Pronóstico de precios del aguacate Hass y Papelillo en Colombia mediante métodos de series temporales

Yenny Xiomara Morales Rodríguez[1]
Karol Ximena Melo Bayona[2]
Orlando Stiven Jaramillo Piza[3]
Mario José Pacheco López[4]

Resumen

En Colombia, la producción agropecuaria representa una parte importante del Producto Interno Bruto y muchas familias encuentran el sustento en las diversas actividades que la componen. Sin embargo, los precios de algunos productos suelen ser volátiles, como en productos perecederos tales como el aguacate, generando incertidumbre alrededor de cuándo es el mejor momento para vender y sobre la rentabilidad de la producción. En este trabajo, se aborda el caso específico del aguacate papelillo y aguacate Hass, con el objetivo de comparar diferentes técnicas de modelamiento para la predicción de los precios de estos productos, incluyendo técnicas de modelamiento estadístico o, en particular de *Machine Learning*, usando datos recolectados por el DANE sobre los precios nacionales semanales disponibles del aguacate Hass y el aguacate papelillo entre el 10 de noviembre de 2012 y el 27 de diciembre de 2024. Entre los principales resultados se observó que los modelos de *Machine Learning*, particularmente los modelos basados en árboles de decisión son más

[1] Universidad El Bosque; Maestría en Estadística Aplicada y Ciencia de Datos; yxmorales@unbosque.edu.co.

[2] Universidad El Bosque; Maestría en Estadística Aplicada y Ciencia de Datos; kmelob@unbosque.edu.co.

[3] Universidad El Bosque; Maestría en Estadística Aplicada y Ciencia de Datos; ojaramillop@unbosque.edu.co.

[4] Universidad El Bosque; Docente Maestría en Estadística Aplicada y Ciencia de Datos; mpachecol@unbosuqe.edu.co.

eficientes en estas predicciones, superando el rendimiento de los modelos estadísticos. En el caso del aguacate papelillo, el modelo Decision Tree, registró los menores errores, con un RMSE de 1 245.95, MAE de 1 221.45, y un MAPE de 14.09%. De forma similar, para el aguacate Hass este mismo modelo se destacó, obteniendo las menores métricas de error, con un RMSE de 942.43, MAE de 829.07 y MAPE de 15.50%.

1. Introducción

En Colombia, el mercado de los productos agrícolas se encuentra concentrado en pequeños productores, y representa un rol estratégico en la economía y bienestar social, no solo por su participación en el Producto Interno Bruto, sino por su incidencia en la mano de obra rural y la seguridad alimentaria. Uno de los cultivos más importantes en este contexto es el aguacate, en particular las variedades Hass y Papelillo. El aguacate Hass destaca por su creciente demanda interna como externa, lo que lo convierte en fruta clave para el desarrollo del sector, mientras que el aguacate Papelillo por ser parte de la dieta de los colombianos es importante de la seguridad alimentaria del país.

Los precios de productos agrícolas, incluido el del aguacate, presentan una alta volatilidad por factores como estacionalidad, condiciones climáticas, costos de la cadena de suministro, así como la relación de oferta y demanda (Zhao et al., 2025; De la Torre-Torres et al., 2024). Lo anterior, repercutiendo directamente en la calidad de vida de los habitantes del país y su estabilidad social. Como se mencionó, a diferencia de otros bienes, los precios agrícolas están influenciados por factores de mayor complejidad, difíciles de cuantificar, y muestran comportamientos no estacionarios y no lineales. Dado que el sector agropecuario constituye una actividad estratégica para la producción nacional y el sustento de numerosos hogares, contar con pronósticos confiables de precios agrícolas resulta esencial para que los actores públicos y privados puedan anticipar

cambios en las condiciones del mercado, promoviendo la estabilidad de los bienes, así como el desarrollo rural y económico.

El pronóstico de precios agrícolas ha sido abordado anteriormente por diversos autores. Ramírez Castañeda et al. (2021) analizan las fluctuaciones en los precios de los cultivos de maracuyá y granadilla, hallando que el modelo *ARIMA* es el que mejor se adapta a los datos analizados, lo que sugiere su eficacia en la predicción de precios en este contexto agrícola. Ruiz Hernández et al. (2019) usaron metodología Box-Jenkins para el pronóstico del precio promedio mensual de la manzana en México, donde encontraron que el modelo más adecuado para obtener información futura sobre este producto fue un SARIMA$(2,1,0)(1,1,0)_{12}$. En trabajos como los de Sal (2024), se usaron técnicas de *machine learning*, y de redes neuronales para predecir el precio del azúcar en Argentina. Dentro de esta investigación se encontró que el modelo *XGBoost* fue el que obtuvo un mejor rendimiento. En otras investigaciones, como la realizada por Gómez Arango & Diaz Valencia (2024), se destaca la efectividad de los modelos de *Boosting* como *PCA Random Forest* y *PCA XGB Boosting*, en la predicción de los cultivos objeto de estudio.

Sin embargo, algunas metodologías de pronósticos de precios no siempre capturan la complejidad de estas fluctuaciones. En este contexto, surge el cuestionamiento: ¿Cuál de los enfoques, entre los métodos tradicionales, tales como ARIMA, SARIMA y regresiones penalizadas, y los métodos de *machine learning*, como árboles y SVM, se ajusta mejor a los pronósticos de los precios del aguacate Hass y aguacate Papelillo?

La obtención de un pronóstico preciso de los precios del aguacate es fundamental para apoyar la toma de decisiones relacionadas con la viabilidad y sostenibilidad de sus cultivos en diferentes regiones del país. Debido a la alta demanda y la importancia

económica del aguacate Hass y del aguacate Papelillo, estos pronósticos permitirán evaluar calificaciones de riesgo y optimizar los recursos destinados al establecimiento y sostenimiento de áreas productivas, fomentando el crecimiento del sector agrícola.

Desde una perspectiva teórica, disponer de predicciones acertadas de los precios del aguacate permitirá que productores, entidades financieras y responsables de políticas públicas se anticipen a las fluctuaciones del mercado. Esto facilitará la toma de decisiones informadas y ayudará a mitigar el impacto de la volatilidad, contribuyendo así a la seguridad alimentaria y a la optimización de los recursos disponibles, generando competitividad y estabilidad.

Teniendo en cuenta lo anterior, esta investigación tiene por objetivo evaluar métodos de series de tiempo para la proyección semanal de los precios nacionales del aguacate Hass y el aguacate Papelillo, a través de los criterios de evaluación *RMSE*, *MAPE* y *MAE*. Para el cumplimiento de este, se fijaron dos objetivos específicos: primero, identificar los modelos y sus hiperparámetros que proporcionen las predicciones más precisas para la proyección de precios semanales del aguacate, empleando métodos de la estadística clásica y modelos de *machine learning*; y segundo, comparar el rendimiento de los métodos tradicionales de series temporales con las técnicas de *machine learning* en la predicción semanal de precios nacionales del aguacate, utilizando los criterios de evaluación RMSE, MAPE y MAE.

Este artículo está formado por seis secciones: en primer lugar, la presente introducción; seguido del marco de referencia, en el que se exponen investigaciones previas relacionadas; en tercer lugar, la metodología usada para lograr los objetivos y el flujo de trabajo implementado; posteriormente, los resultados encontrados y, por último,

las conclusiones y análisis de resultados del trabajo de investigación, así como recomendaciones para trabajos futuros.

2. Marco de referencia

2.1 Características del Aguacate y factores que afectan el mercado agrícola

El aguacate es una fruta que muchos alrededor del mundo conocen, y Colombia es un productor con alto potencial. Existen diferentes variedades de este fruto, sin embargo, las que más se comercializan y consumen en el país son las variedades Hass y Papelillo. La variedad Hass alcanzó un valor de exportaciones de USD 309.4 millones con un volumen de ventas externas 138.315 toneladas netas (Analdex, 2025), mientras que, según la información reportada por la Unidad de Planificación Rural Agropecuaria (UPRA), para el 2023 la participación en la siembra de aguacate Hass y Papelillo fue de 52 % y 6 % respectivamente. Cada una de ellas tiene propiedades y características propias, en cuanto a su aspecto físico y cualidades nutricionales. Se sabe que Colombia es un consumidor de aguacate Papelillo por su reconocimiento en la cultura gastronómica y el conocimiento de los hogares sobre cómo elegirlos y consumirlos; el aguacate Hass es una variedad que ha venido tomando importancia dentro de los mercados nacionales y es la variedad de exportación que ya cuenta con su propia partida arancelaria, por lo que el precio de cualquiera de estas variedades tiene impacto sobre los colombianos.

Al revisar la evolución de los precios, la volatilidad de los productos agrícolas suele ser impredecible, ya que se ven afectados por eventualidades como fluctuaciones del precio del petróleo, los efectos del cambio climático, la escasez de agua, plagas, desastres naturales, etc. Esta incertidumbre en los precios puede comprometer el acceso de los consumidores a los alimentos, conduciendo a la seguridad alimentaria, hambruna y malnutrición (Chen et al., 2021; Rana et al., 2024; Özden, 2023).

2.2 Fundamentos de los métodos de pronóstico de series de tiempo

Un proceso estocástico es un conjunto de variables aleatorias organizadas cronológicamente. Estos procesos ayudan a describir fenómenos que cambian con el tiempo, siguiendo reglas de probabilidad. Así, una serie temporal se puede ver como un ejemplo concreto de un proceso estocástico, mostrando cómo evoluciona un fenómeno que presenta observaciones que son dependientes entre sí (Gujarati & Down, 2009, p. 740).

2.2.1 Características de las series temporales

Cuando se hace referencia a un proceso estocástico estacionario, significa que sus características estadísticas, como la media y la varianza, no cambian con el tiempo. Cuando una serie temporal incumple con al menos una de las características antes mencionadas, se dice que es una serie no estacionaria. Esta estabilidad en las características estadísticas de los procesos estacionarios facilita su estudio y permite hacer predicciones más precisas (Gujarati & Down, 2009, pp. 740 - 741; Box et al., 2016, p. 24).

Las series temporales pueden tener cuatro componentes (Gujarati & Down, 2009, p. 290): el estacional, que se refiere a la aparición regular de ciertos cambios en una variable a lo largo de intervalos específicos, que generalmente no superan un año; cíclico, del que se habla en contextos de tiempo más amplios, sin embargo, las variaciones cíclicas son menos comunes en comparación con las estacionales; el componente de tendencia refleja los movimientos a largo plazo o de baja frecuencia en una serie temporal, con una duración que supera los ocho años; y el aleatorio, que son aquellas fluctuaciones que no provienen de ninguno de los anteriores componentes, también conocidos como ruido.

La selección de la frecuencia de muestreo y la longitud de la serie temporal en el análisis de series temporales permiten optimizar la capacidad predictiva de los modelos. La frecuencia de muestreo determina el intervalo de tiempo entre cada observación, lo cual impacta directamente en la granularidad de los datos. Por otro lado, la longitud de la serie temporal se refiere a cuántos puntos de datos históricos se utilizan para el análisis y modelado. La longitud óptima depende de la disponibilidad de los datos, la naturaleza de la serie (si es estacionaria o no), y el horizonte de predicción deseado (Yu et al., 2024).

2.2.2 Preprocesamiento: descomposición, diferenciación y suavizado

La descomposición y el suavizado son técnicas fundamentales en el pronóstico de series temporales, especialmente para analizar precios de productos agropecuarios. La descomposición es clave para entender patrones complejos de los datos, dividiéndolos en componentes más simples, como la tendencia, estacionalidad y ciclos. Algunos de estos métodos son la descomposición estacional y la descomposición por *wavelets*.

El suavizado se utiliza para reducir el ruido y las fluctuaciones de los datos, facilitando la identificación de patrones subyacentes, como, por ejemplo, el suavizado exponencial. Estas técnicas mejoran el rendimiento de los modelos de pronóstico de precios agropecuarios, permitiendo una predicción más precisa de los patrones complejos en los datos (Sun et al., 2023).

2.3. Modelado de series temporales

2.3.1 Métodos tradicionales

Entre los métodos comunes se encuentran: (a) modelos autorregresivos (*AR*), que asumen que las observaciones actuales son una combinación lineal de observaciones

pasadas; (b) modelos de media móvil (*MA*), que capturan la aleatoriedad a través de promedios ponderados de errores y valores rezagados; (c) ARIMA, que combina *AR* y *MA*, y añade operaciones de diferencia para tratar con tendencias y estacionalidad; (d) SARIMA, que es una extensión del modelo ARIMA y añade un componente estacional al modelo para capturar este tipo de patrones; y (e) suavizado exponencial (*ES*), que ajusta pesos a datos históricos para mejorar la precisión (Sun et al., 2023, Adineh et al., 2021).

2.3.2 Aprendizaje automático o Machine learning en series de tiempo

El aprendizaje automático, o *machine learning*, proporciona mecanismos mediante los cuales el ordenador es capaz de aprender por sí mismo a resolver un problema, aprovechando para ello los datos a los que tenga acceso y las heurísticas de aprendizaje incorporadas en el algoritmo de aprendizaje creado por el programador (Berzal, 2018, p. 27). Esta capacidad de aprendizaje autónomo se ha vuelto relevante en el análisis de series temporales, debido a que diversos modelos de machine learning ofrecen alternativas avanzadas para modelar patrones complejos con tendencias no lineales y con dinámicas temporales, permitiendo incorporar características de las series, como los rezagos e inclusión de variables exógenas, lo que contribuye a generar predicciones más robustas. En este sentido, la flexibilidad de los algoritmos de aprendizaje autónomo y su capacidad de adaptación destaca frente a los métodos estadísticos clásicos.

El modelado de las series temporales mediante algoritmos de machine learning requiere incluir la transformación de la serie para que esta pueda ser tratada como un problema de aprendizaje supervisado, donde se utilizan las observaciones pasadas (rezagos) como variables predictoras. A diferencia de los modelos clásicos, estos algoritmos aprenden automáticamente patrones a partir de los datos mediante la

optimización de los hiperparámetros específicos por cada modelo. Además de necesitar el preprocesamiento de la serie, es fundamental el uso de una validación cruzada temporal y la validación del desempeño mediante las métricas de error, para determinar la precisión de las predicciones y la robustez de los resultados (Mariño Villalba, 2023).

3. Metodología

Este estudio tiene un alcance explicativo, ya que su objetivo es evaluar diferentes métodos de series de tiempo para la predicción de precios de dos variedades de aguacate. La investigación buscó identificar el método que mejor rendimiento tenga en la captura de los patrones y tendencias de los precios de este producto agrícola. Se aplicó un diseño no experimental de tipo longitudinal; este diseño es adecuado debido a la estacionalidad de los precios de los productos agrícolas y a la dependencia del precio en un momento dado de sus rezagos en el tiempo. Con base en lo anterior, se planteó la siguiente hipótesis: los modelos de *machine learning* proporcionan una mayor precisión en la proyección semanal de los precios nacionales del aguacate Hass y Papelillo en comparación con los métodos tradicionales de series de tiempo.

La población objeto de estudio está compuesta por los precios nacionales históricos disponibles del aguacate Hass y el aguacate Papelillo desde el 10 de noviembre de 2012 hasta el 27 de diciembre de 2024, informados semanalmente por el Departamento Administrativo Nacional de Estadística (DANE) a través del Sistema de Información de Precios (SIPSA). Esta información es consolidada para distintas plazas de mercado mayoristas monitoreadas por el DANE a nivel nacional, donde las plazas consultadas no son necesariamente las mismas en cada periodo y por lo tanto presentan variaciones regionales como lo puede ser en costos logísticos, márgenes de comercialización, disponibilidad de oferta, entre otros.

Para obtener las series temporales de las variedades Hass y Papelillo, se recolectaron, concatenaron y agruparon los archivos Excel anexos en el boletín semanal mayorista emitido por el DANE. No se observaron valores faltantes por lo que no se requirió aplicar técnicas de imputación o ajustes adicionales.

Como parte del preprocesamiento de los datos se utilizó la media aritmética como medida de tendencia central para representar el precio semanal de las variedades Hass y Papelillo, con el fin de resumir la diversidad de los precios a través de las distintas regiones, reducir el impacto de valores atípicos y así facilitar la detección de tendencias o patrones generales.

Esta metodología es consistente con las prácticas empleadas por organizaciones como la *Food and Agriculture Organization* (FAO) en su *Food Price Index* , que mide los cambios mensuales en precios internacionales de alimentos, o los boletines del DANE, donde el uso de precio medio se usa para el monitoreo del comportamiento económico del sector agrícola en Colombia.

Ahora, para realizar un modelo ARIMA es importante que se tenga la estacionariedad como criterio, es decir, se asume que el comportamiento de las series de tiempo permanece en equilibrio estadístico, es decir que la media y la varianza de estas series, permanecen constantes a través del tiempo (Box et al., 2016).

Por esta razón, para cumplir con este criterio de estacionariedad, indispensable para modelos clásicos ARIMA cómo diversos algoritmos de Machine Learning se llevó a cabo un análisis de estacionariedad a las series de tiempo aguacate Papelillo y Hass, con la prueba de Dickey-Fuller aumentado (ADF) para detectar estacionariedad, la prueba Ljung-Box para validar el supuesto de independencia temporal en los residuos, prueba

ARCH y Breusch-Pagan para indicar si la varianza cambia con el tiempo, obteniendo los resultados que se muestran en la Tabla 1.

Tabla 1

Pruebas de varianza antes de la transformación de la serie.

	Precio Papelillo		Precio Hass	
Prueba	*Valor p*	*Resultado*	*Valor p*	*Resultado*
ADF	0.0865	No estacionaria	0.9406	No estacionaria
Ljung-Box	0	Varianza no estable	0	Varianza no estable
ARCH	0	Varianza no estable	0	Varianza no estable
Breusch-Pagan	0	Varianza no estable	0	Varianza no estable

Dado que ambas series presentaron una varianza no estable, para cumplir el criterio de estacionariedad y estabilizar la varianza, fueron transformadas aplicando normalización móvil con ventanas de cinco y cuatro periodos para el Papelillo y Hass, respectivamente.

Entonces para la transformación se tuvo que, siendo P_t como el precio medio en el tiempo t, y $\sigma_t^{(w)}$la desviación estándar móvil de P_t en una ventana de tamaño w, la variable normalizada P_t se presenta como:

$$\tilde{P}_t = \frac{P_t}{\sigma_t^{(w)}}$$

Donde,

- $\tilde{P}_t$ representa el precio medio normalizado en el tiempo t.
- $\sigma_t^{(w)} = std(P_{t-w+1}, \dots, P_t)$ es la desviación estándar calculada sobre una ventana de tamaño w.

- $\tilde{P}_t$ representa el precio medio normalizado.

Una vez estabilizada la varianza, se procedió a aplicar una diferenciación a ambas series para lograr la estacionariedad. La diferenciación se realizó con un rezago de 26 periodos y así eliminar tendencia y estacionalidad presentes en las series.

A partir de volver a aplicar las pruebas expuestas en la Tabla 1 a las series ya transformadas, se observaron los resultados que se muestran en la Tabla 2.

Tabla 2

Pruebas de varianza después de la transformación de la serie.

	Precio Papelillo		Precio Hass	
Prueba	*Valor p*	*Resultado*	*Valor p*	*Resultado*
ADF	0	Estacionaria	0	Estacionaria
Ljung-Box	0	Varianza no estable	0	Varianza no estable
ARCH	0.1909	Varianza estable	0.6723	Varianza estable
Breusch-Pagan	0.2445	Varianza estable	0.2286	Varianza estable

Por lo cual se observa que la normalización móvil logró hacer que las propiedades probabilísticas de media y varianza se estabilizaran a través del tiempo, mientras que la diferenciación eliminó la tendencia y la estacionalidad, dejando series estacionarias para realizar los modelados.

Además de estas transformaciones también se exploraron otras alternativas, como la transformación logarítmica, raíz cuadrada, Box-Cox y Yeo-Johnson. A pesar de esto, ninguna de estas transformaciones mostró resultados efectivos como la normalización móvil combinada con la diferenciación. Por lo tanto, se optó por esta última para el análisis y modelado posterior.

Para el análisis, se utilizaron las siguientes herramientas: *Python* versión 3.11.12 y *R* versión 4.3.2 dentro de un entorno de *Google Colab* versión 1.0.0, empleando librerías como *Pandas*, *Numpy*, *Scikit-learn*, *Statsmodel*, *Matplotlib*, *Seaborn*, *Skforecast*, *Optuna* en *Python* y *Tidyverse*, *TSeries y Forecast* en *R*. Las técnicas de análisis incluyeron métodos de predicción de series temporales como, *SARIMA*, modelos de regresión lineal regularizados y algoritmos de *machine learning* como modelos de árboles de decisión y modelos tipo *boosting*. Se destacan los métodos de series temporales por la naturaleza longitudinal de los datos, caracterizados por su dependencia y secuencia temporal. Los métodos seleccionados, basados en la revisión de literatura, son aquellos que mostraron mayor frecuencia de uso y mejores resultados en cada enfoque. El análisis incluyó una comparación de métricas. Se utilizaron las siguientes métricas para evaluar los modelos: error cuadrático medio (*RMSE)*, error absoluto medio (*MAE)* y error porcentual absoluto medio (*MAPE)*, las fórmulas de cada métrica se presentan en las ecuaciones (1), (2) y (3), respectivamente:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{1}$$

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

$$\mathrm{MAPE} = \frac{100\backslash\%}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3}$$

donde y_i es el valor real observado, $\hat{y}_i$ es el valor predicho y n es el número de observaciones.

Para evaluar la robustez de los resultados, se utilizó validación cruzada, lo cual permitió controlar el sobreajuste de los modelos y aprovechar todos los datos disponibles.

3.1 Implementación de modelos

Los modelos evaluados se clasificaron en dos grupos principales, modelos fundamentados en técnicas de estadística clásica junto con modelos de regresión lineal regularizada y modelos basados en algoritmos de *machine learning*.

a. Modelos de estadística clásica y de regresión lineal regularizada: *SARIMA*, *Ridge*, *Bayesian Ridge*, *Lasso*, *LassoLars*, *ElasticNet*. Estos modelos asumen una estructura funcional explícita y dependen del cumplimiento de ciertos supuestos estadísticos. Por lo general se consideran más adecuados para horizontes de predicción de corto plazo, (James, Witten, Hastie, & Tibshirani, 2013).
b. Modelos de *Machine learning*: *Orthogonal Matching Pursuit (OMP)*, *Support Vector Regressor (SVR)*, *Random Forest*, *Extra Trees*, *AdaBoost*, *CatBoost*, *LightGBM*, *XGBoost*, *GradientBoosting* y *DecisionTree*. Estos modelos se centran en la detección de patrones complejos, priorizando el rendimiento predictivo sobre la interpretabilidad, y han demostrado ser más eficaces en horizontes de predicción más largos.

Todos los modelos a excepción del SARIMA, fueron implementados en Python, para lo cual se llevó a cabo una búsqueda de hiperparámetros, incluyendo la selección de los mejores rezagos, lo que permitió obtener mayor estabilidad y precisión de las predicciones.

El modelo SARIMA (*Seasonal Autoregressive Integrated Moving Average*) relaciona las observaciones de un momento dado con valores rezagados en el tiempo mediante componentes autorregresivos (AR) y de medias móviles (MA), incorporando las diferenciaciones necesarias para hacer la serie estacionaria (I). El modelo se entrenó con la función auto.arima del paquete *forecast* en R, la cual selecciona automáticamente el mejor modelo ARIMA o SARIMA para una serie temporal. Para ello, optimiza los hiperparámetros del modelo utilizando criterios de evaluación como AIC, AICc o BIC.

El modelo SARIMA con estacionalidad de 26 semanas (seis meses) se empleó como modelo de referencia, dado que es un modelo tradicionalmente utilizado en *forecasting* reconocido por su efectividad, debido a que combina componentes autorregresivos y de media móvil que le confiere una alta capacidad para modelar la estacionalidad capturando así los patrones temporales. Este modelo es ampliamente utilizado gracias a su interpretabilidad y facilidad de implementación.

3.1.1 Búsqueda de hiperparámetros mediante Optuna *en* Skforecaster

Se implementó una búsqueda bayesiana que además de ajustar los hiperparámetros, obtiene los mejores rezagos temporales como variables predictoras. Este proceso se beneficia de la información de iteraciones previas a través de un modelo probabilístico, con el objetivo de minimizar las métricas de validación, en este caso, el error cuadrático medio (*MSE*). Este esquema se desarrolló mediante *Skforecast*, una biblioteca en *Python*, diseñada para la predicción de series temporales que permite utilizar modelos de *machine learning* compatibles con la API *Scikit-learn* (Amat & Escobar Ortiz, 2025).

Para potenciar esta búsqueda, se utilizó la biblioteca *Optuna*, una herramienta de código abierto en *Python*, que permite realizar optimización bayesiana (Akiba et al.,

2019). logrando encontrar configuraciones óptimas de hiperparámetros y reduciendo el tiempo de búsqueda. La evaluación del rendimiento de los modelos en la búsqueda de las mejores combinaciones de hiperparámetros, se efectuó utilizando la validación cruzada temporal mediante la técnica de *TimeSeriesFold*, la cual conserva la naturaleza temporal de los datos, previniendo el uso de información futura durante el entrenamiento.

Todos los modelos evaluados mediante *Skforecast*, emplearon la clase *ForecasterRecursive*, ya que esta requiere de menos recursos computaciones en comparación con *ForecasterDirect.* Este enfoque recursivo utiliza la predicción de los pasos previos para obtener pronósticos futuros. La optimización de la búsqueda se realizó explorando mil iteraciones, identificando las combinaciones que minimizan la métrica objetivo.

Además de la optimización de hiperparámetros a través de *optuna*, se realizó búsqueda de hiperparámetros con *GridSearchCV* minimizando el error cuadrático medio (MSE) calculado a partir de tres particiones (*splits*) diferentes. La grilla de búsqueda se aplicó a catorce combinaciones de rezagos, que incluían hasta cuatro rezagos no estacionales y hasta dos rezagos estacionales. Entre los modelos resultantes con hiperparámetros optimizados para cada combinación de rezagos, también se comparó el desempeño a través del MSE y se seleccionó la combinación de rezagos más eficiente. Los modelos realizados a través de este flujo de trabajo fueron: *SVR, Lineal, Ridge, ElasticNet, OMP, Bayesian Ridge, Decision tree, Lasso* y *Lasso LARS*. Sin embargo, los hiperparámetros encontrados con *optuna* llevaban, en general, a mejores resultados.

Tras el entrenamiento de los modelos con la optimización de hiperparámetros y el mejor conjunto de rezagos temporales para cada algoritmo, se evaluaron los modelos de acuerdo con las métricas de error ya mencionadas.

Se seleccionaron las configuraciones idóneas para cada modelo, optimizando los hiperparámetros clave y logrando las mejores predicciones. Para los modelos de regresión lineal regularizada, como Lasso, LassoLars, Ridge, Bayesian Ridge y ElasticNet, se ajustó principalmente el parámetro alpha (o lambda), el cual controla la fuerza de la regularización. Estos modelos utilizan penalizaciones L1, L2 o una combinación de ambas para restringir la magnitud de los coeficientes, con el fin de mejorar la generalización y prevenir el sobreajuste.

Para los modelos de *boosting* como XGBoost, CatBoost, LightGBM, AdaBoost y GradientBoosting, se optimizaron hiperparámetros clave como n_estimators (número de árboles), learning_rate (tasa de aprendizaje), max_depth (profundidad de los árboles) y subsample (proporción de datos usada por árbol). Estos ajusten buscan un equilibrio entre capacidad predictiva y control del sobreajuste según las particularidades de cada modelo.

En los algoritmos de *bagging*, como Random Forest y Extra Trees, se ajustaron parámetros como n_estimators, max_depth, min_samples_split, min_samples_leaf y max_features, con el objetivo de generar árboles suficientemente diversos a partir de muestras bootstrap y mejorar la estabilidad del modelo por medio de agregación.

Por su parte, el modelo OMP, se centró en el ajuste de n_nonzero_coefs, que define el número máximo de coeficientes distintos de cero (es decir, la cantidad de rezagos que puede utilizar el modelo), y tol, que indica la tolerancia del error de ajuste.

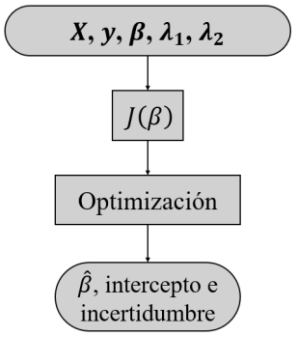
En el modelo SVR, se seleccionaron los hiperparámetros C (penalización del error), epsilon (margen de tolerancia del error) y *kernel* (función de transformación de los datos), con el objetivo de ajustar una función que permita predecir dentro de un margen de error aceptable sin sobreajustar.

Finalmente, en el modelo Decision Tree Regressor, se ajustaron los hiperparámetros max_depth, min_samples_split y min_samples_leaf, controlando así la complejidad del árbol para evitar sobreajuste y mejorar la interpretación.

3.1.2 Funcionamiento de los algoritmos

Los modelos de regresión lineal regularizada comienzan definiendo X, y, β, y los hiperparámetros λ_1, λ_2. A partir de los valores definidos se construye una función objetivo $J(\beta)$ que combina el ajuste del modelo y con término de regularización. Para realizar la optimización, Ridge aplica una solución cerrada, mientras que Lasso y ElasticNet utilizan descenso por coordenadas y Bayesian Ridge usa la inferencia bayesiana. Para llegar a la convergencia, el algoritmo itera hasta que el cambio en $J(\beta)$ sea menor que un umbral. Como salida, se obtienen los coeficientes $\hat{\beta}$, el intercepto y, si aplica, una medida de incertidumbre (scikit-learn developers, s.f.). La Figura 1 muestra la estructura de este proceso.

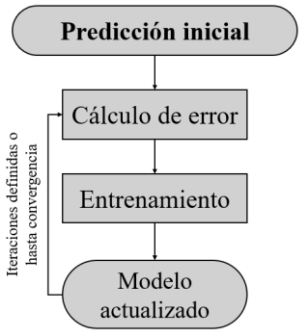
Figura 1
Diagrama representativo de las regresiones lineales regularizadas



La Figura 2 muestra el funcionamiento de los algoritmos de boosting. Estos modelos comienzan con una predicción inicial que es constante (por ejemplo, la media de la variable objetivo). A continuación, se calcula el error entre las predicciones estimadas y los valores reales para entrenar un nuevo modelo base (generalmente un árbol débil) sobre los residuos. El modelo global se actualiza sumando la nueva predicción ponderada por con un factor de aprendizaje η.

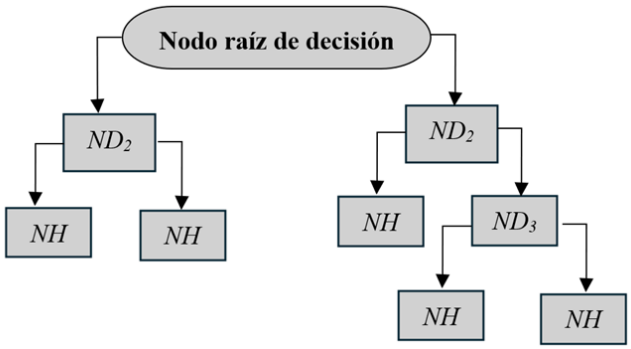
El proceso se repite por un número definido de etapas o hasta llegar a la convergencia (Amazon, n.d.).

Figura 2
Diagrama representativo de los modelos de boosting



La Figura 3 representa la estructura del modelo Decisión Tree o árbol de decisión, donde se realizan divisiones progresivas del conjunto de datos con el fin de generar predicciones específicas.

Figura 3
Diagrama representativo del árbol de decisión



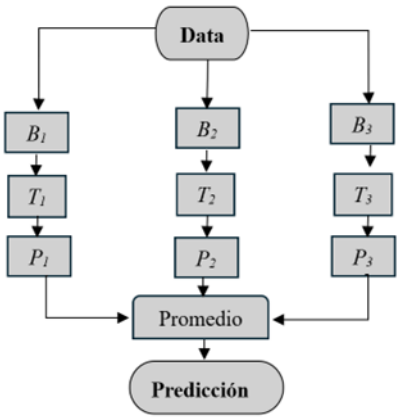
El proceso inicia con el nodo raíz, donde el modelo selecciona las características con mayor capacidad informativa, para dividir los datos de manera eficiente, minimizando la heterogeneidad de los grupos formados. Esta división, conduce a nuevos nodos de decisión intermedios (ND_2), que siguen segmentando los datos en función de

otras variables. Finalmente, se generan nodos hojas (NH), que representan las salidas como predicciones finales del modelo (Bartz, Bartz-Beielstein, Zaefferer & Mersmann, 2022).

En cuanto a los modelos que implementan la técnica de *bagging*, tal como se ilustra en la Figura 4, a partir del conjunto completo de datos (Data), generan diferentes subconjuntos por medio de un muestro aleatorio con reemplazo, conocido como *bootstrap* (B_1, B_2 y B_3). Cada muestra es usada para entrenar un modelo independiente. En algoritmos de ensamble como Random Forest, los modelos bases corresponden a árboles de decisión (T_1, T_2, y T_3). Cada uno de los modelos genera su propia predicción (P_1, P_2 y P_3), posteriormente, estas predicciones se integran a través del mecanismo de agregación basado en el promedio de las predicciones. Como resultado del proceso, se obtiene la predicción final optimizada buscando que sea generalizable y robusta (Nayak, Sharma, Bhakar & Tiwari, 2025).

Figura 4

Diagrama representativo del enfoque bagging



El algoritmo Random Forest utiliza el principio de bagging, pero también incorpora una aleatoriedad adicional, ya que, en cada nodo, selecciona un subconjunto de

variables, a diferencia del bagging tradicional, que utiliza todas las variables. Esto da como resultado árboles con mayor diversidad.

Por su parte, el modelo Extra trees, al igual que Random Forest, también se basa en árboles de decisión, sin embargo, se diferencia en que no realiza boostrap, sino que emplea todo el conjunto de datos para entrenar cada árbol, seleccionando aleatoriamente el umbral de corte en cada nodo, en lugar de buscar el umbral óptimo.

En todos los modelos, se incorporó un proceso para la selección de rezagos tanto estacionales como no estacionales como parte fundamental del modelado, dado que estos definen la estructura temporal de entrada y tienen un impacto directo en la precisión de las predicciones. En desempeño, destacaron los modelos: AdaBoost, Decision Tree, ElasticNet, Extra Trees, Gradient Boosting, LightGBR, Random Forest y Ridge.

3.2 Estimación de predicciones e intervalos de confianza

Para llevar a cabo la estimación de las proyecciones y sus intervalos de confianza al 95%, se implementó un método basado en *Bootstrap*, debido a que las predicciones no cumplían con el supuesto de normalidad, evaluado a través de una prueba *Shapiro-Wilk*. Se realizaron mil remuestreos aleatorios con reemplazo del conjunto de datos**,** entrenando el modelo en cada iteración, con los hiperparámetros que ya habían sido optimizados anteriormente. A partir de cada modelo reentrenado, se generaron proyecciones recursivas para un horizonte de predicción de cuatro semanas, replicando la estructura de rezagos utilizada en el modelo final.

Lo anterior permitió la construcción de una distribución empírica de las distribuciones, de las que se estimaron los percentiles 2.95 y 97.5 para definir los límites del intervalo de confianza. La predicción puntual se obtuvo del cálculo del promedio de

las predicciones generadas por el *Bootstrap*. A través de este enfoque es posible capturar un acercamiento a la incertidumbre propia de las estimaciones sin estar sujeto a los supuestos de las distribuciones paramétricas.

En este estudio el horizonte de predicción fue de cuatro semanas, ya que las series temporales para el aguacate Papelillo y Hass se trataron con frecuencia semanal. Para esto se tuvo en cuenta la necesidad de capturar tendencias de mercado inmediatas y el hecho de asegurar la estabilidad de las condiciones de los precios para contar con predicciones confiables. Además, al pronosticar un periodo mensual, se respeta la dinámica del mercado en cuanto a productos agrícolas perecederos, facilitando la toma de decisiones en prácticas de logística, comercialización y planificación.

Los horizontes superiores al mes incrementaron considerablemente el error del pronóstico debido a la acumulación de incertidumbres y finalmente, un horizonte mensual permite el equilibrio entre precisión en la predicción y la utilidad operativa del sector agrícola (Apéndice B).

4. Resultados

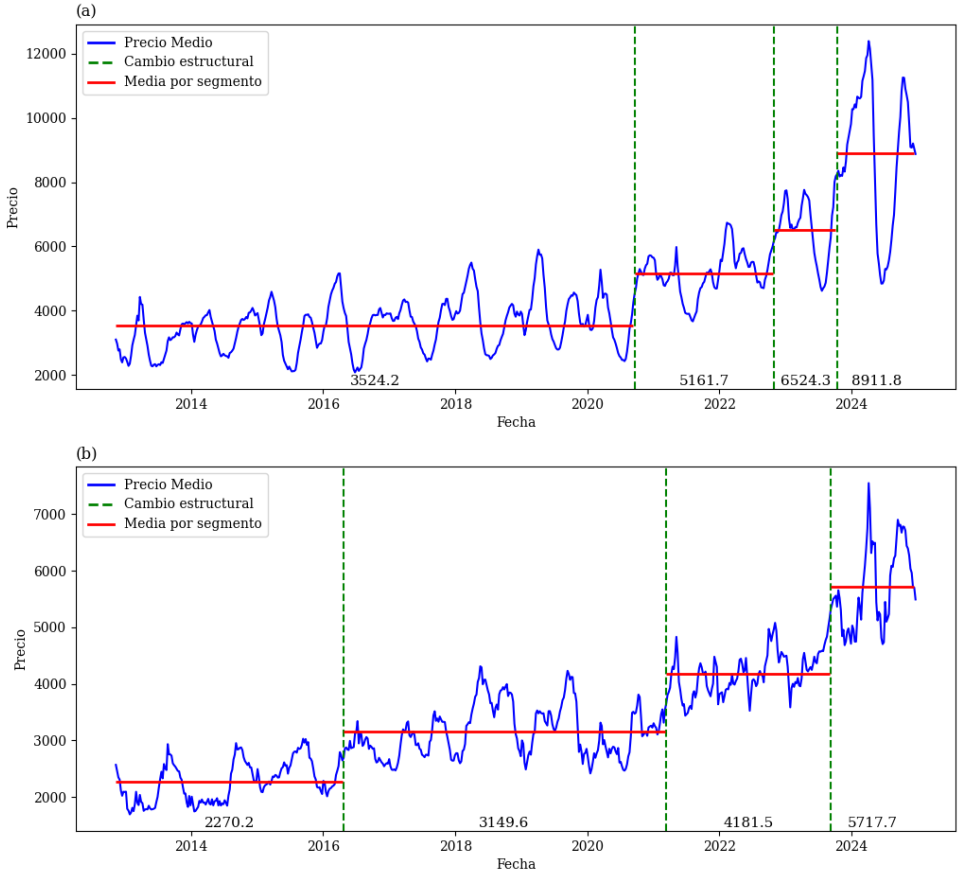
4.1 Análisis exploratorio de los datos

En la Figura 5 se observa el precio nacional de las variedades de aguacate papelillo y Hass y su evolución a través del tiempo. La media del precio del aguacate papelillo es de $4 581.9, con una desviación estándar de $1 988.8, un coeficiente de variación (CV) de 43.4% y desviación de retornos logarítmicos de 4.96%. Por otra parte, la variedad Hass presenta una media de $3 387.0, $1 137.4 de desviación estándar, representando un CV de 33.6% y desviación de retornos logarítmicos de 4.5%. Al comparar las series, de forma

gráfica y desde las medidas de variabilidad, se observó que el precio del aguacate papelillo fue más volátil que el del Hass.

Figura 5

(a) Evolución del precio del aguacate papelillo. (b) Evolución del precio del aguacate Hass.



En los últimos 12 años, a los que corresponde el periodo de observación, el precio mayorista de ambos productos ha presentado fuertes cambios en su tendencia, que se ve reflejado en cómo las medias de las series van aumentando a través del tiempo, con una tendencia al crecimiento más marcada en la variedad papelillo que en Hass. A través del método de segmentación binaria[4] (binseg), que detecta puntos o momentos en los que una

[4] El método binseg trabaja de forma secuencial. Primero, encuentra un punto en los datos donde hay un cambio de propiedades estadísticas y allí segmenta la serie en dos subseries. Luego, repite el proceso dentro de cada una de las subseries y continúa hasta que no encuentre más cambios significativos (Zdansky, J., 2006). En el caso específico del algoritmo *ruptures* implementado en Python, es posible indicar el número de segmentaciones que deben buscarse.

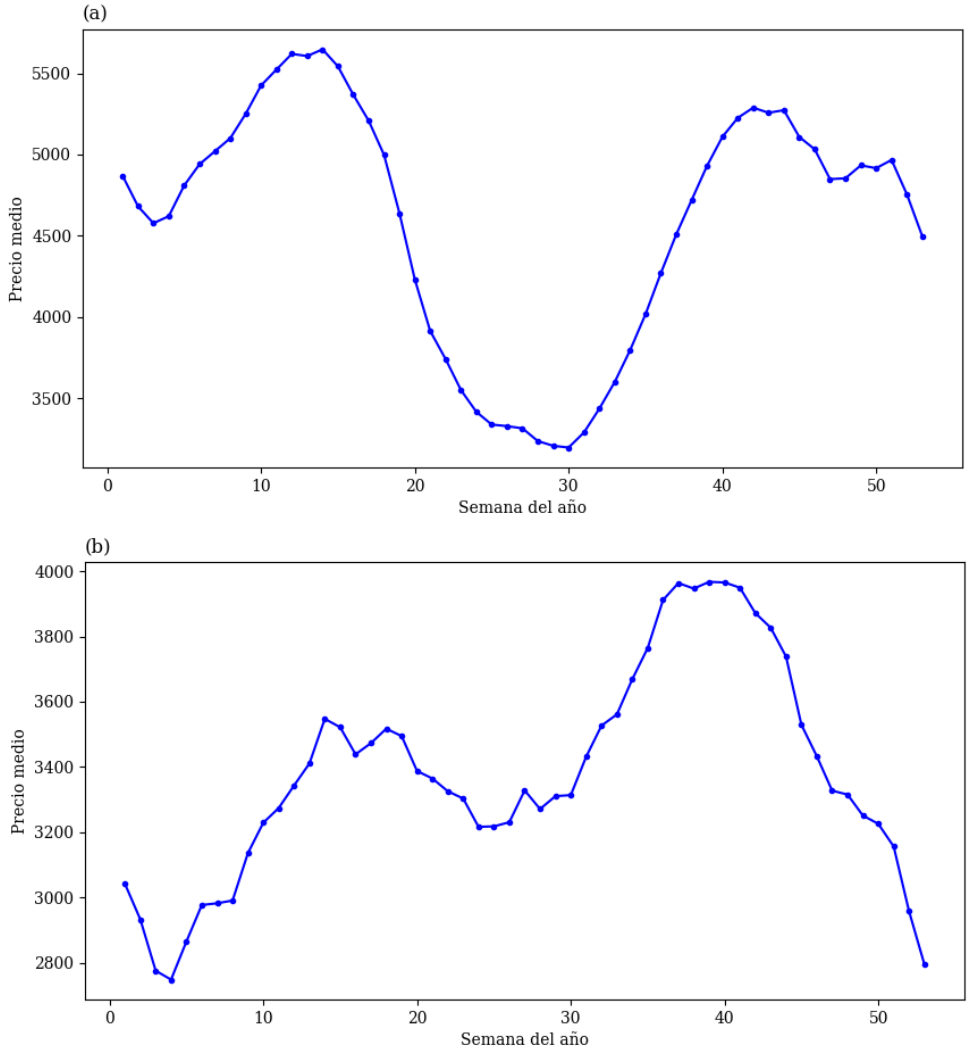
serie tiene cambios significativos en sus propiedades estadísticas, se identificaron tres cambios estructurales en cada una de las series. La serie de papelillo presenta un incremento de 36.49%, en promedio entre los cambios estructurales, que fueron identificados en los puntos correspondientes a las fechas: 19/09/20, 29/10/22 y 14/10/2023. Por parte del aguacate Hass, los cambios se observaron en 23/04/23, 13/03/21 y 09/09/23, presentando una variación promedio de la media de 35.73% entre cambios estructurales. Además de la tendencia, se ve una marcada estacionalidad que, dada la naturaleza del producto, corresponde a los diferentes picos y caídas en la cosecha, que afectan la oferta y, por ende, el precio del producto en el mercado.

4.2 Detección de estacionalidad

Como se observa en la Figura 6, tanto el aguacate papelillo como el Hass presentan una estacionalidad de 26 semanas, lo que corresponde a, aproximadamente, seis meses. Como la mayoría de los productos agropecuarios, su precio responde en gran medida a la oferta que haya disponible en el mercado y qué tanto suple la demanda. En el caso del aguacate, hay dos temporadas de cosecha: la principal, que se da entre octubre y diciembre, y la traviesa, que se recolecta en abril. Lo anterior lleva a que, cuando se comienzan las cosechas, el precio inicie el descenso (aproximadamente en la semana 14 y en la 40-42, que corresponden a los meses de cosecha), mientras que durante la época que no hay una fuerte salida de producción el precio vuelve a elevarse, subiendo más a medida que se aleja de la época de cosecha y el desabastecimiento se hace más fuerte.

Figura 6

(a) Precio promedio por semana del aguacate papelillo 2013-2024. (b) Precio promedio por semana del aguacate Hass 2013-2024.



Esta estacionalidad es más marcada en el caso de la variedad papelillo, ya que esta variedad no genera producción en los intervalos de tiempo entre ambas cosechas del año, mientras que el Hass reduce notablemente su producción, pero continúa habiendo salida de producto para abastecer, al menos, una parte del mercado. Además, el papelillo es el que más se comercializa y consume en el país, lo que contribuye a que los choques de oferta y demanda sean más marcados en su caso.

Dado lo anterior, se realizó una exploración a través de modelos SARIMA, en los que se encontró que aquellos modelos con estacionalidad de 26 semanas eran los que, de

forma más recurrente, presentaban mejores métricas. Lo anterior se realizó a través de una grilla de búsqueda que entrenó más de 500 modelos para cada una de las variedades de aguacate trabajadas.

4.3 Precio del aguacate en Colombia en el contexto del COVID-19.

La pandemia de COVID-19, declarada el 11 de marzo de 2020, tuvo efectos significativos en la economía, afectando el sector agrícola. En Colombia, la cuarentena nacional, implementada para mitigar la propagación del virus, impactó seriamente la cadena productiva.

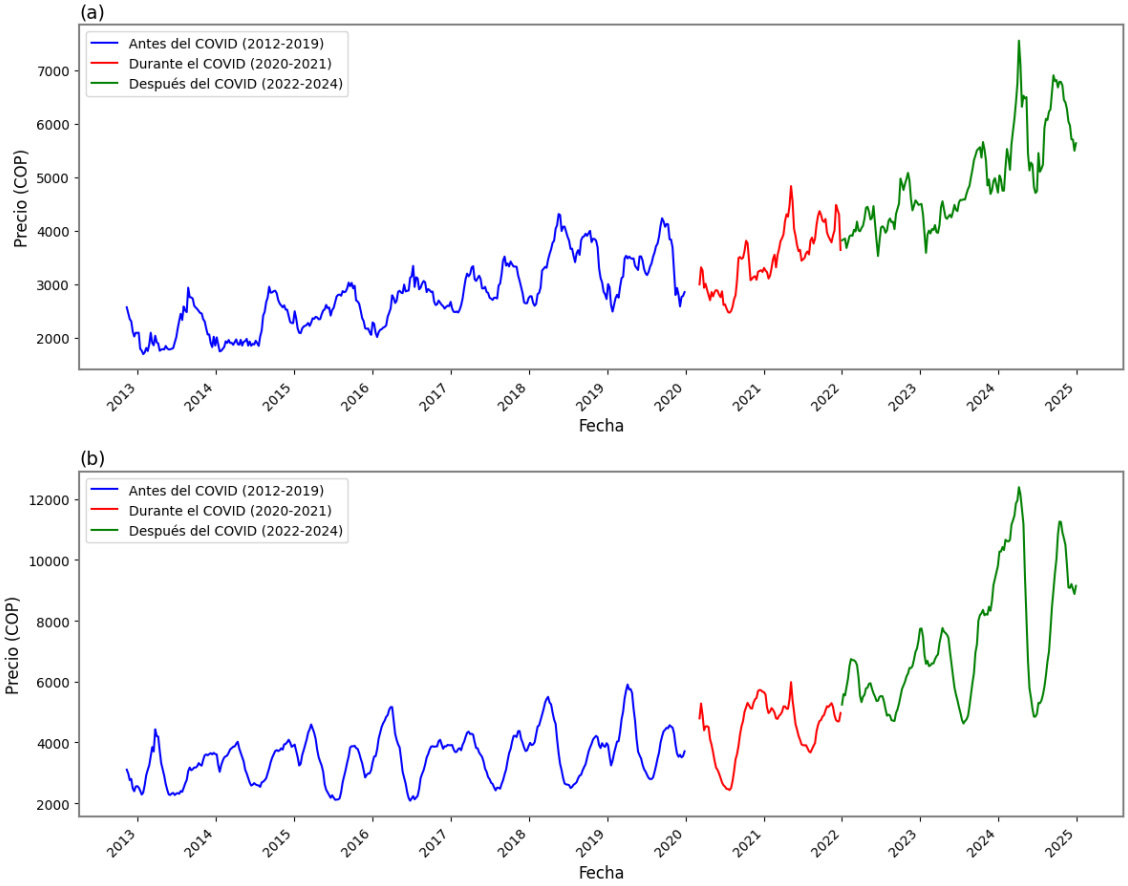
Según el informe "*Análisis de afectación de las cadenas productivas agropecuarias derivada de la emergencia sanitaria por COVID 19*", elaborado por la Dirección de Cadenas Agrícolas y Forestales del Ministerio de agricultura (2020), las restricciones durante la cuarentena generaron incrementos en los costos de producción (insumos y fertilizantes), fluctuaciones en los precios y alteraciones en las cosechas.

Si bien el aguacate no tuvo afectaciones severas, se evidenció un aumento en los costos de producción, compensados por un ligero incremento en el precio nacional, sin cambios relevantes en la estacionalidad de las cosechas.

El comportamiento de la serie, como se evidencia en la Figura 7, indica que, durante la pandemia, la estructura estacional y la tendencia de los precios se conservaron estables. Sin embargo, a partir de noviembre de 2023 y hasta el año 2024, se observa un alza considerable en los precios del aguacate Hass y Papelillo, atribuida no a efectos postpandemia, sino al impacto del fenómeno El Niño, que redujo la oferta de ambas variedades, y aumentó el precio de venta en promedio, para el aguacate Hass en un 19.6% y para el aguacate papelillo en un 35.1%.

Figura 7

(a) Impacto del COVID-19 en el precio promedio semanal del aguacate Papelillo (b) Impacto del COVID-19 en el precio promedio semanal del aguacate Hass



4.4 Resultados aguacate Papelillo

Los modelos que mostraron un mejor desempeño, al evaluar las métricas de error durante el entrenamiento fueron: Extra Trees, SARIMA, Ridge, ElasticNet, Decision Tree y LightGBR. Estos resultados se obtuvieron a partir de datos transformados y se muestran en la Tabla 3.

En la Tabla 4 se pueden observar los modelos con mejores resultados, esta vez con métricas de error evaluadas a través de validación externa con los valores reales de la serie para las primeras semanas del 2025, teniendo en cuenta un horizonte de predicción de cuatro periodos. El mejor modelo se seleccionó basado en el desempeño conjunto en

términos de RMSE, MAE y MAPE (tanto en entrenamiento como en validación externa) y del tamaño del intervalo de predicción al 95% de confianza.

Tabla 3

Métricas en entrenamiento para modelos de aguacate papelillo

Modelo	Hiperparámetros	Rezagos	RMSE	MAE	MAPE
Extra Trees	'n_estimators': 450, 'max_depth': 19, 'min_samples_split': 3, 'min_samples_leaf': 1, 'max_features': 0.5740875608309076, 'bootstrap': True	26, 52	21.55	11.647	249%
SARIMA	(1,0,0) (2,0,0) [26]		32,28	16.09	321.64%
Ridge	'alpha': 38.51037649951748, 'fit_intercept': False, 'solver': 'saga', 'tol': 0.0003185732960505859, 'max_iter': 926	1, 26, 52	33.63	17.54	364%
Elastic Net	'alpha': 1.000005091758481e-06, 'fit_intercept': False, 'max_iter': 639, 'tol': 1.2545674446215401e-05	1, 26, 52	33.63	17.5439	365%
Decision Tree	'max_depth': 3, 'min_samples_split': 4, 'min_samples_leaf': 8, 'max_features': 'sqrt'	26	36.63	20.36	474%
LightGBR	'n_estimators': 300, 'num_leaves': 80, 'learning_rate': 0.037115934721279525, 'max_depth': 4, 'min_child_samples': 27, 'subsample': 0.9624347534298316, 'colsample_bytree': 0.8577263560429845, 'max_bin': 35	26	38.73	20.7914	326%

Tabla 4

Métricas en validación externa para modelos de aguacate papelillo

Modelo	RMSE	MAE	MAPE	Tamaño del intervalo
Decision Tree	1 245.96	1 221.45	14.09%	4 518.26

Modelo	RMSE	MAE	MAPE	Tamaño del intervalo
LightGBR	1 318.18	1 262.57	14.60%	4 966.95
SARIMA	1 779.97	1 397.52	15.92%	42 855.92
ExtraTrees	1 877.07	1 528.97	17.18%	4 210.22
Ridge	1 631.88	1 586.52	18.55%	6 571.242
ElasticNet	1 653.25	1 608.75	18.81%	6 170.016

Al revisar las métricas de los modelos de Machine Learning durante el entrenamiento, se observa que extra Trees presenta un mejor desempeño comparado con los demás, con un RMSE de 21.55, MAE de 11.647 y un MAPE de 249%, seguido de Ridge y ElasticNet, que son modelos basados en regresiones lineales penalizadas y, por último, el modelo Decision Tree, con un RMSE de 36.26, MAE de 20.36 y un MAPE de 474%, considerablemente más alto que el de los demás modelos.

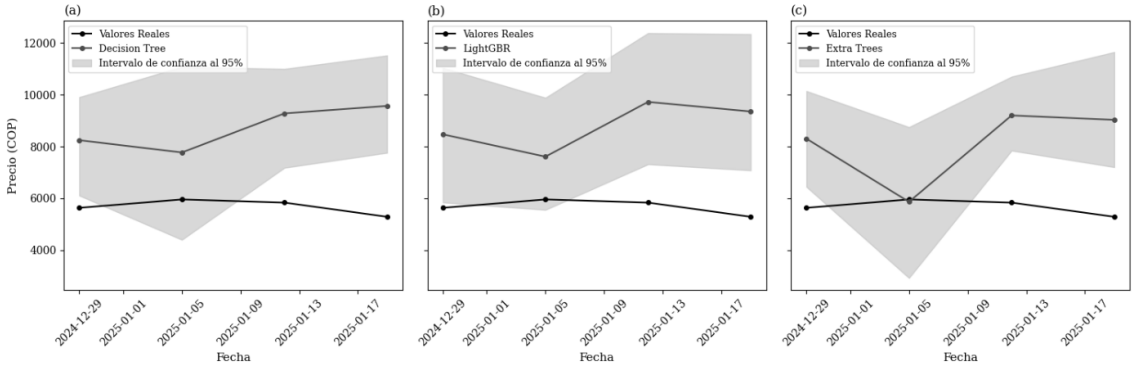
Sin embargo, al realizar validación externa de los modelos (calculando las mismas métricas para las predicciones con los valores reales del mercado, en escala original), la situación es bastante diferente. En esta validación, el modelo Decision Tree pasa a tener el mejor desempeño con un RMSE de 1 245.96, un MAE de 1 221.45 y un MAPE de 14.09%, mostrando equilibrio entre precisión y estabilidad, al contar con un tamaño de intervalo relativamente reducido de 4 518.26. Por otra parte, Extra Trees, que se había presentado como el mejor modelo, muestra degradación en la validación externa, con un RMSE de 1 877.07 y un MAPE de 17.18%, lo que da indicios de un posible sobreajuste a los datos; situación similar ocurrió con los modelos Ridge y ElasticNet, que vieron su desempeño relativo reducido. En cuanto al modelo LightGBR, resalta que, aunque no se ubicó entre los tres mejores durante el entrenamiento, sus métricas destacan durante la validación externa, subiendo cuatro posiciones en desempeño y superando a todos aquellos que hacen parte de la familia de los modelos lineales; lo anterior, con un intervalo de confianza relativamente más alto que el del Decision Tree, pero que se mantiene de un tamaño de amplitud viable (4 966.95).

Por parte del modelo SARIMA, aunque inicialmente mostró mejores métricas que Decision Tree, en la validación presenta métricas de error más altas, acompañadas de un intervalo de confianza que es excesivamente amplio (42 855.92). Este valor, en un contexto donde los precios rondan una media de 9 079.53 (en su predicción), resulta excesivo, llegando incluso a proyectar límites inferiores negativos, lo que evidencia que, a pesar de su ajuste aceptable en error (MAPE de 15.92%), no ofrece predicciones confiables para fines puntuales. Lo anterior sugiere que, si bien el modelo SARIMA logra capturar parte de la tendencia, su capacidad al momento de generar predicciones no conocidas hacia adelante es limitada, probablemente por su estructura lineal que no alcanza a modelar la complejidad de los datos que recibió.

Por otra parte, al comparar puntualmente las predicciones (con sus respectivos intervalos de confianza) con los valores reales, solo la segunda semana hacia adelante logra mantenerse dentro del intervalo de confianza de la predicción para el caso de los modelos Decision Tree, LightGBR y Extra Trees (ver Figura 8), que fueron aquellos que presentaron las mejores métricas de error en validación externa. A pesar de lo anterior, el modelo que más logra acercarse es el LightGBR, que presenta menores distancias entre los límites del intervalo y el valor real (para aquellos puntos que no se encuentran dentro del intervalo.

Figura 8

(a) Comparación de las predicciones del modelo Decision Tree para el aguacate papelillo respecto a los valores reales. (b) Comparación de las predicciones del modelo LightGBR para el aguacate papelillo respecto a los valores reales. (c) Comparación de las predicciones del modelo Extra Trees para el aguacate papelillo respecto a los valores reales.



Estos resultados resaltan que, al comparar, los modelos basados en árboles logran capturar las tendencias de una mejor forma frente a modelos tradicionales de series de tiempo, para este escenario específico, siendo el modelo LightGBR el que se consolida como el mejor al presentar un equilibrio entre métricas en entrenamiento, validación externa y estabilidad de las predicciones.

4.5 Resultados aguacate Hass

Los modelos con mejor desempeño en cuánto a las métricas de error evaluadas durante el entrenamiento fueron: Decision Tree, XGBoost, GradientBoosting, ExtraTree, Random Forest, SARIMA y AdaBoost. Estos resultados se obtuvieron con los datos transformados y se resumen en la tabla 5.

En la Tabla 6 se presentan los modelos con mejores resultados en la validación externa incluido el modelo referente SARIMA, para la predicción semanal del precio del aguacate Hass, considerando un horizonte de un mes, correspondientes a las primeras semanas de 2025. El criterio de selección de los modelos se basó en el desempeño

conjunto en términos de RMSE, MAE y MAPE y del tamaño del intervalo de predicción al 95% de confianza.

Tabla 5

Métricas en entrenamiento para modelos de aguacate Hass

Modelos	Hiperparámetros	Rezagos	RMSE	MAE	MAPE
XGBoost	n_estimators: 700, max_depth: 3, learning_rate: 0.010092248633727858, subsample: 0.5176263790087087, colsample_bytree: 0.736998639654112, reg_alpha: 0.33675294750019896, reg_lambda: 0.5382157802692751.	26, 52	38.36	25.21	313.00%
SARIMA	ARIMA (1,0,0) (2,0,0) [26]		42.68	24.11	325.38%
Gradient Boosting	n_estimators: 300, max_depth: 3, learning_rate: 0.01005345460479184, subsample: 0.8315168873694121, min_samples_split: 9, min_samples_leaf: 3.	26, 52	43.39	25.56	311.00%
AdaBoost	n_estimators: 100, learning_rate: 0.010986012314572854	1	45.05	27.04	317.00%
Extra Trees	n_estimators: 200, max_depth: 6, min_samples_split: 6, min_samples_leaf: 1, max_features: 0.7110099367851404, bootstrap: False	26	45.46	26.99	236.00%
Random Forest	n_estimators: 450, max_depth: 3, min_samples_split: 2, min_samples_leaf: 3, max_features: None,	26	47.64	26.89	313.00%

Modelos	Hiperparámetros	Rezagos	RMSE	MAE	MAPE
Decision Tree	max_depth: 3, min_samples_split: 7, min_samples_leaf: 5, max_features: sqrt	26	48.42	27.36	339.00%

Tabla 6

Métricas en validación externa para modelos de aguacate Hass

Modelo	RMSE	MAE	MAPE	Tamaño del intervalo
Decision Tree	942.43	829.07	15.50%	2 090.93
Random Forest	895.07	961.21	15.51%	1 634.43
Extra Trees	1 276.71	1 219.00	21.21%	1 384.68
Gradient Boosting	1 284.69	1 232.92	21.63%	5 622.98
Ada Boost	1 592.01	1 566.33	27.41%	3 434.50
SARIMA	4 464.87	3 157.73	58.01%	28 762.54

Los resultados muestran que los modelos de machine learning junto con el modelo SARIMA, presentan un mejor desempeño durante la fase de entrenamiento. De estos modelos, XGBoost destacó por su rendimiento en términos de RMSE de 38.36, lo que sugiere que de todos los modelos es el que obtiene la más baja desviación cuadrática en sus predicciones, con el segundo mejor MAE de 25.21. No obstante, su MAPE de 313% es más elevado que otros modelos, sugiriendo variabilidad relativa respecto a los valores reales. Los modelos como AdaBoost, GradientBoosting, Random Forest y Extra Trees, generan métricas como RMSE y MAE relativamente mayores a XGBoost, junto con errores porcentuales también elevados, lo cual indica inestabilidad relativa en las predicciones, a excepción de Extra Tress que, por el contrario, obtuvo un MAPE inferior del 236%.

De acuerdo con los resultados de la Tabla 6, es evidente que los modelos de machine learning tienen un mejor comportamiento para predicciones futuras. Esto ratifica que los modelos basados en árboles de decisión pueden ser efectivos para predicción de series de tiempo. Sin embargo, al relacionar los mejores modelos obtenidos durante el

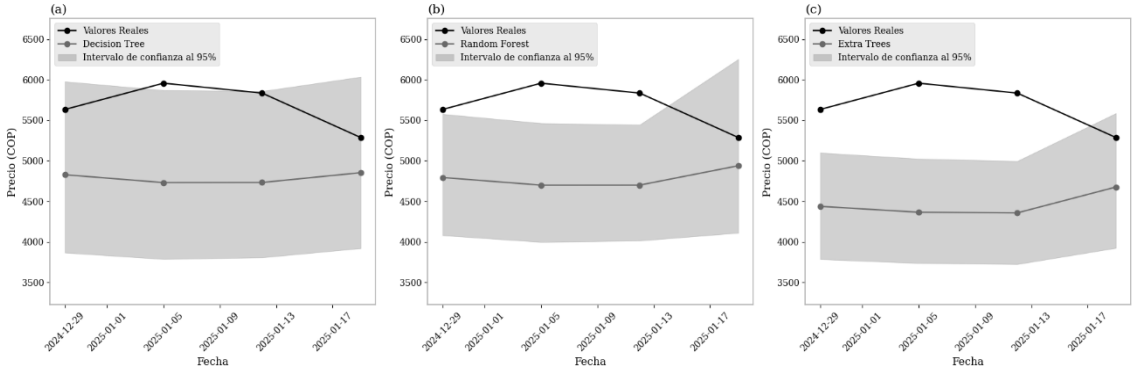
entrenamiento frente a los mejores modelos obtenidos durante la validación externa, se evidencia que XGBoost y SARIMA mostraron degradación en su precisión; por tanto, no se encuentran entre los mejores modelos finales (Apéndice A1). Se demostró que modelos como Decision Tree, Random Forest, y Extra Trees, optimizaron su desempeño, obteniendo métricas de error competitivas y tamaños de intervalos de predicción aceptables.

Dado lo anterior, se destaca el modelo Decision Tree, como se observa en la Figura 9, debido a que su intervalo de predicción es aceptable y abarca la mayoría de los valores reales. Además, sus métricas de error, como MAE y MAPE, son las más bajas entre los modelos evaluados.

Por otro lado, Random Forest se posiciona como un modelo equilibrado, con métricas de error competitivas y una incertidumbre relativamente inferior en sus predicciones. Esto refleja una mayor estabilidad, ya que la combinación de múltiples árboles de decisión contribuye a la reducción de la varianza, lo que se evidencia en menores fluctuaciones en las estimaciones. Extra Trees presentó el intervalo de predicción más estrecho. No obstante, sus métricas de error fueron significativamente más altas, con incrementos del 42.6% en RMSE, 26.8% en MAE y 36.7% en MAPE respecto a Random Forest, lo que evidencia una menor precisión en las estimaciones.

Figura 9.

(a) Comparación de las predicciones del modelo Decision tree para el aguacate Hass respecto a los valores reales. (b) Comparación de las predicciones del modelo Random Forest para el aguacate Hass respecto a los valores reales. (c) Comparación de las predicciones del modelo Exxtra trees para el aguacate Hass respecto a los valores reales.



Modelos como Gradient Boosting y AdaBoost, a pesar de ofrecer errores absolutos y porcentuales moderados, generaron intervalos de predicción excesivamente amplios, comprometiendo la utilidad práctica de sus resultados. La magnitud de estas incertidumbres limita su aplicabilidad para proyecciones confiables en el mercado del aguacate Hass, lo que dificulta la toma de decisiones fundamentadas.

En cuanto al modelo de referencia SARIMA, este fue competitivo durante el entrenamiento en términos absolutos (RMSE y MAE); sin embargo, obtuvo el MAPE más alto de 325.38%, lo que evidencia limitaciones para capturar la estructura de la serie. Esta deficiencia se acentúa aún más al evaluar sus predicciones, pues sus métricas de error son las más altas y su intervalo de predicción, con un valor de $28 762.54 COP, resulta poco confiable, llegando incluso a generar valores negativos.

Según los resultados, Decision Tree es el modelo que mejor equilibra tanto las métricas de error como la precisión de sus predicciones y la confiabilidad de sus intervalos.

5. ANÁLISIS Y CONCLUSIONES

En Colombia a marzo de 2025, el índice Precios del Consumidor (IPC) cerró en 5.09%, mostrando una tendencia a la baja en comparación a marzo 2024 y marzo 2023, donde el IPC de año corrido fue 13.34% y 7.36% respectivamente (Departamento Administrativo Nacional de Estadística – DANE, 2025a). Aunado a esto, para finales de abril de 2025, la junta directiva del Banco de la República decidió por unanimidad reducir en 25 puntos básicos la tasa de interés de la política monetaria, dada la tendencia decreciente de la inflación, así como el crecimiento de la actividad económica en el país y un momento de alta volatilidad e incertidumbre en el mercado internacional (Banco de la República, 2025).

Ahora, habiendo mencionado el crecimiento en la actividad económica del país, el PIB creció 1,7% en el 2024 respecto a 2023, donde la agricultura destacó como una de las actividades económicas que más contribuyeron a la dinámica del valor agregado (DANE, 2025b).

Por un lado, Colombia ha tenido un crecimiento notable en las exportaciones aguacate Hass, dado que para 2024 las exportaciones fueron de 138.315 toneladas, aumentando en 20.8% respecto al 2023 (ANALDEX, 2025), razón por la cual organizaciones como CorpoHass debaten estrategias para fortalecer la competitividad y expansión del sector (CorpoHass, s.f.; Portafolio, 2024).

Por otro lado, el aguacate Papelillo hace parte de las variedades de piel verde, criollos o antillanos, que para 2020 representaban el 66% de la siembra de aguacate en Colombia, según los datos de la Unidad de Planificación Rural Agropecuaria (UPRA) a 2023 representaban el 6.43%, y que actualmente no se exportan de manera representativa como la variedad Hass (El Colombiano, 2024a). Hay que mencionar, que estas variedades

se ven afectadas en Colombia, por los fenómenos de El niño y La niña, los cuales inciden al crear “baches” entre cosechas afectando la disponibilidad de la fruta y por ende la relación oferta demanda (El Colombiano, 2024b).

Dado lo anterior, se hace énfasis en las afectaciones climáticas que inciden en los tiempos de floración de la traviesa de aguacate como con el fenómeno de El Niño (ADAMA Colombia, s.f.) y se hace importante contar con políticas de sostenibilidad y seguridad alimentaria que vayan de la mano con pronósticos de precios y así poder tomar decisiones que permitan mejorar la competitividad de los agricultores.

Los modelos tradicionales de serie de tiempo, como el *ARIMA*, suponen relaciones lineales y frecuentemente falla al momento de capturar relaciones no lineales como ocurre en el precio de los productos agrícolas. En los últimos años el desarrollo de los modelos de machine learning, modelos no lineales como el *SVM, RF, TCN y XGBoost* otorgan predicciones más precisas y estables (Zhao et al., 2025). Es importante destacar que estos modelos tienen riesgo de sobreajuste y son sensibles a las transformaciones que tenga la serie de tiempo, siendo así que la fase en la que se ajustan los hiperparámetros se es crucial (Zhang et al., 2024). Lo anterior en concordancia con los resultados de este artículo, donde se observaron con mejores resultados los modelos Decision Tree, LightGBR, Random Forest y Extra Trees.

RECOMENDACIONES, TRABAJO FUTURO, ETC

Teniendo en cuenta los resultados obtenidos, se sugiere que futuras investigaciones contemplen la exploración de nuevos enfoques avanzados y de mayor complejidad, tales como el uso de redes neuronales profundas, modelos de ensamble con implementación de *stacking*, y redes LSTM (*Long Short-Term Memory*).

Asimismo, se recomienda evaluar la posibilidad de incorporar dentro de los modelos predictivos variables exógenas, incluyendo factores climáticos o datos referentes sobre la oferta y la demanda de las variedades de aguacate. La incorporación de estas variables y la implementación de nuevos modelos podría mejorar la precisión de las predicciones de los precios nacionales del aguacate en Colombia.

Referencias

ADAMA Colombia. (s.f.). El Niño y la agricultura en Colombia. https://www.adama.com/colombia/es/el-nino-y-la-agricultura-en-colombia

Adineh, A. H., Narimani, Z., & Satapathy, S. C. (2021). Importance of data preprocessing in time series prediction using SARIMA: A case study. International Journal of Knowledge-Based and Intelligent Engineering Systems, 24(4), 331–342. https://doi.org/10.3233/KES-200065

Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1907.10902

Amat Rodrigo, J., & Escobar Ortiz, J. (2025). skforecast (Version 0.16.0) [Computer software]. https://doi.org/10.5281/zenodo.8382788

Amazon. (s.f.). *¿Qué es el boosting?* Recuperado el 10 de mayo de 2025, de https://aws.amazon.com/es/what-is/boosting/

Analdex. (2025, 17 de febrero). *Informe exportaciones de aguacate Hass 2024*. https://analdex.org/2025/02/17/informe-exportaciones-de-aguacate-hass-2024/

Asociación Nacional de Comercio Exterior – ANALDEX. (2025, febrero 17). Informe de exportaciones de aguacate Hass 2024. https://analdex.org/2025/02/17/informe-exportaciones-de-aguacate-hass-2024/

Bartz, E., Bartz-Beielstein, T., Zaefferer, M., & Mersmann, O. (Eds.). (2022). Hyperparameter tuning for machine and deep learning with R: A practical guide. Springer.

Berzal F. (2018). Redes neuronales & Deep learning.

Banco de la República. (2025, abril 30). La Junta Directiva del Banco de la República reduce la tasa de interés en abril de 2025. https://www.banrep.gov.co/es/noticias/junta-directiva-abril-2025

Box, G. E. P., Jenkins, G. M., Reinsen, G. C., & Ljung, G. M. (2016). Time Series Analysis: Forecasting and Control (5th ed.).

Chen, Z., Goh, H. S., Sin, K. L., Lim, K., Chung, N. K. H., & Liew, X. Y. (2021). Automated Agriculture Commodity Price Prediction System with Machine Learning Techniques. Advances in Science, Technology and Engineering Systems Journal, 6(4), 376–384. https://doi.org/10.25046/aj060442

CorpoHass. (s.f.). Iniciativas. https://www.corpohass.com/iniciativas/

Departamento Administrativo Nacional de Estadística – DANE. (2025a). Índice de Precios al Consumidor (IPC). https://sitios.dane.gov.co/ipc/visorIPC/#!/

Departamento Administrativo Nacional de Estadística – DANE. (2025b). Producto Interno Bruto trimestral – Información técnica. https://www.dane.gov.co/index.php/estadisticas-por-tema/cuentas-nacionales/cuentas-nacionales-trimestrales/pib-informacion-tecnica

El Colombiano. (2024a, marzo 1). ¿Por qué el aguacate está caro en Colombia? https://www.elcolombiano.com/negocios/agro/por-que-el-aguacate-esta-caro-en-colombia-BC24151176

El Colombiano. (2024b, abril 15). Empleos verdes en Colombia: Formación sostenible. https://www.elcolombiano.com/empleos/contenidos/empleos-verdes-colombia-formacion-sostenible-NM27092539

Gómez Arango, M. A., & Diaz Valencia, E. (2024). Predicción del rendimiento de cultivos agrícolas en los cinco corregimientos de la ciudad de Medellín, utilizando modelos de Machine Learning.

Gujarati, D., & Down, P. (2009). Econometría (5th ed.).

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: with applications in R (2nd ed.). Springer.

Mariño Villalba, J. A. (2023). Una comparación entre modelos estadísticos y de Machine Learning para la predicción de series de tiempo multivariadas (Informe de investigación). Universidad Nacional de Colombia, Facultad de Ciencias, Departamento de Estadística. Bogotá, Colombia

Ministerio de Agricultura y Desarrollo Rural. (2020). *A*nálisis de afectación de las cadenas productivas agropecuarias derivada de la emergencia sanitaria por COVID-19. https://sioc.minagricultura.gov.co/Boletines/Forms/AllItems.aspx

Nayak, A. K., Sharma, K. C., Bhakar, R., & Tiwari, H. (2025). Probabilistic online learning framework for short-term wind power forecasting using ensemble bagging regression model. Energy Conversion and Management, 323(A), 115-130. https://doi.org/10.1016/j.enconman.2024.115130

Özden, C. (2023). Comparative Analysis of CNN, LSTM And Random Forest for Multivariate Agricultural Price Forecasting. Black Sea Journal of Agriculture, 6(4), 422–426. https://doi.org/10.47115/bsagriculture.1304625

Portafolio. (2024, marzo 27). Colombia impulsa el crecimiento del aguacate Hass: Exportaciones aumentan un 15,7 % en 2024. https://www.portafolio.co/economia/agro/colombia-impulsa-el-crecimiento-del-aguacate-hass-exportaciones-aumentan-un-15-7-en-2024-618559

Ramírez Castañeda, L. N., Cristancho Cruz, S. L., & Cleves Leguizamo, J. A. (2021). Modelos de Fluctuaciones de Precios Agricolas - Estudio Comparativo de Frutas Tropicales en Colombia. Revista de Ciencias Sociales, XXVII(Especial 4), 197–212. https://dialnet.unirioja.es/servlet/articulo?codigo=8145517

Rana, H., Farooq, M. U., Kazi, A. K., Baig, M. A., & Akhtar, M. A. (2024). Prediction of Agricultural Commodity Prices using Big Data Framework. Engineering, Technology & Applied Science Research, 14(1), 12652–12658. https://doi.org/10.48084/etasr.6468

Ruiz Hernández, J. A., Barrios Puente, G., & Gómez Gómez, A. A. (2019). Análisis del precio de la manzana mediante un modelo SARIMA. Revista Mexicana de Ciencias Agrícolas, 10(2), 225–237.

Sal, M. (2024). Predicción del precio de azúcar en Argentina a partir de modelos de Machine Learning. https://repositorio.utdt.edu/handle/20.500.13098/12972

scikit-learn developers. (s.f.). *Linear models — scikit-learn 1.3.0 documentation*. Recuperado el 10 de mayo de 2025, de https://scikit-learn.org/stable/modules/linear_model.html

Sun, F., Meng, X., Zhang, Y., Wang, Y., Jiang, H., & Liu, P. (2023). Agricultural Product Price Forecasting Methods: A Review. In Agriculture (Switzerland) (Vol. 13, Issue 9). Multidisciplinary Digital Publishing Institute (MDPI). https://doi.org/10.3390/agriculture13091671

Torre-Torres, O. V. D. l., del Río-Rama, M. d. l. C., & Álvarez-García, J. (2024). Non-commodity agricultural price hedging with minimum tracking error portfolios: The case of Mexican Hass avocado. Agriculture, 14(10), 1692. https://doi.org/10.3390/agriculture14101692

Unidad de Planificación Rural Agropecuaria (UPRA). (2025). Visor geográfico del sector agropecuario: Aguacate. https://upra.gov.co/

Yu, X., Liu, B., & Lai, Y. (2024). Monthly Pork Price Prediction Applying Projection Pursuit Regression: Modeling, Empirical Research, Comparison, and Sustainability Implications. Sustainability (Switzerland), 16(4). https://doi.org/10.3390/su1604146

Zdansky, J. (2006). BINSEG: an efficient speaker-based segmentation technique. *Interspeech 2006*, paper 1459-Thu1A1O.2-0. https://doi.org/10.21437/Interspeech.2006-567

Zhang, N., An, Q., Zhang, S., & Ma, H. (2024). Price prediction for fresh agricultural products based on a boosting ensemble algorithm. Mathematics, 13(1), 71. https://doi.org/10.3390/math13010071

Zhao T, Chen G, Suraphee S, Phoophiwfa T, Busababodhin P. A hybrid TCN-XGBoost model for agricultural product market price forecasting. PLoS One. 2025 May 2;20(5):e0322496. doi: 10.1371/journal.pone.0322496. PMID: 40315255; PMCID: PMC12047798.

APÉNDICES

Apéndice A. Modelos descartados

- **Aguacate papelillo**

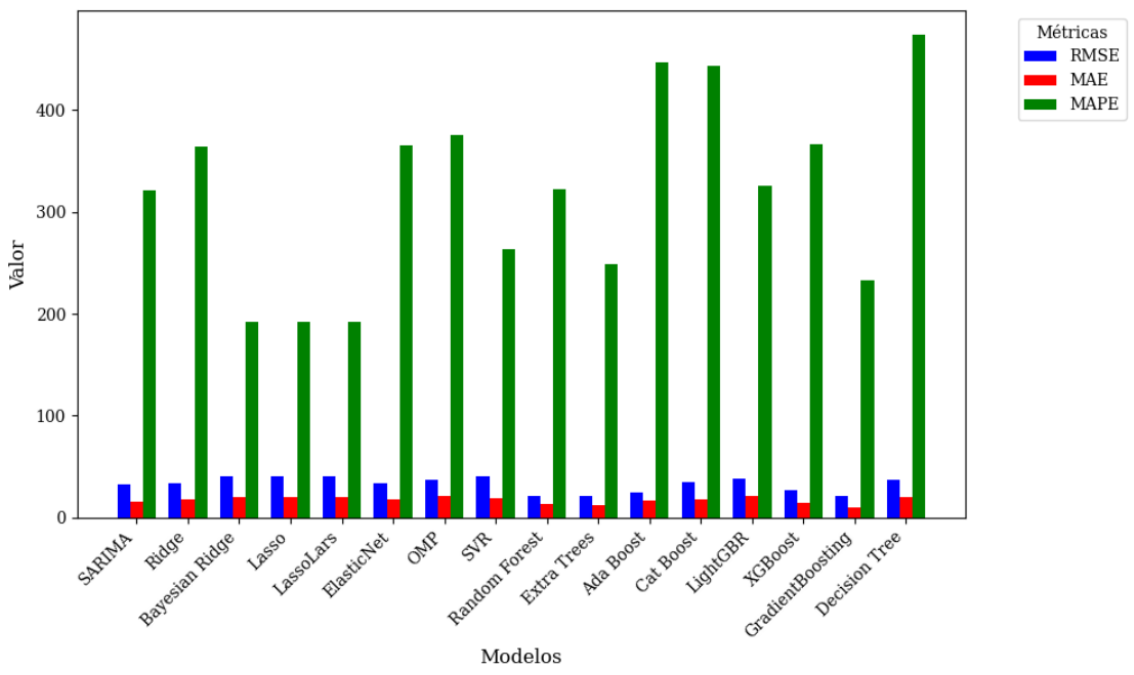
Los modelos que se descartaron para aguacate papelillo, que se observan en la Figura A1, fueron los modelos de regresión regularizada: *Lasso*, *LassoLars*, *ElasticNet*, *Ridge* y *Bayesian Ridge*, que presentaron resultados muy similares entre sí, posiblemente convergiendo a lo mismo dada su naturaleza lineal. Los modelos Ridge y ElasticNet presentaron métricas iguales, con un RMSE de 33.63 entrenamiento, mientras que Bayesian Ridge, Lasso y Lasso LARS, tuvieron una métrica de 40.4. Por parte de la validación externa, entre ellos hubo un RMSE que osciló entre 1 631.88 y 4 419.03.

A pesar de que ElasticNet y Ridge destacaron entre los modelos mejores desempeñados durante el entrenamiento, sus métricas se degradaron durante la validación externa, sumado a sus intervalos de confianza considerablemente más altos respecto a los árboles. Igual situación se presentó con modelos como XGBoost y GradientBoosting. Otros modelos se descartaron por su desempeño ampliamente peor que los seleccionados, tales como OMP, AdaBoost y CatBoost.

Por su parte, el modelo SARIMA, que presentó un buen desempeño tanto en entrenamiento como en validación externa, fue finalmente descartado por su intervalo de confianza excesivamente más amplio que todos los demás modelos (alcanzando incluso valores negativos en su límite inferior), generando predicciones que, si bien presentan buenas métricas, no poseen una confianza significativa.

Figura A1

Comparación de métricas de modelos de predicción de precios del aguacate Hass



- **Aguacate Hass**

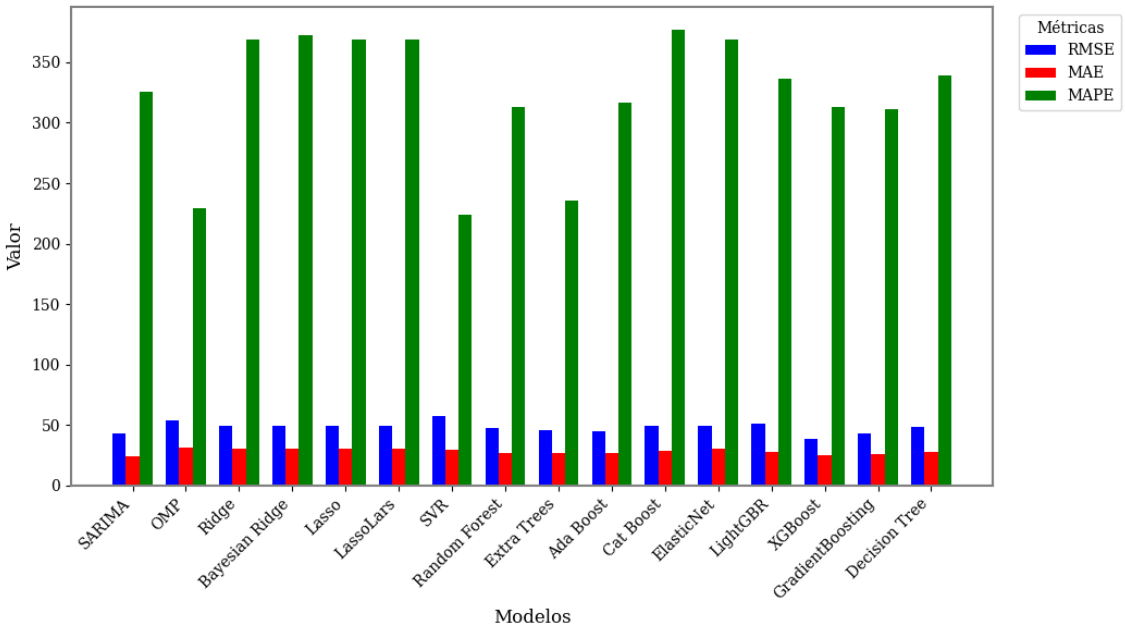
Los modelos descartados para aguacate Hass, que se observan en la Figura A2, fueron los modelos de regresión regularizada como *Lasso*, *LassoLars*, *ElasticNet*, *Ridge* y *Bayesian Ridge* que mostraron resultados similares entre sí y no destacaron frente a los demás modelos, probablemente debido a su naturaleza lineal y similitud en sus hiperparámetros, con regularizaciones bastante débiles. Obteniendo resultados durante la validación externa de cuatro pasos adelante con un RMSE entre 4 061.69 y 4 151.01, un MAE entre 2 461.44 y 2 505.61 y un MAPE entre 45.66% y 46.51%.

Los otros modelos descartados como *SVR*, y OMP tuvieron métricas de error menores a los métodos de regresión regularizada, en promedio un RMSE de 2251.84, MAE de 2237.20 y MAPE de 39.37% y el tamaño del intervalo de predicción fue el más pequeño, lo que puede indicar subestimación de incertidumbre. Por su parte Cat Boost, aunque generó métricas de error similares a las anteriores su intervalo de predicción fue el mayor con un valor de $13 825.65 COP.

XGBoost y SARIMA, si bien, presentaron resultados óptimos durante la fase de entrenamiento, sus métricas de error disminuyeron obteniendo valores de RMSE de 1 942.06 y 4 464.87, MAE de 1 879.66 y 3157.73 y MAPE de 32.38% y 58.01% respectivamente. No obstante, su mayor dificultad se encontró en los intervalos de predicción generados de 5 957.11 para XGBoost y 28 762.54 para SARIMA.

Figura A2

Comparación de métricas de modelos de predicción de precios del aguacate Hass

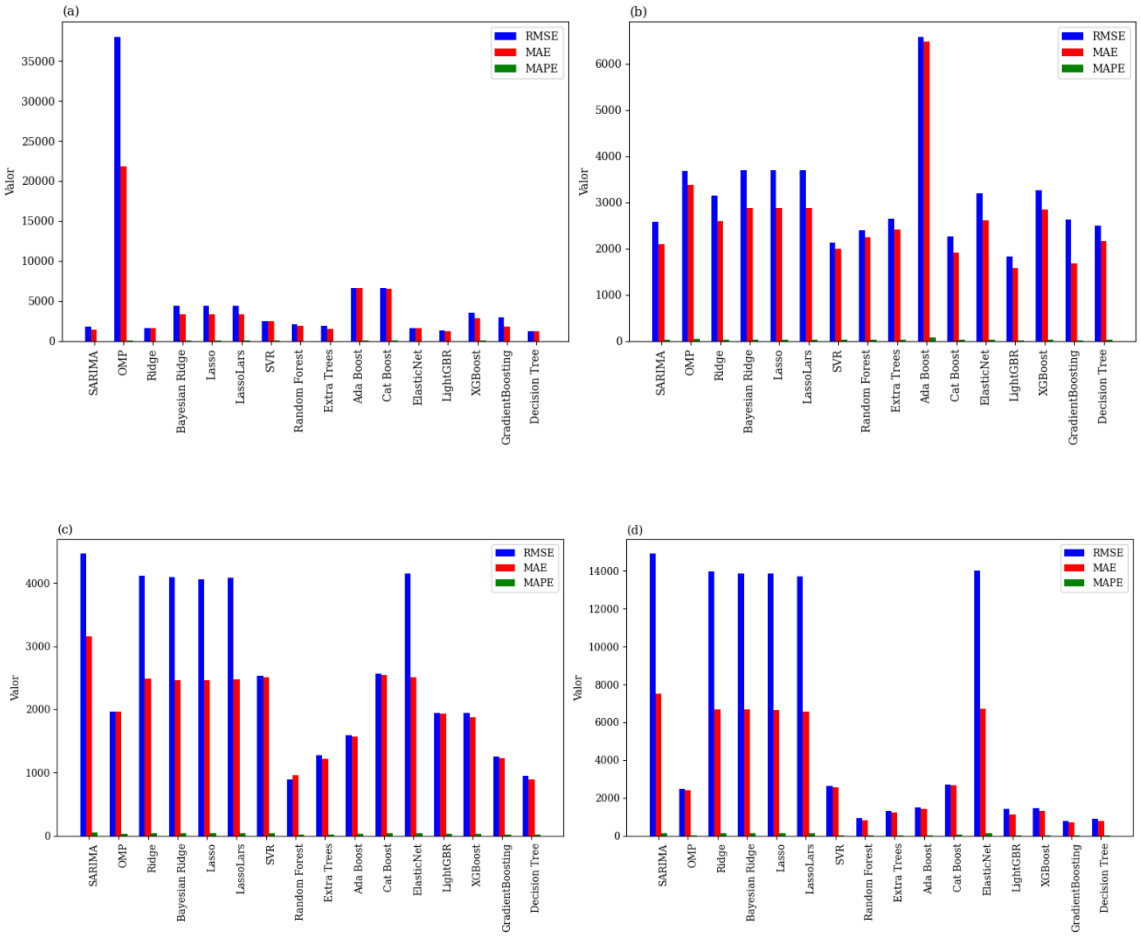


Apéndice B. Horizonte de predicción

Con los modelos entrenados, se generaron predicciones semanales del precio del aguacate papelillo y Hass para dos horizontes de pronóstico: cuatro y ocho pasos hacia adelante, correspondientes a las primeras semanas de 2025. La evaluación de la capacidad predictiva de los modelos se realizó sobre la serie revertida a su escala original (pesos colombianos), utilizando métricas de error (RMSE, MAE, MAPE), por lo que se compararon las métricas en ambos horizontes para determinar cuál presentaba mejor predictibilidad. En la Figura B1 se muestran los resultados de las métricas para los dos horizontes de predicción.

Figura B1

(a) Métricas de error con horizonte de predicción de cuatro periodos en aguacate papelillo (b) Métricas de error con horizonte de predicción de ocho periodos en aguacate papelillo (c) Métricas de error con horizonte de predicción de cuatro periodos en aguacate Hass (d) Métricas de error con horizonte de predicción de ocho periodos en aguacate Hass



Si bien algunos modelos presentan mejor predicción en un horizonte de cuatro pasos y otros en el de ocho (resultando un 50/50), en la evaluación de aquellos que se mostraron como los mejores modelos resalta que cuatro es siempre el mejor horizonte y sus intervalos de confianza son más conservadores. Teniendo en cuenta lo anterior, se definió un horizonte de predicción de cuatro periodos semanales, con el objetivo de no perder predictibilidad en los modelos.