PROPUESTA DE UN MODELO DE PREDICCIÓN DE INVENTARIO DE UNA EMPRESA DEDICADA A LA VENTA DE DISPOSITIVOS DE ALMACENAMIENTO DE ENERGÍA, BASADO EN MODELOS DE MACHINE LEARNING.

Luz Angélica Norato Díaz
Andrés Monsalve Rodríguez

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

Director
Julián Gil González

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, JUNIO 9 DE 2025

TABLA DE CONTENIDO

1 DEFINICIÓN DEL PROBLEMA 3

1.1 PLANTEAMIENTO DEL PROBLEMA 3

1.2 FORMULACIÓN DEL PROBLEMA 4

2 OBJETIVOS DEL PROYECTO 6

2.1 OBJETIVO GENERAL 6

2.2 OBJETIVOS ESPECÍFICOS 6

3 MARCO TEÓRICO Y ANTECEDENTES 7

3.1 MARCO TEÓRICO 7

3.1.1 Concepto de Machine Learning 7
3.1.2 Modelos Lineales en Machine Learning 8
3.1.3 Modelos Basados en Arboles de Decisión 9
3.1.3.1 Arboles de Decisión 9
3.1.3.2 Random Forest 9
3.1.3.3 XGBoost (Extreme Gradient Boosting) 10
3.1.4 Componentes Temporales en Aprendizaje Automático: Rezagos y Medias Móviles 10
3.1.4.1 Rezagos (Lags) 10
3.1.4.2 Medias Móviles 11
3.1.5 Modelos Híbridos y Enfoque Top-Down con Proporciones. 11
3.1.6 Métricas de Desempeño 13
3.1.6.1 Métrica MSE (Error Cuadrático Medio) 13
3.1.6.2 Métrica RMSE (Raíz del Error Cuadrático Medio) 14
3.1.6.3 Métrica MAPE (Error Porcentual Absoluto Medio) 15
3.1.6.4 Métrica sMAPE (Error Porcentual Absoluto Medio Simétrico) 15

3.1.6.5 Ventajas del sMAPE 16
3.1.7 Validación Cruzada 17
3.1.8 Prueba de Wilcoxon 18
3.1.9 Metodología CRISP-DM 18

3.2 ANTECEDENTES 21

4 PREPARACIÓN DE DATOS 26

4.1 LIMPIEZA Y FILTRADO DE VENTAS 27

4.1.1 Descripción del conjunto de datos 27
4.1.2 Proceso de limpieza: pasos aplicados y fundamentos 28
4.1.2.1 Análisis estadístico descriptivo por ítem 28
4.1.2.2 Eliminación de ítems con alta proporción de ceros 28
4.1.2.3 Tratamiento de valores negativos 28
4.1.2.4 Filtro por participación relativa en ventas 29
4.1.2.5 Detección y tratamiento de valores atípicos 29
4.1.3 Resultado final del proceso de limpieza 29

5 METODOLOGÍA USADA EN LOS MODELOS 30

5.1 PREPARACION Y PREPROCESAMIENTO DE DATOS 30

5.2 Modelado con Enfoques Clásicos de Series de Tiempo 31

5.3 Modelado con Enfoques Supervisados de Regresión 31

5.4 ENFOQUE TOD_DOWN: MODELO HÍBRIDO PARA LA DESAGREGACIÓN DE VENTAS POR ÍTEM 34

6 EVALUACIÓN DE LOS MODELOS 36

6.1 RESULTADOS MODELADOS CON ENFOQUE CLASICOS DE SERIES DE TIEMPO 36

6.1.1 ARIMA 36
6.1.2 SARIMA 38

6.2 RESULTADOS MODELADOS DE MACHINE LEARNING 39

6.2.1 Regresión Lineal 39
6.2.2 Árboles de Decisión (Regresión) 40

6.2.3 Random Forest 41
6.2.4 XGBoost 43

6.3 SELECCIÓN DEL MODELO 44

6.3.1 Evidencia Cuantitativa: Superioridad en Métricas de Error 44
6.3.2 Justificación Cualitativa y operativa 45

6.4 ANÁLISIS COMPARATIVO DE MODELOS: XGBOOST VS RANDOM FOREST 45

6.5 RESULTADOS MODELO HÍBRIDO （Random Forest + Proporción Media Exponencial）.47

7 CONCLUSIONES Y TRABAJOS FUTUROS 49

7.1 CONCLUSIONES 49

7.2 TRABAJOS FUTUROS 50

8 BIBLIOGRAFÍA 51

9 ANEXOS 54

LISTA DE FIGURAS

Figura 1 Ciclo de metodología del proyecto. 19
Figura 2 Modelo ARIMA (1,0,1) 36
Figura 3 Modelo ARIMA (1,1,1) 37
Figura 4 Modelo SARIMA 38
Figura 5 Modelo de Regresión Lineal 39
Figura 6 Modelo de Árboles de Decisión 40
Figura 7 Modelo de Árboles de Decisión 42
Figura 8 Modelo XGBoost 43
Figura 9 Comparación de $MAPE^2$ entre XGBoost y Random Forest 46

LISTA DE TABLAS

Tabla 1 Métricas de Evaluación. 17
Tabla 2 Modelos utilizados en prediiciones 25
Tabla 3 Ventas totales por país 2023 26
Tabla 4 Resumen de Actividades de Preparación y Procesamiento de Datos 30
Tabla 5 Configuración de Modelos Clásicos de Series de Tiempo 31
Tabla 6 Configuración de Modelos Supervisados de Regresión 32
Tabla 7 Resultados validación cruzada modelo Random Forest 33
Tabla 8 Resultados validación cruzada modelo XGBoost 34
Tabla 9 Métricas Modelo ARIMA 36
Tabla 10 Métricas Modelo SARIMA 38
Tabla 11 Métricas Modelo de Regresión Lineal 39
Tabla 12 Métricas Modelo de Árboles de Decisión 40
Tabla 13 Métricas Modelo Random Forest 41
Tabla 14 Métricas Modelo XGBoost 43
Tabla 15 Comparación de Métricas de Desempeño Modelos Random Forest y XGBoost 44
Tabla 16 Comparación Beneficios de Modelos Random Forest y XGBoost 45
Tabla 17 Métricas Modelo Híbrido Top-Down por Item 47

LISTA DE ANEXOS

Anexo 1 Pruebas de estacionalidad, estacionaridad y tendencia variable Total 54
Anexo 2 Predicciones por item usando modelo hibrido. 55
Anexo 3 Gráficas de las predicciones por item 57

INTRODUCCIÓN

La incertidumbre en la demanda de productos representa un desafío estratégico para empresas con presencia en múltiples países, especialmente cuando deben gestionar ciclos logísticos complejos y cadenas de suministro globalizadas. La empresa ABC, dedicada a la distribución internacional de dispositivos de almacenamiento de energía, ha enfrentado dificultades operativas causadas por la acumulación excesiva de inventario en sus centros de distribución en México, Colombia y Costa Rica. Esta situación ha provocado sobrecostos en almacenamiento, pérdida de eficiencia en la reposición de productos y una considerable inmovilización de capital.

Frente a esta problemática, la generación de pronósticos precisos de la demanda se torna una herramienta crítica para fortalecer la toma de decisiones y optimizar la gestión de inventarios. En este contexto, el presente proyecto se enfocó en el desarrollo de un modelo de predicción de ventas mediante técnicas de aprendizaje automático, con el fin de anticipar el comportamiento semanal de la demanda en los tres países mencionados. El enfoque combinó análisis de series temporales con modelos supervisados de regresión, utilizando como insumo datos históricos semanales recopilados entre 2021 y 2024.

Bajo el marco metodológico CRISP-DM, se llevó a cabo un proceso riguroso que incluyó la comprensión del negocio, exploración y preparación de datos, modelado, evaluación y documentación de resultados. Durante esta etapa, se exploraron y compararon múltiples algoritmos de predicción, incluyendo modelos clásicos y de ensamble como la regresión lineal, árboles de decisión, XGBoost y Random Forest. Este último fue el modelo que demostró mejor desempeño en el conjunto de prueba, alcanzando métricas robustas como un error cuadrático medio (RMSE) de 1048.25, un error porcentual absoluto medio ajustado (MAPE) de 24.84% y un error porcentual absoluto medio simétrico (SMAPE) de 22.36%.

El modelo desarrollado cumplió con los requisitos técnicos definidos en el proyecto y permitió generar predicciones semanales de ventas. Los resultados indican que este tipo de herramienta podría ser utilizada como apoyo en la planificación de inventarios y en la programación de reposiciones. Asimismo, el trabajo dejó una base metodológica que puede ser adaptada para otros

contextos o líneas de producto. Los resultados de este proyecto ofrecen elementos que podrían ser considerados en futuras aplicaciones de técnicas de ciencia de datos dentro de la organización.

1 DEFINICIÓN DEL PROBLEMA

1.1 PLANTEAMIENTO DEL PROBLEMA

La empresa ABC, con sede en EE. UU., se consolida como un referente en soluciones de energía segura y equipos de almacenamiento y acondicionamiento. Su portafolio de productos abarca desde computadoras personales hasta redes de comunicación LAN, WAN y globales. Con una estructura internacional bien establecida, ABC opera en 10 países a través de 12 oficinas de ventas y más de 200 centros de servicio distribuidos en más de 40 naciones. La producción de sus productos se concentra en plantas ubicadas estratégicamente en China, desde donde se despachan los contenedores hacia las bodegas de distribución en México, Colombia y Costa Rica, los principales centros logísticos para cubrir los mercados de América Latina.

En estas bodegas, no solo se almacenan los productos, sino que también se registran las ventas correspondientes a las operaciones de distribución en sus respectivas regiones. A pesar de la sólida infraestructura de ABC, la empresa enfrenta dificultades para alinear los procesos de producción y distribución con la demanda real de estos mercados clave.

Uno de los principales síntomas del problema es la acumulación excesiva de productos en las bodegas de México, Colombia y Costa Rica. Debido a la falta de precisión en la previsión de la demanda, una parte significativa de los productos que llegan en contenedores desde China no se vende dentro de los tiempos proyectados, lo que genera un exceso de inventario. Esta situación provoca una ocupación innecesaria de espacio en las bodegas, afecta la capacidad operativa de la empresa y representa un capital inmovilizado que no genera retorno inmediato, incluyendo en ocasiones algunas devoluciones.

Además, la acumulación prolongada de inventario implica un aumento en los costos de almacenamiento, tanto por el uso continuo del espacio en las bodegas como por los recursos necesarios para mantener el inventario y para su distribución desde China. Estos costos

adicionales impactan negativamente la rentabilidad de la empresa. Aunque los productos no pierden su valor de mercado inmediatamente, su permanencia prolongada en las bodegas disminuye la eficiencia de las operaciones logísticas y comerciales.

El desajuste entre la llegada de productos desde China y la venta efectiva en las bodegas de México, Colombia y Costa Rica genera una falta de alineación en la cadena de suministro. La falta de un pronóstico de ventas más preciso dificulta la planificación efectiva de las operaciones de distribución, como señala Philip Kotler en su libro *Dirección de Marketing* (2006), un pronóstico de ventas preciso es fundamental para anticipar la demanda de productos y facilitar la asignación eficiente de recursos, la planificación de la producción y la toma de decisiones estratégicas [1]. La falta de este tipo de pronósticos conduce a problemas operativos y financieros, como es el caso de ABC en estos mercados clave.

La raíz del problema radica en la falta de un sistema predictivo que permita anticipar con mayor precisión las fluctuaciones en la demanda de productos en las bodegas de México, Colombia y Costa Rica. Sin un modelo de pronóstico de ventas efectivo, la planificación de la producción y distribución de productos desde China se vuelve inconsistente, lo que genera excesos de inventario.

Por tal razón, se desarrolló un modelo predictivo de ventas para los productos de la empresa ABC, el cual se alimentó con la base de datos del historial de ventas de sus bodegas en México, Colombia y Costa Rica.

1.2 FORMULACIÓN DEL PROBLEMA

En este contexto, la acumulación de inventarios en las bodegas de México, Colombia y Costa Rica pone de manifiesto la dificultad que enfrenta la empresa ABC para prever con precisión las fluctuaciones en la demanda de los mercados de América Latina. Esta problemática ha generado costos adicionales de almacenamiento y ha afectado la eficiencia operativa de la compañía. En este escenario, surge la siguiente pregunta principal:

¿Cómo desarrollar un modelo predictivo para la venta de productos de ABC en las bodegas de México, Colombia y Costa Rica utilizando técnicas de Machine Learning?

A partir de esta interrogante, surgen otras preguntas relacionadas que orientan el enfoque del problema:

- ¿Cómo preparar los datos existentes de las ventas de ABC en México, Colombia y Costa Rica para el entrenamiento y la evaluación de modelos predictivos?
- ¿Qué técnicas de ciencia de datos son adecuadas para entrenar modelos predictivos de ventas?
- ¿Cómo entrenar diferentes modelos utilizando técnicas de ciencia de datos para crear un modelo predictivo de las ventas en México, Colombia y Costa Rica?
- ¿Cuál de los modelos entrenados ofrece el mejor desempeño en la predicción de las ventas?

2 OBJETIVOS DEL PROYECTO

2.1 OBJETIVO GENERAL

Desarrollar un modelo predictivo para pronosticar las ventas de productos de ABC en las bodegas de México, Colombia y Costa Rica, utilizando técnicas avanzadas de Machine Learning.

2.2 OBJETIVOS ESPECÍFICOS

- Preparar los datos existentes de las ventas de ABC en México, Colombia y Costa Rica para el entrenamiento y evaluación de los modelos predictivos.

- Identificar las técnicas de ciencia de datos que se utilizan para entrenar modelos predictivos de ventas.

- Entrenar diferentes modelos, utilizando técnicas de ciencia de datos para crear un modelo de predicción de las ventas de ABC en México, Colombia y Costa Rica.

- Evaluar el rendimiento de los modelos entrenados con el fin de determinar cuál de ellos ofrece los mejores resultados.

3 MARCO TEÓRICO Y ANTECEDENTES

3.1 MARCO TEÓRICO

3.1.1 Concepto de Machine Learning

El aprendizaje automático (Machine Learning) es una disciplina de la inteligencia artificial (IA) que permite a los sistemas computacionales mejorar su desempeño mediante el análisis de datos, sin depender de programación explícita [1]. Según Sandoval (2017), se define como "un campo que dota a las computadoras de la habilidad de aprender a partir de datos, identificando patrones para realizar predicciones o clasificaciones" [2]. Esta capacidad se estructura en dos enfoques principales:

1. Aprendizaje supervisado: Utiliza datos etiquetados para entrenar modelos predictivos en tareas específicas como clasificación de riesgos crediticios o predicción de ventas futuras [3]. Este enfoque es esencial cuando se dispone de datos históricos claramente definidos.

2. Aprendizaje no supervisado: Detecta relaciones intrínsecas en datos sin etiquetas, útil en escenarios como la segmentación de clientes o análisis exploratorio donde se requiere descubrir estructuras ocultas en grandes volúmenes de datos [4].

Un hito histórico relevante fue el trabajo de Arthur L. Samuel en 1959, quien desarrolló un algoritmo autónomo capaz de aprender a jugar damas, sentando las bases para sistemas complejos actuales como motores de recomendación utilizados en plataformas digitales como Netflix o sistemas avanzados para vehículos autónomos [5].

El funcionamiento general del machine learning se fundamenta en tres etapas iterativas:

- Proceso de decisión: El algoritmo genera predicciones utilizando datos de entrada, como el historial de compras.

- Función de error: Evalúa la precisión de las predicciones comparándolas con datos reales.

- Optimización: Ajusta los parámetros internos para minimizar errores, repitiendo el ciclo hasta alcanzar una precisión óptima [6].

Además, el machine learning emplea diversos modelos avanzados como redes neuronales artificiales (que simulan estructuras cerebrales), máquinas de soporte vectorial y técnicas de ensamble como Random Forest y Extreme Gradient Boosting (XGBoost) para abordar problemas complejos desde diagnósticos médicos hasta campañas de marketing personalizadas [7].

3.1.2 Modelos Lineales en Machine Learning

Los modelos lineales son fundamentales en el aprendizaje supervisado por su interpretabilidad y eficiencia computacional. El modelo de regresión lineal expresa una relación lineal entre variables independientes (predictoras) y la variable dependiente (respuesta), formulado como:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \varepsilon$$

Donde: $\hat{y}$ es el valor estimado de la variable dependiente, β_0 es el intercepto, β_i son los coeficientes del modelo, x_i son las variables independientes o predictoras del modelo y ε es el término de error, que recoge la variabilidad no explicada por el modelo, incluyendo el ruido aleatorio o factores no considerados.

Estos modelos son particularmente útiles por su simplicidad y capacidad de interpretación directa, aunque presentan limitaciones al no capturar adecuadamente relaciones complejas o no lineales, lo

que ha llevado al desarrollo de métodos más avanzados como modelos polinomiales, redes neuronales y métodos basados en árboles de decisión [3].

3.1.3 Modelos Basados en Arboles de Decisión

Además de los modelos lineales, el aprendizaje automático incluye modelos no paramétricos como los árboles de decisión y sus variantes mejoradas, que se han destacado por su capacidad para capturar relaciones no lineales y realizar segmentaciones complejas en el espacio de características.

3.1.3.1 Arboles de Decisión

Un árbol de decisión es una estructura jerárquica en la que los datos se dividen mediante reglas basadas en condiciones booleanas hasta llegar a un resultado o predicción. Cada nodo interno representa una condición sobre una variable, y cada hoja representa un valor de salida. Son especialmente útiles por su interpretación y visualización clara del proceso de decisión. Sin embargo, presentan riesgo de sobreajuste al trabajar con datos ruidosos o altamente dimensionales [8].

3.1.3.2 Random Forest

Random Forest es un modelo de ensamblaje basado en árboles de decisión. Consiste en construir múltiples árboles (generalmente cientos o miles) entrenados sobre subconjuntos aleatorios del conjunto de datos y del espacio de características. La predicción final se realiza mediante votación mayoritaria (para clasificación) o promedio (para regresión). Este enfoque reduce la varianza del modelo y mitiga el sobreajuste, siendo especialmente útil cuando se dispone de muchas variables con potenciales interacciones [9].

3.1.3.3 XGBoost (Extreme Gradient Boosting)

XGBoost es una implementación eficiente del algoritmo de gradient boosting, que construye árboles de manera secuencial. Cada nuevo árbol intenta corregir los errores cometidos por los anteriores mediante el ajuste del gradiente del error. XGBoost incorpora regularización L1 y L2, lo cual previene el sobreajuste, y su eficiencia computacional lo ha convertido en una herramienta popular en competencias de ciencia de datos y en aplicaciones de predicción de ventas [10].

Estos modelos han demostrado ser altamente eficaces en contextos donde las relaciones entre las variables son complejas y no lineales. Además, permiten la extracción de importancia de variables, lo que proporciona interpretabilidad y facilita la toma de decisiones estratégicas.

3.1.4 Componentes Temporales en Aprendizaje Automático: Rezagos y Medias Móviles

El análisis de series temporales es esencial en múltiples dominios de la ciencia de datos, desde finanzas y economía hasta logística e inteligencia de negocios. En el contexto del aprendizaje automático, los componentes temporales como los rezagos y las medias móviles permiten capturar patrones secuenciales y dependencias históricas que no pueden ser modeladas eficazmente con variables independientes tradicionales.

3.1.4.1 Rezagos (Lags)

Los rezagos representan los valores pasados de una variable y permiten modelar la autocorrelación dentro de una serie temporal [11]. Son fundamentales en modelos autorregresivos (AR), donde el valor actual de la serie depende linealmente de sus valores anteriores. Un modelo autorregresivo de orden p, denotado AR(p), se representa como:

$$y_t = c + \phi_1 y_1 + \phi_2 y_2 + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

Donde y_t es el valor actual de la serie temporal en el tiempo t, c es una constante, φ_i son los coeficientes del modelo, y_{t-i} son los rezagos o valores pasados de la serie temporal y ε_t representa el término de error aleatorio. Este tipo de modelos ha sido ampliamente estudiado y aplicado, tal como se discute en [5] [12].

En aprendizaje automático, los rezagos se integran como variables predictoras en modelos supervisados (por ejemplo, regresión lineal, Random Forest o XGBoost) para capturar dinámicas temporales sin necesidad de aplicar modelos clásicos como ARIMA o SARIMA.

3.1.4.2 Medias Móviles

Las medias móviles: Son técnicas de suavizamiento utilizadas para eliminar el ruido de corto plazo y resaltar tendencias de largo plazo [13]. Las formas más comunes son:

$$SMA_t = \frac{1}{k}\sum_{i=0}^{k-1} y_{t-i}$$

$$EMA_t = \alpha y_t + (1-\alpha)EMA_{t-1}$$

Donde k es el número de periodos considerados para la media simple y $\alpha \in (0,1)$ es un parámetro de suavizado en la media exponencial. La EMA da mayor peso a los valores más recientes, lo cual la hace especialmente útil en entornos donde los patrones de corto plazo son relevantes [13].

Estas técnicas también se emplean para construir nuevas variables que ayudan a enriquecer modelos predictivos, permitiendo que algoritmos como XGBoost o redes neuronales aprendan patrones más robustos de comportamiento temporal.

3.1.5 Modelos Híbridos y Enfoque Top-Down con Proporciones.

Los modelos híbridos combinan diversas técnicas de modelado con el fin de mejorar la precisión

y la generalización del pronóstico. En particular, integran enfoques lineales tradicionales —como la regresión lineal— con algoritmos no lineales más potentes, tales como árboles de decisión, Random Forest o XGBoost [14] [15]. Esta integración permite capturar tanto relaciones globales como interacciones complejas entre variables predictoras.

Dentro del ámbito del pronóstico de ventas, una de las aplicaciones más efectivas de modelos híbridos es el enfoque top-down, el cual comienza con la predicción del valor agregado (por ejemplo, el total de ventas semanales) y luego distribuye dicha predicción entre ítems individuales utilizando proporciones históricas [16] [17]. Este método es especialmente útil en contextos donde las series desagregadas presentan una alta proporción de ceros, baja estabilidad o ruido significativo, pero la serie total muestra una señal clara y consistente.

Una implementación destacada de este enfoque es el método "**Total and Split**", propuesto por Spiliotis, Petropoulos y Assimakopoulos [18]. En dicho esquema, se utiliza un modelo predictivo robusto para estimar la serie total, y posteriormente se realiza la descomposición en niveles inferiores mediante proporciones dinámicas. Estas proporciones pueden ser obtenidas a través de técnicas adaptativas, como el suavizamiento exponencial, que asignan mayor peso a las observaciones recientes manteniendo al mismo tiempo la memoria histórica del sistema [18] [5].

Este enfoque híbrido basado en proporciones presenta varias ventajas:

- Facilita la incorporación indirecta de estacionalidad y tendencia a través de la serie total.
- Mejora la estabilidad de las predicciones cuando las series individuales son ruidosas o escasas.
- Reduce la complejidad computacional frente a la modelación simultánea de múltiples series desagregadas.
- Asegura la coherencia entre las predicciones agregadas y las desagregadas, cumpliendo principios de reconciliación jerárquica [7].

Debido a estas propiedades, el enfoque *top-down* híbrido se ha posicionado como una herramienta eficaz para escenarios en los que se requiere preservar la estructura relativa de los componentes sin perder robustez en el nivel agregado.

3.1.6 Métricas de Desempeño

En el ámbito del forecasting o pronóstico de series temporales, la selección de métricas de evaluación desempeña un papel crítico en la validación de la capacidad predictiva de los modelos. A diferencia de los modelos de regresión estática, donde indicadores como el coeficiente de determinación (R^2) son comunes, en el contexto del pronóstico el interés se centra en la precisión out-of-sample, es decir, en la capacidad del modelo para predecir datos futuros no observados. En esta línea, autores como Hyndman y Athanasopoulos [5], así como Makridakis et al. [14], han destacado que el objetivo fundamental del forecasting es anticipar con precisión el comportamiento futuro de la serie, más allá de su ajuste a los datos históricos.

Una limitación crítica del R^2 es que se calcula sobre el conjunto de entrenamiento (in-sample), lo cual puede generar una falsa percepción de buen desempeño cuando en realidad el modelo se encuentra sobreajustado (overfitting) y pierde capacidad de generalización [2]. Adicionalmente, el R^2 es altamente dependiente de la varianza de la serie, lo que impide su uso como métrica comparativa entre series con diferentes escalas o niveles de volatilidad. Esta propiedad lo convierte en un indicador inadecuado en contextos de análisis multiserie, como el pronóstico simultáneo de ventas por producto o por zona geográfica.

Por consiguiente, en esta investigación se opta por un conjunto de métricas recomendadas por la comunidad científica especializada en pronóstico, particularmente aquellas utilizadas en las múltiples ediciones de las M-Competitions. Estas métricas permiten una evaluación más robusta y comparativa, y se clasifican en tres grandes grupos:

3.1.6.1 Métrica MSE (Error Cuadrático Medio)

El Error Cuadrático Medio (MSE, por sus siglas en inglés Mean Squared Error) es una de las métricas más comunes para evaluar la precisión de los modelos de regresión en la predicción de valores continuos, como en el caso de la predicción de ventas. Esta métrica mide la media de las

diferencias al cuadrado entre los valores reales y las predicciones del modelo, lo que permite penalizar más fuertemente los errores grandes. Matemáticamente, se define como:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

donde y_i son los valores observados, $\hat{y}_i$ las predicciones del modelo, y n el número total de observaciones. Debido a su naturaleza cuadrática, el MSE es sensible a las grandes discrepancias entre las predicciones y los valores reales, lo cual lo convierte en una métrica adecuada para aplicaciones donde los errores grandes son indeseables, como en la predicción de ventas, donde los pronósticos incorrectos pueden tener consecuencias económicas importantes. Sin embargo, una limitación del MSE es que su valor se encuentra en las unidades cuadradas de la variable de salida, lo que puede hacer que su interpretación sea menos intuitiva [12].

3.1.6.2 Métrica RMSE (Raíz del Error Cuadrático Medio)

La raíz del error cuadrático medio (RMSE, por sus siglas en inglés Root Mean Squared Error) es otra métrica comúnmente utilizada que representa la raíz cuadrada del MSE. Se define como:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Esta métrica mantiene las unidades originales de la variable de respuesta, lo que facilita su interpretación directa en el contexto del problema. El RMSE penaliza los errores grandes más que los pequeños, pero de forma menos severa que el MSE. Es especialmente útil para evaluar modelos donde se desea interpretar la magnitud media del error en las mismas unidades que la variable objetivo [52].

3.1.6.3 Métrica MAPE (Error Porcentual Absoluto Medio)

El Error Porcentual Absoluto Medio (MAPE, por sus siglas en inglés Mean Absolute Percentage Error) es una métrica ampliamente utilizada para evaluar la precisión de los modelos de regresión, especialmente en contextos empresariales donde se requiere interpretar los errores en términos relativos o porcentuales. Esta métrica calcula el valor absoluto del error entre las predicciones y los valores reales, expresándolo como un porcentaje del valor observado, y luego promedia estos porcentajes a lo largo del conjunto de datos. Se define matemáticamente como:

$$MAPE = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

donde y_i representa el valor real, $\hat{y}_i$ la predicción del modelo y n el número de observaciones.

El MAPE es especialmente valioso por su interpretabilidad: un MAPE de 10% indica que, en promedio, las predicciones se desvían un 10% de los valores reales. Esta propiedad lo convierte en una herramienta preferida en aplicaciones como la predicción de ventas, donde es importante comunicar los errores de forma comprensible a los responsables de toma de decisiones. Sin embargo, presenta limitaciones cuando existen valores cercanos a cero en la variable objetivo, ya que puede conducir a errores porcentuales extremadamente grandes o indefinidos [13].

3.1.6.4 Métrica sMAPE (Error Porcentual Absoluto Medio Simétrico)

El Error Porcentual Absoluto Medio Simétrico (sMAPE, por sus siglas en inglés *Symmetric Mean Absolute Percentage Error*) es una variante del MAPE diseñada para abordar algunas de sus limitaciones, particularmente en casos donde los valores reales son cercanos a cero o cuando las predicciones y observaciones pueden ser asimétricas. A diferencia del MAPE, que solo considera el error relativo respecto al valor real, el sMAPE introduce una normalización simétrica que incluye tanto el valor real como el pronosticado en el denominador, evitando así divisiones por cero o errores excesivamente grandes.

La fórmula del sMAPE se define como:

$$sMAPE = \frac{100\%}{n} \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}$$

donde y_i representa el valor real, $\hat{y}_i$ la predicción del modelo y n el número de observaciones

3.1.6.5 Ventajas del sMAPE

Simetría: Al incluir tanto el valor real como el pronosticado en el denominador, el sMAPE penaliza de manera equilibrada los errores por exceso y por defecto, lo que lo hace más justo en comparación con el MAPE cuando las magnitudes de las predicciones y los valores reales difieren significativamente [19].

Robustez frente a valores cercanos a cero: A diferencia del MAPE, el sMAPE no tiende a infinito cuando se aproxima a cero, ya que el denominador nunca será cero.

A pesar de sus ventajas, el sMAPE también presenta algunas desventajas:
Sensibilidad a errores pequeños: Cuando tanto son pequeños pero el error absoluto es significativo en relación con ellos, el sMAPE puede inflarse artificialmente [20].

Interpretabilidad menos intuitiva: Aunque el resultado sigue siendo un porcentaje, su cálculo simétrico puede dificultar la interpretación directa en comparación con el MAPE, especialmente para audiencias no técnicas [21].

El sMAPE es ampliamente utilizado en campos como la predicción de demanda, pronósticos financieros y evaluación de modelos de series temporales, donde se busca una métrica balanceada que evite los sesgos del MAPE tradicional [22].

Métrica	Interpretación	Ventajas / Desventajas
MAE	Promedio de los errores absolutos	Robusto ante outliers; fácil de interpretar
RMSE	Raíz cuadrada del promedio de errores al cuadrado	Penaliza errores grandes
MAPE	Porcentaje medio del error absoluto	Escalable e interpretable; sensible a valores cercanos a cero
sMAPE	Versión simétrica del MAPE	Más estable que MAPE, evita divisiones por cero

Tabla 1 Métricas de Evaluación.

3.1.7 Validación Cruzada

La validación cruzada es una técnica estadística utilizada para evaluar la capacidad de generalización de un modelo de aprendizaje automático. Su objetivo es proporcionar una estimación más robusta del desempeño del modelo sobre datos no vistos, evitando así el sobreajuste (overfitting), que ocurre cuando el modelo se ajusta demasiado bien a los datos de entrenamiento y pierde capacidad predictiva frente a nuevos datos [3][13].

El método más común es la validación cruzada k-fold, que consiste en dividir el conjunto de datos en k subconjuntos (o pliegues) de tamaño similar. En cada iteración, uno de los pliegues se utiliza como conjunto de validación y los k − 1 restantes se utilizan para entrenar el modelo. Este proceso se repite k veces, rotando el pliegue de validación, y al final se promedian las métricas de desempeño obtenidas en cada iteración para estimar la eficacia del modelo.

Formalmente, si se tiene un conjunto de datos $D = \{(x_i, y_i)\}$ para $i = 1$ hasta n, este se divide en D_1, D_2, ..., D_k y el modelo se entrena y evalúa de forma iterativa sobre estos subconjuntos.

Una variante útil cuando los datos tienen una estructura temporal es la validación cruzada con

series temporales, que respeta el orden cronológico y permite simular condiciones reales de pronóstico. Este enfoque es crucial en contextos como la predicción de ventas, donde se busca replicar escenarios futuros sin usar información del futuro en el entrenamiento [4].

La validación cruzada es considerada una técnica estándar en la evaluación de modelos y ha sido extensamente recomendada en la literatura científica como parte de las buenas prácticas en ciencia de datos y estadística aplicada [3][13].

3.1.8 Prueba de Wilcoxon

La prueba de Wilcoxon para rangos con signo es una técnica estadística no paramétrica utilizada para comparar dos muestras relacionadas, como los resultados de un mismo conjunto de datos evaluado con dos modelos distintos. Esta prueba permite evaluar si existe una diferencia significativa en la mediana de las diferencias entre ambas condiciones, sin asumir normalidad en los datos [23]. Esto la convierte en una opción ideal cuando se utilizan métricas asimétricas o con valores atípicos, como el MAPE (Mean Absolute Percentage Error), donde las distribuciones suelen estar sesgadas [24] [25].

En la comparación de modelos predictivos, el MAPE puede variar ampliamente entre observaciones, por lo que la prueba de Wilcoxon ofrece una manera robusta de identificar si uno de los modelos supera sistemáticamente al otro [25]. Al trabajar con pares de datos (por ejemplo, los MAPE obtenidos por dos modelos en los mismos subconjuntos), esta prueba analiza la magnitud y el signo de las diferencias sin verse afectada por la no normalidad o por valores extremos [23] [26]. Además, es adecuada incluso con muestras pequeñas, lo que la hace especialmente útil en procesos de validación cruzada.

3.1.9 Metodología CRISP-DM

La metodología CRISP-DM (*Cross-Industry Standard Process for Data Mining*) es un modelo de proceso estándar ampliamente reconocido y utilizado en el campo de la ciencia de datos y la minería de datos. Su popularidad radica en su flexibilidad, robustez y aplicabilidad en diversos contextos empresariales, ofreciendo un enfoque estructurado para la planificación y ejecución de proyectos de análisis de datos. Esta metodología proporciona una estructura sistemática que guía el desarrollo del proyecto desde la comprensión inicial del problema y los objetivos del negocio hasta la implementación final de la solución analítica, permitiendo una adaptación iterativa basada en los hallazgos y requerimientos que surgen durante el proceso [25], [26].

CRISP-DM se compone de seis fases principales interrelacionadas, las cuales no se ejecutan necesariamente de forma estrictamente secuencial, sino que pueden interactuar cíclicamente entre sí, permitiendo retroalimentación y ajustes a lo largo del ciclo de vida del proyecto. La naturaleza iterativa y las dependencias comunes entre estas fases se ilustran a menudo en diagramas de flujo de trabajo. A continuación, se describen las fases de esta metodología:

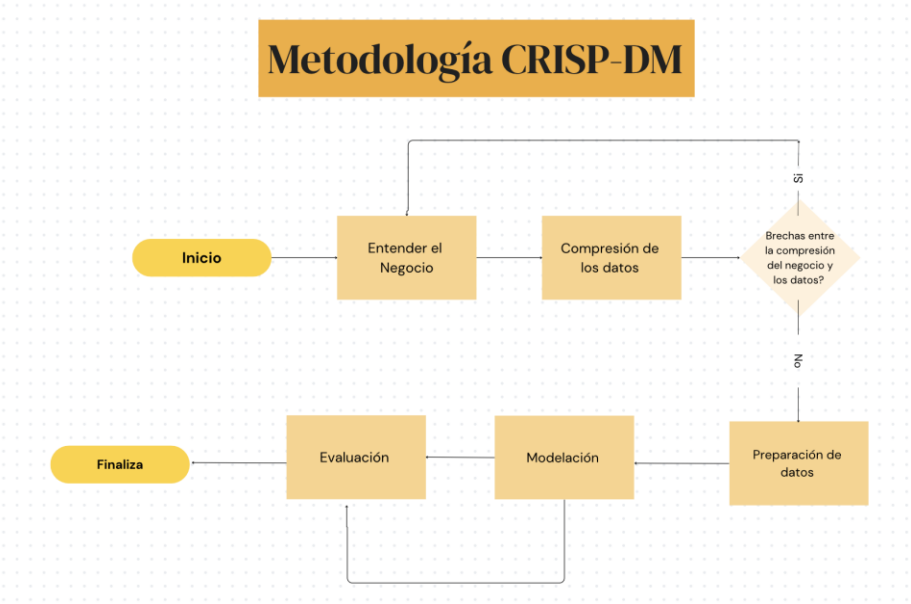


Figura 1 Ciclo de metodología del proyecto.

1. Comprensión del negocio

Esta fase inicial tiene como objetivo comprender los requerimientos del negocio y definir claramente el problema a resolver. Se establecen los objetivos del proyecto desde la perspectiva organizacional y se desarrolla un plan de acción orientado al análisis de datos. Esta etapa permite contextualizar los indicadores clave y definir criterios de éxito medibles [26].

2. Comprensión de los datos

En esta fase se recolectan, describen y exploran los datos relevantes. Se evalúa su calidad, completitud y estructura, y se identifican posibles inconsistencias, valores atípicos o sesgos. Esta exploración preliminar es esencial para detectar limitaciones y oportunidades analíticas en los datos [25][3].

3. Preparación de los datos

La preparación de los datos incluye tareas como limpieza, integración, selección de variables, imputación de valores faltantes y transformación. Estas acciones son críticas para garantizar que el conjunto de datos sea consistente, representativo y apropiado para el modelado posterior. Se aplican técnicas como normalización o creación de variables derivadas [26][3].

4. Modelado

Durante esta etapa se seleccionan las técnicas de modelado adecuadas (por ejemplo, regresión lineal, Random Forest, XGBoost, ARIMA o Prophet) y se ajustan sus hiperparámetros. Se crean modelos predictivos usando conjuntos de entrenamiento y validación, y se evalúan mediante métricas de desempeño como RMSE, MAPE y R^2. Esta fase es iterativa y puede retroalimentar las fases anteriores [25] [27].

5. Evaluación

Una vez construido el modelo, se evalúa su calidad y utilidad práctica. Se verifica que los resultados cumplan con los objetivos del negocio, se comparan modelos y se identifican posibles limitaciones.

Se realiza un análisis crítico de desempeño y se valida con stakeholders antes de su implementación definitiva [13].

6. Implementación

La fase final consiste en desplegar la solución analítica en el entorno operativo. Incluye documentar el modelo, crear scripts automatizados, integrar con sistemas existentes y planificar el monitoreo de resultados. Es fundamental establecer mecanismos para el mantenimiento y ajuste del modelo conforme se actualicen los datos o cambien las condiciones del negocio [26].

3.2 ANTECEDENTES

Los antecedentes que se presentan en esta sección tienen como objetivo contextualizar y respaldar metodológicamente este proyecto de grado en relación con la aplicación de técnicas avanzadas de aprendizaje automático para la predicción de ventas. Cada antecedente seleccionado aporta perspectivas metodológicas y prácticas valiosas, destacando la efectividad de modelos predictivos en diferentes sectores empresariales y evidenciando tendencias actuales en la implementación de enfoques híbridos, explicables y automatizados.

Antecedente 1: Demand Forecasting with Machine Learning.

El estudio realizado por Tseng y Turkmen en el MIT (2024) desarrolló un modelo predictivo para productos Heat-not-Burn (HNB) en mercados emergentes, específicamente en Italia. Utilizando el enfoque metodológico CRISP-DM, los autores partieron de un análisis profundo del problema empresarial seguido de una cuidadosa transformación y expansión de los datos disponibles, incorporando características temporales relevantes. Se compararon métodos tradicionales como Holt-Winters y ARIMA contra modelos de aprendizaje automático más avanzados, tales como XGBoost y Prophet. Este último destacó significativamente por su capacidad para modelar tendencias y patrones estacionales, alcanzando un error porcentual absoluto medio (MAPE) del 7.05% después del ajuste de hiperparámetros. El estudio también consideró variables externas,

incluyendo menciones en redes sociales y factores macroeconómicos, incrementando así la precisión predictiva y facilitando una comprensión integral de los determinantes de la demanda [28].

La relevancia de este antecedente radica en su similitud con el contexto del presente trabajo, al abordar desafíos en la predicción de ventas bajo condiciones de alta incertidumbre y datos limitados. Las estrategias metodológicas empleadas, especialmente la transformación temporal de datos y la integración de factores externos mediante modelos explicables, aportan elementos metodológicos clave para el desarrollo del modelo predictivo planteado para la empresa ABC.

Antecedente 2: Forecasting Nike's Sales using Facebook Data.

Boldt et al. en su trabajo presentado en la Conferencia IEEE Big Data 2016 investigaron la posibilidad de predecir ventas globales de Nike mediante el análisis de interacciones en Facebook. Utilizando el marco conceptual AIDA (Atención, Interés, Deseo y Acción) y técnicas de Social Set Analysis, el estudio analizó métricas diarias provenientes de redes sociales, tales como 'me gusta', comentarios y publicaciones, correlacionándolas con datos trimestrales de ventas extraídos de reportes financieros oficiales. Se observó que modelos simples de regresión proporcionaron alta precisión predictiva, mientras que los modelos multivariados enfrentaron dificultades debido a problemas de multicolinealidad. Asimismo, se aplicó una herramienta de visualización llamada Social Set Visualizer (SoSeVi) para identificar eventos específicos que impactaron notablemente en la interacción social vinculada a lanzamientos y campañas publicitarias [29].

Este antecedente es particularmente relevante por su metodología innovadora en la integración de datos no estructurados para la predicción de ventas, una práctica aplicable para ABC, especialmente en contextos donde el análisis del comportamiento del consumidor puede ofrecer ventajas estratégicas en la planificación de ventas y marketing.

Antecedente 3: Automated Machine Learning: AI-Driven Decision Making in Business Analytics.

El estudio realizado por la Universidad de Oxford en 2023 analizó el potencial del aprendizaje automático automatizado (AutoML) en el contexto empresarial, evaluando la herramienta H2O AutoML frente a modelos manualmente ajustados. Se utilizaron tres conjuntos de datos reales, mostrando que AutoML alcanzó resultados comparables con una menor necesidad de intervención humana, facilitando así su adopción práctica en diversos sectores industriales [30].

La relevancia para esta tesis reside en cómo la automatización del aprendizaje automático puede democratizar y agilizar el proceso de desarrollo de modelos predictivos, especialmente relevante para ABC en términos de reducción de costos y aumento de la eficiencia operativa.

Antecedente 4: Deep Learning for Time Series Prediction & Decision Making Over Time.

En su tesis doctoral, Lim exploró técnicas avanzadas de aprendizaje profundo para la predicción y toma de decisiones sobre series temporales. Se introdujeron innovadoras arquitecturas, como el Filtro Neuronal Recurrente (RNF), y modelos híbridos combinando técnicas clásicas y profundas. Además, la extracción avanzada de características incrementó considerablemente la precisión predictiva en contextos de comercio minorista y gestión de la demanda [31].

Este estudio ofrece metodologías robustas aplicables directamente a la empresa ABC, potenciando la precisión de sus pronósticos y optimizando decisiones estratégicas relacionadas con su cadena de suministro.

Antecedente 5: Implementación de un Método para el Pronóstico de Demanda de Medicamentos.

En esta investigación de la Universidad Nacional de Colombia, se desarrolló un modelo híbrido inteligente para predecir ventas a mediano plazo en el sector minorista de moda, empleando técnicas como Extreme Learning Machine y Harmony Search. El modelo propuesto optimizó significativamente el inventario y redujo costos asociados a excesos o faltantes [32].

Este antecedente proporciona evidencia práctica sobre la eficacia de modelos híbridos en la predicción de demanda, siendo directamente transferible al contexto de ABC para mejorar su gestión logística y eficiencia operativa.

Además de aportar fundamentos metodológicos valiosos, los antecedentes revisados permitieron identificar y comparar modelos utilizados en la predicción de ventas, tanto tradicionales como avanzados. Esta revisión incluyó enfoques estadísticos, algoritmos de aprendizaje automático, modelos híbridos y técnicas de aprendizaje profundo.

La siguiente matriz de comparación presenta una síntesis de los principales modelos utilizados en la predicción de ventas, incluyendo algoritmos tradicionales como la regresión lineal y ARIMA, técnicas de árboles de decisión como Random Forest y XGBoost, y enfoques híbridos que combinan medias móviles con aprendizaje automático. Cada modelo se describe según su definición, ventajas, limitaciones y nivel de explicabilidad, permitiendo una evaluación comparativa de su idoneidad en diferentes escenarios de negocio.

Modelo de ML / Estadístico	Definición	Ventajas	Desventajas	Explicabilidad
Regresión Lineal	Modelo supervisado que ajusta una relación lineal entre variables independientes y dependientes.	Fácil de implementar, interpretable, útil como línea base.	No captura relaciones no lineales ni interacciones complejas.	Alta
Árboles de Decisión	Modelo basado en particiones sucesivas del espacio de entrada para predecir una salida.	Intuitivo, fácil de visualizar, maneja variables categóricas.	Propenso al sobreajuste, especialmente con muchos niveles.	Alta
Random Forest	Conjunto de árboles de decisión entrenados con bagging.	Reduce el sobreajuste, robusto frente a outliers, generaliza bien.	Menor interpretabilidad que un solo árbol, mayor costo computacional.	Media-Alta
XGBoost	Algoritmo de boosting que optimiza errores residuales con árboles de decisión.	Precisión alta, manejo eficiente de datos faltantes y outliers.	Requiere ajuste fino de hiperparámetros, riesgo de sobreajuste si no se regula.	Media
Modelos Híbridos (ML + proporciones)	Combinan modelos de ML con proporciones históricas para capturar patrones globales y distribuir	Capturan dinámica global, mejoran estabilidad en series ruidosas, y aseguran	Requieren diseño cuidadoso y tratamiento especial de proporciones para evitar fuga de datos.	Media

	predicciones a nivel desagregado.	coherencia entre niveles agregados y desagregados.		
ARIMA	Modelo estadístico autorregresivo que incluye términos de diferencia e integración.	Adecuado para series estacionarias, buena capacidad explicativa.	No captura estacionalidad ni relaciones exógenas.	Alta
SARIMA	Extensión de ARIMA que incluye términos estacionales.	Maneja estacionalidad explícita, útil en series con patrones cíclicos.	Supone estacionariedad, requiere análisis de residuos.	Alta

Tabla 2 Modelos utilizados en prediiciones

.

4 PREPARACIÓN DE DATOS

La presente investigación se apoya en una base de datos detallada suministrada por la empresa ABC, que contiene el registro semanal de ventas de 475 ítems clasificados en diversas categorías de productos, tales como AVR, Data Center, Inverter, Line Interactive UPS, Online UPS, Solar, Storage Batteries, Surge y UPS Accessories. El período cubierto por esta base de datos abarca desde el 4 de enero de 2021 hasta el 1 de mayo de 2024 y se encuentra organizada por orden de relevancia comercial de cada ítem para la empresa, lo cual facilita la identificación de los productos clave.

Para el desarrollo del modelo predictivo, se seleccionaron los primeros 80 ítems más representativos, los cuales concentraron el 90.24% de las ventas totales (USD 16.882.844) y el 90.06% de las utilidades (USD 5.949.037) durante el año 2023. Los ítems restantes, con menor volumen de ventas, fueron categorizados como "ventas por pedido", lo que permite una gestión más eficiente del inventario y una segmentación más adecuada de la demanda.

La distribución geográfica de las ventas refleja la presencia de ABC en seis países: México, Colombia, Guatemala, Costa Rica, Nicaragua y Miami (EE. UU.). La Tabla 2 resume el comportamiento de ventas de los 80 ítems en estas seis regiones.

País	Conteo de ventas	USD
México	131,749	6,441,879
Colombia	31,501	2,172,075
Costa Rica	68,168	3,353,242
Guatemala	65,588	1,622,381
Miami	34,021	1,632,526
Nicaragua	21,059	788,953

Tabla 3 Ventas totales por país 2023

Ventas totales 2023 (475 ítems): USD 18,708,542

Utilidades totales 2023 (475 ítems): USD 6,324,257

Del total de regiones, se determinó que México, Colombia y Costa Rica concentran la mayor parte de las operaciones comerciales, por lo que el análisis se focalizó en estas tres bodegas. Durante la inspección de los datos, se identificó una alta presencia de ceros en los registros de ventas, indicando demanda intermitente o ausente para varios ítems. Esta característica es especialmente relevante en la construcción de modelos predictivos, pues influye directamente en la selección de métodos adecuados para manejar series con ventas esporádicas. Además, los valores negativos que representan devoluciones fueron transformados en ceros, con el fin de preservar la consistencia analítica y evitar distorsiones en los indicadores de desempeño. Esta decisión permitió centrar el estudio en las ventas efectivas.

Como resultado del proceso de segmentación, se elaboró una base de datos separada para cada uno de los países seleccionados (Colombia, México y Costa Rica), y una base de datos consolidada que integra los tres. Esta última contiene las ventas semanales de los 80 ítems priorizados entre el 10 de enero de 2021 y el 1 de mayo de 2024.

4.1 LIMPIEZA Y FILTRADO DE VENTAS

La limpieza y preparación de los datos constituyó una etapa fundamental dentro del proceso de modelado estadístico y predictivo. Su propósito principal fue garantizar la calidad, integridad y adecuación de la información utilizada para construir modelos confiables. Un conjunto de datos con errores, valores atípicos o inconsistencias podría haber inducido sesgos, disminuido la precisión del modelo y comprometido la validez de las conclusiones. Por tanto, en este trabajo, se aplicaron técnicas ampliamente documentadas en la literatura especializada, con el objetivo de asegurar la robustez del conjunto de datos que alimentó el modelo de pronóstico de ventas.

4.1.1 Descripción del conjunto de datos

La base de datos original estuvo compuesta por registros semanales de ventas correspondientes a 80 ítems a lo largo de varios años. Su estructura incluyó una columna de fechas (Fecha), una

columna para cada ítem (Item 1, Item 2, ..., Item 80) y una columna adicional (Total) que representó la suma semanal de todas las ventas. Esta disposición permitió analizar tanto el comportamiento global como la evolución individual de cada producto.

4.1.2 Proceso de limpieza: pasos aplicados y fundamentos

A continuación, se describen las acciones realizadas para depurar el conjunto de datos, junto con su respectiva justificación técnica.

4.1.2.1 Análisis estadístico descriptivo por ítem

Se calcularon medidas como la media, mediana, moda, desviación estándar, varianza y cuartiles para cada ítem. Este análisis permitió identificar comportamientos erráticos, alta dispersión y escasa actividad comercial en ciertos productos. Según Hyndman y Athanasopoulos [13], esta etapa inicial de exploración es crucial para evaluar la viabilidad de aplicar técnicas de pronóstico sobre series temporales.

4.1.2.2 Eliminación de ítems con alta proporción de ceros

Se estimó el porcentaje de semanas con ventas nulas por cada ítem. Aquellos productos que no registraron ventas en más del 40 % de las semanas fueron eliminados. Esta decisión se fundamentó en estudios sobre demanda intermitente, que advierten que una frecuencia baja de ventas reduce la efectividad de los modelos de pronóstico [34] [33].

4.1.2.3 Tratamiento de valores negativos

Se identificaron y reemplazaron los valores negativos de ventas por ceros, bajo la consideración de que representaban errores de digitación, inconsistencias y devoluciones. Esta práctica es común en

ambientes empresariales donde las ventas negativas no son esperadas salvo que se documente lo contrario [35].

4.1.2.4 Filtro por participación relativa en ventas

Se estimó la participación promedio de cada ítem sobre el total semanal de ventas. Los productos cuya contribución fue inferior al 1 % en promedio fueron descartados, dado que su influencia en la dinámica general era marginal. Este criterio se alineó con los principios del análisis ABC y con los enfoques jerárquicos de pronóstico (top-down) [37].

4.1.2.5 Detección y tratamiento de valores atípicos

Se utilizó el rango intercuartílico (IQR) para identificar outliers. Los valores que se encontraban fuera del intervalo [Q1−1.5×IQR, Q3+1.5×IQR][Q1 - 1.5 \times IQR,\ Q3 + 1.5 \times IQR][Q1−1.5×IQR, Q3+1.5×IQR] fueron considerados atípicos y reemplazados por la mediana del ítem correspondiente. La elección de la mediana respondió a su robustez frente a valores extremos, evitando distorsiones en la estructura central de la serie [32] [38].

4.1.3 Resultado final del proceso de limpieza

Después de aplicar el proceso de limpieza, que incluyó análisis estadístico descriptivo, eliminación de ítems con alta proporción de ceros, tratamiento de valores negativos, filtrado por baja participación en ventas y detección de valores atípicos, se conservaron 45 ítems. Esta base depurada se utilizó como insumo para la construcción y evaluación de los modelos de predicción desarrollados en el proyecto.

5 METODOLOGÍA USADA EN LOS MODELOS

Este capítulo detalla el procedimiento técnico y analítico implementado para el desarrollo de los modelos de predicción de ventas semanales en la empresa ABC, considerando sus operaciones en México, Colombia y Costa Rica. La estrategia metodológica se articuló en dos fases principales: la primera se enfocó en la predicción del volumen total de ventas agregadas, mientras que la segunda se concentró en la estimación de ventas por ítem individual mediante la aplicación de un enfoque híbrido.

5.1 PREPARACION Y PREPROCESAMIENTO DE DATOS

La base del análisis fue un conjunto de datos históricos de ventas, consolidado en el archivo **ventas_3_paises.** El proceso de preparación de datos comprendió varias etapas cruciales, desde la transformación inicial hasta la depuración y generación de características. Estas actividades se resumen detalladamente en la Tabla 4

Etapa/Actividad	**Descripción**	**Detalles / Variables Generadas o Criterios Aplicados**
Fuente de Datos	Carga del conjunto de datos históricos.	Archivo: ventas_3_paises.
Transformación Inicial de Fecha	Conversión de la columna Fecha a un formato datetime estándar de Pandas.	Facilitó la manipulación temporal y la ingeniería de características.
Ingeniería de Características Temporales	Generación de nuevas columnas predictoras a partir de la variable Fecha.	Año, Mes, Día, Semana del año.

Tabla 4 Resumen de Actividades de Preparación y Procesamiento de Datos.

Para el entrenamiento y la evaluación de los modelos, el conjunto de datos se segmentó cronológicamente con el fin de preservar la estructura temporal inherente y prevenir la fuga de información (*data leakage*). Se estableció como fecha de corte principal el 7 de mayo de 2023, definiendo los siguientes subconjuntos:

- **Conjunto de Entrenamiento**: Datos registrados hasta el 7 de mayo de 2023.
- **Conjunto de Prueba**: Datos a partir del 7 de mayo de 2023 en adelante.

Esta segmentación se aplicó de manera consistente en el desarrollo de los modelos supervisados para asegurar una evaluación robusta de su capacidad de generalización.

5.2 Modelado con Enfoques Clásicos de Series de Tiempo

En una primera aproximación, se exploraron modelos univariados de series de tiempo aplicados directamente sobre la serie de ventas totales (Total). Estos modelos se entrenaron utilizando únicamente la información histórica de la propia serie, sin incorporar variables exógenas. La configuración específica para cada modelo se detalla en la Tabla 5

Modelo	**Hiperparámetros / Configuración Clave**	**Periodo de Entrenamiento**
ARIMA	Orden (p,d,q) = (1,0,1). Y (1,1,1)	Hasta 07/Mayo/2023
SARIMA	Orden no estacional (p,d,q) = (1,0,1). Orden estacional (P,D,Q,s) = (1,0,1,52).	Hasta 07/Mayo/2023

Tabla 5 Configuración de Modelos Clásicos de Series de Tiempo

5.3 Modelado con Enfoques Supervisados de Regresión

Paralelamente, se evaluaron diversos algoritmos de aprendizaje supervisado para la predicción de ventas totales. Estos modelos utilizaron las características temporales y de series de tiempo generadas en la etapa de preprocesamiento. La Tabla 6 resume las variables predictoras principales y los hiperparámetros relevantes para cada modelo.

Modelo	**Variables Predictoras Principales**	**Observaciones / Hiperparámetros Relevantes**
Regresión Lineal (con Fecha)	Año, Mes, Semana del año, Día del mes	División cronológica: Entrenamiento (70%), Validación (20%), Prueba (10%).
Regresión Lineal (con rezagos y M.M.)	lag_1, lag_2, rolling_mean_3, rolling_mean_4	**Advertencia:** Cálculo inicial sin desfase temporal, lo que resultó en fuga de datos.
Regresión Lineal Corregida	lag_1, lag_2, rolling_mean_3 (shifted), rolling_mean_4 (shifted)	Corrección del desfase en promedios móviles (desplazados una semana hacia atrás) para evitar fuga de datos. División cronológica

		estándar (Entrenamiento/Validación/Prueba) .
Árboles de Decisión	Características temporales y de series de tiempo corregidas.	Hiperparámetros por defecto de la librería scikit-learn. Sin técnicas de regularización explícitas.
Random Forest	Características temporales y de series de tiempo corregidas.	n_estimators= 124 max_depth= 7, min_samples_leaf=10, random_state=42.
XGBoost	Características temporales y de series de tiempo corregidas (excluyendo Fecha directa).	n_estimators= 98, max_depth= 3, learning_rate=0.012, reg_alpha=1.05 (L1), reg_lambda=0,8 (L2), random_state=42,

Tabla 6 Configuración de Modelos Supervisados de Regresión

Para optimizar el rendimiento del modelo Random Forest, se implementó un proceso de optimización de hiperparámetros mediante la técnica RandomizedSearchCV, la cual permite explorar combinaciones aleatorias dentro de un espacio de búsqueda definido. Esta búsqueda incluyó valores posibles para los parámetros más influyentes del modelo: el número de árboles (n_estimators entre 50 a 200), la profundidad máxima del árbol (max_depth 3, 5, 7, Ninguna) y el número mínimo de muestras por hoja (min_samples_leaf entre 1 a 11). Se evaluaron 20 combinaciones diferentes utilizando validación cruzada de 5 pliegues, lo que resultó en un total de 100 ajustes del modelo. Como resultado, la mejor configuración encontrada fue: n_estimators = 124, max_depth = 7 y min_samples_leaf = 10. Estos valores fueron seleccionados automáticamente al minimizar el error cuadrático medio durante la validación, permitiendo entrenar un modelo final optimizado que mejora la capacidad de generalización y reduce el riesgo de sobreajuste. Es importante aclarar que los métodos utilizados para la evaluación del desempeño durante el proceso de selección de hiperparámetros no tuvieron en cuenta la estructura temporal de los datos en la evaluación cruzada y selección de los k-fold.

En la siguiente tabla 7 se muestra los resultados del top 10 de las mejores combinaciones de hiperparametros para el modelo Random Forest.

n_estimators	max_depth	min_samples_leaf	Mean Test Score (neg MSE)
124	7	10	-2739287.063
98	5	9	-2744948.185
64	7	10	-2746356.57
70	7	8	-2749083.499
180	Ninguno	8	-2750049.473
166	7	8	-2752341.055
100	3	3	-2752595.093
58	3	2	-2766881.292
57	Ninguno	7	-2795509.366
179	Ninguno	6	-2830136.985

Tabla 7 Resultados validación cruzada modelo Random Forest

Para optimizar el rendimiento del modelo XGBoost, se aplicó una búsqueda de hiperparámetros mediante la técnica RandomizedSearchCV, evaluando 20 combinaciones distintas con validación cruzada de 5 pliegues, lo que resultó en un total de 100 entrenamientos. El espacio de búsqueda incluyó parámetros clave como el número de árboles (n_estimators entre 50 a 200), la profundidad máxima (max_depth de 3 a 10), la tasa de aprendizaje (learning_rate entre 0.01 a 0.3), así como los términos de regularización L1 (reg_alpha entre 0 a 2) y L2 (reg_lambda 0 a 2). El modelo óptimo identificado por este proceso fue configurado con: n_estimators = 98, max_depth = 3, learning_rate ≈ 0.01, reg_alpha ≈ 1.05 y reg_lambda ≈ 0.80. Esta combinación fue seleccionada al minimizar el error cuadrático medio durante la validación cruzada. Es importante aclarar que los métodos utilizados para la evaluación del desempeño durante el proceso de selección de hiperparámetros no tuvieron en cuenta la estructura temporal de los datos en la evaluación cruzada y selección de los k-fold.

En la siguiente tabla 8 se muestra los resultados del top 10 de las mejores combinaciones de hiperparametros para el modelo XGBoost.

n_estimators	max_depth	learning_rate	reg_alpha	reg_lambda	Mean Test Score (neg MSE)
98	3	0.012119892	1.049549321	0.799721943	-2776379.875
89	3	0.068794859	1.689067697	1.49464022	-3281229.15
64	6	0.023999699	0.912139968	1.570351923	-3299047.05
120	3	0.059580082	0.846802961	0.789763036	-3468181.35
195	7	0.180310098	1.550265647	1.878997883	-3995881.15
199	5	0.052860045	0.112823158	1.443997545	-4001912.2
86	8	0.121245476	1.459212357	1.275114943	-4072863.45
54	5	0.247052662	1.852601757	1.302154051	-4081825.1
113	9	0.069902135	0.933525786	1.719880813	-4083275.7
188	9	0.098046452	0.397684808	1.422683905	-4084788.9

Tabla 8 Resultados validación cruzada modelo XGBoost

5.4 ENFOQUE TOD_DOWN: MODELO HÍBRIDO PARA LA DESAGREGACIÓN DE VENTAS POR ÍTEM

En la segunda fase, se implementó un procedimiento de desagregación de las predicciones agregadas a nivel de ítem, siguiendo un enfoque top-down híbrido. Este procedimiento se desarrolló en tres etapas consecutivas: generación de proporciones dinámicas por ítem, aplicación de dichas proporciones sobre la serie total predicha, y evaluación cuantitativa del desempeño por producto.

A partir de los datos históricos, se calcularon las proporciones diarias de cada ítem con respecto al total de ventas, mediante la relación:

$$Proporción_{i,t} = \frac{Ventas\ del\ ítem\ i\ en\ t}{Ventas\ totales\ en\ t}$$

A diferencia del enfoque inicial que utilizaba una media móvil simple, en esta versión se optó por un promedio móvil exponencial (EMA) con un parámetro span=5.

En segundo lugar, se tomó como insumo la serie de predicciones generada por el modelo Random Forest entrenado en la fase 1, correspondiente al total de ventas, y se descompuso multiplicando cada valor predicho por la proporción móvil correspondiente a cada ítem. El resultado fue una serie de estimaciones por producto para cada fecha del conjunto de prueba.

Finalmente, se calcularon métricas de desempeño individuales para cada ítem utilizando los valores observados y sus respectivas predicciones. Las métricas consideradas fueron: **MAE** (error absoluto medio), **RMSE** (raíz del error cuadrático medio), **MAPE** (error porcentual absoluto medio) y **sMAPE** (error porcentual absoluto medio simétrico). A partir de estos indicadores, se identificaron los tres ítems con mejor ajuste, y se graficaron sus series reales y predichas para análisis visual.

6 EVALUACIÓN DE LOS MODELOS

6.1 RESULTADOS MODELADOS CON ENFOQUE CLASICOS DE SERIES DE TIEMPO

6.1.1 ARIMA

Métricas:

Modelo	RMSE	MSE	MAPE (%)	sMAPE (%)
ARIMA(1,0,1)	1,143.25	1,307,013.64	30.91	25.98
ARIMA(1,1,1)	1,144.96	1,310,924.37	30.79	25.94

Tabla 9 Métricas Modelo ARIMA

Gráficas:

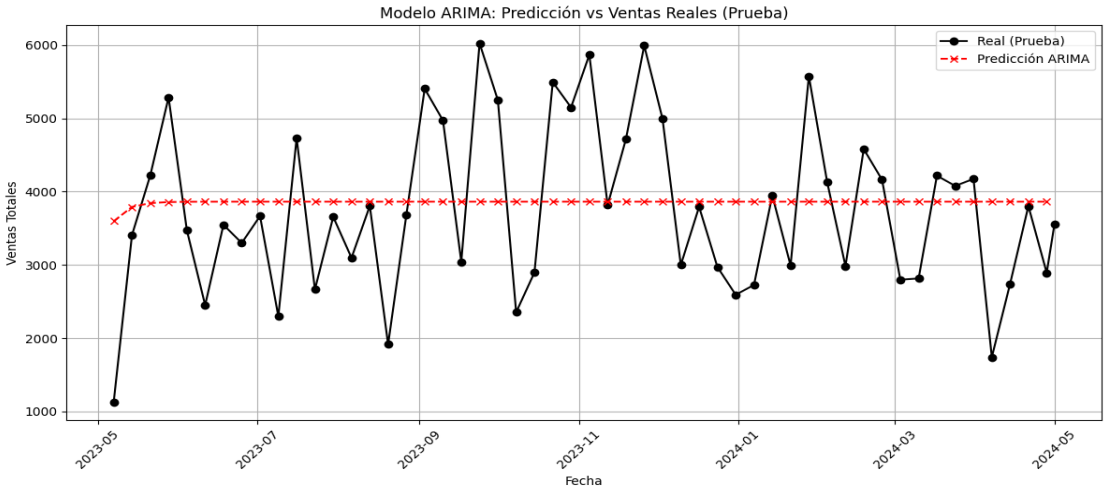


Figura 2 Modelo ARIMA (1,0,1)

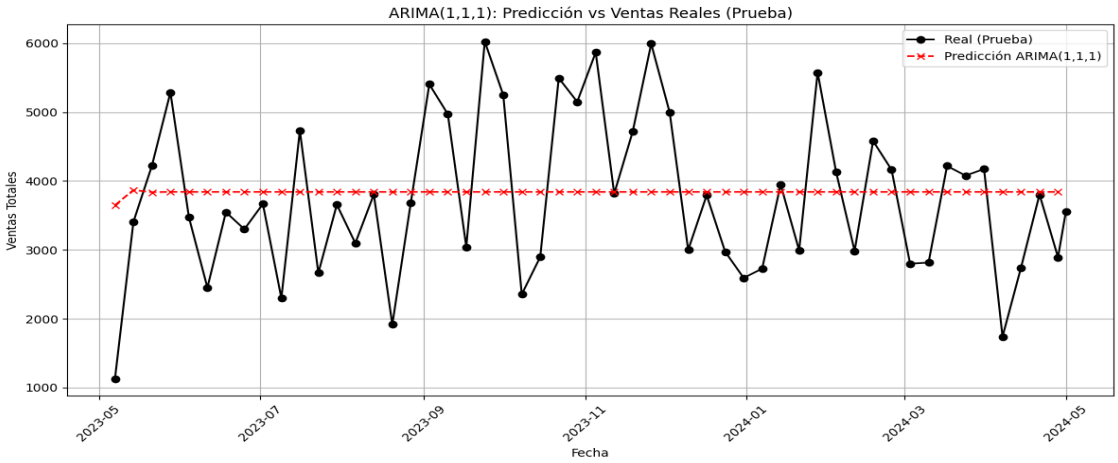


Figura 3 Modelo ARIMA (1,1,1)

Análisis:

La evaluación de los modelos ARIMA para la serie de ventas muestra que, aunque el modelo ARIMA(1,1,1) introduce una diferenciación para abordar la no estacionariedad, ofrece solo una mejora marginal en la precisión predictiva frente al modelo ARIMA(1,0,1). Las métricas MAPE y sMAPE muestran ligeras reducciones en el error porcentual, pero los valores de RMSE y MSE se mantienen elevados, lo que evidencia una débil capacidad de predicción en términos absolutos.

Ambos modelos presentan una clara deficiencia: no logran capturar la alta volatilidad y los patrones dinámicos presentes en los datos reales, como lo evidencian las gráficas de comparación. El resultado son pronósticos suavizados, cercanos a una media general, sin responder adecuadamente a picos o caídas abruptas en las ventas.

Este comportamiento sugiere la existencia de componentes complejos en la serie temporal —como estacionalidad, ciclos irregulares o la influencia de variables externas— que los modelos ARIMA básicos no están considerando.

6.1.2 SARIMA

Métricas:

Modelo	RMSE	MSE	MAPE (%)	sMAPE (%)
SARIMA(1,0,1)(1,0,1,52)	1,150.92	1,324,623.16	30.46	26.54

Tabla 10 Métricas Modelo SARIMA

Gráfica:

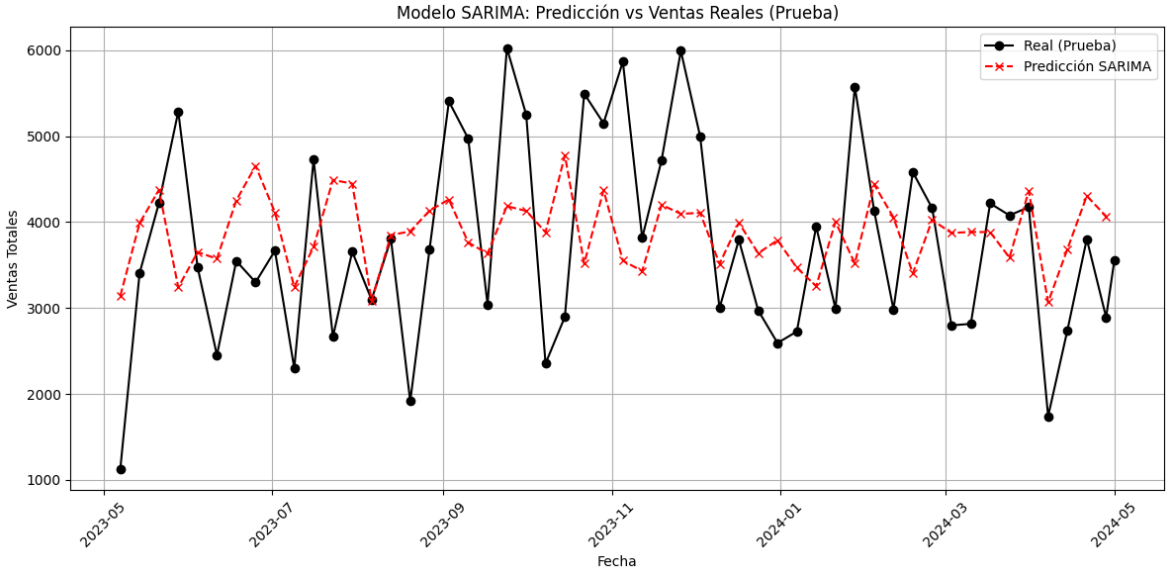


Figura 4 Modelo SARIMA

Análisis:

El modelo SARIMA(1,0,1)(1,0,1,52) fue evaluado sobre la serie de ventas con el objetivo de capturar patrones estacionales que los modelos ARIMA básicos no lograban reflejar adecuadamente. Sin embargo, los resultados indican que, a pesar de incorporar estacionalidad semanal, el modelo no presenta una mejora sustancial en comparación con los modelos ARIMA.

Las métricas obtenidas muestran un RMSE de 1,150.92 y un MAPE del 30.46%, con un sMAPE del 26.54%. Estos valores indican que el error porcentual sigue siendo elevado y que el modelo

continúa generando predicciones suavizadas que no capturan con precisión la alta volatilidad de la serie de ventas.

Este comportamiento sugiere que, si bien la estacionalidad es un componente presente en la serie, por sí sola no explica la variabilidad observada. Es probable que existan factores exógenos o relaciones no lineales que este modelo no logra representar.

Los modelos ARIMA y SARIMA demostraron ser inadecuados para capturar la complejidad de la serie de ventas. Sus errores porcentuales elevados y predicciones suavizadas confirman que la dinámica de la serie no puede explicarse únicamente por tendencia o estacionalidad.

6.2 RESULTADOS MODELADOS DE MACHINE LEARNING

6.2.1 Regresión Lineal

Métricas:

Modelo	RMSE	MSE	MAPE (%)	sMAPE (%)
Regresión Lineal	1,111.43	1,235,272.89	30.13	25.26

Tabla 11 Métricas Modelo de Regresión Lineal

Gráfica:

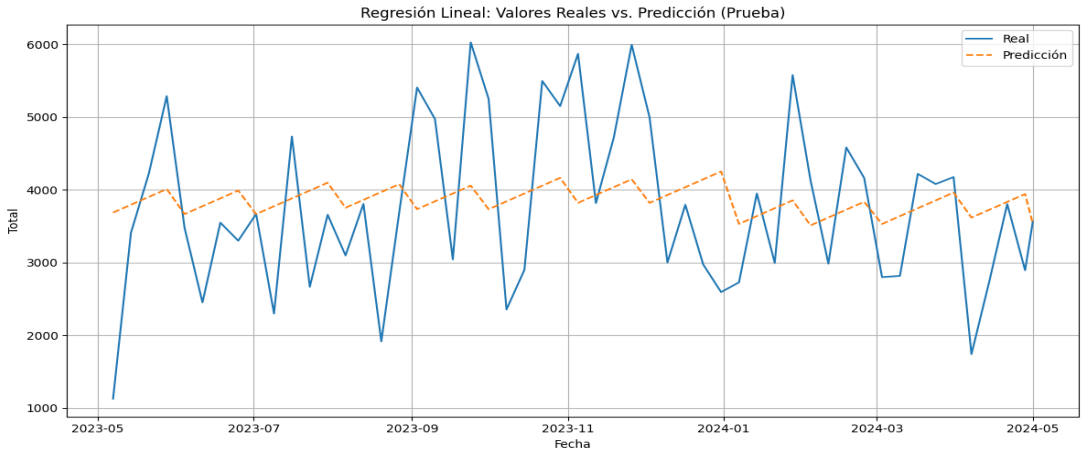


Figura 5 Modelo de Regresión Lineal

Análisis:

El modelo de regresión lineal aplicado con variables temporales extraídas de la fecha (año, mes, semana del año y día del mes) fue evaluado mediante una división cronológica del 70% para entrenamiento y 30% para prueba. Las métricas obtenidas indican una precisión limitada, con un MAPE de 30.13% y un sMAPE de 25.26%, lo que refleja una capacidad moderada del modelo para predecir el comportamiento de las ventas.

La gráfica comparativa evidencia que las predicciones generadas por el modelo son suavizadas y no logran capturar adecuadamente la alta variabilidad observada en los datos reales. Aunque las variables temporales aportan estructura, su poder explicativo resulta insuficiente frente a los cambios abruptos y la complejidad inherente de la serie.

6.2.2 Árboles de Decisión (Regresión)

Métricas:

Modelo	RMSE	MSE	MAPE (%)	sMAPE (%)
Árbol de Decisión	1,948.15	3,795,286.46	45.84	44.99

Tabla 12 Métricas Modelo de Árboles de Decisión

Gráfica:

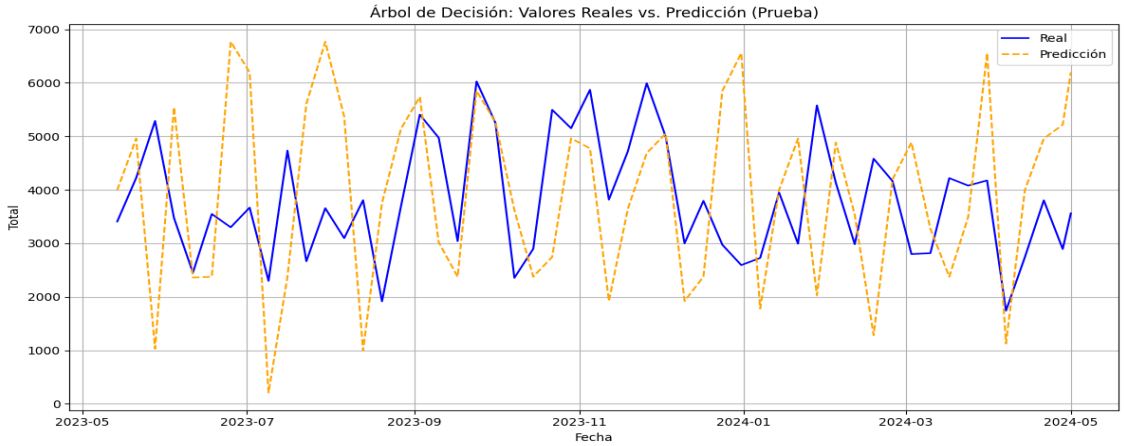


Figura 6 Modelo de Árboles de Decisión

Análisis:

Se implementó un modelo de Árbol de Decisión utilizando variables temporales derivadas de la fecha y una división basada en corte fijo: los datos registrados hasta el 7 de mayo de 2023 fueron utilizados para el entrenamiento, mientras que las observaciones posteriores formaron el conjunto de prueba.

Los resultados obtenidos muestran un desempeño deficiente, con un MAPE del 45.84% y un sMAPE del 44.99%, indicando una alta desviación de las predicciones respecto a los valores reales. La gráfica evidencia un comportamiento de predicción altamente errático, característico del sobreajuste, donde el modelo intenta memorizar el entrenamiento sin lograr una generalización adecuada.

Aunque los árboles de decisión pueden capturar relaciones no lineales, en contextos de series temporales con alta variabilidad, su desempeño puede degradarse sin una adecuada regularización o sin considerar secuencias temporales.

6.2.3 Random Forest

Métricas:

Modelo	RMSE	MSE	MAPE (%)	sMAPE (%)
Random Forest	1,048.25	1,098,828.06	24.84	22.36

Tabla 13 Métricas Modelo Random Forest

Gráfica:

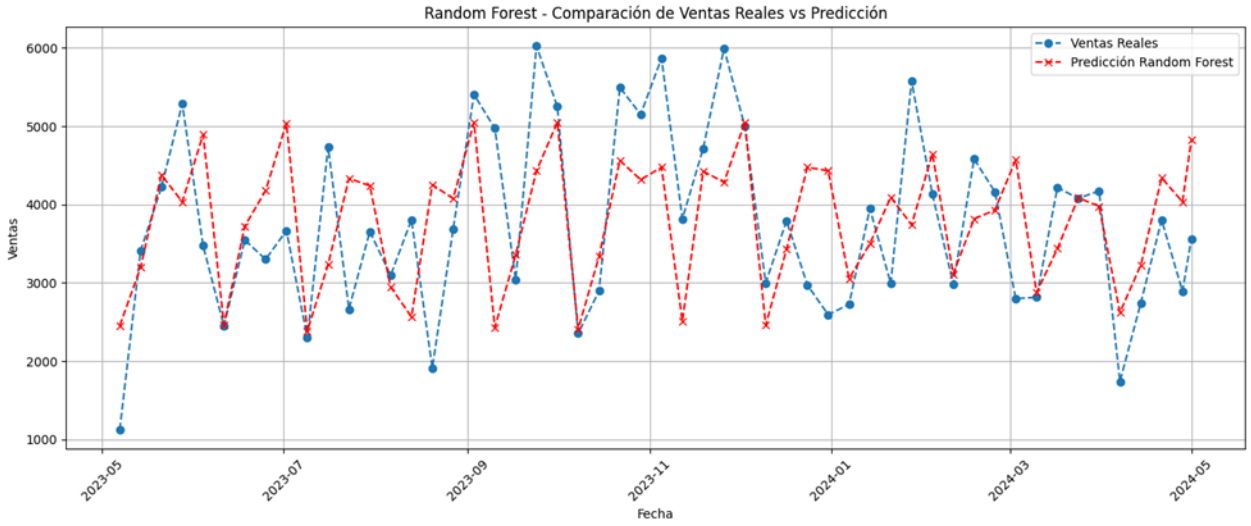


Figura 7 Modelo de Árboles de Decisión

Análisis:

Se implementó un modelo Random Forest para la predicción de ventas semanales utilizando variables temporales extraídas de la fecha (año, mes, día). El conjunto de entrenamiento incluyó los datos hasta el 7 de mayo de 2023, y el conjunto de prueba abarcó las semanas posteriores.

Las métricas obtenidas muestran un sólido desempeño del modelo, con un RMSE de 1,048.25, un MAPE de 24.84 % y un sMAPE de 22.36 %. Estos resultados reflejan una buena capacidad del modelo para aproximarse a los valores reales, tanto en términos absolutos como relativos.

El Random Forest demostró ser robusto ante la variabilidad de los datos y eficiente en términos de generalización, sin requerir una optimización extensa de hiperparámetros. Por ello, se posiciona como una alternativa fuerte para problemas de pronóstico con características similares.

6.2.4 XGBoost

Métricas:

Modelo	RMSE	MSE	MAPE (%)	sMAPE (%)
XGBoost	1,229.15	1,510,809.72	28.64	26.12

Tabla 14 Métricas Modelo XGBoost

Gráfica:

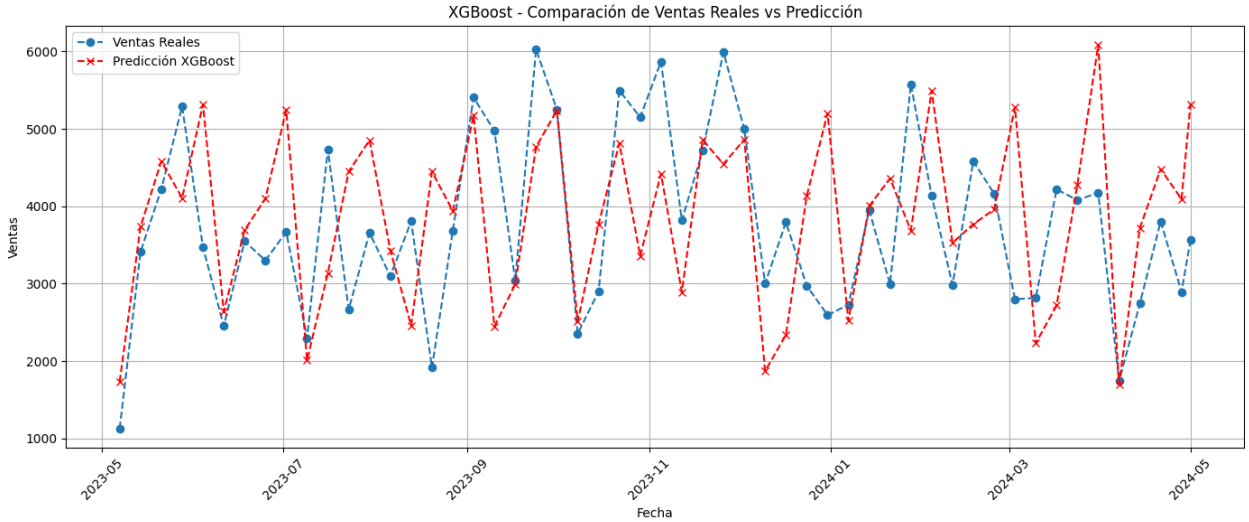


Figura 8 Modelo XGBoost

Análisis:

El modelo XGBoost fue evaluado bajo las mismas condiciones que Random Forest, utilizando como predictores las variables temporales derivadas de la fecha. El conjunto de entrenamiento incluyó los registros hasta el 7 de mayo de 2023 y la evaluación se realizó sobre las semanas posteriores.

Los resultados muestran un desempeño inferior al de Random Forest, con un RMSE de 1,229.15, un MAPE de 28.64 % y un sMAPE de 26.12 %. Aunque XGBoost ofrece ventajas en términos de regularización y prevención del sobreajuste, en este escenario no logró superar a Random Forest.

6.3 SELECCIÓN DEL MODELO

Métrica	Random Forest	XGBoost
MSE	1098828.06	1510809.72
RMSE	1048.25	1229.15
MAPE	24.84%	28.64%
SMAPE	22.36%	26.12%

Tabla 15 Comparación de Métricas de Desempeño Modelos Random Forest y XGBoost

En la fase inicial del modelado, se seleccionaron los algoritmos Random Forest y XGBoost para la predicción del total de ventas semanales, debido a que obtuvieron las mejores métricas de error y los resultados gráficos más consistentes frente a otros modelos evaluados. Ambos fueron considerados apropiados para esta etapa final por su capacidad para trabajar con variables temporales y adaptarse a la estructura de los datos.

6.3.1 Evidencia Cuantitativa: Superioridad en Métricas de Error

Los resultados obtenidos muestran que Random Forest superó a XGBoost en todas las métricas analizadas:

- Random Forest: MAPE de 24.84 %, RMSE de 1,048.25
- XGBoost: MAPE de 28.64 %, RMSE de 1,229.15

La ventaja también se refleja en los valores de MSE y sMAPE, lo que indica que Random Forest generó predicciones más precisas y cercanas a los valores reales.

6.3.2 Justificación Cualitativa y operativa

La selección final también considera aspectos operativos clave, resumidos en la Tabla 16:

- Robustez y simplicidad de ajuste: Random Forest mostró buen desempeño con una configuración básica, mientras que XGBoost fue más sensible a la necesidad de ajuste fino.
- Control del sobreajuste: Random Forest generalizó mejor en los datos de prueba. XGBoost presentó indicios de sobreajuste bajo la configuración utilizada.
- Eficiencia e interpretabilidad: Random Forest ofrece entrenamiento más rápido y una estructura más fácil de comprender en comparación con el proceso secuencial de boosting de XGBoost.

Criterio	Random Forest (RF)	XGBoost (XGB)
Precisión (cuando está bien ajustado)	Superior en este caso, menor MSE, MAPE y SMAPE	Buena, pero menos precisa en este caso específico
Velocidad de entrenamiento	Más rápido y simple	Más lento debido a su naturaleza secuencial (boosting)
Tolerancia a ruido/datos faltantes	Robusto, aunque algo sensible a outliers	Mejor manejo, gracias a técnicas internas de regularización
Overfitting	Controlado por max_depth y n_estimators; buena generalización observada	Regularización interna (reg_alpha, reg_lambda), pero sobreajustó ligeramente en este caso
Tuning (ajuste de parámetros)	Menos sensible; rendimiento razonable con pocos ajustes	Requiere tuning para mejorar resultados, altamente dependiente de hiperparámetros
Interpretabilidad	Fácil de interpretar (árboles simples visibles)	Más complejo, especialmente con boosting secuencial

Tabla 16 Comparación Beneficios de Modelos Random Forest y XGBoost

6.4 ANÁLISIS COMPARATIVO DE MODELOS: XGBOOST VS RANDOM FOREST

Para evaluar si existía una diferencia estadísticamente significativa en el rendimiento de los modelos XGBoost y Random Forest, se llevaron a cabo 10 ejecuciones independientes de cada uno. En cada ejecución se utilizaron combinaciones distintas de hiperparámetros seleccionados aleatoriamente. Esta estrategia buscó representar un abanico más amplio de posibles escenarios de entrenamiento y evitar sesgos asociados a una única configuración de modelo.

La métrica empleada para comparar el rendimiento fue el MAPE (Mean Absolute Percentage Error), ampliamente utilizada en contextos de predicción de ventas debido a su capacidad para capturar el error relativo de forma interpretable. Dado que los valores de MAPE no siguen necesariamente una distribución normal, se aplicó la prueba de Wilcoxon para muestras pareadas, un método no paramétrico ideal para comparar dos muestras relacionadas.

El resultado de la prueba de Wilcoxon arrojó un estadístico $W = 0.0$ y un valor $p = 0.00195$. Este valor p es significativamente inferior al umbral comúnmente utilizado de 0.05, lo que indica evidencia sólida para rechazar la hipótesis nula. Por lo tanto, se concluye que existe una diferencia estadísticamente significativa en el desempeño entre ambos modelos.

En todas las ejecuciones, Random Forest presentó valores de MAPE consistentemente menores que XGBoost. Esta diferencia no solo es estadísticamente significativa, sino también práctica, ya que implica una mejor capacidad del modelo para realizar predicciones más precisas en términos relativos. En el contexto de la predicción de ventas semanales con base únicamente en variable temporales, el modelo Random Forest demuestra ser más robusto y confiable.

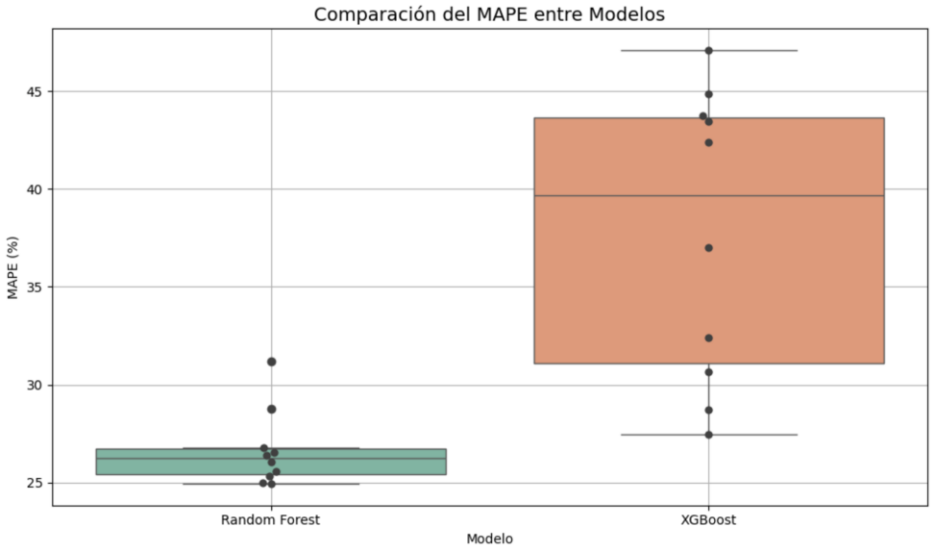


Figura 9 Comparación de MAPE[2] entre XGBoost y Random Forest

La Figura 9 proporciona una representación visual complementaria de los valores de MAPE para ambos modelos. Como se observa, el modelo Random Forest no solo obtiene valores de MAPE inferiores en promedio, sino que además presenta una menor dispersión en su rendimiento. En cambio, XGBoost muestra una mayor variabilidad y una mediana más alta de error porcentual. Esta evidencia gráfica refuerza los hallazgos estadísticos, destacando la ventaja competitiva del modelo Random Forest tanto en precisión en la consistencia para la predicción de las ventas.

6.5 RESULTADOS MODELO HÍBRIDO （Random Forest + Proporción Media Exponencial）

Métricas:

Item	R^2	RMSE	MAPE (%)	sMAPE (%)
Item 2	0.36083	277.41190	35.41289	32.97303
Item 1	0.21558	298.38540	39.67502	35.96576
Item 3	0.05382	317.74160	43.32424	36.33831
Item 10	0.57883	193.89900	46.61491	50.57910
Item 9	0.65825	24.66077	75.91813	76.85385
Item 62	0.46030	42.97635	77.19903	83.15309
Item 5	0.42420	21.78475	84.82753	57.40725
Item 39	0.33064	54.26408	85.03282	60.08879
Item 22	0.46230	52.42906	86.00951	71.24552
Item 49	0.56327	43.36834	95.46860	79.70778
Item 6	0.50286	43.64670	95.81626	53.10888
Item 24	0.40454	220.95410	107.75400	69.49027
Item 15	0.37881	65.13882	137.36590	68.97287
Item 30	0.54578	43.61881	151.94850	64.68657
Item 11	0.39480	184.08910	288.98980	67.14864

Tabla 17 Métricas Modelo Híbrido Top-Down por Item

Análisis:

El modelo híbrido, permitió evaluar la capacidad del enfoque top-down para distribuir las predicciones del total de ventas entre los diferentes ítems, basados en las predicciones obtenidas con el modelo Random Forest. Como se observa en la tabla 17 de métricas, los mejores resultados se obtuvieron en ítems con una participación más estable.

El mejor desempeño se alcanzó con el Item 2 (MAPE de 35,41 %), seguido por el Item 1 y el Item 3. Aunque los errores son aún significativos, estos tres productos representan los casos en los que la proporción suavizada se ajustó mejor al comportamiento real.

Otros ítems, como el Item 9 y el Item 11, mostraron errores extremadamente altos, lo que evidencia limitaciones del modelo para capturar ítems con alta variabilidad, estacionalidad no explícita o posibles outliers.

Es posible afirmar que el modelo híbrido con EMA representa un avance respecto a métodos más simples al capturar tendencias recientes, pero su capacidad predictiva es aún limitada para productos individuales sin considerar factores adicionales. Se sugiere complementar este enfoque con modelos más complejos que integren variables exógenas o estructuras no lineales específicas.

7 CONCLUSIONES Y TRABAJOS FUTUROS

7.1 CONCLUSIONES

El presente trabajo permitió comprobar que la aplicación de técnicas de aprendizaje automático en contextos logísticos reales, como la gestión de inventarios, es no solo viable, sino estratégica cuando se cuenta con datos históricos estructurados y un enfoque metodológico riguroso. La formulación precisa del problema, junto con el cumplimiento de los objetivos generales y específicos planteados, permitió desarrollar una solución predictiva aplicable y replicable en la operación de la empresa ABC.

- El modelo desarrollado respondió de forma efectiva a la pregunta de investigación, cumpliendo con el objetivo de construir un sistema de pronóstico de ventas semanales capaz de mejorar la planificación logística en las bodegas de México, Colombia y Costa Rica. Se logró integrar una solución analítica que permite anticipar la demanda y reducir el impacto negativo del exceso de inventario.

- Entre los modelos evaluados, el algoritmo Random Forest fue el que presentó el mejor desempeño predictivo para la serie temporal de ventas totales. El modelo final alcanzó un RMSE de 1048.25, un MAPE ajustado de 24.84% y un SMAPE de 22.36% en el conjunto de prueba. Estos resultados superaron a los obtenidos por modelos tradicionales y por otros enfoques supervisados, validando así su selección como modelo definitivo.

- La distribución semanal de las predicciones permitió preservar la coherencia temporal y evitar fugas de información, asegurando la robustez del entrenamiento y de las métricas de validación. Esto fue clave para garantizar que los resultados fueran realistas y aplicables en el entorno operativo real de la empresa.

- El proceso de limpieza, preparación y transformación de variables fue determinante para asegurar la calidad del conjunto de datos. La extracción de atributos temporales derivados de la

fecha (año, mes y día) resultó útil para capturar patrones subyacentes, y la segmentación geográfica por país permitió al modelo adaptarse a dinámicas comerciales específicas.

- Como resultado del proyecto, la empresa ABC cuenta ahora con una herramienta analítica capaz de respaldar decisiones de planeación de la demanda con mayor precisión. Su aplicación práctica puede traducirse en una reducción significativa de costos logísticos, mejor rotación de inventario y optimización del capital de trabajo, fortaleciendo así su eficiencia operativa y competitividad en el mercado.

7.2 TRABAJOS FUTUROS

A partir de los resultados alcanzados y de las oportunidades detectadas durante el desarrollo del presente proyecto, se plantean las siguientes líneas de trabajo futuro con el objetivo de mejorar, escalar y consolidar el sistema de predicción implementado:

- Se recomienda la implementación de un pipeline automatizado de MLOps que permita el reentrenamiento periódico del modelo Random Forest con datos actualizados. Esta estrategia asegurará que el modelo se mantenga vigente ante posibles cambios en los patrones de demanda. El pipeline debe contemplar etapas de ingestión, limpieza, validación y despliegue automático, además de un sistema de monitoreo del rendimiento que permita detectar señales de concept drift y ejecutar acciones correctivas, como la reconfiguración o sustitución del modelo en producción.

- Para facilitar la implementación operativa del modelo en los procesos internos de la empresa, se sugiere avanzar en su integración con los sistemas de gestión empresarial (ERP). Esta conexión permitiría automatizar flujos clave como la consulta de pronósticos, la actualización dinámica de niveles de inventario, la generación de alertas tempranas ante riesgos de sobreabastecimiento o desabastecimiento, y la sincronización con los calendarios de compras y abastecimiento. Esta integración consolidaría un ecosistema de toma de decisiones asistido por datos, generando valor operativo de forma continua.

8 BIBLIOGRAFÍA

[1] T. M. Mitchell, Machine Learning, New York, NY, USA: McGraw-Hill, 1997.

[2] C.Sandoval, Introducción al Machine Learning, Lima, Perú: Editorial Cientifica, 2017.

[3] W. H. T. James, An Intrduction to Statistical Learning, 2nd ed., New York, NY, USA: Springer, 2021.

[4] R. J. H. a. G. Athanasopoulos, Forecasting: Priciples and Practice, 2nd ed., Melbourne, Australia: OTexts, 2018.

[5] R. D. P.J. Brockwell, Introduction to Time Series and Forecasting, 3rd ed., New York, USA: Springer, 2016.

[6] A. B. R.S. Sutton, Reinfrocement Learning: An Introduction, 2nd ed, Cambridge, MA, USA: MIT, 2018.

[7] Y. B. a. A. C. I. Goodfellow, Deep Learning, Cambridge, MA, USA: MIT, 2016.

[8] R. T. a. J. F. T. Hastie, Data Mining, Inference, and Prediction, 2nd ed., New York, NY, USA, 2009.

[9] L. Breiman, "Random Forest," *Mach,* vol. 45, no. 1, pp. 5 - 32, 2001.

[10] T. C. a. C. Guestrin, "XGBoost: A scalable tree boosting system," in *22nd ACM SIGKDD int. Conf. Knowl. DIcov. Data Mining* , San Francisco, CA, USA.

[11] D. S. R.H. Shumway, Time Series Analysis and Its Applications: With R Examples, 4th ed, New York, NY, USA, 2017.

[12] D. W. T. H. a. R. T. G. James, An Introduction to Statistical Learning with Applications in R, 2nd ed., New York, NY, USA: Springer, 2021.

[13] R. J. H. a. G. Athanasopoulos, Forecasting: Principles and Practice, 3rd ed., Melbourne, Australia , 2021.

[14] R. K. a. M. Ross, The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling, 3rd ed., Wiley, 2013.

[15] T. C. a. C. Guestrin, " "XGBoost: A scalable tree boosting system," i," *22nd ACM SIGKDD, ,* no. San Francisco, USA, p. 785–794, 2016.

[16] G. Lapide, "“Forecasting hierarchies”," *Journal of Business Forecasting,,* vol. 30, no. 2, p. 18–26, 2011.

[17] M. G. Sahin and E. Dincer, " “Inventory Forecasting and Demand Classification in the Supply Chain,”," *Journal of Supply Chain Management,* vol. 9, no. 3, pp. 17 - 29, 2015.

[18] F. P. a. V. A. E. Spiliotis, "“Forecasting with decomposition: A total and split approach to forecasting hierarchical and grouped time series,”," *International Journal of Forecasting,,* vol. 36, no. 4, pp. 1348 - 1366, 2020.

[19] J. S. Armstrong and F. Collopy, "Error measures for generalizing about forecasting methods: Empirical comparisons,," *International Journal of Forecasting,* vol. 8, no. 1, pp. 69-80, 1992.

[20] P. G. a. R. Lawton, "On the asymmetry of the symmetric MAPE," *International Journal of Forecasting,* vol. 15, no. 4, p. 405–408, 1999.

[21] S. K. a. H. Kim, "A new metric of absolute percentage error for intermittent demand forecasts,," *International Journal of Forecasting,* vol. 32, no. 3, pp. 669 - 679, 2016.

[22] E. S. a. V. A. S. Makridakis, "The M5 competition: Background, organization, and implementation,," *International Journal of Forecasting,* vol. 38, no. 4, pp. 1325 - 1336, 2022.

[23] S. S. a. N. J. Castellan, Nonparametric Statistics for the Behavioral Sciences, 2nd ed., New York: McGraw-Hill, 1988.

[24] F. Wilcoxon, ""Individual comparisons by ranking methods,"," *Biometrics Bulletin,* vol. 1, no. 6, pp. 80-83, 1945.

[25] W. J. Conover, Practical Nonparametric Statistics, 3rd, ed. New York: Wiley,, 1999.

[26] J. D. Gibbons and S. Chakraborti, Nonparametric Statistical Inference, 5th ed., Boca Raton: CRC Press, 2010.

[27] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA,, 2016.

[28] C. S. T. a. T. Turkmen, "Demand Forecasting with Machine Learning," Capstone Project, Massachusetts Institute of Technology, Jun 2024. [Online]. Available: https://ctl.mit.edu/sites/ctl.mit.edu/files/theses/Demand%20Forecasting%20with%20Machine%20Learning.pdf.

[29] V. V. F. W. M. S. M. E. R. R. M. N. B. L. B. F. A. H. y. R. V. L. C. Boldt, "Forecasting Nike's Sales using Facebook Data," in *Proceedings of the 2016 IEEE International Conference on Big Data (Big Data)*, Washington, DC, USA, 2016.

[30] M. Schmitt, "Automated Machine Learning: AI-Driven Decision Making in Business Analytics," *Intelligent Systems with Applications,* vol. 18, no. 200188, 2023.

[31] B. Lim, "Deep Learning for Time Series Prediction & Decision Making Over Time," University of Oxford, Oxford, UK, 2020. [Online]. Available: https://ora.ox.ac.uk/objects/uuid:38843c7a-2123-4b5d-8454-2ae35a7ed6ce.

[32] B. Usta, "Implementación de un método para el pronóstico de demanda de medicamentos,," Trabajo de grado, Universidad Nacional de Colombia, 2021. [Online]. Available: https://repositorio.unal.edu.co/handle/unal/84189.

[33] S. J. Taylor y B. Letham, "Forecasting at Scale," *The American Statistician,* vol. 72, no. 1, p. 37–45, 2018.

[34] V. A. a. S. M. E. Spiliotis, "Total and split: A simple hybrid approach for hierarchical forecasting,," *J. Oper. Res. Soc.,* vol. 71, no. 12, pp. 1927 - 1939, 2020.

[35] D. C. Montgomery, Montgomery, Design and Analysis of Experiments, 9th ed., Hoboken, NJ, USA: Wiley, 2017.

[36] J. C. R. K. T. K. T. R. C. S. y. R. W. P. Chapman, "CRISP-DM 1.0: Step-by-Step Data Mining Guide," CRISP-DM Consortium, SPSS Inc, 2000. [Online]. Available: https://mineracaodedados.files.wordpress.com/2012/12/crisp-dm-1-0.pdf.

[37] A. Géron, Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd ed, Sebastopol, CA, USA: O'Reilly Media, 2019.

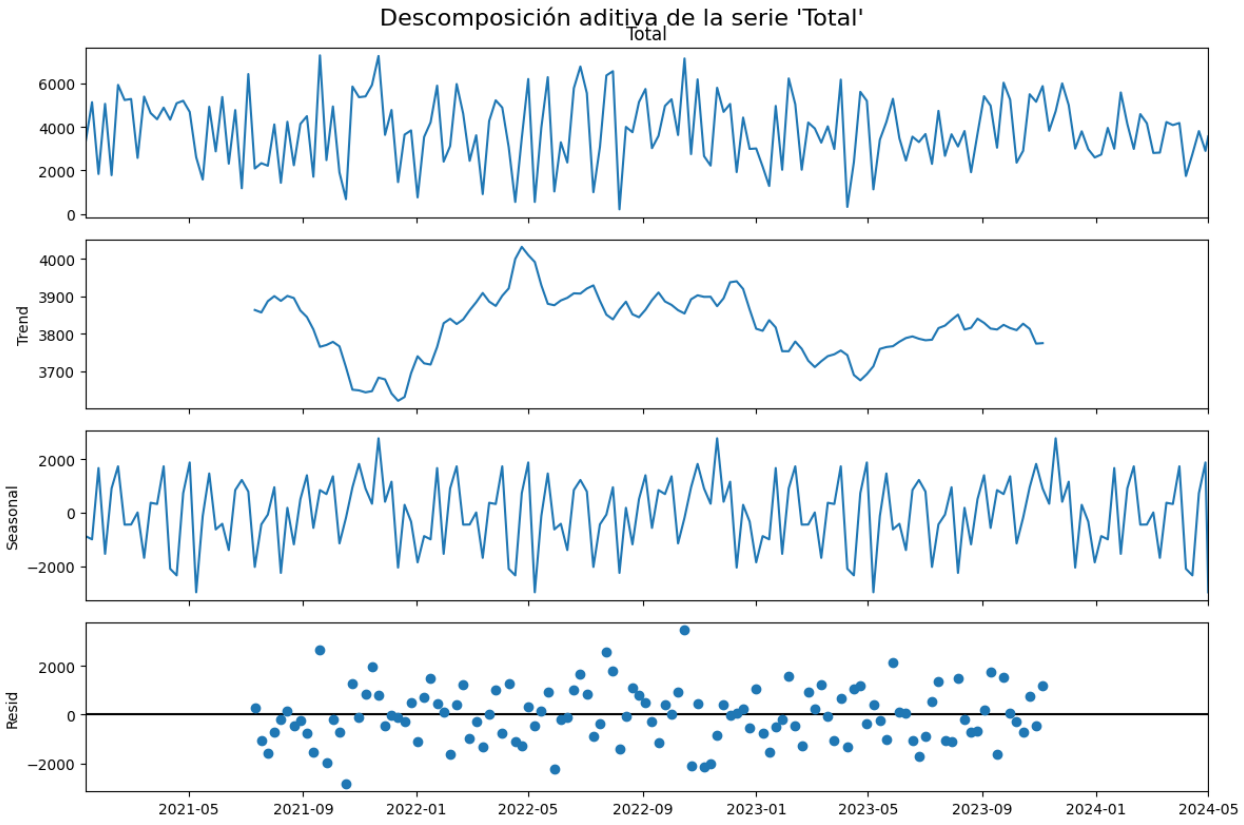
[38] F. K. y. J. M. G. C. Schröer, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Computer Science,* vol. 181, p. 526–534, 2021.

[39] J. Ngugi, "Handling Missing Values in Data Science," *Medium,* 2021.

[40] M. K. y. J. P. J. Han, Data Mining: Concepts and Techniques, 3ª ed., Waltham, MA, EE. UU: Morgan Kaufmann, 2011.

9 ANEXOS

Anexo 1 Pruebas de estacionalidad, estacionaridad y tendencia variable Total

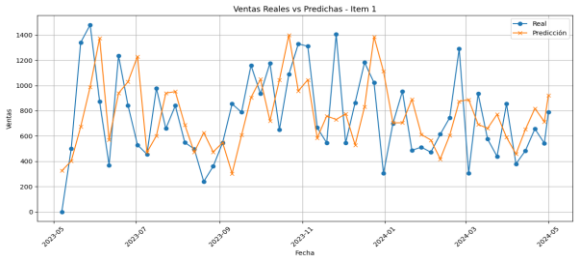


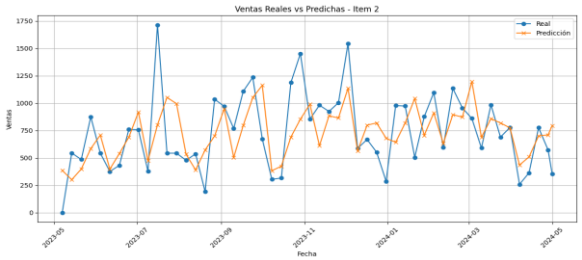
Anexo 2 Predicciones por item usando modelo hibrido.

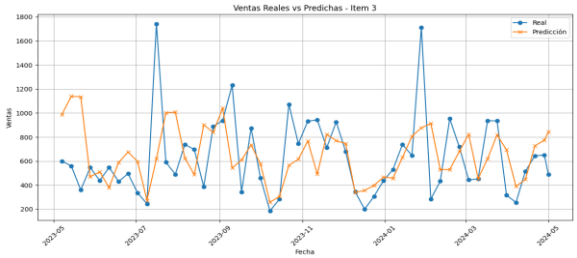
Fecha	Total Real	Total Predicho
2023-05-07 00:00:00	1127	1731.759888
2023-05-14 00:00:00	3408	3735.173096
2023-05-21 00:00:00	4223	4581.155762
2023-05-28 00:00:00	5288	4107.030273
2023-06-04 00:00:00	3474	5315.755371
2023-06-11 00:00:00	2451	2638.080566
2023-06-18 00:00:00	3548	3690.937744
2023-06-25 00:00:00	3300	4100.184082
2023-07-02 00:00:00	3668	5244.150391
2023-07-09 00:00:00	2298	2008.410645
2023-07-16 00:00:00	4733	3138.990234
2023-07-23 00:00:00	2665	4447.908203
2023-07-30 00:00:00	3657	4853.380859
2023-08-06 00:00:00	3097	3430.83252
2023-08-13 00:00:00	3806	2459.459229
2023-08-20 00:00:00	1915	4447.908203
2023-08-27 00:00:00	3686	3939.799072
2023-09-03 00:00:00	5406	5170.597656
2023-09-10 00:00:00	4973	2448.752441
2023-09-17 00:00:00	3042	2992.301758
2023-09-24 00:00:00	6027	4767.030273
2023-10-01 00:00:00	5250	5238.310547
2023-10-08 00:00:00	2354	2511.095703
2023-10-15 00:00:00	2898	3766.915527
2023-10-22 00:00:00	5496	4814.344727
2023-10-29 00:00:00	5151	3357.008301
2023-11-05 00:00:00	5870	4420.750977
2023-11-12 00:00:00	3817	2891.420166
2023-11-19 00:00:00	4718	4856.144531
2023-11-26 00:00:00	5994	4545.558105
2023-12-03 00:00:00	4996	4862.80957
2023-12-10 00:00:00	2999	1869.728027
2023-12-17 00:00:00	3794	2337.609863
2023-12-24 00:00:00	2973	4134.109375
2023-12-31 00:00:00	2593	5203.809082
2024-01-07 00:00:00	2727	2523.462891
2024-01-14 00:00:00	3949	4003.025879

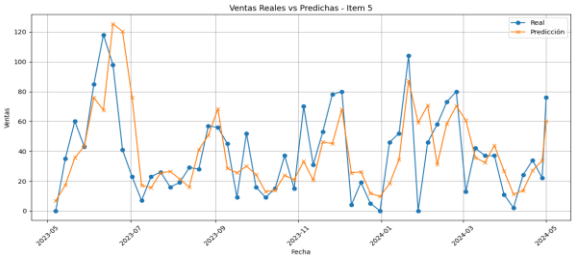
2024-01-21 00:00:00	2993	4360.116211
2024-01-28 00:00:00	5578	3687.688232
2024-02-04 00:00:00	4132	5496.040527
2024-02-11 00:00:00	2984	3528.032471
2024-02-18 00:00:00	4582	3768.333008
2024-02-25 00:00:00	4163	3960.307617
2024-03-03 00:00:00	2799	5281.713867
2024-03-10 00:00:00	2816	2229.976563
2024-03-17 00:00:00	4219	2718.82373
2024-03-24 00:00:00	4079	4281.60498
2024-03-31 00:00:00	4175	6086.1875
2024-04-07 00:00:00	1740	1697.718506
2024-04-14 00:00:00	2743	3720.921143
2024-04-21 00:00:00	3804	4480.546875
2024-04-28 00:00:00	2893	4092.779053
2024-05-01 00:00:00	3561	5315.755371

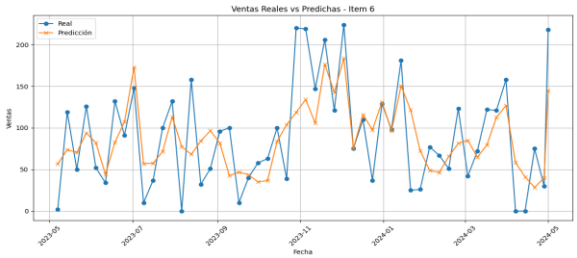
Anexo 3 Gráficas de las mejores 15 predicciones por item

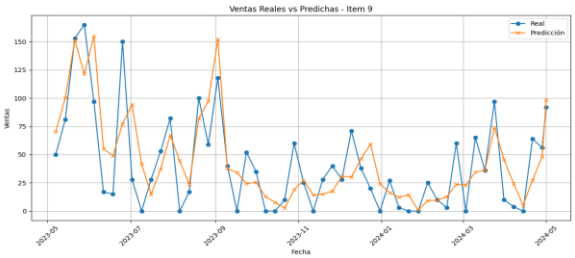


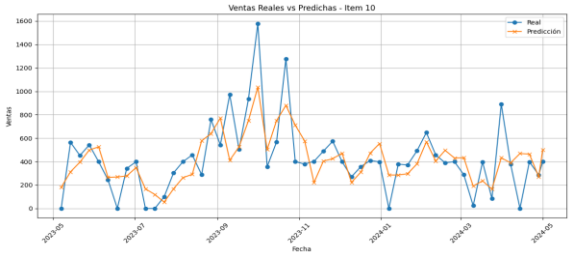


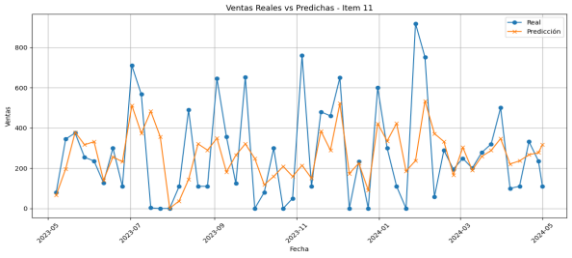


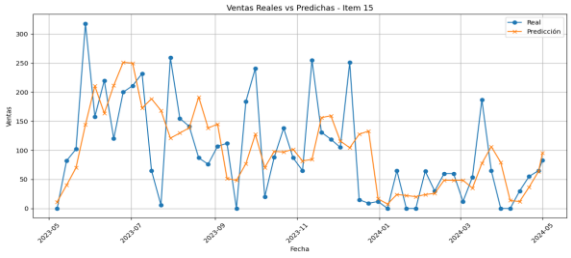


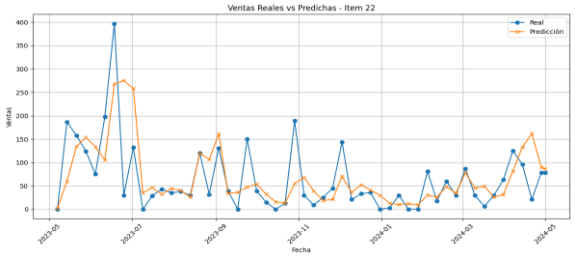


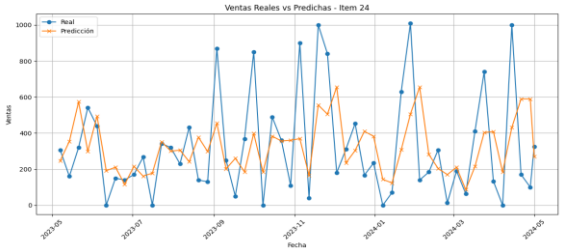


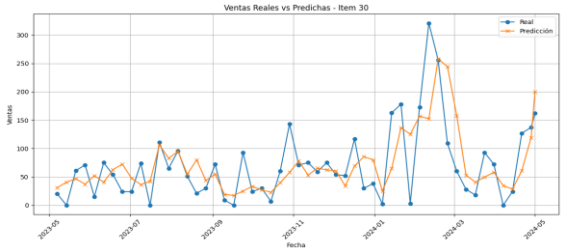


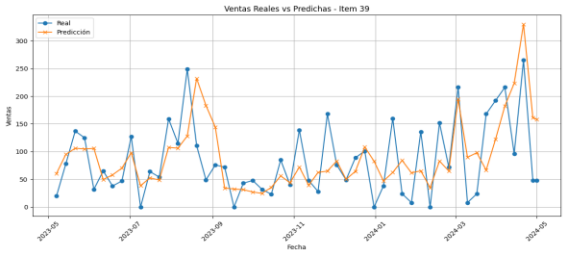


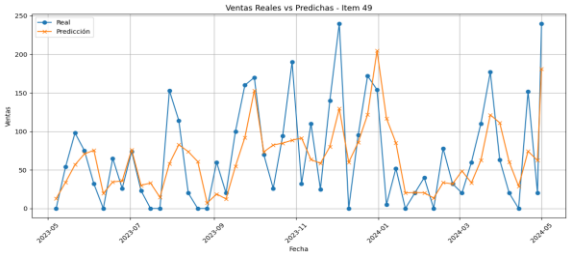


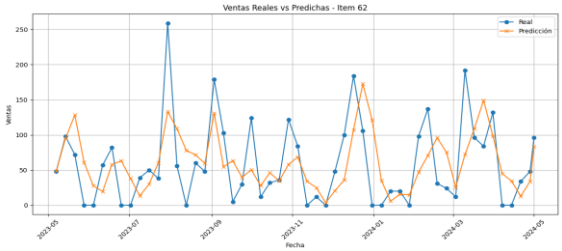



Ventas Reales vs Predichas - Item 24
Real
Predicción
Ventas
Fecha


Ventas Reales vs Predichas - Item 30
Real
Predicción
Ventas
Fecha


Ventas Reales vs Predichas - Item 39
Real
Predicción
Ventas
Fecha


Ventas Reales vs Predichas - Item 49
Real
Predicción
Ventas
Fecha


Ventas Reales vs Predichas - Item 62
Real
Predicción
Ventas
Fecha