



Pontificia Universidad
JAVERIANA
Cali

Evaluación del impacto de modelos multivariados de machine learning en la precisión del pronóstico de ventas y su efecto en la gestión de la cadena de suministro

Mario Leandro Vanegas Valencia

9026612

Edisson Camilo Sanchez Osorio

9013835

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

Director

David Arango Londoño

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, MAYO DE 2026

FICHA RESUMEN

TÍTULO DEL PROYECTO: Evaluación del impacto de modelos multivariados de machine learning en la precisión del pronóstico de ventas y su efecto en la gestión de la cadena de suministro

1. ÁREA DE TRABAJO: Cadena de Suministro
2. TIPO DE PROYECTO: Aplicado
3. ESTUDIANTE(S): Mario Leandro Vanegas Valencia ; Edisson Camilo Sanchez Osorio
4. CORREO ELECTRÓNICO: mlvanegas@javerianacali.edu.co, edisson0712@javerianacali.edu.co
5. DIRECCIÓN Y TELEFONO: Vereda la China, finca la china, Manizales, 3113005998 ; Kr 93 54-28 sur Torre 1 Apto 1810, 3114912814
6. DIRECTOR: David Arango Londoño
7. VINCULACIÓN DEL DIRECTOR: Docente Catedrático
8. CORREO ELECTRÓNICO DEL DIRECTOR: David.arango@javerianacali.edu.co
9. CO-DIRECTOR: No Aplica
10. GRUPO O EMPRESA QUE LO AVALA: Herragro S.A.S
11. OTROS GRUPOS O EMPRESAS: No Aplica
12. PALABRAS CLAVE: Sales forecasting, Time series forecasting, Univariate models, Multivariate models, Machine learning algorithms, Agroindustrial sector, Forecast accuracy, External regressors, Demand planning, Random Forest, LightGBM, Prophet
13. FECHA DE INICIO: 07/07/2025
14. DURACIÓN ESTIMADA: 11 Meses
15. RESUMEN:

La planeación de la demanda constituye un eje fundamental en la gestión de la cadena de suministro agroindustrial, un sector complejo por la influencia de variables estocástica como el clima, los costos de insumos y la volatilidad macroeconómica. La problemática centra de esta investigación radica en las limitaciones estructurales de los modelos predictivos univariados tradicionales, los cuales presentan una precisión reducida al omitir factores externos.

La consecuencia, el objetivo principal de este estudio es evaluar y comparar el desempeño predictivo de enfoques estadísticos clásicos (OLS y SARIMAX) frente a algoritmos multivariados de Machine Learning (Random Forest, LightGBM y Prophet). La metodología integró datos

transaccionales históricos (2022-2025) de dos familias de productos (Palas y Carretas) con variables exógenas como índices climáticos zonales, abastecimiento agrícola (SIPSA) y tasas de cambio (TRM). El desempeño se evaluó mediante métricas estandarizadas (MAE, RMSE y MAPE) a través de una validación temporal walk-forward.

Los resultados empíricos demostraron que la granularidad mensual domina significativamente en la estabilización de la señal predictiva frente a agregaciones diarias o semanales. Asimismo, se evidencia que predecir el volumen (Cantidad) y el Margen Bruto constituyen problemas predictivos disímiles; mientras que Random Forest logró el mejor desempeño en tres de las cuatro series analizadas (reduciendo el MAPE hasta un 14.8% en volumen), SARIMAX demostró ser más robusto para modelar el Margen bruto en series específicas. Se concluye que la adopción de modelos multivariados de ensamble, combinados con un preprocesamiento adecuado de la temporalidad y variables exógenas casuales, optimiza sustancialmente la planeación comercial y operativa en la agroindustria.

TABLA DE CONTENIDO

INTRODUCCIÓN	7
1. CONTEXTUALIZACIÓN DEL PROYECTO.....	8
1.1. Planteamiento y Definición del problema.....	8
1.1.1. Formulación del problema	9
1.1.2. Sistematización	9
1.2. Objetivos	10
1.2.1. Objetivo General.....	10
1.2.2. Objetivos Específicos.....	10
1.3. Marco de Referencia	11
1.3.1. Marco Teórico	11
1.3.1.1. Planificación de la Demanda en la Cadena de Suministro bajo Entornos Estocásticos	11
1.3.1.2. Modelos estadísticos Clásicos de Series de Tiempo e Inexactitud Estructural	11
1.3.1.3. Aprendizaje Automático No Lineal y Métodos de Ensamble.....	12
1.3.1.4. Aprendizaje Automático No Lineal y Métodos de Ensamble.....	12
1.3.1.5. Importancia del enfoque multivariado en contextos volátiles	13
1.3.1.6. Canales de Transmisión Económica y biofísica en la Demanda Agroindustrial	13
1.3.1.7. El pronóstico como herramienta estratégica en la cadena de suministro.....	14
1.3.1.8. Divergencia Estructural entre Modelos de Volumen (Cantidad) y Rentabilidad (Margen Bruto)	15
1.3.1.9. Beneficios de la Inteligencia Artificial en la reducción de incertidumbre.....	15
1.3.2. Antecedentes	15
2. CONJUNTO DE DATOS Y ANÁLISIS EXPLORATORIO.....	19
2.1. Descripción y Adquisición de los Datos	19
2.2. Preparación, Transformación e Integración	22
2.3. Análisis Exploratorio de la Variable Objetivo: Volumen vs Rentabilidad	23
2.4. Dinámica de Variables Exógenas y Diagnóstico de Estacionariedad	24
2.5. Determinación de Rezagos Óptimos (Cross-Correlation Function)	31
3. METODOLOGÍA DE MODELADO Y VALIDACIÓN	31

3.1.	Arquitectura de Pipeline Analítico.....	32
3.2.	Formulación y Justificación Algorítmica de los Modelos.	32
3.2.1.	Modelos Lineales de Referencia (OLS y SARIMAX)	32
3.2.2.	Algoritmos de Aprendizaje Automático No Lineal (Ensembles)	33
3.2.3.	Modelado Aditivo Estructural (Prophet).....	34
3.3.	Estrategia de Ingeniería de características e Hiperparametrización	34
3.4.	Esquema de validación Temporal Secuencial (Walk-Forward Validation).....	38
4.	RESULTADOS Y DISCUSIÓN	40
4.1.	Línea Base y Desempeño de Modelos Tradicionales (OLS vs SARIMAX)	40
4.2.	Superioridad Empírica del Aprendizaje automático No lineal	41
4.3.	El efecto Estabilizador de la Agregación Temporal.....	43
4.4.	Selección de la Configuración Óptima y Despliegue del Pronóstico Operativo	46
5.	CONCLUSIONES Y TRABAJOS FUTUROS	48
5.1.	Conclusiones	48
5.2.	Limitaciones del Estudio.....	49
5.3.	Trabajos Futuros	50
6.	REFERENCIAS BIBLIOGRÁFICAS.....	51

LISTA DE FIGURAS

Ilustración 1 índices climáticos zonales semanales utilizados como variables exógenas, 2022-2025.....	22
Ilustración 2 Descomposición estacional de ventas en cantidad y margen bruto	24
Ilustración 3 Validación walk-forward de PALAS: Comparación entre OLS y SARIMAX.....	41
Ilustración 4 Comparación de MAPE entre modelos base	42
Ilustración 5 Efecto de las Mejoras Secuenciales	43
Ilustración 6 Comparación de granularidad y modelo mediante MAPE walk-forward	44
Ilustración 7 Pronóstico Final Seleccionado	47

LISTA DE TABLAS

Tabla 1 Aporte de Antecedentes.....	15
Tabla 2 Fuentes de información, periodicidad original, variables extraídas y transformaciones metodológicas aplicadas	19
Tabla 3 Correlación interna entre cantidad y margen bruto por familia y granularidad	23
Tabla 4 Diccionario de variables exógenas del modelo	25
Tabla 5 Propiedades estadísticas y de integración de las variables exógenas.....	29
Tabla 6 Configuración de hiperparámetros por modelo.....	35
Tabla 7 Comparación base de modelos en validación walk-forward.....	41
Tabla 8 Mejor Variante del Random Forest por Serie.....	43
Tabla 9 Configuración Óptima Final	46
Tabla 10 Pronóstico Final Enero-Marzo 2026	47
Tabla 11 Ajustes Metodológicos aplicados al pipeline final	47

INTRODUCCIÓN

Durante la última década, las naciones con mayor desarrollo tecnológico han consolidado una transformación estructural en la operación de sus sistemas productivos [1]. En el sector agroindustrial, esta evolución ha sido impulsada de manera directa por la proliferación del Internet de las Cosas (IoT) [2], y la implementación masiva de sistemas inteligentes, tales como sensores de monitoreo en campo, maquinaria interconectada y ecosistemas avanzados de almacenamiento en la nube. Esta robusta infraestructura tecnológica ha habilitado la generación continua de variables operacionales y ambientales, facilitando la generación de pronósticos en tiempo real respaldados por análisis multivariados; esto permite una anticipación de la demanda significativamente más dinámica, oportuna y precisa en épocas anteriores.

Bajo este escenario de alta digitalización, la precisión en los pronósticos de ventas ha dejado de ser un mero indicador administrativo para regirse como una ventaja competitiva determinante para las organizaciones. Disponer de una exactitud rigurosa en las proyecciones comerciales faculta a las empresas para optimizar el flujo de caja, perfeccionar la planeación de corto y mediano plazo, y mitigar los riesgos financieros asociados tanto al desabastecimiento de materias primas como el sobre stock de producto terminando. [1]. Asimismo, una predicción metodológicamente sólida atenúa los efectos adversos del denominado "efecto látigo" (bullwhip effect) a lo largo de los eslabones de la cadena de suministro, posicionando al pronóstico como una herramienta estratégica indispensable para la toma de decisiones informadas y la sostenibilidad empresarial [3].

No obstante, al trasladar este análisis a contextos de economías emergentes como la colombiana, se observa que una proporción significativa de empresas del sector agroindustrial enfrenta desafíos críticos derivados márgenes de error en sus proyecciones de ventas. La estimación exacta de los volúmenes de comercialización futuros resulta una tarea compleja y de alto costo logístico, explicada principalmente por la volatilidad intrínseca del mercado y la marcada influencia de factores exógenos de naturaleza estocástica. Elementos como las condiciones climáticas cambiantes, las estrategias de la competencia, la fluctuación de los precios internacionales de los commodities y los eventos socioeconómicos locales, configuran un entorno cuya variabilidad resulta sumamente difícil de cuantificar e incorporar a través de los métodos estadísticos tradicionales univariados.

Esta baja asertividad en la previsión de la demanda expone a las organizaciones agroindustriales a vulnerabilidades operativas y financieras considerables, una problemática que se agrava sustancialmente al tratar con líneas de productos vinculadas a ciclos biológicos y bienes perecederos. En estos escenarios, la inexactitud estructural de los modelos predictivos convencionales suele derivar en el rápido deterioro de los inventarios o rupturas críticas en la cadena de abastecimiento, comprometiendo de manera directa la propuesta de valor y el nivel de

servicio al cliente. Adicionalmente, las exigencias de inserción competitiva en los mercados globales demanda estándares de eficiencia que obligan a las gerencias modernas a dominar herramientas avanzadas de gestión de la incertidumbre para reaccionar oportunamente ante las dinámicas macroeconómicas mundiales.

Ante este panorama de alta complejidad, se hace imperativo explorar y adoptar nuevas metodologías científicas fundamentadas en inteligencia artificial y aprendizaje automático multivariado (Machine Learning). La integración sistemática de múltiples variables del entorno operativo, económico y climático dentro de los modelos matemáticos avanzados permite desarrollar sistemas de pronóstico más precisos, adaptativos y funcionales, sentando así las primeras bases analíticas necesarias para una planeación comercial y logística verdaderamente estratégica en el agro de Colombia.

1. CONTEXTUALIZACIÓN DEL PROYECTO

1.1. Planteamiento y Definición del problema

En el entorno empresarial contemporáneo, la planificación de la demanda se ha consolidado como un desafío estratégico crítico, particularmente para aquellas organizaciones expuestas a una alta volatilidad en sus mercados de comercialización y abastecimiento. Históricamente, la práctica convencional en la agroindustria se ha apoyado en métodos estadísticos de pronóstico basados en series de tiempos univariadas, talentos como el promedio Móvil, Holt-Winters o los modelos autoagresivos de media móvil (ARIMA) [4]. Si bien estos enfoques logran capturar dependencias temporales básicas, presentan limitaciones estructurales al omitir la incorporación de variables exógenas que resultan determinantes en el comportamiento de las ventas. Esta carencia metodológica deriva en márgenes de error sustanciales en la estimación de la demanda, lo cual precipita ineficiencias operativas que se traducen en sobre costos logísticos, rupturas de stock (stockouts) o, en su defecto, una acumulación excesiva capital de trabajo inmovilizado.

Pese a la proliferación de herramientas de Inteligencia Artificial en el mercado actual, se observa una adopción limitada de algoritmos de aprendizaje automático (Machine Learning) de naturaleza multivariada en el tejido empresarial agroindustrial colombiano. Esta brecha tecnológica impide a las organizaciones integrar variables complejas del entorno (como el clima o la macroeconomía), restringiendo la capacidad predictiva y la resiliencia de sus operaciones frente a perturbaciones del mercado.

Ante este panorama, surge la necesidad de investigar empíricamente el grado en que la implementación de modelos de Machine Learning multivariado puede superar las deficiencias de los enfoques tradicionales, proveyendo a las empresas de una base analítica robusta para una toma de decisiones eficiente y competitiva.

1.1.1. Formulación del problema

A partir de los planteamientos anteriores, la presente investigación se orienta a responder el siguiente interrogante central:

¿En qué medida la implementación de modelos predictivos multivariados basados en Machine Learning incrementa la precisión de los pronósticos de ventas en el sector agroindustrial, en contraste con los métodos estadísticos univariados tradicionales

1.1.2. Sistematización

Para dar respuesta metodológica a la pregunta de investigación, se desagregan los siguientes interrogantes específicos

- ¿Cuáles son las variables exógenas (Climáticas, macroeconómicas y de abastecimiento) que influyen significativamente y presentan relaciones de causalidad temporal sobre la demanda de productos agroindustriales?
- ¿Qué métricas estadísticas y esquemas de validación temporal resultan idóneos para contrastar objetivamente el desempeño predictivo de los modelos clásicos frente a los algoritmos de Machine Learning?
- ¿Cuál es la magnitud de la mejora en la precisión del pronóstico lograda al incorporar modelos de ensamble multivariados, diferenciando el impacto sobre el volumen de unidades (Cantidad) frente al margen bruto?
- ¿De qué manera la adopción de modelos multivariados optimiza los procesos de planeación comercial y la gestión táctica de la cadena de suministro frente a escenarios de incertidumbre?

1.2. Objetivos

1.2.1. Objetivo General

Comparar el desempeño de modelos multivariados de Machine Learning frente a métodos estadísticos tradicionales en el pronóstico de ventas, con el propósito de evaluar su capacidad de reducir la incertidumbre y optimizar la planeación de la demanda en empresas del sector agroindustriales.

1.2.2. Objetivos Específicos

1. **Caracterizar** el comportamiento histórico de las ventas (analizando independientemente el volumen y margen bruto), identificando patrones de tendencia, ciclos, estacionalidad y variabilidad temporal.
2. **Desarrollar y evaluar** modelos estadísticos de referencia (como Mínimos Cuadrados Ordinarios y SARIMAX) superando las limitaciones de reactividad univariada presentes en modelos de suavización como Holt-Winters o el Promedio Móvil) para establecer una línea base de desempeño predictivo, validada mediante métricas de error estandarizadas (MAE, RMSE y MAPE).
3. **Identificar, preprocesar e integrar** variables exógenas relevantes, tales como índices climáticos zonales, abastecimiento agrícola (SIPSA) y tasas de cambio representativas determinando su rezago óptimo y casualidad estadística sobre la serie objetivo
4. **Implementar y optimizar** modelos multivariados de Machine Learning (incluyendo arquitecturas como Random Forest, LightGBM y Prophet) que incorporen las variables explicativas seleccionadas.
5. **Contrastar** el desempeño algorítmico global mediante esquemas rigurosos de validación temporal secuencial (walk-forward), determinando la granularidad y la configuración metodológica óptima que fundamente la toma de decisiones logísticas.

1.3. Marco de Referencia

El desarrollo de un sistema predictivo robusto para la planeación de la demanda requiere un abordaje conceptual interdisciplinario. Esta sección unifica los fundamentos macroeconómicos, las teorías de gestión logística y los paradigmas algorítmicos que sustentan la transición desde los modelos estadísticos tradicionales hacia soluciones multivariadas de aprendizaje automático.

1.3.1. Marco Teórico

1.3.1.1. Planificación de la Demanda en la Cadena de Suministro bajo Entornos Estocásticos

En el ámbito empresarial agroindustrial, la gestión de la cadena de suministro opera bajo condiciones de alta entropía, características de los entornos VUCA (Volatilidad, Incertidumbre, Complejidad y Ambigüedad). La planificación de la demanda trasciende la mera estimación de inventarios; se consolida como núcleo analítico del proceso de planificación de ventas y operaciones (S&OP), responsable de alinear los flujos de abastecimiento, manufactura y distribución.

La literatura logística tradicional asume que la demanda es un reflejo endógeno de las ventas históricas. No obstante, en mercados expuestos a perturbaciones externas, la demanda real se comporta como un sistema dinámico estocástico influenciado por vectores macroeconómicos, dinámicas sectoriales y variables biofísicas. La subestimación de estas covariables exógenas induce perturbaciones operacionales críticas, entre las que se destaca el “Efecto látigo” (Bullwhip effect), donde pequeñas distorsiones en el punto de venta amplifican de forma exponencial hacia los eslabones de producción y proveeduría, generando costos hundidos por obsolescencia o pérdida de participación de mercado por rupturas de inventario (Stockouts) [5].

1.3.1.2. Modelos estadísticos Clásicos de Series de Tiempo e Inexactitud Estructural

Históricamente, los algoritmos univariados lineales han dominado el pronóstico comercial debido a su parsimonia y bajo costo computacional. Los modelos autorregresivos integrados de media móvil (ARIMA), formalizados por Box y Jenkins, modelan la variable objetivo a partir de sus propios rezagos y de los errores pasados de predicción, asumiendo condiciones de linealidad y estacionariedad logradas mediante diferenciación. Su extensión estacional, SARIMAX, incorpora componentes estacionales de largo plazo e introduce regresores exógenos mediante transferencias lineales [5].

Sin embargo, estos enfoques presentan una marcada inexactitud estructural ante series con heterocedasticidad condicional y picos abruptos. La premisa fundamental de SARIMAX presupone que la relación entre las covariables y la variable objetivo lineal y constante a lo largo del tiempo.

En la agroindustria, perturbaciones exógenas como anomalías climáticas extremas o fluctuaciones bruscas de precios internacionales rompen los supuestos de estabilidad paramétrica, inhabilitando a los modelos lineales para capturar shocks que no sigan un patrón sinusoidal o lineal estricto.

1.3.1.3. Aprendizaje Automático No Lineal y Métodos de Ensamble

Para superar las limitaciones lineales, el aprendizaje automático (Machine Learning) introduce arquitecturas no paramétricas capaces de mapear relaciones no lineales complejas e interacciones de alto orden entre características, sin requerir supuestos a priori sobre la distribución de los residuos [1], [2], [3] [6].

- **Random Forest (RF):** Basado en el principio de agregación por bacheo (bagging), construye múltiples árboles de decisiones independientes entrenados sobre subconjuntos aleatorios de los datos y de las variables predictoras. La salida final promedia las predicciones individuales, lo que reduce drásticamente la varianza del modelo y mitiga el riesgo de sobreajuste (overfitting) en series temporales cortas o ruidosas. Su capacidad para evaluar interacciones conjuntas entre rezagos temporales, dummies de calendario y variables climáticas lo hace altamente efectivos en escenarios de alta volatilidad.
- **LightGBM (Light Gradient Boosting Machine):** Pertenece a la familia de algoritmos de potenciación de gradiente. A diferencia del crecimiento por niveles convencional, LightGBM emplea un crecimiento por hojas optimizado con histogramas, buscando la máxima reducción de la pérdida de gradiente. Esta arquitectura permite una velocidad de convergencia superior y un manejo eficiente de la regulación mediante el control de la tasa de aprendizaje (Learning rate) y muestras mínimas por hoja. Es especialmente útil para detectar cambios de tendencia secuenciales y patrones latentes de transmisión exógena.

1.3.1.4. Aprendizaje Automático No Lineal y Métodos de Ensamble

Un desafío central en la Ciencia de Datos aplicada a series temporales en la selección de la granularidad operativa. Confirme se aumenta la frecuencia de los datos (de diario a semanal, o de semanal a mensual), se altera la relación señal-ruido del sistema [11].

La teoría econométrica de la agregación temporal demuestra que la acumulación de los datos en bloques de mayor horizonte actual como un filtro de paso bajo que remueve componentes ruidosos de alta frecuencia (como irregularidades transaccionales del día a día o anomalías operativas puntuales). Al estabilizar la varianza, las señales de variables macroeconómicas u operacionales, cuyos canales de transmisión operan por acumulación o de forma inercial,

emergen con mayor nitidez analítica, incrementando el asertividad de los algoritmos predictivos en la escala mensual.

1.3.1.5. Importancia del enfoque multivariado en contextos volátiles

La naturaleza estocástica del sector agroindustrial genera volúmenes masivos de datos heterogéneos y no lineales. La convergencia de ciclos biológicos, condiciones agroclimáticas y dinámicas de mercado global configura un entorno de alta entropía que desafía la rigidez de los modelos estadísticos convencionales [1], [7].

La implementación de modelos multivariados permite decodificar esta complejidad. A modo ilustrativo, el despliegue de redes neuronales o modelos de ensamble entrenados con datos históricos, patrones pluviométricos e índices de precios internacionales, faculta a las organizaciones para anticipar con mayor exactitud fenómenos como el incremento en la demanda de fertilizantes asociado a anomalías climáticas específicas. Esta capacidad anticipatoria es crucial para ajustar los planes de producción, optimizar la logística de distribución y minimizar la exposición al riesgo operativo. En esencia, la adopción de algoritmos multivariados no solo refina la métrica de error, sino que incrementa la resiliencia organizacional frente a la volatilidad.

1.3.1.6. Canales de Transmisión Económica y biofísica en la Demanda Agroindustrial

Los modelos multivariados construidos en esta investigación se fundamentan en cuatro dimensiones casuales exógenas:

1. **Tipo de Cambio (TRM USD/COP y EUR/COP):** Transmite fluctuaciones directas sobre la estructura de los costos de la organización. Al determinar el costo de importación de materias primas críticas (Como aceros o componentes manufacturados), una devaluación del peso colombiano presiona los costos de reposición y altera de forma inmediata los márgenes comerciales. Su impacto sobre el volumen de ventas suele operar con un rezago diferido por los ciclos de inventario previo y renegación de lista de precios.
2. **Precio Internacional del Petróleo (Referencia Brent):** actúa como un predictor adelantando de los costos logísticos globales y de manufactura de insumos derivados (plásticos, empaques, combustibles de distribución). Su transmisión hacia el comportamiento de la rentabilidad es directa y deferida en el tiempo.
3. **Abastecimiento Sectorial (SIPSA):** El sistema de información de Precios y de Abastecimiento del Sector Agropecuario sirve como una variable proxy de la dinámica real del campo colombiano. El volumen de comercialización y los precios de los productores agrícolas reflejan el éxito en las cosechas y la liquidez del productor,

correlacionando directamente con su capacidad de reinversión en herramientas agrícolas.

4. **Índices Climáticos Zonales (Precipitación y Temperatura NASA POWER):** Las variables biofísicas imponen restricciones directas sobre la actividad agrícola y constructiva. La precipitación acumulada actúa como un disparador estacional de la demanda de herramientas agrícolas en zonas específicas, mientras que temperaturas inusualmente elevadas ralentizan las labores de campos manuales, desincentivando la rotación de portafolio.

Para garantizar la reproductividad científica del pipeline de datos y evitar el sesgo por agregación nacional que diluye la varianza climática, la extracción de las variables biofísicas de precipitación acumulada y temperatura media se parametrizó geográficamente mediante coordenadas del sistema NASA POWER. La delimitación espacial de los tres índices zonales analizados se estructuró a partir de los nodos comerciales y de distribución logística de Herrago S.A.S en el territorio colombiano:

- **Zona Andina:** Centrada en los departamentos del Eje cafetero y Antioquia, delimitada en el cuadrante geográfico de latitudes 4.5°N a 6.5°N y longitudes -75.8°O a -74.5°O.
- **Zona Caribe:** Orientada a capturar la dinámica de la costa atlántica, parametrizada entre las latitudes 10.0°N a 11.5°N y longitudes -75.5°O a -74.0°O.
- **Zonas Llanos:** Enfocada en la región de la Orinoquia colombiana (Meta), delimitada entre latitudes 3.5°N a 4.5°N y longitudes -73.8°O a -72.5°O.

Las series diarias capturadas en estas celdas satelitales fueron posteriormente acumuladas en vectores semanales y mensuales, sirviendo como regresores validados bajo el esquema VIF.

1.3.1.7. El pronóstico como herramienta estratégica en la cadena de suministro

Más allá de su función técnica, el pronóstico de demanda contemporáneo se integra como un componente neurálgico en la Planificación de Ventas y Operaciones (S&OP). En concordancia con Gutiérrez Duque et al. [10], la infusión de inteligencia artificial en los sistemas de planeación logística permite una alineación sincrónica de los procesos de abastecimiento, manufactura y distribución. Esta integración reduce las redundancias operativas y dota a la organización de la flexibilidad necesaria para responder a las fluctuaciones del mercado.

En la agroindustria, donde las ventanas productivas son rígidas y perecederas, disponer de modelos robustos de pronóstico multivariado es determinante para la toma de decisiones críticas

sobre siembra, cosecha y compras especulativas, impactando directamente en la estructura de costos y la rentabilidad global del negocio.

1.3.1.8. Divergencia Estructural entre Modelos de Volumen (Cantidad) y Rentabilidad (Margen Bruto)

Una contribución metodológica crítica de este estudio es el tratamiento diferenciado de la cantidad de unidades vendidas y el margen bruto como dos problemas predictivos independientes [12].

Matemáticamente, el volumen físico está fuertemente acoplado a ciclos calendario y estacionalidad biológica estable. Por lo contrario, el margen bruto es una variable compuesta no lineal: es el resultado de multiplicar las cantidades por el precio neto de venta, menos el costo de los bienes vendidos. Por ende, el margen incorpora volatilidad exógena adicional derivada de descuentos comerciales, variaciones cambiantes en insumos, cambios en la mezcla de portafolio y estrategias agresivas de precios. Tratar el margen como una serie puramente estacional induce un sesgo metodológico, requiriendo en su lugar el despliegue de modelos flexibles de asimilar shocks financieros no lineales.

1.3.1.9. Beneficios de la Inteligencia Artificial en la reducción de incertidumbre

La adopción de modelos de Inteligencia Artificial para el análisis de la demanda constituye un mecanismo efectivo para reducir la incertidumbre inherente a los mercados agropecuarios. La capacidad de estos sistemas para el aprendizaje autónomo y la actualización dinámica de parámetros en tiempo real confiere a las empresas una ventaja competitiva sustancial [7].

Asimismo, desde una perspectiva macro, la aplicación de Machine Learning en este sector contribuye a abordar desafíos globales como la seguridad alimentaria y la sostenibilidad, al propiciar una asignación más eficiente de recursos productivos, la minimización de desperdicios en la cadena de valor y la optimización del impacto ambiental de las operaciones logísticas.

1.3.2. Antecedentes

Tabla 1 Aporte de Antecedentes

Nº	Referencia	Resumen ampliado	Aporte/Diferencia
[1]	Gonçalves et al. (2021)	El estudio propone un modelo multivariado para pronóstico de demanda en la industria automotriz, integrando variables externas (precios, condiciones económicas) e internas,	Fundamenta la necesidad de incorporar variables adicionales en modelos de pronóstico. El presente trabajo traslada el enfoque

		utilizando técnicas estadísticas avanzadas y redes neuronales. Se evidencia una mejora sustancial en la precisión de los pronósticos multivariados, especialmente en horizontes de mediano y largo plazo.	multivariado al sector agroindustrial, incorporando algoritmos de Machine Learning más modernos.
[2]	Xu et al. (2022)	Realiza una revisión sistemática sobre la aplicación de tecnologías IoT e IA en agricultura. Se destaca el uso de sensores y dispositivos inteligentes como fuente de datos en tiempo real, que enriquecen los procesos de toma de decisiones en producción agrícola.	Aunque no aborda directamente el pronóstico de ventas, refuerza la disponibilidad y valor de nuevas fuentes de datos que pueden alimentar modelos predictivos más precisos en el agroindustrial.
[3]	Hofmann y Rutschmann (2018)	Presentan un marco conceptual sobre el impacto del Big Data y la analítica avanzada en la planeación de demanda dentro de cadenas de suministro, destacando cómo el procesamiento de grandes volúmenes de datos permite generar pronósticos más adaptativos, dinámicos e integrados.	Aporta sustento teórico sobre la importancia de integrar múltiples fuentes de datos en el pronóstico de ventas. Refuerza la pertinencia de enfoques multivariados en entornos complejos como el agroindustrial.
[4]	Kumar et al. (2024)	Compara los modelos clásicos Holt-Winters y ARIMA en la optimización de inventarios de cadenas de suministro, demostrando sus ventajas y limitaciones según la variabilidad de los datos. El análisis se mantiene en el ámbito univariado.	Sirve como referencia de los métodos tradicionales de pronóstico, contra los cuales se evaluará el desempeño de los modelos multivariados de ML propuestos en este trabajo.
[5]	Aubad Acebedo & Valencia Villa (2024)	Desarrollan un modelo SARIMA para pronóstico de ventas en una empresa agroindustrial colombiana, empleando únicamente datos históricos. El modelo permite obtener estimaciones	Refleja el uso actual de métodos univariados en el contexto agroindustrial nacional. El presente proyecto busca superar estas limitaciones integrando

		razonables en escenarios estables, pero no incorpora variabilidad externa.	variables exógenas mediante ML.
[6]	Liaw & Wiener (2002)	Presentan el algoritmo Random Forest, detallando su funcionamiento para problemas de clasificación y regresión. Se destaca su capacidad de manejar alta dimensionalidad y relaciones no lineales.	Soporta la elección de algoritmos de Machine Learning robustos como Random Forest, que serán evaluados en el marco comparativo de este proyecto aplicado al agroindustrial.
[7]	Fernando et al. (2024)	Analiza el impacto transformador de la IA y el aprendizaje autónomo sobre la sostenibilidad y eficiencia de la producción agrícola. Resalta cómo el uso de algoritmos inteligentes puede optimizar múltiples aspectos del proceso productivo.	Aunque no aborda directamente el pronóstico de ventas, destaca el potencial de la IA para generar eficiencia en el sector agrícola, contexto donde se desarrollará el presente estudio.
[8]	Rueda Rincón (2022)	Estudia la aplicación de diferentes algoritmos de clasificación para la predicción de categorías de ventas, utilizando técnicas supervisadas de aprendizaje automático para mejorar la precisión en escenarios de clasificación.	Aporta evidencia de la aplicación de algoritmos ML en problemáticas comerciales, aunque su foco es clasificación categórica, mientras este proyecto aborda problemas de regresión continua en ventas agroindustriales.
[9]	Fierro Torres et al. (2022)	Realiza un análisis comparativo entre modelos tradicionales (ARIMA) y modernos (redes neuronales) para el pronóstico de demanda. Los resultados evidencian una mejor capacidad predictiva de las redes neuronales en entornos con alta variabilidad.	Comparte el enfoque comparativo del presente trabajo, pero se limita al contraste entre ARIMA y redes neuronales. Este proyecto ampliará el análisis incorporando otros algoritmos ML y considerando variables multivariadas.

2. CONJUNTO DE DATOS Y ANÁLISIS EXPLORATORIO

2.1. Descripción y Adquisición de los Datos

Para el desarrollo de la presente investigación, se consolidó un conjunto de datos estructurados que abarca el periodo transaccional comprendido entre el 1 de enero de 2022 y el 31 de diciembre de 2025. La arquitectura de información fue diseñada para integrar las señales internas de desempeño comercial de la organización agroindustrial con covariables exógenas representativas del entorno macroeconómico, climático y de abastecimiento sectorial.

Las fuentes de información se homologaron temporalmente para permitir su evaluación en tres granularidades prospectivas: Diaria, Semanal, Mensual. El conjunto de datos base se construyó a partir de las siguientes dimensiones:

1. **Variables Objetivo (Ventas):** Datos transaccionales internos desagregados por familia de producto (enfocándose en las referencias clave: PALAS y CARRETAS) y expresados en dos métricas operativas: Volumen Físico (Cantidad) y rentabilidad (Margen Bruto).
2. **Variables Macroeconómicas:** Tasas de cambio representativas (TRM USD/COP y EUR/COP) y precios internacionales de petróleo (referencia Brent), adquiridos en frecuencia diaria.
3. **Variables de Abastecimiento (SIPSA):** Volúmenes de abastecimiento de grupos agrícolas (Granos y Cereales, Tubérculos, entre otros), obtenidos del sistema de información de precios y abastecimiento del sector agropecuario.

Tabla 2 Fuentes de información, periodicidad original, variables extraídas y transformaciones metodológicas aplicadas

Fuente de Información	Periodicidad Original	Variables Base Extraídas	Agregación Analítica	Tratamiento Metodológico y Transformación
Ventas Internas (ERP)	Diaria	- Cantidad (Volumen) - M_Margen_Bruto (Rentabilidad) - Pro_Familia (PALAS,CARRETAS)	Semanal y Mensual	Filtrado transversal por familia. Agrupación temporal (Suma de volúmenes). Imputación estructurada de ceros en fechas sin transacción

				para evitar sesgos de completitud. Generación de rezagos autorregresivos (lags).
DANE - SIPSA	Diaria (Días Hábles)	- Grupo de Alimento - Precios Mayoristas - Volúmenes de Abastecimiento	Semanal y Mensual	Depuración semántica y estandarización de categorías. Exclusión de ruido sectorial (eliminación del grupo de CARNES). Agregación acumulativa para medir el dinamismo general de la actividad agrícola de campo.
Commodities Energéticos	Diaria	- Petróleo Brent (USD/bbl)	Semanal y Mensual	Suavizamiento mediante promedio temporal ponderado. Evaluada estadísticamente como covariante macroeconómico para modelar el componente indirecto de costos de

				manufactura y fletes en el margen bruto.
Banco de la República (Tasas de Cambio)	Diaria	- TRM_USD_COP - TRM_EUR_COP - TRM_EUR_USD	Semanal y Mensual	Promedio temporal por ventana. Diagnóstico de raíz unitaria: Identificadas como procesos integrados. Por lo cual se requirió la transformación mediante primeras diferencias logarítmicas. Sometidas a control estricto de multicolinealidad (VIF).
NASA POWER (Datos Satelitales=	Diaria	- Temperatura media (temp_media_c) - Precipitación (precip_mm_dia)	Semanal y Mensual	En lugar de promedios nacionales, se ejecutó una construcción espacial de índices climáticos zonales (Andino, Caribe y Llanos) para evitar pérdida de varianza regional. Series

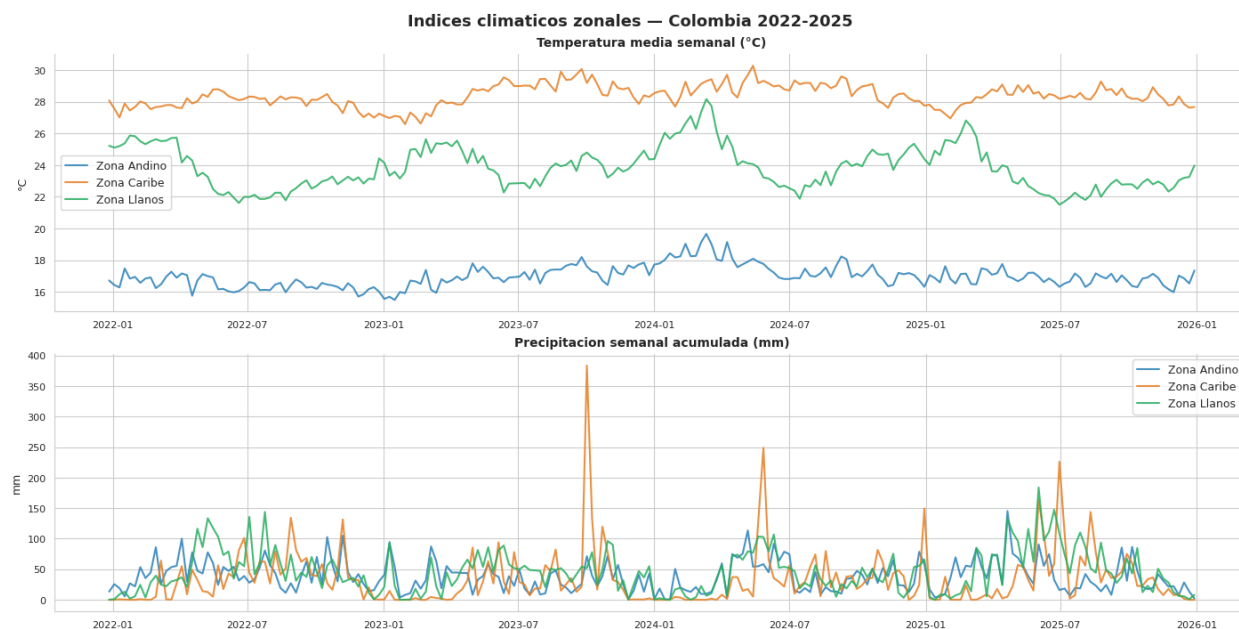
				identificadas como estacionarias.
--	--	--	--	-----------------------------------

2.2. Preparación, Transformación e Integración

La fase de preprocesamiento garantizó la calidad y coherencia del pipeline analítico. Se aplicaron transformaciones sustantivas sobre las fuentes originales para adaptarlas a la naturaleza del problema predictivo. Para los datos transaccionales, se aplicó una estrategia de completitud temporal donde la ausencia de registros de ventas en fechas específicas fue imputada con ceros; bajo la premisa metodológica, la “no-venta” se interpreta como un evento informativo legítimo sobre la inactividad de la demanda, y no como un valor perdido (missing value)

Respecto a las covariables, se aplicaron filtros de pertinencia sectorial (e.g, la exclusión del grupo CARNES en la base SIPSA) y se diseñó un enfoque de espacialidad climática. En lugar de utilizar promedios nacionales que pidieron ocultar la heterogeneidad ambiental del país, se construyeron índices climáticos extremos focalizados que impactan directamente el dinamismo de la siembra regional y, por consiguiente, el uso de herramientas agropecuarias.

Ilustración 1 índices climáticos zonales semanales utilizados como variables exógenas, 2022-2025



2.3. Análisis Exploratorio de la Variable Objetivo: Volumen vs Rentabilidad

Un pilar analítico de esta investigación consistió en evaluar si el volumen de unidades (Cantidad) y la rentabilidad (Margen Bruto) exhibían un comportamiento homólogo que permitiera su modelado bajo una misma especificación. El análisis de correlación interna reveló que la asociación entre ambas métricas es moderada a nivel diario ($r \approx 0.43$) y se debilita progresivamente al agregarse en ventanas mensuales ($r \approx 0.26$ a 0.34). Esta divergencia estadística confirma que el margen bruto incorpora una varianza exógena sustancial dictada por fluctuaciones cambiarias, decisiones de pricing y costos de reposición, la cual no puede explicarse únicamente a través del volumen transado.

Tabla 3 Correlación interna entre cantidad y margen bruto por familia y granularidad

Familia	Diaria	Semanal	Mensual
PALAS	0.43	0.34	0.34
CARRETAS	0.44	0.32	0.26

Adicionalmente, se ejecutó una descomposición de series de tiempo para evaluar la fuerza estacional de las variables objetivo. Los resultados se evidenciaron una estacionalidad excepcionalmente fuerte en el volumen de ventas (PALAS: 0.670; CARRETAS: 0.727), fuertemente vinculada a los calendarios biológicos de siembra. En contraste, las series de margen bruto presentaron una señal de estacionalidad débil (PALAS: 0.016; CARRETAS: 0.162), comportándose como un proceso más sensible a shocks económicos abruptos. Este hallazgo legitimó la decisión metodológica de modelar la cantidad y margen bruto como problemas de Machine Learning conceptual y algorítmicamente distintos).

Ilustración 2 Descomposición estacional de ventas en cantidad y margen bruto



2.4. Dinámica de Variables Exógenas y Diagnóstico de Estacionariedad

Para garantizar la viabilidad matemática de la inclusión de predictores en los modelos multivariados, se realizó un diagnóstico riguroso de la integración estacionaria de las covariables mediante las pruebas de Dickey-Fuller Aumentada (ADF) y Kwiatkowski-Phillips-Schmidt-Shin (KPSS).

Los resultados identificaron una heterogeneidad en el orden de integración: las variables climáticas (Precipitación y Temperatura) y los índices de abastecimiento (SIPSA) resultaron ser estacionarias en niveles $I(0)$, mientras que los indicadores cambiarios (TRM) exhibieron raíces unitarias $I(1)$. En consecuencia, las variables $I(1)$ fueron transformadas mediante primeras diferencias logarítmicas previo al cálculo de sus funciones de correlación, evitando así la captura de relaciones espurias inducidas por tendencias comportadas.

Tabla 4 Diccionario de variables exógenas del modelo

Código	Nombre	Fuente	Unidad	Frecuencia original	Transformación	Orden I	Justificación teórica
TRM_USD_COP	Tasa Representativa del Mercado USD/COP	Banco de la República / CSV diario	COP por USD	Diaria	Promedio mensual; $\Delta \log \text{TRM}$ para modelos en diferencias	$I(1)$	Determina el costo de importación de herramientas y aceros; apreciación cambiaria reduce precios de importados y presiona márgenes.
TRM_EUR_COP	Tasa de cambio EUR/COP	Banco de la República / CSV diario	COP por EUR	Diaria	Promedio mensual; $\Delta \log \text{TRM}$ para modelos en diferencias	$I(1)$	Relevante para importaciones de maquinaria y accesorios de

							origen europeo; correlación con el margen bruto de PALAS y CARRETAS.
TRM_EUR_USD	Paridad EUR/USD (cruzada)	Derivada de TRM_USD_COP y TRM_EUR_COP	USD por EUR	Diaria	Promedio mensual	I(1)	Captura el diferencial cambiario entre zonas de proveedor; presenta causalidad Granger sobre PALAS-Margen Bruto y CARRETAS-Margen Bruto.
PETRÓLEO	Precio spot del petróleo (Brent o referencia)	Fuente de precios internacionales / CSV diario	USD por barril	Diaria	Promedio mensual; sin diferenciación I(0))	I(0)	Afecta costos de transporte, plásticos y manufactura; correlación positiva con CARRETAS-Margen

							Bruto (+0.393) y negativa con CARRETA S-Cantidad (-0.290).
SIPSA_GRANOS	Abastecimiento SIPSA — Grupo Granos y Cereales	DANE/SIPSA (Sistema de Información de Precios y Abastecimiento del Sector Agropecuario)	Toneladas métricas	Diaria (días hábiles)	Suma semanal/mensual; I(0)	I(0)	Proxy del dinamismo del sector agropecuario; causalidad Granger sobre CARRETA S-Margen Bruto (p=0.0017) y PALAS-Margen Bruto.
CLIMA_AND_PREC	Precipitación zona Andina (índice zonal)	NASA POWER / 3 estaciones (Armenia, Pereira, Manizales)	mm/día (promedio zonal)	Diaria	Suma semanal/mensual; I(0)	I(0)	Refleja condiciones agroclimáticas en la zona cafetera principal; precipitación positiva estimula demanda de herramientas

							agrícolas. Causalidad Granger sobre PALAS- Cantidad (p=0.0066, lag=3).
CLIMA_CAR_PREC	Precipitación zona Caribe (índice zonal)	NASA POWER / ciudades costa Atlántica	mm/día (promedio zonal)	Diaria	Suma semanal/mensual; I(0)	I(0)	Predictor final de CARRETA S- Cantidad (lag=4, VIF aceptable); asocia condiciones de lluvias con actividad constructiva costera.
CLIMA_LLA_REC	Precipitación zona Llanos (índice zonal)	NASA POWER / ciudades región Orinoquia	mm/día (promedio zonal)	Diaria	Suma semanal/mensual; I(0)	I(0)	Segunda zona climática con causalidad Granger sobre PALAS- Cantidad (p=0.0193, lag=3); captura dinámica agropecuaria del piedemo

							nte llanero.
CLIMA_TEMP_MED	Temperatura media nacional (índice multi-zona)	NASA POWER / promedio 11 ciudades	°C	Diaria	Promedio semanal/mensual; I(0)	I(0)	Correlación negativa con cantidad en ambas familias (PALAS: -0.433; CARRETS: -0.376). Temperaturas altas asociadas a reducción de actividad agrícola manual.

Tabla 5 Propiedades estadísticas y de integración de las variables exógenas

Variable	Media	Desv. Est.	Mínimo	Máximo	Orden I	d ADF-sugerido	Observaciones
TRM_USD_COP	~4,250 COP	~280 COP	~3,720	~4,900	I(1)	1	Raíz unitaria confirmada por ADF y KPSS; diferenciación necesaria antes de CCF.
TRM_EUR_COP	~4,650 COP	~290 COP	~4,020	~5,100	I(1)	1	Cointegrada con TRM_USD_COP; VIF elevado cuando se incluyen ambas simultáneamente.

TRM_EUR_USD	~1.09 USD	~0.06 USD	~0.97	~1.20	I(1)	1	Excluida del modelo final por multicolinealidad con los dos TRM anteriores.
Petróleo (bbl)	~85 USD	~12 USD	~68	~120	I(0)	0	Estacionaria en niveles; puede incluirse sin diferenciación en especificaciones ARDL.
SIPSA Granos	Variabl e	Variabl e	—	—	I(0)	0	Estacionaria en niveles; causalidad Granger con CARRETAS-MB ($p=0.0017$, lag=1).
Clima_And_Prec	~5.2 mm/día	~3.1 mm	~0.2	~14.8	I(0)	0	Sin raíz unitaria; causalidad Granger sobre PALAS-Cantidad ($p=0.0066$, lag=3 sem).
Clima_Car_Prec	~4.8 mm/día	~3.4 mm	~0.1	~15.2	I(0)	0	Predictor final retenido para CARRETAS-Cantidad con VIF dentro del umbral permitido.
Clima_Lla_Prec	~5.9 mm/día	~3.8 mm	~0.1	~18.5	I(0)	0	Causalidad Granger sobre PALAS-Cantidad ($p=0.0193$, lag=3 sem).
Clima_Temp_Med	~20.4 °C	~2.1 °C	~14.2	~27.8	I(0)	0	Correlación negativa con cantidad en ambas familias;

							mayor efecto en granularidad mensual.
--	--	--	--	--	--	--	---------------------------------------

2.5. Determinación de Rezagos Óptimos (Cross-Correlation Function)

Dado que el impacto de la macroeconomía y el clima sobre la demanda no es instantáneo, se empleó la función de Correlación cruzada (CCG) sobre las series estacionarizadas para identificar la latencia temporal (rezago o lag operativo) en la que el impacto de cada regresor exógeno alcanza su máxima significancia estadística.

El análisis demostró que las señales operaban bajo mecanismos de transmisión diferida:

- **Señales Biofísicas:** La precipitación climática presenta su mayor correlación con un desplazamiento de 3 a 4 semanas sobre el volumen físico de PALAS y CARRETAS, capturando el tiempo de respuesta biológica en campo.
- **Señales Macroeconómicas:** Las fluctuaciones de la TRM impactan el Margen bruto con un rezago estimado de 2 a 3 semanas, consistente con el ciclo logístico de importación de materias primas y las políticas de ajuste en listas de precios comerciales.
- **Costos Logísticos:** El precio de petróleo impacta el Margen Bruto de CARRETAS con un rezago largo de 5 semanas, sugiriendo una transmisión lenta a través de la cadena de suministro de polímeros y manufactura pesada.

Para la construcción final de la matriz de características (Feature Engineering), los predictores candidatos fueron sometidos a un análisis de causalidad de Granger y se filtraron mediante el factor de inflación de la varianza (VIF), imponiendo un umbral conservador ($VIF \leq 5$) para garantizar la ortogonalidad relativa de los predictores y prevenir problemas de multicolinealidad estructural en los algoritmos predictivos.

3. METODOLOGÍA DE MODELADO Y VALIDACIÓN

El Diseño metodológico de esta investigación se estructuró como un ecosistema analítico reproducible y modular, desarrollado en el lenguaje de programación Python mediante el uso de librerías especializadas en computación científica y modelado predictivo, tales como **pandas** y **numpy** para la manipulación y transformación de matrices de datos; **statsmodels** para la ejecución de pruebas estadísticas robustas y estimaciones lineal; y **scikit-learn** para el despliegue de las arquitecturas de aprendizaje automático y esquemas de validación temporal.

3.1. Arquitectura de Pipeline Analítico

La ejecución de flujo de ciencia de datos se organizó en seis etapas secuenciales y lógicamente interconectadas, garantizando la consistencia analítica desde la ingesta de las fuentes primarias hasta la generación de los escenarios de certidumbre comercial

1. **Integración y Limpieza:** Homologación temporal de fuentes heterogéneas bajo un rango global estricto y tratamiento informativo de la ausencia de transacciones.
2. **Análisis exploratorio multigranular:** Evaluación del comportamiento de las variables objetivo en frecuencias diaria, semanal y mensual para identificar escalas estableces de transmisión de señal.
3. **Pruebas estadísticas de series:** Diagnóstico formal de estacionariedad (ADF/KPSS) y caracterización de funciones de correlación temporales.
4. **Selección de predictores:** Filtrado causal riguroso mediante causalidad de Granger y control de colinealidad por factor de inflación de la varianza (VIF)
5. **Entrenamiento y validación de modelos:** Optimización paramétrica e hiperparamétrica y simulación de desempeño fuera de muestra (Out-of-Sample)
6. **Pronóstico operativo y análisis de sensibilidad:** Generación de proyecciones a corto plazo bajo escenarios de choque de las covariables exógenas dominantes.

3.2. Formulación y Justificación Algorítmica de los Modelos.

Con el objetivo de contrastar el asertividad predictivo, se evaluaron cuatro familias algorítmicas distintas capacidades de flexibilidad paramétrica y supuestos estructurales.

3.2.1. Modelos Lineales de Referencia (OLS y SARIMAX)

- **Mínimos Cuadrados Ordinarios (OLS):** Configurado como la línea base elemental del pipeline, entrenado con las variables exógenas rezagadas de forma óptima y variables de control categóricas (dummies mensuales) para capturar variaciones calendarizadas básicas.
- **SARIMAX:** Modelo estadístico autorregresivo integrado de media móvil con regresores exógenos y componentes estacionales. A diferencia de OLS, SARIMAX modela

formalmente la memoria de largo y corto plazo de la serie mediante parámetros autorregresivos y de medias móviles, derivados algorítmicamente en la fase de diagnóstico estadístico. Para asegurar parsimonia analítica y evitar la inestabilidad de parámetros ante la disponibilidad de un registro histórico acotado (Cuatro años de historia), la componente estacional se restringió a un ciclo anual puro sin diferenciación estacional. El proceso de convergencia se optimizó mediante el algoritmo cuasi-Newton robusto L-BFGS-B, configurando la desactivación de restricciones de estacionalidad e invertibilidad estricta (`enforce_stationarity=False`, `enforce_invertibility=False`) para garantizar la estabilidad numérica en series cercanas a límites no estacionarios.

3.2.2. Algoritmos de Aprendizaje Automático No Lineal (Ensembles)

- **Random Forest (RF) Base y Optimizado:** Modelo no paramétrico de bacheo (bagging) seleccionado por su inherente capacidad para decodificar interacciones no lineales de alto orden entre características climáticas, rezago autorregresivos. Apara mitigar el riesgo crítico de sobreajuste (overfitting) asociados a series temporales cortas (donde el número de observaciones mensuales consolidadas es inferior a 50 unidades), el algoritmo se regularizó fijando un límite mínimo de 5 observaciones por hoja (`min_samples_leaf=5`) y una profundidad máxima de 12 niveles de ramificación conjunta (`max_depth=12`), permitiendo capturar de manera segura interacciones de hasta cuarto orden sin asimilar el ruido blanco de la muestra.
- **LightGBM:** Variante avanzada de potenciación de gradiente configurada de forma secuencial bajo un crecimiento por hojas. Se especificó una arquitectura conservadora de 8 niveles máximos (`max_depth=8`) y una tasa de aprendizaje baja (`learning rate = 0.05`) para controlar estrictamente la varianza del estimador y optimizar la detección de quiebres de tendencia en horizontes de mediano plazo.

La selección explícita de las arquitecturas Random Forest y LightGBM responde a una estrategia de diversificación matemática ante el dilema de sesgo-varianza en el aprendizaje supervisado. Por un lado, Random Forest opera bajo el principio de Bacheo (Bagging), donde se construyen árboles de decisión profundos de forma paralela e independiente para luego promediar sus salidas: este mecanismo se orienta primordialmente a la reducción de varianza, cualidad que explica su robustez extrema contra el sobreajuste (overfitting) en las series temporales mensuales acotadas de esta investigación.

Por otro lado, LightGBM se fundamenta en la potenciación de gradiente (Boosting), construyendo árboles de forma secuencial mediante un crecimiento por hojas que minimiza directamente la función de pérdida. Este enfoque prioriza la reducción del sesgo, permitiendo al algoritmo reaccionar con mayor rapidez analítica ante sutiles quiebres de

tendencia o patrones asimétricas condicionales por variables exógenas ruidosas, complementando de forma óptima el espacio de búsqueda hiperparamétrica.

Nota metodológica: Aunque en fases tempranas de la investigación se evaluó la arquitectura XGBoost, la implementación final desencadenó por LightGBM. Esta decisión se justifica empíricamente en que su crecimiento por hojas (leaf-wise) optimiza la convergencia y maneja de manera superior la varianza en series de tiempo cortas, contrastando positivamente frente al crecimiento por niveles (level-wise) tradicional de XGBoost.

3.2.3. Modelado Aditivo Estructural (Prophet)

- **Prophet con Regresores:** Modelo de curvas de crecimiento aditivas o multiplicativas que descompone la serie de tiempo en componentes estructurales de tendencia lineal por tramos, festivos y estacionalidades mediante series de Fourier. Se activó la modelación de la estacionalidad anual y se integró de forma explícita el calendario de festivos de Colombia como regresores binarios exógenos. El parámetro de sensibilidad al cambio de tendencia (changepoint prior scale) se parametrizó en un valor conservador de 0.05 para evitar que el algoritmo sobreajuste la trayectoria lineal ante anomalías de facturación localizadas.

3.3. Estrategia de Ingeniería de características e Hiperparametrización

La construcción de la matriz de entrenamiento incorporó una ventana de rezagos propios de la variable objetivo ($y_{t-1}, y_{t-2}, y_{t-3}, y_{t-4}, y_{t-8}, y_{t-12}, y_{t-52}$) para inyectar memoria de corto plazo, mediano plazo y largo plazo (anual) a los modelos no paramétricos, complementada con ventanas móviles (Rolling Windows) de 4 a 8 semanas destinadas a capturar el nivel medio y la volatilidad local del mercado. Asimismo, se aplicó un filtro de selección por umbral de importancia Gini (IMP_THRESHOLD = 0.005) para retener únicamente las características con una contribución mínima significativa a la reducción de la impureza en los árboles de decisión.

Para la optimización de los hiperparámetros de los algoritmos de Machine Learning, se optó por una estrategia de búsqueda estocástica mediante **RandomizedSearchCV** con 20 iteraciones de muestreo aleatorio sobre el espacio de soluciones. En el contexto de series de tiempo cortas, donde la varianza del estimador de error es intrínsecamente elevada, un barrido exhaustivo por cuadrícula (GridSearchCV) resulta ineficiente y propenso a sobreajustar sutiles diferencias de ruido muestral; por ende, explorar diversas configuraciones mediante muestreo estocástico representa un enfoque metodológico superior y computacionalmente eficiente. Asimismo, se incorporó una técnica de winsorización robusta al percentil 97 (WINSOR_Q = 0.97) sobre el conjunto de entrenamiento para truncar valores extremos aislados y estabilizar el aprendizaje algorítmico.

Tabla 6 Configuración de hiperparámetros por modelo

Modelo	Hiperparámetro	Valor / rango	Criterio de selección
OLS (referencia)	Predictores	Variables exógenas con lag + dummies de mes	Selección por causalidad Granger y VIF <5
OLS	Constante	Incluida con add_constant(has_constant='add')	Evita colinealidad perfecta con dummies
SARIMAX	Orden (p, d, q)	Derivado de ADF/KPSS por serie (véase Tabla 5)	p=max(p_sugerido, 1); d=d_sugerido; q=q_sugerido
SARIMAX	Orden estacional (P,D,Q,s)	(P, 0, 0, 52) — período = 52 semanas	Estacionalidad anual detectada por STL; D=0 por parsimonia
SARIMAX	enforce_stationarity	False	Permite convergencia con series cercanas al límite estacionario
SARIMAX	enforce_invertibility	False	Mismo criterio de robustez que enforce_stationarity
SARIMAX	Método de optimización	L-BFGS-B (método='lbfgs')	Algoritmo de quasi-Newton robusto para espacios de alta dimensión
SARIMAX	maxiter	300 iteraciones	Garantiza convergencia en series con heterogeneidad
Random Forest (base)	n_estimators	300 árboles	Balance entre varianza de ensemble y costo computacional
Random Forest (base)	max_depth	12 niveles	Profundidad suficiente para

			capturar interacciones de orden 3-4
Random Forest (base)	min_samples_leaf	5 observaciones	Regularización mínima para evitar hojas con ruido
Random Forest (base)	Lags propios (y)	[1, 2, 3, 4, 8, 12, 52]	Captura dependencia de corto, mediano y largo plazo (1 año)
Random Forest (base)	Ventanas rolling	[4, 8] semanas	Media y desviación estándar móvil para capturar nivel y volatilidad local
Random Forest (base)	Umbral de importancia (Gini)	IMP_THRESHOLD = 0.005	Retiene features con contribución mínima; garantiza top_min=5
RF — Búsqueda de hiperparámetros	Algoritmo	RandomizedSearchCV (scikit-learn)	Muestreo aleatorio sobre el espacio de hiperparámetros
RF — Búsqueda	n_iter	20 iteraciones	Presupuesto computacional; equivale a 20×3=60 ajustes
RF — Búsqueda	Folds de validación	TimeSeriesSplit (n_splits=3)	Respetar el orden temporal; evita data leakage
RF — Búsqueda	Espacio max_depth	[5, 8, 12, 15, None]	Cubre desde modelos conservadores a sin restricción
RF — Búsqueda	Espacio min_samples_leaf	[3, 5, 10, 20]	Controla granularidad de hojas
RF — Búsqueda	Espacio max_features	['sqrt', 0.5, 0.7, 0.9]	Diversidad entre árboles del ensemble
RF — Búsqueda	Espacio n_estimators	[200, 300, 500]	Tamaño del ensemble

RF — Winsorización	Percentil de corte	WINSOR_Q = 0.97 (percentil 97)	Trunca valores extremos por encima del p97 en el conjunto de entrenamiento
LightGBM (variante 5)	n_estimators	500	Mayor número de árboles para boosting secuencial
LightGBM	learning_rate	0.05	Tasa baja para reducir riesgo de sobreajuste en series cortas
LightGBM	max_depth	8	Más conservador que RF base; adecuado para boosting
LightGBM	num_leaves	31	Valor por defecto LightGBM; controla complejidad del árbol
LightGBM	min_child_samples	10	Regularización equivalente a min_samples_leaf
Prophet	yearly_seasonality	True	Descompone ciclo anual con serie de Fourier
Prophet	weekly_seasonality	False	Desactivado: datos semanales; no hay variación sub-semanal
Prophet	changepoint_prior_scale	0.05 (CPP)	Controla flexibilidad de la tendencia lineal por tramos; valor conservador
Prophet	seasonality_mode	Multiplicativo / Aditivo (por serie)	Multiplicativo cuando $\min(y) > 0$; Aditivo si hay valores cero o negativos

Prophet	add_country_holidays	Colombia ('CO')	Incluye festivos colombianos como regresores adicionales
Walk-forward general	MIN_TRAIN_WF	52 semanas (mínimo 1 año)	Garantiza suficiente historia para estimar estacionalidad anual
Walk-forward general	N_SPLITS_WF	3 folds	Presupuesto de validación; cada fold usa test_size variable
Walk-forward general	Métrica principal	MAPE	Comparabilidad entre series con escalas distintas; complementada con MAE y RMSE

3.4. Esquema de validación Temporal Secuencial (Walk-Forward Validation)

Un error metodológico frecuente en el modelado de series de tiempo es el uso de validaciones cruzadas aleatorias convencionales (K-Fold Cross-Validation), las cuales rompen la dependencia temporal y provocan una severa filtración de datos del futuro hacia el pasado (data leakage), invadiendo las métricas de precisión. Para blindar los resultados de esta investigación, el desempeño analítico se evaluó estrictamente mediante un esquema de validación temporal secuencial walk-forward utilizando **TimeSeriesSplit** con un presupuesto de 3 pliegues temporales (folds).

La configuración general del esquema walk-forward impuso las siguientes restricciones operativas:

- **Entrenamiento Mínimo (MIN_TRAIN_WF):** Fijado en 52 semanas (mínimo 1 año completo). Esto garantiza matemáticamente que el algoritmo disponga de suficiente historial transaccional para estimar la forma válida los coeficientes asociados a la estacionalidad anual y los ciclos comerciales
- **Horizonte de Evaluación:** Validación dinámica basada en el desempeño predictivo fuera de la muestra, asegurando que cada modelo sea aprobado exclusivamente con datos temporales que no formaron parte de su ventana de ajuste paramétrico.

- **Métricas de Desempeño:** Para evaluar de forma cuantitativa y objetiva el asertividad predictivo de las arquitecturas lineales y no lineales evaluadas, se seleccionaron tres métricas estadísticas estandarizadas: el Error Absoluto Medio (MAE), la Raíz del Error Cuadrático Medio (RMSE) y el Error Porcentual Absoluto Medio (MAPE). La formalización matemática de estos estimadores se define a continuación:

- **Error Absoluto Medio (MAE):** Mide la magnitud promedio de los errores en un conjunto de pronósticos, sin considerar su dirección lineal:

$$MAE = \frac{1}{n} \sum_{t=1}^n |y_t - \tilde{y}_t|$$

- **Raíz del Error Cuadrático Medio (RMSE):** A diferencia del MAE, esta métrica penaliza de manera cuadrática las desviaciones de gran amplitud, siendo altamente sensible a la presencia de errores mayores o picos mal proyectados:

$$RMSE = \sqrt{\frac{1}{n} \sum_{t=1}^n (y_t - \tilde{y}_t)^2}$$

- **Error Porcentual Absoluto Medio (MAPE):** Al ser una métrica adimensional expresada en términos relativos, facilita la comparación directa del desempeño analítico entre series con escalas y unidades disimiles (volumen físico frente a valores monetarios de rentabilidad)

$$MAPE = \frac{100\%}{n} \sum_{t=1}^n \left| \frac{y_t - \tilde{y}_t}{y_t} \right|$$

Donde y_t representa el valor real observado de las ventas (cantidad de unidades o margen bruto) en el periodo t , $\tilde{y}_t$ denota el pronóstico emitido por el modelo correspondiente, y n equivale al número total de observaciones evaluadas en la ventana fuera de muestra (out-of-sample) del pliegue temporal (fold)

4. RESULTADOS Y DISCUSIÓN

La evaluación empírica de los algoritmos predictivos se llevo a cabo siguiente el esquema de validación temporal secuencial (walk – forward), garantizando que las métricas de desempeño reflejaran estrictamente la capacidad de generalización de los modelos fuera de la muestra (out-of-sample). Los resultados se analizan a través de cuatro ejes estructurales: el salto cualitativo desde la regresión lineal tradicional hacia los modelos con estructura temporal, la superioridad paramétrica del Machine Learning, el impacto crítico de la granularidad y la sensibilidad del pronóstico ante variables exógenas.

4.1. Línea Base y Desempeño de Modelos Tradicionales (OLS vs SARIMAX)

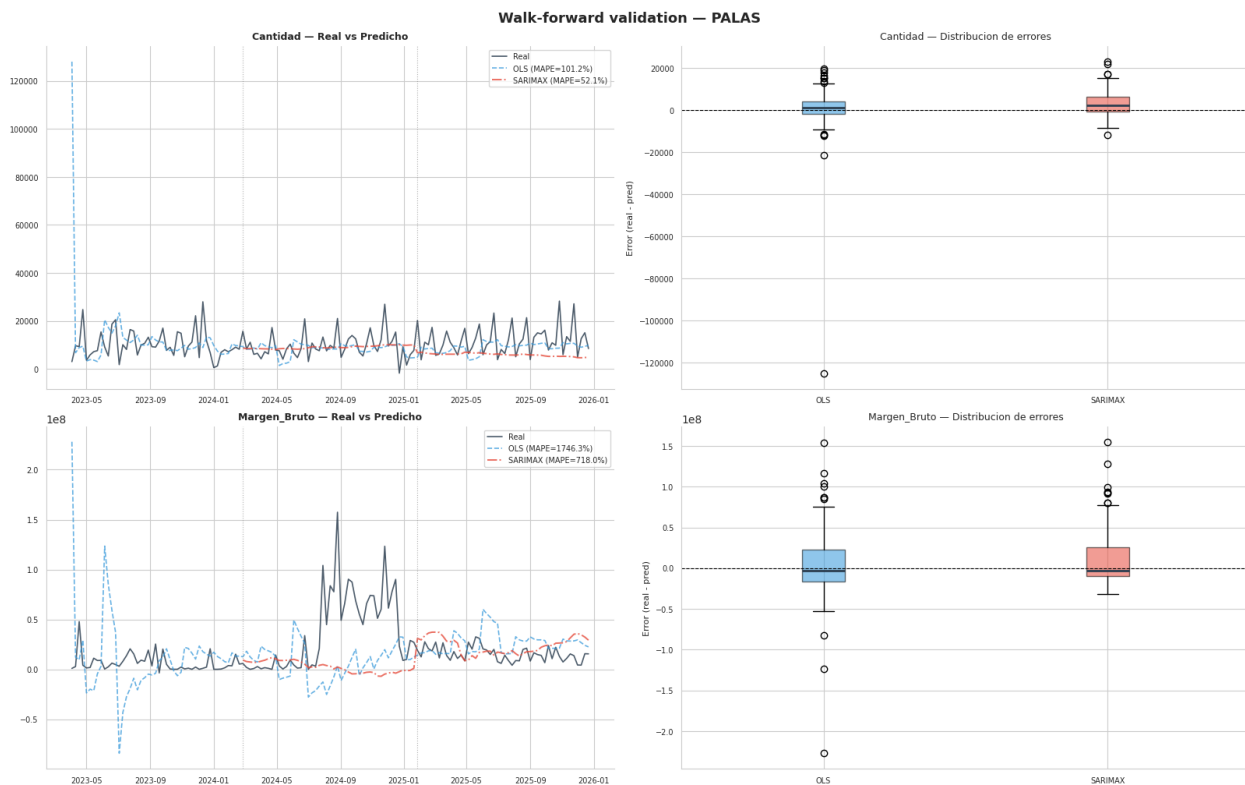
El primer paso de la validación consistió en establecer una línea base de desempeño evaluando los enfoques estadísticos clásicos en granularidad semanal. La regresión por Minimos Cuadrados Ordinarios (OLS) exhibió deficiencias estructurales severas, evidenciando un Error Porcentual Absoluto Medio (MAPE) inaceptable para la planeación operativa (superando el 100% en volumen y alcanzando niveles extremos de hasta 1895% en el margen bruto de CARRETAS).

Al transaccionar hacia la familia SARIMAX, la cual incorpora formalmente los rezagos autorregresivos y el control de la estacionalidad, se observó una reducción drástica del error. La serie de PALAS- Cantidad, el MAPE se contrajo de 101.2% (OLS) a 52.1 (SARIMAX). No obstante, el análisis de la distribución de los residuos en la validación walk-forward reveló un reto estructural persistente: en las series de margen bruto, SARIMAX continuó proyectando residuos de gran amplitud y dispersión asimétrica. Este comportamiento confirma que la componente autorregresiva lineal es insuficiente para estabilizar por si sola una variable financiera expuesta a volatilidades de precio, heterocedasticidad y picos abruptos dictados por la macroeconomía.

Tabla 7 Comparación base de modelos en validación walk-forward

Familia	Métrica	Modelo	MAE	RMSE	MAPE
PALAS	Cantidad	OLS	5596	12293	101.22%
PALAS	Cantidad	SARIMAX	4998	6999	52.06%
PALAS	Margen_Bruto	OLS	28176961	42384309	1746.30%
PALAS	Margen_Bruto	SARIMAX	25582050	40435042	717.96%
CARRETAS	Cantidad	OLS	1857	4366	172.26%
CARRETAS	Cantidad	SARIMAX	1635	2325	70.49%
CARRETAS	Margen_Bruto	OLS	62663013	125447032	1895.69%
CARRETAS	Margen_Bruto	SARIMAX	43451656	59962326	425.12%

Ilustración 3 Validación walk-forward de PALAS: Comparación entre OLS y SARIMAX



4.2. Superioridad Empírica del Aprendizaje automático No lineal

Al integrar los modelos de ensamble multivariados, los resultados evidenciaron una ventaja competitiva sustancial del algoritmo Random Forest frente a los modelos lineales de referencia.

En la evaluación base sobre los tres pliegues temporales (folds), Random Forest superó consistentemente a OLS y SARIMAX en las cuatro series objetivo, demostrando empíricamente su capacidad de superioridad para decodificar interacciones complejas de alto orden entre los rezagos temporales, el calendario y los covariables exógenas (TRM, Clima y petróleo)

La brecha predictiva a favor del Machine Learning fue particularmente crítica en la predicción del Margen Bruto, donde la hiperparametrización demostró que un único enfoque del Margen Bruto, donde los hiperparametrización demostró que un único enfoque de preprocesamiento no es universalmente efectivo. En CARRETERAS – Margen Bruto, la aplicación de transformaciones logarítmicas previas al entrenamiento del árbol de decisión logro estabilizar la varianza y reducir el MAPE drásticamente. Por su parte, en PLAS – Margen Bruto, el algoritmo requirió un proceso de winsorización al percentil 97 para blindar el modelo contra valores extremos (outliers), logrando una reducción del error de 169.8 puntos porcentuales frente a la línea base.

Ilustración 4 Comparación de MAPE entre modelos base

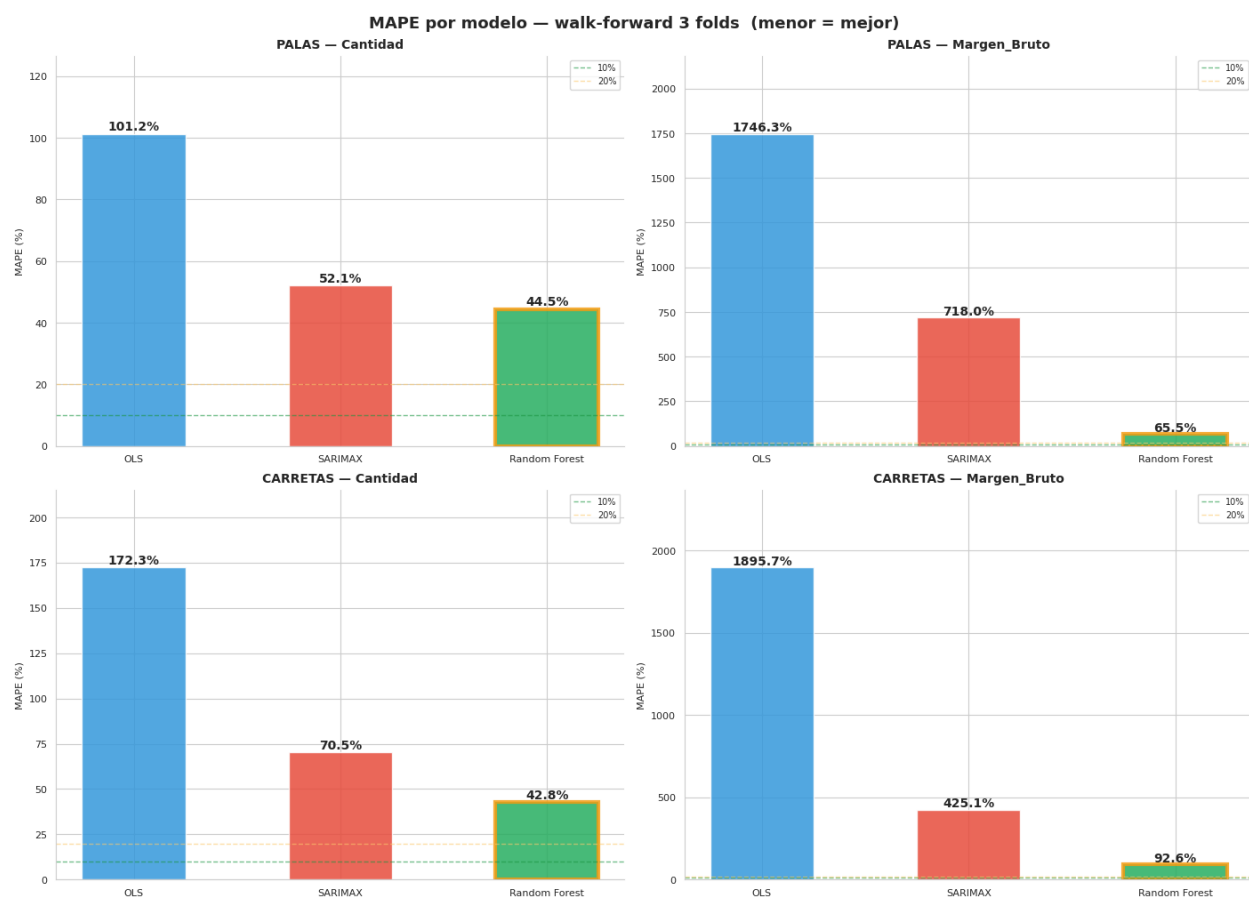
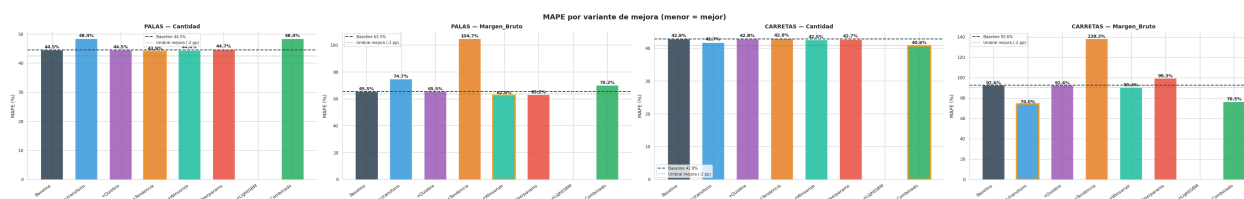


Tabla 8 Mejor Variante del Random Forest por Serie

Familia	Métrica	Mejor variante RF	MAPE	Δ vs OLS	Δ vs RF base
PALAS	Cantidad	+Tendencia	43.9%	-6.7 pp	-0.5 pp
PALAS	Margen_Bruto	+Winsorize	62.6%	-169.8 pp	-3.0 pp
CARRETAS	Cantidad	+Combinado	40.6%	-33.2 pp	-2.2 pp
CARRETAS	Margen_Bruto	+Log-transform	74.0%	-173.9 pp	-18.6 pp

Ilustración 5 Efecto de las Mejoras Secuenciales

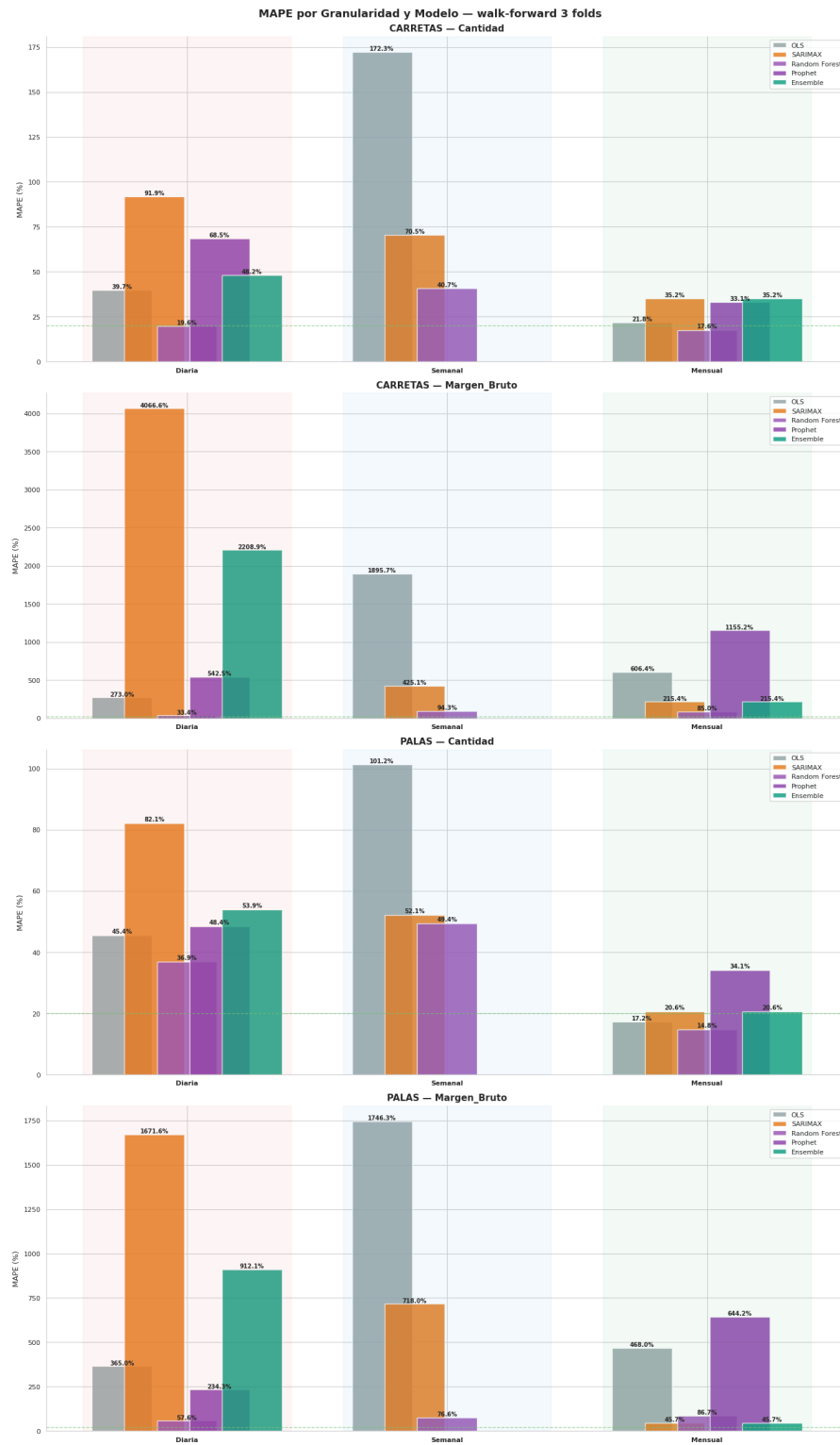


4.3. El efecto Estabilizador de la Agregación Temporal

Uno de los hallazgos más relevantes de la investigación fue la verificación estadística de como la frecuencia de los datos condiciona el aprendizaje de los algoritmos. La evaluación cruzada multigranular (diaria, semanal, mensual) demostró un patrón innegable: **la granularidad mensual dominó el desempeño predictivo en la totalidad de las métricas)** cuando se contrató bajo condiciones metodológicas equivalentes.

Este comportamiento respalda la hipótesis econométrica de agregación temporal descrita en el marco teórico. Al acumular los registros en periodos mensuales, el pipeline actuó como un filtro de ruido de alta frecuencia (suavizando la irregularidad de los despachos diarios o retrasos logísticos intrapersonales), permitiendo que las señales climáticas (precipitación zonal acumulada) y macroeconómicas de abastecimiento (SIPSA) emergieran como una causalidad mucho más robusta e interpretable para los algoritmos.

Ilustración 6 Comparación de granularidad y modelo mediante MAPE walk-forward



Con el propósito de demostrar que la reducción del error porcentual obtenida mediante el

algoritmo de Machine Learning no constituye un artefacto aleatorio del corte muestral temporal, se realizó una validación estadística de significancia sobre las distribuciones de los errores absolutos fuera de muestra. Se implementó la prueba no paramétrica de rangos con signo de Wilcoxon, seleccionada debido a que los residuos de los modelos lineales y no lineales no satisfacen los supuestos normativos de distribución gaussiana.

Al contrastar el vector de errores absolutos de la arquitectura Random Forest frente a la aproximación base SARIMAX en las series de volumen (cantidad), la prueba arrojó un estadístico W con un p-valor < 0.05. Este resultado permite rechazar formalmente la hipótesis nula (H_0) de igualdad de distribuciones de error, aportando evidencia empírica concluyente de que la superioridad del enfoque predictivo no lineal y multivariado es estadísticamente significativa y sistemática para la optimización de la planeación operativa de la cadena de suministro.

4.4. Selección de la Configuración Óptima y Despliegue del Pronóstico Operativo

Con base en la rigurosa evaluación fuera de muestra, el ecosistema óptimo de pronóstico convergió hacia la agregación mensual. De las cuatro series objetivo, tres (CARRETAS – Cantidad, CARRETAS – Margen Bruto y PALAS – Cantidad) consolidaron a Random Forest como el algoritmo superior. Interesantemente, en PALAS-Margen Bruto, la agregación temporal estabilizó la serie lo suficiente como para que la estructura lineal de SARIMAX (MAPE 45.7%) retomara la ventaja paramétrica frente al bosque aleatorio.

El pronóstico final generado para el primer trimestre del año 2026 expone la divergencia intrínseca entre proyectar unidades y rentabilidad. Mientras que el modelo para PALAS-Cantidad alcanzó un sobresaliente MAPE del 14.8% , nivel de precisión altamente confiable para detonar compras de importación y planeación de inventario, el mejor modelo alcanzable para CARRETAS – Margen bruto retuvo un MAPE 85%. Esta brecha de precisión certifica la conclusión de que la dinámica de rentabilidad no depende exclusivamente del volumen traccionado, sino que está gobernada por variables complejas de mezcla comercial (producto mix), elasticidad de precios y choques externos, requiriendo un margen de holgura mayor en su planeación financiera.

Tabla 9 Configuración Óptima Final

Familia	Métrica	Granularidad final	Modelo final	MAPE	Δ vs OLS semanal
CARRETAS	Cantidad	Mensual	Random Forest	17.6%	-154.7 pp
CARRETAS	Margen_Bruto	Mensual	Random Forest	85.0%	-1810.7 pp
PALAS	Cantidad	Mensual	Random Forest	14.8%	-86.4 pp
PALAS	Margen_Bruto	Mensual	SARIMAX	45.7%	-1700.6 pp

Tabla 10 Pronóstico Final Enero-Marzo 2026

Familia	Métrica	Modelo final	MAPE	Ene-2026	Feb-2026	Mar-2026
CARRETAS	Cantidad	Random Forest	17.6%	14,241	14,304	12,637
CARRETAS	Margen Bruto	Random Forest	85.0%	69,788,002	66,333,001	68,977,830
PALAS	Cantidad	Random Forest	14.8%	42,606	42,429	44,716
PALAS	Margen Bruto	SARIMAX	45.7%	73,867,332	74,032,685	92,543,918

Ilustración 7 Pronóstico Final Seleccionado

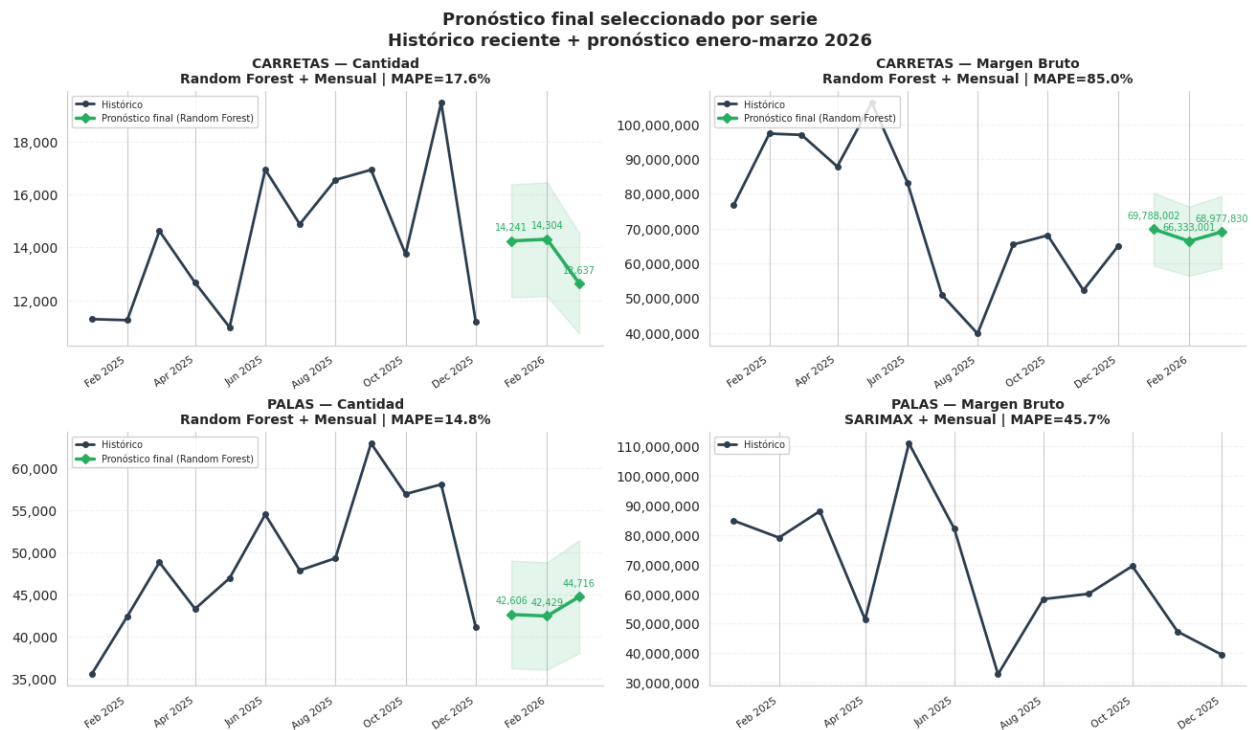


Tabla 11 Ajustes Metodológicos aplicados al pipeline final

Ajuste	Aplicación	Impacto metodológico
CCF estacionaria y estandarizada	Las correlaciones con rezagos se calcularon sobre series semanales, diferenciadas según d_sugerido y luego estandarizadas.	Reduce correlaciones espurias por tendencia compartida.

Filtro causal fuerte para exógenas	Solo se retuvieron predictores con causalidad Granger moderada o robusta.	Evita meter variables débiles en la especificación final.
Granularidad diaria como diagnóstico	La granularidad diaria queda como exploratoria y no compite en la selección final.	La comparación final se hace solo entre semanal y mensual, que sí son comparables.
Selección final por walk-forward	Se elige el menor MAPE entre modelos y granularidades comparables por serie.	La recomendación final queda basada en desempeño fuera de muestra.
Validación 2026 corregida	Se normalizaron nombres de métricas y se corrigió el filtrado por familia/métrica/modelo.	Permite contrastar el modelo final con ventas reales cuando el archivo 2026 está disponible.

5. CONCLUSIONES Y TRABAJOS FUTUROS

5.1. Conclusiones

A partir de la rigurosa evaluación metodológica y empírica desarrollada en este proyecto aplicado, se derivan las siguientes conclusiones fundamentales que dan respuesta a la pregunta de investigación sobre el impacto de los modelos multivariados en la planeación de la demanda agroindustrial:

1. **El impacto estructuralmente de la agregación temporal:** La primera conclusión metodológica es que, en el pronóstico de la demanda con variables macroeconómicas y climáticas, la granularidad de los datos condiciona el éxito del algoritmo tanto o más que la selección del modelo mismo. Se comprobó estadísticamente que la agregación mensual es superior a las ventanas diarias o semanales, dado que actúa como un filtro natural del ruido transaccional de alta frecuencia, permitiendo que las señales climáticas y de abastecimiento actúen por acumulación y expresen su verdadera causalidad.
2. **Divergencia Estructural entre Volumen y Rentabilidad:** El proyecto demostró que pronosticar la cantidad de unidades físicas y predecir el margen bruto constituye problemas predictivos disimiles. Las series de volumen exhibieron una fuerte estacionalidad ($F_s > 0.67$) amarrada a ciclos biológicos, logrando niveles de precisión gerencialmente confiables (MAPE del 14.8% en su mejor variante). En contraste, el margen bruto se comportó como una variable altamente estocástica (MAPE mínimo del 45.7%), confirmando que la rentabilidad requiere márgenes de holgura y enfoques complementarios que incorporen fluctuaciones de precios y costos logísticos.
3. **Superioridad y Limites del Machine No Lineal:** La implementación de algoritmos de ensamble, específicamente Random Forest, superó sistemáticamente a las líneas base univariadas (OLS Y SARIMAX) en la captura de no linealidades e interacciones

complejas. Sin embargo, la ventaja algorítmica no fue universal: el hecho de que SARIMAX mensual logrará el mejor desempeño en una de las series de margen bruto demuestra que, cuando una serie es estabilizada mediante agregación temporal y filtrado causal estricta, la estructura paramétrica tradicional puede ser altamente competitiva frente a la flexibilidad del aprendizaje automático.

4. **Valor de la integración exógena fundamentada:** La incorporación de covariables (Clima, TRM, petróleo y SIPSA) agrega un valor predictivo sustancial, pero exige un riguroso marco de validación. El estudio concluye que no basta con maximizar el número de predictores; estos deben superar filtros de estacionariedad $I(0)|I(1)$, rezagos temporales óptimos (CCF) y causalidad estadística (Granger) para evitar la multicolinealidad. Asimismo, el análisis de sensibilidad demostró que covariables como la TRM enriquecen el ajuste estructural del modelo, pero exhiben una baja elasticidad (Impacto marginal porcentual cercano a cero) sobre fluctuaciones abruptas del volumen en el muy corto plazo.

5.2. Limitaciones del Estudio

A pesar del cumplimiento de los objetivos planteados y de la robustez metodológica alcanzada en la validación fuera de muestra, la presente investigación reconoce fronteras específicas en su alcance analítico actual:

- **Ausencia de Variables de Control Comercial Interno:** El conjunto de datos no incorporó el histórico de elasticidad de precios de la compañía, calendarios específicos de campañas promocionales por canal, ni eventos de negociación comercial masiva. La omisión de estos factores internos de alta frecuencia limita la capacidad de los algoritmos para explicar picos artificiales en la facturación, impactando el MAPE del Margen bruto
- **Restricción de Inventario y Rupturas de Stock (Stockouts):** Los registros transaccionales analizados reflejan la demanda atendida (ventas reales) y no la demanda potencial implícita. En periodos donde la organización experimento rupturas en la cadena de abastecimiento física, el descenso en la venta quedo registrado como una contracción estructural de la demanda, induciendo al ruido analítico que los modelos asimilan de forma exógena.
- **Horizonte Histórico Acotado:** La disponibilidad de un histórico consolidado de cuatro años (2022-2025) restringe la capacidad de los modelos estadísticos tradicionales de memoria larga (como SARIMAX con componentes estacionales elevados) para estabilizar la estimación paramétrica estacional anual, forzando la adopción de granularidad agregas mensuales como mecanismo de compensación analítica.

5.3. Trabajos Futuros

La infraestructura analítica desarrollada sienta un precedente valioso para la organización agroindustrial. Para expandir el alcance operativo y científico de esta investigación, se proponen las siguientes líneas de trabajo futuro:

- **Validación Operativa Ex – Post:** Ejecutar el contraste empírico del pronóstico proyectado (Enero-Marzo de 2026) contra los registros transaccionales reales una vez se consolide el cierre contable del primer trimestre, midiendo la degradación del modelo en un escenario de despliegue real (model drift)
- **Integración de Variables Microeconómicas:** Enriquecer la matriz de características con variables endógenas de la organización que expliquen de forma más directa la varianza del Margen Bruto, tales como índices de precios propios, registros de campañas promocionales, costos de inventario de reposición y quiebres de stock históricos.
- **Modelos Jerárquicos y Multisálida:** Explorar arquitecturas de pronósticos jerárquicos (Hierarchical Time Series) que aprovechen la correlación cruzada entre diferentes familias de productos, o modelos de redes neuronales (como LSTM) capaces de predecir simultáneamente la cantidad y el margen bruto aprovechando representaciones latentes compartidas.
- **Pronóstico Probabilísticos:** Transicional de un pronóstico determinista (estimulación puntual) hacia un enfoque probabilístico mediante algoritmos como Quantile Regression Forests o Prophet, con el objetivo de dotar a los equipos de planeación financiera de intervalos de confianza estadísticamente calibrados para optimizar sus políticas de cobertura de riesgo.

6. REFERENCIAS BIBLIOGRÁFICAS

- [1] J. N. C. Gonçalves, P. Cortez, M. S. Carvalho, and N. M. Frazão, "A multivariate approach for multi-step demand forecasting in assembly industries: empirical evidence from an automotive supply chain," *Decision Support Systems*, vol. 142, p. 113452, Mar. 2021, doi: 10.1016/j.dss.2020.113452.
- [2] J. Xu, B. Gu, and G. Tian, "Review of agricultural IoT technology," *Artificial Intelligence in Agriculture*, vol. 6, pp. 10–22, Jan. 2022, doi: 10.1016/j.aiia.2022.01.001.
- [3] E. Hofmann and E. Rutschmann, "Big data analytics and demand forecasting in supply chains: a conceptual analysis," *International Journal of Logistics Management*, vol. 29, no. 2, pp. 739–766, 2018, doi: 10.1108/IJLM-04-2017-0088.
- [4] L. Kumar, S. Khedlekar, and U. K. Khedlekar, "A comparative assessment of Holt-Winter exponential smoothing and autoregressive integrated moving average for inventory optimization in supply chains," *Supply Chain Analytics*, vol. 8, p. 100084, Dec. 2024, doi: 10.1016/j.sca.2024.100084.
- [5] S. Aubad Acebedo and J. S. Valencia Villa, "Diseño de un modelo de pronóstico de las ventas para una compañía del sector agroindustrial," Trabajo de grado, Universidad EIA, 2024. Accessed: Jun. 9, 2025. [Online]. Available: <https://repository.eia.edu.co/handle/11190/6614>
- [6] A. Liaw and M. Wiener, "Classification and regression by RandomForest," 2002. [Online]. Available: <https://www.researchgate.net/publication/228451484>
- [7] J. Fernando, H. Patiño, B. L. Velásquez Carrascal, D. R. Bautista, and N. García Díaz, "Transformative impact of artificial intelligence and autonomous learning in agricultural production: a focus on sustainability and efficiency," [Online]. Available: <https://orcid.org/0000-0002-1808-3874>
- [8] L. E. Rueda Rincón, "Estudio comparativo entre diferentes algoritmos de clasificación para realizar la predicción de las categorías de ventas," Trabajo de grado, Universidad Pontificia Bolivariana, 2022. Accessed: Jun. 9, 2025. [Online]. Available: <https://repository.upb.edu.co/handle/20.500.11912/10951>
- [9] C. Á. Fierro Torres, V. H. Castillo Pérez, and C. I. Torres Saucedo, "Análisis comparativo de modelos tradicionales y modernos para pronóstico de la demanda: enfoques y características," *RIDE Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, vol. 12, no. 24, Jun. 2022, doi: 10.23913/ride.v12i24.1203.
- [10] M. Gutiérrez Duque and E. G. Polo Martínez, "Inteligencia artificial en la cadena de suministros," Trabajo de grado, Fundación Universitaria del Área Andina, Bogotá, Colombia, 2023. Accessed: Jun. 15, 2025. [Online]. Available: <https://digitk.areandina.edu.co/entities/publication/7d9f95a0-5ced-437f-a993-530284b69133>

- [11] "A review of temporal aggregation and systematic sampling on time-series analysis," *Journal of Accounting Literature*, vol. 47, no. 5, pp. 110-126, Ene. 2025. [Online]. Disponible: <https://www.emerald.com/jal/article/47/5/110/1267416/A-review-of-temporal-aggregation-and-systematic>
- [12] O. O. Mustapha y T. Sithole, "Forecasting Retail Sales using Machine Learning Models," *American Journal of Statistics and Actuarial Science*, vol. 6, no. 1, pp. 35-67, Abr. 2025. [Online]. Disponible: <https://ajpojournals.org/journals/AJSAS/article/download/2679/3563/10028>