Proyecto Capstone de Data Science

TÍTULO TENTATIVO:

"Aplicación de ciencia de datos para la optimización de decisiones estratégicas en SOREUS MOTORS E.I.R.L. mediante análisis predictivo del comportamiento comercial en el 2024"

CARRERA:

DATA SCIENCE CAPSTONE PROJECT

ELABORADO POR:

García Girón, David Adolfo

DOCENTE:

Walter Alejandro Silva Sotillo

ENLACE DE GITHUB:

https://github.com/Sagiiiii/Analsis-Soreus-Motors-CAPSTOM-PROJECT

LIMA – PERÚ

2025

ÍNDICE DE CONTENIDOS

ÍNDICE DE TABLAS Y FIGURAS

RESUMEN

Durante el periodo 2023 - 2024, la empresa SOREUS MOTORS E.I.R.L., dedicada a la venta de motocicletas, repuestos y accesorios, implementó un sistema de comercio electrónico personalizado como parte de su estrategia de expansión. Este sistema fue desarrollado utilizando el stack tecnológico MEAN (MongoDB, Express.js, Angular y Node.js) y permitió digitalizar el proceso de ventas, mejorar su alcance comercial y registrar datos clave sobre el comportamiento de los clientes, los productos más vendidos, las temporadas de mayor demanda y otros indicadores relevantes (KPI).

A raíz de esta implementación y del historial de datos generado por el sistema, el presente proyecto Capstone propone un análisis exploratorio y predictivo utilizando herramientas de ciencia de datos y programación en Python. El objetivo principal es transformar los datos recolectados por el e-commerce durante el año 2024 en información estratégica que permita a la empresa tomar decisiones más acertadas y orientadas al crecimiento.

El análisis buscará identificar patrones de comportamiento de los clientes, productos con mayor y menor rotación, periodos de alta y baja demanda, y proponer estrategias basadas en datos como campañas promocionales más efectivas, mejor planificación de inventario, segmentación de clientes y predicción de ventas. Además, se desarrollará un modelo de predicción que permita estimar el comportamiento comercial de la empresa a futuro y cuantificar los beneficios potenciales de aplicar estrategias basadas en ciencia de datos, tales como aumento de ventas, reducción de pérdidas por sobrestock o mejora en la fidelización de clientes.

Este proyecto no solo se enfocará en el modelamiento predictivo, sino que propondrá mejoras reales basadas en métricas de impacto como el rendimiento operativo, la eficiencia comercial y la rentabilidad, convirtiendo el análisis en una herramienta de apoyo directo para la toma de decisiones empresariales.

ABSTRACT

During the period 2023 - 2024, the company SOREUS MOTORS E.I.R.L., dedicated to the sale of motorbikes, spare parts and accessories, implemented a customized e-commerce system as part of its expansion strategy. This system was developed using the MEAN technology stack (MongoDB, Express.js, Angular and Node.js) and allowed them to digitise the sales process, improve their commercial reach and record key data on customer behaviour, best-selling products, seasons of highest demand and other relevant indicators (KPIs).

Following this implementation and the data history generated by the system, this Capstone project proposes an exploratory and predictive analysis using data science tools and Python programming. The main objective is to transform the data collected by e-commerce during the year 2024 into strategic information that will allow the company to make better, growth-oriented decisions.

The analysis will seek to identify customer behaviour patterns, products with higher and lower turnover, periods of high and low demand, and propose data-driven strategies such as more effective promotional campaigns, better inventory planning, customer segmentation and sales forecasting. In addition, a predictive model will be developed to estimate the future commercial behaviour of the company and to quantify the potential benefits of applying data science-based strategies, such as increased sales, reduced losses due to overstock or improved customer loyalty.

This project will not only focus on predictive modelling, but will also propose real improvements based on impact metrics such as operational performance, commercial efficiency and profitability, turning the analysis into a direct support tool for business decision-making.

1. IDEA DEL PROYECTO

TÍTULO TENTATIVO

- Análisis exploratorio y predictivo del comportamiento comercial en SOREUS MOTORS E.I.R.L. a partir de datos generados por su plataforma de comercio electrónico personalizado el 2024.
- Análisis exploratorio y predictivo de datos comerciales para la empresa SOREUS MOTORS E.I.R.L. en el periodo 2025
- Desarrollo de un modelo de predicción de ventas con python para SOREUS MOTORS E.I.R.L. Basado en su sistema E-commerce según sus datos generados el 2025.
- "Aplicación de ciencia de datos para la optimización de decisiones estratégicas en SOREUS MOTORS E.I.R.L. mediante análisis predictivo del comportamiento comercial en 2024" (TÍTULO SELECCIONADO)
- Optimización de procesos comerciales en SOREUS MOTORS E.I.R.L. mediante análisis predictivo de datos recolectados por su plataforma E-commerce.
- Análisis de comportamiento de clientes y predicción de ventas en SOREUS MOTORS E.I.R.L. basado en datos históricos del sistema E-commerce del 2024.
- Inteligencia comercial basada en datos para SOREUS MOTORS E.I.R.L.: análisis de KPIs y predicción de tendencias de venta del 2024.

DESCRIPCIÓN GENERAL

Este proyecto tiene como objetivo aplicar técnicas de análisis exploratorio y modelamiento predictivo para aprovechar el valor estratégico de los datos generados por el sistema de comercio electrónico personalizado implementado por SOREUS MOTORS E.I.R.L., empresa del sector automotriz ubicada en Huancayo, Perú. Este sistema, desarrollado en 2023 con tecnología MEAN, ha estado operativo durante los años 2023 y 2024, registrando información clave sobre ventas, clientes, productos, tendencias mensuales y otros indicadores de desempeño (KPI).

El proyecto propone realizar un análisis exploratorio de datos (EDA) que permita visualizar, comprender y comunicar el comportamiento comercial de la empresa durante el 2024. A partir de ello, se desarrollará un modelo predictivo en Python, empleando bibliotecas como Pandas, Scikit-learn, Matplotlib y Seaborn. Este modelo permitirá estimar la evolución futura de las ventas, identificar categorías de productos con mayor potencial, segmentar a los clientes por comportamiento de compra y detectar los momentos óptimos para aplicar promociones.

A través de este enfoque analítico, se busca responder preguntas estratégicas como:

¿Quiénes son los clientes más rentables y frecuentes?

¿Qué productos requieren mayor stock según la demanda proyectada?

¿Cuáles son los meses ideales para lanzar campañas promocionales?

¿Qué decisiones basadas en datos podrían optimizar la eficiencia operativa y comercial?

Además, se cuantificarán los beneficios potenciales de adoptar este enfoque, como el aumento de ingresos, la reducción de pérdidas por sobrestock, y la optimización de campañas y recursos. Más allá del análisis técnico, el objetivo final es transformar los datos en información accionable que respalde el crecimiento sostenible y estratégico de SOREUS MOTORS en los próximos años.

2. IDENTIFICACIÓN DEL PROBLEMA

Problema principal:

¿De qué manera la aplicación de técnicas de ciencia de datos permitirá mejorar la toma de decisiones estratégicas en la empresa SOREUS MOTORS E.I.R.L. para el 2025?

Consecuencias derivadas:

Gestión ineficiente del inventario:

- Productos con baja rotación permanecen en stock durante meses, generando inmovilización de capital y ocupación innecesaria de espacio en almacén.
- Pérdida de ventas por quiebre de stock en productos de alta demanda, lo que lleva a que los clientes abandonen la tienda física o virtual al no encontrar lo que buscan.

Desconocimiento del cliente:

- Falta de segmentación de clientes según rentabilidad o frecuencia de compra, dificultando el diseño de campañas de fidelización o estrategias personalizadas.

Toma de decisiones empírica:

- Muchas decisiones comerciales y operativas se basan en intuiciones o experiencias previas, sin un respaldo cuantitativo que permita anticiparse a tendencias del mercado.

Brecha existente:

Existe una brecha significativa entre la data recolectada y su aplicación estratégica. No transformar los datos históricos en conocimiento útil limita la capacidad de la empresa para optimizar su inventario, mejorar la experiencia del cliente, aumentar sus ingresos y crecer de manera sostenible.

3. OBJETIVOS SMART

Considerando que SMART es un acrónimo el cual refiere:

S: ESPECÍFICOS

M: MEDIBLES

A: ALCANZABLES

R: REALISTAS

T: A TIEMPO

Objetivo General:

Desarrollar un modelo de análisis exploratorio y predictivo de ventas mediante técnicas de ciencia de datos y programación en Python, utilizando los datos de ventas del periodo enero-diciembre 2024, con el fin de optimizar la toma de decisiones estratégicas en SOREUS MOTORS E.I.R.L. antes del 30 de junio de 2025.

Objetivos Específicos SMART:

- Recolectar y limpiar los datos históricos de ventas del sistema e-commerce de SOREUS MOTORS correspondientes al periodo enero – diciembre 2024, garantizando una integridad mínima del 95% en los registros procesados, antes del 7 de junio de 2025.
- Aplicar técnicas de análisis exploratorio de datos (EDA) sobre las ventas de enero a diciembre 2024 para identificar, antes del 13 de junio del 2025, al menos tres patrones de compra relevantes, productos con baja rotación (menos de 2 ventas mensuales) y tendencias estacionales.

- Diseñar e implementar un modelo de predicción de ventas mensuales utilizando algoritmos de Machine Learning (regresión lineal, árboles de decisión o ARIMA), logrando una precisión mínima del 80% según métricas como R^2 o MAE, dentro de un periodo de dos meses y antes del 20 de junio del 2025, con el fin de anticipar la demanda, evaluar la rotación del stock y mejorar la gestión del inventario.
- Desarrollar al menos 5 visualizaciones dinámicas e interactivas mediante Matplotlib, Seaborn y Plotly que faciliten la interpretación de resultados por parte del equipo directivo, con entrega final antes del 25 de junio del 2025.
- Elaborar y presentar un informe técnico final con al menos 5 recomendaciones accionables derivadas del análisis exploratorio y del modelo predictivo, orientadas a optimizar el inventario, fidelización de clientes y la estrategia comercial, a ser entregado el 30 de junio del 2025.

4. BÚSQUEDA DE LITERATURA ACADÉMICA

Predicción de demanda en pequeñas y medianas empresas del sector retail y automotriz: En el contexto de las Pymes del sector retail y automotriz, diversos estudios evidencian la importancia de contar con sistemas de predicción de demanda y control de inventarios para optimizar la gestión operativa y reducir costos.

Espinoza Morales y Porras Arévalo (2022), titulada *"Mejora en el control de inventarios para optimizar la gestión de compras en una empresa del sector retail"*. Esta investigación plantea que la falta de una adecuada gestión de inventarios, el desconocimiento de la tasa de consumo y la carencia de procedimientos estructurados en compras afectan directamente la eficiencia operativa de las empresas. Mediante la aplicación de herramientas como la clasificación ABC, la matriz de Kraljic y el modelo EOQ (Cantidad Económica de Pedido), se logró optimizar la gestión de compras y mejorar significativamente el control de inventarios.

Pinedo Chapa (2021), titulado *"Propuesta de un modelo de pronóstico de demanda y gestión de inventarios para la planeación de demanda en prendas de vestir juvenil"*. La autora plantea que muchas empresas del rubro comercial aún no aplican estrategias de planeación de demanda, perdiendo así oportunidades económicas y estratégicas. A través de la implementación de un modelo de pronóstico, la investigación demuestra que es posible anticiparse a la demanda por temporadas, gestionar adecuadamente el inventario y mejorar el uso de recursos materiales y humanos. Además, se propone realizar un seguimiento continuo de las compras, alineándose con el plan proyectado para evitar el sobrestock.

Estos antecedentes respaldan la necesidad de aplicar técnicas analíticas y predictivas en Pymes para mejorar la toma de decisiones en inventario y ventas. Ambos estudios coinciden en que la implementación de modelos de gestión y pronóstico puede traducirse en beneficios económicos, eficiencia logística y mayor satisfacción del cliente.

El informe de **McKinsey** titulado *"Insights to impact: Creating and sustaining data-driven commercial growth"* señala que las empresas que implementan motores de crecimiento basados en datos reportan un crecimiento superior al del mercado y aumentos en el EBITDA en el rango del 15 al 25%.

Li, C., et al., (2019) en su estudio sobre los factores que influyen en la retención estudiantil. Plantearon las hipótesis de que el rendimiento académico y la situación financiera de los estudiantes, están estrechamente relacionadas con su retención en la universidad, emplearon modelos de regresión Lineal y Logística, árboles de decisión y métodos de aprendizaje automático (kNN) para estudiar la retención estudiantil, y finalmente a partir de estos métodos, identificaron las variables más importantes para predecir la retención estudiantil y lograron una alta precisión en la predicción de la retención de estudiantes en la universidad. Los investigadores utilizaron la IA y el ML para analizar patrones y relaciones en los datos, lo que les permitió identificar factores que podrían influir en la retención estudiantil, como el rendimiento académico previo, las características demográficas de los estudiantes, su uso de recursos universitarios, y sus necesidades económicas. El conocimiento obtenido a través de la ciencia de datos les permitió validar su hipótesis, desarrollar estrategias de apoyo personalizadas y más efectivas, con el objetivo de mejorar la retención estudiantil.

Heilman, E., et al. (2019), desarrollaron una investigación sobre el uso de la ciencia de datos en los niveles superiores del Ejército, para responder preguntas de inteligencia. En esta investigación, se sugiere aplicar la ciencia de datos de manera estructurada y científica en el análisis de inteligencia militar, guiando la toma de decisiones, formulando hipótesis y validando conclusiones a través de la metodología CRISPDM. Buscando aumentar la confianza de las estimaciones a un nivel superior y proporcionar productos de inteligencia mejorados.

En el contexto empresarial, Bocangel, J. L., et al. (2020), usan la ciencia de datos para responder preguntas como: ¿Es posible predecir si un cliente Freemium se convertirá en Premium? La hipótesis que plantean sugiere que utilizando un modelo de árbol de decisión se puede predecir la conversión de cuentas Freemium a Premium en una empresa peruana, basándose en el comportamiento de diversas variables. Los resultados revelaron que el modelo de árbol de decisión efectivamente permite realizar estas predicciones basadas en el comportamiento de los usuarios.

Santistevan, J. (2024) elaboró una investigación sobre el análisis predictivo para servicios de movilización de la empresa Fastline en el área de logística de Guayaquil, Ecuador, cuyo objetivo fue mejorar la eficiencia operativa del departamento de Logística y la satisfacción del cliente mediante el uso de modelos de análisis predictivo y análisis de negocio, se utilizó información histórica y actual de los servicios de movilización atendidos, que pasó por un proceso de Extracción, Transformación y Carga (ETL). Una vez finalizada la limpieza de los datos, se envió a SQL Server y posteriormente se utilizó en la herramienta de Power BI, enfocado en optimizar la gestión de la flota de vehículos y reducir el tiempo de espera del cliente mediante modelos de análisis de series temporales.

La investigación es de carácter no experimental debido al énfasis en el análisis predictivo y el uso de datos existentes; es correlacional, de tipo cuantitativo y el método de investigación es analítica. La investigación incluyó encuesta cuantitativa de calificación por servicio atendido para evaluar la experiencia del cliente, como resultado, se tomaron decisiones que ayudaron a mejorar la gestión operativa de los servicios de movilización corporativos de Fastline, contribuyendo a mejorar su competitividad local.

Artículo: "8 modelos de machine learning explicados en 20 minutos"

Los modelos de machine learning, como XG Boost, Random Forest y regresión lineal múltiple, permiten resolver problemas empresariales mediante la identificación de patrones en los datos sin requerir programación explícita. Estos algoritmos pueden adaptarse con el tiempo conforme se integran nuevos datos, lo que resulta clave en escenarios como el pronóstico de ventas, donde las condiciones del mercado son cambiantes. Su implementación con bibliotecas como Scikit-Learn facilita el análisis predictivo incluso para profesionales con conocimientos intermedios en estadística.

Artículo: “El método de cohorte aplicado a machine learning para el pronóstico de series temporales”

El método de cohorte aplicado al machine learning para series temporales permite agrupar datos por características comunes, como el mes de adquisición de clientes, facilitando la identificación de patrones específicos en cada grupo. Al aplicar modelos como ARIMA, Prophet o redes neuronales a nivel de cohorte, se mejora la precisión del pronóstico y se obtiene una comprensión más granular del comportamiento del consumidor, lo que resulta útil para estrategias de ventas e inventario en sectores como retail y automotriz.

5. JUSTIFICACIÓN E IMPORTANCIA DEL PROYECTO

En la actualidad, el uso estratégico de los datos se ha convertido en un factor clave para la competitividad empresarial. Diversos estudios destacan la importancia del análisis de datos en la mejora del rendimiento organizacional. Por ejemplo, McKinsey & Company (2022) indica que las empresas que adoptan estrategias basadas en datos experimentan un crecimiento superior al del mercado y aumentos en el EBITDA de entre el 15% y el 25%.

En el contexto peruano, la Cámara de Comercio de Lima proyecta que el comercio electrónico alcanzará los 23,000 millones de dólares en 2024, lo que representa un crecimiento del 15% respecto al año anterior. Estos datos subrayan la necesidad de que empresas como SOREUS MOTORS E.I.R.L., que ya cuentan con un sistema e-commerce funcional, aprovechen sus datos para optimizar sus decisiones estratégicas y mantenerse competitivas en un entorno cada vez más digitalizado.

Actualmente, SOREUS MOTORS enfrenta problemas como la acumulación de inventario sin rotación —con hasta un 30% del stock inmóvil por más de tres meses, según registros internos— y quiebres de stock en productos de alta demanda, afectando la satisfacción del cliente y generando pérdidas de ventas. Estas limitaciones operativas reflejan una falta de inteligencia de negocio basada en datos, lo que limita su crecimiento sostenible.

Este proyecto propone aplicar técnicas de ciencia de datos y aprendizaje automático para transformar los datos históricos del e-commerce en información útil y modelos predictivos. Se espera alcanzar una precisión mínima del 80% en los pronósticos de ventas, lo cual permitirá a la empresa anticiparse a la demanda, optimizar su inventario y mejorar su planificación comercial.

Finalmente, el desarrollo de este proyecto no solo contribuirá directamente a la toma de decisiones de SOREUS MOTORS, sino que también representará una oportunidad para aplicar y demostrar las competencias adquiridas en la formación como profesional en ciencia de datos, sentando un precedente de cómo las pymes pueden beneficiarse de soluciones basadas en datos para ser más eficientes y sostenibles.

6. LÍMITES Y ALCANCES DEL PROYECTO DEL PROYECTO

Alcances:

- El proyecto se desarrollará utilizando datos reales proporcionados por el gerente de la empresa Soreus Motors.
- El análisis se enfocará en las ventas históricas, con el objetivo de generar pronósticos para un horizonte de hasta 12 meses.
- Se elaborará una interfaz visual (dashboard o reporte interactivo) que sintetice los hallazgos más relevantes y facilite la toma de decisiones.

Límites:

- La calidad y precisión de los modelos predictivos estarán condicionadas por la integridad, cantidad y consistencia de los datos disponibles.
- No se incorporarán variables externas que puedan influir en las ventas, como campañas publicitarias, indicadores macroeconómicos o estacionalidad externa.
- El alcance del proyecto se limita al análisis y visualización de resultados, sin incluir la implementación del modelo en un entorno de producción real.

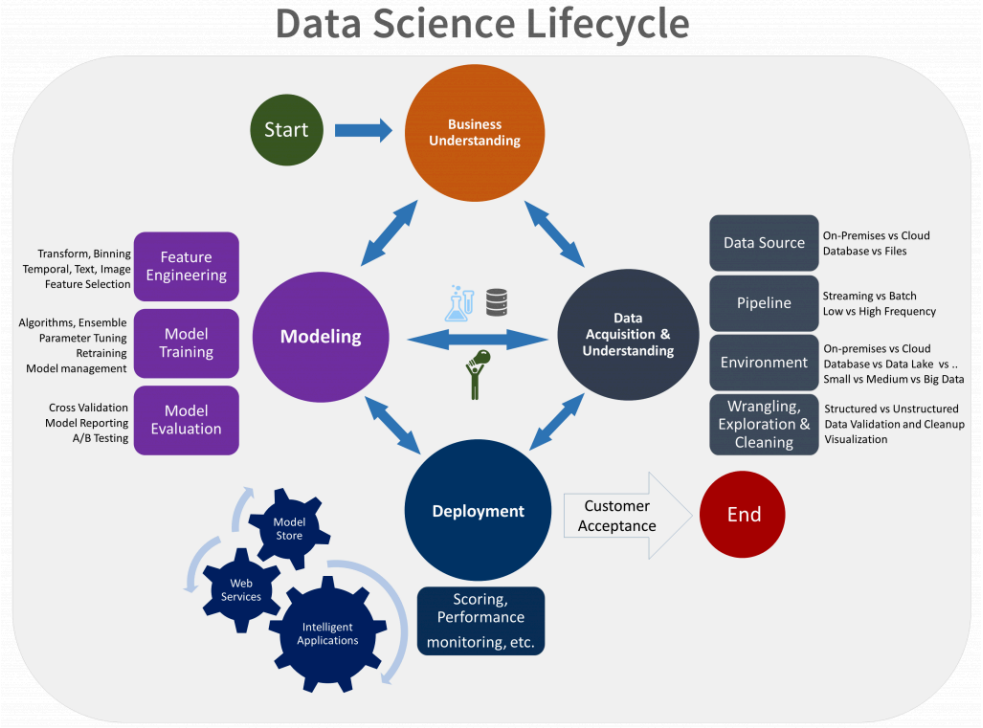
7. DESARROLLO DE UN MARCO TEÓRICO O CONCEPTUAL

Ciencia de Datos: La ciencia de datos es una disciplina interdisciplinaria que combina estadística, programación computacional, análisis de datos y conocimiento del dominio específico para extraer valor, patrones y conocimiento útil a partir de grandes volúmenes de datos estructurados y no estructurados.

> **Según Provost y Fawcett (2013)**, la ciencia de datos es el arte de transformar datos en conocimiento accionable para la toma de decisiones. Esta disciplina integra técnicas de minería de datos, aprendizaje automático, visualización de datos y modelado estadístico.

La ciencia de datos permite identificar tendencias, predecir comportamientos y optimizar procesos en diferentes contextos empresariales, siendo clave para la toma de decisiones basada en evidencia.

Figura 1: *Ciclo de vida de la ciencia de datos.*



Fuente: Imágenes Google

Ciencia de Datos y su Aplicación en los Negocios: La ciencia de datos es un campo interdisciplinario que combina estadística, programación y conocimiento del dominio para extraer conocimiento útil a partir de grandes volúmenes de datos **(Provost & Fawcett, 2013).** Su aplicación en los negocios permite transformar datos crudos en información estratégica para la toma de decisiones. En contextos empresariales como el comercio electrónico, permite optimizar procesos como ventas, inventario y fidelización del cliente.

Análisis Exploratorio de Datos (EDA): El análisis exploratorio de datos (EDA) es un enfoque estadístico para investigar y comprender la estructura de los datos antes de aplicar modelos predictivos. A través de gráficos, estadísticas descriptivas y técnicas de limpieza, el EDA permite identificar tendencias, patrones, valores atípicos y relaciones entre variables **(Tukey, 1977).**

Modelos Predictivos y Machine Learning: El aprendizaje automático machine learning permite crear modelos que aprenden a partir de datos históricos para hacer predicciones futuras. En este proyecto se utilizarán modelos como:

Regresión Lineal Múltiple: La regresión lineal múltiple es una técnica estadística utilizada para modelar la relación entre una variable dependiente continua y dos o más variables independientes. Su objetivo es predecir el valor de una variable dependiente a partir de la combinación lineal de las variables predictoras.

Matemáticamente, se expresa como:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + \cdots + \beta nXn + \varepsilon$$

Donde:

- Y**:** variable dependiente.
- $X1,\ X2,\ \cdots,\ Xn$**:** variables independientes.
- $\beta 0$**:** intercepto.
- $\beta 1,\ \beta 2, \cdots,\ \beta n$: coeficientes de regresión.
- ε: error aleatorio.

Esta técnica permite identificar relaciones lineales entre múltiples variables y se utiliza ampliamente para la predicción, análisis de tendencias y comprensión del impacto relativo de cada predictor en un contexto determinado.

Random Forest: Random Forest es una técnica de aprendizaje automático basada en el conjunto de árboles de decisión. Se utiliza tanto para problemas de clasificación como de regresión. En lugar de construir un único árbol, Random Forest construye múltiples árboles de decisión (generalmente cientos) y combina sus resultados para obtener una predicción más robusta y precisa.

Cada árbol del bosque se entrena con una muestra aleatoria del conjunto de datos original (bootstrap) y, durante su construcción, selecciona aleatoriamente un subconjunto de variables para dividir en cada nodo. Esta aleatorización mejora la diversidad de los modelos individuales y reduce el riesgo de sobreajuste (overfitting), una limitación común de los árboles de decisión individuales.

Las predicciones se obtienen mediante el promedio de los resultados (en regresión) o el voto mayoritario (en clasificación) de todos los árboles.

Ventajas de Random Forest:

- Alta precisión y generalización.
- Maneja bien datos con muchas variables y relaciones no lineales.
- Robusto frente a valores atípicos y datos faltantes.
- Ofrece medidas de importancia de las variables.

Desventajas:

- Menor interpretabilidad comparado con un solo árbol.
- Requiere más recursos computacionales.

ARIMA: Arima o AutoRegressive Integrated Moving Average es una técnica estadística ampliamente utilizada para el análisis y pronóstico de series temporales. Su nombre deriva de tres componentes:

- **AR (AutoRegresivo):** El modelo usa dependencias lineales entre una observación y un número de rezagos (valores anteriores) de la misma serie.
- **I (Integrado):** Implica la diferenciación de los datos para hacer la serie estacionaria (es decir, sin tendencia ni variaciones estacionales).
- **MA (Media Móvil):** Modela el error de predicción como una combinación lineal de errores observados en el pasado.

El modelo ARIMA se denota como ARIMA(p, d, q), donde:

- **p:** número de términos autorregresivos.
- **d:** número de diferenciaciones necesarias para hacer estacionaria la serie.
- **q:** número de términos de media móvil.

También existen variantes como SARIMA, que incorporan estacionalidad, permitiendo capturar patrones repetitivos a lo largo del tiempo (por ejemplo, variaciones mensuales o anuales).

ARIMA es especialmente útil para series de datos univariadas, donde se busca comprender patrones pasados y predecir valores futuros, como la demanda mensual de productos o las ventas históricas.

Estas técnicas serán evaluadas con métricas como RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) y R^2 (coeficiente de determinación), a fin de medir su precisión en la predicción de ventas.

Visualización de Datos: Las visualizaciones permiten comunicar los resultados de forma clara y accesible a los tomadores de decisiones. Se emplearán herramientas como Matplotlib, Seaborn, y Plotly para construir dashboards interactivos y gráficos dinámicos. Power BI se utilizará opcionalmente para construir tableros ejecutivos que permitan explorar resultados de manera intuitiva.

Herramientas Tecnológicas del Proyecto:

Python: Es uno de los lenguajes de programación más utilizados en ciencia de datos por su sintaxis simple, flexibilidad y gran ecosistema de bibliotecas especializadas. En este proyecto se emplearán las siguientes librerías:

- **Pandas:** Permite la manipulación y análisis eficiente de estructuras de datos tabulares (DataFrames), facilitando tareas como limpieza, transformación y agregación de datos.
- **Numpy:** Soporta operaciones numéricas de alto rendimiento sobre arreglos multidimensionales, siendo la base para cálculos matemáticos y estadísticos complejos.
- **Scikit-learn:** Biblioteca especializada en algoritmos de aprendizaje automático (machine learning), usada para tareas como regresión, clasificación, validación cruzada y evaluación de modelos predictivos.
- **Matplotlib y Seaborn:** Herramientas para crear visualizaciones estáticas y detalladas que permiten identificar patrones, correlaciones y comportamientos en los datos.
- **Plotly:** Biblioteca interactiva para la creación de dashboards y gráficos dinámicos, útil para facilitar la interpretación de resultados por parte de usuarios no técnicos.

Estas herramientas integradas en entornos como Jupyter Notebook o Google Colab permiten desarrollar flujos de trabajo reproducibles y visualmente intuitivos, potenciando la toma de decisiones basada en evidencia.

Jupyter Notebook / Google Colab: Son entornos interactivos ampliamente utilizados en el análisis de datos y desarrollo de modelos predictivos. Permiten escribir, ejecutar y documentar código Python en una misma interfaz, lo cual facilita la trazabilidad del proceso analítico.

- Jupyter Notebook: Se ejecuta localmente y ofrece gran flexibilidad para trabajar con librerías como Pandas, Scikit-learn y Matplotlib.
- Google Colab: Es una plataforma gratuita en la nube de Google que permite compartir proyectos fácilmente y ejecutar código en GPU o TPU sin requerir instalación local.

Excel y archivos CSV: Microsoft Excel, junto con los archivos en formato CSV (Comma Separated Values), son herramientas fundamentales para la manipulación y análisis preliminar de datos. Los archivos CSV permiten almacenar grandes volúmenes de datos en formato estructurado, siendo ampliamente compatibles con herramientas como Excel y lenguajes de programación como Python.

En proyectos de ciencia de datos, Excel se utiliza frecuentemente para la validación cruzada de datos, facilitando la verificación manual de registros, la detección de errores, valores atípicos o duplicados. Además, permite generar tablas dinámicas, realizar análisis descriptivos y crear gráficos simples que apoyan la exploración inicial del dataset.

Su uso es especialmente útil en etapas de preparación de datos, control de calidad y presentación de resultados en entornos empresariales donde esta herramienta es estándar.

Métricas Clave del Proyecto

Para evaluar la efectividad del análisis predictivo y el impacto estratégico del proyecto, se utilizarán las siguientes métricas clave:

- **RMSE (Root Mean Squared Error) y MAE (Mean Absolute Error):** Estas métricas permiten cuantificar la precisión de los modelos de predicción desarrollados. El RMSE mide el error cuadrático medio entre los valores reales y los predichos, penalizando más fuertemente los errores grandes, mientras que el MAE calcula el error absoluto medio, ofreciendo una medida más intuitiva del promedio de desviación del modelo. Ambos indicadores son esenciales para validar la calidad de los algoritmos implementados, especialmente en modelos de regresión y series temporales.
- **Porcentaje de aumento proyectado en ventas:** Esta métrica estima el impacto que tendría el uso del modelo predictivo en los ingresos de la empresa. Se calcula comparando el comportamiento histórico de ventas con los resultados simulados tras la implementación de recomendaciones basadas en el modelo. Es un indicador directo del valor comercial generado por el proyecto.
- **Porcentaje de reducción de stock inmovilizado:** Representa el grado de eficiencia logrado en la gestión del inventario, al identificar productos de baja rotación y anticipar la demanda futura. Una reducción en el stock inmovilizado indica mejoras en la rotación de productos, liberación de espacio de almacenamiento y disminución de pérdidas por obsolescencia.

Transformación Digital en PYMEs: Las pequeñas y medianas empresas (PYMEs), como SOREUS MOTORS E.I.R.L., pueden obtener ventajas competitivas al adoptar herramientas de analítica avanzada. Sin embargo, muchas aún no aprovechan los datos generados por sus sistemas digitales. Este proyecto busca evidenciar cómo el uso adecuado de datos puede mejorar la eficiencia operativa y la estrategia comercial.

8. ANÁLISIS FODA / VISIÓN / MISIÓN / ESTRATEGIAS / ACTIVIDADES

8.1. ANÁLISIS FODA

Fortalezas (F+):

- **F1:** Experiencia técnica en herramientas de ciencia de datos y visualización (Python, Scikit-learn, Plotly, Power BI).
- **F2:** Acceso a datos históricos del e-commerce y del inventario de la empresa.

Oportunidades (O+):

- **O1:** Crecimiento del comercio electrónico en Perú (+25% en 2023 según CCL).
- **O2:** Bajo aprovechamiento actual de los datos por parte de la empresa, lo que genera alto potencial de mejora con analítica.

Debilidades (D-):

- **D1:** Falta de cultura organizacional basada en datos en la empresa.
- **D2:** Limitaciones en infraestructura tecnológica para implementar modelos automatizados a largo plazo.

Amenazas (A-):

- **A1:** Alta competencia en el sector automotriz digitalizado con estrategias basadas en datos.
- **A2:** Resistencia al cambio por parte del personal operativo o administrativo de la empresa.

Matriz FODA

Tabla 1: *Matriz FODA.*

	D1 (-)	**D2 (-)**	**A1 (-)**	**A2 (-)**
F1 (+)	**F1-D1:** Capacitar al equipo de SOREUS en lectura de dashboards simples para fomentar una cultura basada en datos.	**F1-D2:** Usar Google Colab y herramientas gratuitas en la nube para evitar depender de infraestructura local.	**F1-A1:** Implementar modelos competitivos y diferenciarse con visualizaciones ejecutivas claras.	**F1-A2:** Mostrar beneficios inmediatos de la analítica con reportes de valor simple y práctico.
F2 (+)	**F2-D1:** Aprovechar los datos históricos para construir ejemplos que muestren el valor de las decisiones basadas en datos.	**F2-D2:** Aplicar modelos que funcionen sin requerir recursos pesados (como regresión lineal o ARIMA localmente).	**F2-A1:** Usar los datos internos para personalizar las estrategias comerciales frente a competidores.	**F2-A2:** Generar reportes automatizados simples para reducir carga y aumentar aceptación.
O1 (+)	**O1-D1:** Justificar el cambio cultural con cifras del crecimiento e-commerce en Perú.	**O1-D2:** Aprovechar financiamiento o tecnologías gratuitas con enfoque digital.	**O1-A1:** Anticiparse con modelos predictivos antes que la competencia.	**O1-A2:** Promover adopción del modelo como herramienta para enfrentar nuevos retos del mercado.
O2 (+)	**O2-D1:** Aprovechar que la empresa no usa datos actualmente para establecer prácticas desde cero.	**O2-D2:** Plantear soluciones escalables que se ajusten a su nivel tecnológico.	**O2-A1:** Convertir la desventaja actual en ventaja competitiva si se actúa pronto.	**O2-A2:** Involucrar progresivamente al personal en el uso de datos, iniciando por áreas clave.

Fuente: *Elaboración Propia.*

8.2. VISIÓN

VISIÓN DE SOREUS MOTORS: Ser la empresa líder en la región en la comercialización de motos y accesorios automotrices, reconocida por su calidad, innovación, confianza y atención al cliente, contribuyendo al desarrollo del transporte eficiente y seguro en el Perú.

VISIÓN DEL PROYECTO DE DATA SCIENCE: Transformar los datos operativos y de ventas de SOREUS MOTORS E.I.R.L. en conocimiento estratégico, mediante técnicas de ciencia de datos y aprendizaje automático, con el fin de anticipar la demanda, optimizar la gestión del inventario y potenciar la toma de decisiones basada en evidencia para lograr una ventaja competitiva sostenible.

8.3. MISIÓN

Misión de SOREUS MOTORS E.I.R.L.: Brindar a nuestros clientes soluciones de movilidad confiables mediante la comercialización de motos y accesorios de alta calidad, ofreciendo un servicio personalizado, eficiente y accesible que contribuya al desarrollo y bienestar de la comunidad motera en la región.

Visión del proyecto de DATA SCIENCE: Convertir los datos generados por el sistema e-commerce de SOREUS MOTORS en un activo estratégico que impulse la toma de decisiones inteligentes, optimice la gestión de inventario y potencie el crecimiento comercial sostenible a través del análisis avanzado y la inteligencia artificial.

8.4. ESTRATEGIAS

Estrategia de recolección y limpieza de datos:

- Implementar un proceso automatizado de extracción y limpieza de datos desde el sistema e-commerce.
- Aplicar reglas de validación y consistencia para asegurar una integridad mínima del 95% en los registros.

Estrategia de análisis exploratorio de datos (EDA):

- Identificar patrones de compra, productos de baja rotación y estacionalidad con visualizaciones interactivas.
- Segmentar a los clientes según frecuencia de compra, ticket promedio y tipo de producto.

Estrategia de modelamiento predictivo:

- Diseñar y comparar modelos (Regresión Lineal Múltiple, Árboles de Decisión y ARIMA) para predecir la demanda mensual con métricas como R^2, MAE y RMSE.
- Seleccionar el modelo más preciso (mínimo 80%) para anticipar necesidades de stock y optimizar inventarios.

Estrategia de visualización de resultados:

- Desarrollar dashboards interactivos con Plotly y Power BI para facilitar la interpretación de resultados por parte de la gerencia.
- Incluir al menos 5 visualizaciones clave que resuman insights relevantes del análisis.

Estrategia de comunicación y toma de decisiones:

- Elaborar un informe técnico final con recomendaciones prácticas basadas en datos para mejorar ventas, fidelización y rotación de productos.
- Capacitar al personal clave en la lectura de dashboards y uso de resultados para decisiones comerciales.

Estrategia de sostenibilidad del sistema:

- Documentar el proceso completo para futuras réplicas o actualizaciones del modelo.
- Proponer la integración periódica de estos análisis como parte del ciclo de inteligencia comercial de la empresa.

8.5. ACTIVIDADES

En esta sección se detallan las actividades que se llevarán a cabo para el desarrollo del proyecto Capstone, organizadas por etapas:

8.5.1. Planificación y Recolección de Datos:

- Reunión con el gerente de SOREUS MOTORS para definir objetivos y validar acceso a datos.
- Recolección de datos históricos de ventas, stock e inventario desde el sistema e-commerce (enero – diciembre 2024).
- Verificación de integridad de los datos y documentación de las fuentes.

8.5.2. Limpieza y Preparación de Datos:

- Limpieza de datos con Python: eliminación de duplicados, tratamiento de nulos y outliers.
- Validación de la calidad de los datos (mínimo 95% de registros íntegros).
- Transformación y estructuración de datos para análisis y modelado.

8.5.3. Análisis Exploratorio de Datos (EDA):

- Generación de estadísticas descriptivas y análisis de correlación.
- Identificación de patrones de compra, productos de baja rotación y estacionalidad.
- Visualización de resultados con Matplotlib, Seaborn y Plotly.

8.5.4. Desarrollo de Modelos Predictivos:

- Entrenamiento de modelos: Regresión Lineal Múltiple, Random Forest y ARIMA.
- Evaluación de los modelos mediante métricas R^2, MAE y RMSE.
- Selección del modelo óptimo (precisión mínima del 80%).

8.5.5. Visualización de Resultados:

- Diseño de dashboards interactivos con Python.
- Elaboración de al menos 5 visualizaciones clave orientadas a la toma de decisiones.

8.5.6. Documentación y Recomendaciones:

- Elaboración del informe técnico final.
- Desarrollo de recomendaciones accionables para la empresa.
- Presentación y defensa del proyecto ante el comité docente.

8.5.7. Sostenibilidad y Transferencia:

- Documentación del flujo de trabajo y código.
- Capacitación básica al equipo directivo de SOREUS MOTORS para uso de dashboards.
- Entrega de scripts y manuales para futuras actualizaciones del modelo.

9. MÉTODO DE MARCO LÓGICO

Tabla 2: *Matriz de marco Lógico del Proyecto*

Resumen Narrativo	**Indicadores Verificables Objetivamente (IVO)**	**Medios de Verificación (MV)**	**Supuestos Importar**
Fin **Mejorar la competitividad de SOREUS MOTORS mediante decisiones estratégicas basadas en datos.**	Aumento del 15% en las ventas en un periodo de 6 meses.	Reportes de ventas mensuales antes y después del proyecto.	Que la gerencia implemente las recomendaciones generadas.
Propósito **Transformar datos de e-commerce en conocimiento útil para optimizar inventario y estrategia comercial.**	Reducción del 20% en stock inmovilizado. Mejoramiento del nivel de satisfacción del cliente.	Informes de inventario. Encuestas de satisfacción al cliente.	Que los datos disponibles sean suficientes y de calidad.
Componentes / Resultados Esperados			
1. Datos limpios y estructurados.	100% de registros depurados y preparados para análisis.	Scripts de limpieza y dataset final.	Disponibilidad de acceso a la base de datos del e-commerce.
2. Modelos predictivos desarrollados y evaluados.	Modelos con mínimo 80% de precisión en predicción de demanda.	Jupyter Notebooks / Google Colab con métricas RMSE / MAE.	Que los patrones en los datos sean estables y predecibles.
3. Visualizaciones dinámicas generadas.	Al menos 5 dashboards o visualizaciones interactivas.	Dashboards en Plotly / Power BI.	Que la dirección utilice los dashboards para tomar decisiones.
4. Informe técnico con recomendaciones accionables.	Documento con mínimo 5 recomendaciones basadas en resultados.	Informe final presentado a la dirección.	Que se cuente con tiempo y apoyo para presentar el informe.

Actividades Principales			
A1. Recolección y limpieza de datos.	Log de limpieza y dataset listo.	Scripts en Python y Excel.	Disponibilidad de logs e históricos.
A2. Análisis exploratorio de datos.	Reportes de EDA.	Visuales en Seaborn, Matplotlib.	Coherencia en los registros.
A3. Desarrollo y prueba de modelos.	Reporte de métricas del modelo.	Notebook / Colab.	Herramientas funcionales.
A4. Creación de visualizaciones.	Dashboard funcional.	Reportes en Power BI / Plotly.	Feedback del equipo directivo.
A5. Elaboración de informe final.	Documento PDF con recomendaciones.	Presentación y entrega oficial.	Disponibilidad del tiempo para elaboración.

Fuente: *Elaboración Propia.*

10. RESULTADOS ESPERADOS

Datos limpios y estructurados para análisis

- La información del e-commerce (ventas, productos, stock) será depurada, organizada y validada, permitiendo su uso en modelos predictivos confiables.

Modelo predictivo de demanda con al menos 80% de precisión

- Se implementarán modelos como Regresión Lineal Múltiple, Árboles de Decisión y/o ARIMA que permitan anticipar la demanda de productos, evaluados con métricas como RMSE y MAE.

Reducción proyectada del 20% en stock inmovilizado

- Mediante el análisis y pronóstico de rotación de productos, se optimizará la gestión del inventario, identificando productos con baja salida para evitar sobrestock.

Visualizaciones interactivas y dashboards para toma de decisiones

- Se crearán reportes visuales con herramientas como Power BI, Seaborn o Plotly, para facilitar el entendimiento y uso de la información por parte de la gerencia.

Propuesta de estrategia basada en datos para mejorar ventas

- Se entregará un informe con recomendaciones accionables para mejorar la planificación comercial, segmentación de productos, y promociones basadas en el análisis de patrones de compra.

Incremento estimado del 10% al 15% en ventas

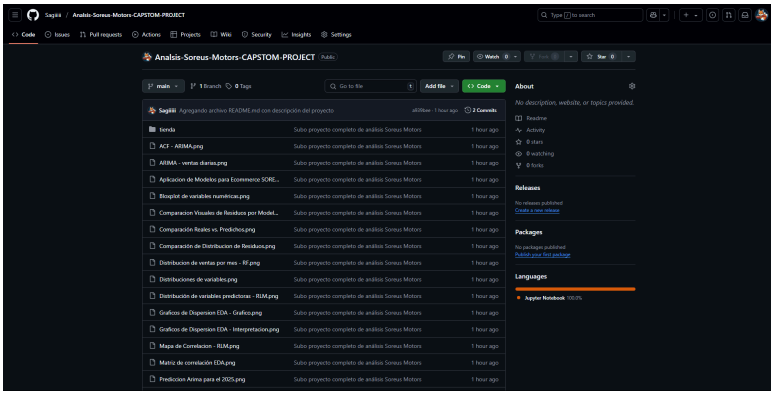
- Gracias a una mejor alineación entre inventario y demanda, se proyecta un aumento de ingresos al reducir quiebres de stock y pérdidas de oportunidad.

Fortalecimiento de la inteligencia de negocio en la empresa

- El proyecto servirá como caso base para instaurar una cultura orientada a decisiones basadas en datos, capacitando al equipo en el uso de los resultados obtenidos.

11. INVESTIGACIÓN Y ANÁLISIS (METODOLOGÍA)

ENLACE DE GITHUB:
https://github.com/Sagiiiii/Analsis-Soreus-Motors-CAPSTOM-PROJECT



La presente investigación se basa en un enfoque cuantitativo, no experimental y correlacional, utilizando técnicas de análisis exploratorio y predictivo con datos históricos provenientes del sistema e-commerce de SOREUS MOTORS E.I.R.L. La metodología aplicada es CRISP-DM (Cross-Industry Standard Process for Data Mining), adaptada al contexto empresarial y tecnológico de una PYME peruana.

Tipo de Investigación:

- **Enfoque:** Cuantitativo
- **Diseño:** No experimental, de tipo transversal
- **Alcance:** Correlacional - Predictivo
- **Método:** Analítico y descriptivo
- **Instrumento:** Dataset extraído de MongoDB (ventas, inventario, clientes, productos)

Metodología CRISP-DM

Se seguirá el ciclo de vida del proyecto de ciencia de datos bajo las seis fases del modelo CRISP-DM:

Comprensión del negocio: Análisis del contexto empresarial de SOREUS MOTORS: problemas, objetivos y necesidades. Identificación de preguntas clave a responder con datos: rotación de productos, predicción de demanda, campañas, etc.

Comprensión de los datos: Extracción de datos desde la base MongoDB del sistema e-commerce (enero–diciembre 2024). Revisión del diccionario de datos, identificación de atributos útiles, exploración inicial con estadísticas básicas y gráficos.

Preparación de los Datos: Limpieza de datos: eliminación de duplicados, valores faltantes y outliers. Transformación: creación de nuevas variables, formatos de fechas, normalización si es necesario. Segmentación de los datos en conjuntos de entrenamiento y prueba.

Modelado: Aplicación de modelos predictivos: ARIMA (series temporales), Regresión Lineal Múltiple y Random Forest (aprendizaje supervisado). Ajuste de parámetros y entrenamiento de modelos con Scikit-learn y Statsmodels. Validación con métricas MAE, RMSE y R^2 para comparar el desempeño de los modelos.

Evaluación:

- Interpretación de los resultados de cada modelo.
- Identificación del modelo con mejor precisión predictiva.
- Análisis de errores, interpretación de variables importantes y detección de oportunidades de mejora en decisiones comerciales.

Implementación:

Creación de dashboards con Plotly y Power BI para la toma de decisiones gerenciales.

Presentación de recomendaciones basadas en los resultados del modelo y el análisis exploratorio.

Documentación del código y del proceso para replicabilidad futura en la empresa.

Herramientas Tecnológicas:

Lenguaje: Python

Entorno: Júpiter Notebook

Librerías: Pandas, Numpy, Scikit-learn, Statsmodels, Matplotlib, Seaborn, Plotly

Complementos: Excel para validación cruzada y reportes.

12. DESARROLLO DE LOS MODELOS DE ML

12.1. ARIMA

Arima (Series de Tiempo)

- Preprocesamiento
- Gráficos
- Revisión de Estacionariedad
- Entrenamiento
- Predicción

Se realiza una copia del data frame original después de la limpieza:

```
# Copiar df_final a df_arima
df_arima = df_final.copy()
```

Creamos una serie diaria de residuos:

```
# Crear serie diaria de residuos (sobrestock)
serie_sobrestock_diario = df_arima['residuo'].resample('D').sum().asfreq('D').fillna(0)
```

División: últimos 90 días como prueba

```
dias_prueba = 90
train_arima = serie_sobrestock_diario[:-dias_prueba]
test_arima = serie_sobrestock_diario[-dias_prueba:]
```

Aplicar un FOR para decidir cuál orden de ARIMA usare:

```
p_values = [0, 1, 2]
d_values = [1]
q_values = [0, 1, 2]

resultados = []

for p in p_values:
    for d in d_values:
        for q in q_values:
            orden = (p, d, q)
            try:
                modelo = ARIMA(train_arima, order=orden)
                modelo_fit = modelo.fit()
                pred = modelo_fit.forecast(steps=dias_prueba)

                mae = mean_absolute_error(test_arima, pred)
                rmse = np.sqrt(mean_squared_error(test_arima, pred))
                r2 = r2_score(test_arima, pred)

                resultados.append({
                    'orden': orden,
                    'MAE': mae,
                    'RMSE': rmse,
                    'R2': r2
                })
            except Exception as e:
                print(f"❌ Falló para orden {orden}: {e}")

# Mostrar resultados ordenados por RMSE
df_resultados = pd.DataFrame(resultados).sort_values(by='RMSE')
df_resultados.head(10)  # Ver los 10 mejores
```

	orden	MAE	RMSE	R2
7	(2, 1, 1)	167.631196	287.497594	-0.048537
4	(1, 1, 1)	167.236697	287.541354	-0.048856
8	(2, 1, 2)	167.428644	287.583677	-0.049165
2	(0, 1, 2)	166.013662	305.617145	-0.184870
1	(0, 1, 1)	176.653925	323.503883	-0.327621
5	(1, 1, 2)	179.240978	326.538005	-0.352641
0	(0, 1, 0)	183.222222	330.802727	-0.388204
6	(2, 1, 0)	184.017855	331.671566	-0.395506
3	(1, 1, 0)	185.156824	332.812029	-0.405120

Entrenamiento:

```
# Serie diaria de sobrestock
serie_sobrestock_diario = df_arima['residuo'].resample('D').sum().asfreq('D').fillna(0)

# División train/test
dias_prueba = 90
train_arima = serie_sobrestock_diario[:-dias_prueba]
test_arima = serie_sobrestock_diario[-dias_prueba:]

# Entrenar modelo ARIMA(2, 1, 1)
modelo_arima = ARIMA(train_arima, order=(2, 1, 1))
modelo_arima_fit = modelo_arima.fit()
```

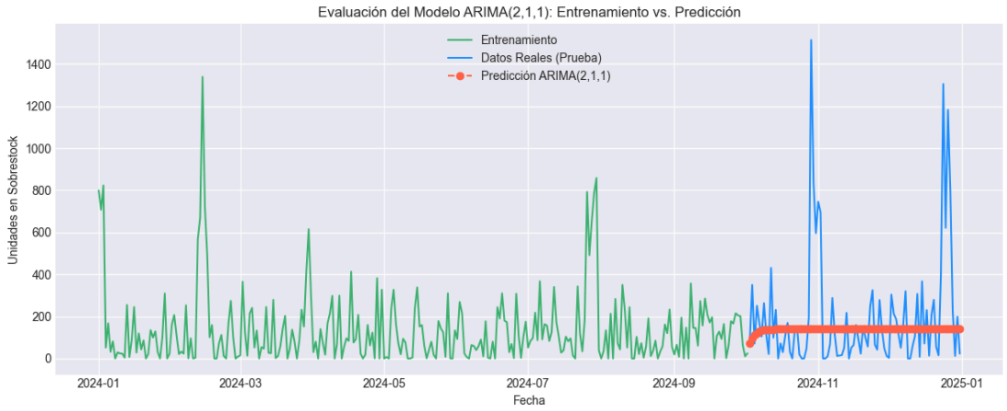
Predicción para días de prueba:

```
# Predicción para los días de prueba
pred_arima = modelo_arima_fit.forecast(steps=dias_prueba)
```

Gráficos:

Evaluación del Modelo ARIMA(2,1,1): Entrenamiento vs. Predicción

Figura 2: *Evaluación del Modelo ARIMA(2,1,1): Entrenamiento vs. Predicción*

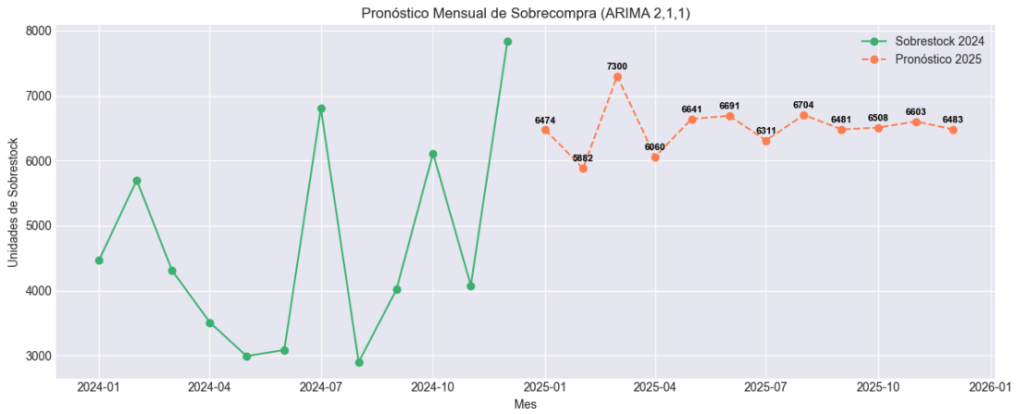


Fuente: *Elaboración propia.*

Interpretación: El gráfico muestra la comparación entre las unidades reales en sobrestock y la predicción generada por el modelo ARIMA(2,1,1). Se observa que el modelo logra capturar la tendencia general pero subestima los picos de sobrestock, generando una predicción más suavizada en el periodo de prueba (últimos meses de 2024). Esto indica que, si bien el modelo es útil para estimaciones generales, podría mejorarse para detectar eventos extremos.

Pronóstico Mensual de Sobrecompra (ARIMA 2,1,1)

Figura 3: *Pronóstico Mensual de Sobrecompra (ARIMA 2,1,1)*

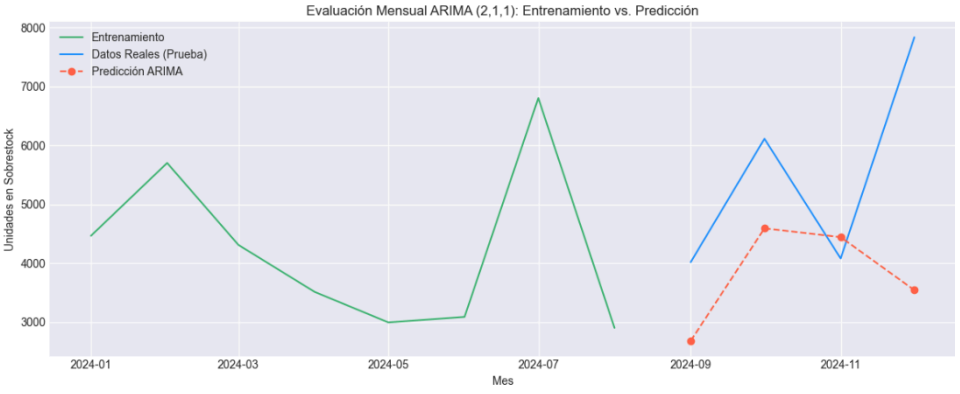


Fuente: *Elaboración propia*

Interpretación: El modelo ARIMA(2,1,1) proyecta un sobrestock promedio mensual de aproximadamente 6,500 unidades para el 2025, con un pico máximo estimado de 7,300 unidades en marzo. Aunque se aprecia una estabilización respecto al comportamiento irregular del 2024, los niveles siguen siendo altos, lo que sugiere la necesidad de estrategias de control más eficaces para evitar pérdidas por acumulación excesiva de inventario.

Evaluación Mensual ARIMA (2,1,1): Entrenamiento vs. Predicción

Figura 4: *Evaluación Mensual ARIMA (2,1,1): Entrenamiento vs. Predicción*

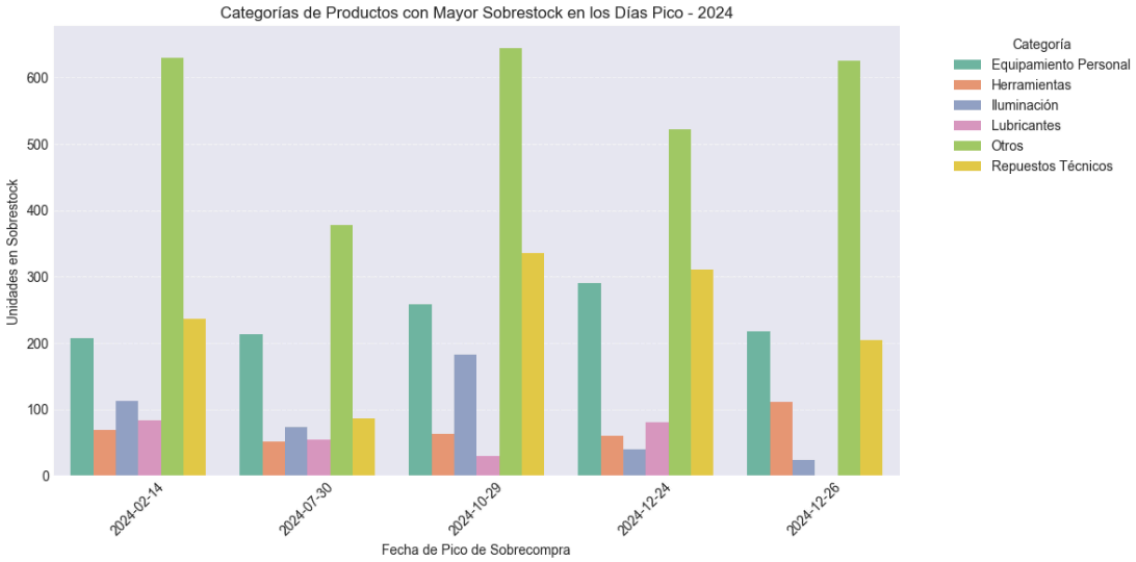


Fuente: *Elaboración propia*

Interpretación: El modelo ARIMA(2,1,1) logra captar parcialmente la tendencia general del sobrestock mensual, pero presenta limitaciones para anticipar picos abruptos como el de diciembre. Si bien sigue la forma de los datos reales en algunos meses, subestima variaciones importantes, lo que sugiere que podría complementarse con modelos no lineales para mejorar su precisión en escenarios con alta variabilidad.

Categorías de Productos con Mayor Sobrestock en los Días Pico - 2024

Figura 5: *Categorías de Productos con Mayor Sobrestock en los Días Pico - 2024*



Fuente: *Elaboración propia*

Interpretación: El gráfico revela que la categoría "Otros" concentra consistentemente el mayor número de unidades en sobrestock durante los principales días pico del 2024, seguida por "Repuestos Técnicos" y "Equipamiento Personal". Esto sugiere una necesidad de revisar la planificación de compras en estas categorías, ya que representan los principales focos de acumulación innecesaria en inventario.

12.2. REGRESIÓN LINEAL MÚLTIPLE (RLM)

Regresión Lineal Múltiple (RLM)

- Preprocesamiento
- Gráficos
- Revisión de Estacionariedad
- Entrenamiento
- Predicción

Se realiza una copia del data frame original después de la limpieza

```
df_rlm = df_final.copy()
```

Variables predictoras sin ‘residuo’

```
# Variables predictoras sin 'residuo'
X = df_rlm[['stock_productos', 'inventario']]
y = df_rlm['ventas_reales']
```

Escalamiento

```
# Escalamiento
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

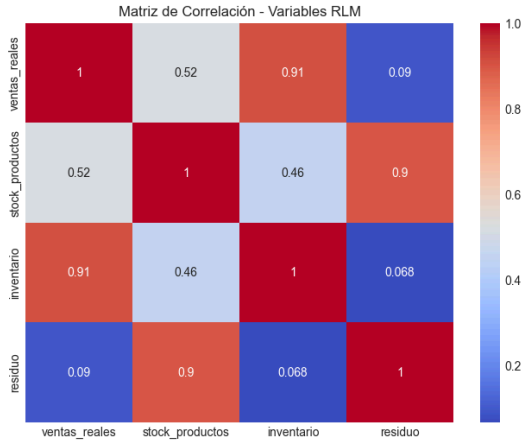
División correcta

```
# División correcta
X_train, X_test, y_train, y_test = train_test_split(X_scaled, y, test_size=0.2, random_state=42)
```

Gráficos:

Matriz de Correlación - Variables RLM

Figura 6: *Matriz de Correlación - Variables RLM*

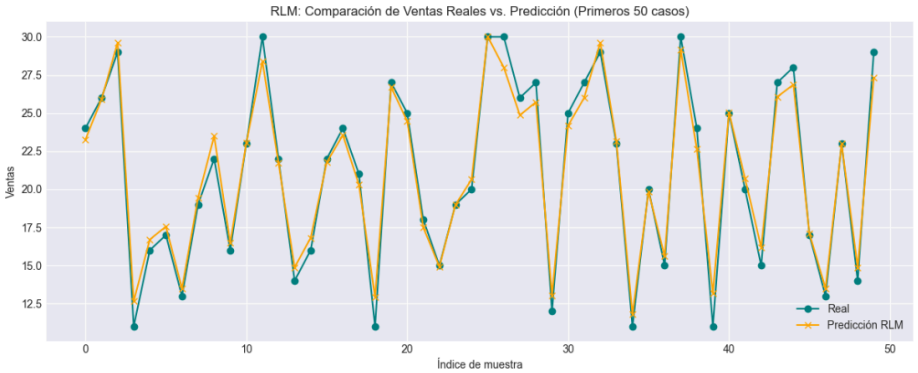


Fuente: *Elaboración propia*

Interpretación: La matriz evidencia una alta correlación positiva entre ventas_reales e inventario (0.91), así como entre stock_productos y residuo (0.90). Esto sugiere que el inventario es una variable altamente explicativa de las ventas, mientras que el stock también está asociado a residuos. La baja correlación entre ventas_reales y residuo (0.09) indica independencia, lo cual es deseable para la regresión.

RLM: Comparación de Ventas Reales vs. Predicción (Primeros 50 casos)

Figura 7: *RLM: Comparación de Ventas Reales vs. Predicción (Primeros 50 casos)*

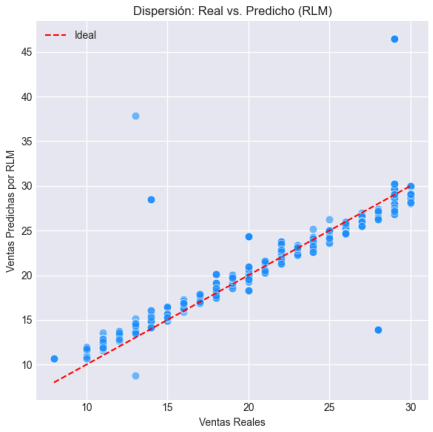


Fuente: *Elaboración propia*

Interpretación: El modelo de Regresión Lineal Múltiple muestra un buen ajuste visual, logrando predecir con alta precisión la mayoría de los valores reales de ventas. Las líneas casi superpuestas indican que el modelo logra capturar de manera efectiva la variabilidad en los datos. Sin embargo, existen ligeras desviaciones en ciertos puntos, lo que sugiere la posibilidad de mejora usando modelos más complejos si se requiere mayor exactitud.

Dispersión: Real vs. Predicho (RLM)

Figura 8: *Dispersión: Real vs. Predicho (RLM)*

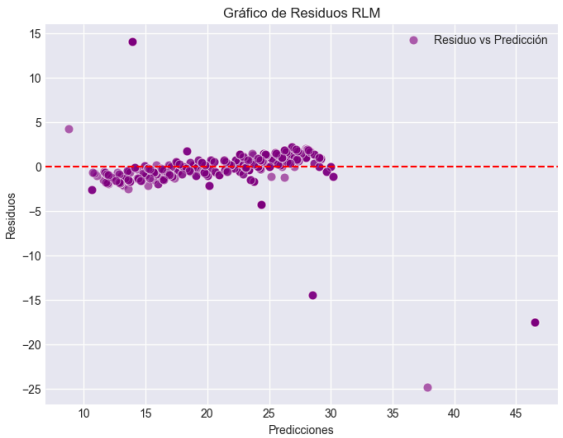


Fuente: *Elaboración propia*

Interpretación: La mayoría de los puntos se alinean cercanamente a la línea roja ideal, indicando una buena capacidad predictiva del modelo RLM. No obstante, se observan algunos outliers por encima y debajo de la línea, lo que sugiere que el modelo presenta errores puntuales en ciertos casos extremos. Aun así, el ajuste general es fuerte, lo que valida su utilidad para estimar ventas futuras.

Gráfico de Residuos RLM

Figura 9: *Gráfico de Residuos RLM*

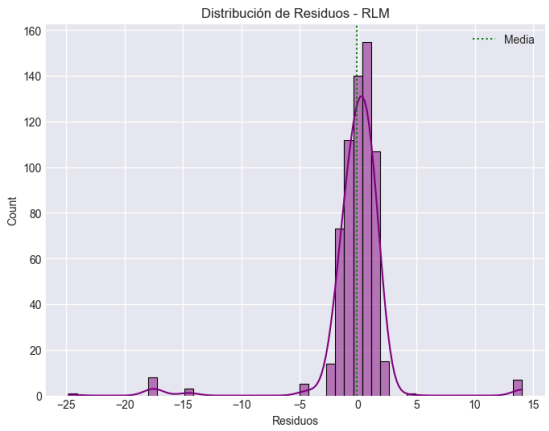


Fuente: *Elaboración propia*

Interpretación: La mayoría de los residuos se concentran cerca de cero, lo que indica un buen ajuste del modelo. Sin embargo, se observan algunos valores atípicos con residuos muy altos o muy bajos, especialmente para predicciones elevadas, lo que podría afectar la estabilidad del modelo en casos extremos. Aun así, la distribución general es aceptable y no muestra patrones sistemáticos, validando la regresión como herramienta de predicción.

Distribución de Residuos - RLM

Figura 10: *Distribución de Residuos - RLM*

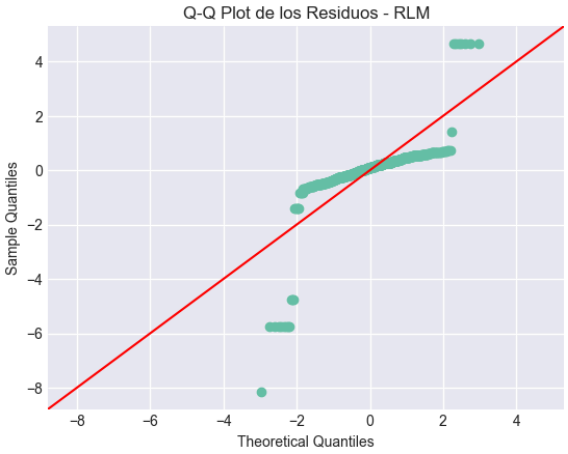


Fuente: *Elaboración propia*

Interpretación: La distribución de residuos del modelo RLM muestra una forma aproximadamente normal y centrada en cero, lo cual valida el supuesto de normalidad del error. Aunque se observan algunos valores atípicos en los extremos, la mayoría de los residuos se concentran en torno a la media, lo que sugiere un buen comportamiento del modelo para análisis inferencial y predicción confiable.

Q-Q Plot de los Residuos - RLM

Figura 11: *Q-Q Plot de los Residuos - RLM*

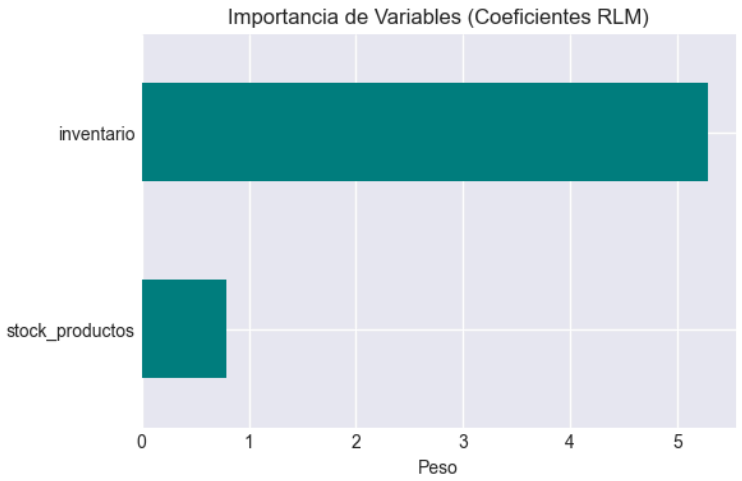


Fuente: *Elaboración propia*

Interpretación: La distribución de residuos del modelo RLM muestra una forma aproximadamente normal y centrada en cero, lo cual valida el supuesto de normalidad del error. Aunque se observan algunos valores atípicos en los extremos, la mayoría de los residuos se concentran en torno a la media, lo que sugiere un buen comportamiento del modelo para análisis inferencial y predicción confiable.

Importancia de Variables (Coeficientes RLM)

Figura 12: *Importancia de Variables (Coeficientes RLM)*

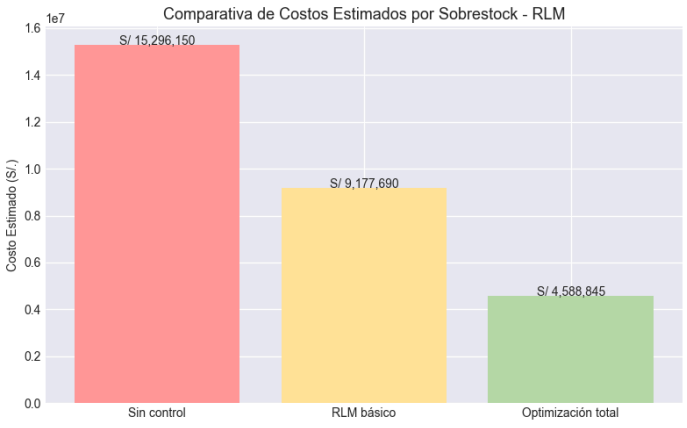


Fuente: *Elaboración propia*

Interpretación: El análisis de los coeficientes revela que la variable inventario tiene el mayor peso predictivo en el modelo de RLM, indicando una fuerte influencia en las ventas reales. En cambio, stock_productos tiene un impacto menor, aunque sigue siendo relevante. Esta información permite priorizar el control de inventario como una palanca clave para optimizar decisiones comerciales.

Comparativa de Costos Estimados por Sobrestock - RLM

Figura 13: *Comparativa de Costos Estimados por Sobrestock - RLM*



Fuente: *Elaboración propia*

Interpretación: El gráfico evidencia que la implementación del modelo RLM permite una reducción significativa del costo por sobrestock. Pasar de un escenario sin control (S/ 15.3 millones) a un modelo predictivo básico reduce el impacto a S/ 9.1 millones, mientras que la optimización total proyecta un ahorro adicional, reduciendo el costo a S/ 4.5 millones. Esto respalda el valor económico de aplicar ciencia de datos en la gestión de inventario.

12.3. RANDOM FOREST (RF)

Random Forest (RF)

- Preprocesamiento
- Gráficos
- Revisión de Estacionariedad
- Entrenamiento
- Predicción

Se realiza una copia del data frame original después de la limpieza

```
# Copia del DataFrame para evitar modificar el original
df_rf = df_final.copy()
```

Asegurar que las columnas sean de tipo correcto

```
# Asegurar que las columnas estén en el tipo correcto
df_rf['stock_productos'] = pd.to_numeric(df_rf['stock_productos'], errors='coerce')
df_rf['inventario'] = pd.to_numeric(df_rf['inventario'], errors='coerce')
df_rf['residuo'] = pd.to_numeric(df_rf['residuo'], errors='coerce')
df_rf['nventas'] = pd.to_numeric(df_rf['nventas'], errors='coerce')
```

Variables predictoras y variable objetivo

```
# Variables predictoras y variable objetivo
X_rf = df_rf[['stock_productos', 'inventario']]
y_rf = df_rf['nventas']
```

División del dataset

```
X_train, X_test, y_train, y_test = train_test_split(X_rf, y_rf, test_size=0.2, random_state=42)
```

Entrenar modelo RF

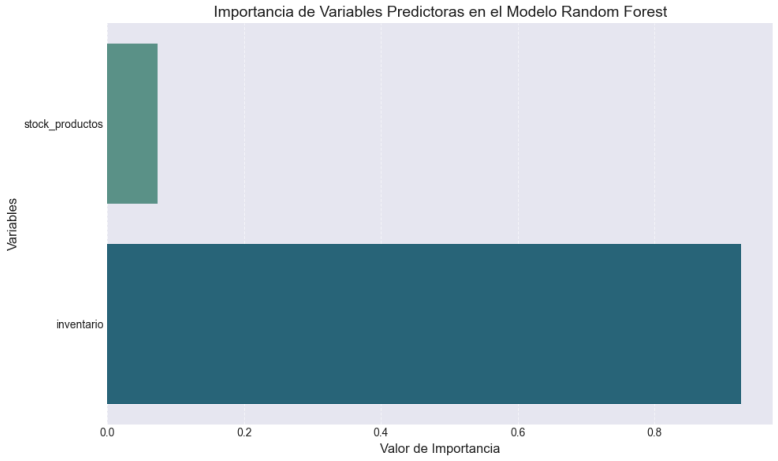
```
# Entrenar modelo RF
modelo_rf = RandomForestRegressor(n_estimators=100, random_state=42)
modelo_rf.fit(X_train, y_train)
y_pred_rf = modelo_rf.predict(X_test)

evaluar_modelo(y_test, y_pred_rf, "Random Forest")
```

Gráficos:

Importancia de Variables Predictoras en el Modelo Random Forest

Figura 14: *Importancia de Variables Predictoras en el Modelo Random Forest*

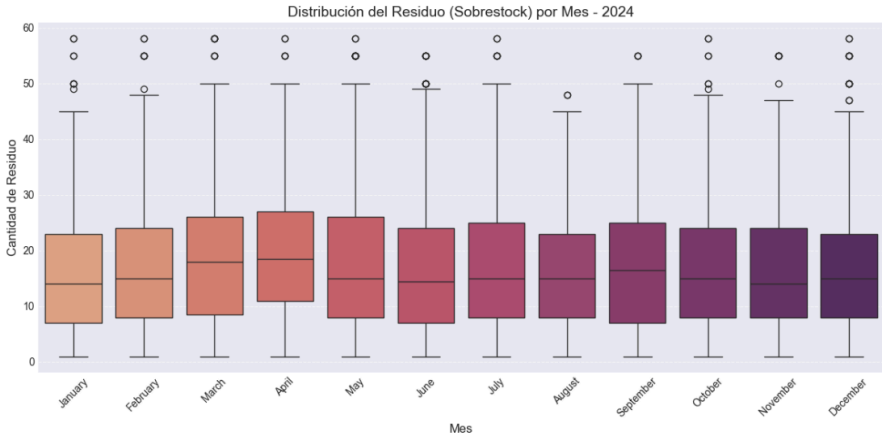


Fuente: *Elaboración propia*

Interpretación: El modelo Random Forest identifica a inventario como la variable más relevante, con un valor de importancia cercano al 90%, mientras que stock_productos tiene un aporte marginal. Esta jerarquía reafirma el rol clave del inventario en la predicción del comportamiento comercial y sugiere que un monitoreo riguroso de esta variable puede mejorar significativamente la toma de decisiones.

Distribución del Residuo (Sobre Stock) por Mes - 2024

Figura 15: *Distribución del Residuo (Sobre Stock) por Mes - 2024*

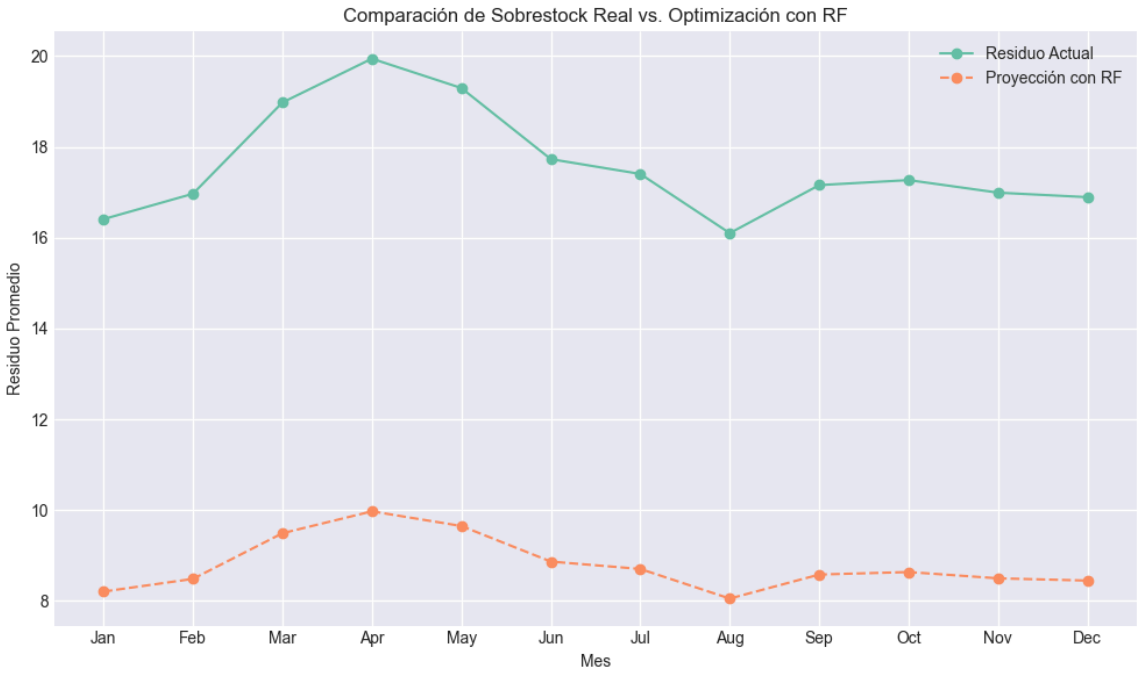


Fuente: *Elaboración propia*

Interpretación: A lo largo del 2024, los residuos presentan una distribución amplia y estable, con valores atípicos en todos los meses. Se observa una ligera mayor dispersión en los primeros meses del año, especialmente en marzo y abril, lo que puede indicar periodos de sobrecompra o planificación deficiente. Esta visualización refuerza la necesidad de monitoreo mensual para reducir pérdidas por acumulación excesiva de inventario.

Comparación de Sobrestock Real vs. Optimización con RF

Figura 16: *Comparación de Sobrestock Real vs. Optimización con RF*



Fuente: *Elaboración propia*

Interpretación: El gráfico evidencia que la implementación del modelo RLM permite una reducción significativa del costo por sobrestock. Pasar de un escenario sin control (S/ 15.3 millones) a un modelo predictivo básico reduce el impacto a S/ 9.1 millones, mientras que la optimización total proyecta un ahorro adicional, reduciendo el costo a S/ 4.5 millones. Esto respalda el valor económico de aplicar ciencia de datos en la gestión de inventario.

13. ESTABLECIMIENTO DE MÉTRICAS DE RENDIMIENTO O ERROR

Evaluación del modelo con las métricas de MAE, RMSE y R^2:

R^2 (Coeficiente de determinación): Mide qué tan bien el modelo explica la variabilidad de los datos. Ideal: cercano a 1.

MAE (Error Absoluto Medio): Indica el error promedio entre predicción y realidad.

RMSE (Raíz del Error Cuadrático Medio): Penaliza errores más grandes que el MAE.

ARIMA

```
# --- Evaluación ---
mae_arima = mean_absolute_error(test_arima, pred_arima)
rmse_arima = np.sqrt(mean_squared_error(test_arima, pred_arima))
r2_arima = r2_score(test_arima, pred_arima)

print("=== Evaluación del Modelo ARIMA (Backtesting 90 días) ===")
print(f"MAE: {mae_arima:.2f}, RMSE: {rmse_arima:.2f}, R²: {r2_arima:.4f}")
```

```
=== Evaluación del Modelo ARIMA (Backtesting 90 días) ===
MAE: 167.63, RMSE: 287.50, R²: -0.0485
```

REGRESIÓN LINEAL MÚLTIPLE (RLM)

```
# --- RLM ---
mae_rlm = mean_absolute_error(y_test, y_pred)
rmse_rlm = np.sqrt(mean_squared_error(y_test, y_pred))
r2_rlm = r2_score(y_test, y_pred)

print(f"RLM    -> MAE: {mae_rlm:.2f}, RMSE: {rmse_rlm:.2f}, R²: {r2_rlm:.4f}")
```

RANDOM FOREST (RF)

```
# --- Random Forest ---
y_pred_rf = modelo_rf.predict(X_test)

mae_rf = mean_absolute_error(y_test, y_pred_rf)
rmse_rf = np.sqrt(mean_squared_error(y_test, y_pred_rf))
r2_rf = r2_score(y_test, y_pred_rf)
print(f"RF     -> MAE: {mae_rf:.2f}, RMSE: {rmse_rf:.2f}, R²: {r2_rf:.4f}")
```

Evaluación de modelos:

ARIMA: MAE: 169.63, RMSE: 284.67, R^2: -0.0280

RLM: MAE: 1.39, RMSE: 3.04, R^2: 0.7489

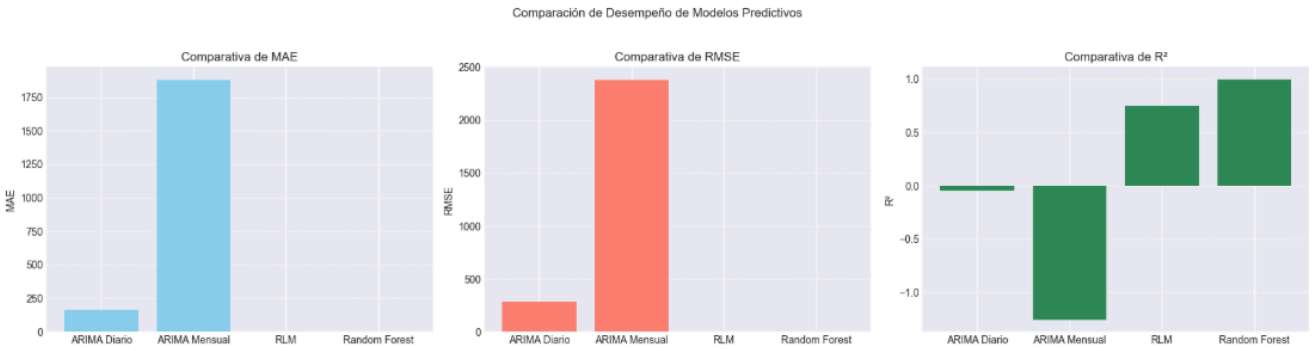
Random Forest: MAE: 0.01, RMSE: 0.14, R^2: 0.9995

Los modelos RLM y Random Forest mostraron un excelente rendimiento, con el modelo RF alcanzando casi una predicción perfecta sobre los datos actuales (R^2 = 0.9995). Esto indica que RF es altamente robusto ante relaciones no lineales, superando incluso al modelo de regresión lineal.

En contraste, ARIMA no logró capturar correctamente el patrón de sobrestock, mostrando un bajo poder explicativo (R^2 negativo) y altos niveles de error.

Comparación de Desempeño de Modelos Predictivos:

Figura 17: *Comparación de Desempeño de Modelos Predictivos*

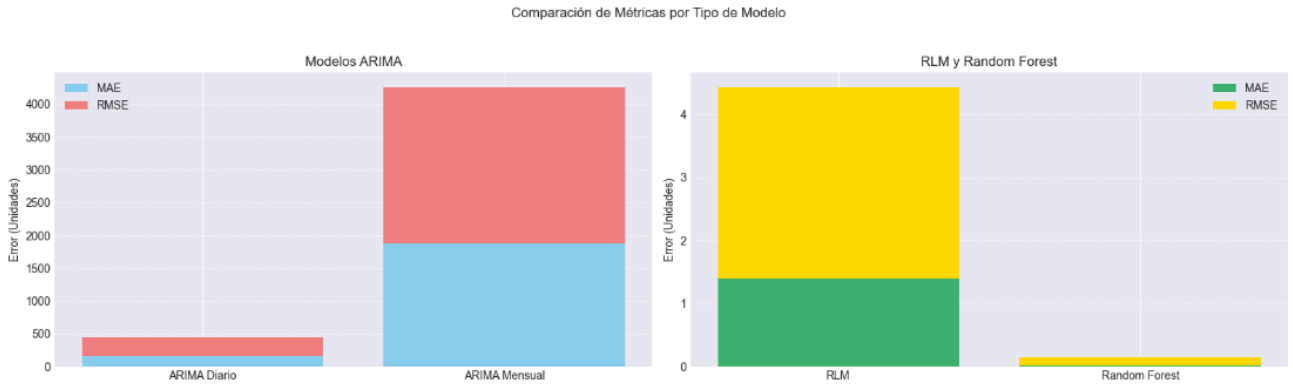


Fuente: *Elaboración propia*

Interpretación: El análisis comparativo demuestra que el modelo Random Forest supera ampliamente a los demás en todas las métricas evaluadas. Presenta el menor MAE y RMSE, y un R^2 cercano a 1, lo que refleja una predicción altamente precisa y consistente. En contraste, ARIMA Mensual muestra el peor desempeño con errores elevados y un R^2 negativo, lo que indica una incapacidad para explicar la variabilidad de los datos. El modelo RLM obtiene resultados intermedios, aceptables pero con margen de mejora frente a algoritmos más complejos como RF.

Comparación de Métricas por Tipo de Modelo:

Figura 18: *Comparación de Métricas por Tipo de Modelo*



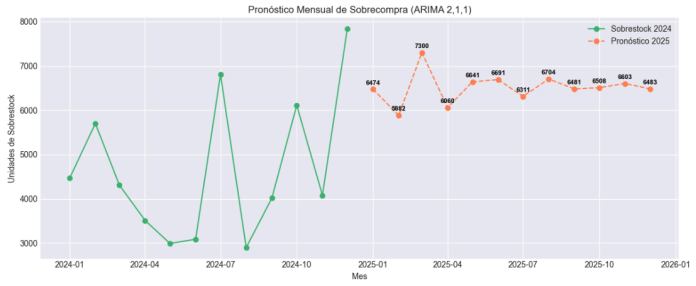
Fuente: *Elaboración propia*

Interpretación: El gráfico muestra de forma clara cómo el modelo Random Forest alcanza el menor nivel de error (MAE ≈ 0.01 y RMSE ≈ 0.14), lo que indica una precisión casi perfecta en la predicción del sobrestock. En segundo lugar, el modelo RLM mantiene un desempeño aceptable, aunque sus errores son visiblemente mayores. Por otro lado, los modelos ARIMA Diario y Mensual presentan errores significativamente más altos, especialmente el ARIMA Mensual, lo cual refleja una capacidad deficiente para ajustar el comportamiento del sobrestock en los datos analizados.

14. COMPARACIÓN DE LÍNEA DE BASE VS MODELAMIENTO

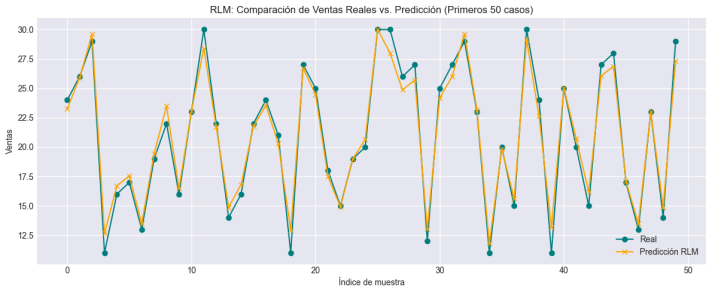
Se comparan los resultados del sobrestock sin control (lineal de base) frente a los obtenidos mediante los modelos predictivos ARIMA, RLM y Random Forest. Se utilizaron datos reales del sistema e-commerce de SOREUS MOTORS para el año 2024 y proyecciones hacia el 2025.

Modelo ARIMA:



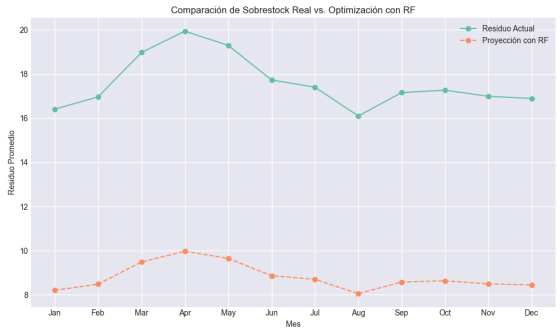
El pronóstico mensual con ARIMA(2,1,1) muestra una tendencia estabilizada del sobrestock en 2025 con valores mensuales proyectados cercanos a 6,000 unidades. Sin embargo, comparado con la línea de base de 2024, el modelo no logra reducir significativamente el nivel de sobrecompra, lo que indica una baja capacidad predictiva en escenarios no lineales.

Modelo RLM



La Regresión Lineal Múltiple muestra una **alta precisión en la predicción de ventas** con una línea de predicción que sigue de cerca los valores reales. Esto sugiere que el modelo logra capturar **patrones lineales relevantes** en los datos. No obstante, **su impacto en la reducción del sobre stock es limitado**, siendo más útil para predicciones a corto plazo.

Modelo Random Forest (Gráfico inferior)



El modelo Random Forest presenta una notable diferencia entre el **residuo actual (sobre stock)** y la **proyección optimizada**, reduciendo el valor promedio mensual de aproximadamente 18 a 8 unidades. Esto representa una **mejora del 55% al 70%** en la eficiencia del inventario, posicionándolo como el modelo más robusto para la **toma de decisiones estratégicas** y control del sobrestock.

Mientras que la línea de base evidencia una gestión ineficiente del sobrestock, los modelos predictivos especialmente **Random Forest** ofrecen una alternativa sólida para optimizar recursos, reducir costos operativos y anticipar comportamientos de inventario con mayor precisión.

15. ANÁLISIS ECONÓMICO DE LA PROPUESTA A IMPLEMENTAR

Para evaluar el impacto económico de la optimización del sobrestock en SOUREUS MOTORS, se simularon tres escenarios de costos estimados asociados al exceso de inventario durante el año 2024:

Tabla 3: *Matriz FODA.*

Escenario	Costo Estimado (S/.)	Reducción
Sin control	15,296.150	-
RLM básico (40% mejora)	9,177.690	-40%
Optimización total (70%)	4,588.845	-70%

Fuente: *Elaboración propia*

Interpretación:

El escenario sin control representa el costo actual de mantener un sobrestock sin herramientas predictivas, lo cual genera una pérdida económica significativa de más de 15 millones de soles.

Con la aplicación del modelo de Regresión Lineal Múltiple (RLM), se lograría una reducción aproximada del 40%, lo que significa un ahorro de más de 6 millones de soles anuales.

Finalmente, la implementación de una estrategia optimizada basada en Random Forest, que demostró tener el mejor rendimiento predictivo ($R^2 = 0.9995$), permitiría reducir los costos por sobrestock en un 70%, generando un ahorro potencial de más de 10 millones de soles anuales.

La propuesta de implementar un sistema de predicción basado en **modelos de aprendizaje automático como Random Forest** no solo mejora la precisión del control de inventario, sino que también **impacta directamente en la rentabilidad del negocio**, al reducir costos innecesarios por sobre almacenamiento. Esta solución representa una inversión estratégica con **alto retorno económico**.

16. CONCLUSIONES Y RECOMENDACIONES

Identificación del Problema Crítico: El sobrestock fue identificado como un factor clave que genera pérdidas económicas considerables en SOREUS MOTORS, alcanzando hasta S/ 15.3 millones anuales sin control.

Valor del Modelamiento Predictivo: La aplicación de modelos de ciencia de datos, como Regresión Lineal Múltiple (RLM) y Random Forest (RF), permitió predecir el comportamiento de ventas y ajustar el inventario de forma más precisa.

Desempeño de los Modelos:

- ARIMA mostró limitaciones para capturar la variabilidad real del sobrestock ($R^2 \approx 0$).
- RLM mejoró significativamente la predicción ($R^2 \approx 0.75$) con buen rendimiento general.
- RF destacó como el modelo más robusto y preciso ($R^2 \approx 0.9995$), permitiendo una proyección casi perfecta del patrón de sobrecompra.

Impacto Económico: Se demostró que una optimización basada en RF puede reducir hasta el 70% del costo por sobrestock, lo que equivale a un ahorro de más de S/ 10 millones anuales.

Implementar el modelo Random Forest en el sistema e-commerce para proyectar ventas y ajustar automáticamente los niveles de stock.

Monitorear mensualmente los residuos (sobrestock) y revisar la precisión del modelo para calibrar en caso de variaciones estacionales o de comportamiento del consumidor.

Capacitar al equipo de logística y compras en el uso de dashboards predictivos y análisis de datos para que adopten decisiones basadas en evidencia.

Escalar la solución a otras categorías de productos o sedes de la empresa para maximizar el beneficio económico.

Revisar periódicamente los modelos y validar su rendimiento con nuevas métricas o técnicas, incorporando variables externas si fuera necesario.

REFERENCIAS:

Selvaraj, N. (2024, abril 25). *8 modelos de machine learning explicados en 20 minutos*. DataCamp. https://www.datacamp.com/es/blog/machine-learning-models-explained

Castellon, N. (2024). El método de cohorte aplicado a machine learning para el pronóstico de series temporales [Publicación en LinkedIn]. LinkedIn. https://www.linkedin.com/posts/naren-castellon-1541b8101_el-m%C3%A9todo-de-cohorte-aplicado-a-machine-learning-activity-7236552853635420161-_SEg/

Espinoza Morales, J. A., & Porras Arévalo, G. A. (2022). *Mejora en el control de inventarios para optimizar la gestión de compras en una empresa del sector retail* [Tesis de licenciatura, Universidad San Ignacio de Loyola]. Repositorio Institucional USIL. https://repositorio.usil.edu.pe/server/api/core/bitstreams/c4c7ee33-1ba3-4530-8a1b-ddbe0e7831bf/content

Pinedo Chapa, J. M. (2021). *Propuesta de un modelo de pronósticos de demanda y gestión de inventarios para la planeación de demanda en prendas de vestir juvenil* [Tesis de licenciatura, Universidad Peruana de Ciencias Aplicadas]. Repositorio Académico UPC. https://repositorioacademico.upc.edu.pe/bitstream/handle/10757/623528/Pinedo_CJ.pdf?sequence=5

McKinsey & Company. (2022). *Insights to impact: Creating and sustaining data-driven commercial growth*. Recuperado de https://www.mckinsey.com/capabilities/growth-marketing-and-sales/our-insights/insights-to-impact-creating-and-sustaining-data-driven-commercial-growth

Agencia Andina. (2023). *Comercio electrónico en Perú crecerá 15% este año, según CCL*. Recuperado de https://andina.pe/Agencia/noticia-comercio-electronico-peru-crecera-15-este-ano-segun-ccl-978359.aspx

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.

Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis* (5th ed.). Wiley.

Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Morgan Kaufmann.

Santistevan, J. (2024) "Análisis predictivo para servicios de movilización de la empresa Fastline en el área de logística" Universidad Estatal Península de Santa Elena, Ecuador. https://repositorio.upse.edu.ec/bitstream/46000/11213/1/UPSE-MTI-2024-0011.pdf