Escuela Colombiana de Ingeniería Julio Garavito

Asunto: Solicitud de Autorización para la entrega del trabajo de grado ajustado, firmado y aceptado por el director.

Fecha: 12/01/2024

Solicitan: Rubén Darío Ferro Rugeles, Nicolás Fernández Moncada, Gonzalo Cossio Escobar,

Estimado Profesor José Fernando Jiménez,

Espero que este mensaje le encuentre bien. Nos dirigimos a usted con el propósito de solicitar su autorización para la entrega del documento del trabajo de grado "Pronóstico de Ventas a través de una Aplicación Web Aplicando Machine Learning y Variables Exógenas", con las correcciones y ajustes solicitados.

Dado que usted ha sido el experto y director del trabajo de grado durante todo el proceso, nos gustaría solicitar su autorización formal para la entrega del documento. Su apoyo y orientación han sido fundamentales para el desarrollo exitoso de este proyecto, y su aprobación es de suma importancia para avanzar en el proceso de graduación. Agradezco sinceramente su tiempo y consideración. En el presente documento podrá poner su autorización, la cual le agradeceríamos que firme en señal de aprobación. Quedamos a la espera de su respuesta y estamos profundamente agradecidos por su valioso apoyo en este proceso académico.

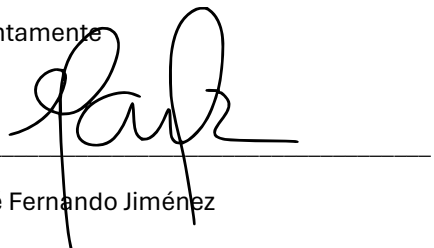
Atentamente, Nicolás Fernández Moncada, Gonzalo Cossio Escobar, y Rubén Darío Ferro Rugeles

Estudiantes de la Maestría en Ciencia de Datos, Universidad Escuela Colombiana de Ingeniería.

Autorización

Yo José Fernando Jiménez por medio del presente documento autorizo a los estudiantes Nicolás Fernández Moncada, Gonzalo Cossio Escobar y Rubén Darío Ferro Rugeles estudiantes de la Maestría en Ciencia de Datos de la Universidad Escuela Colombiana de Ingeniería, de los cuales fui Director del Trabajo de Grado "Pronóstico de Ventas a través de una Aplicación Web Aplicando Machine Learning y Variables Exógenas", para la entrega del documento con los ajustes realizados con base en la sustentación y comentarios de los jurados. He supervisado y guiado a los estudiantes durante todo el proceso de desarrollo de la tesis, y considero que han cumplido con los requisitos y estándares académicos requeridos entregar el documento. Por lo tanto, otorgo mi autorización para que los estudiantes puedan presentar y entregar su documento relacionado con la tesis de grado.

Atentamente



_____

José Fernando Jiménez

Director de Trabajo de Grado

Universidad Escuela Colombiana de Ingeniería

Pronóstico de ventas a través de una aplicación web aplicando *machine learning* y variables exógenas

Rubén Darío Ferro Rugeles
Nicolás Fernández Moncada
Gonzalo Cossio Escobar

Trabajo de grado

Director trabajo de grado:
Prof. Dr. José Fernando Jiménez Gordillo

UNIVERSIDAD
ESCUELA COLOMBIANA DE INGENIERÍA JULIO GARAVITO
MAESTRÍA EN CIENCIA DE DATOS
BOGOTÁ D.C
2023

Agradecimientos

Queremos expresar nuestro más sincero agradecimiento a todas las personas que contribuyeron de manera significativa en la realización de nuestro trabajo de grado. Sin su apoyo y colaboración, este logro no habría sido posible. En primer lugar, deseamos agradecer profundamente a nuestro director de trabajo de grado, el profesor José Fernando Jimenez. Su orientación experta, dedicación incansable y valioso apoyo fueron fundamentales en cada etapa de este proyecto. Nos brindó su valioso tiempo y nos inspiró a superar desafíos, siempre guiándonos hacia nuestras metas académicas. Agradecemos a la Universidad Escuela Colombiana de Ingeniería por brindarnos la oportunidad de cursar nuestros estudios superiores. A todos los profesores y al personal de la universidad, les agradecemos su dedicación y compromiso en proporcionar una educación de calidad. Su apoyo académico y las oportunidades de aprendizaje que nos brindaron fueron invaluables en nuestra formación.

Resumen

Actualmente, el entorno empresarial y organizacional, se enfrenta a una problemática generalizada que trasciende sectores: la gestión ineficiente de los pronósticos de ventas y su falta de implementación. La realización de las predicciones de ventas, dentro de las organizaciones, se realiza de manera manual y utilizando métodos obsoletos, lo cual consume recursos valiosos, tanto en tiempo como en dinero, con consecuencias económicas bastante graves. Se habla de cifras impactantes, cuyos valores oscilan entre 568 millones de pesos a nivel de una sola organización. Estas pérdidas, representan desafíos reales para las empresas, ya que estos recursos podrían destinarse a la expansión, innovación e inversión en otros aspectos cruciales del negocio aumentando su crecimiento.

En este contexto, la precisión en las proyecciones de ventas es vital, ya que cualquier margen de error puede tener un impacto significativo en los resultados financieros. En un mundo empresarial altamente competitivo, la capacidad de identificar patrones y relaciones entre indicadores y ventas es un activo invaluable. La implementación de soluciones de pronóstico de ventas precisas y eficientes no solo permite recuperar pérdidas, sino que también impulsa el crecimiento sostenible y la toma de decisiones informadas, lo que convierte esta área en una prioridad en cualquier organización que busque el éxito en el mercado global.

Con base en las razones y el contexto expuesto anteriormente, este proyecto de grado tiene como objetivo desarrollar una solución, que incluya un modelo de predicción de ventas, el cual emplee técnicas de *machine learning* previamente seleccionados e investigadas, para la predicción de ventas, incorporando diferentes variables exógenas que impactan significativamente en el mercado. Para los modelos seleccionados, se incluirán variables relevantes tanto del entorno empresarial como del contexto externo. La solución se visualizará a través de una aplicación web, desarrollada para el trabajo de grado.

El ambiente de desarrollo será *Google Colab* y se utilizará una aplicación web desarrollada en Angular como plataforma de visualización de la data. Inicialmente, se extraerán indicadores macroeconómicos del mercado mundial de una plataforma llamada Trading Economics y la página del Banco Mundial, posteriormente se realizara un análisis exploratorio de los datos y se decidirá que variables tanto internas como externas serán de importancia para la realización del modelo.

A continuación, se realizará la evaluación de diferentes modelos de *machine learning* aplicados a Series Temporales, donde se espera predecir las ventas en Euros. En este contexto, se evaluará el modelo con los datos de las ventas de una compañía del sector de tecnología. El modelo deberá aprender de las múltiples variables y pronosticar las ventas. Por último, se realizará la visualización de los datos a través de una aplicación web, donde se mostrará la precisión de predicción del modelo, la data histórica, data pronosticada y las variables de alto impacto en el modelo.

Palabras clave: *Aplicación web, Indicadores macroeconómicos, Machine learning, Pérdida de dinero, Predicción de ventas, Series temporales, Variables exógenas, Variables externas, Variables internas, Visualización de datos.*

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

INTRODUCCIÓN

Este primer capítulo introduce la motivación que impulsa el desarrollo de este trabajo de grado, así como la justificación que se plantea para abordar este problema de investigación. El proyecto se enfoca en desarrollar una solución de analítica predictiva de ventas de una empresa del sector tecnológico que, mediante la utilización de técnicas de machine learning y la consideración de variables exógenas, permitan estimar las ventas de la compañía a corto, mediano y largo plazo. Estos elementos se integran en una aplicación web diseñada para visualizar los datos a través de una serie temporal. A partir de este punto, se detallan los objetivos del proyecto y se resalta tanto su impacto como las contribuciones que se obtendrán con su culminación. Por último, se presenta la estructura y organización del presente documento.

1.1 Motivación

En este trabajo de grado, el enfoque se centra en la aplicación de modelos de machine learning con el propósito de permitir a las empresas tomar decisiones fundamentadas en proyecciones de ventas basadas en datos. Estas proyecciones se presentarán de manera visual a través de una aplicación web, donde los datos que se mostrarán exhibirán su comportamiento y variaciones a lo largo del tiempo, a través de una serie temporal.

La intención es que las empresas puedan anticipar y predecir sus niveles de ventas, en la moneda de su país, en diversos sectores, tipos de clientes y áreas de negocio, lo que les brindará la capacidad de identificar oportunidades en el mercado antes que sus competidores. Asimismo, se busca que estas empresas puedan adoptar medidas y tomar decisiones proactivas en situaciones en las que el mercado indique una disminución en sus ventas. Esta anticipación permitirá una respuesta más rápida y efectiva ante cambios en las condiciones del mercado.

La predicción de ventas es indispensable para cualquier empresa, ya que les permite tomar decisiones informadas en cuanto a la gestión de sus recursos, planificación de producción y estrategias de marketing [1]. Sin embargo, las predicciones basadas en los métodos tradicionales pueden no ser lo suficientemente precisas, lo que conlleva a errores y a una planificación deficiente. En este contexto, el uso de técnicas de machine learning ofrece una gran oportunidad para mejorar la precisión de las predicciones de ventas [2].

Por estos motivos, al utilizar algoritmos de machine learning, se logra procesar grandes cantidades de datos históricos de ventas, así como de otros factores que influyen en las ventas, como el panorama de la compañía, el comportamiento del consumidor y la situación económica. De esta manera, se puede identificar patrones y tendencias que pueden ayudar a predecir

de manera más precisa las ventas [1] [2].

Para las empresas, con métodos tradicionales de planeación, les es más urgente predecir sus ventas para evitar la acumulación de *stock* o satisfacer una demanda inesperada. Sin embargo, en un entorno empresarial tan competitivo y dinámico, predecir las ventas con precisión a través de una solución digital es todo un reto [3].

Basándonos en este contexto, la predicción de ventas demanda una exploración minuciosa de las variables internas y externas de la empresa. Esto permite entender su funcionamiento interno y llevar a cabo un análisis exhaustivo de las variables del entorno. Estas últimas son elementos externos a la empresa, como, por ejemplo, datos como el Producto Interno Bruto (PIB), el Índice de Precios al Consumidor (IPC), el índice de inversión, entre otros. Estos índices son herramientas que facilitan la evaluación de cómo evolucionará el mercado y cuál será la respuesta de la empresa ante estas cambiantes condiciones [3] [4].

Por estas razones, la creación de modelos de regresión basados en conjuntos de datos finitos se posiciona como el enfoque predominante en el ámbito estadístico para la predicción de ventas [5]. No obstante, el desafío del sobre ajuste es una preocupación notoria dentro de este enfoque de modelos de regresión. Adicionalmente, la necesidad de implementar modelos de aprendizaje automático que sean capaces de aprender a partir de variables e historiales introducidos, permitiendo así derivar conclusiones comerciales, adquiere un nivel de utilidad significativo en el contexto de la transformación e innovación digital que están experimentando las empresas [6] [7].

1.2 Justificación

El problema que se quiere abordar es demostrar si se puede generar un modelo escalable de pronóstico de ventas aplicando machine learning y haciendo uso de variables exógenas, por lo cual se evaluarán diferentes algoritmos de machine learning para encontrar aquellos modelos con mayor precisión y concluir si pronostican de manera correcta las ventas de la empresa a corto, mediano y largo plazo.

El motivo por el cual las empresas no utilizan modelos escalables de pronóstico de ventas, es la baja implementación de la inteligencia de negocios *(BI)*, la analítica de datos *(DA)* y ciencia de datos [8] [9]. Las pequeñas y medianas empresas (*PYMEs*) se ven particularmente afectadas por este problema debido a la falta de recursos y la resistencia al cambio [9]. La adopción de tecnologías, herramientas de redes sociales, comercio electrónico y otros tipos de tecnología se enfrenta a múltiples desafíos, incluyendo la falta de habilidades técnicas y eficiencia, problemas de seguridad, falta de apoyo organizacional, costos de tecnología e infraestructura y falta de apoyo gubernamental [10].

Dentro de los desafíos más importantes que presenta la implementación de *BI* y *DA*, por parte de las organizaciones, se encuentra la falta de financiamiento, bajas competencias del personal, el tiempo de transición digital y la infraestructura desactualizada, lo que impide una implementación efectiva de *BI* y *DA* en los proyectos empresariales [9] [10]. En general, esta baja adopción tiene un gran impacto en las empresas, ya que se estima que el 70 % de los proyectos de *BI* fracasan en proporcionar los beneficios esperados [11].

Además, se estima que solo el 21 % de las empresas a nivel global emplean *BI* y *DA* en el contexto de predicción de sus ventas [8]. Los desafíos asociados con la adopción y el uso de los sistemas de *BI* incluyen la resistencia a su uso, falta de motivación, miedo a perder el control sobre la información, falta de conocimientos y competencias técnicas, problemas de

infraestructura, comunicación insuficiente entre el personal de TI y los usuarios de negocios, y problemas de los sistemas implementados [9].

En consecuencia, la escasa utilización de las herramientas anteriormente mencionadas, restringe la capacidad de las empresas para mejorar la toma de decisiones y elevar su ventaja competitiva [12]. Entre los desafíos identificados destaca la falta de aplicación de la analítica de datos en la predicción de ventas, lo cual limita la capacidad de las empresas para planificar y tomar decisiones fundamentadas [12] [13].

Con base en el contexto anterior, para abordar esta problemática, se propone el desarrollo de una solución que junta el BI, DA, Ciencia de datos y desarrollo web en una única solución digital para realizar predicción de ventas que aproveche tanto los datos históricos de ventas como las variables exógenas. La solución permitiría a las empresas identificar patrones en los datos y tomar decisiones organizacionales informadas sobre la planificación de producción, la gestión de inventarios y la asignación de recursos.

Además, al proporcionar una visión más clara de las tendencias de ventas futuras, las empresas podrían anticiparse a las fluctuaciones del mercado y ajustar sus estrategias en consecuencia. Por otra parte, la solución también ayudaría a abordar los desafíos anteriormente mencionados al ser fácil de usar, intuitiva y accesible para todas las partes interesadas en las empresas. En este contexto se plantea la pregunta, ¿Por qué escoger usar la ciencia de datos para resolver un problema del negocio?, porque desde un punto de vista investigativo y científico, la ciencia de datos ha demostrado ser el mejor campo para modelar situaciones complejas y reales brindando una solución efectiva y confiable.

Con respecto a los modelos de machine learning, tienen la capacidad de analizar y aprender en profundidad de grandes cantidades de datos para identificar patrones que probablemente un humano no podría, y con esto dar una solución o recomendación más precisa para tomar acciones en el momento adecuado y con información [14]. Finalmente, el desarrollo web permitirá dar una escalabilidad a los modelos implementados y probados, a tal nivel, que cualquier empresa sin la experticia en ciencia de datos, sea capaz de interpretar la información visualizada, es decir: Variables de alta correlación y predicción de las ventas, para poder tomar decisiones basadas en información con el fin de aumentar sus ventas o reducir sus pérdidas. Una empresa que maneja su incertidumbre puede generar múltiples estrategias en función de diferentes predicciones realizadas en el tiempo y así prepararse lo antes posible para diferentes escenarios positivos o negativos. En resumen, la solución desarrollada reunirá diferentes campos del área de datos y desarrollo web para trasladarlos al mundo real.

1.3 Objetivos

1.3.1. General

Desarrollar una solución de analítica predictiva de ventas (en Euros) de una empresa del sector tecnológico que, mediante la utilización de técnicas de machine learning y la consideración de variables exógenas cuantitativas, permitan estimar las ventas de la compañía a corto, mediano y largo plazo.

1.3.2. Específicos

- Elaborar un modelo de *web scrapping* que capture la variables macroeconómicas más importantes validadas por expertos de la compañía.

- Realizar un análisis exploratorio de datos (EDA) para identificar y caracterizar la evolución de las ventas y las variables exógenas asociadas con la solución.
- Seleccionar modelos de machine learning de series temporales que permita predecir el comportamiento de las ventas teniendo en cuenta múltiples variables internas de la compañía y externas del mercado.
- Desarrollar un código escalable que permita la estimación de ventas de otra compañía a corto, mediano y largo plazo.
- Desarrollar una herramienta de visualización en una aplicación web que permita presentar el histórico y la estimación de ventas de la compañía a corto, mediano y largo plazo.

1.4 Contribución

Este trabajo de investigación presenta las siguientes contribuciones en el campo de la predicción de ventas, inteligencia de negocios y analítica de datos:

- **Mejora de la precisión en predicciones de ventas:** Investigar y desarrollar técnicas avanzadas de machine learning con el objetivo de perfeccionar los modelos de predicción de ventas, centrándose en la optimización de estos modelos para tener mejores resultados en los pronosticos futuros.
- **Mejora de la toma de decisiones organizacionales:** Utilizar análisis de datos y machine learning para extraer conclusiones valiosas a partir de patrones ocultos en los datos de ventas, proporcionando información relevante para comprender las tendencias del mercado y formular estrategias comerciales efectivas, que permitan mejorar la toma de decisiones de las ventas.
- **Desarrollo de una aplicación web de visualización de datos:** Proporcionar una aplicación web que permita visualizar de manera interactiva los datos de ventas y las predicciones generadas por las técnicas de machine learning, proporcionará a los usuarios una manera fácil y eficiente de explorar y entender los *insights* extraídos de los datos.
- **Desarrollo de un modelo escalable:** Desarrollar un código que seleccione el mejor modelo de predicción multivariada de ventas (en Euros) según el comportamiento de las ventas de otra compañía, teniendo en cuenta los errores MAPE y RMSE.

1.5 Organización del documento

En esta sección se presenta como está organizado el documento escrito de este trabajo de grado.

- El capítulo 2, realiza una revisión profunda de la literatura relacionada con series temporales, así como con la predicción de ventas y el uso de machine learning en este contexto. De la misma manera, la justificación de las variables exógenas, internas y externas utilizadas para la predicción. Finalmente, la importancia de la visualización de datos y las

herramientas y formas para realizar predicción de ventas. Para esto, se analizan estudios previos para comprender las tendencias y los desafíos existentes en el campo, y se extraen conocimientos fundamentales para respaldar el desarrollo de la metodología y los resultados del presente trabajo.

- El capítulo 3 se centra en la metodología utilizada en este trabajo.

  - Comienza con la búsqueda y selección de datos relevantes para el estudio del mercado. Luego, se identifican las variables internas y externas que más influyen en las ventas.
  - A continuación, se presenta el análisis exploratorio de datos, que incluye la preparación y limpieza de los mismos para análisis posteriores.
  - Luego, se procede a la selección de modelos de machine learning, donde se eligen los algoritmos más adecuados. Estos modelos se someten a evaluación, entrenamiento y refinamiento, detallando el proceso de ajuste y optimización. Posteriormente, se analizan y presentan los resultados de los modelos, explicando su interpretación y comparación.
  - Por último, se describe el diseño y desarrollo de la aplicación web para la visualización de datos, incluyendo la implementación de la interfaz de usuario para una presentación efectiva de la información.

- El capítulo 4, presenta los resultados de la información organizada y procesada de los datos seleccionados y obtenidos del mercado. A continuación, se detallan los resultados del análisis exploratorio de datos con sus respectivas conclusiones y hallazgos encontrados en los datos. Posteriormente, se describen los modelos seleccionados y los criterios que se tuvieron en cuenta y se realiza su respectiva evaluación utilizando diversas métricas. Finalmente, se presenta la aplicación diseñada e implementada.

- El capítulo 5, lleva a cabo un análisis y discusión profunda de los resultados obtenidos en el marco de este proyecto de investigación. Este análisis se centra en contrastar y comparar los resultados obtenidos en el desarrollo del modelo de predicción de ventas con las investigaciones y hallazgos previamente documentados en la literatura. Mediante esta comparativa, se evalúa la coherencia y consistencia de los resultados alcanzados en este estudio.

  Este ejercicio de comparación enriquece la comprensión de la relevancia y validez de los resultados del proyecto en el contexto más amplio de la investigación en predicción de ventas y el uso de técnicas de machine learning en este dominio. Asimismo, brinda una visión integral de cómo los nuevos conocimientos obtenidos contribuyen al avance de la disciplina y a la toma de decisiones empresariales fundamentadas.

- El capítulo 6, expone las conclusiones derivadas de los resultados obtenidos en este proyecto de investigación. Estas conclusiones se basan en un análisis profundo de los datos y los hallazgos que se han presentado en los capítulos anteriores. A partir de la evaluación detallada de los resultados y su comparación con los objetivos planteados en la investigación, se extraen las implicaciones más relevantes y se formulan las conclusiones.

- El capítulo 7, explora las posibles direcciones para futuras investigaciones, considerando los hallazgos obtenidos en este estudio. Se detallarán áreas de estudio que podrían beneficiarse de un análisis más profundo y se sugieren aspectos específicos que podrían abordarse en futuros trabajos.

Capítulo 2

ESTADO DEL ARTE

En este capítulo, se aborda una perspectiva teórica fundamental en el ámbito de la predicción de ventas utilizando series temporales y modelos de machine learning. Se exploran conceptos clave relacionados con la predicción de ventas y su relevancia en el entorno empresarial actual, considerando las variables exógenas, internas y externas que impactan en este proceso. Además, se presentan los fundamentos teóricos detrás de los modelos de machine learning utilizados en la literatura, destacando su importancia en la mejora de la precisión en las predicciones de ventas.

A medida que se avanza en el capítulo, se introduce una revisión exhaustiva sobre las metodologías y enfoques previos en la literatura que se han sido aplicados exitosamente en la predicción de ventas. A través de esta búsqueda literaria, se busca establecer una base sólida para comprender la evolución y las tendencias en este campo, y cómo influirán en la metodología adoptada en el presente trabajo.

En este proyecto, para la revisión bibliográfica inicialmente, se identifico que metodología utilizar para la búsqueda de la literatura. Particularmente, se adopto la estrategia PICO como enfoque principal de la búsqueda literaria. Se definió la población objetivo (P), como empresas, startups y compañías interesadas en la predicción de ventas. En cuanto a la intervención (I), se analizo los diversos procesos y métodos empleados para realizar estas predicciones. Se realizo una comparación (C), detallada basada en los diferentes métodos, algoritmos y herramientas digitales que se emplean en la actualidad para la predicción de ventas. Finalmente, los resultados de investigación (O), se enfocaron en determinar cuáles son los métodos y estrategias más efectivas y precisas para llevar a cabo predicciones de ventas. Para la búsqueda de la literatura se planteo y utilizó la siguiente ecuación 1 basada en la lógica Booleana.

$$\begin{aligned} &(\textit{“Sales Forecasting”}\ \text{OR}\ \textit{“Sales Prediction”})\ \text{AND} \\ &(\textit{“Machine Learning”}\ \text{OR}\ \textit{“Predictive Analytics”})\ \text{AND} \\ &(\textit{“Small and Medium-sized Enterprises”}\ \text{OR}\ \textit{“SMEs”})\ \text{AND} \\ &(\textit{“Time Series Analysis”}\ \text{OR}\ \textit{“Time Series Forecasting”}) \end{aligned} \tag{1}$$

Con respecto a los criterios de inclusión, se consideraron artículos que proporcionaban información detallada sobre el funcionamiento de sistemas de predicción de ventas, sus características clave y métodos de control, así como aquellos que estaban enfocados en ayudar a empresas a mejorar la precisión de sus pronósticos de ventas. Para los criterios de exclusión, se elimi-

naron artículos que trataban sobre métodos distintos de predicción de ventas, tecnologías y estrategias diferentes a las de análisis de datos aplicado a predicción de ventas.

La búsqueda bibliográfica se llevó a cabo en varias bases de datos académicas relevantes. Se incluyeron las siguientes bases de datos: *ACM Digital Library*, *Springer*, *Science Direct* y *Google Scholar*. A través de un proceso de revisión metodológica y análisis de la literatura encontrada, se identificaron un total de 222 artículos y fuentes literarias. En *Google Scholar* 154, en *Springer* 52, en *Science Direct* 7, y finalmente en *ACM Digital* Library 9. Después, al aplicar nuestros criterios de selección, se incorporaron un total de 90 artículos y fuentes literarias que proporcionaban información completa y detallada sobre el problema de investigación que se abordaría en este trabajo de grado.

2.1 Series temporales

Las series temporales son un tipo de datos que se recopilan a lo largo del tiempo y se utilizan para analizar patrones y tendencias en un proceso en particular [15]. Pueden ser utilizadas en una variedad de campos, incluyendo finanzas, economía, marketing y ciencia de datos. Una serie temporal representa una serie de órdenes basadas en el tiempo [16]. El tiempo se puede representar en años, meses, semanas, días, horas, minutos o segundos, y así sucesivamente. Es una observación de la secuencia de tiempo discreto de intervalos sucesivos [15]. Se tienen dos tipos de series temporales:

- **Estacionaria:** un conjunto de datos debe seguir las siguientes reglas, el valor medio de ellos debe ser completamente constante en los datos durante el análisis, la varianza debe ser constante con respecto al marco de tiempo, la covarianza mide la relación entre dos variables [17].
- **No estacionaria:** si la varianza media o la covarianza cambian con respecto al tiempo, el conjunto de datos se denomina no estacionario [17].

La predicción utilizando series temporales es un tema ampliamente reconocido que ha generado interés en diversas comunidades de investigación, como la estadística, el aprendizaje automático, la econometría, la investigación operativa, las bases de datos, la minería de datos y las redes, entre otros [18]. Estas previsiones se obtienen a través de modelos matemáticos que capturan una relación parametrizada entre valores históricos y futuros. Estos modelos buscan expresar el comportamiento y las características de una serie temporal utilizando parámetros. Dichos parámetros son estimados utilizando un conjunto de datos de entrenamiento, con el objetivo de ajustarse a las particularidades de la serie temporal y minimizar el error de predicción [19].

Con bastante frecuencia, las series temporales presentan una o varias características, denominadas componentes, que ayudan a explicar su comportamiento en el tiempo. Se denomina tendencia o componente tendencial de una serie temporal a su comportamiento o movimiento a largo plazo. La tendencia, es una representación de la dirección en la que se mueven los valores de la serie a medida que transcurre el tiempo. En el contexto de análisis de series temporales, la tendencia puede ser creciente, decreciente o incluso mantenerse constante [20].

Para identificar y cuantificar esta tendencia, se utilizan diversos métodos, como el cálculo de la línea recta de regresión que mejor se ajusta a los puntos de la serie temporal. Esta línea recta puede proporcionar información sobre la tasa de cambio promedio en los valores de la

serie en función del tiempo, lo que a su vez permite entender si hay un aumento o disminución gradual en los datos a lo largo de la serie temporal [20]. En la figura 1, se expone la serie temporal, donde la línea recta, representa la tendencia que es creciente en este caso particular.

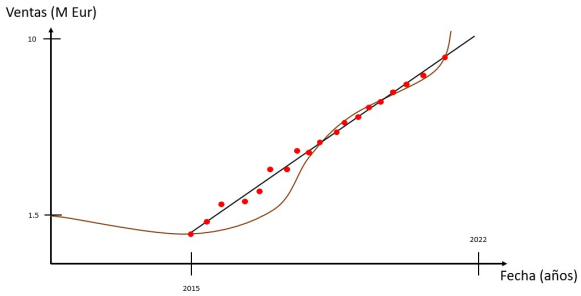


Figura 1: Modelo de línea recta en una serie temporal [20].

Otro elemento de las series temporales es el componente de fluctuación cíclica, este se refiere a las variaciones de mediano plazo que no siguen patrones estacionales predecibles ni son simplemente aleatorias. Este componente evidencia ciclos y tendencias de crecimiento y declive que no se repiten en intervalos regulares. Al aislar esta fluctuación cíclica de otros factores como la tendencia y la estacionalidad, se obtiene una comprensión más clara de los movimientos a mediano plazo en los datos [21]. La figura 2, expone una serie temporal donde se ve la fluctuación cíclica, que puede ser tanto creciente como decreciente en diferentes momentos.

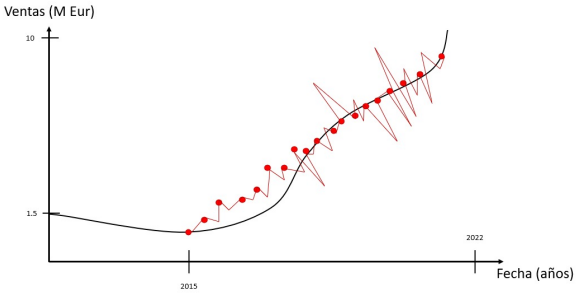


Figura 2: Ejemplo de fluctuación cíclica en una serie temporal [20].

2.2 Diferencia un pronóstico de series temporales de otro tipo de predicción

La predicción de series temporales se enfoca en predecir el valor futuro de una serie de datos ordenados cronológicamente, basándose en los patrones y tendencias observados en los datos históricos. En comparación con otros tipos de aprendizaje automático, la predicción de series temporales tiene algunas características únicas, como la dependencia temporal y la estacionalidad [22].

La dependencia temporal se refiere a que los valores en una serie temporal están correlacionados entre sí y su orden temporal importa. Es decir, el valor actual de la serie temporal depende del valor anterior, lo que significa que la predicción del valor futuro debe considerar el valor anterior [22] [23].

La estacionalidad se refiere a patrones que se repiten en la serie temporal a intervalos regulares, como los patrones de consumo en las temporadas de vacaciones o los patrones climáticos en las diferentes estaciones del año. Para realizar la predicción de series temporales, se utilizan técnicas de modelado estadístico y matemático, dentro de los más comunes están *AR*, *ARMA*, *ARIMA*, modelos de regresión, entre otros [24].

Parte del objetivo del proyecto es aplicar modelos de aprendizaje automático, que aprendan del comportamiento de diferentes variables que tengan relación con la variable a predecir, las ventas. Estos modelos se prueban en tres ventanas temporales: corto, mediano y largo plazo, con el fin de seleccionar el mejor modelo posible dependiendo la necesidad de predicción [22].

Los retos a los que se enfrentan las series temporales a diferencia de otros modelos no medidos en el tiempo son el ruido y los valores atípicos (*outliers*). Más que todo en aquellos modelos donde las ventas son ocasionales y no recurrentes, por lo que se debe analizar cual es el mejor rango de tiempo a tomar, diario, mensual, anual, etc [25].

La dependencia temporal, en las series temporales, implica que los datos están correlacionados en el tiempo, lo que significa que los patrones y tendencias en los datos históricos pueden ser utilizados para predecir los valores futuros. Sin embargo, esta dependencia temporal también puede ser un desafío en la construcción de modelos precisos, ya que los datos pueden ser más difíciles de analizar y de predecir [26].

La selección de características adecuadas puede ser una tarea de gran complejidad en las series temporales debido a la gran cantidad de variables que pueden estar influyendo en la predicción. La selección de características incorrectas puede afectar negativamente la precisión del modelo, por eso es de vital importancia realizar análisis de multicolinealidad, para determinar cuáles serán aquellas variables que soportarán la predicción de las ventas de manera mas adecuada [26].

Las series temporales pueden tener valores faltantes (*gaps*) en los datos debido a problemas de registro, eventos impredecibles o incluso momentos en los que no hay ventas. La falta de datos puede afectar la precisión del modelo y puede requerir la utilización de técnicas de interpolación de datos [25] [26].

Las series temporales a menudo están sujetas a cambios en el comportamiento, como cambios estacionales, ciclos económicos y eventos impredecibles. La detección y modelización de estos cambios es desafío para los modelos de predicción. Es por esto que el modelo planteado en el proyecto analizará variables internas y externas de la empresa de tecnología a la cual se le aplicará el modelo [18].

2.3 Predicción o pronostico

El pronóstico, se presenta inicialmente desde una perspectiva teórica. En primer lugar, se presenta una definición del término y se destaca su importancia en el contexto del proyecto. El pronóstico se refiere a las estimaciones de valores futuros de una variable a partir de datos históricos y modelos matemáticos, el cual guía la formulación de proyectos de innovación y desarrollo tecnológico [27].

El pronóstico, es una técnica que consiste en la estimación de variables futuras utilizando técnicas estadísticas y de machine learning, por lo cual es una tarea fundamental dentro de la ciencia de datos en una organización [28]. La precisión y confiabilidad de la predicción es de gran importancia para la toma de decisiones organizacionales, entre los indicadores

de presición se encuentran, el error cuadrático medio, error absoluto medio y r2 [29]. Estos indicadores, se explicaran a detalle en el capitulo 3, en la sección 3.3.

La necesidad de las predicciones es inherente y se evidencia para la mayoría de las empresas, independientemente de su sector o actividad, ya que todas ellas deben planificar y gestionar sus recursos de manera eficiente para alcanzar sus objetivos y satisfacer las necesidades de sus clientes [29]. Por ejemplo, una empresa de fabricación debe prever la demanda de sus productos para poder ajustar su producción y no incurrir en costos innecesarios, mientras que una empresa de servicios debe estimar su capacidad y demanda para poder asignar sus recursos de manera más adecuada. [29].

Por estas razones, muchas empresas invierten importantes recursos financieros en la contratación de investigadores operativos y estadísticos, así como en la adquisición de programas informáticos especializados en realizar predicciones, con la finalidad de mejorar la calidad de sus pronósticos de interés [27] [29]. Por otra parte, la combinación, o agregación, de predicciones en las empresas, lo cual es importante resaltar que no es una idea nueva, ha recibido recientemente una mayor atención en la comunidad y ha demostrado dar buenos resultados en las organizaciones [27].

A continuación, se introducen la predicción enfocado a las ventas, detallando sus características, uso y propósito. En este sentido, la predicción de ventas se utiliza para prever el comportamiento de las ventas futuras de una organización, lo que resulta fundamental para la planificación y toma de decisiones empresariales. Finalmente, se presenta una revisión profunda sobre el uso de machine learning aplicado a la predicción de ventas, detallando los diferentes modelos, sus finalidades e impacto en el proceso de predicción de ventas. Asimismo, se mencionan algunas variables externas como datos macroeconómicos y de entorno que pueden ser de utilidad para el proyecto.

2.4 Predicción de ventas

La predicción de ventas, hace referencia a la actividad habitual en la mayoría de las empresas que afecta a las operaciones, el *marketing* y la planificación de las empresas [30]. Por lo cual, utilizar técnicas de predicción de ventas precisas es fundamental para seleccionar con éxito un emplazamiento comercial. Sin embargo, los métodos tradicionales de previsión de ventas minoristas han adolecido de una excesiva subjetividad en el análisis y de una incapacidad para considerar simultáneamente los efectos de múltiples variables [31].

Por otra parte, una de las ventajas más significativas de usar predicción de ventas a través de modelos computarizados e informatizados, es que permite definir estadísticamente las relaciones entre las ventas de las tiendas y las variables influyentes, como las características del emplazamiento, demográficas y competitivas [31].

En el mercado actual se le ha permitido a empresas como Meta tomar acciones con el suficiente tiempo de antelación para evitar pérdidas o aumentar sus ventas. Es así como el dueño de Meta explica que los primeros meses del 2023 su empresa tuvo una caída del 55 % pero gracias a los pronósticos realizados, pudieron contener esta bajada y reducirla a un guía para la formulación de proyectos de innovación y desarrollo tecnológico en un 12 % [32].

Actualmente la predicción de ventas se realiza comúnmente con modelos univariados o regresiones lineales que limitan la predicción a solo un comportamiento o tendencia [33]. Sin embargo. las empresas requieren contratar especialistas para desarrollar estos modelos y es por esto por lo que la propuesta en este proyecto, consiste en realizar un modelo que utilice

machine learning para no limitarnos a una función lineal y univariada, y a través del uso de *Web Scraping*, obtener las variables del mercado que predigan como será el comportamiento y la exactitud de Pronóstico de las ventas en tres (3) ventanas de tiempo diferentes. Las ventanas de tiempo son a corto, mediano y largo plazo [33] [34]. La figura 3, expone algunos de los beneficios que ofrecen los pronósticos de ventas a nivel organizacional y empresarial.

Figura 3: Diagrama acerca de los beneficios de la predicción de ventas dentro de una organización [35].

2.5 Modelos de machine learning

En el ámbito del análisis de series temporales, los algoritmos de machine learning desempeñan un papel fundamental al permitirnos descubrir patrones complejos y extraer información valiosa de datos secuenciales [36]. Estos modelos son especialmente útiles para comprender la dinámica de fenómenos que evolucionan en el tiempo, como las ventas en un negocio. Los métodos supervisados de machine learning son herramientas poderosas para abordar problemas de series temporales [36].

Estos algoritmos utilizan un enfoque basado en ejemplos previamente etiquetados para aprender patrones y relaciones en los datos [36] [37]. Recientemente los avances teóricos en estadística han dado lugar a diferentes métodos para realizar la predicción de series temporales. Se exponen a continuación diferentes modelos de machine learning obtenidos de la literatura la tabla 1, expone el modelo, la aplicación, la técnica, la importancia y la dificultad de los modelos.

1. La regresión de vectores de soporte (SVR) por sus siglas en ingles, perteneciente al área de la minería de datos [6]. La *SVR* es una extensión de la máquina de vectores soporte (*SVM*) y se desarrolla en base al objetivo de minimización del riesgo estructural (*SRM*). Estudios previos han demostrado que *SVR* tiene una buena capacidad de generalización [6] [7].

2. Por otra parte, el modelo *WaveNet*, es un modelo de red neuronal convolucional profunda que se utiliza para modelar series temporales secuenciales. *WaveNet* ha demostrado ser eficaz en la generación de voz y música, así como en la predicción de series temporales de alta resolución [38]. El modelo, se adapta específicamente para el problema de pronóstico de ventas, considerando una arquitectura de secuencia a secuencia donde el codificador y el decodificador no comparten parámetros para manejar errores acumulativos [39].

   Además, se aplican técnicas de suavizado, como promedios móviles y promedios móviles exponenciales, para mejorar la precisión de las predicciones y reducir el sobre ajuste. Este enfoque basado en *WaveNet* demostró ser altamente efectivo en la resolución del desafío de pronóstico de ventas en la competencia, y se posiciona como una herramienta valiosa para abordar problemas de series temporales complejas [38] [39].

3. Otro modelo que se presenta, es *DeepAR*, un modelo de red neuronal recurrente diseñado para la predicción de series temporales multivariadas. Su enfoque principal radica en capturar las relaciones interdependientes entre diferentes variables de entrada, lo que le permite pronosticar con precisión los valores futuros de cada una de estas variables [40]. Esta capacidad para modelar las complejas interacciones entre series de datos lo convierte en una herramienta valiosa para abordar problemas de pronóstico en los que múltiples aspectos están intrínsecamente conectados [40].

   El modelo *DeepAR* aborda el desafío de la predicción de series temporales en un contexto en el que se enfrenta a la necesidad de pronosticar miles o millones de series temporales relacionadas. A diferencia de los métodos tradicionales que se enfocan en pronosticar series individuales o en grupos pequeños, *DeepAR* aprende de un modelo global a partir de los datos históricos de todas las series temporales en el conjunto de datos. Para esto, se basa en una arquitectura de red neuronal recurrente autorregresiva que incorpora una distribución negativa binomial para datos de conteo y un tratamiento especial para lidiar con la variabilidad amplia en las magnitudes de las series temporales [40].

4. Por otra parte, se expone el Modelo de auto regresión Vectorial (*VAR*), el cual es un enfoque estadístico ampliamente utilizado para modelar simultáneamente múltiples series temporales. Este modelo se aplica comúnmente en contextos financieros y económicos para capturar las interrelaciones entre diversas variables económicas [41]. *VAR* ofrece una forma efectiva de analizar cómo las diferentes series temporales interactúan y se influyen mutuamente, lo que lo convierte en una herramienta valiosa para la comprensión y predicción de fenómenos complejos en estos campos [41].

   La aplicación de los modelos de auto regresión vectorial (*VAR*) y auto regresión vectorial bayesiana de Litterman (*LBVAR*) en el contexto de las ventas pronostica un avance significativo en la precisión de las predicciones para la industria tecnológica. Estos modelos multivariados consideran no solo los comportamientos de los clientes, sino también el impacto de factores macroeconómicos en la demanda del mercado. Al incorporar datos

macroeconómicos en el proceso de pronóstico, los modelos *VAR* y *LBVAR* capturan relaciones más completas y realistas, lo que resulta en pronósticos más precisos y adaptativos [42].

5. A continuación se expone el modelo *TCN* (*Temporal Convolutional Network*), el cual es un modelo de red neuronal convolucional que se utiliza para modelar series temporales secuenciales. El modelo, es capaz de aprender patrones a largo plazo y puede ser utilizado para predecir valores futuros de una serie temporal [43]. En el ámbito de la predicción de ventas, predecir la demanda de diferentes productos es crucial para optimizar la planificación de inventario, los procesos de adquisición y las operaciones logísticas [44]. El modelo aprovecha las arquitecturas de redes neuronales convolucionales (*CNN*) para capturar patrones temporales y correlaciones entre los datos de ventas de diferentes productos. Esto es altamente relevante en escenarios minoristas donde se deben pronosticar simultáneamente miles o incluso millones de productos [44].

6. Ahora, se expone el modelo Neural *ODE* (*Neural Ordinary Differential Equation*), el cual es un modelo de red neuronal que utiliza ecuaciones diferenciales ordinarias (*ODE*) para modelar series temporales. Neural ODE es capaz de modelar series temporales con dinámicas complejas y no lineales [45]. *Neural ODE* es un enfoque de modelado flexible que se especializa en predecir la evolución de sistemas dinámicos a lo largo del tiempo. En el contexto de las ventas, las series temporales de datos pueden considerarse como sistemas dinámicos que cambian con el tiempo debido a factores como temporadas, promociones y tendencias del mercado. En este contexto, se podría modelar estas dinámicas de manera más precisa y flexible que los enfoques tradicionales con el uso de este modelo [46].

7. A continuación, se expone el modelo pronostico autorregresivo. Estos modelos de pronostico autorregresivo se utilizan en investigaciones para predecir series temporales. Estos modelos son introducidos por box jenkins y se componen del termino autorregresivo (AR), el termino media móvil (NA) y el nivel de integración (I), para componer el termina ARIMA. Las series de tiempo expresadas por este tipo de modelos se pueden modelar por una función de p con un error.

   De la ecuación 2, se tiene que P es el orden en el cual se ubuca el modelo, Z_t es el valor de la serie de tiempo en un tiempo t y A_t es un error que sigue una distribución normal con promedio 0 en el tiempo [47]. En el pronostico autorregresivo se predice un valor futuro de la serie temporal modelando su comportamiento con base en el pasado de la misma variable o empleando variables exógenas.

$$Z_t = \phi_1 Z_{t-1} + \phi_2 Z_{t-2} + ... + \phi_p Z_{t-p} + A_t \tag{2}$$

8. A continuación se expone el modelo Prophet, el cual es un modelo de series temporales desarrollado por Facebook. El modelo, se utiliza para predecir series temporales con tendencias estacionales. Prophet se basa en un modelo aditivo en el que las tendencias estacionales y no estacionales se modelan por separado y se combinan para hacer predicciones. Prophet se destaca por su capacidad para manejar tendencias no lineales y cambios en la tendencia a lo largo del tiempo, así como por su capacidad para manejar valores faltantes y outliers en los datos de series temporales [48].

9. PyAF es una biblioteca Python de código abierto para pronostico automático, construida sobre módulos de python de ciencia de datos populares, cómo: NumPy, SciPy, Pandas y scikit-learn. PyAF funciona como un proceso automatizado para predecir valores futuros de una señal utilizando un enfoque de aprendizaje automático para realizar la tarea de pronóstico. Ofrece un conjunto de funciones comparable al de algunos productos comerciales de predicción automática. Comienza construyendo un modelo de series temporales basado en valores pasados (proceso de entrenamiento) y luego utiliza este modelo para generar los valores futuros (pronostico). PyAF combina métodos de descomposición, modelos ARIMA y modelos de Fourier para proporcionar un enfoque automatizado y eficiente para el pronóstico de series de tiempo en Python [49].

10. XGBoost, acrónimo de Extreme Gradient Boosting, es una implementación del algoritmo de stochastic gradient boosting. Además de que la librería XGBoost incluye la clase XGBRegressor que sigue la libería de scikit learn y, por lo tanto, es compatible con skforecast. XGBoost Regressor es una técnica de aprendizaje utilizada para problemas de regresión. Está basado en el algoritmo de Gradient Boosting y se caracteriza por su capacidad para manejar problemas, como el forecast multivariado. XGBoost Regressor aborda el forecast multivariado al combinar varios árboles de decisión débiles y mejorarlos de forma iterativa [50].

    Cada árbol se construye secuencialmente, y en cada iteración, se ajusta a los errores residuales del árbol anterior. De esta manera, los árboles posteriores se enfocan en corregir las deficiencias de los árboles anteriores, mejorando gradualmente la precisión del modelo. Una de las principales ventajas de XGBoost Regressor es su capacidad para manejar relaciones no lineales y capturar patrones complejos en los datos. Además, utiliza técnicas de regularización para evitar el sobreajuste y mejorar la generalización del modelo [50].

11. El modelo LightGBM es una biblioteca de aprendizaje automático de código abierto que implementa el algoritmo de Gradient Boosting Decision Tree (GBDT). Este algoritmo es conocido por su capacidad para mejorar gradualmente la precisión de un modelo a través de la combinación de múltiples árboles de decisión débiles. LightGBM se destaca en términos de eficiencia y rendimiento, lo que lo hace especialmente adecuado para aplicaciones de gran escala y conjuntos de datos complejos, cómo lo puede un modelo multivariado. Una de las características clave de LightGBM es el uso de técnicas avanzadas para mejorar su rendimiento [51].

    Por ejemplo, utiliza Gradient-based One-Side Sampling (GOSS) para seleccionar ejemplos de entrenamiento con gradientes significativos, lo que permite un entrenamiento más eficiente y reduce la probabilidad de sobreajuste. Además, utiliza Exclusive Feature Bundling (EFB) para reducir la dimensionalidad de los datos, lo que ayuda a mejorar la velocidad de entrenamiento y la eficiencia del modelo. Otra característica importante de LightGBM es su enfoque en la construcción de árboles de decisión. Utiliza una estrategia llamada leaf-wise splitting, que selecciona la mejor división de un nodo en función de la ganancia máxima en lugar de seguir una estrategia de nivel por nivel. Esto permite una construcción más rápida de los árboles y, en general, mejora la precisión del modelo [51].

    Además, LightGBM utiliza un algoritmo de histograma para calcular los gradientes de

manera más eficiente. En lugar de calcular los gradientes para cada ejemplo individualmente, agrupa los datos en intervalos y calcula los gradientes para cada intervalo. Esto reduce significativamente el costo computacional y mejora el rendimiento general del modelo. Todas las características expuestas anteriormente hacen de LightGBM un modelo para aplicaciones científicas y de investigación [51].

12. El modelo Hist Gradient Boosting Regressor es una técnica de aprendizaje automático basada en árboles de decisión que se utiliza para realizar tareas de regresión. Este modelo utiliza una variante del algoritmo de Gradient Boosting para mejorar su rendimiento. En lugar de trabajar con los datos en su forma original, el modelo utiliza un enfoque llamado histogram binning. El histogram binning es una técnica que divide los datos en intervalos llamados bins y representa cada bin mediante un histograma. En lugar de utilizar cada valor individualmente, el modelo utiliza la información contenida en los histogramas de los bins para realizar las predicciones. Este enfoque permite reducir la complejidad computacional y mejorar la eficiencia del modelo [52].

    El modelo utiliza una estrategia de boosting, que consiste en entrenar una serie de árboles de decisión débiles de forma secuencial. Cada árbol se entrena para corregir los errores cometidos por los árboles anteriores. De esta manera, el modelo aprende gradualmente a mejorar sus predicciones a medida que se agregan más árboles. Además, el modelo utiliza una técnica llamada histogram gradient boosting, que utiliza los gradientes del error de los histogramas para encontrar la mejor dirección para actualizar los valores de los bins. Esto permite que el modelo realice ajustes más precisos y mejore su capacidad de generalización [52].

Modelo	**Aplicación**	**Técnica**	**Importancia**	**Dificultad**
Regresión de Vectores de Soporte (SVR) [6] [7].	Predicción de Series Temporales	Aprendizaje Supervisado	Buena capacidad de generalización en pronósticos.	Moderada
WaveNet [39].	Predicción de Series Temporales	Redes Neuronales Convolucionales	Eficaz en pronósticos de alta resolución.	Alta
DeepAR [40].	Predicción de Series Temporales Multivariadas	Redes Neuronales Recurrentes	Captura relaciones interdependientes entre variables.	Alta
Modelo de Auto regresión Vectorial (VAR) [41].	Modelado de Múltiples Series Temporales	Estadística	Ampliamente utilizado en contextos financieros y económicos.	Moderada

Auto regresión Vectorial Bayesiana de Litterman (LBVAR) [41].	Modelado de Múltiples Series Temporales	Estadística	Captura relaciones completas y realistas con datos macroeconómicos.	Moderada
Temporal Convolutional Network (TCN) [44].	Predicción de Ventas de Múltiples Productos	Redes Neuronales Convolucionales	Capaz de manejar múltiples productos y capturar correlaciones.	Alta
Neural ODE [46].	Modelado de Series Temporales con Dinámicas Complejas	Ecuaciones Diferenciales Ordinarias	Modela dinámicas complejas y no lineales con flexibilidad.	Alta
Prophet [48]	Predicción de Series Temporales con Tendencias Estacionales	Modelo Aditivo	Maneja tendencias no lineales y cambios en la tendencia.	Moderada
PyAF [49]	Predicción de Series Temporales	Aprendizaje Automático, descomposición, ARIMA, Fourier	Enfoque automatizado y eficiente para pronóstico de series de tiempo	Moderada
XGBoost [50].	Predicción de Series Temporales	Gradient Boosting	Problemas multivariados, relaciones no lineales y patrones complejos	Alta
LightGBM [51]	Predicción de Series Temporales	Gradient Boosting Decision Tree	Eficiencia y rendimiento, técnicas avanzadas para mejorar el rendimiento	Alta
Hist Gradient Boosting Regressor [52].	Predicción de Series Temporales	Decision Tree	Histogram binning, gradient boosting	Alta

Tabla 1: Comparación de Modelos para Series Temporales

2.6 Predicción de ventas aplicando machine learning

Como se mencionó en la sección 2.4, la predicción de ventas, es una tarea crítica en cualquier negocio, ya que permite a las empresas planificar y tomar decisiones estratégicas en función de la proyección de ingresos futuros. En este contexto, el uso de técnicas de machine learning puede ayudar a mejorar la precisión de las predicciones de ventas y proporcionar *insights* valiosos para la toma de decisiones [53].

Debido a esto, el uso de machine learning, es una parte crucial de la inteligencia empresarial contemporánea, ya que puede presentar dificultades considerables, sobre todo en casos de falta de datos, valores perdidos o presencia de *outliers*. Dado que las ventas suelen ser interpretadas como series temporales, se han creado diversos modelos para el análisis de dichas series, tales como *Holt-Winters*, *ARIMA*, *SARIMA*, *SARIMAX*, *GARCH*, entre otros. Es importante mencionar que estos modelos pueden resultar complejos y requieren un enfoque riguroso y detallado para lograr una adecuada implementación [36].

En un entorno de negocio a negocio (*B2B*), prever la demanda futura es esencial para la planificación de la producción y el suministro. La tecnología de aprendizaje automático ofrece oportunidades para mejorar estas previsiones, esto permite una priorización más efectiva de los recursos y un aumento en las ventas potenciales [54]. No obstante, en el contexto de cadenas de suministro extendidas, donde la demanda puede sufrir distorsiones debido a la falta de integración y al efecto látigo, la precisión de los pronósticos sigue siendo un desafío. Aquí es donde entran en juego los modelos de aprendizaje automático [55].

Estas potentes técnicas, pueden analizar datos complejos y no estructurados, como las solicitudes de cotización (*RFQs*), para predecir la probabilidad de que una *RFQ* se convierta en una venta efectiva. Al abordar la distorsión de la señal de demanda, estos modelos ofrecen la posibilidad de optimizar la asignación de recursos y mejorar la toma de decisiones en un contexto *B2B* [55]. En la figura 4, se puede observar un ejemplo del comportamiento de las predicciones de ventas. Estas predicciones se basan en datos históricos y utilizan un enfoque autorregresivo, donde se pronostica el valor futuro de una serie temporal en función de su comportamiento previo [56].

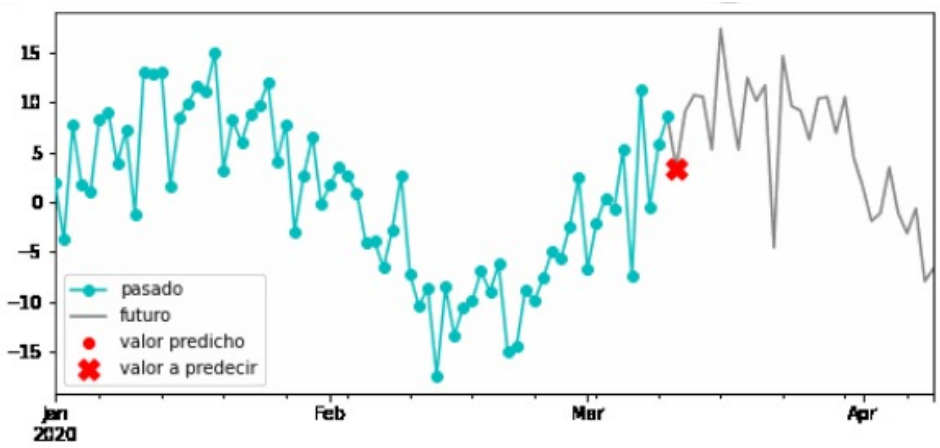


Figura 4: Predicción utilizando series temporales [56].

En la figura 5, se presenta visualmente la implementación de variables exógenas en la serie

temporal, utilizadas para llevar a cabo el proceso de predicción de ventas [56]. La inclusión de variables exógenas en este contexto es de gran relevancia, ya que permite enriquecer el análisis al considerar factores externos que podrían influir en las ventas. Estas variables exógenas, también conocidas como covariables, son aquellas que no son inherentes a la serie de tiempo que se está pronosticando, pero que se cree que tienen un impacto en su comportamiento [53].

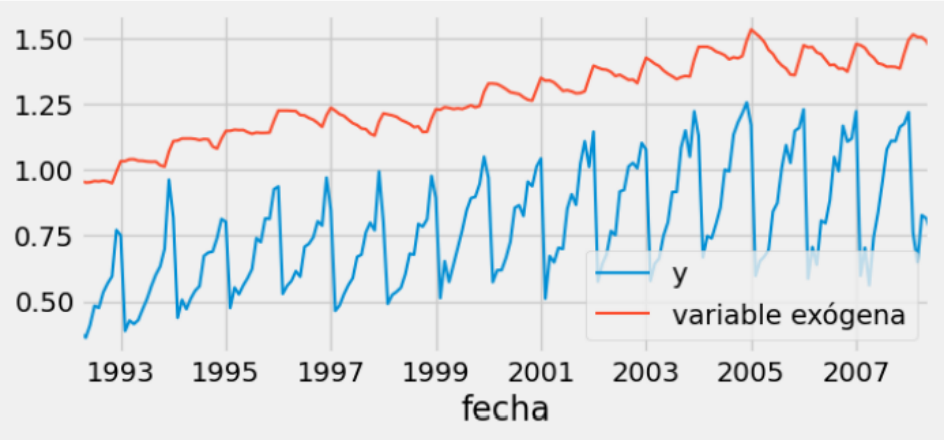


Figura 5: Predicción utilizando de series temporales con variable exógena [56].

2.6.1. Variables exógenas

Las variables exógenas, son aquellas que encarnan factores que yacen más allá del ámbito de influencia o manipulación directa por parte de la empresa o del ejecutor del análisis. Por el contrario, las variables endógenas abarcan elementos que se gestan desde el núcleo de la organización y, por ende, quedan susceptibles a una intervención activa [57].

Estas variables exógenas, aunque se sitúan fuera del control directo de la empresa o del análisis, juegan un rol esencial en el panorama de decisiones estratégicas. Actúan como factores externos que pueden tener un impacto significativo en la operación y el rendimiento de la empresa. Al considerar su influencia, se logra una comprensión más completa de los desafíos y oportunidades que se presentan en el entorno empresarial [57].

En el contexto de la medición de la eficiencia empresarial, las variables exógenas son factores externos a la empresa que influyen en su desempeño operativo y productivo, pero que están fuera del control directo de la dirección. Estas variables son determinantes clave de las condiciones en las que opera la empresa y que pueden afectar su nivel de producción y consumo de insumos [58].

Se sugiere que, al considerar la eficiencia de una empresa, es fundamental reconocer la influencia de estas variables exógenas para evitar atribuir erróneamente ineficiencias observadas a la gestión interna de la empresa. Ejemplos de variables exógenas podrían incluir factores económicos, condiciones medioambientales, regulaciones gubernamentales, condiciones del mercado, disponibilidad de recursos naturales, entre otros [58].

La forma en que se incorporan y abordan estas variables en la medición de la eficiencia puede variar según el enfoque metodológico utilizado, ya sea integrándolas directamente en la estimación de la función frontera o a través de un análisis de regresión en una segunda etapa. Estos enfoques permiten comprender cómo el entorno influye en el desempeño de la empresa

y cómo separar los efectos de las variables exógenas de la eficiencia productiva interna [59].

2.6.2. Variables internas

El análisis de variables internas en la predicción de ventas es crucial. Al considerar factores dentro de la empresa, se establece una base sólida para pronósticos precisos. Esto permite comprender patrones de ventas que pueden no depender de factores externos. Como resultado, se toman decisiones estratégicas más informadas, desde inventario hasta precios y fuerza de ventas [60]. Las variables internas de una empresa desempeñan un papel crucial en el proceso de pronóstico [61].

Estas variables, que pueden incluir datos relacionados con la producción, inventario, estrategias de marketing y desempeño histórico de ventas, ofrecen información invaluable para comprender las tendencias y patrones que afectan las ventas futuras. Al integrar estas variables internas en el marco de modelado, se logra mejorar la precisión de los pronósticos de ventas, realizando un análisis detallado de estas variables internas e integrando su impacto en el proceso de predicción de ventas para la toma de decisiones informadas en la planificación, la producción y la estrategia empresarial [61].

Algunos modelos se distinguen al considerar exclusivamente variables internas observables de la organización. Esta particularidad simplifica su implementación en empresas similares. El acceso a datos internos y la investigación que destaca la influencia de factores organizativos respaldan este enfoque. Además, el modelo puede explicar de manera efectiva las ventas [59]. Esto fortalece las técnicas de pronóstico y mejora la toma de decisiones en áreas clave como precios, diversificación de clientes y planificación de producción, proporcionando ventajas competitivas basadas en las características internas de la empresa [60]. La tabla 2, presenta variables internas relacionadas con la predicción de ventas, las cuales fueron obtenidas de la empresa de tecnología piloto.

Variable Interna	**Definición**	**Ejemplo**
Quantity_sell_in	Cantidad de productos vendidos	28000 Unidades de producto vendidos
Sales_sell_in	Ventas de productos y servicios	3000000 de Euros vendidos
Credit_limit	Cupo de credito de los clientes	20000 Euros de cupo
Sales_Oppt_win	Monto de oportunidades ganadas	3000000 Euros ganados por un servicio
Quantity_Oppt_win	Cantidad de oportunidades ganadas	40000 Servicios completados
Sales_Oppt_lost	Monto de oportunidades perdidas	1200000 Euros perdidos por un servicio
Quantity_Oppt_lost	Cantidad de oportunidades perdidas	22000 Servicios no completados

Tabla 2: Variables internas relacionadas con la predicción de ventas obtenidas de la empresa de tecnología piloto.

2.6.3. Variables externas

En el campo de la predicción de ventas, como un proceso de gran importancia en la gestión empresarial, se ve profundamente influenciado por diversas variables externas. Estas variables, se encuentran más allá del alcance de control directo de la empresa, ejercen una influencia directa en los patrones de demanda de los consumidores. Al incorporar estas variables externas en los modelos de previsión de ventas, las empresas pueden obtener una comprensión más precisa y holística de las tendencias de demanda. Esto a su vez les permite tomar decisiones informadas y estratégicas para la asignación de recursos y la planificación operativa [62].

La influencia de factores externos en el rendimiento empresarial ha adquirido un papel crucial en el análisis contemporáneo de la gestión empresarial. Según Harvard Business Review, aproximadamente un 85 % del desempeño de una empresa está estrechamente relacionado con variables externas [63]. Sin embargo, a pesar de la abundancia de conjuntos de datos disponibles, muchas organizaciones enfrentan desafíos para identificar fuentes relevantes que permitan discernir con precisión qué elementos externos impactan su desempeño [62].

Estos indicadores ofrecen una visión integral de la economía y pueden proporcionar información valiosa para prever cambios en la demanda futura [64]. Con el avance tecnológico y la mayor accesibilidad a datos macroeconómicos están permitiendo una transición hacia enfoques más cuantitativos y basados en datos, lo que permite incorporar estos indicadores en los modelos de pronóstico de ventas, mejorando así la precisión y la capacidad de anticipar cambios en el entorno empresarial [65] [64]. Un ejemplo, es la actividad económica, la cual se menciona como un índice para el pronostico del consumo de energía. Por otra parte, otros ejemplos como el PIB, el cual es uno de los indicadores de mayor representación en términos de actividad económica de un país, sin embargo la actividad económica también se puede medir a través de otros sectores como el de hidrocarburos, agricultura y construcción [66].

En este contexto, para la selección de variables que tengan incidencia en el pronóstico de ventas, se explica que la demanda de los productos de una compañía suele ser influenciada por la demanda de los bienes y servicios de la economía. La actividad económica de un país es medida a través del Producto Interno Bruto (PIB) por lo que esta variable es crucial para los pronósticos [67]. Aunque cada negocio, industria y región responde de manera única a diversos factores, es pertinente destacar algunos elementos recurrentes que se han encontrado en la literatura. Algunas de las variables externas más significativas en la literatura se exponen en la tabla 3.

Variable externa	**Definición**	**Ejemplo**
Mes [61]	El mes en el que se registra la actividad.	Mes de enero
Año [61]	El año en el que se registra la actividad.	Año 2023
Tasas de interés [68]	El costo del dinero prestado o ganado por inversiones.	Porcentaje que refleja el costo de tomar prestado dinero o el rendimiento ganado al invertirlo.
Tipo de cambio de divisas [63]	Valor relativo entre diferentes monedas.	1 pem es igual a 0.26 USD

Precio de la gasolina [63]	Costo por unidad de volumen de gasolina.	Costo que los consumidores pagan por un litro o galón de combustible en una estación de servicio.
PIB [69]	Valor total de bienes y servicios producidos en un país.	7196 USD para el 01/12/2022
IPC [69]	Medida de la inflación y cambio en los precios.	7 % Anual para Perú.
Ingresos tributarios [70]	Ingresos recaudados por el gobierno a través de impuestos.	Recaudo en los ingresos tributarios que afecta las finanzas públicas y el gasto.
Tendencias económicas globales [71]	Patrones de crecimiento económico a nivel mundial.	Recesión global que afecta la demanda de bienes de lujo.
Competencia en el mercado [71]	Otras empresas que ofrecen productos similares.	Situación en la que múltiples empresas compiten para ofrecer servicios a los consumidores.
Tarifas energéticas residenciales [71]	Precio por unidad de consumo de energía eléctrica en hogares.	Costos por el consumo de energía eléctrica en los hogares.
Laboralidad [66] [65]	El calendario laboral tiene una incidencia en el consumo de energía eléctrica para los días con mayor actividad laboral.	Indica la cantidad de días de trabajo por mes.

Tabla 3: Variables externas encontradas en la literatura relacionadas con la predicción de ventas

2.7 Visualización de datos en predicción de ventas

La visualización de datos es un ámbito crucial en la actualidad debido a la gran cantidad de información que las organizaciones generan y que está disponible en la web. Esta técnica nos permite representar datos de manera gráfica y comprensible, lo que facilita a los tomadores de decisiones entender patrones, identificar información relevante y formarse opiniones. Al convertir datos complejos en gráficos visuales, se mejora la interpretación y se agiliza el proceso de comprensión, lo que resulta especialmente útil en campos científicos. La visualización de datos es una herramienta poderosa para convertir información en conocimiento y mejorar la comunicación de ideas en una sociedad cada vez más orientada por los datos [72].

La visualización de datos desempeña un papel crucial en la predicción de ventas. A medida que las empresas lidian con grandes cantidades de datos de diversas fuentes, la capacidad de comprender y comunicar patrones, tendencias y relaciones dentro de estos datos se vuelve esencial. La visualización de datos permite transformar datos complejos en representaciones visuales claras y comprensibles, lo que facilita la identificación de *insights* y la toma de

decisiones informadas [73].

En este contexto, la visualización de datos mediante Excel emerge como un aspecto útil en el proceso de visualización de predicción de ventas. Con la capacidad de convertir números en gráficos comprensibles, Excel se convierte en una herramienta para interpretar tendencias futuras. Una vez que las estimaciones de ventas han sido obtenidas, Excel simplifica la creación de gráficos de líneas, de barras y otros tipos, permitiendo una representación visual vívida de las proyecciones. La personalización de colores, etiquetas y estilos garantiza una adaptación precisa a los requerimientos del análisis [74].

Otra herramienta de visualización de predicción de ventas es Python, en el cual destaca la importancia de la visualización de los datos como una herramienta útil. A través Python y diversas bibliotecas, se logra transformar los datos en representaciones gráficas esclarecedoras. Esta visualización provee una comprensión inmediata y reveladora de los datos de previsión de ventas, facilitando la identificación de patrones y tendencias relevantes [75].

En el ámbito de análisis de ventas, Tableau emerge como una poderosa herramienta para la visualización de datos. Su interfaz intuitiva y amigable permite representar de manera efectiva las proyecciones de ventas. Mediante gráficos interactivos, gráficos de barras y líneas, mapas de calor y dashboards personalizados, Tableau posibilita transformar datos complejos en representaciones visuales claras y comprensibles. Por ejemplo, en el contexto de previsión de ventas, Tableau facilita la identificación de patrones y tendencias en los datos históricos, permitiendo a los usuarios explorar las proyecciones futuras con un enfoque centrado en la visualización [76].

Por otra parte, Power BI emerge como una herramienta eficaz para la visualización de datos. Mediante la creación de un panel de control específico para predicción de ventas, Power BI permite explorar de manera interactiva y visual el rendimiento de ventas. La herramienta posibilita la construcción de gráficos dinámicos, tablas y visualizaciones interactivas que representan las proyecciones de ventas de manera clara y comprensible. Estas visualizaciones pueden incluir gráficos de tendencias, análisis comparativos entre períodos, mapas geográficos de ventas, y otros elementos que proporcionan una visión rápida y efectiva de las previsiones de ventas [77].

Por otra parte, el uso de aplicaciones web de pronóstico de ventas se ofrece una herramienta esencial para las Pequeñas y Medianas Empresas de Confección. Estas aplicación permiten pronósticos de ventas precisos y oportunos. Esto evita desafíos como la sobreproducción o la falta de stock, permitiendo a las empresas ajustar la producción y optimizar los recursos. La accesibilidad en línea garantiza que los usuarios puedan acceder a pronósticos actualizados en cualquier momento y lugar, facilitando decisiones ágiles y una gestión eficiente [78].

Una aplicación desarrollada es *eTIFIS*, la cual permite a las empresas realizar pronósticos precisos de series temporales en línea. Los usuarios pueden elegir entre diversas técnicas de pronóstico, como el Modelo *Theta*. Al aprovechar la arquitectura web y los estándares abiertos de Internet, la aplicación facilita la integración de procesos empresariales. Sus beneficios incluyen acceso en línea desde cualquier ubicación, variedad de técnicas de pronóstico, mejor precisión en los pronósticos, una interfaz de usuario intuitiva, capacidad de integración con otras aplicaciones y potencial reducción de costos en comparación con soluciones especializadas. De la misma manera la personalización de la aplicación con base en las necesidades y requerimientos que el cliente desee agregar al la aplicación [79].

En un estudio, realizado por Dalrymple en 1987, se encuestaron un total de 860 compañías norteamericanas dirigidas principalmente a los gerentes de marketing y de pronósticos para que estas fueran contestadas por personas con conocimiento en proyección de ventas, de estas encuestas se observó que solo el 19 % de los que respondieron suelen utilizar mas de un solo método de predicción y que se plantean diferentes escenarios para cada pronóstico, adicionalmente a esto, a pesar de que según la literatura, siempre es mejor proporcionar intervalos de confianza al momento de hacer un pronóstico, el 77 % de las empresas que respondieron afirman no hacer uso de estos [80].

Para finalizar, el 40 % de las encuestas que respondieron afirman utilizar diferentes tipos de encuestas tanto internas de la compañía como externas para poder apoyar le proceso de pronóstico de ventas, el 39 % utilizan métodos estadísticos como media móvil y tasa porcentual de cambio y solo el 21 % de estas empresas utiliza análisis de regresión múltiple y modelos econométricos (cabe recalcar que estos porcentajes no suman el 100 % debido a que una empresa puede utilizar varios enfoques), a partir de esto se puede observar que el porcentaje de empresas que utilizan modelos estadísticos para el pronóstico de ventas es bajo.

En conclusión, se mencionan varios modelos de pronóstico utilizados en diferentes estudios. Algunos de estos modelos incluyen regresión polinómica, k means junto con árboles de decisión, regresión lineal múltiple, modelos de series temporales como ARIMA, redes neuronales convolucionales (CNN) y redes neuronales recurrentes (LSTM), entre otros. Comparativamente, el modelo Prophet se destaca por su capacidad para manejar tendencias no lineales y cambios en la tendencia a lo largo del tiempo en series temporales con tendencias estacionales. A diferencia de algunos modelos mencionados, Prophet está diseñado específicamente para trabajar con series temporales y capturar patrones estacionales de manera eficiente.

Por ejemplo, en el caso de la regresión polinómica o la regresión lineal múltiple, Prophet puede ser más adecuado para capturar cambios no lineales y tendencias estacionales más complejas sin necesidad de que un experto ajuste manualmente el modelo. Respecto a modelos basados en redes neuronales como CNN y LSTM, si bien pueden tener capacidades poderosas de aprendizaje, a veces pueden requerir más tiempo de entrenamiento y configuración adecuada, lo cual puede no ser óptimo para aplicaciones que necesitan predicciones rápidas y precisas.

En resumen, Prophet se diferencia por su capacidad para manejar automáticamente tendencias no lineales y estacionales en comparación con modelos tradicionales como regresión polinómica o regresión lineal múltiple, y puede ofrecer una alternativa más rápida y precisa en comparación con modelos basados en redes neuronales en ciertos contextos de predicción de series temporales. Esta comparación resalta las fortalezas y diferencias específicas entre Prophet y otros modelos mencionados en la literatura, destacando así la relevancia y contribución del modelo implementado en la tesis.

Capítulo 3

DESARROLLO DE LA METODOLOGÍA

En este capítulo se presenta de manera sistemática la metodología que sustenta el desarrollo de este trabajo de grado de investigación. En un principio, se lleva a cabo una exhaustiva investigación y selección de datos del mercado con el propósito de identificar fuentes relevantes que proporcionen información crucial sobre el mercado. A continuación, se aborda la tarea de identificación de las variables internas relevantes, resaltando aquellas que ejercen un impacto significativo en las ventas. Subsecuentemente, se identifican las variables externas y se explora el uso del *Web Scraping* como método para obtener datos adicionales y complementarios.

Posteriormente, se realiza el proceso de preprocesamiento y limpieza de datos a través de un análisis exploratorio de datos, el cual es un paso de gran importancia, encaminado a preparar los datos para el análisis subsiguiente. A continuación, se emprende la etapa de selección de los modelos de machine learning, durante la cual se exploran y eligen los algoritmos más apropiados para la tarea de predicción de ventas. Después de ello, se procede con la evaluación, entrenamiento y refinamiento de los modelos. Este paso aborda con detalle el proceso de ajuste y optimización de los modelos. A continuación, se lleva a cabo el análisis y presentación de resultados de los modelos, donde se expone cómo se interpretan y comparan los resultados obtenidos.

Finalmente, se centra en el diseño y desarrollo de la aplicación web para la visualización de los datos obtenidos de los modelos seleccionados para la predicción de ventas. En esta etapa, se describe la implementación de la interfaz de usuario, los frameworks y librerías utilizadas para el desarrollo de la aplicación y como se realizara el despliegue de esta para que los usuarios tengan acceso.

Es importante mencionar que para el desarrollo del trabajo de grado, se hizo uso de una metodología ágil llamada *Scrum*. Esta es una metodología ágil popular que se enfoca en el trabajo en equipo, la colaboración y la entrega de resultados iterativos y continuos, para producir un producto final de la mejor calidad [81]. Para esta metodología, se entregan partes funcionales del producto de manera constante y no se espera hasta que todo el producto esté completo. Esto permite tener un producto probado y aprobado por el cliente en la fase final del proyecto. En el desarrollo del proyecto de grado, se tendrá retroalimentación del director del trabajo de grado y la empresa de tecnología con la cual se realizará la prueba de los algoritmos. En *Scrum*, se trabaja en ciclos de tiempo fijo, llamados *sprints*, en los que el equipo trabaja

en un conjunto de tareas específicas para producir un resultado utilizable y potencialmente entregable. Los *sprints* son generalmente de dos a cuatro semanas de duración, aunque la duración exacta puede variar según las necesidades del equipo y del proyecto.

Durante cada *sprint*, se selecciona un conjunto de tareas a realizar y distribuyen dentro de los miembros del equipo. Al final de cada *sprint*, se revisa los resultados y se realiza una demostración del trabajo completado. Los *sprints* permiten enfocarse en un conjunto específico de tareas durante un período de tiempo definido y producir resultados iterativos y continuos.

De esta manera, se permite al director del trabajo de grado y a la empresa de tecnología proporcionar retroalimentación constante. Esto, a su vez, contribuye a mejorar el producto y garantizar que este cumpla con las necesidades y expectativas del proyecto de grado. Para aplicar *Scrum* en el trabajo de grado, se exponen las historias de usuario, con el fin de definir las diferentes tareas y entregables para el desarrollo del trabajo de grado. A continuación se expone en la figura 6, el diagrama de flujo del desarrollo de cada *sprint* y las historias de usuario dentro de cada *sprint*.

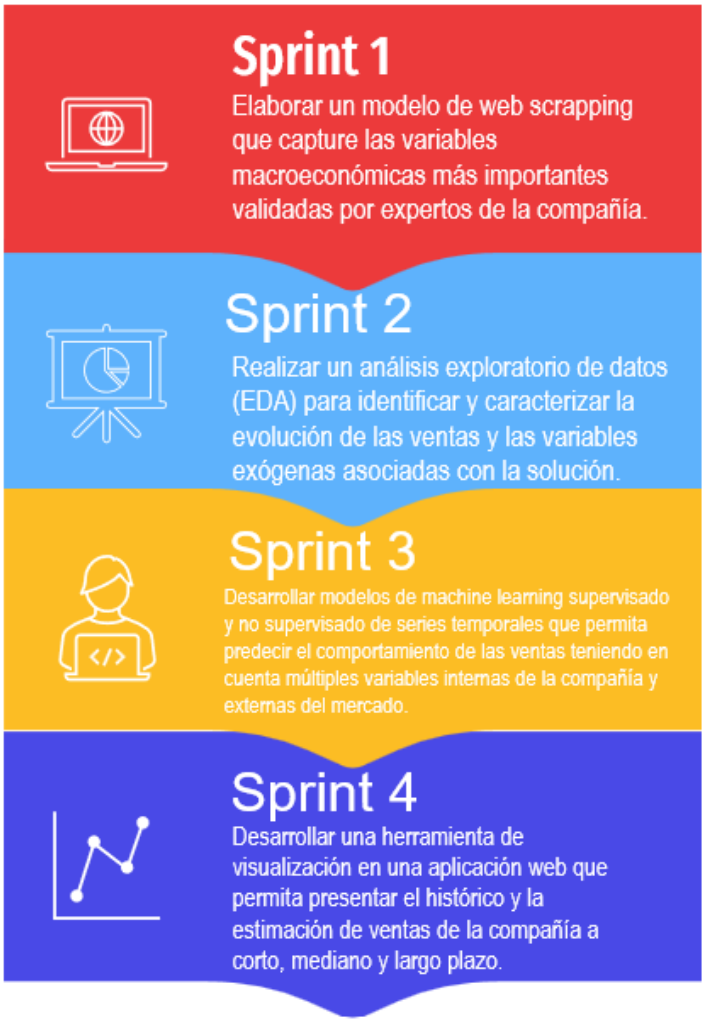


Figura 6: Diagrama de los Sprints del desarrollo del proyecto.

Una vez definidas las historias de usuario, se trabaja en cada una de ellas durante cada *sprint* y se llevan a cabo reuniones semanales de *Scrum* para revisar el progreso y hacer ajustes si es necesario. Cada *sprint* tiene una duración de 2 a 4 semanas, dependiendo de la complejidad de las tareas involucradas. Al final de cada se presenta los resultados tangibles y funcionales que permitan avanzar en el proyecto de manera iterativa y progresiva. Además, cada *sprint* es revisado por el equipo para identificar oportunidades de mejora y ajustar el plan de trabajo para el siguiente *sprint*. *Scrum* al ser una metodología ágil, es flexible y adaptable, por lo que se ajusta según las necesidades y la naturaleza del trabajo de grado.

3.1 Investigación y selección de los datos

En esta sección, se abordará el proceso de investigación y selección de datos, que comprende la selección de variables internas y la selección de variables externas. La importancia de estos procedimientos radica en optimizar los modelos de machine learning utilizados para la predicción de ventas. Esto implica la identificación y utilización de las variables más relevantes y significativas, tanto internas como externas, con el fin de mejorar el rendimiento y la precisión de dichos modelos. A continuación, se detallarán los métodos y enfoques empleados en la elección de estas variables, con el objetivo de lograr resultados más precisos en la predicción de ventas.

Para la selección de las variables internas se abordará la tarea de identificar y elegir las variables internas mas relevantes, esto se realizo por medio de entrevistas a la empresa de tecnología. Un grupo de interesados de la empresa de tecnología, provenientes de diferentes empresas, identificaron que se compartía esta misma problemática en sus anteriores experiencias laborales. A partir de esto, se planteó que una solución de pronóstico de ventas podría ser la solución.

Se agendan entrevistas cortas (veinte minutos) a los Presidentes de Colombia, Ecuador, Venezuela, Perú y Bolivia, con el fin de entender el proceso actual. Estas entrevistas permiten comprender que en el proceso de pronóstico, los presidentes consolidan la información de los once gerentes de ventas, que se dividen por unidad de negocio o tipo de cliente, reciben esta información en formato Excel y deben comenzar a unificar la información. Es un proceso muy manual y que consume muchas horas de trabajo de los presidentes.

Antes de enviar la información a los presidentes, los gerentes de ventas de cada país reciben la información de cada uno de los integrantes de sus equipos, a estos los llaman comerciales (80 en total), y recopilan cuanto espera vender cada uno, juntan la información de su equipo sumando las ambiciones de cada comercial y con esto pueden decir cuanto será la venta para la unidad de negocio o tipo de cliente para el que trabajan. Se nota como este proceso consume el tiempo de mucho niveles de la organización por falta de optimización y digitalización.

Finalmente se entrevistaron grupos de comerciales, donde las preguntas eran relacionadas a cómo planifican sus ventas, qué herramientas usan para este pronóstico, cómo les gustaría que una herramienta los apoyara en este proceso y qué variables internas y externas usan para realizar este proceso de pronóstico, las entrevistas duraron una hora. Se halla que los comerciales pronostican sus ventas por mes, trimestre y año. Cada comercial tiene en cuenta indicadores internos cómo el crédito financiero del cliente, las oportunidades ganadas y perdidas, y las ventas de su cliente (llamado Sell Out). Por medio de este análisis se determinan cuales son las variables internas que se seleccionarán. Finalmente, con la ayuda de los presidentes de país, gerentes de ventas y comerciales, se seleccionan cuales son esas variables

económicas que inciden en la dinámica de ventas, es por medio de las indicaciones que da la empresa de tecnología que se seleccionan 20 variables que pueden contribuir a un mejor pronóstico de las ventas.

La solución propuesta para la problemática es una herramienta web, que permita visualizar las ventas futuras a corto, mediano y largo plazo, aplicando Machine Learning y Web Scraping para la selección de las variables externas. Con esta solución se espera reducir las 11.160 horas que se invierten actualmente en la planificación de vetas y aumentar el forecast accuracy (hasta en un +20 %). En la Figura 7, se describe con detalles como se realizan las entrevistas:

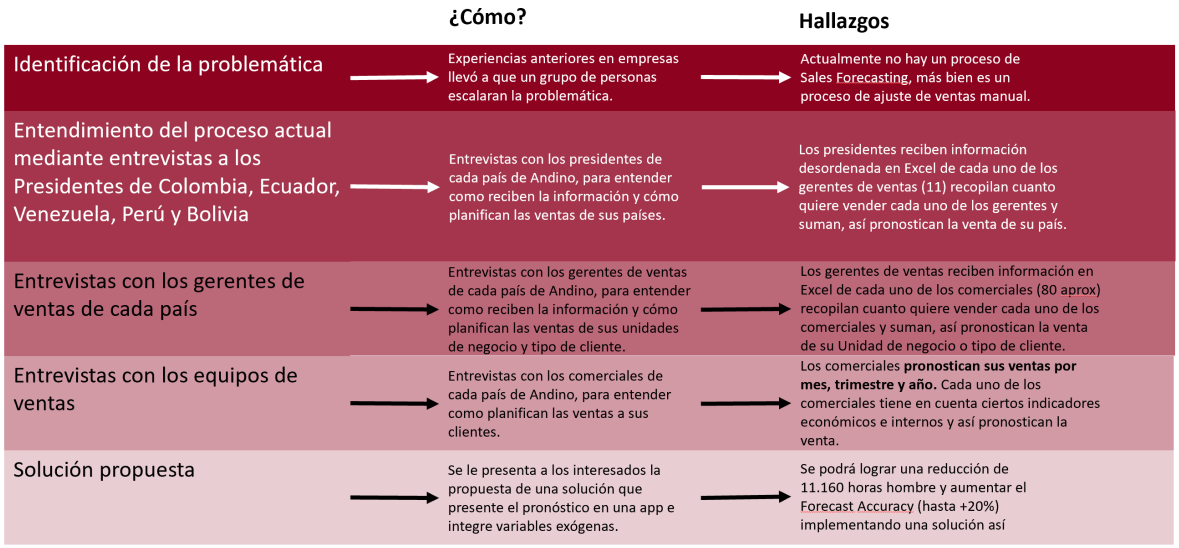


Figura 7: Proceso de realización de las entrevistas.

Todas las variables seleccionadas son provenientes de las bases de datos disponibles. Estas bases de datos incluyen información sobre ventas por oportunidades y proyectos, así como ingresos totales, junto con los límites de crédito financieros otorgados a los clientes. A continuación, se describe cada una de las variables seleccionadas por base de datos.

- **Base de datos de oportunidades:** Contiene información correspondiente a las ventas por oportunidades / proyectos realizados en Perú, por cada comercial, cliente, tipo de cliente, unidad de negocio.

  1. **Año**: Contiene el año en el que se cerró la oportunidad / proyecto. Es una variable cuantitativa discreta. Su unidad es años.
  2. **M**es: Contiene el mes en el que se cerró la oportunidad / proyecto. Es una variable cuantitativa discreta. Su unidad es meses.
  3. ***Customer***: Indica el cliente con el cual se cerró la oportunidad. (Alcantarillado de Lima, Petróleos del Perú, Manelsa, etc.). Es una variable cualitativa categórica.
  4. ***Golden ID***: Es el código con el cual se relaciona el cliente con las demás bases de datos. Es una variable cualitativa categórica.
  5. ***Customer category***: Es el tipo de cliente con el cual se realizó el proyecto. Podemos tener usuarios finales, instaladores, fabricante de equipos, etc. Es una variable cualitativa categórica.

6. **BU (*Business Unit*)**, es la unidad de negocio que está realizando el proyecto. Es una división interna de Schneider Electric e indica que área / unidad de negocio realizó el proyecto en conjunto con el cliente. Es una variable cualitativa categórica.
7. **Sales_Oppt:** Indica los ingresos producidos con esta oportunidad / proyecto cerrado. Es una variable cuantitativa discreta. Su unidad es Euros.
8. **Quantity_Oppt:** Indica la cantidad de oportunidades de negocio / proyectos entregados cuantitativa discreta. Su unidad es unidades de proyectos vendidos.

- **Base de datos de Sell in:** Al igual que la base de datos de oportunidades, contiene los ingresos por cliente, tipo de cliente, unidad de negocio por año y por mes; con la diferencia que las ventas que se tienen acá son las ventas totales medidas en Euros y es una variable cuantitativa discreta, es decir la venta por productos transaccionales y de proyectos.

- **Base de datos de crédito financiero:** Esta base de datos contiene la información correspondiente al cupo de crédito que tienen los clientes para comprar en la compañía, la unidad de medida es el Euro y es una variable cuantitativa discreta. Es un dato esencial ya que indica el límite de compra de cada uno de los clientes.

  1. **Año**: Contiene el año en el que se le otorga el monto de cupo de crédito. Es una variable cuantitativa discreta. Su unidad es años.
  2. **Mes**: Contiene el mes en el que se le otorga el monto de cupo de crédito. Es una variable cuantitativa discreta. Su unidad es meses.
  3. ***Customer***: Indica el cliente con el cual se cerró la oportunidad. (Alcantarillado de Lima, Petróleos del Perú, Manelsa, etc.). Es una variable cualitativa categórica.
  4. ***Amount***: Indica el monto máximo que tienen los clientes para realizar compras con la empresa. Es una variable cuantitativa discreta. Su unidad de medida es el Euro.

- **Base de datos de Indicadores Macroeconómicos:** Esta base de datos se consolidan todos los indicadores económicos de importancia, previamente seleccionados por la empresa de tecnología, organizados por año y mes. Al ser indicadores, son valores adimensionales.

  1. **Año**: Contiene el año del indicador a analizar. Es una variable cuantitativa discreta. Su unidad es años.
  2. **Mes**: Contiene el mes del indicador a analizar. Es una variable cuantitativa discreta. Su unidad es meses.
  3. ***Indicador***: Valor del indicador económico para este año y mes.

Con este proceso de selección de variables internas se busca comprender y cuantificar cómo factores como el tipo de cliente, la unidad de negocio y la relación histórica con los clientes impactan en las ventas y los ingresos generados, lo cual permite generar un modelo efectivo que refleje con precisión el comportamiento y las tendencias de las ventas.

Para la selección de las variables externas se aborda el proceso de selección de variables externas. Así como existen variables internas de la misma compañía que explican el comportamiento de las ventas, existen variables externas que cumplan este mismo propósito. Para la

selección de variables externas se toma como punto de partida la revisión de literatura que se expone en el capítulo 2, en la sección 2.6, en la subsección 2.6.3.

A partir de investigaciones que se realizaron previamente se pueden identificar variables de alto impacto en la variable de interés, las ventas, es por esto por lo que a través del proceso de revisión de literatura se identificaron que variables externas son candidatas para utilizar en el modelo de pronóstico.

La identificación de estas variables permite generar un modelo robusto y con una mayor explicación del comportamiento de las ventas en el futuro. Finalmente, una vez identificadas y seleccionadas las variables externas de interés, y ya que, dentro de los objetivos del proyecto, es necesario construir un algoritmo que permita realizar la extracción de las variables. Todas estas variables se encuentran en dos páginas web, una es la correspondiente a Trading Econmics y las otras del Banco Central de Reserva, desde acá mediante un código de web scraping se extrae la información lo más actualizada posible, en su mayoría, mensual.

La figura 8 describe en el proceso de extracción y selección de las variables exógenas. El proceso de extracción de variables externas, identifica las variables a predecir y las variables que pueden influir el comportamiento de las ventas. Luego, las variables internas y externas son seleccionada de diversas fuentes. Finalmente, se limpian y combinan las variables internas y externas en un único conjunto de datos. En el caso específico de las entrevistas, el proceso se lleva a cabo en Bogotá, Colombia. Las entrevistas con CEOs, Gerentes y comerciales se realizan para obtener información adicional sobre las variables internas y externas. En resumen, el proceso consiste en identificar, seleccionar, extraer, limpiar y combinar las variables internas y externas.

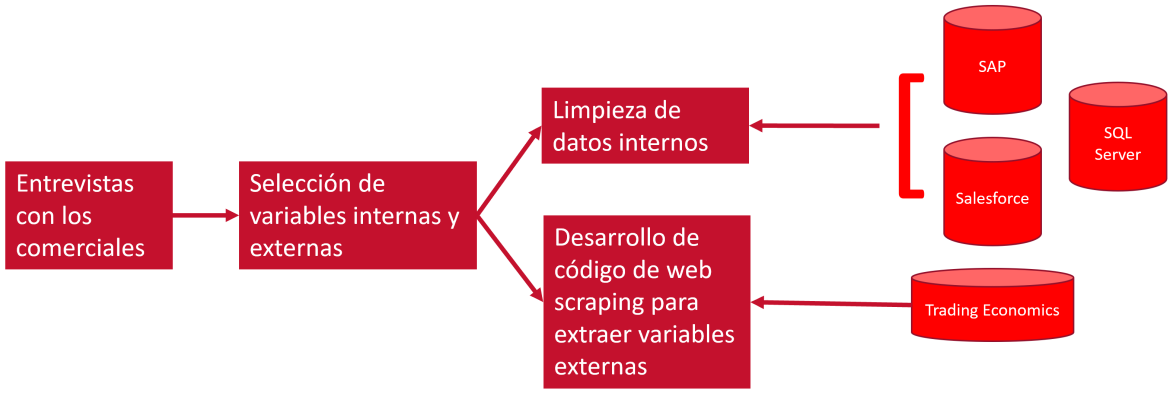


Figura 8: Diagrama de flujo - Proceso de extracción de las variables externas.

El Preprocesamiento, se realizo con la finalidad de seleccionar aquellas variables que servirán para alimentar y robustecer el modelo, se procede a buscar los accesos a aquellas bases de datos que servirán para extraer los data frames. Las Bases de datos se pueden visualizar en la figura 9, donde se han seleccionado cuatro. Una es Salesforce donde se encuentra la base de datos de Oportunidades, la siguiente es SAP donde se encontrará esa información que se carga de manera automática correspondiente a las ordenes de compra de los clientes de la empresa de tecnología, esto es lo que determinarán las ventas de la empresa, a este data frame se le llama Sell in, de SAP también proviene el data frame de crédito financiero el cual viene por cliente y es configurado por el equipo de finanzas el cual determina el cupo financiero que

tendrá cada cliente para comprar a la empresa, el siguiente es un SQL Server el cual contiene la información de las ventas realizadas por el cliente, a este data frame se le llama Sell out y contiene la información de las ventas de los clientes, finalmente se consolida, con el código de web scraping, y capturando la información de Trading Economics, las variables económicas que se seleccionaron en las entrevistas.

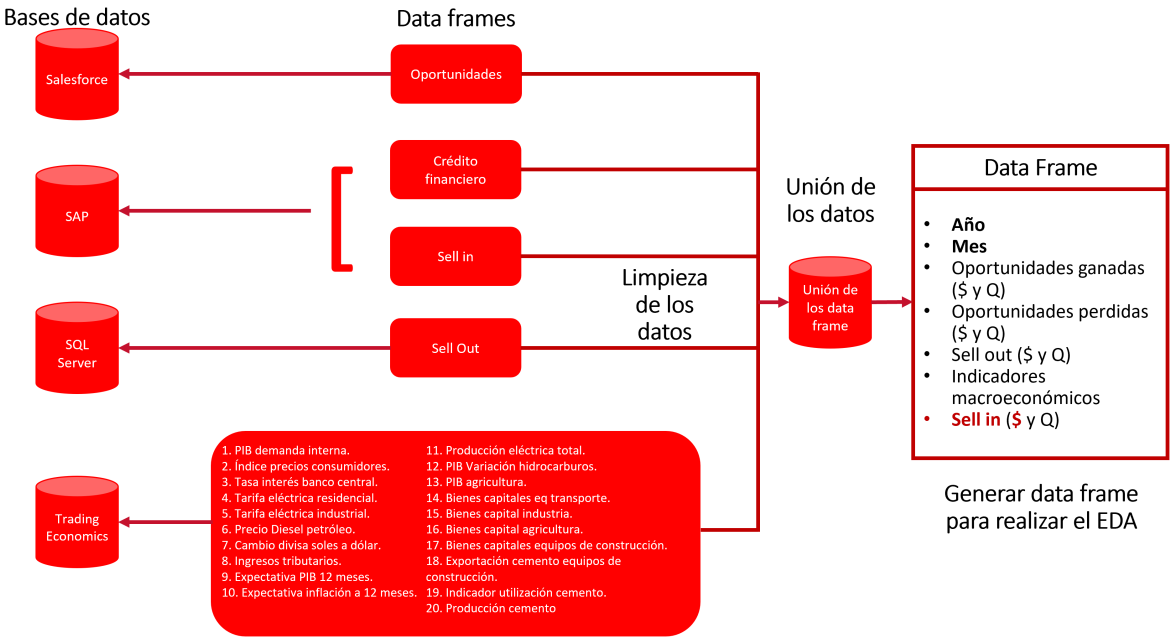


Figura 9: Generación de data frame maestro.

Los data frames que contienen las variables internas son los siguentes:

1. **Oportunidades**
2. **Crédito financiero**
3. **Sell in**
4. **Sell out**

Estos dataframes, contienen información ordenada por día y cliente, por lo que se planteó que mediante un group by, se suman los valores cuantitativos para resumir la información por mes. Al tener estos data frames resumidos por mes, se procedió a organizarlos por Unidad de negocio y Tipo de cliente, ya que cada cliente tiene una única categoría de estas, esto permitiendo llegar a unos data frames que se pueden relacionar. El data frame que contiene variables externas, ya contenía la información organizada por mes y por año, junto a la variable seleccionada. Se generan 5 data frames por cada una de las categorías que se analizan, para proceder a unir los datos, en la figura 10, se puede ver cual fue la variable usada para unir cada uno de los data frames y el resultado después de la limpieza, Al tener el data frame maestro, se comienza a realizar un análisis exploratorio de la data.

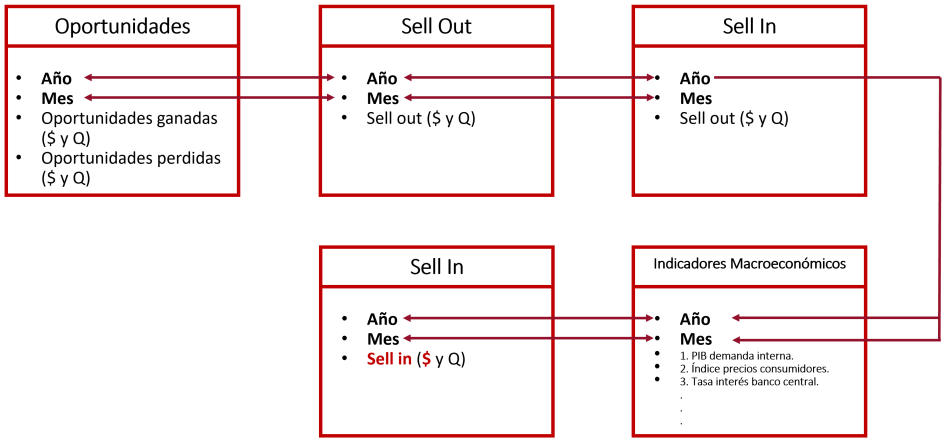


Figura 10: Relación de los data frames.

El funcionamiento del código para la realización de este procedimiento, es mediante la librería Beautiful Soup, esta es una biblioteca de Python utilizada para extraer datos de HTML y XML. Es muy útil para analizar páginas web y extraer información específica de ellas. Para usarla primero se debe importar Beautiful Soup en el notebook de Python y posterior a esto obtener la URL del contenido que se desea extraer, para seleccionar el HTML que contienen las variables macroeconómicas. Con el código requests se descarga este contenido HTML de las páginas web. Una vez capturado el contenido HTML, se crea un objeto Beautiful Soup pasándole ese contenido, con soup = BeautifulSoup(html_content, 'html.parser'). Con los métodos find() y find_all() se buscan las etiquetas HTML específicas y atributos. Una vez guardados los datos, con get_text() se obtiene el texto dentro de la etiqueta HTML. Todas estas variables se consolidaron en un formato que fuera afín a las demás bases de datos, es decir por mes y año, al menos desde el año 2015 porque fueron los datos facilitados por la empresa de tecnología, finalmente se obtiene una tabla en excel. En la figura 11, se puede ver DataFrame de las variables externas:

	Month	Year	pib_demanda_interna	indice_precios_consumidor	tasa_interes_banco_central	tarifa_electrica_residencial
0	1	2012	13.8	74.24	4.25	101.0
1	2	2012	13.8	74.48	4.25	105.0
2	3	2012	20.2	75.05	4.25	104.0
3	4	2012	14.7	75.45	4.25	104.0
4	5	2012	20.3	75.48	4.25	101.0
...	...	...	...	...	...	...
121	2	2022	-2.5	100.35	3.50	136.0
122	3	2022	1.4	101.84	4.00	134.0
123	4	2022	4.9	102.82	4.50	133.0
124	5	2022	0.2	103.21	5.00	133.0
125	6	2022	5.1	104.44	5.50	132.0

Figura 11: DataFrame de las variables externas.

Las variables contenidas en esta tabla de excel son:

1. Año.
2. Mes.
3. PIB.
4. Demanda interna.
5. Índice precios consumidores.
6. Tasa interés banco central.
7. Tarifa eléctrica residencial.
8. Tarifa eléctrica industrial.
9. Precio Diesel petróleo.
10. Cambio divisa soles a dólar.
11. Ingresos tributarios.
12. Expectativa PIB 12 meses.
13. Expectativa inflación a 12 meses.
14. Producción eléctrica total.
15. PIB Variación hidrocarburos.
16. PIB agricultura.
17. Bienes capitales eq transporte.
18. Bienes capital industria.
19. Bienes capital agricultura.
20. Bienes capitales equipos de construcción.
21. Exportación cemento equipos de construcción.
22. Indicador utilización cemento.
23. Producción cemento.

3.2 Análisis exploratorio de datos (EDA)

El análisis exploratorio de datos (EDA) se llevó a cabo con el propósito de comprender mejor la naturaleza de los datos y las relaciones entre las variables antes de proceder con un análisis más profundo de selección y evaluación de los modelos de machine learning. Esta etapa es de gran importancia en el análisis de datos, ya que ayuda a identificar patrones, tendencias, valores atípicos y posibles correlaciones, lo que puede guiar la selección y construcción de modelos posteriores. Inicialmente, se seleccionaron las top 3 unidades de negocio junto a 2 los top tipos de cliente, ya que representan un porcentaje alto de las ventas, 72 % y 75 % respectivamente. En la figura 12, se observa el diagrama de pareto para la selección de las tres unidades de negocio y los dos tipos de cliente.

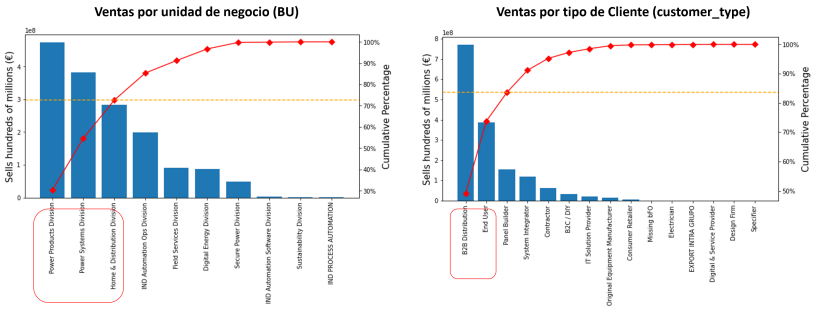


Figura 12: Diagrama de pareto para la selección de las ventas.

El EDA continuo con un resumen estadístico de las variables, incluyendo medidas de tendencia central y dispersión, como media, desviación estándar, mediana, mínimo y máximo. Este resumen proporcionó una idea inicial sobre la distribución y variabilidad de las variables. A continuación, se generaron gráficos, de diagrama de caja y mapas de calor de correlación, para visualizar la distribución de los datos y las relaciones entre las variables. Los diagramas de caja permitieron identificar valores atípicos y la variabilidad de las variables a lo largo del tiempo, mientras que el mapa de calor de correlación ayudó a identificar patrones de multicolinealidad y las relaciones de las variables con la variable objetivo (ventas).

Se exploraron las relaciones temporales al analizar la correlación de las variables con diferentes retrasos (*delays*) en el tiempo. Es importante presentar la definición de retrasos, que hace referencia a los valores pasados de las variables analizadas [82], lo que permitió identificar patrones de relación entre las variables y las ventas en diferentes momentos. Esto ayudó a entender cómo los valores pasados de las variables influyen en las ventas actuales.

De la misma manera, se aplicaron técnicas de visualización, como gráficos de dispersión y métricas de correlación, para comprender mejor las relaciones entre pares de variables. Estas visualizaciones permitieron identificar tendencias lineales y no lineales, así como la fuerza y dirección de las correlaciones. Este enfoque proporcionó una base sólida para tomar decisiones informadas sobre qué variables incluir en análisis posteriores y modelos predictivos. En la figura 13, se expone el diagrama con las fases realizadas para el análisis exploratorio de los datos de las variables que se utilizarían para la predicción de ventas.

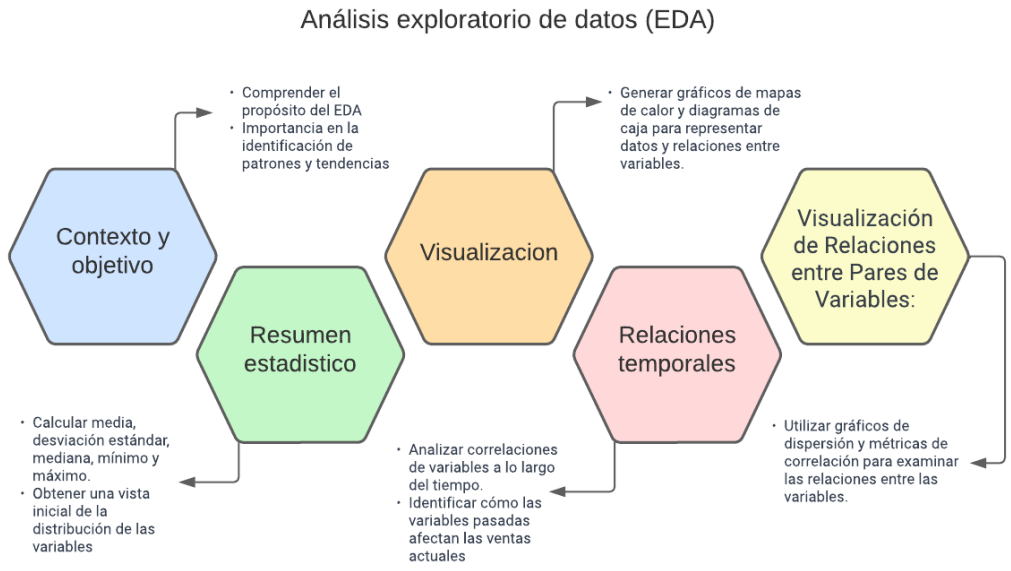


Figura 13: Diagrama de fases del análisis exploratorio de datos.

3.3 Selección y evaluación de los modelos

Los modelos de machine learning para pronosticar series de tiempo requieren de los valores históricos de las variables para poder pronosticar los valores futuros. Estos modelos parten del hecho de que el valor de la variable a predecir en un punto de tiempo determinado depende de una combinación lineal de todos los valores pasados de la misma variable; adicionalmente existen modelos que no solamente toman valores pasados de la variable a predecir sino también de otras variables externas. En las ecuaciones 3 y 4, se exponen las ecuaciones de un modelo de pronostico de serie de tiempo para dos variables, el cual puede expresarse de la siguiente manera:

$$Y_{1,t} = \alpha_1 + \beta_{1,1,1}Y_{1,t-1} + \beta_{1,2,1}Y_{2,t-1} + \beta_{1,1,2}Y_{1,t-2} + \beta_{1,2,2}Y_{2,t-2} + \varepsilon_{1,t} \tag{3}$$

$$Y_{2,t} = \alpha_2 + \beta_{2,1,1}Y_{1,t-1} + \beta_{2,2,1}Y_{2,t-1} + \beta_{2,1,2}Y_{1,t-2} + +\beta_{2,2,2}Y_{2,t-2} + \varepsilon_{2,t} \tag{4}$$

En donde α es el intercepto de la función y β es el coeficiente de cada uno de los retrasos de las variables Y_1 y Y_2; estos coeficientes se pueden interpretar como el efecto que tienen los retrasos de las variables Y_1 y Y_2 en la variable objetivo de la función. Por último ε es el error el cual es considerado como ruido blanco.

Cada uno de los modelos implementados realiza una validación de las variables a disposición y requiere una serie de pruebas para elegir las variables que finalmente se utilizarán como predictoras de la variable de interés. Cada modelo tiene sus ventajas y desventajas, por lo que se debe elegir el que mejor se adapte a las tres (3) ventanas temporales. El objetivo es

que el algoritmo sea capaz de seleccionar aquel modelo que mayor precisión presente en cada una de las ventanas temporales y datos compartidos por la empresa.

Posteriormente la evaluación de los modelos, es una de las etapas finales en lo que se refiere a los modelos de machine learning, para la definición del mejor modelo. Una vez es aplicado el modelo a las series temporales se evalúa el comportamiento de los residuos para definir si este es adecuado para los datos proporcionados; principalmente se analizan métricas, que definen cual es el modelo que mejor predice el comportamiento de la variable de interés. A continuación se exponen las métricas.

1. Debe haber evidencia suficiente para concluir que no existe correlación entre los residuos (errores de predicción del modelo), esto se puede hacer a través del Test de Durbin - Watson.

2. Una vez se compruebe que el modelo definido si es adecuado para los datos se valida que tan acertado fue este; para poder cuantificar el ajuste del modelo se pronostíca una porcion de los valores históricos ya conocidos y se utiliza una de las siguientes medidas de error:

   - **Error de Porcentaje Medio Absoluto (MAPE):**
     Es una medida de error relativa que usa valores absolutos, evitando que se eliminen valores positivos y negativos. La ecuación 5, hace referencia al número total de observaciones, A el valor real de la serie de tiempo y F el valor que predijo el modelo [83].

$$MAPE = \frac{1}{n}\sum_{t-1}^{n}\left|\frac{A_t - F_t}{A_t}\right| \tag{5}$$

   - **Error de Raíz Cuadrada Media (RMSE):**
     Corresponde a la desviación estandar de los valores residuales de predicción (diferencia entre el valor real y el valor predicho). En la ecuacion 6, n hace referencia al número total de observaciones, A el valor real de la serie de tiempo y F el valor que predijo el modelo [84].

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(A_t - F_t)^2} \tag{6}$$

Debido a que el alcance de este trabajo de grado, tiene como propósito proyectar las ventas teniendo en cuenta los siguientes filtros respectivamente: Customer Category, tendrá en cuenta los valores B2B Distribution y End User; para BU solo se usarán los valores de Power Products Division, Power Systems Division y Home & Distribution. Los modelos que se utilizaron fueron Prophet, Vectores Autorregresivos (VAR), Forecast autorregresivo, PyAf y AutoARIMA. Adicionalmente se considero el uso de modelos que parten de redes neuronales (Wave Net, Deep AR), pero debido a que la cantidad de datos no era suficientemente amplia, no fue posible hacer uso de estos modelos ya que solo se cuenta con 88 datos mensuales desde enero del 2015 hasta agosto del 2022.

Una vez se definieron los modelos que se iban a utilizar y los filtros que se iban a implementar, se realizo un preprocesamiento de los datos con el fin de poder identificar cuales

eran las variables a tener en cuenta para proyectar los modelos. Los datos fueron cargados a un NoteBook en Python haciendo uso de la libreria Pandas y los algoritmos de pronósticos multivariados como VAR y Forecast autorregresivo se implementaron para proyectar mas de una serie de tiempo, las cuales tienen una influencia sobre la otra. Cada una de las variables se modelo en función de los valores pasados de si misma y de los demas, pero fue necesario realizar una serie de pruebas las cuales permitieron confirmar la existencia de dicha influencia entre las variables. La pruebas que se realizaron se exponen a continuacion.

- **Prueba de causalidad de Granger:**

  Esta prueba tiene como hipótesis nula que los coeficientes de los valores pasados en la ecuación de regresión son iguales a cero (es decir, que los valores pasados no tienen ningún efecto en la variable a predecir); por lo que, si existe suficiente evidencia estadística para rechazar la hipótesis nula, se pude afirmar que los valores pasados si tienen un efecto causal en la variable a proyectar [85].

- **Prueba de cointegración:**

  Soren Johanssen (1991) desarrolló un procedimiento para implementar la prueba de cointegración, esta consiste principalmente en determinar si existe suficiente evidencia estadística para afirmar que dos series de tiempo están relacionadas a lo largo del tiempo; esta hipótesis es el fundamento principal sobre el que se basan los modelos VAR [86].

- **Prueba de Dickey-Fuller aumentada:**

  Esta prueba determina si las series de tiempo son estacionarias. En caso de no serlo, es necesario realizar una transformación a todas las variables, por ejemplo, restar el valor de cada periodo con el inmediatamente anterior. Una vez realizada la transformación se vuelve a realizar la prueba [87].

- **Prueba de Durbin Watson:** Despues de realizar las tres (3) pruebas expuestas anteriormente, se descartan las variables que no tengan influencia en las ventas a proyectar, así como aquellas que a pesar de las transformaciones no sean estacionarias. Una vez aplicado el modelo a las variables de interés, se realiza una última prueba llamada Durbin Watson, la cual toma los valores residuales del modelo con el fin de analizar correlación entre estos, en caso de que se tenga un valor de correlación positivo o negativo, es necesario evaluar si se deben descartar nuevamente las variables para obtener un mejor modelo [88].

Modelos de Pronóstico

A continuación se presenta la metodología realizada para cada uno de los modelos seleccionado y utilizados en el proyecto para la predicción de ventas.

- **Pronóstico Autorregresivo:**

  La primera modificación que se necesita hacer, para aplicar modelos de Forecasting es realizar una separación en datos de entrenamiento y datos de prueba. Posterior a esto, se transforma la serie temporal en una matriz, en la que cada serie temporal está asociada a una ventana (retrasos) que le precede [47]. A continuación, se crea un objeto llamado Forecaster, utilizando el codigo ForecasterAutorreg para realizar predicciones utilizando un modelo de Arima.

Como regresor interno del forecaster se proporciona un RandomForestRegressor, esto significa que se utilizará un modelo de busqueda aleatoreo para realizar las predicciones. Se especificasn 12 retrasos, dado que se quiere predecir los 12 meses siguientes, los indican cada uno de los tiempos que se utilizaran en la serie temporal como variables de entrada para predecir el siguiente valor.

De ahí, la definición de forecast autorregresivo donde las 12 observaciones se utilizan como caracteristicas para predecir los 12 siguientes valores, se ajusta el forecaster utilizando los valores de entrenamiento e indicando que la variable objetivo será Sales_Sell_In. Con el objeto de Forecaster entrenado, se utiliza el metodo predict() del forecaster y como argumento se ponen los 12 pasos a predecir. El resultado de la predicción se guardará en la variable de predicciones. Este proceso se realizó para los customer_type y BU anteriormente mencionados.

- **Vectores Autorregresivos:**

  Para la aplicación del modelo de vectores autorregresivos es necesario aplicar las pruebas descritas anteriromente a los datos (prueba de causalidad de Granger, prueba de cointegración y prueba de Dickey - Fuller) con el fin de poder seleccionar las variables que servirán de soporte para predecir las ventas según el filtro que se esté aplicando por Customer Category o BU.

  Una vez seleccionadas las variables y en caso de ser necesario aplicar las transformaciones correspondientes, se aplica el modelo de vectores autorregresivos a los datos, para una correcta implementación del modelo, es necesario seleccionar la cantidad de retrasos que debe tener en cuenta cada modelo. Al finalizar la ejecución del modelo, se revisa si existe autocorrelación en los residuos haciendo uso de la prueba Durbin-Watson, luego se predice la cantidad de periodos a futuro que se desee para este modelo.

- **Prophet:**

  Prophet es un modelo desarrollado por Meta, este es un modelo univariado por lo que no requiere pasar por el proceso de selección de variable, para implementar este modelo en la proyección de las ventas se separan los datos en un 80 % de entrenamiento y un 20 % de prueba.

  Una vez aplicado el modelo se recorre toda la serie de datos y por cada dato en el grupo de datos de entrenamiento se genera una predicción del periodo siguiente, se compara el valor real con la predicción y se ajusta el modelo a partir del error. Una vez realizado este proceso para todos los datos de entrenamiento, se realiza el pronóstico de los datos de prueba y se comparan para poder definir si el pronostico generado logra predecir correctamente la serie de tiempo.

- **PyAF:**

  Para el modelo de PyAF se crea un nuevo dataframe llamado data que contiene solo las columnas Year_Month_str y Sales_Sell_in, que son las únicas dos variables necesarias para ejecutar el modelo. Se importa la clase cForecastEngine del módulo pyaf.ForecastEngine y se crea una instancia de cForecastEngine llamada lEngine. Se entrena el modelo utilizando los datos de data y se especifica que la columna Year_Month_str

es la variable independiente y Sales_Sell_in es la variable dependiente. Además, se establece un parámetro de temporada de 12, lo que indica que hay una estacionalidad mensual en los datos. Con lEngine.getModelInfo() se obtiene información sobre el modelo entrenado, como los componentes y las métricas de rendimiento. Finalmente con lEngine.forecast(data, 4): Se realiza un pronóstico utilizando el modelo entrenado para los próximos 4 períodos. Los resultados del pronóstico se almacenan en el dataframe data_forecast. Para visualizar de mejor manera el modelo, se utiliza plotly.graph_objects as go: Se importa la clase go del módulo plotly.graph_objects para crear gráficos interactivos. Se crean cuatro trazas en el gráfico utilizando la función add_trace de go.Figure(). Cada traza representa una línea en el gráfico con diferentes variables del dataframe data_forecast.

- **XG BoosT Regressor:**

  XG BoosT utiliza la clase ForecasterAutoreg de la biblioteca sktime en Python para realizar un pronóstico de series de tiempo con variables exógenas.

  En primer lugar, se crea una instancia del modelo ForecasterAutoreg con los siguientes parámetros: regressor: Se utiliza el algoritmo XGBRegressor del paquete xgboost como regresor base para realizar las predicciones. Este algoritmo es un modelo de refuerzo (boosting) basado en árboles de decisión optimizados para obtener un mejor rendimiento en términos de velocidad y precisión.

  Se establecen 12 , lo que significa que el modelo utilizará las 12 observaciones anteriores de la serie de tiempo como variables de entrada para predecir el siguiente valor. Se ajusta el modelo utilizando el método fit, pasando los argumentos de entrenamiento, 'Sales Sell in' del dataframe datos train, las variables exógenas de entrenamiento, representadas por el dataframe variables exog train.

  El modelo se entrena utilizando la serie de tiempo y las variables exógenas proporcionadas, permitiendo que el algoritmo XGBRegressor aprenda los patrones y relaciones entre las variables para realizar pronósticos precisos. Una vez que el modelo ha sido ajustado, se devuelve la instancia del modelo forecaster, que ahora está listo para realizar predicciones utilizando el método predict.

- **LightGBM:**

  El modelo descrito crea un "forecaster.º pronosticador utilizando la clase ForecasterAutoreg del paquete sktime. Este pronosticador se construye utilizando un regresor de LightGBM (LGBMRegressor) como modelo base y se configura para utilizar 12 (retrasos) anteriores como variables predictoras. Después de crear el pronosticador, se ajusta (entrena) utilizando los datos de entrenamiento. La variable objetivo (y) se especifica como Sales_Sell_in de los datos de entrenamiento, mientras que las variables exógenas (exog) se obtienen de variables_exog_train. El ajuste del pronosticador implica entrenar el regresor de LightGBM utilizando las variables predictoras y la variable objetivo proporcionadas. Una vez que se completa el ajuste, el pronosticador está listo para realizar predicciones utilizando los datos de prueba o nuevos datos. Así finalmente se construye el modelo de predicción multivariada.

- **Hist Gradient Boosting Regressor:**

El ForecasterAutoreg se entrena utilizando una variable objetivo y variables exógenas (exog). En este caso, la variable objetivo es Sales_Sell_in y las variables exógenas son variables_exog_train, las cuales se escogieron en procesos anteriores. El parámetro se establece en 12, lo que significa que el modelo utilizará las últimas 12 observaciones de la variable objetivo y las variables exógenas para realizar las predicciones. Esto permite que el modelo capture las dependencias temporales y utilice la información histórica para predecir valores futuros. Una vez que el modelo se entrena con los datos de entrenamiento, se puede utilizar para realizar predicciones futuras. Esto implica proporcionar valores de las variables exógenas correspondientes al período de pronóstico deseado. El modelo utilizará los datos históricos y las características exógenas para generar las predicciones.

3.4 Desarrollo de la aplicación web de visualización

Finalmente, la ultima fase de este proyecto, consiste en el desarrollo de la aplicación web para la visualización de datos, con el objetivo de presentar de manera efectiva el histórico y las estimaciones de ventas de compañías a corto, mediano y largo plazo. Para el desarrollo, el framework Angular fue utilizado para el desarrollo, aprovechando las capacidades de Angular Material para la interfaz de usuario y Highcharts para la representación gráfica de los datos. El despliegue de la aplicación fue realizado a traves de Firebase, garantizando su disponibilidad en línea y capacidad de escalabilidad. A lo largo de este proceso, se dio especial atención al diseño de la interfaz de usuario para proporcionar una experiencia de usuario intuitiva y eficiente en la visualización de información relacionada con las ventas de las compañías.

El desarrollo de la aplicación web se dividió en varias etapas clave. En primer lugar, se configuró el entorno de desarrollo, asegurándose de que Node.js, npm y Angular CLI estuvieran correctamente instalados. Luego, se creó un nuevo proyecto Angular y se integraron las librerías de Angular Material y Highcharts para aprovechar sus funcionalidades. La integración de datos y la lógica de negocio se llevaron a cabo de manera muy cuidadosa para asegurarse de que las visualizaciones sean precisas y coherentes.

Para lograrlo, se utilizó un servicio, el cual utiliza la biblioteca HttpClient para hacer una solicitud a una URL específica donde se encontraba en un formato JSON la información almacenada de la salida de los modelos de machine learning. De esta manera, se obtienen los datos relacionados con las predicciones de ventas en la aplicación mediante una petición tipo GET. Estos datos luego se utilizan en la aplicación para crear visualizaciones precisas y significativas. Además, se diseñaron y construyeron componentes específicos para presentar el historial y las estimaciones de ventas en diferentes plazos, asegurando una experiencia de usuario fluida y atractiva.

En resumen, el desarrollo de esta aplicación en Angular, representa una solución sólida y efectiva para el análisis de ventas a corto, mediano y largo plazo. Esta metodología de desarrollo permitió crear una interfaz de usuario moderna y atractiva, junto con gráficos interactivos que facilitan la comprensión de los datos. La elección de Firebase para el despliegue garantiza la disponibilidad en línea y la capacidad de adaptación a medida que la aplicación crece. En conjunto, este enfoque proporciona una herramienta valiosa para la toma de decisiones basada en datos en el contexto empresarial. En la figura 14, se presenta el *wireframe* que desempeñó un papel inicial como diseño preliminar y esquema visual para el diseño de la aplicación. Este proporcionó una representación gráfica de la estructura y disposición de los elementos clave de la interfaz de usuario.

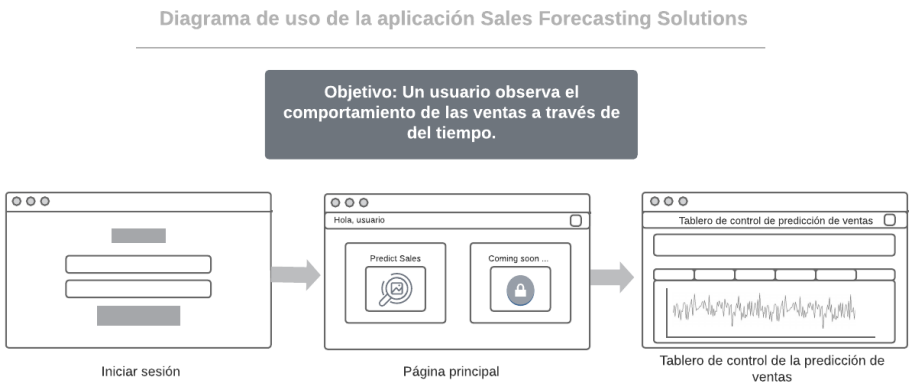


Figura 14: Wireframe de la aplicación web.

Capítulo 4

RESULTADOS

En este capítulo se detallan de manera específica y puntual los resultados obtenidos en este trabajo de grado. Inicialmente, se exponen los resultados obtenidos durante la investigación y selección de los datos, posteriormente del análisis exploratorio de datos, que proporciona una visión detallada de las características y patrones presentes en los datos. Seguidamente, se describe en detalle el proceso de selección de modelos, mostrando los criterios y las justificaciones detrás de cada elección. Igualmente, se presentan los resultados de la evaluación de los modelos seleccionados, resaltando su desempeño y su adecuación a los objetivos del proyecto. Finalmente, se aborda el resultado del desarrollo de la aplicación web, detallando su diseño, funcionalidades implementadas y su interacción.

4.1 Investigación y selección de los datos

Esta sección, presenta los resultados obtenidos en la fase de investigación y selección de datos para la elaboración del modelo de pronóstico de ventas utilizando técnicas de machine Learning y haciendo uso de variables exógenas. Este proceso es esencial para garantizar la precisión y la confiabilidad de las proyecciones de ventas, lo que, a su vez, contribuirá a una mejor toma de decisiones y planificación estratégica en la empresa, es por esto que la investigación se centró en la obtención de datos tanto internos como externos con base en la experiencia y experticia de los equipos comerciales, gerenciales y de presidencia que servirán como variables exógenas para alimentar el modelo de pronóstico de ventas.

Estas variables exógenas son fundamentales, ya que representan factores internos, los cuales representan indicadores bajo su control de cómo la empresa se comporta y toma sus decisiones, y externos a la empresa que pueden influir en las ventas y que no están directamente bajo su control. La correcta selección y tratamiento de estas variables es crucial para obtener pronósticos precisos y útiles. La selección de las variables exógenas se llevó a cabo mediante un proceso de colaboración entre el equipo de desarrollo de la solución, el equipo comercial y los presidentes de país de la empresa.

Este enfoque colaborativo fue esencial para garantizar que las variables seleccionadas reflejen de manera precisa el entorno en el que opera la empresa y tengan un impacto significativo en las ventas. Para esto se realizó una revisión exhaustiva de los métodos tradicionales que sigue la empresa para la realización de su pronóstico de ventas y realización de estrategias de ventas, y de todas las posibles variables exógenas que podrían tener un alto impacto en las

ventas de la empresa. Esto incluyó factores económicos, crediticios, de proyectos y de ventas que influyen en el comportamiento de las ventas de la empresa.

Las variables potenciales se diferencian en dos tipos, externas e internas, y estas provienen de diferentes bases de datos, finalmente con las variables seleccionadas se fabrican unas tablas de excel, estas son:

- **Crédito Financiero:** Acá se encuentra el cupo que tiene cada cliente para realizar compras a la empresa. Hace parte de las variables internas. En la figura 15, se expone el Dataframe de crédito financiero.
- **Oportunidades:** Contiene la información correspondiente a las oportunidades en curso, perdidas, ganadas y el monto de la oportunidad. Hace parte de las variables internas. En la figura 16, se expone Dataframe de oportunidades.
- **Indicadores Macroeconómicos:** Cómo resultado de la conexión con web scraping a la página del Banco Centra de Reserva, se obtuvieron las 23 variables económicas que se usarían para predecir de manera correcta las ventas. En la figura 17, se expone Dataframe de indicadores macroeconómicos.
- **Ventas:** Finalmente el resultado de la extracción de la base de datos de ventas, es una tabla con la información de las ventas organizada por mes y año con las ventas por unidad de negocio. En la figura 18, se expone el Dataframe de ventas.

	Customer ID	Year	Month	Credit Limit
0	3480300	2022	3	71693.45
1	11065200	2022	7	12934.95
2	11065200	2022	6	36841.75
3	11065200	2022	4	1473.20
4	11065200	2022	3	68299.85
...	...	...	...	...
2525	56420000	2022	6	285510.14
2526	56420100	2022	1	7844.05
2527	56420100	2021	10	2808.14
2528	56420100	2021	9	41517.99
2529	56420100	2021	8	0.00

Figura 15: Dataframe de credito financiero.

	Year	Month	Customer ID	Customer Category	BU	Sales_Oppt_win	Quantity_Oppt_win	Sales_Oppt_lost	Quantity_Oppt_lost
0	2010	1	56331600	End User	IND Automation Ops Division	39362.13	1	0.00	0
1	2010	1	56324800	Contractor	Power Systems Division	17369.80	2	0.00	0
2	2010	2	56324800	Contractor	Power Systems Division	18703.50	2	0.00	0
3	2010	3		End User	IND Automation Ops Division	85068.34	1	0.00	0
4	2010	3	56324800	Contractor	Power Systems Division	11381.13	2	0.00	0
...	...	...	...	...	...	...	...	...	...
6887	2022	9	56417800	End User	Power Products Division	56418.56	1	0.00	0
6888	2022	9	56329400	End User	Field Services Division	19266.34	1	0.00	0
6889	2022	9	56415800	End User	IND Automation Ops Division	251862.95	2	0.00	0
6890	2022	9	10518	IT Solution Provider	Secure Power Division	8750.88	1	1592.00	1
6891	2022	9	56326800	Panel Builder	Power Systems Division	73941.30	1	194565.44	2

Figura 16: Dataframe de crédito oportunidades.

	Month	Year	pib_demanda_interna	indice_precios_consumidor	tasa_interes_banco_central	tarifa_electrica_residencial	tarifa_electrica_insutrial	precio_diesel_petroleo
0	1	2012	13.8	74.24	4.25	101.0	104.0	117.0
1	2	2012	13.8	74.48	4.25	105.0	109.0	117.0
2	3	2012	20.2	75.05	4.25	104.0	109.0	121.0
3	4	2012	14.7	75.45	4.25	104.0	108.0	120.0
4	5	2012	20.3	75.48	4.25	101.0	103.0	121.0
...	...	...	...	...	...	...	...	...
119	12	2021	-8.9	100.00	2.50	135.0	160.0	108.0
120	1	2022	-0.6	100.04	3.00	135.0	159.0	107.0
121	2	2022	-2.5	100.35	3.50	136.0	161.0	109.0
122	3	2022	1.4	101.84	4.00	134.0	159.0	112.0
123	4	2022	4.9	102.82	4.50	133.0	157.0	102.0

Figura 17: Dataframe de crédito indicadores macroeconómicos.

	Year	Month	Customer Category	Customer ID	Customer	Quantity_Sell_in	Sales_Sell_in	BU
0	2015	1	B2B Distribution	56301900	A M P INGENIEROS S.A.C.	31.0	3070.06	Home & Distribution Division
1	2015	1	B2B Distribution	56301900	A M P INGENIEROS S.A.C.	165.0	13459.13	Power Products Division
2	2015	1	B2B Distribution	56301900	A M P INGENIEROS S.A.C.	16.0	2523.15	IND Automation Ops Division
3	2015	1	B2B Distribution	56303900	BELLCORP REPRESENTACIONES S.A.C.	252.0	3894.93	Home & Distribution Division
4	2015	1	B2B Distribution	56303900	BELLCORP REPRESENTACIONES S.A.C.	33.0	2526.87	Power Products Division
...	...	...	...	...	...	...	...	...
28257	2022	8	System Integrator	56417600	INTELSAC S.A.C.	4.0	10717.12	IND Automation Ops Division
28258	2022	8	System Integrator	56417600	INTELSAC S.A.C.	4.0	2180.17	Power Products Division
28259	2022	8	System Integrator	56419600	DEMMPRO S.A.C.	3.0	487.68	IND Automation Ops Division
28260	2022	8	System Integrator	56422700	SEAL TELECOM COMERCIO E SERVICIOS	22.0	49165.20	Digital Energy Division
28261	2022	8	System Integrator	56422700	SEAL TELECOM COMERCIO E SERVICIOS	20.0	18216.02	Power Products Division

Figura 18: Dataframe de ventas.

Una vez extraídas, transformadas y cargadas las variables potenciales seleccionadas, se realiza una validación nuevamente con los expertos de la empresa para obtener retroalimentación. Como resultado de esta retroalimentación, se obtiene que se puede continuar con el proyecto con estas variables como el mínimo producto viable. Estas variables se consideraron como las más relevantes y significativas en función de su impacto en las ventas históricas y la comprensión experta. Las variables exógenas seleccionadas, evaluadas y aprobadas son:

- Ventas
- Crédito financiero
- Monto de oportunidades ganadas
- Monto de oportunidades perdidas
- Cantidad de oportunidades ganadas
- Cantidad de oportunidades perdidas
- PIB demanda interna
- Índice precios al consumidor
- Tasa interés banco central
- Tarifa eléctrica residencial
- Tarifa eléctrica industrial
- Precio diésel petróleo
- Cambio divisa soles a dólar
- Ingresos tributarios
- Expectativa PIB 12 meses
- Expectativa inflación a 12 meses
- Producción eléctrica total
- PIB var hidrocarburos
- Agricultura
- Bienes capitales equipo de transporte
- Bienes capital industria
- Bienes capital agricultura
- Bienes capital equipos de construcción
- Exportación cemento equipos de construcción
- Indicador utilización cemento
- Producción cemento

La selección cuidadosa de las variables exógenas es un paso crítico en la construcción de un modelo de pronóstico de ventas preciso. Estas variables representan factores externos que pueden influir en las ventas de una empresa y, por lo tanto, son fundamentales para comprender y predecir el comportamiento del mercado. Los resultados de esta fase de investigación y selección de datos proporcionan una base sólida sobre la cual construir el modelo de pronóstico.

Un modelo de pronóstico efectivo permite a la empresa anticipar y planificar sus ventas futuras en función de estos factores externos relevantes. Esto tiene un valor estratégico significativo, ya que permite a la empresa tomar decisiones informadas sobre la gestión de la producción, el inventario, la asignación de recursos y otras áreas clave del negocio. Además, la capacidad de prever las ventas con precisión puede mejorar la toma de decisiones estratégicas y operativas, lo que a su vez puede tener un impacto positivo en la rentabilidad y el crecimiento de la empresa.

La próxima sección del estudio se centrará en el análisis exploratorio de los datos, utilizando modelos estadísticos para validar y comprender mejor las relaciones entre las variables seleccionadas. Este análisis permitirá evaluar cómo las variables exógenas influyen en las ventas y qué patrones o tendencias pueden identificarse en los datos. Estas observaciones y conclusiones ayudarán a afinar y ajustar el modelo de pronóstico, garantizando que sea una herramienta efectiva para la planificación y la toma de decisiones estratégicas en la empresa.

4.2 Analisis exploratorio de datos (EDA)

Inicialmente, se procedió a realizar un resumen que abarca la totalidad de las variables. El resumen de las variables con sus métricas mas relevantes, se encuentra presentado de manera detallada en la tabla 4. Cabe destacar que las variables han sido dispuestas en una secuencia determinada por el coeficiente de variación. Este coeficiente es calculado mediante la división de la desviación estándar entre la media de cada variable.

Variable	**Media**	**Des. Estandar**	**Mediana**	**Min**	**Max**	**Coef. Variacion**
pib_demanda _interna	14.01	93.63	4	-90.1	990	6.684
url_pib_ var_hidrocarb	3.25	12.44	1.9	-45.8	67.1	3.824
Producción_ eléctrica_total	4.17	7.00	4.75	-29.3	38.6	1.678
url_pib_ agricultura	3.91	4.25	3.65	-5.1	16.4	1.085
Ventas	2831652.9	2765076.1	2188567	685007.3	27602572	0.976
Month	6.37	3.49	6	1	12	0.548
Expectativa_ PIB_12_meses	4.20	1.84	3.885	-3.05	9.25	0.438
tasa_interes _banco_central	3.07	1.34	3.5	0.25	4.5	0.434

Numero_de _Proyectos	72.33	23.84	70.5	24	147	0.330
ingresos_ tributarios	8676.33	2244.87	8132	4598	21136	0.259
url_bienes _capital _agricultura	12.55	2.93	12	7	21	0.234
url_exportacion _cemento_eq _construccion	12.69	2.78	13	3	18	0.219
url_bienes_ capital_eq _transporte	244.06	51.05	243	84	359	0.209
Expectativa _inflacion_a _12_meses	2.67	0.55	2.675	1.4	4.62	0.205
url_bienes_ capital_eq_ construcción	109.25	20.66	109.5	59	174	0.189
precio_diesel _petróleo	93.55	16.34	90.5	65	121	0.175
url_indicador _utilización _cemento	72.30	11.79	72.05	3	100	0.163
url_producción _cemento	135.41	22.07	134.9	5.7	187.3	0.163
tarifa_eléctrica _industrial	129.92	17.25	131	102	162	0.133
url_bienes_ capital_ industria	662.98	80.33	665	434	833	0.121
cambio_divisa _soles_a_dolar	3.22	0.38	3.2735	2.552	4.108	0.118
tarifa_eléctrica _residencial	116.06	10.57	118	98	137	0.091
indice_precios _consumidor	86.68	7.13	87.98	74.24	102.82	0.082
Year	2016.68	3.00	2017	2012	2022	0.001

Tabla 4: Estadísticas de las variables seleccionadas.

De la tabla anterior se extrae y presenta las siguiente información relacionada con las variables de interes que fueron investigadas y seleccionadas previamente, como se presento en la sección 4.1.

1. La media de ventas es de 2.831.652.96 euros y el máximo de ventas es de 27.602.572

euros que se debe a un proyecto puntual que hubo en Perú.

2. El PIB en Perú ha sido el indicador macroeconómico que más ha variado a través de los años, queriendo decir que la economía de Perú ha sido volátil en los últimos 10 años.

3. La tasa de cambio de sol a dólar, ha sido estable través de los años, queriendo decir que la moneda peruana se ha mantenido fuerte frente al dólar.

4. El índice de precios al consumidor se mantiene con una desviación estándar moderada. Esto indica que las fluctuaciones en los precios de consumo en Perú han sido controladas en general durante el periodo analizado.

5. La expectativa de inflación a 12 meses tiene una desviación estándar baja, lo que sugiere que las expectativas sobre la inflación en Perú se han mantenido estables en el período de estudio.

6. El número de proyectos muestra una mediana superior a la media, lo que indica que hubo períodos en los que se llevaron a cabo más proyectos de lo habitual. Esto puede reflejar cambios en la inversión y desarrollo en el país.

7. El precio del diesel de petróleo tiene una desviación estándar moderada, lo que sugiere variabilidad en el precio del combustible en el mercado peruano, pero en general no experimenta fluctuaciones.

8. La tarifa eléctrica industrial tiene una mediana baja con respecto a la media, lo que podría indicar que en algunos años hubo un aumento significativo en las tarifas eléctricas para la industria en Perú.

9. El índice de utilización de cemento presenta una variabilidad significativa, con valores que van desde 3 hasta 100. Esto podría indicar que la demanda de cemento en la construcción en Perú ha experimentado fluctuaciones a lo largo del tiempo.

10. La expectativa de PIB a 12 meses tiene una desviación estándar baja, lo que sugiere que las expectativas sobre el crecimiento económico a corto plazo en Perú han sido consistentes en el período de análisis.

11. El año tiene una desviación estándar baja y una mediana cercana al valor máximo, lo que indica que el conjunto de datos se centra en un rango estrecho de años, lo que concuerda con el período de estudio específico.

12. El coeficiente de variación más alto se encuentra en la variable pib_demanda_interna, lo que indica que esta variable ha experimentado fluctuaciones significativas a lo largo de los años, lo que podría deberse a cambios en la demanda interna del país.

13. La producción de cemento y su utilización en la construcción tienen coeficientes de variación bajos, lo que sugiere estabilidad en esta industria a lo largo del período analizado.

14. La tarifa eléctrica residencial tiene una mediana cercana al valor máximo, lo que indica que en general, las tarifas residenciales de electricidad en Perú tienden a ser significativamente altas.

15. La variable precio_diesel_petróleo muestra una mediana cercana al valor máximo, lo que sugiere que los precios del diesel en Perú tienden a ser altos.

16. La variable cambio_divisa_soles_a_dolar tiene una desviación estándar moderada y una mediana cercana a la media, lo que indica que la tasa de cambio entre el sol peruano y el dólar estadounidense ha mantenido estabilidad a lo largo del tiempo.

17. La variable indice_precios_consumidor tiene una desviación estándar baja y una mediana cercana a la media, lo que sugiere que los precios al consumidor en Perú han experimentado variaciones moderadas durante el período de estudio.

18. La variable año muestra que el período de estudio abarca desde 2015 hasta 2022, lo que proporciona una visión de la evolución de estas variables a lo largo de una década.

En la figura 19, se muestra una grafica de las ventas respecto a los meses. En la gráfica se puede ver un pico de las ventas entre el mes 80 y 100. El mes cero es el que corresponde al mes desde el cual se tomaron los datos, es decir enero del 2012. Las ventas están en un orden de 10 a la 7 y su unidad son los euros.

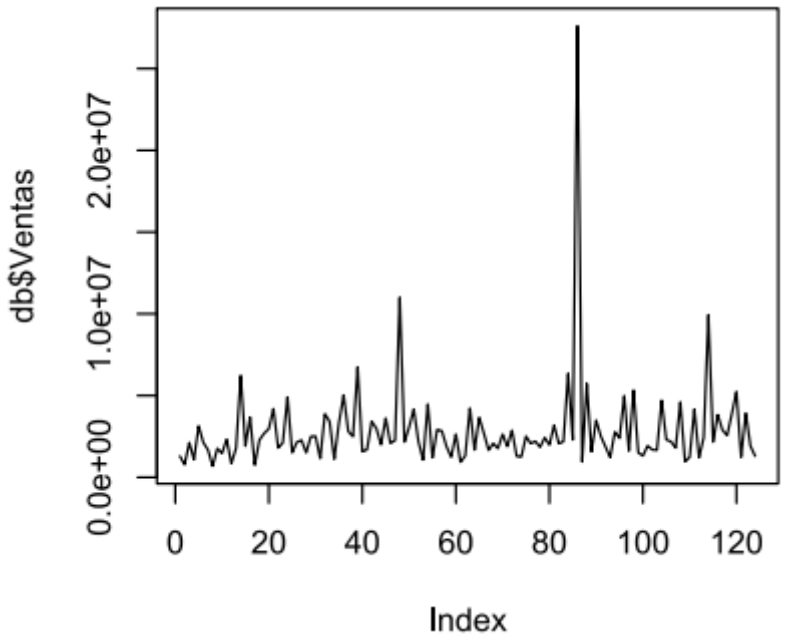


Figura 19: Gráfico de ventas a través de los meses.

Teniendo en cuenta los hallazgos anteriores, se identificaron un total de cinco (5) valores atípicos o datos dispersos. Estos valores atípicos son puntos de datos que se desvían significativamente de la tendencia general de los datos y pueden tener un impacto desproporcionado en los resultados del análisis. Identificar y comprender estos valores atípicos es fundamental, ya que pueden indicar eventos excepcionales o anomalías que afectaron las variables en esos momentos específicos.

Posteriormente, como parte del proceso de análisis, se llevó a cabo una evaluación de la correlación entre las variables seleccionadas. Esto se refleja en la figura 20, que muestra las

relaciones de correlación entre las variables. El objetivo principal de este análisis fue identificar posibles problemas de multicolinealidad, es decir, la existencia de altos niveles de correlación entre pares de variables independientes en el modelo, lo que podría dificultar la interpretación de los efectos individuales de estas variables en las ventas.

La figura 20 también proporciona información valiosa sobre la relación de cada variable con la variable de ventas, que es de particular interés en este estudio. Al examinar la correlación de cada variable con las ventas, es posible identificar aquellas que tienen una relación más fuerte o más débil con las ventas. Esto puede ayudar a determinar qué variables son más relevantes para predecir o explicar las fluctuaciones en las ventas, lo que a su vez puede guiar la selección de variables para incluir en un modelo de regresión o análisis predictivo.

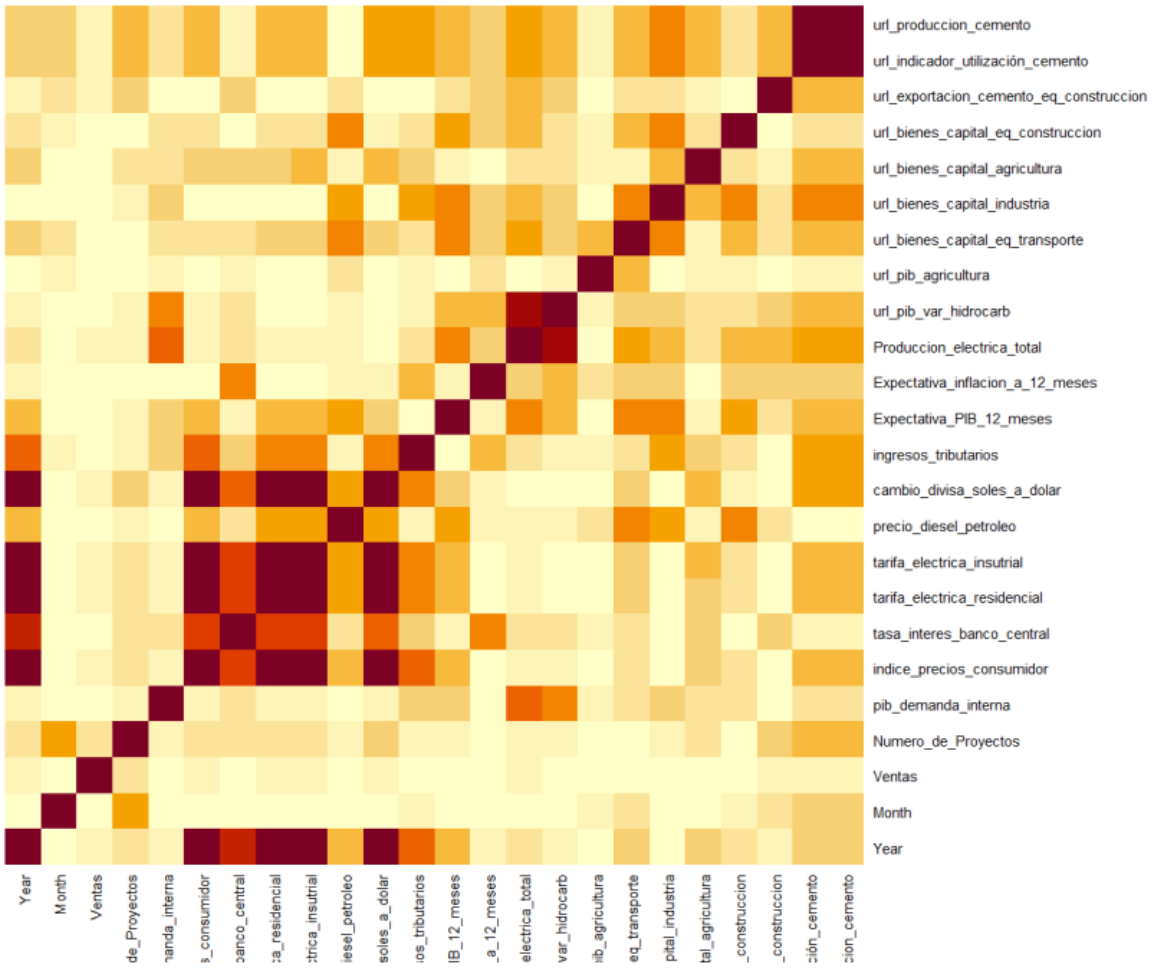


Figura 20: Mapa de calor de las variables correlacionadas entre sí.

En este análisis, se evalúa la relación entre las ventas y otras variables a lo largo del tiempo. Para comprender mejor cómo estas variables afectan a las ventas. Se aplico el concepto de retraso (delay), el cual implica desplazar los datos de las variables independientes es decir las que se estan analizando en el tiempo con respecto a las ventas. El objetivo de añadir un

retraso a las variables es observar cómo las condiciones pasadas afectan a las ventas actuales. Esto es útil cuando sospechamos que hay un efecto retardado en el comportamiento de las ventas debido a ciertas variables.

Puede ser que las decisiones de tomadas en un mes tengan un impacto en las ventas de los meses siguientes. La figura 21, muestra los resultados de este análisis de correlación. En el eje x, representamos los diferentes retrasos (meses anteriores) que aplicamos a las variables independientes. En el eje y, mostramos la medida de correlación entre cada variable y las ventas. La correlación es un indicador estadístico que mide la relación lineal entre dos variables y puede variar de -1 (correlación negativa perfecta) a 1 (correlación positiva perfecta), con 0 indicando una falta de correlación. De la misma manera, se analizó correlación de las variables con las ventas teniendo un retraso de meses en el eje x, y poniendo 3 escenarios de correlación con las ventas, en rojo la variable con mayor correlación, en verde la variable con menor correlación y en azul el promedio de correlaciones de todas las variables.

Figura 21: Métricas de correlación de las ventas respecto a las variables en los meses anteriores.

Posteriormente se llevo a cabo un análisis de correlación entre las ventas y múltiples variables independientes, considerando distintos períodos de retraso temporal, comúnmente denominados retrasos. Este enfoque metodológico tiene como propósito discernir el efecto de las variables en las ventas en función de la distancia temporal que media entre su ocurrencia y su impacto en el desempeño de ventas. Los resultados indican que se han observado correlaciones significativas entre las ventas y ciertas variables tanto en ausencia de retraso (0 meses) como en un horizonte temporal de retraso de seis meses.

Estos hallazgos insinúan la existencia de factores con efectos inmediatos sobre las ventas, así como la presencia de variables cuyo impacto en las ventas se manifiesta de manera más gradual y se consolida después de un período de seis meses. La representación gráfica de estas correlaciones se presenta en la figura 22, donde se examinan los coeficientes de correlación entre las ventas y las variables independientes a lo largo de un rango de retraso que se extiende desde 0 hasta 24 meses.

Un ejemplo de ello es la variable url_exportacion_cemento_eq_construccion, que exhibe

una correlación especialmente relevante con las ventas en el período de seis meses subsiguientes a su influencia. Este enfoque analítico tiene la finalidad de proporcionar una comprensión más detallada y estructurada de la relación temporal entre las variables independientes y las ventas, lo que, a su vez, puede servir de fundamento para la toma de decisiones estratégicas en el ámbito empresarial, tales como la planificación de estrategias de marketing o la gestión de inventario, basadas en las correlaciones identificadas.

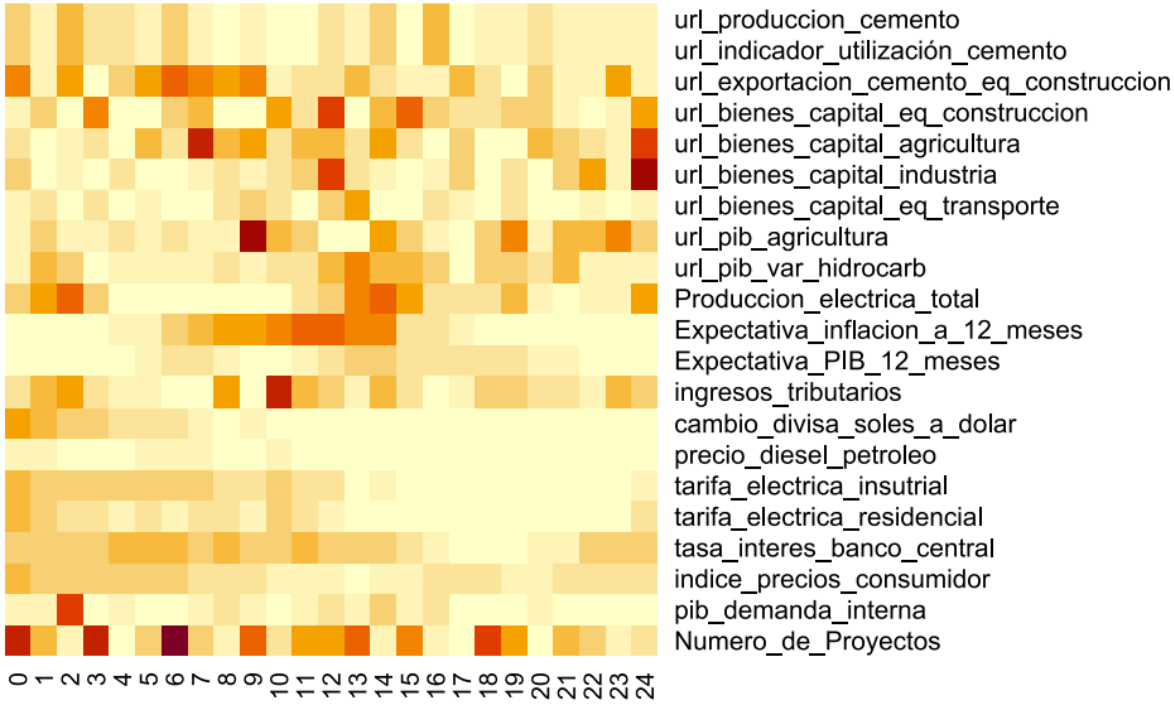


Figura 22: Mapa de calor de la correlación entre las variables y las ventas con retraso.

Siguiendo los análisis efectuados, se procedió a la exclusión de aquellas variables que presentaban multicolinealidad o cuya correlación con las ventas parecía ser limitada. Un ejemplo destacado de este proceso se puede apreciar en la figura 20, donde se evidencia que las variables relacionadas con la tarifa eléctrica residencial, tarifa eléctrica industrial e índice de precio al consumidor exhiben multicolinealidad en el modelo analítico. Una vez depurado el conjunto de variables, se llevaron a cabo nuevos análisis empleando únicamente las variables restantes, con el propósito de obtener un modelo más robusto y preciso.

Este enfoque de selección de variables es fundamental para reducir la complejidad del modelo, mejorar su interpretabilidad y asegurar que las variables incluidas en el análisis sean genuinamente significativas en la explicación de las variaciones en las ventas. Este proceso de refinamiento y selección de variables es esencial en la investigación empírica, ya que permite concentrar la atención en las relaciones más significativas entre las variables independientes y la variable dependiente. Las variables que se mantienen después de este proceso son aquellas que se considera que aportan un valor real al modelo analítico y contribuyen de manera más efectiva a la comprensión de los factores que influyen en las ventas.

Por otra parte, en las figuras 23 y 24 se muestran los diagramas de caja con las variables seleccionadas. Se puede observar como la variación y dispersión de las variables es baja. De

la misma manera, se observan los valores extremos de la caja que corresponden a los valores atípicos que no cumplen ciertos requisitos de heterogeneidad de los datos. Estos datos atípicos principalmente sobresalen en las variables de producción eléctrica total, observada en la figura 23 y a la producción de cemento de la figura 24 respectivamente.

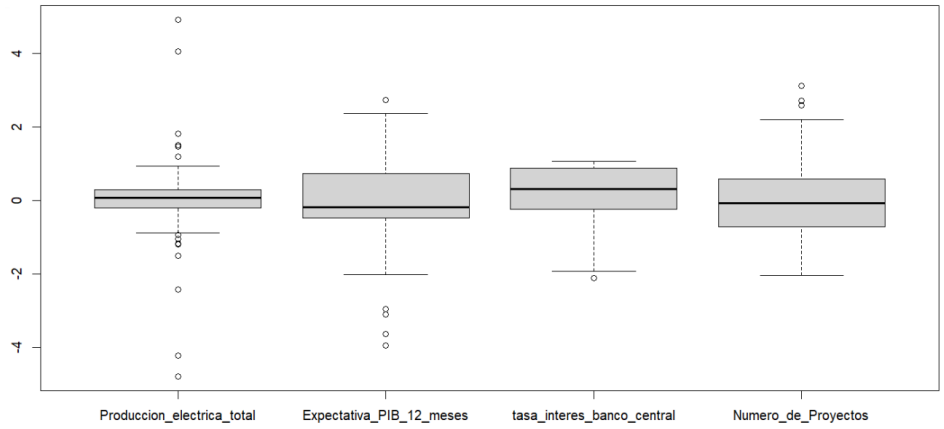


Figura 23: Diagrama de caja de bigotes 1° parte.

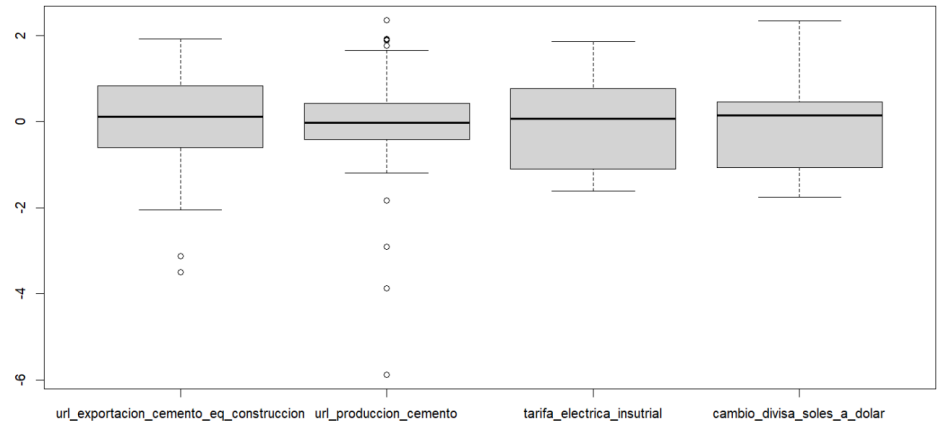


Figura 24: Diagrama de caja de bigotes 2° parte.

En la Figura 25, se expone el mapa de calor con el objetivo de evaluar la presencia de multicolinealidad entre las variables. Se observa que la multicolinealidad entre las variables se ha reducido significativamente debido a la eliminación de aquellas variables que mostraban una fuerte relación entre sí. Como resultado de este proceso de selección, el modelo se vera simplificado ya que consta de un conjunto de nueve (9) variables.

Un aspecto a destacar es el caso de las variables cambio de divisa y tarifa eléctrica industrial, apesar de que presentan una correlación elevada entre ellas, se ha tomado la decisión de mantener ambas en el modelo. Esta elección se ha basado en la consideración de que ambas variables poseen información relevante y única que puede contribuir a la resolución del problema de investigación.

La decisión de retener estas variables a pesar de su correlación alta se debe a la importancia

que cada una aporta al contexto del estudio. En ocasiones, incluso variables altamente correlacionadas pueden proporcionar perspectivas complementarias o capturar diferentes aspectos de un fenómeno. Esto resalta la necesidad de una evaluación cuidadosa de la multicolinealidad y la selección de variables en función de los objetivos y la lógica de la investigación.

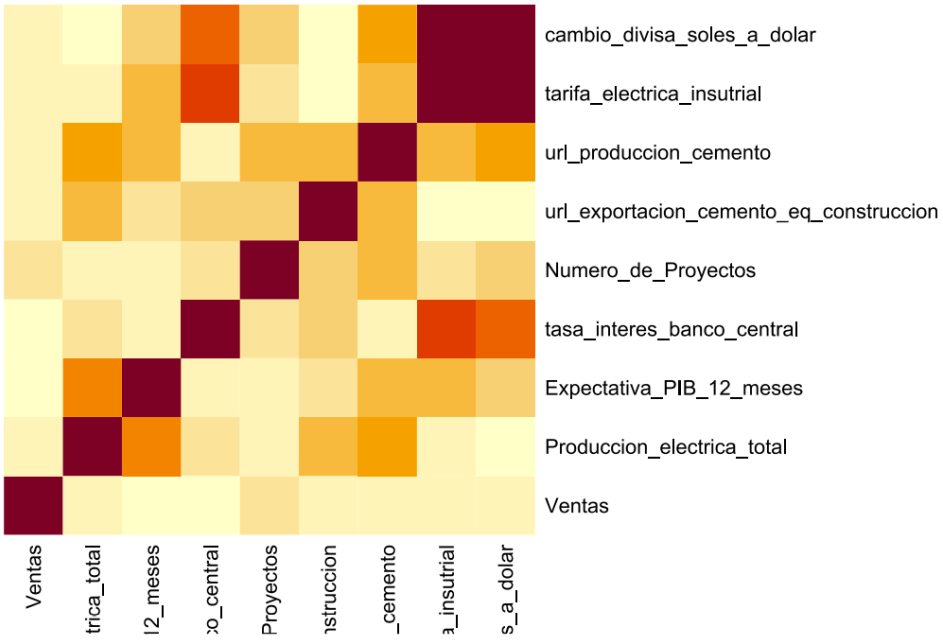


Figura 25: Mapa de calor reducido.

Posteriormente, se procedió a investigar las relaciones entre las variables que permanecieron en el análisis. Con este propósito, se generó un gráfico de dispersión que comparó cada variable con todas las demás, como se ilustra en la Figura 26. En este gráfico, se representaron las relaciones entre pares de variables, lo que permitió visualizar cómo se relacionan entre sí. Además de la representación gráfica de las relaciones, en la figura 26 se incluye una representación numérica del valor de correlación entre las variables en el triángulo superior derecho.

Esta información adicional proporciona una medida cuantitativa de la fuerza y la dirección de las relaciones entre las variables. La creación de gráficos de dispersión y la evaluación de correlaciones son pasos esenciales en el análisis de datos para comprender mejor la estructura subyacente de las relaciones entre las variables. Estos análisis pueden revelar patrones, tendencias y asociaciones que ayudan a fundamentar las conclusiones y las recomendaciones en una investigación. Si hay más detalles que desees explorar o alguna pregunta específica relacionada con esta parte del análisis, por favor, indícamela, y estaré encantado de proporcionar más información.

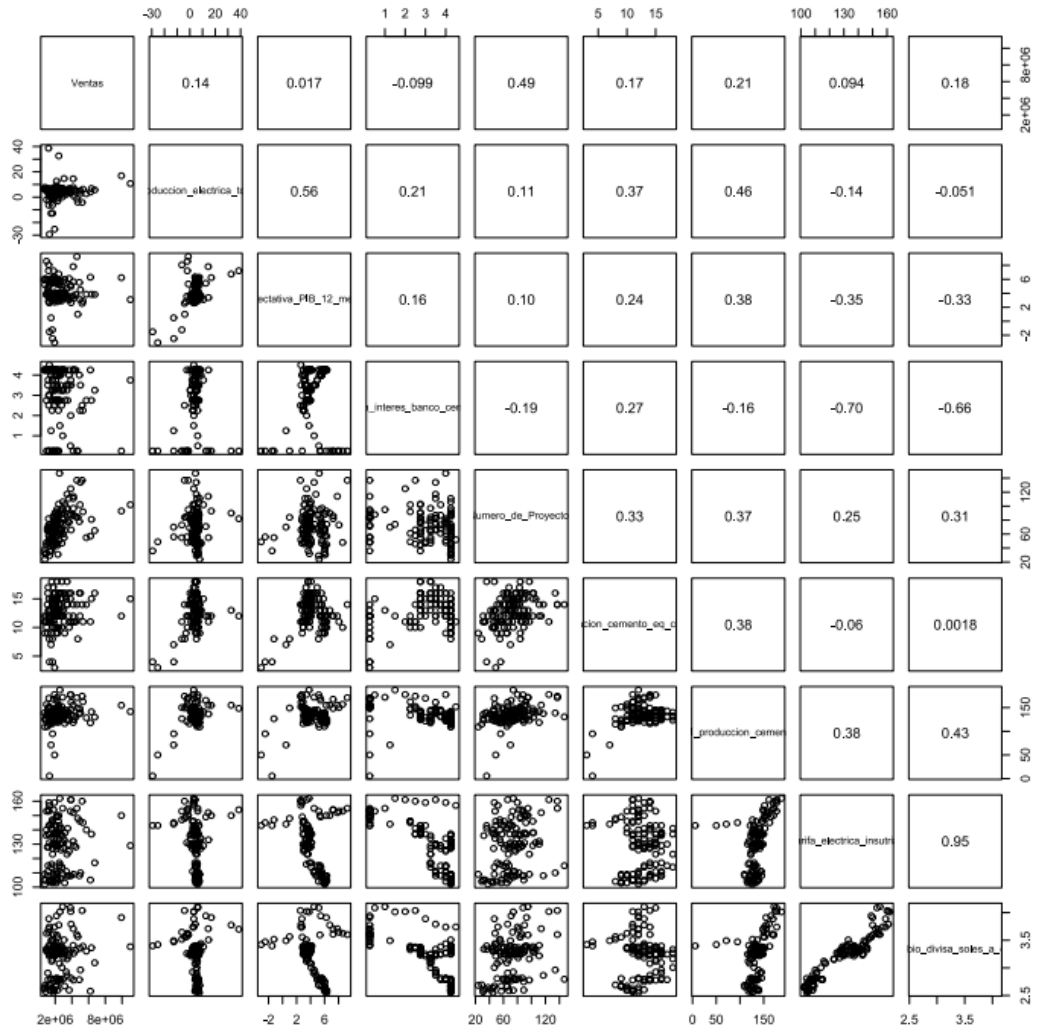


Figura 26: Grafica de las relaciones pareadas de las 9 variables seleccionadas.

En la figura 26, se realiza un análisis visual de la correlación de varias variables con las ventas. Se observa que hay una tendencia lineal positiva entre el cambio de divisa de soles a dólares y la tarifa eléctrica industrial. Del mismo modo, la producción de cemento muestra una correlación positiva tanto con la tarifa eléctrica industrial como con el cambio de divisa. Estas relaciones sugieren que cambios en la tasa de cambio de la moneda local y en la tarifa eléctrica pueden influir en la producción de cemento y, por ende, en las ventas.

La figura 26 también muestra tres escenarios de correlación de las variables seleccionadas con las ventas. En estos escenarios, las variables se destacan en colores diferentes según su correlación con las ventas. Se observa que los escenarios "mes sin retrasoz "6 meses de retrasoesultan en una mayor correlación de las variables con las ventas, lo que indica que estas dos temporalidades pueden ser especialmente relevantes para comprender los patrones de ventas en el contexto de análisis. Este análisis visual proporciona información importante sobre las relaciones entre las variables y su impacto en las ventas, lo que puede ser fundamental para tomar decisiones informadas basadas en datos.

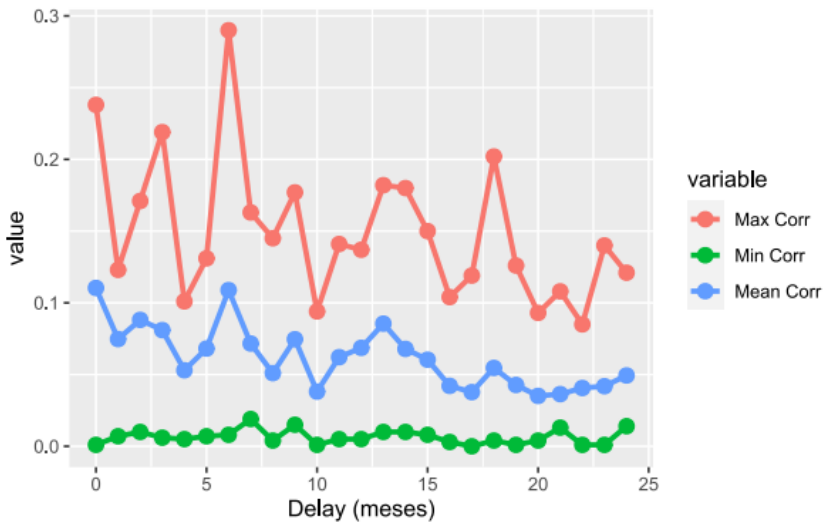


Figura 27: Métricas de correlación de las ventas respecto a las variables seleccionadas en los meses anteriores.

En la figura 27, se destacan dos observaciones significativas con respecto a las correlaciones de las variables con retrasos de 0 y 6 meses con las ventas de la compañía de tecnología. En primer lugar, se destaca que la variable con la correlación más alta, tanto en el escenario de ningún retraso como en un retraso de 6 meses, es el número de proyectos. Esto sugiere que la cantidad de proyectos en curso o completados tiene una influencia en las ventas de la compañía, tanto de manera inmediata como en el medio plazo.

Además, se resalta la alta correlación de las variables de producción eléctrica y exportación de cemento y equipos de construcción con las ventas de la compañía. Esto implica que el crecimiento en la producción de energía eléctrica y la exportación de materiales de construcción, como el cemento y equipos de construcción, también están fuertemente relacionados con las ventas de la empresa. Estos hallazgos indican que las ventas dependen en gran medida de la actividad en el sector de proyectos de construcción y el crecimiento del sector energético en el país.

Finalmente el análisis, que se expone en la figura 28, presenta las variables seleccionadas. Este mapa de calor proporciona una representación visual adicional de las correlaciones entre las variables restantes en el modelo. La figura 28 es de utilidad para evaluar y comprender las interacciones entre las variables seleccionadas, lo que contribuirá a una mejor comprensión de los factores que influyen en las ventas de la compañia. Este análisis es de gran importancia para identificar los impulsores clave de las ventas y respaldar la toma de decisiones estratégicas en el contexto de la predicción de ventas.

Además, el análisis de las correlaciones en el mapa de calor puede ayudar a identificar posibles patrones o tendencias que no son evidentes a simple vista. Al examinar la intensidad y dirección de las correlaciones entre las variables, es posible detectar relaciones causales y dependencias que son cruciales para desarrollar estrategias efectivas de marketing, gestión de inventario o cualquier otro aspecto relacionado con las ventas de la empresa. En última instancia, este análisis contribuirá a una toma de decisiones más informada y a la formulación de estrategias que maximicen el rendimiento y la rentabilidad de la compañía en el mercado.

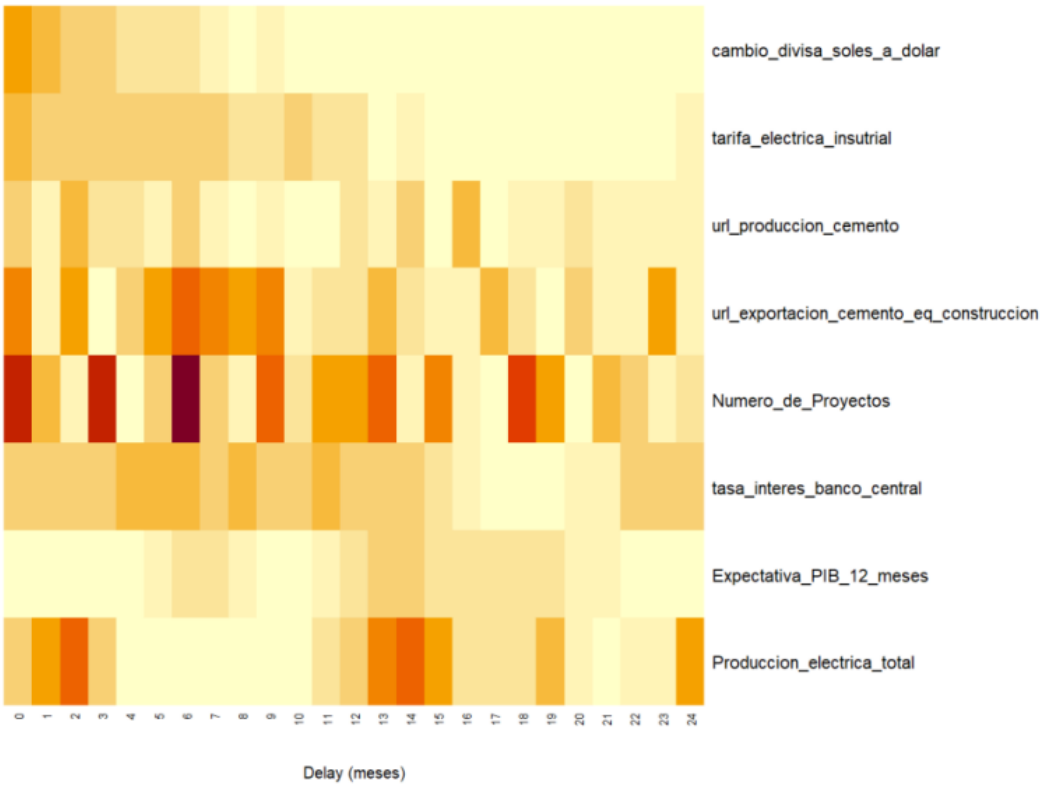


Figura 28: Mapa de calor de las variables seleccionadas.

4.3 Selección y evaluación de los modelos

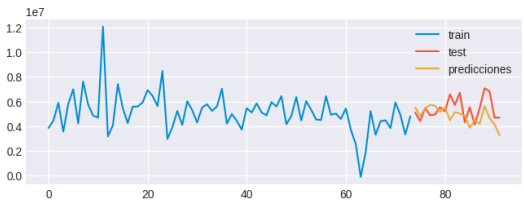
Esta sección presenta los resultados obtenidos a través de la implementación de varios modelos de pronóstico utilizados en el contexto de este trabajo de grado para predecir ventas. Los modelos que se han considerado son el modelo de pronóstico autorregresivo, los vectores autorregresivos, el modelo Prophet, XG Boost Regressor, LightGBM, Hist Gradient Boosting Regressor y PyAF. La evaluación de estos modelos se ha llevado a cabo teniendo en cuenta su precisión, que se mide a través de dos indicadores fundamentales: el Error Porcentual Absoluto Medio (MAPE) y la Raíz del Error Cuadrático Medio (RMSE).

Esta comparación en función de MAPE y RMSE permite discernir las variaciones en el desempeño de los diferentes modelos, permitiendo determinar cuál modelo se ajusta de manera más efectiva para predecir ventas. La presentación de estos resultados contribuirá significativamente a la selección del modelo de pronóstico más adecuado. La evaluación de estos modelos es importante para garantizar la precisión en las predicciones de ventas, ya que una elección adecuada puede tener un impacto significativo en la toma de decisiones empresariales. El eje X, representa los meses y el eje Y representa las ventas totales por periodo para la predicción.

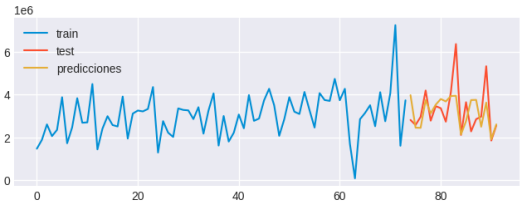
- **Prónostico autorregresivo (Multivariado):**

  La figura 34 y la tabla 5 muestran las predicciones, el MAPE y el RMSE para cada filtro. La tabla expone los valores de los errores para 12 meses, 3 meses y 1 mes que son

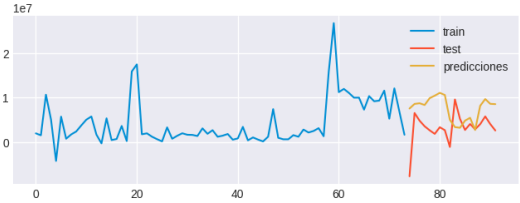
las ventanas temporales. Sin embargo, la gráfica solo se muestra para 12 meses, ya que involucraría los demás meses. El eje X, representa los meses y el eje Y representa las ventas totales por periodo.



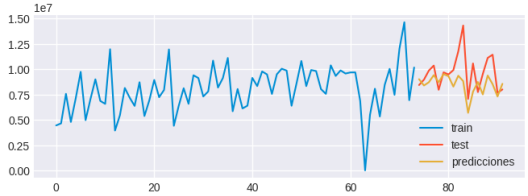
(a) Predicción de Power Products Division.



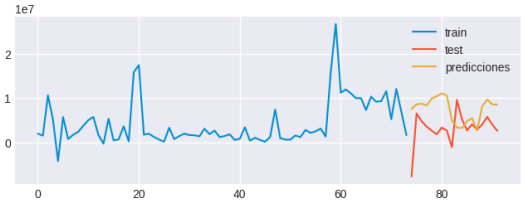
(b) Predicción de Home and Distribution.



(c) Predicción de Power Systems Division.



(d) Predicción de B2B Distribution.



(e) Predicción de End User.

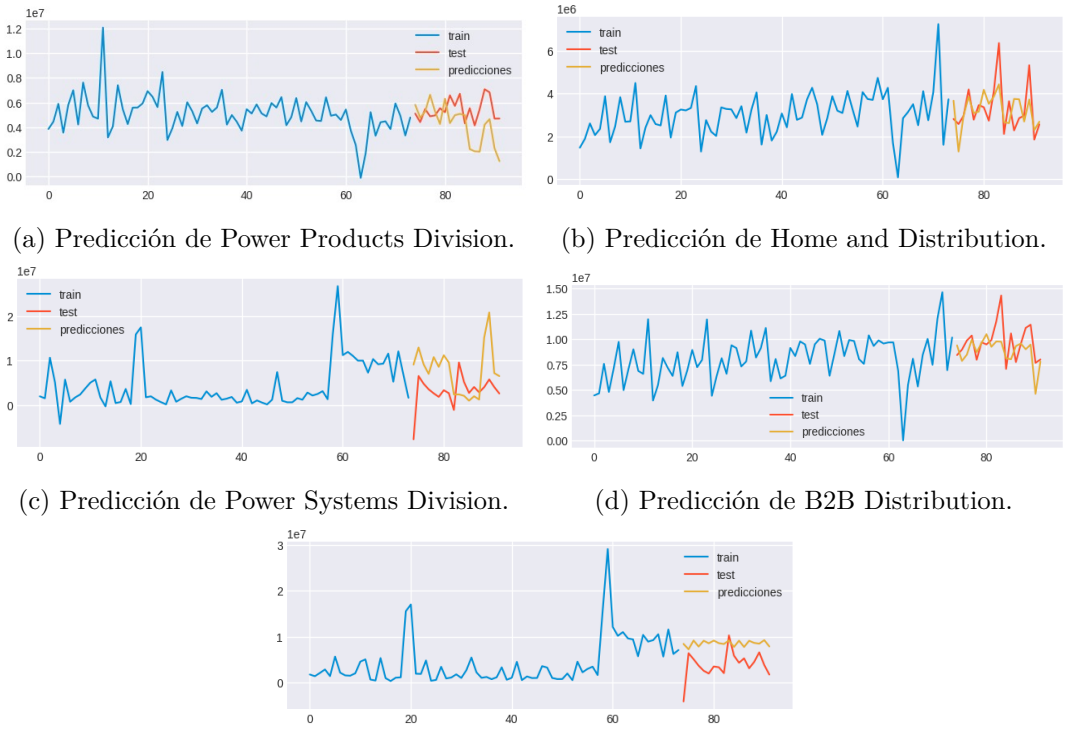
Figura 29: Predicción del modelo prónostico autorregresivo para cada filtro aplicado a los datos.

Filtro	**MAPE**			**RMSE**		
	12	3	1	12	3	1
Power products division	28.24 %	24.29 %	28.35 %	594,100	1'206,798	1'447,373
Home and distribution division	23.35 %	19.81 %	50.44 %	56,652	394,015	1'422,045
Power system division	179.87 %	119.7 %	201.87 %	4'917,983	7'947,190	15'228,811
B2B distribution	14.03 %	11.93 %	10.52 %	854,414	514,143	889,225
End user	150.09 %	129.95 %	297.91 %	4'491,540	579,845	12'509,551

Tabla 5: MAPE y RMSE modelo de pronostico autorregresivo para cada filtro aplicado a los datos para 12, 3 y 1 mes respectivamente.

- **XGBoost Regressor (Multivariado):**

  La figura 30 y la tabla 6 muestran las predicciones, el MAPE y el RMSE para cada unidad de negocio. La tabla expone los valores de los errores para 12 meses, 3 meses y 1 mes que son las ventanas temporales. Sin embargo, la grafica solo se muestra para 12 meses, ya que involucraría los demás meses.



(a) Predicción de Power Products Division.

(b) Predicción de Home and Distribution.

(c) Predicción de Power Systems Division.

(d) Predicción de B2B Distribution.

(e) Predicción de End User.

Figura 30: Predicción del modelo XGBoost Regressor para cada filtro aplicado a los datos.

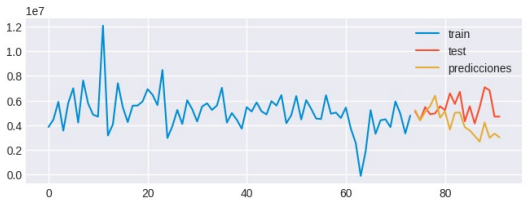
Filtro	**MAPE**			**RMSE**		
	12	3	1	12	3	1
Power products division	31.66 %	8.59 %	13.98 %	1'165,334	290,989	713,710
Home and distribution division	23.13 %	26.92 %	30.26 %	41,808	137,200	853,151

Power system division	167.71 %	135.62 %	220.85 %	4'502,494	9'123,536	16'662,539
B2B distribution	14.21 %	12.56 %	11.23 %	814,522	517,975	948,945
End user	154.94 %	131.59 %	300.34 %	4'671,063	5'854,355	12'621,678

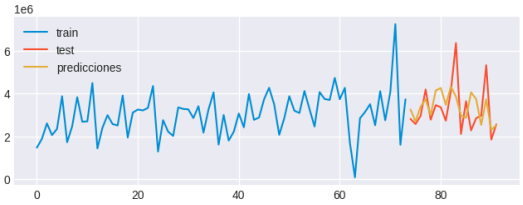
Tabla 6: MAPE y RMSE modelo XGBoost Regressor (multivariado) para cada filtro aplicado a los datos para 12, 3 y 1 mes respectivamente.

- **LightGBM (Multivariado):**

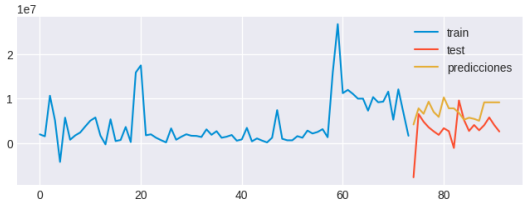
  La figura 31 y la tabla 7 muestran las predicciones, el MAPE y el RMSE para cada unidad de negocio. La tabla expone los valores de los errores para 12 meses, 3 meses y 1 mes que son las ventanas temporales. Sin embargo, la gráfica solo se muestra para 12 meses, ya que involucraría los demás meses.



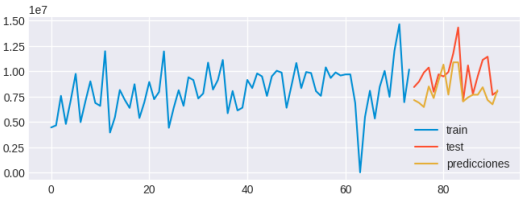
(a) Predicción de Power Products Division.



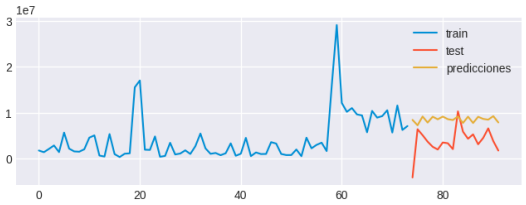
(b) Predicción de Home and Distribution.



(c) Predicción de Power Systems Division.



(d) Predicción de B2B Distribution.



(e) Predicción de End User.

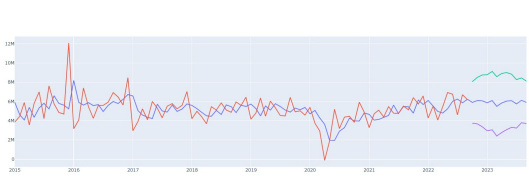
Figura 31: Predicción del modelo LightGBM para cada filtro aplicado a los datos.

Filtro	MAPE			RMSE		
	12	3	1	12	3	1
Power products division	24.29 %	3.34 %	2.15 %	1'143,683	105,443	109,538
Home and distribution division	22.98 %	11.63 %	15.6 %	107,118	330,988	439,747
Power system division	155.54 %	70.61 %	156.17 %	4'080,044	4'940,149	11'782,740
B2B distribution	16.48 %	24.24 %	15.19 %	1'562,732	2'252,798	1'283,436
End user	147.57 %	124.86 %	290.59 %	4'429,350	5'595,198	12'202,268

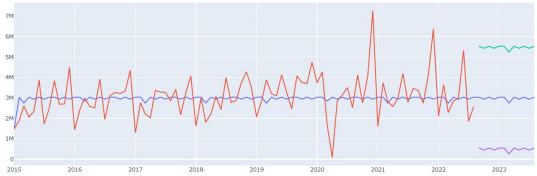
Tabla 7: MAPE y RMSE modelo LightGBM (multivariado) para cada filtro aplicado a los datos para 12, 3 y 1 mes respectivamente.

- **PyAF (Univariado):**

  La figura 32 y la tabla 8 muestran las predicciones, el MAPE y el RMSE para cada unidad de negocio. En este caso, solo se presentan los de 12 meses, debido a que la metodología utilizada para la predicción de l;os datos, limita el calculo de errores, para diferentes ventanas temporales.

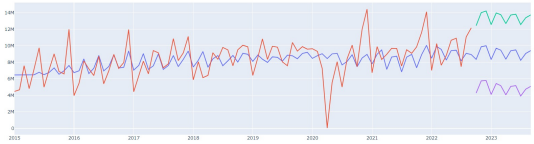


(a) Predicción de Power Products Division.



(b) Predicción de Home and Distribution.

(c) Predicción de Power Systems Division.



(d) Predicción de B2B Distribution.

(e) Predicción de End User.

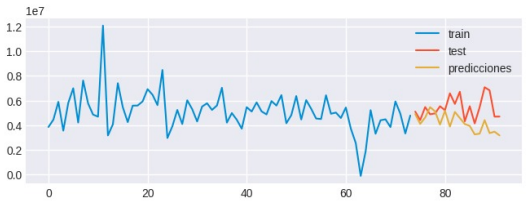
Figura 32: Predicción del modelo PyAF para cada filtro aplicado a los datos.

Filtro	MAPE	RMSE
Power Products Division	70 %	69,925
Home and Distribution Division:	61 %	104,287
Power Systems Division	196 %	463,123
B2B Distribution	294 %	836,117
End User	293 %	836,117

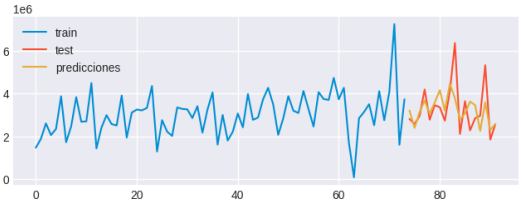
Tabla 8: MAPE y RMSE del modelo modelo PyAF para cada filtro aplicado a los datos.

- **Hist Gradient Boosting Regressor (Multivariado):**

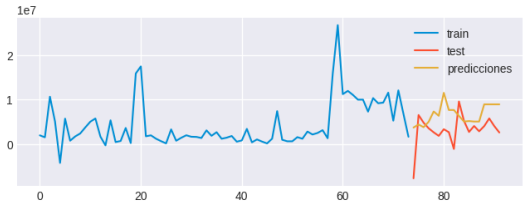
  La figura 33 y la tabla 9 muestran las predicciones, el MAPE y el RMSE para cada unidad de negocio. La tabla expone los valores de los errores para 12 meses, 3 meses y 1 mes que son las ventanas temporales. Sin embargo, la gráfica solo se muestra para 12 meses, ya que involucraría los demás meses.



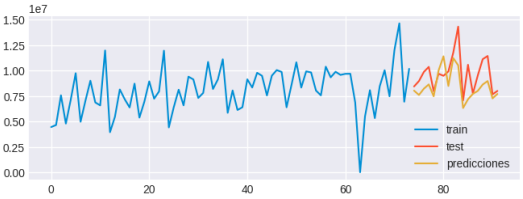
(a) Predicción de Power Products Division.



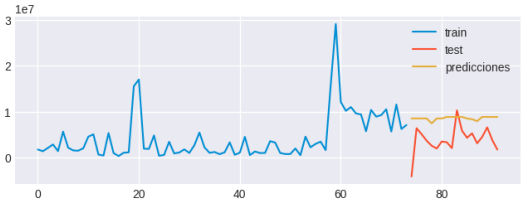
(b) Predicción de Home and Distribution.



(c) Predicción de Power Systems Division.



(d) Predicción de B2B Distribution.



(e) Predicción de End User.

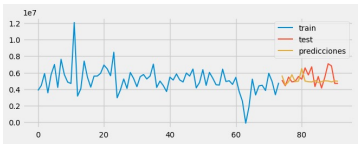
Figura 33: Predicción del modelo Hist Gradient Boosting Regressor para cada filtro aplicado a los datos.

Filtro	MAPE			RMSE		
	12	3	1	12	3	1
Power products division	21.94 %	8.81 %	3.91 %	1'186,481	452,681	199,530
Home and distribution division	19.5 %	8.92 %	13.82 %	34,288	130,115	389,530
Power system division	149.25 %	67.64 %	150.15 %	3'440,429	2'735,906	11'327,199
B2B distribution	13.33 %	12.21 %	4.81 %	1'138,124	1'138,223	406,493
End user	155.64 %	134.69 %	302.44 %	4'702,708	6'092,647	12'699,717

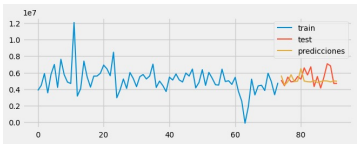
Tabla 9: MAPE y RMSE modelo Hist Gradient Boosting Regressor (Multivariado) para cada filtro aplicado a los datos para 12, 3 y 1 mes respectivamente.

- **Prónostico autorregresivo (Univariado):**

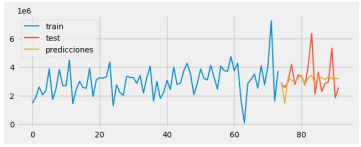
  La figura 34 y la tabla 10 muestran las predicciones, el MAPE y el RMSE para cada filtro. La tabla expone los valores de los errores para 12 meses, 3 meses y 1 mes que son las ventanas temporales. Sin embargo, la gráfica solo se muestra para 12 meses, ya que involucraría los demás meses.



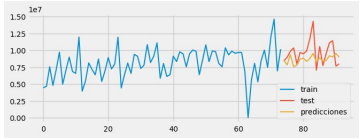
(a) Predicción de Power Products Division.



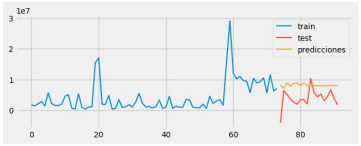
(b) Predicción de Home and Distribution.



(c) Predicción de Power Systems Division.



(d) Predicción de B2B Distribution.



(e) Predicción de End User.

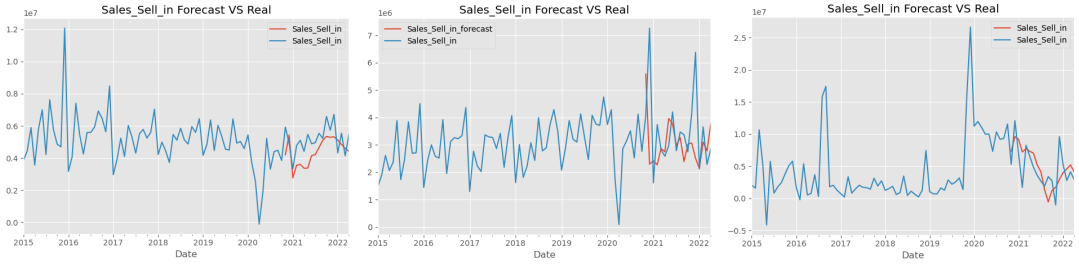
Figura 34: Predicción del modelo prónostico autorregresivo para cada filtro aplicado a los datos.

Filtro	MAPE			RMSE		
	12	3	1	12	3	1
Power products division	14.83 %	12.06 %	6.68 %	454,688	139,295	340,885
Home and distribution division	24.13 %	26.24 %	21.93 %	368,554	580,121	618,371
Power system division	80.26 %	72.15 %	180.44 %	237,606	3'758,649	13'613,445
B2B distribution	18.39 %	20.53 %	2.2 %	6,710	1'791,706	18,116
End user	146.62 %	127.33 %	294.65 %	4'277,972	5'649,087	12'372,805

Tabla 10: MAPE y RMSE modelo Pronostico autorregresivo para cada filtro aplicado a los datos para 12, 3 y 1 mes respectivamente.

- **Vectores Autorregresivos (Multivariado):**

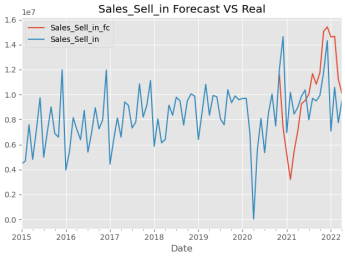
  La figura 35 y la tabla 11 muestran las predicciones, el MAPE y el RMSE para cada unidad de negocio. La tabla expone los valores de los errores para 12 meses, 3 meses y 1 mes que son las ventanas temporales. Sin embargo, la grafica solo se muestra para 12 meses, ya que involucraría los demás meses.



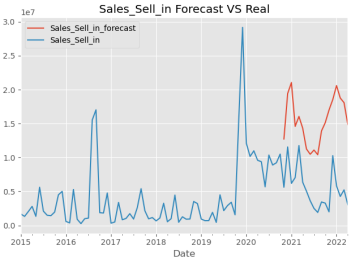
(a) Predicción de Power Products Division.

(b) Predicción de Home and Distribution.

(c) Predicción de Power Systems Division.



(d) Predicción de B2B Distribution.



(e) Predicción de End User.

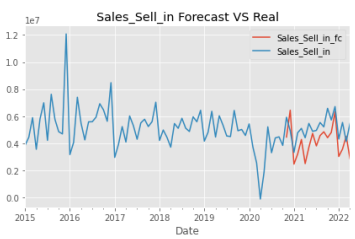
Figura 35: Predicción del modelo vectores Autorregresivos para cada filtro aplicado a los datos.

Filtro	MAPE			RMSE		
	12	3	1	12	3	1
Power products division	18.34 %	18.8 %	29.6 %	1'075,447	1'087,697	1'723,156
Home and distribution division	25.18 %	51.72 %	36.95 %	1'656,879	3'029,210	1'503,922
Power system division	70.57 %	30.8 %	38.6 %	2'725,902	2'277,197	2'048,216
B2B distribution	29.3 %	26.24 %	3.6 %	3'571,381	4'342,428	432,895
End user	250.39 %	144.49 %	12.6 %	10'481,090	10'533,191	7'101,742

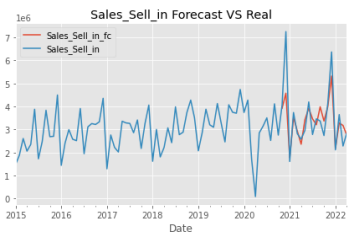
Tabla 11: MAPE y RMSE modelo Vectores Autorregresivos para cada filtro aplicado a los datos para 12, 3 y 1 mes respectivamente.

- **Prophet (Univariado):**

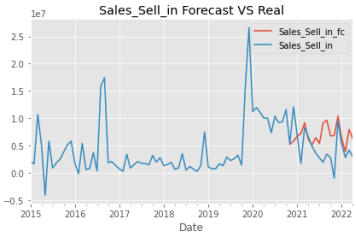
  La figura 36 y la tabla 12 muestran las predicciones, el MAPE y el RMSE para cada unidad de negocio. La tabla expone los valores de los errores para 12 meses, 3 meses y 1 mes que son las ventanas temporales. Sin embargo, la gráfica solo se muestra para 12 meses, ya que involucraría los demás meses.



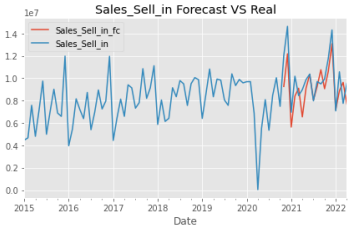
(a) Predicción de Power Products Division.



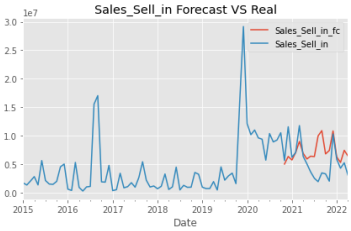
(b) Predicción de Home and Distribution.



(c) Predicción de Power Systems Division.



(d) Predicción de B2B Distribution.



(e) Predicción de End User.

Figura 36: Predicción del modelo Prophet para cada filtro aplicado a los datos.

Filtro	**MAPE**			**RMSE**		
	12	3	1	12	3	1
Power products division	24.03 %	27.14 %	24.7 %	1'404,246	1'316,148	1'464,256
Home and distribution division	13.33 %	16.91 %	3.63 %	780,994	1'550,551	147,771
Power system division	129.21 %	18.95 %	1.89 %	3'949,317	3'636,734	100,146
B2B distribution	12.54 %	19.59 %	22.98 %	1'504,382	22'674,408	2'764,501
End user	85.29 %	21.29 %	11.02 %	3'630,047	3'044,427	620,891

Tabla 12: MAPE y RMSE modelo Prophet para cada filtro aplicado a los datos para 12, 3 y 1 mes respectivamente.

4.4 Desarrollo de la aplicación web de visualización

Esta sección presenta los resultados obtenidos al desarrollar la aplicación web como una herramienta para la visualización de las predicciones de ventas. El propósito del desarrollo de la aplicación se centra en ofrecer una experiencia de usuario sofisticada y completa. Los usuarios son recibidos en el proceso de autenticación en la página de inicio de sesión, lo cual garantiza el acceso personalizado a la aplicación. Una vez que los usuarios ingresan, son dirigidos a la página principal, donde pueden acceder a la funcionalidad de visualizar las predicciones de ventas. Además, la opción de 'Próximamente' proporciona una primera aproximación a la escalabilidad de la aplicación, donde se incluirán futuras características que se agregarán a la aplicación, manteniéndola en constante evolución. Las figuras 37 y 38 presentan el login y pagina principal de la aplicación.

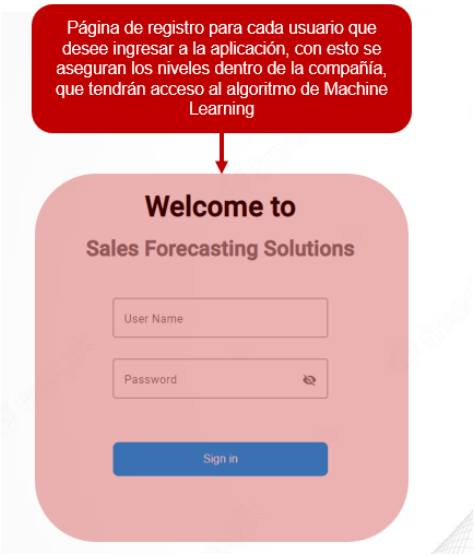


Figura 37: Log In de la aplicación web.

Figura 38: Pagina Principal de la aplicación web.

La funcionalidad de visualizar las predicciones de ventas incluye un tablero de control, donde los usuarios pueden monitorear indicadores clave, filtrar por unidades de negocio y

visualizar las series temporales con el histórico de las ventas y la predicción de las ventas, comparadas con el histórico, utilizando dos colores: verde y azul. El valor del histórico está dividido en histórico total, último año, último mes, y las proyecciones de las ventas están divididas en pestañas de corto, mediano y largo plazo. De la misma manera, la aplicación incluye un filtro que se muestra a través de un menú desplegable, el cual permite seleccionar cada unidad de negocio para actualizar los datos de los indicadores y de la serie temporal. Este tablero brinda una visión profunda tanto de las ventas pasadas como de las proyecciones futuras, proporcionando una herramienta de útil para apoyar la toma de decisiones basada en los datos del contexto empresarial. Las figuras 39, 40 y 41, exponen las partes del tablero de conntrol descritas con anterioridad.

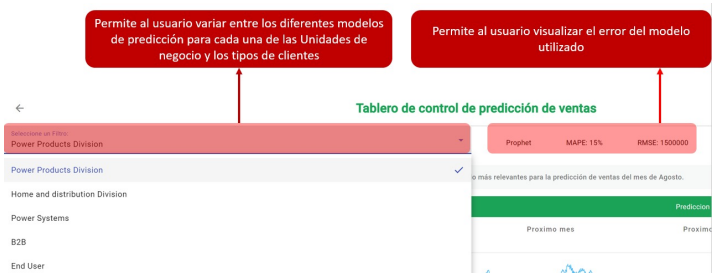


Figura 39: Filtros del tablero de control.

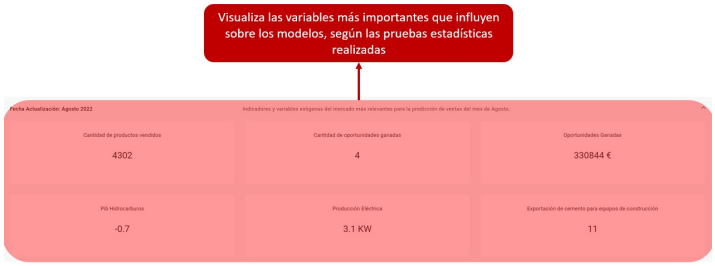


Figura 40: Variables e indicadores del tablero de control.

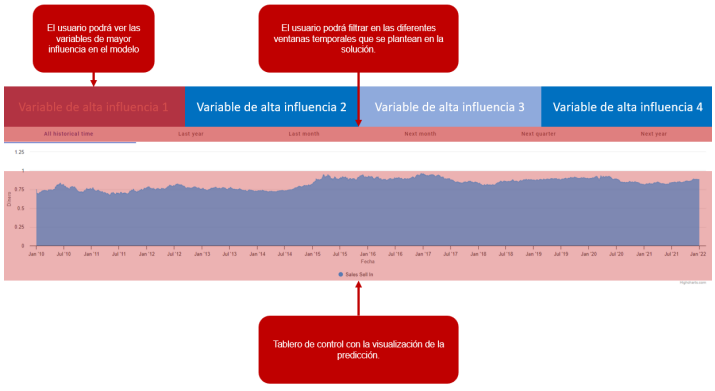


Figura 41: Visualización de la serie temporal del tablero de control.

La arquitectura de la aplicación se basa en el framework Angular y se compone de varios módulos y componentes. El módulo principal, 'AppModule', se encarga de coordinar y ensamblar los diferentes elementos de la aplicación. Hace uso de módulos de Angular como 'BrowserModule' para la funcionalidad principal de la aplicación, 'HttpClientModule' para la comunicación con servicios web, 'AppRoutingModule' para el enrutamiento de páginas, y 'HighchartsChartModule' para la integración de gráficos interactivos. Además, la aplicación hace uso de módulos de Material Design, como 'MatTabsModule', 'MatDialogModule', 'MatExpansionModule', 'MatInputModule', 'MatFormFieldModule', 'MatTooltipModule', y 'MatIconModule', para garantizar una interfaz de usuario moderna y atractiva.

Los componentes desempeñan un papel crucial en la arquitectura. 'AppComponent' actúa como el componente raíz de la aplicación, mientras que otros componentes como 'TimeSeriesComponent', 'LogInComponent', 'HomePageComponent', y otros, se encargan de gestionar las diferentes vistas y funcionalidades de la aplicación, como la visualización de series temporales, el inicio de sesión, y la página principal. Esta arquitectura modular y componentizada permite una organización eficiente y una comunicación fluida entre los elementos de la aplicación, garantizando una experiencia de usuario atractiva y funcional.

Capítulo 5

DISCUSIONES

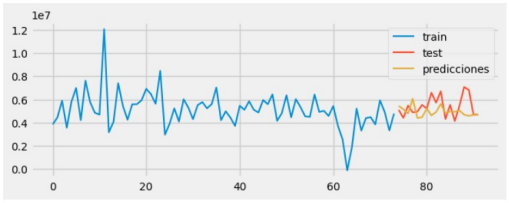
En este capítulo, se presenta la discusión de los resultados más relevantes obtenidos y expuestos en el capítulo 4. Inicialmente se presenta la discusión de los modelos, para poder comparar el rendimiento de los modelos se tiene en cuenta las medidas de error MAPE y RMSE; adicionalmente a estas medidas también es necesario tener en cuenta el comportamiento de la predicción observando las gráficas, es decir, a pesar de que una predicción pudo haber tenido una peor calificación a nivel de las medidas de error, es posible que a través de la gráfica se pueda observar como la predicción logra replicar el comportamiento de los valores reales a pesar de que los valores del pronóstico no sean acertados. A partir de lo anterior para cada uno de los filtros se presentan en la tabla 13, los mejores modelos con base en el MAPE y RMSE.

Filtro	**Modelo**	**MAPE**	**RMSE**
Power products division	Pronostico autorregresivo (univariado)	15 %	454,688
Home and distribution division	Prophet (univariado)	13 %	780,994
Power system division	VAR (Multivariado)	70 %	2'725,902
B2B distribution	Prophet (univariado)	13 %	1'504,381
End user	Prophet (univariado)	156 %	3'630.470

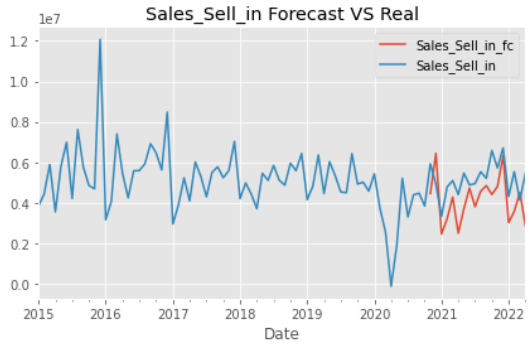
Tabla 13: Selección del mejor modelo para cada filtro a partir de las medidas de error.

De la tabla anterior, el modelo prophet es el seleccionado para predecir 3 de los 5 filtros de las series de tiempo de ventas, pero al comparar el comportamiento de las predicciones del pronostico autorregresivo y VAR con respecto a las predicciones generadas por prohet en la figura 42, se puede observar que las predicciones del modelo prophet logran imitar en mayor medida el comportamiento de los datos reales, siguiendo los cambios abruptos entre periodos,

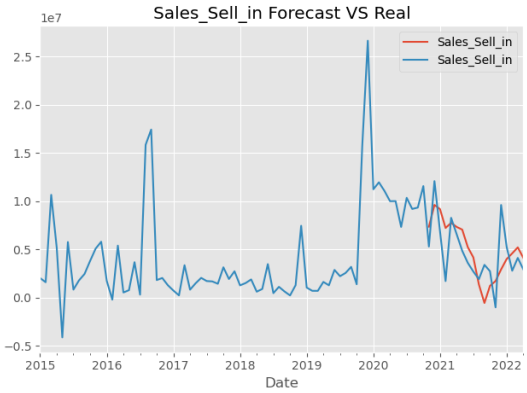
mientras que los demás modelos solo generan una tendencia general de los datos a lo largo del tiempo.



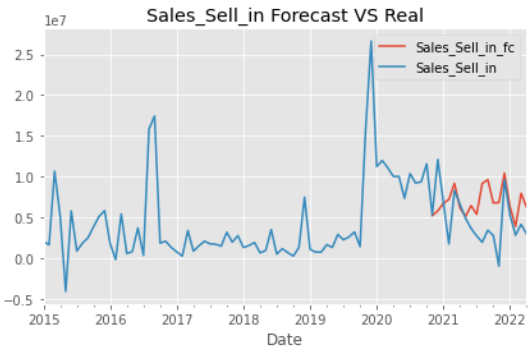
(a) Power Products Division - Pronóstico autorregresivo.



(b) Power Products Division - Prophet.



(c) Power Systems Division - VAR.



(d) Power Systems Division - Prophet

Figura 42: Comparación del comportamiento de las predicciones para los modelos prophet, VAR y pronóstico autorregresivo.

Por otra parte, en la actual era de la analítica de datos y la toma de decisiones comerciales, la elección entre desarrollar una aplicación web personalizada para la predicción de ventas y utilizar herramientas de tablero de control establecidas, como Power BI o Tableau, es una cuestión crítica para las organizaciones. Este debate gira en torno a encontrar el equilibrio entre la personalización y flexibilidad que brinda una aplicación web a medida, frente a la conveniencia y la orientación hacia la visualización de datos que ofrecen las soluciones de tablero de control preexistentes [89].

Por un lado, una aplicación web personalizada ofrece un control total sobre la funcionalidad, la interfaz de usuario y la lógica de procesamiento de datos, permitiendo adaptarla a las necesidades específicas de una empresa. Esto implica la capacidad de combinar datos de diversas fuentes y la integración de algoritmos de predicción personalizados. Además, su escalabilidad y la posibilidad de integración con sistemas empresariales crean un flujo de trabajo eficiente y una mayor seguridad de datos [90].

Sin embargo, es importante reconocer que este enfoque conlleva un costo y tiempo de desarrollo más elevados, junto con la necesidad de un equipo técnico altamente capacitado. Además, la aplicación web requiere mantenimiento constante y actualizaciones de seguridad

a lo largo del tiempo. Por otro lado, las herramientas de tablero de control se destacan por su enfoque en la visualización de datos y la creación de informes interactivos, lo que puede ser más adecuado si la prioridad principal es la presentación de datos en lugar de la automatización de predicciones [89] [90].

En última instancia, la elección entre estos enfoques dependerá de la situación particular de cada empresa, sus recursos disponibles y sus objetivos comerciales. La personalización y flexibilidad de una aplicación web son invaluables para organizaciones con necesidades analíticas complejas, mientras que las herramientas de tablero de control ofrecen una solución más rápida y accesible para aquellos centrados en la visualización y presentación de datos. Esta decisión debe tomarse cuidadosamente, evaluando las ventajas y desventajas de cada enfoque, para asegurar una alineación óptima con los objetivos de la empresa.

Por lo tanto, para este trabajo de grado la decisión de desarrollar una aplicación web personalizada se basa en una estrategia que pone en el centro la propiedad intelectual de los desarrolladores del trabajo de grado. Esto permite evitar costos elevados asociados con licencias de tableros de control existentes, al tiempo que garantiza una personalización total para satisfacer las necesidades específicas de cada cliente. La aplicación web ofrece un control completo sobre la estética, los colores y los detalles, además de brindar escalabilidad y la capacidad de incorporar funcionalidades adicionales a los tableros de control tradicionales.

En última instancia, al desarrollar esta aplicación, se ha creado un valioso activo que no solo satisface las demandas únicas de los clientes, sino que se adapta a las cambiantes necesidades comerciales y permite la expansión futura sin depender de terceros. Esto resulta en una solución poderosa y adaptable que protege la propiedad intelectual y fortalece la capacidad de ofrecer análisis de ventas de vanguardia a los clientes.

Capítulo 6

CONCLUSIONES

En este capitulo se exponen las conclusiones relacionadas al trabajo de grado. La selección del modelo de predicción de ventas ha sido un proceso meticuloso y fundamentado en la evaluación de dos factores cruciales. En primera instancia, se realizó un análisis exhaustivo para identificar aquel modelo que generara la menor medida de error. Esta métrica, siendo un indicador objetivo de la precisión del modelo, se selecciona como un criterio esencial para garantizar la fiabilidad de las proyecciones.

En segundo lugar, se consideró esencial incorporar un enfoque visual a través de la observación de las gráficas de predicción. Este segundo factor permitió evaluar no solo la precisión cuantitativa, sino también la calidad cualitativa de las predicciones. Se buscó específicamente que el comportamiento de las predicciones se asemejara de manera óptima al comportamiento de los datos reales, prestando especial atención a la capacidad del modelo para capturar y reflejar cambios abruptos entre periodos. Es decir, se evaluó la capacidad del modelo para predecir las tendencias estacionarias de los datos reales.

El análisis conjunto de estos dos factores proporcionó una base robusta para la elección del modelo de predicción de ventas. La combinación de la precisión cuantitativa y la capacidad del modelo para adaptarse a los cambios bruscos en los datos brinda confianza en la fiabilidad de las proyecciones. En última instancia, este enfoque integral no solo busca la exactitud numérica, sino también la capacidad del modelo para capturar la complejidad y la dinámica inherentes al comportamiento de las ventas. Este proceso de selección garantiza que el modelo elegido no solo sea una herramienta precisa, sino también una representación fiel y adaptable del entorno comercial en estudio, dando como resultado la selección del modelo prophet para Home and distribution Division, B2B distribution y End User; el Pronóstico autoregresivo (univariado) para Power products division, y finalmente, VAR (Multivariado) para Power system divison.

Se puede concluir, cómo los modelos univariados representan de mejor manera aquellas BU y Customer Type que realizan sus ventas a partir de proyectos, es decir su tendencia es las ventas no es estacionaria y se presentan muchos picos. Esto puede deberse a que el modelo de pronóstico univariado puede ser más efectivo para series de tiempo inestables cuando no hay influencias externas o variables relacionadas significativas que se están usando en el modelo para predecir estos proyectos inusuales que generan estos picos en las ventas. Pero por el contrario, se puede ver como un modelo multivariado, predice con mayor exactitud las ventas en aquellos casos donde las ventas tienen un comportamiento estacionario.

Además, se destaca la importancia de las variables exógenas en el proceso de pronóstico.

Al incorporar información externa relevante, como datos económicos y datos importantes en el sector energetico, se logró mejorar aún más la precisión de los modelos, para los casos de Power Products, Home and distribution y B2B distribution debido a su comportamiento estacionario.Los modelos multivariados, como VAR, XGBoost Regressor, LightGBM y Hist Gradient Boosting Regressor, mostraron un alto rendimiento.

A partir del análisis de los resultados de la predicción de ventas a corto (1 mes) y largo (12 meses) se nota que los modelos con mejor predicción son aquellos univariados y a mediano (3 meses). Se concluye que es dado a que a mediano plazo las variables exógenas están aportando de gran manera a la predicción de las ventas, pero a largo y corto plazo este efecto desaparece. Se nota como Prophet es el modelo que en la mayor parte de los casos está realizando una predicción más acertada a los datos, según los criterios de selección del mejor modelo. Es por esto, que Prophet es el modelo escogido para visualizar en la aplicación web. Los criterios de selección del modelo que apoyaron la selección de los modelos fueron los errores, se seleccionaron aquellos modelos que mostraran un menor error como criterio clave para fiabilidad en las proyecciones y se comparo el comportamiento de las predicciones con el comportamiento de los valores reales. A partir de las pruebas para la selección de variables, se puede notar como no todas las variables seleccionadas en las entrevistas tienen una alta correlación. El presentimiento de los involucrados realmente no está tan acercado a la realidad.

Por otra parte, con respecto al desarrollo de la aplicación web para presentar la predicción de ventas, representa una herramienta significativa en el ámbito de la analítica de datos y la toma de decisiones comerciales. Debido a su enfoque en la propiedad intelectual y la capacidad de adaptación a las necesidades específicas de cada cliente, se ha logrado evitar los costos asociados con las licencias de tableros de control preexistentes. La aplicación proporciona un control completo sobre la estética, la funcionalidad y la lógica de procesamiento de datos, lo que resulta en escalabilidad y la capacidad de integrar características adicionales en el futuro basadas en las necesidades de los clientes.

La experiencia del usuario se ha basado en la implementación de una página de inicio de sesión que garantiza un acceso personalizado a la aplicación. Una vez que los usuarios ingresan, son dirigidos a la página principal, desde donde pueden acceder a una variedad de funcionalidades para visualizar las predicciones de ventas. Además, la posibilidad de agregar nuevas funcionalidades a futuro, ofrece una visión de la escalabilidad de la aplicación y las características futuras que se pueden agregar, manteniéndola en constante evolución.

Con respecto a la funcionalidad de visualización de la predicción de ventas a través del tablero de control incluido en la aplicación, permite a los usuarios monitorear indicadores clave y examinar series temporales de ventas en pestañas que abarcan desde el corto hasta el largo plazo. lo cual proporciona una visión profunda de las ventas pasadas y las proyecciones futuras, lo que se traduce en una herramienta valiosa para respaldar la toma de decisiones basada en datos en el contexto empresarial.

En última instancia, la aplicación web representa una herramienta estratégica valiosa, ya que no solo cumple con las necesidades de los clientes, sino que también se adapta a los cambios comerciales y asegura la no dependencia de terceros. Esto permite ofrecer análisis de ventas protegiendo la propiedad intelectual de la herramienta.

El desarrollo de una aplicación web, permitió una personalización del contenido según la necesidad del usuario. La retroalimentación final, permite que se realicen ajustes de manera rápida en la interfaz. La respuesta del ingreso a la página, los filtros, y visualización de los valores, es rápida en comparación a otras herramientas de visualización. Las respuestas en la

aplicación web se dan en menos de un segundo, mientras en otras herramientas el tiempo de espera es superior del segundo.

La aplicación web permite visualizar de manera rápida y sencilla la información requerida por los presidentes de país, gerentes de venta y comerciales para planificar sus ventas. Se logra una reducción de las horas hombre invertidas adicional a una información disponible 25/7, sin errores y en un mismo estándar. Como conclusión, el objetivo de desarrollar una solución de analítica predictiva de ventas se logró, estimando las ventas de la compañía piloto a corto, mediano y largo plazo. Además de encontrar que este modelo es un modelo escalable para otras soluciones, siempre y cuando se cumplan las limitantes en la entrega de los datos.

Capítulo 7

RECOMENDACIONES Y TRABAJOS FUTUROS

El presente trabajo de grado expone un conjunto de recomendaciones y trabajos futuros que tienen el potencial de llevar esta investigación hacia un nivel de mayor profundidad en la predicción de ventas. En primer lugar, se sugiere una expansión significativa de la base de datos utilizada, con un enfoque tanto en la cantidad como en la calidad de los datos. El aumento en la cantidad de datos permitirá explorar una diversidad más amplia de modelos de machine learning, mientras que la mejora en la calidad de los datos garantizará la confiabilidad de las predicciones de ventas. Esta iniciativa puede resultar en un desempeño más preciso y robusto en la predicción de ventas, un factor crítico para la toma de decisiones estratégicas basadas en la explotación de los datos.

Por otra parte, se sugiere continuar con el mejoramiento de la la funcionalidad de la aplicación web. Enfocándose, en la implementación de nuevas características que enriquezcan la experiencia del usuario, tales como la capacidad de subir y descargar datos de manera sencilla, la posibilidad de descargar reportes del tablero de control con sus indicadores, la incorporación de un panel de control más intuitivo y accesible, y la integración de herramientas de visualización avanzadas. De la misma manera, enfocarse en la seguridad de la aplicación web también debe ser una prioridad; la inclusión de protocolos de seguridad robustos y medidas de protección de los datos garantizará la confidencialidad y la integridad de la información.

Por otra parte, un enfoque de gran importancia es la inversión en la capacitación de los empleados de las empresa en temas de análisis de datos y técnicas de inteligencia de negocios. Esto no solo fortalecerá la comprensión y el uso efectivo de las herramientas de análisis, como la que se plantea para este trabajo de grado, sino que también fomentará una cultura empresarial orientada hacia la analítica y la toma de decisiones basadas en datos. Además, este enfoque en el desarrollo de capacidades internas abrirá nuevas oportunidades para la empresa en el ámbito de la creación, mejoramiento y uso de productos digitales orientados a la predicción de ventas, lo que podría diversificar sus fuentes de ingresos y ampliar su presencia en el mercado.

En resumen, estas recomendaciones y trabajos futuros representan una visión integral para el corto, mediano y largo plazo. Ofrecen un camino claro hacia la mejora continua de la precisión y el valor de las predicciones de ventas, al tiempo que fortalecen la posición de las empresas en el mercado a través de una aplicación web más eficiente, una mayor inversión en

el desarrollo de habilidades analíticas internas y una mayor diversificación de las ofertas de productos basados en datos.

Bibliografía

[1] X. Yu, Z. Qi e Y. Zhao, «Support vector regression for newspaper/magazine sales forecasting,» *Procedia Computer Science*, vol. 17, págs. 1055-1062, 2013.

[2] S. Hanifi, X. Liu, Z. Lin y S. Lotfian, «A critical review of wind power forecasting methods—past, present and future,» *Energies*, vol. 13, n.º 15, pág. 3764, 2020.

[3] D. Khazanchi y S. G. Sutton, «Assurance services for business-to-business electronic commerce: a framework and implications,» *Journal of the Association for Information Systems*, vol. 1, n.º 1, pág. 1, 2000.

[4] Z. Qi, Y. Tian e Y. Shi, «Robust twin support vector machine for pattern classification,» *Pattern recognition*, vol. 46, n.º 1, págs. 305-316, 2013.

[5] Z. Qi, Y. Tian, Y. Shi y X. Yu, «Cost-sensitive support vector machine for semi-supervised learning,» *Procedia Computer Science*, vol. 18, págs. 1684-1689, 2013.

[6] Y. Qi, C. Li, H. Deng, M. Cai, Y. Qi e Y. Deng, «A deep neural framework for sales forecasting in e-commerce,» en *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019, págs. 299-308.

[7] Z. Li, X. Zhong y Z. Cui, «Evaluating forecasting algorithm of realistic datasets based on machine learning,» en *Proceedings of the 2nd International Conference on Innovation in Artificial Intelligence*, 2018, págs. 72-76.

[8] M. Schnegg y K. Möller, «Strategies for data analytics projects in business performance forecasting: a field study,» *Journal of Management Control*, vol. 33, n.º 2, págs. 241-271, 2022.

[9] D. A. A. Shaikh, M. A. Kumar, D. A. A. Syed y M. Z. Shaikh, «A two-decade literature review on challenges faced by SMEs in technology adoption,» *Academy of Marketing Studies Journal*, vol. 25, n.º 3, 2021.

[10] A. Perdana, H. H. Lee, S. Koh y D. Arisandi, «Data analytics in small and mid-size enterprises: Enablers and inhibitors for business value and firm performance,» *International Journal of Accounting Information Systems*, vol. 44, pág. 100 547, 2022.

[11] A. Perdana, H. H. Lee, D. Arisandi y S. Koh, «Accelerating data analytics adoption in small and mid-size enterprises: A Singapore context,» *Technology in Society*, vol. 69, pág. 101 966, 2022.

[12] S. Lodemann, S. Lechtenberg, K. Wesendrup, B. Hellingrath, K. Hoberg y W. Kersten, «Supply Chain Analytics: Investigating Literature-Practice Perspectives and Research Opportunities,» *Logistics Research*, vol. 15, n.º 1, 2022.

[13] R. Simon y M. T. Suarez, «Examining the Behavioral Intention of Philippine MSMEs Toward Business Intelligence Adoption.,» *Journal of Business & Management*, vol. 28, n.º 1, 2022.

[14] X. Fu y H. Asorey, «Data-driven product innovation,» en *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, págs. 2311-2312.

[15] S. Pandian. «Time Series Analysis and Forecasting | Data-Driven Insights.» Obtenido de Analytics Vidhya. (oct. de 2021).

[16] B. Lim y S. Zohren, «Time-series forecasting with deep learning: a survey,» *Philosophical Transactions of the Royal Society A*, vol. 379, n.º 2194, pág. 20 200 209, 2021.

[17] O. B. Sezer, M. U. Gudelek y A. M. Ozbayoglu, «Financial time series forecasting with deep learning: A systematic literature review: 2005–2019,» *Applied soft computing*, vol. 90, pág. 106 181, 2020.

[18] C. Faloutsos, V. Flunkert, J. Gasthaus, T. Januschowski e Y. Wang, «Forecasting big time series: Theory and practice,» en *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, págs. 3209-3210.

[19] L. Dannecker, R. Lorenz, P. Rösch, W. Lehner y G. Hackenbroich, «Efficient forecasting for hierarchical time series,» en *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, 2013, págs. 2399-2404.

[20] Universidad de Valladolid, *Componentes de una serie temporal*, `http://www5.uva.es/estadmed/datos/series/series1.htm`, Consultado el 23 de agosto de 2023.

[21] F. J. P. Rodríguez, *Estadística y Machine Learning con R: Ejercicios resueltos con R (Spanish Edition)*, Español, Edición en Español. Editorial, 2017, pág. 288, ISBN: 978-6202252164.

[22] A. Bajaj, «Time Series Prediction: How Is It Different From Other Machine Learning?» *neptune.ai*, abr. de 2023.

[23] J. Brownlee, «What Is Time Series Forecasting?» *Time Series*, dic. de 2016, Last Updated on August 15, 2020.

[24] W. Polasek, *Time series analysis and its applications: With r examples*, 2013.

[25] I. P. Androulakis, E. Yang y R. R. Almon, «Analysis of time-series gene expression data: methods, challenges, and opportunities,» *Annu. Rev. Biomed. Eng.*, vol. 9, págs. 205-228, 2007.

[26] S. Aminikhanghahi y D. J. Cook, «A survey of methods for time series change point detection,» *Knowledge and information systems*, vol. 51, n.º 2, págs. 339-367, 2017.

[27] F. Petropoulos, D. Apiletti, V. Assimakopoulos et al., «Forecasting: theory and practice,» *International Journal of Forecasting*, vol. 38, n.º 3, págs. 705-871, 2022.

[28] S. J. Taylor y B. Letham, «Forecasting at scale,» *The American Statistician*, vol. 72, n.º 1, págs. 37-45, 2018.

[29] L. Kurzak et al., «Importance of forecasting in enterprise management,» *Advanced Logistic Systems*, vol. 6, n.º 1, págs. 173-182, 2012.

[30] R. Fildes, S. Bretschneider, F. Collopy et al., «Researching sales forecasting practice: Commentaries and authors' response on "Conducting a Sales Forecasting Audit" by MA Moon, JT Mentzer & CD Smith,» *International Journal of Forecasting*, vol. 19, n.º 1, págs. 27-42, 2003.

[31] D. Rogers, «A review of sales forecasting models most commonly applied in retail site evaluation,» *International Journal of Retail & Distribution Management*, vol. 20, n.º 4, 1992.

[32] BLOOMBERG. «Meta cae mientras pronóstico de ventas muestra debilidad del mercado de publicidad.» (oct. de 2022).

[33] J. R. Trapero, N. Kourentzes y R. Fildes, «On the identification of sales forecasting models in the presence of promotions,» *Journal of the operational Research Society*, vol. 66, n.º 2, págs. 299-307, 2015.

[34] C.-W. Chu y G. P. Zhang, «A comparative study of linear and nonlinear models for aggregate retail sales forecasting,» *International Journal of production economics*, vol. 86, n.º 3, págs. 217-231, 2003.

[35] Freshworks, *Una guía completa sobre pronóstico de ventas*, URL: `https://www.freshworks.com/latam/crm/sales/pronostico-de-ventas/`, Accedido el 28 de agosto de 2023, 2023.

[36] B. M. Pavlyshenko, «Machine-learning models for sales time series forecasting,» *Data*, vol. 4, n.º 1, pág. 15, 2019.

[37] Z. Shilong et al., «Machine learning model for sales forecasting by using XGBoost,» en *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, IEEE, 2021, págs. 480-483.

[38] P. X. Tan, T. N. Duc, C. M. Tran y E. Kamioka, «Continuous QoE prediction based on WaveNet,» en *Proceedings of the 2020 12th International Conference on Computer and Automation Engineering*, 2020, págs. 80-84.

[39] G. Kechyn, L. Yu, Y. Zang y S. Kechyn, «Sales forecasting using WaveNet within the framework of the Kaggle competition,» *arXiv preprint arXiv:1803.04037*, 2018.

[40] D. Salinas, V. Flunkert, J. Gasthaus y T. Januschowski, «DeepAR: Probabilistic forecasting with autoregressive recurrent networks,» *International Journal of Forecasting*, vol. 36, n.º 3, págs. 1181-1191, 2020.

[41] W. B. Nicholson, I. Wilms, J. Bien y D. S. Matteson, «High dimensional forecasting via interpretable vector autoregression,» *The Journal of Machine Learning Research*, vol. 21, n.º 1, págs. 6690-6741, 2020.

[42] Y.-C. Chiu y J. Z. Shyu, «Applying multivariate time series models to technological product sales forecasting,» *International Journal of Technology Management*, vol. 27, n.º 2-3, págs. 306-319, 2004.

[43] P. Lara-Benıtez, M. Carranza-Garcıa, J. M. Luna-Romera y J. C. Riquelme, «Temporal convolutional networks applied to energy-related time series forecasting,» *applied sciences*, vol. 10, n.º 7, pág. 2322, 2020.

[44] Y. Chen, Y. Kang, Y. Chen y Z. Wang, «Probabilistic forecasting with temporal convolutional neural network,» *Neurocomputing*, vol. 399, págs. 491-501, 2020.

[45] X. Xie, A. K. Parlikad y R. S. Puri, «A neural ordinary differential equations based approach for demand forecasting within power grid digital twins,» en *2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids (SmartGridComm)*, IEEE, 2019, págs. 1-6.

[46] X. Chen, F. A. Araujo, M. Riou et al., «Forecasting the outcome of spintronic experiments with neural ordinary differential equations,» *Nature communications*, vol. 13, n.º 1, pág. 1016, 2022.

[47] S. Papadopoulos e I. Karakatsanis, «Short-term electricity load forecasting using time series and ensemble learning methods,» en *2015 IEEE Power and Energy Conference at Illinois (PECI)*, IEEE, 2015, págs. 1-6.

[48] I. Yenidoğan, A. Çayir, O. Kozan, T. Dağ y Ç. Arslan, «Bitcoin forecasting using ARIMA and PROPHET,» en *2018 3rd international conference on computer science and engineering (UBMK)*, IEEE, 2018, págs. 621-624.

[49] S. Y. Shah, D. Patel, L. Vu et al., «AutoAI-TS: AutoAI for time series forecasting,» en *Proceedings of the 2021 International Conference on Management of Data*, 2021, págs. 2584-2596.

[50] A. Shehadeh, O. Alshboul, R. E. Al Mamlook y O. Hamedat, «Machine learning models for predicting the residual value of heavy construction equipment: An evaluation of modified decision tree, LightGBM, and XGBoost regression,» *Automation in Construction*, vol. 129, pág. 103 827, 2021.

[51] T. Deng, Y. Zhao, S. Wang y H. Yu, «Sales Forecasting Based on LightGBM,» en *2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, IEEE, 2021, págs. 383-386.

[52] C. Persson, P. Bacher, T. Shiga y H. Madsen, «Multi-site solar power forecasting using gradient boosted regression trees,» *Solar Energy*, vol. 150, págs. 423-436, 2017.

[53] J. P. U. Cadavid, S. Lamouri y B. Grabot, «Trends in machine learning applied to demand & sales forecasting: A review,» en *International conference on information systems, logistics and supply chain*, 2018.

[54] D. Rohaan, E. Topan y C. G. Groothuis-Oudshoorn, «Using supervised machine learning for B2B sales forecasting: A case study of spare parts sales forecasting at an after-sales service provider,» *Expert systems with applications*, vol. 188, pág. 115 925, 2022.

[55] R. Carbonneau, K. Laframboise y R. Vahidov, «Application of machine learning techniques for supply chain demand forecasting,» *European journal of operational research*, vol. 184, n.º 3, págs. 1140-1154, 2008.

[56] J. A. Rodrigo y J. E. Ortiz, *Skforecast: Forecasting Series Temporales con Python y Scikitlearn*, Enlace: `https://www.cienciadedatos.net/py27-forecasting-series-temporales-python-scikitlearn.html`, Disponible bajo la licencia Attribution 4.0 International (CC BY 4.0), Año no proporcionado.

[57] S. DELSOL, *Variable exógena*, `https://www.sdelsol.com/glosario/variable-exogena/`, Accedido el 18 de agosto de 2023.

[58] C. Yu, «The effects of exogenous variables in efficiency measurement—a Monte Carlo study,» *European journal of operational research*, vol. 105, n.º 3, págs. 569-580, 1998.

[59] J. M. Cordero, F. Pedraja y D. Santın, «Alternative approaches to include exogenous variables in DEA measures: A comparison using Monte Carlo,» *Computers & Operations Research*, vol. 36, n.º 10, págs. 2699-2706, 2009.

[60] J. Pinheiro, «A Sales Forecasting Model Based on Internal Organizational Variables,» *Available at SSRN 2214543*, 2013.

[61] J. T. Luxhøj, J. O. Riis y B. Stensballe, «A hybrid econometric—neural network modeling approach for sales forecasting,» *International Journal of Production Economics*, vol. 43, n.º 2-3, págs. 175-192, 1996.

[62] A. Lasek, N. Cercone y J. Saunders, «Smart restaurants: survey on customer demand and sales forecasting,» *Smart cities and homes*, págs. 361-386, 2016.

[63] Prevédere. «The Top 10 External Factors That Impact Forecast Accuracy.» Accedido el 22 de agosto de 2023. (2023).

[64] G. Verstraete, E.-H. Aghezzaf y B. Desmet, «A leading macroeconomic indicators' based framework to automatically generate tactical sales forecasts,» *Computers & Industrial Engineering*, vol. 139, pág. 106 169, 2020.

[65] D. Mügge, «Studying macroeconomic indicators as powerful ideas,» *Journal of European Public Policy*, vol. 23, n.º 3, págs. 410-427, 2016.

[66] M. Fernández de Mesa Bustelo, «Análisis y mejora de la predicción de la demanda eléctrica en periodos de alto ECM,» 2016.

[67] J. G. Hauk et al., «Teorıa y modelos en los pronósticos de ventas,» *Revista Universidad EAFIT*, vol. 1, n.º 1, págs. 12-31, 1965.

[68] J. T. Rothe, «Effectiveness of sales forecasting methods,» *Industrial Marketing Management*, vol. 7, n.º 2, págs. 114-118, 1978.

[69] C. M. Torobeo Chávez y M. A. Hernández Ballena, «Mejora del modelo de la demanda en el canal masivo de una empresa de empaques,»

[70] G. A. Grizzle y W. E. Klay, «Forecasting state sales tax revenues: comparing the accuracy of different methods,» *State & Local Government Review*, págs. 142-152, 1994.

[71] J. Liu, C. Liu, L. Zhang e Y. Xu, «Research on sales information prediction system of e-commerce enterprises based on time series model,» *Information Systems and e-Business Management*, vol. 18, págs. 823-836, 2020.

[72] M. Sadiku, A. E. Shadare, S. M. Musa, C. M. Akujuobi y R. Perry, «Data visualization,» *International Journal of Engineering Research And Advanced Technology (IJERAT)*, vol. 2, n.º 12, págs. 11-16, 2016.

[73] A. S. Harsoor y A. Patil, «Forecast of sales of Walmart store using big data applications,» *International Journal of Research in Engineering and Technology*, vol. 4, n.º 6, págs. 51-59, 2015.

[74] P. Willingham. «Introduction to Sales Forecasting with Excel.» (feb. de 2023).

[75] C. Sahyaja y T. Shenoy, «A Study on Sales Forecasting of Reliance Retail Stores Using Python,» *LATEST TRENDS IN MULTIDISCIPLINARY RESEARCH & DEVELOPMENT*, pág. 60, 2023.

[76] B. Jena, «An Approach for Forecast Prediction in Data Analytics Field by Tableau Software.,» *International Journal of Information Engineering & Electronic Business*, vol. 11, n.º 1, 2019.

[77] M. Raje, P. Jain y V. Chole, «SALES ANALYSIS AND PREDICTION DASHBOARD USING POWER BI,»

[78] I. K. Suwintana, I. O. Sudiadnyani y N. Saptarini, «Developing Web-Based Application of Sales Forecasting System Using Triple Exponential Smoothing Method For Small and Medium Garment Enterprises,» en *International Conference on Science and Technology (ICST 2018)*, Atlantis Press, 2018, págs. 1068-1071.

[79] E. Tavanidou, K. Nikolopoulos, K. Metaxiotis y V. Assimakopoulos, «eTIFIS: An innovative e-forecasting web application,» *International Journal of Software Engineering and Knowledge Engineering*, vol. 13, n.º 02, págs. 215-236, 2003.

[80] D. J. Dalrymple, «Sales forecasting practices: Results from a United States survey,» *International journal of Forecasting*, vol. 3, n.º 3-4, págs. 379-391, 1987.

[81] P. Agarwal, «Continuous scrum: agile management of saas products,» en *Proceedings of the 4th India Software Engineering Conference*, 2011, págs. 51-60.

[82] A. Gelman, «Exploratory data analysis for complex models,» *Journal of Computational and Graphical Statistics*, vol. 13, n.º 4, págs. 755-779, 2004.

[83] A. De Myttenaere, B. Golden, B. Le Grand y F. Rossi, «Mean absolute percentage error for regression models,» *Neurocomputing*, vol. 192, págs. 38-48, 2016.

[84] T. Chai y R. R. Draxler, «Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature,» *Geoscientific model development*, vol. 7, n.º 3, págs. 1247-1250, 2014.

[85] J. C. Alonso, «Estimación de modelos var, prueba de causalidad de granger y función impulso respuesta empleando easyreg,» Universidad Icesi, inf. téc., 2011.

[86] S. Johansen, «Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models,» *Econometrica: journal of the Econometric Society*, págs. 1551-1580, 1991.

[87] J. C. Alonso, «Tutorial para pruebas de raıces unitarias: Dickey-Fuller aumentado y Phillips-Perron en easyreg,» Universidad Icesi, inf. téc., 2010.

[88] K. J. White, «The Durbin-Watson test for autocorrelation in nonlinear models,» *The Review of Economics and Statistics*, págs. 370-373, 1992.

[89] T. M. McCarthy, D. F. Davis, S. L. Golicic y J. T. Mentzer, «The evolution of sales forecasting management: A 20-year longitudinal study of forecasting practices,» *Journal of Forecasting*, vol. 25, n.º 5, págs. 303-324, 2006.

[90] H. Song, S. F. Witt y X. Zhang, «Developing a Web-based tourism demand forecasting system,» *Tourism Economics*, vol. 14, n.º 3, págs. 445-468, 2008.

ANEXOS

- **Anexo A:** Primeros códigos implementados para el pronóstico de las ventas.
- **Anexo B:** Segunda parte de códigos para el pronóstico de las ventas.
- **Anexo C:** Códigos de la aplicación web.