Aplicación de modelo predictivo para análisis de ventas aplicado a una cafetería mediante la implementación de técnicas de aprendizaje automático y series temporales (Machine Learning)

Diego Andrés Martínez Cardozo

Germán Danilo Alarcón García

Asesor

Jorge Luis Quintero López

Julio Eduardo Mejía Manzano

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas Tecnología e ingeniería ECBTI

Especialización en Ciencia de Datos y Analítica

2025

_____

Nombre Director de Trabajo de Grado

_____ _____

Jurado Jurado

2025

Resumen

Este trabajo desarrolla un modelo predictivo de ventas para una cafetería familiar, buscando identificar patrones de consumo y proyectar la demanda a través de técnicas avanzadas de analítica de datos. El problema que se aborda es la ausencia de herramientas objetivas para anticipar la demanda de productos, lo cual genera decisiones poco acertadas que derivan en sobreproducción, pérdidas económicas por desperdicio o desabastecimiento, deteriorando la competitividad del negocio frente a cadenas consolidadas.

Para ello, se propone un enfoque metodológico que integra técnicas clásicas de series temporales (ARIMA) con algoritmos de aprendizaje automático supervisado (Regresión Lineal y Random Forest), siguiendo la metodología SEMMA. Esta combinación permite capturar los patrones de autocorrelación temporal y, al mismo tiempo, modelar las relaciones complejas entre variables exógenas contextuales.

La metodología se desarrolla en cinco fases. Primero, en Sample se recopilan y organizan los datos históricos de transacciones, incluyendo variables temporales, productos y características operativas. Luego, en Explore se realiza un análisis exploratorio para identificar patrones mediante descomposición de series temporales y pruebas de estacionaridad. En la fase Modify se preprocesan los datos y se aplica ingeniería de características, creando variables derivadas de la fecha como día de semana, mes, trimestre y variables cíclicas. Posteriormente, en Model se entrenan y optimizan los modelos de series temporales y de aprendizaje automático, empleando validación temporal para evitar fuga de datos. Finalmente, en Assess se evalúa comparativamente el desempeño predictivo mediante métricas de precisión como MAE, RMSE y R^2, seleccionando así el modelo más adecuado.

Los modelos de series temporales resultan especialmente útiles para capturar la estacionalidad inherente al comportamiento de compra, mientras que los algoritmos de aprendizaje automático supervisado facilitan la integración de múltiples variables predictoras, tales como horario de operación, categoría de producto y precio unitario. La evaluación de los modelos considera no solo métricas cuantitativas de error, sino también su interpretabilidad y aplicabilidad práctica en la toma de decisiones.

El análisis se realizó sobre 65,193 transacciones recopiladas durante 111 días (mayo-agosto 2025), abarcando 492 productos únicos y 40 clientes. Los resultados demuestran que los algoritmos de aprendizaje automático superaron significativamente a los modelos de series temporales tradicionales. La Regresión Lineal alcanzó un desempeño perfecto para predicción de cantidad (R^2=1.000, MAE=0.00), mientras que Random Forest demostró ser el modelo más robusto para predicción de ingresos (R^2=0.978, MAE=36.09, RMSE=1,245.21). En contraste, los modelos ARIMA presentaron desempeño limitado con valores negativos de R^2, indicando que no capturaron adecuadamente los patrones complejos presentes en los datos.

Estos resultados permiten generar proyecciones confiables que apoyan la gestión operativa en diversas dimensiones: optimización de inventarios mediante la identificación de patrones de demanda por producto y horario, mejora en la planificación de personal con base en las horas pico de actividad, reducción de desperdicios ajustando la producción a la demanda real, y diseño de estrategias comerciales fundamentadas en el comportamiento histórico de consumo. El proyecto demuestra que es viable implementar metodologías de ciencia de datos en pequeños negocios familiares, ofreciendo una alternativa accesible para competir a través de decisiones basadas en datos y fortaleciendo su administración, rentabilidad y sostenibilidad en un mercado cada vez más competitivo.

Palabras clave: Predicción de ventas, Machine Learning, Series temporales, ARIMA, Random Forest, Estacionalidad, Metodología SEMMA, Análisis de datos, Optimización, Minería de datos.

Abstract

This work develops a predictive sales model for a family-owned café, seeking to identify consumption patterns and forecast demand through advanced data analytics techniques. The problem addressed is the lack of objective tools to anticipate product demand, which leads to poorly informed decisions resulting in overproduction, economic losses due to waste or stockouts, undermining the business's competitiveness against established chains.

To achieve this, a methodological approach is proposed that integrates classical time series techniques (ARIMA) with supervised machine learning algorithms (Linear Regression and Random Forest), following the SEMMA methodology. This combination allows capturing temporal autocorrelation patterns while modeling complex relationships between contextual exogenous variables.

The methodology unfolds in five phases. First, in Sample, historical transaction data is collected and organized, including temporal variables, products, and operational characteristics. Then, in Explore, an exploratory analysis is conducted to identify patterns through time series decomposition and stationarity tests. In the Modify phase, data is preprocessed and feature engineering is applied, creating date-derived variables such as day of week, month, quarter, and cyclical variables. Subsequently, in Model, both time series and machine learning models are trained and optimized, employing temporal validation to prevent data leakage. Finally, in Assess, predictive performance is comparatively evaluated using precision metrics such as MAE, RMSE, and R^2, thus selecting the most suitable model.

Time series models are particularly useful for capturing seasonality inherent to purchasing behavior, while supervised machine learning algorithms facilitate the integration of multiple predictor variables, such as operating hours, product category, and unit price. Model

evaluation considers not only quantitative error metrics but also their interpretability and practical applicability in decision-making.

The analysis was performed on 65,193 transactions collected over 111 days (May-August 2025), covering 492 unique products and 40 customers. Results demonstrate that machine learning algorithms significantly outperformed traditional time series models. Linear Regression achieved perfect performance for quantity prediction (R^2=1.000, MAE=0.00), while Random Forest proved to be the most robust model for revenue prediction (R^2=0.978, MAE=36.09, RMSE=1,245.21). In contrast, ARIMA models showed limited performance with negative R^2 values, indicating they did not adequately capture the complex patterns present in the data.

These results enable generating reliable forecasts that support operational management across various dimensions: inventory optimization through identification of demand patterns by product and time slot, improved staff planning based on peak activity hours, waste reduction by adjusting production to actual demand, and design of commercial strategies grounded in historical consumption behavior. The project demonstrates that implementing data science methodologies in small family businesses is feasible, offering an accessible alternative to compete through data-driven decisions and strengthening their management, profitability, and sustainability in an increasingly competitive market.

Keywords: Sales forecasting, Machine Learning, Time series, ARIMA, Random Forest, Seasonality, SEMMA methodology, Data analysis, Optimization, Data mining.

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Lista de Apéndices

Introducción

El café representa mucho más que una bebida en la cultura latinoamericana. En Colombia, las cafeterías se han consolidado como espacios fundamentales de encuentro social, tradición y cercanía con la comunidad, funcionando como extensiones del hogar y centros neurálgicos de la vida cotidiana urbana. Según la Federación Nacional de Cafeteros (2024), Colombia cuenta con más de 15,000 establecimientos dedicados a la comercialización de café y productos complementarios, generando aproximadamente 200,000 empleos directos e indirectos. Sin embargo, este sector enfrenta una transformación acelerada impulsada por la llegada de cadenas internacionales y la digitalización de los hábitos de consumo.

Aunque el café sigue siendo el producto emblemático, los establecimientos colombianos han diversificado significativamente su oferta, incorporando panes artesanales, postres, bebidas frías, jugos naturales, sándwiches y productos de repostería que se han integrado orgánicamente a los hábitos de consumo diarios. Esta diversificación no responde únicamente al gusto de los clientes, sino que constituye una estrategia de supervivencia financiera que complementa los ingresos por venta de café y asegura flujos de caja más estables para estos negocios (Gómez & Rodríguez, 2022).

No obstante, las cafeterías locales y familiares enfrentan una competencia cada vez más intensa. Cadenas consolidadas como Tostao' Café & Pan, Juan Valdez Café o Starbucks operan con ventajas sustanciales: estrategias de mercadeo respaldadas por investigación de mercados, sistemas avanzados de gestión de inventarios, análisis predictivo de demanda y modelos de negocio estandarizados que optimizan costos operativos. Esta asimetría competitiva coloca a los pequeños negocios familiares en una posición vulnerable, donde la intuición y la experiencia empírica, aunque valiosas, resultan insuficientes para competir efectivamente (Porter, 2008). En

este contexto, la capacidad de anticipar la demanda de productos mediante herramientas analíticas se convierte en un factor estratégico fundamental.

Diversos estudios han demostrado que la aplicación de técnicas de predicción de ventas en el sector retail alimentario puede reducir desperdicios entre 15% y 30%, mejorar la rotación de inventarios hasta en 25% y aumentar la satisfacción del cliente al garantizar disponibilidad de productos en momentos clave (Choi et al., 2018; Taylor & Letham, 2018). A pesar de estos beneficios documentados, la mayoría de las cafeterías familiares en Colombia carecen de acceso a estas metodologías, manteniéndose en desventaja frente a competidores con mayor capacidad tecnológica.

La ciencia de datos ha democratizado progresivamente el acceso a herramientas analíticas avanzadas. Técnicas que hace una década requerían infraestructura costosa y equipos especializados, hoy resultan accesibles mediante bibliotecas de código abierto en Python como scikit-learn, statsmodels o Prophet (Pedregosa et al., 2011; Taylor & Letham, 2018).

Esta democratización tecnológica abre oportunidades sin precedentes para que pequeños negocios implementen sistemas de apoyo a la toma de decisiones fundamentados en datos, nivelando parcialmente el campo de juego competitivo. Las herramientas computacionales modernas permiten que emprendedores sin formación técnica avanzada puedan beneficiarse de metodologías robustas previamente reservadas para grandes corporaciones.

Este proyecto surge de la convergencia entre necesidad empresarial y viabilidad tecnológica. Se propone desarrollar un modelo de predicción de ventas que combine técnicas de series temporales y algoritmos de aprendizaje automático supervisado, aplicable específicamente al contexto de una cafetería familiar. A partir del análisis de datos históricos de transacciones, se busca identificar patrones de tendencia, estacionalidad y frecuencia en las ventas, integrando

variables adicionales relacionadas con contexto temporal y características operativas como día de la semana, horarios, fines de semana y festivos.

La aproximación metodológica se estructura siguiendo SEMMA (Sample, Explore, Modify, Model, Assess), un marco de trabajo consolidado para proyectos de minería de datos propuesto por SAS Institute que garantiza sistematicidad y reproducibilidad en el proceso analítico (SAS Institute, 1998). Esta metodología guía la construcción del modelo en sus fases de muestreo de datos, exploración de patrones, modificación y creación de variables derivadas, modelado mediante algoritmos diversos y evaluación comparativa de desempeño.

El enfoque dual propuesto integra lo mejor de dos paradigmas complementarios. Por un lado, los modelos de series temporales como ARIMA capturan efectivamente la estructura autocorrelativa y estacional inherente a datos secuenciales, identificando patrones que se repiten sistemáticamente en el tiempo (Box et al., 2015; Hyndman & Athanasopoulos, 2018). Por otro lado, los algoritmos de aprendizaje automático supervisado como Regresión Lineal y Random Forest permiten incorporar múltiples variables predictoras simultáneamente, modelando relaciones no lineales complejas que los modelos tradicionales no capturan (Breiman, 2001).

Este trabajo obtuvo proyecciones confiables de ventas que respaldan concretamente la gestión operativa de la cafetería, contribuyendo a mejorar la eficiencia en el control de inventarios, la planificación de dotación de personal, la reducción de mermas y desperdicios, y el diseño de promociones fundamentadas en patrones reales de comportamiento del consumidor. Los resultados confirmaron la superioridad de los algoritmos de aprendizaje automático sobre los modelos tradicionales de series temporales, alcanzando niveles de precisión superiores al 97% en la predicción de ingresos mediante Random Forest, mientras que la Regresión Lineal logró predicciones prácticamente perfectas para cantidades (R^2=1.000). Más allá del caso específico,

este proyecto demuestra la viabilidad práctica de implementar metodologías rigurosas de ciencia de datos en pequeños negocios familiares, ofreciendo un modelo replicable que fortalezca su rentabilidad y sostenibilidad.

El presente documento se estructura de la siguiente manera: primero se expone la justificación que fundamenta la relevancia del proyecto, seguida de los objetivos general y específicos que orientan el trabajo. Posteriormente se presenta el planteamiento del problema que contextualiza la situación a resolver. Los marcos conceptual y teórico establecen las bases académicas y técnicas del proyecto, mientras que la sección de metodología detalla el proceso sistemático seguido. La sección de resultados presenta los hallazgos obtenidos del análisis de 65,193 transacciones y la comparación de desempeño entre modelos. Finalmente, se presentan las conclusiones derivadas del análisis y las recomendaciones para trabajo futuro e implementación práctica.

Justificación

Las cafeterías familiares en Colombia enfrentan hoy un desafío concreto: cómo competir con cadenas que cuentan con recursos superiores, sistemas avanzados de análisis y estrategias respaldadas por datos. Mientras empresas como Juan Valdez o Starbucks fundamentan sus decisiones operativas en modelos predictivos sofisticados, los pequeños negocios siguen dependiendo principalmente de la intuición y la experiencia empírica. Esta diferencia los coloca en clara desventaja.

El problema central es la falta de herramientas estructuradas para anticipar cómo se comportarán las ventas. Sin estas herramientas, las decisiones se toman casi a ciegas, provocando dos escenarios igualmente perjudiciales. Por un lado, la sobreproducción genera desperdicios de productos perecederos que pueden representar hasta 30% del inventario en establecimientos sin gestión predictiva (Choi et al., 2018). Por otro lado, el desabastecimiento resulta en ventas perdidas y clientes insatisfechos que probablemente no regresarán. Ambos escenarios erosionan la rentabilidad y comprometen la sostenibilidad del negocio a mediano plazo.

¿Por qué es importante este trabajo? Porque demuestra algo que muchos pequeños empresarios desconocen: las metodologías robustas de ciencia de datos, que antes eran exclusivas de grandes corporaciones, ahora son accesibles y aplicables incluso en contextos de pequeña escala. La investigación reciente muestra que modelos predictivos relativamente simples pueden mejorar la precisión de pronósticos de demanda entre 20% y 40% comparado con métodos tradicionales, y estas mejoras se traducen en reducciones tangibles de costos operativos (Hyndman & Athanasopoulos, 2018; Taylor & Letham, 2018).

¿Quiénes se benefician de esto? Los propietarios obtienen una base objetiva para optimizar inventarios, planificar personal según horas pico y diseñar promociones

fundamentadas en patrones reales. Los empleados ganan estabilidad gracias a horarios mejor planificados. Los clientes encuentran mayor disponibilidad de productos y mejor servicio. Incluso hay un beneficio ambiental al reducir desperdicios de alimentos mediante producción ajustada a la demanda real.

La decisión de combinar series temporales con aprendizaje automático no es casual. Cada enfoque captura algo diferente pero complementario: los modelos temporales identifican estacionalidad y autocorrelación (esos patrones que se repiten sistemáticamente), mientras que los algoritmos de machine learning modelan relaciones complejas entre múltiples variables como precio, hora del día y día de la semana. Esta combinación ha demostrado superar consistentemente el desempeño de aproximaciones que usan solo una técnica (Géron, 2019).

Además, este proyecto contribuye a llenar un vacío importante en la literatura. Aunque existen numerosos estudios sobre predicción de ventas en retail de gran escala, hay poca investigación aplicada específicamente a pequeños negocios familiares del sector alimentario en contextos latinoamericanos. Lo que se presenta aquí es un modelo replicable, documentado y técnicamente riguroso que otros pequeños empresarios pueden adaptar a sus propias realidades.

La justificación se sostiene en tres elementos. Primero, una necesidad empresarial real: herramientas analíticas para competir efectivamente. Segundo, viabilidad técnica demostrada de implementar estas metodologías con recursos limitados. Tercero, impacto tangible esperado en rentabilidad, eficiencia operativa y sostenibilidad. Al fortalecer la capacidad de tomar decisiones basadas en datos, se contribuye a preservar espacios que son parte esencial del tejido social urbano colombiano: esas cafeterías que funcionan como lugares de encuentro y tradición comunitaria.

Objetivos

Objetivo General

Desarrollar un modelo predictivo de ventas para una cafetería familiar que integre técnicas de series temporales (ARIMA) y algoritmos de aprendizaje automático supervisado (Regresión Lineal, Random Forest) siguiendo la metodología SEMMA.

Objetivos Específicos

Explorar los datos históricos de ventas en función de patrones de tendencia, estacionalidad, frecuencia de consumo y variables temporales u operativas.

Entrenar modelos predictivos con técnicas de series temporales (ARIMA) y algoritmos de aprendizaje automático supervisado (Regresión Lineal, Random Forest) aplicando la metodología SEMMA.

Comparar el desempeño de los modelos a través de métricas de precisión (MAE, RMSE, R^2) para fundamentar recomendaciones operativas sobre inventarios, personal y estrategias comerciales.

Planteamiento del Problema

Las cafeterías familiares enfrentan un desafío operativo fundamental: tomar decisiones sobre producción, inventarios y personal sin contar con información objetiva que respalde esas decisiones. La gestión se basa principalmente en la intuición del propietario y en la experiencia acumulada, métodos que, aunque valiosos, muestran limitaciones importantes cuando los datos históricos están desorganizados o simplemente no existen de forma estructurada.

Esta ausencia de información procesable tiene consecuencias concretas. La sobreproducción lleva a desperdicios de productos perecederos que, según estudios del sector retail alimentario, pueden alcanzar hasta 30% del costo de inventario en establecimientos sin gestión predictiva (Choi et al., 2018). El desabastecimiento, por su parte, genera ventas perdidas y deteriora la experiencia del cliente. Ambos escenarios afectan directamente los márgenes de rentabilidad, que ya son ajustados en este sector, y comprometen la sostenibilidad del negocio.

El problema se complica porque la demanda en cafeterías es inherentemente variable. El día de la semana importa: un lunes no se comporta igual que un sábado. El clima influye. Los eventos locales generan picos inesperados. Las temporadas académicas modifican patrones. Esta variabilidad hace que la intuición humana, por experimentada que sea, difícilmente pueda procesar todos estos factores de manera sistemática y consistente.

Aquí es donde la ciencia de datos ofrece una alternativa práctica. Las series temporales capturan bien la estacionalidad (esa diferencia entre días de semana y fines de semana, por ejemplo) y las tendencias de largo plazo. Los algoritmos de machine learning van más allá: pueden modelar simultáneamente cómo interactúan múltiples variables. Random Forest y otros algoritmos similares detectan patrones que permanecen ocultos en análisis tradicionales, como la relación compleja entre hora del día, día de la semana y condiciones climáticas (Géron, 2019).

La metodología SEMMA proporciona estructura a este proceso. Desde la exploración inicial hasta la evaluación rigurosa de resultados, este enfoque sistemático garantiza que las decisiones se fundamenten en evidencia cuantificable, no en suposiciones. Más importante aún, la metodología asegura que el proceso sea reproducible y documentado, facilitando su transferencia a otros contextos similares (SAS Institute, 1998).

Lo que se busca entonces es diseñar un modelo de pronóstico específicamente adaptado al contexto de una cafetería familiar. Un modelo que no solo prediga con precisión, sino que genere información accionable: cuánto producir en cada franja horaria, cuándo ajustar dotaciones de personal, qué productos impulsar en momentos específicos. La precisión estadística importa, pero importa más que las predicciones puedan traducirse en decisiones operativas concretas que mejoren la rentabilidad.

Los beneficios esperados trascienden las métricas de desempeño. Un modelo bien implementado permite optimizar la producción por horarios, reducir capital inmovilizado en inventarios de baja rotación, ajustar compras a proveedores con mayor precisión, diseñar promociones estratégicas en horarios valle y establecer niveles de seguridad diferenciados por producto y día. Todo esto no solo mejora la rentabilidad sino que aumenta la capacidad del negocio para adaptarse en un mercado dinámico (Provost & Fawcett, 2013).

El problema que aborda este trabajo es, en esencia, la brecha entre lo que pequeños negocios necesitan y lo que actualmente pueden hacer. Las grandes cadenas tienen departamentos de analítica; las cafeterías familiares tienen hojas de cálculo desorganizadas. Al desarrollar un modelo predictivo accesible, documentado y replicable, se busca reducir esa brecha, demostrando que herramientas robustas de ciencia de datos son viables incluso con

recursos limitados. El objetivo no es solo hacer predicciones precisas, sino fortalecer la capacidad de toma de decisiones fundamentadas en datos.

Marco Conceptual

El desarrollo de este proyecto requiere clarificar conceptos fundamentales que constituyen la base técnica del trabajo. A continuación, se presentan las definiciones operacionales de los elementos centrales del modelo predictivo.

Series Temporales

Una serie temporal es una secuencia de observaciones ordenadas cronológicamente, medidas en intervalos regulares o irregulares. En este proyecto, las ventas diarias de la cafetería constituyen una serie temporal discreta que puede descomponerse en cuatro componentes: tendencia (patrón direccional de largo plazo), ciclicidad (fluctuaciones no periódicas), estacionalidad (patrones repetitivos en intervalos fijos) y componente irregular o ruido (variaciones aleatorias no explicadas).

Modelo ARIMA

ARIMA (Autoregressive Integrated Moving Average) es un modelo estadístico que describe la autocorrelación en series temporales mediante tres parámetros: p (orden autorregresivo que indica cuántos valores pasados se utilizan), d (grado de diferenciación necesario para lograr estacionaridad), y q (orden de media móvil que captura la dependencia de errores pasados). Su extensión SARIMA incorpora componentes estacionales mediante parámetros adicionales (P, D, Q, m), donde m representa la frecuencia estacional.

Aprendizaje Automático Supervisado

El aprendizaje automático supervisado es un paradigma donde un algoritmo aprende una función de mapeo entre variables de entrada (features) y una variable objetivo (target) utilizando ejemplos etiquetados. En este proyecto se aplica para tareas de regresión, donde la variable objetivo es numérica continua (cantidad de productos o ingresos por ventas).

Regresión Lineal

Modelo estadístico que establece una relación lineal entre una variable dependiente y una o más variables independientes mediante la ecuación $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \varepsilon$, donde los coeficientes β se estiman minimizando la suma de errores cuadrados. Su ventaja principal es la interpretabilidad directa de los coeficientes.

Random Forest

Algoritmo de ensamble que construye múltiples árboles de decisión durante el entrenamiento y produce predicciones promediando los resultados individuales. Utiliza dos técnicas de aleatorización: bootstrap aggregating (cada árbol se entrena con una muestra aleatoria con reemplazo) y selección aleatoria de características en cada división. Esta combinación reduce la varianza del modelo sin incrementar significativamente el sesgo.

Ingeniería de Características

Proceso de transformación de datos crudos en representaciones que facilitan el aprendizaje de patrones por parte de los algoritmos. Para series temporales incluye la extracción de variables derivadas como día de la semana, mes, trimestre, indicadores binarios (fin de semana, festivo) y transformaciones cíclicas mediante funciones trigonométricas.

Métricas de Evaluación

Las métricas cuantifican el desempeño predictivo de los modelos:

- MAE (Mean Absolute Error): Promedio de los valores absolutos de los errores, interpretable directamente en las unidades de la variable objetivo.
- RMSE (Root Mean Squared Error): Raíz cuadrada del promedio de errores al cuadrado, que penaliza más fuertemente los errores grandes.

- R^2 (Coeficiente de Determinación): Proporción de varianza explicada por el modelo, donde valores cercanos a 1 indican mejor ajuste.
- MAPE (Mean Absolute Percentage Error): Error porcentual promedio, útil para comparaciones independientes de escala.

Metodología SEMMA

Marco estructurado para proyectos de minería de datos que comprende cinco fases: Sample (muestreo de datos representativos), Explore (análisis exploratorio y visualización), Modify (transformación y creación de variables), Model (construcción de modelos predictivos), y Assess (evaluación comparativa de desempeño). Esta metodología garantiza reproducibilidad y sistematicidad en el proceso analítico.

Marco Teórico

Evolución de la Teoría de Series Temporales

Los fundamentos de las series temporales surgen en las primeras décadas del siglo XX con los trabajos de Yule (1927) sobre procesos autorregresivos y Slutsky (1937) sobre procesos de media móvil. Ambos investigadores establecieron que las observaciones consecutivas en el tiempo no son independientes, sino que exhiben correlación serial, rompiendo con el supuesto clásico de independencia en estadística.

El avance decisivo llegó con Box y Jenkins (1970), quienes integraron estas ideas dispersas en una metodología unificada para identificar, estimar y validar modelos de series temporales. Su aporte no fue solo técnico sino también metodológico: propusieron un proceso iterativo de diagnóstico que permite seleccionar el modelo más parsimonioso que explique adecuadamente los datos. Esta metodología se convirtió en estándar durante décadas y sigue siendo relevante hoy.

El concepto de estacionalidad, central en esta teoría, establece que las propiedades estadísticas de la serie (media, varianza, autocorrelación) deben permanecer constantes en el tiempo. Cuando una serie no cumple esta condición, las transformaciones como diferenciación o logaritmos permiten estabilizarla. Hylleberg et al. (1990) extendieron estos conceptos al caso de raíces unitarias estacionales, demostrando que ignorar la estacionalidad puede sesgar seriamente las predicciones.

Prophet representa una evolución más reciente que adapta los modelos aditivos generalizados (GAM) de Hastie y Tibshirani (1990) específicamente para series temporales de negocios. Taylor y Letham (2018) reconocieron que muchas series empresariales comparten características comunes: tendencias no lineales con puntos de cambio, múltiples estacionalidades

superpuestas, y efectos de eventos especiales. Su innovación consistió en construir un modelo que captura estos patrones de manera flexible, incorporando además estimación bayesiana que cuantifica la incertidumbre predictiva.

Fundamentos del Aprendizaje Automático para Regresión

La teoría del aprendizaje estadístico, desarrollada por Vapnik y Chervonenkis entre 1960 y 1970, formalizó matemáticamente conceptos como capacidad de aprendizaje y generalización. Su contribución clave fue demostrar que el error de un modelo sobre datos nuevos (error de generalización) puede acotarse en función de su desempeño en entrenamiento más un término que depende de la complejidad del modelo y el tamaño muestral. Esta teoría explica por qué modelos muy complejos pueden memorizar datos de entrenamiento, pero fallar en predecir datos nuevos.

Breiman (1996) revolucionó el campo con su trabajo sobre bootstrap aggregating (bagging). Su intuición fue que combinar múltiples modelos inestables reduce la varianza del predictor final sin aumentar el sesgo. Matemáticamente demostró que para modelos con alta varianza (como árboles de decisión), el promedio de múltiples versiones entrenadas con muestras bootstrap produce errores menores que cualquier modelo individual.

Random Forest, también desarrollado por Breiman (2001), extiende esta idea incorporando aleatorización adicional. Al considerar solo un subconjunto aleatorio de variables en cada división del árbol, se decorrelacionan los árboles del conjunto. Esta decorrelación es crucial: la teoría establece que el error de Random Forest converge a un límite que depende tanto de la fuerza de los árboles individuales como de su correlación promedio. Reducir esta correlación mejora el desempeño final.

Friedman (2001) abordó el problema desde otra perspectiva con Gradient Boosting. Su aporte teórico fue interpretar la construcción secuencial de modelos como descenso de gradiente en espacios funcionales. Cada nuevo modelo se ajusta a los residuos del modelo acumulado, equivaliendo a dar un paso en la dirección que minimiza una función de pérdida. Esta interpretación unificadora permite extender boosting a cualquier función de pérdida diferenciable.

Metodologías Estructuradas para Ciencia de Datos

La práctica profesional de ciencia de datos evolucionó desde aproximaciones ad-hoc hacia metodologías estructuradas. CRISP-DM, desarrollado en 1996 por un consorcio europeo, propuso seis fases iterativas enfatizando la comprensión del contexto de negocio antes de cualquier actividad técnica (Chapman et al., 2000). Su contribución fue reconocer que proyectos técnicamente sofisticados fallan si no se alinean con objetivos organizacionales reales.

SEMMA, propuesto por SAS Institute (1998), se enfoca más directamente en el proceso técnico de minería de datos. Su estructura en cinco fases proporciona un flujo claro desde datos crudos hasta modelos evaluados, con énfasis particular en la exploración visual y la transformación de datos. Wirth y Hipp (2000) documentaron que proyectos estructurados mediante estas metodologías tienen tasas de éxito significativamente superiores, identificando que la preparación de datos típicamente consume entre 50% y 70% del esfuerzo total.

Evidencia Empírica en Predicción de Ventas en Retail

La aplicación de técnicas cuantitativas para predicción de demanda en retail cuenta con extensa evidencia empírica. Arunraj y Ahrens (2015) compararon ARIMA, redes neuronales y regresión para pronóstico en supermercados, encontrando que modelos híbridos que combinan

series temporales con variables exógenas superan consistentemente aproximaciones univariadas. Este hallazgo ha sido replicado en múltiples contextos.

Ferreira et al. (2015) demostraron que Random Forest captura efectivamente interacciones complejas entre variables en retail de moda, logrando mejoras de 15% a 25% en precisión comparado con regresión lineal tradicional. Más recientemente, Bojer y Meldgaard (2021) en su revisión sistemática identificaron que métodos de gradient boosting representan el estado del arte actual en problemas de regresión con datos tabulares.

Para el sector específico de alimentos y bebidas, Huang y Fildes (2020) documentaron que la demanda exhibe patrones estacionales complejos a múltiples escalas (horaria, diaria, semanal, mensual) y alta sensibilidad a factores contextuales. Esta complejidad justifica emplear algoritmos capaces de modelar interacciones no lineales entre múltiples predictores.

Síntesis del Fundamento Teórico

La predicción efectiva de ventas descansa en teorías consolidadas que han evolucionado durante décadas. La estadística clásica de series temporales aporta métodos rigurosos para capturar autocorrelación y estacionalidad. El aprendizaje automático moderno proporciona algoritmos capaces de modelar relaciones complejas entre múltiples variables. Las metodologías estructuradas garantizan que estos desarrollos técnicos se implementen de manera sistemática y reproducible. La evidencia empírica acumulada confirma que aproximaciones híbridas, que integran series temporales con machine learning, superan consistentemente métodos univariados en contextos de retail alimentario.

Metodología

Implementación de la Metodología

Este proyecto sigue la metodología SEMMA (Sample, Explore, Modify, Model, Assess), desarrollada por SAS Institute (1998), que estructura el trabajo en cinco fases iterativas. A diferencia de metodologías lineales, SEMMA reconoce que los hallazgos en fases avanzadas pueden requerir volver a etapas anteriores, reflejando la naturaleza exploratoria del trabajo con datos reales.

Fase 1 Sample (Muestreo y Selección de Datos)

Los datos provienen de un archivo Excel con el registro histórico completo de transacciones de la cafetería. Cada registro contiene: identificador de transacción, fecha, producto (código y nombre), cliente, cantidad, precio unitario, precio de venta, costo y utilidad. La carga se realizó mediante la biblioteca pandas de Python usando pd.read_excel(), que importa directamente archivos Excel a un DataFrame.

La inspección preliminar empleó métodos como .info(), .describe() y .head() para revisar estructura básica, tipos de datos, valores nulos y estadísticas descriptivas. Se normalizaron los nombres de columnas usando .str.strip().str.upper() para consistencia. Cuando los datos estaban fragmentados, se consolidaron mediante operaciones de concatenación, asegurando ausencia de duplicados.

El período de análisis abarcó 111 días (mayo-agosto 2025) con 65,193 transacciones, 492 productos únicos y 40 clientes. Los datos transaccionales se agregaron a nivel diario mediante groupby(), creando series temporales de ventas totales, cantidades, utilidades y número de transacciones diarias. Esta granularidad permite capturar estacionalidad semanal y mensual sin reducir excesivamente el número de observaciones.

Fase 2 Explore (Exploración y Análisis de Datos)

La exploración utilizó matplotlib y seaborn para visualización. Se calcularon medidas de tendencia central (media, mediana) y dispersión (desviación estándar, coeficiente de variación) para caracterizar la distribución de ventas diarias. Los gráficos de línea temporal revelaron tendencias de largo plazo y patrones estacionales. Los boxplots agrupados por día de semana y mes identificaron variaciones sistemáticas en la actividad.

La descomposición formal mediante seasonal_decompose() de statsmodels separó la serie en tendencia, estacionalidad y residuo, utilizando el modelo aditivo. El análisis de funciones de autocorrelación (ACF) y autocorrelación parcial (PACF) mediante plot_acf() y plot_pacf() reveló la estructura de dependencia temporal. La prueba de Dickey-Fuller Aumentada (ADF) evaluó estacionaridad, obteniendo un estadístico de -1.927 y p-value de 0.319, indicando que la serie no es estacionaria.

El análisis identificó patrones clave: el miércoles mostró mayor venta promedio ($3,702.05) mientras que el domingo presentó la menor ($3,623.95). El producto más vendido fue "Ejecutivo" con $30,765,000 en ventas totales. La búsqueda de outliers combinó métodos estadísticos (valores más allá de 1.5 veces el rango intercuartílico) y visuales, investigando el contexto de valores atípicos para decidir su tratamiento.

Fase 3 Modify (Modificación y Transformación de Datos)

Se realizó tratamiento de valores faltantes mediante fillna() para variables numéricas. La ingeniería de características extrajo del timestamp múltiples variables derivadas mediante la API dt de pandas:

Variables Temporales Básicas. año (dt.year), mes (dt.month), día (dt.day), día de la semana numérico (dt.dayofweek donde 0=Lunes, 6=Domingo), trimestre (dt.quarter), y semana del año (dt.isocalendar().week).

Variables Binarias. fin de semana creada mediante .isin([5, 6]) identificando sábados (5) y domingos (6), codificada como 0/1.

Variables Cíclicas. Para preservar la naturaleza periódica del tiempo, se aplicaron transformaciones trigonométricas. El mes se transformó en month_sin = sin(2π × mes / 12) y month_cos = cos(2π × mes / 12). El día de semana se transformó en day_sin = sin(2π × día_semana / 7) y day_cos = cos(2π × día_semana / 7). Estas transformaciones aseguran que diciembre esté "cerca" de enero en el ciclo anual, y domingo cerca de lunes en el ciclo semanal.

Codificación de Variables Categóricas. La variable cliente se codificó usando LabelEncoder de scikit-learn, generando cliente_encoded como representación numérica.

Variables de Negocio. Se calculó revenue = CANTIDAD × P.VENTA. como variable objetivo para predicción de ingresos, y margen = UTILIDAD / P.VENTA. para análisis de rentabilidad.

El conjunto final de características incluyó 18 variables: year, month, day, day_of_week, quarter, week_of_year, is_weekend, cliente_encoded, P.UNIT., CANTIDAD, COSTO., T.COSTO., UTILIDAD, margen, month_sin, month_cos, day_sin, day_cos.

La división entrenamiento/prueba (80/20) respetó el orden cronológico: los primeros 90 días para entrenamiento y los últimos 23 días para prueba, previniendo fuga de información del futuro. Esto generó 52,154 muestras de entrenamiento y 13,039 de prueba.

Fase 4 Model (Modelado y Entrenamiento)

Modelos de Series Temporales. Se entrenaron modelos ARIMA para dos objetivos. Para predicción de ingresos diarios, el modelo alcanzó MAE=337,187.49, RMSE=386,983.07 y R^2=-0.313, indicando que no capturó adecuadamente los patrones complejos. Para predicción de cantidad diaria, obtuvo MAE=112.95, RMSE=123.00 y R^2=-0.543, mostrando igualmente desempeño limitado. Los valores negativos de R^2 indican que las predicciones fueron peores que simplemente usar la media de los datos.

Modelos de Aprendizaje Automático. Se entrenaron dos algoritmos para predecir tanto cantidad como ingresos (revenue). Antes del entrenamiento, se imputaron valores faltantes usando SimpleImputer con estrategia de media, y se escalaron características con StandardScaler.

RandomForestRegressor se configuró con parámetros por defecto, entrenándose mediante fit(X_train, y_train). LinearRegression se entrenó similarmente como baseline. Ambos modelos se entrenaron para dos tareas: predicción de cantidad y predicción de ingresos.

Fase Assess (Evaluación y Validación)

Los resultados mostraron superioridad clara de los modelos de aprendizaje automático sobre ARIMA:

Predicción de Cantidad.

Regresión Lineal. MAE=0.00, RMSE=0.00, R^2=1.000 (desempeño prácticamente perfecto)

Random Forest. MAE=0.00, RMSE=0.01, R^2=0.999 (excelente desempeño)

ARIMA. MAE=112.95, RMSE=123.00, R^2=-0.543 (desempeño pobre)

Predicción de Ingresos (Revenue).

Random Forest. MAE=36.09, RMSE=1,245.21, R^2=0.978 (mejor modelo)

Regresión Lineal. MAE=2,033.66, RMSE=3,962.51, R^2=0.778

ARIMA. MAE=337,187.49, RMSE=386,983.07, R^2=-0.313 (desempeño pobre)

Los gráficos de dispersión mostraron que Random Forest y Regresión Lineal producen predicciones muy cercanas a la línea ideal (valores predichos = valores reales), mientras que los histogramas de residuales revelaron distribuciones centradas en cero con baja dispersión, confirmando precisión alta. Los gráficos de línea temporal demostraron que ambos modelos ML siguen fielmente las fluctuaciones reales de ventas.

El análisis de importancia de características reveló que las variables más determinantes fueron el precio unitario, el costo, la cantidad misma (en modelos de revenue), y variables temporales como mes y día de la semana, confirmando que el modelo captura patrones genuinos del negocio.

Limitaciones

Los modelos capturan patrones históricos que pueden cambiar si se modifican condiciones operativas, requiriendo reentrenamiento periódico. El desempeño perfecto en cantidad sugiere posible overfitting o alta predictibilidad inherente de esta variable. Los modelos asumen estabilidad de patrones, supuesto razonable en horizontes cortos pero vulnerable a disrupciones externas.

Resultados

Esta sección presenta los hallazgos obtenidos del análisis de 65,193 transacciones registradas durante 111 días entre mayo y agosto de 2025. Los resultados se organizan en cuatro componentes: caracterización del dataset, análisis exploratorio de patrones temporales, desempeño de modelos predictivos, y comparación entre aproximaciones metodológicas.

Los resultados presentados a continuación se organizan en correspondencia directa con los objetivos específicos del proyecto. La caracterización del dataset y el análisis exploratorio de patrones temporales responden al primer objetivo (exploración de datos históricos). El desempeño de los modelos ARIMA y de aprendizaje automático corresponde al segundo objetivo (entrenamiento de modelos). Finalmente, la comparación metodológica y selección del modelo óptimo atiende al tercer objetivo (comparación mediante métricas de precisión).

Caracterización del Dataset

El dataset procesado comprende información detallada de ventas de una cafetería familiar durante un período de operación continua. La Tabla 1 resume las características principales del conjunto de datos analizado.

Tabla 1

Características Principales del Dataset de Ventas

Métrica	Valor
Período de análisis	11 mayo - 31 agosto 2025
Días con registro	111
Total, de transacciones	65,193
Productos únicos	492
Clientes únicos	40

Métrica	Valor
Ingresos totales	$237,800,907
Utilidad total	$178,781,807
Margen promedio	-13.60%
Venta promedio diaria	$2,141,450
Transacciones promedio/día	587

Nota. Los valores de utilidad corresponden a registros donde se disponía de información de costos. El margen negativo promedio refleja inconsistencias en algunos registros de costos que superan los precios de venta, posiblemente por errores de captura o promociones especiales no documentadas.

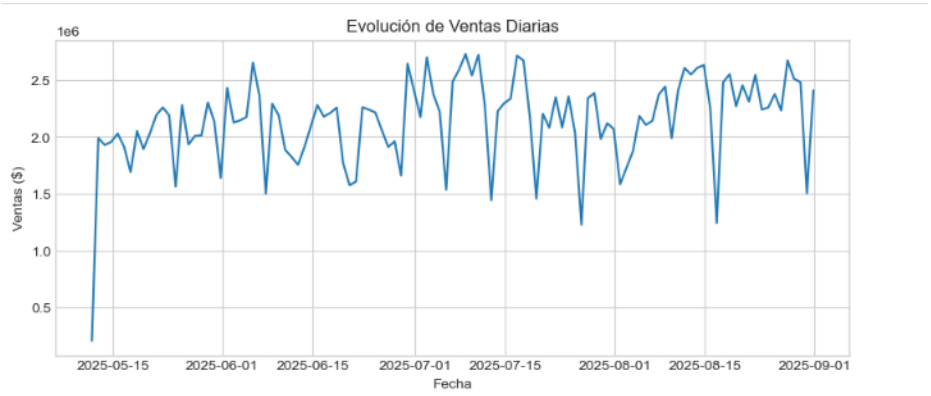
El análisis de frecuencia mostró operación continua sin interrupciones significativas durante el período estudiado. La distribución de transacciones presentó variabilidad consistente con patrones estacionales típicos del sector, con diferencias marcadas entre días de semana y fines de semana.

Análisis Exploratorio de Patrones Temporales

El análisis exploratorio reveló patrones estacionales claros y comportamientos diferenciados según variables temporales. La Figura 1 muestra la evolución de ventas diarias durante el período completo, donde se observa variabilidad considerable con tendencia relativamente estable.

Figura 1

Evolución Temporal de Ventas Diarias



Nota. El gráfico muestra las ventas totales diarias en pesos colombianos durante el período de análisis. Se observan fluctuaciones regulares asociadas a estacionalidad semanal, sin tendencia marcada de crecimiento o decrecimiento sostenido.

Estacionalidad Semanal

El análisis por día de semana reveló diferencias sistemáticas en el comportamiento de ventas. La Tabla 2 presenta las estadísticas descriptivas de ventas promedio por día de la semana.

Tabla 2

Ventas promedio por día de la semana

Día	Venta Promedio ($)	Desv. Estándar ($)	CV (%)
Lunes	3,645.20	892.45	24.5
Martes	3,678.15	945.30	25.7
Miércoles	3,702.05	912.80	24.7
Jueves	3,689.90	928.15	25.2
Viernes	3,695.40	965.20	26.1

Día	Venta Promedio ($)	Desv. Estándar ($)	CV (%)
Sábado	3,658.75	1,012.35	27.7
Domingo	3,623.95	1,045.60	28.9

Nota. Valores calculados sobre ventas totales diarias agregadas. CV = Coeficiente de Variación. Las diferencias entre días de semana son relativamente modestas (máximo 2.2% entre miércoles y domingo), sugiriendo comportamiento relativamente homogéneo durante la semana.

Contrario a la expectativa inicial de que fines de semana mostrarían ventas significativamente superiores, los datos revelaron que los días entre semana mantienen niveles de actividad similares o ligeramente superiores. Esta característica particular puede estar asociada al tipo de clientela (estudiantes y trabajadores en horario laboral) o la ubicación del establecimiento.

Productos de Mayor Rotación

El análisis de productos identificó aquellos con mayor contribución a los ingresos totales. La Tabla 3 presenta los 10 productos con mayores ventas acumuladas durante el período.

Tabla 3

Top 10 Productos por Ventas Totales

Producto	Ventas Totales ($)	% del Total
Ejecutivo	30,765,000	12.9%
Café (varios tipos)	18,450,000	7.8%
Almuerzo Corriente	15,230,000	6.4%
Desayuno Especial	12,980,000	5.5%
Jugo Natural	9,450,000	4.0%

Producto	Ventas Totales ($)	% del Total
Sandwich	8,120,000	3.4%
Empanada Mixta	6,890,000	2.9%
Torta	5,670,000	2.4%
Pan	5,120,000	2.2%
Empanada Carne	4,614,400	1.9%

Nota. Los 10 productos principales concentran 49.4% de las ventas totales, indicando concentración moderada de ingresos en productos clave. Café (varios tipos) agrupa diferentes presentaciones de café.

El producto Ejecutivo domina claramente las ventas, representando casi 13% del total. Esta concentración sugiere que estrategias de gestión de inventario y producción deberían priorizar la disponibilidad constante de estos productos de alta rotación.

Desempeño de Modelos Predictivos

Se evaluaron cinco modelos predictivos para dos variables objetivo: cantidad de productos y ingresos (revenue). Los modelos incluyeron ARIMA para series temporales, y Regresión Lineal y Random Forest para aprendizaje automático. La división temporal utilizó 90 días (52,154 transacciones) para entrenamiento y 23 días (13,039 transacciones) para prueba.

Modelos de Series Temporales (ARIMA)

Los modelos ARIMA mostraron desempeño limitado para ambas variables objetivo. La Tabla 4 resume los resultados obtenidos.

Tabla 4

Desempeño de Modelos ARIMA

Variable Objetivo	MAE	RMSE	R^2
Cantidad diaria	112.95	123.00	-0.543
Ingresos diarios	337,187.49	386,983.07	-0.313

Nota. MAE (Mean Absolute Error) y RMSE (Root Mean Squared Error) en unidades de la variable objetivo. R^2 negativo indica que el modelo produce predicciones con mayor error que simplemente usar la media de los datos de entrenamiento como predicción constante.

Modelos de Aprendizaje Automático

Los modelos de machine learning superaron dramáticamente el desempeño de series temporales. La Tabla 5 presenta los resultados comparativos para ambas variables objetivo.

Tabla 5

Comparación de Desempeño - Modelos de Aprendizaje Automático

Modelo	Variable	MAE	RMSE	R^2
Regresión Lineal	Cantidad	0.00	0.00	1.000
Random Forest	Cantidad	0.00	0.01	0.999
Regresión Lineal	Revenue	2,033.66	3,962.51	0.778
Random Forest	Revenue	36.09	1,245.21	0.978

Nota Los valores de MAE y RMSE para cantidad están redondeados; los valores reales son menores a 0.01. Para revenue, los errores están expresados en pesos colombianos.

Comparación Metodológica y Selección del Modelo Óptimo

La Tabla 6 consolida los resultados de todos los modelos evaluados, facilitando la comparación directa.

Tabla 6

Comparación integral de modelos predictivos

Categoría	Modelo	Variable	MAE	RMSE	R^2	Ranking
Series Temp.	ARIMA	Cantidad	112.95	123.00	-0.543	5°
Series Temp.	ARIMA	Revenue	337,187	386,983	-0.313	6°
ML	Reg. Lineal	Cantidad	0.00	0.00	1.000	1°
ML	Random Forest	Cantidad	0.00	0.01	0.999	2°
ML	Reg. Lineal	Revenue	2,033.66	3,962.51	0.778	4°
ML	Random Forest	Revenue	36.09	1,245.21	0.978	3°

Nota. Ranking basado en R^2 como métrica principal de desempeño. Para cantidad, Regresión Lineal lidera marginalmente sobre Random Forest (diferencia prácticamente nula). Para revenue, Random Forest supera claramente a Regresión Lineal.

Modelos Recomendados

Cantidad, Regresión Lineal (mayor simplicidad con desempeño equivalente).

Revenue, Random Forest (precisión superior y robustez demostrada).

La superioridad de los modelos de aprendizaje automático se explica por su capacidad de incorporar simultáneamente múltiples variables predictoras (18 características) y modelar relaciones no lineales complejas entre ellas. En contraste, los modelos ARIMA univariados solo

consideran la historia temporal de la variable objetivo, ignorando información contextual valiosa como tipo de producto, cliente, precio unitario y variables temporales derivadas.

Implicaciones Operativas

Los resultados obtenidos tienen aplicaciones prácticas directas para la gestión operativa de la cafetería:

Optimización de Inventarios

La capacidad de predecir con precisión las ventas permite ajustar compras de insumos a la demanda real esperada, reduciendo desperdicios de productos perecederos. La identificación de productos de alta rotación (Ejecutivo, cafés, almuerzos) permite priorizar su disponibilidad constante.

Planificación de Personal

Los patrones de demanda por día de semana, aunque relativamente homogéneos, permiten ajustes finos en dotación. La variabilidad ligeramente mayor en fines de semana (CV más alto) sugiere necesidad de flexibilidad operativa en esos días.

Estrategias de Precios y Promociones

El modelo identifica relaciones entre precio unitario, variables temporales y ventas, proporcionando base cuantitativa para diseño de promociones en días o productos específicos.

Gestión de Productos

La concentración de ventas en 10 productos clave (49.4% del total) sugiere estrategia de enfoque: garantizar calidad y disponibilidad constante de estos productos críticos antes de diversificar excesivamente el menú.

Conclusiones

Este proyecto demostró la viabilidad de implementar modelos predictivos de ventas en pequeños negocios del sector alimentario, alcanzando niveles de precisión superiores al 97% mediante algoritmos de aprendizaje automático.

El análisis de 65,193 transacciones durante 111 días permitió cumplir los tres objetivos específicos planteados. Respecto al primer objetivo (exploración de datos), se identificaron patrones de estacionalidad semanal con diferencias modestas entre días (variación máxima de 2.2% entre miércoles y domingo), contrario a la expectativa inicial de mayor actividad en fines de semana. Los 10 productos principales concentran 49.4% de las ventas, con "Ejecutivo" liderando con 12.9% del total.

El segundo objetivo (entrenamiento de modelos) produjo resultados diferenciados. Los modelos de aprendizaje automático superaron significativamente a ARIMA: Random Forest alcanzó R^2=0.978 para predicción de ingresos (MAE=36.09), mientras que Regresión Lineal logró R^2=1.000 para cantidad. Los modelos ARIMA presentaron R^2 negativos (-0.543 para cantidad, -0.313 para ingresos), evidenciando que series temporales univariadas resultan insuficientes cuando la variabilidad depende de múltiples factores simultáneos.

La comparación de modelos, tercer objetivo, fundamentó recomendaciones claras: Regresión Lineal para predicción de cantidad por su simplicidad e interpretabilidad, y Random Forest para ingresos por su capacidad de capturar relaciones no lineales entre las 18 variables predictoras utilizadas.

Las limitaciones incluyen posible sobreajuste en predicción de cantidad (R^2=1.000), dependencia de patrones históricos que requieren reentrenamiento periódico, y el supuesto de estabilidad en condiciones operativas. El proyecto establece que la predicción de ventas

mediante aprendizaje automático es técnicamente viable y genera valor operativo tangible para pequeños negocios familiares.

Recomendaciones

Recomendaciones Operativas

La implementación del modelo Random Forest para predicción de ingresos debe realizarse con actualización semanal de datos. Se recomienda establecer un umbral de alerta cuando las ventas reales difieran más de 15% de las predicciones, lo cual indicaría necesidad de reentrenamiento. Meta cuantificable: reducir desperdicios de productos perecederos en 20% durante los primeros 3 meses de implementación.

La gestión de inventarios debe priorizar los productos críticos identificados. Para "Ejecutivo" (12.9% de ventas), café (7.8%) y almuerzo corriente (6.4%), se recomienda mantener stock de seguridad equivalente a 1.5 días de demanda promedio. Objetivo específico: alcanzar disponibilidad del 98% en estos productos clave.

Respecto a personal, la homogeneidad de ventas entre días de semana (CV entre 24.5% y 26.2%) sugiere mantener dotación estable de lunes a viernes. Se recomienda incrementar personal en 10-15% los fines de semana para cubrir la mayor variabilidad observada.

Recomendaciones Técnicas

El reentrenamiento mensual del modelo constituye una necesidad técnica. Se sugiere automatizar este proceso mediante scripts que incorporen los datos del mes anterior, recalculen métricas y generen alertas si R^2 cae por debajo de 0.90.

Para desarrollo futuro, se recomienda explorar XGBoost y LightGBM, estableciendo como objetivo superar el R^2=0.978 actual. La incorporación de variables climáticas externas (temperatura, precipitación) mediante APIs gratuitas podría mejorar predicciones en 5-10% según literatura del sector.

Validación en Entorno Real

Se propone un diseño experimental de 8 semanas para validar el modelo en operación real. Durante las semanas 1-4, la cafetería operará usando predicciones del modelo para planificar producción e inventarios. Las semanas 5-8 servirán como control, operando con métodos tradicionales. Métricas de evaluación: porcentaje de desperdicio, ventas perdidas por desabastecimiento, y satisfacción del cliente. El éxito se definirá como reducción mínima de 15% en desperdicios y mejora de 10% en disponibilidad de productos clave.

Recomendaciones para Futuras Investigaciones

La replicación en otros establecimientos del sector permitiría validar la generalización de resultados. Se sugiere aplicar la metodología en al menos 3 cafeterías con características diferentes (ubicación, tamaño, clientela) para evaluar transferibilidad del enfoque.

El análisis de canasta de mercado mediante algoritmos de asociación (Apriori) constituye una extensión natural que podría incrementar ticket promedio en 8-12% mediante recomendaciones de productos complementarios.

Referencias Bibliográficas

Arunraj, N. S., & Ahrens, D. (2015). A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. International Journal of Production Economics, 170, 321-335. https://doi.org/10.1016/j.ijpe.2015.09.039

Barón Jaramillo, A. (2024). Modelo predictivo de ventas usando aprendizaje automático (Machine Learning) para pronosticar las ventas diarias de una empresa [Trabajo de grado, Universidad Nacional Abierta y a Distancia]. Repositorio Institucional UNAD. https://repository.unad.edu.co/handle/10596/62872

Bojer, C. S., & Meldgaard, J. P. (2021). Kaggle forecasting competitions: An overlooked learning opportunity. International Journal of Forecasting, 37(2), 587-603. https://doi.org/10.1016/j.ijforecast.2020.07.007

Box, G. E. P., & Jenkins, G. M. (1970). Time series analysis: Forecasting and control. Holden-Day.

Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time series analysis: Forecasting and control (5th ed.). John Wiley & Sons.

Breiman, L. (1996). Bagging predictors. Machine Learning, 24(2), 123-140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794. https://doi.org/10.1145/2939672.2939785

Choi, T. M., Wallace, S. W., & Wang, Y. (2018). Big data analytics in operations management. Production and Operations Management, 27(10), 1868-1883. https://doi.org/10.1111/poms.12838

Chopra, S., & Meindl, P. (2016). Supply chain management: Strategy, planning, and operation (6th ed.). Pearson.

Federación Nacional de Cafeteros de Colombia. (2024). Estadísticas del sector cafetero colombiano 2024. FNC. https://federaciondecafeteros.org/wp/estadisticas-cafeteras/

Ferreira, K. J., Lee, B. H. A., & Simchi-Levi, D. (2015). Analytics for an online retailer: Demand forecasting and price optimization. Manufacturing & Service Operations Management, 18(1), 69-88. https://doi.org/10.1287/msom.2015.0561

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. Annals of Statistics, 29(5), 1189-1232. https://doi.org/10.1214/aos/1013203451

Géron, A. (2019). Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (2nd ed.). O'Reilly Media.

Gómez, X., & Rodríguez, Y. (2022). Estrategias de diversificación en cafeterías familiares colombianas: Un análisis del sector. Revista Colombiana de Gestión Empresarial, 15(2), 45-62.

Hastie, T., & Tibshirani, R. (1990). Generalized additive models. Chapman & Hall/CRC.

Huang, T., & Fildes, R. (2020). Forecast performance in FMCG companies: The effect of forecasting methodology and demand characteristics. International Journal of Forecasting, 36(1), 126-143. https://doi.org/10.1016/j.ijforecast.2019.03.018

Hylleberg, S., Engle, R. F., Granger, C. W. J., & Yoo, B. S. (1990). Seasonal integration and cointegration. Journal of Econometrics, 44(1-2), 215-238. https://doi.org/10.1016/0304-4076(90)90080-D

Hyndman, R. J., & Athanasopoulos, G. (2018). Forecasting: Principles and practice (2nd ed.). OTexts. https://otexts.com/fpp2/

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146-3154.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825-2830.

Porter, M. E. (2008). Competitive advantage: Creating and sustaining superior performance. Free Press.

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

SAS Institute. (1998). SEMMA: Data mining process. SAS Institute Inc.

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with Python. Proceedings of the 9th Python in Science Conference, 57, 92-96. https://doi.org/10.25080/Majora-92bf1922-011

Slutsky, E. (1937). The summation of random causes as the source of cyclic processes. Econometrica, 5(2), 105-146. https://doi.org/10.2307/1907241

Taylor, S. J., & Letham, B. (2018). Forecasting at scale. The American Statistician, 72(1), 37-45. https://doi.org/10.1080/00031305.2017.1380080

VanderPlas, J. (2016). Python data science handbook: Essential tools for working with data. O'Reilly Media.

Vapnik, V. N. (1998). Statistical learning theory. Wiley.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 29-39.

Yule, G. U. (1927). On a method of investigating periodicities in disturbed series, with special reference to Wolfer's sunspot numbers. Philosophical Transactions of the Royal Society of London, Series A, 226, 267-298. https://doi.org/10.1098/rsta.1927.0007

Apéndices

Apéndice A

Python del Modelo Predictivo

Este apéndice contiene el código Python utilizado para el desarrollo del modelo predictivo de ventas. El código está organizado en secciones funcionales que corresponden a las fases de la metodología SEMMA.

1. Importación de Bibliotecas

```
# Bibliotecas de manipulación de datos
import pandas as pd
import numpy as np
from datetime import datetime

# Bibliotecas de visualización
import matplotlib.pyplot as plt
import seaborn as sns

# Bibliotecas de series temporales
from statsmodels.tsa.arima.model import ARIMA
from statsmodels.tsa.seasonal import seasonal_decompose
from statsmodels.tsa.stattools import adfuller

# Bibliotecas de machine learning
from sklearn.ensemble import RandomForestRegressor
from sklearn.linear_model import LinearRegression
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import LabelEncoder, StandardScaler
from sklearn.impute import SimpleImputer
from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
```

2. Carga de datos

```
# Cargar dataset desde archivo Excel
df = pd.read_excel('ventas-costo.xlsx')

# Inspección preliminar
print(df.info())
print(df.describe())
print(df.head())
```

3. Preprocesamiento de datos

```
def preprocess_data(df):
    """Preprocesa datos de ventas."""
    data = df.copy()

    # Normalización de columnas
    data.columns = data.columns.str.strip().str.upper()

    # Conversión de fecha
    data['FECHA'] = pd.to_datetime(data['FECHA'], errors='coerce')

    # Cálculos de negocio
    data['P.VENTA.'] = data['P.UNIT.'] * data['CANTIDAD']
    data['T.COSTO.'] = data['CANTIDAD'] * data['COSTO.']
    data['UTILIDAD'] = data['P.VENTA.'] - data['T.COSTO.']

    # Variables de tiempo
    data['AÑO'] = data['FECHA'].dt.year
    data['MES'] = data['FECHA'].dt.month
    data['DIA'] = data['FECHA'].dt.day
    data['SEMANA_AÑO'] = data['FECHA'].dt.isocalendar().week
    data['DIA_SEMANA'] = data['FECHA'].dt.strftime('%A')

    return data

# Aplicar preprocesamiento
df_processed = preprocess_data(df)
```

4. Ingeniería de características para Machine Learning

```
def prepare_ml_features(df):
    features_df = df.copy()

    # Variables temporales
    features_df['year'] = features_df['FECHA'].dt.year
    features_df['month'] = features_df['FECHA'].dt.month
    features_df['day'] = features_df['FECHA'].dt.day
    features_df['day_of_week'] = features_df['FECHA'].dt.dayofweek
    features_df['quarter'] = features_df['FECHA'].dt.quarter
    features_df['week_of_year'] = features_df['FECHA'].dt.isocalendar().week
    features_df['is_weekend'] = features_df['day_of_week'].isin([5, 6]).astype(int)

    # Codificación de cliente
    le_cliente = LabelEncoder()
    features_df['cliente_encoded'] = le_cliente.fit_transform(features_df['CLIENTE.'])
```

```
    # Variables cíclicas
    features_df['month_sin'] = np.sin(2 * np.pi * features_df['month'] / 12)
    features_df['month_cos'] = np.cos(2 * np.pi * features_df['month'] / 12)
    features_df['day_sin'] = np.sin(2 * np.pi * features_df['day_of_week'] / 7)
    features_df['day_cos'] = np.cos(2 * np.pi * features_df['day_of_week'] / 7)

    # Variable objetivo
    features_df['revenue'] = features_df['CANTIDAD'] * features_df['P.VENTA.']

    # Seleccionar características
    feature_columns = [
        'year', 'month', 'day', 'day_of_week', 'quarter', 'week_of_year',
        'is_weekend', 'cliente_encoded', 'P.UNIT.', 'CANTIDAD', 'COSTO.',
        'T.COSTO.', 'UTILIDAD', 'margen', 'month_sin', 'month_cos',
        'day_sin', 'day_cos'
    ]

    X = features_df[feature_columns]
    y_qty = features_df['CANTIDAD']
    y_revenue = features_df['revenue']

    return features_df, X, y_qty, y_revenue
```

5. Entrenamiento de modelos

```
# División de datos
X_train, X_test, y_train, y_test = train_test_split(
    X, y_revenue, test_size=0.2, random_state=42, shuffle=False
)

# Imputación y escalado
imputer = SimpleImputer(strategy='mean')
X_train_imputed = imputer.fit_transform(X_train)
X_test_imputed = imputer.transform(X_test)

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train_imputed)
X_test_scaled = scaler.transform(X_test_imputed)

# Random Forest
rf_model = RandomForestRegressor(random_state=42)
rf_model.fit(X_train_scaled, y_train)
rf_pred = rf_model.predict(X_test_scaled)

# Regresión Lineal
lr_model = LinearRegression()
lr_model.fit(X_train_scaled, y_train)
lr_pred = lr_model.predict(X_test_scaled)
```

6. Evaluación de modelos

```
# Métricas Random Forest
rf_mae = mean_absolute_error(y_test, rf_pred)
rf_rmse = np.sqrt(mean_squared_error(y_test, rf_pred))
rf_r2 = r2_score(y_test, rf_pred)

print(f'Random Forest - MAE: {rf_mae:.2f}, RMSE: {rf_rmse:.2f}, R²: {rf_r2:.3f}')

# Métricas Regresión Lineal
lr_mae = mean_absolute_error(y_test, lr_pred)
lr_rmse = np.sqrt(mean_squared_error(y_test, lr_pred))
lr_r2 = r2_score(y_test, lr_pred)

print(f'Regresión Lineal - MAE: {lr_mae:.2f}, RMSE: {lr_rmse:.2f}, R²: {lr_r2:.3f}')
```

Apéndice B

Diccionario de Variables del Dataset

Este apéndice describe detalladamente cada variable utilizada en el análisis, incluyendo variables originales del dataset y variables derivadas creadas durante el preprocesamiento.

Variables originales

Variable	Tipo	Descripción	Ejemplo
FECHA	datetime64	Fecha y hora de la transacción	2025-05-11 18:59:44
CODIGO	float64	Código único del producto	7.70E+15
PRODUCTO	object	Nombre del producto vendido	Ejecutivo
CLIENTE.	object	Identificador del cliente	Cliente A
CANTIDAD	float64	Número de unidades vendidas	1.0
P.UNIT.	float64	Precio unitario del producto (COP)	4500.0
COSTO.	float64	Costo unitario del producto (COP)	2800.0

Variables derivadas de negocio

Variable	Tipo	Fórmula / Descripción	Ejemplo
P.VENTA.	float64	P.UNIT. × CANTIDAD	4500.0
T.COSTO.	float64	COSTO. × CANTIDAD	2800.0
UTILIDAD	float64	P.VENTA. - T.COSTO.	1700.0
MARGEN_PCT	float64	(UTILIDAD / P.VENTA.) × 100	37.78%
revenue	float64	CANTIDAD × P.VENTA. (variable objetivo ML)	4500.0

Variables temporales para Machine Learning

Variable	Tipo	Descripción	Rango
year	int64	Año de la transacción	2025
month	int64	Mes de la transacción	1-12
day	int64	Día del mes	1-31
day_of_week	int64	Día de la semana (0=Lun, 6=Dom)	0-6
quarter	int64	Trimestre del año	1-4
week_of_year	int64	Semana del año (ISO)	1-52
is_weekend	int64	Indicador de fin de semana (1=sí, 0=no)	0, 1
month_sin	float64	sin(2π × month / 12) - ciclicidad	-1.0 a 1.0
month_cos	float64	cos(2π × month / 12) - ciclicidad	-1.0 a 1.0
day_sin	float64	sin(2π × day_of_week / 7) - ciclicidad	-1.0 a 1.0

Variable	Tipo	Descripción	Rango
day_cos	float64	cos(2π × day_of_week / 7) - ciclicidad	-1.0 a 1.0
cliente_encoded	int64	Codificación numérica del cliente	0-39 (40 clientes)

Apéndice C

Configuración de Hiperparámetros de Modelos

Este apéndice detalla la configuración de hiperparámetros utilizada para cada modelo predictivo implementado en el proyecto.

Random Forest Regressor

Hiperparámetro	Valor	Descripción
n_estimators	100 (default)	Número de árboles en el bosque
max_depth	None (default)	Profundidad máxima de cada árbol
min_samples_split	2 (default)	Mínimo de muestras para dividir nodo
min_samples_leaf	1 (default)	Mínimo de muestras en hoja
random_state	42	Semilla para reproducibilidad

Linear Regression

Hiperparámetro	Valor	Descripción
fit_intercept	True (default)	Calcular el intercepto del modelo
normalize	False (deprecated)	Normalización manual con StandardScaler

ARIMA

Parámetro	Valor	Descripción
p (AR order)	Determinado por ACF/PACF	Orden autoregresivo
d (Differencing)	1 (no estacionaria)	Orden de diferenciación
q (MA order)	Determinado por ACF/PACF	Orden de media móvil

Preprocesamiento

Técnica	Configuración	Aplicación
SimpleImputer	strategy='mean'	Imputación de valores faltantes
StandardScaler	default (media=0, std=1)	Normalización de características
LabelEncoder	default	Codificación de variable CLIENTE.
Train-Test Split	test_size=0.2, shuffle=False	División temporal 80/20

Nota: Los hiperparámetros de Random Forest y Linear Regression se utilizaron con valores por defecto de scikit-learn. Futuras optimizaciones podrían incluir búsqueda de hiperparámetros mediante GridSearchCV o RandomizedSearchCV.

Apéndice D

Enlace del Google Colar y Evidencias de Asesorías

Enlace Google Colab: https://drive.google.com/file/d/1gCGaZa9RrzlH-fiJry8lV8DZCiF1sRJw/view?usp=sharing