Pontificia Universidad
JAVERIANA
Cali

Modelo de pronósticos de demanda para la optimización de la cadena de suministros en supermercados MEGATIENDAS

Aldair Blanco Segura

Angie Cantillo Martinez

Proyecto Aplicado para optar al título de

Magister en Ciencia de Datos

Director

Cristhian Kaori Valencia Marin

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, DICIEMBRE 05 DEL 2024

FICHA RESUMEN

TÍTULO: Modelo de pronósticos de demanda para la optimización de la cadena de suministros MEGATIENDAS.

1. ÁREA DE TRABAJO: Ciencia de Datos

2. TIPO DE PROYECTO (Aplicado, Innovación, Investigación): Aplicado

3. ESTUDIANTE(S): Aldair Blanco Segura y Angie Cantillo Martinez

4. CORREO ELECTRÓNICO: ajblanco@javerianacali.edu.co, cantilloaa@javerianacali.edu.co.

5. DIRECTOR: Cristhian Kaori Valencia Marin

6. VINCULACIÓN DEL DIRECTOR:

7. CORREO ELECTRÓNICO DEL DIRECTOR: ckvalencia@utp.edu.co

8. GRUPO O EMPRESA QUE LO AVALA (Si aplica): Supermercados Megatiendas

9. PALABRAS CLAVE: Almacén, Gestión de inventarios, pronósticos, demanda.

10. FECHA DE INICIO: 01/2024

11. DURACIÓN ESTIMADA: 11 meses

RESUMEN

Megatiendas, la cadena de supermercados, se enfrentó a desafíos logísticos significativos en la gestión del suministro para las 25 tiendas que tiene abiertas a la fecha. Debido a la inexactitud en los pronósticos, la cadena enfrentaba compras excesivas, desperdicio de productos perecederos y un aumento en los costos operativos. El objetivo central de este proyecto fue desarrollar e implementar un modelo de aprendizaje automático que permitiera prever la demanda diaria de manera precisa, optimizando la gestión de inventarios. El proceso comenzó con la extracción, transformación y limpieza de datos de ventas históricos, lo cual se logró mediante la creación de un módulo en R para automatizar la gestión de datos. Se extrajeron variables clave, como ventas por tienda, producto y promociones, para garantizar la calidad de los datos empleados en los modelos predictivos. El modelo se construyó utilizando una combinación de métodos estadísticos y de aprendizaje automático, centrándose en tres enfoques principales: ARMA, SARIMAX y Gradient Boosting. Tras la implementación y evaluación de estos modelos, se comprobó que Gradient Boosting superó significativamente a los modelos ARMA y SARIMAX en términos de precisión, mostrando los mejores resultados en todas las métricas clave, como el MAE, RMSE y R^2. Esto permitió a Megatiendas mejorar sus pronósticos, alcanzando un nivel de precisión superior al 85%. Los resultados de Gradient Boosting no solo ayudaron a reducir el exceso de inventario y el desperdicio de productos, sino que también mejoraron la toma de decisiones estratégicas en la cadena de suministro. El análisis de resultados confirmó que la correcta implementación de un modelo de aprendizaje automático optimiza la cadena de suministro, permitiendo a Megatiendas gestionar de manera más eficiente sus productos frescos, reducir costos y mejorar la disponibilidad de los mismos. El proyecto no solo ofrece una solución inmediata a los problemas de Megatiendas, sino que también establece un marco de trabajo para futuras mejoras, con recomendaciones para incorporar más variables y datos en el modelo para aumentar su precisión y adaptabilidad. En resumen, el desarrollo de este modelo de pronósticos de demanda representa un paso clave para optimizar la gestión de inventarios en la cadena de supermercados Megatiendas, mejorando su eficiencia operativa y su competitividad en el mercado.

Tabla de contenido

LISTA DE FIGURAS

LISTA DE TABLAS

LISTA DE ANEXOS

INTRODUCCIÓN

Megatiendas, la cadena de supermercados, enfrenta desafíos significativos en su cadena de suministro, especialmente en la gestión de FRUVER para sus 25 tiendas. La problemática identificada radicó en la incapacidad para evaluar con precisión el volumen de abastecimiento necesario, resultando en compras excesivas y desalineadas con las necesidades del mercado. Estas dificultades generaron problemas logísticos y pérdida de ventas debido a la sobrecompra de productos. A pesar de esfuerzos previos, los modelos desarrollados alcanzaron solo un 70% de precisión, insuficiente para satisfacer las necesidades operativas y estratégicas de la compañía.

Para abordar este desafío, se ejecutó un proyecto integral que implicó el desarrollo e implementación de un modelo de aprendizaje automático. Este modelo se enfocó en proveer diariamente la demanda en unidades a partir de datos históricos, con una atención particular en la unidad estratégica de negocio FRUVER en Megatiendas. El objetivo principal fue optimizar y mejorar la gestión de inventarios, reducir costos y pérdidas de productos, y determinar de manera efectiva la necesidad por tienda en el área de FRUVER.

Para abordar esta investigación, fue necesario esclarecer conceptos clave como la gestión de inventarios, los pronósticos y el manejo de almacenes, dado que parte del problema radicó en la deficiente gestión de estos aspectos. La gestión precisa de inventarios es esencial en la industria minorista, permitiendo una planificación estratégica y proporcionando una visión actualizada de la disponibilidad de productos [1]. En el caso de Megatiendas, la falta de claridad sobre la capacidad de almacenamiento, rotación y demanda condujo a compras excesivas de productos perecederos, como FRUVER, resultando en una considerable pérdida de producto y capital. Asimismo, la ausencia de pronósticos precisos de la demanda futura dificulta la proyección de suministros y presupuestos necesarios.

Este documento presenta una metodología detallada en tres fases para abordar objetivos específicos relacionados con la extracción, transformación, análisis y predicción de datos en el contexto de Megatiendas, centrándose en FRUVER. Para ello, se dividió el proyecto en tres fases: la primera fase se enfocó en la extracción de datos desde diversas fuentes de Megatiendas; la segunda fase se centró en la transformación y limpieza de los datos; y la tercera fase abordó el desarrollo del modelo predictivo, destacando la importancia de la selección de variables relevantes y la investigación exhaustiva de modelos disponibles en el mercado.

Este proyecto no solo buscó mejorar la gestión de inventarios y la planificación de la cadena de suministro para FRUVER, sino que también aspiró a ampliar el conocimiento sobre la predicción de la demanda en la costa caribe colombiana. Se esperaba que esta iniciativa no solo resolviera un desafío operativo para Megatiendas, sino que también posicione a la empresa como un referente innovador y eficiente en la gestión de su cadena de suministro. Sus resultados previeron generar eficiencia operativa, reducir costos asociados al exceso de inventario y minimizar el desperdicio, beneficiando tanto a la empresa como a su entorno operativo, mejorando la eficiencia y la sostenibilidad a largo plazo en Megatiendas.

1. DEFINICIÓN DEL PROBLEMA

1.1. Planteamiento Del Problema

Los supermercados desempeñan un papel fundamental en la economía y la vida cotidiana, gracias a su diversidad de productos, capacidad para ahorrar tiempo, ofrecer precios competitivos, generar empleo, distribución eficiente e innovación. Este modelo de negocio, si bien fue conocido por ser rentable y generar buenos ingresos, requirió atención especial en la gestión de inventarios, control de pronósticos de demanda y reducción de desperdicios. Estos aspectos fueron críticos para mantener la eficiencia operativa y maximizar la rentabilidad, garantizando la disponibilidad adecuada de productos sin excesos ni faltantes, y minimizando el desperdicio de productos perecederos.

Los pronósticos de la demanda son una herramienta vital para el desarrollo y la gestión eficiente de cualquier empresa. Permitiendo la toma de decisiones informadas en áreas que iban desde la administración de inventarios hasta la optimización de la cadena de suministro. Dado que los mercados son dinámicos y cambiantes, es fundamental que las empresas cuenten con métodos precisos para anticipar la demanda [2]. Además, un buen sistema de pronóstico también debe considerar los modelos estadísticos más adecuados para prever la demanda futura, como lo demuestra el uso del modelo ARIMA (0,1,1) en estudios de optimización de inventarios y desempeño en cadenas de suministro [3].

En este estudio, nos enfocamos en las necesidades identificadas en los sistemas de pronóstico para la cadena de supermercados Megatiendas, líder en la costa colombiana con 25 ubicaciones, la cual enfrentó un desafío crítico en su cadena de suministro de productos frescos, específicamente en la gestión de frutas y verduras ("FRUVER"). El problema principal radicó en la incapacidad para evaluar correctamente el volumen de abastecimiento necesario para las 25 tiendas, lo que resultó en sobrecompras desalineadas con las necesidades del mercado. Esto generó dificultades logísticas para satisfacer la demanda y riesgos de pérdida de ventas al comprar en exceso. Estas dificultades se derivaron de la ausencia de respaldo tecnológico que permitiera pronosticar la demanda de manera efectiva.

Para abordar la problemática mencionada anteriormente, se constituyó un equipo multidisciplinario integrado por un analista de datos, un estadista y un economista, con el objetivo de desarrollar un modelo de pronóstico inicial. Este equipo logró generar proyecciones iniciales con una precisión que osciló entre el 60% y el 65%. A pesar de los esfuerzos para mejorar, solo se pudo alcanzar un 70% de precisión en los pronósticos. Aunque este nivel de precisión fue significativo, resultó insuficiente para satisfacer las necesidades operativas y estratégicas de la compañía. Por ende, el modelo fue implementado durante aproximadamente dos meses, pero finalmente fue descartado debido a que un 70% de precisión no cumplía con los estándares necesarios para la optimización de la cadena de suministros. Esta decisión tuvo repercusiones negativas tanto en la toma de

decisiones estratégicas como en la gestión de inventarios, evidenciándose un aumento en la destrucción o donación de productos. La falta de alineación entre los pronósticos generados y las demandas reales del mercado contribuyó a esta problemática, llevando a la empresa a dejar de utilizar el modelo por completo.

Los desafíos experimentados por Megatiendas sugirieron una necesidad inminente de mejorar sus sistemas de gestión de pronósticos. Los métodos y modelos actuales no estaban produciendo los resultados deseados, lo que resultó en pérdidas significativas para la empresa. Un ejemplo destacado fue el proyecto llevado a cabo por Jesús Belalcazar y Alejandra Cárdenas en Cali. Presentaron una propuesta para mejorar el sistema de pronósticos del Supermercado Punto Mercar S.A. Su enfoque implicó el desarrollo de una herramienta en Excel que les permitió analizar y evaluar sistemas de pronóstico aplicables a las SKU más importantes de la organización. Utilizaron métodos como la clasificación ABC, Diagramas SIPOC y flujogramas para comprender el entorno operativo del supermercado y, así, mejorar el proceso de pronóstico de la demanda [4].

Siguiendo este ejemplo, fue crucial para Megatiendas desarrollar un nuevo modelo que permitiera optimizar el control de inventario y reducir el desperdicio de productos frescos. Esto se logró mediante la implementación de herramientas estadísticas para analizar datos, similar a lo realizado por Belalcazar y Cárdenas [4], además del uso de lenguajes de programación que facilitaron el procesamiento y la visualización de información. La aplicación de modelos predictivos avanzados como el aprendizaje automático también ha demostrado ser eficaz en la optimización de inventarios, lo que demuestra ser un enfoque viable para mejorar la precisión de los pronósticos [5].

1.2. Formulación Del Problema

Preguntas de Sistematización:

1.1. ¿Qué estrategias de optimización podrían implementarse para mejorar la precisión del modelo de aprendizaje automático a medida que se acumulan más datos o cambian los patrones de ventas en FRUVER?

1.2. ¿Cómo podemos garantizar la actualización y mantenimiento constante de los datos utilizados para el modelo de aprendizaje automático en FRUVER?

1.3. ¿Cómo se puede desarrollar un modelo de aprendizaje automático utilizando los historiales de ventas en unidades de la unidad estratégica de negocio FRUVER, con una atención especial a la desagregación por tiendas, día y productos?

1.4. ¿De qué manera se puede incorporar un módulo en R que permita medir la precisión de los resultados del modelo de aprendizaje automático a partir de los datos reales, considerando métricas y medidas de desempeño comúnmente utilizadas en el estado del arte para evaluar su eficacia y precisión?

Pregunta de Investigación:

¿Cómo se puede diseñar e implementar un modelo de aprendizaje automático que permita prever con precisión la demanda para la unidad estratégica de negocio FRUVER de manera diaria con una precisión igual o superior al 85% mediante el uso exclusivo de datos históricos para lograr optimizar los inventarios y reducir el desperdicio de productos perecederos?

2. OBJETIVOS DEL PROYECTO

2.1 Objetivo General

Desarrollar e implementar un modelo de aprendizaje automático para la predicción diaria de la demanda en unidades, basándose en datos históricos, con una precisión superior a los modelos implementados anteriormente, focalizado en la unidad estratégica de negocio FRUVER en Megatiendas.

2.2 Objetivos Específicos

1. Desarrollar una estrategia para la extracción precisa de datos detallados de ventas en unidades, desglosados por tiendas, día, horas y productos, incorporando de manera integral variables cruciales como promociones y eventos.
2. Implementar un módulo en R para transformar, limpiar e integrar datos brutos de las diferentes fuentes de megatiendas, con el fin de tener un mejor manejo de los datos garantizando que siempre estén a nuestra disposición.
3. Desarrollar un modelo de aprendizaje automático utilizando los historiales de ventas en unidades de la unidad estratégica de negocio FRUVER, considerando la desagregación por tiendas, día y productos.
4. Incorporar un módulo en R que permita medir la precisión de los resultados del modelo a partir de los datos reales. En este módulo se considerarán métricas y medidas de desempeño utilizadas comúnmente en el estado del arte.

3. MARCO TEÓRICO Y ANTECEDENTES

3.1. Marco Teórico

En la actualidad, la eficiencia operativa en la industria minorista, particularmente en el ámbito de los supermercados, se ha convertido en un desafío estratégico crucial. La gestión eficaz de inventarios y la capacidad para prever con precisión la demanda de productos frescos son elementos fundamentales para garantizar un equilibrio óptimo entre la oferta y la demanda. En este contexto, la cadena de supermercados Megatiendas, líder en la costa colombiana, se enfrenta a desafíos significativos en su cadena de suministro, particularmente en la gestión de frutas y verduras ("FRUVER"). La imprecisión en los pronósticos de demanda ha generado exceso de inventario, dificultades logísticas y un impacto negativo tanto en la toma de decisiones estratégicas como en la gestión de inventarios.

Para abordar esta problemática, se requiere un enfoque integral que combine métodos avanzados de pronóstico adaptativos a los cambios del mercado, herramientas tecnológicas innovadoras y estrategias de gestión de inventarios que minimicen el desperdicio sin comprometer la disponibilidad para los clientes. El marco teórico propuesto se basa en la comprensión de la gestión estratégica de inventarios, la adaptabilidad en los pronósticos de demanda y la integración efectiva de herramientas tecnológicas. Estos elementos conforman la base para el diseño y la implementación de un modelo de predicción personalizado que responda a las necesidades específicas de Megatiendas, con el objetivo de mejorar la eficiencia en toda la cadena de suministro y la toma de decisiones en su unidad de negocio FRUVER.

A continuación, definiremos el concepto de almacén para entender mejor el contexto.

Almacén

El propósito del almacén se centra en varios objetivos clave: agilizar las entregas, mantener un control preciso sobre las mercancías almacenadas, optimizar el uso del espacio disponible, reducir la manipulación y transporte de productos, disminuir las devoluciones debido a errores y mejorar la eficiencia en términos de costos. Estas funciones no solo garantizan un flujo eficiente de mercancías dentro de la cadena de suministro, sino que también contribuyen a la minimización de costos operativos y la mejora del servicio al cliente, como lo destaca Richards en su guía integral sobre gestión de almacenes [6].

El almacenamiento es de vital importancia en cualquier empresa de suministros, ya que permite ahorrar tiempo y dinero. Sin embargo, esta práctica puede convertirse en un arma de doble filo si no se gestiona adecuadamente el inventario, lo que podría resultar en confusión y pérdidas significativas. El desarrollo efectivo de la gestión en centros de distribución implica dos pasos esenciales: primero, definir el perfil de actividad de cada producto; luego, evaluar y garantizar el almacenamiento adecuado para aprovechar al máximo las ubicaciones donde se encuentran los productos [6].

En este proyecto, el almacenamiento juega un papel crucial en el desarrollo del modelo. Como se

mencionó anteriormente, el almacén no solo proporcionará un resumen y una visión general inicial del estado de almacenamiento de los productos frescos en Megatiendas, sino que también será fundamental para controlar y mantener un registro preciso de la rotación de los productos.

3.1.1 Pronósticos de Demanda

En la actualidad, los pronósticos se han vuelto imprescindibles para el análisis y la planificación de cualquier negocio. Estas proyecciones nos ofrecen una visión futura fundamentada en datos históricos de producción y ventas. Por esta razón, es fundamental para cualquier empresa comprender sus proyecciones a largo plazo. El objetivo principal de pronosticar la demanda es diseñar estrategias más efectivas a fin de responder eficazmente a las exigencias del mercado [7].

Hay dos formas de clasificar los pronósticos: una según el período de tiempo (corto, mediano o largo plazo) y otra según el enfoque utilizado, que puede ser cuantitativo o cualitativo **[8]**, **[9]**.

3.1.1.1 Modelos Cualitativos

Los pronósticos cualitativos se basan en información que puede ser menos relevante o carecer de una estructura analítica claramente definida [9]. Algunos de ellos son:

- Delfos: Útil para pronósticos a largo plazo, planeación de capacidad e innovación tecnológica, facilita la obtención de consensos y minimiza conflictos interpersonales [8].
- Juicio de expertos: Depende de la experiencia y conocimientos de los expertos, es una técnica simple y rápida [9].
- Redacción del escenario: Requiere un juicio razonable y claro entendimiento de suposiciones para generar escenarios futuros, pero puede presentar una amplia gama de resultados [8].
- Enfoques intuitivos: Se basa en la capacidad de procesamiento de la información de un grupo, útil para situaciones donde los datos son difíciles de cuantificar [9].

3.1.1.2 Modelos Cuantitativos

Los métodos cuantitativos ofrecen una aproximación más rigurosa y basada en datos para la generación de pronósticos. Al emplear técnicas analíticas y estadísticas, permiten una proyección más precisa y objetiva hacia el futuro. Esta precisión suele ser mayor en comparación con los métodos cualitativos, ya que se apoya en la estructura y análisis de datos históricos para anticipar tendencias y patrones futuros con mayor exactitud [8].

3.1.1.3 Modelos de series de tiempo

En este proyecto, nos valdremos de métodos cuantitativos, especialmente un modelo de series de tiempo, para identificar tendencias a partir del histórico de ventas de las tiendas. Es crucial analizar los datos históricos y discernir los patrones y tendencias diarias, lo cual hace que el uso de modelos de series de tiempo sea sumamente pertinente para este proyecto.

Las series de tiempo son métodos ampliamente utilizados y seguros para pronosticar la demanda

de productos, tal como lo señala Stephen N. Chapman en su libro "Planificación y control de la producción" [10]. Chapman destaca que estos pronósticos se basan en la suposición de que los patrones pasados de la demanda pueden analizarse y emplearse para proyectar la demanda futura, partiendo del supuesto de que esos patrones se mantienen consistentes. En este sentido durante el proyecto identificamos los siguientes patrones:

- Patrones aleatorios
- Patrones de tendencia
- Patrones estacionales

Estos tres factores lo podemos definir de la siguiente manera:

- **Patrones aleatorios**

Se basa en la premisa de que la demanda siempre tiene una componente aleatoria. Esta premisa refleja la naturaleza no uniforme y no completamente predecible del comportamiento de los clientes al demandar bienes y servicios de una empresa [10].

- **Patrones de tendencia**

A pesar de las fluctuaciones aleatorias en los datos recopilados a lo largo de uno o varios períodos, es factible identificar cambios graduales que indican una tendencia en el comportamiento de la demanda. Los factores que influyen en esta tendencia son numerosos y pueden incluir la introducción de nuevos productos, mejoras en la calidad, cambios de precios o eventos inesperados [11]. Además, la capacidad de identificar y comprender las tendencias en el comportamiento de la demanda es esencial para tomar decisiones estratégicas y adaptarse a los cambios del mercado.

- **Patrones estacionales**

Estos patrones se consideran cíclicos, ya que podrían o no corresponderse con las estaciones. En esencia, los patrones cíclicos son aquellos que siguen un ciclo de demanda, ya sea creciente o decreciente [12]. Comprender y reconocer estos patrones permite a las empresas anticipar y planificar adecuadamente estas variaciones cíclicas en la demanda, adaptando sus estrategias de oferta, inventario y marketing para satisfacer las necesidades de los consumidores durante estos ciclos recurrentes.

- **Suavización**: Los métodos de suavización representan una herramienta eficaz para reducir la variabilidad aleatoria en series de tiempo estables, lo que permite obtener pronósticos más estables y precisos, especialmente a corto plazo. Su facilidad de uso y capacidad para ajustarse a patrones temporales hacen que sean una opción valiosa en situaciones donde se requiere una predicción rápida y precisa.
- **Proyección de tendencia:** Es útil para analizar datos históricos, identificando tendencias subyacentes en la demanda mientras se consideran los componentes estacionales. La

desestacionalización es clave para aislar estas tendencias, permitiendo una comparación más precisa entre períodos y facilitando la planificación estratégica a largo plazo [13].

- **Proyección de tendencia ajustada:** La estacionalidad es un factor clave en la predicción de la demanda para ciertos productos. Los cambios estacionales en el comportamiento del consumidor, ya sea por variaciones climáticas o festividades, generan patrones predecibles en la demanda. Ignorar esta estacionalidad puede llevar a pronósticos inexactos y dificultar la planificación adecuada de inventarios y recursos. Por ello, es fundamental tener en cuenta este componente en los modelos de pronóstico para productos que muestran este tipo de comportamiento estacional. Integrar esta estacionalidad en los modelos de pronóstico permite una planificación más precisa y ajustada a las necesidades cambiantes a lo largo del año [14].

3.1.1.4 Modelos ARMA y SARIMAX

3.1.1.4.1 Modelo ARMA

El modelo ARMA (Autoregressive Moving Average) es una técnica utilizada en series de tiempo que combina dos componentes: el autorregresivo (AR) y el de promedio móvil (MA). Este modelo es útil para analizar y predecir datos donde los valores futuros están influenciados tanto por los valores pasados (componente AR) como por los errores pasados (componente MA) [15].

Componente AR: El modelo AR es el componente autorregresivo, este modelo se basa en la idea de que el valor actual de una serie temporal puede explicarse por una combinación lineal de sus valores anteriores y un término de error [15]. El cual está definido por la siguiente función.

Componente MA: Este componente modela la dependencia de errores pasados en la serie de tiempo. Este modelo considera que el valor actual de la serie depende de los últimos términos de error, capturando así la influencia de perturbaciones anteriores en la evolución actual de la serie [15]. La estructura de este modelo permite entender y prever cómo los choques pasados afectan el comportamiento presente de la serie, proporcionando una herramienta útil para la modelización y el análisis de series temporales con patrones de errores auto-correlacionados. Este modelo está definido por la siguiente función.

Formulación del Modelo ARMA (p, q):

El valor de la serie temporal en un momento t está determinado por una combinación lineal de sus valores anteriores y los errores anteriores [16]. La formulación matemática del modelo ARMA (p, q) se expresa como:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + .. + \phi_p y_{t-p} + \theta_1 v_{t-1} + \theta_2 v_{t-2} + .. + \theta_q v_{t-q} + v_t \quad (3.1)$$

Donde:

- y_t: valor de la serie en el tiempo t.
- ϕ_i: coeficientes del componente AR.

- p: orden del componente AR.
- θ_i: coeficientes del componente MA.
- q: orden del componente MA.

Los coeficientes $\phi_1,\phi_2,\ldots,\phi_p$ y $\theta_1, \theta_2, \ldots, \theta_p$ Se estiman a partir de los datos históricos de la serie de tiempo utilizando técnicas estadísticas como el método de Mínimos Cuadrados o el método de Máxima Verosimilitud. La estimación de los coeficientes en un modelo ARMA se realiza utilizando datos históricos de la serie de tiempo, aplicando técnicas estadísticas como el método de Mínimos Cuadrados o el método de Máxima Verosimilitud. La elección de los valores de p y q, que representan el orden de los componentes autorregresivos (AR) y de media móvil (MA), se basa en análisis de autocorrelación y autocorrelación parcial de los residuos, lo que permite determinar cuántos valores pasados y errores deben incluirse en el modelo. Estos modelos ARMA son efectivos para capturar la dependencia temporal compleja y autocorrelación en series de tiempo, lo que los hace útiles para la modelización y pronóstico [16].

3.1.1.4.2 Modelo SARIMAX

El modelo SARIMAX (Seasonal AutoRegressive Integrated Moving Average with eXogenous variables) es una extensión del modelo ARIMA que incorpora tanto la estacionalidad como variables exógenas. Es una poderosa herramienta para modelar series temporales cuando se tienen en cuenta factores externos que podrían influir en la variable de interés [17].

El modelo SARIMAX es una extensión del modelo SARIMA(p,d,q)(P,D,Q)m que incorpora variables exógenas al modelo. Esto significa que el valor presente de la serie temporal y_t se puede representar como un modelo SARIMA al que se le agregan una o más variables exógenas X_t. Tanto SARIMA como SARIMAX son modelos lineales, ya que son combinaciones lineales de valores pasados de la serie y términos de error, junto con las variables exógenas en el caso de SARIMAX [17].

$$y_t = SARIMA(p,d,q)(P,D,Q)_m + \sum_{i=1}^{n} \beta_i X_t^i \quad (3.2)$$

3.1.1.5 Modelo Gradient Boosting

El modelo **Gradient Boosting** es un método de aprendizaje supervisado no paramétrico basado en árboles de decisión, que se utiliza para tareas tanto de clasificación como de regresión. A través de un proceso iterativo, cada modelo se entrena para corregir los errores del modelo anterior, mejorando el rendimiento general. El **Gradient Boosting** es una extensión del algoritmo **AdaBoost** que permite utilizar cualquier función de coste diferenciable, lo que le otorga gran flexibilidad y permite su aplicación en diversos problemas, como regresión y clasificación múltiple. El modelo se entrena de manera secuencial: cada iteración ajusta un modelo a los residuos del anterior, buscando minimizar los errores acumulados [18].

$$F_M(x) = F_{M-1}(x) + \lambda \cdot h_M(x) \quad (3.3)$$

Donde:

- $F_M(x)$ Es el modelo predictivo final tras M iteraciones.
- $F_{M-1}(x)$ Es el modelo obtenido después de M−1 iteraciones.
- $h_M(x)$ Es el modelo débil ajustado en la M-ésima iteración para corregir los residuos del modelo anterior.
- λ es el **learning rate** o parámetro de regularización que controla la influencia de cada modelo débil en el conjunto del ensemble.

3.1.2 Clusterización

Antes del desarrollo de los modelos predictivos, es fundamental realizar la clusterización de tiendas. La clustering permite identificar patrones y relaciones ocultas entre los datos, facilitando el análisis y la toma de decisiones. Al crear grupos homogéneos, se pueden abordar problemas con una mayor precisión y optimizar estrategias basadas en las similitudes identificadas [19]. En nuestro caso nos permitirá identificar similitudes entre las tiendas, facilitando el agrupamiento y el tratamiento del desarrollo del modelo por bloques.

Esta medida permite comparar series temporales de distinta longitud y localizar similitudes entre ellas, ajustando las discrepancias en el tiempo [20]. Existen diversas técnicas para comparar las similitudes de series temporales, siendo la DTW y la distancia euclidiana dos de las más comunes.

DTW (Dynamic Time Warping)

La DTW, es capaz de ajustar discrepancias temporales al permitir que ciertos puntos de una serie se alineen con múltiples puntos de otra. Esto hace que la DTW sea especialmente útil para comparar series temporales que tienen patrones similares, pero no están perfectamente sincronizadas [21]. Por ejemplo, si dos tiendas tienen picos de ventas en diferentes momentos del día o de la semana, la DTW puede identificar estos picos como similares, mientras que la distancia euclidiana podría considerarlos diferentes debido a la desalineación temporal.

Distancia euclidiana

La distancia euclidiana es una medida que calcula la distancia en línea recta entre dos puntos en un espacio euclidiano. En el aprendizaje automático, se utiliza para medir la similitud o disimilitud entre puntos de datos, lo cual es clave en tareas como agrupación, clasificación y detección de anomalías. Esta métrica se emplea para comparar vectores de características, que representan puntos de datos en espacios de alta dimensión, ayudando a identificar qué tan similares o diferentes son [22]. Sin embargo, esta técnica no es adecuada cuando hay variaciones en la velocidad o en el tiempo de los eventos en las series.

3.1.3 Gestión Estratégica de Inventarios

La gestión precisa de inventarios en la industria minorista, especialmente en supermercados, es fundamental para equilibrar la oferta y la demanda. Es esencial para la planificación estratégica y para tener una visión en tiempo real de la disponibilidad de productos, comprendiendo cuánto hay y cuánto se necesita. La gestión de inventarios se concentra normalmente en mantenerlos dentro de la empresa, pero en ocasiones pueden hallarse fuera de la misma o en condiciones que requieren cuidado especial, lo que demanda un manejo diferente. Además de los tipos de productos usuales mencionados previamente (materias primas, provisiones, componentes, producto en proceso y producto terminado), es esencial ampliar esta clasificación para incluir otros materiales como producto en tránsito, producto en consignación e inventarios en cuarentena [23].

3.1.4 Error del Pronóstico

El error del pronóstico es una medida que indica cuán precisos son los pronósticos en relación con los resultados reales. Este error es fundamental para evaluar la calidad de un modelo de pronóstico o método utilizado, ya que permite medir cuán cerca o lejos están las predicciones de la realidad [24]. Las técnicas más comunes para el cálculo del error del pronóstico son:

3.1.4.1 Media aritmética del error del pronóstico

La media aritmética del error del pronóstico se utiliza para calcular el error promedio de un modelo de pronóstico. Esta métrica se obtiene sumando los errores de pronóstico de los distintos valores y dividiendo el resultado entre el número total de observaciones. Es una forma sencilla de evaluar cuán cerca o lejos están las predicciones de los valores reales [25].

3.1.4.2 Error cuadrático

El cálculo del error cuadrático busca amplificar la dimensión del error, ofreciendo así un valor más representativo y amplificado del error en la estimación. Esta medida, al elevar al cuadrado la diferencia entre el pronóstico y el valor real, resalta de manera más significativa las discrepancias entre ambos, ofreciendo una perspectiva más detallada de la magnitud del error en la predicción [26].

Para este proyecto usaremos la siguiente fórmula para el cálculo del error del pronóstico:

$$MPE = \frac{1}{n}\sum_{t=1}^{n}\frac{(y-y_t)}{y_t} \tag{3.4}$$

En donde:

- MPE: Porcentaje Promedio de Error Absoluto. (Mean Absolute Percentage Error)
- y: Demanda real
- y_t: Demanda pronosticada en el periodo t.
- n : número de periodos.

Error Porcentual Medio (MPE) proporciona una indicación sobre el sesgo de un modelo cuantitativo de pronóstico. Si el modelo no está sesgado, el MPE se acercará a cero. Sin embargo, si existe sesgo, un MPE con porcentajes positivos grandes (cercanos a 100%) indica una subestimación, mientras que porcentajes negativos grandes (cercanos a -100%) indican una sobreestimación [27].

3.1.4.3 Error Cuadrático Medio (MSE)

El MSE es una métrica eficaz para evaluar modelos de pronóstico, particularmente útil para identificar grandes errores. Sin embargo, su dependencia de la escala lo limita en algunos contextos, especialmente cuando los datos varían significativamente en magnitud. Es similar a la varianza, ya que mide la incertidumbre del pronóstico, pero no es adecuada para comparar modelos cuando las variables tienen diferentes escalas. El MSE (Mean Squared Error) es una métrica utilizada para evaluar la precisión de los pronósticos en comparación con los resultados reales de una serie temporal. La función MSE(X, F) toma dos argumentos: X, que representa los valores reales, y F, que son los valores pronosticados. Ambos deben tener el mismo tamaño y las series de tiempo deben ser homogéneas y espaciadas de manera equitativa. Las observaciones con valores faltantes se excluyen del cálculo. La ecuación del MSE eleva al cuadrado los errores (la diferencia entre el valor real y el pronosticado) y luego los promedian, dando más peso a los errores grandes, lo que lo hace útil para detectar valores atípicos [28] .

3.1.4.4 Raíz del Error Cuadrático Medio (RMSE)

La raíz del error cuadrático medio (RMSE) es una métrica clave para evaluar el rendimiento de un modelo predictivo de regresión. Mide la diferencia promedio entre los valores predichos por el modelo y los valores reales observados, proporcionando una indicación de la precisión del modelo. Un valor de RMSE más bajo indica un mejor ajuste del modelo, y un RMSE de 0 representaría un modelo perfecto [29]. Esta métrica es útil para comparar modelos, especialmente cuando se necesita evaluar la precisión en términos cuantitativos (por ejemplo, dólares, unidades). Un RMSE bajo indica que el modelo tiene un buen rendimiento predictivo y puede ajustarse adecuadamente a los datos observados.

3.1.4.5 Error absoluto medio (MAE)

Es una métrica que mide el promedio de la diferencia absoluta entre los valores observados y los valores predichos. Es un puntaje lineal, lo que implica que todas las diferencias entre las predicciones y los valores reales se tratan de manera equitativa en el cálculo del promedio [30]. En otras palabras, cada error contribuye de manera proporcional a la métrica final.

3.1.4.6 Error de porcentaje medio absoluto (MAPE)

El MAPE (Mean Absolute Percentage Error) es una métrica utilizada para evaluar la precisión de modelos predictivos, especialmente en series temporales y modelos de regresión. Representa el error promedio absoluto en términos porcentuales, lo que facilita la interpretación de la precisión

del modelo en relación con el valor real. Existen variaciones del MAPE, como el **SMAPE** (MAPE Simétrico), que ajusta el cálculo para mejorar la comparación entre series. El **SMAPE** tiene la ventaja de estar limitado entre 0% y 200%, y ofrece una medición más equilibrada entre el sobre pronóstico y el pronóstico insuficiente [31] .

3.1.4.7 R2 (R cuadrado)

El **R^2** o coeficiente de determinación es una métrica que mide qué tan bien un modelo de regresión se ajusta a los datos reales. Se utiliza para evaluar la precisión general de un modelo, especialmente en el caso de modelos de regresión, como los árboles CHAID en IBM Cognos Analytics [32]. Su valor se encuentra entre 0 y 1, donde:

- Un valor de **1** indica que el modelo predice perfectamente los resultados.
- Un valor de **0** significa que el modelo no tiene capacidad predictiva.

En situaciones simples con una sola variable de entrada, **R^2** es equivalente al cuadrado de la correlación de Pearson. El **R^2** es una herramienta clave para evaluar el rendimiento de modelos de regresión. Un valor más cercano a 1 indica un buen ajuste del modelo, mientras que un valor cercano a 0 indica que el modelo no es útil para la predicción. Aunque R^2 es una métrica útil, no siempre refleja toda la calidad del modelo, por lo que debe complementarse con otras métricas para una evaluación más exhaustiva [33].

3.2. ANTECEDENTES

En el dinámico entorno empresarial actual, caracterizado por cambios rápidos y una creciente complejidad, la capacidad de prever eventos futuros se ha convertido en un activo estratégico indispensable. Es de aquí donde nacen los modelos de pronósticos, que han desempeñado un papel fundamental en el panorama empresarial al ofrecer una perspectiva valiosa para comprender y mejorar la eficiencia en los procesos operativos de las organizaciones. Estos modelos, basados en análisis estadísticos y algoritmos avanzados, permiten a las empresas anticipar tendencias, identificar patrones y tomar decisiones informadas, no sólo ayudando a anticipar la demanda de productos, sino que también permiten optimizar la gestión de inventarios, mejorar la planificación de la cadena de suministro y minimizar los riesgos operativos.

El artículo "Planificación y Control de la Producción" de 2006, escrito por Stephen N. Chapman [10], resalta que la actividad de planificación de ventas y operaciones, en su mayoría, no se emplea directamente en la programación de la producción. Su función principal radica en planificar y coordinar recursos, abarcando tipo, cantidad y relevancia. El horizonte temporal de esta planificación se ajusta al momento futuro en el que la empresa necesitará estimar las demandas de recursos para garantizar su disponibilidad.

En este contexto, la planificación de ventas y operaciones emerge como una fuente crucial para la planificación de diversos aspectos, incluyendo niveles de inventario, flujo de efectivo, necesidades de recursos humanos, necesidades de capital, niveles de producción, planificación de la capacidad (como equipos) y actividades de ventas y marketing, que abarcan desde promociones y publicidad hasta fijación de precios, introducción de nuevos productos y expansión de mercados.

Un estudio de 2011 titulado "Supply Chain Management Strategies Based on Demand Planning in Colombia" [34], publicado por la revista politécnica, resalta la importancia de la logística y las cadenas de suministro como generadoras de ventajas competitivas en las empresas colombianas. Además, advierte en varias ocasiones sobre el riesgo de utilizar técnicas informales y con un alto componente empírico en la proyección de la demanda. Estas prácticas podrían exponer a la compañía a niveles elevados de inventarios, con los consiguientes costos asociados, o, en su contraste, a un considerable desabastecimiento que resultaría en pérdidas de ventas, ineficacia de planes de marketing y la posible pérdida de clientes [34].

Ingrid Hernández y Luis Torres de la universidad de los andes, realizaron una propuesta para la mejora de los modelos de predicción de la demanda utilizados por la empresa Electrisol en el año 2021 [35], la empresa manejaba aproximadamente 39.418 artículos clasificados en tres grupos A, B y C según su sistema de ERP, de los cuales solo los artículos tipo A, alrededor del 19.05% (unos 7.509 artículos), eran productos de alta rotación y representaban más del 90% en ventas mensuales, ellos enfocaron su propuesta a los artículos tipo A, entre los datos históricos lograron identificar patrones de comportamientos de la demanda de cada artículo, revelando tendencias positivas y negativas que afectarán la demanda. Se realizaron pruebas en modelos como ARIMA, Facebook Prophet, Deep Learning en redes neuronales, Machine Learning basado en árboles, estos dos últimos demostramos un mejor comportamiento comparando los resultados de los modelos con las

ventas reales de la compañía, obteniendo un menor error en datos nuevos por los cuales los hicieron más precisos que otros modelos [35].

Jesús Belalcazar y Alejandra Cárdenas llevaron a cabo un proyecto en Cali, presentando una propuesta para mejorar el sistema de pronósticos del Supermercado Punto Mercar S. A [4]. Su enfoque se centró en perfeccionar un modelo tradicional mediante el desarrollo de una herramienta en Excel. Esta herramienta permitió analizar y evaluar sistemas de pronóstico aplicables a las SKU más relevantes de la organización, buscando aumentar la precisión de los pronósticos y optimizar la gestión de inventarios para estas líneas de productos clave. Además, el proyecto destacó el uso de métodos como la clasificación ABC, Diagramas SIPOC y flujogramas para comprender el contexto operativo del supermercado [4].

El acto de pronosticar implica la anticipación de eventos futuros o el desarrollo de un proceso, fundamentándose en criterios lógicos, científicos o análisis de datos disponibles. En el contexto de la predicción de la demanda, este proceso puede llevarse a cabo mediante métodos simples o complejos. La importancia clave de la pronosticación radica en la capacidad de reducir la incertidumbre que enfrentan los líderes empresariales, especialmente en el ámbito de los supermercados enfocados en la venta de productos perecederos. La habilidad para prever y comprender estos patrones se convierte en un elemento esencial para la toma de decisiones efectiva y la gestión estratégica en cualquier tipo de organización [36].

4. ESTRATEGIA PARA LA EXTRACCIÓN DE DATOS

La necesidad de una planificación detallada y organizada se volvió evidente para asegurar que cada objetivo específico fuera abordado de manera eficiente y con resultados exitosos. Para dar cumplimiento a nuestro **objetivo específico 1 (OE1)**, se desarrolló una estrategia para la extracción precisa de datos detallados de ventas en unidades, desglosados por tiendas, día, horas y productos, incorporando de manera integral variables cruciales como promociones y eventos. Primero, se realizó la solicitud formal de los datos, seguida por un desglose y análisis exhaustivo de las promociones y parámetros a utilizar.

4.1 Obtención de datos.

- **Solicitud Formal de Credenciales:**

El proceso de obtención de los datos comenzó con una solicitud formal dirigida a **Megatiendas**, en la que se solicitó la provisión de credenciales de usuario y los permisos necesarios para extraer información desde sus diversas fuentes de datos. Dicha solicitud fue realizada a través de un correo oficial y posteriormente aprobada. Como parte del proceso, se nos remitió un contrato de confidencialidad y tratamiento de datos. Este documento fue firmado y enviado de vuelta, cumpliendo con los requerimientos legales y normativos, tal como se detalla en el **Anexo 3**.

- **Fuentes de Datos:**

Megatiendas proporcionó los datos en formato de hojas de cálculo Excel, facilitando además el **diccionario de datos** de su principal *data warehouse*. Este diccionario, representado en la **Tabla 1**, describe las características principales de las bases de datos, incluyendo información clave sobre tablas relacionadas con promociones y eventos comerciales futuros. Los datos entregados están orientados específicamente a la **unidad estratégica de negocio FRUVER**, lo que permite un análisis detallado y específico del sector.

Tabla 1

Marco de datos inicial (Elaboración propia).

Variable	Descripción	Tipo de datos
Source.nave	Nombre del archivo de origen	Character
Fecha movto.	Fecha del movimiento	datetime
C.O	Código de operación	numeric
Desc. C.O.	Descripción del código de operación	character

Secciones	secciones de productos	character
Unidad de negocio	Unidad de negocio	character
Categorías	categorías de productos	character
Familia	Familia de productos	character
Item	Ítem relacionado	numeric
Desc. item	Descripción de item	character
Valor subtotal local	Valor subtotal en moneda local	numeric
Cantidad inv.	Cantidad inventariada	numeric
Año	Año de los datos	numerci
Mes	Mes de los datos	character
Semana	Semana de los datos	character

- **Segmentación de Datos:**

Los datos se segmentaron por regiones y tiendas:

- **Región Bolívar:** Incluye 18 tiendas (Tienda 101 a Tienda 118).
- **Región Atlántico:** Incluye 7 tiendas (Tienda 201 a Tienda 208).

Esta segmentación permitió una organización más clara y facilitó el análisis por bloques, enfocándose en patrones locales y regionales.

4.2. Parámetros Relevantes para el Análisis

Una vez obtenidos los datos, se definieron los parámetros clave para la extracción y análisis. Estas variables se eligieron en función de su impacto en la demanda y su relevancia para el modelo predictivo. A continuación, se explican las principales:

1) **Ventas por unidad:** Los datos de ventas se analizaron desglosados por tienda, producto y fecha. Esto permitió identificar patrones y tendencias específicas para cada punto de venta.
2) **Promociones:** Un aspecto central de la información proporcionada corresponde a las **promociones y eventos comerciales**, que tienen un impacto significativo en los patrones de ventas. Cabe destacar que los días miércoles se realiza una promoción especial, lo que genera picos notables en las ventas. Estos patrones serán analizados en detalle, como se observa en la

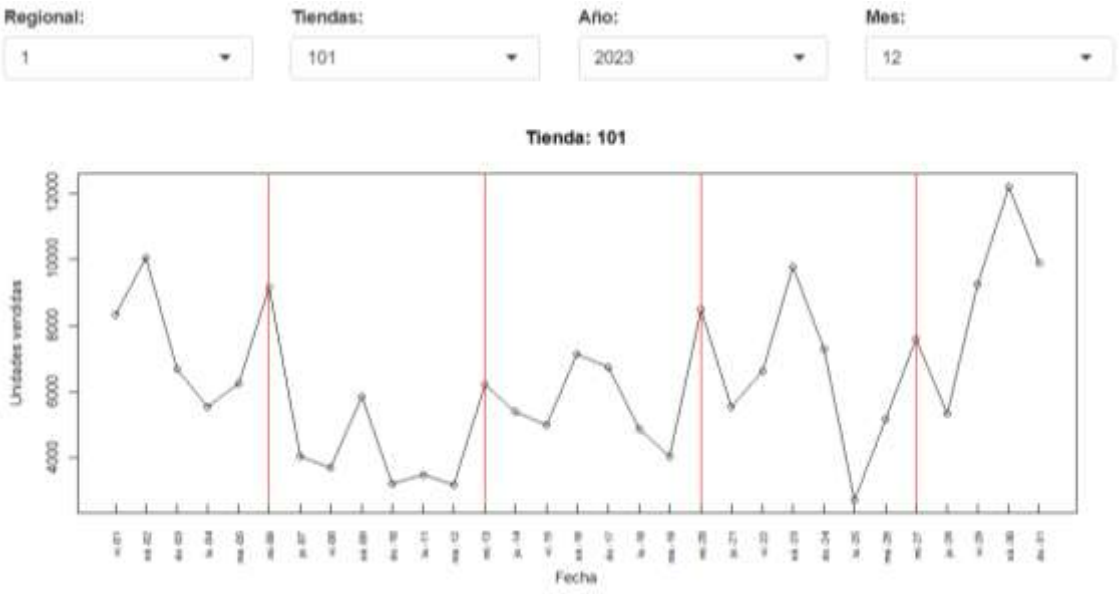

Figura 1: Tendencia de datos para la regional 1, tienda 101, en diciembre de 2023 (Elaboración propia en R).

Figura 1, que muestra el comportamiento de las ventas durante estos días.

3) **Región geográfica**: Los datos se segmentaron por región (Bolívar y Atlántico) y tienda, permitiendo un análisis granular adaptado a las necesidades locales.

Estos parámetros no solo facilitaron el análisis exploratorio, sino que también sirvieron como insumos clave para el desarrollo de los modelos predictivos descritos en capítulos posteriores.

4.3 Uso de Herramientas para la Extracción

Dado que no se permitió la extracción directa de datos desde nuestro entorno de desarrollo en R mediante VPN, el uso de Google Drive como repositorio de archivos XLSX actualizados mensualmente se convirtió en una opción viable. Este paquete nos permite interactuar con Google Drive desde R, facilitando la automatización del proceso de descarga y carga de archivos de datos. El paquete google drive de R es fundamental en nuestro flujo de trabajo debido a la solución adoptada para la extracción y carga de datos desde Google Drive, resultado de las restricciones de seguridad impuestas por MEGATIENDAS.

Con este paquete, podemos:

• **Descargar archivos:** Cada primer día del mes, utilizamos google drive para descargar automáticamente el archivo XLSX actualizado con las ventas del mes anterior.

• **Cargar archivos**: En cada etapa del proyecto, especialmente al finalizar cada una, guardamos los resultados obtenidos en nuestra cuenta de Google Drive. Este enfoque garantiza que todos los avances y resultados intermedios estén organizados y accesibles, permitiendo una continuidad en

el flujo de trabajo y asegurando que la información esté siempre disponible para futuras consultas o análisis.
• **Gestión de carpetas**: Aprovechando las funcionalidades de creación y eliminación de carpetas y archivos, optimizamos la organización de nuestros datos y proyectos en Google Drive. Esto nos permite, por ejemplo, asignar una carpeta específica para cada tienda tras la transformación de datos, donde almacenaremos archivos en formato XLSX que serían las entradas para nuestros modelos.

4.4 Exclusión de Variables y Limpieza Inicial

Exclusión de Variables

Se realizó una evaluación detallada de las columnas disponibles para identificar aquellas que no aportaban valor al modelo.

- **Horas específicas:** Aunque se disponía de datos horarios, se decidió excluirlos ya que los modelos de series de tiempo se centran en identificar patrones como tendencias y estacionalidades, más que en horas específicas. Además, muchos modelos de series temporales requieren estacionariedad, y agregar horas podría romper esta característica sin aportar valor predictivo adicional. Los modelos suelen funcionar mejor al analizar la relación temporal entre los puntos de datos, en lugar de enfocarse en horas exactas. En nuestro caso, el modelo de pronóstic de demanda para la optimización de la cadena de suministros en supermercados MEGATIENDAS está orientado a predecir la demanda diaria, semanal o mensual. Dado que las decisiones de inventario y reposición se basan en estas escalas temporales, la hora específica de compra no añade valor. Incluirla podría introducir ruido y complicar el modelo, ya que la demanda sigue patrones diarios o estacionales, no variaciones horarias precisas. Por lo tanto, excluiremos las horas en nuestro modelo de pronósticos.

Tratamiento de Datos Faltantes

Identificamos que menos del 0,01% del total datos contenía valores nulos, los cuales fueron imputados.

Codificación De la información

En el desarrollo de nuestro proyecto de grado, hemos priorizado la confidencialidad y protección de la información utilizada. Por esta razón, hemos decidido modificar los nombres de las tiendas reales involucradas en los análisis, reemplazándolos por nombres ficticios como se muestra en la tabla 2. Además, los datos numéricos(enteros) presentados en tablas, como ventas, ingresos u otros indicadores, se han generado de manera ficticia, utilizando valores enteros que no corresponden a las cifras reales. Estas medidas garantizan el cumplimiento de los estándares éticos y la privacidad en el manejo de datos, asegurando que ninguna entidad o negocio se vea afectado por los resultados o conclusiones presentados en este trabajo.

Tabla 2

Nombres de las tiendas por región (Elaboración propia).

Desc. C.O	C.O	Regional
Tienda 1	101	Bolívar
Tienda 2	102	Bolívar
Tienda 3	103	Bolívar
Tienda 4	104	Bolívar
Tienda 5	105	Bolívar
Tienda 6	106	Bolívar
Tienda 7	107	Bolívar
Tienda 8	108	Bolívar
Tienda 9	109	Bolívar
Tienda 10	110	Bolívar
Tienda 11	111	Bolívar
Tienda 12	112	Bolívar
Tienda 13	113	Bolívar
Tienda 14	114	Bolívar
Tienda 15	115	Bolívar
Tienda 16	116	Bolívar
Tienda 17	117	Bolívar
Tienda 18	118	Bolívar
Tienda 19	201	Atlántico
Tienda 20	202	Atlántico
Tienda 21	203	Atlántico
Tienda 22	205	Atlántico
Tienda 23	206	Atlántico
Tienda 24	207	Atlántico
Tienda 25	208	Atlántico

4.5. Validación del Dataset Final

Una vez completada la limpieza y transformación de los datos, se consolidó un dataset final que incluía todas las variables relevantes. Este archivo se validó mediante:

1. **Revisión manual de las primeras y últimas filas del dataset** para confirmar su coherencia.
2. **Visualización de gráficos exploratorios preliminares**, como histogramas y boxplots, para verificar la distribución de los datos.
3. **Exportación en formato .rds** para garantizar un manejo eficiente dentro de R.

Impacto y Conexión con los Objetivos

La implementación de esta estrategia permitió cumplir con el **objetivo específico 1 (OE1)**, asegurando que los datos obtenidos fueran de alta calidad, consistentes y relevantes para el análisis y modelado predictivo. Este enfoque sentó las bases para:

- El análisis exploratorio presentado en el siguiente capítulo.
- El desarrollo de modelos predictivos detallados en el Capítulo 6.

5. MÓDULO DE EXTRACCIÓN Y TRANSFORMACIÓN DE DATOS

Tras haber implementado una estrategia precisa para la extracción de datos de ventas, desglosando unidades por tienda, producto y fecha, e incorporando variables clave como promociones y eventos, logramos garantizar la obtención eficiente y precisa de los datos necesarios para el análisis de la unidad estratégica de negocio FRUVER. Como siguiente paso, para dar cumplimiento a nuestro **objetivo específico 2 (OE2)** desarrollamos un módulo en R para la transformación, limpieza e integración de datos brutos de las diferentes fuentes de megatiendas, con el fin de tener un mejor manejo de los datos garantizando que siempre estén a nuestra disposición.

Cabe destacar que, conforme a la política de uso de datos acordada, toda la información recopilada es estrictamente confidencial y manejada bajo los más altos estándares de seguridad. Los datos se utilizan exclusivamente para los fines específicos previstos y no se comparten con terceros sin el consentimiento expreso de los titulares. Además, se implementan medidas técnicas y organizativas adecuadas para proteger la información frente a accesos no autorizados, pérdida, alteración o destrucción.

5.1. Extracción de Datos

Obtención de Credenciales y Seguridad

El primer paso fue gestionar credenciales y permisos de acceso a las fuentes de datos de Megatiendas. Como se mencionó en el capítulo 4 y se detalla en el Anexo 3, este proceso incluyó la firma de un acuerdo de confidencialidad que asegura el cumplimiento de normas éticas en el manejo de información sensible.

Fuentes de Datos

Los datos se obtuvieron de hojas de cálculo en formato Excel proporcionadas por Megatiendas. Estas incluían:

- Ventas desglosadas por tienda, producto y fecha.
- Promociones y eventos comerciales.

Herramientas Utilizadas

Dado que la extracción directa desde el sistema de Megatiendas no fue posible, implementamos un flujo de trabajo con el paquete googledrive de R para interactuar con Google Drive, donde se almacenaron los archivos. Las funciones empleadas se muestran a continuación.

Funciones del Paquete Google Drive a Utilizar en la Extracción y a lo Largo del Proyecto

El paquete google drive proporciona una serie de funciones útiles para gestionar archivos. A continuación, se detallan algunas de las funciones clave que utilizaremos para la extracción y manejo de datos a lo largo del proyecto:

1. Autenticación y Configuración

• drive_auth(): Auténtica la conexión con Google Drive, permitiendo a R acceder a los archivos de la cuenta, Además se abrirá una ventana en el navegador donde se podrá ingresar las credenciales de la cuenta [37]. Una vez autenticado, estas credenciales se guardan en el entorno de desarrollo, permitiendo un acceso continuo a los recursos sin necesidad de reingresar los datos cada vez.

2. Gestión de Archivos

• drive_get(): Obtiene información sobre archivos específicos en Google Drive. Esta función recupera metadatos de archivos especificados mediante su ID o ruta es relativamente sencilla cuando se utiliza el ID [38].

• drive_download(): Esta función permite descargar archivos de Google Drive. Para los tipos de archivos nativos de Google, como Google Docs, Google Sheets y Google Slides, es necesario exportarlos a un formato de archivo convencional antes de la descarga [39]. Esta función es esencial para obtener los archivos XLSX actualizados mensualmente.

• drive_upload(): Cargar archivos desde el entorno local. Utilizada para guardar los resultados de los modelos y otros archivos generados.

• drive_rm(): Elimina archivos. Se usa esta función para gestionar y limpiar archivos antiguos.

3. Gestión de Carpetas

• drive_mkdir(): Esta función nos permitió organizar los archivos en carpetas específicas, como las correspondientes a cada tienda.

• drive_find(): Permite obtener una lista de archivos almacenados, similar a las operaciones que se pueden realizar en la interfaz web de Google Drive [40]. Se puede filtrar sus búsquedas por tipo de archivo o propiedad y trabajar con unidades compartidas, lo que proporciona una flexibilidad considerable. Junto con la función drive_get(), esta herramienta es esencial para identificar archivos relevantes para futuras tareas.

• drive_mv(): Estas función nos permitió mantener el almacenamiento organizado, facilitando el acceso y la identificación de documentos.

Una vez comprendidas las funciones disponibles, se procedió a la extracción de datos. Hasta este momento logramos descargar nuestros datos en un archivo temporal de datos, a continuación, entra en juego una librería importante para el proyecto.

5.2. Transformación y limpieza de datos.

Para dar inicio a la transformación y limpieza de datos se desarrolló un módulo en R el cual

permitiera transformar, limpiar e integrar datos brutos de las diferentes fuentes de megatiendas. En este módulo de R, se implementaron diversas tareas para garantizar un procesamiento eficiente de los datos provenientes de múltiples hojas de cálculo en formato Excel. Para ello, se utilizó el paquete **readxl**, que permite la lectura de archivos Excel de manera rápida y sin dependencias externas.

Inicialmente, se especificó la ruta de cada archivo y se automatizó la lectura de múltiples hojas correspondientes a diferentes períodos de ventas en la categoría FRUVER. Cada hoja fue convertida en un **data frame** individual, facilitando la manipulación de los datos por separado. Durante este proceso, se identificó que las columnas de los distintos data frames presentaban discrepancias en los nombres, lo que podía generar inconsistencias en el análisis. Para resolver esto, se unificaron los nombres de las columnas, garantizando la compatibilidad y coherencia entre los data frames.

Posteriormente, todos los data frames se consolidaron en una única estructura de datos, permitiendo un análisis integral de las ventas. Antes de su almacenamiento, se realizó una validación exhaustiva de la estructura del **data frame** combinado, verificando la correcta alineación y formato de las variables clave. Finalmente, el archivo consolidado fue guardado en formato **.rds**, un estándar nativo de R que asegura la integridad y eficiencia en el almacenamiento y recuperación de los datos.

Estas acciones no solo permitieron organizar los datos de manera estructurada y consistente, sino que también prepararon el dataset para etapas posteriores de análisis y modelado predictivo, asegurando una manipulación más sencilla y confiable en el futuro, el proceso más detallado se encuentra en el Anexo 1.

Procesamiento de Datos

Los datos extraídos se transformaron utilizando el paquete dplyr, con los siguientes pasos:

1. **Unificación de Columnas**: Homologación de nombres en todas las hojas para evitar discrepancias.

2. **Eliminación de Valores Faltantes**: Identificamos que menos del 0,01% del total datos contenía valores nulos, los cuales fueron imputados.

3. **Tratamiento de Outliers**: En el presente análisis, se identificaron y excluyeron ciertos registros atípicos con el fin de mejorar la precisión del modelo de pronóstico. Este tratamiento se centró en dos puntos fundamentales: la exclusión de ventas mayoristas y devoluciones.

   Exclusión de ventas mayoristas: Se excluyeron las ventas mayoristas, que corresponden a compras realizadas por un único comprador de grandes cantidades de un producto en una única factura. Estas ventas no son comunes y no reflejan el comportamiento típico del cliente, por lo que podrían distorsionar los resultados del modelo. Aunque estas

transacciones son permitidas en casos excepcionales, como para evitar el desperdicio de productos, su inclusión podría generar previsiones inexactas y contribuir a sobrecompras o excesos de inventario.

Exclusión de devoluciones: También se excluyeron los registros de devoluciones, que se identifican por tener cantidades negativas, lo que indica que el producto fue devuelto en lugar de vendido. En total, los datos consisten en 3'943.083 registros, de los cuales aproximadamente 132,286 son devoluciones, lo que representa un 3,35% del total. Dado que este porcentaje es relativamente pequeño, se decidió proceder con la eliminación de estos registros, ya que las devoluciones no reflejan ventas regulares y podrían afectar la precisión del análisis de demanda. Excluirlas ayuda a evitar que influyan negativamente en las previsiones de ventas y mantiene la integridad del modelo.

Conversión de Tipos de Datos: Se normalizaron fechas al formato yyyy-mm-dd y se transformaron categorías en factores para facilitar el análisis.

Codificación de Datos Sensibles

Para cumplir con las políticas de privacidad, los nombres de tiendas y productos fueron reemplazados por códigos ficticios. Además, se generaron valores ficticios para datos numéricos en las tablas presentadas en este documento como se mencionó en el capítulo 4.

5.2.1. Instalación de Paquetes R Necesarios

readxl

El paquete **readxl** es una herramienta eficiente para la extracción de datos de archivos de Excel y su integración en R. A diferencia de otros paquetes como **gdata**, **xlsx** o **xlsReadWrite**, readxl se destaca por no tener dependencias externas, lo que simplifica su instalación y uso en diferentes sistemas operativos. Este paquete está especialmente diseñado para trabajar con datos tabulares, facilitando la manipulación y el análisis de estos datos en R [41].

dplyr

Es un paquete de R que se centra en la manipulación de datos, ofreciendo una gramática clara y consistente a través de un conjunto de verbos que simplifican las tareas comunes de transformación y análisis de datos. Estos verbos permiten realizar operaciones como filtrar, seleccionar, agrupar y resumir datos de manera intuitiva, lo que facilita la comprensión y la aplicación de técnicas de manipulación de datos [42]. Nos ayudó para agrupar, filtrar y resumir los datos de ventas, así como para realizar transformaciones necesarias para los gráficos.

shiny

Es un paquete de R que permite a los usuarios crear aplicaciones web interactivas de manera sencilla y directa desde R. Este paquete facilita la integración de código R en una interfaz web [43]. Se empleó para seleccionar parámetros y visualizar gráficos dinámicos basados en los datos de

ventas.

lubridate

Es un paquete diseñado para simplificar el manejo de fechas y horas en R, proporcionando funciones que permiten realizar operaciones que R no soporta de manera nativa [44]. En este caso nos ayudó para manipular y extraer información temporal (como años y meses) de las fechas en los datos de ventas.

plotly

Es una biblioteca de visualización de datos muy versátil, permite crear una amplia variedad de visualizaciones interactivas, incluidas gráficas complejas y mapas [45].

ggplot2

Esta librería de visualización de datos nos permite crear gráficos complejos de manera intuitiva. Es un sistema de visualización de datos en R que utiliza un enfoque declarativo para la creación de gráficos, basado en la gramática de los gráficos [46].

scales

Proporciona herramientas para escalar datos en visualizaciones de ggplot2.Una de las dificultades principales en la creación de gráficos radica en el escalado, que implica convertir valores de datos en propiedades perceptuales [47].

5.3. Análisis exploratorio de datos

Con el dataset limpio y consolidado, se realizó un análisis exploratorio para identificar patrones, tendencias y posibles relaciones en las variables clave. Este análisis sirvió como una etapa preliminar para comprender mejor los datos y guiar la construcción de los modelos predictivos.

Este análisis exploratorio se llevó a cabo utilizando dos herramientas especializadas en el análisis de datos: **R** y **Power BI**. El análisis en **R** se realizó con ayuda de la librería **Shiny**, que permite la creación de gráficos interactivos y facilita la interacción con los usuarios. No obstante, debido a las restricciones de confidencialidad establecidas en el contrato del proyecto, solo se incluyen imágenes estáticas en este documento. Simultáneamente, se realizó un análisis más detallado en **Power BI**, explorando una variedad de gráficos que permitieron alcanzar un nivel de detalle superior. Este informe en **Power BI** fue desarrollado exclusivamente para **MEGATIENDAS** y está destinado únicamente a los involucrados en el desarrollo del proyecto.

Ventas por departamento tendencia mes a mes departamento de Bolívar

Antes de comenzar el siguiente análisis, es importante destacar que el departamento de Bolívar cuenta actualmente con 18 tiendas, lo que representa la mayor participación de MEGATIENDAS en comparación con el departamento del Atlántico. Podemos observar que, durante los últimos tres años, las ventas han mostrado una tendencia al alza. Es importante destacar que el 2021, año

después de la pandemia, registró las cifras más bajas en todos los meses en comparación con los otros dos años. Además, la figura 2 muestra que, tradicionalmente, diciembre ha sido el mes con las ventas más altas. Sin embargo, en 2023 se observó un cambio en esta tendencia. Aunque diciembre mantuvo cifras elevadas, fueron los meses de julio y agosto los que registraron las mayores ventas del año.

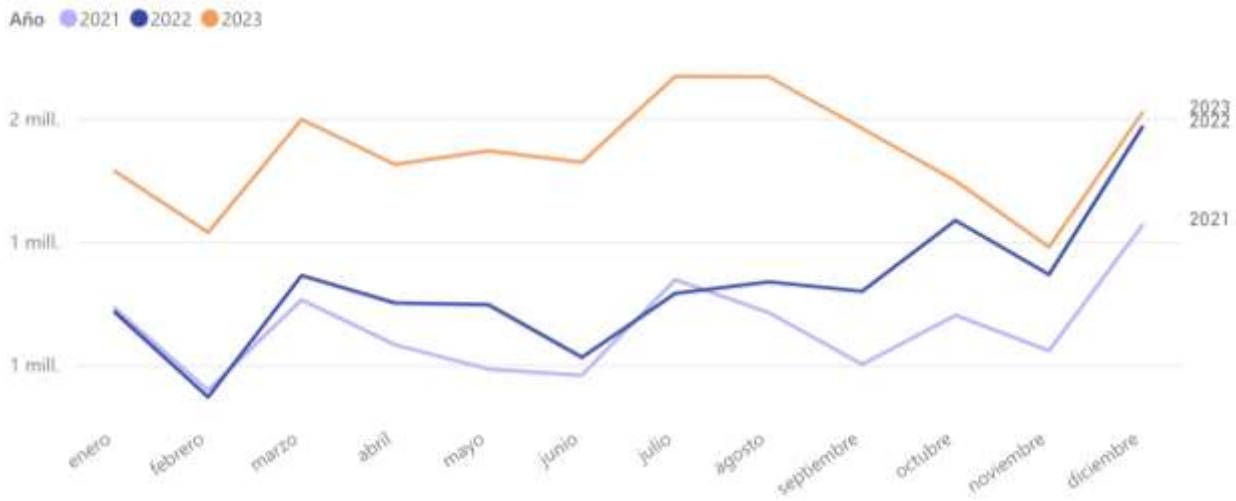

*Figura 2: Unidades vendidas por mes año en las tiendas del departamento de **Bolívar** (Elaboración propia en Power BI)*

En la figura 3, se muestra el porcentaje de participación en ventas por secciones. En este caso, durante los últimos tres años en el departamento de Bolívar, la sección 0062 ha registrado el mayor porcentaje de ventas, con un 46.66% de participación, seguida de cerca por la sección 0061, que alcanza un 39.02%.

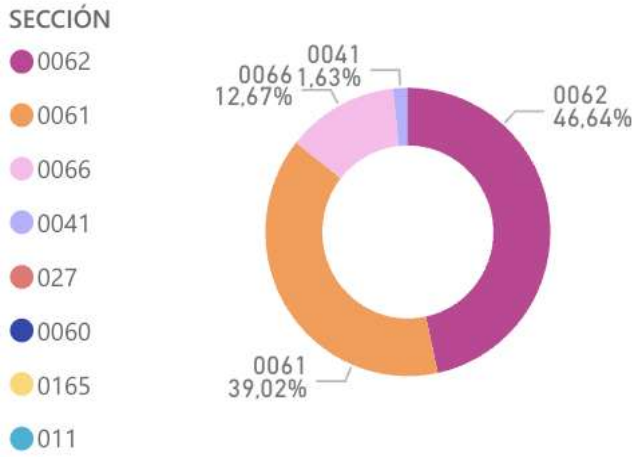

Figura 3: Porcentaje de unidades vendidas por sección del departamento de Bolívar (Elaboración propia en Power BI).

Al analizar las ventas por tienda, se observa que la tienda 101 registró las mayores ventas durante los últimos tres años, seguida de la tienda 103. Todas las tiendas muestran una tendencia de ventas similar a lo largo de estos tres años como lo vemos en la figura 4. Un ejemplo notable es la tienda 107, que consistentemente ha sido la que reporta las menores ventas en este período.

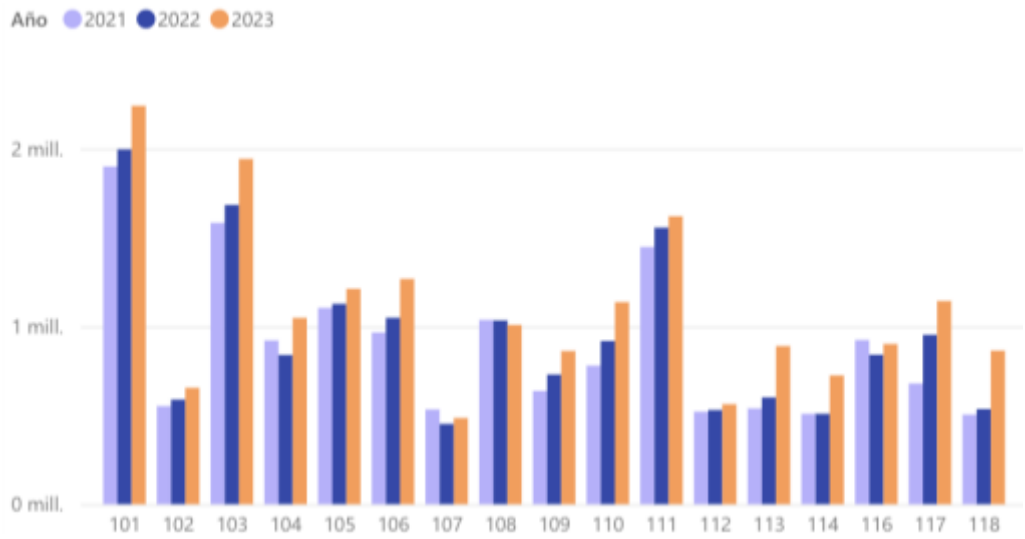

Figura 4: Unidades vendidas por tiendas del departamento de Bolivar (Elaboración propia en Power BI).

Ventas por departamento tendencia mes a mes departamento del Atlántico

En comparación con la región de Bolívar, el departamento del Atlántico registró ventas muy similares de enero a agosto durante los años 2021 y 2022, como se muestra en la figura 5. Observemos que, en los meses de junio y julio, las ventas incluso superaron las de 2022. Por otro lado, se mantiene la tendencia de diciembre como el mes con mayores ventas. Sin embargo, a diferencia de la región de Bolívar, en Atlántico los meses de julio y agosto no alcanzaron las cifras de diciembre, aunque mantuvieron niveles similares, sin llegar a ser los meses con mayores ventas.

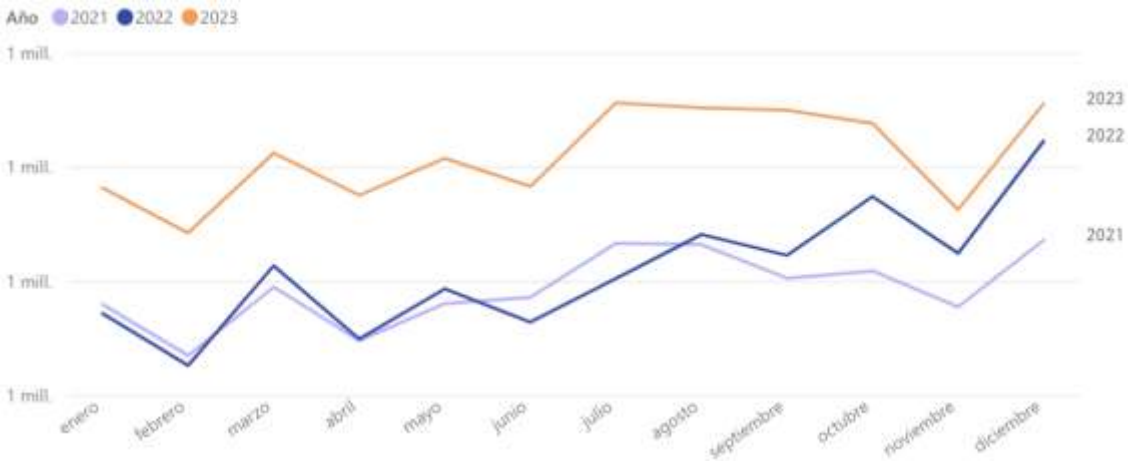

Figura 5: Unidades vendidas por mes año en las tiendas del departamento del Atlántico (Elaboración propia en Power BI.

En la figura 6, se muestra el porcentaje de participación en ventas por secciones. En este caso, durante los últimos tres años en el departamento del Atlántico, la sección 0062 ha registrado el mayor porcentaje de ventas, con un 44.73% de participación, seguida de cerca por la sección

0061, que alcanza un 38.18%. Sigue la misma tendencia del departamento de Bolívar.

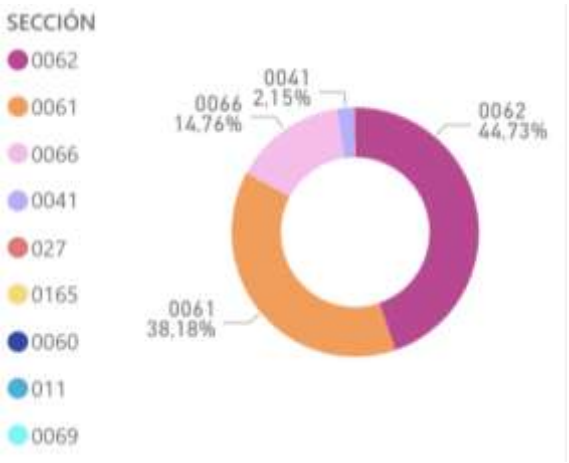

Figura 6: Porcentaje de unidades vendidas por sección del departamento del Atlántico (Elaboración propia en Power BI).

Al analizar las ventas por tienda, se observa que la tienda 201 registró las mayores ventas durante los últimos tres años, seguida de la tienda 205. Aunque la mayoría de las tiendas muestran una tendencia de ventas similar en este período, la tienda 207 destaca en 2023, mostrando un aumento significativo en sus ventas, alcanzando niveles comparables a las demás tiendas como se observa en la figura 7.

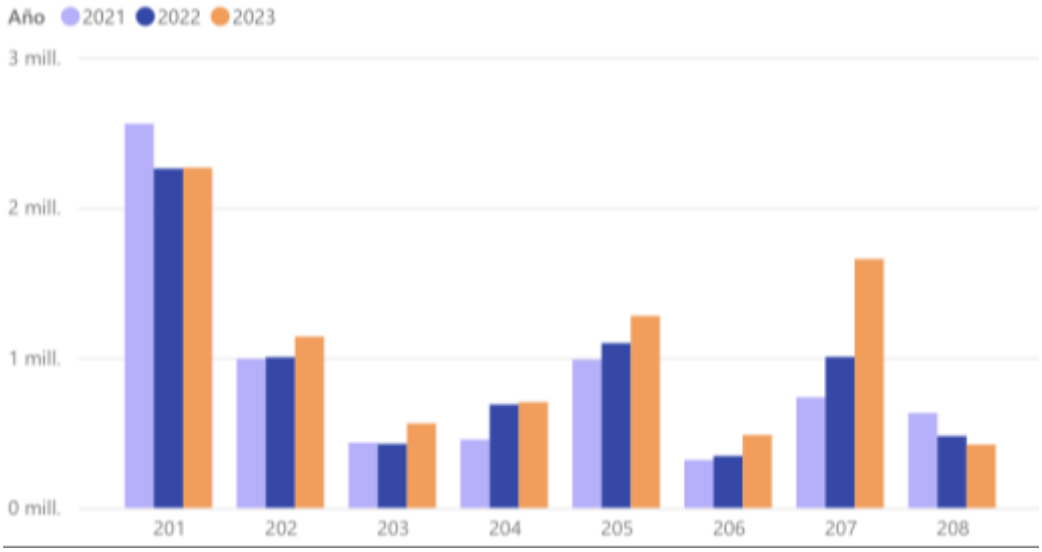

Figura 7: Unidades vendidas por tiendas del departamento del Atlántico (Elaboración propia en Power BI).

De lo anterior, podemos anticipar que los datos seguirán tendencias y patrones similares, lo que facilitará la elaboración de pronósticos. Sin embargo, es importante tener en cuenta que también pueden surgir situaciones particulares o datos atípicos que no se ajusten a estos patrones esperados. Este análisis nos permitió visualizar cómo están distribuidos los datos y qué podemos anticipar durante las etapas de limpieza y transformación de los mismos.

5.3.1. Identificar y excluir ventas al por mayor

En los supermercados Megatiendas, se consideran ventas mayoristas aquellas en las que un solo comprador adquiere una cantidad máxima de un solo producto en una única factura. Estas ventas deben ser excluidas en esta fase, ya que no son muy comunes y son más bien oportunidades excepcionales. De hecho, existen políticas en Megatiendas que prohíben estas ventas. Sin embargo, las tiendas a veces las realizan para evitar que los productos se desperdicien. Para nosotros, estas ventas irregulares podrían causar distorsiones en nuestros modelos de pronóstico. Por ejemplo, en un modelo de pronóstico de demanda, la inclusión de estas ventas mayoristas podría resultar en previsiones inexactas, afectando la precisión y efectividad del modelo. Además, esto podría contribuir a sobrecompras de productos, generando un exceso de inventario y potenciales pérdidas financieras.

Actualmente, estas ventas mayoristas están presentes en nuestros datos y se distinguen de las demás tiendas por tener un centro de operación (C.O.) igual a 1. En nuestro DataFrame final, filtramos solo las tiendas de los departamentos de Bolívar y Atlántico (C.O < 300), excluyendo aquellas con centro de operación 1.

5.3.2. Excluir devoluciones

Las devoluciones, como su nombre lo indica, se refieren a productos que el cliente compró y luego decidió regresar a la tienda. Este proceso puede ocurrir por diversas razones. Una de las más comunes es que, al revisar su compra, el cliente descubre que el producto está en estado de descomposición. Sin embargo, las devoluciones también pueden ocurrir por otros motivos, En estos casos, el cliente tiene la opción de devolver el producto a la tienda, el cliente puede elegir entre varias opciones. Puede recibir un reembolso del dinero pagado, cambiar el producto por otro de la misma categoría o familia de productos, o incluso optar por un producto de una categoría diferente, siempre y cuando el precio sea el mismo.

Las devoluciones no están específicamente marcadas en los datos enviados, lo que puede dificultar su identificación directa. Sin embargo, existe un patrón que nos permite reconocerlas: **las devoluciones suelen aparecer con una cantidad vendida en negativo. Esto indica que el producto ha sido retornado y no representa una venta regular.**

En la siguiente sección de código, abordaremos cómo identifica estas devoluciones utilizando esta característica distintiva. Analizaremos los datos para encontrar entradas con cantidades negativas y, posteriormente, las clasificaremos como devoluciones. Dependiendo del porcentaje de datos que representen las devoluciones, se determinará cómo manejar estos datos.

Si las devoluciones no constituyen la mayoría de los datos, es decir, si están por debajo del 10% del total de los datos, procederemos con su eliminación. Esta estrategia nos permitirá mantener la precisión del análisis de ventas y evitar que las devoluciones influyan desproporcionadamente en los resultados. En cambio, si las devoluciones representan un porcentaje significativo, se implementará un tratamiento especial para integrarlas adecuadamente en el análisis.

El número de registros de devoluciones asciende a aproximadamente 132,286 filas de datos, lo que

representa un 3.35% del total del DataFrame (3’943.083). Este porcentaje indica que las devoluciones son una parte pequeña del conjunto de datos.

Dado que hemos establecido un umbral del 10% para decidir si se debe dar un tratamiento especial a las devoluciones en nuestro análisis, y considerando que estas devoluciones no alcanzan dicho umbral, hemos decidido proceder con la eliminación de todos estos registros del DataFrame.

5.3.3 Eliminación de valores NAS

En el marco de datos presentado inicialmente en la Tabla 1, observamos que la última columna, “Número de NAs”, muestra la cantidad de valores faltantes (NAs) en cada columna. Estos NAs son muy escasos, representando menos del 0.01% del total de los datos. Por lo tanto, procederemos a eliminarlos. Seguimos con nuestro DataFrame obtenido a través de nuestra extracción de datos “Ventas_2021_2024”, el cual hasta este momento consta de 3’943.083 observaciones.

5.4. Consolidación del Dataset Final

Para iniciar el desarrollo del modelo predictivo, finalizamos generando un dataframe limpio, excluyendo las devoluciones y las ventas al por mayor, ya que estos datos son poco relevantes para nuestro proyecto.

Este capítulo detalló cómo el módulo en R permitió extraer, transformar y analizar los datos de manera eficiente. El análisis exploratorio no solo proporcionó insights valiosos sobre las características de los datos, sino que también validó su calidad para etapas posteriores de modelado predictivo. Esto representó un avance significativo hacia el cumplimiento del OE2 y estableció una base sólida para el éxito del proyecto. Además, se confirmó que los datos procesados cumplían con los estándares necesarios para alimentar los modelos predictivos. Algunos hallazgos clave incluyeron:

- **Diferencias regionales:** Las tiendas en la región Atlántico mostraron mayores fluctuaciones en ventas en comparación con Bolívar.
- **Impacto de promociones:** Estas resultaron ser una variable crítica, especialmente para productos perecederos como frutas y verduras.

6. DESARROLLO DEL MODELO PREDICTIVO

En este capítulo, Nos centramos en seleccionar tres modelos prometedores basados en su rendimiento teórico y aplicabilidad. Posteriormente, implementamos estos modelos utilizando nuestros datos históricos de ventas desagregados por fecha, tiendas y productos. Realizamos pruebas rigurosas para evaluar el desempeño de cada modelo en la predicción diaria de la demanda en unidades de FRUVER en Megatiendas. Para asegurar la validez de nuestras conclusiones, diseñamos un esquema de validación robusto que nos permite comparar objetivamente la precisión de cada modelo. Este enfoque nos permite determinar cuál de los modelos, o qué combinación de ellos, ofrece la mejor precisión en el pronóstico de la demanda, asegurando así una mejora significativa respecto a los modelos existentes.

El desarrollo del modelo siguió un flujo de trabajo bien estructurado, dividido en las siguientes etapas principales:

1. **Selección de Variables:**
    - Identificación de variables clave como ventas históricas, promociones y estacionalidad.
    - Análisis de la correlación entre variables para seleccionar las más relevantes.
2. **Preprocesamiento de Datos:**
    - Normalización y escalado de variables para garantizar consistencia en el modelado.
    - Transformación de variables categóricas en factores o dummies según el modelo.
3. **Construcción de Modelos:**
    - Implementación de tres enfoques: ARMA, SARIMAX y Gradient Boosting.
    - Entrenamiento inicial en subconjuntos de datos para ajustar parámetros y comparar rendimiento.
4. **Validación y Evaluación:**
    - División de los datos en conjuntos de entrenamiento y prueba (80%-20%).
    - Cálculo de métricas como MAE, RMSE, R^2 y MAPE para cada modelo.
    - Comparación de resultados para seleccionar el modelo más eficiente.

6.1. Selección de variables

Se identificaron las variables relevantes para el modelo, no limitándonos únicamente a las unidades vendidas, sino también evaluando el impacto que una dinámica comercial tiene al aumentar o disminuir la cantidad de unidades vendidas, contemplando la tienda, día y el producto. Para ello seleccionamos 4 variables particulares que son:

C.O: El código de operación o código de la tienda es fundamental para que el modelo pueda entender y diferenciar el comportamiento de ventas entre distintas tiendas. Esto permite al modelo identificar patrones específicos y categorizar las ventas y tendencias según las características individuales de cada tienda. Incluir esta variable es crucial para comprender variaciones en el desempeño, lo que ayuda a mejorar la precisión del pronóstico y a optimizar estrategias de ventas y distribución a nivel local.

item: Código de operación o código de producto. Permite al modelo entender y diferenciar el comportamiento de ventas entre diferentes productos. Esto es crucial para prever cómo las ventas de cada producto pueden verse afectadas por diversas variables.

Fecha movto.: Fecha de movimiento. Ayuda a capturar la estacionalidad y las variaciones temporales en las ventas, como fluctuaciones diarias, semanales o estacionales. Esto puede mejorar la precisión de las predicciones al considerar factores como días festivos, promociones y otras variaciones temporales.

Cantidad inv: Cantidad inventariada o cantidad vendida. Es fundamental para predecir la cantidad de productos que se venderán en el futuro. Al modelar esta variable, se puede ajustar la oferta y planificar la demanda de manera más efectiva.

6.2. Clusterización de tiendas

Antes de desarrollar modelos predictivos, es esencial realizar la clusterización de tiendas para identificar patrones y relaciones en los datos. Esto permite crear grupos homogéneos y abordar problemas con mayor precisión. Para nuestro caso, proponemos el uso de la medida de similitud DTW (Dynamic Time Warping), que ajusta discrepancias temporales y compara series temporales de distinta longitud [48]. Aunque la distancia euclidiana es común en el aprendizaje automático, no es adecuada para series con desalineación temporal, ya que solo mide la distancia directa entre puntos. Por el contrario, la DTW alinea patrones similares en distintas escalas de tiempo, siendo más flexible y precisa para agrupar tiendas con comportamientos similares, optimizando así el desarrollo de modelos predictivos por bloques. En la tabla 3 veremos las ventajas y desventajas de DTW y Distancia Euclidiana.

Tabla 3

Ventajas y desventajas de DTW y Distancia Euclidiana (Elaboración propia).

Característica	**DTW (Dynamic Time Warping)**	**Distancia Euclidiana**
Ventajas		
Maneja discrepancias temporales	Sí	No
Comparación flexible	Sí	No
Identifica patrones similares	Sí, incluso si no están sincronizados en el tiempo	Solo si están perfectamente alineados
Adecuado para series de distinta longitud, requiere series de igual longitud	Sí	No, requiere series de igual longitud
Aplicaciones	Útil en reconocimiento de patrones y series temporales no alineadas	Útil en análisis de datos alineados y métricas simples

Desventajas Complejidad computacional	Alta, debido a la necesidad de calcular alineaciones óptimas	Baja, cálculo directo
Sensibilidad al ruido	Puede ser sensible al ruido, aunque menos que la distancia euclidiana si se aplica suavizado	Alta, puede ser muy sensible al ruido y a pequeñas variaciones
Interpretación	Más compleja, requiere comprensión de la alineación temporal	Simple, diferencia directa.
Necesidad de preprocesamiento	Puede requerir preprocesamiento para reducir ruido y optimizar la alineación	Generalmente menos preprocesamiento necesario

6.2.1. Análisis con DTW

1 - Agrupamos las unidades vendidas por tienda y fecha.

Este paso es crucial porque la estructura de nuestro DataFrame final, tras la selección de variables, incluye la fecha, el código de tienda, el artículo y la cantidad de inventario. Por lo tanto, para analizar la similitud entre las tiendas, es absolutamente necesario agrupar los datos por tienda, creando un nuevo DataFrame, diferente del DataFrame final, en el que conservamos la estructura de fecha, código de tienda y la suma de las cantidades de inventario. En el anexo 1 podemos encontrar detalladamente el proceso descrito anteriormente.

2 - Pivotar el Data Frame para reorganizar los datos.

Estructurar los datos con fechas como columnas y tiendas como filas permite representar cada tienda como una serie temporal independiente, facilitando la comparación de similitudes entre tiendas mediante DTW. Esto es esencial para calcular las similitudes, realizar clustering y mejorar la interpretabilidad y visualización de los resultados, ya que este formato es ideal para analizar y comparar el comportamiento de las tiendas a lo largo del tiempo. En el anexo 1 podemos encontrar detalladamente el proceso descrito anteriormente.

3 - Eliminamos la columna de tiendas para crear la matriz de datos.

Convertir la columna de tiendas en nombres de filas y luego transformar los datos a una matriz tiene como objetivo estructurarlos de manera que sean directamente utilizables por el algoritmo DTW. Esta organización optimizo la ejecución de dichos algoritmos, asegurando que cada serie temporal esté correctamente identificada y asociada con su tienda correspondiente. En el anexo 1 podemos encontrar detalladamente el proceso descrito anteriormente.

4 - Matriz de distancia y aplicación del algoritmo DTW.

Calculamos la matriz de distancias entre cada par de tiendas utilizando la métrica DTW, que mide

la distancia entre sus series temporales permitiendo desalineaciones en el tiempo [49]. A partir de esta matriz de distancias, aplicamos el algoritmo de clustering jerárquico para agrupar las tiendas según la similitud de sus series temporales.

5 - Guardamos los resultados obtenidos.

Se calcula la asignación de 4 clusters para todas las tiendas, se añade esta información al Data Frame original para facilitar la visualización y se crea un nuevo Data Frame ClusteringTiendas con los nombres de las tiendas y sus números de clúster, útil para análisis y visualización. En el anexo 1 podemos encontrar detalladamente el proceso descrito anteriormente.

6 - Visualizamos los Clusters.

El análisis de clústeres jerárquico es una técnica de agrupamiento que permite organizar objetos o casos en una jerarquía de grupos (clusters). Los gráficos más comunes en este análisis son los **dendrogramas** y los **diagramas de témpanos**. El dendrograma representa visualmente cómo se agrupan los casos y qué tan cohesionados están los clústeres, lo que facilita la selección del número adecuado de clústeres [50]. El diagrama de témpanos muestra cómo los casos se combinan en diferentes iteraciones del análisis, y puede presentarse tanto en orientación vertical como horizontal [51]. En este caso generamos un dendrograma mostrado en la figura 8 utilizando la función fviz_dend del paquete factoextra, que permite visualizar cómo se agrupan las observaciones en diferentes clusters. La línea horizontal en el gráfico indica el nivel de corte utilizado para formar los clusters. En el anexo 1 podemos encontrar detalladamente el proceso descrito anteriormente.

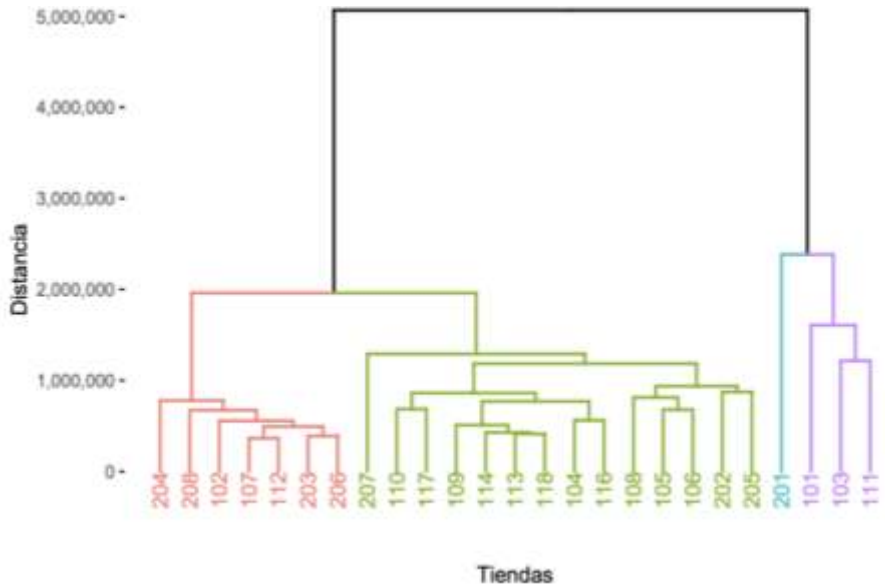

Figura 8: Dendograma del Clustering de Tiendas (Elaboración propia en R).

El análisis de similitud de series temporales mediante el uso del algoritmo DTW fue altamente efectivo, permitiendo agrupar las tiendas en 4 clusters distintos. Estos clusters se formaron en función de las similitudes identificadas en las series temporales de unidades vendidas en cada

tienda, lo que facilita una comprensión más profunda de los patrones de comportamiento entre los diferentes grupos. Este proceso no sólo permitió categorizar las tiendas según sus patrones de unidades vendidas a lo largo del tiempo, sino que también sienta las bases para la estrategia que desarrollaremos a continuación.

6.2.2. Muestra intencionada basada en la clasificación de unidades vendidas.

Para desarrollar el pronóstico de unidades vendidas, seleccionaremos una tienda representativa de cada clúster utilizando el dendrograma mostrado en la Figura 8. Este enfoque nos permitirá identificar patrones de comportamiento y tendencias específicas dentro de cada grupo, facilitando así una mejor estimación de las ventas. Dentro de cada una de las tiendas, se elegirá una muestra representativa de los productos con más unidades vendidas. De un total de aproximadamente 150 productos por tienda, se seleccionarán los 30 productos con mayores unidades vendidas. Dentro de este grupo, se hará una selección intencionada para garantizar la inclusión de productos con diferentes niveles de ventas: los dos productos más vendidos, dos productos con ventas intermedias (posiciones 15 y 16), y los dos productos con las ventas más bajas dentro del Top 30 (posiciones 29 y 30). Este enfoque no solo permitirá evaluar el desempeño de los modelos de pronóstico en productos con distintos niveles de demanda, desde los más populares hasta los menos vendidos, sino que también facilitará una retroalimentación rápida, lo que nos ahorrará tiempo, permitiéndonos enfocarnos en los ajustes necesarios para optimizar nuestros modelos de pronóstico.
Para llevar a cabo el procedimiento anterior, se desarrolló una función en R denominada Funct_TopItems. A continuación, se presentará la función en su totalidad, seguida de una explicación detallada de cada una de sus partes y su respectiva funcionalidad. En el anexo 1 podemos encontrar detalladamente el proceso en R descrito a continuación.

1. Cálculo de la fecha de inicio: En la variable primer_dia_hace_dos_meses, calculamos el primer día del mes que corresponde a dos meses antes de la fecha máxima en UnidadesVendidas. Por ejemplo, si la última fecha es “2024-06-30”, primer_dia_hace_dos_meses será “2024-05-01”, lo que nos permite filtrar las ventas de los últimos dos meses. La variable DatosTienda recibe el identificador de una tienda y utiliza primer_dia_hace_dos_meses para filtrar UnidadesVendidas, generando un dataframe con las ventas de los últimos dos meses para esa tienda.

2. Filtrado de datos por tienda y fecha: Una vez filtrados los datos de unidades vendidas por tienda para los últimos dos meses, se agrupan por producto (ítem) y se suman las unidades vendidas de cada uno. Luego, se ordenan en orden descendente según la cantidad total de unidades vendidas, seleccionando los 30 productos más vendidos utilizando head (30).

3. Creación de un gráfico: En el anterior segmento de código lo que hacemos es crear un gráfico de barras con los 30 productos más vendidos utilizando ggplot2, a partir de la data que ya viene filtrada del paso anterior. Además, se destacan con rectángulos rojos las posiciones de los productos seleccionados los dos más vendidos (1,2), los dos de ventas medianas (14,15), y los dos menos vendidos (29,30).

4. Selección de los productos para el pronóstico: ItemsApronosticar guarda las posiciones de los

datos filtrados para posteriormente cargar un archivo CSV en nuestro drive, como se explica a continuación

5. Gestión de archivos en Google Drive: carpeta_existente guarda el identificador único creado por Google Drive para cada carpeta. Cada carpeta lleva el mismo nombre que la tienda ingresada en la función. En la línea del if, se valida si la carpeta ya existe; si el identificador para esa carpeta existe, se procede a eliminar dicha carpeta. La variable carpeta_proyecto almacena la ruta donde se guardarán los productos a pronosticar. carpeta_tienda crea una nueva carpeta con el nombre de la tienda en la ruta especificada por carpeta_proyecto. archivo_temp crea un archivo temporal en memoria, y write.csv escribe el contenido de ItemsApronosticar en este archivo temporal. Luego, drive_upload sube el archivo temporal a la nueva carpeta creada en Google Drive, utilizando el nombre de la tienda para el archivo CSV. Finalmente, unlink(archivo_temp) elimina el archivo temporal de la memoria.

6. Gráfico: Ejecutamos la función para cada una de nuestras tiendas elegidas en nuestra clusterización y validemos los resultados obtenidos, con la creación de la carpeta en google drive en la ruta Proyecto R/TransformacionDeDatos con los identificadores de los productos a pronosticar. A continuación, en la figura 9 mostramos el gráfico almacenado en el paso 3.

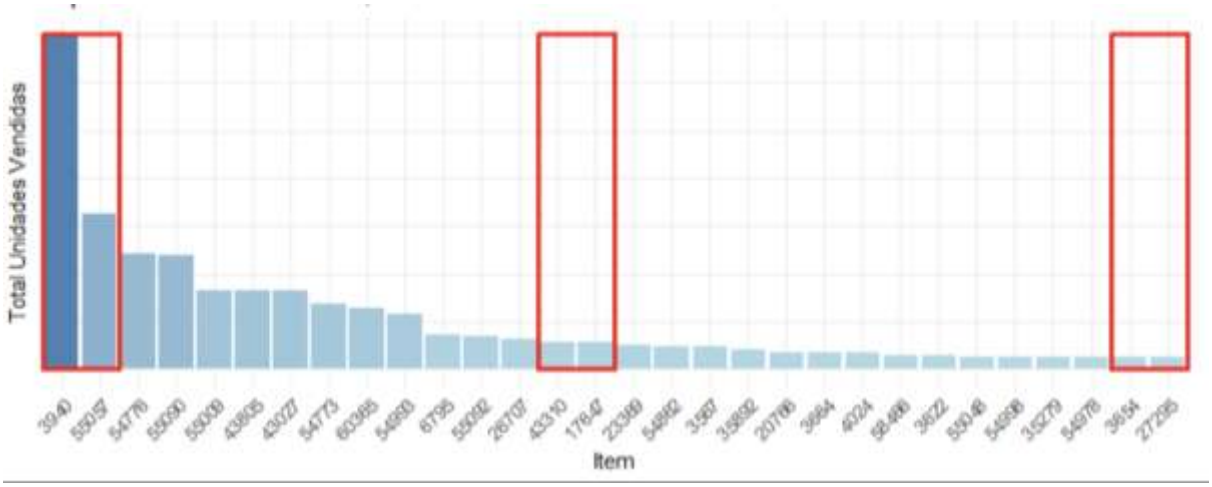

Figura 9: Gráfico resultado de la función Funct_TopItems(102) (Elaboración propia en R).

6.3. Investigación y evaluación de modelos

6.3.1. Modelo ARMA

La estimación y el pronóstico utilizando un modelo ARMA (AutoRegressive Moving Average) es una técnica ampliamente utilizada en el análisis de series temporales. Los modelos ARMA combinan componentes autoregresivos (AR), que capturan la dependencia de un valor presente con valores pasados, con componentes de promedios móviles (MA), que modelan la relación de un valor presente con errores pasados. Esta combinación permite a los modelos ARMA captar patrones complejos en los datos, proporcionando una base robusta para la estimación y el

pronóstico [16]. Al ajustar un modelo ARMA a una serie temporal, es posible identificar y predecir las tendencias y fluctuaciones futuras con mayor precisión, lo que lo convierte en una herramienta esencial en la planificación y toma de decisiones en diversos campos, como la economía, la ingeniería y las ciencias sociales [16].El flujograma presentado a continuación en la figura 10, muestra los pasos seguidos, que incluyen la preselección de parámetros (p, q), el ajuste del modelo, la evaluación de residuos y la validación mediante métricas estándar. Este enfoque permitió establecer una línea base para comparar el rendimiento de los modelos más avanzados.

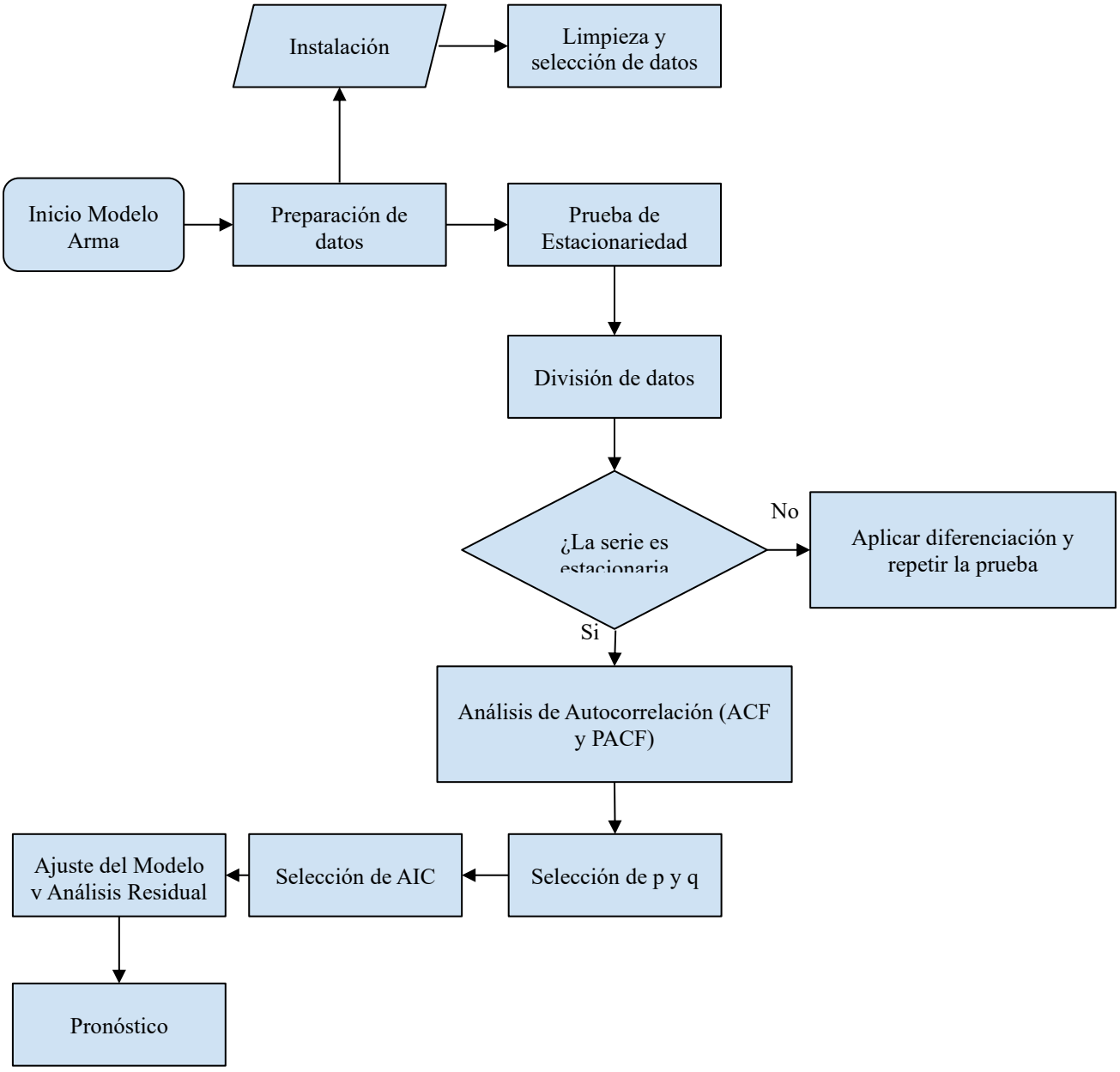

Figura 10: Diagrama de flujo modelo ARMA

- **Librerías**:

Comenzamos instalando las librerías necesarias: Pandas, NumPy, Matplotlib y Statsmodels, las cuales nos permitieron manipular los datos de manera eficiente. Luego, procedimos con la limpieza de las columnas y filtramos las variables de interés, que en nuestro caso fueron Fecha de Venta y Unidades Vendidas, centrando el análisis en las tiendas seleccionadas.

6.3.1.2 Prueba de Estacionariedad:

La estacionariedad es un concepto fundamental en el análisis de series temporales, ya que muchas técnicas de modelado y pronóstico asumen que las series son estacionarias. La identificación de la no estacionariedad es crucial para aplicar transformaciones adecuadas, como la diferenciación, que pueden ayudar a estabilizar la serie y facilitar su análisis. El uso de pruebas de estacionariedad permite a los analistas determinar la naturaleza de la serie y seleccionar el modelo más apropiado para el análisis y pronóstico. Para nuestro caso usaremos el método de Dickey-Fuller.

Test de Dickey-Fuller:

La prueba aumentada de Dickey-Fuller evalúa la estacionariedad de una serie temporal proporcionando hipótesis, estadísticas de prueba y valores p. La estadística de la prueba indica si se debe rechazar la hipótesis nula de que los datos no son estacionarios, mientras que el valor p mide la evidencia contra esta hipótesis. Para decidir si se deben diferenciar los datos, se compara la estadística de prueba con el valor crítico y el valor p con niveles de significación específicos (0.01, 0.05, 0.10). Si la estadística de prueba es menor o igual al valor crítico o el valor p es menor o igual al nivel de significación, se rechaza la hipótesis nula, lo que quiere decir que los datos son estacionarios. Si no, se considera la diferenciación [52].

Posteriormente, realizamos una prueba de estacionariedad utilizando el Test de Dickey-Fuller Aumentado (ADF) para verificar si las series temporales eran estacionarias. Para la tienda 102 obtuvimos un p-valor de 0.0009446 que es menor a 0.05, lo que indica que la serie es estacionaria. Este análisis se realizó para las 4 tiendas que pronosticamos, y en todas se confirmó que las series temporales eran estacionarias.

- **División de datos**:

Cabe resaltar que inicialmente, contábamos con 1025 registros para cada uno de los ítems a modelar. Sin embargo, durante las pruebas del pronóstico, observamos que, al reducir el número de datos, el ajuste del modelo mejoraba. Finalmente, obtuvimos mejores resultados utilizando series de entre 200 y 300 registros por tienda. Para la división de los datos, empleamos el 80% para el entrenamiento del modelo y el 20% restante para pruebas. Los datos de prueba abarcan el último mes disponible en la base de datos, correspondiente a junio. Esta estrategia permitió un ajuste preciso y eficiente en el pronóstico.

6.3.1.3 Análisis de Autocorrelación:

A continuación, realizamos un análisis de Autocorrelación (ACF) y Autocorrelación Parcial (PACF) con el objetivo de identificar si los valores pasados de la serie están correlacionados con los valores futuros. En el gráfico de ACF, observamos que los primeros retardos muestran una alta correlación, lo que indica que las ventas recientes tienen una influencia considerable sobre las ventas actuales. A medida que aumenta el número de días, la correlación disminuye gradualmente, lo que sugiere que las ventas de días anteriores aún ejercen cierto efecto, pero con menor intensidad. Este patrón de autocorrelación sugiere una fuerte dependencia a corto plazo, lo que es crucial para modelar la serie adecuadamente. Por otro lado, el gráfico de PACF revela que los primeros rezagos tienen una autocorrelación significativa, lo que confirma que las ventas recientes impactan de manera notable en los valores actuales. Sin embargo, esta influencia disminuye rápidamente conforme aumentan los rezagos, lo que sugiere que la dependencia de los valores pasados es limitada a corto plazo. Este comportamiento indica que un modelo con un componente AR de bajo orden podría ser suficiente para capturar de manera efectiva la dinámica de las ventas.

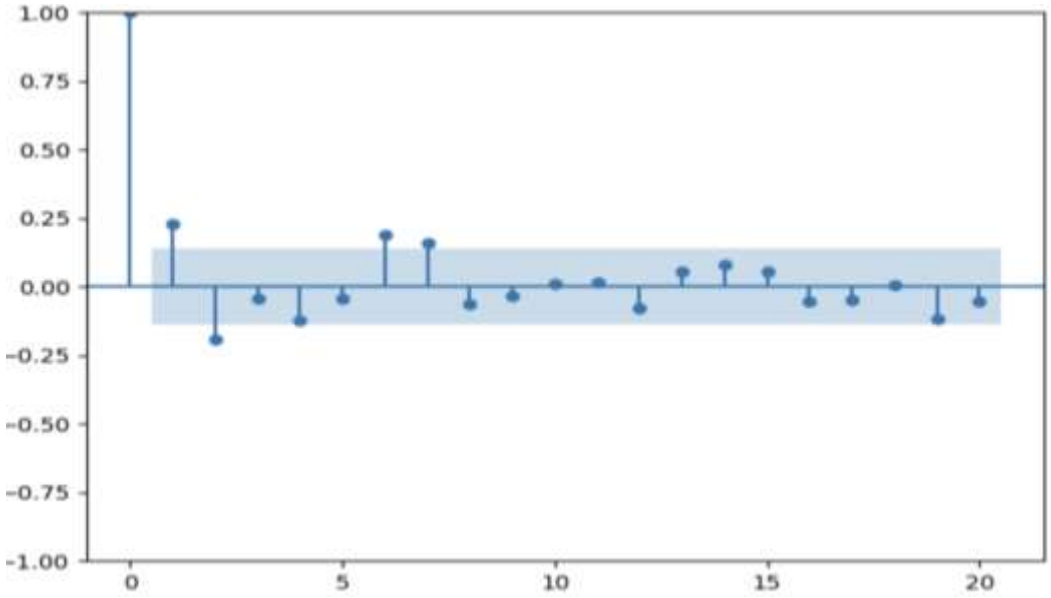

Figura 11: Autocorrelación parcial (Elaboración propia en R).

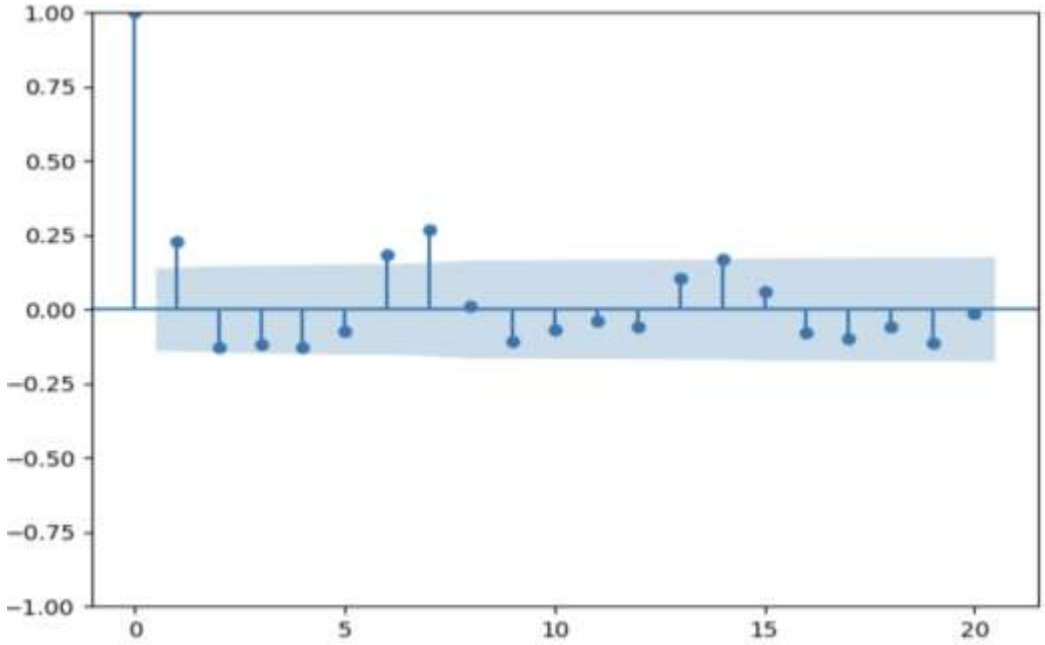

Figura 12: Autocorrelación (Elaboración propia en R).

- **Selección de AIC:**

Para la selección de los valores óptimos de p y q, utilizamos la función optimize_ARMA, la cual nos permitió evaluar diferentes combinaciones de los parámetros autorregresivos (AR) y de media móvil (MA). El criterio utilizado para seleccionar el mejor modelo fue el AIC (Criterio de Información de Akaike), que ayuda a equilibrar el ajuste del modelo y su complejidad, seleccionando el modelo que minimiza este valor. De esta manera, logramos identificar el modelo adecuado con un buen nivel de precisión y sin sobreajuste. De este modo, se obtuvieron los valores óptimos de p = 5 y q = 7, logrando un AIC de 1280.889828.

6.3.1.4 Análisis Residual

Una vez seleccionado el modelo con los mejores parámetros (p, q), se ajusta y se realiza un análisis residual. Los residuos deben comportarse como ruido blanco, lo que significa que el modelo captura adecuadamente la estructura de los datos como se muestra en la figura 13.

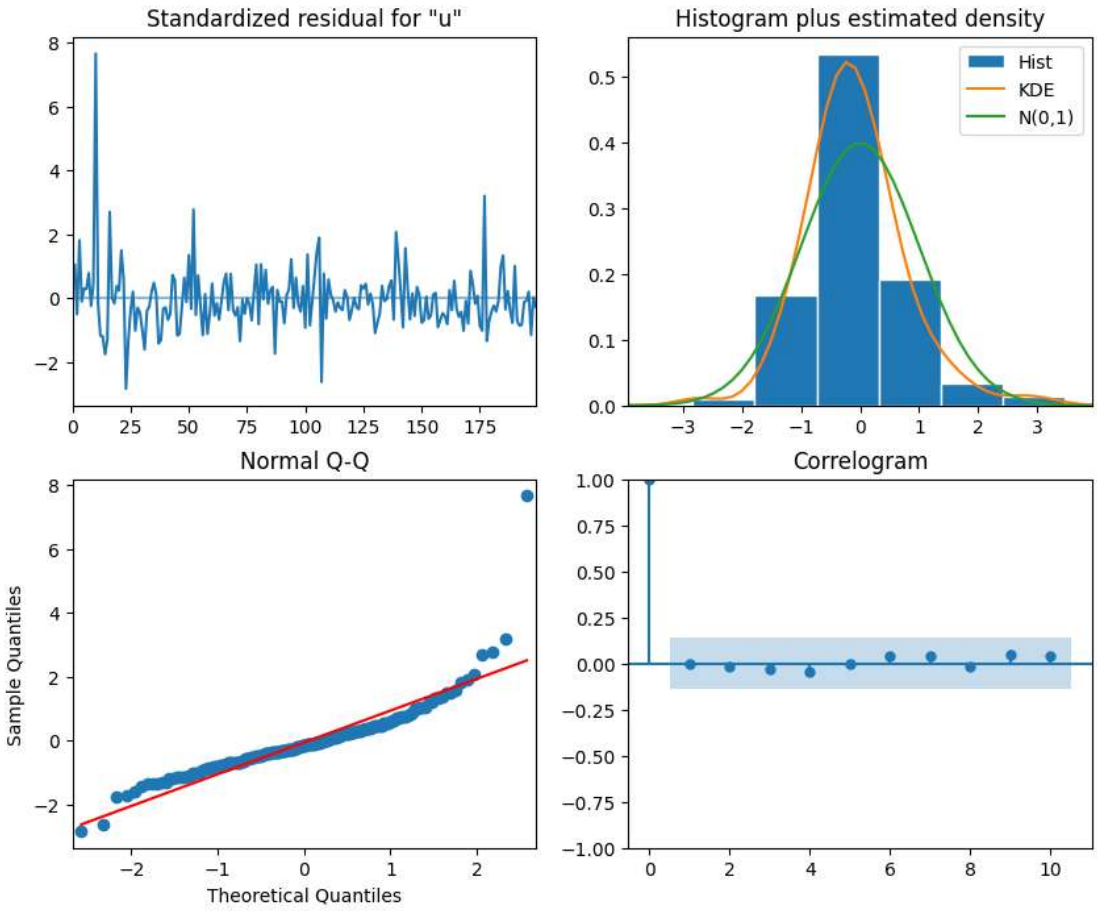

Figura 13: Figura 13: Gráfico de residuos (Elaboración propia en R).

Residuals (Residuos):

- La primera gráfica muestra los residuos a lo largo del tiempo. Podemos observar que los residuos se comportan como ruido blanco, lo que significa que no presentan patrones claros. Esto quiere decir que el modelo está capturando adecuadamente la estructura de los datos.

Histograma de los Residuos y Densidad:

- La forma del histograma muestra una distribución que se aproxima a la normalidad, con un pico alrededor de cero y colas en ambas direcciones. La línea de la densidad KDE sigue el patrón del histograma y se alinea relativamente bien con la curva normal, por lo tanto, los residuos son aproximadamente normales. Esto es un buen indicio de que el modelo es apropiado.

Q-Q Plot (Gráfico Cuantil-Cuantil):

- La mayoría de los puntos están alineados cerca de la línea roja, lo que significa que los residuos siguen una distribución normal. Sin embargo, hay algunas desviaciones en los extremos

(colas). Por lo tanto, podemos decir que los residuos son en su mayoría normales

Correlograma (ACF) de los Residuos:

- La mayoría de las autocorrelaciones se encuentran dentro del intervalo de confianza, lo que significa que no hay correlaciones significativas entre los residuos en diferentes retardos. Por lo tanto, el modelo ha capturado todas las relaciones temporales relevantes en los datos.

Prueba Ljung-Box: Esta prueba verifica si las autocorrelaciones en los residuos son significativas. En este caso, los p-valores obtenidos son mayores a 0.05, lo que indica que los residuos son independientes y, por lo tanto, el modelo es adecuado para realizar pronósticos.

6.3.2. Modelo SARIMAX

El modelo SARIMAX proporciona una herramienta poderosa y flexible para el pronóstico de series temporales al permitir la inclusión de variables exógenas. Esto amplía las capacidades del modelo SARIMA, permitiendo capturar efectos externos que puedan influir en la serie temporal [13]. Al comprender las diferentes combinaciones de estacionalidad y variables exógenas, se pueden adaptar modelos específicos a problemas concretos, lo que mejora la precisión y relevancia de los pronósticos. La correcta implementación y codificación de variables exógenas es clave para maximizar el rendimiento de estos modelos. El flujograma presentado a continuación en la figura 14, detalla el flujo de trabajo utilizado, que incluye la selección de parámetros estacionales, la incorporación de variables exógenas y la validación de las predicciones. Este modelo permitió capturar mejor las fluctuaciones regulares y los factores externos que influyen en la demanda.

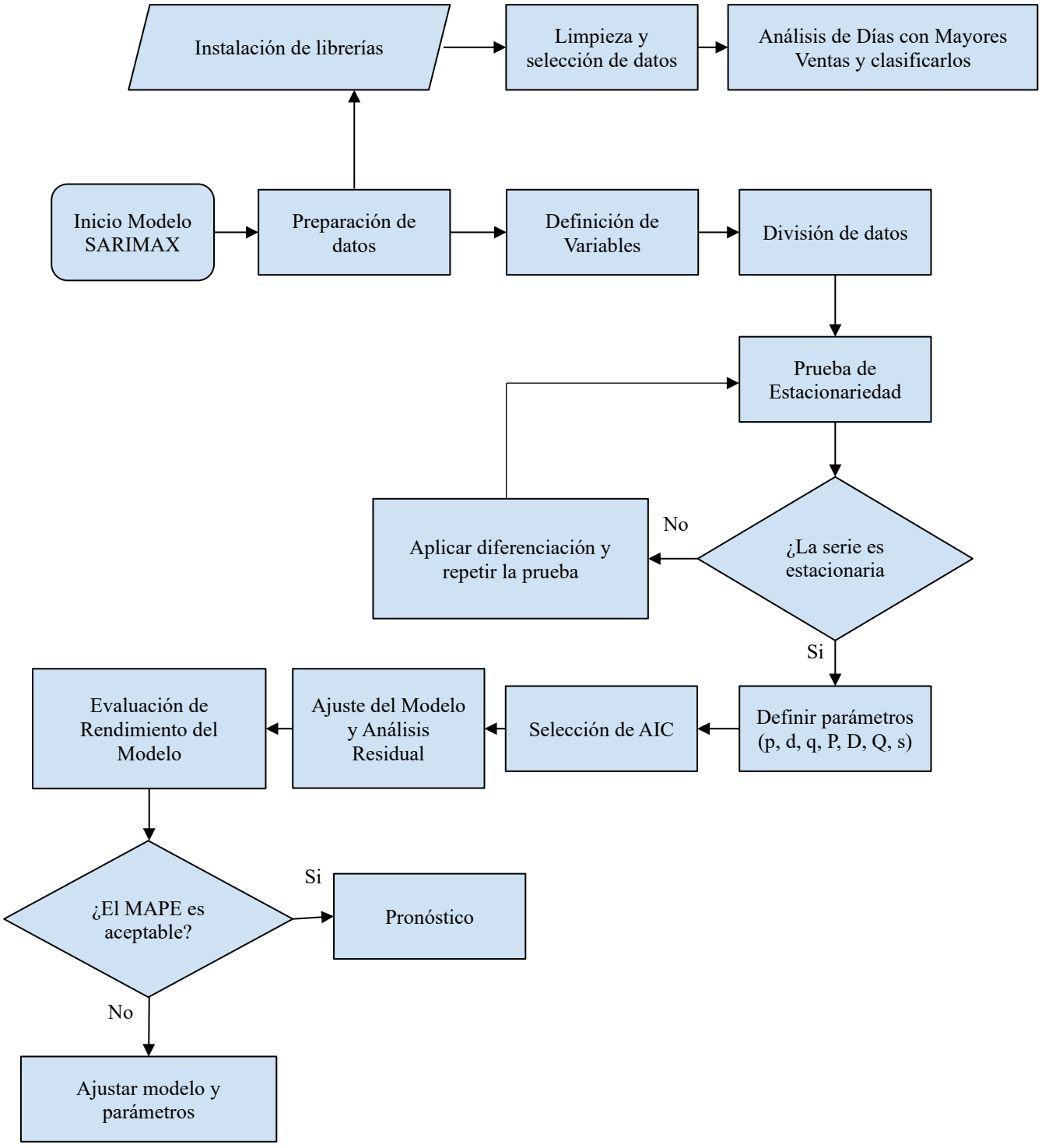

Figura 14: Diagrama de flujo modelo SARIMAX

6.3.2.1 Estimación y Pronóstico con un Modelo SARIMAX

Para la estimación del modelo SARIMAX, es fundamental tener en cuenta que, a diferencia del modelo ARMA, este modelo incorpora variables exógenas. En nuestro caso, consideraremos las promociones vigentes, lo cual abordaremos más adelante. Para el desarrollo del modelo, se seguirán los siguientes pasos:

Inicialmente, realizamos un análisis de los días con mayores ventas, lo que nos proporciona una visión clara de los patrones semanales de demanda. Para ello, creamos una nueva columna llamada 'weekday', que contiene el número correspondiente a cada día de la semana (del 0 al 6). Esta columna nos permite identificar los días en los que un producto específico presenta un rendimiento superior en términos de unidades vendidas. A continuación, agrupamos los datos por día de la semana y sumamos las unidades vendidas para cada uno de estos días. Este enfoque nos ayuda a determinar cuáles son los días más propicios para la venta de cada producto.

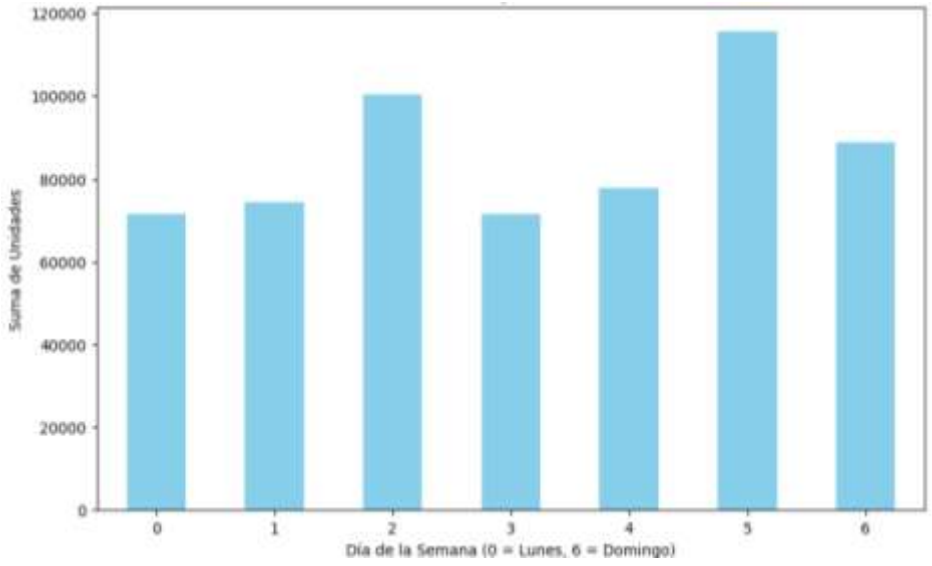

Figura 15: Unidades vendidas por día de la semana (Elaboración propia en R).

La figura 10 muestra la suma de unidades vendidas por día de la semana, donde se observa un patrón claro de demanda. Los días miércoles (2) y sábado (5) destacan como los días con mayor volumen de ventas, superando significativamente las 100,000 unidades, mientras que el lunes (0) y el jueves (3) presentan volúmenes menores, pero aún considerables. Los días martes (1), viernes (4) y domingo (6) muestran un comportamiento intermedio, con ventas que rondan las 70,000 a 80,000 unidades. Esto quiere decir que las estrategias de inventario y promoción podrían enfocarse en estos días de alta demanda para maximizar los ingresos. Finalmente, se calcula la media de ventas diarias, y cada día se clasifica como un día de "alta" o "baja" venta en función de si sus ventas están por encima o por debajo de esta media, información que luego se utiliza como una variable exógena en el modelo SARIMAX para mejorar la precisión de las predicciones. Si su valor es superior a la media, se marcará en el data frame que se pasará al modelo con un 1; de lo contrario, se marcará con un 0.

- **Definición de Variables Objetivo y Exógenas:**

En esta fase, se definen las variables objetivo y exógenas que alimentarán el modelo de predicción. La variable objetivo, denominada target, corresponde a la serie temporal de las unidades vendidas (unidades), que se utilizará para pronosticar futuras ventas. Se selecciona un subconjunto de los datos históricos para establecer esta variable, limitando su longitud a un número específico de filas (NumeroFilas) igual a 1247. Por otro lado, la variable exógena, llamada exog, se deriva de la columna DiasVentas, que indica si el día de la semana en que ocurrió la venta es considerado un día de alta demanda, basado en el análisis previo. Estas variables son fundamentales para entrenar el modelo SARIMAX, que busca capturar tanto las tendencias estacionales como los efectos exógenos en las ventas futuras.

6.3.2.2 Verificación de Estacionariedad:

Se verifica si la serie temporal es estacionaria, Este paso es fundamental porque muchos modelos de series temporales, incluido SARIMAX, asumen que los datos son estacionarios. Detectar y corregir la no estacionariedad es vital para evitar errores de modelado y asegurar que las predicciones sean confiables. Para nuestra prueba, aplicamos la prueba de Dickey-Fuller definido en el subcapítulo anterior, para determinar si la serie temporal es estacionaria. Dependiendo del resultado, se decide el grado de diferenciación (valor d) que se utilizará en el modelo. En nuestro caso obtuvimos como resultado lo siguiente:

P-valor menor a 0.05: El valor p (0.002) es menor a 0.05, lo que nos permite rechazar la hipótesis nula de que la serie temporal tiene una raíz unitaria. Por lo tanto, la serie temporal es estacionaria sin necesidad de aplicar transformaciones adicionales como la diferenciación.

6.3.2.3 Determinación de Parámetros de Diferenciación (Set d and D):

Una vez que la serie es estacionaria, se establece el valor del parámetro d, que representa el número de diferenciaciones necesarias para eliminar tendencias. También se determina D, que indica cuántas diferenciaciones estacionales son necesarias para remover los patrones estacionales. Estos parámetros son esenciales para ajustar correctamente el modelo a la estructura subyacente de los datos, mejorando así su capacidad de predicción.

- **p = range(0, 52, 1):** Define un rango de valores para el parámetro p (el orden de la autoregresión). Los valores irán desde 0 hasta 51.
- **d = 1:** Fija el parámetro d en 1, lo que significa que se aplicará una sola diferenciación no estacional para hacer que la serie sea estacionaria.
- **q = range(0, 52, 1):** Define un rango de valores para el parámetro q (el orden de la media móvil). Los valores irán desde 0 hasta 51.
- **P = range(0, 52, 1):** Define un rango de valores para el parámetro P (el orden de la autoregresión estacional). Los valores irán desde 0 hasta 51.
- **D = 1:** Fija el parámetro D en 1, lo que significa que se aplicará una sola diferenciación estacional para eliminar la estacionalidad.

- **Q = range(0, 52, 1):** Define un rango de valores para el parámetro Q (el orden de la media móvil estacional). Los valores irán desde 0 hasta 51.
- **s = 52:** Establece la periodicidad estacional en 52, lo que podría indicar una estacionalidad semanal en un conjunto de datos anual

- **Selección de AIC:** La función product de la biblioteca itertools se utiliza para generar todas las combinaciones posibles de los parámetros dentro de los rangos especificados, creando una lista completa de configuraciones a evaluar. Esta exhaustiva búsqueda de parámetros busca encontrar la mejor combinación que minimice el criterio de información Akaike (AIC), optimizando así el rendimiento del modelo SARIMAX para la serie temporal dada. En él se enumeran posibles valores para los parámetros p, q, P, y Q. Estos representan, respectivamente, los órdenes del modelo autoregresivo (AR), la media móvil (MA) y sus componentes estacionales correspondientes [13]. Para ello se probó diferentes combinaciones de estos parámetros permitiendo explorar múltiples estructuras de modelado, ayudando a identificar la configuración que mejor captura las dinámicas de la serie temporal. En nuestro caso obtuvimos los siguientes valores para nuestros parámetros p = 2, q = 1, d = 0, P = 1, Q = 0, D = 0, s = 7 con un AIC de 1660.642.

6.3.2.4 Análisis de residuos:

Una vez seleccionado el modelo con los mejores parámetros, se ajusta y se realiza un análisis residual. Los residuos deben comportarse como ruido blanco, lo que significa que el modelo captura adecuadamente la estructura de los datos, para ello obtenemos lo siguiente:

Gráfico de Residuos Estandarizados:

Los residuos parecen estar distribuidos aleatoriamente alrededor de cero, sin patrones obvios. En general, la ausencia de un patrón claro lo que quiere decir que el modelo está capturando adecuadamente la dinámica de la serie temporal.

Histograma y Densidad Estimada:

La distribución de los residuos parece aproximarse a una normal, ya que el histograma muestra una forma similar a una campana. No obstante, hay ligeras asimetrías y colas que podrían indicar que los residuos no son perfectamente normales.

Gráfico Q-Q (Cuantil-Cuantil):

Los puntos en el gráfico Q-Q siguen aproximadamente la línea roja diagonal, lo que significa que los residuos están distribuidos de manera similar a una distribución normal en la mayoría de los cuantiles.

Correlograma (ACF) de los Residuos:

La mayoría de las autocorrelaciones se encuentran dentro del intervalo de confianza, lo que significa que no hay autocorrelación significativa en los residuos. Por lo tanto, el modelo ha

capturado todas las dependencias temporales en los datos como se muestra en la figura 16.

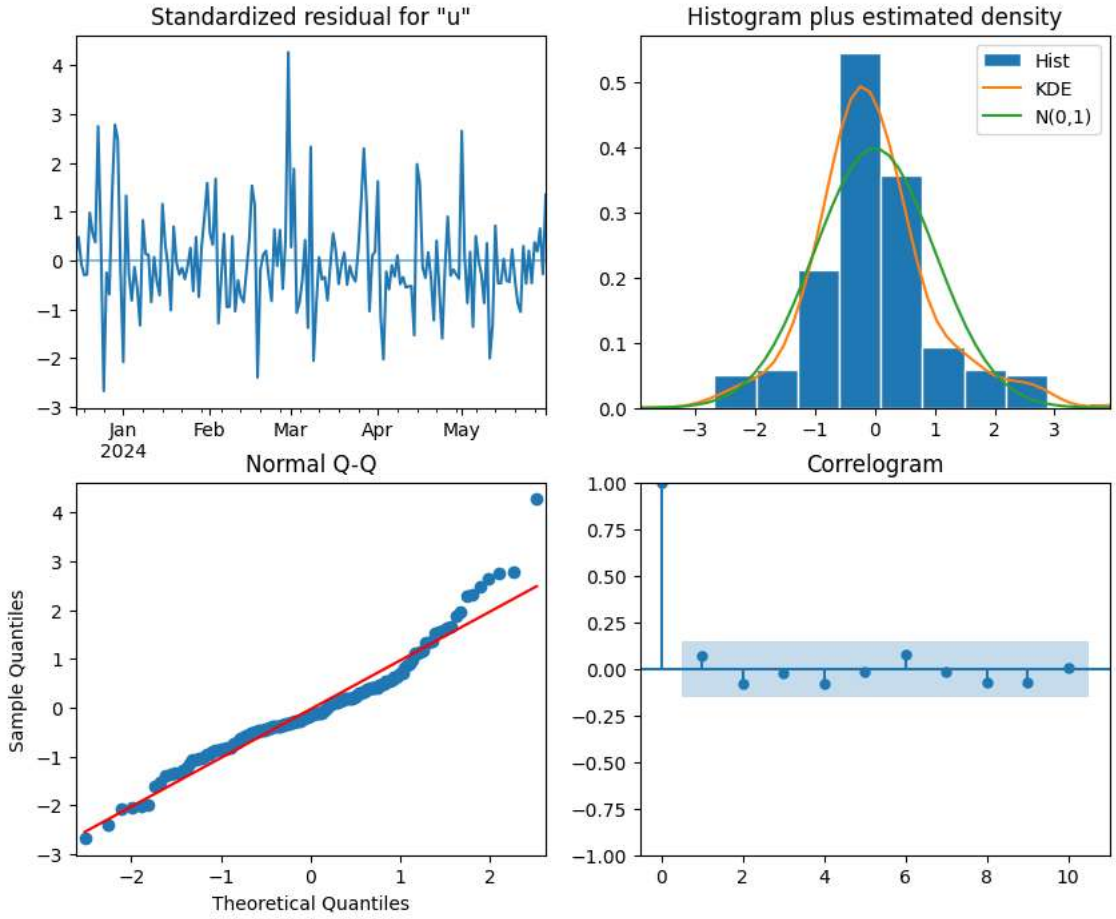

Figura 16: Diagnóstico de Residuos del Modelo SARIMAX (Elaboración propia en R).

Los gráficos muestran que el modelo SARIMAX ha capturado adecuadamente la estructura de los datos. Los residuos parecen ser aproximadamente normales y no muestran autocorrelación significativa, aunque hay algunas pequeñas desviaciones que podrían considerarse, especialmente en los extremos. Esto refuerza la idea de que el modelo seleccionado es una buena opción para predecir las unidades vendidas, con un ajuste sólido a los datos.

- Luego de lo anteriormente mencionado corremos el modelo SARIMAX.
- El MAPE obtenido para el modelo SARIMAX, que es del 53.55%, Lo que quiere decir que el modelo presenta un margen de error elevado en sus predicciones comparado con los valores reales. Este alto MAPE indica que el modelo no está capturando de manera adecuada la serie temporal.

6.3.3. Gradient Boosting

Gradient Boosting es un enfoque de aprendizaje automático que combina varios modelos débiles para crear un modelo predictivo robusto. En este proyecto, se empleó XGBoost, reconocido por su eficiencia y precisión en grandes conjuntos de datos. El modelo **Gradient Boosting** es un método de aprendizaje supervisado no paramétrico basado en árboles de decisión, que se utiliza para tareas tanto de clasificación como de regresión. A través de un proceso iterativo, cada modelo se entrena para corregir los errores del modelo anterior, mejorando el rendimiento general. El modelo se entrena de manera secuencial: cada iteración ajusta un modelo a los residuos del anterior, buscando minimizar los errores acumulados [53].

Este método ha ganado relevancia y popularidad por su capacidad de adaptación a múltiples tipos de problemas y su rendimiento en términos de precisión. No obstante, su éxito también depende de un cuidadoso ajuste de sus hiper parámetros y de su implementación para evitar problemas como el sobreajuste. El flujograma presentado en la figura 17, describe un proceso iterativo que incluye la preparación de los datos, la optimización de hiperparámetros mediante validación cruzada y la evaluación de las métricas clave. Este enfoque se caracterizó por su capacidad para modelar relaciones no lineales y ofrecer predicciones altamente precisas, superando las limitaciones de los modelos estadísticos.

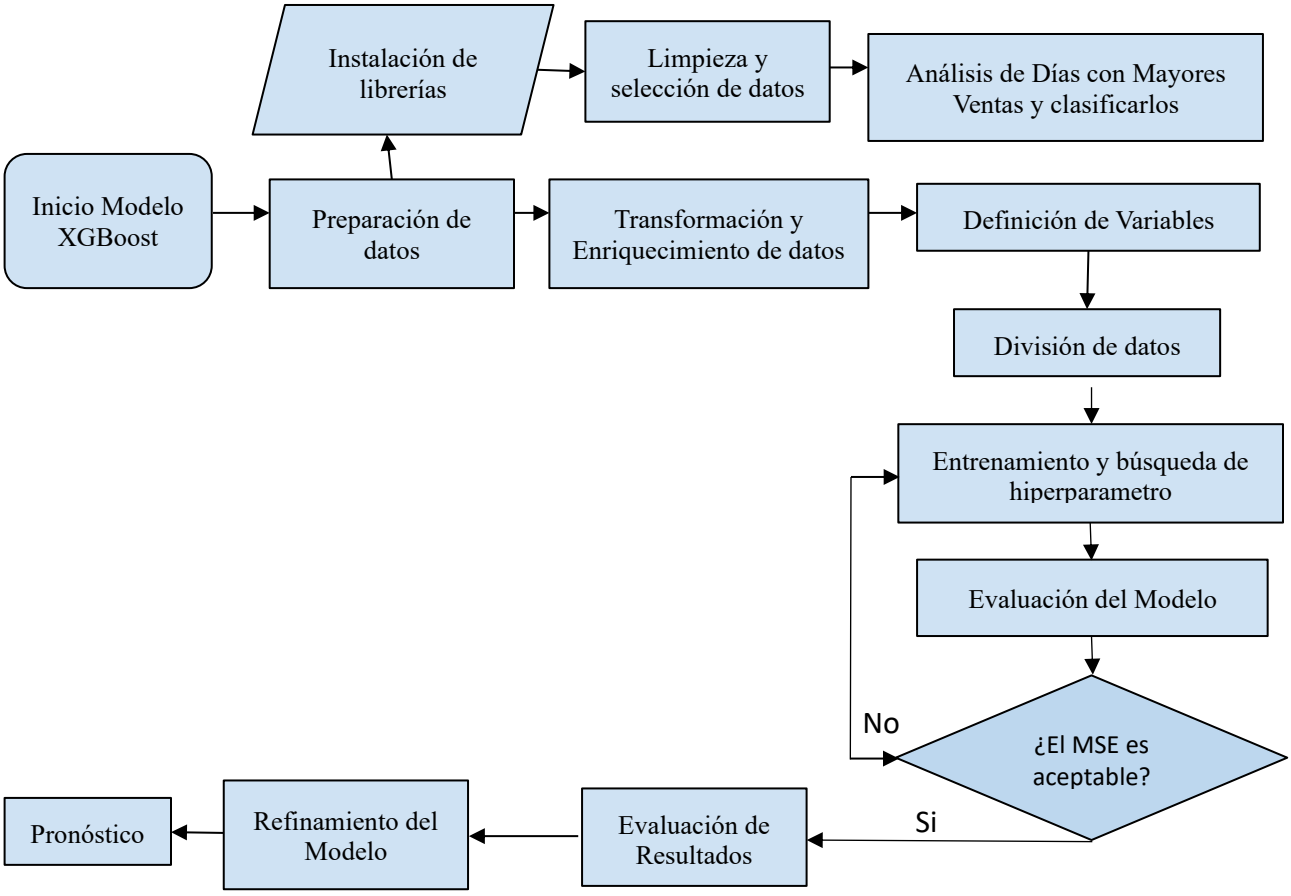

Figura 17: Diagrama de flujo modelo Gradient Boosting (Elaboración propia).

6.3.3.1 Estimación y Pronóstico con un Modelo Gradient Boosting

Para la estimación del modelo **Gradient Boosting**, es crucial tener en cuenta que, a diferencia de los modelos **ARMA** y **SARIMAX**, este utiliza un proceso iterativo. En cada iteración, el modelo se entrena para corregir los errores del modelo anterior, lo que permite mejorar continuamente el rendimiento global.

Siguiendo el enfoque utilizado en el modelo SARIMAX, analizamos los días con mayores ventas para obtener una visión clara de los patrones semanales de demanda. Agrupamos los datos por día de la semana y sumamos las unidades vendidas para cada día. Este análisis nos permite identificar los días más favorables para la venta de cada producto. Luego, calculamos la media diaria de ventas y clasificamos cada día como de 'alta' o 'baja' venta, dependiendo de si sus ventas están por encima o por debajo de dicha media. Si las ventas de un día superan la media, se marca en el data frame con un 1; de lo contrario, con un 0. Estos valores se incorporan posteriormente al modelo, lo descrito anteriormente se puede observar en la figura 13, ubicada en la sección de SARIMAX.

Aunque utilizamos el mismo enfoque del modelo SARIMAX inicial tanto para el análisis de los días con mayor venta como para la identificación de patrones semanales de la demanda, el modelo Gradient Boosting requiere un análisis con mayor detallade y profundización. Esto se debe a que, mientras que SARIMAX se basa en la identificación de estacionalidades y tendencias mediante modelos autorregresivos, Gradient Boosting utiliza una metodología diferente, que combina varios modelos débiles para mejorar la predicción a través de un enfoque secuencial.

Inicialmente, utilizamos la variable 'Mejor día de ventas de la semana' y la fecha en el formato '2024-06-01' (Año-Mes-Día) para realizar el análisis. Con este enfoque, logramos obtener alrededor de un 65.5% de explicación en la variable objetivo, es decir, las unidades vendidas. Sin embargo, al profundizar en el comportamiento de los datos, nos dimos cuenta de que usar la fecha en un formato único limitaba el poder predictivo del modelo. Esto es porque el modelo no podía capturar adecuadamente las interacciones entre los componentes temporales (año, mes y día) lo que podría estar influyendo en las ventas. Para mejorar esta limitación, decidimos descomponer la fecha en sus componentes individuales: año, mes y día, creando una columna separada para cada uno. Esto permitió que el modelo identificara patrones con mayor precisión en las ventas relacionadas con las variaciones temporales. Al probar este nuevo enfoque, observamos una mejora significativa en la capacidad predictiva del modelo, incrementando la explicación de la variable 'unidades vendidas' a un 71.2%.

Este avance no solo mejoró la precisión del modelo, sino que también nos proporcionó un entendimiento más profundo sobre cómo las diferentes dimensiones temporales influyen en la demanda. La descomposición de la fecha nos permitió identificar patrones que antes pasaban desapercibidos. Además, descubrimos una nueva característica clave: los días con mayor o mejor venta mensual en el historial del producto. Usando la variable 'Día', logramos analizar el comportamiento de las ventas de una manera más detallada. Utilizando esta variable, logramos agrupar los datos por día del mes y sumar las unidades vendidas correspondientes a cada grupo, lo que nos proporcionó una visión más clara de cómo las ventas fluctuaban a lo largo del tiempo.

Este análisis reveló patrones importantes en los días específicos del mes en que las ventas eran más altas, brindándonos una nueva perspectiva para refinar nuestras estrategias de predicción y ajuste del modelo. La siguiente tabla simula los resultados de esta agrupación, para darnos una idea de cómo las ventas variaron según el día del mes, y nos ofreció un insumo crucial para continuar mejorando el desempeño del modelo. Recordemos que los días del mes van del 1 al 31, la siguiente tabla 4 es solo una muestra por lo tanto solo mostrará de 1 al 5.

Tabla 4

Agrupación del data frame por días (Elaboración propia).

Dia	Unidades
1	275.552
2	274.423
3	273.153
4	272.503
5	270.173

A partir de este análisis, calculamos el promedio de ventas para cada día del mes sumando la cantidad total de ventas de todos los días y dividiendo este valor entre el número de días. Este cálculo nos permitió obtener un promedio histórico de ventas para cada día del mes. Posteriormente, incorporamos este promedio al sub-dataframe de ventas diarias, añadiendo una nueva columna que indica si las ventas diarias superan o no dicho promedio. Esta columna asigna un valor de 0 si las ventas del día no superan el promedio histórico, y un valor de 1 si las ventas están por encima. Así, creamos una métrica clara para identificar fácilmente los días con ventas superiores al promedio. A continuación, en la tabla 5 podremos observar una simulación de nuestro sub-dataframe:

Tabla 5

Sub Dataframe Promedio Ventas Diarias (Elaboración propia).

Dia	Unidades	PromedioVentasDiaria	Días de ventas
1	275.552	273.102	1
2	274.423	273.102	1
3	273.153	273.102	1
4	272.503	273.102	0
5	270.173	273.102	0

Con nuestro **dataframe final**, realizamos una unión tipo **left join** con un sub-dataframe utilizando la variable "Día" como clave común. De este modo, añadimos únicamente la columna **Días de ventas**, previamente calculada, al data frame principal. Ahora incluía la información del día con mejores ventas del mes, siendo así introducido en nuestro modelo predictivo. Además, los resultados mostraron una mejora notable: el coeficiente de determinación (R^2) aumentó de un 71.2% a un **79.8%** en la explicación de la variable **unidades vendidas**, simplemente al incluir esta

nueva característica.

Este nuevo análisis nos llevó a explorar la creación de más características derivadas de la variable fecha, con el objetivo de mejorar aún más el modelo. Uno de los atributos adicionales fue **NumeroSemanaMes**, que representa el número de semanas dentro del mes. Implementamos un proceso similar al anterior: primero, creamos un sub-dataframe que calculaba el número de la semana y sumaba las ventas históricas de cada semana. Y, por último, calculamos el promedio de ventas por semana a lo largo del historial. A continuación, añadimos una columna binaria que indica si las ventas semanales superan o no el promedio histórico para esa semana. Si las unidades vendidas eran superiores al promedio, asignamos un valor de **1**; en caso contrario, un **0**. Este procedimiento puede verse claramente en la tabla 6:

Tabla 6

Sub Dataframe Mejor semana de ventas (Elaboración propia).

Numero Semana	Unidades	Promedio Unidades semana	Mejor semana de ventas
1	504.120	496.500	1
2	495.10	496.500	0
3	480.370	496.500	0
4	506.203	496.500	1

Siguiendo el mismo enfoque que utilizamos para identificar los días con mejores ventas, realizamos una unión entre este sub-dataframe y el data frame final. Como resultado, obtuvimos dos nuevas variables:

1. **SemanaDelMes**: numera las semanas del mes de 1 a 4.

2. **MejorSemanaDelMes**: identifica la semana con mejores ventas históricas para cada producto. A esta variable se le asigna un valor de **1** si corresponde a la mejor semana, y **0** en caso contrario.

De esta forma, el dataframe final que se introdujo en el modelo quedó estructurado con estas nuevas características adicionales, aportando más información histórica y temporal que mejoró el rendimiento predictivo del modelo. En la siguiente tabla 7 se muestra como quedo el Sub Dataframe.

Tabla 7

Sub Dataframe SemanaDelMes y MejorSemanaDelMes (Elaboración propia).

Año	Mes	Dia	Días Mejor Ventas	Nu.Dia Semana	Mejor Dia Semana	Numero Semana del mes	Mejor Semana del mes
2024	6	1	1	5	1	1	1
2024	6	2	1	6	1	1	1
2024	6	3	1	0	0	1	1

Este enfoque arrojó resultados positivos. Inicialmente, nuestro modelo explicaba el **79.8%** de la variable objetivo utilizando solo seis variables. Sin embargo, al incorporar dos nuevas características clave: **NúmeroSemanaDelMes** y **MejorSemanaDelMes**, la capacidad explicativa del modelo se incrementó significativamente, alcanzando un **84.6%**. Estas mejoras fueron fundamentales para aumentar la precisión en la predicción de las unidades vendidas. Motivados por estos resultados, decidimos agregar dos variables adicionales: **Número de semana del año**, que varía de la semana 1 a la 52, y **Número de día del año**, que indica el día específico del año, del 1 al 365. Con estas inclusiones, el dataframe final quedó estructurado de la siguiente manera, como se ve en la tabla 8:

Tabla 8

Data frame final (Elaboración propia).

Año	Mes	Dia	Días Mejor Ventas	Nu.Dia Semana	Mejor Dia Semana	Num. Semana del mes	Mejor Semana del mes	Num. Semana del año	Num. de día del año
2024	6	1	1	5	1	1	1	22	152
2024	6	2	1	6	1	1	1	22	153
2024	6	3	1	0	0	1	1	22	154

Aunque este último enfoque no generó un impacto tan grande como esperábamos, aún logramos mejorar la precisión del modelo en un **2,65%**, incrementando su capacidad explicativa del **84.6%** al **87.25%**. Este resultado fue significativo, ya que superamos nuestra meta inicial de alcanzar al menos un **85%** de precisión, lo cual considerábamos un hito importante.

6.3.3.2 Definición de la variable objetivo (y) y las características (X):

X: Aquí seleccionas un subconjunto de columnas de df_filtrado que serán utilizadas como características o variables predictoras para entrenar el modelo. Las columnas elegidas son:'mes': El mes del año, 'día': El día del mes, 'año': El año, 'weekday': El día de la semana (0=lunes,

6=domingo), 'DiasVentas': Días de mejores ventas mensuales, 'DiasVentasSemana': Días de mejores ventas semanales. Estas columnas (X) son los inputs o variables independientes que el modelo utilizará para hacer predicciones.

y: Aquí defines la variable objetivo o dependiente que quieres predecir. En este caso, es la columna 'unidades' del DataFrame df_filtrado, que representa las unidades vendidas. Esta columna contiene los valores que el modelo intentará predecir basándose en las características (X).

- **División del conjunto de datos:**

Se lleva a cabo la **separación del conjunto de datos** en dos partes: un conjunto de **entrenamiento** y un conjunto de **prueba.** Esta separación es esencial para garantizar que el modelo sea capaz de generalizar a nuevos datos y no se limite a memorizar los patrones presentes en el conjunto de entrenamiento. Para realizar esta división, utilizamos la función train_test_split, que es parte de la biblioteca sklearn.model_selection en Python. Esta función permite dividir el conjunto de datos de manera aleatoria y eficiente, garantizando que ambos subconjuntos mantengan la misma distribución de las variables y que no haya sesgos en la selección de los datos. En este caso, se establece un **90%** del conjunto de datos para el entrenamiento y un **10%** para la prueba.

6.3.3.3 Entrenamiento del modelo y búsqueda de los mejores hiperparametros:

En este paso, se inicializa un modelo XGBRegressor de la biblioteca XGBoost, diseñado específicamente para abordar problemas de regresión. Este modelo se configura con el objetivo de minimizar el error cuadrático medio (ECM) durante su entrenamiento, utilizando los datos de entrenamiento previamente preparados. XGBoost, o Extreme Gradient Boosting, es un algoritmo de machine learning basado en el principio del boosting de árboles de decisión. A diferencia de métodos más simples, como la regresión lineal, XGBoost crea un modelo predictivo robusto al combinar múltiples árboles de decisión, donde cada nuevo árbol corrige los errores cometidos por los árboles anteriores. Este enfoque de ensamble no solo mejora la precisión de las predicciones, sino que también permite manejar datasets de gran tamaño y complejidad, lo cual es crucial en aplicaciones del mundo real que pueden involucrar millones de registros y múltiples características. Aquí es donde el modelo se entrena o ajusta a los datos. La función fit() toma como entrada los datos de entrenamiento (X_train) y los valores objetivo correspondientes (y_train). Durante este proceso, el modelo aprende las relaciones entre las características (X_train) y la variable objetivo (y_train) optimizando el error cuadrático (en este caso, minimizando el MSE).

En el proceso de optimización de hiperparámetros para un modelo de regresión, se emplea la librería Optuna junto con XGBoost. Primero, se divide el conjunto de datos en entrenamiento y prueba utilizando train_test_split, con un 10% de los datos reservados para las pruebas.

Luego, se define una función objetivo que describe el proceso de optimización. En esta función, se especifican los hiperparámetros que se desean optimizar, tales como el número de estimadores (n_estimators), la profundidad máxima de los árboles (max_depth), la tasa de aprendizaje (learning_rate), entre otros. Optuna sugiere valores para estos hiperparámetros dentro de rangos predefinidos, con el fin de mejorar el rendimiento del modelo.

Para cada combinación de hiperparámetros propuesta, se crea un modelo de XGBRegressor con los valores actuales. Posteriormente, el modelo se evalúa utilizando validación cruzada (con 3 particiones), y se calcula el error cuadrático medio negativo como métrica de evaluación, lo que permite valorar la calidad del modelo en función de su capacidad predictiva.

Finalmente, se crea un estudio de Optuna y se inicia el proceso de optimización, donde se realizan múltiples iteraciones (100 en este caso) para encontrar la combinación óptima de hiperparámetros que minimice el error del modelo. La optimización se realiza con el objetivo de maximizar la puntuación obtenida en la validación cruzada.

6.3.3.4 Evaluación del Modelo:

En esta etapa, se aplica la métrica mean_squared_error (MSE) para evaluar el rendimiento del modelo, comparando las predicciones realizadas por el modelo con los valores reales del conjunto de prueba. Esta métrica permite cuantificar la precisión del modelo al determinar cuán cercanas están las predicciones a los resultados observados. La evaluación de un modelo es un paso crítico en el proceso de desarrollo de machine learning. Utilizar el mean_squared_error como métrica proporciona una forma rigurosa de medir el rendimiento del modelo en un contexto de regresión. El MSE calcula la media de los cuadrados de las diferencias entre las predicciones del modelo y los valores reales, lo que significa que errores más grandes tienen un impacto desproporcionado en el resultado. Esto es especialmente útil en escenarios donde grandes desviaciones son inaceptables, como en la predicción de ventas, donde una estimación errónea puede conllevar a pérdidas significativas.

Este capítulo es el corazón del proyecto, ya que define cuál modelo es el más efectivo para abordar el problema de Megatiendas. La selección del modelo Gradient Boosting da cumplimiento a nuestro **(OE3)**, basado en su superior desempeño, es un paso fundamental hacia la optimización de la cadena de suministro. Su implementación permitirá a Megatiendas mejorar la precisión de sus pronósticos de demanda, con el fin de reducir el desperdicio de productos y optimizar la gestión de inventarios.

7. ANÁLISIS DE RESULTADOS

7.1. Métricas

Las métricas permiten simplificar y centralizar el análisis de datos complejos en indicadores que tienen un significado claro para la gestión de la calidad de datos. Al reducir la cantidad de resultados detallados a unas pocas métricas claves, se mejora la comprensión y la toma de decisiones relacionadas con la calidad y coherencia de los datos, optimizando el monitoreo y el control en procesos de análisis de información [43]. Su función principal es consolidar múltiples mediciones detalladas en una única métrica significativa, facilitando la evaluación global de la calidad de los datos.

Evaluación de las métricas

Para la evaluación de las métricas, se incorporó un módulo en R que permite medir la precisión de los resultados del modelo a partir de los datos reales. En este módulo se considerarán métricas y medidas de desempeño utilizadas comúnmente en el estado del arte.

Configuración Inicial

Se configura knitr para ejecutar el código y se definen funciones para verificar la instalación de librerías necesarias, lo que garantiza que los paquetes necesarios se descarguen y se carguen automáticamente.

Carga de Librerías

Se cargan varias librerías como ggplot2, readxl, shiny, lubridate, plotly, y otras específicas para análisis de series temporales y clustering, como dtwclust. También se utiliza googledrive para acceder a archivos almacenados en Google Drive, y se autentica el acceso con drive_auth().

Resumen del Proyecto

El proyecto se enfoca en analizar los resultados de los pronósticos generados por los modelos ARMA, SARIMA y Gradient Boosting, utilizando métricas clave (MSE, RMSE, MAE, MAPE, R^2) para evaluar su desempeño. Los resultados se presentan en un tablero interactivo que ofrece visualizaciones detalladas por región, tienda y producto, facilitando el análisis granular del rendimiento de los modelos.

Extracción de Datos

Se descargan los archivos de pronósticos desde Google Drive para los modelos ARMA, SARIMA y XGBOOST. Los datos se combinan en un solo dataframe llamado Resultados, que contendrá los resultados de todos los modelos para su posterior análisis.

Transformación de Datos

Se realizan algunas transformaciones en el dataframe combinado, como la conversión de fechas y la corrección de errores específicos en los datos del modelo ARMA.

Dashboard Interactivo con Shiny

El dashboard permite al usuario seleccionar una región, tienda, modelo y producto para analizar los pronósticos y las unidades vendidas. La visualización principal es un gráfico interactivo de líneas generado con plotly que muestra la comparación entre unidades reales y pronosticadas, Como se muestra en el Anexo 2. Además, se muestran métricas como MAE, MSE, RMSE, MAPE y R^2 en tarjetas para facilitar la interpretación del rendimiento del modelo.

Se incluyen tablas resumen que desglosan los resultados por región, tienda y modelo, lo que permite un análisis detallado de los pronósticos.

Funcionalidad Dinámica

El código incluye observadores que actualizan dinámicamente las opciones de tiendas, modelos y productos disponibles, según las selecciones del usuario, garantizando que las visualizaciones sean relevantes para los datos filtrados.

Métricas de Evaluación

Se calculan y muestran las métricas de evaluación del modelo (MAE, MSE, RMSE, MAPE, R^2), lo que permite medir el desempeño de los modelos de pronóstico implementados. Este módulo es un sistema completo para analizar y visualizar los resultados de modelos de pronóstico en la optimización de una cadena de suministro, con una interfaz interactiva y funcionalidades de análisis granular.

7.1.1. Tabla de Promedio de Unidades Vendidas por Tienda y Producto

La siguiente tabla presenta el promedio de unidades vendidas por tienda y producto en el período analizado. Este análisis preliminar no solo ofrece una visión general del comportamiento de la demanda, sino que también sirve como punto de referencia clave para la evaluación de las métricas de los modelos predictivos presentados posteriormente.

Al comparar las métricas como el MAE, RMSE y MAPE con estos promedios, es posible contextualizar los errores de predicción en términos prácticos, permitiendo interpretar si las desviaciones son significativas o aceptables dentro del rango de operación. Este enfoque ayuda a garantizar que los resultados de los modelos sean relevantes y accionables para la toma de decisiones en Megatiendas.

Tabla 9

Promedio de Unidades Vendidas Diarias por Tienda y Producto (Elaboración propia).

Tienda	Producto	Promedio de unidades
207	55057	528,6
	3940	334,3
	54862	137,0
	6665	97,4
	4055	34,3
	3656	34,0
101	3940	461,1
	4002	297,9
	17647	103,6
	4048	101,3
	4024	55,2
	44548	3,9
201	4002	403,5
	3940	224,5
	50682	104,9
	3946	101,1
	3884	60,4
	3844	58,8
102	3940	274,2
	55057	141,8
	17647	25,5
	43310	22,3
	3654	11,1
	27295	10,2
Total, general		**151,1**

Consideraciones a tener en cuenta.

- Los siguientes análisis de gráficas de líneas se centrarán exclusivamente en la tienda 101 y el producto 3940. Los resultados de porcentaje de cumplimiento estarán enfocados en la tienda 101 y los seis productos evaluados. Por último, las métricas presentadas serán generales por modelo y tienda

- Antes de iniciar, para tener una mejor interpretación y análisis de las métricas, tomaremos como dato de referencia el promedio diario de unidades vendidas en el último año mostradas en la tabla 9, este nos ayudará a tener un nivel de comparación más sólido en la interpretación de resultados del MAE y MSE.

- Por último, al realizar las pruebas y análisis, se detectaron algunos datos atípicos en el mes de junio para ciertas tiendas y productos, como la tienda 102 con el producto 27295 y la tienda 207 con el producto 5507. Según investigaciones realizadas con la empresa, estos datos atípicos se debieron a una ruptura de inventario ocasionada por un desabastecimiento. Como se puede observar en las figuras 15 y 16, los resultados para estas variables específicas se verán afectados por lo mencionado anteriormente.

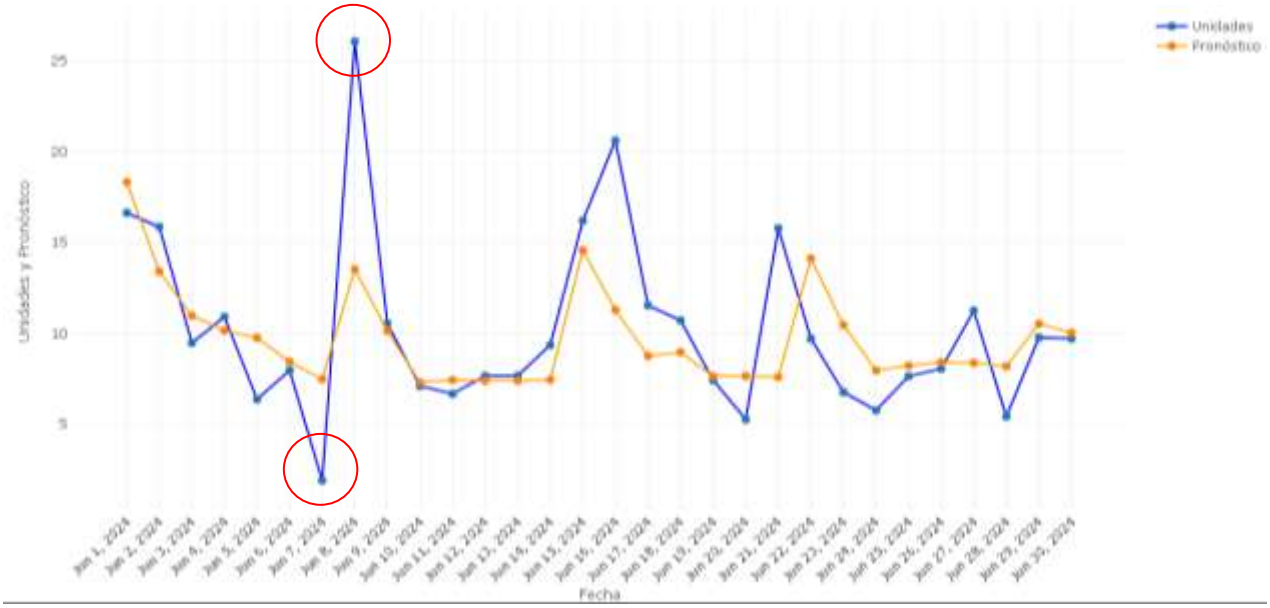

Figura 18: Ruptura de inventario tienda 102 producto 272295 modelo Gradient Boosting (Elaboración propia modulo en R).

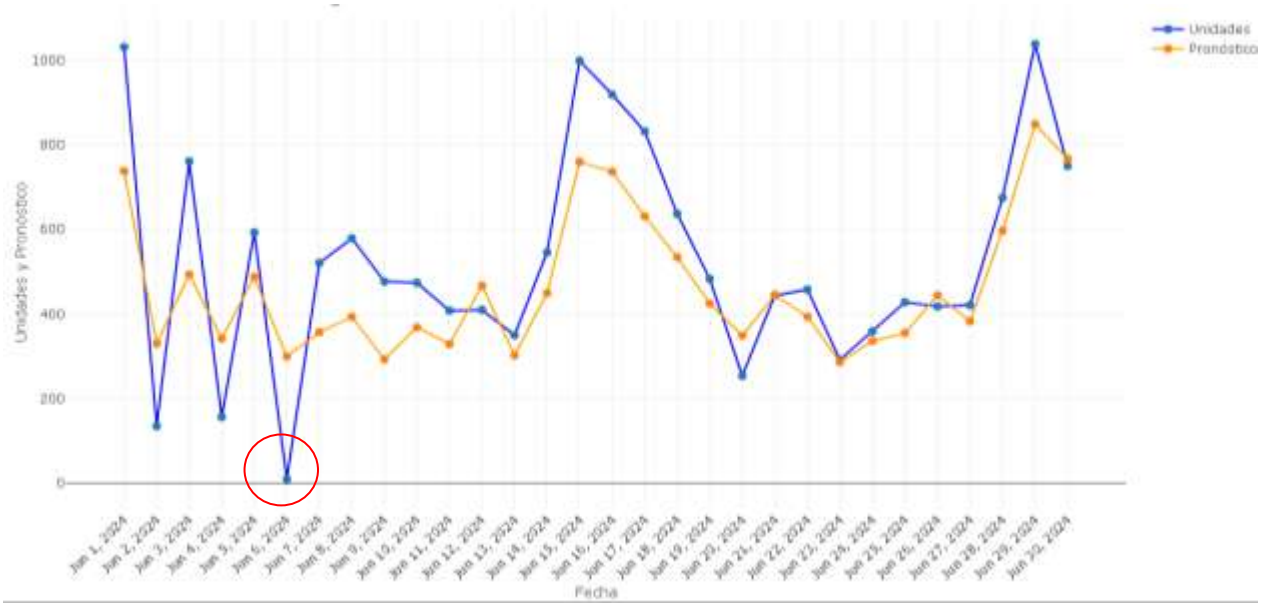

Figura 19: Ruptura de inventario tienda 207 producto 5507 modelo Gradient Boosting (Elaboración propia modulo en R):

7.2 Modelo ARMA

Como se mencionó en la sección 6.2, para el desarrollo de los pronósticos se seleccionó una tienda representativa por cada clúster. En este caso, las tiendas elegidas para las pruebas fueron: tienda 101 y tienda 102, pertenecientes al departamento de Bolívar, y tienda 201 y tienda 207, que representan al departamento del Atlántico. Para cada tienda, se seleccionaron los dos productos más vendidos, dos productos con ventas intermedias y dos productos con las ventas más bajas. Además, la proyección se evaluó para el mes de junio del 2024.

La figura 15 muestra una comparación entre las Unidades reales y el Pronóstico para la Tienda 101 en la Regional Bolívar, correspondiente al Producto 3940 durante el mes de junio de 2024. El modelo ARMA presenta un ajuste moderado para este producto en esta tienda, siendo capaz de predecir la tendencia general, aunque con algunos errores en la predicción de ciertos cambios bruscos de tendencia. Podemos observar que la Tendencia General en ambas líneas sigue un comportamiento cíclico, con picos y valles que coinciden en varias ocasiones. Esto indica que el modelo ARMA captura de manera razonable la estacionalidad y las fluctuaciones en las unidades vendidas.

Tienda 101

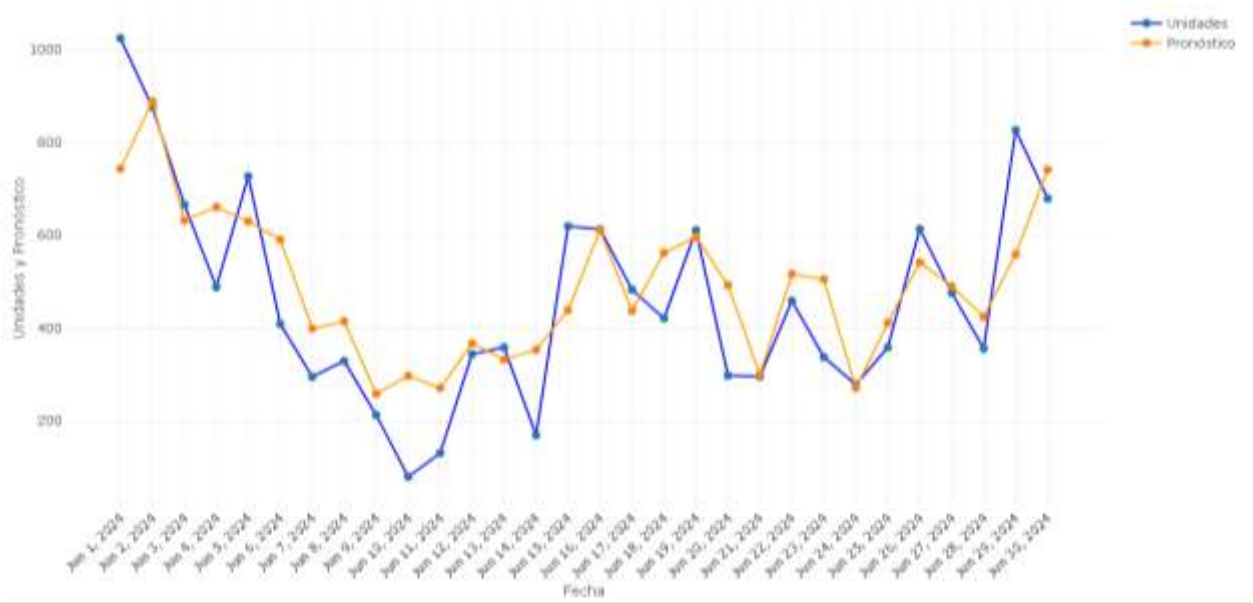

Figura 20: Unidades vendidas vs proyección arma mes de junio tienda 101 producto 3940 modelo ARMA (Elaboración propia modulo en R).

En varios puntos (por ejemplo, a mediados de junio y hacia el final del mes), las predicciones siguen bastante bien las variaciones reales de las unidades vendidas. Pero hay momentos donde la línea naranja del pronóstico se desvía notablemente de la azul (por ejemplo, al principio del mes), lo que muestra que el modelo ARMA no siempre capta de forma precisa las caídas abruptas o los incrementos repentinos en las ventas.

Tabla 10

Porcentaje de cumplimiento proyección ARMA por producto, tienda 101 (Elaboración propia).

FECHA	3940	4002	4024	4048	17647	44548	Total general
01/06/2024	138%	123%	159%	159%	171%	189%	142%
02/06/2024	99%	215%	121%	95%	90%	108%	109%
03/06/2024	105%	99%	180%	158%	116%	132%	113%
04/06/2024	74%	254%	66%	57%	46%	66%	100%
05/06/2024	115%	6594%	150%	125%	168%	119%	175%
06/06/2024	69%	126%	97%	77%	47%	89%	79%
07/06/2024	74%	257%	116%	122%	95%	87%	114%

08/06/2024	79%	147%	81%	45%	90%	132%	85%
09/06/2024	82%	414%	49%	73%	34%	44%	88%
10/06/2024	27%	54%	97%	148%	160%	79%	60%
11/06/2024	48%	42%	93%	71%	86%	83%	55%
12/06/2024	94%	59%	145%	129%	86%	76%	81%
13/06/2024	108%	47%	52%	29%	71%	65%	66%
14/06/2024	48%	66%	104%	214%	103%	113%	70%
15/06/2024	141%	71%	181%	168%	158%	141%	116%
16/06/2024	100%	53%	97%	99%	133%	138%	90%
17/06/2024	110%	60%	186%	239%	119%	63%	94%
18/06/2024	75%	138%	86%	60%	59%	100%	95%
19/06/2024	102%	184%	103%	161%	132%	102%	125%
20/06/2024	60%	96%	65%	54%	65%	75%	66%
21/06/2024	99%	58%	77%	124%	83%	87%	81%
22/06/2024	89%	27%	32%	66%	73%	64%	60%
23/06/2024	67%	14%	34%	55%	78%	85%	42%
24/06/2024	103%	19%	63%	124%	94%	58%	46%
25/06/2024	87%	26%	38%	126%	163%	69%	53%
26/06/2024	113%	33%	208%	171%	109%	154%	75%
27/06/2024	97%	40%	60%	91%	96%	56%	62%
28/06/2024	84%	31%	69%	91%	56%	115%	52%
29/06/2024	148%	36%	132%	137%	226%	250%	92%
30/06/2024	92%	24%	101%	88%	101%	74%	62%
Total general	96%	45%	95%	110%	102%	97%	74%

Podemos observar que, aunque el modelo ARMA parece capturar la tendencia general, las diferencias en ciertos días indican que podría no estar considerando factores que afectan las ventas, como promociones o eventos especiales. Para analizar los resultados de los diferentes modelos ARMA aplicados a las distintas tiendas (centro_op), evaluamos varios indicadores clave: **MAE**, **MSE**, **RMSE**, **MAPE** y **R^2**.

Tienda 101:

- **MAE**: 86.85. Indica que el modelo ARMA tiene un error absoluto promedio de 86.85 unidades, lo que es relativamente bajo comparado con las ventas promedio diarias (13,000).
- **MSE**: 29,375.14. El error cuadrático medio es bastante alto, lo que significa que hay desviaciones significativas en los errores más grandes.
- **RMSE**: 171.39. Este valor indica que, en promedio, el modelo se desvía en 171.39 unidades, lo que es aceptable pero no óptimo.
- **MAPE**: 55.11%. El error porcentual absoluto medio es alto, quiere decir que el modelo tiene dificultades para predecir de manera precisa.
- **R^2**: 0.53. El modelo explica solo el 53% de la variabilidad en los datos, lo cual es un porcentaje de cumplimiento muy bajo para lo que estamos buscando.

Tienda 102:

- **MAE**: 82.30. Un buen valor para el error promedio. Es relativamente bajo comparado con las ventas promedio diarias.
- **MSE**: 41,368.35. Este valor alto indica que los errores son grandes cuando ocurren.
- **RMSE**: 203.39. El error promedio es mayor que en el Centro 101, lo que indica un peor rendimiento.
- **MAPE**: 77.45%. Un valor muy alto, quiere decir que el modelo no es confiable para esta tienda.
- **R^2**: 0.57. El modelo explica el 57% de la variabilidad, algo mejor que en el Centro 101 pero sigue siendo un porcentaje de cumplimiento muy bajo para lo que estamos buscando.

Tienda 201:

- **MAE**: 340.08. Muy alto a comparación del promedio de unidades diarias vendidas, quiere decir que el modelo tiene un error muy alto.
- **MSE**: 778,133.91. Errores muy grandes.
- **RMSE**: 882.12. Un valor extremadamente alto que refleja un mal rendimiento del modelo para esta tienda.
- **MAPE**: 122.59%. El error porcentual es muy alto, lo que hace que el modelo sea inadecuado.
- **R^2**: 0.41. El modelo explica muy poco de la variabilidad, por lo tanto, es muy bajo su porcentaje de cumplimiento.

Tienda 207:

- **MAE**: 139.29. Alto para el error promedio.
- **MSE**: 56,200.70. Bastante elevado.
- **RMSE**: 237.07. Un error grande en términos absolutos.
- **MAPE**: 314.59%. Altísimo, lo que indica que el modelo no es adecuado para esta tienda.
- **R^2**: 0.38. El modelo tiene un muy bajo porcentaje de asertividad.

En general las tiendas 101 y 102 tienen el mejor rendimiento, con un R^2 de más de 50% y valores relativamente bajos de MAE y RMSE, aunque con MAPE relativamente altos, por lo tanto, podemos decir que las predicciones absolutas son más confiables que las relativas. El centro de operación 201 tiene el peor rendimiento, con un error absoluto y porcentual muy elevado, por lo tanto, el modelo es inadecuado para esta tienda. La tienda 207 muestra un rendimiento pobre, especialmente en términos de MAPE y R^2, lo que indica baja precisión y un ajuste deficiente. Por lo tanto, este modelo no nos da resultados satisfactorios para la predicción diaria de las unidades vendidas por tienda.

7.3 Modelo SARIMAX

Tienda 101

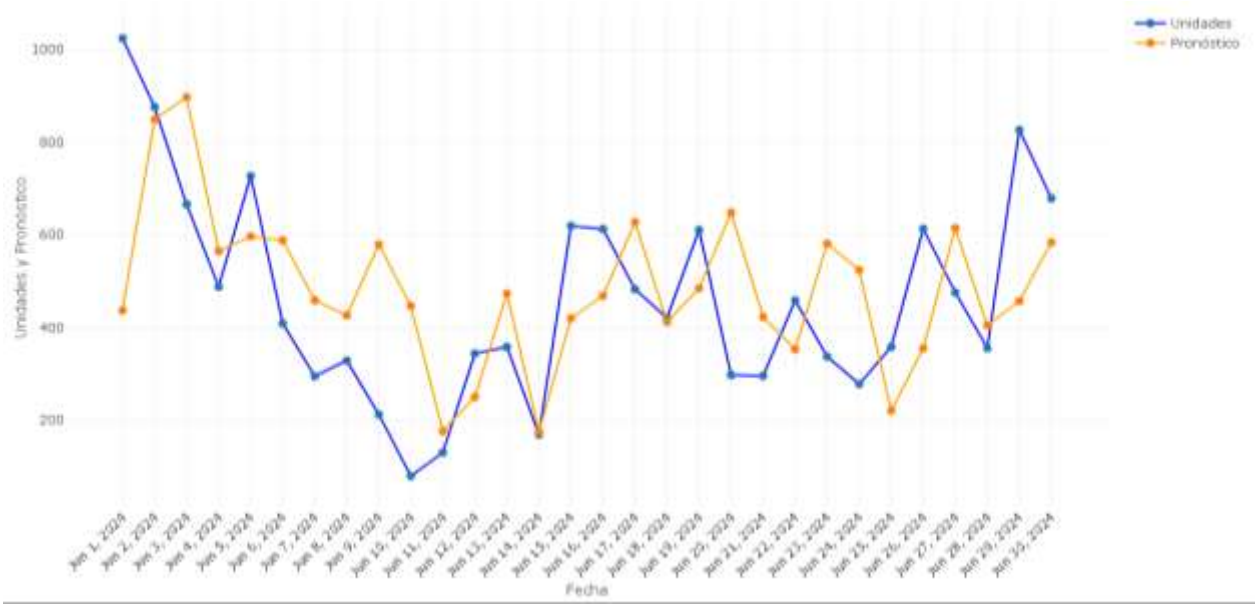

Figura 21: Unidades vendidas vs proyección mes de junio tienda 101 producto 3940 modelo SARIMAX (Elaboración propia modulo en R).

En la figura 18 se observan dos series de datos: las Unidades reales (línea azul) y el Pronóstico generado por el modelo SARIMAX (línea naranja) para el producto 3940 en la tienda 101 de la región Bolívar, en el mes de junio de 2024. La tendencia general muestra que las unidades y pronóstico siguen una tendencia similar, con algunos momentos en que el pronóstico se desvía notablemente de las unidades reales, pero generalmente están alineados. Hay una alta fluctuación en ambos conjuntos de datos, quiere decir que las ventas o la demanda del producto varían significativamente día a día. En la primera parte del gráfico (1 al 5 de junio), el modelo sobreestima las unidades reales, ya que la línea naranja está por encima de la línea azul. A partir del 10 de junio, las predicciones comienzan a ajustarse mejor a los valores reales, aunque hay diferencias ocasionales, especialmente a mitad del mes. En la última semana de junio, se nota un buen ajuste entre el pronóstico y las unidades, pero con algunas excepciones donde el modelo sigue subestimando o sobreestimando en algunos puntos.

El modelo SARIMAX ofrece un ajuste razonable, aunque hay algunas discrepancias en días específicos, particularmente al inicio del mes. Podemos decir que, aunque el modelo está capturando correctamente la tendencia general, hay espacio para mejorar en términos de ajuste a la variabilidad diaria.

Tabla 11:

Porcentaje de cumplimiento proyección SARIMAX por producto, tienda 101 (Elaboración propia).

FECHA	3940	4002	4024	4048	17647	44548	Total general
01/06/2024	234%	186%	215%	215%	244%	228%	219%
02/06/2024	103%	74%	98%	97%	104%	160%	98%
03/06/2024	74%	115%	153%	119%	107%	106%	92%
04/06/2024	86%	159%	65%	68%	76%	67%	103%
05/06/2024	122%	172%	167%	151%	112%	90%	135%
06/06/2024	70%	53%	90%	66%	52%	102%	65%
07/06/2024	64%	88%	87%	93%	66%	78%	77%
08/06/2024	77%	41%	91%	66%	112%	143%	71%
09/06/2024	37%	41%	35%	50%	26%	43%	39%
10/06/2024	18%	62%	87%	147%	107%	66%	50%
11/06/2024	74%	48%	86%	64%	77%	121%	64%
12/06/2024	137%	104%	197%	167%	94%	132%	125%
13/06/2024	76%	58%	37%	20%	62%	54%	60%
14/06/2024	97%	120%	67%	89%	112%	120%	107%
15/06/2024	147%	126%	263%	261%	164%	216%	156%
16/06/2024	131%	68%	98%	98%	175%	110%	109%
17/06/2024	77%	154%	120%	176%	117%	70%	105%
18/06/2024	102%	215%	100%	68%	107%	83%	1335
19/06/2024	126%	117%	140%	159%	104%	118%	125%
20/06/2024	46%	29%	50%	51%	67%	76%	45%
21/06/2024	70%	64%	86%	86%	60%	78%	70%

22/06/2024	130%	42%	53%	97%	107%	90%	89%
23/06/2024	58%	26%	22%	42%	43%	67%	45%
24/06/2024	53%	76%	73%	114%	104%	89%	68%
25/06/2024	162%	81%	60%	105%	131%	115%	113%
26/06/2024	173%	100%	301%	256%	181%	266%	158%
27/06/2024	77%	98%	62%	71%	74%	69%	81%
28/06/2024	88%	105%	43%	67%	60%	95%	86%
29/06/2024	181%	117%	169%	148%	135%	322%	159%
30/06/2024	116%	69%	104%	90%	143%	96%	102%
Total general	95%	90%	101%	102%	101%	108%	96%

El rendimiento general parece cercano a lo esperado en promedio, con algunas desviaciones notables. Hay días en los que los rendimientos superan el 200% o caen por debajo del 50%, podría ser consecuencia de eventos atípicos o problemas con el ajuste del modelo para ciertos días o productos. Los productos 3940 y 44548 presentan mayor volatilidad en sus rendimientos, mientras que los productos 4024 y 4002 muestran una menor variabilidad, pero con un rendimiento general más bajo. A comparación con el modelo ARMA tiende a tener un mejor asertividad con respecto a los días, pero aun así no llega al porcentaje de confiabilidad que queremos.

Tienda 101:

- **MAE**: 70.28. El error promedio es bastante bajo.
- **MSE**: 13,102.05. Significativamente menor que el de ARMA, lo que indica menos errores grandes.
- **RMSE**: 114.46. El error cuadrático es menor, lo que refleja un mejor ajuste.
- **MAPE**: 50.60%. Aunque sigue siendo alto, es mejor que en ARMA.
- **R^2**: 0.65. Un buen valor, indicando que el modelo explica un 65% de la variabilidad.

Tienda 102:

- **MAE**: 29.54. Un error bastante bajo.
- **MSE**: 3,677.74. Muy bajo comparado con otros modelos, lo que es positivo.
- **RMSE**: 60.64. Mucho menor, por lo tanto, es un buen ajuste.
- **MAPE**: 48.92%. Aún alto, pero mejor que el ARMA.
- **R^2**: 0.71. Un buen nivel explicativo, mejor que en otros modelos.

Tienda 201:

- **MAE**: 68.19. El error promedio es bajo.

- **MSE**: 12,393.90. Relativamente bajo.
- **RMSE**: 111.33. Muestra un buen ajuste.
- **MAPE**: 51.76%. Aunque sigue alto, es mejor que ARMA.
- **R^2**: 0.57. Explica más que el modelo ARMA.

Tienda 207:

- **MAE**: 83.92. Algo elevado.
- **MSE**: 20,883.94. Bastante alto, lo que indica errores grandes.
- **RMSE**: 144.51. Indica un error considerable.
- **MAPE**: 272.58%. Altísimo, lo que refleja un pobre rendimiento.
- **R^2**: 0.60. Algo mejor que ARMA, pero aún no es ideal.

En general, los resultados más favorables se observan en la tienda 102, mientras que la tienda 207 muestra un desempeño notablemente inferior esto podría ser por lo anteriormente mencionado sobre la ruptura de inventario que hubo durante los primeros días del mes. Podemos decir entonces que, aunque el modelo SARIMAX supera al modelo ARMA gracias a la inclusión de variables exógenas, todavía no alcanza el nivel de confiabilidad deseado.

7.4 Modelo Gradient Boosting

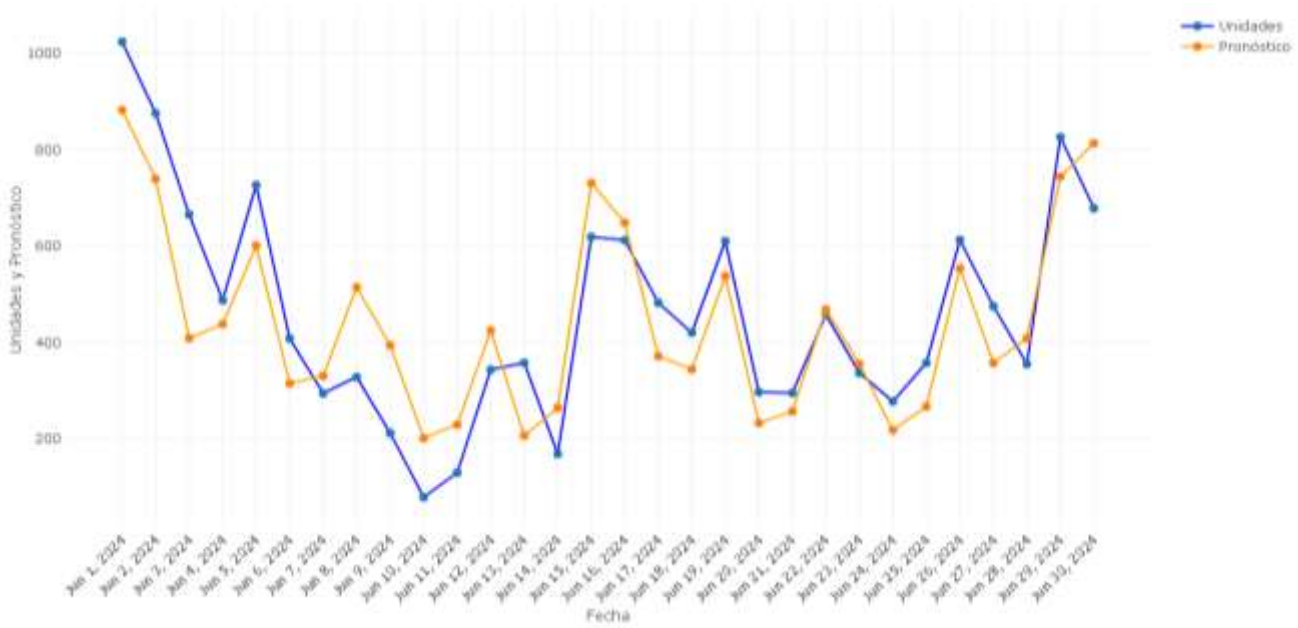

Figura 22: Unidades vendidas vs proyección mes de junio tienda 101 producto 3940 modelo Gradient Boosting (Elaboración propia modulo en R).

Se observa que en muchas partes de la figura 19, las Unidades reales y el Pronóstico siguen patrones similares, aunque con algunas desviaciones, especialmente entre el 1 de junio y el 10 de junio, donde el pronóstico es más alto que las unidades reales. Después del 15 de junio, las líneas tienden a coincidir mucho más, con diferencias menores entre los valores de pronóstico y las unidades reales. El gráfico muestra una disminución general al inicio de junio, seguida de fluctuaciones más

pequeñas a lo largo del mes. Entre el 15 y el 22 de junio, hay un pico claro en ambos conjuntos de datos. En comparación con los modelos ARMA y SARIMAX, este modelo ofrece una mayor precisión en las predicciones diarias.

Tabla 12

Porcentaje de cumplimiento proyección Gradient Boosting por producto, tienda 101 (Elaboración propia).

FECHA	17647	3940	4002	4024	4048	44548	Total general
1/06/2024	156%	116%	137%	88%	262%	101%	128%
2/06/2024	128%	118%	83%	81%	98%	102%	106%
3/06/2024	246%	163%	73%	120%	142%	121%	131%
4/06/2024	103%	111%	236%	56%	77%	99%	137%
5/06/2024	158%	121%	155%	99%	107%	113%	129%
6/06/2024	73%	130%	91%	127%	93%	105%	106%
7/06/2024	93%	89%	139%	105%	127%	110%	109%
8/06/2024	105%	64%	48%	94%	65%	114%	64%
9/06/2024	43%	54%	92%	70%	106%	96%	67%
10/06/2024	130%	39%	57%	81%	114%	98%	66%
11/06/2024	112%	57%	86%	87%	83%	105%	77%
12/06/2024	93%	81%	120%	123%	114%	97%	100%
13/06/2024	88%	173%	129%	72%	37%	95%	126%
14/06/2024	125%	64%	184%	88%	106%	104%	115%
15/06/2024	148%	85%	131%	127%	135%	102%	106%
16/06/2024	164%	94%	106%	110%	127%	108%	105%
17/06/2024	185%	130%	119%	102%	196%	97%	134%
18/06/2024	121%	122%	276%	93%	83%	103%	162%
19/06/2024	146%	113%	105%	96%	143%	95%	115%
20/06/2024	103%	128%	52%	97%	96%	105%	91%

21/06/2024	103%	115%	103%	106%	135%	97%	111%
22/06/2024	90%	98%	54%	51%	81%	92%	81%
23/06/2024	71%	95%	50%	45%	82%	97%	77%
24/06/2024	134%	127%	70%	57%	99%	90%	97%
25/06/2024	198%	134%	124%	57%	120%	95%	130%
26/06/2024	145%	111%	116%	145%	130%	106%	119%
27/06/2024	119%	133%	170%	97%	108%	90%	137%
28/06/2024	75%	87%	131%	57%	71%	101%	94%
29/06/2024	128%	111%	102%	97%	98%	113%	108%
30/06/2024	115%	83%	86%	89%	87%	99%	87%
Total general	124%	104%	113%	94%	110%	101%	108%

Se muestra un desempeño sólido con una precisión promedio del 108%, lo que indica una buena capacidad para ajustarse a los datos.

Tienda 101:

- **MAE**: 42.96. El error promedio es bajo comparado con el promedio de unidades vendidas diarias.
- **MSE**: 5,497.10. Bajo comparado con los otros modelos.
- **RMSE**: 74.14. Buen resultado, con un error mucho menor que en ARMA y SARIMAX.
- **MAPE**: 25.34%. Mucho más bajo que en otros modelos, quiere decir que tiene buena capacidad predictiva.
- **R^2**: 0.86. El modelo explica un 86% de la variabilidad, muy buen resultado.

Tienda 102:

- **MAE**: 20.19. Error muy bajo comparado con el promedio de unidades vendidas diarias. Lo que quiere decir que es un muy buen indicador.
- **MSE**: 1,673.14. Muy bajo, lo que indica buen ajuste.
- **RMSE**: 40.90. Refleja un excelente ajuste.
- **MAPE**: 29.30%. Aceptable y significativamente mejor que otros modelos.
- **R^2**: 0.89. Explica un 89% de la variabilidad, el mejor resultado.

Tienda 201:

- **MAE**: 36.51. Un error bajo comparado con el promedio de unidades vendidas diarias.
- **MSE**: 4,432.50. Bajo, lo que refleja un buen ajuste.

- **RMSE**: 66.58. Muy buen ajuste.
- **MAPE**: 20.79%. Muy bajo, lo que muestra un buen rendimiento.
- **R^2**: 0.86. Explica el 86% de la variabilidad, excelente resultado.

Tienda 207:

- **MAE**: 49.37. El error promedio es bajo comparado con el promedio de unidades vendidas diarias.
- **MSE**: 6577.11. Moderado comparado con las otras tiendas.
- **RMSE**: 81.10. Muestra un buen rendimiento.
- **MAPE**: 79.97%. Alto, pero significativamente mejor que los otros modelos.
- **R^2**: 0.88. Explica el 88% de la variabilidad, excelente resultado

En general, el modelo de Gradient Boosting ha demostrado un rendimiento excelente en cada una de las tiendas, logrando los mejores resultados en todas las métricas: MAE, MSE, RMSE y R^2. Además, el MAPE del modelo de Gradient Boosting es significativamente más bajo que el de los modelos ARMA y SARIMAX, por lo tanto, es más confiable para predecir las ventas diarias. Aunque SARIMAX representa una opción viable con un desempeño razonable, ARMA queda rezagado, mostrando errores más grandes y una menor capacidad explicativa en todos los centros analizados. En resumen, Gradient Boosting es la mejor opción para la predicción de ventas, proporcionando resultados con mejor precisión y confiabilidad en comparación con los otros modelos evaluados. Obteniendo en promedio una asertividad del 87,25% el cual es un excelente indicador dando así cumplimiento a nuestro objetivo principal.

Además, cabe destacar que este modelo identifica tendencias y patrones muy precisos, como podemos observar en las **Figuras 15 y 16**, las rupturas de inventario provocaron variaciones extremas en las ventas, generando picos muy bajos debido al desabastecimiento y picos muy elevados por el sobreabastecimiento posterior. Estas fluctuaciones no solo afectaron la estabilidad de las ventas, sino que también subrayan la importancia de contar con herramientas predictivas como el modelo que aquí se presenta. A pesar de estas variaciones, es notable que el modelo de pronóstico sigue una tendencia bastante precisa en la mayoría de los casos. Aunque no captura con exactitud los valores extremos, como las caídas abruptas por desabastecimiento, sí predice adecuadamente las tendencias generales y ofrece una señal clara de alerta en momentos críticos, como la caída de ventas alrededor del segundo día del mes en la figura 16, señalada en las gráficas.

En un escenario hipotético en el que esta herramienta hubiera estado disponible en tiempo real, los responsables de la gestión de inventario podrían haber anticipado estos problemas y tomado medidas correctivas a tiempo. Por ejemplo, al prever la caída en las ventas proyectada por el modelo, habrían podido ajustar los niveles de inventario y evitar la falta de productos en los puntos de venta, mejorando así la experiencia del cliente y reduciendo el impacto financiero de las rupturas de stock.

Tabla 13

Resumen Modelos por Tiendas (Elaboración propia).

Modelo	centro_op	MAE	MSE	RMSE	MAPE	R2
ARMA	101	86.85	29375.14	171.39	55.11	0.53
ARMA	102	82.30	41368.35	203.39	77.45	0.57
ARMA	201	340.08	778133.91	882.12	122.59	0.41
ARMA	207	139.29	56200.70	237.07	314.59	0.38
SARIMAX	101	70.28	13102.05	114.46	50.60	0.65
SARIMAX	102	29.54	3677.74	60.64	48.92	0.71
SARIMAX	201	68.19	12393.90	111.33	51.76	0.57
SARIMAX	207	83.92	20883.94	144.51	272.58	0.60
XGBOOST	**101**	**42.96**	**5497.10**	**74.14**	**25.34**	**0.86**
XGBOOST	**102**	**20.19**	**1673.14**	**40.90**	**29.30**	**0.89**
XGBOOST	**201**	**36.51**	**4432.50**	**66.58**	**20.79**	**0.86**
XGBOOST	**207**	**49.37**	**6577.11**	**81.10**	**79.97**	**0.88**

El análisis de resultados es fundamental para validar el desempeño de los modelos y justificar la selección de Gradient Boosting como el enfoque más efectivo. Este capítulo confirma que el modelo propuesto no solo supera en rendimiento a los modelos tradicionales, sino que también ofrece un impacto tangible en la reducción de costos y en la eficiencia operativa de Megatiendas. Este análisis establece una base sólida para la implementación del modelo en el contexto real. Dando así cumplimiento a nuestro último objetivo **(OE4)**.

8. CONCLUSIONES Y TRABAJOS FUTUROS

8.1. CONCLUSIONES

Para concluir este proyecto, se ha implementado y evaluado un modelo de pronósticos de demanda con el fin de optimizar la cadena de suministros de la unidad estratégica de negocio FRUVER en Megatiendas. El enfoque ha sido mejorar la gestión de inventarios y reducir el desperdicio de productos perecederos a través de la implementación de diferentes modelos predictivos, incluyendo ARMA, SARIMAX y Gradient Boosting.

El proyecto cumplió con éxito los objetivos planteados, logrando la implementación de un modelo de aprendizaje automático que optimiza la predicción de la demanda en la unidad estratégica de negocio FRUVER de Megatiendas. Los principales logros incluyen:

1. **Estrategia para la extracción de datos**: Se desarrolló una estrategia eficiente para extraer, limpiar y transformar los datos de ventas históricos. Esto incluyó la integración de variables clave como promociones y eventos, permitiendo un análisis más preciso.

Este capítulo describe los métodos y herramientas utilizados para la obtención y manejo de los datos. Incluyendo la solicitud de credenciales y la implementación de bases de datos proporcionadas por Megatiendas. Se explica cómo se limpian y transforman los datos para ser utilizados en el análisis. La correcta extracción y transformación de los datos fue clave para garantizar la calidad y precisión del modelo predictivo. Este proceso sentó las bases para un análisis detallado de los datos históricos, asegurando que se pueda confiar en la información utilizada para la predicción.

2. **Módulo de Extracción de datos**: Se implementó un módulo funcional en R que facilitó el manejo de los datos extraídos, asegurando su disposición continua y actualizada para el análisis y la predicción de la demanda.

En este capítulo se desarrolló un módulo en R que permite la transformación y limpieza de los datos extraídos. También se utilizó el paquete Google Drive para la descarga y almacenamiento continuo de los datos. Y se detalla la creación de un DataFrame limpio y listo para análisis. El módulo de extracción y limpieza de datos resultó ser fundamental para organizar los datos de forma adecuada, eliminando devoluciones y ventas irregulares que podrían distorsionar los pronósticos. Esto permitió obtener un conjunto de datos listo para la modelación.

3. **Desarrollo del modelo Predictivo:** Se evaluaron varios modelos, destacándose el modelo **Gradient Boosting** como el más preciso para la predicción de la demanda, superando a los modelos ARMA y SARIMAX en términos de métricas de error y precisión.

Se realizó la selección de variables relevantes y la clusterización de tiendas utilizando la técnica DTW (Dynamic Time Warping) para identificar similitudes en las series temporales de ventas. A partir de ello, se implementaron tres modelos predictivos: ARMA, SARIMAX y Gradient

Boosting. La implementación de los tres modelos fue esencial para explorar diferentes enfoques predictivos. La clusterización de tiendas permitió optimizar los pronósticos, mientras que la evaluación de los tres modelos brindó una comprensión clara de cuál ofrecía los mejores resultados para Megatiendas.

4. **Análisis de Resultados:** Se incorporó un módulo en R para medir la precisión del modelo predictivo, usando métricas estándar como MAE, MSE, RMSE, MAPE y R^2, lo que permitió seleccionar el modelo más adecuado para las necesidades de la empresa.

Este capítulo presenta los resultados obtenidos de los tres modelos predictivos. Se utilizaron métricas de evaluación como MAE, RMSE, MSE, MAPE y R^2 para comparar los modelos. Gradient Boosting se destacó como el modelo más preciso en la predicción de la demanda. El análisis de resultados mostró que el modelo **Gradient Boosting** es el más adecuado para la predicción de la demanda en Megatiendas. Este modelo ofrece mayor precisión, reduciendo el error en comparación con los modelos ARMA y SARIMAX, lo que lo convierte en la solución ideal para la gestión de inventarios.

8.2. TRABAJOS FUTUROS

En los próximos pasos de este proyecto, es fundamental explorar diversas áreas de mejora para seguir optimizando la precisión y eficiencia del modelo de pronósticos de demanda. Algunas líneas de trabajo futuro incluyen:

1. **Incorporación de más variables externas**: Aunque el modelo actual se basa principalmente en datos históricos de ventas, agregar variables externas como condiciones climáticas, tendencias económicas, festividades y eventos locales podría mejorar aún más la precisión de las predicciones. Estas variables pueden tener un impacto significativo en la demanda de productos perecederos como los FRUVER.
2. **Expansión a otras categorías de productos**: Si bien este proyecto se centró en la unidad de FRUVER, el enfoque y los modelos utilizados pueden adaptarse y extenderse a otras categorías de productos dentro de Megatiendas. Esto permitiría optimizar la cadena de suministro en una mayor variedad de productos, reduciendo aún más los costos y las pérdidas por exceso de inventario.

Estas mejoras no solo fortalecerán la capacidad predictiva de Megatiendas, sino que también la posicionarán como una empresa más ágil y adaptable a los cambios del mercado.

9. REFERENCIAS BIBLIOGRÁFICAS

[1] J. A. Z. Cortes, «Fundamentos de la gestión de inventarios,» *Centro Editorial Esumer*, p. 68, 2014.

[2] Spyros Makridakis, Steven C. Wheelwright, Rob J. Hyndman, Forecasting: Methods and Applications, New York: John Wiley & Sons, 1998.

[3] Babai, M. Z., Ali, M. M., Boylan, J. E., & Syntetos, A. A., «Forecasting and inventory performance in a two-stage supply chain with ARIMA(0,1,1) demand,» *International Journal of Production Economics,* p. 74-83, 2015.

[4] J. D. Belalcázar y A. H. Cárdenas, «Propuesta de un sistema de pronósticos para el supermercado Punto Mercar S.A,» *Universidad ICESI,* p. 53, 2020.

[5] Martínez, M. A., & Aguilar, A, «Optimización de inventarios en empresas minoristas mediante pronósticos basados en aprendizaje automático,» *Revista de Ingeniería Industrial,* 2020.

[6] R. Gwynne, Warehouse Management: A Complete Guide to Improving Efficiency and Minimizing Costs in the Modern Warehouse, Londres: Kogan Page Publishers, 2017.

[7] J. S. Armstrong, Principles of Forecasting: A Handbook for Researchers and Practitioners, 1 ed., Philadelphia: The Wharton School, University of Pennsylvania, Springer, 2001.

[8] R. J. H. y. G. Athanasopoulos, Forecasting: Principles and Practice, 2 ed., Australia: OTexts, 2018.

[9] N. R. Sanders, Supply Chain Management: An Integrated Approach, 6 ed., Boston: Pearson, 2021.

[10] S. N. Chapman, Planificación y control de la producción, México: Pearson Educación, 2006.

[11] J. L. B. B. y. L. M. Á. Posada, «Caracterización de la Gestión de pronósticos de demanda empresarial,» *Universidad del Rosario,* 2013.

[12] R. J. Hyndman, «Cyclic and seasonal time series,» *robjhyndman.com,* 14 diciembre 2011.

[13] V. L. a. Y. R. Gel, «Time Series Analysis: Lecture Notes with Examples in R,» University of Maryland, Septiembre 2023. [En línea]. Available: https://vlyubchich.github.io/tsar/. [Último acceso: 5 Junio 2024].

[14] A. Borucka, «Seasonal Methods of Demand Forecasting in the Supply Chain as Support for the Company's Sustainable Growth,» *Sustainability,* p. 21, 2023.

[15] A. M. Marcos Rubio, «Estimación de modelos ARMA usando el modelo espacio de los estados,» *E.T.S.I. Industriales (UPM),* p. 54, Junio 2023.

[16] M. Foley, «Time Series Analysis,» bookdown.org, 6 Noviembre 2021. [En línea]. Available: https://bookdown.org/mpfoley1973/time-series/. [Último acceso: 8 6 2024].

[17] M. Peixeiro, Time Series Forecasting in Python, New York: Manning Publications, 2022.

[18] J. A. Rodrigo, «cienciadedatos.net,» Octubre 2020 . [En línea]. Available: https://cienciadedatos.net/documentos/py09_gradient_boosting_python. [Último acceso: 15 6 2024].

[19] G. L. Y. S. a. X. L. W. Wang, «Time Series Clustering Based on Dynamic Time Warping,» *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS),* pp. 487-490, 2018.

[20] R. M. a. R. Angryk, «Distance and Density Clustering for Time Series Data,» *2017 IEEE International Conference on Data Mining Workshops (ICDMW),* pp. 25-32, 2017.

[21] G. H. N. S. G. L. a. C. -H. C. B. Cai, «Efficient Time Series Clustering by Minimizing Dynamic Time Warping Utilization,» *IEEE Access,* vol. 9, p. 11, 2021.

[22] R. Mussabayev, «Optimizing Euclidean Distance Computation,» *Mathematics,* p. 36, 2024.

[23] D. J. C. M. B. C. J. C. B. Donald J. Bowersox, Supply Chain Logistics Management, 5 ed., New York: McGraw-Hill Education, 2020, pp. 145-157.

[24] F. X. Diebold, Elements of Forecasting, New York: South-Western, 2007.

[25] K. S. P. D. R. S. D. R. K. P. Udbhav Vikas, «A Comprehensive Study on Demand Forecasting Methods and Algorithms for Retail Industries,» *Journal of University of Shanghai for Science and Technology,* vol. 23, p. 420, 2021.

[26] E. S. V. A. Spyros Makridakis, «Statistical and Machine Learning forecasting methods: Concerns and ways forward,» *PLoS ONE,* p. 28, 2018.

[27] M. A. S. Romero, «Modelo de pronóstico para la estimación de la utilización y confiabilidad de equipos dinámicos,» *GPE-RMNE,* vol. 53, nº 5, p. 11, 2013.

[28] C. J. M. K. Douglas Montgomery, Introduction to Time Series Analysis and Forecasting, 2 ed., Hoboken: John Wiley & Sons, Inc, 2015.

[29] T. Chai, «Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature,» *Geoscientific Model Development,* vol. 7, p. 1247–1250, 2014.

[30] J. R. Camarillo-Peñaranda, A. J. Saavedra-Montes y C. A. Ramos-Paja, «Recomendaciones para Seleccionar Índices para la Validación de Modelos,» *Tecno Lógicas [en linea],* pp. 109-122, octubre 2013.

[31] O. G. G. M, «Predicción de la presión de burbujeo utilizando aprendizaje automático,» *Innovación y Software,* vol. 4, nº 1, pp. 204-218, 2023.

[32] IBM, «R^2,» IBM Cognos Analytics Documentation, 2024. [En línea]. Available: https://www.ibm.com/docs/en/cognos-analytics/11.2.0?topic=terms-r2.

[33] E. M. RODRÍGUEZ, «Errores frecuentes en la interpretación del coeficiente de determinación lineal,» *Anuario Jurídico y Económico Escurialense,* pp. 315-332, 2005.

[34] P. M. E. G. A. Zuluaga, «Supply chain management strategies based on demand planning in colombia,» *Revista politécnica,* 2021.

[35] I. H. y. L. Torres, «Modelo de pronóstico de demanda para productos del sector eléctrico, Universidad de los Andes,» *Universidad de los Andes,* p. 35, 2021.

[36] A. E. Moszkowitz, «Pronosticar la demanda¿ejercicio de adivinación o fundamento de la planificación operativa?,» *Revista de Antiguos Alumnos del IEEM,* vol. 4, nº 1, pp. 74-80, 2001.

[37] J. B. Lucy D'Agostino McGowan, «Authorize googledrive,» Googledrive package

documentation, junio 2023. [En línea]. Available: https://googledrive.tidyverse.org/reference/drive_auth.html. [Último acceso: 2024].

[38] J. B. Lucy D'Agostino McGowan, «Get Drive files by path or id,» Googledrive package documentation, Diciembre 2022. [En línea]. Available: https://googledrive.tidyverse.org/reference/drive_get.html. [Último acceso: 3 7 2024].

[39] J. B. Lucy D'Agostino McGowan, «Download a Drive file,» Googledrive package documentation, Diciembre 2022. [En línea]. Available: https://googledrive.tidyverse.org/reference/drive_download.html#arguments. [Último acceso: 4 7 2024].

[40] R. Documentation, «Find files on Google Drive,» CRAN R Project Documentation, 2024. [En línea]. Available: https://search.r-project.org/CRAN/refmans/googledrive/html/drive_find.html. [Último acceso: 4 7 2024].

[41] H. W. a. J. Bryan, «readxl: Read Excel Files in R,» 6 7 2023. [En línea]. Available: https://cran.r-project.org/web/packages/readxl/index.html. [Último acceso: 5 7 2024].

[42] R. F. L. H. a. K. M. Hadley Wickham, «dplyr: A Grammar of Data Manipulation,» 17 11 2023. [En línea]. Available: https://cran.r-project.org/web/packages/dplyr/index.html. [Último acceso: 5 7 2024].

[43] J. C. B. S. Y. X. C. S. J. A. J. M. A. D. a. B. B. Winston Chang, «shiny: Web Application Framework for R,» 1 8 2024. [En línea]. Available: https://cran.r-project.org/web/packages/shiny/index.html. [Último acceso: 5 7 2024].

[44] G. G. H. W. Vitalie Spinu, «Package 'lubridate',» 27 September 2023. [En línea]. Available: https://cran.r-project.org/web/packages/lubridate/lubridate.pdf. [Último acceso: 6 7 2024].

[45] C. Sievert, «Interactive web-based data visualization with R, plotly, and shiny,» 19 12 2019. [En línea]. Available: https://plotly-r.com/. [Último acceso: 6 7 2024].

[46] H. Wickham, «ggplot2: Elegant Graphics for Data Analysis,» 2016. [En línea]. Available: https://ggplot2.tidyverse.org/. [Último acceso: 6 7 2024].

[47] T. L. P. Hadley Wickham, «scales: Scale Functions for Visualization,» 28 11 2023. [En línea]. Available: https://cran.r-project.org/web/packages/scales/index.html. [Último acceso: 6 7 2024].

[48] P. E. Puspita y Zulkarnain, «A Practical Evaluation of Dynamic Time Warping in Financial Time Series Clustering,» 23 11 2020. [En línea]. Available: https://ieeexplore.ieee.org/document/9263123/authors#authors. [Último acceso: 1 8 2027].

[49] E. E. Özkoç, «Clustering of Time-Series Data,» *IntechOpen,* p. 20, 2020.

[50] J. D, «Hierarchical Cluster Analysis,» University of Washington Pressbooks, 2024. [En línea]. Available: https://uw.pressbooks.pub/appliedmultivariatestatistics/chapter/hierarchical-cluster-analysis/. [Último acceso: 1 8 2024].

[51] IBM, «Análisis de clústeres jerárquico: Gráficos,» IBM Corporation, 2024. [En línea]. Available: https://www.ibm.com/docs/es/spss-statistics/saas?topic=analysis-hierarchical-

cluster-plots. [Último acceso: 2 8 2024].

[52] Minitab, «Resultados clave para Prueba de Dickey-Fuller aumentada,» Minitab, 2024. [En línea]. Available: https://support.minitab.com/es-mx/minitab/help-and-how-to/statistical-modeling/time-series/how-to/augmented-dickey-fuller-test/interpret-the-results/key-results/. [Último acceso: 2 8 2024].

[53] J. A. Rodrigo, «Gradient Boosting con Python,» Septiembre 2023. [En línea]. Available: https://cienciadedatos.net/documentos/py09_gradient_boosting_python. [Último acceso: 2 8 2024].

10. ANEXOS

ANEXO 1: Modelo de pronósticos de demanda para la optimización de la cadena de suministros: https://github.com/ModelosMegatiendas/ModeloPDF/blob/main/Modelo.pdf

ANEXO 2: Dashboard Interactivo, resultados de pronósticos.
https://drive.google.com/file/d/1qz5oe7fbQ6ebuHfb-sLRmdEbFK-7zB2Q/view