ESCUELA SUPERIOR POLITÉCNICA DEL LITORAL

Facultad de Ingeniería en Electricidad y Computación

Modelo de optimización del safety stock basado en aprendizaje automático para una empresa del sector retail en Ecuador

PROYECTO DE TITULACIÓN

Previo la obtención del Título de:

Magister en Ciencias de Datos

Presentado por:
Alejandro Leonidas Alarcón Lamota
Yelena Maribel Lucas Aguirre

GUAYAQUIL - ECUADOR
Año: 2024

DEDICATORIA

Dedico este proyecto a mi esposo, Oswaldo Carrión, mis padres, y agradezco a mis amigos y profesores por su motivación y apoyo. Su influencia fue esencial para culminar con éxito este proceso.

Yelena Maribel Lucas Aguirre

Dedico esta tesis al Todopoderoso, el sublime creador de todo y a mi querida esposa Gabriela Villamar, cuya luz y sostén inquebrantable han sido mi faro. A mis padres, cuyo amor sin límites ha nutrido mi alma y espíritu. Este triunfo es un testimonio de vuestra presencia y amor inagotable en mi viaje académico.

Leonidas Alejandro Alarcon Lamota

AGRADECIMIENTOS

En primer lugar, agradezco a Dios, en segundo lugar, a mi esposo por su apoyo incondicional siempre, por su buena enseñanza y confiar siempre en mí, y finalmente, muchas gracias a la Mgtr. Karen Calva, quien fue la mayor guía para que todo esto sea posible. A todos ustedes, les expreso mi gratitud por esto y mucho más.

Yelena Maribel Lucas Aguirre

Agradezco profundamente a Dios por ser mi todo, a mi esposa por su amor y apoyo constante y a nuestra guía Karen Calva por ser una luz orientadora en este proceso. Mi gratitud hacia cada uno de ustedes es inmensa por creer en mí y contribuir a hacer este sueño realidad.

Leonidas Alejandro Alarcon Lamota

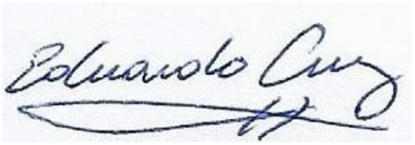
DECLARACIÓN EXPRESA

“Los derechos de titularidad y explotación, nos corresponden conforme al reglamento de propiedad intelectual de la institución; Alejandro Leonidas Alarcón Lamota y Yelena Maribel Lucas Aguirre damos nuestro consentimiento para que la ESPOL realice la comunicación pública de la obra por cualquier medio con el fin de promover la consulta, difusión y uso público de la producción intelectual"

Alejandro L. Alarcón Lamota

Yelena M. Lucas Aguirre

COMITÉ EVALUADOR



Karen Priscilla Calva Yaguana	**Nombre del Profesor**
PROFESOR TUTOR	PROFESOR EVALUADOR

RESUMEN

La gestión eficiente del inventario es crucial para las empresas, ya que influye directamente en la rentabilidad y en la satisfacción del cliente. El equilibrio entre evitar faltantes y reducir los costos asociados al exceso de stock es un desafío constante en este sector. Esta tesis se enfoca en la implementación de un modelo de optimización del safety stock utilizando técnicas de aprendizaje automático supervisado en una cadena de supermercado en Ecuador. El objetivo principal es predecir el safety stock de productos para reducir faltantes y maximizar la rotación de mercadería. Se identificaron productos perecibles de alta importancia, se seleccionaron algoritmos de aprendizaje automático adecuados y se integró el modelo en una plataforma para su uso práctico. Los resultados indican que el modelo Random Forest Regressor es la opción más robusta. Las conclusiones sugieren continuar utilizando este modelo, realizar actualizaciones periódicas, fomentar la colaboración entre usuarios y el sistema de predicción, y monitorear de cerca las condiciones cambiantes en el mercado minorista.

Palabras Clave: Safety stock, Supermercado, Aprendizaje Automático Supervisado, Predicción

ABSTRACT

Efficient inventory management is of paramount importance for supermercado companies as it directly influences profitability and customer satisfaction. Striking a balance between avoiding stockouts and reducing the costs associated with excess stock is an ongoing challenge in this sector. This thesis focuses on the implementation of a safety stock optimization model using supervised machine learning techniques in a supermercado chain in Ecuador. The primary objective is to predict safety stock for products to reduce stockouts and maximize merchandise turnover. High-importance perishable products were identified, suitable machine learning algorithms were selected, and the model was integrated into a practical platform. Results indicate that the Random Forest Regressor model is the most robust option. The conclusions suggest continuing to use this model, making periodic updates, promoting collaboration between users and the prediction system, and closely monitoring changing market conditions in supermercado.

Keywords: *Safety Stock, Reatil, Supervised Machine learning, Forecast*

ÍNDICE GENERAL

ABREVIATURAS

SS	Stock de Seguridad
SKU	Stock-keeping unit
R2	Coeficiente de determinación
RMSE	Error de raíz cuadrada media
ML	Machine Learning
AI	Artificial Intelligence

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

CAPÍTULO 1

1. INTRODUCCIÓN

La gestión de inventarios es esencial para el funcionamiento y rentabilidad de las empresas minoristas. En un entorno empresarial donde la eficiencia en la gestión de recursos es crucial, el equilibrio entre un inventario excesivo y uno insuficiente es un desafío constante. Estas polaridades en lo referente al inventario pueden afectar la experiencia del cliente y la salud financiera de la empresa.

La industria de productos de consumo masivo se enfrenta a un entorno volátil, donde las ventas cambian rápidamente y las fluctuaciones son impredecibles. Esto requiere una adaptabilidad inmediata para evitar problemas de desabastecimiento o exceso de inventario.

En este contexto, la gestión de parámetros como lo es el safety stock en los modelos de abastecimiento es fundamental para un rendimiento óptimo. Esta investigación busca analizar este problema y proponer soluciones basadas en la ciencia de datos para optimizar el abastecimiento, reducir errores y mejorar la eficiencia en la gestión de inventarios en empresas minoristas.

1.1. Descripción del problema

La gestión eficiente del inventario es un aspecto crítico en el sector supermercado, especialmente en el contexto de un entorno volátil como el de productos de consumo masivo. Un equilibrio adecuado en el nivel de inventario, manteniendo el safety stock en un punto óptimo, es fundamental para garantizar la rentabilidad y la satisfacción del cliente. Un exceso de inventario conduce a costos adicionales y riesgos de obsolescencia, mientras que un inventario insuficiente aumenta el riesgo de desabastecimiento y pérdida de ventas, afectando negativamente la lealtad del cliente.

La predicción precisa de la demanda es vital para la gestión de inventarios, como señalan Zhu, Geng, & Sarkis (2019) y Tang, Guo, & Liu (2020). Estudios indican que errores en la predicción pueden resultar en significativas pérdidas económicas y en la disminución de la satisfacción del cliente. En la empresa de supermercado, actualmente dieciséis planificadores de abastecimiento se encargan de manejar los parámetros de abastecimiento, incluyendo el safety stock, basándose en su experiencia y análisis de tendencias del mercado. Sin embargo, con la previsión de un aumento de 250 a 300 sucursales en los próximos cinco años, esta metodología basada en la experiencia y el manejo manual de datos está cada vez más desafiada por el volumen creciente de datos a procesar.

Cada planificador debe analizar aproximadamente seis mil combinaciones de productos por sucursal diariamente, cifra que se espera aumente en el futuro. Además, la actualización de parámetros se realiza mediante el uso de archivos planos, lo cual está sujeto a errores y genera incertidumbre en el proceso de abastecimiento. Esto podría llevar a una caída en la productividad y eficiencia debido a la falta de herramientas de control adecuadas.

Por lo tanto, es imperativo modernizar y mejorar nuestra metodología de gestión de inventarios. La adopción de soluciones avanzadas basadas en la ciencia de datos y la automatización puede ayudar a optimizar el proceso de abastecimiento, reducir errores, y adaptarse de manera más eficaz a las fluctuaciones en la demanda, asegurando así un rendimiento óptimo en el abastecimiento y la satisfacción continua del cliente.

1.2. Justificación del problema

Se busca utilizar técnicas de aprendizaje automático supervisado para mejorar la eficiencia en la gestión de inventarios, reduciendo costos y aumentando la rentabilidad. Esto garantiza la disponibilidad de productos, mejora la satisfacción del cliente y permite una toma de decisiones fundamentada en el análisis de datos.

Los modelos de aprendizaje automático tienen la capacidad de adaptarse ante las variaciones en el mercado y a las fluctuaciones en la demanda. Esto simplifica la

tarea de adaptar las estrategias de inventario y permite que estas estrategias sean escalables a medida que la empresa crece.

1.3. Solución propuesta

La solución propuesta consiste en desarrollar un sistema de gestión de inventarios semiautomatizado supervisado que busca alcanzar varios objetivos clave.

En primer lugar, se implementará un sólido modelo de pronóstico de la demanda, que permitirá realizar predicciones precisas de las necesidades futuras de inventario. A continuación, el sistema generará recomendaciones diarias para determinar el nivel de inventario de seguridad (safety stock) basándose en los pronósticos, los niveles de inventario actuales y los parámetros de los diversos sistemas de abastecimiento.

Posteriormente, se integrará con los diferentes modelos de abastecimiento utilizados por la compañía, lo que facilitará la comunicación con los sistemas de pedidos y la distribución de inventario hacia las sucursales. Por último, se permitirá la entrada manual de ajustes en el nivel de inventario de seguridad lo que dará a los usuarios la flexibilidad necesaria para abordar eventos imprevistos.

En conjunto, esta solución tiene como objetivo optimizar la gestión de inventarios, garantizando una mayor precisión en los niveles de inventario y como resultado una mayor satisfacción del cliente.

1.4. Objetivos

1.4.1. Objetivo General

Predecir el safety stock de productos en una cadena de supermercado mediante técnicas de aprendizaje automático supervisado para reducir faltantes y maximizar la rotación de productos.

1.4.2. Objetivos Específicos

1. Identificar los productos perecibles de mayor importancia en el surtido en función de su rentabilidad y rotación con el fin de dar prioridad al análisis y la gestión.
2. Seleccionar los algoritmos de aprendizaje automático más adecuados para aplicar la predicción de demanda y determinación del safety stock en función de las características y dinámicas de la cadena de supermercado.
3. Integrar el modelo de aprendizaje automático en una plataforma para que los usuarios puedan examinar y confirmar la sugerencia del safety stock para integrarlo a los diversos modelos de abastecimiento.
4. Reprocesar los modelos de predicción tras identificar las combinaciones de sucursal y estadístico que requieren safety stock.

1.5. Metodología

En el presente proyecto se emplearán técnicas de aprendizaje automático supervisado para evaluar parámetros y seleccionar modelos predictivos que permita obtener resultados efectivos que aborden las necesidades específicas relacionadas con la gestión del safety stock en una empresa de supermercado.

La primera fase consiste en la recolección de datos históricos relacionados con el safety stock, las ventas históricas y otras variables relevantes en el contexto de la empresa del sector supermercado en Ecuador. Esta recopilación abarcará al menos los últimos dos años y permitirá obtener patrones de comportamiento de los productos. Esta información se utilizará para crear un conjunto de datos como base del estudio.

En la segunda fase, se llevará a cabo el tratamiento y análisis de los datos recopilados, esto incluye la limpieza de datos para eliminar valores nulos, faltantes y puntos atípicos, también se procederá a normalizar los datos y se realizará un análisis descriptivo para comprender las características obtenidas. Se evaluará la correlación entre diferentes variables y su influencia en la predicción de la variable dependiente, que en este caso es el safety stock.

En la tercera fase, se diseñará y desarrollará un modelo de pronóstico de demanda robusto utilizando técnicas de aprendizaje automático supervisado como: random forest, extreme gradient boosting machine (XGBoost) y regresión lineal. Estos modelos permitirán predecir con precisión las necesidades futuras de inventario en función de los datos históricos y las variables relevantes. El objetivo es alcanzar una cobertura óptima y adaptativa ante las fluctuaciones de ventas. Los resultados de experimentar con diferentes técnicas de aprendizaje automático permitirán determinar si alguno de los modelos sugeridos puede asistir en el proceso de abastecimiento con el objetivo principal de optimizar la ecuación.

$$SS = RMSE * z * \sqrt{n}\,.$$

Para el modelado del safety stock (SS) se consideran tres variables las cuales son error de pronóstico, factor multiplicador y lead time.

La última fase implica la integración del modelo de pronóstico junto con la recomendación del safety stock previsto con los sistemas de pedidos y distribución de mercadería hacia las sucursales. Además, se permitirá la entrada manual para ajustes del safety stock por partes de los usuarios lo que brindará flexibilidad para abordar eventos imprevistos.

1.6. Resultados esperados

El producto final es un sistema de recomendación con características de ser semiautomatizado y supervisado para la gestión de inventarios que:

1. Predice la demanda con un margen de error menor al 5%.

2. Proporciona recomendaciones semanales para definir el safety stock.
3. Se integra con los diversos modelos de abastecimiento de la compañía mediante plataforma.
4. Permite la entrada manual para ajustes del safety stock basados en la experiencia del usuario en el área para eventos no previstos.

1.7. Tabla

A diario la empresa registra cerca de doscientos veinte mil ventas de sus clientes, estas transacciones (tickets) se encuentran dentro de base de datos. De este total de transacciones el sector cárnico representa un aproximado de 5.6% lo que aproximadamente es once mil transacciones por día.

Este proyecto se trabajará en base a un conjunto de datos con aproximadamente doscientos millones de transacciones de venta efectuadas en doscientos cincuenta y dos sucursales a nivel nacional. Estas ventas están comprendidas entre enero de 2021 y agosto de 2023.

El conjunto de datos cuenta con las siguientes variables: 1 de tipo espacial, 7 categóricas y 6 numéricas.

Tabla 1.1 Variables del modelo

#	VARIABLES	DESCRIPCIÓN	TIPO VARIABLE
1	**Periodo**	Fecha en la que se realizó la venta en la sucursal.	Espacial
2	**Sucursal**	Código con el que se identifica la sucursal.	Categórica
3	**SKU**	Código con el que se identifica el producto.	Categórica
4	**Descripción**	Nombre del producto.	Categórica
5	**Marca**	Etiqueta que indica el estado del artículo (activo, inactivo, suspendido, nuevo)	Categórica
6	**Consumo**	Número de unidades vendidas.	Numérico
7	**Stock**	Número de unidades que posee el producto.	Numérico
8	**PVP**	Precio de venta del producto.	Numérico
9	**Costo**	Costo de adquisición del producto.	Numérico
10	**Clase**	Categoría de productos que comparten características comunes.	Categórica

11	**Subclase**	Segmento dentro de una clase que agrupa productos basados en sus características.	Categórica
12	**Promoción**	Etiqueta que indica si el producto está o no en promoción.	Categórica

Fuente: Elaboración propia

CAPÍTULO 2

2. ESTADO DEL ARTE

En el dinámico mundo del supermercado la adecuada gestión del inventario emerge como uno de los pilares fundamentales para garantizar operaciones eficientes y rentables. El stock de seguridad también conocido como "safety stock", actúa como un sistema de compensatorio de inventario contra las variaciones imprevistas en la demanda o en el suministro previniendo tanto el exceso de stock que conlleva costos innecesarios, así como el desabastecimiento que puede resultar en la pérdida de ventas y la insatisfacción del cliente (Axsäter, 2015).

En el marco de esta investigación centrada en la gestión del inventario y en particular en la estimación del safety stock óptimo es común recurrir a enfoques tradicionales que se apoyan en métodos determinísticos o reglas basadas en la experiencia previa. No obstante, en la era actual de la digitalización y el abundante flujo de datos se divisa el aprendizaje automático supervisado como una herramienta de gran potencial para ofrecer predicciones más precisas y adaptativas en esta área (David Diaz, 2021). Esta rama de la inteligencia artificial tiene la capacidad de identificar patrones a partir de extensos conjuntos de datos históricos lo que permite generar estimaciones altamente ajustadas y personalizadas las cuales son específicas para las condiciones y particularidades de cada sucursal.

El sector minorista se beneficia especialmente de estas técnicas, dado su manejo constante de las fluctuaciones en la demanda, derivadas de factores como la estacionalidad, las promociones y los cambios en las tendencias del mercado (Aggarwal & Agrawal, 2019). En este contexto, la flexibilidad y precisión del aprendizaje automático supervisado lo posiciona como una solución prometedora para abordar estos desafíos y optimizar la gestión del safety stock.

2.1. Métodos tradicionales versus métodos modernos

Históricamente, la predicción de demanda en supermercado se basaba en métodos estadísticos convencionales. Técnicas como la media móvil, la suavización

exponencial y la descomposición estacional eran ampliamente utilizadas. Estos métodos si bien pueden ser efectivos en condiciones estables y con patrones de demanda consistentes, pero pueden no adaptarse bien a situaciones volátiles o a cambios abruptos en las tendencias.

Con el auge de la ciencia de datos y la disponibilidad de grandes conjuntos de datos, el aprendizaje automático supervisado emerge como una herramienta poderosa para la predicción de demanda. Estos enfoques que incluyen técnicas como regresión lineal, árboles de decisión, máquinas de soporte vectorial (SVM) y redes neuronales tienen la capacidad de modelar relaciones complejas y no lineales entre variables y de adaptarse rápidamente a nuevos patrones en los datos. Además, estos métodos pueden integrar una variedad más amplia de fuentes de datos, como información de redes sociales, datos meteorológicos y eventos promocionales, mejorando así la precisión de las predicciones.

2.2. Aplicaciones en supermercado

En la actualidad, existen diversos modelos de aprendizaje supervisado que son aplicables a problemas con múltiples variables de entrada (características) con el propósito de predecir la variable de salida (objetivo) basándose en estas características.

La adopción de técnicas avanzadas de predicción de demanda ha transformado a muchas empresas de supermercado a nivel mundial. Por ejemplo, Walmart, una de las mayores cadenas de tiendas minoristas del mundo ha implementado sistemas de aprendizaje automático para mejorar sus previsiones lo que les ha permitido reducir significativamente su inventario mientras mantienen e incluso mejoran sus niveles de servicio (Yi, 2023).

En un estudio llevado a cabo Walmart, se evaluaron tres algoritmos de aprendizaje automático destacando la eficacia del modelo XGBoost en la predicción de tendencias de precios minoristas, alcanzando un MAE de 0.124 y una puntuación R^2 de 0,983 (Yang, 2023).

Otro estudio explora cómo el método Extreme Gradient Boosting (XGBoost) puede utilizarse para predecir las ventas para un total de 45 tiendas en Walmart. Los

resultados obtenidos a través de medidas como el error absoluto medio, R2 y RMSE revelaron que la técnica de aprendizaje automático XGBoost puede utilizarse de manera efectiva para predecir las ventas, lo que puede guiar a los gerentes de ventas y otros profesionales en la toma de decisiones relacionadas con el abastecimiento de productos (Yetunde Faith, 2022).

En el estudio "Comparative Analysis of Supervised Machine Learning Techniques for Sales Forecasting" realizado en Walmart, se evaluaron diversas técnicas de aprendizaje automático supervisado, destacándose el algoritmo de Random Forest por su eficacia en la predicción de ventas. Este modelo se aplicó para predecir las ventas en tiendas minoristas en distintas ubicaciones geográficas, tomando en cuenta factores como datos de ventas previas, eventos promocionales, semanas festivas, temperatura, precio del combustible, Índice de Precios al Consumidor (CPI) y tasa de desempleo.

El estudio "Solar radiation forecasting using MARS, CART, M5, and random forest" (Rachit Srivastava, 2019) presenta un análisis detallado de la predicción de la radiación solar utilizando varios modelos de aprendizaje automático. Para evaluar los modelos se emplearon el error cuadrático medio (RMSE) y el error absoluto medio (MAE). Se realizó la predicción de radiación solar desde 1 día hasta 6 días por adelantado para cada mes de 2017. El RMSE varió desde un 83.9% para pronósticos de 1 día por adelantado hasta un 90.92% para pronósticos de 6 días por adelantado indicando una precisión razonable en las predicciones a corto plazo

El estudio "Electricity consumption forecasting in Italy using linear regression models" (Vincenzo Bianco, 2009) se enfoca en desarrollar un modelo de pronóstico a largo plazo para el consumo eléctrico en Italia, empleando la regresión lineal y tomando en cuenta factores económicos y demográficos. Los modelos mostraron MAE del ±2% en comparación con los datos históricos.

En resumen, la predicción de demanda en el sector supermercado ha evolucionado considerablemente en las últimas décadas. La integración de técnicas modernas impulsadas por el aprendizaje automático supervisado ha permitido a las empresas adaptarse mejor a un entorno irregular y mejorar su eficiencia operativa. La adopción

de estas técnicas combinada con una implementación adecuada promete ser un diferenciador clave en la competitividad y éxito de las empresas de supermercado en el futuro.

2.3. Aprendizaje automático en la estimación del safety stock

El aprendizaje automático ha revolucionado muchos campos esto incluye a la gestión de inventarios dentro del sector supermercado. Diversos algoritmos han mostrado ser efectivos para adaptarse a esta dinámica dentro de ellos tenemos.

2.3.1. Random forest

La esencia de Random Forest radica en su habilidad para construir y utilizar múltiples árboles de decisión para realizar predicciones más precisas y confiables se basa en la idea de “bosques aleatorios” o conjunto de árboles de decisión. (Punia, 2020).

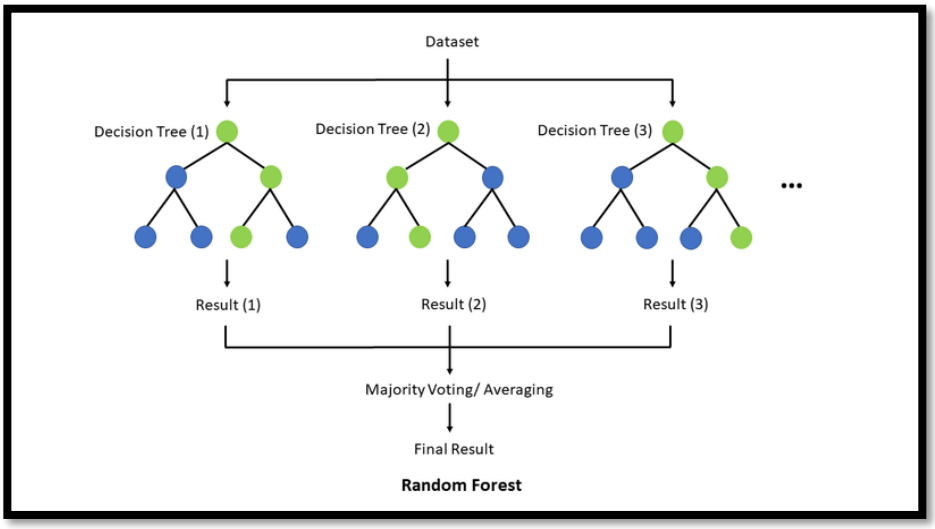


Figura 2.1 Random forest

Fuente: https://en.wikipedia.org/wiki/File:Random_forest_explain.png

2.3.1.1. Bootstrap aggregating (bagging)

El primer paso en Random Forest es el muestreo con reemplazo, conocido como "bootstrap". Para cada árbol que se va a construir se toma una muestra aleatoria con reemplazo del conjunto de datos original. Esto introduce diversidad en los árboles, ya que cada uno se entrena con un subconjunto ligeramente diferente de datos.

2.3.1.2. Selección de características

En lugar de considerar todas las características (o variables) en cada división del árbol, Random Forest selecciona un subconjunto aleatorio de estas. Esta aleatoriedad asegura que los árboles no estén correlacionados y amplía aún más la diversidad en el bosque, reduciendo la varianza y ayudando a evitar el sobreajuste.

2.3.1.3. Construcción de árboles

Con el conjunto de datos muestreado y la selección de características, el algoritmo construye un árbol de decisión. A diferencia de otros métodos que podan el árbol, Random Forest permite que cada árbol crezca a su máxima profundidad, lo que le otorga gran flexibilidad.

2.3.1.4. Predicción

Una vez entrenado el bosque, al ingresar una nueva entrada para predicción, cada árbol en el bosque emite una predicción. En problemas de regresión se toma el promedio de las predicciones de todos los árboles.

La capacidad de Random Forest para manejar grandes cantidades de características y su resistencia al sobreajuste lo hacen ideal para el sector supermercado, donde los datos suelen ser extensos y complejos. Puede adaptarse fácilmente a las fluctuaciones de demanda, tendencias estacionales y características no lineales en los datos, permitiendo a las empresas obtener una estimación precisa de la demanda futura y ajustar sus estrategias de inventario en consecuencia.

2.3.2. Extreme gradient boosting

XGBoost (Extreme Gradient Boosting) es otro algoritmo popular que se puede utilizar en el contexto de predicciones para un posterior cálculo de safety stock basado en modelos de aprendizaje supervisado. A continuación, se detalla cómo opera XGBoost en este contexto:

2.3.2.1. Gradient boosting

Al igual que Random Forest, XGBoost también utiliza un enfoque de ensamble, pero se basa en el boosting en lugar del bagging. En el boosting, los árboles se construyen secuencialmente y cada árbol se ajusta para corregir los errores cometidos por los árboles anteriores. Esto permite que el modelo se concentre en las áreas donde comete errores, lo que puede ser beneficioso para predecir ya que puede adaptarse a patrones complejos y sutiles en los datos.

2.3.2.2. Regularización

XGBoost incluye técnicas de regularización que ayudan a prevenir el sobreajuste. Esto es importante en el contexto de las predicciones donde es crucial evitar estimaciones excesivamente optimistas o pesimistas que podrían llevar a problemas en la gestión del inventario. Las técnicas de regularización en XGBoost incluyen la limitación de la profundidad de los árboles y la penalización de los árboles que tienen hojas con poca importancia.

2.3.2.3. Selección de características

Al igual que Random Forest, XGBoost también permite la selección aleatoria de características en cada árbol. Esto introduce diversidad en el modelo y evita la correlación entre los árboles, lo que puede mejorar la calidad de las predicciones.

2.3.2.4. Construcción de árboles

XGBoost construye árboles de decisión de manera similar a Random Forest pero utiliza técnicas de optimización más avanzadas para encontrar la mejor estructura de árbol. Esto puede llevar a modelos más eficientes y precisos.

2.3.2.5. Predicción

Cuando se utiliza XGBoost para predecir, cada árbol en el ensamble emite una predicción y estas predicciones se promedian para obtener la estimación final. Esta estimación es más robusta y precisa debido

a la naturaleza secuencial del boosting y las técnicas de regularización.

En el contexto de las predicciones, XGBoost también se adapta bien a datos extensos y complejos. Puede capturar relaciones no lineales y patrones sutiles en los datos, lo que es esencial para estimar de manera precisa las necesidades de inventario en entornos de demanda fluctuante y características estacionales. En resumen, XGBoost es una herramienta valiosa para mejorar la precisión de las estimaciones de safety stock en el sector supermercado y optimizar las estrategias de inventario.

2.3.3. Regresión lineal

Aunque es uno de los métodos más simples y antiguos la regresión lineal puede ser muy efectiva, especialmente cuando las relaciones entre las variables independientes y dependientes son lineales. En el contexto del safety stock, si se tiene un histórico claro de cómo ciertas variables (como promociones o eventos especiales) afectan la demanda la regresión lineal puede ser una herramienta poderosa.

2.3.3.1. Modelo

La regresión lineal múltiple utiliza la ecuación $y = b_0 + b_1 x_1 + \cdots + b_n x_n$ (Gujarati, 2015), donde *y* es la variable dependiente, $x_1 + x_2 \ldots + x_n$ son las variables independientes y $b_1 + b_2 \ldots + b_n$ son los coeficientes por estimar.

2.3.3.2. Estimación de coeficientes

Mediante técnicas como el método de mínimos cuadrados, se buscan los coeficientes $b_0, b_1, \ldots, b_n$ que minimicen la suma de los cuadrados de las diferencias entre las observaciones reales y las predichas.

En el contexto del supermercado, la regresión lineal múltiple puede ser usada para modelar y predecir ventas basadas en múltiples factores como precios, promociones, día de la semana, estacionalidades, entre otros. Ayuda a descomponer el impacto de

cada factor y a tomar decisiones estratégicas basadas en un análisis multidimensional.

2.3.3.3. Verificación de supuestos

La regresión lineal múltiple también se basa en suposiciones de linealidad, independencia, homocedasticidad y normalidad de errores. Además, se debe tener cuidado con la multicolinealidad, donde dos o más variables independientes están altamente correlacionadas entre sí.

2.3.4. Variables relevantes

Las variables como el error de pronóstico, el factor multiplicador y el lead time juegan roles esenciales en la formulación de estrategias efectivas para la estimación del safety stock. Cada una de estas variables representa una dimensión específica de la problemática en la gestión de inventarios y, por lo tanto, su adecuada comprensión y manejo es indispensable para diseñar modelos de aprendizaje automático que permitan optimizar la gestión del safety stock. A continuación, se explica de una forma más detalla cada una de ellas.

- **Error de pronóstico:** Este refleja la diferencia entre la demanda pronosticada y la demanda real. Es una medida esencial porque cualquier desviación en las previsiones se traduce directamente en la necesidad de un safety stock adicional para compensar ese error y evitar faltantes.
- **Factor multiplicador (z):** Este factor se utiliza para escalar el error de pronóstico y se basa generalmente en la confianza deseada o en el nivel de servicio. Por ejemplo, si se desea un nivel de servicio del 95%, el factor z podría ser 1.645 (para una distribución normal). Cuanto mayor sea el factor z, mayor será el safety stock, reflejando una mayor precaución contra posibles faltantes.
- **Lead time (tiempo de espera):** Es el tiempo que transcurre desde que se realiza un pedido hasta que se recibe. Durante este tiempo, la empresa debe confiar en su inventario para satisfacer la demanda. Si

el lead time es largo y hay variabilidad en la demanda o en el tiempo de entrega se necesita un mayor safety stock para evitar faltantes durante ese período.

2.4. Métricas de evaluación

Al entrenar un modelo para series de tiempo es fundamental evaluar su desempeño y determinar qué tan precisas son sus predicciones. Existen diferentes métricas para evaluar la eficacia de un modelo dentro de las más comunes tenemos:

- **Error Absoluto Medio (MAE):** Es el promedio de los valores absolutos de los errores y representa cuánto se desvía, en promedio, las predicciones del modelo de los valores reales.

$$\mathrm{MAE}=\frac{1}{n}\sum_{i=1}^{n}\left|Yi - \widehat{Yi}\right|$$

(Levin, 2004)

Donde *Yi* es el valor real observado y $\widehat{Yi}$ es el valor predicho por el modelo, y n es el número total de observaciones.

Un MAE pequeño sugiere que el modelo tiene un buen desempeño en sus predicciones. Sin embargo, al ser un promedio, puede ser sensible a valores extremos.

- **Error Cuadrático Medio (MSE):** Es el promedio de los errores al cuadrado y castiga más a los errores grandes que el MAE.

$$\mathrm{MSE}=\frac{1}{n}\sum_{i=1}^{n}\left(Yi - \widehat{Yi}\right)^2$$

(Levin, 2004)

Donde *Yi* es el valor real observado y $\widehat{Yi}$ es el valor predicho por el modelo, y n es el número total de observaciones.

Un MSE pequeño indica que el modelo hace predicciones cercanas a los valores reales. Un MSE grande sugiere que el modelo puede estar teniendo problemas con algunos valores atípicos o no captura bien la tendencia de los datos.

- **Raíz del Error Cuadrático Medio (RMSE):** Es la raíz cuadrada del MSE y proporciona el error en las mismas unidades que la variable objetivo.

$$RMSE = \sqrt{MSE}$$

Al igual que con el MSE, un RMSE pequeño indica predicciones precisas mientras que un RMSE grande sugiere posibles problemas con el modelo.

- **Error Porcentual Absoluto Medio (MAPE):** Representa el error como un porcentaje lo cual puede ser útil para comparar modelos entre diferentes conjuntos de datos o escalas.

$$\text{MAPE}=\frac{100}{n}\sum_{i=1}^{n}\frac{Yi - \widehat{Yi}}{Yi}$$

(Levin, 2004)

Donde *Yi* es el valor real observado y $\widehat{Yi}$ es el valor predicho por el modelo, y n es el número total de observaciones.

Un MAPE cercano a 0% indica que el modelo tiene un excelente desempeño mientras que un MAPE alto sugiere que el modelo puede no estar capturando adecuadamente las dinámicas de la serie temporal.

2.5. Librerías y software por utilizar

2.5.1. Librerías

La tabla 2.1 detalla las librerías como: pandas, NumPy, scikit-learn, XGBoost, Random Forest y Regresión Lineal. Reconocidas por su eficacia en el tratamiento, depuración y modelado de los datos.

Tabla 2.1 Resumen de librerías

Librería	Descripción
Pandas	Manipulación y transformación de datos
	Carga de archivos CSV
	Manejo de valores faltantes
NumPy	Cálculos matemáticos avanzados
	Procesamiento de datos esenciales
Scikit-learn	Modelado de datos

	Búsqueda de hiperparámetros
XGBoost	Modelado avanzado con regresión basada en árboles de decisión optimizados
Random Forest	Potente técnica de modelado con regresión basada en múltiples árboles de decisión
Regresión Lineal	Modelo básico de regresión para relaciones lineales entre variables

Fuente: Elaboración propia

2.5.2. Software

La tabla 2.2 muestra los softwares a utilizar como: Phyton para realizar el modelo de predicción de stock de seguridad, Power BI se empleará para diversas opciones de representación visual y SQL server para realizar el almacenamiento de los datos.

Tabla 2.2 Resumen Software

Software	Descripción
Python	Lenguaje de programación versátil y potente fue usado para realizar el modelo de predicción.
Power BI	Herramienta de visualización y análisis de datos.
SQL Server	Sistema de gestión de bases de datos relacionales.

Fuente: Elaboración propia

2.5.3. Herramientas

En la tabla 2.3 se observa los componentes de la laptop que se usará en este proyecto.

Tabla 2.3 Resumen de componentes de laptop

Componente	Descripción
Procesador	Intel(R) Core(TM) i7-10510U CPU @ 1.80GHz
Memoria RAM	16GB
Sistema Operativo	Procesador x64.

Fuente: Elaboración propia

2.5.4. Resumen del Proceso

En el proceso de cálculo de predicciones, inicialmente se lleva a cabo una fase de preprocesamiento. Durante esta etapa, se realizan diversas

transformaciones y limpiezas en el conjunto de datos. Esto incluye conversiones de tipo de dato, identificación y manejo de outliers, y la imputación de valores faltantes. Además, se ejecuta una segmentación y análisis profundo para identificar y retener las combinaciones de 'SUCURSAL' y 'SKU' que representan hasta el 80% de participación de consumo.

Posteriormente, en la fase de modelado, se establecen hiperparámetros para los modelos de predicción. Se itera sobre combinaciones únicas de 'SUCURSAL' y 'SKU', dividiendo los datos en conjuntos de entrenamiento y prueba para cada combinación. Durante este proceso, se entrenan y validan tres modelos principales: RandomForest, Regresión Lineal y, por supuesto, XGBoost. Se utiliza la herramienta `RandomizedSearchCV` para optimizar los hiperparámetros de estos modelos. Una vez que los modelos están entrenados, se evalúa su precisión utilizando el cálculo del RMSE, MAE, MAPE y R^2. Finalmente, se seleccionan las predicciones del modelo que presenta el menor MAE luego para cada combinación se procede con el cálculo del safety stock y se consolidan en una tabla para su posterior análisis o visualización.

CAPÍTULO 3

3. DISEÑO E IMPLEMENTACIÓN

3.1. Exploración y validación de datos

Para el presente proyecto se obtuvo el registro de datos recolectados de los últimos 2 años comprendiendo desde el año 2021 al 2023.

Tabla 3.1 Número de entidades a usar en el modelo

Entidad	Cantidad
Años	2
Sucursales	253
SKU	234
Semanas	143

Fuente: Elaboración propia

Conforme la tabla 3.1, los datos obtenidos son previamente de ventas realizadas en cada sucursal en el Ecuador, se trabajará con datos de 2 años para analizar posibles tendencias estacionales y cambios a lo largo del tiempo en la demanda y el safety stock. Además, se considerarán un total de 253 sucursales para obtener una visión integral de las necesidades de inventario en toda la empresa, junto con 234 productos (SKU) diferentes que requieren una gestión precisa del safety stock para prevenir faltantes y maximizar las ventas. Asimismo, se examinarán 143 semanas en el análisis lo que permitirá evaluar cómo varía la demanda a lo largo del tiempo.

3.1.1. Procesos de limpieza y transformación

En el análisis de los datos, es esencial llevar a cabo procedimientos de limpieza y transformación para garantizar que los datos estén en un formato adecuado y libre de anomalías que puedan afectar los resultados. Los siguientes son los pasos llevados a cabo para este propósito:

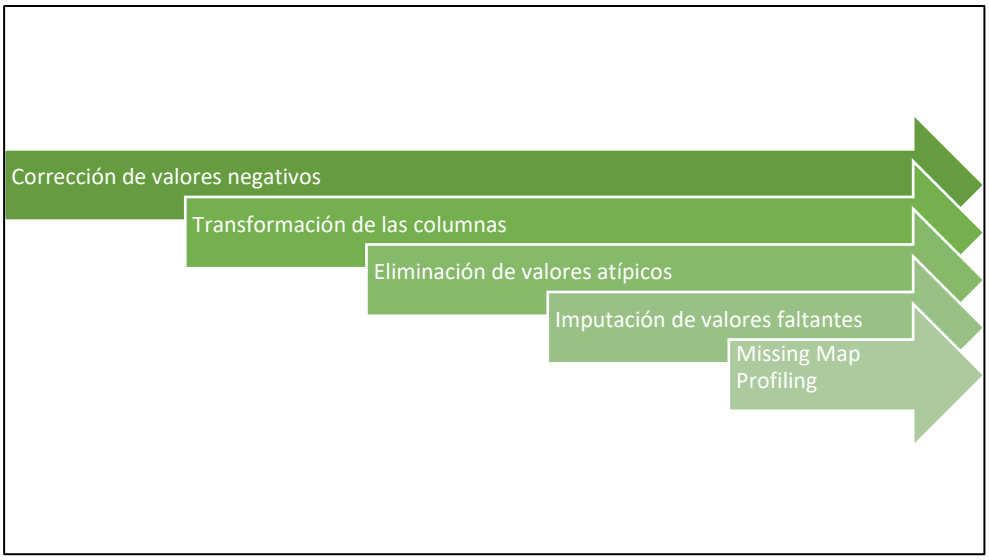


Figura 3.1 Proceso de limpieza y transformación de datos
Fuente: Elaboración propia

1. Corrección de valores negativos dado que el consumo no puede ser negativo en un contexto real, todos los valores negativos en la columna 'CONSUMO' fueron reemplazados con ceros lo que indica la ausencia de consumo durante esos períodos.
2. Transformación de la columna 'PERIODO', que representa una fecha la misma que transformada para ajustarla al formato de fecha estándar (año – semana). Esta transformación es crucial para realizar análisis temporales o de series temporales.
3. Para garantizar que el análisis no estuviera sesgado por valores extremos, se empleó el método del rango intercuartil (IQR). Los valores que estaban fuera de este rango, considerados valores atípicos, fueron excluidos del conjunto de datos.
4. Es común encontrar datos incompletos en conjuntos de datos reales. En lugar de descartar estos registros, se optó por la imputación, reemplazando los valores faltantes con la mediana de cada columna correspondiente.

3.2. Selección de características relevantes

En la tabla 3.2, presenta la clasificación de las variables de este proyecto en dos categorías: independientes y dependientes. Aquí hay una descripción de cada parte de la tabla.

- **Independientes:** Estas variables son aquellas que se utilizarán para predecir o explicar la variabilidad en la variable dependiente, "CONSUMO". Cada fila bajo "Independientes" representa una característica específica o atributo de tus datos, como el PERIODO, SUCURSAL, SKU, DESCRIPCIÓN, MARCA, STOCK, PVP, COSTO, CLASE, SUBCLASE y PROMOCIÓN.
- **Dependientes:** Esta columna enumera las variables que son dependientes, con "Consumo" como la variable principal. La variable dependiente es aquella que estás tratando de entender, predecir o explicar en función de las variables independientes.

Tabla 3.2 Clasificación de Variables en el Modelo de Optimización del Safety Stock

Independientes	Dependientes
Periodo	Consumo
Sucursal	
SKU	
Descripción	
Marca	
Consumo	
Stock	
PVP	
Costo	
Clase	
Subclase	
Promoción	

Fuente: Elaboración propia

En la figura 3.2 se observa La matriz de correlación examina las relaciones lineales entre las variables de un conjunto de datos. Correlaciones fuertes indican conexiones importantes. En nuestro análisis, se encontró una correlación moderada positiva (0.287) entre CONSUMO y STOCK, sugiriendo su relevancia para modelar la demanda. Además, la correlación alta (0.767) entre 'PVP' y 'COSTO' señala una conexión crucial en estrategias de precios y márgenes de beneficio.

Sin embargo, las variables SUCURSAL y SKU, al ser códigos, tienen correlaciones no prácticas. En resumen, el análisis inicial revela correlaciones significativas entre "CONSUMO" y 'STOCK', 'PVP' y 'COSTO', indicando su importancia para análisis posteriores.

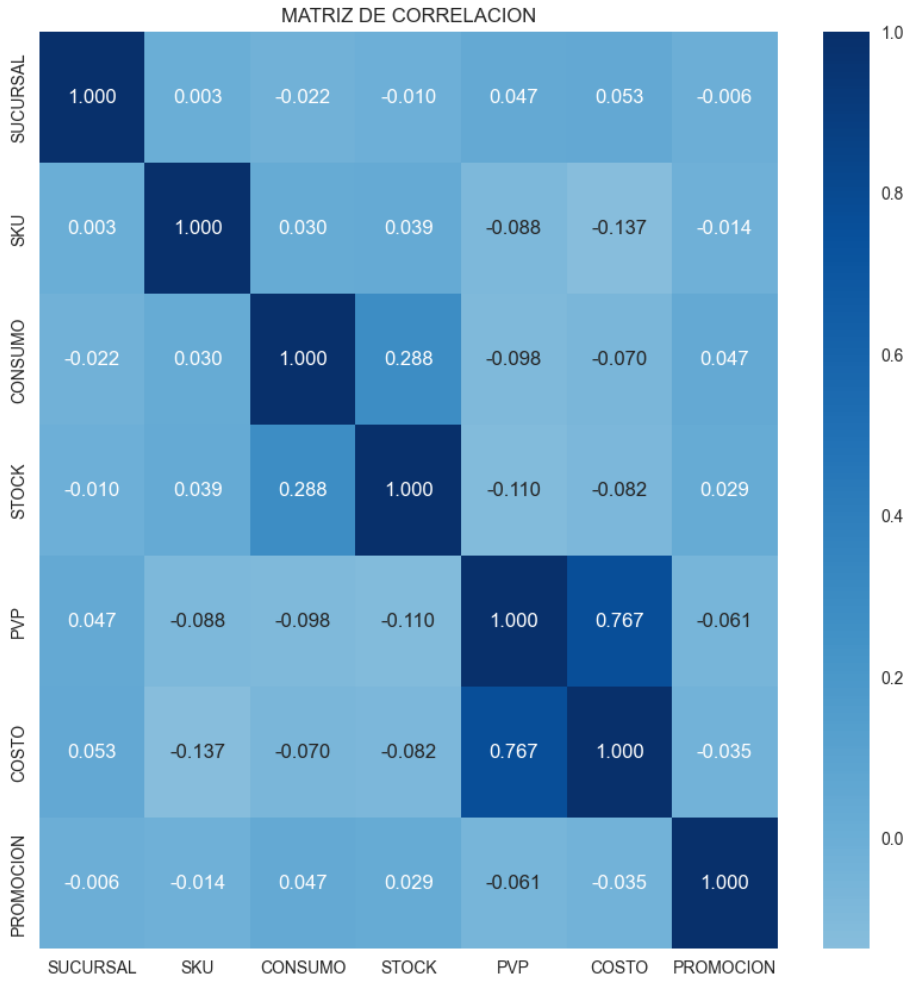


Figura 3.2 Matriz de correlación

Fuente: Elaboración propia

3.3. Segmentación de datos

La segmentación de datos está basada en la participación de los ‘SKU’ en 'SUCURSAL' basados en el principio de Pareto por lo que se llevó a cabo un proceso para identificar qué combinaciones de 'SUCURSAL' y ‘SKU’ eran más relevantes en términos de consumo. Luego se calculó la participación de cada ‘SKU’ en su respectiva 'SUCURSAL' como un porcentaje del consumo total y posteriormente se seleccionaron aquellas combinaciones que representaban hasta el 80% del consumo

acumulado. Esta selección permitió centrarse en las combinaciones más relevantes para el análisis logrando que el estudio se centrara en las áreas de mayor impacto. Después de realizar este proceso, el número inicial de 2’212.369 registros del tabla se redujo a 1’240.026. Este conjunto reducido será la base de datos con la que trabajaremos y sobre la cual aplicaremos diversos modelos de aprendizaje automático.

3.3.1. Exploración y validación de datos de ventas

El barplot de la figura 3.3 muestra las ventas de diferentes categorías de productos en una empresa minorista en Ecuador. Las aves lideran con 4.31 millones de dólares en ventas, seguidas por la carne de res con 2.07 millones, y la carne de cerdo con 1.47 millones, aunque estas cifras son más bajas. El pescado tiene las ventas más bajas, con 0.18 millones de dólares. Estos datos son cruciales para decisiones estratégicas en la gestión de inventario y la oferta de productos en el sector minorista de Ecuador. Las aves son el segmento más popular y exitoso, seguido de cerca por la carne de res.

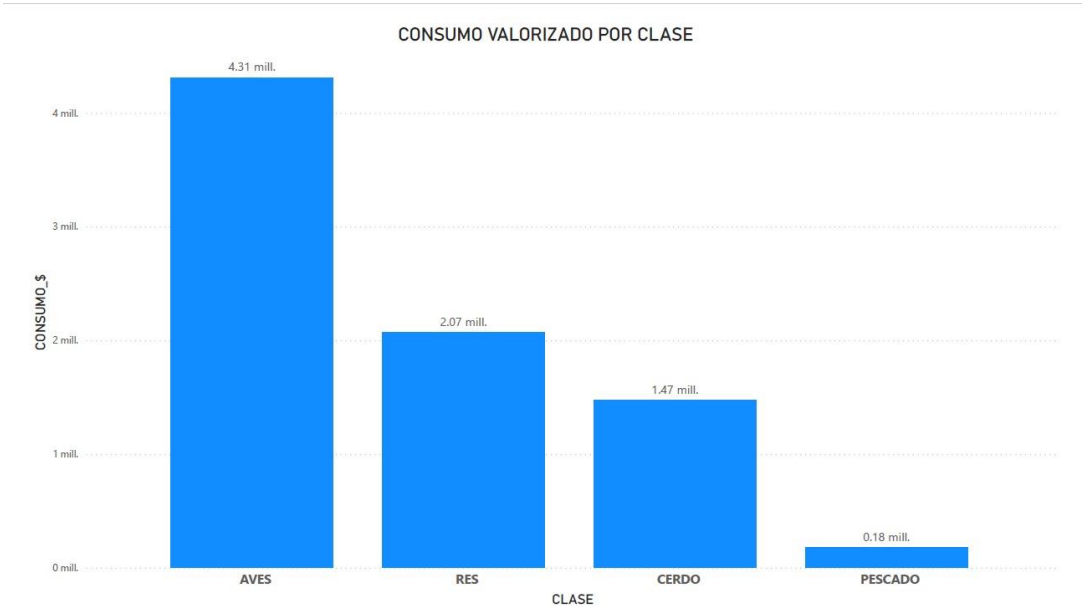


Figura 3.3 Barplot de consumo por clase
Fuente: Elaboración Propia

En la Figura 3.4, el gráfico de líneas que muestra un notable incremento en las ventas durante los últimos tres meses de cada año, mientras que en los meses

anteriores se mantiene una tendencia relativamente constante. Además, es evidente que las ventas en lo que va del año 2023 muestran un crecimiento continuo.

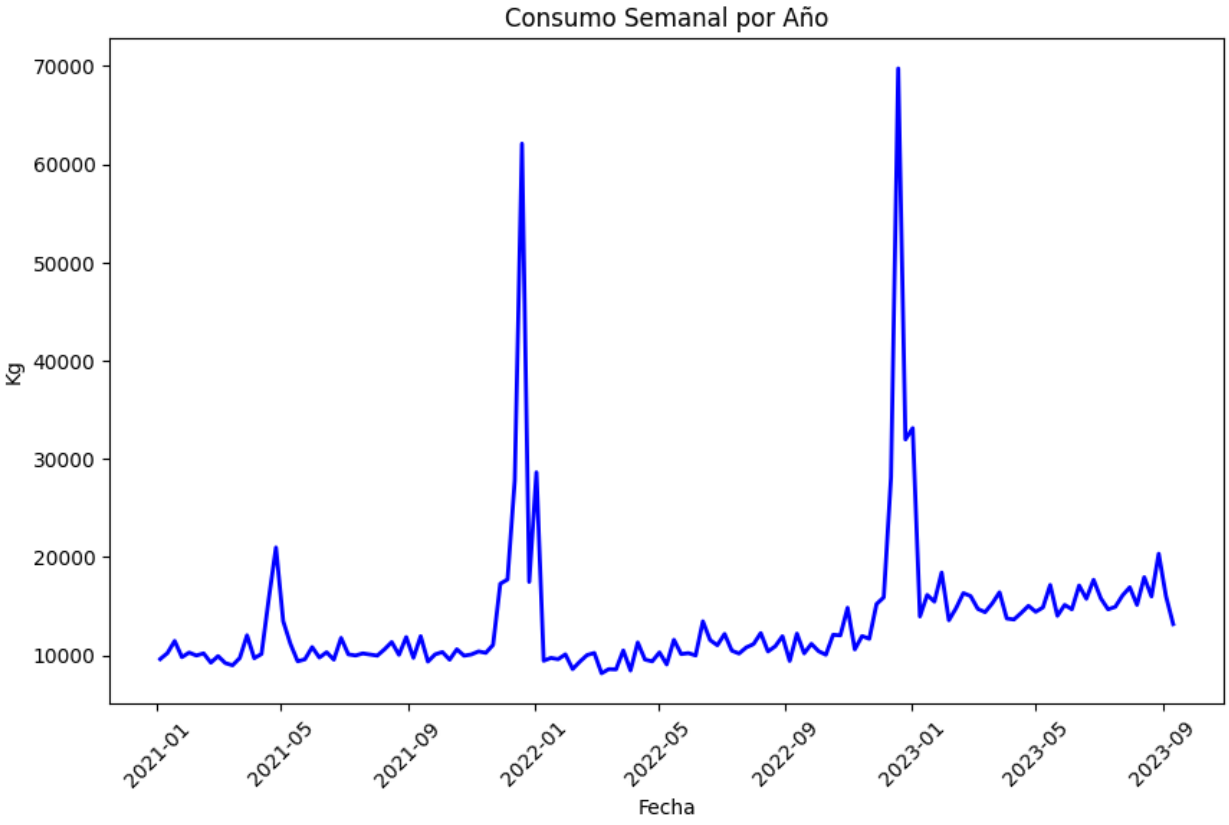


Figura 3.4 Evolución de ventas semanales
Fuente: Elaboración propia

En la Figura 3.5 se un crecimiento constante en la tendencia desde valores cercanos a 5,000 hasta superar los 15,000. Hacia el final del periodo hay un incremento drástico que supera los 18,000 sugiriendo un cambio significativo o un evento que causó un aumento en los valores medidos.

Con relación a la estacionalidad esta se mantiene dentro de un rango fijo de aproximadamente -600 a +600, indicando un patrón estacional consistente y repetitivo a lo largo del tiempo sin un cambio aparente en la magnitud o la frecuencia de estos ciclos estacionales.

Lo residuos varían alrededor de cero con algunos picos y caídas, mostrando valores que se extienden aproximadamente de -4,000 a +3,000. Estos podrían

representar el ruido o las fluctuaciones aleatorias que no son explicadas por la tendencia o la estacionalidad.

La serie muestra fluctuaciones estacionales regulares y una tendencia creciente similar al primer gráfico. Al final, hay un pico significativo que alcanza cerca de 70,000 lo que destaca un evento extraordinario (temporada navideña) o una acumulación de variaciones que no se ven en los otros componentes descompuestos.

Figura 3.5 Estacionalidad del consumo en el tiempo

Fuente: Elaboración propia

3.4. Prototipo de algoritmo y modelos

La implementación de los modelos de aprendizaje automático se llevó a cabo utilizando diversas bibliotecas de Python especializadas. Para el modelo Random Forest, se utilizó la biblioteca sklearn. En el caso de XGBoost, se empleó la biblioteca xgboost. Y, para la Regresión Lineal, se aprovechó nuevamente sklearn

(sklearn.linear_model). Estas herramientas fueron seleccionadas debido a su robustez y eficiencia en tareas similares, permitiendo una experimentación precisa y efectiva en el contexto de la ciencia de datos.

Tabla 3.3 Parámetros de procesamiento de datos

Descripción	Valor
Target	CONSUMO
Target type	Regression
Original data shape	(141, 2)
Transformed data shape	(141, 4)
Transformed train set shape	(133, 4)
Transformed test set shape	(8, 4)
Preprocess	True
Imputation type	simple
Numeric imputation	mean
Categorical imputation	mode
Fold Generator	TimeSeriesSplit
Fold Number	5
CPU Jobs	-1
Use GPU	False
Log Experiment	False
Experiment Name	reg-default-name

Fuente: Elaboración propia

- **Target:** El objetivo del análisis se llama "CONSUMO," y se trata de una tarea de regresión, lo que sugiere que el objetivo es predecir valores numéricos.
- **Original data shape:** El conjunto de datos original tenía una forma de (141, 2), lo que significa que contenía 141 filas y 2 columnas.
- **Transformed train set shape:** El conjunto de entrenamiento transformado tiene 133 filas y 4 columnas que representa un 94,32%.
- **Transformed test set shape:** El conjunto de prueba transformado tiene 8 filas y 4 columnas que representa un 5,6%.
- **Imputation type:** Se utilizó imputación valores faltantes de los datos.
- **Numeric imputation**: Los valores faltantes en las variables numéricas se imputaron utilizando la media.

- **Categorical imputation:** Los valores faltantes en las variables categóricas se imputaron utilizando la moda (el valor más común).
- **Fold Generator:** Se utilizó la técnica de división de series temporales para generar pliegues (conjuntos de datos de entrenamiento y prueba) en la validación cruzada.
- **Fold Number**: Se realizaron 5 pliegues en la validación cruzada.
- **CPU Jobs:** No se especifica el número de trabajos de CPU utilizados, probablemente se determina automáticamente.
- **Use GPU:** No se utilizó una unidad de procesamiento gráfico (GPU) para el análisis.
- **Log Experiment:** No se registraron los resultados del experimento.
- **Experiment Name:** El nombre del experimento es "reg-default-name."

3.4.1. Modelos

Para el desarrollo de la predicción de la variable "CONSUMO" basada en características como "STOCK", "PVP", "COSTO" y "PROMOCION", se optó por una variedad de modelos que incluyen a Random Forest (Bosques Aleatorios) que es un modelo basado en ensamblar múltiples árboles de decisión. También se empleó Linear Regression (Regresión Lineal) ya que es modelo matemático utilizado para predecir una variable dependiente a partir de una o más variables independientes. Y, finalmente XGBoost que es un algoritmo optimizado de aumento de gradiente que es altamente eficiente y versátil.

3.4.2. Parámetros de los modelos

A continuación, se presenta la sintaxis con sus respectivos parámetros para estructurar el algoritmo de la biblioteca scikit-learn de Python, cada parámetro es presentado y descrito en la tabla 3.2.

Tabla 3.42 Hiperparámetros del Modelo de Random forest

Parámetro	Valor	Descripción
bootstrap	False	Si se utiliza bootstrapping al construir árboles. Si es falso, se utiliza todo el conjunto de datos

ccp_alpha	0	Parámetro de complejidad utilizado para la poda mínima de costo-complejidad
criterion	absolute_error	Medida de calidad de la división
max_depth	9	La máxima profundidad del árbol. Si es None, los nodos se expanden hasta que todas las hojas sean puras
max_features	log2	Número de características a considerar al buscar la mejor división
min_impurity_decrease	0.1	Un nodo se dividirá si esta división induce una disminución de la impureza mayor o igual a este valor
min_samples_leaf	6	El número mínimo de muestras requeridas para ser un nodo hoja
min_samples_split	7	El número mínimo de muestras requeridas para dividir un nodo interno
min_weight_fraction_leaf	0	La fracción mínima ponderada de la suma total de muestras requeridas para ser un nodo hoja
n_estimators	50	El número de árboles en el bosque
n_jobs	-1	Número de trabajos para ejecutar en paralelo
oob_score	False	Si se utiliza muestras fuera de la bolsa para estimar el R^2 en datos no vistos
random_state	42	Controla la aleatoriedad para la reproducción del resultado
verbose	0	Controla la verbosidad del proceso de construcción del árbol
warm_start	False	Reutiliza la solución de la llamada anterior para ajustar y agregar más estimadores al conjunto

Fuente: Elaboración propia

El modelo de Random Forest fue configurado para manejar un equilibrio entre capacidad de aprendizaje y prevención del sobreajuste, al limitar la profundidad del árbol y configurar otros hiperparámetros como min_samples_leaf y min_samples_split. Estos hiperparámetros, cuando se ajustan adecuadamente, permiten que el modelo tenga un mejor desempeño.

Tabla 3.5 Hiperparámetros del Modelo de Regresión lineal

Parámetro	Valor	Descripción
copy_X	True	Si X se sobrescribirá o no durante la estimación
fit_intercept	True	Si se calcula el término de intercepción para este modelo
n_jobs	-1	Número de trabajos para ejecutar en paralelo
positive	False	Si se forza a los coeficientes a ser positivos

Fuente: Elaboración propia

En la tabla 3.5, la regresión lineal se utiliza como un modelo de base o punto de referencia, ya que es fácil de entender e interpretar. Los hiperparámetros seleccionados están diseñados para asegurar un cálculo eficiente y correcto.

Tabla 3.6 Hiperparámetros del Modelo de XGBoost

Parámetro	Valor	Descripción
n_jobs	-1	Número de trabajos para ejecutar en paralelo
verbosity	0	Nivel de verbosidad
learning_rate	0.05	Tasa de aprendizaje
reg_lambda	0.1	Peso para la regularización L2
subsample	0.3	Fracción de muestras que se usarán para el próximo árbol
reg_alpha	0.4	Peso para la regularización L1
colsample_bytree	0.5	Fracción de submuestreo de columnas para construir cada árbol
min_child_weight	4	Suma mínima de ponderación de instancia (hessian) necesaria en un hijo
max_depth	8	Máxima profundidad de un árbol
scale_pos_weight	17.5	Controla el equilibrio de clases positivas y negativas
random_state	42	Controla la aleatoriedad para la reproducción del resultado
n_estimators	130	Número de árboles a ajustar
tree_method	auto	Algoritmo utilizado para construir árboles
enable_categorical	False	Habilita el soporte experimental para variables categóricas directas
booster	gbtree	Tipo de booster a utilizar
objective	reg:squarederror	Función de aprendizaje a utilizar

Fuente: Elaboración propia

En la tabla 3.6, XGBoost es conocido por su eficiencia y capacidad para manejar tablas con grandes volúmenes de datos. Los hiperparámetros como max_depth, min_child_weight, y subsample ayudan a controlar el sobreajuste. La tasa de aprendizaje y la regularización (reg_alpha y reg_lambda) son cruciales para garantizar que el modelo aprenda de manera efectiva sin ser demasiado complejo. Estos hiperparámetros, en conjunto, permiten que XGBoost sea robusto y altamente adaptativo a diferentes estructuras de datos.

3.4.3. Evaluación y Validación

Tras la implementación y entrenamiento de tres modelos de aprendizaje automático distintos Random Forest (rf), Extreme Gradient Boosting (xgboost)

y Linear Regression (lr), se obtuvieron los siguientes resultados en base a diversas métricas de evaluación.

Tabla 3.7 Evaluación y validación de los modelos

Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
Random Forest Regressor	0.3899	13.699	11.622	0.5804	0.3321	0.5882	194.550
Extreme Gradient Boosting	0.4095	15.467	12.323	0.5328	0.3339	0.6403	13.320
Linear Regression	0.4319	13.145	11.400	0.5844	0.3436	0.6202	0.1670

Fuente: Elaboración propia

En la table 3.7, el modelo Random Forest tiene un RMSE de 11.62, lo que es notablemente más bajo que el RMSE de XGBoost (12.32) y Linear Regression (11.40). El RMSE es una métrica especialmente relevante ya que da más peso a errores grandes, lo que significa que Linear Regression tiene menos errores en sus predicciones en comparación con los otros modelos.

En cuanto al MAPE, el valor más bajo lo presenta Random Forest con 0.588, lo que indica que, en términos porcentuales, las predicciones de este modelo son, en promedio, las más cercanas al valor real.

Dadas todas estas métricas, podemos concluir que el modelo Random Forest Regressor es el más robusto y consistente en general. Aunque XGBoost tiene un MAE ligeramente más bajo, Random Forest supera a los otros modelos en términos de RMSE y MAPE, lo que lo convierte en la mejor opción para este conjunto de datos en particular.

En la figura 3.6, ilustra la eficacia de nuestro modelo para prever el safety stock en una empresa del sector supermercado en Ecuador. Con un coeficiente de determinación (R^2) de 0.945, el modelo muestra un alto grado de precisión en sus predicciones. El "best fit" del modelo, destacado mediante la función de identidad, revela una alineación cercana entre las predicciones y los valores reales de safety stock. Estos resultados respaldan la validez y utilidad del modelo propuesto, sugiriendo que puede ser una herramienta eficaz para optimizar el nivel de inventario y garantizar la seguridad en el

suministro de productos en el contexto específico del sector supermercado en Ecuador.

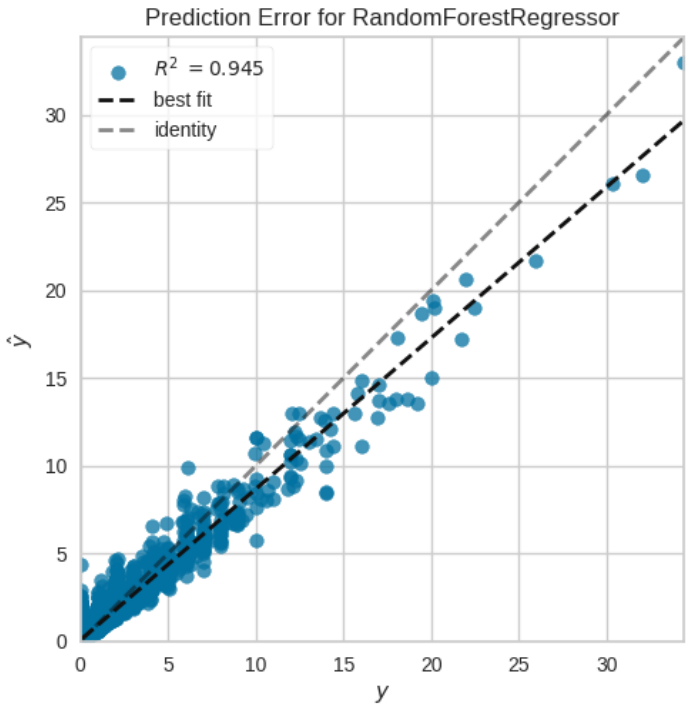


Figura 3.6 Predicción de error para Random Forest Regressor

Fuente: Elaboración propia

Este análisis de residuales respalda la solidez de nuestro modelo Random Forest Regressor para prever el safety stock en el sector supermercado de Ecuador. En el grafico 3.7, Con coeficientes de determinación (R^2) notables de 0.942 en el conjunto de entrenamiento y 0.945 en el conjunto de prueba, el modelo demuestra una capacidad para explicar la variabilidad en los datos. El examen de residuales revela que las discrepancias entre las predicciones y los valores reales son mínimas y no muestran patrones sistemáticos. La homocedasticidad y la distribución de los residuales sugieren que el modelo generaliza de manera efectiva, manteniendo una constante precisión en diferentes niveles de la variable de respuesta. Estos hallazgos respaldan la confianza en la capacidad del modelo para realizar predicciones precisas y su capacidad para generalizar a datos no vistos.

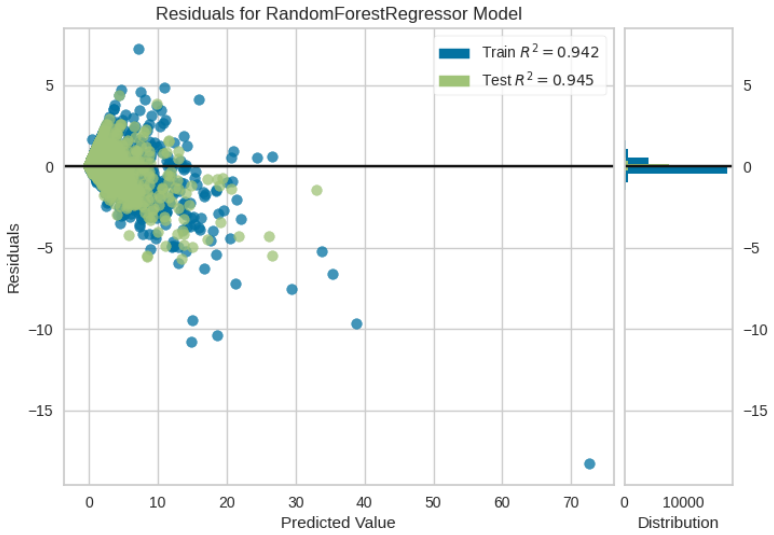


Figura 3.7 Residuales para Random Forest Regressor

Fuente: Elaboración propia

3.4.4. Aplicación del Safety Stock en base al RMSE

Dado que el RMSE (Root Mean Squared Error) es una medida del error entre los valores pronosticados y los valores observados, se puede utilizar para cuantificar la variabilidad o incertidumbre en las predicciones del modelo. En otras palabras, el RMSE nos da una idea de cuánto podríamos esperar que varíen nuestras predicciones del valor real.

La fórmula para el cálculo del safety stock es:

$$\mathrm{SS} = \left(\mathrm{RMSE} \times z \times \sqrt{n}\right) \times \left(\frac{3}{7}\right)$$

Donde:

RMSE: es el error cuadrático medio raíz obtenido del modelo, representando la desviación estándar de las predicciones.

z: Es el factor de nivel de servicio, que se refiere a la probabilidad de no quedarse sin stock. Por ejemplo, un valor de z de 1.96 se correspondería con un nivel de servicio del 95%.

n: Es el tiempo de suministro en días.

El factor **“3/7”** es una constante que ajusta el safety stock basado en la proporción de días de la semana que se quiere cubrir.

Una vez calculado el safety stock, este valor se sumará al inventario base para determinar el nivel óptimo de inventario que debe mantenerse en reserva. Esto garantiza que, incluso en caso de variaciones inesperadas en la demanda o demoras en el suministro, hay suficiente inventario en reserva para satisfacer la demanda durante el tiempo que se necesita reabastecer.

3.5. Infraestructura para procesamiento y almacenamiento

3.5.1. Arquitectura del sistema

En la figura 3.8, representa un flujo de trabajo integral para el desarrollo de una aplicación de predicción que involucra múltiples etapas esenciales:

1. **Ingesta y Preparación de Datos:** Esta fase marca el punto de partida del proceso, donde se exploran y preparan los datos de entrada. Se destaca el uso de Python para la exploración de datos, lo que sugiere un enfoque detallado en la comprensión de los datos antes de proceder con el modelo.
2. **Entrenamiento y Ejecución del Modelo:** En esta fase, se realiza el núcleo del proceso de aprendizaje automático. Python y Jupyter son las herramientas elegidas para desarrollar y entrenar el modelo. Esto implica la selección de algoritmos, la configuración de hiperparámetros y la ejecución del modelo para realizar predicciones.
3. **Almacenamiento:** Después de obtener las predicciones del modelo, se destaca la importancia de almacenar estos datos de manera efectiva. SQL Server es la plataforma de almacenamiento designada, lo que garantiza la persistencia y la disponibilidad de los resultados del modelo.
4. **Visualización de Datos:** La última fase se centra en la presentación efectiva de los resultados al usuario final. Power BI es la tecnología utilizada para desarrollar la interfaz de usuario y crear visualizaciones de datos comprensibles que permitan a los usuarios interpretar y aprovechar las predicciones del modelo.

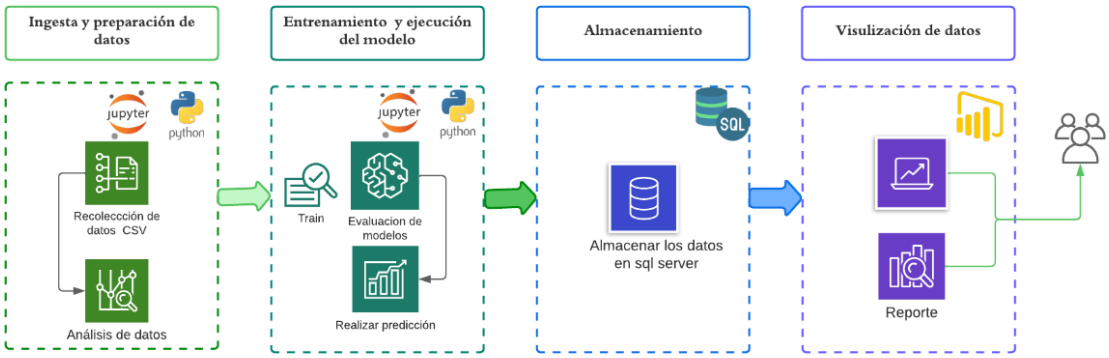


Figura 3.8: Arquitectura de Aplicación de Predicción

Fuente: Elaboración propia

En la figura 3.9, El diagrama de flujo representa el proceso del modelo de optimización de safety stock basado en aprendizaje automático supervisado para una empresa del sector supermercado en Ecuador. Comienza con la entrada de etiquetas, donde se ingresan los datos etiquetados relacionados con el safety stock. Luego, los datos de entrenamiento y prueba son procesados para prepararlos adecuadamente. A continuación, se lleva a cabo el entrenamiento de tres algoritmos diferentes: XGBoost, Random Forest y Regresión Lineal, utilizando el conjunto de datos de entrenamiento.

Una vez que los modelos están entrenados, se procede a la evaluación de su rendimiento mediante métricas clave como el RMSE, MAPE, MAE y R^2. Con base en estas métricas, se selecciona el modelo final que mejor se ajusta a los criterios de evaluación establecidos.

Después de elegir el modelo final, se generan predicciones y posteriormente el safety stock. Estas predicciones se utilizan para proporcionar recomendaciones específicas de safety stock para cada combinación SUCURSAL y SKU

Finalmente, el proceso concluye con la generación de recomendaciones y la salida de etiquetas de safety stock.

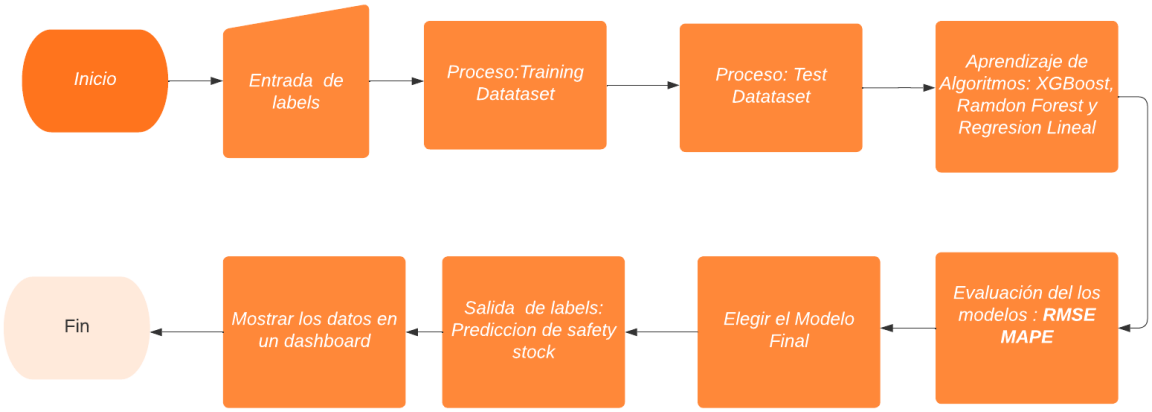


Figura 3.9: Diagrama de flujo del sistema
Fuente: Elaboración propia

3.6. Plataforma y prototipo de visualización

En la figura 3.10, se han desarrollado dos componentes clave para la gestión eficiente del inventario en la cadena de supermercado en Ecuador. El primero consiste en un gráfico de línea de tiempo que abarca los años 2021 a 2023, donde el eje “x” representa los años y el eje “y” muestra el consumo de productos, resaltando picos de ventas al finalizar el año. Estos picos pueden relacionarse con eventos estacionales o estrategias exitosas. El segundo componente presenta un listado de los 10 productos más rentables en el negocio, calculados considerando ingresos y costos asociados. Estos productos son prioritarios debido a su contribución significativa al margen de beneficio de la empresa. Ambos elementos son esenciales para respaldar decisiones estratégicas y la gestión óptima del safety stock.

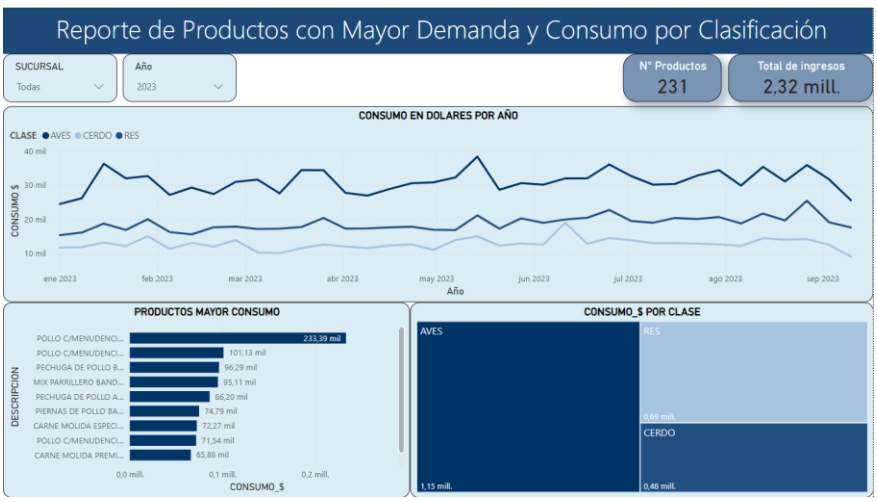


Figura 3.10: Visualización de los datos
Fuente: Elaboración propia

La figura 3.11 ilustra una comparativa temporal entre el consumo real de la serie (denominado "Consumo") y el pronóstico de dicho consumo (denominado "Forecast"). La representación visual de ambas series de tiempo permite identificar con claridad la precisión y eficacia del modelo predictivo empleado.

Desde una perspectiva general, la proximidad entre las dos series a lo largo del período de tiempo evaluado es indicativa de la capacidad del modelo para capturar con precisión la tendencia subyacente del consumo.

Si las líneas "Consumo" y "Forecast" se mueven juntas, significa que las predicciones están alineadas el consumo real y es un indicador de que el modelo está captando correctamente las tendencias de consumo.

Ahora, si existen fuertes desviaciones entre "Consumo" y "Forecast" en ciertos puntos, esto puede indicar momentos en los que nuestras predicciones no coinciden con el consumo real. Estas desviaciones pueden deberse a eventos inesperados o factores que el modelo no pudo considerar.

Cualquier pico o caída significativa en "Consumo" que no se vea reflejada en "Forecast" sugiere eventos o situaciones especiales que afectaron el consumo y que no estaban previstos por el modelo.

Basado en lo expuesto con anterioridad se puede afirmar que esta visualización es útil para mostrar la precisión de las predicciones e identificar áreas donde podríamos necesitar ajustar el enfoque predictivo del abastecimiento mediante la incorporación del safety stock.

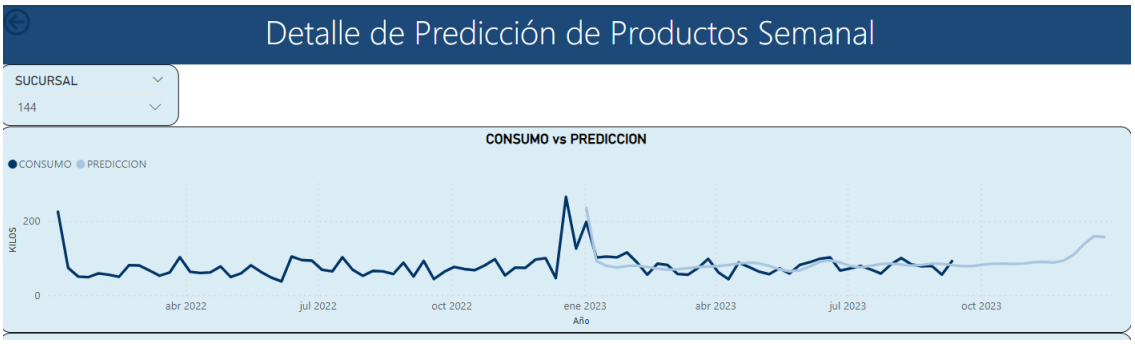
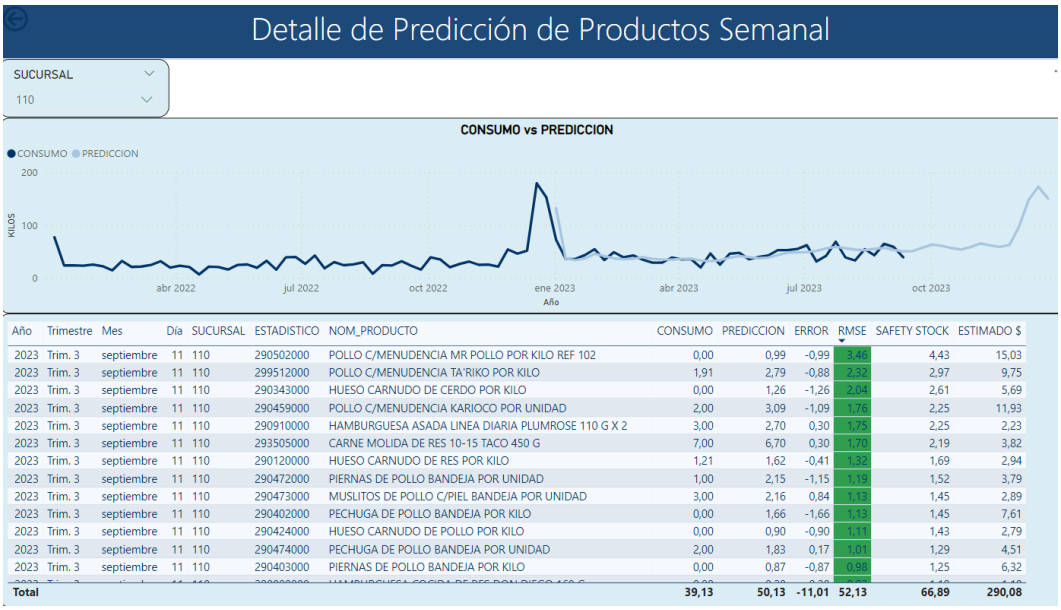


Figura 3.11: Comparativo consumo vs forcecast
Fuente: Elaboración propia

La tabla de la figura 3.12 proporciona una visión detallada de las variables que generan el abastecimiento para cada SKU en diferentes sucursales. Con esta información, es posible:

1. **Identificar SKUs de alto movimiento:** Estos son productos que tienen una alta demanda y, por lo tanto, requieren una gestión cuidadosa del inventario para evitar desabastecimiento.
2. **Revisar niveles de SS:** Si un SKU tiene un SS elevado, puede ser necesario revisar las estrategias de abastecimiento o las predicciones para ese producto en particular.
3. **Optimizar el inventario:** Al comparar el consumo real con el forecast y el SS, es posible tomar decisiones sobre qué productos reponer, en qué cantidad y en qué momento, asegurando así una gestión eficiente del inventario.



Año	Trimestre	Mes	Día	SUCURSAL	ESTADISTICO	NOM_PRODUCTO	CONSUMO	PREDICCION	ERROR	RMSE	SAFETY STOCK	ESTIMADO $
2023	Trim. 3	septiembre	11	110	290502000	POLLO C/MENUDENCIA MR POLLO POR KILO REF 102	0,00	0,99	-0,99	3,46	4,43	15,03
2023	Trim. 3	septiembre	11	110	299512000	POLLO C/MENUDENCIA TA'RIKO POR KILO	1,91	2,79	-0,88	2,32	2,97	9,75
2023	Trim. 3	septiembre	11	110	290343000	HUESO CARNUDO DE CERDO POR KILO	0,00	1,26	-1,26	2,04	2,61	5,69
2023	Trim. 3	septiembre	11	110	290459000	POLLO C/MENUDENCIA KARIOCO POR UNIDAD	2,00	3,09	-1,09	1,76	2,25	11,93
2023	Trim. 3	septiembre	11	110	290910000	HAMBURGUESA ASADA LINEA DIARIA PLUMROSE 110 G X 2	3,00	2,70	0,30	1,75	2,25	2,23
2023	Trim. 3	septiembre	11	110	293505000	CARNE MOLIDA DE RES 10-15 TACO 450 G	7,00	6,70	0,30	1,70	2,19	3,82
2023	Trim. 3	septiembre	11	110	290120000	HUESO CARNUDO DE RES POR KILO	1,21	1,62	-0,41	1,32	1,69	2,94
2023	Trim. 3	septiembre	11	110	290472000	PIERNAS DE POLLO BANDEJA POR UNIDAD	1,00	2,15	-1,15	1,19	1,52	3,79
2023	Trim. 3	septiembre	11	110	290473000	MUSLITOS DE POLLO C/PIEL BANDEJA POR UNIDAD	3,00	2,16	0,84	1,13	1,45	2,89
2023	Trim. 3	septiembre	11	110	290402000	PECHUGA DE POLLO BANDEJA POR KILO	0,00	1,66	-1,66	1,13	1,45	7,61
2023	Trim. 3	septiembre	11	110	290424000	HUESO CARNUDO DE POLLO POR KILO	0,00	0,90	-0,90	1,11	1,43	2,79
2023	Trim. 3	septiembre	11	110	290474000	PECHUGA DE POLLO BANDEJA POR UNIDAD	2,00	1,83	0,17	1,01	1,29	4,51
2023	Trim. 3	septiembre	11	110	290403000	PIERNAS DE POLLO BANDEJA POR KILO	0,00	0,87	-0,87	0,98	1,25	6,32
Total							**39,13**	**50,13**	**-11,01**	**52,13**	**66,89**	**290,08**

Figura 3.12: Dashboard Semanal
Fuente: Elaboración propia

La visualización de la figura 3.13 destaca un aspecto crucial de la gestión de inventario que es el costo de oportunidad asociado a las ventas perdidas debido a niveles bajos de stock. Esta perspectiva no sólo cuantifica las ventas perdidas semanales en términos de unidades sino también en términos de ingresos potencialmente perdidos, lo que refleja directamente en la salud financiera y la rentabilidad de la empresa.

La tabla proporciona una desagregación detallada de las ventas y los ingresos perdidos para cada SKU en diferentes sucursales. Usando esta tabla es posible:

1. **Priorizar Reposiciones:** Si un SKU tiene ingresos perdidos significativos, podría ser prioritario abastecer ese producto en particular para evitar futuras pérdidas.
2. **Estrategias de Abastecimiento:** Los SKUs con altas ventas perdidas recurrentes pueden requerir una revisión de las estrategias de abastecimiento y parámetros de los diversos modelos de abastecimiento en la empresa.

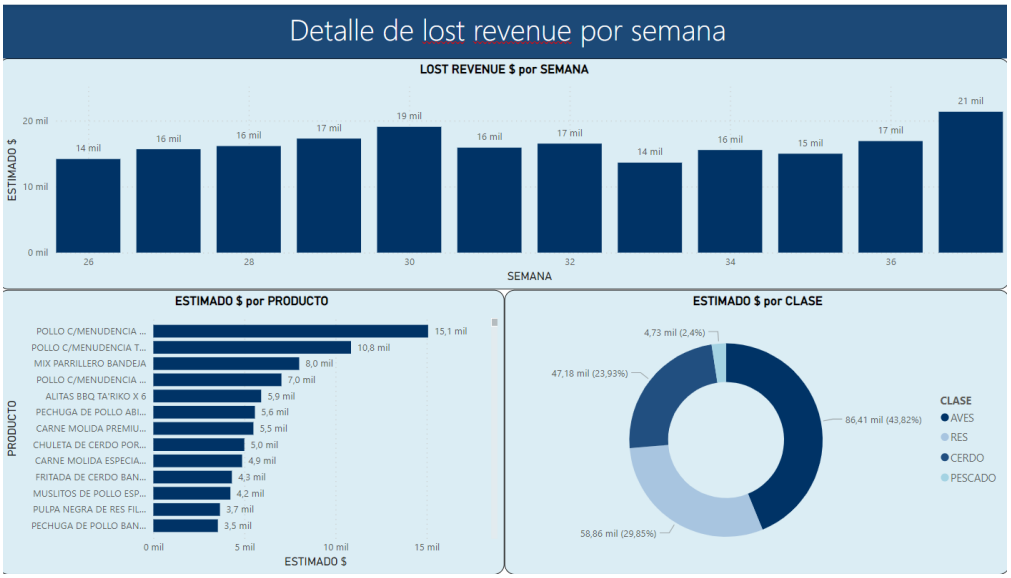


Figura 3.13: Lost Revenue
Fuente: Elaboración propia

3.7. Métricas y comunicación de resultados

En la tabla 3.6, se muestra datos detallados sobre el consumo real y los pronósticos asociados para varios períodos específicos Cada fila de la tabla corresponde a un período determinado, con la fecha indicada. La columna "Consumo" refleja las cifras reales del consumo para cada período, proporcionando una visión de la demanda real en la empresa. Por otro lado, la columna "Forecast" muestra los valores pronosticados o estimados para el mismo período, indicando las predicciones anticipadas para el consumo.

Esta tabla 3.8, es esencial para evaluar la precisión de los pronósticos en comparación con los valores reales de consumo. Facilita la identificación de posibles variaciones entre las predicciones y la realidad, lo que puede ser crucial para el perfeccionamiento y ajuste continuo de los modelos de pronóstico utilizados en la optimización del safety stock. La comparación sistemática de estas cifras a lo largo del tiempo puede ayudar a mejorar la eficacia del modelo y proporcionar información valiosa para la toma de decisiones relacionadas con la gestión de inventarios y la planificación estratégica.

Tabla 3.8 Comparación entre Consumo Real y Pronóstico por Período

PERIODO	CONSUMO	FORECAST
26/06/2023	17,678	14,536
03/07/2023	15,786	14,088
10/07/2023	14,671	14,234
17/07/2023	14,924	14,423
24/07/2023	16,083	14,062
31/07/2023	16,913	13,700
07/08/2023	15,124	14,060
14/08/2023	17,943	14,151
21/08/2023	15,987	13,933
28/08/2023	20,335	13,685
04/09/2023	16,021	13,998
11/09/2023	13,162	14,133

Fuente: Elaboración propia

En la figura 3.14, se muestra las variables el consumo real de productos perecederos a lo largo del año 2023 y la predicción del consumo por semana representa la estimación de cómo se espera que se comporte el consumo a lo largo del año 2023, como ambas líneas son muy similares y se superponen en gran medida, esto indica que el modelo de predicción ha tenido un buen desempeño al estimar el consumo real.

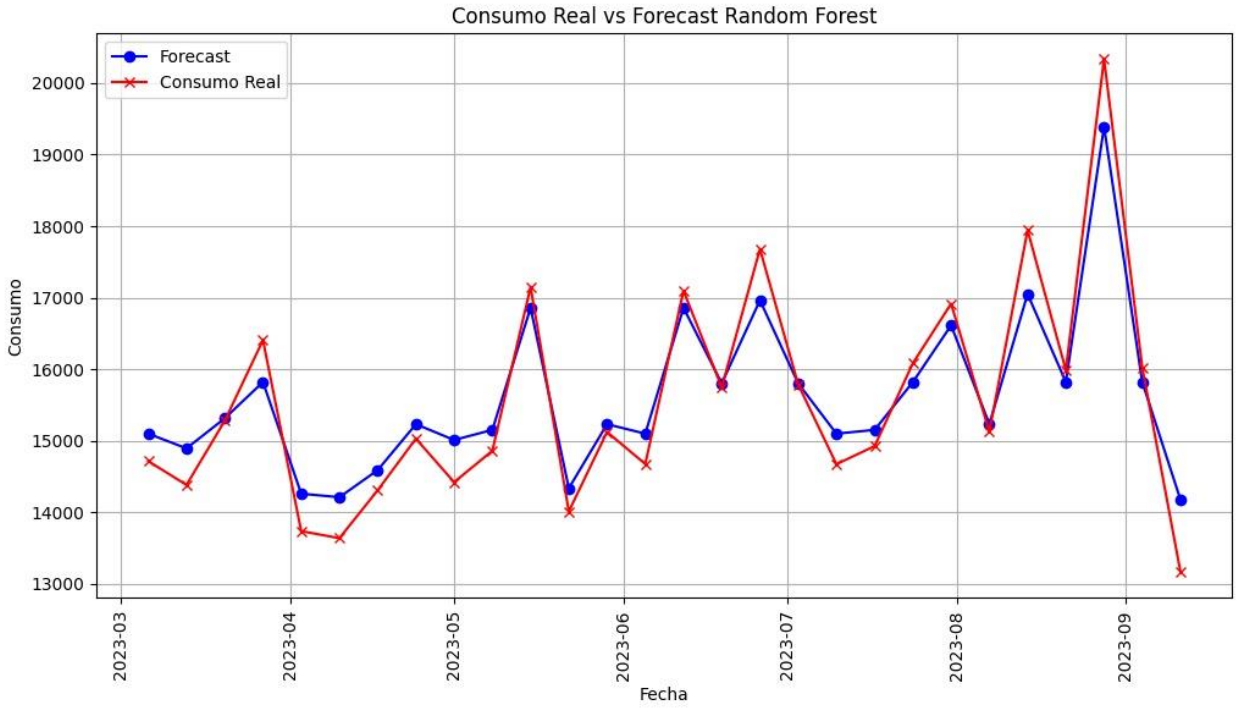


Figura 3.14 Comparación del consumo vs forcecast

Fuente: Elaboración propia

CAPÍTULO 4

4. ANÁLISIS DE RESULTADOS

4.1. Recolección de datos y estrategia para visualización del proyecto

4.1.1. Recolección de datos

Los datos que alimentan el proyecto provienen de múltiples fuentes internas de la organización, específicamente de sistemas de gestión empresarial y sistemas de punto de venta. Estos sistemas recopilan y almacenan información relacionada con las ventas, inventario, promociones y otros aspectos relevantes de la operación diaria.

Para la extracción de estos datos, se establecieron conexiones seguras y automatizadas que permiten el acceso a bases de datos específicas. A través de herramientas de ETL (Extract, Transform, Load), se extrajo la información relevante, la cual se sometió a procesos de limpieza y transformación, adecuándola para su posterior análisis y modelado. Las extracciones se programaron para ser realizadas periódicamente, garantizando así que el modelo siempre contara con la información más actualizada.

Figura 4.1: Diagrama de recolección de datos

Fuente: Elaboración propia

4.1.2. Calidad, integridad y relevancia de los datos recopilados

Uno de los desafíos iniciales fue garantizar la calidad e integridad de los datos extraídos. Durante las primeras etapas del proyecto, se identificaron ciertas inconsistencias y valores faltantes en los registros. Estos problemas se abordaron mediante técnicas de imputación y corrección, basadas en reglas de negocio y criterios estadísticos.

La relevancia de los datos es crítica para el éxito del proyecto. Durante la fase de exploración, se llevó a cabo un análisis detallado para identificar qué variables y registros eran cruciales para el modelado predictivo. Variables como "STOCK", "PVP", "COSTO" y "PROMOCION" se destacaron como esenciales debido a su directa correlación con el "CONSUMO", la variable objetivo.

Es vital destacar que la calidad de los datos es fundamental para obtener modelos predictivos precisos. Si bien se han realizado esfuerzos significativos para asegurar la integridad y relevancia de los datos, es una tarea continua. El monitoreo regular y la validación de los datos son esenciales para mantener la eficacia y precisión del sistema en el tiempo.

4.1.3. Validación de los datos

Para garantizar la precisión y la integridad de los datos, se implementó un proceso de validación multidisciplinario que utiliza tanto técnicas manuales como automatizadas. Una de las primeras herramientas empleadas fue el análisis exploratorio de datos (EDA) a través de paquetes de software como Pandas y Seaborn. Esto permitió visualizar distribuciones, identificar valores atípicos y detectar posibles anomalías en los datos.

Además, se desarrollaron scripts automatizados que realizan verificaciones de consistencia, como la detección de valores faltantes, la validación de rangos para campos numéricos y la comprobación de formatos para datos categóricos. Estos scripts se ejecutan periódicamente, asegurando que cualquier anomalía sea detectada a tiempo.

También se empleó la técnica de validación cruzada durante el modelado, lo que ayudó a identificar si los datos presentaban problemas de sobreajuste o subajuste, garantizando así la robustez del modelo frente a datos no vistos.

4.1.4. Desafíos encontrados

Durante el proceso de validación, se encontraron varios desafíos. Uno de los más comunes fue la presencia de valores faltantes en ciertas columnas. En lugar de eliminar estos registros, se emplearon técnicas de imputación basadas en la mediana.

Además, se detectaron ciertas inconsistencias en datos categóricos, donde diferentes terminologías se utilizaban para representar la misma categoría. Para abordar esto, se creó un mapa de terminologías estandarizadas y se realizó una limpieza y transformación exhaustiva.

Otro desafío fue la presencia de valores atípicos, especialmente en las variables relacionadas con el consumo. Estos se trataron mediante técnicas de suavizado, asegurando que no afectaran de manera desproporcionada los modelos predictivos.

A pesar de estos desafíos, el proceso de validación y limpieza permitió mejorar significativamente la calidad del tabla, proporcionando una base sólida para el análisis y modelado posterior.

4.2. Puesta en marcha y funcionamiento

4.2.1. Despliegue de Modelos Predictivos

Los modelos predictivos, tras una minuciosa fase de entrenamiento y validación, fueron implementados utilizando Python, un lenguaje de programación conocido por su robustez en tareas de ciencia de datos. Se recurrió a bibliotecas especializadas como Scikit-learn para modelos convencionales y XGBoost para el modelo de boosting.

Una vez que los modelos estuvieron entrenados y ajustados se procedió a generar predicciones sobre el conjunto de datos. Estas predicciones se almacenaron en archivos CSV para garantizar la portabilidad y facilidad de uso. Luego, estos archivos CSV se importaron a una base de datos SQL. La elección de SQL se basó en su confiabilidad, eficiencia y compatibilidad con diversas herramientas de visualización y análisis.

4.2.2. Optimización realizada después de la implementación inicial

Posterior al despliegue inicial, y basándose en las necesidades de integración con Power BI, se realizó una estructuración específica de los datos en la base de datos SQL. Esto garantizó que la conexión entre SQL y Power BI fuera fluida y que los datos estuvieran dispuestos de una manera que facilitara la creación de visualizaciones y paneles en Power BI.

4.2.3. Interacción con Power BI

Una vez que las predicciones de los modelos se almacenaron en la base de datos SQL, se estableció una conexión directa con Power BI. Esta herramienta de visualización permite a los usuarios interactuar con los datos, visualizar las predicciones y compararlos con los valores reales, ofreciendo así insights valiosos sobre el rendimiento del modelo y las tendencias del consumo.

4.3. Pruebas de funcionalidad

Las pruebas de funcionalidad son esenciales para asegurarse de que el sistema funciona según lo esperado y cumple con las necesidades propuestas. Estas pruebas se realizaron en diferentes etapas del desarrollo, considerando tanto la precisión de las predicciones de los modelos como la interacción y presentación de resultados en Power BI.

4.3.1. Resultados de las Pruebas

Durante la fase de pruebas, se evaluaron varios aspectos, desde la capacidad de los modelos para predecir el consumo con precisión hasta la facilidad de uso e interacción del dashboard en Power BI.

Las predicciones de los modelos se compararon con los valores reales, utilizando métricas estándar como el RMSE. Estas métricas proporcionaron una visión clara de la precisión y fiabilidad de cada modelo.

Se llevó a cabo una evaluación exhaustiva de la interfaz, garantizando que las visualizaciones fueran claras, intuitivas y proporcionaran insights valiosos para los usuarios.

4.3.2. Presentación de los resultados claves al usuario

Los modelos mostraron una precisión prometedora en las predicciones, lo que indica una selección adecuada de características y un buen ajuste de hiperparámetros. Referente a los paneles diseñados en Power BI fueron bien recibidos, con visualizaciones claras que ayudan a los usuarios a interpretar rápidamente los datos.

Aunque los modelos tuvieron un buen desempeño general, ciertas situaciones específicas o eventos atípicos no fueron capturados adecuadamente, lo que sugiere la necesidad de futuras optimizaciones o la inclusión de más variables explicativas.

Algunos usuarios sugirieron personalizaciones adicionales o características específicas en el dashboard de Power BI para adaptarse mejor a sus flujos de trabajo individuales.

El feedback de los usuarios fue invaluable durante el proceso de prueba. Los comentarios sobre la precisión de las predicciones ayudaron a identificar áreas donde los modelos podrían necesitar ajustes o mejoras adicionales.

Por otro lado, las recomendaciones relacionadas con la interfaz de Power BI llevaron a la incorporación de nuevas visualizaciones, así como a ajustes en la presentación de ciertas métricas para garantizar una interpretación más intuitiva.

Este feedback directo de los usuarios finales garantizó que el sistema no solo fuera técnicamente sólido, sino también alineado con las necesidades y expectativas del usuario, maximizando su utilidad y eficacia.

4.4. Análisis costo/beneficio

En esta sección, se realizará un análisis costo/beneficio para evaluar la implementación del modelo de optimización del safety stock en la empresa de supermercado en Ecuador, expresado en términos de ganancias en dólares.

4.4.1. Costos de Implementación

Los costos asociados con la implementación del modelo incluyen:

Desarrollo del Modelo: La inversión en el desarrollo del modelo, incluyendo la contratación de expertos en aprendizaje automático y la adquisición de software especializado, ascendió a $ 64,800.

Tabla 4.1 Costo de desarrollo

Costo del desarrollo	Valor
Programador Python	$ 21,600
Programador SQL	$ 21,600
Programador Power BI	$ 15,000
Costo total	**$ 64,800**

Fuente: Elaboración propia

Capacitación: Se destinaron $10,000 para la capacitación del personal de la empresa en el uso del modelo y las herramientas asociadas.

Mantenimiento y licencias: Se invirtió $ 56,100 en la actualización de la infraestructura tecnológica necesaria para ejecutar el modelo de manera eficiente.

Costo Total de Implementación: $ 94,800.

Tabla 4.2 Costo de solución

Costo de la Solución	Valor
Costo del Desarrollo	$ 64,800
Capacitación	$ 10,000
Mantenimientos y Licencias	$ 56,100
Costo total	**$ 130,900**

Fuente: Elaboración propia

4.4.2. Beneficios Esperados

Los beneficios esperados de la implementación del modelo se estiman en términos de ganancias adicionales en dólares:

1. Se estima que la optimización del stock debido a un abastecimiento mucho más preciso en un 5% de la venta anual lo que representa una reducción de los costos de inventario por obsolescencia de $ 116,156 al año.
2. Se espera un aumento de las ventas de $ 197,166 por trimestre lo que representa alrededor de $ 788,664 anuales debido a la disponibilidad de productos en sucursales.

Tabla 4.2 Costo de solución

Beneficio Esperado	Monto Anual
Mejora en la gestión del inventario	$ 116,156
Aumento de las ventas	$ 788,664
Beneficio Total Anual	**$ 904,820**

Fuente: Elaboración propia

El análisis financiero presentado revela el impacto significativo y tangible que la implementación del sistema de predicción tiene sobre la gestión

empresarial. Con una mejora estimada en la gestión del inventario que asciende a $116,156 anuales y un aumento en las ventas de $788,664, el valor acumulado de estos beneficios refleja un beneficio total anual de $904,820.

Estos resultados reafirman la importancia de invertir en tecnologías de análisis de datos y aprendizaje automático para impulsar la toma de decisiones informadas y maximizar los beneficios operativos y financieros.

4.4.3. Evaluación del Costo/Beneficio

Para evaluar el costo/beneficio en términos de ganancias en dólares, calcularemos el Retorno de la Inversión (ROI) esperado en función de los costos de implementación y los beneficios esperados:

$$ROI = \frac{(Beneficio\ Neto\ -Costo\ Implementación\ )}{Costo\ Implementación} x100\%$$

$$ROI = \frac{(\$\ 904{,}820\ -\$\ 130{,}900)}{\$\ 130{,}900} * 100 = 591\%$$

4.4.4. Resultados Esperados

Basado en el análisis costo/beneficio, se espera que la implementación del modelo de optimización del safety stock genere un ROI del 591 %. Esto indica que, por cada dólar invertido en la implementación, se espera un retorno de $ 5.91 en términos de ganancias adicionales y reducción de costos al año.

5. CONCLUSIONES

- El problema de la gestión de inventario en una cadena de supermercado es un desafío crítico que requiere un equilibrio preciso entre evitar faltantes y reducir costos asociados al exceso de stock.
- La implementación de un modelo de predicción del safety stock utilizando técnicas de aprendizaje automático supervisado ha demostrado ser eficaz para abordar este desafío.
- Se identificó que el modelo Random Forest Regressor es la opción más sólida y consistente en términos de métricas de rendimiento, como RMSE y MAPE, lo que lo convierte en la elección óptima para este conjunto de datos.
- La predicción del safety stock de productos perecibles de mayor importancia en función de su rentabilidad y rotación es fundamental para asegurar una gestión eficiente del inventario en la cadena de supermercado.

6. RECOMENDACIONES

- Continuar utilizando el modelo Random Forest Regressor como la principal herramienta para la predicción del safety stock en la cadena de supermercado. Este modelo ha demostrado un rendimiento sólido y consistente, lo que contribuye a una gestión de inventario más precisa y eficiente.
- Realizar un seguimiento constante y actualizaciones periódicas del modelo de aprendizaje automático para adaptarse a las fluctuaciones en la demanda y cambios en las dinámicas de la cadena de supermercado. La evolución de los patrones de compra y las tendencias del mercado deben reflejarse en el modelo.
- Implementar un sistema de integración que permita a los usuarios examinar y confirmar las sugerencias de safety stock generadas por el modelo. Esto asegurará una mayor colaboración entre el equipo de gestión y el sistema de predicción.
- Realizar un monitoreo continuo de las combinaciones de sucursales y estadísticas que requieren ajustes en el safety stock. Las condiciones pueden variar en diferentes ubicaciones y momentos, por lo que es esencial adaptar el modelo en consecuencia.

- Considerar la posibilidad de implementar un sistema de alerta temprana que informe a los responsables de inventario sobre desviaciones significativas en la demanda o en los niveles de stock. Esto permitirá una acción rápida y eficaz en caso de cambios inesperados en las condiciones del mercado.

BIBLIOGRAFÍA

Abdullah Al Imran, Zaman Wahid, Alpana Akhi Prova and Md. Hannan. (2022). Harnessing the meteorological effect for predicting the retail price of rice in Bangladesh. *Department of Computer Science, Islamic University of Technology*.

Carrillo, D. P. (2019). Repositorio de la Universidad Santiago de Compostela. Retrieved from Universidad Santiago de Compostela. Obtenido de http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_1680.pdf

David Diaz, R. M. (2021). A Machine Learning Approach for Modeling Safety Stock Optimization. *Facultad de Ingenieria, Universidad Panamericana, Ciudad de Mexico*, 14.

Gujarati, D. N. (2015). *Econometría (5ta ed.).* McGraw-Hill.

Levin, R. I. (2004). *Estadística para Administración y Economía (7ma ed.).* Pearson Educación.

Punia, S. (2020). Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail. *International journal of production research*, 58(16).

Rachit Srivastava, A. T. (2019). Solar radiation forecasting using MARS, CART, M5, and random forest. *ELSEVIER*.

Vincenzo Bianco, O. M. (2009). Electricity consumption forecasting in Italy using linear regression models. *ELSEVIER*.

Yang, T. (2023). Sales Prediction of Walmart Sales Based on OLS, Random. *Research Gate*.

Yetunde Faith, J. I. (2022). Application of XGBoost Algorithm for Sales Forecasting Using Walmart Dataset. *Springer*.

Yi, S. (2023). Walmart Sales Prediction Based on Machine Learning. *Research Gate*.

GLOSARIO

- **Outliers**: Observaciones atípicas o excepcionales en un conjunto de datos, que se desvían significativamente del comportamiento general de la muestra.

- **Hiperparámetros**: Variables externas que influyen en el comportamiento de un modelo de aprendizaje automático y que deben ser ajustadas manualmente antes de iniciar el proceso de entrenamiento del modelo.

- **Tabla**: Una estructura de datos tabular bidimensional que organiza los datos en filas y columnas, similar a una tabla de una base de datos o una hoja de cálculo.

- **SKU (Stock Keeping Unit):** Un código único asignado a cada producto para facilitar la gestión y seguimiento de inventarios. Cada SKU representa un artículo específico con sus propias características y detalles.

- **RMSE (Root Mean Squared Error):** Una métrica que cuantifica la precisión de un modelo al calcular la raíz cuadrada del promedio de los cuadrados de las diferencias entre los valores predichos y los valores reales.

- **Safety Stock:** Un nivel adicional de inventario que se mantiene para evitar agotamientos inesperados debido a variaciones en la demanda o en el suministro. Este stock de seguridad actúa como un amortiguador para asegurar que haya suficiente inventario disponible incluso en condiciones imprevistas.