life time value?	¿Cuál es la cantidad y el monto vendido? ¿Cantidad de clientes frecuentes? ¿Clientes de mayor consumo? ¿Dónde se encuentran la mayoría de mis clientes? ¿Qué están comentando nuestros clientes? ¿Cuál es el sentimiento de nuestros clientes con el servicio del restaurante? ¿Cuánto aproximadamente se va vender el próximo mes?

Tabla 1: Preguntas clave a resolver con el proyecto

7.2 Recolección de datos

En estas sesiones se identificó que la fuente principal de los datos de la empresa TuChat son los chatbots de donde provienen algunos datos estructurados tabulares en filas y columnas y otros datos no estructurados como los comentarios en lenguaje natural y ubicación del cliente de cada pedido que se realiza.

En la siguiente figura se representa la recolección de los datos de TuChat:

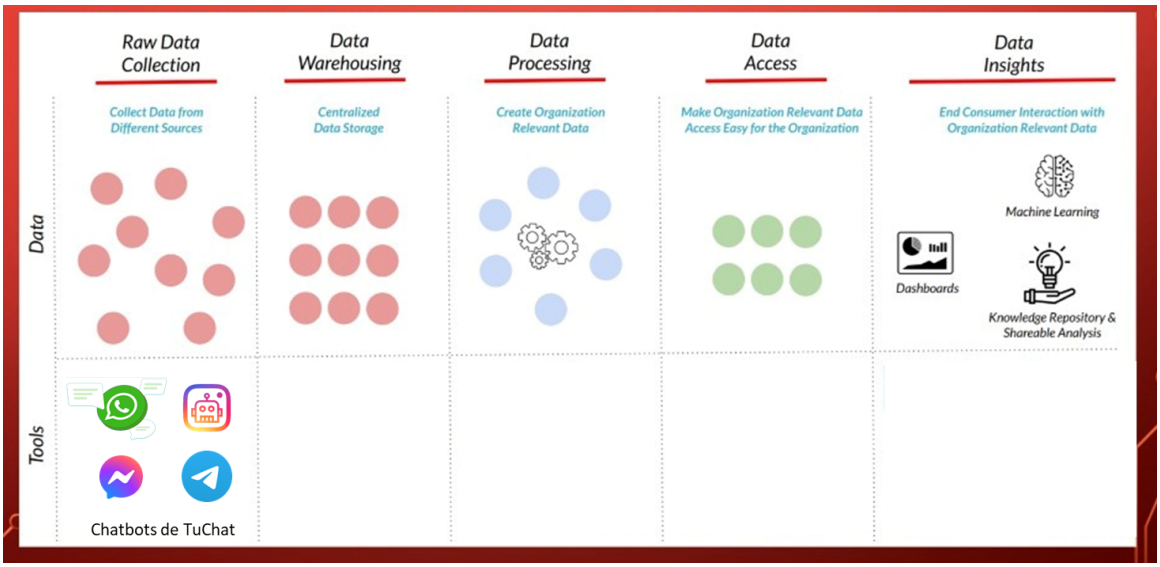


Figura 6: Notas de webinar de Data Camp. Flujo de datos en una organización.

7.3 Almacenamiento de los datos

Al inicio de la operación, TuChat le construye a cada cliente las siguientes bases de datos estructuradas: DB_Pedidos, DB_usuarios y DB_Traking y una base no estructurada llamada DB_Comentarios. En estas bases de datos se recopilan todos los pedidos y comentarios que los clientes realizan a la empresa por medio del chatbot. Estas bases contienen las siguientes variables y sus componentes:

Descripción	**Fuente de datos 1:**	**Fuente de datos 2:**	**Fuente de datos 3:**	**Fuente de datos 4:**
¿Nombre de la fuente de datos?	Bd_pedidos	Bd_usuarios	Bd_comentarios	Bd_traking
¿Qué contiene? / atributos	Fecha, hora, nombre, celular, dirección, id pedido, pedido, total, comanda, método de pago, canal y sede.	Fecha, celular, nombre y dirección.	fecha, celular, nombre, comentario	fecha, celular, nombre, canal, código, detalle

¿Con qué frecuencia se recopilan los datos?	En tiempo real a ritmo de negocio	En tiempo real a ritmo de negocio	En tiempo real a ritmo de negocio	En tiempo real a ritmo de negocio
¿Tienen datos únicos para construir llaves con otras fuentes?	Celular	Celular	Celular	Celular
¿Propietario interno de los datos?	TuChat	TuChat	TuChat	TuChat
¿Dónde se almacenan los datos?	Google Cloud	Google Cloud	Google Cloud	Google Cloud

Tabla 2: Análisis comparativo de las bases de datos

Continuando con la comprensión de los datos y completando la figura 8, se comprende que la empresa TuChat tiene centralizados los datos en Google Cloud por medio de bases de datos en google sheets que son hojas de cálculo de excel en la nube de Google. En la siguiente figura se representa la centralización de los datos de TuChat:

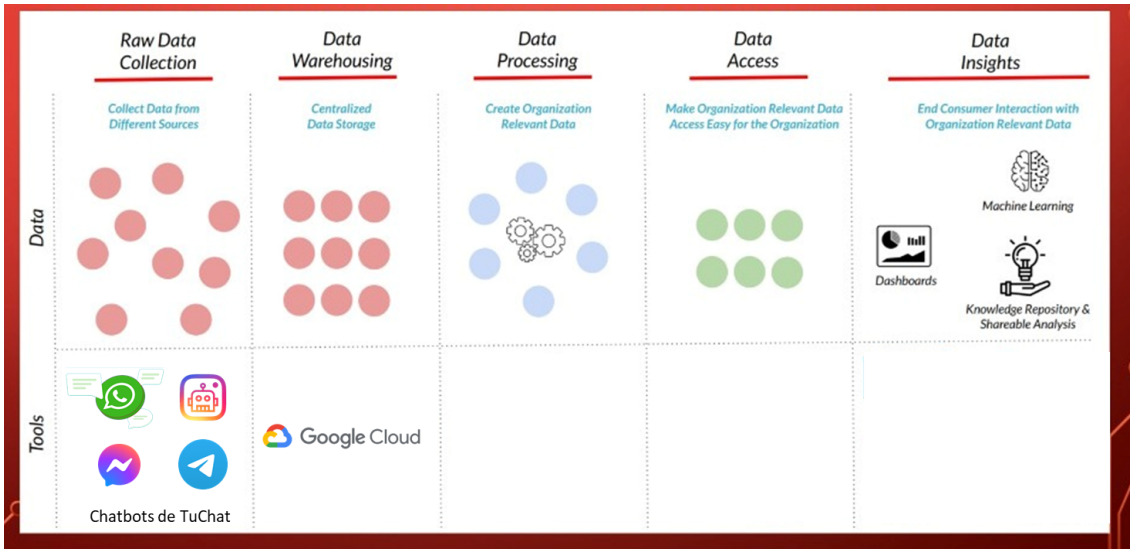


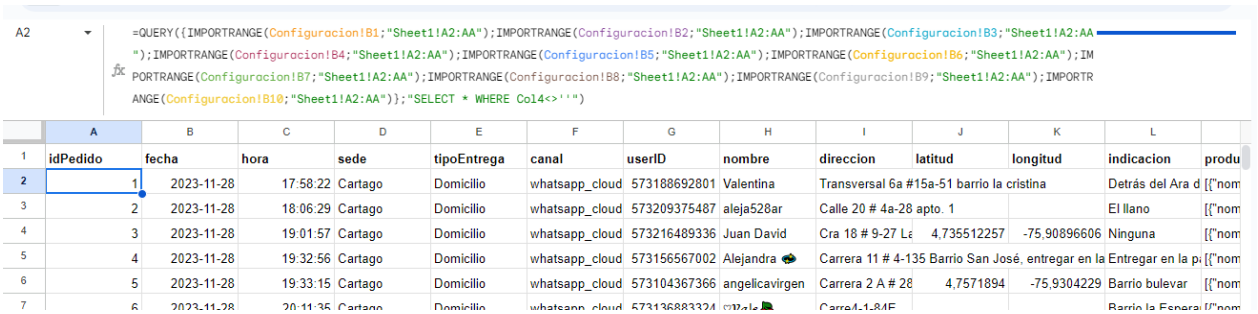
Figura 7: Almacenamiento de los datos. Elaboración personalizada para la empresa TuChat.

El cliente desea ver los resultados de las 3 sedes, en un único tablero de control y para eso se debía unificar todas las bases de datos en tiempo real a ritmo de negocio. Esto se realizó de la siguiente manera:

Al ser bases de datos estructuras tienen un beneficio y es que los datos siempre van a estar en el mismo orden, es por ello que La DB_Pedidos siempre va tener las mismas 27 columnas y para unificarlas nos apoyamos a de la teoría y conexión con la función IMPORTRANGE de Google, la cual Importa un rango de celdas de una hoja de cálculo específica [21] y de esta forma conectamos todas las bases de datos en una única DB_Pedidos Consolidada que se actualiza en línea a ritmo de negocio.

La consolidación había quedado correcta, sin embargo, se tenían, miles de espacios en blanco que se trasladaban en la importación de los datos y esto generaba menor rendimiento en la consolidación e importación de la información. Para solucionar esto y lograr que solo nos importara los campos con datos, es decir, que no tomara los campos vacíos, nos apoyamos con la función QUERY de Google, la cuál aparte de ayudarnos con la eliminación de los vacíos, el orden de las consulta de acuerdo al criterio dado, nos permite más adelante hacer filtros personalizados en las bases de datos con un mejor rendimiento y usabilidad.

El uso y combinación de estas dos funciones se puede observar en la siguiente figura:



	A	B	C	D	E	F	G	H	I	J	K	L	
1	idPedido	fecha	hora	sede	tipoEntrega	canal	userID	nombre	direccion	latitud	longitud	indicacion	produ
2	1	2023-11-28	17:58:22	Cartago	Domicilio	whatsapp_cloud	573188692801	Valentina	Transversal 6a #15a-51 barrio la cristina			Detrás del Ara d	[{"nom
3	2	2023-11-28	18:06:29	Cartago	Domicilio	whatsapp_cloud	573209375487	aleja528ar	Calle 20 # 4a-28 apto. 1			El llano	[{"nom
4	3	2023-11-28	19:01:57	Cartago	Domicilio	whatsapp_cloud	573216489336	Juan David	Cra 18 # 9-27 La	4,735512257	-75,90896606	Ninguna	[{"nom
5	4	2023-11-28	19:32:56	Cartago	Domicilio	whatsapp_cloud	573156567002	Alejandra	Carrera 11 # 4-135 Barrio San José, entregar en la			Entregar en la pi	[{"nom
6	5	2023-11-28	19:33:15	Cartago	Domicilio	whatsapp_cloud	573104367366	angelicavirgen	Carrera 2 A # 28	4,7571894	-75,9304229	Barrio bulevar	[{"nom
7	6	2023-11-28	20:11:35	Cartago	Domicilio	whatsapp_cloud	573136883324		Carre4-1-84E			Barrio la Esperal	[{"nom

Figura 8: Uso personalizado de importrange y query para la empresa TuChat.

7.4 Procesamiento de los datos

De acuerdo a la metodología CRISP-DM, continuamos con el procesamiento de los datos y completando la figura 8, esto se realizó en el entorno y lenguaje de programación estadístico "R" [18]. En la siguiente figura se representa la herramienta usada para la preparación de los datos de TuChat:

Figura 9: Procesamiento de los datos, Elaboración personalizada para la empresa TuChat.

7.4.1 Análisis exploratorio de los datos

A la Bd_pedidos se le realizó un análisis exploratorio para ampliar este entendimiento y preparación de los datos. En este proceso se logró cumplir con el objetivo específico número uno apoyado por el entorno y lenguaje de programación estadístico R [18] se obtuvieron los siguientes resultados:

Tabla de Indicadores Importantes

```
total_registros=nrow(Data_tesis)
atributos_por_registro = ncol(Data_tesis)
precio_promedio=mean(Data_tesis$total,na.rm = TRUE)
mediana_precio=median(Data_tesis$total,na.rm = TRUE)
promedio_pizza=mean(Data_tesis$totalPizza,na.rm = TRUE)
promedio_plancha=mean(Data_tesis$totalPlancha,na.rm = TRUE)
promedio_bebidas=mean(Data_tesis$totalBebida,na.rm = TRUE)
#cantidad_pedidos=length(Data_tesis$total,na.rm = TRUE))

Resultado = data.frame(total_registros, atributos_por_registro, precio_promedio, mediana_precio,promedio_pizza, promedio_pla
ncha, promedio_bebidas)
Resultado
```

```
##   total_registros atributos_por_registro precio_promedio mediana_precio
## 1            1870                     19        25394.28          22000
##   promedio_pizza promedio_plancha promedio_bebidas
## 1             NA               NA               NA
```

Figura 10: Indicadores descriptivos de las variables numéricas. Fuente: Elaboración propia

En la figura anterior se puede observar que el precio promedio de un pedido en el restaurante es de $25.394 pesos y la mediana es $22.000, esto es un insumo para usar en estrategias comerciales a futuro para el dueño del restaurante.

En la siguiente figura se identifica el tipo de cada atributo para conocer si ya están listos para usar o cuál técnica se debe seleccionar en el próximo paso al hacer uso de la variable en algún modelo o análisis que se realice

Indentificar de qué tipo son los atributos:

```
tipos_atributos=sapply(Data_tesis, class)
tipos_atributos
```

```
## $fecha
## [1] "POSIXct" "POSIXt"
##
## $hora
## [1] "character"
##
## $nombre
## [1] "character"
##
## $celular
## [1] "numeric"
##
## $direccion
## [1] "character"
##
```

Figura 11: Identificación del tipo de atributo. Fuente: Elaboración propia

En la siguiente figura se realiza una tabla descriptiva de las variables numéricas, destacando los resultados de la variable “total” que va ser utilizada más adelante. En ella se observa y se confirma el promedio mencionado anteriormente de $25.394, sin embargo, se obtiene una desviación

estándar de $14.612, es decir, es muy variable el costo de los pedidos realizados por los clientes y se obtiene un punto máximo de $110.000. Si se quisiera realizar una estrategia para la mayoría, los datos nos están mostrando que la mayoría se encuentra entre $22.000 a $31.500.

Tabla de estadística descriptiva

```
library(summarytools)

tabla_descrip = descr(Data_tesis)

tabla_descrip
```

```
## Descriptive Statistics
## Data_tesis
## N: 1870
##
##                        celular   idPedido       total        X1        X2
## ----------------- ------------- ---------- ----------- --------- ---------
##              Mean   3117184396.22       9.96    25394.28    -76.48      3.52
##           Std.Dev    328436658.72       7.70    14612.42      0.15      0.56
##               Min     44634110.00       1.00        0.00    -76.61      3.33
##                Q1   3113659869.00       4.00    15500.00    -76.50      3.46
##            Median   3154729724.00       8.00    22000.00    -76.50      3.47
##                Q3   3177840588.00      14.00    31500.00    -76.49      3.47
##               Max   5731734689.00      44.00   110000.00    -74.14      9.36
##               MAD     41607426.94       7.41    11119.50      0.01      0.01
##               IQR     64180719.00      10.00    16000.00      0.01      0.01
##                CV            0.11       0.77        0.58      0.00      0.16
##          Skewness           -5.48       1.18        1.72     10.56     10.14
##       SE.Skewness            0.06       0.06        0.06      0.06      0.06
##          Kurtosis           66.77       1.45        4.65    124.63    101.88
##           N.Valid         1870.00    1867.00     1869.00   1870.00   1870.00
##         Pct.Valid          100.00      99.84       99.95    100.00    100.00
```

Figura 12: Tabla estadística descriptiva. Fuente: Elaboración propia

En la siguiente figura se observa que el método de pago más usado es el efectivo, seguido por el datáfono.

Graficos de variables categóricas

```
library(ggplot2)

# Seleccionar la variable categórica
variable_categorica = Data_tesis$metodoPago

# Calcular las frecuencias de la variable categórica
frecuencias = table(variable_categorica)

# Crear el gráfico de torta
grafico_torta <- ggplot(data.frame(frecuencias), aes(x = "", y = Freq, fill = variable_categorica)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar("y", start = 0) +
  theme_void()

# Mostrar el gráfico
print(grafico_torta)
```

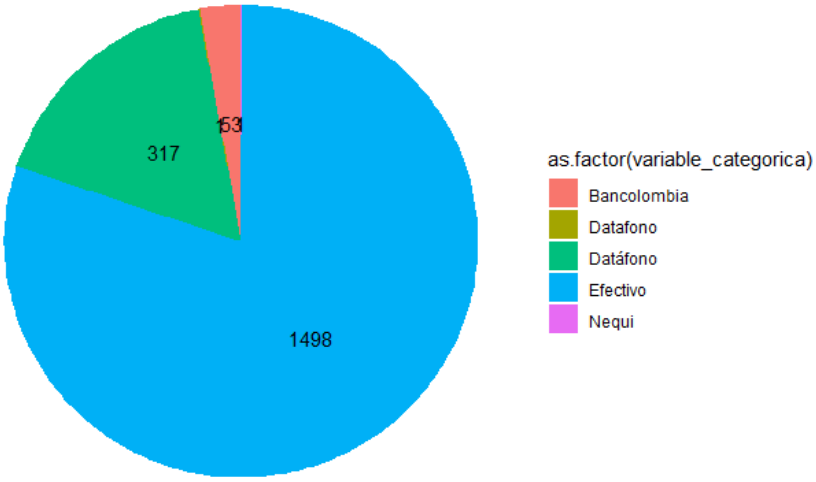


Figura 13: Gráfico de variable categórica. Fuente: Elaboración propia

Se analizan los datos vacíos, ya que estos afectan más adelante los resultados y se les debe hacer un tratamiento antes para que no afecten. Los atributos con datos faltantes son los siguientes

```
columnas_con_faltantes
```

```
## [1] "idPedido"     "pedido"       "totalPizza"   "totalPlancha" "totalBebida"
## [6] "pagaCon"      "cambio"       "total"        "comanda"
```

Figura 14: Identificación de variables con datos vacíos. Fuente: Elaboración propia

Todos los análisis detallados se pueden observar en la siguiente publicación en Rpubs: https://rpubs.com/sgcifuentes/1060540

7.4.2 Limpieza de los datos

Se realizó la limpieza, construcción y transformación de los datos necesarios para cumplir con los objetivos específicos propuestos.

Uno de los pasos clave para realizar la preparación de los datos fue la limpieza de estos; puntualmente en este caso y para cumplir los objetivos se necesita realizar una limpieza general del conjunto de datos, pero también preparar los siguientes atributos: hora, dirección y comentarios ya que a futuro se van a utilizar en la etapa de modelado.

El conjunto de datos como normalmente sucede tiene contenido innecesario como por ejemplo caracteres especiales como hashtags, menciones y URLs, se debe realizar la sustitución de abreviaturas y jergas, hay palabras conjugadas que hay que llevar a sus raíces, se debe realizar la eliminación de palabras vacías o también conocidas como palabras que no aportan valor al contexto por ejemplo de un comentario (stopwords). Este proceso se realizó de la siguiente manera:

```
mycorpus_url_numb_punct_symb <- tokens(mycorpus, what = "word", remove_url = T, remove_numbers = T,
                                       remove_punct = T, remove_symbols = T, remove_separators = T,)

## Paso 4 - STOPWORDS
# a="EL PROBLEMA"
# tolower(a)
#stopwords.es <- stopwords(language = "ES")
# stopwords.en <- stopwords(language = "en")
# stopwords.es=c(stopwords.es,"etc","q","k","problema",stopwords.en)
mycorpus_sw <- tokens_remove(mycorpus_url_numb_punct_symb)
mycorpus_sw
```

Figura 15: Limpieza del conjunto de datos. Fuente: Elaboración propia

7.5. Geocodificación

En la Bd_pedidos se encuentra la variable dirección y esta tiene una característica y es que la información puede venir mixta. Es decir, en algunas ocasiones el cliente digita la dirección con muchas más indicaciones y explicaciones coloquiales para orientar al domiciliario y en otras ocasiones envía por el chatbot la ubicación del celular. Este último caso es mucho más preciso ya que se puede capturar la longitud y latitud y con estas coordenadas se realiza el gráfico de localización más rápido. Sin embargo para el cliente Las Ricuras de Sebastian, la mayoría de las direcciones eran digitadas por el mismo cliente y esto llevaba a realizar una preparación de los datos más ardua. Se realizó una limpieza y depuración de la información, datos atípicos, datos

vacíos, caracteres especiales, textos y demás datos basura que bloqueara la geocodificación de la dirección para poder utilizarla. Este proceso se conoce como geocodificación de acuerdo a la teoría de análisis espacial para extraer la longitud y la latitud de la ubicación dada en la dirección y esto con el objetivo de tener los datos necesarios para identificar la ubicación de las solicitudes en un mapa geográfico [19]. Se realizó de la siguiente manera:

Dirección - Proceso de Geocodificación

```
require(tmaptools)

locs=matrix(NA,nrow =dim(Data_tesis)[1],ncol = 2 )
for(i in 1:dim(Data_tesis)[1]){
loc=geocode_OSM(q = paste(gsub("#","",Data_tesis$direccion[i]), ",cali,colombia"))
if(length(loc)>0){
locs[i,1]=loc$coords[1]
locs[i,2]=loc$coords[2]
}
#print(i)
}

locs=data.frame(locs)
locs2=na.omit(locs)
require(leaflet)
leaflet() %>% addTiles() %>% addCircleMarkers(lng = locs2$X1,lat = locs2$X2)
```

Figura 16: Proceso de geo-codificación. Fuente: Elaboración propia

Una vez limpia la data después de todo este proceso se pudo cumplir con el objetivo que es identificar esos lugares considerados de alta demanda, es decir, desde donde el cliente está realizando la mayoría de sus pedidos, lo cual es insumo para que la empresa pueda dirigir sus campañas de mercadeo para obtener mayores beneficios, como se observa el resultado en la siguiente gráfica.

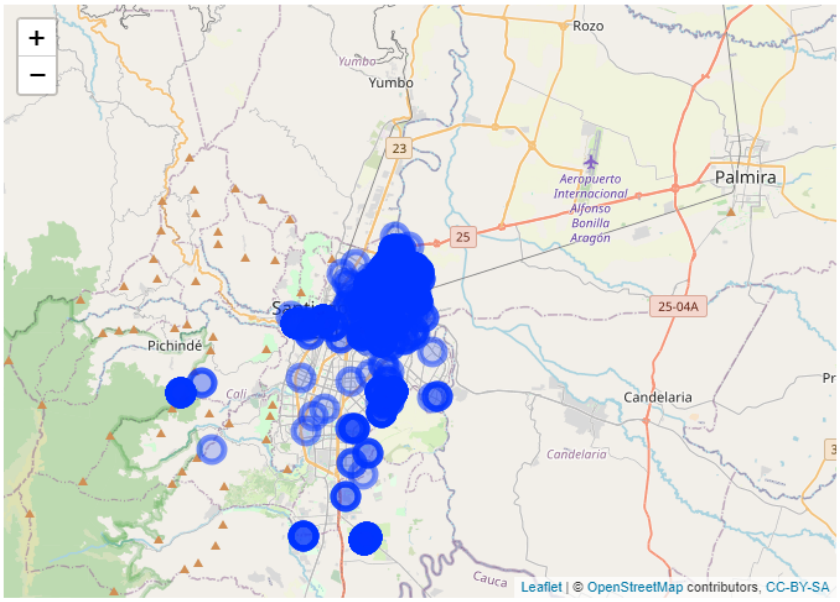


Figura 17: Georeferenciación de las solicitudes de los clientes. Fuente: Elaboración propia

7.6 Análisis de sentimiento

En la DB_Comentarios se tienen datos que son observaciones que se reciben del cliente en lenguaje natural acerca de la experiencia que este vivió con el restaurante. Esta DB tiene un atributo que se llama comentarios y allí se encuentran todas las observaciones mencionadas por los clientes.

Para conocer a detalle y masivamente lo que dice allí en esta variable se apoyó de la técnica del análisis del lenguaje natural, con base en la minería de textos, este análisis es denominado análisis de sentimiento [20], el cual indica de cierta forma lo que están escribiendo los clientes y cuál es su sentimiento con la empresa, suministrando así esta información a las empresas aliadas y que sea insumo de sus campañas de mercadeo y retención de clientes.

El concepto general se resume en la siguiente imagen, donde se muestra la matriz que se construye con los comentarios realizados por los clientes, se hace la partición de palabras en cada documento evaluado y luego se realiza una puntuación de cada palabra para hacer finalmente un conteo y obtener el resumen de palabras más representativas.

CORPUS	BAG OF WORDS					
	Palabra 1	Palabra 2	Palabra 3	Palabra 4	Palabra 5	Palabra 6
Documento 1	x	x		x		
Documento 2		x	x		x	
Documento 3		x				x
Documento 4	x		x	x		x
Documento 5	x	x	x	x	x	
Documento 6		x	x			x
Documento 7	x		x		x	

Figura 18. Matriz de documento para análisis de sentimiento [20]

El primer paso del modelo es crear el corpus, este se construye con cada uno de los comentarios que deja el cliente al restaurante por medio del chatbot. Esta construcción de corpus en el modelo se observa de la siguiente forma:

```
Console   Background Jobs
R 4.3.0 · ~/
> ## Paso 1 - ahora creo el CORPUS
> require(quanteda)
> mycorpus <- corpus(Data_tesis$Comentario)
> mycorpus
Corpus consisting of 1,870 documents.
text1 :
"Muy buen servicio. Gracias"

text2 :
"Bueno"

text3 :
"Excelente"

text4 :
"Las respuestas muy lentas"

text5 :
"Súper bueno"

text6 :
"Excelente ❤"
```

Figura 19. Generación de corpus para análisis de sentimiento [20]

El paso 2 es realizar la tokenización la cual es partir cada corpus en cada una de sus palabras, en la construcción del modelo este proceso se observa de la siguiente manera:

Figura 20. Tokenización de corpus para análisis de sentimiento [20]

En el paso 3 y 4 el proceso consiste en normalizar y eliminar las palabras que no aportan al contexto del comentario. En la construcción del modelo se observa así:

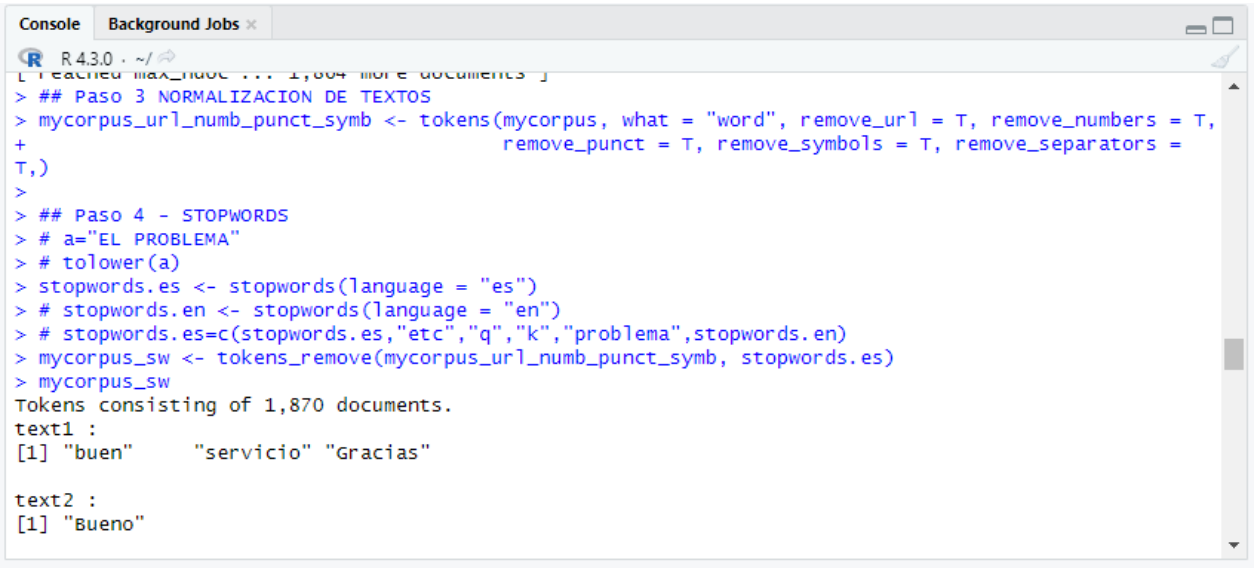


Figura 21. Normalización y eliminación de stopwords del corpus para análisis de sentimiento [20]

Después de eliminar estas palabras ya se tiene el corpus limpio con las palabras que se van a usar en la técnica de análisis de sentimiento.

Se realiza un conteo y agrupación de todas las palabras, para mostrarlas en una gráfica donde cada vez que el modelo encuentra la misma palabra la muestra en un tamaño más grande y así las palabras más repetidas se observan más grandes que las demás.

Se realizó este procedimiento a la base de datos del proyecto y los resultados se muestran en la figura:

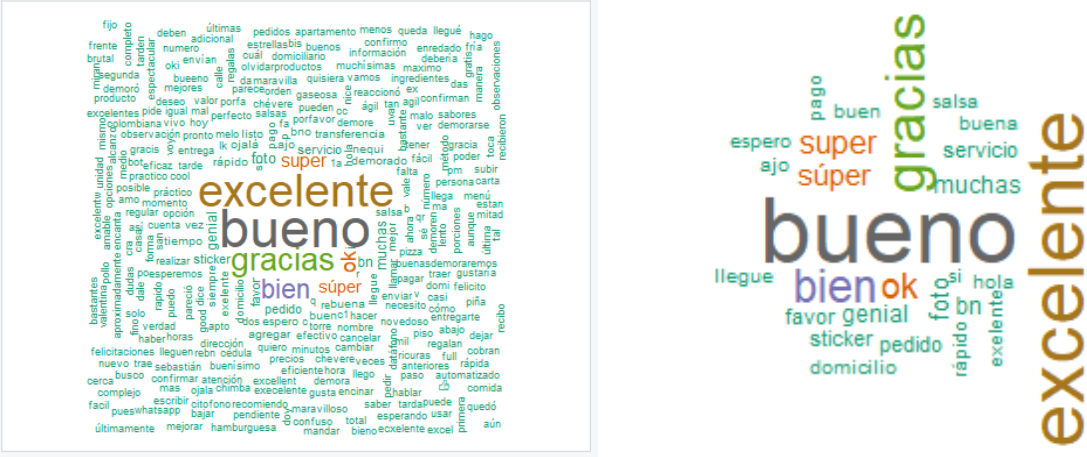


Figura 22. Nube de palabras para análisis de sentimiento [20]

Se continúa con el paso 5 y consiste en que cada palabra del corpus es buscada en un diccionario con millones de palabras puntuadas en una escala de -5 a 5, donde -5 es una palabra que representa un comentario negativo y 5 es una palabra que representa un comentario positivo y de acuerdo a la puntuación que se le otorga a cada palabra, se obtiene este histograma del sentimiento expresado por los clientes en los comentarios. En este se puede observar que la mayoría de los clientes tienen un sentimiento positivo con una puntuación promedio de 3, sin embargo, el análisis sirve para gestionar con los que tienen un sentimiento negativo y lograr convertirlos en clientes positivos.

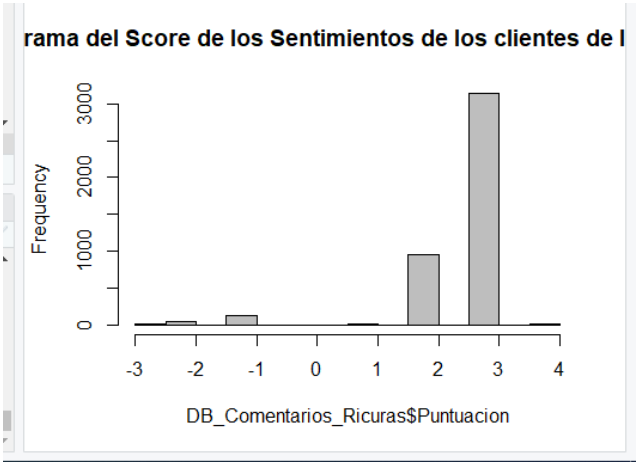


Figura 23. Gráfica de barras para análisis de sentimiento [20]

Finalmente, podemos concluir el sentimiento de los clientes con una gráfica de serie de tiempo donde se ve el comportamiento.

Allí se puede observar que al inicio del bot en el 2020 el sentimiento fue disminuyendo porque era una nueva forma de pedir a domicilio, pero con el paso del tiempo fue creciendo y ahora va en aumento, es decir, que a los clientes les gusta y van adecuándose a la forma de hacer su pedido a domicilio.

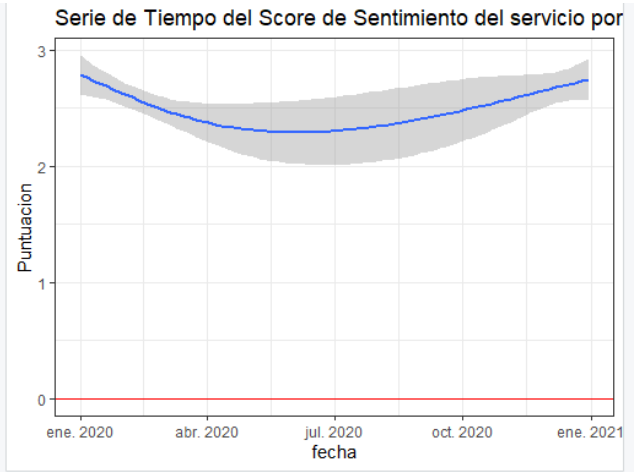


Figura 24. Comportamiento del sentimiento del análisis de sentimiento [20]

En la Db_Pedidos se tiene la variable “hora”, esta trae consigo un reto técnico y es que viene capturada por el bot con la hora del celular del cliente y se encuentra en un formato que hay que hacerle un arreglo para que se deje manipular y tratar como formato hora. El objetivo de esta variable es determinar esos horarios calientes y fríos en donde los clientes pueden hacer estrategias puntuales para ofrecer ciertos productos o sacar ciertas promociones. El resultado se puede observar en la siguiente imagen donde el eje “x” representa la hora en formato de 24 horas y el eje “y” representa el volumen de pedidos. En ella se puede observar que el comportamiento es un reflejo de la realidad de acuerdo a lo conversado con el dueño del restaurante donde las horas pico son de 6pm a 9pm.

Formato Hora

```
hora_dia=substring(Data_tesis$hora,1,2)
jornada=substring(Data_tesis$hora,7,8)
hora_dia=as.numeric(hora_dia)
validos=which(jornada=="AM"|jornada=="PM")
hora_dia=hora_dia[validos]
jornada=jornada[validos]

hora_dia[jornada=="PM"]=hora_dia[jornada=="PM"]+12

barplot(table(hora_dia))
```

Figura 25. Comportamiento del sentimiento del análisis de sentimiento [20]

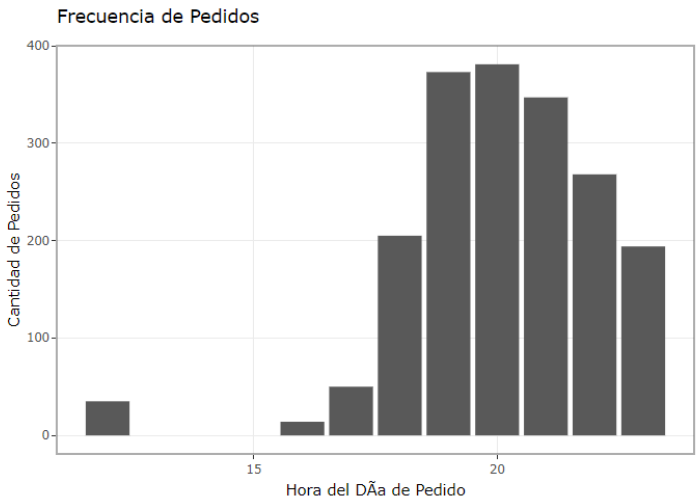


Figura 26. Comportamiento del volumen de pedidos de acuerdo al horario. Fuente: Elaboración propia

7.7. Pronóstico ARIMA

En la DB_pedidos se tiene la variable “fecha” que indica el mes, día y año que un cliente realizó un pedido y la variable “total” que representa el monto pagado por el cliente de acuerdo a su pedido. Con estas dos variables se logró responder a la siguiente pregunta del cliente: ¿cuánto aproximadamente se va vender el próximo mes? para ir realizando la planeación financiera y operativa requerida.

Utilizando el análisis de series temporales e implementando un modelo de pronóstico ARIMA, en el entorno y lenguaje de programación estadístico R [18], se logra obtener los valores futuros de la serie con los datos del pasado, generando esa relación entre los datos para obtener un pronóstico de ventas mensual en base a su misma historia.

Este modelo se desarrolló de la siguiente manera:

De la DB_pedidos se extrae un nuevo conjunto de datos con las dos variables necesarias para realizar el modelo: fecha y total.

Este procesamiento de datos se realizó así:

```
{r message=FALSE, warning=FALSE}
library(fpp2)
library(readxl)
library(dplyr)
library(lubridate)
library(tidyr)

Data_tesis <- read_excel("C:/Users/User/Desktop/Data_tesis.xlsx")
View(Data_tesis)

nrow(Data_tesis)
dim(Data_tesis)

attach(Data_tesis)
Data_tesis$fecha = date(Data_tesis$fecha)

na_count <- colSums(is.na(Y))
print(na_count)

Y <- na.omit(Y)


columnas_seleccionadas = Data_tesis[, c("fecha", "total")]

fecha_venta = data.frame(columnas_seleccionadas)

View(fecha_venta)
```

Figura 27. Extracción del conjunto de datos para uso del modelo ARIMA. Fuente: Elaboración propia

Continuando con la transformación de los datos, se realizó una agrupación del conjunto de datos por mes y año, es decir, el conjunto de datos estaba compuesto por las siguientes columnas: Año, mes y total. Pasando así de unas transacciones diarias a una agrupación por año y mes.

Después de realizar la agrupación, se creó una serie temporal llamada “Y”, cual estaba compuesta por el conjunto de datos agrupado en año y mes y unos parámetros que le indicaban que los datos iniciaban en Agosto 2020 y finalizaban en Marzo 2023.

```
columnas_seleccionadas = Data_tesis[, c("fecha", "total")]

fecha_venta = data.frame(columnas_seleccionadas)

View(fecha_venta)

Ventas_mensuales = fecha_venta %>%
  mutate(month = format(fecha, "%m"), year = format(fecha, "%Y"))%>%
  group_by(month, year) %>%
  summarise(total = sum(total))


Ventas_mensuales = Ventas_mensuales[with(Ventas_mensuales, order(Ventas_mensuales$year)),]

Ventas_mensuales
Ventas_mensuales = na.omit(Ventas_mensuales)

Y = ts(Ventas_mensuales[,3],start = c(2020,8), end = c(2023,3), frequency=12)
```

Figura 28. Creación de una serie temporal llamada “Y”para uso del modelo ARIMA. Fuente: Elaboración propia

Ejecutando este proceso se realizó la serie de tiempo temporal la cual se ve de la siguiente manera:

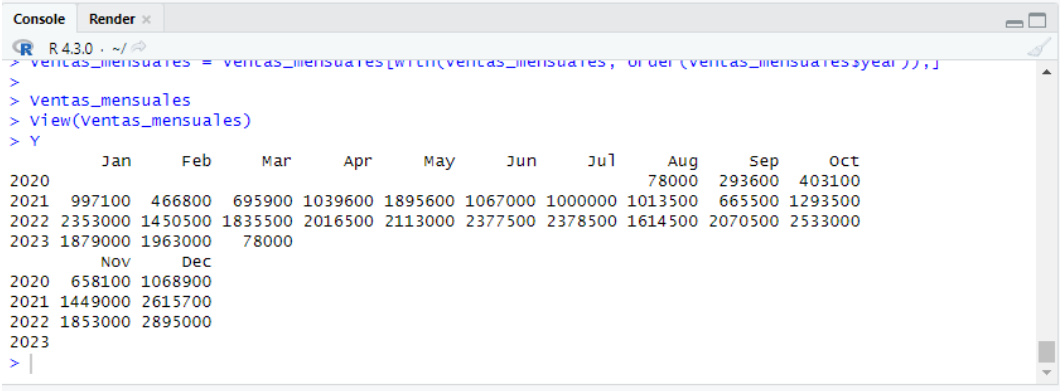


Figura 29. Resultado de la serie temporal llamada “Y” para uso del modelo ARIMA. Fuente: Elaboración propia

El próximo paso fue realizar una exploración de esta serie temporal construida para validar gráficamente la evolución de las ventas en el tiempo. La siguiente figura expresa este comportamiento de las ventas del restaurante en el tiempo:

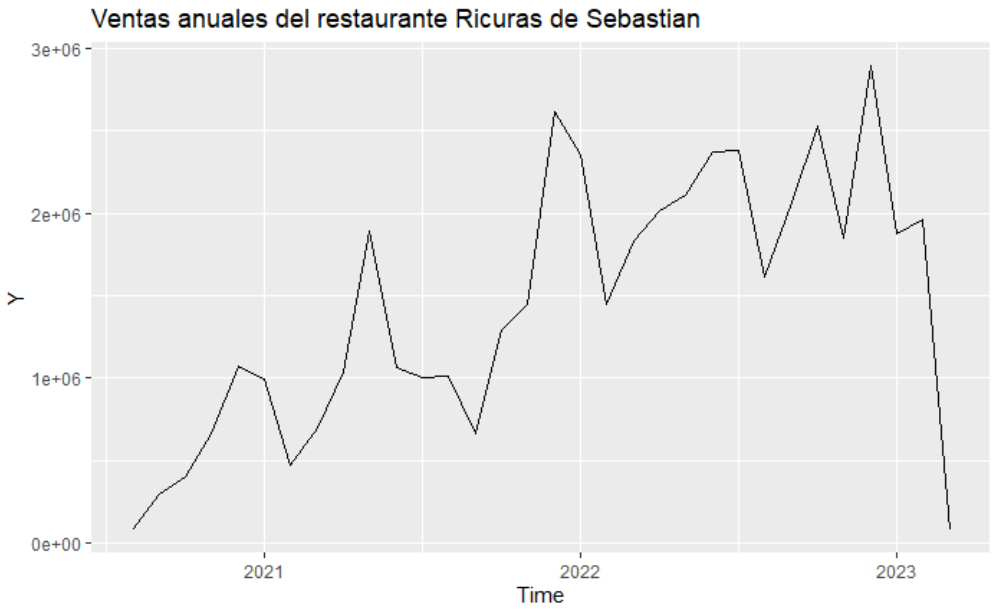


Figura 30. Comportamiento de las ventas del restaurante en el tiempo. Fuente: Elaboración propia

En la anterior figura se puede observar que las ventas tienen una tendencia creciente con el pasar de los años, sin embargo, llama la atención que se tienen picos muy similares de un año a otro por allí al finalizar cada año y esto es un claro indicio que las ventas tienen estacionalidad.

Para validar esta estacionalidad se realiza un proceso con la función descompose en el entorno y lenguaje de programación estadístico R [18] y se logra observar las siguientes figuras:

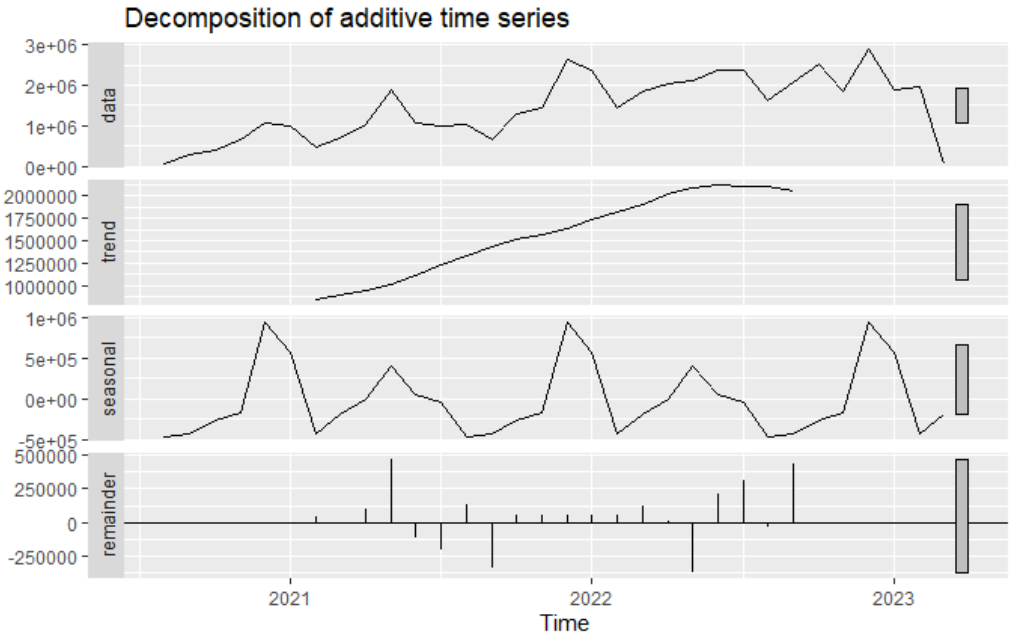


Figura 31. Comportamiento de las ventas del restaurante en el tiempo. Fuente: Elaboración propia

Observando de arriba hacia abajo, en la primera figura se encuentra el comportamiento de las ventas de acuerdo a la serie temporal creada, después observamos la tendencia creciente claramente, luego observamos la estacionalidad donde se ratifica lo mencionado anteriormente y se evidencia que hay un claro comportamiento estacional, lo cual es muy coherente con la realidad del comportamiento del negocio. En la última figura observamos los residuos que va dejando el modelo cuando disminuye el volumen y cuando aumenta.

Se continuó con la creación del modelo ARIMA apoyándonos de la función auto.arima del lenguaje de programación estadístico R [18]. Este modelo realiza varias iteraciones hasta encontrar el mejor modelo que se adapte a los datos para ser usado más adelante en el pronóstico. El procesamiento de los datos se realizó de la siguiente manera:

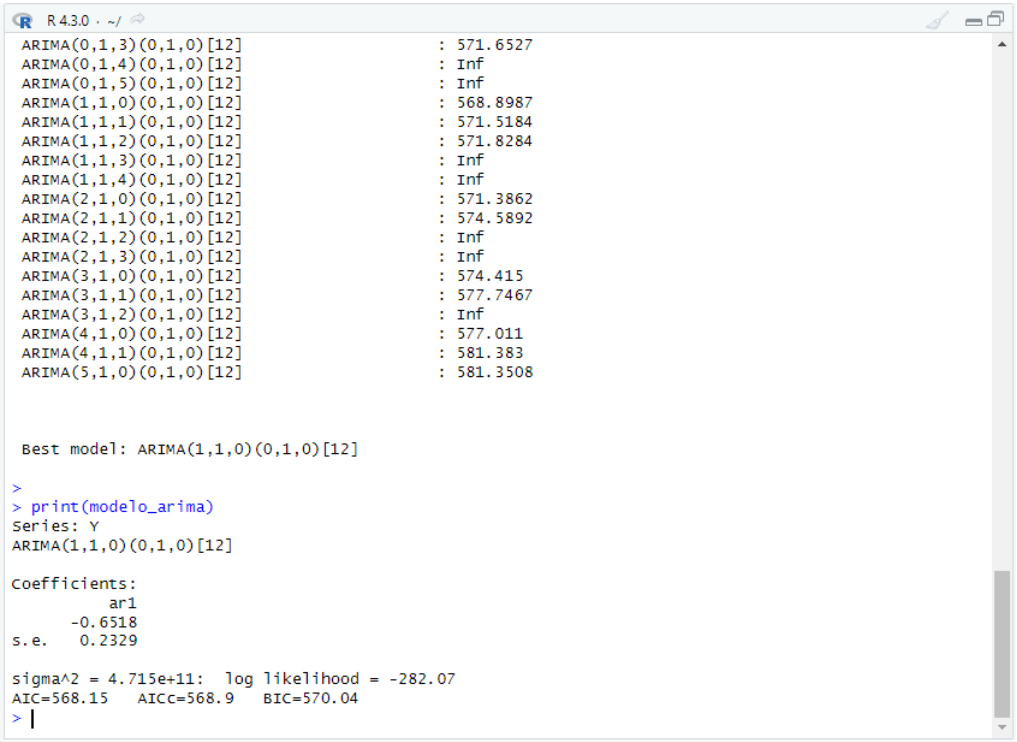


Figura 32. Creación del modelo ARIMA apoyado de la función auto.arima en R. Fuente: Elaboración propia

Esta función nos indicó que el modelo que más se ajusta a los datos fue:

Best model: ARIMA(1,1,0)(0,1,0)[12]

ya que lo que trata de determinar el modelo auto.arima es determinar la estructura y los parámetros indicados para lograr una estimación rápida y así determinar que esta combinación es la que más se ajustó.

El siguiente paso fue realizar un chequeo de los resultados del modelo y acá lo relevante fue que se observó que los valores residuos están dentro de los límites de significancia y la distribución de los residuos tiende a centrarse en cero, lo cual significa que ese margen de error de la predicción va tendiendo a cero logrando una distribución normal.

Estos resultados se pueden observar en la siguiente figura:

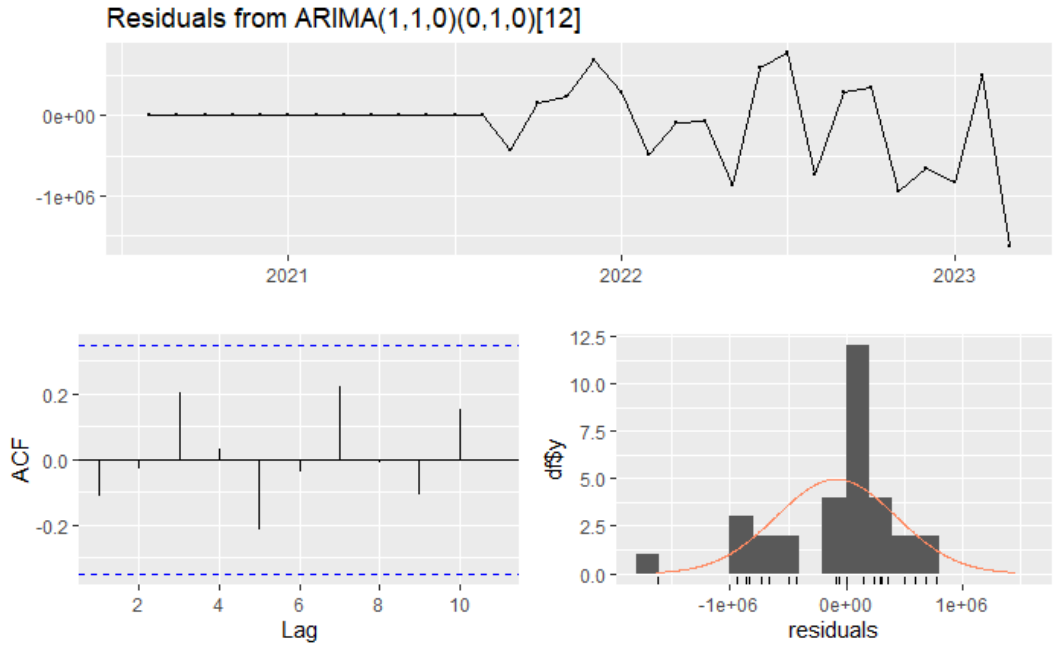


Figura 33. Chequeo de resultados del modelo ARIMA. Fuente: Elaboración propia

Por último, el siguiente paso es crear en base a este modelo que se desarrolló una función de pronóstico para predecir el comportamiento de los próximos 6 meses con un nivel de significancia del 95% en base al comportamiento pasado.

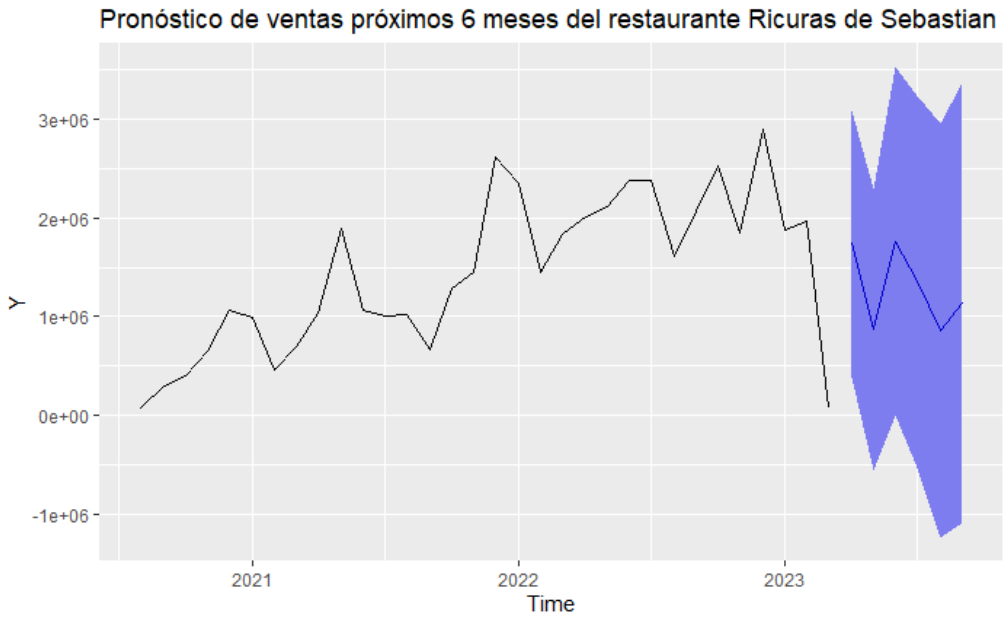


Figura 34. Pronóstico basado en modelo ARIMA. Fuente: Elaboración propia

En la figura anterior se observa gráficamente el pronóstico que está prediciendo el modelo basado en los datos y se observa un comportamiento similar al comportamiento del pasado, donde se mantiene la estacionalidad y la tendencia creciente a lo largo de los años.

Para finalizar se obtiene una tabla donde se resume el pronóstico promedio y los grados de certeza del 95% alto y bajo emitido por el modelo, que nos permitió responder a la pregunta de negocio del cliente.

	Point Forecast <dbl>	Lo 95 <dbl>	Hi 95 <dbl>
Apr 2023	1738566.4	392785.448	3084347
May 2023	870697.8	-554337.010	2295733
Jun 2023	1763765.0	-615.529	3528145
Jul 2023	1355070.3	-531422.211	3241563
Aug 2023	858105.7	-1236653.584	2952865
Sep 2023	1140054.3	-1085687.641	3365796

Tabla 3. Tabla pronóstica basado en modelo ARIMA. Fuente: Elaboración propia

7.8. Acceso a los datos y descubrimientos relevantes

Continuando con la metodología CRISP-DM, pasamos a la etapa de evaluación e implementación y sincronizándolo con el flujo de los datos de la organización, continuamos con el acceso a los datos por medio de GoogleCloud y la interacción del cliente con los descubrimientos relevantes a través de los datos se realiza en la herramienta Looker Studio de Google.

En la figura 8, se representa cómo la empresa TuChat facilita el acceso a los datos relevantes de la empresa cliente Las Ricuras de Sebastian para observar los resultados y datos relevantes del procesamiento de datos anteriormente realizado:

Figura 35: Acceso a los datos y descubrimientos relevantes. Elaboración personalizada para la empresa TuChat.

Este acceso a los datos de la empresa cliente Las Ricuras de Sebastian para observar los resultados y descubrimientos relevantes del procesamiento de datos anteriormente realizado, nos lleva a cumplir con el objetivo específico número dos, donde se debía organizar y estructurar un repositorio o lago de datos. Para ello, es indispensable validar y utilizar el trabajo exploratorio realizado anteriormente. Allí se evidencia que las bases de datos de estudio están relacionadas entre sí, ya sea por el número de celular, nombre, dirección o una mezcla de estas variables. El siguiente esquema expresa la relación que hay entre las bases de datos y cómo al momento de explorar se pueden realizar multifiltros entre todas.

A continuación, explicaré la relación que hay entre las bases de datos creadas para cada cliente por medio de un diagrama entidad - relación.

En la DB_Usuarios se tiene una llave principal que es el Celular del cliente y esta base de datos tiene una relación de uno a uno o muchos con la DB_Traking, ya que un usuario se encuentra una única vez en la DB_usuarios pero puede tener varios tracking de varios pedidos en diferentes días en la DB_Traking. Ahora, esta DB_Usuarios tiene una relación de uno a cero o muchos con la DB_Pedidos ya que un único usuario puede haber realizado muchos pedidos o no tener ningún

pedido. En esta DB_Pedidos la llave foranea es el célular ya que en la DB_usuarios es la llave principal.

La llave principal en la DB_Pedidos es el “id”; este representa un consecutivo que se le otorga a cada pedido para identificarlo y cada día se reinicia a cero.
La DB_Pedidos tiene una relación uno a uno con su llave principal que es el “Id” con la DB_Comentarios, ya que para un “id” de pedido sólo existe un único comentario.

En la siguiente figura se puede observar las relaciones explicadas que hay entre una base de datos y otra.

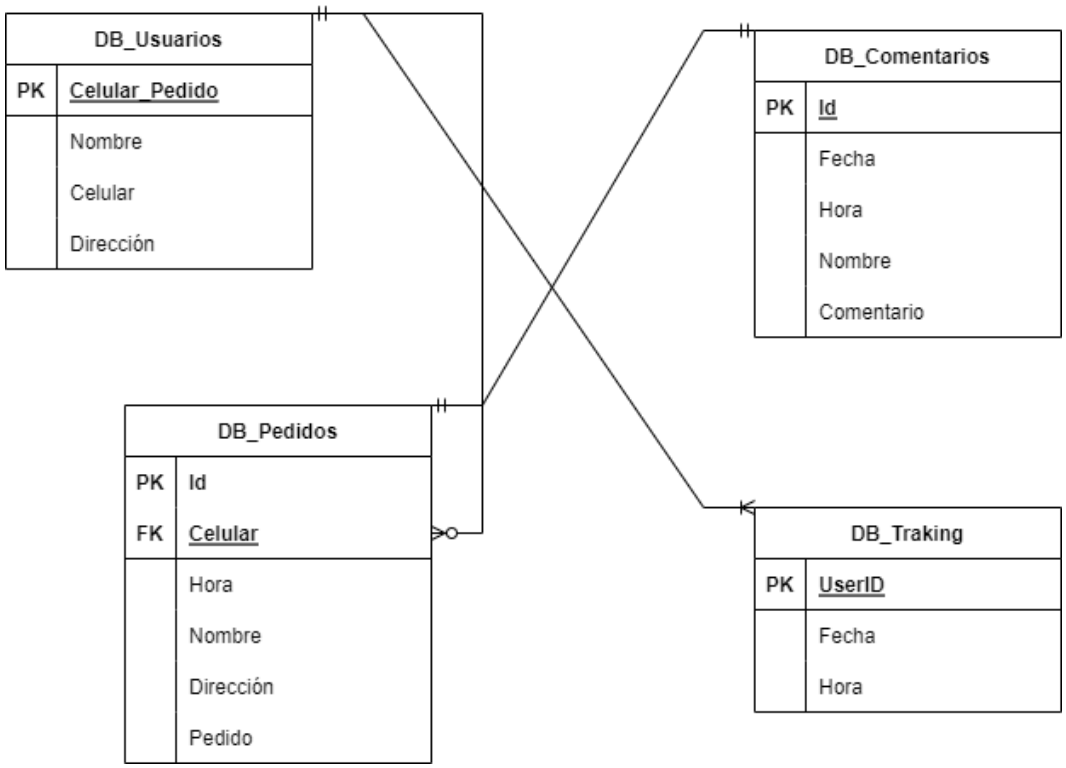


Figura 36: Diagrama entidad - relación entre las bases de datos. Fuente: Elaboración propia

De esta forma, el repositorio se debe estructurar bajo un modelo integrado teniendo en cuenta los siguientes componentes: la seguridad, las fuentes de datos, los procesos de ingesta de datos, el procesamiento, análisis de los datos y la visualización de estos.

El modelo conceptual representado en la figura 39 detalla las fuentes de datos que se consideran como entrada y salida del modelo.

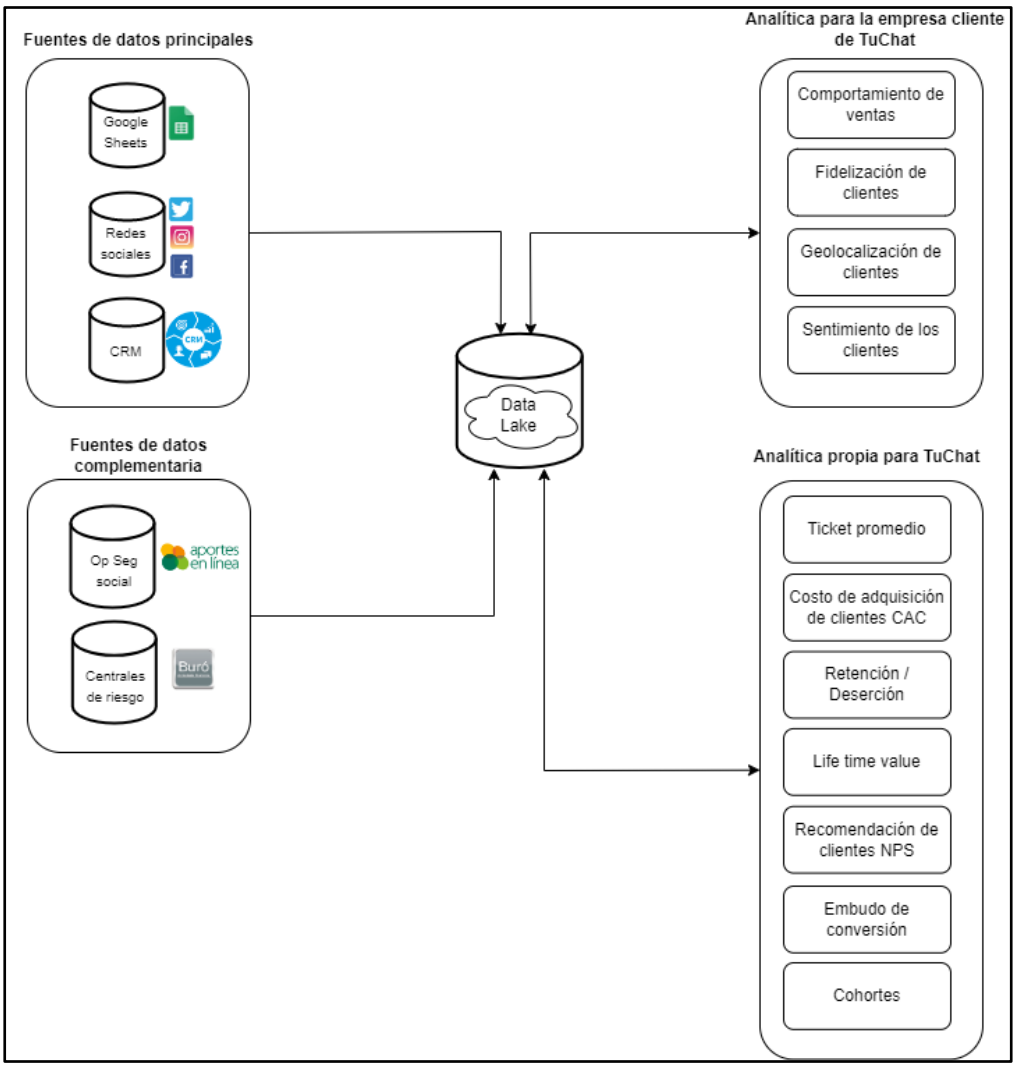


Figura 37: Modelo conceptual detallado de fuentes de datos y salidas.

Fuente: Elaboración propia

Finalizando con la metodología CRISP-DM y de acuerdo al flujo de datos en una organización finalizamos con la etapa de la implementación y el acceso a los datos relevantes y descubrimientos con los datos para dar cumplimiento al objetivo específico número tres, donde se logró potenciar todo este análisis exploratorio de los datos y depuración de las bases de datos actuales, obteniendo indicadores clave como el total de ventas, la cantidad de pedidos, el ticket promedio y la cantidad de clientes que han realizado un pedido, así mismo graficar la cantidad, monto y comportamiento de ventas mes a mes.

Relacionando la base de datos de usuarios con la base de datos de los pedidos, se puede analizar que tan frecuente es un cliente, a cuál sede es la que más realiza pedidos el cliente, cuál es el monto que compra en un rango de tiempo dicho cliente, logrando así analizar ya sea por su celular, nombre o dirección.

Las siguientes imágenes muestran la propuesta de tablero integrador para la visualización de las empresas donde podrán comprender la información generada con los datos y convertir los datos en información estratégica para la toma de sus decisiones:

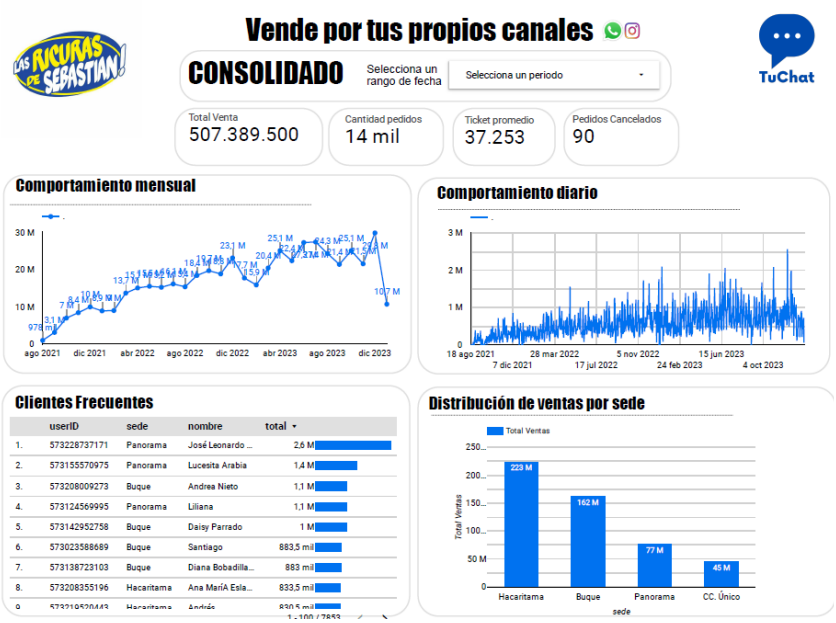


Figura 38: Propuesta de tablero integrador para visualización de información espacial y temporal.

Fuente: Elaboración propia

En la gráfica anterior se puede observar que la propuesta de tablero integrador tiene varias secciones: Inicia con un campo para seleccionar el rango de fecha deseado para observar, una vez el usuario selecciona el rango de fecha, todos los datos se actualizan en base a ese rango seleccionado. Continúa con los cuadros de resultados, que son: El total general de la venta, la cantidad de pedidos realizados por los clientes, el ticket promedio y la cantidad de pedidos cancelados. En la siguiente sección se continúan dos gráficas de serie de tiempo, una mensual y la otra diaria, para observar cómo ha sido el comportamiento de las ventas de todas las sedes del restaurante en el rango de fecha seleccionado por el usuario. Se continúa con una tabla que organiza a los clientes en forma descendente de acuerdo a sus compras al restaurante, obteniendo así el cliente que más ha comprado en la primera posición; con esta información el restaurante puede crear una campaña de fidelización por sede y premiar a esos clientes que son los que más consumen y solicitan pedidos a domicilio al restaurante.

En la siguiente página del tablero integrador se obtiene el pronóstico de ventas de los próximos seis meses, donde se identifica una tendencia creciente con la misma estacionalidad marcada en los meses anteriores y el análisis de sentimiento del cliente, después de vivir la experiencia de solicitar un pedido por Whatsapp. Allí se identifica que la mayoría de los comentarios acerca de la experiencia vivida es positiva; esto se confirma en la nube de palabras donde se observa que las palabras de mayor tamaño son las más repetidas, como: excelente, bueno, gracias, super, ok. También se observa que la calificación del sentimiento promedio es de tres (3) y este tiene una escala dentro del modelo de menos cinco a cinco positivo, lo que significa un sentimiento positivo. Allí también se observa el comportamiento de este sentimiento a lo largo del tiempo, al inicio empezó decreciendo pero con el tiempo y experimentación con el canal de venta nuevo por Whatsapp, este sentimiento fue creciendo. Esta información se puede observar en la siguiente figura.

Figura 39: Propuesta de tablero integrador para visualización de información espacial y temporal.

Fuente: Elaboración propia

En la tercera página del tablero integrador se observan varias capas del proceso de georeferenciación realizado, donde por medio de las direcciones que son escritas en lenguaje natural, se extrae la latitud y longitud para poder realizar estos mapas e identificar las zonas de donde los clientes más solicitan pedidos a domicilio por medio del canal de venta Whatsapp

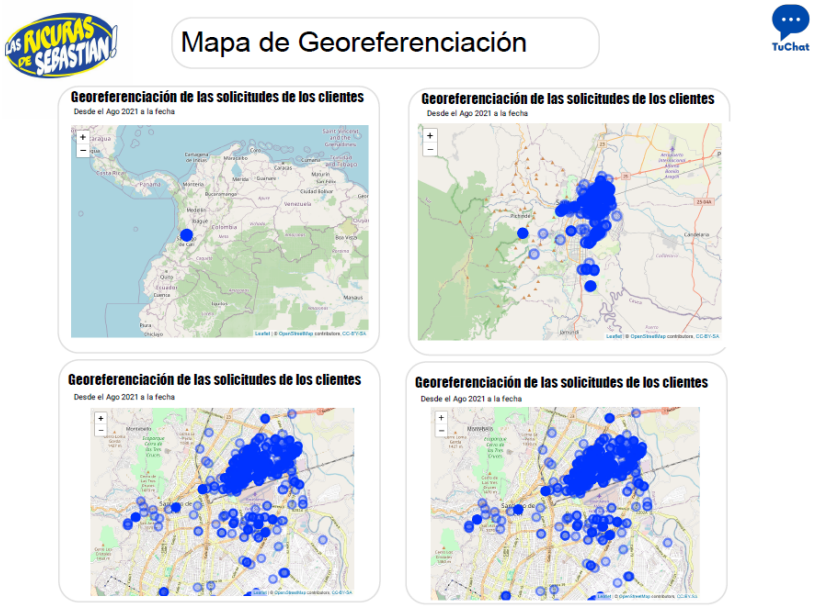


Figura 40: Propuesta de tablero integrador para visualización de información espacial y temporal.

Fuente: Elaboración propia

La propuesta interactiva del tablero integrador para visualización de información espacial y temporal, se puede observar escaneando el siguiente código QR:

Figura 41: Propuesta interactiva de tablero integrador para visualización de información espacial y temporal

Finalmente, cumpliendo con el objetivo general del proyecto, se logra desarrollar un modelo de analítica de datos que potencie, tanto para TuChat como para las empresas clientes, cada uno de los datos que son almacenados actualmente y lograr convertirlos en información estratégica para el negocio. Este modelo inicia con la recolección de los datos donde el insumo son los chatbots, de allí se alimentan las bases de datos que están centralizadas en Google Cloud, se continúa con la consolidación de las bases de datos, el análisis exploratorio de los datos, la limpieza de los datos, una vez finalizados estos procesos se continúa con el modelado de procesos de geocodificación para ubicar a los clientes, la ejecución del análisis de sentimiento identificando si el cliente está feliz o molesto y un modelo de pronóstico para predecir las ventas de los meses futuros. Una vez la data ya está lista, se realiza la conexión y visualización de los datos y se le entrega al usuario para que observe lo que está sucediendo en tiempo real, tome decisiones informadas y actúe para el bien de su negocio.

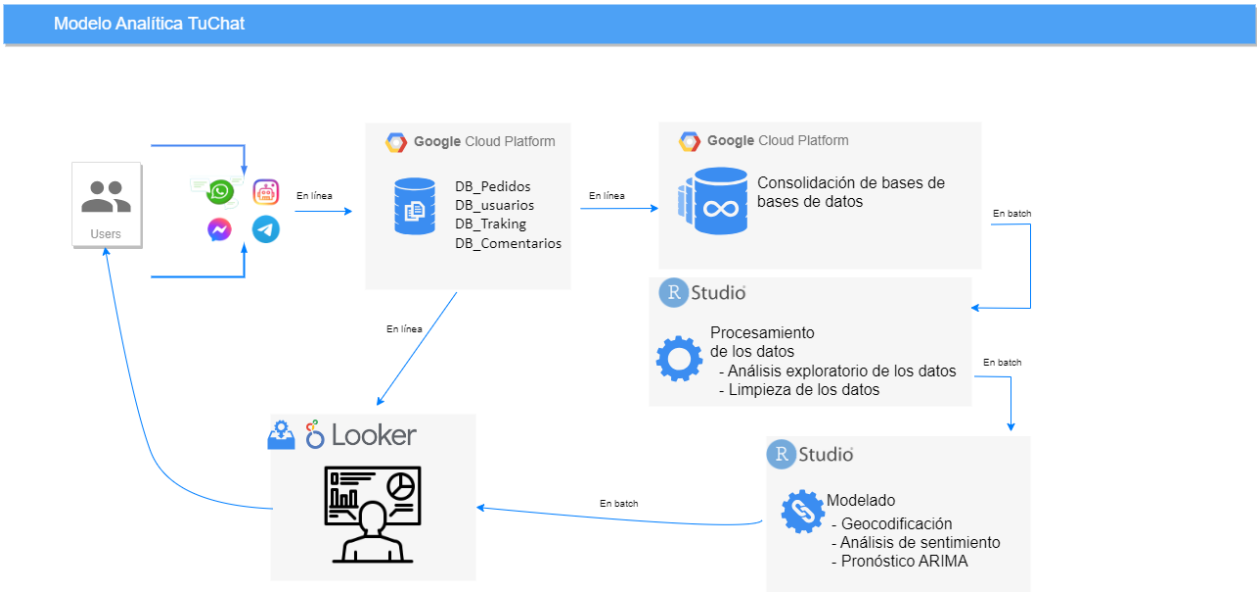


Figura 42: Modelo de analítica TuChat. Fuente: Elaboración propia

8. CONCLUSIONES Y TRABAJOS FUTUROS

En el transcurso de este proyecto, se abordó el desafío estratégico de TuChat, una empresa especializada en chatbots, al enfrentarse a la necesidad de transformar datos conversacionales en información valiosa para sus clientes. El enfoque analítico integral propuesto se puso a prueba con "Las Ricuras de Sebastián", un cliente de TuChat, con el objetivo de organizar, analizar y visualizar eficientemente la información obtenida a través de chatbots. A lo largo de este proyecto, se llevó a cabo un proceso estructurado que incluyó análisis exploratorio, limpieza y consolidación de bases de datos, revelando patrones significativos en el comportamiento de los clientes. La implementación de modelos avanzados, como la geocodificación, análisis de sentimiento y modelos ARIMA, permitió optimizar la oferta de productos, evaluar la satisfacción del cliente y pronosticar las ventas futuras. Las conclusiones presentadas a continuación resumen no solo los logros alcanzados, sino también las áreas de mejora identificadas, destacando la relevancia de una analítica de datos efectiva en la toma de decisiones y la optimización de operaciones comerciales.

8.1. CONCLUSIONES

1. Con la integración de las diferentes fuentes de datos se lograron realizar análisis más completos que permiten conocer mejor el comportamiento de los clientes a nivel temporal (horarios y proyecciones de ventas), espacial (zonas con mayor concentración de clientes) y percepción del servicio (comentarios y análisis de sentimiento). Con este nuevo servicio, los clientes pueden tener una visión más integral de su negocio y tomar decisiones que les permitan crecer y mejorar en los aspectos que aparecen como limitantes.

2. Se identificaron aspectos que se pueden mejorar en el proceso de captura de datos espaciales ya que el proceso de geocodificación requiere datos muy estandarizados en las direcciones. Con este punto se realizaron mejoras en los bots que permiten capturar directamente la coordenada desde el GPS del celular y con esto el análisis espacial se realiza de manera más apropiada y automatizada.

3. Se realizó un análisis de sentimiento tradicional basado en un léxico de palabras con sentimientos previamente codificados y este arroja unos resultados interesantes y muy útiles para evaluar las percepciones de los servicios de los clientes. Sin embargo, concluimos que es importante avanzar en este aspecto integrando modelos de inteligencia artificial que permitan predecir con mejor desempeño la realidad de los sentimientos de los usuarios.

4. Se identificó que los clientes más frecuentes se encuentran a máximo 3 km a la redonda de los restaurantes. Esto es curioso, ya que estando tan cerca de éste, pagan el costo del domicilio.

5. Se logró desarrollar un modelo de analítica de datos que permitió organizar y aprovechar los datos almacenados por TuChat y las empresas asociadas, generando valor a través de la obtención de información estratégica. Esto se traduce en una mejora en la atención al cliente, la identificación de clientes de alto valor y el potencial de expansión del modelo a otras empresas.

6. Se demuestra que los datos almacenados por una empresa y sus empresas asociadas tienen un gran potencial para generar valor. Al organizar y estructurar un repositorio de datos, se podrá obtener información estratégica que permita tomar decisiones basadas en datos y optimizar las operaciones del negocio. Además, la aplicación de técnicas de análisis estadístico y minería de textos permitirá identificar patrones y tendencias que pueden ser aprovechados para campañas de marketing y retención de clientes.

7. El modelo de analítica de datos desarrollado en este proyecto puede ser aplicado a otras empresas que enfrenten necesidades similares. La estructura y metodología utilizada pueden ser adaptadas y replicadas en diferentes industrias, lo que brinda la oportunidad de expandir el alcance y generar impacto en un mayor número de organizaciones.

8.2. TRABAJOS FUTUROS

Aunque se cumplió con lo planteado inicialmente para el desarrollo del proyecto, hay algunos aspectos que se pueden llevar a cabo en un futuro para mejorar la calidad del proyecto o para ampliar los servicios que se pueden ofrecer:

1. Implementar la integración con otras fuentes de datos relevantes, como datos de redes sociales, datos demográficos o datos de fuentes externas. Esto ampliará el alcance del análisis y brindaría una visión más completa del negocio y sus clientes.

2. Explorar el uso de técnicas de procesamiento de lenguaje natural avanzadas para mejorar el análisis de sentimiento y la comprensión del texto de los clientes. Esto podría incluir el uso de modelos de lenguaje pre entrenados, análisis de emociones o análisis de intención del cliente.

3. Mantener la mejora continua del modelo de analítica de datos, con la recopilación de más datos y realizando un análisis más profundo, para identificar nuevas métricas e indicadores clave que brinden aún más valor tanto a TuChat como a las empresas asociadas. Se puede trabajar en la refinación y ampliación del modelo para obtener información estratégica más precisa y de mayor valor.

4. Explorar modelos no supervisados para descubrir patrones ocultos y segmentos de clientes que no sean evidentes a simple vista. Esto puede ayudar a personalizar aún más las estrategias de marketing y mejorar la experiencia del cliente.

9. REFERENCIAS BIBLIOGRÁFICAS

1. E. Doudchitzky, Comercio conversacional ¿Qué es y qué beneficios tiene?, chattigo. Acceso: [En línea]. Disponible: https://blog.chattigo.com/comunicacion-omnicanal/comercio-conversacional
2. Pedro Jiménez Martín, Jesús Sánchez Allende. "De Eliza a Siri: La evolución, tecnología y desarrollo, volumen XIII. 2015 [En línea]. Disponible: https://revistas.uax.es/index.php/tec_des/article/view/616/572
3. Eleni Adamopoulou, Lefteris Moussiades. "Aprendizaje automático con aplicaciones", volumen 2, Dic 2020 [En línea]. Disponible: https://www.sciencedirect.com/science/article/pii/S2666827020300062
4. J. Silge, D. Robinson, Text Mining with R, Boston, 2017.
5. T. Wilches, b2chat. Mayo 15, 2023.[En línea]. Disponible: https://www.b2chat.io/blog/chatbots/ejemplos-de-chatbots-exitosos-en-diferentes-industrias/
6. Dipanjan Sarkar "Text Analytics with Python", 2019. pag -583-646.
7. Estado de la mensajería 2020 [En línea]: disponible en: https://www.zendesk.com/service/messaging/state-of-messaging-2020/
8. Auribox Training. ¿Qué es un Data Warehouse | Business Intelligence? Junio de 2017. [En línea]: disponible en: https://www.youtube.com/watch?v=jFsRdTcljeU [accedido en 2020]
9. O. Salcedo, R. Milena, L. Rodriguez."Metodología crisp para la implementación Data Warehouse". Scielo Articles. Jun 2010. [En línea]. Disponiblehttp://www.scielo.org.co/scielo.php?pid=S0123-921X2010000100005&script=sci_arttext
10. Dipanjan Sarkar "Text Analytics with Python", 2019. pag -583-646
11. D.ortiz, cyberclick, Online Marketing & Digital marketing Data, Nov. 24, 2023. [En línea]. https://www.cyberclick.es/numerical-blog/que-es-un-dashboard
12. M. Intelligence™, mordorintelligence, Tamaño del mercado y análisis de acciones, 2023. [En línea]. https://www.mordorintelligence.com/es/industry-reports/global-business-

intelligence-bi-vendors-market-industry
13. Fundamentals power bi, Microsoft, USA, Marz. 22, 2023.[En línea]. Disponible: https://learn.microsoft.com/es-es/power-bi/fundamentals/power-bi-overview
14. Tableau, USA, Dic. 2023.[En línea]. Disponible: https://www.tableau.com/es-es/why-tableau/what-is-tableau
15. Looker, Google Cloud Platform , USA, Dic. 2023.[En línea]. Disponible: https://lookerstudio.google.com/overview
16. E.Paredes, M. Velasco, "Análisis y Diseño de Sistemas de Información", Unipamplona, 2010. [En línea]. https://www.unipamplona.edu.co/unipamplona/portalIG/home_109/recursos/octubre2014/administraciondeempresas/semestre7/11092015/analisisydisenosistinformacion.pdf
17. Dr. Rodríguez Cortés Francisco. Teoría Análisis Geográfico con Patrones Espaciales, Julio 2018. Simposio Internacional de Estadística, Universidad Nacional de Colombia.
18. The R Project for Statistical Computing [En línea]: disponible en: https://www.r-project.org/
19. Dr. Rodríguez Cortés Francisco. Teoría Análisis Geográfico con Patrones Espaciales, Julio 2018. Simposio Internacional de Estadística, Universidad Nacional de Colombia.
20. J. Silge, D. Robinson, Text Mining with R, Boston, 2017.
21. Google, Comunidad de ayuda Google, Ene. 2024. [En línea] Disponible: https://support.google.com/docs/answer/3093340?hl=es-419#zippy=%2Cpermisos-y-acceso%2Crendimiento%2Cactualidad-de-los-datos%2Cdetalles-t%C3%A9cnicos-y-pr%C3%A1cticas-recomendadas
22. Google, Comunidad de ayuda Google, Ene. 2024. [En línea] Disponible: https://support.google.com/docs/answer/3093343?hl=es-419
23. K.Beck, M.Beedle, agilemanifesto, 2001.[En línea]. Disponible: https://agilemanifesto.org/
24. M. Olga, S. José. Lean Six Sigma. Definición método Lean Six Sigma, Science Direct. Sept 2012. [En línea]: https://www.sciencedirect.com/science/article/pii/S0123592312702140
25. Coursera, Jun. 15, 2023.[En línea]. Disponible: https://www.coursera.org/mx/articles/what-does-a-data-engineer-do-and-how-do-i-become-one
26. D. Sarkar, Análisis de Texto con Python, 2da edición, Bangalore, India, Apress,2019, pp. 23
27. Datta, P. and Nwankpa, J. K. (2021). Digital transformation and the covid-19 crisis continuity planning. Journal of Information Technology Teaching Cases, 11:81–89.
28. Brown, M., & Wilson, G. (2020). The Rise of Chatbots: A Timeline. ChatGPT Blog. [https://dl.acm.org/doi/abs/10.1145/3064663.3064672].
29. Lee, S., & Lee, S. (2019). Chatbot technology in the restaurant industry. [https://www.sciencedirect.com/science/article/pii/S1447677020302102#sec1].
30. Inuk D. Ene. 2024. Chaty Blog. [En línea] Disponible: https://chaty.app/es/blog/5-customer-support-options-for-small-businesses-in-2024/
31. Smith, J., & Johnson, K. (2021). Leveraging Chatbot Data for Business Insights [En línea] Disponible: https://researchberg.com/index.php/eqme/article/view/46

32. Bögershausen, J. Castelo, N. Hildebrando, C. P. Henkel. (2023). Case Study: Crear bots de servicio al cliente que la gente no odie. [En línea] Disponible: https://hbr.org/2023/10/creating-customer-service-bots-that-people-dont-hate