Pontificia Universidad
JAVERIANA
Cali

# PREDICCIÓN DE PRECIOS DE ACTIVOS DE RENTA VARIABLE DE LA BVC MEDIANTE MODELOS DE APRENDIZAJE SUPERVISADO

*Genjis Alberto Ossa González*
*Código: 9013889*
*Paolo Andrés de la Hoz Vicari*
*Código: 9015063*

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

Director(a)

Mario Julián Mora Cardona

FACULTAD DE INGENIERÍA Y CIENCIAS
MAESTRÍA EN CIENCIA DE DATOS
SANTIAGO DE CALI, NOVIEMBRE 26 DE 2025

## RESUMEN

El presente trabajo desarrolla un modelo integral para la predicción de retornos logarítmicos y precios de activos de renta variable de la Bolsa de Valores de Colombia (BVC) empleando metodologías de aprendizaje supervisado. Dado que los mercados financieros operan en un entorno marcado por la alta incertidumbre, la volatilidad y la presencia de dinámicas no lineales, anticipar correctamente el comportamiento de los activos continúa siendo un desafío central tanto para investigadores como para gestores de inversión. Con este propósito, se construyó un pipeline para el tratamiento y modelado de series temporales financieras, que abarcó la depuración de datos, el enriquecimiento mediante indicadores de calidad, volumen, microestructura de mercado, volatilidad e indicadores técnicos, y la posterior implementación de modelos predictivos.

La investigación compara el desempeño de modelos clásicos y avanzados, incluyendo dos baselines el Naive0 y Persist, algoritmos basados en árboles de decisión (Random Forest, XGBoost y LightGBM) y redes neuronales recurrentes tipo LSTM, diseñadas para capturar dependencias temporales de largo plazo. Todos los modelos fueron evaluados bajo un esquema de partición temporal estricta (TRAIN–VALID–TEST), evitando la fuga de información y asegurando una medición realista del desempeño fuera de muestra.

En cuanto a la predicción de retornos, los resultados muestran que la naturaleza altamente ruidosa y volátil del mercado dificulta la obtención de patrones estables y consistentes a lo largo del tiempo. En este escenario, los modelos basados en boosting, especialmente LightGBM, lograron mejoras moderadas respecto a los baselines en términos de MAE, RMSE y tasas de acierto direccional. Asimismo, las redes LSTM demostraron una capacidad superior para capturar señales direccionales, aun cuando sus métricas de error no siempre superaron a los modelos más simples.

Bajo esta estructura, el modelo Naive0 resultó ser el más efectivo, al obtener los menores valores de MAE, RMSE y MAPE, junto con los mayores niveles de $R^2$ en todos los emisores. Esto indica que, para pronosticar el nivel absoluto del precio, la mejor aproximación consiste en asumir que el precio futuro será similar al del periodo anterior, lo que explica por qué los modelos complejos no lograron superarlo de manera consistente. El segundo mejor desempeño correspondió a LightGBM, que mostró una mayor estabilidad y generalización que el modelo Persist y que otros algoritmos supervisados, posicionándose como la alternativa más robusta entre los modelos de aprendizaje automático más avanzados.

# TABLA DE CONTENIDO

## LISTA DE ILUSTRACIONES

## LISTA DE FIGURAS

## LISTA DE TABLAS

## LISTA DE ANEXOS

# INTRODUCCIÓN

La predicción de precios de activos financieros es un desafío clave en el contexto del mercado bursátil, debido a su alta volatilidad y comportamiento no lineal – lo cual dificulta en cierta medida el análisis del mismo bajo aplicaciones estadísticas clásicas. En particular, la Bolsa de Valores de Colombia (BVC) representa un contexto donde las fluctuaciones[1] en los precios de activos de renta variable (o no variable) son fundamentales para la toma de decisiones de inversión. Este proyecto busca abordar esta problemática mediante la implementación de técnicas avanzadas de aprendizaje supervisado para mejorar la precisión en la predicción de precios, facilitando estrategias de inversión en función del pronóstico.

En este contexto, la evolución de los métodos estadísticos y computacionales ha impulsado el desarrollo de modelos predictivos más robustos, capaces de capturar tendencias, episodios de volatilidad y patrones estocásticos característicos de los mercados financieros. Entre ellos, los algoritmos de machine learning como Random Forest, XGBoost y las redes neuronales han adquirido un protagonismo creciente en el campo de las finanzas cuantitativas, gracias a su capacidad de procesar grandes volúmenes de información histórica y generar aproximaciones más precisas sobre el comportamiento futuro de los activos.

El presente trabajo tiene como objetivo aplicar y evaluar diferentes metodologías de aprendizaje automático para la estimación de retornos a partir de la información histórica de cinco activos financieros cotizados en la BVC. Adicionalmente, se propone el desarrollo de una herramienta de visualización interactiva que permita a los usuarios explorar las proyecciones de precios y comprender el comportamiento de cada activo, facilitando la implementación de estrategias de diversificación de inversión. Como resultados esperados se contemplan: un conjunto de datos debidamente procesado, modelos predictivos ajustados, un prototipo funcional de visualización y un informe técnico detallado que documente el proceso y sus hallazgos principales.

[1] Entiéndase como volatilidad o en términos formales la desviación estándar.

# 1. DEFINICIÓN DEL PROBLEMA

## 1.1. PLANTEAMIENTO DEL PROBLEMA

En el ámbito de las finanzas y las inversiones, la predicción de precios de activos de renta variable se ha convertido en un tema de gran relevancia para los analistas y gestores de portafolio. Los inversores, al depender de la precisión en la estimación de precios futuros a través del análisis fundamental o técnico buscan mitigar el riesgo y maximizar los rendimientos, lo cual es especialmente complejo debido a la alta volatilidad y fluctuaciones inherentes a los mercados financieros. La toma de decisiones estratégicas en la inversión de activos de renta variable se enfrenta constantemente a la incertidumbre del mercado, es decir al riesgo, que ciertamente después del año 2020 el término "*risk finance*" o el "*variable income*" han tenido un creciente interés en la web [34] que muy seguramente surge a partir de factores diversos afectando los precios, por tal motivo, se hace necesaria una metodología que permita adelantarse con mejor precisión a la predicción de estos valores con base a su recorrido histórico.

Y es por lo anterior y tal cual como lo argumenta [35] el mercado financiero adopta indirectamente una disciplina como la ciencia de datos, que, desde la extracción de los datos y la formulación de hipótesis intentan anticiparse en el tiempo soportándose en métodos mucho más robustos y potentes que la modelación tradicional. Generalmente los modelos estadísticos lineales han sido utilizados para prever los movimientos de precios, sin embargo, estas metodologías presentan limitaciones significativas cuando se enfrentan a patrones no lineales y a interacciones con desviaciones considerables que rompen con los supuestos de la linealidad. Además, la velocidad a la que se generan y cambian los datos en los mercados financieros presenta un reto adicional: los modelos convencionales suelen carecer de la flexibilidad necesaria para adaptarse rápidamente a estos cambios.

Lo anterior puede reflejarse en predicciones menos exactas y, por lo tanto, en una mayor exposición al riesgo para los inversores. Con el avance de las técnicas de aprendizaje supervisado, es viable abordar este problema mediante el uso de modelos de Machine Learning, que pueden aprender de grandes volúmenes de datos y que ciertamente lo hace un tema aplicable a la ciencia de datos dado que cuenta con muy buena historicidad para poder detectar patrones complejos que escapan a las técnicas tradicionales.

El problema que se pretende abordar en este proyecto es la falta de precisión y adaptabilidad en las predicciones de retornos de precios de activos de renta variable. A través de la implementación de modelos de aprendizaje supervisado, este proyecto busca mejorar la exactitud de las predicciones no lineales, optimizando así la toma de decisiones en estrategias de inversión y permitiendo una mejor gestión del riesgo y del rendimiento.

### 1.2. FORMULACIÓN DEL PROBLEMA

¿Cómo desarrollar modelos de aprendizaje supervisado para predecir precios de activos de renta variable de la BVC?

### 1.3. SISTEMATIZACIÓN DEL PROBLEMA

- ¿Cómo seleccionar, preprocesar y preparar los datos históricos financieros para garantizar su calidad y relevancia en la construcción de modelos de aprendizaje supervisado?
- ¿Cuáles modelos de aprendizaje supervisado son los más adecuados para la predicción de precios en función del tipo de activo y su comportamiento en el mercado?
- ¿Cómo entrenar modelos de aprendizaje supervisado a partir de datos, ajustando sus configuraciones y optimizando los hiperparámetros para mejorar la precisión?
- ¿De qué manera evaluar la efectividad de los modelos mediante métricas de rendimiento que garanticen la precisión y capacidad de generalización en distintos entornos de mercado?
- ¿Cómo diseñar un sistema de visualización que permita a los usuarios analizar las proyecciones de precios y aplicar estrategias de diversificación de inversión?

# 2. OBJETIVOS DEL PROYECTO

## 2.1. OBJETIVO GENERAL

Desarrollar modelos de aprendizaje supervisado para predecir los precios de activos de renta variable, optimizando la precisión de las proyecciones para mejorar las estrategias de inversión.

## 2.2. OBJETIVOS ESPECÍFICOS

1. Identificar los activos a modelar mediante el análisis de variables como volatilidad y rentabilidad.
2. Implementar y entrenar modelos de aprendizaje supervisado, ajustando cada modelo para la predicción de precios de acuerdo con el tipo de activo.
3. Evaluar la efectividad de los modelos predictivos mediante métricas de rendimiento, midiendo la precisión y capacidad de generalización en la predicción de precios.
4. Elaborar visualizaciones que permitan analizar el comportamiento de cada activo financiero de acuerdo con su mejor modelo

# 3. MARCO TEÓRICO Y ANTECEDENTES

## 3.1. MARCO TEÓRICO

El presente marco teórico aborda los fundamentos conceptuales y técnicos necesarios para sustentar el desarrollo del proyecto. Se estructura en torno a los principios del aprendizaje automático, los modelos clásicos de series temporales, y las técnicas modernas de machine learning aplicadas al análisis financiero.

### *3.1.1 Mercado de Renta Variable*

El mercado de renta variable es el espacio donde compradores y vendedores realizan transacciones de acciones y otros títulos representativos de participación, operaciones que se concretan por medio de los comisionistas de bolsa, quienes actúan como intermediarios autorizados [32]. En Colombia, su principal escenario es la Bolsa de Valores de Colombia (BVC), una empresa privada listada en el propio mercado de valores. La BVC administra plataformas de negociación para Renta Variable, Renta Fija y Derivados Estandarizados, y, mediante filiales creadas a través de alianzas estratégicas con otras compañías, opera además mercados de commodities [31].

Estos productos anteriormente mencionados son sujetos de inversión, por lo cual siempre se hace imperante la necesidad de protegerse contra el riesgo y ciertamente adelante a la rentabilidad. Por lo anterior, y dándole profundidad a este marco teórico. Una acción o producto de renta variable se define como inversiones en las que los rendimientos no son fijos y pueden variar, como las acciones [33]. No obstante, muchas de las acciones de empresas colombianas también pueden estar indexadas o negociadas de forma simultánea en otras plataformas internacionales, lo que permite una mayor integración y visibilidad en los mercados globales.

### *3.1.2 Series temporales*

Una serie de tiempo se define como una secuencia de observaciones de una variable ordenadas cronológicamente y registradas a intervalos regulares o irregulares en el tiempo [43]. El análisis de series temporales tiene como objetivo describir, modelar y predecir la evolución temporal de un fenómeno, teniendo en cuenta la dependencia existente entre observaciones consecutivas [44]. En el ámbito financiero, las series temporales representan la evolución dinámica de variables como precios de activos, retornos, tasas de interés, volatilidad y volúmenes de negociación. Estas series constituyen un objeto de estudio central en las finanzas cuantitativas,

debido a que los mercados financieros son sistemas dinámicos, complejos y sujetos a alta incertidumbre [41]. Una característica distintiva de las series temporales financieras es su comportamiento estocástico [45], en particular, los precios de los activos suelen presentar no estacionariedad, mientras que los retornos financieros tienden a ser estacionarios en media, aunque exhiben propiedades empíricas bien documentadas como volatilidad agrupada, heterocedasticidad condicional, asimetría y colas pesadas, lo que implica desviaciones significativas respecto a la distribución normal [46].

### *3.1.3 Aprendizaje automático (Machine Learning)*

El aprendizaje automático (Machine Learning, ML) es un subcampo de la inteligencia artificial que se centra en desarrollar algoritmos capaces de aprender patrones en datos masivos para realizar predicciones o tomar decisiones automáticas [1] [2]. En este proyecto, se busca utilizar ML para abordar el problema del pronóstico de series temporales financieras, una tarea que implica lidiar con datos altamente volátiles, no estacionarios y de naturaleza estocástica.

En relación con [3] las técnicas de Machine Learning comprenden un gran número de algoritmos que, basados en diversos modelos matemáticos, permiten encontrar patrones en los datos; dichos algoritmos pue den ser clasificados en tres grupos principales a partir de la forma en que estos aprenden: aprendizaje supervisado, no supervisado y por refuerzo, en este caso nos atañe solamente el supervisado.

De acuerdo con [10] el mercado de renta variable se caracteriza por el comportamiento volátil que tienen el precio de las acciones y que se ajusta a las normas que riesgo el mercado en cada momento. Dicho comportamiento volátil es de naturaleza no lineal, algo que dificulta mucho el proceso de análisis y estimación del precio futuro. De hecho, una de las características para determinar la rentabilidad diaria no obedece al tipo de variación relativa de:

$$\Delta = \frac{P_t}{P_{t-1}} - 1 \tag{1}$$

Si no más bien, que, dado la volatilidad de los movimientos, las diferencias pueden ser significativas, y los rendimientos logarítmicos captan con mejor sensibilidad dicha variación.

$$Ln_{\Delta} = Ln\left(\frac{P_t}{P_{t-1}}\right) \tag{2}$$

Ahora bien, existen modelos estadísticos base para diseñados específicamente para series temporales el pronóstico de series con financieras, se podría comenzar con Modelos clásicos. Por

otro lado, para el cálculo de la volatilidad de un activo financiero suele emplearse la desviación estándar:

$$\sigma = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} \tag{2.1}$$

#### *3.1.1.1 Modelos base*

En el contexto del modelado predictivo, los modelos baseline o modelos de referencia son el punto de partida más simple para la evaluación comparativa del desempeño de algoritmos más sofisticados. Su propósito es establecer un nivel mínimo de rendimiento que un modelo debe superar para considerarse útil o con capacidad de aprendizaje. Los modelos baseline son esenciales para validar que las mejoras obtenidas por modelos más complejos no se deben al azar, sino a una verdadera captura de patrones en los datos.

- **Naive:**

El modelo Naive asume que el retorno esperado en el siguiente período es igual a cero $\Delta_{t+1} = 0$, En términos de precios, esta hipótesis equivale a considerar que el precio futuro será igual al precio actual, es decir $P_{t+1} = P_t$. Esta dirección permite determinar si los modelos predictivos más avanzados logran capturar alguna estructura temporal o si simplemente replican un comportamiento aleatorio.

- **Persist:**

El modelo Persist parte de la premisa de que el retorno observado en el último período se repetirá en el siguiente. Formalmente, el pronóstico se define $Ln_\Delta = \Delta$. donde $Ln_\Delta$ representa el retorno logarítmico del período $\Delta$. Este modelo se fundamenta en la inercia temporal de las series financieras, asumiendo que el comportamiento reciente es un buen indicador de la dinámica inmediata.

#### *3.1.1.2 Modelos de Machine Learning.*

El aprendizaje supervisado consiste en desarrollar algoritmos capaces de generalizar a partir de datos etiquetados, asignando con precisión entradas a salidas para realizar predicciones sobre nuevas instancias. Este proceso incluye una fase de entrenamiento, donde el modelo aprende la relación entre pares de entrada y salida optimizando una métrica específica, y una fase de prueba, en la que se evalúa su desempeño utilizando datos no vistos para medir su capacidad de generalización.

- **Gradient Boosting Machines (GBMs):**

Las Gradient Boosting Machines son modelos de aprendizaje supervisado ampliamente utilizados en tareas de predicción como regresión y clasificación [29]. Funcionan combinando modelos débiles en un modelo fuerte mediante un enfoque iterativo que minimiza los errores de predicción, destacándose por su capacidad para capturar relaciones no lineales y manejar datos categóricos o continuos con poco preprocesamiento.

- **LightGBM con Validación Walk-Forward (WFV):**

El modelo principal implementado en este estudio corresponde a un Gradient Boosted Decision Trees (GBDT), desarrollado mediante la librería LightGBM. Este algoritmo pertenece a la familia de los métodos de boosting [2]de árboles, los cuales construyen secuencialmente múltiples modelos débiles (árboles de decisión) que se combinan para formar un modelo predictivo más robusto y preciso. El objetivo del modelo es predecir el retorno logarítmico futuro $Ln_{\Delta+1}$ de cada activo financiero a partir de un conjunto de indicadores derivados en etapas previas (variables transformadas MINMAX SAFE).

Posteriormente, el retorno estimado se transforma en precio predicho mediante la relación exponencial:

$$\widehat{P_{t+1}} = P_t[e^{\widehat{Ln_{\Delta+1}}}] \tag{3}$$

Es decir, que el precio futuro será igual al precio actual multiplicado por el factor de crecimiento exponencial del retorno logarítmico esperado.

- **Extreme Gradient Boosting (XGBoost):**

XGBoost es una implementación avanzada de Gradient Boosted Machines (GBMs) basada en árboles de decisión. Este modelo destaca por incorporar regularización para mejorar la generalización y minimizar problemas como el sobreajuste [17]. Su optimización se basa en una función objetivo que combina una función de pérdida con un término de regularización, con el fin de controlar la complejidad del modelo y obtener un equilibrio entre precisión y generalización en tareas de regresión y clasificación.

El algoritmo de acuerdo con [18] emplea un ensamblado secuencial de árboles de decisión conocidos como CART (Classification and Regression Trees). Cada árbol se construye para corregir

[2] Técnica de aprendizaje automático que combina múltiples modelos simples (aprendices débiles) para crear uno solo, más fuerte y preciso.

los errores del anterior, utilizando un enfoque de gradiente descendente hasta que los errores no puedan reducirse más.

De acuerdo con [19] algoritmo XG Boost tiene las siguientes características:

a. Se obtiene un árbol inicial $F_0$ para predecir la variable objetivo $\hat{Y}_i$, el resultado se asocia con un residual $\left(\hat{Y}_i - F_0\right)$
b. Se obtiene un nuevo árbol $h_1$ que ajusta al error del paso previo.
c. Los resultados de $F_0$ y $h_1$ se combinan para obtener el árbol $F_1$, donde el error cuadrático medio de $F_1$ será menor que el de $F_0$:

$$F_1(x) < -F_0(x) + h_1(x) \tag{4}$$

d. Este proceso se sigue iterativamente hasta que el error es minimizado lo más posible de la siguiente forma:

$$F_m(x) < - F_{m-1}(x) + h_m(x) \tag{4.1}$$

- **Random Forest:**

Es un algoritmo de aprendizaje de conjuntos utilizado para clasificación y regresión debido a su alta precisión y robustez frente al sobreajuste [4]. Este algoritmo combina múltiples árboles de decisión para mejorar la precisión de la predicción. Se ha demostrado que supera a otros modelos, como las máquinas de vectores de soporte y las redes neuronales, en diversas aplicaciones [5]. Para una entrada $x$, el resultado de Random Forest se calcula combinando las predicciones individuales de $M$ árboles:

$$\hat{Y}_i = \frac{1}{M}\sum_{m=1}^{M} f_j(x_i) \tag{5}$$

Ahora bien, los árboles de decisión utilizan un diagrama de flujo como una estructura de árbol para mostrar las predicciones que resultan de una serie de divisiones basadas en características. Comienza con un nodo raíz y termina con una decisión tomada por las hojas.

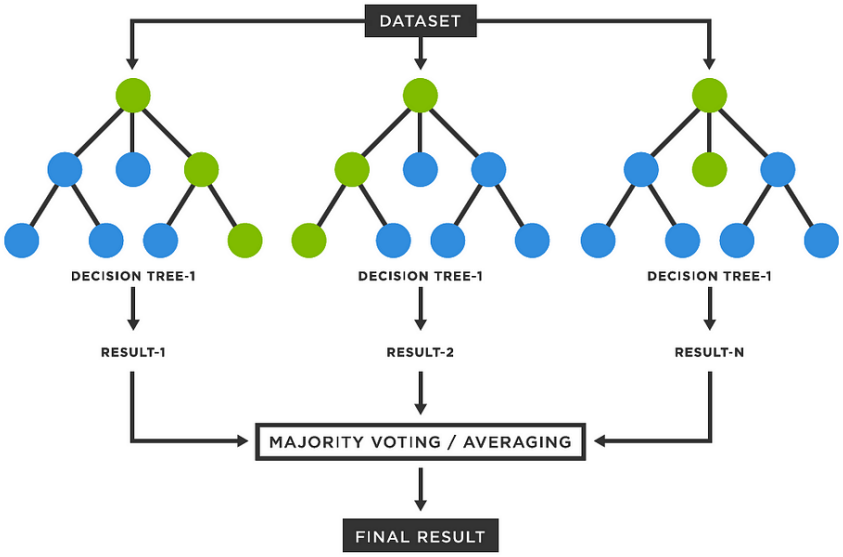


**Ilustración 1:** Estructura bosque aleatorio.

Y para evaluar necesitamos la impureza de nuestro conjunto de datos y tomaremos esa característica como el nodo raíz que proporciona la impureza más baja o, por ejemplo, cuál tiene el índice de Gini más bajo. Matemáticamente, el índice de Gini se puede escribir como:

$$\mathrm{G}_t = \sum_{c=1}^{C} P_C(1 - P_C) \tag{5.1}$$

La función objetivo para seleccionar la mejor división en un nodo $t$ se basa en maximizar la ganancia de información o minimizar dicha métrica.

#### *3.1.1.3 Redes neuronales*

Las redes neuronales, específicamente las redes neuronales artificiales (ANN), son modelos computacionales inspirados en la estructura y función del cerebro humano. Consisten en neuronas interconectadas que procesan información en paralelo, lo que las hace adecuadas para tareas complejas de resolución de problemas [6]. De acuerdo con [20] la RNN función con que la capa de entrada recibe cada uno de los elementos del vector de entrada $x_t$ y los transmite a la primera capa oculta. Las capas ocultas calculan sus valores de salida o señales, y las transmiten como vector de entrada a la siguiente capa.

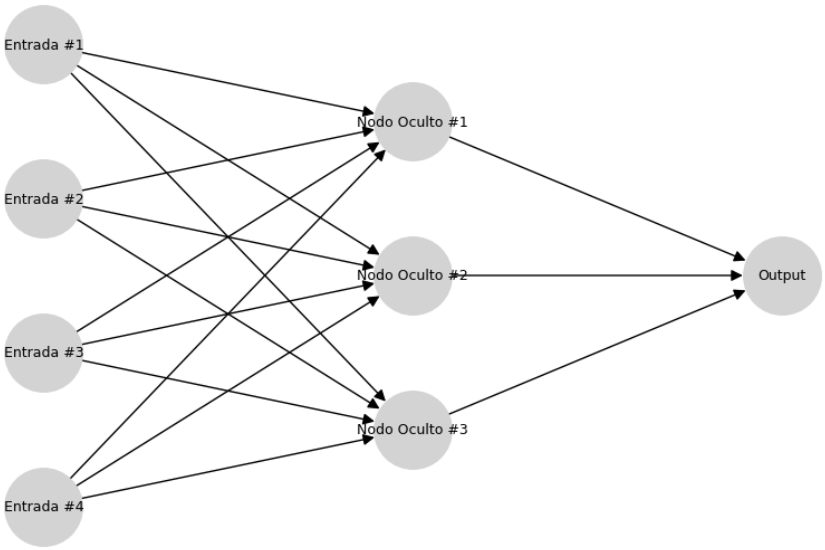


**Ilustración 2:** Representación de una red neuronal artificial.

En cuanto a los tipos de redes neuronales se tienen las siguientes:

- **Redes Neuronales Recurrentes (RNN):**

Son una clase de redes neuronales artificiales diseñadas para manejar datos secuenciales manteniendo un estado interno que captura información sobre entradas anteriores. Esto los hace particularmente efectivos para tareas que involucran datos de series temporales, procesamiento de lenguaje natural y otros tipos de datos secuenciales [7]. La característica de este tipo de red es que no tiene una estructura de capas definida, sino que permiten conexiones arbitrarias entre las neuronas, incluso pudiendo crear ciclos, con esto se consigue crear la temporalidad, permitiendo que la red tenga memoria.

De acuerdo con [11] una red recurrente es una red neuronal donde la matriz $X_{Nxn}$ de datos de entrenamiento se compone de datos secuenciales, esto es, de vectores indexados temporalmente. En una red recurrente, a diferencia de una red convencional, los vectores de la matriz de datos de entrenamiento son procesados secuencialmente uno a uno. Esto permite que el vector de capa oculta $S_{t-1}$ pueda ser usado recursivamente en el periodo $t$ para $N$ periodos de la serie. Su representación matemática es la siguiente:

$$S_t = g(x_t W_s + S_{t-1} S_s + u_t U_s + b_s) \tag{6}$$

$$Y_t = h(W_y S_t + b_y) \tag{6.1}$$

donde, $Y_t$ es el vector de salida del periodo $t$. $u_t$ es, por otra parte, un vector de entradas externas del periodo $t$ y $W_s$, $S_s$ $U_s$, $b_s$, $W_y$ y $b_y$ son las matrices de pesos y los vectores de interceptos de la capa oculta y el vector de salida. $g()$ y $h()$ son funciones de activación no lineal.

- **LSTM (Long Short-Term Memory):**

LSTM es un tipo especializado de red neuronal recurrente (RNN) diseñada para modelar secuencias temporales y dependencias de largo alcance de manera más efectiva que las RNN convencionales [8].

La ilustración 3 representa de acuerdo con lo descrito por [20] una unidad o neurona $A$ con una conexión recurrente, acompañada de un vector de entrada $x_t$ y una señal de salida $h_t$. La conexión recurrente se traduce en la formación de una secuencia de unidades $A$, donde cada una envía una señal adicional a la siguiente mientras se procesan las observaciones del conjunto de datos de entrenamiento.

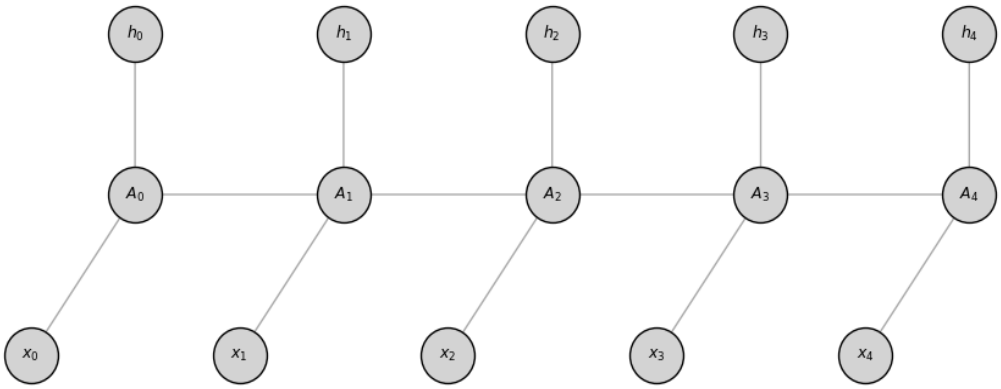


**Ilustración 3:** Representación de una red neuronal recurrente.

Las puertas en una red LSTM (Long Short-Term Memory) son componentes clave que controlan el flujo de información dentro de la celda, lo que permite que las LSTM manejen dependencias a largo plazo de manera eficiente, según lo expuesto por [20]:

- Puerta de olvido (*forget gate* $f_t$): Controla qué información del estado previo se olvida.
- Puerta de entrada (*input gate* $i_t$): Decide qué nueva información se almacena en el estado.
- Puerta de salida (*output gate* $o_t$): Genera la salida basada en el estado actualizado.

Para más detalle, se una capa oculta de una red LSTM se puede descomponer en dos partes. La primera es aquella en donde se actualizan los componentes del estado de celda en el periodo $t$:

$$f_t = \sigma(W_f[h_{t-1}, x_t] + b_f)$$

(7)

En la literatura, $f_t$ se conoce como *forget gate*, ya que, mediante una función de activación sigmoide[3] $\sigma$, se decide que información mantener del vector de la capa oculta del periodo anterior $h_{t-1}$. $W_f$ es la matriz de pesos asociada a $f_t$. La diferencia $[h_{t-1}, x_t]$ es la concatenación del vector de la capa oculta del periodo anterior $h_{t-1}$ y el vector de entrada actual $x_t$ y $b_f$ hace referencia al vector de sesgos asociado a $f_t$.

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i)$$

(8)

$\mathrm{i_t}$ controla qué nueva información debe actualizarse en el estado de celda, y para este caso la función sigmoide $\sigma$ decide qué proporción de la nueva información candidata $\tilde{\mathrm{C}}_\mathrm{t}$ incorporar.

$$\tilde{C}_t = tanh(W_i[h_{t-1}, x_t] + b_i)$$

(9)

El término $\tilde{C}_t$ es el vector de valores candidatos para actualizar el estado de celda. Se calcula utilizando una función de activación tangente hiperbólica $tanh$ que restringe los valores al rango [-1,1]. El vector resultante de la multiplicación entre el vector de la puerta del olvido y el estado celda en el periodo anterior $C_{t-1}$ más el resultado del vector resultante de la multiplicación entre el vector de la puerta de entrada y $\tilde{C}_t$ da como resultado el nuevo estado de celda $C_t$. Finalmente, la nueva información modulada por $\mathrm{i_t}$ y a información pasada filtrada por $\tilde{\mathrm{C}}_\mathrm{t}$ interactúan mediante el producto de Hadamard $\odot$:

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t$$

(10)

- $f_t \odot C_{t-1}$*: Conserva partes relevantes de* $C_{t-1}$.
- $i_t \odot \tilde{C}_t$*:* Incorpora nueva información relevante a $C_t$

La segunda parte de la capa oculta se calcula con el nuevo estado de celda $\mathrm{C_t}$ de la siguiente forma:

$$o_t = \sigma(W_0[h_{t-1}, x_t] + b_0)$$

(11)

$$h_t = o_t \odot tanh(C_t)$$

(12)

Donde, $\mathrm{o_t}$ es una función de activación sigmoide del vector de la capa oculta de periodo anterior $\mathrm{h_{t-1}}$ y la información nueva $\mathrm{x_t}$, mientras que $\mathrm{h_t}$ es el vector de capa oculta del periodo

[3] Función de activación sigmoide, que restringe los valores de salida al intervalo [0,1]

t. La estimación de los parámetros de la red LSTM, se realiza con el algoritmo de gradiente descendente estocástico. Estimados los parámetros, la predicción se realiza para una nueva unidad de entrada $\mathrm{x}_1$ de la siguiente manera.

$$o_{t+1} = \sigma(\widehat{W}_0[h_t, x_1] + \hat{b}_0) \tag{13}$$

- **Red neuronal de tipo GRU**

La GRU constituye una variante simplificada y eficiente de las LSTM, que mantiene la capacidad de capturar dependencias temporales en secuencias, pero con una arquitectura más compacta y menor costo computacional [28]. A diferencia de las LSTM, que incorporan tres compuertas (entrada, olvido y salida), las GRU utilizan únicamente dos compuertas principales, la primera compuerta de actualización (update gate), que controla la proporción de información anterior que debe mantenerse; y la segunda compuerta de reinicio (reset gate), que decide cuánto de la información pasada se debe olvidar.

- **Redes Convolucionales Unidimensionales (CNN-1D)**

El modelo CNN-1D (Convolutional Neural Network unidimensional) está diseñada para capturar patrones locales y estructuras espaciales-temporales en series financieras mediante filtros convolucionales. A diferencia de las redes recurrentes (LSTM o GRU), que procesan secuencias de forma iterativa, las CNN aplican ventanas deslizantes (kernels) [22] sobre la serie, detectando automáticamente regularidades o formas características en los retornos históricos.

$$\widehat{Ln_{\Delta+1}} = \mathrm{fCNN1D}[X_t; \theta] \tag{14}$$

Donde $Ln_{\Delta+1}$ es la estimación del cambio logarítmico de precio entre $t$ y $t_{+1}$ generada por la red CNN-1D. La siguiente variable es $\mathrm{fCNN1D}$ que es la función de la red convolucional que transforma las entradas $X_t$ en una predicción de retorno. Así pues $X_t$ es el conjunto de variables predictoras derivadas del histórico de precios e indicadores técnicos (medias móviles, volatilidades) y finalmente $\theta$ Conjunto de pesos y sesgos aprendidos durante el entrenamiento-

#### *3.1.1.4 Métricas de error:*

Las métricas de error son herramientas esenciales en el análisis de regresión y en los modelos de predicción basados en aprendizaje automático. Estas métricas se utilizan para medir qué tan cerca está el resultado predicho del resultado real [15]. A continuación, se exponen métricas que hasta el momento se considerar el acorde a proyecto:

- **Error medio:**

El Error Medio (ME) es una métrica que calcula el promedio de los errores entre los valores reales $y_i$ y los valores predichos $\hat{y}_i$ en un conjunto de datos.

$$ME = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i) \tag{15}$$

- **RMSE (Error cuadrático medio):**

El error cuadrático medio (MSE) es una métrica común que se utiliza para medir el promedio de los cuadrados de los errores, que son las diferencias entre los valores estimados y el valor real.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{16}$$

- **Theil's U:**

El estadístico U de Theil es una medida utilizada principalmente en el contexto de la precisión de los pronósticos. Está diseñado para comparar la precisión de un modelo de pronóstico dado con un modelo ingenuo, generalmente un paseo aleatorio o un promedio simple.

$$U = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}y_i^2} + \sqrt{\frac{1}{n}\sum_{i=1}^{n}\hat{y}_i^2}} \tag{17}$$

- **$R^2$ (Coeficiente de Determinación):**

El coeficiente de determinación, denotado como "$R^2$" es una medida estadística que se utiliza en el análisis de regresión para evaluar qué proporción de la variabilidad en la variable dependiente puede ser explicada por el modelo de regresión. En otras palabras, $R^2$ indica cuán bien el modelo de regresión se ajusta a los datos observados.

$$\mathrm{R}^2 = 1 - \frac{\sum(\hat{x} - x)^2}{\sum(x - \bar{x})^2} \tag{18}$$

- **Hit Rate:**

Métrica que mide el porcentaje de éxitos dentro de un conjunto de intentos, pudiendo traducirse como tasa de acierto o porcentaje de éxito [21]. Se utiliza en diversos campos como el de ventas, para evaluar la efectividad de los vendedores, o en informática, para medir la proporción de solicitudes que la caché atiende correctamente.

$$DA = \frac{1}{n}\sum_{t+1}^{n} 1[sign(y_t) = sign(\hat{y}_t)] \tag{19}$$

#### *3.1.1.5 Entrenamiento en Machine Learning*

Como se ha mencionado previamente, el aprendizaje automático se basa en la capacidad de los algoritmos para identificar patrones, sin embargo, para que este proceso sea eficiente, es fundamental preparar los datos adecuadamente y entrenar los modelos de forma estructurada. Dos aspectos cruciales en este contexto son la normalización de los datos y el entrenamiento del modelo.

La normalización o estandarización de datos se refiere al proceso de transformar los valores de las características (variables) en un rango o escala común. en este caso si llegase a la necesidad de ello, se transformarían los datos para que tengan media cero y desviación estándar unitaria:

$$x^{'} = \frac{x - \mu}{\sigma} \tag{20}$$

Y el entrenamiento es el proceso mediante el cual un modelo de Machine Learning ajusta sus parámetros para minimizar una función de pérdida en un conjunto de datos de entrenamiento. Esto se logra identificando patrones y relaciones en los datos.

- Dividir los datos en conjuntos de entrenamiento, validación y prueba.
- Realizar normalización o estandarización según sea necesario.
- Selección del algoritmo más adecuado para el problema.
- Utilizar un optimizador, como el gradiente descendente, para ajustar los parámetros minimizando la función de pérdida.

## 3.2 ANTECEDENTES

Hay una gran variedad de antecedentes relacionadas con la predicción de precios del mercado bursátil. Siempre ha sido un gran tema de interés el poder anticipar subidas o bajadas en el valor de las acciones, ya que esto permite a los inversionistas tomar decisiones informadas y estratégicas para maximizar sus ganancias o minimizar sus riesgos. Con la evolución de la tecnología y la creciente disponibilidad de datos, se han desarrollado numerosas investigaciones y metodologías que buscan mejorar la precisión en estas predicciones. Desde enfoques tradicionales basados en modelos estadísticos hasta técnicas avanzadas de aprendizaje automático y profundo, la predicción de precios bursátiles sigue siendo un campo de estudio dinámico que combina elementos de finanzas, matemáticas y computación.

**Predicción del precio de acciones en el mercado de valores con machine Learning [12].**

En el ámbito académico, destaca el proyecto de [12], cuyo objetivo fue desarrollar un algoritmo de regresión para predecir el precio a corto plazo de acciones individuales, utilizando indicadores y datos históricos. La metodología se dividió en tres etapas principales, la primera consistió en descargar datos del S&P500 desde Yahoo Finance. La segunda en el procesamiento y transformación de los datos mediante ETL, utilizando el paquete Pandas en Python y la tercera etapa fue la aplicación de herramientas de aprendizaje automático: NLTK y ScikitLearn. NLTK se empleó por sus modelos pre-entrenados de análisis de lenguaje natural, específicamente VADER (*Valence Aware Dictionary for Sentiment Reasoning*), por su capacidad de evaluar polaridad e intensidad emocional. Por su parte, *ScikitLearn* contiene algoritmos de machine learning y herramientas para optimización de hiperparámetros y evaluación del modelo.

**Predicción del precio de acciones de la bolsa de valores estadounidense utilizando técnicas de aprendizaje automático basadas en datos de análisis técnico y fundamental [13].**

Otro proyecto es el de [13] él tenía como objetivo predicción de un conjunto de acciones de los índices más importantes el S&P 500, en el cual se seleccionan tres técnicas de aprendizaje automático como la Regresor de Vectores de Soporte (SVR), los Bosques Aleatorios y finalmente el Perceptrón Multicapa (MLP). En la primera etapa, se generaron aproximadamente 180 modelos basados en las técnicas seleccionadas y se evaluaron utilizando métricas de error como RMSE y MAE. Como resultados se obtuvo que, para el análisis técnico, el modelo que mejor desempeño tuvo fue el Regresor de Vectores de Soporte (SVR), con un RMSE promedio de 1.07 y un MAE promedio de 0.5. Para el análisis fundamental, el Perceptrón Multicapa (MLP) mostró el mejor desempeño, con un RMSE promedio de 4.75 y un MAE promedio de 3.1.

**Stock market prediction using data mining techniques [14].**

El documento [14] implementaron un enfoque para predecir tendencias del mercado bursátil utilizando técnicas de minería de datos y aprendizaje automático, específicamente los modelos Random Forest y Support Vector Machine (SVM). La metodología combina análisis de sentimientos, que evalúa noticias y publicaciones en redes sociales como Reddit para determinar su impacto positivo, negativo o neutral sobre los precios de las acciones, con datos históricos del índice Dow Jones Industrial Average (DJIA). Para procesar el texto, se emplearon modelos como "Bag of Words" y n-gramas, que convierten la información textual en vectores de características.

El objetivo principal del estudio es demostrar que la integración de datos históricos con análisis de sentimientos mejora la precisión en la predicción de movimientos del mercado bursátil. Los resultados indican que Random Forest supera a SVM en precisión, alcanzando un 86.2% con el modelo 2-gram. SVM también muestra un rendimiento competitivo, con su modelo no lineal logrando una precisión del 85.1%.

**Portfolio optimization with return prediction using deep learning and machine learning [16].**

El trabajo llevado a cabo por [16] analiza cómo la integración de predicciones de rentabilidad, utilizando modelos avanzados de aprendizaje automático y profundo para datos históricos de 9 años del índice China Securities 100, mejora el rendimiento en la formación de carteras en comparación con métodos tradicionales de series temporales. Se emplearon dos modelos de aprendizaje automático (Random Forest y Support Vector Regression) y tres de aprendizaje profundo (LSTM, DMLP y redes neuronales convolucionales) para preseleccionar acciones antes de optimizar las carteras con los enfoques de media-varianza (MV) y omega. Los resultados muestran que el modelo Random Forest con media-varianza ajustada tiene el mejor desempeño, incluso después de considerar costos por alta rotación.

**Pronóstico de volatilidad de la TRM mediante un modelo híbrido LSTM-GARCH [20]**

Otro antecedente relevante en la investigación de pronóstico financiero es el modelo híbrido LSTM-GARCH desarrollado por [20] para predecir la volatilidad de la Tasa Representativa del Mercado (TRM). Este modelo combina redes neuronales recurrentes LSTM con coeficientes provenientes de modelos de series temporales GARCH, EGARCH y EWMA, siguiendo la metodología de [30]. El enfoque utiliza datos históricos de la TRM desde 2008 hasta 2018 para entrenar los modelos, y evalúa su desempeño en el periodo de julio 2018 a julio 2019. Los resultados muestran que el modelo híbrido supera en precisión al modelo LSTM estándar, al incorporar la información estadística de los coeficientes GARCH y EGARCH. La evaluación se llevó a cabo utilizando múltiples medidas de error, tanto lineales como no lineales.

# 3. METODOLOGÍA

En el proyecto titulado “Predicción de Precios de Activos de Renta Variable de la BVC Mediante Modelos de Aprendizaje Supervisado”, se implementó un enfoque metodológico descriptivo con orientación cuantitativa, articulado bajo el estándar internacional CRISP-DM (Cross Industry Standard Process for Data Mining). Esta metodología permitió estructurar de manera sistemática todas las fases del proceso analítico, desde la comprensión del problema y de los datos, hasta la construcción de los modelos y la obtención de resultados verificables.

En consonancia con la fase Comprensión del Negocio, se definió como objetivo principal la predicción del comportamiento de activos listados en la Bolsa de Valores de Colombia (BVC) mediante técnicas de aprendizaje supervisado. Para ello, se observaron y analizaron variables como precios históricos, volatilidad y liquidez, permitiendo establecer el fundamento estadístico del estudio.

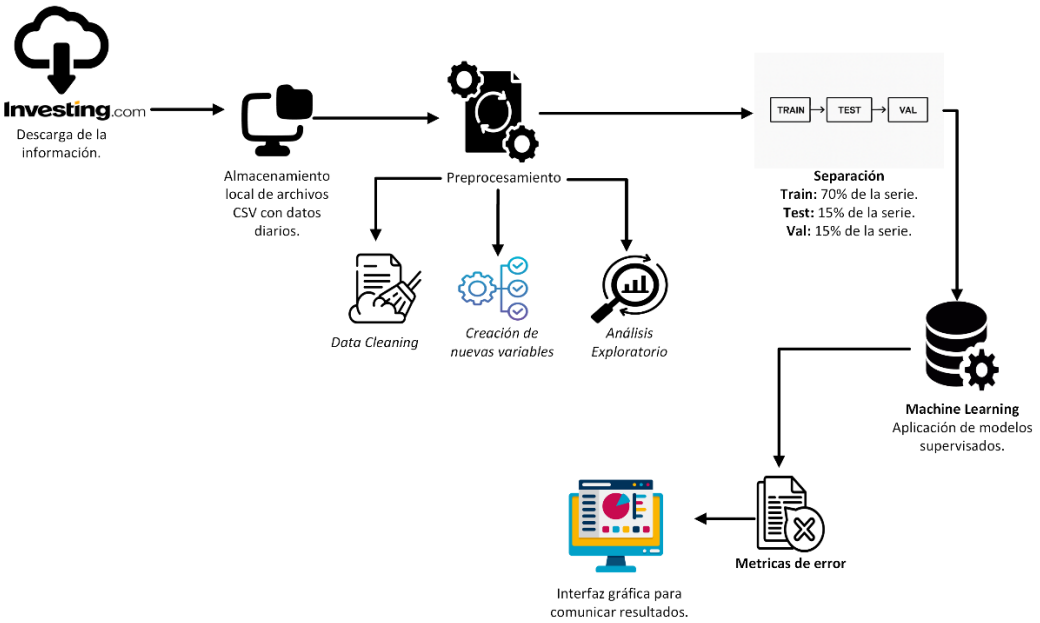


**Ilustración 4.** Proceso metodológico.

En la fase de Comprensión de los Datos, la obtención de información se realizó mediante el API de Investing.com, seleccionando un periodo temporal comprendido entre enero del 2021 y octubre del 2025. Esta selección garantizó trabajar con series históricas amplias, actualizadas y de alta calidad. Las tareas fueron desarrolladas en Python. Como resultado de esta fase, los datos quedaron consolidados en estructuras tabulares limpias, coherentes y listas para las etapas subsiguientes de Modelado, Evaluación y Despliegue, fases que completan el ciclo metodológico CRISP-DM aplicado al proyecto.

La metodología se desarrolla en 5 etapas principales:

| Hito | Tarea | Resultado | CRISP-DM |
|---|---|---|---|
| 1 | **Identificar los activos a modelar mediante el análisis de variables como volatilidad, liquidez y rentabilidad.** | | |
| | Recopilar datos históricos de activos financieros desde Investing y/o BVC. | Recopilación de los datos | Comprensión de los Datos |
| | Justificar la selección de activos basándose en indicadores como volatilidad Y rentabilidad. | Selección de los datos | |
| | Realizar un análisis exploratorio de los datos (estadísticos descriptivos). | Descripción y exploración de los datos | |
| 2 | **Preparar y preprocesar los datos, garantizando su calidad mediante la limpieza, normalización y selección de características relevantes para el modelado.** | | |
| | Normalizar y limpiar los datos, eliminando valores faltantes o inconsistentes. | Selección de los datos | Preparación de los Datos |
| | Construir variables adicionales como retornos logarítmicos y volatilidad promedio. | Construcción de variables adicionales | |
| | Estructurar los datos preprocesados en un formato adecuado para su uso en los modelos. | Preprocesamiento | |
| 3 | **Implementar y entrenar modelos de aprendizaje supervisado, ajustando cada modelo para la predicción de precios de acuerdo con el tipo de activo.** | | |
| | Implementar modelos de aprendizaje supervisado. | Construcción del modelo | Modelado |
| | Configurar los hiperparámetros de los modelos mediante técnicas de optimización. | Configuración de hiperparámetros | |
| | Entrenar los modelos con los datos preprocesados y generar predicciones en los conjuntos de prueba. | Entrenar modelos | |
| 4 | **Evaluar la efectividad de los modelos predictivos mediante métricas de rendimiento, midiendo la precisión y capacidad de generalización en la predicción de precios.** | | |
| | Métricas para cada modelo y activo. | Evaluación de los resultados | Evaluación |
| | Comparar el desempeño de los diferentes modelos y seleccionar el mejor modelo para cada activo. | Selección del mejor modelo | |
| 5 | **Emplear una herramienta de visualización que permita visualizar el comportamiento de cada activo financiero de acuerdo a su mejor modelo.** | | |
| | Dashboard interactivo con datos, tendencias y métricas. | Creación de herramientas de visualización | Despliegue |
| | Documento técnico con el análisis, modelos seleccionados y conclusiones del proyecto. | Informe final | |

Tabla 1. Descripción de las etapas del proyecto.

# 4. RESULTADOS

## 4.1 Recopilación de datos

El proceso de descarga de la información corresponde a los activos de renta variable que conforman el índice COLCAP[4], en el COLCAP cotizan y se incluyen las empresas más líquidas del mercado accionario colombiano, es decir, las acciones que se negocian con mayor frecuencia y volumen en la Bolsa de Valores de Colombia (BVC).

| Empresa | Sector | Código Superfinanciera |
|---|---|---|
| CEMENTOS ARGOS S.A. | Materiales básicos | COD38PAAO005 |
| CELSIA S.A. E.S.P. | Suministros | COT60PAAO005 |
| CANACOL ENERGY LTD | Productores de petróleo y gas | CAC13PAAO008 |
| CORPORACIÓN FINANCIERA COLOMBIANA S.A. | Financiero | COJ12PAAO007 |
| BANCOLOMBIA S.A. | Financiero | COB07PAAO012 |
| BOLSA DE VALORES DE COLOMBIA S.A. | Financiero | COR01PAAO011 |
| BANCO DE BOGOTÁ S.A. | Financiero | COB01PAAO006 |
| ECOPETROL S.A. | Productores de petróleo y gas | COC04PAAO008 |
| GRUPO ENERGIA BOGOTA S.A. E.S.P. | Electricidad | COE01PAAO018 |
| EMPRESA DE TELECOMUNICACIONES DE BOGOTA S.A. ESP | Tecnología | COI13PAAO007 |
| GRUPO ARGOS S.A. | Materiales de construcción | COT09PAAO003 |
| GRUPO DE INVERSIONES SURAMERICANA S.A. | Servicios financieros | COT13PAAO004 |
| INTERCONEXIÓN ELÉCTRICA S.A. E.S.P. | Suministros | COE15PAAO001 |
| GRUPO NUTRESA S.A. | Productos alimenticios | COT04PAAO003 |
| GRUPO AVAL ACCIONES Y VALORES S.A. | Servicios financieros | COT29PAAD012 |
| CEMENTOS ARGOS S.A. | Materiales y construcción | COD38PAAD115 |
| BANCO DAVIVIENDA SA | Financiero | COB51PAAD096 |
| GRUPO ARGOS S.A. | Materiales y construcción | COT09PAAD006 |
| GRUPO DE INVERSIONES SURAMERICANA S.A. | Servicios financieros | COT13PAAD031 |
| PROMIGAS S.A. E.S.P. | Gas, agua y servicios múltiples | COI04PAAO006 |
| CORPORACION FINANCIERA COLOMBIANA S.A. | Servicios financieros | COJ12PAAD000 |
| MINEROS S.A. | Minería | COC07PAAO002 |
| ORGANIZACION TERPEL S.A. | Equipos, servicios y distribución de petróleo | COG20PAAO006 |
| GRUPO BOLIVAR S.A. | Servicios financieros | COT23PAAO003 |

Tabla 2. Descripción formal de las empresas consideradas.
Fuente: Investing y BVC (2025)

[4] Representa el comportamiento de las acciones más líquidas y de mayor importancia del mercado accionario colombiano.

El conjunto de datos utilizado en este estudio contiene series temporales diarias de precios de cierre para las 25 acciones más representativas de acuerdo al MCSI COLCAP como se muestra en (Ver tabla 3) del mercado de renta variable colombiano que va desde el sector financiero, construcción, energía y telecomunicaciones. La duración del período de muestra está dictada por el período comprendido entre el 1 de enero del 2019 al 11 de abril de 2025, es decir alrededor de 4.3 años bursátiles.

| Nemotécnico | Fecha Inicio | Fecha Fin | Registros |
|---|---|---|---|
| GRUPOARGOS | 5/01/2021 | 10/10/2025 | 1134 |
| PFGRUPOARG | 5/01/2021 | 10/10/2025 | 1116 |
| BOGOTA | 5/01/2021 | 10/10/2025 | 1090 |
| PFBCOLOM | 5/01/2021 | 10/10/2025 | 1159 |
| BVC | 5/01/2021 | 10/10/2025 | 864 |
| CEMARGOS | 5/01/2021 | 10/10/2025 | 1163 |
| PFCEMARGOS | 5/01/2021 | 10/10/2025 | 999 |
| CELSIA | 5/01/2021 | 10/10/2025 | 1152 |
| CORFICOLCF | 5/01/2021 | 10/10/2025 | 1154 |
| PFCORFICOL | 5/01/2021 | 10/10/2025 | 1018 |
| CIBEST | 5/01/2021 | 10/10/2025 | 1168 |
| CNEC | 5/01/2021 | 10/10/2025 | 1145 |
| PFDAVVNDA | 5/01/2021 | 10/10/2025 | 1149 |
| ECOPETROL | 5/01/2021 | 10/10/2025 | 1166 |
| ETB | 5/01/2021 | 10/10/2025 | 693 |
| PFAVAL | 5/01/2021 | 10/10/2025 | 1150 |
| GEB | 5/01/2021 | 10/10/2025 | 1159 |
| ISA | 5/01/2021 | 10/10/2025 | 1165 |
| MINEROS | 5/01/2021 | 10/10/2025 | 1108 |
| NUTRESA | 5/01/2021 | 10/10/2025 | 1018 |
| PROMIGAS | 5/01/2021 | 10/10/2025 | 1001 |
| GRUBOLIVAR | 5/01/2021 | 10/10/2025 | 1013 |
| GRUPOSURA | 5/01/2021 | 10/10/2025 | 1092 |
| PFGRUPSURA | 5/01/2021 | 10/10/2025 | 1147 |
| TERPEL | 5/01/2021 | 10/10/2025 | 1002 |

Tabla 3. Registros de las cotizaciones.
Fuente: Investing y BVC (2025)

### 4.1.1 Descripción de data set

Posteriormente se hace un proceso de consolidación por cada activo de renta variable, donde se muestra el data set original:

| Índice | Emisor | Fecha | Cierre | Apertura | Máximo | Mínimo | Volumen |
|---|---|---|---|---|---|---|---|
| 1 | Banco de Bogotá | 2021-01-04 | 76.780 | 75.600 | 76.780 | 75.000 | 774000 |
| 2 | Banco de Bogotá | 2021-01-05 | 77.400 | 76.780 | 77.400 | 76.600 | 331000 |
| 3 | Banco de Bogotá | 2021-01-06 | 77.400 | 77.390 | 77.500 | 77.000 | 447000 |
| 4 | Banco de Bogotá | 2021-01-07 | 76.050 | 76.990 | 76.990 | 76.050 | 32000 |
| 5 | Banco de Bogotá | 2021-01-08 | 77.000 | 77.000 | 77.000 | 77.000 | 417000 |
| 6 | Banco de Bogotá | 2021-01-12 | 77.300 | 77.460 | 77.820 | 77.300 | 469000 |
| 7 | Banco de Bogotá | 2021-01-13 | 77.500 | 77.800 | 77.800 | 77.020 | 743000 |

Tabla 4. Registros de las cotizaciones diarias para activo BBO – Banco de Bogotá.
Fuente: Investing y BVC (2025)

## 4.2 Justificación de la elección de los datos

Para lo expuesto en la tabla 4 la cual expone los criterios de selección se desarrolló de la siguiente manera:

### 4.2.1 Explicación cálculo de la rentabilidad y volatilidad promedio.

Para la primera columna de rentabilidad diaria se aplicó el cálculo de las variaciones relativas, es decir el cambio porcentual que hay en el momento actual $t_1$ con respecto al día anterior $t_{-1}$, pero como trabajamos con una serie temporal, lo que se hizo fue promediarlo para obtener el promedio diario y luego multiplicar ese promedio por 252 (ver ecuación 1 y 1.1) para poder obtener la rentabilidad anual que es el resultado de la segunda columna de la tabla 5.

$$\Delta = \frac{\Delta_t}{\Delta_{t-1}} - 1 \rightarrow \overline{\Delta} = (1+\Delta)^{252} - 1 \tag{1.1}$$

Ahora, se preguntarán porque se multiplica por 252 el promedio y no por 365 si se esta tratando de anualizar una serie, pues la respuesta a ello radica en que las bolsas cotizan alrededor de 252 a 255 días hábiles al año. Por lo que trabajar con un promedio global de 252 es un estándar en el campo de las finanzas.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(\Delta_i - \overline{\Delta})^2}{n-1}} \rightarrow \sigma_* = \sigma * [\sqrt{252}] \tag{2.2}$$

Para la primera y segunda columna se obtiene los procesos diarios y anuales de distancia entre cada punto (cotización) individual Δ con respecto a la media $\overline{\Delta}$. Lo anterior se le conoce como volatilidad, y estadísticamente es la desviación estándar $\sigma$. Y para ello se obtiene la raíz cuadrada de las diferencias de las sumatorias al cuadrado de los datos individuales con respecto a la media y se divide entre $n-1$ dado que nuestra serie es una muestra. Finalmente $\sigma$ se multiplica por la raíz de 252 para anualizarla, porque la varianza escala linealmente con el tiempo mientras que la desviación estándar (volatilidad) escala con la raíz del número de días de negociación en el año.

| Nemotécnico | Volatilidad Diaria | Volatilidad Anual | Rentabilidad diaria | Rentabilidad Anual |
|---|---|---|---|---|
| GRUPOARGOS | 2,406% | 38,200% | 0,021% | 5,340% |
| PFGRUPOARG | 2,584% | 41,025% | 0,007% | 1,892% |
| BOGOTA | 2,161% | 34,302% | -0,072% | -16,609% |
| PFBCOLOM | 1,741% | 27,635% | 0,036% | 9,464% |
| BVC | 2,095% | 33,253% | 0,005% | 1,307% |
| CEMARGOS | 2,174% | 34,505% | 0,043% | 11,545% |
| PFCEMARGOS | 2,440% | 38,737% | 0,103% | 29,559% |
| CELSIA | 1,786% | 28,352% | 0,000% | 0,045% |
| CORFICOLCF | 1,811% | 28,755% | -0,058% | -13,538% |
| PFCORFICOL | 1,891% | 30,017% | -0,044% | -10,573% |
| CNEC | 2,810% | 44,607% | -0,179% | -36,259% |
| PFDAVVNDA | 18,237% | 289,500% | -0,031% | -7,400% |
| ECOPETROL | 2,251% | 35,735% | -0,022% | -5,449% |
| ETB | 2,807% | 44,556% | -0,209% | -40,963% |
| PFAVAL | 1,840% | 29,209% | -0,044% | -10,590% |
| GEB | 1,889% | 29,986% | 0,006% | 1,580% |
| ISA | 2,279% | 36,171% | -0,009% | -2,151% |
| MINEROS | 2,208% | 35,043% | 0,105% | 30,353% |
| NUTRESA | 3,373% | 53,544% | 0,216% | 72,253% |
| PROMIGAS | 1,924% | 30,544% | -0,020% | -4,844% |
| GRUBOLIVAR | 2,252% | 35,755% | 0,018% | 4,512% |
| GRUPOSURA | 2,913% | 46,243% | 0,052% | 13,856% |
| PFGRUPSURA | 2,222% | 35,272% | 0,045% | 12,054% |
| TERPEL | 1,900% | 30,163% | 0,061% | 16,657% |

Tabla 5. Cálculo de la rentabilidad y volatilidad promedio.
Fuente: Cálculos propios con base a Investing y BVC (2025)

La tabla 5 muestra el perfil riesgo–retorno de las acciones analizadas. En términos de volatilidad anual, la mayoría de los activos se ubican entre el 28% y el 45%, lo cual expone un riesgo típico del mercado colombiano; sin embargo, destacan casos extremos (dato atípico) como PFDAVVNDA, con casi 290%. En contraste, emisores como PFBCOLOM, CELSIA, GEB y PROMIGAS exhiben menor volatilidad, funcionando como activos más defensivos. En rentabilidad anual, sobresalen NUTRESA, MINEROS y PFCEMARGOS, con retornos superiores al 25%, compensando adecuadamente su nivel de riesgo. Por el contrario, títulos como ETB, CNEC, CORFICOLCF y BOGOTÁ presentan pérdidas significativas pese a su volatilidad elevada, reflejando una relación riesgo–retorno desfavorable.

| Nemotécnico | Volatilidad Diaria | Volatilidad Anual |
|---|---|---|
| PFDAVVNDA | 18,237% | 289,500% |
| NUTRESA | 3,373% | 53,544% |
| GRUPOSURA | 2,913% | 46,243% |
| CNEC | 2,810% | 44,607% |
| ETB | 2,807% | 44,556% |

Tabla 6. Acciones escogidas por mayor volatilidad.
Fuente: Cálculos propios con base a Investing y BVC (2025)

### 4.3 Realizar un análisis exploratorio de los datos y construcción de variables.

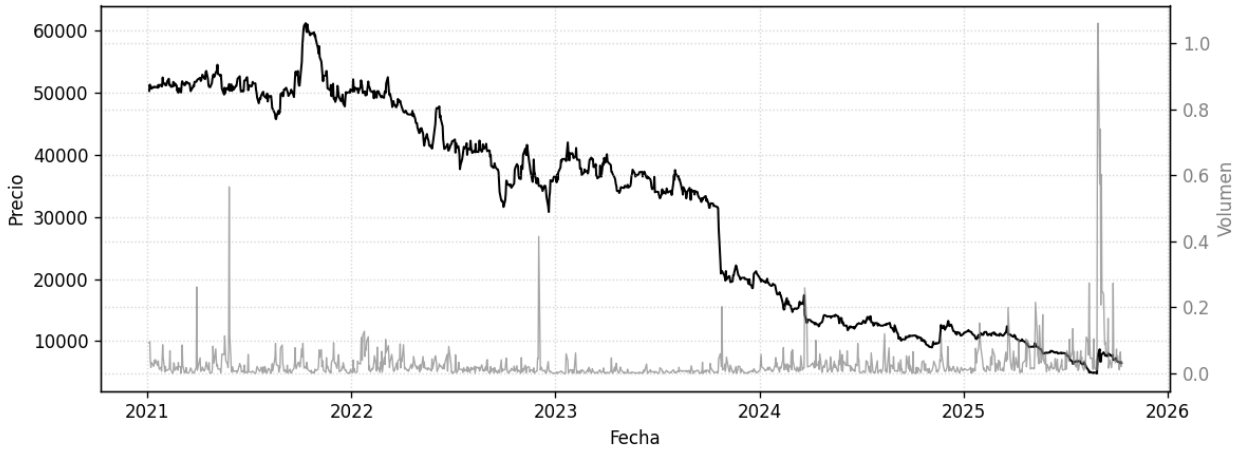


**Figura 1.** Tendencia histórica CANACOL.

| Variable | Conteo | Media | Desviación | Percentil 25% | Percentil 95% |
|---|---|---|---|---|---|
| Precio | 1146 | 30901 | 16889 | 12500 | 52250 |
| Apertura | 1146 | 30903 | 16861 | 12500 | 52250 |
| Máximo | 1146 | 31160 | 16990 | 12600 | 52600 |
| Mínimo | 1146 | 30628 | 16749 | 12445 | 51787 |
| Variación relativa | 1145 | -0.00141 | 0.02840 | -0.01380 | 0.037025 |
| Variación logarítmica | 1145 | -0.001787 | 0.028100 | -0.013878 | 0.036413 |

Tabla 7. Estadísticas descriptivas CANACOL.

Fuente: Cálculos propios con base a Investing y BVC (2025)

El precio promedio de CANACOL se sitúa alrededor de $30.900, con una desviación estándar cercana a $16.889, indicando fluctuaciones significativas en la cotización. Los percentiles vigorizan esta dispersión: mientras el 25% de los precios se ubican por debajo de $12.500, el 95% supera los $52.250, lo que evidencia episodios de fuertes apreciaciones. Los valores de apertura, máximo y mínimo mantienen patrones similares. En cuanto a la variación relativa y logarítmica, ambas presentan medias negativas, lo que insinúa una ligera tendencia bajista en el rendimiento diario.

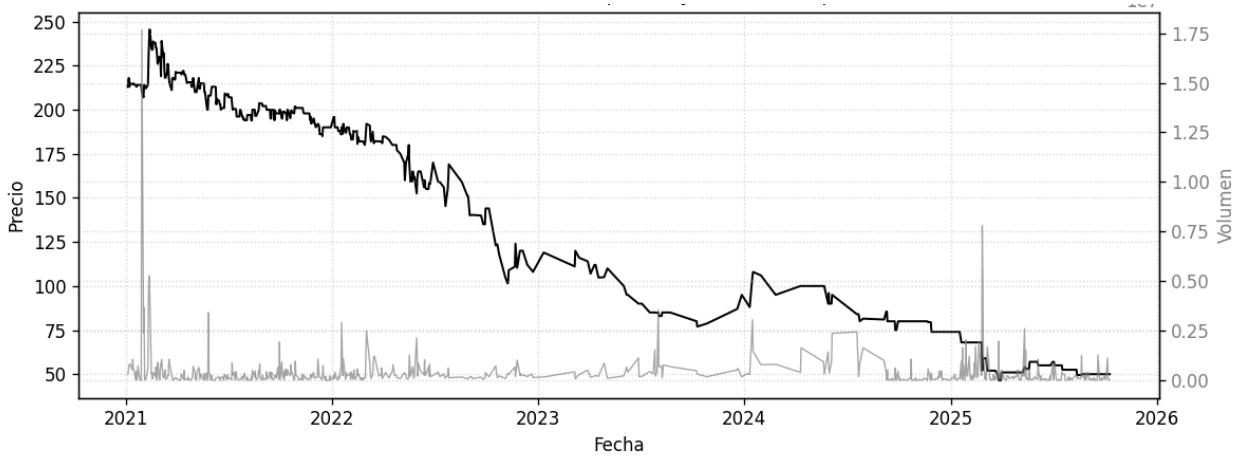


**Figura 2.** Tendencia histórica ETB.

| Variable | Conteo | Media | Desviación | Percentil 25% | Percentil 95% |
|---|---|---|---|---|---|
| Precio | 1019 | 58731 | 37250 | 34750 | 129962 |
| Apertura | 1019 | 58304 | 37017 | 33840 | 129962 |
| Máximo | 1019 | 59086 | 37249 | 34900 | 129980 |
| Mínimo | 1019 | 60243 | 214559 | 1610 | 238436 |
| Variación relativa | 1019 | 0.002751 | 0.036326 | -0.007400 | 0.047540 |
| Variación logarítmica | 1018 | 0.002158 | 0.033729 | -0.007445 | 0.046525 |

Tabla 8. Estadísticas descriptivas ETB.

Fuente: Cálculos propios con base a Investing y BVC (2025)

Las estadísticas descriptivas de ETB muestran que el precio promedio ronda los $58.731, pero la desviación estándar es muy elevada ($37.250), lo que indica fuertes fluctuaciones en la serie. Esto se confirma con sus percentiles: mientras el 25% de los valores se sitúan alrededor de $34.750, el 95% supera los $129.962, demostrando episodios de variaciones extremas. El comportamiento de apertura, máximo y mínimo mantiene esta dinámica. En cuanto a los rendimientos, tanto la variación relativa como la logarítmica presentan medias positivas, lo que indica una ligera tendencia alcista diaria.

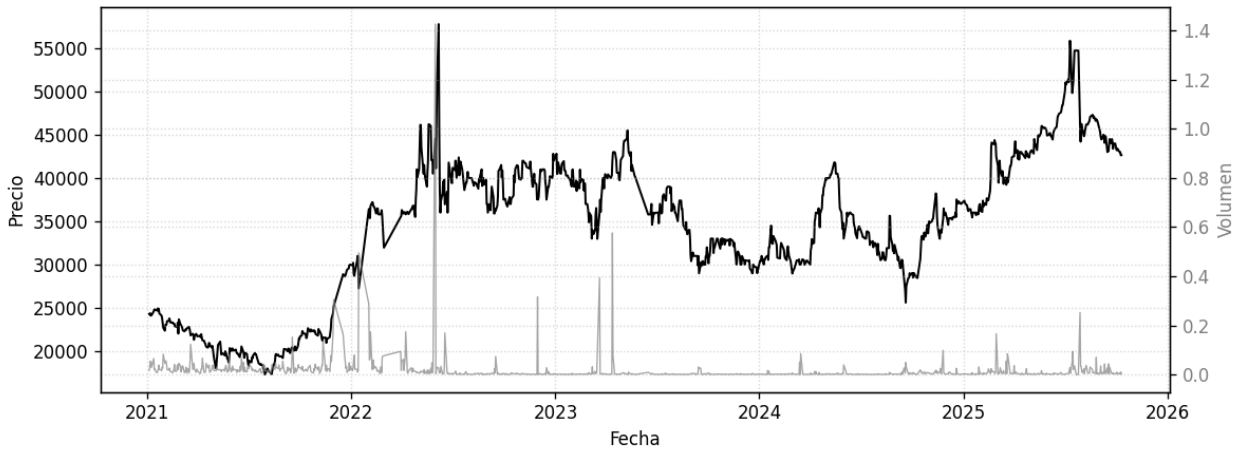


**Figura 3.** Tendencia histórica SURAMERICANA.

| Variable | Conteo | Media | Desviación | Percentil 25% | Percentil 95% |
|---|---|---|---|---|---|
| Precio | 1093 | 34184 | 8316 | 29900 | 45852 |
| Apertura | 1093 | 34060 | 8251 | 29500 | 4585 |
| Máximo | 1093 | 34571 | 8408 | 30000 | 46200 |
| Mínimo | 1093 | 33619 | 8138 | 29000 | 45380 |
| Variación relativa | 1093 | 0.000905 | 0.029451 | -0.010600 | 0.042340 |
| Variación logarítmica | 1092 | 0.000515 | 0.029130 | -0.010668 | 0.041515 |

Tabla 9. Estadísticas descriptivas SURAMERICANA.
Fuente: Cálculos propios con base a Investing y BVC (2025)

En relación a la tabla 9, se muestra un comportamiento relativamente estable dentro del mercado colombiano. El precio promedio se sitúa alrededor de $34.184, acompañado de una desviación estándar de $8.316. En términos de retornos, tanto la variación relativa como la logarítmica presentan medias muy cercanas a cero 0.000905 y 0.000515, lo que apunta que el activo no exhibe una tendencia alcista o bajista marcada en el corto plazo.

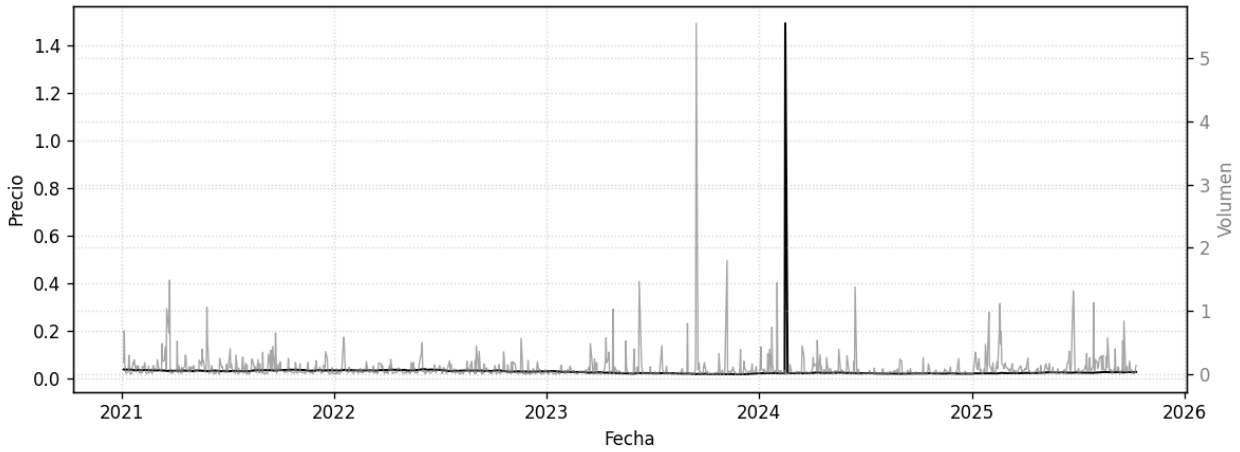


**Figura 4.** Tendencia histórica PFDAVVNDA.

| Variable | Conteo | Media | Desviación | Percentil 25% | Percentil 95% |
|---|---|---|---|---|---|
| Precio | 1150 | 26298 | 43715 | 19860 | 33761 |
| Apertura | 1150 | 26258 | 44168 | 19820 | 33711 |
| Máximo | 1150 | 26543 | 44176 | 20000 | 34000 |
| Mínimo | 1150 | 25968 | 43704 | 19700 | 33235 |
| Variación relativa | 1150 | 0.066454 | 2.289519 | -0.010600 | 0.030410 |
| Variación logarítmica | 1149 | -0.000305 | 0.182370 | -0.010643 | 0.029938 |

Tabla 10. Estadísticas descriptivas PFDAVVNDA.
Fuente: Cálculos propios con base a Investing y BVC (2025)

Las estadísticas descriptivas de PFDAVVNDA evidencian un comportamiento altamente inusual y volátil. Aunque el precio promedio se ubica alrededor de $26.298, las desviaciones estándar tanto del precio como de las demás medidas de cotización superan ampliamente los valores medios, lo que indica fluctuaciones extremas y atípicas en el periodo analizado. En cuanto a los retornos, la variación relativa presenta una media de 0.066, acompañada de una desviación extraordinariamente alta (2.289). La variación logarítmica muestra una media cercana a cero, pero nuevamente una desviación muy elevada (0.182), lo cual refleja inestabilidad pronunciada en los rendimientos diarios.


**Figura 5.** Tendencia histórica NUTRESA.

| Variable | Conteo | Media | Desviación | Percentil 25% | Percentil 95% |
|---|---|---|---|---|---|
| Precio | 1019 | 58731 | 37250 | 34750 | 129962 |
| Apertura | 1019 | 58304 | 37017 | 33840 | 129962 |
| Máximo | 1019 | 59086 | 37249 | 34900 | 129980 |
| Mínimo | 1019 | 57794 | 36921 | 33760 | 129962 |
| Variación relativa | 1019 | 0.002751 | 0.036326 | -0.007400 | 0.047540 |
| Variación logarítmica | 1018 | 0.002158 | 0.033729 | -0.007445 | 0.046525 |

Tabla 11. Estadísticas descriptivas NUTRESA.
Fuente: Cálculos propios con base a Investing y BVC (2025)

Las estadísticas descriptivas de NUTRESA muestran un activo con alta volatilidad y amplios movimientos en el precio. El valor promedio es de $58.731, pero la elevada desviación estándar indica fluctuaciones fuertes, con precios que van desde $34.750 (percentil 25) hasta más de $129.962 (percentil 95), reflejando episodios de valorización significativa. En cuanto a los rendimientos, tanto la variación relativa como la logarítmica presentan medias positivas, señalando una tendencia general al alza; sin embargo, sus desviaciones estándar muestran que dicha tendencia ocurre en medio de movimientos diarios amplios, característicos de un activo sensible a eventos del mercado.

### 4.4 Auditoría de colinealidad

El proceso de construcción, evaluación y depuración de las variables explicativas se desarrolló mediante un enfoque escalonado que combina criterios financieros, diagnósticos estadísticos y técnicas especializadas de preprocesamiento. El objetivo consistió en obtener un conjunto de predictores estable, interpretable y adecuado para diferentes familias de modelos utilizados en el análisis: modelos basados en árboles de decisión y modelos neuronales secuenciales. El tratamiento aplicado abarcó desde la creación de indicadores derivados de la microestructura del mercado, pasando por auditorías de correlación y análisis de inflación de varianza, hasta la reducción dimensional mediante componentes principales.

En una primera etapa se generó un conjunto amplio de indicadores provenientes de dos fuentes principales. Por un lado, se incorporaron variables asociadas a la microestructura del mercado y por otro lado, se construyeron variables financieras tradicionales, entre las cuales se incluyen medias móviles exponenciales, medidas de momentum, osciladores, volatilidades estimadas mediante métodos de Parkinson, Rogers-Satchell y ventanas móviles, así como medidas de aceleración y pendiente de tendencia.

### 4.5 Auditoría inicial de correlaciones y eliminación de redundancias

Dados los múltiples indicadores generados, se realizó una auditoría global de correlación con el propósito de identificar relaciones casi perfectas entre variables. Con un umbral de correlación igual o mayor 98%, se evidenció que algunos precios OHLC (apertura, máximo, mínimo) resultaban esencialmente duplicados frente al precio de cierre. Con base en estos resultados, se eliminaron explícitamente variables como SMA_5, SMA_10, SMA_20, percent_b y RSI_20, por considerarse redundantes, mientras que se conservaron versiones más estables o representativas de cada familia de indicadores.

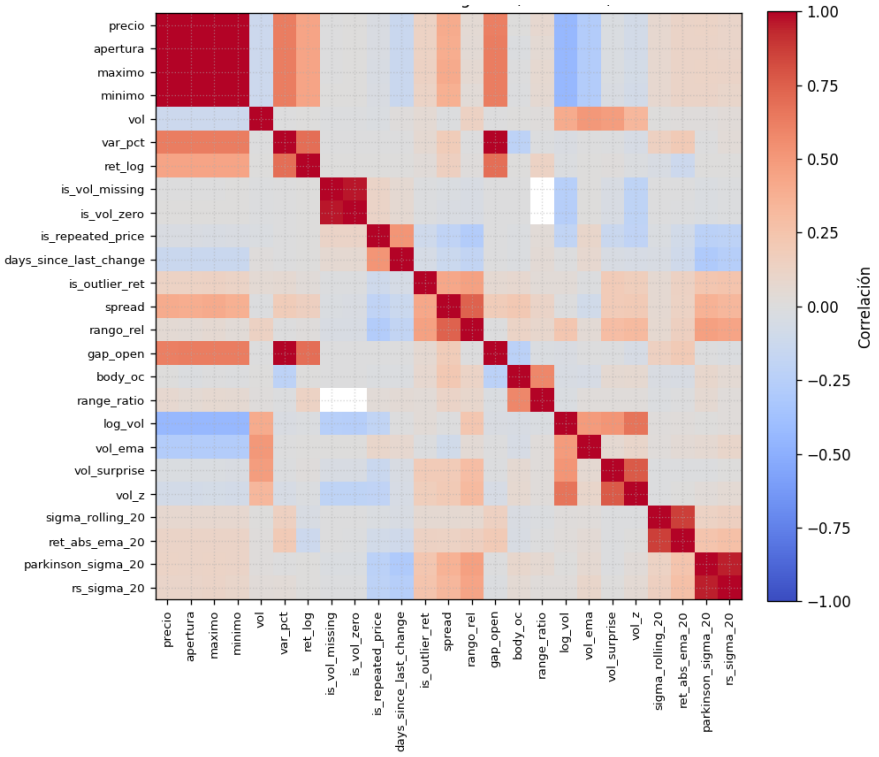


**Figura 6.** Matriz de correlación numéricas ex ante de imputación.

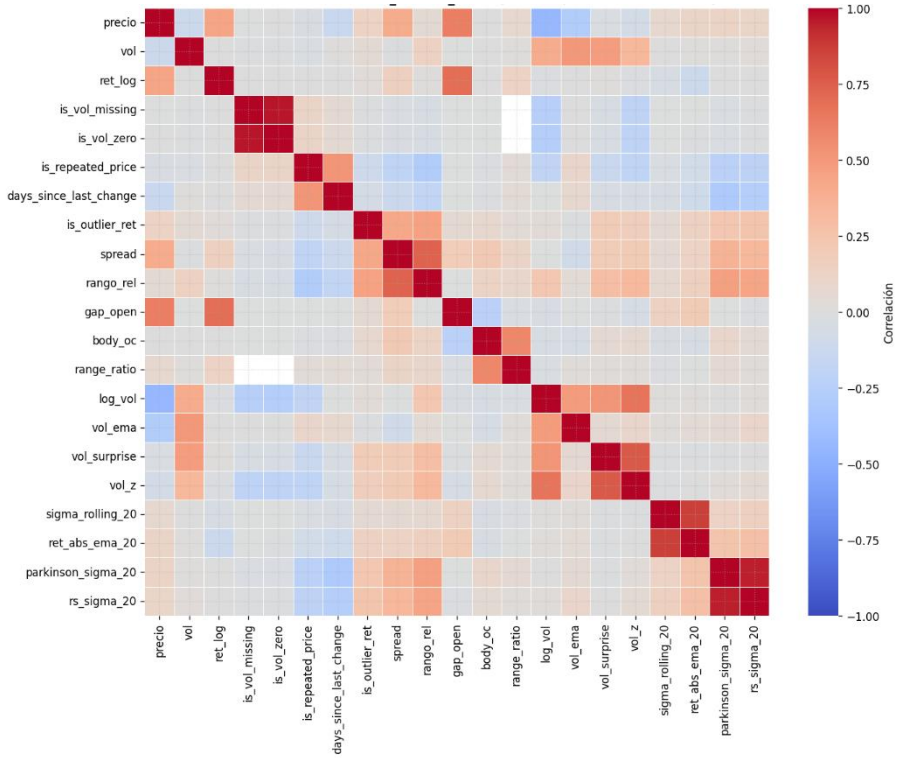


**Figura 7.** Matriz de correlación numéricas pos imputación.

### 4.6 Diagnóstico de multicolinealidad mediante VIF

Con el conjunto depurado se realizó un diagnóstico formal de multicolinealidad mediante el Factor de Inflación de Varianza (VIF). El análisis confirmó la presencia de colinealidad extrema entre numerosos indicadores derivados de la misma estructura subyacente, los indicadores de momentum, las razones precio[5], y las medidas de volatilidad de ventana larga.

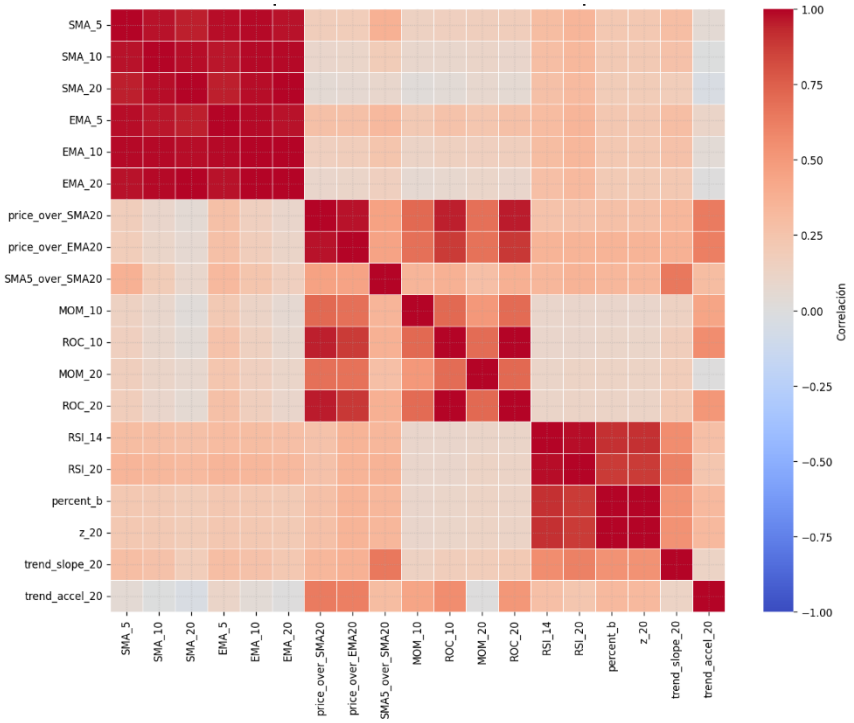


**Figura 8.** Matriz de correlación con SMA y EMA.

Cabe mencionar que, aunque el retorno logarítmico presentó un VIF relativamente alto (≈ 13), se decidió preservarlo explícitamente dentro del grupo de variables estables, debido a su importancia económica como variable fundamental en modelos financieros, siguiendo recomendaciones de la literatura especializada en series temporales financieras [36].

La siguiente tabla presenta la clasificación de variables según el diagnóstico de multicolinealidad obtenido mediante el Factor de Inflación de Varianza (VIF). El Grupo A está conformado por aquellas variables que muestran baja colinealidad (VIF < 5) o que, aun presentando un VIF superior, fueron protegidas debido a su alta relevancia económica e interpretativa dentro del modelo, como es el caso de ret_log.

[5] Tanto para SMA y EMA

Por su parte, el Grupo B reúne las variables que exhiben niveles moderados o severos de colinealidad, operacionalmente definidas como aquellas con VIF ≥ 5.

En lugar de eliminar estas variables colineales, se implementó un PCA selectivo exclusivamente sobre el Grupo B, con el fin de eliminar redundancia manteniendo la información esencial. Como resultado, las 23 variables colineales fueron transformadas en 8 componentes principales ortogonales, los cuales retienen más del 90% de la varianza original. De esta manera, se preserva la estructura informativa del conjunto, pero en una representación matemáticamente estable y adecuada para modelos neuronales.

| Grupo | Criterio | Variables incluidas |
|---|---|---|
| **Grupo A** (Variables estables) | VIF < 5 o protegidas por relevancia económica | vol, is_repeated_price, days_since_last_change, is_outlier_ret, body_oc, range_ratio, log_vol, vol_ema, vol_surprise, vol_z, MOM_10, **ret_log** (protegida) |
| **Grupo B** (Variables colineales) | VIF ≥ 5 (no protegidas) | precio, is_vol_missing, is_vol_zero, spread, rango_rel, gap_open, sigma_rolling_20, ret_abs_ema_20, parkinson_sigma_20, rs_sigma_20, EMA_5, EMA_10, EMA_20, price_over_SMA20, price_over_EMA20, SMA5_over_SMA20, ROC_10, MOM_20, ROC_20, RSI_14, z_20, trend_slope_20, trend_accel_20 |

Tabla 12. Diagnóstico de multicolinealidad mediante VIF.
Fuente: Cálculos propios con base a Investing y BVC (2025)

### 4.7 Reducción dimensional mediante PCA selectivo

Debido a que las redes neuronales son especialmente sensibles a la colinealidad y a escalas heterogéneas entre variables, se implementó un procedimiento de reducción dimensional mediante Análisis de Componentes Principales (PCA), aplicado exclusivamente al Grupo B. El PCA se ejecutó sobre las variables previamente estandarizadas y el número de componentes retenidos se definió a partir del criterio de explicar al menos el 90% de la varianza total. El proceso redujo 23 variables colineales a solo 8 componentes principales, los cuales capturan la estructura latente conjunta de tendencia, volatilidad y momentum.

Así pues, para concluir este proceso. Se evaluaron las 35 variables numéricas generadas en los bloques anteriores mediante el Factor de Inflación de Varianza (VIF). El análisis mostró la presencia de colinealidad extrema en un subconjunto importante de indicadores técnicos, especialmente aquellos derivados de medias móviles exponenciales, osciladores, volatilidad implícita y transformaciones basadas en precios. Sin embargo, con el fin de preservar completamente la información disponible para los modelos basados en árboles de decisión —los

cuales son intrínsecamente robustos a la escala y colinealidad— no se eliminó ninguna variable del conjunto original.

Para los modelos neuronales, cuya estabilidad numérica sí se ve afectada por la colinealidad, se adoptó una estrategia diferente que no implica eliminación sino transformación estructural. A partir del conjunto de 23 variables clasificadas como colineales (Grupo B), se aplicó un PCA selectivo que permitió reducirlas a 8 componentes ortogonales que explican más del 90% de la varianza acumulada.

De este modo, la información se conserva sin pérdida sustancial, pero se evita la redundancia algebraica que afecta la convergencia de redes neuronales. Las 12 variables clasificadas como no colineales (Grupo A) se mantuvieron sin modificaciones, incluyendo explícitamente la variable *ret_log*, que se protegió por su relevancia económica pese a presentar VIF elevado.

### 4.8 Definición del conjunto de entrenamiento y validación

Como se mencionó anteriormente, la división en conjuntos de entrenamiento y prueba se empleó un 70-15-15 con el fin de mantener uniformidad en resultados y de este modo, facilitar una comparación más certera con el modelo base. 70% modelo aprende patrones, 15% Validación y 15% Test.

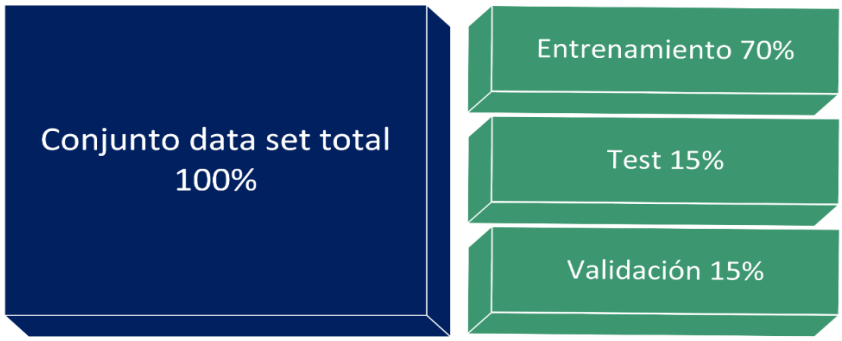


**Ilustración 5.** Separación de la data set.

Para la fase de modelado se implementó un esquema de partición temporal estricto, adecuado para series financieras y coherente con las mejores prácticas de validación en contextos donde no puede existir fuga de datos. La división se realizó por cada emisor, respetando el orden cronológico natural de la serie y evitando mezclar información futura en los conjuntos de entrenamiento o validación.

El procedimiento aplicado fue el siguiente:

| Paso | Definición |
|---|---|
| Ordenamiento temporal por emisor. | Todas las observaciones se organizaron por *emisor_key* y *fecha*, garantizando que los modelos únicamente aprendieran patrones a partir de información disponible en el pasado. |
| Eliminación de observaciones con objetivo no disponible. | Como el objetivo (*y_target*) es el retorno logarítmico del día siguiente (*t+1*), las últimas filas de cada emisor no pueden ser utilizadas y se removieron. |
| Aplicación de un warmup[6] causal de 20 días. | Para asegurar que las variables derivadas de ventanas móviles y EMAs estuvieran estabilizadas, se descartaron los primeros 20 días válidos de cada emisor. |

Tabla 12. Estadísticas descriptivas NUTRESA.
Fuente: Cálculos propios con base a Investing y BVC (2025)

Después de toda la depuración, construcción de indicadores y eliminación de colinealidad, el dataset final contiene 35 variables numéricas para entrenar los modelos.

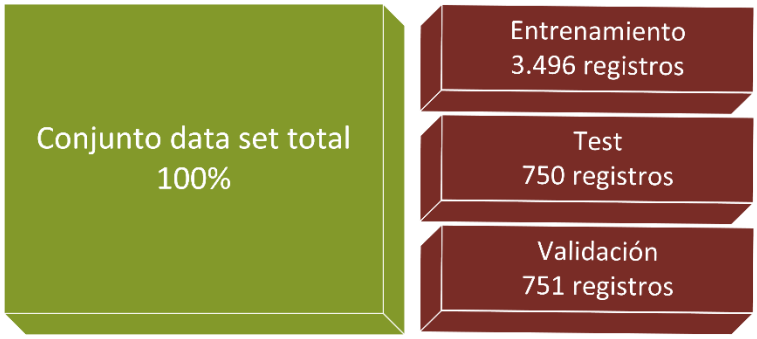


**Ilustración 6.** Separación de la data set en registros.

Seguidamente, se realizaron transformaciones de escalado exclusivamente sobre las variables numéricas (35 en total), manteniendo sin modificar las columnas de identificación. El escalado generó dos versiones del dataset:

- Una versión RAW (sin normalización) destinada a modelos basados en árboles, que no requieren normalización.
- Una versión STD (variables normalizadas) para modelos neuronales, particularmente LSTM, GRU y CNN, que dependen de magnitudes normalizadas para un aprendizaje estable.

[6] Elimina los primeros días donde los indicadores y ventanas móviles todavía no tienen suficiente historial para producir valores confiables

### 4.9 Creación de señales basadas en el análisis técnico fundamental.

En el proceso de modelaje predictivo se empleó la totalidad de las variables generadas y depuradas. Estas variables fueron incorporadas como predictores en los modelos de aprendizaje supervisado. Es decir, todas las variables numéricas finales, excepto la variable objetivo[7], fueron utilizadas como insumos en cada modelo, sin exclusiones adicionales. Esta decisión metodológica se justifica tanto desde el punto de vista estadístico como desde el enfoque de ingeniería de atributos adoptado en este proyecto.

En primer lugar, las variables de microestructura de mercado (spread, rango relativo, gap de apertura, cuerpo vela, range ratio) permiten capturar fricciones, asimetrías intradías y señales subyacentes que suelen anticipar cambios direccionales, lo que aporta información relevante para un modelo de retornos.

En segundo lugar, las transformaciones de volumen suministran medidas robustas del comportamiento transaccional, la presión de compra/venta y la actividad anómala del mercado, variables tradicionalmente asociadas con la evolución de retornos futuros.

Tercero, las medidas de incorporan información sobre la intensidad del riesgo y la magnitud esperada de los movimientos de precios, lo cual es fundamental en modelos orientados a predecir variaciones logarítmicas. Asimismo, los indicadores técnicos (EMAs, MACD, ROC, RSI, fuerza direccional, etc.) se integraron como señales derivadas que complementan la lectura del comportamiento del activo desde una perspectiva técnica ampliamente utilizada en finanzas cuantitativas.

| Variable | Tipo | Descripción |
|---|---|---|
| precio | Precio | Precio de cierre del activo. |
| vol | Volumen | Volumen transado diario. |
| ret_log | Retorno | Retorno logarítmico diario. |
| is_vol_missing | Flag | Indica si el volumen original venía faltante. |
| is_vol_zero | Flag | Indica si el volumen reportado es igual a cero. |
| is_repeated_price | Flag | Señala si el precio de cierre es igual al del día anterior. |
| days_since_last_change | Flag | Días consecutivos sin variación en precio. |
| is_outlier_ret | Flag | Retorno marcado como atípico mediante MAD. |
| spread | Microestructura | Diferencia entre máximo y mínimo diarios. |
| rango_rel | Microestructura | Spread relativo al precio (spread / precio). |
| gap_open | Microestructura | Diferencia entre apertura y cierre previo. |
| body_oc | Microestructura | Diferencia entre cierre y apertura. |

[7] Retorno logarítmico del activo.

| Variable | Tipo | Descripción |
|---|---|---|
| range_ratio | Microestructura | (Cierre − apertura) / (Máximo − mínimo). |
| log_vol | Volumen | log(1 + vol). |
| vol_ema | Volumen | Media móvil exponencial del volumen. |
| vol_surprise | Volumen | vol / vol_ema. |
| vol_z | Volumen | Z-score móvil del volumen. |
| sigma_rolling_20 | Volatilidad | Volatilidad móvil de 20 días (STD). |
| ret_abs_ema_20 | Volatilidad | EMA del retorno absoluto. |
| parkinson_sigma_20 | Volatilidad | Estimador de volatilidad de Parkinson. |
| rs_sigma_20 | Volatilidad | Estimador de Rogers–Satchell. |
| EMA_5 | Técnico | Media móvil exponencial de 5 días. |
| EMA_10 | Técnico | Media móvil exponencial de 10 días. |
| EMA_20 | Técnico | Media móvil exponencial de 20 días. |
| price_over_SMA20 | Técnico | Relación precio / SMA20. |
| price_over_EMA20 | Técnico | Relación precio / EMA20. |
| SMA5_over_SMA20 | Técnico | Relación SMA5 / SMA20. |
| MOM_10 | Técnico | Momentum de 10 días. |
| ROC_10 | Técnico | Tasa de cambio de 10 días. |
| MOM_20 | Técnico | Momentum de 20 días. |
| ROC_20 | Técnico | Tasa de cambio de 20 días. |
| RSI_14 | Técnico | Índice de fuerza relativa (14 días). |
| z_20 | Técnico | Z-score móvil del precio. |
| trend_slope_20 | Tendencia | Pendiente del log-precio (20 días). |
| trend_accel_20 | Tendencia | Aceleración de la tendencia (derivada segunda estimada). |

Tabla 13. Variables finales.

Fuente: Cálculos propios con base a Investing y BVC (2025)

La construcción del conjunto final de 35 variables permitió consolidar un panel robusto, homogéneo y estructuralmente adecuado para el modelado predictivo. Las variables abarcan dimensiones clave del comportamiento financiero lo que garantiza una representación amplia y bien balanceada de la dinámica diaria de los activos. Así pues, está la tabla previa expone un proceso de ingeniería de variables cuidadosamente diseñado, donde cada feature aporta una pieza distinta de información económica, técnica o microestructural.

## 4.10 Implementar y entrenar modelos de aprendizaje supervisado.

**Random Forest**

En este bloque se emplearon tres modelos con el fin de evaluar la capacidad predictiva sobre los retornos logarítmicos de las acciones y los precios reconstruidos. El primer modelo fue Naive0, un enfoque base que asume que el retorno futuro será igual a cero, es decir, que el precio del activo permanecerá constante. El segundo modelo utilizado fue Persist, un baseline que supone que el retorno del siguiente periodo será similar al retorno observado en el periodo actual. Y finalmente, el modelo principal estimado fue un Random Forest Regressor, un algoritmo de aprendizaje automático no lineal basado en el ensamblaje de múltiples árboles de decisión.

Se ejecutó una búsqueda en rejilla (Grid Search) para identificar la combinación que minimizara el error absoluto medio (MAE) sobre el conjunto de validación, evaluando un total de 54 configuraciones.

| Hiperparámetros | Valor |
|---|---|
| n_estimators | 200, 400, 800 |
| max_depth | None, 5, 10 |
| max_features | sqrt, log2 |
| min_samples_leaf | 1, 3 y 5 |

Tabla 14. Hiperparámetros del modelo Random Forest inicial
Fuente: Cálculos propios.

El modelo óptimo correspondió a un Random Forest poco profundo (max_depth=5), con 400 árboles y selección aleatoria de predictores por raíz cuadrada; una estructura que favorece generalización y menor sobreajuste.

| Hiperparámetros | Valor |
|---|---|
| n_estimators | 400 |
| max_depth | 5 |
| max_features | sqrt |
| min_samples_leaf | 1 |

Tabla 15. Hiperparámetros del modelo Random Forest final.
Fuente: Cálculos propios.

Se utilizó una división temporal en TRAIN, VALIDATION y TEST para evitar fuga de información y garantizar una evaluación realista del desempeño predictivo. Sin embargo, los resultados obtenidos muestran que, en el conjunto de prueba (TEST), el modelo Random Forest no supera consistentemente a los modelos base Naive0 y Persist en la predicción de retornos logarítmicos. Esto confirma la dificultad inherente a predecir retornos diarios, los cuales permanecen dominados por ruido de mercado.

En la reconstrucción de precios, el modelo Random Forest obtiene un desempeño comparable, pero no superior al modelo más simple (Naive0), lo cual es consistente con la literatura financiera que señala la baja predictibilidad de los retornos en horizontes cortos y la eficiencia relativa de los mercados. Además, la presencia de sobreajuste es evidente al observar que el rendimiento del modelo en entrenamiento es superior al observado en prueba, sugiriendo que la complejidad del Random Forest no se traduce en mayor capacidad predictiva.

| Split | Emisor | Modelo | MAE | RMSE | $R^2$ | HitRate |
|---|---|---|---|---|---|---|
| TEST | Canacol | Naive0 | 0.023299 | 0.044623 | -0.004890 | 0.118343 |
| | | Persist | 0.033189 | 0.057550 | -0.671432 | 0.414201 |
| | | RandomForest | 0.025271 | 0.045937 | -0.064955 | 0.378698 |
| | Davivienda | Naive0 | 0.011052 | 0.015587 | -0.010892 | 0.094118 |
| | | Persist | 0.015709 | 0.022076 | -1.027611 | 0.470588 |
| | | RandomForest | 0.011246 | 0.015717 | -0.027786 | 0.423529 |
| | ETB | Naive0 | 0.002924 | 0.012125 | -0.002269 | 0.930693 |
| | | Persist | 0.005848 | 0.017147 | -1.004538 | 0.861386 |
| | | RandomForest | 0.007028 | 0.014616 | -0.456446 | 0.049505 |
| | Nutresa | Naive0 | 0.011317 | 0.021690 | -0.033785 | 0.473333 |
| | | Persist | 0.018208 | 0.029310 | -0.887756 | 0.446667 |
| | | RandomForest | 0.026570 | 0.057255 | -6.203483 | 0.293333 |
| | Suramericana | Naive0 | 0.013427 | 0.023298 | -0.001204 | 0.105590 |
| | | Persist | 0.019192 | 0.030710 | -0.739528 | 0.465839 |
| | | RandomForest | 0.014323 | 0.025171 | -0.168682 | 0.459627 |
| TRAIN | Canacol | Naive0 | 0.015545 | 0.023566 | -0.004783 | 0.063532 |
| | | Persist | 0.021581 | 0.031414 | -0.785386 | 0.448539 |
| | | RandomForest | 0.015589 | 0.022942 | 0.047742 | 0.513342 |
| | Davivienda | Naive0 | 0.024592 | 0.219612 | -0.000006 | 0.063291 |
| | | Persist | 0.041215 | 0.380350 | -1.999552 | 0.478481 |
| | | RandomForest | 0.018554 | 0.098856 | 0.797372 | 0.502532 |
| | ETB | Naive0 | 0.015246 | 0.029452 | -0.005895 | 0.443737 |
| | | Persist | 0.025633 | 0.044474 | -1.293767 | 0.435244 |
| | | RandomForest | 0.015288 | 0.028646 | 0.048364 | 0.367304 |
| | Nutresa | Naive0 | 0.018484 | 0.037422 | -0.000744 | 0.053009 |
| | | Persist | 0.028160 | 0.052512 | -0.970530 | 0.441261 |
| | | RandomForest | 0.017753 | 0.033978 | 0.174993 | 0.565903 |
| | Suramericana | Naive0 | 0.019545 | 0.031559 | -0.000344 | 0.081333 |
| | | Persist | 0.029357 | 0.045819 | -1.108639 | 0.457333 |
| | | RandomForest | 0.018796 | 0.028240 | 0.199010 | 0.528000 |

Tabla 16. Métricas promedio por emisor Random Forest (Retornos, TRAIN y TEST)
Fuente: Cálculos propios.

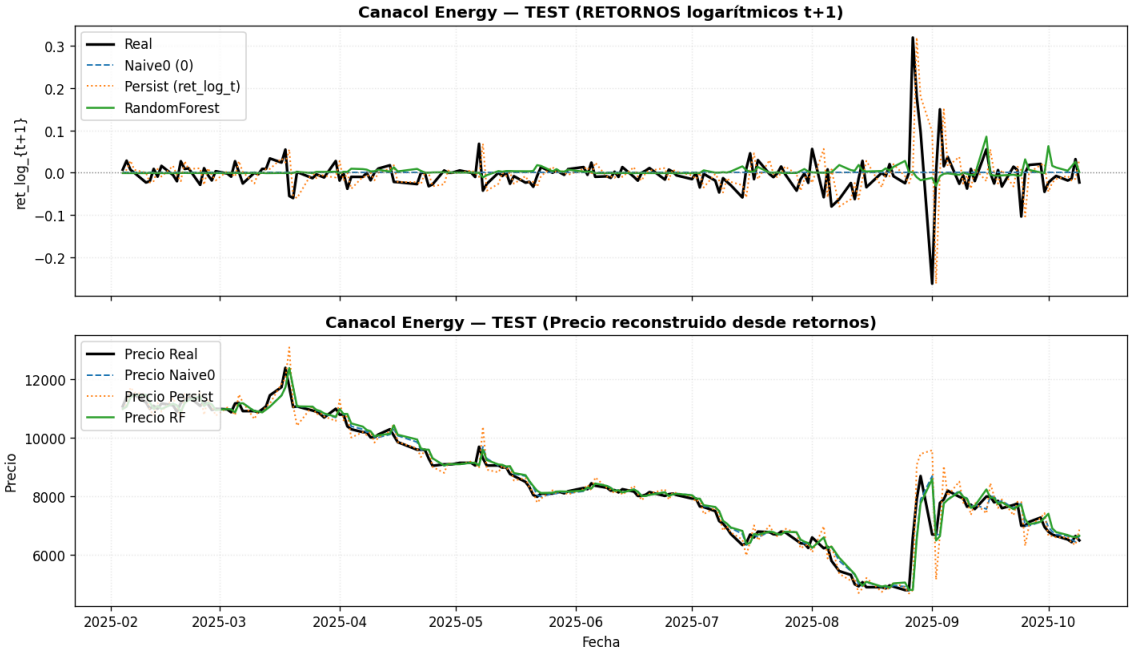


**Figura 9.** Comparación modelos de retornos y precios para CANACOL.

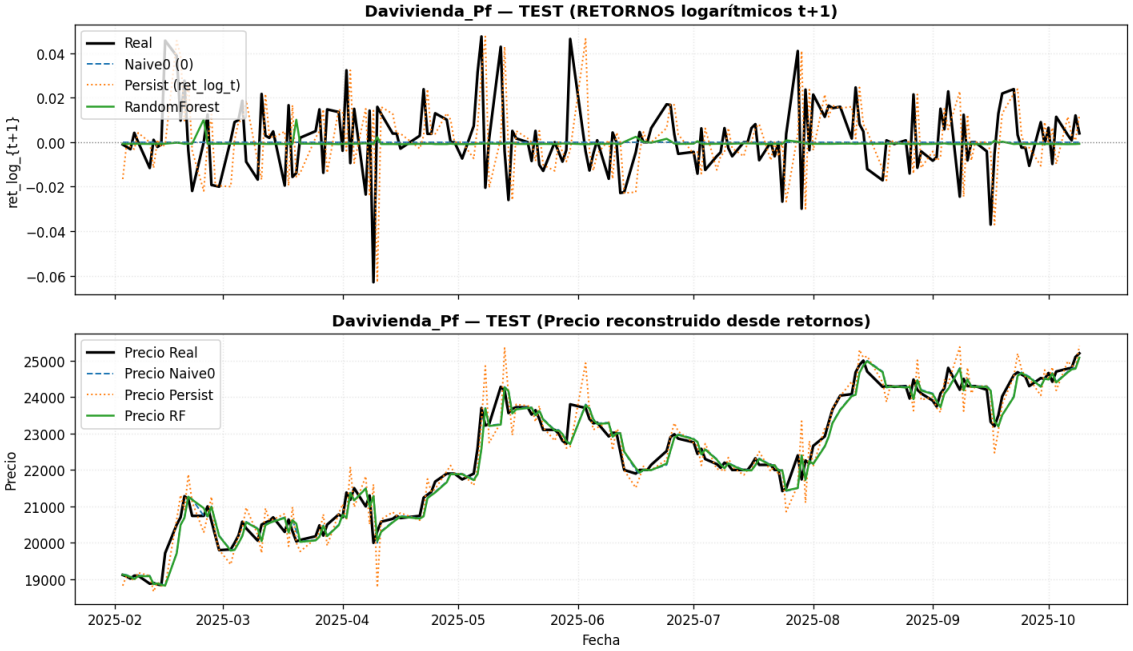


**Figura 10.** Comparación modelos de retornos y precios para DAVIVIENA.

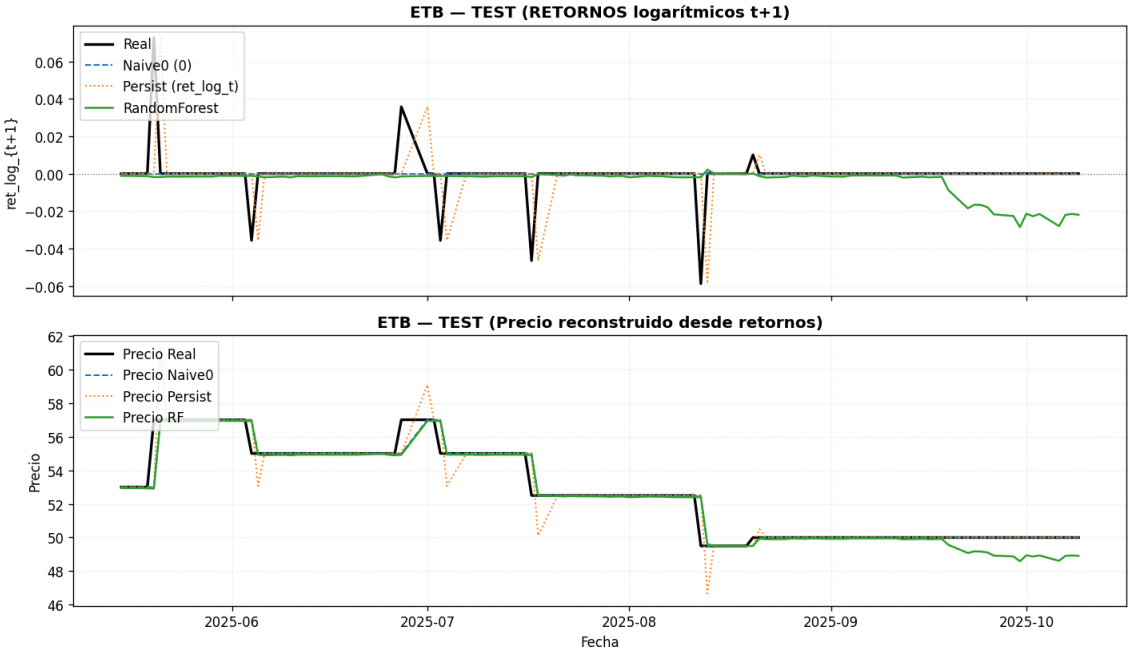


**Figura 11.** Comparación modelos de retornos y precios para ETB.

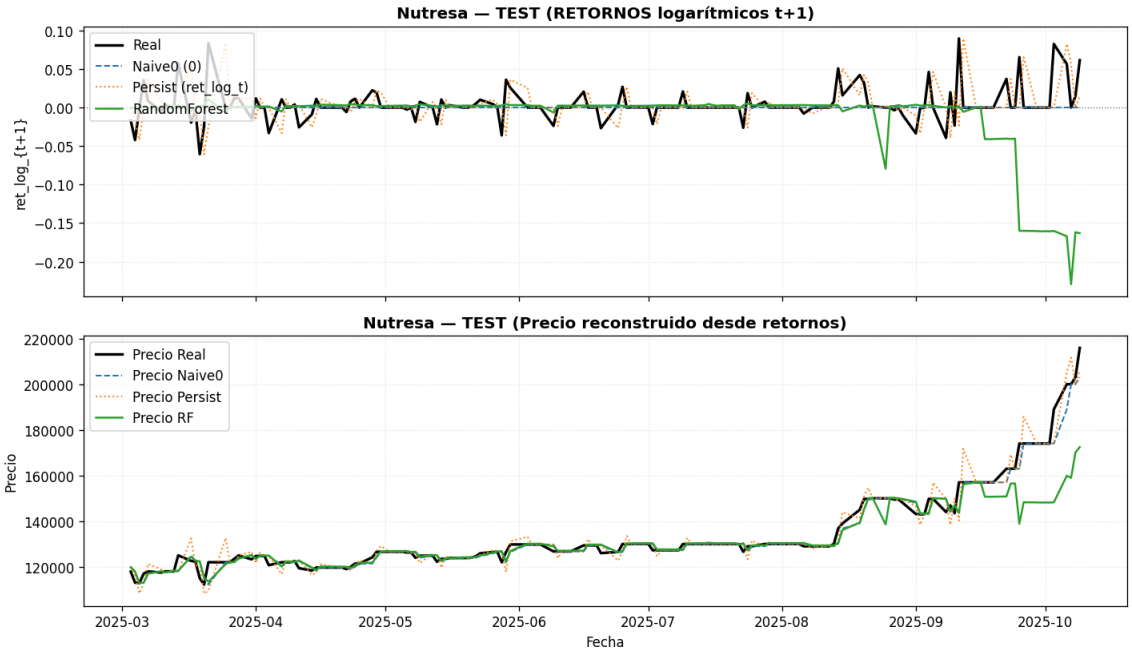


**Figura 12.** Comparación modelos de retornos y precios para NUTRESA.

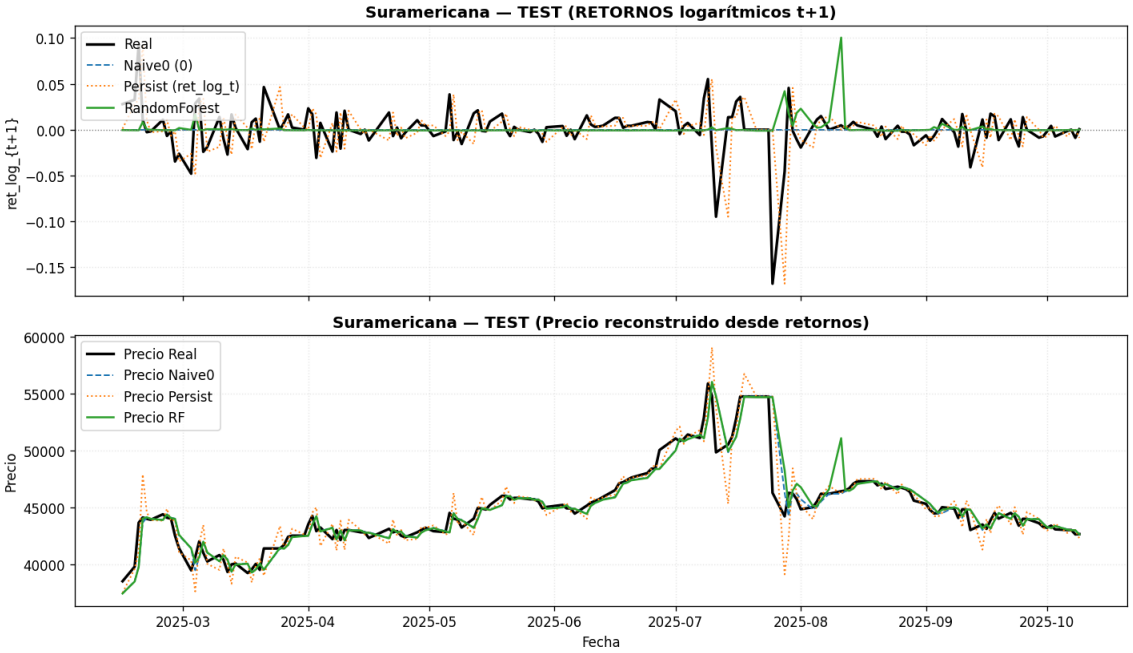


**Figura 13.** Comparación modelos de retornos y precios para SURAMERICANA.

### XGBoost

En este modelaje se estimó un algoritmo de gradiente reforzado del tipo XGBoost Regressor para la predicción de los retornos logarítmicos. Se utilizaron los mismos conjuntos definidos en el pipeline de árboles, conservando su estructura temporal y evitando cualquier fuga de información. A diferencia del Random Forest, la calibración del XGBoost se llevó a cabo mediante una búsqueda aleatoria de hiperparámetros (Random Search). Dicho espacio incluyó parámetros clave como el número de árboles, la profundidad máxima, la tasa de aprendizaje, el muestreo por filas y columnas, y las penalizaciones L1 y L2.

| Hiperparámetros | Valor |
|---|---|
| n_estimators | 200, 400, 600, 800 |
| max_depth | 3, 4, 5, 6 |
| learning_rate | 0.01, 0.03, 0.05, 0.1 |
| subsample | 0.7, 0.8, 0.9, 1.0 |
| colsample_bytree | 0.6, 0.8, 1.0 |
| min_child_weight | 1, 3, 5, 10 |
| reg_lambda (L2) | 0.0, 1.0, 5.0, 10.0 |
| reg_alpha (L1) | 0.0, 0.1, 0.5 |
| min_samples_leaf | 1, 3 y 5 |

Tabla 17. Hiperparámetros iniciales para XGBoost inicial
Fuente: Cálculos propios.

Con los hiperparámetros óptimos obtenidos en la búsqueda, el modelo XGBoost se reentrenó utilizando la combinación TRAIN + VALID y posteriormente fue evaluado sobre el conjunto de prueba (TEST). Su desempeño se comparó frente a dos modelos base: Naive0 (retorno nulo) y Persist (retorno del periodo anterior). Al igual que en el caso del Random Forest, se reportaron métricas tanto para retornos (MAE, RMSE, $R^2$ y Hit Rate) como para los precios reconstruidos a partir de dichos retornos.

| Hiperparámetros | Valor |
|---|---|
| subsample | 0.7 |
| reg_lambda (L2) | 10.0 |
| reg_alpha (L1) | 0.5 |
| n_estimators | 400 |
| min_child_weight | 1 |
| max_depth | 4 |
| learning_rate | 0.01 |
| colsample_bytree | 0.6 |
| MAE VALID | 0.015776 |

Tabla 18. Hiperparámetros del modelo XGBoost final.
Fuente: Cálculos propios.

Finalmente, se presentan las tablas de métricas por emisor y modelo (MAE, RMSE, $R^2$ y Hit Rate), así como los resultados agregados para retornos, que permiten comparar el desempeño del XGBoost respecto a los modelos base.

La comparación entre modelos evidencia que XGBoost no supera de manera consistente a los modelos base en la predicción de retornos logarítmicos. En el conjunto TEST, su desempeño es intermedio: mejora ligeramente al modelo Persist en la mayoría de emisores, pero rara vez supera al modelo Naive0, el cual obtiene errores similares o inferiores en varios casos. Asimismo, las métricas negativas de $R^2$ indican que ninguno de los modelos logra capturar adecuadamente la variabilidad real de los retornos. Sin embargo, en el conjunto TRAIN, XGBoost muestra una mejor capacidad de ajuste (MAE y RMSE ligeramente menores y $R^2$ positivos en algunos activos), lo que confirma que aprende patrones locales, pero estos no se traducen en capacidad predictiva fuera de muestra, revelando un problema de overfitting típico en mercados financieros.

| Split | Emisor | Modelo | MAE | RMSE | $R^2$ | HitRate |
|---|---|---|---|---|---|---|
| TEST | Canacol | Naive0 | 0.023299 | 0.044623 | -0.004890 | 0.118343 |
| | | Persist | 0.033189 | 0.057550 | -0.671432 | 0.414201 |
| | | XGBoost | 0.025584 | 0.048284 | -0.176506 | 0.420118 |
| | Davivienda | Naive0 | 0.011052 | 0.015587 | -0.010892 | 0.094118 |
| | | Persist | 0.015709 | 0.022076 | -1.027611 | 0.470588 |
| | | XGBoost | 0.011080 | 0.015512 | -0.001108 | 0.458824 |
| | ETB | Naive0 | 0.002924 | 0.012125 | -0.002269 | 0.930693 |
| | | Persist | 0.005848 | 0.017147 | -1.004538 | 0.861386 |
| | | XGBoost | 0.005461 | 0.012281 | -0.028198 | 0.039604 |
| | Nutresa | Naive0 | 0.011317 | 0.021690 | -0.033785 | 0.473333 |
| | | Persist | 0.018208 | 0.029310 | -0.887756 | 0.446667 |
| | | XGBoost | 0.011857 | 0.021579 | -0.023238 | 0.253333 |
| | Suramericana | Naive0 | 0.013427 | 0.023298 | -0.001204 | 0.105590 |
| | | Persist | 0.019192 | 0.030710 | -0.739528 | 0.465839 |
| | | XGBoost | 0.013858 | 0.023641 | -0.030878 | 0.465839 |
| TRAIN | Canacol | Naive0 | 0.016346 | 0.024385 | -0.004371 | 0.077406 |
| | | Persist | 0.023047 | 0.033192 | -0.860898 | 0.442469 |
| | | XGBoost | 0.016388 | 0.024268 | 0.005216 | 0.516736 |
| | Davivienda | Naive0 | 0.022168 | 0.199415 | -0.000008 | 0.070907 |
| | | Persist | 0.036703 | 0.345310 | -1.998538 | 0.468196 |
| | | XGBoost | 0.020203 | 0.160397 | 0.353029 | 0.539103 |
| | ETB | Naive0 | 0.014102 | 0.030458 | -0.006459 | 0.524476 |
| | | Persist | 0.024006 | 0.046479 | -1.343720 | 0.505245 |
| | | XGBoost | 0.014744 | 0.029616 | 0.048406 | 0.325175 |
| | Nutresa | Naive0 | 0.017716 | 0.035802 | -0.002739 | 0.130896 |
| | | Persist | 0.027227 | 0.050361 | -0.984078 | 0.438679 |
| | | XGBoost | 0.017254 | 0.034153 | 0.087471 | 0.524764 |
| | Suramericana | Naive0 | 0.018863 | 0.030272 | -0.000310 | 0.084523 |
| | | Persist | 0.028263 | 0.043683 | -1.082968 | 0.451153 |
| | | XGBoost | 0.018370 | 0.028724 | 0.099356 | 0.544457 |

Tabla 19. Métricas promedio por emisor y modelo XGBoost (Retornos, TRAIN y TEST)
Fuente: Cálculos propios.

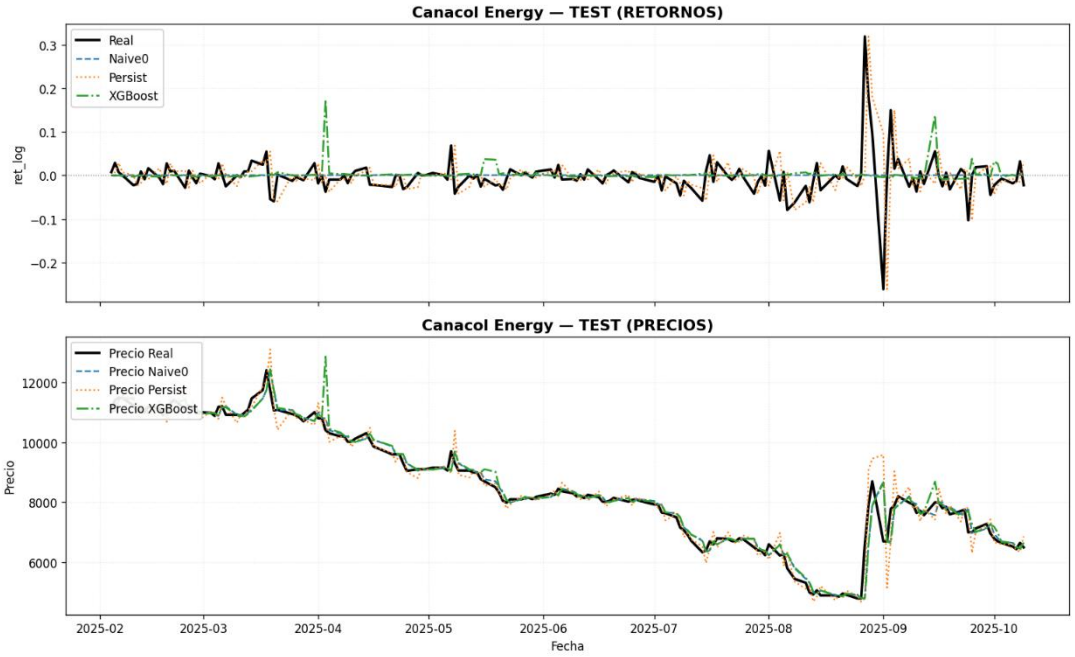


**Figura 14.** Comparación modelos de retornos y precios para CANACOL.

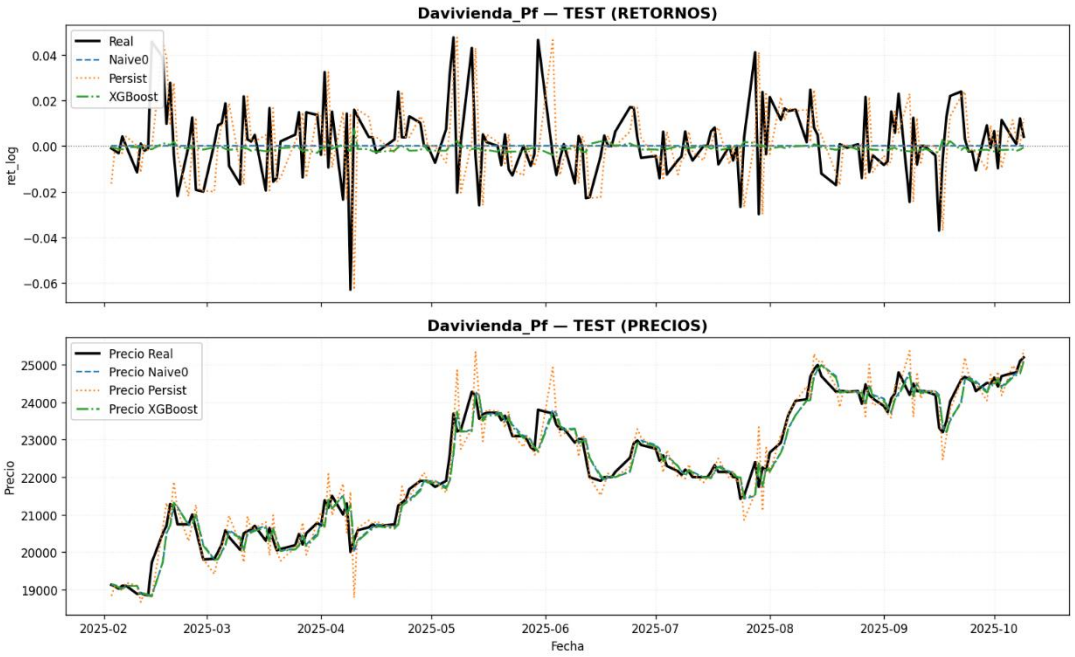


**Figura 15.** Comparación modelos de retornos y precios para DAVIVIENDA.

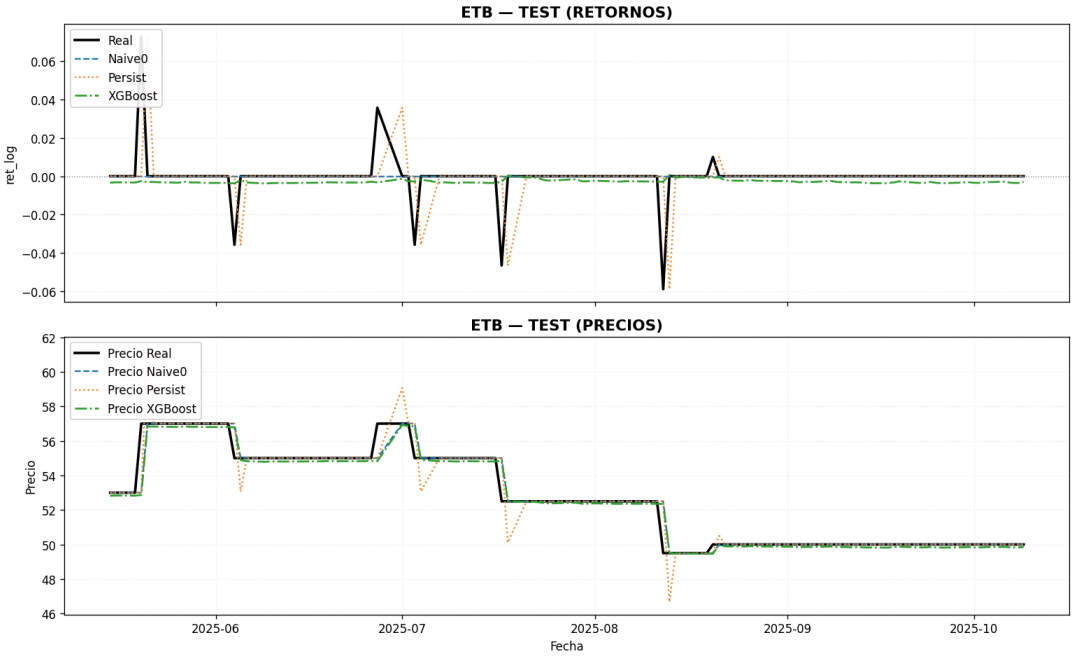


**Figura 16.** Comparación modelos de retornos y precios para ETB.

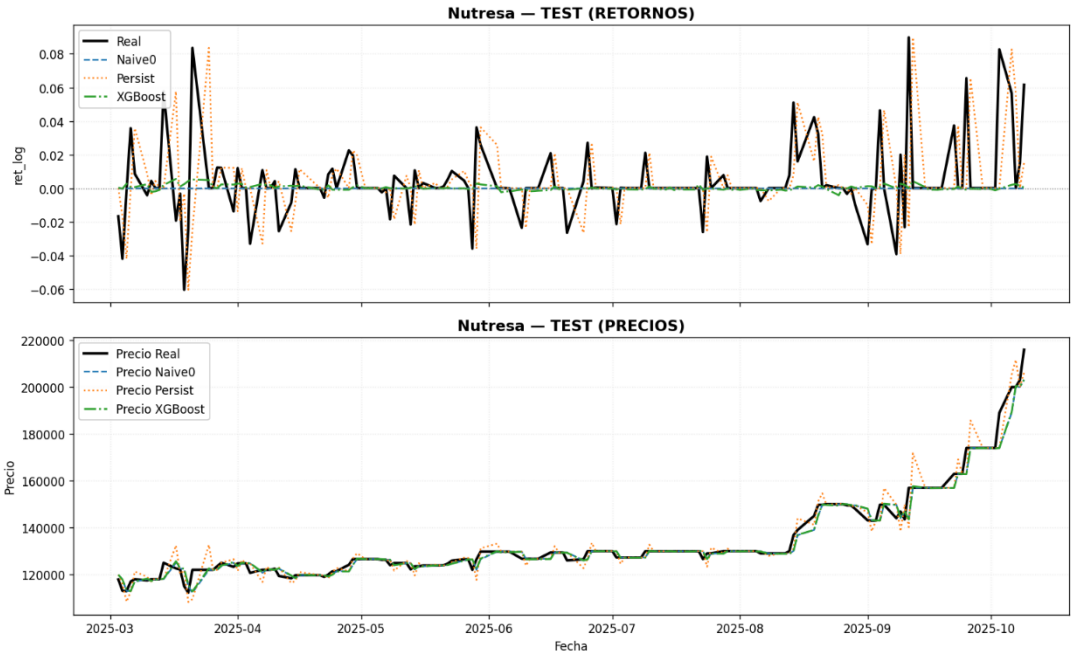


**Figura 17.** Comparación modelos de retornos y precios para NUTRESA.

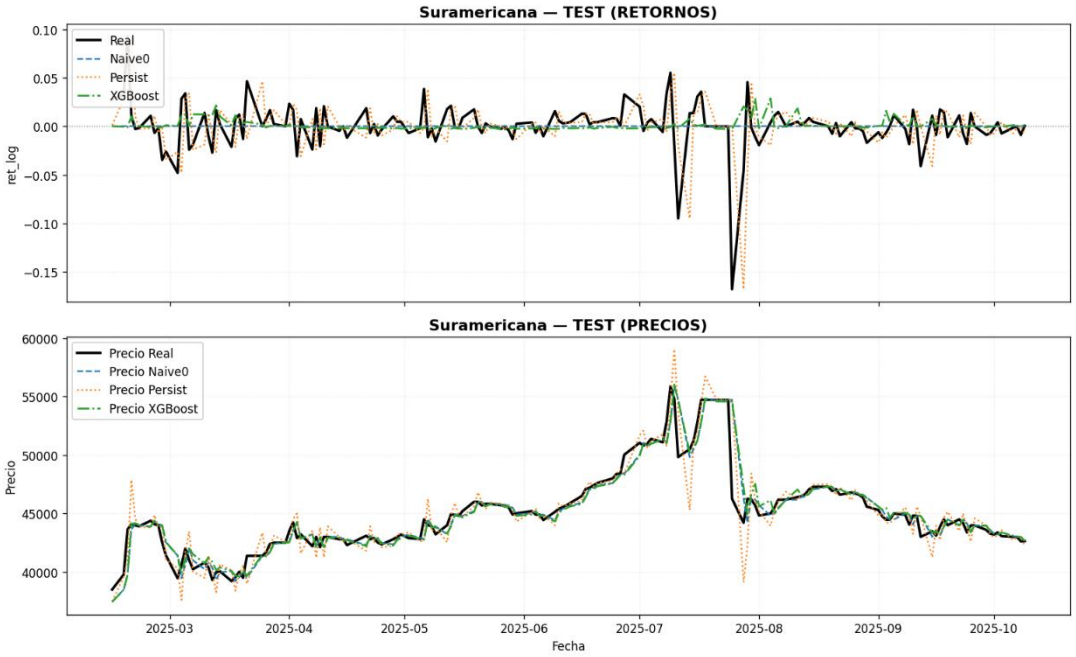


**Figura 18.** Comparación modelos de retornos y precios para SURAMERICANA.

**LightGBM**

En una tercera etapa se estimó un modelo de gradiente reforzado basado en árboles del tipo LightGBM Regressor para la predicción de los retornos logarítmicos. Se reutilizaron los mismos conjuntos de entrenamiento, validación y prueba (TRAIN, VALID y TEST) definidos para los modelos basados en árboles, preservando la estructura temporal de la serie.

La calibración del modelo se realizó mediante una búsqueda exhaustiva reducida (grid search) sobre un conjunto de hiperparámetros que incluía el número de hojas por árbol (num_leaves), profundidad máxima, tasa de aprendizaje (learning rate), número de árboles (n_estimators), proporción de filas y columnas utilizadas en cada iteración (subsample y colsample_bytree), el tamaño mínimo de muestras por hoja (min_child_samples) y penalizaciones L1 y L2 sobre los pesos del modelo (reg_alpha y reg_lambda). En total se evaluaron 60 combinaciones de parámetros y el criterio de selección fue la minimización del error absoluto medio (MAE) en el conjunto de validación.

| Hiperparámetros | Valor |
|---|---|
| num_leaves | 31, 63 |
| max_depth | -1, 4, 6 |
| learning_rate | 0.01, 0.05 |
| n_estimators | 200, 400 |
| subsample | 0.7, 1.0 |
| colsample_bytree | 0.7, 1.0 |
| min_child_samples | 20, 40 |
| reg_lambda (L2) | 0.0, 5.0 |
| reg_alpha (L1) | 0.0, 0.5 |

Tabla 20. Hiperparámetros iniciales para LightGBM inicial
Fuente: Cálculos propios.

El mejor modelo obtenido presentó un MAE de validación de 0,01999 y se caracterizó por una configuración relativamente conservadora: 200 árboles, tasa de aprendizaje baja (0,01), 31 hojas por árbol, sin límite explícito de profundidad máxima, y regularización moderada (penalización L2 igual a 5,0 y L1 igual a 0,5). Con estos hiperparámetros, el modelo final se reentrenó sobre la combinación TRAIN + VALID y posteriormente se evaluó en el conjunto de prueba, comparando su desempeño frente a dos *baselines* sencillos: Naive0 (retorno nulo) y Persist (retorno igual al del periodo previo).

| Hiperparámetros | Valor |
|---|---|
| num_leaves | 31 |
| max_depth | -1 |
| learning_rate | 0.01 |
| n_estimators | 200 |
| subsample | 0.7 |
| colsample_bytree | 0.7 |
| min_child_samples | 40 |
| reg_lambda (L2) | 5.0 |
| reg_alpha (L1) | 0.5 |

Tabla 21. Hiperparámetros iniciales para LightGBM final
Fuente: Cálculos propios.

| Split | Emisor | Modelo | MAE | RMSE | $R^2$ | HitRate |
|---|---|---|---|---|---|---|
| TRAIN | Canacol | Naive0 | 0.015545 | 0.023566 | -0.004783 | 0.000000 |
| | | Persist | 0.021574 | 0.031413 | -0.785255 | 0.440915 |
| | | LightGBM | 0.015547 | 0.023161 | 0.029498 | 0.561626 |
| | Davivienda | Naive0 | 0.024592 | 0.219612 | -0.000006 | 0.000000 |
| | | Persist | 0.041221 | 0.380350 | -1.999556 | 0.468354 |
| | | LightGBM | 0.024149 | 0.211987 | 0.068227 | 0.574684 |
| | ETB | Naive0 | 0.015246 | 0.029452 | -0.005895 | 0.000000 |
| | | Persist | 0.025562 | 0.044448 | -1.291043 | 0.178344 |
| | | LightGBM | 0.016050 | 0.028032 | 0.088736 | 0.411890 |
| | Nutresa | Naive0 | 0.018484 | 0.037422 | -0.000744 | 0.000000 |
| | | Persist | 0.028158 | 0.052512 | -0.970517 | 0.439828 |
| | | LightGBM | 0.018271 | 0.035153 | 0.116943 | 0.601719 |
| | Suramericana | Naive0 | 0.019545 | 0.031559 | -0.000344 | 0.000000 |
| | | Persist | 0.029317 | 0.045805 | -1.107297 | 0.442667 |
| | | LightGBM | 0.018618 | 0.029339 | 0.135446 | 0.612000 |
| TEST | Canacol | Naive0 | 0.023299 | 0.044623 | -0.004890 | 0.000000 |
| | | Persist | 0.033178 | 0.057549 | -0.671344 | 0.402367 |
| | | LightGBM | 0.025279 | 0.046264 | -0.080160 | 0.443787 |
| | Davivienda | Naive0 | 0.011052 | 0.015587 | -0.010892 | 0.000000 |
| | | Persist | 0.015624 | 0.022044 | -1.021732 | 0.441176 |
| | | LightGBM | 0.011121 | 0.015525 | -0.002837 | 0.452941 |
| | ETB | Naive0 | 0.002924 | 0.012125 | -0.002269 | 0.000000 |
| | | Persist | 0.005848 | 0.017147 | -1.004538 | 0.000000 |
| | | LightGBM | 0.007161 | 0.014685 | -0.470179 | 0.039604 |
| | Nutresa | Naive0 | 0.011317 | 0.021690 | -0.033785 | 0.000000 |
| | | Persist | 0.018208 | 0.029310 | -0.887756 | 0.166667 |
| | | LightGBM | 0.017581 | 0.027198 | -0.625477 | 0.326667 |
| | Suramericana | Naive0 | 0.013427 | 0.023298 | -0.001204 | 0.000000 |
| | | Persist | 0.019199 | 0.030716 | -0.740199 | 0.434783 |
| | | LightGBM | 0.015137 | 0.023926 | -0.055893 | 0.484472 |

Tabla 22. Métricas promedio por emisor y modelo LightGBM (Retornos, TRAIN y TEST)
Fuente: Cálculos propios.

Las métricas obtenidas muestran que LightGBM supera consistentemente al modelo Persist y, en la mayoría de los casos, iguala o mejora ligeramente al baseline Naive0 en términos de MAE y RMSE, especialmente en el conjunto de entrenamiento. En emisores como Davivienda, Nutresa y Suramericana, LightGBM logra las mayores mejoras en $R^2$ y en la tasa de aciertos (Hit Rate), evidenciando una capacidad moderada para capturar dirección y magnitud del retorno. Sin embargo, en emisores de muy baja volatilidad como ETB, el modelo tiende a sub ajustarse y presenta un desempeño inferior al baseline, lo cual es esperable cuando las variaciones son tan pequeñas que un predictor casi constante resulta competitivo.

En general, LightGBM presenta un comportamiento estable y menos propenso al sobreajuste que Persist, logrando $R^2$ positivos en TRAIN y reduciendo la pérdida relativa en TEST. Aunque Naive0 mantiene un MAE competitivo debido a la baja magnitud de los retornos diarios, LightGBM ofrece ventajas en interpretación, estabilidad y capacidad de aprendizaje, mostrando que los modelos de gradiente reforzado pueden capturar estructura adicional en los retornos que los baselines no identifican.

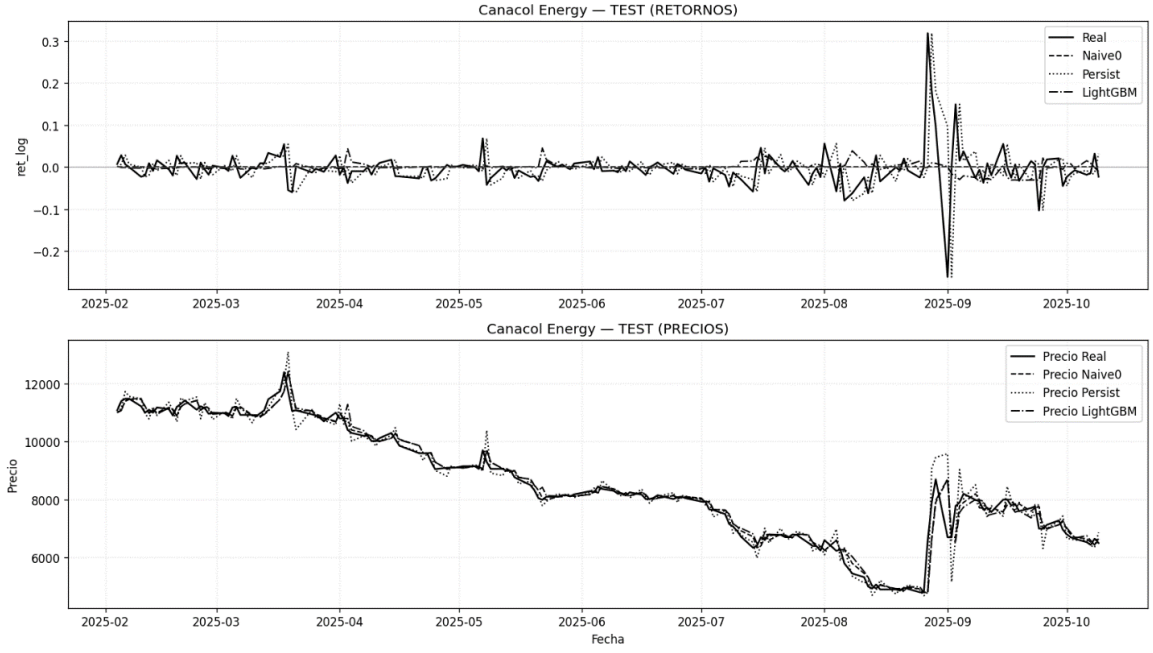


**Figura 19.** Comparación modelos de retornos y precios para CANACOL.

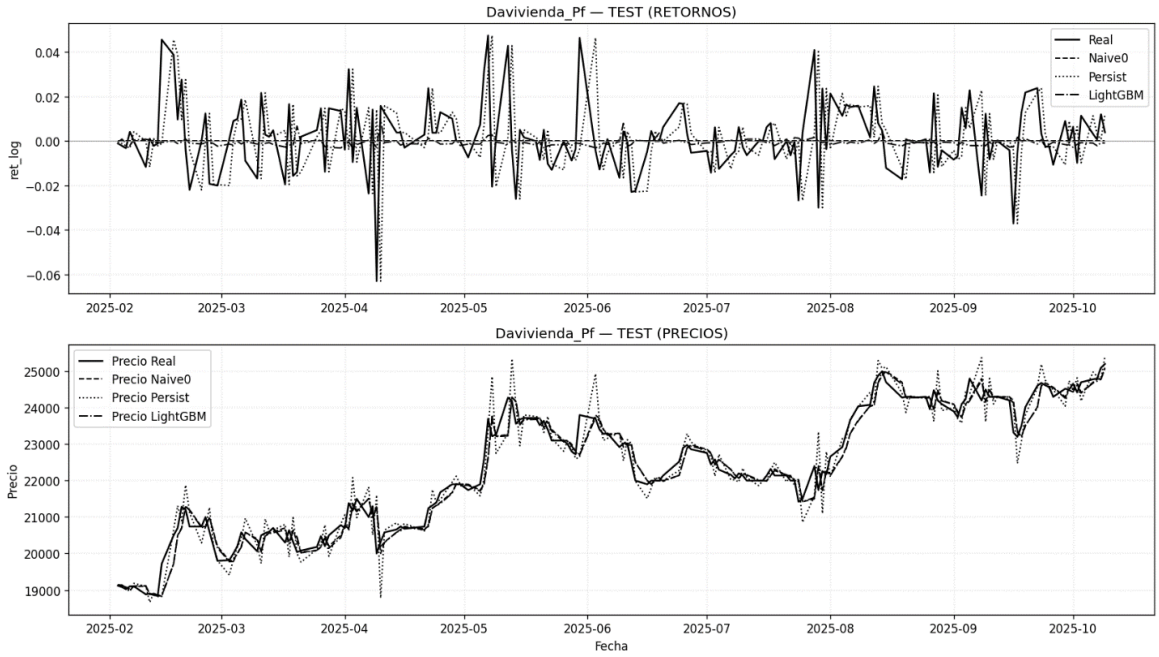


**Figura 20.** Comparación modelos de retornos y precios para DAVIVIENDA.

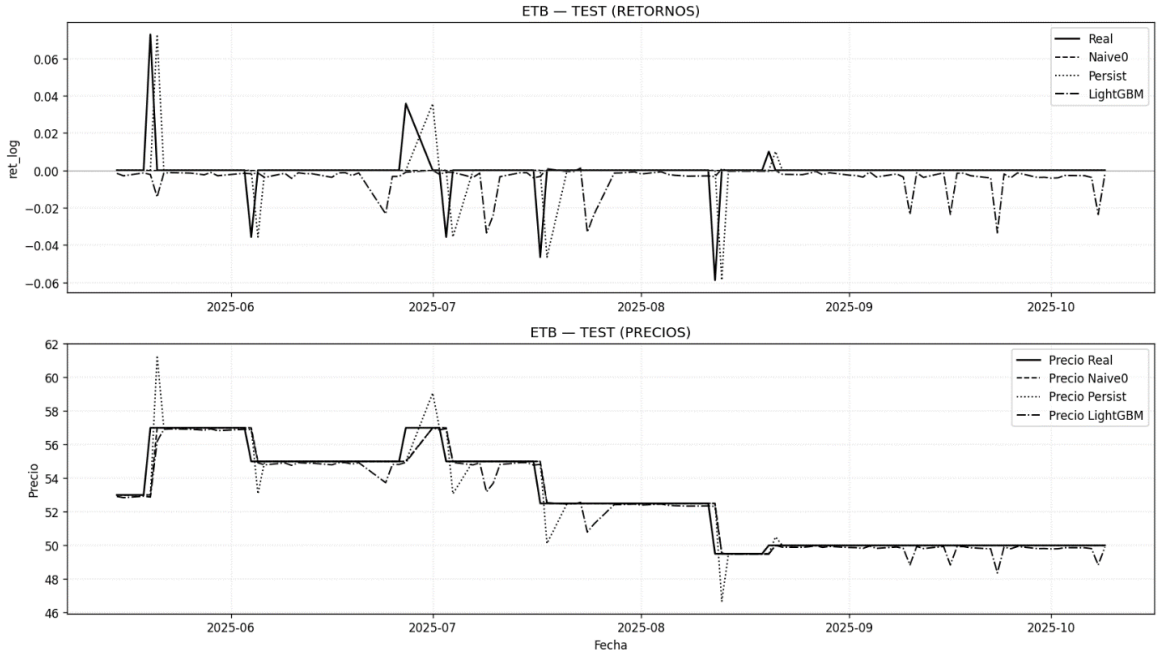


**Figura 21.** Comparación modelos de retornos y precios para ETB.

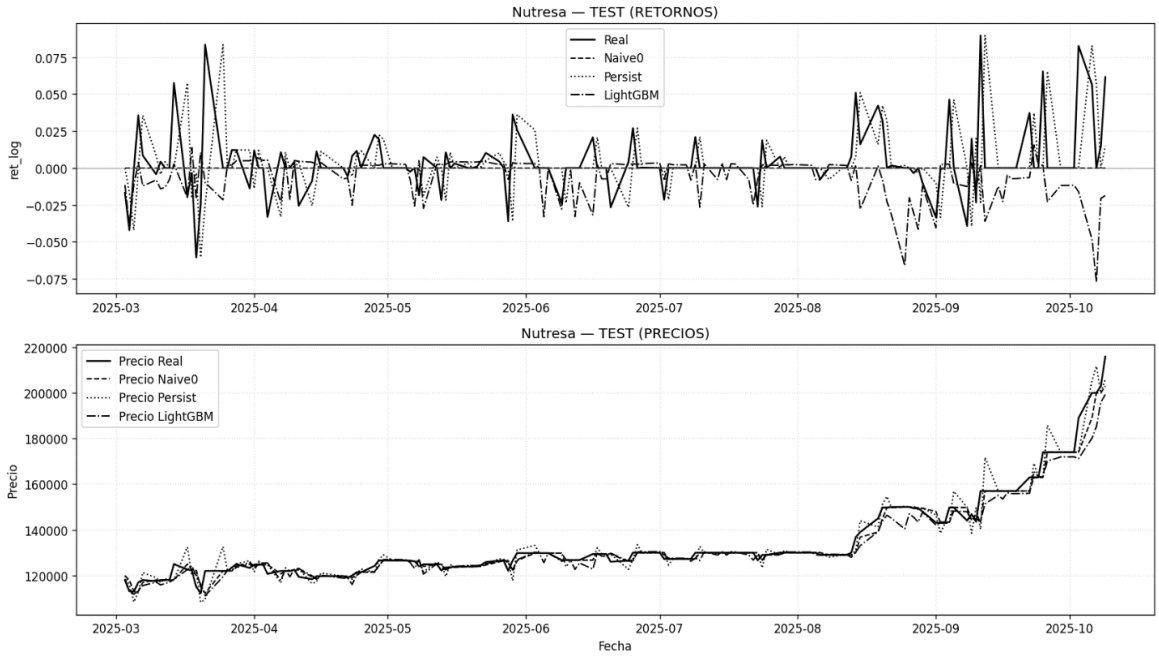


**Figura 22.** Comparación modelos de retornos y precios para NUTRESA.

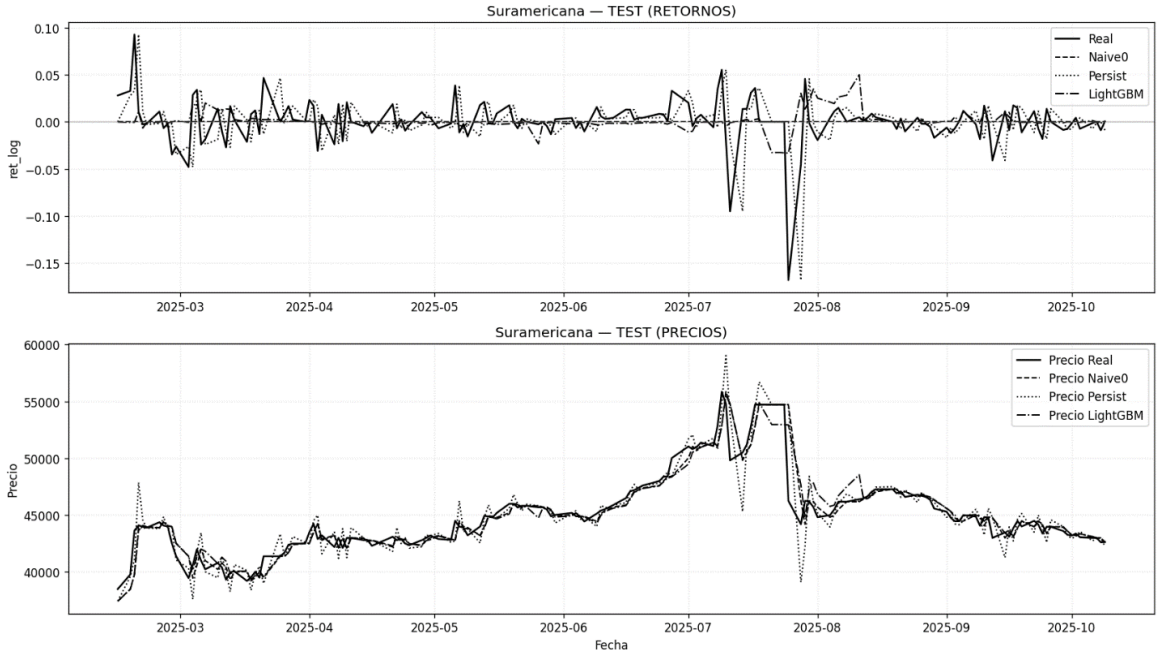


**Figura 23.** Comparación modelos de retornos y precios para SURAMERICANA.

**LSTM**

El modelaje con redes neuronales de tipo LSTM se incorporó como una etapa avanzada del proceso predictivo, con el propósito de capturar dependencias temporales no lineales en los retornos logarítmicos de cada activo. A diferencia de los modelos de aprendizaje basado en árboles como Random Forest, XGBoost y LightGBM, las LSTM permiten procesar secuencias históricas completas, aprovechando la memoria interna de sus celdas para modelar dinámicas persistentes y patrones de largo plazo que no son visibles para los modelos estacionarios. Por ello, se optó por entrenar un modelo LSTM independiente por emisor, manteniendo así la idiosincrasia de cada activo.

Para construir el conjunto de entrada del LSTM, se utilizaron ventanas temporales deslizantes de longitud LOOKBACK que agregan múltiples rezagos de los retornos y variables financieras relacionadas. Cada ventana constituye una secuencia con forma (n_timesteps × n_features), donde la variable objetivo es el retorno logarítmico t+1. Se generaron particiones por emisor para TRAIN, VALID y TEST de modo estrictamente temporal, garantizando ausencia total de leakage en el entrenamiento.

La calibración de cada modelo se realizó mediante una búsqueda aleatoria de hiperparámetros (Random Search), explorando combinaciones que incluyen: número de unidades LSTM en la primera capa, presencia o no de una segunda capa recurrente, tamaño de la capa densa intermedia, tasas de dropout y recurrent dropout, penalización L2, tasa de aprendizaje (learning rate) y tamaño de lote (batch size). Para cada emisor se evaluaron 20 configuraciones distintas, seleccionando la que minimizó el error absoluto medio (MAE) en el conjunto de validación. Con los hiperparámetros óptimos, se reentrenó el modelo final sobre TRAIN+VALID para su evaluación definitiva en TEST.

| | Hiperparámetros | Valor |
|---|---|---|
| 1 | n_units_lstm1 | 32, 64, 96 |
| 2 | use_second_lstm | False, True |
| 3 | n_units_lstm2 | 16, 32, 64 |
| 4 | dense_units | 0, 16, 32 |
| 5 | dropout | 0.0, 0.1, 0.2, 0.3 |
| 6 | recurrent_dropout | 0.0, 0.1, 0.2 |
| 7 | learning_rate | 0.01, 0.005, 0.001 |
| 8 | batch_size | 16, 32, 64 |
| 9 | l2_reg | 0.0, $1\times10^{-4}$, $1\times10^{-3}$ |

Tabla 23. Hiperparámetros iniciales para LSTM inicial
Fuente: Cálculos propios.

Los modelos resultaron heterogéneos entre emisores, reflejando diferencias en volatilidad, estacionalidad y profundidad histórica. Por ejemplo, Canacol Energy convergió hacia una arquitectura relativamente simple (una sola LSTM de 64 unidades y dropout moderado), mientras que Davivienda_Pf y Nutresa favorecieron arquitecturas más profundas con dos capas LSTM. Estos resultados evidencian que cada activo financiero posee dinámicas temporales singulares que requieren arquitecturas especializadas.

| Emisor | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| CANACOL | 64 | No | 0 | 32 | 0,20 | 0,00 | 0,01 | 64 | 0,0001 |
| DAVIVIENDA | 32 | Sí | 64 | 32 | 0,10 | 0,00 | 0,001 | 16 | 0,0000 |
| ETB | 64 | Sí | 16 | 32 | 0,00 | 0,00 | 0,005 | 16 | 0,0010 |
| NUTRESA | 32 | Sí | 64 | 16 | 0,20 | 0,00 | 0,01 | 16 | 0,0010 |
| SURAMERICANA | 96 | No | 0 | 16 | 0,00 | 0,20 | 0,005 | 64 | 0,0000 |

Tabla 24. Hiperparámetros iniciales para LSTM final

Fuente: Cálculos propios.

El análisis de la Tabla 1 evidencia diferencias significativas en el comportamiento de los modelos evaluados. El baseline Naive0, que asume retornos iguales a cero, presenta consistentemente el menor MAE en todos los emisores; sin embargo, su HitRate de 0% confirma que no captura información predictiva, sino que se beneficia de la cercanía natural de los retornos financieros a la media.

En contraste, el modelo Persist logra las mayores tasas de acierto direccional —especialmente en emisores como ETB y Nutresa, pero incurre en los errores más altos (MAE y RMSE), mostrando sensibilidad excesiva a la volatilidad diaria. En este contexto, el LSTM emerge como el modelo con mejor equilibrio general: reduce sustancialmente los errores respecto a Persist, supera ampliamente a Naive0 en acierto direccional y mantiene un comportamiento estable entre emisores tanto en entrenamiento como en prueba, sin evidenciar sobreajuste severo.

| Emisor | Split | Modelo | MAE | RMSE | $R^2$ | HitRate |
|---|---|---|---|---|---|---|
| Canacol | TEST | LSTM | 0.263266 | 0.444757 | -0.046736 | 0.446667 |
| | | Naive0 | 0.226974 | 0.435633 | -0.004229 | 0.000000 |
| | | Persist | 0.322638 | 0.560986 | -0.665308 | 0.500000 |
| | TRAIN | LSTM | 0.138690 | 0.213900 | 0.106801 | 0.633987 |
| | | Naive0 | 0.151963 | 0.226528 | -0.001774 | 0.000000 |
| | | Persist | 0.212774 | 0.307031 | -0.840301 | 0.502179 |
| Davivienda | TEST | LSTM | 0.117801 | 0.157291 | -0.243315 | 0.463576 |
| | | Naive0 | 0.102407 | 0.142185 | -0.015971 | 0.000000 |
| | | Persist | 0.147137 | 0.207829 | -1.170639 | 0.516556 |
| | TRAIN | LSTM | 0.194317 | 1.854347 | 0.033717 | 0.652552 |

| Emisor | Split | Modelo | MAE | RMSE | $R^2$ | HitRate |
|---|---|---|---|---|---|---|
| | | Naive0 | 0.209880 | 1.886421 | 0.000000 | 0.000000 |
| | | Persist | 0.347925 | 3.266656 | -1.998676 | 0.526602 |
| ETB | TEST | LSTM | 0.051552 | 0.106618 | -0.327579 | 0.804878 |
| | | Naive0 | 0.025063 | 0.092764 | -0.004978 | 0.000000 |
| | | Persist | 0.042437 | 0.131747 | -1.027129 | 0.926829 |
| | TRAIN | LSTM | 0.136528 | 0.267847 | 0.136299 | 0.726592 |
| | | Naive0 | 0.136754 | 0.288989 | -0.005430 | 0.000000 |
| | | Persist | 0.231454 | 0.443285 | -1.365678 | 0.614232 |
| Nutresa | TEST | LSTM | 0.162036 | 0.204350 | -0.283312 | 0.786260 |
| | | Naive0 | 0.095697 | 0.185425 | -0.056626 | 0.000000 |
| | | Persist | 0.149428 | 0.244353 | -0.834919 | 0.740458 |
| | TRAIN | LSTM | 0.162956 | 0.323367 | 0.073075 | 0.539506 |
| | | Naive0 | 0.166166 | 0.336575 | -0.004192 | 0.000000 |
| | | Persist | 0.253933 | 0.471907 | -0.974086 | 0.525926 |
| Suramericana | TEST | LSTM | 0.136497 | 0.220025 | -0.167097 | 0.500000 |
| | | Naive0 | 0.112550 | 0.203890 | -0.002194 | 0.000000 |
| | | Persist | 0.168785 | 0.277570 | -0.857410 | 0.535211 |
| | TRAIN | LSTM | 0.161114 | 0.257730 | 0.177084 | 0.607102 |
| | | Naive0 | 0.177730 | 0.284295 | -0.001299 | 0.000000 |
| | | Persist | 0.265284 | 0.408692 | -1.069271 | 0.510882 |

Tabla 25. Métricas promedio por emisor y modelo LSTM (Retornos, TRAIN y TEST)
Fuente: Cálculos propios.

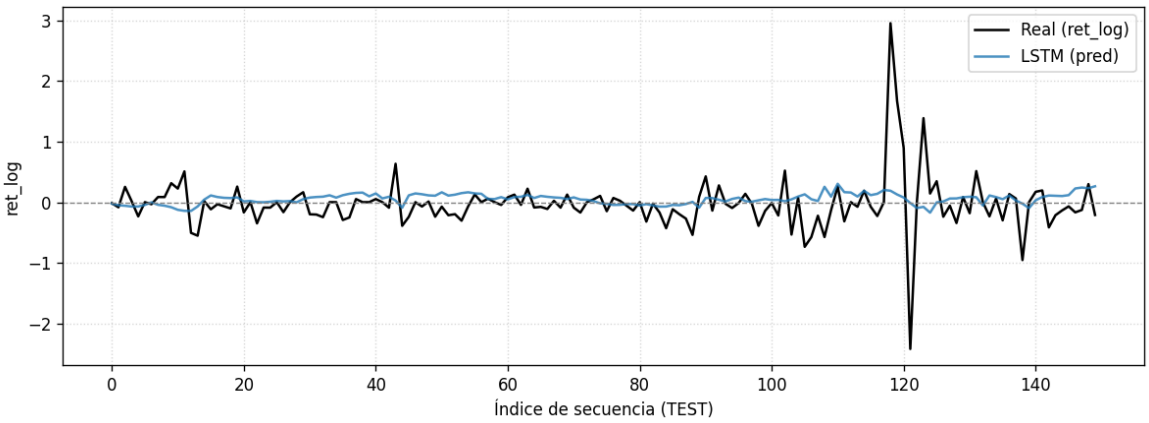


**Figura 24.** Comparación modelos de retornos CANACOL.

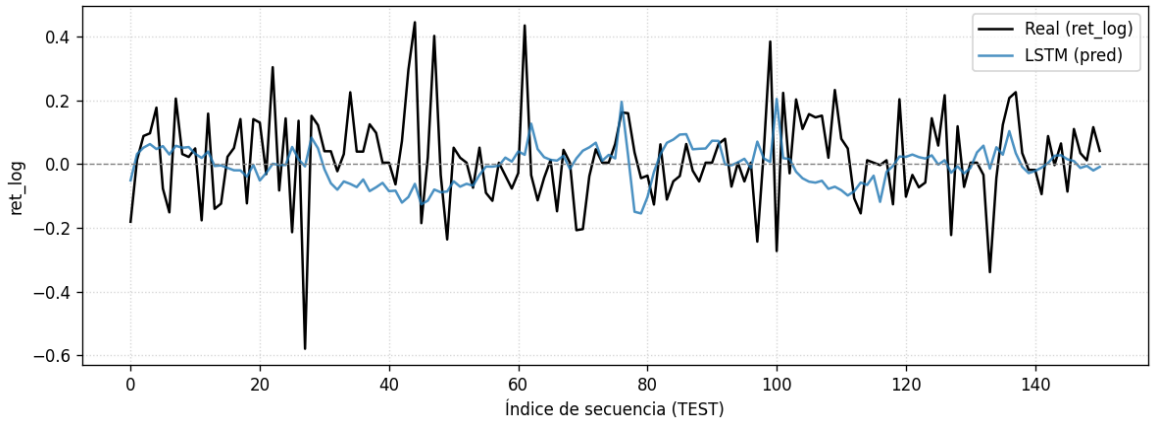


**Figura 25.** Comparación modelos de retornos DAVIVIENDA.

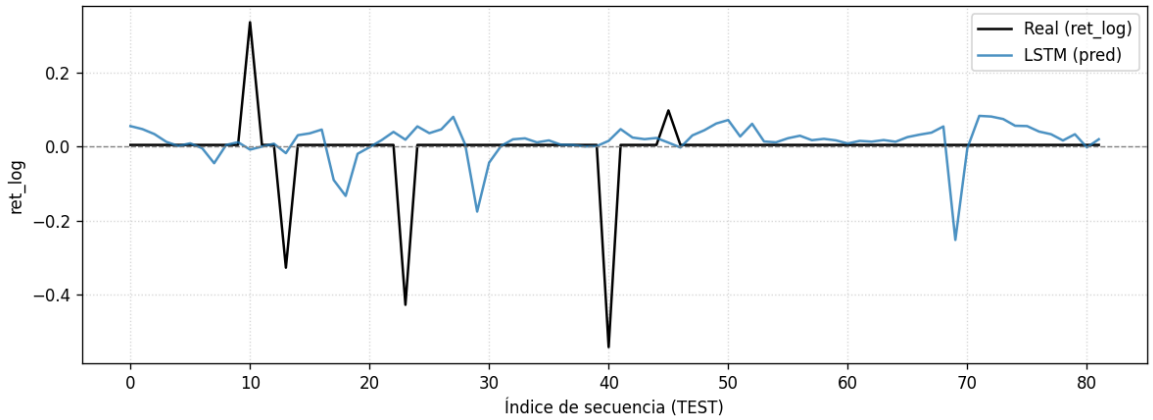


**Figura 26.** Comparación modelos de retornos ETB.

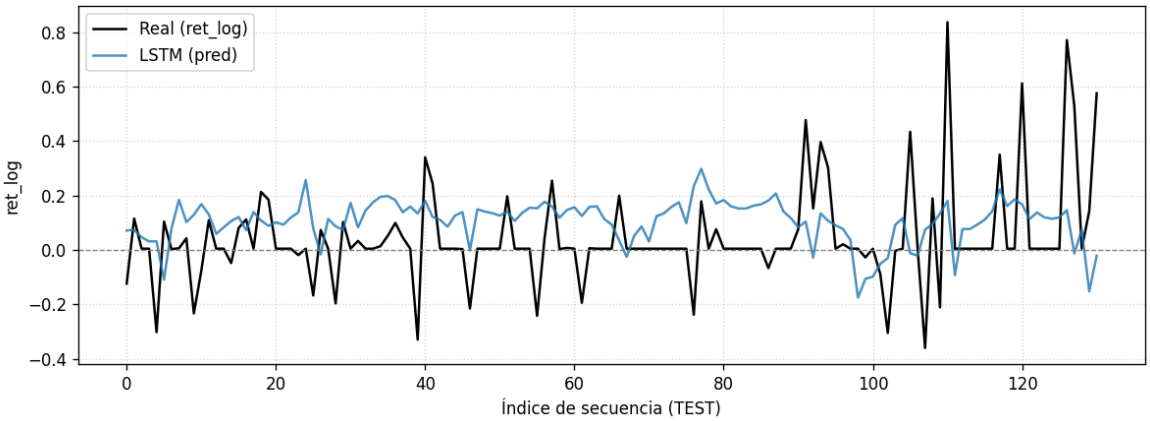


**Figura 27.** Comparación modelos de retornos NUTRESA.

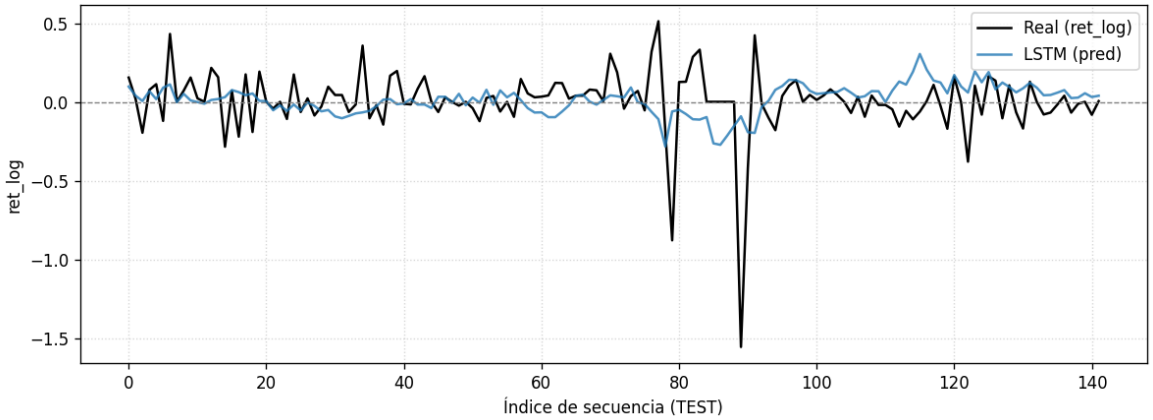


**Figura 28.** Comparación modelos de retornos SURAMERICANA.

**GRU**

Adicionalmente, se implementó un modelo de redes recurrentes del tipo GRU por emisor, siguiendo la misma lógica de construcción secuencial empleada en el LSTM. El modelo utiliza ventanas de 20 días y un conjunto de 20 características híbridas (retornos e indicadores técnicos), preprocesadas y escaladas de manera causal.

Sobre este espacio se evaluaron 12 configuraciones por emisor mediante validación temporal, seleccionando aquella con menor error absoluto medio (MAE) en el conjunto de validación. Posteriormente, el modelo GRU óptimo se reentrenó con los datos de entrenamiento y validación combinados, y su desempeño se comparó frente a los baselines Naive0 y Persist utilizando métricas de regresión (MAE, RMSE, $R^2$) y de clasificación direccional (HitRate).

| | Hiperparámetros | Valor |
|---|---|---|
| 1 | n_units | 32, 64, 96 |
| 2 | num_layers | 1, 2 |
| 3 | dropout_rate | 0.0, 0.2, 0.4 |
| 4 | l2_reg | 0.0, 1e-4, 1e-3 |
| 5 | learning rate (lr) | 0.0005, 0.001 |
| 6 | batch_size | 16, 32 |
| 7 | max_epochs | 60, 80 |
| 8 | patience | 8 |
| 9 | Ventana | 20 días |

Tabla 26. Hiperparámetros iniciales para GRU inicial
Fuente: Cálculos propios.

| Emisor | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| CANACOL | 32 | 1 | 0.0 | 0.0001 | 0.0005 | 32 | 80 | 8 | 20 |
| DAVIVIENDA | 32 | 1 | 0.0 | 0.0000 | 0.0005 | 32 | 60 | 8 | 20 |
| ETB | 32 | 1 | 0.0 | 0.0000 | 0.0010 | 16 | 60 | 8 | 22 |
| NUTRESA | 32 | 1 | 0.0 | 0.0000 | 0.0010 | 32 | 60 | 8 | 16 |
| SURAMERICANA | 32 | 1 | 0.0 | 0.0001 | 0.0005 | 16 | 60 | 8 | 13 |

Tabla 27. Hiperparámetros iniciales para GRU final

Fuente: Cálculos propios.

| Emisor | Split | Modelo | MAE | RMSE | $R^2$ | HitRate |
|---|---|---|---|---|---|---|
| Canacol | TEST | GRU | 0.236430 | 0.440757 | -0.027991 | 0.460000 |
| | | Naive0 | 0.226974 | 0.435633 | -0.004229 | 0.000000 |
| | | Persist | 0.322638 | 0.560986 | -0.665308 | 0.500000 |
| | TRAIN | GRU | 0.146859 | 0.220439 | 0.051361 | 0.584967 |
| | | Naive0 | 0.151963 | 0.226528 | -0.001774 | 0.000000 |
| | | Persist | 0.212774 | 0.307031 | -0.840301 | 0.502179 |
| Davivienda | TEST | GRU | 0.107166 | 0.145592 | -0.065249 | 0.529801 |
| | | Naive0 | 0.102407 | 0.142185 | -0.015971 | 0.000000 |
| | | Persist | 0.147137 | 0.207829 | -1.170639 | 0.516556 |
| | TRAIN | GRU | 0.202222 | 1.831153 | 0.057738 | 0.596091 |
| | | Naive0 | 0.209880 | 1.886421 | 0.000000 | 0.000000 |
| | | Persist | 0.347925 | 3.266656 | -1.998676 | 0.526602 |
| ETB | TEST | GRU | 0.191222 | 0.222414 | -4.777338 | 0.036585 |
| | | Naive0 | 0.025063 | 0.092764 | -0.004978 | 0.000000 |
| | | Persist | 0.042437 | 0.131747 | -1.027129 | 0.926829 |
| | TRAIN | GRU | 0.197958 | 0.291232 | -0.021101 | 0.314607 |
| | | Naive0 | 0.136754 | 0.288989 | -0.005430 | 0.000000 |
| | | Persist | 0.231454 | 0.443285 | -1.365678 | 0.614232 |
| Nutresa | TEST | GRU | 0.265645 | 0.328192 | -2.310078 | 0.152672 |
| | | Naive0 | 0.095697 | 0.185425 | -0.056626 | 0.000000 |
| | | Persist | 0.149428 | 0.244353 | -0.834919 | 0.740458 |
| | TRAIN | GRU | 0.170543 | 0.332790 | 0.018266 | 0.569136 |
| | | Naive0 | 0.166166 | 0.336575 | -0.004192 | 0.000000 |
| | | Persist | 0.253933 | 0.471907 | -0.974086 | 0.525926 |
| Suramericana | TEST | GRU | 0.148773 | 0.256707 | -0.588682 | 0.514085 |
| | | Naive0 | 0.112550 | 0.203890 | -0.002194 | 0.000000 |
| | | Persist | 0.168785 | 0.277570 | -0.857410 | 0.535211 |
| | TRAIN | GRU | 0.162904 | 0.255998 | 0.188109 | 0.616266 |
| | | Naive0 | 0.177730 | 0.284295 | -0.001299 | 0.000000 |
| | | Persist | 0.265284 | 0.408692 | -1.069271 | 0.510882 |

Tabla 25. Métricas promedio por emisor y modelo GRU (Retornos, TRAIN y TEST)

En este contexto, el modelo GRU no logra imponerse de manera sistemática sobre los baselines. En emisores como Canacol y Davivienda, el GRU se sitúa en una posición intermedia: su MAE en prueba es ligeramente superior al de Naive0, pero mejora en términos de HitRate, alcanzando tasas de acierto direccional cercanas al 46–53% y manteniendo errores moderados (por ejemplo, MAE = 0,2364 en Canacol y 0,1072 en Davivienda Pf). No obstante, en emisores como ETB y Nutresa el desempeño del GRU es claramente inferior: presenta MAE y RMSE considerablemente mayores que los baselines y valores de $R^2$ fuertemente negativos (–4,78 y –2,31, respectivamente), acompañados de un HitRate muy bajo (3,7 % en ETB y 15,3 % en Nutresa), lo que indica dificultades del modelo para capturar patrones informativos en estas series.

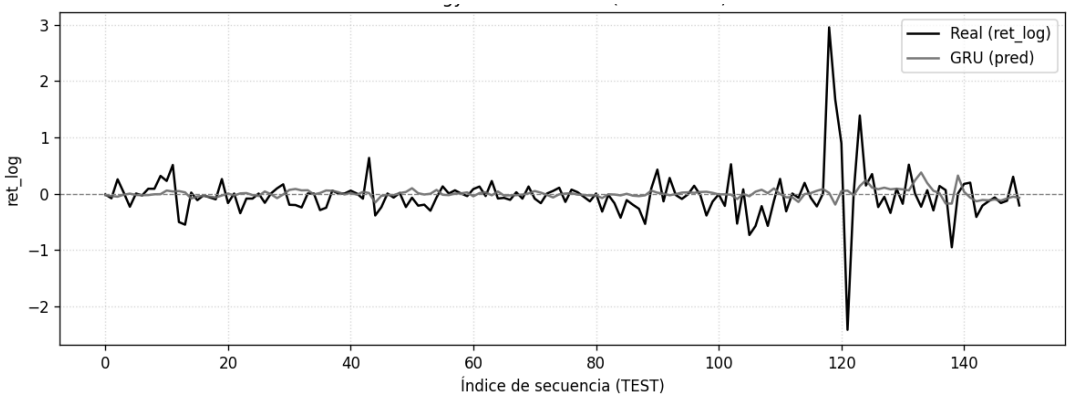


**Figura 29.** Comparación modelos de retornos CANACOL.

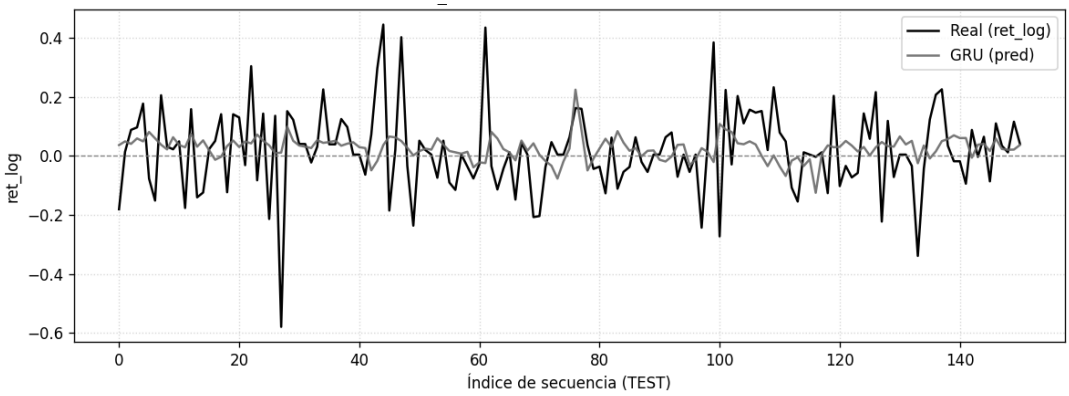


**Figura 30.** Comparación modelos de retornos SURAMERICANA.

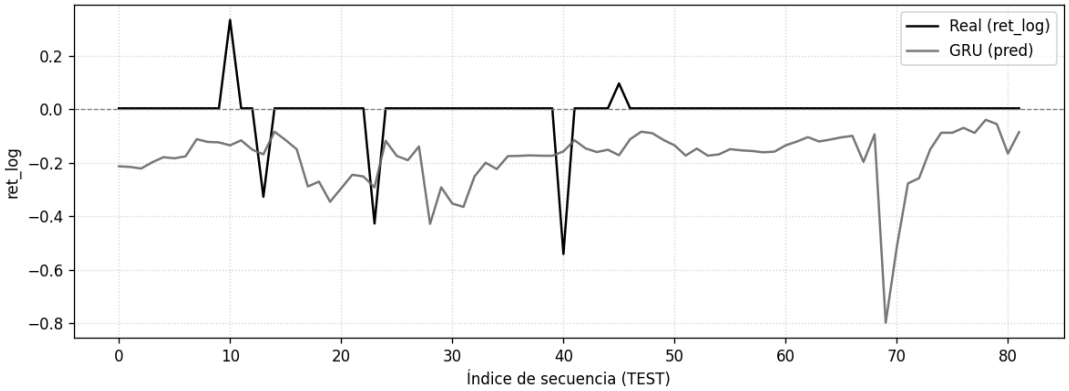


**Figura 31.** Comparación modelos de retornos ETB.

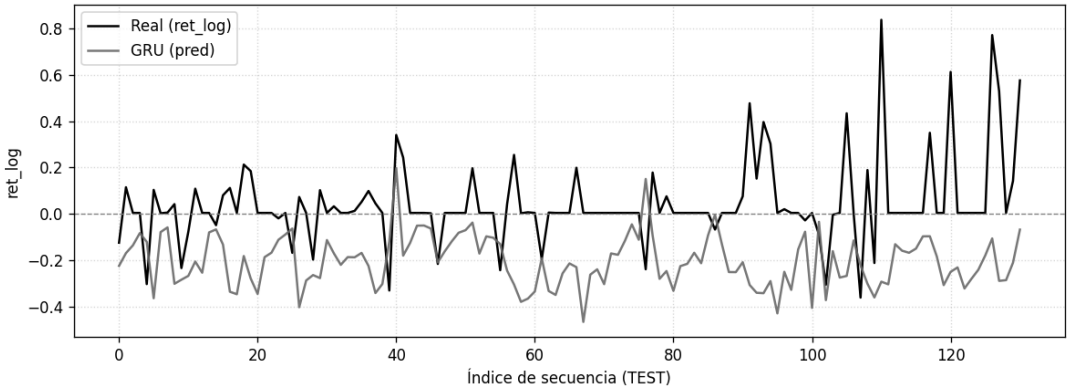


**Figura 32.** Comparación modelos de retornos NUTRESA.

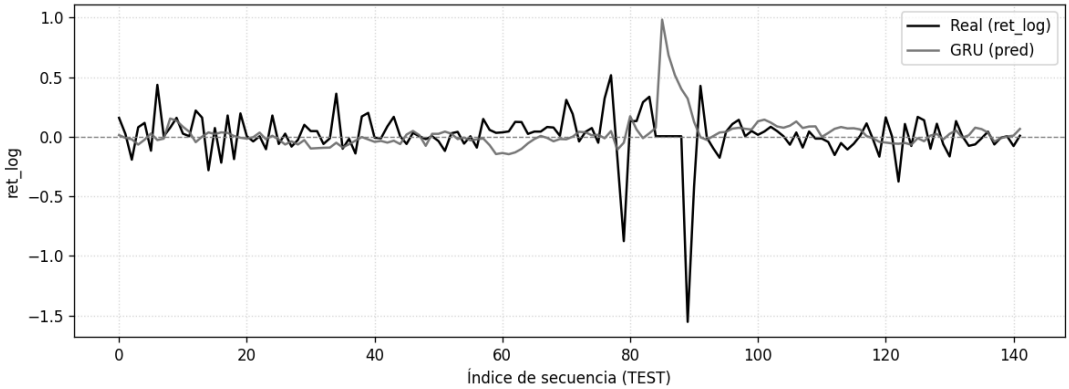


**Figura 33.** Comparación modelos de retornos SURAMERICANA.

**CNN-LSTM**

Para complementar los modelos secuenciales basados en LSTM y GRU, se implementó una arquitectura híbrida CNN-LSTM entrenada de forma desagregada por emisor. El modelo combina una capa convolucional temporal (Conv1D) aplicada sobre ventanas de 20 días de retornos e indicadores técnicos, seguida de un bloque LSTM y una capa densa lineal para estimar el retorno logarítmico del día siguiente. Sobre esta arquitectura se realizó una búsqueda aleatoria de hiperparámetros (Random Search) por emisor, considerando combinaciones de número de filtros convolucionales (16, 32, 48), tamaño de kernel (2, 3, 5), tamaño de ventana de pooling (1 y 2), número de unidades LSTM (32 y 64), tasa de dropout (0 y 0,3), penalización L2 0 y 0,0001 tasa de aprendizaje 0,0005 y 0,001 tamaño de lote (16 y 32) y máximo de épocas (60 y 80), restringiendo el número de ensayos a 12 configuraciones por emisor.

| | Hiperparámetros | Valor |
|---|---|---|
| 1 | n_filters | 16, 32, 48 |
| 2 | kernel_size | 2, 3, 5 |
| 3 | pool_size | 1, 2 |
| 4 | n_units (LSTM) | 32, 64 |
| 5 | dropout_rate | 0.0, 0.3 |
| 6 | l2_reg (Regularización L2) | 0.0, 0,0001 |
| 7 | learning_rate | 0,0005, 0,001 |
| 8 | batch_size | 16, 32 |
| 9 | max_epochs | 60, 80 |
| 10 | patience (Early Stopping) | 8 |

Tabla 27. Hiperparámetros iniciales para CNN-LSTM inicial
Fuente: Cálculos propios.

La selección del mejor modelo se efectuó a partir del MAE en el conjunto de validación, utilizando parada temprana (Early Stopping) con paciencia de 8 épocas y restaurando los mejores pesos. Posteriormente, el modelo óptimo de cada emisor se reentrenó sobre la unión de los conjuntos de entrenamiento y validación y se evaluó en el conjunto de prueba, comparando su desempeño frente a dos líneas base: Naive0 (retorno nulo) y Persist (retorno igual al del día anterior), mediante métricas de MAE, RMSE, $R^2$, sMAPE y tasa de acierto direccional (HitRate).

| Emisor | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|
| CANACOL | 16 | 2 | 1 | 32 | 0.0 | 0.0 | 0.001 | 32 | 80 | 11 |
| DAVIVIENDA | 16 | 2 | 1 | 32 | 0.0 | 0.0001 | 0.0005 | 16 | 80 | 12 |
| ETB | 16 | 2 | 1 | 32 | 0.0 | 0.0 | 0.001 | 16 | 60 | 34 |
| NUTRESA | 16 | 2 | 1 | 32 | 0.0 | 0.0001 | 0.0005 | 16 | 60 | 25 |
| SURAMERICANA | 16 | 2 | 1 | 32 | 0.0 | 0.0 | 0.0005 | 32 | 80 | 22 |

Tabla 28. Hiperparámetros iniciales para CNN-LSTM final
Fuente: Cálculos propios.

En el caso del modelo híbrido CNN-LSTM, la búsqueda aleatoria de hiperparámetros converge sistemáticamente hacia arquitecturas relativamente parsimoniosas para todos los emisores analizados. La mejor configuración para Canacol Energy, Davivienda, ETB, Nutresa y Suramericana coincide en emplear 16 filtros convolucionales con kernel temporal igual a 2, sin max-pooling efectivo adicional (pool_size = 1) y una capa LSTM de 32 unidades.

Asimismo, en todos los casos el modelo opera sin dropout, combinando tasas de aprendizaje moderadas entre 0,0004 y tamaños de lote de 16 o 32 observaciones. La regularización L2 solo resultó relevante para Davivienda y Nutresa (l2_reg = 0,0001), mientras que para los demás emisores el mejor desempeño se obtuvo sin penalización adicional. El número efectivo de épocas de entrenamiento, determinado por el criterio de early stopping[8], osciló entre 11 y 34, lo que indica que el modelo alcanza rápidamente su punto de mínima pérdida en validación sin requerir entrenamientos excesivamente prolongados, favoreciendo así un compromiso razonable entre capacidad de ajuste y control del sobreajuste.

| Emisor | Split | Modelo | MAE | RMSE | $R^2$ | HitRate |
|---|---|---|---|---|---|---|
| Canacol | TEST | CNN-LSTM | 0.233543 | 0.451601 | -0.079197 | 0.506667 |
| | | Naive0 | 0.226974 | 0.435633 | -0.004229 | 0.000000 |
| | | Persist | 0.322638 | 0.560986 | -0.665308 | 0.500000 |
| | TRAIN | CNN-LSTM | 0.145380 | 0.215974 | 0.089401 | 0.618736 |
| | | Naive0 | 0.151963 | 0.226528 | -0.001774 | 0.000000 |
| | | Persist | 0.212774 | 0.307031 | -0.840301 | 0.502179 |
| Davivienda | TEST | CNN-LSTM | 0.109707 | 0.148901 | -0.114212 | 0.562914 |
| | | Naive0 | 0.102407 | 0.142185 | -0.015971 | 0.000000 |
| | | Persist | 0.147137 | 0.207829 | -1.170639 | 0.516556 |
| | TRAIN | CNN-LSTM | 0.203472 | 1.862854 | 0.024831 | 0.572204 |
| | | Naive0 | 0.209880 | 1.886421 | 0.000000 | 0.000000 |
| | | Persist | 0.347925 | 3.266656 | -1.998676 | 0.526602 |

[8] Técnica de regularización que detiene el entrenamiento de un modelo de aprendizaje automático cuando su rendimiento en un conjunto de datos de validación comienza a empeorar.

| Emisor | Split | Modelo | MAE | RMSE | $R^2$ | HitRate |
|---|---|---|---|---|---|---|
| ETB | TEST | CNN-LSTM | 0.049123 | 0.100835 | -0.187474 | 0.951220 |
| | | Naive0 | 0.025063 | 0.092764 | -0.004978 | 0.000000 |
| | | Persist | 0.042437 | 0.131747 | -1.027129 | 0.926829 |
| | TRAIN | CNN-LSTM | 0.138165 | 0.278600 | 0.065563 | 0.765918 |
| | | Naive0 | 0.136754 | 0.288989 | -0.005430 | 0.000000 |
| | | Persist | 0.231454 | 0.443285 | -1.365678 | 0.614232 |
| Nutresa | TEST | CNN-LSTM | 0.108733 | 0.191837 | -0.130965 | 0.465649 |
| | | Naive0 | 0.095697 | 0.185425 | -0.056626 | 0.000000 |
| | | Persist | 0.149428 | 0.244353 | -0.834919 | 0.740458 |
| | TRAIN | CNN-LSTM | 0.169957 | 0.333994 | 0.011152 | 0.519753 |
| | | Naive0 | 0.166166 | 0.336575 | -0.004192 | 0.000000 |
| | | Persist | 0.253933 | 0.471907 | -0.974086 | 0.525926 |
| Suramericana | TEST | CNN-LSTM | 0.130279 | 0.210940 | -0.072700 | 0.514085 |
| | | Naive0 | 0.112550 | 0.203890 | -0.002194 | 0.000000 |
| | | Persist | 0.168785 | 0.277570 | -0.857410 | 0.535211 |
| | TRAIN | CNN-LSTM | 0.168796 | 0.269283 | 0.101652 | 0.594502 |
| | | Naive0 | 0.177730 | 0.284295 | -0.001299 | 0.000000 |
| | | Persist | 0.265284 | 0.408692 | -1.069271 | 0.510882 |

Tabla 26. Métricas promedio por emisor y modelo CNN-LSTM (Retornos, TRAIN y TEST)
Fuente: Cálculos propios.

En el caso de Canacol, el modelo CNN-LSTM presenta un MAE de 0.2335 en TEST, valor muy similar al Naive0 (0.2270), lo que indica que el error absoluto medio no mejora sustancialmente al incluir convoluciones y secuencias recurrentes. Sin embargo, el CNN-LSTM logra un HitRate del 50.66%, mientras que Naive0 no acierta la dirección en ningún caso, lo cual demuestra que el modelo aprende patrones direccionales del retorno que el baseline no capta. Asimismo, el $R^2$ negativo confirma la alta dificultad del problema y la naturaleza ruidosa de los retornos financieros.

Para Davivienda, el CNN-LSTM vuelve a mostrar un error medio similar al Naive0, con un MAE de 0.1097 frente a 0.1024 del baseline. No obstante, consigue un HitRate de 56.29%, superior al Persist (51.66 %) y drásticamente más alto que Naive0 (0 %). En ETB, el beneficio del enfoque CNN-LSTM es más evidente. El modelo logra un MAE en TEST de 0.0491, muy cercano al Naive0 (0.0251), pero con un HitRate sobresaliente del 95.12 %. Este es el mejor resultado entre todos los emisores y apunta que ETB posee patrones temporales más lineales o más estables que pueden ser detectados por la arquitectura híbrida CNN-LSTM. El modelo supera además ampliamente al Persist (0.9268), que ya presenta un desempeño muy alto en esta acción.

En el caso de Nutresa, aunque el Naive0 obtiene el menor error (MAE 0.0957), el modelo CNN-LSTM logra un MAE relativamente bajo (0.1087) y un HitRate de 46.56 %, superando nuevamente al Naive0 (0 %) y mostrando un desempeño direccional moderado, pero efectivo. La magnitud del RMSE y el $R^2$ negativo indican que los retornos de Nutresa presentan mayor ruido e imprevisibilidad relativa. Finalmente, para Suramericana, el CNN-LSTM presenta un MAE de 0.1302, cercano a Naive0 (0.1125) y mejor que Persist (0.1687). Su HitRate de 51.40 % confirma que también aporta valor predictivo en términos direccionales.

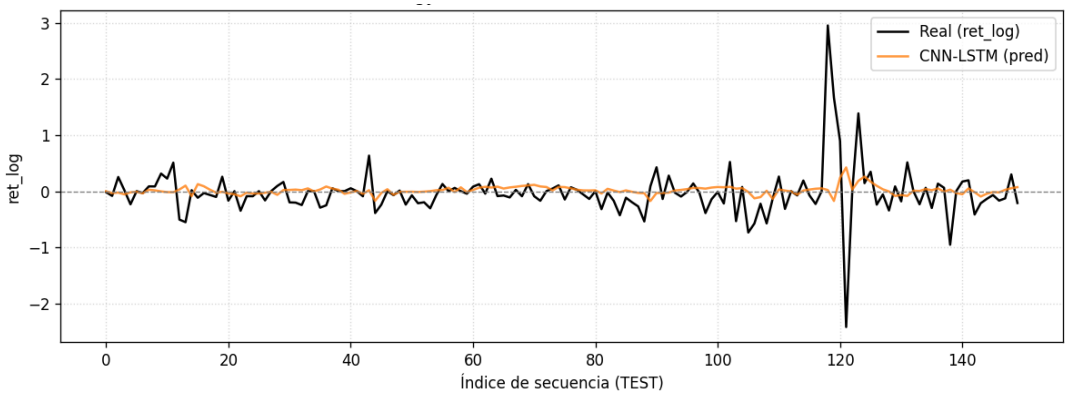


**Figura 34.** Comparación modelos de retornos CANACOL.

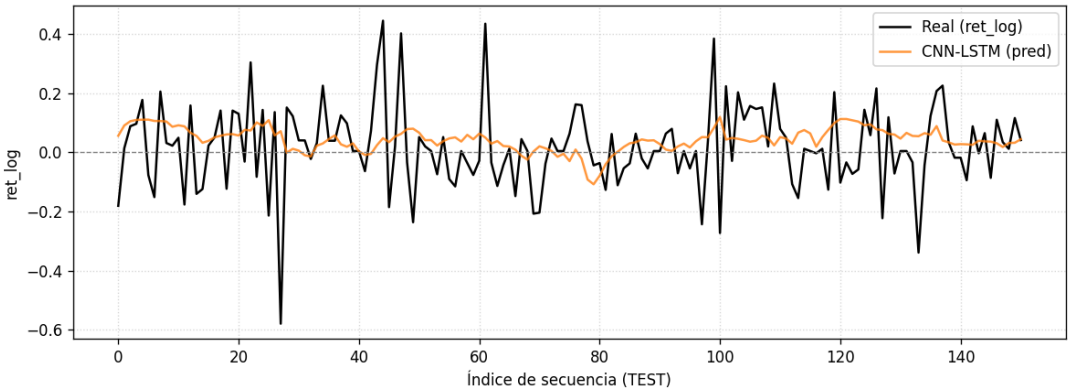


**Figura 35.** Comparación modelos de retornos DAVIVIENDA.

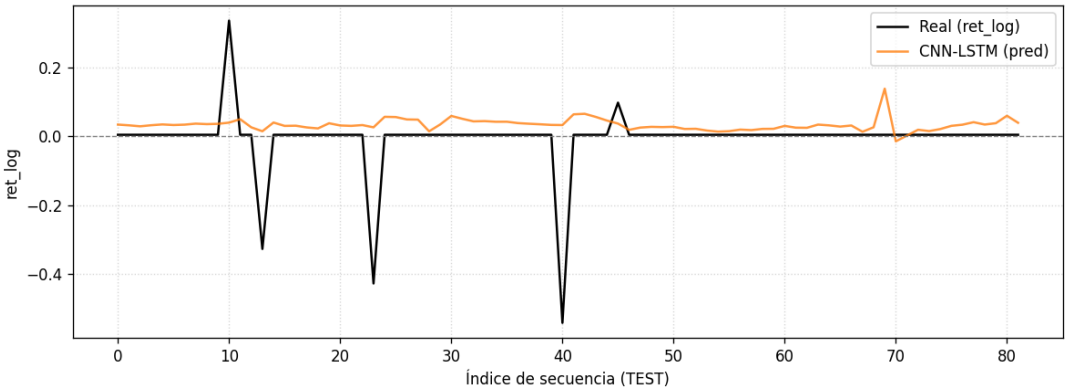


**Figura 36.** Comparación modelos de retornos ETB.

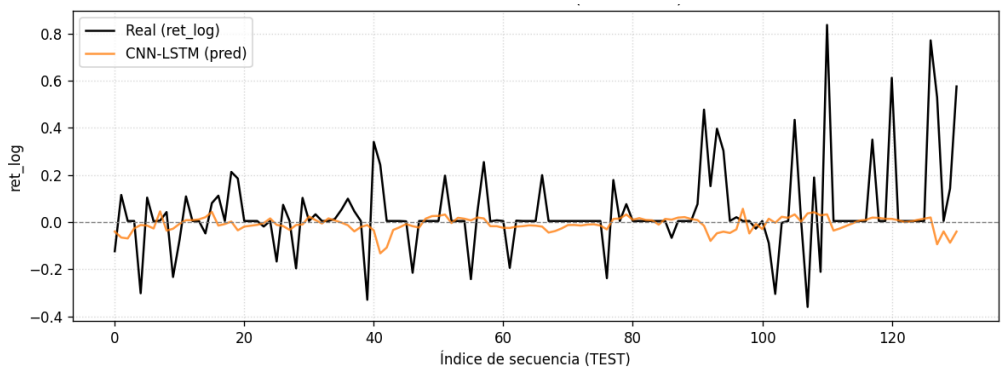


**Figura 37.** Comparación modelos de retornos NUTRESA.

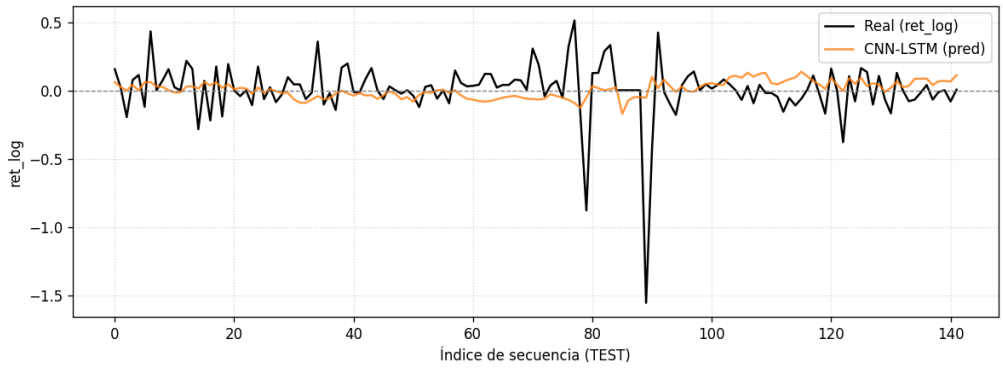


**Figura 38.** Comparación modelos de retornos SURAMERICANA.

**Test de Diebold–Mariano y correlaciones**

| Modelo 1 | Modelo 2 | Estadístico DM | Valor *p* |
|---|---|---|---|
| Random Forest | Naive0 | 3.617177 | 0.000298 |
| Random Forest | Persist | 0.367525 | 0.713228 |

Tabla 27. Métricas Diebold–Mariano Random Forest (Test)
Fuente: Cálculos propios.

El test de Diebold–Mariano aplicado sobre el conjunto de prueba (TEST) indica que el modelo Random Forest presenta un desempeño predictivo estadísticamente superior al modelo Naive0, dado que el estadístico DM es significativo al 1%. En contraste, no se evidencia una diferencia estadísticamente significativa entre Random Forest y el modelo Persist, lo cual sugiere desempeños predictivos comparables entre ambos enfoques.

| Emisor | Modelo | Variable | Correlación |
|---|---|---|---|
| Canacol Energy | Naive0 | Retornos | NaN |
| Canacol Energy | Naive0 | Precios | 0.985689 |
| Canacol Energy | Persist | Retornos | 0.163802 |
| Canacol Energy | Persist | Precios | 0.975186 |
| Canacol Energy | Random Forest | Retornos | 0.015163 |
| Canacol Energy | Random Forest | Precios | 0.985093 |
| Davivienda PF | Naive0 | Retornos | NaN |
| Davivienda PF | Naive0 | Precios | 0.979259 |
| Davivienda PF | Persist | Retornos | -0.009782 |
| Davivienda PF | Persist | Precios | 0.960175 |
| Davivienda PF | Random Forest | Retornos | -0.019038 |
| Davivienda PF | Random Forest | Precios | 0.979048 |
| ETB | Naive0 | Retornos | NaN |
| ETB | Naive0 | Precios | 0.968834 |
| ETB | Persist | Retornos | -0.002269 |
| ETB | Persist | Precios | 0.942878 |
| ETB | Random Forest | Retornos | -0.015517 |
| ETB | Random Forest | Precios | 0.963010 |
| Nutresa | Naive0 | Retornos | NaN |
| Nutresa | Naive0 | Precios | 0.987073 |
| Nutresa | Persist | Retornos | 0.032666 |
| Nutresa | Persist | Precios | 0.975022 |
| Nutresa | Random Forest | Retornos | -0.260564 |
| Nutresa | Random Forest | Precios | 0.913951 |
| Suramericana | Naive0 | Retornos | NaN |
| Suramericana | Naive0 | Precios | 0.951117 |
| Suramericana | Persist | Retornos | 0.130236 |

| Emisor | Modelo | Variable | Correlación |
|---|---|---|---|
| Suramericana | Persist | Precios | 0.929207 |
| Suramericana | Random Forest | Retornos | -0.018517 |
| Suramericana | Random Forest | Precios | 0.943290 |

Tabla 28. Métricas de correlación Random Forest (Test)
Fuente: Cálculos propios.

Antes que nada, es necesario aclarar que el valor NaN en el modelo Naive0 para retornos se debe a que este modelo no genera variación en la predicción de retornos, lo que impide el cálculo de la correlación. Así pues, y en relación a los resultados de la tabla 8, se evidencia ua alta correlación entre los valores observados y predichos en el caso de los precios, lo cual es consistente con la fuerte dependencia temporal de estas series y la naturaleza acumulativa del nivel del precio. En contraste, las correlaciones para los retornos son bajas, cercanas a cero o incluso negativas, evidencia de la dificultad inherente de capturar relaciones lineales estables en series financieras estacionarias y altamente ruidosas.

| Modelo 1 | Modelo 2 | Estadístico DM | Valor *p* |
|---|---|---|---|
| XGBoost | Naive0 | 1.340381 | 0.180122 |
| XGBoost | Persist | -3.089610 | 0.002004 |

Tabla 29. Métricas Diebold–Mariano XGBoost (Test)
Fuente: Cálculos propios.

No existe evidencia estadísticamente significativa para afirmar que el modelo XGBoost supere al modelo Naive0 en términos de precisión predictiva, dado que el valor p es superior a los niveles convencionales de significancia. En contraste, la diferencia entre XGBoost y el modelo Persist indica un desempeño predictivo diferencial entre ambos modelos en la predicción de retornos.

| Emisor | Modelo | Variable | Correlación |
|---|---|---|---|
| Canacol Energy | Naive0 | Retornos | NaN |
| Canacol Energy | Persist | Retornos | 0.163802 |
| Canacol Energy | XGBoost | Retornos | 0.006743 |
| Canacol Energy | Naive0 | Precios | 0.985689 |
| Canacol Energy | Persist | Precios | 0.975186 |
| Canacol Energy | XGBoost | Precios | 0.980638 |
| Davivienda PF | Naive0 | Retornos | NaN |
| Davivienda PF | Persist | Retornos | -0.009782 |
| Davivienda PF | XGBoost | Retornos | 0.192823 |
| Davivienda PF | Naive0 | Precios | 0.979259 |
| Davivienda PF | Persist | Precios | 0.960175 |
| Davivienda PF | XGBoost | Precios | 0.979726 |

| Emisor | Modelo | Variable | Correlación |
|---|---|---|---|
| ETB | Naive0 | Retornos | NaN |
| ETB | Persist | Retornos | -0.002269 |
| ETB | XGBoost | Retornos | 0.067212 |
| ETB | Naive0 | Precios | 0.968834 |
| ETB | Persist | Precios | 0.942878 |
| ETB | XGBoost | Precios | 0.968862 |
| Nutresa | Naive0 | Retornos | NaN |
| Nutresa | Persist | Retornos | 0.032666 |
| Nutresa | XGBoost | Retornos | 0.086711 |
| Nutresa | Naive0 | Precios | 0.987073 |
| Nutresa | Persist | Precios | 0.975022 |
| Nutresa | XGBoost | Precios | 0.987158 |
| Suramericana | Naive0 | Retornos | NaN |
| Suramericana | Persist | Retornos | 0.130236 |
| Suramericana | XGBoost | Retornos | 0.058876 |
| Suramericana | Naive0 | Precios | 0.951117 |
| Suramericana | Persist | Precios | 0.929207 |
| Suramericana | XGBoost | Precios | 0.949176 |

Tabla 30. Métricas de correlación XGBoost (Test)
Fuente: Cálculos propios.

De acuerdo a la tabla 30, y teniendo el mismo patrón de las anteriores tablas, se observa que las correlaciones asociadas a los precios son consistentemente altas, mientras que aquellas correspondientes a los retornos son bajas, lo cual fundamenta la teoría de que existe más dificultad de capturar relaciones lineales estables en series de retornos.

| Modelo 1 | Modelo 2 | Estadístico DM | Valor *p* |
|---|---|---|---|
| LightGBM | Naive0 | 3.420108 | 0.000626 |
| LightGBM | Persist | -3.198151 | 0.001383 |

Tabla 31. Métricas Diebold–Mariano LightGBM (Test)
Fuente: Cálculos propios.

En la tabla 31 se evidencian diferencias significativas en el desempeño predictivo entre el modelo LightGBM y los modelos base. En particular, LightGBM presenta un desempeño distinto frente al modelo Naive0 y al modelo Persist, con significancia estadística al 1%. En relación a la tabla 32 se observa que las correlaciones son sistemáticamente altas para los precios y notablemente menores para los retornos, lo cual es consistente con la evidencia ya expuesta en las tablas previas.

| Emisor | Modelo | Variable | Correlación |
|---|---|---|---|
| Canacol Energy | Naive0 | Retornos | NaN |
| Canacol Energy | Naive0 | Precios | 0.985689 |
| Canacol Energy | Persist | Retornos | 0.163856 |
| Canacol Energy | Persist | Precios | 0.975193 |
| Canacol Energy | LightGBM | Retornos | -0.012205 |
| Canacol Energy | LightGBM | Precios | 0.984379 |
| Davivienda PF | Naive0 | Retornos | NaN |
| Davivienda PF | Naive0 | Precios | 0.979259 |
| Davivienda PF | Persist | Retornos | -0.010902 |
| Davivienda PF | Persist | Precios | 0.960098 |
| Davivienda PF | LightGBM | Retornos | 0.131302 |
| Davivienda PF | LightGBM | Precios | 0.979554 |
| ETB | Naive0 | Retornos | NaN |
| ETB | Naive0 | Precios | 0.968834 |
| ETB | Persist | Retornos | -0.002269 |
| ETB | Persist | Precios | 0.942878 |
| ETB | LightGBM | Retornos | -0.001403 |
| ETB | LightGBM | Precios | 0.958789 |
| Nutresa | Naive0 | Retornos | NaN |
| Nutresa | Naive0 | Precios | 0.987073 |
| Nutresa | Persist | Retornos | 0.032666 |
| Nutresa | Persist | Precios | 0.975022 |
| Nutresa | LightGBM | Retornos | 0.071529 |
| Nutresa | LightGBM | Precios | 0.983766 |
| Suramericana | Naive0 | Retornos | NaN |
| Suramericana | Naive0 | Precios | 0.951117 |
| Suramericana | Persist | Retornos | 0.129903 |
| Suramericana | Persist | Precios | 0.929208 |
| Suramericana | LightGBM | Retornos | 0.155935 |
| Suramericana | LightGBM | Precios | 0.947126 |

Tabla 32. Métricas correlación LightGBM (Test)

Fuente: Cálculos propios.

| Modelo 1 | Modelo 2 | Estadístico DM | Valor *p* |
|---|---|---|---|
| LSTM | Naive0 | 2.551164 | 0.010736 |
| LSTM | Persist | -3.130091 | 0.001748 |

Tabla 33. Métricas Diebold–Mariano LSTM (Test – Retornos)
Fuente: Cálculos propios.

El test en cuestión aplicado al conjunto de prueba indica que el modelo LSTM presenta diferencias frente a ambos modelos de referencia. En particular, la comparación con el modelo Naive0 es significativa al 5%, mientras que frente al modelo Persistente la diferencia es significativa al 1%. Estos resultados exponen que el desempeño predictivo del modelo LSTM difiere de manera sistemática respecto a los bases considerados en la predicción de retornos.

| Emisor | Modelo | Correlación |
|---|---|---|
| Canacol Energy | Naive0 | NaN |
| Canacol Energy | Persist | 0.166939 |
| Canacol Energy | LSTM | 0.074179 |
| Davivienda PF | Naive0 | NaN |
| Davivienda PF | Persist | -0.078555 |
| Davivienda PF | LSTM | -0.054243 |
| ETB | Naive0 | NaN |
| ETB | Persist | -0.013563 |
| ETB | LSTM | 0.000352 |
| Nutresa | Naive0 | NaN |
| Nutresa | Persist | 0.051654 |
| Nutresa | LSTM | 0.025213 |
| Suramericana | Naive0 | NaN |
| Suramericana | Persist | 0.076467 |
| Suramericana | LSTM | 0.036506 |

Tabla 34. Métricas correlación LightGBM (Test – Retorno)
Fuente: Cálculos propios.

| Modelo 1 | Modelo 2 | Estadístico DM | Valor *p* |
|---|---|---|---|
| GRU | Naive0 | 5.805472 | 0.000000 |
| GRU | Persist | -5.841748 | 0.000000 |

Tabla 35. Métricas correlación GRU (Test – Retorno)
Fuente: Cálculos propios.

En ambas comparaciones (GRU con respecto a los modelos base), los valores del estadístico DM y los valores *p* cercanos a cero indican que el desempeño predictivo del modelo GRU difiere de manera contundente frente a los benchmarks considerados en la predicción de retornos,

| Emisor | Modelo | Correlación |
|---|---|---|
| Canacol Energy | GRU | 0.023982 |
| Canacol Energy | Naive0 | NaN |
| Canacol Energy | Persist | 0.166939 |
| Davivienda PF | GRU | 0.023734 |
| Davivienda PF | Naive0 | NaN |
| Davivienda PF | Persist | -0.078555 |
| ETB | GRU | 0.056889 |
| ETB | Naive0 | NaN |
| ETB | Persist | -0.013563 |
| Nutresa | GRU | -0.058644 |
| Nutresa | Naive0 | NaN |
| Nutresa | Persist | 0.051654 |
| Suramericana | GRU | -0.135984 |
| Suramericana | Naive0 | NaN |
| Suramericana | Persist | 0.076467 |

Tabla 36. Métricas correlación GRU (Test – Retorno)
Fuente: Cálculos propios.

En general, las correlaciones observadas para los retornos son bajas y cercanas a cero, e incluso negativas para algunos emisores, lo cual es consistente con la literatura financiera y refleja la dificultad estructural de capturar relaciones lineales estables en la predicción de retornos fuera de la muestra.

| Modelo 1 | Modelo 2 | Estadístico DM | Valor *p* |
|---|---|---|---|
| CNN-LSTM | Naive0 | 1.331161 | 0.183599 |
| CNN-LSTM | Persist | -3.695594 | 0.000238 |

Tabla 37. Métricas correlación CNN-LSTM (Test – Retorno)
Fuente: Cálculos propios.

El test de Diebold–Mariano indica que no existen diferencias estadísticamente significativas entre el modelo CNN-LSTM y el benchmark Naive0 en la predicción de retornos. No obstante, sí se observa una diferencia estadísticamente significativa frente al modelo Persistente al 1%, lo que sugiere que el desempeño predictivo del CNN-LSTM difiere de manera sistemática respecto a este modelo de referencia.

| Emisor | Modelo | Correlación |
|---|---|---|
| Canacol Energy | CNN-LSTM | -0.119472 |
| Canacol Energy | Naive0 | NaN |
| Canacol Energy | Persist | 0.166939 |
| Davivienda PF | CNN-LSTM | -0.012533 |
| Davivienda PF | Naive0 | NaN |
| Davivienda PF | Persist | -0.078555 |
| ETB | CNN-LSTM | 0.091696 |
| ETB | Naive0 | NaN |
| ETB | Persist | -0.013563 |
| Nutresa | CNN-LSTM | -0.063645 |
| Nutresa | Naive0 | NaN |
| Nutresa | Persist | 0.051654 |
| Suramericana | CNN-LSTM | 0.018701 |
| Suramericana | Naive0 | NaN |
| Suramericana | Persist | 0.076467 |

Tabla 38. Métricas correlación CNN-LSTM (Test – Retorno)
Fuente: Cálculos propios.

En general, las correlaciones observadas para los modelos CNN-LSTM son bajas y cercanas a cero, e incluso negativas para algunos emisores, lo cual es consistente con la evidencia empírica sobre la dificultad estructural de predecir retornos financieros fuera de la muestra mediante relaciones lineales o no lineales estables.

**Explicación del coeficiente de determinación**

El coeficiente de determinación $R^2$ es una medida utilizada para evaluar el desempeño de modelos estadísticos, ya que cuantifica la proporción de la variabilidad de la variable dependiente que es explicada por el modelo. No obstante, su interpretación depende fundamentalmente del del contexto en el cual se calcula.

En modelos de carácter explicativo, el $R^2$ se calcula sobre la misma muestra utilizada para estimar el modelo, generalmente mediante el método de Mínimos Cuadrados Ordinarios. En este contexto, el $R^2$ mide la capacidad del modelo para explicar la variabilidad observada de la variable dependiente y, por construcción, toma valores comprendidos entre cero y uno.

- Un valor más alto de $R^2$ indica un mayor grado de ajuste en muestra, aunque no necesariamente implica una mejor capacidad predictiva fuera de la muestra analizada.

Por otro lado, y en diferencia con el anterior, en modelos de carácter predictivo, cuyo objetivo principal es anticipar el comportamiento futuro de una variable, el $R^2$ se evalúa

típicamente fuera de muestra. En este caso, el desempeño del modelo se compara con el de un modelo de referencia, como una predicción basada en la media histórica o un paseo aleatorio.

- Bajo esta formulación, el $R^2$ representa la mejora relativa del modelo propuesto frente a dicho modelo de referencia y puede tomar valores negativos cuando el error de predicción del modelo supera al del modelo de referencia.

La literatura especializada ha documentado ampliamente que, en el ámbito financiero, un bajo o incluso negativo $R^2$ predictivo es un resultado empíricamente común y no invalida necesariamente la utilidad del modelo. Por ejemplo, [47] enfatizan que muchas regresiones predictivas superan el rendimiento promedio histórico, una vez que se imponen restricciones débiles sobre los signos de los coeficientes y los pronósticos de rendimiento. El poder explicativo fuera de la muestra es pequeño, pero, no obstante, es económicamente significativo para los inversores de media-varianza.

**Explicativo**
R2 dentro de la muestra

**Predictivo**
R2 fuera de la muestra

**Table 2**
**Excess return prediction with valuation constraints**

| | Sample Begin | Forecast Begin | In-Sample *t*-statistic | In-Sample *R*-squared | Out-of-Sample *R*-squared with Different Constraints: Unconstrained | Positive Slope, Pos. Forecast | Pos. Intercept, Bounded Slope | Fixed Coefs |
|---|---|---|---|---|---|---|---|---|
| A: Monthly Returns | | | | | | | | |
| Dividend/price | 1872m2 | 1927m1 | 1.25 | 1.12% | −0.66% | 0.08% | 0.19% | 0.42% |
| Earnings/price | 1872m2 | 1927m1 | 2.28 | 0.71 | 0.12 | 0.18 | 0.25 | 0.76 |
| Smooth earnings/price | 1881m2 | 1927m1 | 1.85 | 1.35 | 0.32 | 0.43 | 0.43 | 0.97 |
| Dividend/price + growth | 1891m2 | 1927m1 | 1.40 | 1.03 | −0.05 | 0.20 | 0.17 | 0.63 |
| Earnings/price + growth | 1892m2 | 1927m1 | 1.82 | 0.49 | −0.05 | 0.08 | 0.07 | 0.57 |
| Smooth earnings/price + growth | 1892m2 | 1927m1 | 2.00 | 1.10 | 0.11 | 0.25 | 0.21 | 0.72 |
| Book-to-market + growth | 1936m6 | 1956m6 | 1.61 | 0.33 | −0.35 | −0.34 | −0.34 | 0.33 |
| Dividend/price + growth − real rate | 1891m5 | 1927m1 | 1.47 | 0.86 | −0.02 | 0.21 | 0.18 | 0.41 |
| Earnings/price + growth − real rate | 1892m2 | 1927m1 | 1.53 | 0.36 | 0.00 | 0.12 | 0.09 | 0.39 |
| Smooth earnings/price + growth − real rate | 1892m2 | 1927m1 | 1.97 | 0.84 | 0.15 | 0.26 | 0.23 | 0.52 |
| Book-to-market + growth − real rate | 1936m6 | 1956m6 | 1.68 | 0.36 | −0.45 | −0.45 | −0.42 | 0.24 |
| B: Annual Returns | | | | | | | | |
| Dividend/price | 1872m2 | 1927m1 | 2.69 | 10.89 | 5.53 | 5.63 | 3.76 | 2.20 |
| Earnings/price | 1872m2 | 1927m1 | 2.84 | 6.78 | 4.93 | 4.94 | 4.34 | 5.87 |
| Smooth earnings/price | 1881m2 | 1927m1 | 3.01 | 13.57 | 7.89 | 7.85 | 6.44 | 7.99 |
| Dividend/price + growth | 1891m2 | 1927m1 | 1.77 | 9.30 | 2.49 | 2.99 | 2.67 | 4.35 |
| Earnings/price + growth | 1892m2 | 1927m1 | 1.42 | 4.44 | 1.69 | 2.11 | 1.80 | 3.89 |
| Smooth earnings/price + growth | 1892m2 | 1927m1 | 1.75 | 10.45 | 3.16 | 3.33 | 3.23 | 5.39 |
| Book-to-market + growth | 1936m6 | 1956m6 | 1.97 | 5.45 | −3.53 | −0.64 | −2.39 | 3.63 |
| Dividend/price + growth − real rate | 1891m5 | 1927m1 | 1.46 | 7.69 | 2.87 | 3.24 | 2.95 | 1.89 |
| Earnings/price + growth − real rate | 1892m2 | 1927m1 | 1.13 | 3.27 | 2.01 | 2.05 | 2.04 | 1.85 |
| Smooth earnings/price + growth − real rate | 1892m2 | 1927m1 | 1.53 | 7.90 | 3.35 | 3.35 | 3.38 | 3.22 |
| Book-to-market + growth − real rate | 1936m6 | 1956m6 | 2.03 | 5.77 | −1.73 | −1.12 | −1.82 | 2.33 |

The table presents forecast statistics for value predictors under various constraints. The predictor label "+ growth" indicates that we add an earnings growth forecast to the predictor. The predictor label "− real rate" indicates that we subtract a forecast of the real risk-free rate from the predictor. See the text for details. The "In-Sample" statistics are defined as in Table 1. The "Unconstrained" and "Positive Slope, Pos. Forecast" columns are described in Table 1. "Pos. Intercept, Bounded Slope" indicates that we constrain the intercept to be positive and the slope to be between zero and one. "Fixed Coefs" indicates that we fix the intercept at zero and the slope at one.

Activar Windows
Ve a Configuración para

**Ilustración 7.** Coeficientes de determinación dentro y fuera de la muestra.
Fuente: [47]

En tanto el $R^2$ predictivo puede tomar valores negativos porque se calcula comparando el error del modelo con el error de un modelo de referencia (baseline). Cuando el modelo presenta un mayor error de predicción que el modelo base al evaluarse sobre datos fuera de la muestra, la

mejora relativa es negativa (Error del modelo > Error del modelo base), lo que indica que el modelo no logra superar una predicción básica.

En síntesis, la evaluación dentro de la muestra (in-sample) mide el ajuste del modelo utilizando los mismos datos con los que fue estimado, lo que refleja su capacidad explicativa pero no garantiza un buen desempeño predictivo. En discrepancia, la evaluación fuera de la muestra (out-of-sample) se realiza sobre datos no utilizados en el entrenamiento y permite evaluar la capacidad de generalización del modelo, comparando su desempeño frente a un modelo base. Fuera de la muestra significa evaluar el modelo con datos que NO se usaron para entrenarlo, es decir, los datos de test.

| | |
|---|---|
| Dentro de la muestra | Datos de entrenamiento |
| Fuera de la muestra | Datos de test (parte predictiva) |

Fuente: Elaboración propia (2025)

Finalmente, el Hit Rate muestra un buen desempeño en la predicción direccional del precio, es decir, en la capacidad del modelo para anticipar si el activo sube o baja en el siguiente período. Esta métrica evalúa únicamente si el modelo acierta el signo del cambio, sin considerar la magnitud del error en la predicción numérica del precio. En particular, se considera un acierto cuando el precio efectivamente aumenta y el modelo predice una subida, o cuando disminuye y el modelo anticipa una caída. Por el contrario, no se penalizan errores asociados a la sobreestimación o subestimación del tamaño del movimiento, siempre que la dirección sea correcta.

### 4.11 Visualización de cada activo de acuerdo a su mejor modelo.

A continuación, se presentan las visualizaciones desarrolladas como parte del análisis descriptivo de los activos estudiados. Para este ejercicio se emplearon las bases históricas de las cinco acciones con mayor volatilidad dentro del conjunto seleccionado, las cuales fueron importadas y procesadas en Power BI con el fin de facilitar la exploración interactiva de la información.

El panel cuenta con un filtro por emisor que permite seleccionar individualmente cada activo y analizar su comportamiento específico. Asimismo, se incorporó un rango temporal dinámico que posibilita desplazarse a través de diferentes periodos y ajustar el horizonte de visualización según las necesidades del análisis. Entre los elementos incluidos se encuentra una tabla tipo matriz agrupada por fechas, que permite observar de manera ordenada la evolución temporal de las variables clave.

El tablero integra tres gráficos principales: el primero muestra el promedio del valor de cierre por emisor, permitiendo identificar tendencias generales entre los activos; el segundo presenta el conteo de días disponibles para cada uno de ellos, lo que facilita evaluar la completitud y distribución temporal de la información; y el tercero corresponde a la serie de tiempo del valor de las acciones, fundamental para observar la trayectoria histórica de precios y sus fluctuaciones.

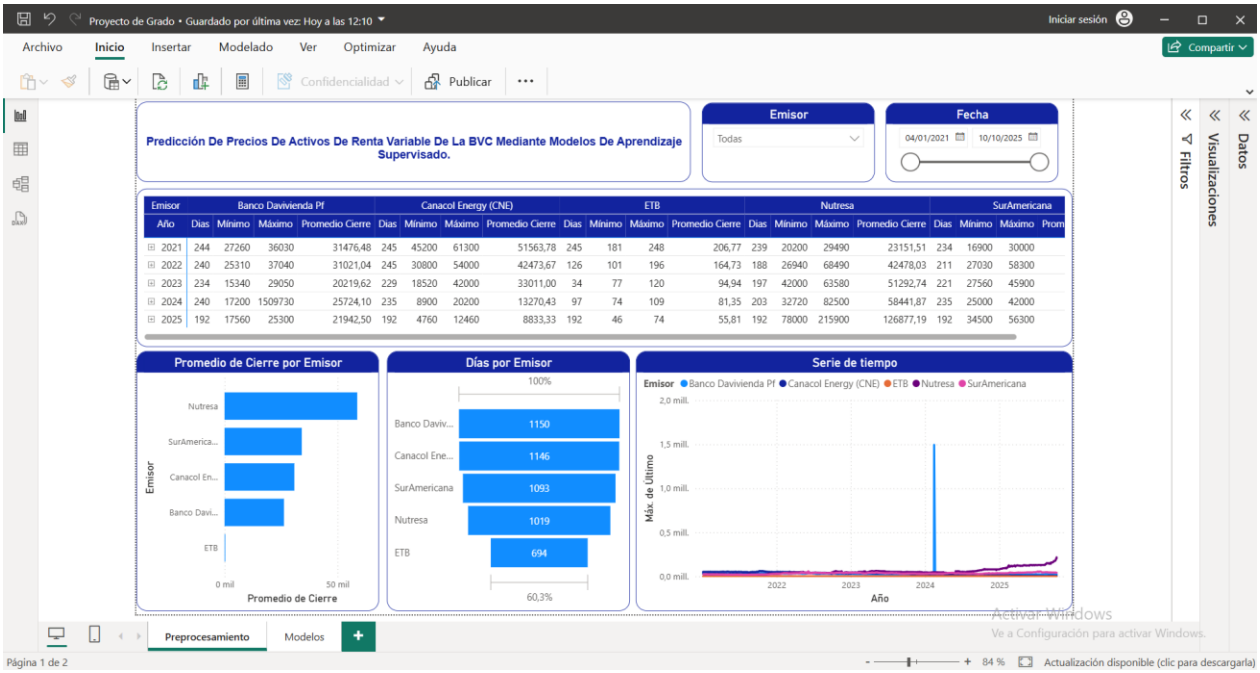


**Ilustración 8.** Filtro de métricas modelos Arboles basados en predicción de Precios.

En la segunda hoja se importan los valores de salida de los modelos supervisados, específicamente aquellos correspondientes a Random Forest, dado que fue el algoritmo que mostró el mejor desempeño al comparar los precios predichos con los valores reales. En la tabla principal se incluyen variables clave como el tipo de partición (TRAIN o TEST), el valor del coeficiente de determinación ($R^2$), el nombre del emisor y otras métricas de error relevantes. Adicionalmente, se incorporaron varios gráficos que facilitan el análisis visual del comportamiento del modelo. Estos permiten observar la distribución del $R^2$ por modelo y por emisor, comparando cómo varía su desempeño entre los conjuntos TRAIN y TEST.

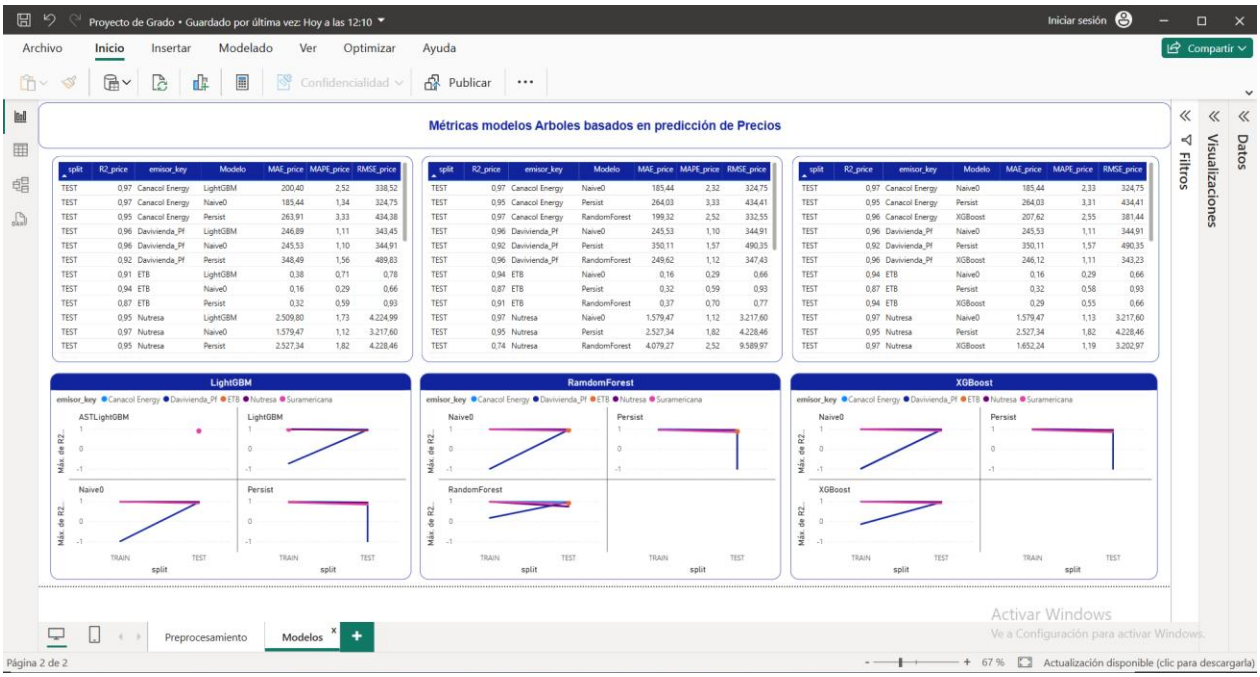


**Ilustración 9.** Valor de salida de los modelos.

**Enlace descarga Power Bi:**

https://drive.google.com/drive/folders/1v9mCuC0Oer5TDnYb_mRPeDZq4SNHXn1u?usp=sharing

# CONCLUSIONES

En el marco de la investigación titulada “Predicción de precios de activos de renta variable de la Bolsa de Valores de Colombia mediante modelos de aprendizaje supervisado”, los resultados obtenidos permiten extraer conclusiones metodológicas y empíricas sobre la capacidad real de distintos algoritmos para anticipar retornos y precios en un mercado emergente. El trabajo articuló un pipeline completo que abarca la depuración y enriquecimiento del panel de datos, la corrección de problemas de microestructura y multicolinealidad, la construcción de indicadores técnicos avanzados y la comparación sistemática entre modelos base, árboles de decisión, algoritmos de boosting y arquitecturas neuronales secuenciales (LSTM, GRU y CNN-LSTM).

Antes que nada, es necesario exponer que los resultados muestran que el modelo captura adecuadamente la dirección del movimiento del precio, aunque no logra estimar con precisión la magnitud del cambio. Esta diferencia explica la coexistencia de un Hit Rate favorable con un $R^2$ negativo en algunos emisores, dado que el primero evalúa aciertos direccionales mientras que el segundo penaliza fuertemente los errores en términos de distancia cuadrática respecto al valor real.

Así pues y, en primer lugar, los modelos base Naive0 y Persist cumplieron su función de referencia al mostrar que, en horizontes diarios, los retornos logarítmicos de las acciones analizadas se comportan de manera altamente volátil y cercana al ruido. Naive0 que asume retorno futuro igual a cero, obtuvo sistemáticamente MAE muy competitivos e incluso el mejor desempeño en varios emisores, mientras que Persist, que replica el retorno del periodo anterior alcanzó Hit Rates elevados, pero a costa de errores (MAE y RMSE) sensiblemente mayores y valores de $R^2$ fuertemente negativos.

En segundo lugar, los modelos basados en árboles de decisión (Random Forest, XGBoost y LightGBM) demostraron una mayor capacidad de ajuste en entrenamiento, aprovechando la riqueza del conjunto de variables construidas (microestructura, volumen, volatilidad e indicadores técnicos). Random Forest y XGBoost lograron mejores métricas en TRAIN y $R^2$ positivos en algunos activos, pero su desempeño en TEST fue más modesto: rara vez superan consistentemente a Naive0 y frecuentemente muestran signos claros de sobreajuste, reflejados en la brecha entre TRAIN y TEST. Dentro de esta familia, LightGBM se posiciona como el modelo más estable: reduce la pérdida relativa frente a Persist, obtiene $R^2$ menos negativos y, en emisores como Davivienda, Nutresa y Suramericana, alcanza un mejor equilibrio entre error (MAE, RMSE) y capacidad de acierto direccional (Hit Rate). No obstante, las mejoras siguen siendo moderadas y fuertemente dependientes de la estructura y volatilidad de cada activo.

En tercer lugar, las redes neuronales secuenciales introdujeron un enfoque más sofisticado para explotar dependencias temporales de mayor longitud. El modelo LSTM, entrenado de manera independiente por emisor y calibrado mediante búsquedas aleatorias de hiperparámetros, mostró un desempeño heterogéneo: si bien en TRAIN logra reducir errores respecto a Persist y alcanzar Hit Rates competitivos, en TEST su capacidad de generalización es limitada y los $R^2$ permanecen en general negativos.

El modelo GRU reproduce este patrón de forma aún más marcada: en algunos emisores se sitúa en una posición intermedia entre Naive0 y Persist, pero en otros (como ETB y Nutresa) su rendimiento es claramente inferior, con errores elevados y muy baja tasa de acierto direccional. Por su parte, la arquitectura híbrida CNN-LSTM destaca como el enfoque secuencial con mejor equilibrio global: aunque su MAE en TEST suele ser cercano al de Naive0 (es decir, no reduce drásticamente el error medio), alcanza Hit Rates superiores al 50 % en la mayoría de emisores y resultados sobresalientes en casos como ETB, donde supera el 90 % de acierto direccional. Esto indica que las redes secuenciales son capaces de capturar señales direccionales débiles que no se reflejan completamente en métricas de error cuadrático, pero que sí pueden ser relevantes en el diseño de estrategias de trading basadas en la dirección del movimiento.

En cuarto lugar, A diferencia de los retornos, la predicción del nivel de precios reveló una estructura marcadamente persistente, consistente con procesos de caminata aleatoria. Al evaluar las métricas de error y ajuste de los diferentes modelos incluyendo los métodos basados en aprendizaje automático y las arquitecturas neuronales se observó que el modelo Naive0 fue sistemáticamente superior. Este modelo, basado en la premisa de que el mejor pronóstico del precio futuro es el precio del periodo previo, obtuvo los menores valores de MAE, RMSE y MAPE, así como los mayores niveles de $R^2$ para la mayoría de emisores en las pruebas realizadas, lo que confirma la alta persistencia del nivel de precios.

Los modelos complejos, tales como Random Forest, XGBoost, LightGBM, LSTM, GRU y CNN-LSTM, no lograron superar al modelo Naive0 de manera consistente fuera de muestra. Entre ellos, LightGBM presentó el comportamiento más estable, con una degradación menor en las métricas y un ajuste relativamente más sólido, posicionándose como el segundo mejor modelo para la predicción de precios. Este resultado es coherente con la teoría financiera[9], que propone

[9] Este resultado es coherente con la literatura clásica de mercados financieros. Según la hipótesis de eficiencia de Fama [37] [38] y la evidencia empírica de [39] y [40], los precios de los activos siguen dinámicas cercanas a una caminata aleatoria, lo que implica que el mejor predictor del precio futuro es el precio actual. Bajo este esquema, los modelos complejos no pueden superar consistentemente a modelos simples como Naive0. Investigaciones posteriores, como las de [41] y los hallazgos de las M-competitions [42], confirman que en series altamente

que los modelos basados en relaciones no lineales pueden capturar ciertos patrones locales, pero no son capaces de vencer la persistencia inherente del proceso generador de precios.

En quinto lugar, el diseño del pipeline de datos e ingeniería de variables constituye una contribución metodológica central de este trabajo. La construcción de indicadores de calidad, microestructura, volumen y volatilidad, junto con indicadore), se complementó con una gestión explícita de la multicolinealidad mediante VIF y PCA selectivo. En particular, se protegieron variables clave como ret_log por su relevancia económica, mientras que un conjunto de 23 variables altamente colineales se condensó en 8 componentes principales ortogonales, utilizados exclusivamente en el pipeline de redes. A ello se suma una imputación causal ligera por emisor (forward-fill sin mirar el futuro) y un recorte de warmup de 20 días por emisor, lo que permitió entrenar los modelos sobre un panel limpio, consistente, sin fugas temporales y adecuado para diferentes familias de algoritmos. Aunque los modelos no logran "vencer" de manera contundente a los baselines, el proceso deja como resultado un marco metodológico replicable para futuros estudios en otros activos, frecuencias u horizontes.

Finalmente, los resultados abren varias líneas de trabajo futuro. En términos de horizonte, sería pertinente (i) explorar predicciones semanales o mensuales, donde los componentes predecibles puedan ser mayores y la señal supere al ruido diario; (ii) incorporar información macroeconómica, de mercado y de libro de órdenes, que complemente el historial puro de precios y volumen; (iii) profundizar en arquitecturas híbridas (por ejemplo, combinaciones de boosting con redes profundas o enfoques tipo Temporal Fusion Transformer) y en técnicas avanzadas de regularización temporal; y (iv) evaluar de manera explícita estrategias de inversión basadas en señales direccionales, más que en la predicción puntual del retorno, integrando costos de transacción y restricciones de liquidez.

persistentes los métodos parsimoniosos suelen ofrecer un desempeño igual o superior al de arquitecturas avanzadas como Random Forest, XGBoost, LightGBM o redes neuronales recurrentes.

## 5. ANEXOS

| Split | Emisor | Modelo | MAE | RMSE | MAPE | R2 |
|---|---|---|---|---|---|---|
| TEST | Canacol | Naive0 | 185.443787 | 324.746600 | 2.320213 | 0.971138 |
| | | Persist | 264.031903 | 434.407600 | 3.331566 | 0.948355 |
| | | RandomForest | 199.320976 | 332.547800 | 2.520549 | 0.969735 |
| | Davivienda | Naive0 | 245.529412 | 344.909200 | 1.102769 | 0.957989 |
| | | Persist | 350.107256 | 490.353200 | 1.572233 | 0.915088 |
| | | RandomForest | 249.620118 | 347.434900 | 1.121701 | 0.957372 |
| | ETB | Naive0 | 0.158416 | 0.656272 | 0.293283 | 0.937832 |
| | | Persist | 0.316379 | 0.931843 | 0.585074 | 0.874662 |
| | | RandomForest | 0.365127 | 0.770322 | 0.699794 | 0.914347 |
| | Nutresa | Naive0 | 1579.466667 | 3217.600000 | 1.119758 | 0.970644 |
| | | Persist | 2527.335169 | 4228.457000 | 1.815741 | 0.949302 |
| | | RandomForest | 4079.270346 | 9589.974000 | 2.515087 | 0.739228 |
| | Suramericana | Naive0 | 607.701863 | 1102.411000 | 1.349625 | 0.900693 |
| | | Persist | 861.808002 | 1413.019000 | 1.920961 | 0.836850 |
| | | RandomForest | 652.285412 | 1194.026000 | 1.444946 | 0.883502 |
| TRAIN | Canacol | Naive0 | 569.606099 | 809.675300 | 1.561451 | 0.995483 |
| | | Persist | 802.318531 | 1119.570000 | 2.163802 | 0.991364 |
| | | RandomForest | 569.484305 | 798.059100 | 1.564932 | 0.995612 |
| | Davivienda | Naive0 | 4090.03493 | 74202.02000 | 11.00594 | -0.99723 |
| | | Persist | 151074 | 4179247 | 760.73 | -6334.68 |
| | | RandomForest | 2143.189651 | 47686.51000 | 1.942830 | 0.175125 |
| | ETB | Naive0 | 2.270064 | 4.033401 | 1.531868 | 0.993966 |
| | | Persist | 3.773827 | 6.063799 | 2.567509 | 0.986362 |
| | | RandomForest | 2.271327 | 3.929664 | 1.537104 | 0.994272 |
| | Nutresa | Naive0 | 769.584527 | 1665.163000 | 1.831262 | 0.982137 |
| | | Persist | 1199.716155 | 2652.737000 | 2.825813 | 0.954665 |
| | | RandomForest | 736.386419 | 1501.817000 | 1.756640 | 0.985470 |
| | Suramericana | Naive0 | 664.560000 | 1181.035000 | 1.951691 | 0.978326 |
| | | Persist | 1007.305847 | 1725.291000 | 2.940178 | 0.953747 |
| | | RandomForest | 635.649185 | 1037.021000 | 1.878266 | 0.983289 |

Tabla 1. Métricas promedio por emisor y modelo RandomForest (Precio, TRAIN y TEST)

| Split | Emisor | Modelo | MAE | RMSE | R2 | MAPE |
|---|---|---|---|---|---|---|
| TEST | Canacol | Naive0 | 0.023299 | 0.044623 | -0.004890 | 176.331171 |
| | | Persist | 0.033189 | 0.057550 | -0.671432 | 149.754239 |
| | | XGBoost | 0.025584 | 0.048284 | -0.176506 | 173.095404 |
| | Davivienda | Naive0 | 0.011052 | 0.015587 | -0.010892 | 181.176086 |
| | | Persist | 0.015709 | 0.022076 | -1.027611 | 141.512049 |
| | | XGBoost | 0.011080 | 0.015512 | -0.001108 | 171.180245 |
| | ETB | Naive0 | 0.002924 | 0.012125 | -0.002269 | 13.861381 |
| | | Persist | 0.005848 | 0.017147 | -1.004538 | 27.722763 |
| | | XGBoost | 0.005461 | 0.012281 | -0.028198 | 198.911080 |
| | Nutresa | Naive0 | 0.011317 | 0.021690 | -0.033785 | 105.332740 |
| | | Persist | 0.018208 | 0.029310 | -0.887756 | 128.149977 |
| | | XGBoost | 0.011857 | 0.021579 | -0.023238 | 193.420677 |
| | Suramericana | Naive0 | 0.013427 | 0.023298 | -0.001204 | 178.881586 |
| | | Persist | 0.019192 | 0.030710 | -0.739528 | 142.376033 |
| | | XGBoost | 0.013858 | 0.023641 | -0.030878 | 164.711859 |
| TRAIN | Canacol | Naive0 | 0.016346 | 0.024385 | -0.004371 | 184.518516 |
| | | Persist | 0.023047 | 0.033192 | -0.860898 | 146.344431 |
| | | XGBoost | 0.016388 | 0.024268 | 0.005216 | 174.627782 |
| | Davivienda | Naive0 | 0.022168 | 0.199415 | -0.000008 | 185.818149 |
| | | Persist | 0.036703 | 0.345310 | -1.998538 | 147.526624 |
| | | XGBoost | 0.020203 | 0.160397 | 0.353029 | 169.359735 |
| | ETB | Naive0 | 0.014102 | 0.030458 | -0.006459 | 95.104728 |
| | | Persist | 0.024006 | 0.046479 | -1.343720 | 113.384769 |
| | | XGBoost | 0.014744 | 0.029616 | 0.048406 | 183.018278 |
| | Nutresa | Naive0 | 0.017716 | 0.035802 | -0.002739 | 173.820203 |
| | | Persist | 0.027227 | 0.050361 | -0.984078 | 149.058586 |
| | | XGBoost | 0.017254 | 0.034153 | 0.087471 | 167.321497 |
| | Suramericana | Naive0 | 0.018863 | 0.030272 | -0.000310 | 183.095080 |
| | | Persist | 0.028263 | 0.043683 | -1.082968 | 150.076062 |
| | | XGBoost | 0.018370 | 0.028724 | 0.099356 | 167.667815 |

Tabla 2. Métricas promedio por emisor y modelo XGBoost (Precio, TRAIN y TEST)

| Split | Emisor | Modelo | MAE | RMSE | MAPE | R2 |
|---|---|---|---|---|---|---|
| TRAIN | Canacol | Naive0 | 569.606099 | 809.675300 | 1.561451 | 0.995483 |
| | | Persist | 801.939557 | 1119.486000 | 2.163063 | 0.991366 |
| | | LightGBM | 568.333890 | 799.640100 | 1.562074 | 0.995595 |
| | Davivienda | Naive0 | 4090.034937 | 74202.020000 | 11.005944 | -0.997232 |
| | | Persist | 151074.7635 | 4179247.0000 | 760.7329 | -6334.6806 |
| | | LightGBM | 3815.5585 | 69008.9900 | 9.635936 | -0.727462 |
| | ETB | Naive0 | 2.270064 | 4.033401 | 1.531868 | 0.993966 |
| | | Persist | 3.758463 | 6.054624 | 2.560329 | 0.986403 |
| | | LightGBM | 2.365381 | 3.845024 | 1.609647 | 0.994516 |
| | Nutresa | Naive0 | 769.5845 | 1665.1630 | 1.831262 | 0.982137 |
| | | Persist | 1199.6732 | 2652.7350 | 2.825639 | 0.954665 |
| | | LightGBM | 764.9041 | 1564.7100 | 1.805256 | 0.984227 |
| | Suramericana | Naive0 | 664.5600 | 1181.0350 | 1.951691 | 0.978326 |
| | | Persist | 1006.371608 | 1725.0800 | 2.936116 | 0.953758 |
| | | LightGBM | 632.100823 | 1096.1330 | 1.859815 | 0.981330 |
| TEST | Canacol | Naive0 | 185.443787 | 324.746600 | 2.320213 | 0.971138 |
| | | Persist | 263.913775 | 434.383100 | 3.330500 | 0.948361 |
| | | LightGBM | 200.398527 | 338.521300 | 2.519003 | 0.968638 |
| | Davivienda | Naive0 | 245.529412 | 344.909200 | 1.102769 | 0.957989 |
| | | Persist | 348.491150 | 489.834300 | 1.563780 | 0.915268 |
| | | LightGBM | 246.894883 | 343.449400 | 1.109103 | 0.958344 |
| | ETB | Naive0 | 0.158416 | 0.656272 | 0.293283 | 0.937832 |
| | | Persist | 0.316379 | 0.931843 | 0.585074 | 0.874662 |
| | | LightGBM | 0.378774 | 0.783625 | 0.712733 | 0.911363 |
| | Nutresa | Naive0 | 1579.4666 | 3217.6000 | 1.119758 | 0.970644 |
| | | Persist | 2527.3359 | 4228.4570 | 1.815741 | 0.949302 |
| | | LightGBM | 2509.7957 | 4224.9850 | 1.729109 | 0.949385 |
| | Suramericana | Naive0 | 607.7018 | 1102.4110 | 1.349625 | 0.900693 |
| | | Persist | 862.0567 | 1413.2020 | 1.921607 | 0.836808 |
| | | LightGBM | 693.3896 | 1124.8640 | 1.520676 | 0.896607 |

Tabla 3. Métricas promedio por emisor y modelo LightGBM (Precio, TRAIN y TEST)

## 7. REFERENCIAS BIBLIOGRÁFICAS


[1] Shaw J, "How Machine Learning Aids Material Selection," 2024 Manufacturing Engineering, Southfield, United States, 2024, pp. 1-26, https://www.scopus.com/record/display.uri?eid=2-s2.0-85195308944&origin=scopusAI

[2] J. J. Cordova Calle, J. X. Farez Villa and R. I. Hurtado Ortiz, "An analysis method for predicting breast cancer using data science processes and machine learning," 2022 IEEE International Autumn Meeting on Power, Electronics and Computing (ROPEC), Ixtapa, Mexico, 2022, pp. 1-4, doi: 10.1109/ROPEC55836.2022.10018755.

[3] Moreno Morales y Quintana Duarte, "Moc-Teml: caso de estudio predicción de tendencia en los índices bursátiles de la Bolsa de Valores de Colombia," 2020 Universidad Distrital Franciso Jose de Caldas, Bogota, Colombia, 2020, pp. 1 – 43, http://hdl.handle.net/11349/25102

[4] Renugadevi A.S., Jayavadivel R., Charanya J., Kaviya P., Guhan R., "Machine Learning Fundamentals," 2025 Artificial Intelligence-Enabled Digital Twin for Smart Manufacturing, pp. 1 – 17, doi: 10.1002/9781394303601.ch

[5] Ignatenko V., Surkov A., Koltcov S., "Random forests with parametric entropybased information gains for classification and regression problems," 2024 PeerJ Computer Science, 10, art. no. e1775, doi: 10.7717/peerj-cs.1775

[6] Katal A., Singh N., "Artificial Neural Network: Models, Applications, and Challenges," 2022 EAI/Springer Innovations in Communication and Computing, pp. 235 - 257.doi: 10.1007/978-3-030-78284-9_11

[7] Khandelwal P., Konar J., Brahma B, "Training RNN and it's Variants Using Sliding Window Technique, "2020 IEEE International Students' Conference on Electrical, Electronics and Computer Science, SCEECS 2020, art. no. 9087240. doi: 10.1109/SCEECS48394.2020.93

[8] Gers F.A., Schraudolph N.N., Schmidhuber J, "Learning precise timing with LSTM recurrent networks," 2003, Journal of Machine Learning Research, 3 (1), pp. 115 – 143. doi: 10.1162/153244303768966139

[9] Hu Q., Zhang S., Xie Z., Mi J., Wan J., "Noise model based v-support vector regression with its application to short-term wind speed forecasting," 2014 Neural Networks, 57, pp. 1 – 11 doi: 10.1016/j.neunet.2014.05.003

[10] E. C. D. Estrada and C. H. F. Toro, "Value and risk in investment portfolios, critical variables calculated from an intelligent hybrid system based on CBR," 2014 9th Iberian Conference on Information Systems and Technologies (CISTI), Barcelona, Spain, 2014, pp. 1-4, doi: 10.1109/CISTI.2014.6876974.

[11] Mariño Villalba, J. A. Una comparación entre modelos estadísticos y de Machine Learning para la predicción de series de tiempo multivariadas (Doctoral dissertation, Universidad Nacional de Colombia).

[12] Estupiñán Romero, S. (2022). Predicción del precio de acciones en el mercado de valores con machine learning. Universidad de los Andes. Disponible en: http://hdl.handle.net/1992/59294

[13] Gutierrez, J., y Garcia, J. (2021). Prediccion del precio de acciones de la bolsa de valores estadounidense utilizando tecnicas de aprendizaje automatico basadas en datos de analisis tecnico y fundamental. Universidad Javeriana Cali. Disponible en: https://vitela.javerianacali.edu.co/handle/11522/2827

[14] S. S. Maini and K. Govinda, "Stock market prediction using data mining techniques," 2017 International Conference on Intelligent Sustainable Systems (ICISS), Palladam, India, 2017, pp. 654-661, doi: 10.1109/ISS1.2017.8389253.

[15] Plevris V., Solorzano G., Bakas N.P., Ben Seghier INVESTIGATION OF PERFORMANCE METRICS IN REGRESSION ANALYSIS AND MACHINE LEARNING-BASED PREDICTION MODELS (2022) World Congress in Computational Mechanics and ECCOMAS Congress DOI: 10.23967/eccomas.2022.155

[16] Ma Y., Han R., Wang W. “Portfolio optimization with return prediction using deep learning and machine learning,” (2021) Expert Systems with Applications, 165, art. no. 113973 doi: 10.1016/j.eswa.2020.113973

[17] Wang J., Zhou S. "Particle swarm optimization-XGBoost-based modeling of radio-frequency power amplifier under different temperaturas," (2024) International Journal of Numerical Modelling: Electronic Networks, Devices and Fields, 37 (2), art. no. e3168 doi: 10.1002/jnm.3168

[18] Berk, Richard A. “An introduction to ensemble methods for data analysis,” (2006) Sociological Methods and Research, 34 (3), pp. 263 - 295. doi: 10.1177/0049124105283119

[19] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, 785–794. https://doi.org/10.1145/2939672.2939785

[20] D. E. Q. Valencia, “Pronóstico de volatilidad de la TRM mediante un modelo híbrido LSTM-GARCH,” Ph.D. dissertation, Univ. del Rosario, 2019.

[21] J. H. Ratcliffe, “Geocoding crime and a first estimate of a minimum acceptable hit rate,” Int. J. Geogr. Inf. Sci., vol. 18, no. 1, pp. 61–72, 2004.

[22] O. A. Montesinos López, A. Montesinos López, and J. Crossa, “Convolutional neural networks,” in *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Cham: Springer International Publishing, 2022, pp. 533–577.

[28] P. Lara Benítez, “Predicción de series temporales en streaming mediante Deep Learning,” Ph.D. dissertation, Univ. de Sevilla, 2022.

[29] Y. A. Bermúdez-Bermúdez and M. H. Bohórquez, “La aplicación de técnicas de investigación en la detección y persecución de delitos económicos: mejores prácticas y desafíos en la cooperación internacional,” *Dixi*, vol. 27, pp. 1–14, 2025.

[30] H. Y. Kim and C. H. Won, “Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models,” *Expert Syst. Appl.*, vol. 103, pp. 25–37, 2018.

[31] Bancolombia, “¿Qué es la Bolsa de Valores de Colombia?” 2025. [Online]. Available: https://www.bancolombia.com/negocios/actualizate/administracion-y-finanzas/todo-sobre-bolsa-valores-de-colombia

[32] G. A. O. González, “Comparación de los modelos de Black-Litterman, Markowitz y CAPM en la estimación de los rendimientos esperados en el mercado de renta variable en Colombia,” *Rev. Estrategia Organizacional*

[33] D. Agudelo, *Inversiones en renta variable: Fundamentos y aplicaciones al mercado accionario colombiano*. Medellín, Colombia: Universidad EAFIT, 2021.

[34] Google Trends, “Riesgo financiero,” 2024. [En línea]. Disponible en: https://trends.google.es/trends/explore?date=today%205-y&geo=CO&q=%2Fm%2F07fpdg&hl=es

[35] R. M. Suárez Rodríguez, P. A. Petro Padilla, M. Financieros, y J. W. Pinedo López, Análisis del mercado de Forex en una economía global. 2024. [En línea]. Disponible en: https://hdl.handle.net/20.500.12494/55878

[36] G. E. P. Box, G. M. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*, 5th ed. Hoboken, NJ, USA: Wiley, 2016.

[37] E. F. Fama, “Efficient Capital Markets: A Review of Theory and Empirical Work,” *Journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970.

[38] E. F. Fama, “Efficient Capital Markets II,” *Journal of Finance*, vol. 46, no. 5, pp. 1575–1617, 1991.

[39] A. W. Lo and A. C. MacKinlay, “Stock Market Prices Do Not Follow Random Walks: Evidence from a Simple Specification Test,” *The Review of Financial Studies*, vol. 1, no. 1, pp. 41–66, 1988.

[40] A. W. Lo and A. C. MacKinlay, *A Non-Random Walk Down Wall Street*. Princeton, NJ: Princeton University Press, 1999.

[41] R. S. Tsay, *Analysis of Financial Time Series*, 3rd ed. Hoboken, NJ: Wiley, 2010.

[42] S. Makridakis, E. Spiliotis, and V. Assimakopoulos, “The M4 Competition: 100,000 Time Series and Forecasting Results,” *International Journal of Forecasting*, vol. 36, no. 1, pp. 54–74, 2020.

[43] J. J. Flores, F. Calderón y J. O. C. Lara, “Aprendizaje de modelos difusos para predicción de series de tiempo,” Aprendizaje de Modelos Difusos para Predicción de Series de Tiempo, vol. 10, 2015.

[44] M. P. G. Casimiro, Análisis de series temporales: Modelos ARIMA, vol. 1, no. 1, pp. 1–169, Universidad del País Vasco, 2009.

[45] L. M. O. Gutiérrez, G. A. Z. Ramírez y P. A. R. Rendón, “Prueba de no linealidad para series temporales financieras,” Scientia et Technica, vol. 1, no. 47, pp. 71–76, 2011.

[46] R. Cont, “Empirical properties of asset returns: Stylized facts and statistical issues,” Quantitative Finance, vol. 1, no. 2, pp. 223–236, 2001.

[47] John Y. Campbell, Samuel B. Thompson, Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average? The Review of Financial Studies, Volume 21, Issue 4, July 2008, Pages 1509–1531, https://doi.org/10.1093/rfs/hhm055