

Predicción del retorno de Bitcoin: el papel del volumen y los indicadores técnicos en modelos de pronóstico avanzados.

Danny Alexander Ramirez Rojas

Trabajo de Grado para Optar el Título de Especialista en Estadística

Director

Carlos Alfonso Mantilla Duarte

M. Sc. en Estadística Aplicada

Universidad Industrial de Santander

Facultad de Ciencias

Escuela de Matemáticas

Especialización en Estadística

Bucaramanga

2026

Tabla de Contenido

	Pag.
Introducción	13
1. Descripción del Problema	14
2. Antecedentes	16
3. Justificación.....	18
4. Objetivos	20
4.1. Objetivo general	20
4.2. Objetivos específicos.....	20
4.2.1. Construir el feature set reproducible	20
4.2.2. Comparar familias de modelos.....	20
4.2.3. Validar con rolling-origin.....	20
4.2.4. Interpretar el aporte del volumen y de los indicadores técnicos	21
4.3. Criterio de comparación empírica entre enfoques univariado y multivariado	21
5. Marco Teórico	22
5.1. Series de tiempo: conceptos básicos	22
5.2. Series de tiempo multivariadas (MTS).....	23
5.2.1. VAR y VARX	24
5.2.2. Modelos ARIMAX y ETSX.....	24
5.2.3. Modelos en espacio de estados y Filtro de Kalman	25
5.2.4. Modelos aditivos (Prophet y NeuralProphet):.....	26
5.2.5. Volatilidad condicional	27
5.2.6. Dimensionalidad y regularización.....	28
5.3. Retornos logarítmicos	28

5.4. Indicadores técnicos y volumen como predictores	30
5.5. Modelos de referencia para pronóstico	31
5.6. Modelos Clásicos:	32
5.7. Modelos con regresores exógenos.....	33
5.8. Modelos en espacio de estados.....	34
5.9. Modelos de machine learning	35
5.10. Modelos Deep learning	36
5.11. Modelos aditivos	37
5.12. Validación temporal y rolling-origin.....	38
5.13. Métricas de evaluación y prueba Diebold-Mariano	40
6. Metodología	41
6.1. Tipo de investigación	41
6.2. Fuente de datos	42
6.3. Población y muestra	43
6.4. Variable dependiente.....	43
6.5. Variables independientes.....	45
6.6. Construcción del feature set	46
6.7. División entrenamiento-validación	50
6.8. Modelos implementados	51
6.9. Esquemas de validación	52
6.10. Métricas de evaluación.....	54
6.11. Herramientas computacionales	55
7. Análisis exploratorio de datos	56

7.1. Evolución del precio de Bitcoin	57
7.2. Distribución de retornos	59
7.3. Volatilidad histórica	60
7.4. Variables de volumen	61
7.5. Indicadores técnicos	63
7.6. Relaciones entre variables, correlaciones y multicolinealidad.....	64
7.6.1. Comparación visual entre variables	65
7.6.2. Matriz de correlación	67
7.6.3. Diagnóstico de multicolinealidad.....	70
7.7. Causalidad de Granger	72
7.8. Síntesis del análisis exploratorio	74
8. Modelos de pronóstico	76
8.1. Modelos de referencia	78
8.2. Modelos clásicos univariados.....	80
8.3. Modelos clásicos multivariados	81
8.4. Modelos en espacio de estados.....	82
8.5. Modelos de machine learning	83
8.6. Modelos de deep learning	85
8.7. Modelos aditivos	86
8.8. Modelos híbridos.....	87
9. Resultados	89
9.1. Resultados Hold-Out.....	90
9.2. Resultados Rolling-Origin $h=1$	94

9.3. Resultados Rolling-Origin $h=7$ multivariado	101
9.4. Comparación por familias de modelos	105
9.4.1. Comparación por métricas de error	106
9.4.2. Comparación por métricas direccionales	108
9.5. Comparación univariado vs. Multivariado.....	111
9.5.1. Comparación por métricas de error	112
9.5.2. Comparación por métricas direccionales	115
9.5.3. Contraste empírico descriptivo entre enfoques univariado y multivariado.....	117
9.6. Análisis de métricas direccionales	119
9.7. Prueba Diebold-Mariano	121
9.8. Selección de mejores modelos	125
9.8.1. Criterio multicriterio para la selección del modelo final.....	127
10. Discusión de resultados	129
10.1. Interpretación general de resultados.....	130
10.2. Aporte del enfoque multivariado.....	131
10.3. Relación entre error, direccionalidad e interpretabilidad	133
10.4. Discusión por condiciones de mercado	133
10.4.1. Construcción de regímenes de mercado.....	134
10.4.2. Desempeño predictivo por régimen de volatilidad	137
10.4.3. Desempeño predictivo por régimen de volumen.....	143
10.4.4. Estabilidad de los modelos entre regímenes	150
10.4.5. Síntesis del análisis por regímenes.....	159
10.5. Implicaciones estadísticas y financieras.....	162

10.6. Contraste con la literatura previa	166
10.7. Limitaciones del estudio.....	168
11. Conclusiones	170
12. Recomendaciones y trabajos futuros.....	173
Referencias Bibliográficas	176
Apéndices	180

Lista de tablas

	Pag.
Tabla 1 Estructura de la matriz multivariada de predictores	48
Tabla 2 Resumen de indicadores técnicos utilizados como variables predictoras	63
Tabla 3 Mejores resultados de la prueba de causalidad de Granger por variable exógena	72
Tabla 4 Síntesis del análisis exploratorio	76
Tabla 5 Familias de modelos implementadas.....	77
Tabla 6 Mejores modelos univariados Hold-Out por familia ordenados por MASE.....	91
Tabla 7 Mejores modelos multivariados Hold-Out por familia ordenados por MASE	92
Tabla 8 Resultados Rolling-Origin $h = 1$ de modelos univariados ordenados por MASE	95
Tabla 9 Resultados Rolling-Origin $h = 1$ de modelos multivariados ordenados por MASE	97
Tabla 10 Resultados Rolling-Origin $h=7$ de modelos multivariados ordenados por MASE.....	102
Tabla 11 Comparación de mejores modelos por familia y esquema de validación	106
Tabla 12 Comparación de mejores modelos por desempeño direccional y esquema de validación.	109
Tabla 13 Comparación entre mejores modelos univariados y multivariados según esquema de validación.	113
Tabla 14 Comparación entre mejores modelos univariados y multivariados según desempeño direccional.	116
Tabla 15 Contraste empírico descriptivo entre el mejor modelo univariado y el mejor modelo multivariado.	118
Tabla 16 Síntesis de resultados de la prueba Diebold-Mariano frente al modelo Naive	122
Tabla 17 Selección final de modelos según criterio de evaluación.....	126

Tabla 18	Criterios de clasificación de regímenes de mercado	136
Tabla 19	Comparación del MASE por modelo según régimen de volatilidad.....	138
Tabla 20	Comparación de la capacidad direccional por modelo según régimen de volatilidad.	142
Tabla 21	Comparación del MASE por modelo según régimen de volumen.....	144
Tabla 22	Comparación de la capacidad direccional por modelo según régimen de volumen ...	148
Tabla 23	Estabilidad del MASE entre regímenes de volatilidad y volumen.....	152
Tabla 24	Robustez del MASE frente a cambios de volatilidad y volumen.....	154
Tabla 25	Modelos más estables según MASE frente a cambios de volatilidad y volumen	157
Tabla 26	Mejores modelos por régimen de mercado según MASE.....	159
Tabla 27	Modelos líderes según error global y estabilidad entre regímenes	160
Tabla 28	Lectura financiera de variables relevantes del enfoque multivariado	164

Lista de figuras

	Pag.
Figura 1 Serie temporal dApéndice 1el precio de cierre de BTC-USD 57	
Figura 2 Serie temporal de retornos logarítmicos de BTC-USD	59
Figura 3 Volatilidad móvil de 7 y 30 días	60
Figura 4 Evolución del volumen de negociación de BTC-USD	61
Figura 5 Comparación visual entre retorno logarítmico y variables explicativas seleccionadas..	65
Figura 6 Matriz de correlaciones de Spearman entre variables del enfoque multivariado	68
Figura 7 Diagnóstico de multicolinealidad mediante el Factor de Inflación de la Varianza (VIF)	70
Figura 8 Pronósticos Hold-Out de modelos univariados seleccionados frente al retorno observado de BTC-USD.....	93
Figura 9 Pronósticos Hold-Out de mejores modelos multivariados por familia frente al retorno observado de BTC-USD.....	94
Figura 10 Comparación Rolling-Origin $h=1$ de modelos univariados seleccionados	99
Figura 11 Comparación Rolling-Origin $h=1$ de modelos multivariados seleccionados	100
Figura 12 Pronósticos Rolling-Origin $h=7$ de modelos multivariados seleccionados frente al retorno observado de BTC-USD.	104
Figura 13 Comparación del MASE por modelo según régimen de volatilidad	140
Figura 14 Comparación del MASE por modelo según régimen de volumen	146
Figura 15 Estabilidad predictiva del MASE frente a cambios de volatilidad y volumen	158
Figura 16 Comparación de importancia de variables: Random Forest vs XGBoost	163

Lista de Apéndices

	Pag.
Apéndice A. Comparación entre correlaciones Pearson y Spearman	180
Apéndice B. Resultados Hold-Out univariados ordenados por MASE	181
Apéndice C. Resultados Hold-Out de modelos multivariados	183
Apéndice D. Métricas completas de desempeño por régimen de volatilidad	185
Apéndice E. Métricas completas de desempeño por régimen de volumen	188
Apéndice F. Interpretabilidad e importancia de variables en modelos multivariados	190

Resumen

Título: Predicción del retorno de Bitcoin: el papel del volumen y los indicadores técnicos en modelos de pronóstico avanzados*

Autor: Danny Alexander Ramírez Rojas**

Palabras clave: Bitcoin, retorno logarítmico, volumen, indicadores técnicos, modelos predictivos, Rolling-Origin, MASE, machine learning

Descripción:

Este trabajo evalúa el aporte del volumen de negociación, la volatilidad histórica y los indicadores técnicos al pronóstico del retorno logarítmico diario de Bitcoin frente al dólar estadounidense (BTC-USD). Para ello, se construyó un conjunto de datos financieros con precios, volumen, rezagos del retorno e indicadores como RSI, MACD, ATR, OBV y CMF. La comparación se desarrolló entre modelos univariados, basados únicamente en la historia del retorno, y modelos multivariados que incorporan variables exógenas de mercado.

La metodología incluyó modelos de referencia, modelos clásicos de series de tiempo, modelos con regresores exógenos, algoritmos de machine learning, modelos de deep learning, modelos aditivos e híbridos. La evaluación se realizó mediante esquemas Hold-Out y Rolling-Origin, con horizontes de uno y siete días. Como métrica principal se utilizó MASE, complementada con MAE, RMSE, sMAPE y métricas direccionales como Accuracy, Precision, Recall, F1-score, MCC y AUC.

Los resultados muestran que los modelos multivariados, especialmente XGBoost y Random Forest, reducen el error predictivo frente a las mejores alternativas univariadas en los esquemas evaluados. También se identificó que algunos modelos, como GAM, presentan un desempeño direccional competitivo, aunque no siempre coinciden con los modelos de menor error en magnitud. En conjunto, el estudio evidencia que la información asociada al volumen, la volatilidad y los indicadores técnicos puede aportar valor predictivo al análisis de BTC-USD, aunque su utilidad depende del horizonte, el esquema de validación y el criterio de evaluación utilizado.

* Trabajo de Grado

** Especialización en Estadística. Facultad de Ciencias. Escuela de Matemáticas. Director: Carlos Alfonso Mantilla Duarte.

Abstract

Title: Bitcoin Return Prediction: The Role of Volume and Technical Indicators in Advanced Forecasting Models*

Author: Danny Alexander Ramírez Rojas**

Key Words: Bitcoin, log return, trading volume, technical indicators, forecasting models, Rolling-Origin, MASE, machine learning

Description:

This study evaluates the contribution of trading volume, historical volatility, and technical indicators to the forecasting of Bitcoin daily log returns against the United States dollar (BTC-USD). A financial dataset was built using prices, volume, return lags, and technical indicators such as RSI, MACD, ATR, OBV, and CMF. The analysis compares univariate models, based only on the historical behavior of returns, with multivariate models that incorporate exogenous market variables.

The methodology includes benchmark models, classical time series models, models with exogenous regressors, machine learning algorithms, deep learning architectures, additive models, and hybrid approaches. Model performance was evaluated using Hold-Out and Rolling-Origin validation schemes, with one-day and seven-day forecasting horizons. MASE was adopted as the main accuracy metric and was complemented with MAE, RMSE, sMAPE, and directional metrics such as Accuracy, Precision, Recall, F1-score, MCC, and AUC.

The results show that multivariate models, particularly XGBoost and Random Forest, reduce forecasting error compared with the best univariate alternatives under the evaluated validation schemes. The findings also indicate that some models, such as GAM, provide competitive directional performance, although they do not always coincide with the models that achieve the lowest magnitude error. Overall, the study provides empirical evidence that volume, volatility, and technical indicators can add predictive value to BTC-USD return forecasting, although their usefulness depends on the forecasting horizon, validation scheme, and evaluation criterion considered.

* Degree Work

** Specialization in Statistics. Faculty of Sciences. School of Mathematics. Advisor: Carlos Alfonso Mantilla Duarte.

Introducción

Durante la última década, Bitcoin (BTC) se ha consolidado como uno de los activos digitales más relevantes del mercado financiero global. Su alta volatilidad, los cambios frecuentes en la liquidez y el componente especulativo de sus movimientos han motivado el desarrollo de modelos cuantitativos para pronosticar sus retornos y apoyar el análisis de riesgo. Sin embargo, anticipar el retorno de Bitcoin sigue siendo una tarea compleja debido a la presencia de relaciones no lineales, episodios de alta variabilidad y cambios en las condiciones del mercado (Kristoufek, 2015; Catania, Grassi, & Ravazzolo, 2019; Bouri, Gil-Alana, Gupta, & Roubaud, 2020).

La literatura ha mostrado que el volumen de negociación, la volatilidad histórica y diversos indicadores técnicos pueden contener señales útiles para estudiar el comportamiento de los mercados de criptomonedas, especialmente en periodos de alta actividad o movimientos abruptos de precio (Vidal-Tomás & Ibáñez, 2023; Anastasiou, Ballis, & Drakos, 2023; Guo, Sang, Tu, & Wang, 2024). Estas variables permiten ampliar la información disponible más allá de los rezagos del retorno. Por ello, los modelos multivariados constituyen una alternativa frente a los enfoques univariados, que utilizan únicamente la historia de la variable objetivo (Shumway & Stoffer, 2025; Hyndman & Athanasopoulos, 2021).

Este trabajo evalúa el aporte del volumen de negociación y de los indicadores técnicos al pronóstico del retorno logarítmico de Bitcoin frente al dólar estadounidense (BTC-USD), para horizontes de uno y siete días. Se comparan modelos univariados y multivariados de distintas familias: modelos de referencia, modelos clásicos de series de tiempo, modelos con regresores exógenos, modelos en espacio de estados, algoritmos de machine learning, arquitecturas de deep learning y modelos aditivos.

La evaluación se desarrolla con validación Hold-Out y Rolling-Origin. Se emplean métricas de error —MAE, RMSE, MASE y sMAPE— y métricas direccionales —Accuracy, Precision, Recall, F1-score, MCC y AUC—. Además, se utiliza la prueba de Diebold-Mariano para contrastar diferencias de precisión entre modelos y el benchmark de referencia cuando corresponde (Hyndman & Koehler, 2006; Diebold & Mariano, 1995).

El análisis se implementó mediante un pipeline reproducible en Python para la descarga, transformación, modelado y validación de los datos financieros. Además de comparar el error de pronóstico, el proyecto examina la estabilidad temporal de los modelos, su comportamiento bajo distintas condiciones de volatilidad y volumen, la capacidad para anticipar la dirección del retorno y la importancia relativa de las variables utilizadas.

De esta manera, el trabajo no se limita a identificar el modelo con menor error. Busca establecer en qué condiciones la información de volumen, volatilidad e indicadores técnicos contribuye al pronóstico de BTC-USD y qué modelos logran un mejor equilibrio entre precisión, estabilidad e interpretación financiera.

1. Descripción del Problema

El mercado de las criptomonedas, y en particular Bitcoin (BTC), se caracteriza por una elevada volatilidad, cambios abruptos en la liquidez y una dinámica de mercado altamente sensible a factores especulativos, tecnológicos y financieros. A diferencia de activos tradicionales, Bitcoin no cuenta con fundamentos convencionales claramente observables que faciliten la predicción de su comportamiento, lo cual dificulta la construcción de modelos de pronóstico estables y confiables. Diversos estudios han señalado que los retornos de Bitcoin presentan alta variabilidad, posibles cambios de régimen y patrones complejos que limitan la capacidad predictiva de los

modelos tradicionales (Kristoufek, 2015; Catania, Grassi, & Ravazzolo, 2019; Bouri, Gil-Alana, Gupta, & Roubaud, 2020).

Tradicionalmente, el análisis de series de tiempo financieras ha utilizado modelos univariados basados exclusivamente en la información histórica del precio o del retorno, como ARIMA, ETS o modelos de referencia tipo naive. Estos enfoques son útiles como punto de comparación y han sido ampliamente documentados en la literatura de pronóstico; sin embargo, pueden resultar limitados en contextos de alta volatilidad, baja autocorrelación lineal y presencia de relaciones no lineales entre variables de mercado (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025). En el caso de Bitcoin, variables como el volumen de negociación, la volatilidad histórica y los indicadores técnicos pueden contener información complementaria sobre la intensidad del mercado, el momentum, la presión compradora o vendedora y los cambios en la liquidez.

En este sentido, indicadores como RSI, MACD, ATR, OBV y CMF, junto con transformaciones del volumen, permiten representar dimensiones adicionales del comportamiento financiero del activo. La literatura reciente ha mostrado que la relación entre precio, volumen e indicadores técnicos puede variar en el tiempo y aportar información útil para el pronóstico de retornos en mercados de criptomonedas, aunque los resultados no siempre son concluyentes (Vidal-Tomás & Ibáñez, 2023; Anastasiou, Ballis, & Drakos, 2023; Guo, Sang, Tu, & Wang, 2024).

Pese a estas evidencias, aún existe incertidumbre sobre el verdadero aporte predictivo del volumen y de los indicadores técnicos frente a modelos basados únicamente en la información histórica del retorno. No es evidente si la inclusión de estas variables mejora de forma significativa la precisión del pronóstico del retorno logarítmico de Bitcoin, tanto en magnitud como en dirección, especialmente en horizontes de corto plazo como uno y siete días. Esta dificultad se acentúa porque

los retornos financieros suelen presentar bajo nivel de dependencia lineal, alta presencia de ruido y cambios frecuentes de signo.

En este contexto, surge la necesidad de evaluar empíricamente si la incorporación de información exógena asociada al volumen, la volatilidad y los indicadores técnicos mejora el desempeño predictivo de modelos multivariados frente a enfoques univariados. Para ello, resulta pertinente comparar distintas familias metodológicas, desde modelos de referencia y modelos clásicos hasta algoritmos de machine learning, deep learning y modelos aditivos, bajo esquemas rigurosos de validación temporal como hold-out y rolling-origin. Esta comparación permitirá determinar si la mayor complejidad de los modelos multivariados se traduce efectivamente en mejoras predictivas y si las variables exógenas aportan información relevante para explicar el comportamiento futuro del retorno logarítmico de Bitcoin.

2. Antecedentes

El mercado de las criptomonedas, y en particular Bitcoin (BTC), se caracteriza por una alta volatilidad, cambios abruptos en la liquidez y una fuerte sensibilidad frente a factores especulativos, tecnológicos y financieros. A diferencia de los activos tradicionales, Bitcoin no cuenta con fundamentos convencionales fácilmente observables que permitan anticipar su comportamiento. Esto dificulta la construcción de modelos de pronóstico estables y confiables. Estudios previos han documentado que sus retornos presentan alta variabilidad, posibles cambios de régimen y patrones complejos que limitan el desempeño de los modelos tradicionales (Kristoufek, 2015; Catania, Grassi, & Ravazzolo, 2019; Bouri, Gil-Alana, Gupta, & Roubaud, 2020).

El análisis de series de tiempo financieras suele partir de modelos univariados basados en la historia del precio o del retorno, como ARIMA, ETS y los modelos de referencia tipo naive. Estos enfoques son necesarios como línea base y cuentan con amplio respaldo en la literatura de pronóstico. Sin embargo, en un mercado como Bitcoin pueden ser insuficientes, debido a la baja autocorrelación lineal de los retornos, la presencia de ruido y las relaciones no lineales entre variables de mercado (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025).

El volumen de negociación, la volatilidad histórica y los indicadores técnicos aportan información que no está contenida únicamente en los rezagos del retorno. Variables como RSI, MACD, ATR, OBV y CMF permiten representar momentum, intensidad de negociación, presión compradora o vendedora, volatilidad y cambios en la liquidez. La literatura reciente ha encontrado que la relación entre precio, volumen e indicadores técnicos puede variar en el tiempo y, en algunos escenarios, aportar señales útiles para el pronóstico de retornos de criptomonedas (Vidal-Tomás & Ibáñez, 2023; Anastasiou, Ballis, & Drakos, 2023; Guo, Sang, Tu, & Wang, 2024).

Aun así, no está claro hasta qué punto estas variables mejoran el pronóstico del retorno logarítmico de Bitcoin frente a modelos que utilizan únicamente su comportamiento histórico. La incertidumbre es mayor en horizontes cortos, como uno y siete días, donde predominan el ruido, los cambios frecuentes de signo y la baja persistencia lineal de los retornos. Además, una mejora en la magnitud del pronóstico no necesariamente se traduce en una mejor predicción direccional.

Por esta razón, el problema central del estudio consiste en evaluar empíricamente si la incorporación de volumen, volatilidad e indicadores técnicos reduce el error predictivo de los modelos multivariados frente a los enfoques univariados. La evaluación incluye modelos de referencia, modelos clásicos, machine learning, deep learning y modelos aditivos, bajo validación temporal Hold-Out y Rolling-Origin. El propósito es identificar en qué escenarios la información

exógena aporta valor predictivo y cuáles familias de modelos logran un mejor desempeño para el retorno logarítmico de BTC-USD.

3. Justificación

El estudio de las criptomonedas, especialmente de Bitcoin (BTC), ha cobrado importancia en la investigación académica y financiera por la volatilidad del activo, los cambios rápidos de liquidez y el peso de factores especulativos. Estas condiciones hacen difícil anticipar sus retornos y plantean retos para la estadística aplicada, la econometría financiera y el aprendizaje automático. Analizar las variables asociadas a sus variaciones resulta útil para la gestión del riesgo y para el desarrollo de herramientas cuantitativas en escenarios de alta incertidumbre (Kristoufek, 2015; Catania, Grassi, & Ravazzolo, 2019; Bouri, Gil-Alana, Gupta, & Roubaud, 2020).

La evolución de Bitcoin no depende solamente de su precio pasado. Kristoufek (2015) relacionó su comportamiento con variables de interés y atención de mercado. Otros estudios han encontrado que los retornos de criptomonedas pueden asociarse con condiciones de mercado, actividad transaccional y relaciones entre activos (Bouri et al., 2020; Guo, Sang, Tu, & Wang, 2024). Por ello, el volumen de negociación, la volatilidad histórica y los indicadores técnicos pueden complementar la información contenida en los rezagos del retorno.

Aun así, el aporte predictivo de estas variables no está completamente resuelto. Indicadores como On-Balance Volume (OBV), Chaikin Money Flow (CMF), Relative Strength Index (RSI), MACD y ATR recogen señales de momentum, presión compradora o vendedora, intensidad de negociación, volatilidad y liquidez. Sin embargo, la evidencia disponible no permite afirmar que su incorporación mejore de manera uniforme el pronóstico del retorno de Bitcoin, especialmente

en horizontes cortos de uno y siete días (Vidal-Tomás & Ibáñez, 2023; Anastasiou, Ballis, & Drakos, 2023; Lahmiri & Bekiros, 2023).

El trabajo se justifica, entonces, en la necesidad de comparar si el volumen y los indicadores técnicos mejoran el desempeño predictivo frente a modelos univariados contruidos solo con la historia del retorno logarítmico. Se consideran modelos de referencia, modelos clásicos, modelos con regresores exógenos, algoritmos de machine learning, arquitecturas de deep learning, modelos híbridos y modelos aditivos. La comparación no busca asumir que los modelos más complejos son siempre mejores; busca establecer qué familias aprovechan mejor la información disponible y bajo qué condiciones.

El análisis se implementa mediante un pipeline reproducible en Python para descargar, preparar, transformar y modelar los datos financieros de Bitcoin. La evaluación utiliza validación temporal Hold-Out y Rolling-Origin, métricas de error y métricas direccionales. También se emplea la prueba de Diebold-Mariano para contrastar diferencias de precisión frente al benchmark de referencia cuando corresponde (Hyndman & Koehler, 2006; Diebold & Mariano, 1995; Hyndman & Athanasopoulos, 2021).

Con este diseño, el proyecto aporta evidencia empírica sobre el papel del volumen, la volatilidad y los indicadores técnicos en el pronóstico del retorno logarítmico de BTC-USD. Los resultados permiten valorar el desempeño de los enfoques multivariados dentro del periodo y los esquemas de validación analizados, y dejan una base reproducible para futuros estudios de predicción, riesgo y análisis cuantitativo de activos digitales.

4. Objetivos

4.1. Objetivo general

Evaluar el aporte predictivo del volumen de negociación y de los indicadores técnicos en el pronóstico del retorno logarítmico de BTC-USD, comparando modelos univariados y multivariados en horizontes de uno y siete días, bajo esquemas de validación temporal hold-out y rolling-origin.

4.2. Objetivos específicos

4.2.1. *Construir el feature set reproducible*

Construir un conjunto reproducible de variables de precio, volumen, volatilidad e indicadores técnicos para modelar el retorno logarítmico de BTC-USD en horizontes de uno y siete días, garantizando alineación temporal y control de fuga de información.

4.2.2. *Comparar familias de modelos*

Comparar modelos univariados y multivariados de distintas familias metodológicas —referencia, clásicos, machine learning, deep learning, híbridos y aditivos— para evaluar su desempeño en el pronóstico del retorno logarítmico de BTC-USD bajo validación temporal.

4.2.3. *Validar con rolling-origin*

Validar el desempeño fuera de muestra de los modelos mediante un esquema rolling-origin, evaluando su estabilidad temporal y capacidad de generalización en horizontes de uno y siete días.

4.2.4. Interpretar el aporte del volumen y de los indicadores técnicos

Evaluar la relevancia del volumen, la volatilidad y los indicadores técnicos como variables predictoras del retorno logarítmico de BTC-USD, considerando su importancia en los modelos multivariados y su desempeño bajo distintos regímenes de mercado.

4.3. Criterio de comparación empírica entre enfoques univariado y multivariado

Con base en el objetivo general del estudio, se estableció un criterio de comparación empírica descriptiva entre el desempeño predictivo fuera de muestra de los modelos univariados y multivariados. Para ello, se utilizó el MASE como métrica principal de evaluación, dado que permite comparar el error relativo de los modelos de pronóstico y que valores menores indican menor error predictivo.

Se define $(MASE_M)$ como el valor del MASE obtenido por el mejor modelo multivariado evaluado, el cual incorpora variables exógenas asociadas al volumen, la volatilidad y los indicadores técnicos. De manera análoga, $(MASE_U)$ corresponde al valor del MASE obtenido por el mejor modelo univariado de comparación, construido únicamente a partir de la información histórica del retorno logarítmico de BTC-USD.

El propósito de esta comparación es verificar descriptivamente si, en los escenarios de validación considerados, la incorporación de variables exógenas se asocia con una reducción del error predictivo frente a los modelos univariados. Para ello, se comparan los valores observados de $(MASE_M)$ y $(MASE_U)$, así como la reducción relativa del error obtenida por el mejor modelo multivariado.

Esta comparación se desarrolla en el subcapítulo 9.5.3 y debe interpretarse como un contraste empírico descriptivo basado en los resultados obtenidos bajo los esquemas Hold-Out y

Rolling-Origin. Por tanto, no constituye una prueba estadística formal de hipótesis ni permite realizar inferencias de significancia estadística agregada entre los enfoques.

5. Marco Teórico

5.1. Series de tiempo: conceptos básicos

Una serie de tiempo es un conjunto de observaciones registradas en un orden cronológico, por lo general con una frecuencia definida. Puede expresarse como $(\{y_t\}_{t=1}^T)$, donde (y_t) representa el valor observado en el instante (t) y (T) corresponde al número total de datos disponibles. La característica esencial de una serie de tiempo es, por tanto, la secuencia temporal de las observaciones.

Una serie puede presentar tendencia, estacionalidad, ciclos, ruido, cambios de nivel o variabilidad cambiante. Estos elementos no hacen parte de su definición, sino que describen comportamientos que deben identificarse en cada aplicación. En las series financieras es frecuente encontrar precios no estacionarios, episodios de alta volatilidad, valores extremos y cambios en las condiciones del mercado.

La estacionariedad es una propiedad relevante para varios modelos, pero no es un requisito para que una secuencia de datos sea considerada una serie de tiempo. En términos generales, una serie estacionaria mantiene una media, una varianza y una estructura de autocovarianza relativamente estables a lo largo del tiempo. Modelos como ARIMA, SARIMA y ETS requieren evaluar esta propiedad o trabajar con transformaciones que permitan aproximarse a ella antes de la estimación (Shumway & Stoffer, 2025; Hyndman & Athanasopoulos, 2021).

En el análisis financiero, los precios suelen mostrar tendencia y cambios importantes de escala. Por esta razón, es habitual transformarlos en retornos, en particular retornos logarítmicos,

para analizar variaciones relativas en lugar de niveles absolutos. La sección 5.3 presenta esta transformación, utilizada como variable objetivo en el presente estudio.

Los retornos de criptomonedas también presentan dificultades propias: heterocedasticidad, colas pesadas, cambios de régimen y posibles relaciones no lineales. Estas características pueden reducir el desempeño de los modelos que dependen únicamente de los rezagos de la serie. Por ello, el proyecto compara modelos univariados, construidos con la historia del retorno logarítmico, con modelos multivariados que incorporan volumen, volatilidad e indicadores técnicos como predictores adicionales (Bouri et al., 2020).

5.2. Series de tiempo multivariadas (MTS)

Una serie de tiempo multivariada permite analizar simultáneamente varias variables observadas a lo largo del tiempo. En este proyecto, el enfoque multivariado se utiliza para incorporar información adicional al retorno logarítmico de Bitcoin, incluyendo variables de precio, volumen, volatilidad e indicadores técnicos como RSI, MACD, ATR, OBV y CMF.

Formalmente, una serie multivariada puede representarse como:

$$y_t = (y_{1,t}, y_{2,t}, \dots, y_{k,t})'. \quad (\text{Ecuación 1})$$

Como se observa en la Ecuación 1, y_t representa un vector de k variables observadas en el tiempo t . Este enfoque permite estudiar relaciones dinámicas, efectos rezagados y posibles interacciones entre variables financieras. En el caso de Bitcoin, resulta especialmente relevante porque el volumen de negociación y los indicadores derivados pueden contener información útil sobre la dinámica futura de los retornos (Guo et al., 2024).

5.2.1. *VAR y VARX*

El modelo **Vector Autorregresivo VAR(p)** permite representar un sistema de variables interrelacionadas, donde cada variable depende linealmente de sus propios rezagos y de los rezagos de las demás variables del sistema. Su forma general es:

$$y_t = c + \sum_{i=1}^p A_i y_{t-i} + u_t, \quad u_t \sim (0, \Sigma) \quad (\text{Ecuación 2})$$

Como se observa en la Ecuación 2, y_t es un vector de dimensión $k \times 1$ que agrupa las variables de interés (por ejemplo, precio, volumen e indicadores técnicos), A_i son matrices de coeficientes, c es un vector de constantes y u_t representa un término de error con matriz de covarianza Σ .

Cuando se incluyen regresores exógenos x_i —como indicadores técnicos o medidas de volumen—, el modelo se extiende a un VARX, que combina la estructura multivariada con información externa. Este marco permite realizar pruebas de causalidad de Granger y analizar funciones impulso–respuesta, útiles para examinar los efectos dinámicos entre el precio y el volumen.

En el contexto de las criptomonedas, su aplicación ha demostrado la existencia de efectos lead–lag (adelantos y rezagos) entre el volumen y el precio, lo que refuerza la pertinencia de los modelos multivariados para capturar interdependencias entre variables de mercado (Shumway & Stoffer, 2025; Guo et al., 2024).

5.2.2. *Modelos ARIMAX y ETSX*

Los modelos ARIMAX (AutoRegressive Integrated Moving Average with Exogenous variables) y ETSX (Error–Trend–Seasonal con regresores) extienden las versiones univariadas clásicas de ARIMA y ETS al incluir variables exógenas que pueden aportar información adicional

relevante, como indicadores técnicos o medidas de volumen. Su estructura general puede representarse como:

$$y_t = \omega(B)x_t + n_t, \quad n_t \sim ARIMA, \quad (\text{Ecuación 3})$$

Como se observa en la Ecuación 3, x_t representa las variables exógenas, $w(B)$ recoge su efecto dinámico mediante rezagos y n_t corresponde al componente temporal modelado mediante una estructura ARIMA.

Estos modelos son relevantes para el proyecto porque permiten evaluar si la incorporación de información externa mejora el pronóstico del retorno logarítmico de BTC-USD frente a modelos univariados. Además, mantienen una estructura interpretable y sirven como punto de comparación frente a modelos más flexibles de machine learning y deep learning (Hyndman & Athanasopoulos, 2021).

5.2.3. Modelos en espacio de estados y Filtro de Kalman

Los modelos en espacio de estados permiten representar procesos dinámicos mediante componentes no observables, como tendencias, niveles o parámetros que varían en el tiempo. Su estructura general se define por dos ecuaciones:

$$\text{Medición: } y_t = Z_t \alpha_t + d_t + \varepsilon_t, \quad \text{Estado: } \alpha_t = T_t \alpha_{t-1} + R_t \eta_t \quad (\text{Ecuación 4})$$

Como se observa en la Ecuación 4, α_t representa el vector de estado (factores latentes), y_t la variable observada, y los términos de error ε_t, η_t son ruidos gaussianos con medias cero y varianzas controladas por matrices de covarianza específicas.

El Filtro de Kalman es el algoritmo recursivo que se emplea para estimar los estados no observables y actualizar las predicciones conforme se incorporan nuevas observaciones. Su flexibilidad lo hace especialmente útil en contextos de alta volatilidad como el de las

criptomonedas, ya que permite el seguimiento dinámico de los cambios en la media y la varianza de los retornos (Shumway & Stoffer, 2025).

En el contexto de Bitcoin, estos modelos son útiles porque permiten capturar cambios dinámicos en la estructura de la serie y adaptarse a contextos de alta volatilidad. En este proyecto, este enfoque se considera dentro de las familias de modelos clásicos y estructurales evaluadas para el pronóstico del retorno logarítmico de BTC-USD.

En el desarrollo empírico del proyecto, este enfoque aparece como **modelo estructural ETSX con variables exógenas**, no como un capítulo independiente de “Filtro de Kalman”. Por tanto, su función dentro del estudio es servir como modelo clásico multivariado interpretable y como punto de comparación frente a ARIMAX, AutoARIMAX, SARIMAX y modelos más flexibles de machine learning y deep learning.

5.2.4. Modelos aditivos (Prophet y NeuralProphet):

El modelo Prophet, desarrollado por Taylor y Letham (2017), se enmarca dentro de los modelos aditivos descomponibles, en los cuales la serie de tiempo se representa como la suma de componentes de tendencia, estacionalidad y efectos de calendario:

$$y_t = g(t) + s(t) + h(t) + \varepsilon_t \quad (\text{Ecuación 5})$$

Como se observa en la Ecuación 5, $g(t)$ representa una tendencia segmentada con posibles puntos de cambio (changepoints), $s(t)$ modela las estacionalidades mediante series de Fourier, $h(t)$ captura efectos de calendario o regresores exógenos, y ε_t es un término de error aleatorio.

Prophet permite incorporar variables externas mediante regresores, mientras que NeuralProphet amplía este enfoque al incluir componentes autorregresivos y redes neuronales. Por su parte, los modelos GAM permiten representar relaciones no lineales mediante funciones suaves. En este proyecto, estos modelos se utilizan como alternativas flexibles e interpretables para evaluar

si el volumen, la volatilidad y los indicadores técnicos mejoran el pronóstico del retorno logarítmico de Bitcoin (Taylor & Letham, 2017; Triebe et al., 2021).

5.2.5. *Volatilidad condicional*

La volatilidad es un componente central en el análisis de series financieras, ya que refleja la magnitud de las fluctuaciones de los retornos. En activos como Bitcoin, es común observar periodos alternados de alta y baja variabilidad, fenómeno conocido como agrupamiento de volatilidad o volatility clustering.

Los modelos ARCH y GARCH permiten representar este comportamiento mediante una varianza condicional que depende de choques pasados y de la volatilidad previa. En particular, los modelos GARCH, propuestos a partir del trabajo de Engle (1982) y extendidos por Bollerslev (1986), expresan la dinámica de la varianza condicional de la siguiente manera:

$$\sigma_t^2 = \alpha_0 + \alpha_1 \cdot \varepsilon_{t-1}^2 + \beta_1 \cdot \sigma_{t-1}^2 \quad (\text{Ecuación 6})$$

Como se observa en la Ecuación 6, σ_t^2 representa la varianza condicional en el tiempo (t), ε_{t-1}^2 recoge el efecto de los choques pasados y σ_{t-1}^2 incorpora la persistencia de la volatilidad previa.

Aunque este proyecto no se centra en la modelación explícita de la varianza mediante modelos GARCH, sí incorpora medidas de volatilidad como variables explicativas, entre ellas rolling_vol_7, rolling_vol_30 y ATR. Estas variables permiten capturar cambios en la intensidad de las fluctuaciones del mercado y evaluar si la volatilidad reciente aporta información útil para el pronóstico del retorno logarítmico de BTC-USD (Engle, 1982; Bollerslev, 1986; Bouri et al., 2020).

5.2.6. Dimensionalidad y regularización

La incorporación de múltiples variables exógenas, rezagos e indicadores técnicos puede aumentar la dimensionalidad del problema y generar riesgos de multicolinealidad o sobreajuste. Por esta razón, resulta importante aplicar criterios de selección, regularización y validación temporal que permitan conservar la capacidad de generalización de los modelos.

En los modelos de machine learning, técnicas como regularización, reducción de complejidad, early stopping, dropout y control de hiperparámetros ayudan a evitar que los modelos aprendan patrones espurios del conjunto de entrenamiento. En particular, métodos como XGBoost incorporan mecanismos de regularización propios, mientras que las redes neuronales pueden utilizar estrategias como dropout y detención temprana para mejorar su estabilidad (Hastie, Tibshirani, & Friedman, 2017; Chen & Guestrin, 2016; Lahmiri & Bekiros, 2023).

En este proyecto, estos elementos son relevantes porque el enfoque multivariado integra variables de volumen, volatilidad, momentum y flujo de mercado. Por tanto, la correcta construcción del feature set, el control de fuga de información y la validación temporal son fundamentales para evaluar de manera confiable el aporte predictivo de las variables exógenas.

5.3. Retornos logarítmicos

En este proyecto, el retorno logarítmico de BTC-USD constituye la variable objetivo del análisis predictivo. Esta transformación permite estudiar las variaciones relativas del precio de cierre de Bitcoin en lugar de trabajar directamente con la serie de precios, la cual suele presentar tendencia, cambios de nivel y comportamiento no estacionario.

El retorno logarítmico se define como:

$$r_t = \log\left(\frac{c_t}{c_{t-1}}\right) \quad (\text{Ecuación 7})$$

Donde se observa en la Ecuación 7, C_t representa el precio de cierre de Bitcoin en el periodo t , y C_{t-1} corresponde al precio de cierre del periodo anterior. Para horizontes superiores, como siete días, el retorno puede extenderse a:

$$r_t^{(h)} = \log\left(\frac{C_{t+h}}{C_t}\right) \quad (\text{Ecuación 8})$$

Como se observa en la Ecuación 8, h representa el horizonte de pronóstico.

La principal ventaja de esta transformación es que permite analizar cambios proporcionales del precio y no variaciones absolutas. Además, los retornos logarítmicos suelen presentar una estructura más estable que los precios en niveles, lo cual facilita la aplicación de modelos estadísticos, algoritmos de machine learning y arquitecturas de deep learning en series financieras (Shumway & Stoffer, 2025; Hyndman & Athanasopoulos, 2021).

En el caso de Bitcoin, trabajar con retornos logarítmicos resulta especialmente pertinente debido a la alta volatilidad del activo y a la presencia de valores extremos, colas pesadas y agrupamientos de volatilidad, características frecuentes en mercados de criptomonedas (Bouri et al., 2020). Por esta razón, el análisis del presente estudio se centra en pronosticar el retorno logarítmico de BTC-USD, tanto en magnitud como en dirección.

A partir de esta variable objetivo se comparan dos enfoques principales: modelos univariados, basados únicamente en la historia del retorno, y modelos multivariados, que incorporan información adicional de volumen, volatilidad e indicadores técnicos. Esta comparación permite evaluar si las variables exógenas aportan capacidad predictiva adicional frente a los modelos que utilizan exclusivamente la dinámica histórica del retorno.

5.4. Indicadores técnicos y volumen como predictores

A partir de las variables financieras básicas OHLCV —apertura, máximo, mínimo, cierre y volumen— es posible construir indicadores técnicos y transformaciones del volumen que resumen información asociada al momentum, la tendencia, la volatilidad, la liquidez y el flujo de mercado. En este proyecto, estos indicadores no se utilizan como reglas de negociación, sino como variables explicativas dentro del enfoque multivariado, con el propósito de evaluar si aportan información adicional para el pronóstico del retorno logarítmico de BTC-USD.

Entre las variables consideradas se incluyen indicadores de momentum y tendencia, como el RSI(14), el MACD y su línea de señal; indicadores de volatilidad, como el ATR(14) y las volatilidades móviles; e indicadores de flujo de mercado, como OBV y CMF(20). Adicionalmente, se incorporan transformaciones del volumen, como el volumen logarítmico, el promedio móvil de 20 días, el `volume_ratio` y el `volume_zscore`, las cuales permiten identificar episodios de alta o baja actividad relativa del mercado.

Desde una perspectiva financiera, el volumen de negociación puede reflejar cambios en la intensidad transaccional, la liquidez y la participación de los agentes del mercado. Por su parte, los indicadores técnicos permiten representar patrones derivados del comportamiento del precio y del volumen, tales como presión compradora o vendedora, aceleración de tendencia, amplitud de movimiento y cambios en la volatilidad. La literatura reciente ha señalado que la relación entre precio, volumen e indicadores técnicos puede ser dinámica y dependiente del régimen de mercado, especialmente en criptomonedas como Bitcoin (Vidal-Tomás & Ibáñez, 2023; Anastasiou, Ballis, & Drakos, 2023; Bouri et al., 2020).

En el marco de este estudio, estas variables se integran al feature set multivariado mediante una adecuada alineación temporal, generación de rezagos y control de fuga de información. Su

incorporación permite comparar modelos basados únicamente en la historia del retorno frente a modelos que incluyen información exógena de volumen, volatilidad e indicadores técnicos. De esta manera, el análisis permite evaluar empíricamente si dichas variables mejoran la precisión y la capacidad direccional del pronóstico del retorno logarítmico de Bitcoin.

5.5. Modelos de referencia para pronóstico

Los modelos de referencia, también denominados modelos ingenuos o baselines, constituyen el punto de comparación básico en todo ejercicio de pronóstico. Su propósito no es capturar estructuras complejas de la serie, sino establecer una línea base mínima que permita evaluar si modelos más sofisticados generan mejoras reales en el desempeño predictivo.

El modelo Naive asume que el mejor pronóstico para el siguiente periodo corresponde simplemente al último valor observado:

$$\hat{y}_{t+1} = y_t$$

Por su parte, el modelo Seasonal Naive o SNaive(m) replica el valor observado hace m periodos. En series diarias, cuando se desea considerar una posible referencia semanal, puede utilizarse (m = 7):

$$\hat{y}_{t+h} = y_{t+h-7}$$

Estos modelos son especialmente relevantes en series financieras, donde los retornos suelen presentar baja autocorrelación lineal y un componente aleatorio elevado. Por tanto, superar consistentemente a un modelo de referencia constituye un criterio mínimo para justificar el uso de modelos más complejos, como ARIMA, modelos con regresores exógenos, algoritmos de machine learning o arquitecturas de deep learning (Hyndman & Athanasopoulos, 2021; Hyndman & Koehler, 2006).

En el presente proyecto, los modelos **Naive** y **SNaive(7)** se utilizan como *benchmarks* para comparar el desempeño de los modelos univariados y multivariados en el pronóstico del retorno logarítmico de BTC-USD. De esta manera, permiten evaluar si la incorporación de variables como volumen, volatilidad e indicadores técnicos aporta una mejora predictiva frente a reglas simples basadas únicamente en la información histórica de la serie.

5.6. Modelos Clásicos:

Los modelos clásicos de pronóstico siguen siendo un referente importante en el análisis de series de tiempo por su interpretación clara, su parsimonia y su amplio uso en aplicaciones estadísticas y financieras. Entre ellos se encuentran ARIMA, SARIMA, ETS y Theta, junto con extensiones que incorporan variables externas, como ARIMAX, SARIMAX y ETSX (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025).

ARIMA representa la dependencia lineal de una serie a partir de sus valores pasados, sus errores y posibles diferenciaciones. Cuando existe un patrón estacional, SARIMA incorpora componentes autorregresivos y de medias móviles asociados a una periodicidad determinada. Los modelos ETS describen la serie mediante componentes de error, tendencia y estacionalidad. Theta, por su parte, ofrece una alternativa clásica sencilla que ha mostrado buen desempeño en diversos ejercicios de pronóstico (Hyndman & Athanasopoulos, 2021).

En este proyecto, los modelos clásicos sirven como referencia para pronosticar el retorno logarítmico de BTC-USD. Los modelos univariados utilizan únicamente la historia del retorno, mientras que ARIMAX, SARIMAX y ETSX incluyen variables de volumen, volatilidad e indicadores técnicos. De esta manera, es posible revisar qué ocurre al pasar de una estructura basada solo en rezagos a otra que incorpora información adicional del mercado.

Su principal ventaja es la facilidad para interpretar los resultados y el menor costo computacional frente a modelos más complejos. Sin embargo, Bitcoin presenta baja autocorrelación lineal, heterocedasticidad, cambios de régimen y relaciones que pueden ser no lineales. Por ello, los modelos clásicos se contrastan con algoritmos de machine learning, arquitecturas de deep learning y modelos aditivos. El interés no está en asumir que un enfoque más complejo será mejor, sino en verificar qué alternativa ofrece menor error fuera de muestra bajo los esquemas de validación definidos.

5.7. Modelos con regresores exógenos

Los modelos con regresores exógenos permiten incorporar variables externas a la serie objetivo con el propósito de mejorar la capacidad explicativa y predictiva del modelo. A diferencia de los modelos univariados, que utilizan únicamente la historia de la variable dependiente, estos enfoques integran información adicional proveniente de variables relacionadas con el comportamiento del mercado, como volumen, volatilidad e indicadores técnicos.

En el contexto de series de tiempo, una formulación general con variables exógenas puede expresarse como:

$$y_t = f(y_{t-1}, y_{t-2}, \dots, x_t, x_{t-1}, \dots) + \varepsilon_t \quad (\text{Ecuación 9})$$

En la Ecuación 9, y_t representa la variable objetivo, x_t corresponde al conjunto de variables exógenas observadas en el tiempo t , y ε_t es el término de error. En este proyecto, la variable objetivo es el retorno logarítmico de BTC-USD, mientras que las variables exógenas incluyen indicadores de volumen, volatilidad, momentum y flujo de mercado.

Entre los modelos clásicos con regresores exógenos se encuentran ARIMAX y SARIMAX. Estos modelos amplían la estructura ARIMA o SARIMA tradicional al incorporar variables explicativas externas, permitiendo evaluar si información adicional al retorno histórico contribuye

al pronóstico. De forma similar, los modelos estructurales con exógenas, como el modelo tipo ETSX implementado mediante estructuras de espacio de estados, permiten combinar componentes dinámicos de la serie con regresores externos (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025).

En el presente estudio, los regresores exógenos corresponden principalmente a variables derivadas del volumen, medidas de volatilidad histórica e indicadores técnicos como RSI, MACD, ATR, OBV y CMF. Su incorporación permite contrastar modelos univariados frente a modelos multivariados y evaluar si estas variables aportan información predictiva adicional para anticipar el retorno logarítmico de Bitcoin.

Desde el punto de vista metodológico, el uso de regresores exógenos exige una adecuada alineación temporal, generación de rezagos y control de fuga de información. Por esta razón, las variables explicativas deben construirse únicamente con información disponible hasta el momento del pronóstico. En este proyecto, este control resulta fundamental para garantizar que la comparación entre modelos sea válida bajo los esquemas de validación temporal hold-out y rolling-origin.

5.8. Modelos en espacio de estados

Los modelos en espacio de estados permiten representar la evolución temporal de una serie mediante componentes observables y no observables, como nivel, tendencia, estacionalidad o factores latentes. Esta formulación resulta útil cuando se requiere modelar procesos dinámicos cuyos parámetros pueden cambiar en el tiempo, una característica frecuente en series financieras altamente volátiles (Shumway & Stoffer, 2025).

De forma general, estos modelos se expresan mediante dos ecuaciones: una ecuación de medición, que relaciona la variable observada con los estados no observados, y una ecuación de

estado, que describe la evolución temporal de dichos componentes. El filtro de Kalman permite estimar y actualizar recursivamente los estados a medida que se incorporan nuevas observaciones, facilitando el seguimiento dinámico de la serie.

En el contexto del presente proyecto, los modelos en espacio de estados se consideran relevantes porque permiten incorporar estructuras dinámicas e incluso regresores exógenos dentro del proceso de pronóstico. En Python, este enfoque se implementa mediante el submódulo `statsmodels.tsa.statespace`, que incluye herramientas como `SARIMAX` y `UnobservedComponents`.

La técnica de espacio de estados se aplicó en el modelo estructural ETSX con variables exógenas, estimado mediante `UnobservedComponents` de `statsmodels`. Este modelo funcionó como una alternativa estructural multivariada que combina nivel local, componente estacional semanal y regresores externos asociados al volumen, volatilidad e indicadores técnicos.

5.9. Modelos de machine learning

Los modelos de machine learning permiten abordar problemas de pronóstico desde una perspectiva flexible, orientada al aprendizaje de patrones a partir de los datos. A diferencia de los modelos estadísticos clásicos, estos enfoques no requieren especificar de forma rígida una estructura funcional previa, sino que aprenden relaciones entre variables predictoras y la variable objetivo mediante algoritmos de optimización y partición del espacio de predictores (Hastie, Tibshirani, & Friedman, 2017).

En el contexto de series financieras, los modelos de machine learning resultan útiles porque pueden capturar relaciones no lineales, interacciones entre variables y efectos complejos que no siempre son identificados por modelos lineales tradicionales. Esta característica es especialmente relevante en mercados como Bitcoin, donde los retornos suelen presentar alta volatilidad, bajo nivel de autocorrelación lineal y posibles cambios de régimen (Lahmiri & Bekiros, 2023).

En este proyecto se utilizan modelos como MLP, Random Forest y XGBoost. El modelo MLP permite aproximar relaciones no lineales mediante una red neuronal multicapa; Random Forest combina múltiples árboles de decisión para mejorar estabilidad y reducir varianza; y XGBoost aplica un esquema de gradient boosting que construye árboles secuenciales orientados a corregir errores de predicción previos, incorporando mecanismos de regularización para controlar el sobreajuste (Chen & Guestrin, 2016).

Dentro del enfoque univariado, estos modelos emplean rezagos del retorno logarítmico como variables predictoras. En el enfoque multivariado, además de los rezagos de la variable objetivo, se incorporan variables de volumen, volatilidad e indicadores técnicos. De esta manera, los modelos de machine learning permiten evaluar si la información exógena mejora la capacidad predictiva frente a modelos basados únicamente en la historia del retorno.

Para garantizar una comparación válida, estos modelos se evalúan bajo esquemas de validación temporal, evitando la mezcla aleatoria de observaciones y respetando el orden cronológico de la serie. Esto resulta fundamental en series financieras, donde el uso de información futura podría generar fuga de información y sobreestimar artificialmente el desempeño predictivo.

5.10. Modelos Deep learning

Los modelos de deep learning secuencial permiten analizar series temporales mediante arquitecturas capaces de aprender dependencias dinámicas a partir de ventanas de observaciones pasadas. Dentro de esta familia se destacan las redes LSTM (Long Short-Term Memory) y GRU (Gated Recurrent Unit), diseñadas para capturar patrones temporales y relaciones no lineales en datos secuenciales (Hochreiter & Schmidhuber, 1997).

Las redes LSTM incorporan mecanismos de memoria y compuertas que regulan el flujo de información a través del tiempo, permitiendo conservar información relevante de periodos

anteriores y mitigar el problema del desvanecimiento del gradiente. Las redes GRU tienen una estructura similar, pero con una arquitectura más compacta, lo que puede reducir el costo computacional y facilitar el entrenamiento en ciertos contextos.

En el análisis financiero, estas arquitecturas resultan útiles porque los retornos pueden presentar relaciones no lineales, episodios de volatilidad agrupada y patrones temporales difíciles de capturar mediante modelos lineales tradicionales. Su aplicación en mercados de criptomonedas ha sido explorada por diversos estudios, especialmente por su capacidad para modelar dinámicas complejas en series altamente volátiles (McNally, Roche, & Caton, 2018; Lahmiri & Bekiros, 2020).

En el presente proyecto, los modelos LSTM y GRU se utilizan tanto en el enfoque univariado como en el multivariado. En el caso univariado, las redes reciben secuencias construidas a partir de rezagos del retorno logarítmico de BTC-USD. En el enfoque multivariado, las secuencias incorporan además variables de volumen, volatilidad e indicadores técnicos, permitiendo evaluar si esta información exógena mejora la capacidad predictiva.

Metodológicamente, estos modelos requieren transformar los datos en estructuras tridimensionales de la forma muestras, ventana temporal y número de variables predictoras. Además, se aplican procedimientos de escalamiento, entrenamiento con validación temporal y técnicas de regularización como early stopping y dropout, con el fin de reducir el sobreajuste y mejorar la capacidad de generalización del modelo.

5.11. Modelos aditivos

Los modelos aditivos permiten representar una serie temporal como la combinación de componentes interpretables, tales como tendencia, estacionalidad, efectos de calendario y

regresores externos. Esta estructura resulta útil en problemas de pronóstico donde se busca combinar flexibilidad predictiva con una lectura clara de los componentes del modelo.

Dentro de esta familia, Prophet modela la serie como una suma de componentes de tendencia, estacionalidad y efectos adicionales. Su formulación general puede expresarse como:

$$y_t = g(t) + s(t) + h(t) + \varepsilon_t \quad (\text{Ecuación 10})$$

En la Ecuación 10, $g(t)$ representa la tendencia, $s(t)$ la estacionalidad, $h(t)$ los efectos adicionales o regresores externos, y $\varepsilon(t)$ el término de error. Prophet permite incorporar variables exógenas mediante regresores, lo cual facilita incluir información de volumen, volatilidad e indicadores técnicos dentro del proceso de pronóstico (Taylor & Letham, 2017).

NeuralProphet extiende la lógica de Prophet al incorporar componentes autorregresivos y redes neuronales, manteniendo una estructura aditiva e interpretable, pero con mayor flexibilidad para capturar patrones temporales complejos (Triebe et al., 2021). Por su parte, los modelos GAM permiten representar relaciones no lineales mediante funciones suaves, lo cual resulta útil cuando el efecto de una variable explicativa sobre el retorno no puede describirse adecuadamente mediante una relación lineal simple.

En el presente proyecto, los modelos Prophet, NeuralProphet y GAM se utilizan como alternativas aditivas para el pronóstico del retorno logarítmico de BTC-USD. En el enfoque multivariado, estos modelos incorporan regresores asociados al volumen, la volatilidad y los indicadores técnicos, permitiendo evaluar si dichas variables aportan información predictiva adicional frente a modelos univariados y enfoques clásicos.

5.12. Validación temporal y rolling-origin

La validación temporal es un componente fundamental en el pronóstico de series de tiempo, ya que permite evaluar el desempeño de los modelos respetando el orden cronológico de las

observaciones. A diferencia de los esquemas tradicionales de partición aleatoria, en series financieras no es adecuado mezclar datos pasados y futuros, debido a que esto puede generar fuga de información y sobreestimar artificialmente la capacidad predictiva del modelo.

En este proyecto se utilizan esquemas de validación fuera de muestra, principalmente hold-out y rolling-origin. La validación hold-out separa un bloque final de observaciones para evaluar los modelos después de entrenarlos con la información histórica disponible. Este procedimiento permite simular un escenario real de pronóstico, donde el modelo se ajusta con datos pasados y se evalúa sobre observaciones futuras.

Por su parte, la validación rolling-origin consiste en realizar múltiples evaluaciones sucesivas desplazando el origen del pronóstico a lo largo del tiempo. En cada corte temporal, el modelo se entrena con la información disponible hasta ese momento y genera pronósticos para un horizonte definido. Este enfoque permite evaluar no solo el error promedio del modelo, sino también su estabilidad temporal y capacidad de generalización en distintos momentos del periodo de validación (Hyndman & Athanasopoulos, 2021).

En el presente estudio, la validación rolling-origin se aplica considerando horizontes de uno y siete días, coherentes con el objetivo de evaluar el pronóstico del retorno logarítmico de BTC-USD en el corto plazo. Este esquema permite comparar modelos univariados y multivariados bajo condiciones más exigentes que una única partición hold-out, reduciendo el riesgo de obtener conclusiones dependientes de un solo periodo de validación.

La combinación de hold-out y rolling-origin permite obtener una evaluación más robusta del desempeño predictivo. Mientras el hold-out ofrece una primera comparación fuera de muestra, el rolling-origin permite analizar la consistencia de los modelos en el tiempo. De esta manera, el proyecto evalúa tanto la precisión en magnitud como la capacidad direccional de los modelos,

garantizando una comparación metodológicamente coherente con la naturaleza secuencial de los datos financieros.

5.13. Métricas de evaluación y prueba Diebold-Mariano

La evaluación del desempeño predictivo requiere utilizar métricas que permitan comparar los valores observados con los pronosticados. En este proyecto se consideran dos tipos de métricas: métricas de error, orientadas a medir la precisión en magnitud, y métricas direccionales, enfocadas en evaluar la capacidad del modelo para anticipar si el retorno será positivo o negativo.

Dentro de las métricas de error se emplean el MAE, RMSE, sMAPE y MASE. el MAE mide el error absoluto promedio, mientras que el RMSE penaliza con mayor intensidad los errores grandes. El sMAPE permite expresar el error en términos porcentuales simétricos, aunque en series de retornos cercanos a cero debe interpretarse con cautela. Por su parte, el MASE compara el error del modelo frente a un modelo ingenuo de referencia, lo cual facilita la comparación entre distintos enfoques de pronóstico (Hyndman & Koehler, 2006; Hyndman & Athanasopoulos, 2021).

Además de la precisión numérica, en series financieras es relevante evaluar la capacidad direccional del modelo. Para ello se utilizan métricas como Accuracy, Precision, Recall, F1-score, MCC y AUC. Estas métricas permiten analizar si el modelo logra anticipar correctamente la dirección del movimiento del retorno logarítmico, aspecto especialmente importante cuando el interés no se limita a estimar la magnitud exacta del cambio, sino también a identificar señales de subida o caída.

La comparación entre modelos se complementa mediante la prueba de Diebold-Mariano, la cual permite evaluar si las diferencias observadas en los errores de pronóstico entre dos modelos son estadísticamente significativas. Esta prueba plantea como hipótesis nula la igualdad de

capacidad predictiva entre dos modelos competidores. Si el valor-p es inferior al nivel de significancia definido, se rechaza la hipótesis nula y se concluye que existen diferencias estadísticamente significativas en el desempeño predictivo (Diebold & Mariano, 1995).

En el presente proyecto, estas métricas y pruebas se aplican para comparar modelos univariados y multivariados en el pronóstico del retorno logarítmico de BTC-USD. De esta forma, la evaluación no se limita a identificar el modelo con menor error promedio, sino que también considera su desempeño direccional, estabilidad temporal y evidencia estadística de diferencias predictivas frente a modelos de referencia.

6. Metodología

6.1. Tipo de investigación

El presente estudio corresponde a una investigación aplicada, cuantitativa, predictiva, no experimental y longitudinal. Es aplicada porque utiliza métodos estadísticos, econométricos y computacionales para abordar un problema concreto de pronóstico financiero: la predicción del retorno logarítmico de BTC-USD.

El enfoque es cuantitativo, dado que se basa en datos numéricos diarios del mercado de Bitcoin, incluyendo precios, volumen e indicadores técnicos. Asimismo, es predictivo porque su propósito principal no es establecer relaciones causales, sino evaluar la capacidad de distintos modelos para anticipar el comportamiento futuro del retorno logarítmico.

El diseño es no experimental, ya que se emplean datos históricos observados sin manipulación directa de variables. También es longitudinal, porque analiza la evolución temporal del mercado de Bitcoin durante el periodo de estudio. Metodológicamente, los modelos se entrenan, validan y comparan bajo esquemas de validación temporal, como hold-out y rolling-origin,

respetando el orden cronológico de la serie y evitando fuga de información (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025).

6.2. Fuente de datos

La fuente de datos utilizada en el presente estudio corresponde a la serie histórica diaria de Bitcoin frente al dólar estadounidense, identificada con el ticker BTC-USD. La información fue obtenida desde Yahoo Finance mediante la librería `yfinance` de Python, la cual permite descargar datos financieros históricos de forma reproducible.

La base de datos incluye información diaria en formato OHLCV, es decir: precio de apertura (Open), precio máximo (High), precio mínimo (Low), precio de cierre (Close) y volumen de negociación (Volume). A partir de estas variables se construyó la variable objetivo del estudio, correspondiente al retorno logarítmico diario de BTC-USD, así como el conjunto de variables exógenas asociadas al volumen, la volatilidad y los indicadores técnicos.

El periodo de análisis comprende observaciones diarias desde el 1 de enero de 2017 hasta el 30 de diciembre de 2025. Esta ventana temporal permite capturar diferentes fases del mercado de Bitcoin, incluyendo periodos de expansión, corrección, alta volatilidad y cambios en la actividad de negociación.

La información descargada fue depurada, ordenada cronológicamente y ajustada a frecuencia diaria. Posteriormente, se verificó la consistencia temporal de la serie, la ausencia de valores faltantes relevantes y la correcta alineación entre la variable objetivo y las variables explicativas. Este procedimiento garantiza que los modelos sean entrenados y evaluados bajo una estructura temporal coherente, evitando la fuga de información y asegurando la reproducibilidad del análisis.

6.3. Población y muestra

La población del estudio corresponde al histórico diario disponible de Bitcoin frente al dólar estadounidense, identificado como BTC-USD, incluyendo información de precios y volumen de negociación. Para el desarrollo empírico se emplea una muestra de datos diarios obtenida desde Yahoo Finance, comprendida entre el 1 de enero de 2017 y el 30 de diciembre de 2025.

La muestra incluye variables financieras en formato OHLCV: precio de apertura, máximo, mínimo, cierre y volumen. A partir de esta información se construye la variable objetivo del proyecto, correspondiente al retorno logarítmico diario de BTC-USD, así como las variables explicativas asociadas al volumen, la volatilidad y los indicadores técnicos.

La división de la muestra se realiza respetando el orden temporal de las observaciones. El conjunto de entrenamiento comprende las observaciones anteriores al 1 de noviembre de 2025, mientras que el conjunto de validación corresponde al periodo comprendido entre el 1 de noviembre de 2025 y el 30 de diciembre de 2025, con un total de 60 observaciones diarias fuera de muestra.

Esta separación temporal permite evaluar los modelos en condiciones similares a un escenario real de pronóstico, evitando la fuga de información y garantizando que las predicciones se realicen únicamente con información disponible hasta el momento de estimación.

6.4. Variable dependiente

La variable dependiente del presente estudio es el retorno logarítmico diario de Bitcoin frente al dólar estadounidense (BTC-USD). Esta variable se construye a partir del precio de cierre diario del activo, debido a que los precios en niveles suelen presentar tendencia, cambios de escala

y comportamiento no estacionario. En el análisis de series financieras, el uso de retornos logarítmicos es habitual porque permite trabajar con variaciones relativas del precio y facilita la modelación estadística de la serie (Shumway & Stoffer, 2025; Hyndman & Athanasopoulos, 2021). El retorno logarítmico se define como:

$$r_t = \log\left(\frac{C_t}{C_{t-1}}\right) \quad (\text{Ecuación 11})$$

En la Ecuación 11, r_t representa el retorno logarítmico en el día t , C_t corresponde al precio de cierre de BTC-USD en el día t , y C_{t-1} representa el precio de cierre del día anterior.

Por propiedad de los logaritmos, esta expresión también puede escribirse como:

$$r_t = \log(C_t) - \log(C_{t-1}) \quad (\text{Ecuación 12})$$

Por tanto, el retorno logarítmico no corresponde a una diferencia absoluta entre precios, sino a una variación relativa expresada en escala logarítmica. Esta transformación permite comparar cambios proporcionales del precio y trabajar con una serie más adecuada para el modelado estadístico y predictivo. En activos altamente volátiles como Bitcoin, esta transformación resulta especialmente pertinente, ya que los retornos permiten analizar de mejor manera la variabilidad del mercado, aunque pueden seguir presentando colas pesadas, valores extremos y heterocedasticidad (Bouri, Gil-Alana, Gupta, & Roubaud, 2020).

En el proyecto, el retorno logarítmico se utiliza como variable objetivo tanto en el enfoque univariado como en el multivariado. En el enfoque univariado, los modelos utilizan únicamente la historia del retorno; mientras que, en el enfoque multivariado, se incorporan variables adicionales asociadas al volumen, la volatilidad y los indicadores técnicos.

Además de predecir la magnitud del retorno, el estudio evalúa la dirección del movimiento, clasificando los retornos como positivos o negativos. Esto permite analizar no solo la precisión

numérica de los modelos, sino también su capacidad para anticipar la dirección futura del retorno de Bitcoin.

6.5. Variables independientes

Las variables independientes del estudio corresponden al conjunto de predictores utilizados para explicar y pronosticar el retorno logarítmico diario de BTC-USD. Estas variables se construyen a partir de la información histórica de precios, volumen, volatilidad e indicadores técnicos derivados de la estructura OHLCV del activo. Su incorporación se fundamenta en la evidencia de que, en mercados de criptomonedas, variables asociadas al volumen, la liquidez, la volatilidad y el momentum pueden contener información útil para el pronóstico de retornos (Vidal-Tomás & Ibáñez, 2023; Anastasiou, Ballis, & Drakos, 2023; Guo, Sang, Tu, & Wang, 2024).

En el enfoque univariado, las variables independientes están representadas por los rezagos del propio retorno logarítmico. Es decir, los modelos utilizan información histórica de la variable objetivo para generar el pronóstico futuro. Esta estructura permite evaluar la capacidad predictiva contenida únicamente en la dinámica pasada del retorno, siguiendo el enfoque tradicional de modelado de series de tiempo financieras (Shumway & Stoffer, 2025; Hyndman & Athanasopoulos, 2021).

En el enfoque multivariado, además de los rezagos del retorno logarítmico, se incorporan variables exógenas asociadas al comportamiento del mercado. Entre ellas se incluyen transformaciones del volumen, como `log_volume`, `volume_ratio` y `volume_zscore`; medidas de volatilidad histórica, como `rolling_vol_7`, `rolling_vol_30` y `ATR_14`; indicadores de momentum y tendencia, como `RSI_14`, `MACD` y `MACD_signal`; e indicadores de flujo de mercado, como `OBV` y `CMF_20`. Estas variables permiten capturar dimensiones complementarias del mercado de

Bitcoin, tales como intensidad de negociación, liquidez, volatilidad reciente, presión compradora o vendedora y cambios en el momentum del precio.

Desde el punto de vista metodológico, la incorporación de múltiples predictores exige controlar problemas como la multicolinealidad, la sobreparametrización y el sobreajuste. Por esta razón, los modelos se evalúan mediante esquemas de validación temporal y, en el caso de algoritmos de machine learning y deep learning, se aplican estrategias de regularización, escalamiento y control de complejidad orientadas a mejorar la capacidad de generalización (Hastie, Tibshirani, & Friedman, 2017; Chen & Guestrin, 2016).

Todas las variables independientes se construyen respetando la secuencia temporal de los datos, evitando el uso de información futura en el proceso de modelado. De esta manera, se garantiza la coherencia metodológica del enfoque predictivo y se reduce el riesgo de fuga de información durante la validación de los modelos.

6.6. Construcción del feature set

La construcción del feature set corresponde al proceso mediante el cual los datos financieros originales de BTC-USD se transforman en un conjunto estructurado de variables para el modelado predictivo. A partir de la información diaria OHLCV —apertura, máximo, mínimo, cierre y volumen— se construyeron variables asociadas al retorno logarítmico, volumen de negociación, volatilidad histórica, momentum, tendencia y flujo de mercado.

En primer lugar, se construyó la variable objetivo del estudio, correspondiente al retorno logarítmico diario de BTC-USD. Esta transformación permite analizar variaciones relativas del precio y reduce el efecto de los cambios de nivel presentes en la serie de precios. El retorno

logarítmico diario fue definido previamente en la Ecuación 11 y expresado de forma equivalente en la Ecuación 12.

Posteriormente, se generaron variables derivadas del volumen, tales como `log_volume`, `volume_sma_20`, `volume_ratio` y `volume_zscore`, con el propósito de capturar cambios en la intensidad de negociación, liquidez relativa y episodios de actividad anormal del mercado.

La inclusión de variables de volumen se justifica por la evidencia de que la relación entre precio, volumen y retornos puede variar en el tiempo y aportar información relevante en mercados de criptomonedas (Vidal-Tomás & Ibáñez, 2023; Guo, Sang, Tu, & Wang, 2024).

Adicionalmente, se incorporaron medidas de volatilidad histórica, como `rolling_vol_7` y `rolling_vol_30`, junto con el indicador `ATR_14`, para representar la intensidad de las fluctuaciones del mercado en ventanas de corto y mediano plazo. Estas variables permiten aproximar cambios en el nivel de riesgo y en la amplitud de los movimientos del activo, aspectos especialmente relevantes en series financieras altamente volátiles como Bitcoin (Bouri, Gil-Alana, Gupta, & Roubaud, 2020).

También se construyeron indicadores técnicos de momentum, tendencia y flujo de mercado, entre ellos `RSI_14`, `MACD`, `MACD_signal`, `MACD_hist`, `OBV` y `CMF_20`. Estos indicadores se utilizaron como variables explicativas dentro del enfoque multivariado, no como reglas de negociación, sino como transformaciones informativas del precio y del volumen que pueden capturar patrones no lineales o relaciones dinámicas relevantes para el pronóstico (Anastasiou, Ballis, & Drakos, 2023; Lahmiri & Bekiros, 2023).

Con el fin de sintetizar la estructura del feature set utilizado en el enfoque multivariado, la Tabla 1 presenta los principales bloques de variables incorporados en la matriz de predictores. Esta tabla permite distinguir entre los rezagos del retorno, las variables de volumen, las medidas de

volatilidad, los indicadores de momentum y tendencia, las variables de flujo de mercado y la variable objetivo. De esta manera, se evidencia que el modelo multivariado no se limita a la información histórica del retorno, sino que integra señales complementarias del comportamiento financiero de BTC-USD.

Tabla 1

Estructura de la matriz multivariada de predictores

Bloque de variables	Variables incluidas	Propósito dentro del modelo
Rezagos del retorno	lag_1, lag_2, ..., lag_28	Capturan la memoria temporal del retorno logarítmico y permiten incorporar dependencia autorregresiva.
Volumen	log_volume, volume_ratio, volume_zscore, volume_sma_20	Representan la intensidad de negociación, el volumen relativo y desviaciones anormales frente al comportamiento reciente.
Volatilidad	rolling_vol_7, rolling_vol_30, ATR_14	Capturan variabilidad reciente, amplitud intradía y condiciones de riesgo del mercado.
Momentum / tendencia	RSI_14, MACD, MACD_signal, MACD_hist	Representan fuerza relativa, señales de sobrecompra/sobreventa y cambios en la tendencia de corto plazo.
Flujo de mercado	OBV, CMF_20	Aproximan acumulación/distribución, presión compradora o vendedora y flujo de capital asociado al precio y volumen.
Variable objetivo	log_return	Corresponde al retorno logarítmico diario de BTC-USD utilizado como variable dependiente del pronóstico.

Nota: La tabla resume la estructura conceptual de la matriz multivariada usada para construir los predictores de los modelos.

A partir de estos bloques de variables se construyó una matriz multivariada de predictores, denotada como (X_t) compuesta por rezagos del retorno logarítmico y variables exógenas observadas hasta el instante (t). De forma general, dicha matriz puede representarse como:

$$X_t = [r_{t-1}, r_{t-2}, \dots, r_{t-p}, Z_{1,t}, Z_{1,t}, \dots, Z_{k,t}] \quad (\text{Ecuación 13})$$

En la Ecuación 13,

$$r_{t-1}, r_{t-2}, \dots, r_{t-p}$$

representan los rezagos del retorno logarítmico, mientras que:

$$Z_{1,t}, Z_{1,t}, \dots, Z_{k,t}$$

corresponden a las variables exógenas de volumen, volatilidad, momentum, tendencia y flujo de mercado.

En este estudio, los rezagos del retorno se incorporaron hasta ($p = 28$), y las variables exógenas incluyeron indicadores como log_volume, volume_ratio, volume_zscore, rolling_vol_7, rolling_vol_30, ATR_14, RSI_14, MACD, MACD_signal, MACD_hist, OBV y CMF_20.

La variable objetivo se definió de acuerdo con el horizonte de pronóstico considerado. Para un horizonte (h), el problema predictivo puede expresarse como:

$$y_{t+h} = r_{t+h} \quad (\text{Ecuación 14})$$

de modo que los modelos buscan aproximar el retorno logarítmico futuro de BTC-USD a partir de la información disponible hasta el instante (t):

$$\hat{y}_{t+h} = f(X_t) \quad (\text{Ecuación 15})$$

En la Ecuación 15, $\hat{y}_{t+h}$ corresponde al retorno pronosticado y $f(X_t)$ representa la función estimada por cada modelo. En el enfoque univariado, (X_t) contiene únicamente rezagos del retorno

logarítmico; en el enfoque multivariado, (X_t) incorpora además las variables exógenas descritas anteriormente.

La matriz multivariada de predictores quedó conformada por 3256 observaciones y 41 variables explicativas, integradas por 28 rezagos del retorno logarítmico y 13 variables exógenas asociadas con volumen, volatilidad, momentum, tendencia y flujo de mercado. Al incluir la variable objetivo `log_return`, la matriz completa de modelado contiene 42 columnas. El número final de observaciones resulta de eliminar los registros iniciales sin información completa, generados por el cálculo de retornos, rezagos y ventanas móviles.

Finalmente, el feature set fue construido respetando estrictamente la secuencia temporal de los datos. Las variables basadas en medias móviles, desviaciones estándar, indicadores técnicos y rezagos se calcularon únicamente con información disponible hasta cada instante temporal, con el fin de evitar fuga de información. Este control metodológico es fundamental en series de tiempo, ya que garantiza que los modelos sean evaluados bajo condiciones realistas de pronóstico fuera de muestra (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025).

6.7. División entrenamiento-validación

La división entrenamiento-validación se realizó respetando estrictamente el orden cronológico de la serie, debido a que en problemas de series de tiempo no es adecuado realizar particiones aleatorias. En este tipo de datos, mezclar observaciones pasadas y futuras puede generar fuga de información y producir una sobreestimación del desempeño predictivo de los modelos (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025).

En el presente proyecto, el conjunto de entrenamiento está conformado por las observaciones históricas anteriores al 1 de noviembre de 2025, mientras que el conjunto de

validación corresponde al periodo comprendido entre el 1 de noviembre de 2025 y el 30 de diciembre de 2025. Esta última ventana contiene 60 observaciones diarias, utilizadas para evaluar el desempeño fuera de muestra de los modelos.

La separación temporal se aplicó tanto en el enfoque univariado como en el multivariado. En el enfoque univariado, la división se realizó sobre la serie de retornos logarítmicos de BTC-USD. En el enfoque multivariado, la misma separación se aplicó sobre la variable objetivo y sobre el conjunto de variables exógenas, garantizando que los predictores estuvieran correctamente alineados con la información disponible en cada fecha.

Este procedimiento permite simular un escenario real de pronóstico financiero, en el cual los modelos son entrenados únicamente con información histórica y posteriormente evaluados sobre datos futuros no utilizados durante el ajuste. De esta forma, la división entrenamiento-validación contribuye a una comparación más objetiva entre modelos univariados y multivariados, reduciendo el riesgo de sobreajuste y fortaleciendo la validez empírica de los resultados.

6.8. Modelos implementados

En el presente proyecto se implementaron y compararon modelos univariados y multivariados para el pronóstico del retorno logarítmico de BTC-USD. La inclusión de distintas familias metodológicas permite evaluar si enfoques más complejos aportan mejoras predictivas frente a modelos simples de referencia, siguiendo el principio de comparación objetiva en ejercicios de pronóstico (Hyndman & Athanasopoulos, 2021; Hyndman & Koehler, 2006).

En primer lugar, se implementaron modelos de referencia, como Naive y SNaive(7), los cuales permiten establecer una línea base mínima de comparación. Posteriormente, se evaluaron modelos clásicos de series de tiempo, entre ellos AutoETS, SARIMA, AutoARIMA y Theta en el

enfoque univariado, así como ARIMAX, AutoARIMAX, SARIMAX y un modelo estructural tipo ETSX con variables exógenas en el enfoque multivariado.

También se incorporaron modelos de machine learning, incluyendo MLP, Random Forest y XGBoost, debido a su capacidad para capturar relaciones no lineales e interacciones entre predictores (Hastie, Tibshirani, & Friedman, 2017; Chen & Guestrin, 2016). En el caso del deep learning, se implementaron arquitecturas secuenciales LSTM y GRU, orientadas al aprendizaje de patrones temporales mediante ventanas de observaciones pasadas (Hochreiter & Schmidhuber, 1997; Lahmiri & Bekiros, 2020).

Adicionalmente, se evaluaron modelos híbridos, como AutoARIMA + MLP y AutoARIMA + XGBoost, en los cuales se combina una estructura clásica con modelos de aprendizaje sobre residuos. Finalmente, se incluyeron modelos aditivos, como Prophet, NeuralProphet y GAM, que permiten representar componentes de tendencia, estacionalidad, relaciones no lineales y regresores externos de forma interpretable (Taylor & Letham, 2017; Triebe et al., 2021).

La implementación de estas familias de modelos permitió comparar el desempeño entre enfoques univariados, basados únicamente en la historia del retorno, y enfoques multivariados, que incorporan variables asociadas al volumen, la volatilidad y los indicadores técnicos. Esta comparación constituye la base empírica para evaluar el aporte predictivo de la información exógena en el pronóstico del retorno logarítmico de Bitcoin.

6.9. Esquemas de validación

La evaluación de los modelos se realizó mediante esquemas de validación temporal, respetando el orden cronológico de la serie. En problemas de series de tiempo financieras no es adecuado aplicar particiones aleatorias, ya que esto puede mezclar información pasada y futura,

generar fuga de información y sobreestimar el desempeño predictivo de los modelos (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025).

En primer lugar, se utilizó un esquema hold-out, en el cual los modelos fueron entrenados con la información histórica disponible antes del periodo de validación y evaluados sobre una ventana final de 60 días. Este procedimiento permitió obtener una primera comparación fuera de muestra entre modelos univariados y multivariados.

Posteriormente, se aplicó validación rolling-origin, también conocida como walk-forward validation. En este esquema, el origen del pronóstico se desplaza progresivamente a lo largo del periodo de validación, reentrenando o actualizando los modelos con la información disponible hasta cada punto temporal. Esto permite evaluar la estabilidad del desempeño predictivo en diferentes cortes de tiempo y simular un proceso real de pronóstico financiero.

La validación rolling-origin se aplicó considerando horizontes de pronóstico de uno y siete días, coherentes con el objetivo del estudio. Para el horizonte $h=1$, se evaluó la capacidad de los modelos para anticipar el retorno diario siguiente; mientras que para $h=7$, se analizó su desempeño en un horizonte semanal de corto plazo.

El uso conjunto de hold-out y rolling-origin permitió comparar los modelos desde dos perspectivas complementarias: el desempeño en una ventana fija de validación y la estabilidad temporal bajo múltiples cortes de evaluación. De esta manera, se fortaleció la validez empírica de los resultados y se redujo el riesgo de obtener conclusiones dependientes de un único periodo de prueba.

6.10. Métricas de evaluación

La evaluación del desempeño predictivo se realizó mediante métricas de error y métricas direccionales, con el fin de analizar tanto la precisión numérica de los pronósticos como la capacidad de los modelos para anticipar la dirección del movimiento del retorno logarítmico de BTC-USD.

Para medir la precisión en magnitud se utilizaron las métricas MAE, RMSE, sMAPE y MASE. El MAE cuantifica el error absoluto promedio entre los valores reales y pronosticados, mientras que el RMSE penaliza con mayor fuerza los errores grandes. El sMAPE permite expresar el error relativo de forma simétrica, aunque su interpretación debe realizarse con cautela en series de retornos cercanos a cero. Por su parte, el MASE compara el desempeño del modelo frente a un modelo ingenuo de referencia, lo cual permite evaluar si el modelo supera una línea base simple de pronóstico (Hyndman & Koehler, 2006; Hyndman & Athanasopoulos, 2021).

Adicionalmente, se incorporaron métricas direccionales como Accuracy, Precision, Recall, F1-score, MCC y AUC. Estas métricas permiten evaluar si el modelo logra anticipar correctamente si el retorno será positivo o negativo, aspecto relevante en aplicaciones financieras donde la dirección del movimiento puede ser tan importante como la magnitud del error.

La comparación entre modelos se complementó con la prueba de Diebold-Mariano, utilizada para contrastar si las diferencias observadas en la precisión predictiva entre dos modelos son estadísticamente significativas. Esta prueba permite evaluar la hipótesis de igualdad de capacidad predictiva entre modelos competidores y aporta un criterio estadístico adicional para seleccionar los mejores enfoques de pronóstico (Diebold & Mariano, 1995).

En conjunto, estas métricas permiten realizar una evaluación integral del desempeño de los modelos, considerando error promedio, sensibilidad frente a errores extremos, capacidad

direccional y evidencia estadística de diferencias predictivas frente al modelo de referencia. De esta manera, la selección de modelos no se basa únicamente en una métrica aislada, sino en una comparación más robusta y coherente con el objetivo del proyecto.

6.11. Herramientas computacionales

El desarrollo del proyecto se realizó mediante un entorno computacional reproducible en Python, utilizando notebooks como estructura principal para la descarga, procesamiento, modelado, validación y comparación de resultados. Esta elección permitió documentar cada etapa del análisis, facilitar la trazabilidad metodológica y conservar evidencia de los procedimientos aplicados durante la construcción de variables, estimación de modelos y generación de resultados.

En la implementación de modelos clásicos y estadísticos se emplearon librerías como statsmodels, pmdarima y statsforecast, utilizadas para modelos ARIMA, SARIMA, AutoARIMA, ETS, Theta, SARIMAX y modelos estructurales de espacio de estados. Para los modelos de machine learning se utilizaron scikit-learn y xgboost, incluyendo MLP, Random Forest y XGBoost. En deep learning, se empleó TensorFlow/Keras para arquitecturas LSTM y GRU. Adicionalmente, se utilizaron Prophet, NeuralProphet y pyGAM para modelos aditivos.

La reproducibilidad del análisis se fortaleció mediante semillas aleatorias, control de advertencias, organización de rutas de trabajo y almacenamiento de resultados intermedios. Asimismo, los procedimientos de escalamiento, construcción de rezagos, generación de variables exógenas y validación temporal fueron implementados de forma sistemática para garantizar coherencia entre los modelos evaluados.

De manera complementaria, durante la elaboración del documento se utilizó ChatGPT como herramienta auxiliar para la revisión de redacción académica, organización de apartados,

depuración de estilo, estructuración de tablas y mejora de la claridad expositiva. Su uso no sustituyó el análisis estadístico, la ejecución de los modelos ni la interpretación técnica de los resultados, desarrollados a partir de los notebooks, bases de datos y procedimientos metodológicos propios. La responsabilidad sobre el contenido final, la verificación de resultados, las citas bibliográficas y las conclusiones corresponde exclusivamente al autor (OpenAI, 2026).

El uso de un pipeline reproducible resulta fundamental en estudios predictivos, ya que permite replicar resultados, comparar modelos bajo condiciones homogéneas y reducir errores asociados a la manipulación manual de datos. Este principio es consistente con las buenas prácticas de evaluación y comparación de modelos de pronóstico en series de tiempo financieras (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025; Makridakis, Spiliotis, & Assimakopoulos, 2022).

7. Análisis exploratorio de datos

El análisis exploratorio de datos tiene como propósito caracterizar el comportamiento histórico de BTC-USD antes de la etapa de modelado predictivo. Esta fase permite identificar patrones relevantes en la serie, revisar la evolución del precio, analizar la distribución de los retornos logarítmicos, explorar la dinámica del volumen y evaluar posibles relaciones entre las variables exógenas construidas.

En el contexto de series financieras, esta etapa es fundamental porque permite reconocer características como tendencia, alta volatilidad, valores extremos, heterocedasticidad, baja autocorrelación lineal y posibles cambios de régimen, aspectos frecuentes en mercados de criptomonedas (Shumway & Stoffer, 2025; Hyndman & Athanasopoulos, 2021; Bouri, Gil-Alana, Gupta, & Roubaud, 2020). En este proyecto, el análisis exploratorio se desarrolla tanto desde una

perspectiva univariada como multivariada, con el fin de justificar la posterior comparación entre modelos basados únicamente en la historia del retorno y modelos que incorporan volumen, volatilidad e indicadores técnicos.

El análisis incluye la revisión de la evolución del precio de Bitcoin, la construcción y distribución de los retornos logarítmicos, la identificación de periodos de alta volatilidad, el comportamiento del volumen de negociación, el análisis de indicadores técnicos, la correlación entre variables, la multicolinealidad y las pruebas de causalidad temporal de Granger. Estos elementos permiten establecer una base empírica para la selección de variables y para la interpretación de los resultados predictivos.

7.1. Evolución del precio de Bitcoin

La evolución histórica del precio de cierre de Bitcoin frente al dólar estadounidense durante el periodo 2017–2025 evidencia un comportamiento altamente dinámico, caracterizado por fases de crecimiento acelerado, correcciones pronunciadas y episodios recurrentes de alta volatilidad. Esta dinámica refleja cambios significativos en el nivel del precio y confirma que la serie en niveles presenta una estructura no estacionaria, por lo cual resulta necesario transformarla antes de su modelado estadístico.

Figura 1

Serie temporal del precio de cierre de BTC-USD



Nota: elaboración propia con datos de Yahoo Finance procesados en Python.

Como se observa en la Figura 1, el precio de BTC-USD presenta cambios significativos de nivel a lo largo del periodo analizado. La serie no mantiene una media estable y evidencia fuertes variaciones en escala, especialmente durante los periodos de expansión y corrección del mercado. Este comportamiento confirma que la serie en precios no resulta adecuada para ser modelada directamente sin una transformación previa.

Por esta razón, en el presente proyecto se utiliza el retorno logarítmico como variable objetivo, ya que permite analizar las variaciones relativas del activo en lugar de trabajar con precios absolutos. Esta transformación facilita el análisis estadístico y permite comparar modelos univariados y multivariados sobre una serie más apropiada para el pronóstico financiero.

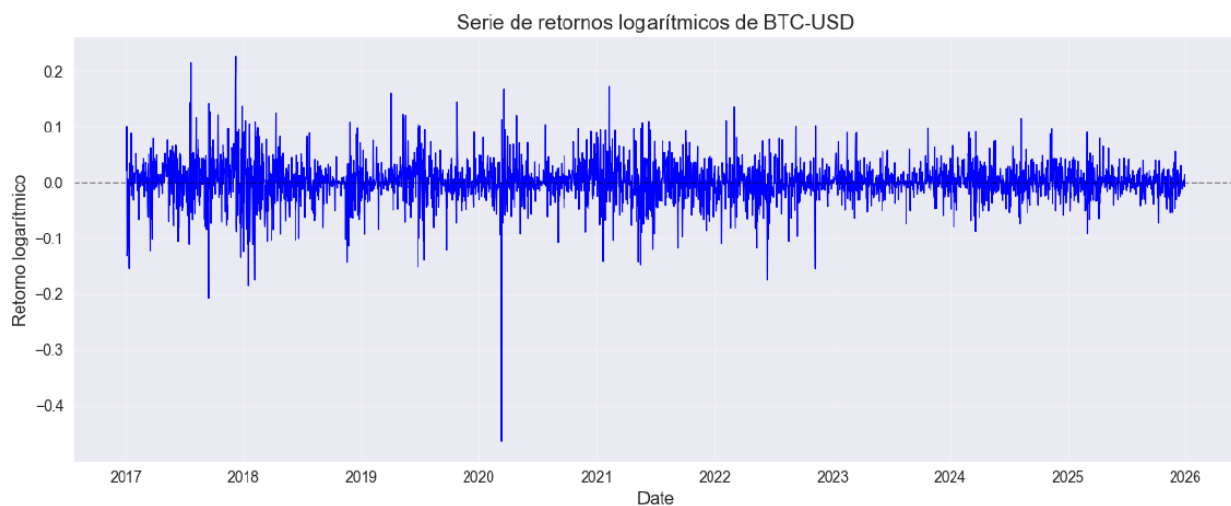
Desde el punto de vista financiero, los cambios abruptos observados en la evolución del precio reflejan la sensibilidad de Bitcoin frente a condiciones de mercado, expectativas de los inversionistas, episodios especulativos, liquidez y actividad transaccional. Esta característica respalda la necesidad de evaluar modelos flexibles y enfoques multivariados que incorporen información adicional al precio histórico, especialmente variables asociadas al volumen, la volatilidad y los indicadores técnicos.

7.2. Distribución de retornos

Los retornos logarítmicos diarios de BTC-USD permiten analizar las variaciones relativas del precio de Bitcoin durante el periodo 2017–2025. A diferencia de la serie de precios en niveles, los retornos oscilan alrededor de cero, lo cual indica que la mayoría de las variaciones diarias corresponden a cambios proporcionales relativamente pequeños.

Figura 2

Serie temporal de retornos logarítmicos de BTC-USD



Nota: elaboración propia con datos de Yahoo Finance procesados en Python.

Como se observa en la Figura 2, la serie de retornos logarítmicos presenta episodios de alta volatilidad agrupada, especialmente en periodos donde se evidencian movimientos abruptos positivos y negativos. Este comportamiento es característico de las series financieras, en las cuales los periodos de alta variabilidad tienden a concentrarse en determinados intervalos de tiempo.

También se identifican valores extremos, destacándose caídas pronunciadas en algunos momentos del periodo analizado. Estos eventos reflejan la sensibilidad del mercado de Bitcoin frente a choques de liquidez, cambios en expectativas, episodios especulativos y condiciones generales del mercado. La presencia de estos movimientos confirma que, aunque los retornos logarítmicos son más adecuados para el modelado que los precios en niveles, la serie conserva

propiedades propias de activos financieros altamente volátiles, como colas pesadas, heterocedasticidad y cambios abruptos de magnitud (Bouri, Gil-Alana, Gupta, & Roubaud, 2020).

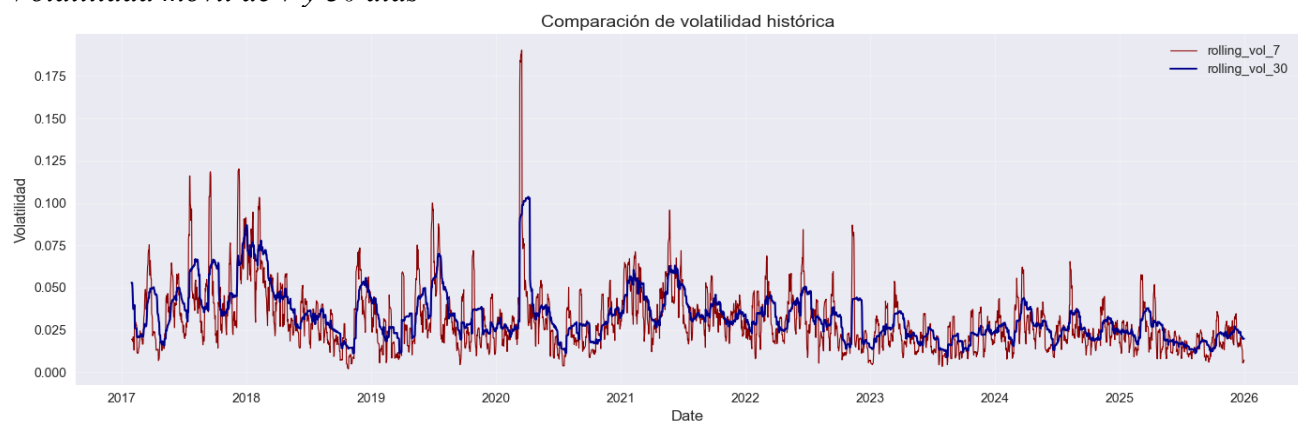
Desde el punto de vista metodológico, esta evidencia justifica la comparación entre modelos clásicos, modelos de machine learning, deep learning y modelos aditivos. La baja estabilidad de la serie y la presencia de episodios extremos hacen necesario evaluar enfoques capaces de capturar patrones lineales, no lineales y posibles efectos asociados al volumen, la volatilidad y los indicadores técnicos.

7.3. Volatilidad histórica

La volatilidad histórica permite analizar la intensidad de las fluctuaciones del retorno logarítmico de BTC-USD a lo largo del tiempo. En este proyecto se consideran dos medidas móviles: la volatilidad de corto plazo, calculada con una ventana de 7 días, y la volatilidad de mediano plazo, calculada con una ventana de 30 días. Estas variables permiten identificar periodos de mayor inestabilidad del mercado y sirven como predictores dentro del enfoque multivariado.

Figura 3

Volatilidad móvil de 7 y 30 días



Nota: elaboración propia con datos de Yahoo Finance procesados en Python.

Como se observa en la Figura 3, la volatilidad de Bitcoin no permanece constante durante el periodo analizado. Se evidencian episodios de alta variabilidad, especialmente entre 2017 y

2018, durante el choque de 2020 y en algunos periodos posteriores de corrección o incertidumbre del mercado. Este comportamiento confirma la presencia de agrupamientos de volatilidad, característica frecuente en series financieras y mercados de criptomonedas (Bouri, Gil-Alana, Gupta, & Roubaud, 2020).

La volatilidad móvil de 7 días presenta movimientos más sensibles y abruptos, por lo que captura cambios rápidos en la dinámica del mercado. En contraste, la volatilidad móvil de 30 días muestra una trayectoria más suavizada, reflejando la persistencia de la variabilidad en ventanas de mediano plazo. Esta diferencia permite complementar el análisis, ya que ambas medidas capturan distintos horizontes de riesgo.

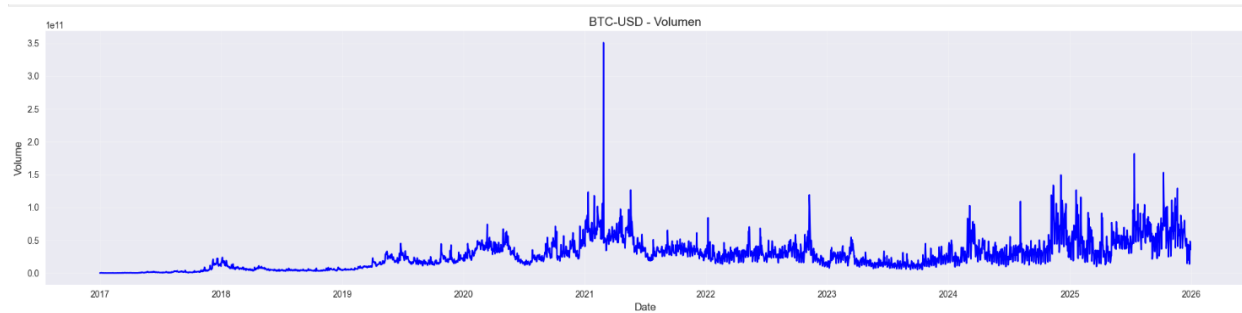
Desde el punto de vista metodológico, la inclusión de `rolling_vol_7` y `rolling_vol_30` en el feature set multivariado permite evaluar si la volatilidad reciente aporta información útil para anticipar el retorno logarítmico de BTC-USD. Su incorporación resulta relevante porque los periodos de alta volatilidad pueden modificar el desempeño de los modelos y afectar tanto la magnitud como la dirección de los pronósticos.

7.4. Variables de volumen

El volumen de negociación constituye una variable central dentro del análisis multivariado, ya que permite aproximar la intensidad de participación del mercado, los cambios en la liquidez y los episodios de mayor actividad transaccional. En mercados como Bitcoin, el volumen puede reflejar aumentos en el interés de los inversionistas, presión compradora o vendedora y posibles cambios en la dinámica del precio (Vidal-Tomás & Ibáñez, 2023; Guo, Sang, Tu, & Wang, 2024).

Figura 4

Evolución del volumen de negociación de BTC-USD



Nota: elaboración propia con datos de Yahoo Finance procesados en Python.

Como se observa en la Figura 4, el volumen de negociación de BTC-USD presenta un comportamiento altamente variable durante el periodo 2017–2025. Se identifican periodos de baja actividad relativa y episodios de fuerte incremento en el volumen, especialmente alrededor de 2021 y en algunos tramos posteriores. Estos picos reflejan momentos de mayor participación del mercado y pueden estar asociados con fases de alta volatilidad, cambios abruptos en el precio o eventos de mayor incertidumbre.

Desde el punto de vista metodológico, esta evidencia justifica la incorporación del volumen como variable explicativa dentro del enfoque multivariado. Sin embargo, debido a que el volumen en niveles puede presentar alta dispersión y valores extremos, en el proyecto se construyeron transformaciones adicionales como `log_volume`, `volume_sma_20`, `volume_ratio` y `volume_zscore`. Estas variables permiten representar la actividad transaccional de forma más estable y comparar el volumen observado frente a su comportamiento reciente.

En consecuencia, el análisis del volumen permite complementar la información contenida en los precios y retornos. Su inclusión en los modelos busca evaluar si los cambios en la intensidad de negociación aportan información predictiva adicional para anticipar el retorno logarítmico de BTC-USD, tanto en magnitud como en dirección.

7.5. Indicadores técnicos

Los indicadores técnicos permiten resumir información relevante del comportamiento histórico de BTC-USD a partir de variables de precio y volumen. En este proyecto, estos indicadores no se emplean como reglas automáticas de negociación, sino como variables predictoras dentro del enfoque multivariado, con el propósito de evaluar si aportan información adicional al pronóstico del retorno logarítmico de Bitcoin.

Tabla 2

Resumen de indicadores técnicos utilizados como variables predictoras

Indicador	Grupo	Interpretación
RSI_14	Momentum	Mide la intensidad relativa de los movimientos recientes del precio.
MACD	Momentum tendencia	/ Resume cambios en la fuerza y dirección de la tendencia.
MACD_signal	Momentum tendencia	/ Suaviza la señal del MACD y permite contrastar cambios de tendencia.
MACD_hist	Momentum tendencia	/ Representa la diferencia entre MACD y su señal, indicando aceleración o desaceleración del momentum.
ATR_14	Volatilidad	Mide la amplitud promedio de oscilación del precio.
OBV	Flujo de mercado	Relaciona volumen y dirección del precio para aproximar acumulación o distribución.
CMF_20	Flujo de mercado	Evalúa presión compradora o vendedora a partir del precio y volumen.

Como se observa en la Tabla 2, los indicadores seleccionados representan distintas dimensiones del mercado. El RSI_14 captura información asociada al momentum reciente del precio, mientras que el conjunto MACD, MACD_signal y MACD_hist permite aproximar cambios

en la fuerza, dirección y aceleración de la tendencia. Estos indicadores son útiles para representar posibles señales de continuidad o cambio en la dinámica reciente del activo.

Por otra parte, el ATR_14 se incorpora como una medida de volatilidad, ya que resume la amplitud promedio de oscilación del precio. Esta variable resulta relevante en el caso de Bitcoin, debido a que el activo presenta episodios frecuentes de alta variabilidad y movimientos abruptos. Finalmente, los indicadores OBV y CMF_20 integran información de volumen y precio, permitiendo aproximar condiciones de acumulación, distribución y presión compradora o vendedora.

Desde el punto de vista metodológico, estos indicadores permiten ampliar el conjunto de información disponible para los modelos multivariados. Su inclusión busca contrastar si variables asociadas al momentum, la volatilidad y el flujo de mercado mejoran la capacidad predictiva frente a modelos univariados basados únicamente en la historia del retorno. Esta decisión es consistente con la literatura reciente, que señala que el volumen y los indicadores técnicos pueden contener información relevante para el pronóstico de retornos en mercados de criptomonedas, aunque su aporte debe evaluarse empíricamente bajo esquemas rigurosos de validación temporal (Vidal-Tomás & Ibáñez, 2023; Anastasiou, Ballis, & Drakos, 2023; Lahmiri & Bekiros, 2023).

En consecuencia, el análisis de indicadores técnicos constituye una etapa clave dentro del proceso exploratorio, ya que permite justificar la construcción del feature set multivariado y preparar la posterior comparación entre modelos con y sin variables exógenas.

7.6. Relaciones entre variables, correlaciones y multicolinealidad

El análisis de relaciones entre variables permite explorar posibles asociaciones entre el retorno logarítmico de BTC-USD y los predictores incorporados al enfoque multivariado. Esta etapa es relevante porque las variables financieras no suelen actuar de forma aislada; por el

contrario, pueden presentar relaciones dinámicas entre volatilidad, volumen, momentum, tendencia y flujo de mercado.

En el contexto de Bitcoin, estas relaciones pueden ser inestables en el tiempo y depender de las condiciones del mercado. Por esta razón, el análisis exploratorio no se limita a revisar estadísticas descriptivas individuales, sino que también examina la interacción entre grupos de variables. Este procedimiento permite identificar posibles patrones de dependencia, asociaciones no lineales, redundancia entre predictores y señales preliminares que pueden ser relevantes para el proceso de modelado multivariado.

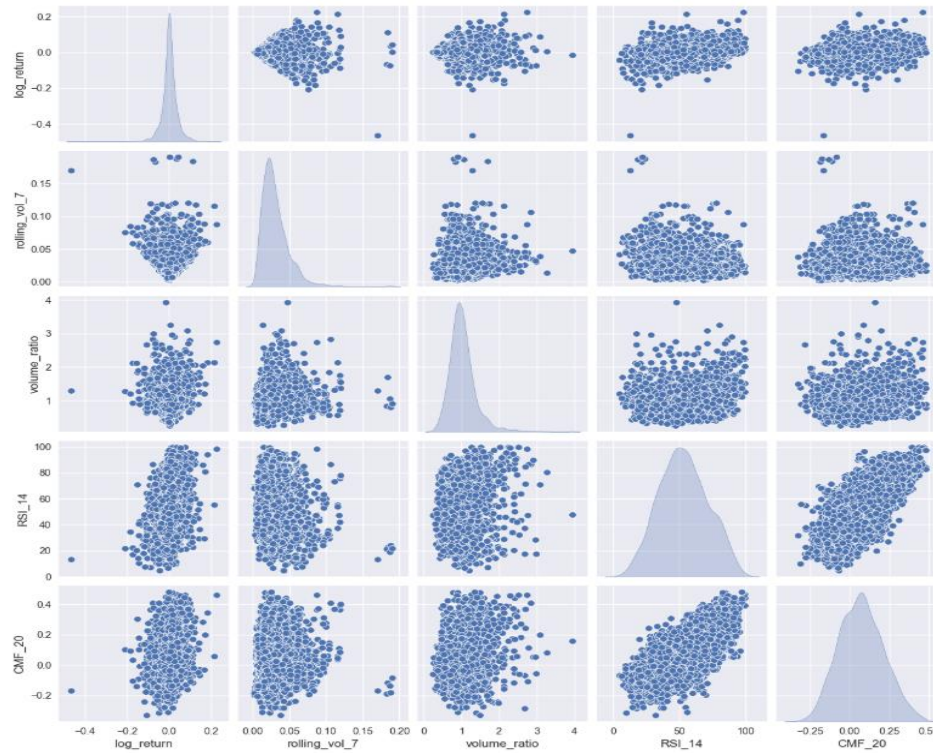
Adicionalmente, el análisis de correlaciones y multicolinealidad permite evaluar si algunas variables explicativas contienen información similar. Este aspecto es importante porque una alta redundancia entre predictores puede afectar la estabilidad de ciertos modelos estadísticos y aumentar el riesgo de sobreajuste, especialmente cuando se trabaja con múltiples indicadores técnicos derivados de las mismas variables base de precio y volumen. Por ello, la revisión conjunta de relaciones visuales, correlaciones y multicolinealidad contribuye a una construcción más robusta del feature set predictivo (Hastie, Tibshirani, & Friedman, 2017; Shumway & Stoffer, 2025).

7.6.1. Comparación visual entre variables

La comparación visual entre variables permite explorar de manera preliminar posibles asociaciones entre el retorno logarítmico de BTC-USD y algunos predictores relevantes del enfoque multivariado. En este análisis se consideran variables asociadas con volatilidad, volumen, momentum y flujo de mercado, tales como `rolling_vol_7`, `volume_ratio`, `RSI_14` y `CMF_20`.

Figura 5

Comparación visual entre retorno logarítmico y variables explicativas seleccionadas



Nota: elaboración propia con base en datos OHLCV de BTC-USD obtenidos de Yahoo Finance; variables calculadas en Python.

Como se observa en la Figura 5, el retorno logarítmico se concentra principalmente alrededor de cero, aunque presenta valores extremos negativos y positivos. La relación entre el retorno y las variables explicativas no muestra patrones lineales fuertes, lo cual es consistente con la naturaleza compleja y volátil del mercado de Bitcoin. Sin embargo, se observan estructuras de dispersión que sugieren posibles relaciones no lineales, especialmente entre el retorno, la volatilidad móvil y algunos indicadores de flujo de mercado.

La figura también permite identificar asociaciones importantes entre variables explicativas. Por ejemplo, RSI_14 y CMF_20 presentan una relación positiva visible, lo que sugiere cierta conexión entre momentum y presión de flujo de mercado. Asimismo, variables como rolling_vol_7 y volume_ratio muestran distribuciones asimétricas y presencia de valores extremos, lo cual

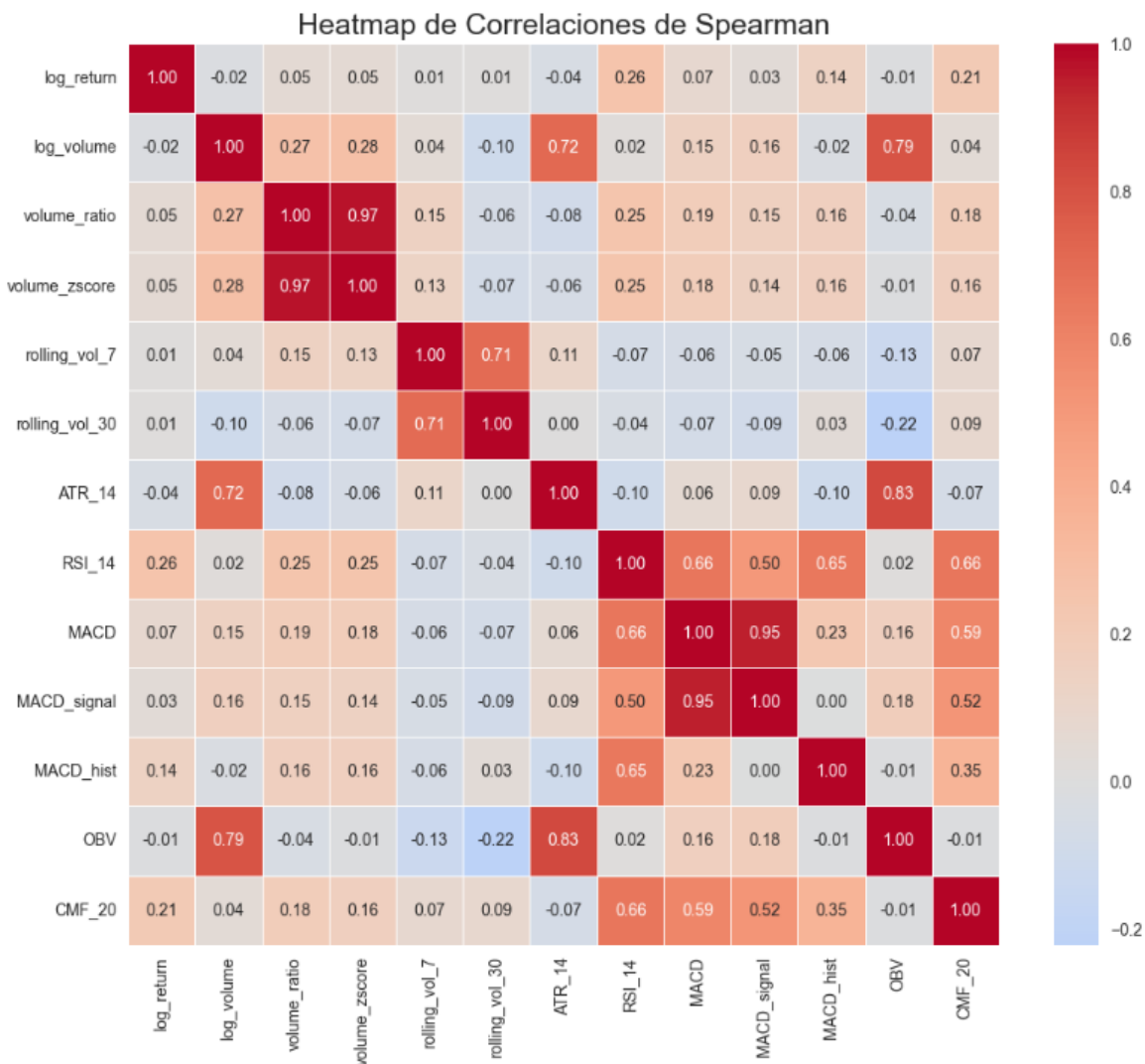
confirma la necesidad de controlar la escala, la dispersión y la posible redundancia entre predictores.

Desde el punto de vista metodológico, esta exploración visual respalda el uso de modelos multivariados y algoritmos flexibles, ya que las relaciones entre las variables no parecen ser exclusivamente lineales. Por tanto, la incorporación de modelos como XGBoost, Random Forest, LSTM, GRU, Prophet, NeuralProphet y GAM permite evaluar si estas estructuras complejas aportan información predictiva adicional frente a los modelos univariados tradicionales.

7.6.2. Matriz de correlación

La matriz de correlación permite evaluar el grado de asociación entre la variable objetivo y las variables explicativas incorporadas al enfoque multivariado. En este análisis se utiliza la correlación de Spearman, debido a que las variables financieras de Bitcoin pueden presentar valores extremos, asimetrías y relaciones no estrictamente lineales. Esta medida resulta adecuada para identificar asociaciones monotónicas entre variables sin depender exclusivamente de relaciones lineales simples.

Como se observa en la Figura 6, el retorno logarítmico de BTC-USD presenta correlaciones bajas o moderadas con la mayoría de las variables explicativas. Este resultado es consistente con la dificultad de pronosticar retornos financieros mediante relaciones lineales simples, especialmente en mercados altamente volátiles como Bitcoin. No obstante, algunas variables muestran asociaciones positivas con el retorno, como RSI_14 y CMF_20, lo cual sugiere que los indicadores de momentum y flujo de mercado pueden aportar información complementaria dentro del enfoque multivariado.

Figura 6*Matriz de correlaciones de Spearman entre variables del enfoque multivariado*

Nota: elaboración propia con base en datos OHLCV de BTC-USD obtenidos de Yahoo Finance; variables calculadas en Python.

La matriz también evidencia correlaciones elevadas entre algunas variables explicativas. Por ejemplo, `volume_ratio` y `volume_zscore` presentan una asociación muy alta, al igual que `MACD` y `MACD_signal`. Este comportamiento es esperable, dado que varias variables provienen de transformaciones relacionadas o de indicadores contruidos a partir de una misma fuente de

información. De igual forma, se observa una relación importante entre ATR_14, log_volume y OBV, lo que refleja la conexión entre volatilidad, volumen e intensidad de negociación.

Desde el punto de vista metodológico, estas correlaciones no invalidan el uso de las variables, pero sí advierten sobre la posible existencia de redundancia entre predictores. Por esta razón, el análisis de correlación se complementa posteriormente con el diagnóstico de multicolinealidad mediante VIF y con la evaluación empírica del desempeño de los modelos. Esta revisión es importante porque la presencia de variables altamente correlacionadas puede afectar la estabilidad de algunos modelos estadísticos, aunque modelos como XGBoost, Random Forest o redes neuronales pueden manejar mejor estructuras de dependencia complejas entre predictores (Hastie, Tibshirani, & Friedman, 2017).

Adicionalmente, se compararon las correlaciones de Pearson y Spearman, cuyos principales resultados se presentan en el Anexo 1. Esta comparación muestra diferencias moderadas en algunos pares de variables, como log_volume–volume_ratio, ATR_14–log_volume, MACD–RSI_14 y ATR_14–OBV. Estas diferencias refuerzan la conveniencia de priorizar Spearman en el cuerpo principal del análisis, dado que captura mejor relaciones monotónicas en presencia de distribuciones asimétricas y valores extremos.

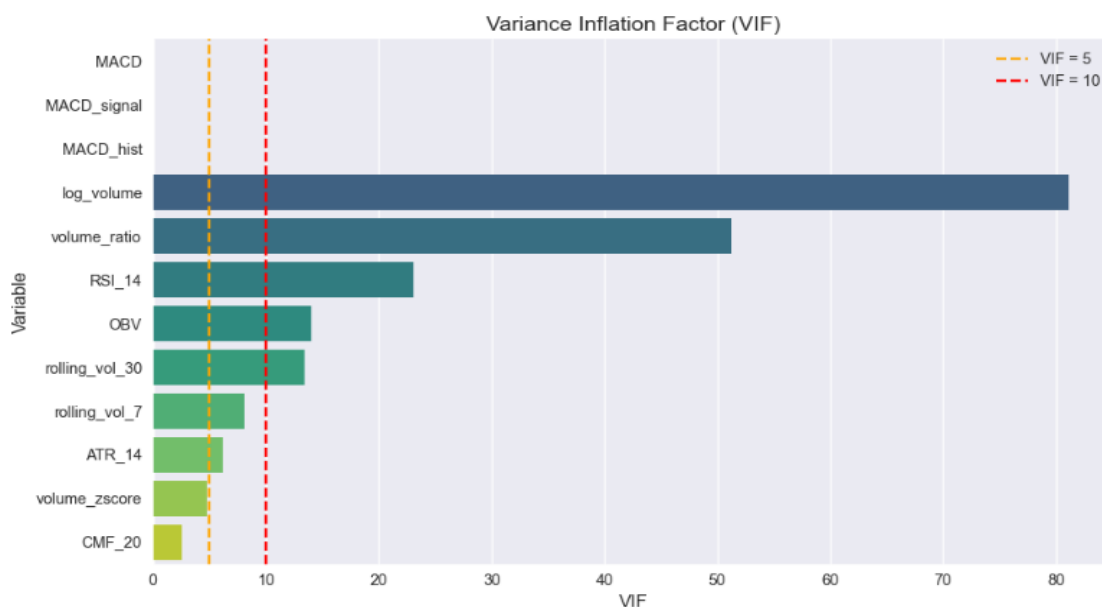
En conjunto, la matriz de correlación permite identificar relaciones relevantes entre los grupos de variables utilizados en el modelo multivariado: volumen, volatilidad, momentum, tendencia y flujo de mercado. Aunque las asociaciones lineales con el retorno son limitadas, la estructura de dependencia entre predictores justifica el uso de modelos flexibles capaces de capturar interacciones no lineales y posibles efectos combinados en el pronóstico del retorno logarítmico de Bitcoin.

7.6.3. Diagnóstico de multicolinealidad

La multicolinealidad se evaluó mediante el Factor de Inflación de la Varianza (VIF), con el fin de identificar posibles relaciones de redundancia entre las variables explicativas del enfoque multivariado. Este diagnóstico es importante porque valores elevados de VIF pueden afectar la estabilidad de algunos modelos estadísticos, dificultar la interpretación de los coeficientes y aumentar la varianza de las estimaciones.

Figura 7

Diagnóstico de multicolinealidad mediante el Factor de Inflación de la Varianza (VIF)



Nota: elaboración propia con base en datos OHLCV de BTC-USD obtenidos de Yahoo Finance; variables calculadas en Python.

Como se observa en la Figura 7, algunas variables presentan niveles elevados de multicolinealidad. En particular, `log_volume` y `volume_ratio` muestran valores de VIF considerablemente altos, lo que sugiere una fuerte redundancia asociada a la información de volumen. Sin embargo, es importante precisar que el VIF no indica la relación de una variable con otra variable específica, sino qué tanto una variable puede ser explicada por el conjunto de las demás variables predictoras. Por ejemplo, el VIF elevado de `log_volume` indica que esta variable

comparte información con varias variables del sistema, especialmente con aquellas asociadas al volumen, la volatilidad y el flujo de mercado, como OBV y ATR_14.

Asimismo, variables como RSI_14, OBV y rolling_vol_30 presentan valores por encima del umbral de referencia de 10, indicando una colinealidad relevante que debe ser considerada en la interpretación de los modelos. Este resultado es coherente con la matriz de correlación, donde se observaron asociaciones importantes entre variables derivadas de información similar, como volumen, volatilidad, momentum y flujo de mercado.

En contraste, variables como CMF_20, volume_zscore, ATR_14 y rolling_vol_7 presentan niveles más moderados de VIF, lo que sugiere una menor redundancia relativa dentro del conjunto de predictores. Por su parte, las variables MACD, MACD_signal y MACD_hist no presentan barra visible en la figura, lo cual puede estar asociado a colinealidad extrema o perfecta, dado que se construyen a partir de la misma estructura de tendencia y momentum.

Desde el punto de vista metodológico, estos resultados no implican necesariamente eliminar todas las variables con VIF alto, pero sí evidencian que el conjunto de predictores contiene información superpuesta. Por esta razón, el diagnóstico de multicolinealidad se interpreta de manera conjunta con la matriz de correlación y con el desempeño empírico de los modelos. La multicolinealidad puede afectar con mayor intensidad a modelos estadísticos lineales o interpretables, mientras que enfoques como Random Forest, XGBoost, LSTM y GRU suelen ser más robustos frente a estructuras complejas de dependencia entre predictores.

En consecuencia, el análisis de VIF permitió identificar redundancias dentro del feature set multivariado y reforzó la necesidad de evaluar los modelos bajo esquemas rigurosos de validación temporal. De esta manera, la selección final de modelos no se basa únicamente en la independencia

entre variables, sino en su capacidad empírica para mejorar el pronóstico del retorno logarítmico de BTC-USD.

7.7. Causalidad de Granger

La prueba de causalidad de Granger se utilizó como herramienta exploratoria para evaluar si los valores rezagados de las variables exógenas contienen información útil para anticipar el retorno logarítmico de BTC-USD. En este contexto, el término causalidad no debe interpretarse como una relación causal estructural, sino como evidencia de precedencia temporal y capacidad predictiva.

Tabla 3

Mejores resultados de la prueba de causalidad de Granger por variable exógena

Bloque	Variable exógena	Rezago	F-statistic	p-value	Significativa
Indicadores técnicos	RSI_14	24	2.289905	0.000346	Sí
Flujo de mercado	CMF_20	1	6.766439	0.009331	Sí
Volumen	log_volume	8	2.104982	0.032133	Sí
Volatilidad	rolling_vol_30	2	3.044650	0.047750	Sí
Volumen	volume_zscore	1	3.701988	0.054435	No
Volatilidad	ATR_14	1	2.915530	0.087827	No
Volumen	volume_ratio	1	2.614556	0.105985	No
Volatilidad	rolling_vol_7	27	1.293048	0.142504	No
Flujo de mercado	OBV	1	1.919610	0.165995	No
Indicadores técnicos	MACD_signal	23	1.274786	0.171149	No
Indicadores técnicos	MACD_hist	23	1.263941	0.179192	No

Bloque	Variable exógena	Rezago	F-statistic	p-value	Significativa
Indicadores técnicos	MACD	23	1.241152	0.197014	No

Nota: elaboración propia con base en datos OHLCV de BTC-USD obtenidos de Yahoo Finance; prueba de Granger calculada en Python.

Los resultados en la Tabla 3 muestran que algunas variables presentan evidencia estadística de capacidad predictiva sobre el retorno logarítmico de Bitcoin. En particular, RSI_14 registra un valor-p de 0.000346 en el rezago 24, lo que sugiere que este indicador de momentum contiene información rezagada relevante para anticipar el comportamiento futuro del retorno. Asimismo, CMF_20, asociado al flujo de mercado, presenta significancia estadística en el rezago 1, indicando que la presión compradora o vendedora reciente puede aportar información predictiva de corto plazo.

Dentro del bloque de volumen, log_volume resulta significativo en el rezago 8, mientras que rolling_vol_30, perteneciente al bloque de volatilidad, muestra significancia en el rezago 2. Estos resultados sugieren que tanto la actividad transaccional como la volatilidad histórica de mediano plazo pueden contener señales útiles para el pronóstico del retorno logarítmico de BTC-USD.

En contraste, variables como volume_zscore, ATR_14, volume_ratio, rolling_vol_7, OBV, MACD_signal, MACD_hist y MACD no presentan evidencia estadística significativa bajo el criterio utilizado. Esto no implica que dichas variables carezcan completamente de utilidad predictiva, sino que, bajo la prueba de Granger aplicada individualmente, sus rezagos no mostraron una relación temporal suficientemente fuerte con el retorno logarítmico.

En conjunto, la prueba de Granger complementa el análisis de correlación y multicolinealidad, ya que permite evaluar no solo asociaciones contemporáneas, sino también

relaciones temporales rezagadas. Los resultados respaldan la inclusión de variables de momentum, flujo de mercado, volumen y volatilidad dentro del enfoque multivariado, aunque su aporte final debe ser evaluado mediante el desempeño empírico de los modelos bajo validación temporal.

7.8. Síntesis del análisis exploratorio

El análisis exploratorio permitió caracterizar la serie BTC-USD antes de la etapa de modelado. El precio de cierre presentó cambios importantes de nivel, fases de crecimiento acelerado, correcciones pronunciadas y episodios de alta volatilidad. Por esta razón, el estudio trabaja con retornos logarítmicos y no con precios en niveles. Esta transformación permite analizar variaciones relativas del activo y facilita el modelado estadístico de la serie (Shumway & Stoffer, 2025; Hyndman & Athanasopoulos, 2021).

Los retornos logarítmicos se concentran alrededor de cero, pero incluyen valores extremos, volatilidad agrupada y cambios abruptos de magnitud. Estas condiciones son habituales en activos financieros altamente volátiles y dificultan el uso de modelos lineales simples. Bitcoin conserva, además, características como heterocedasticidad, colas pesadas y posibles cambios de régimen (Bouri, Gil-Alana, Gupta, & Roubaud, 2020; Catania, Grassi, & Ravazzolo, 2019).

Las volatilidades móviles de 7 y 30 días mostraron comportamientos complementarios. La ventana de 7 días reaccionó con mayor rapidez a los choques recientes, mientras que la de 30 días reflejó variaciones más persistentes. El volumen de negociación también presentó cambios marcados y varios picos durante el periodo analizado. Estas variables se incorporaron al enfoque multivariado porque pueden reflejar condiciones relevantes de actividad e inestabilidad del mercado (Vidal-Tomás & Ibáñez, 2023; Guo, Sang, Tu, & Wang, 2024).

Los indicadores RSI, MACD, ATR, OBV y CMF se incluyeron como predictores, no como reglas de negociación. Cada uno resume dimensiones distintas del mercado: momentum, tendencia,

volatilidad o flujo de capital. Su uso permite ampliar el conjunto de información disponible frente a un modelo basado únicamente en los rezagos del retorno (Anastasiou, Ballis, & Drakos, 2023; Lahmiri & Bekiros, 2023).

La revisión de correlaciones mostró asociaciones relevantes entre algunos predictores. El diagnóstico mediante VIF también evidenció redundancia, sobre todo entre variables relacionadas con volumen, tendencia y momentum. Esto exige cautela al interpretar modelos lineales y refuerza la necesidad de validar el desempeño fuera de muestra, en lugar de seleccionar variables únicamente por sus relaciones individuales (Hastie, Tibshirani, & Friedman, 2017).

La prueba de causalidad de Granger aportó evidencia exploratoria adicional. RSI_14, CMF_20, log_volume y rolling_vol_30 presentaron significancia estadística en algunos rezagos. Estos hallazgos indican que ciertas variables de momentum, flujo de mercado, volumen y volatilidad contienen información temporal útil para anticipar el retorno logarítmico de BTC-USD. No representan causalidad económica directa; muestran precedencia temporal y posible capacidad predictiva, tal como lo plantea la distinción entre predicción y explicación causal de Shmueli (2010).

En conjunto, el análisis exploratorio respalda la construcción del enfoque multivariado. Las variables revisadas no garantizan por sí solas una mejora del pronóstico, pero justifican evaluar si su incorporación reduce el error frente a modelos univariados basados únicamente en la historia del retorno.

La Tabla 4 reúne los hallazgos principales del capítulo y su relación con la etapa de modelado. Allí se sintetizan los resultados descriptivos, gráficos y estadísticos que sustentan la selección del conjunto de variables y la posterior comparación entre modelos univariados y multivariados.

Tabla 4*Síntesis del análisis exploratorio*

Componente analizado	Hallazgo principal	Implicación para el modelado
Precio de BTC-USD	Presenta tendencia, cambios de nivel y alta volatilidad.	Justifica trabajar con retornos logarítmicos.
Retornos logarítmicos	Oscilan alrededor de cero, con valores extremos y volatilidad agrupada.	Requiere modelos robustos y evaluación direccional.
Volatilidad histórica	rolling_vol_7 y rolling_vol_30 capturan inestabilidad de corto y mediano plazo.	Se incorporan como variables exógenas.
Volumen	Presenta picos y cambios fuertes en la actividad de mercado.	Justifica variables como log_volume, volume_ratio y volume_zscore.
Indicadores técnicos	Resumen momentum, tendencia, volatilidad y flujo de mercado.	Amplían el feature set multivariado.
Correlación y VIF	Existen relaciones y redundancias entre algunos predictores.	Requiere interpretar con cautela modelos lineales y comparar con modelos flexibles.
Granger	Algunas variables muestran capacidad predictiva temporal.	Respalda la evaluación empírica de modelos multivariados.

Nota: elaboración propia con base en los resultados del análisis exploratorio desarrollado en Python.

8. Modelos de pronóstico

En este capítulo se presentan las familias de modelos implementadas para el pronóstico del retorno logarítmico de BTC-USD. La estrategia de modelado se estructuró en dos enfoques: un enfoque univariado, basado únicamente en la historia del retorno, y un enfoque multivariado, que incorpora variables exógenas asociadas al volumen, la volatilidad y los indicadores técnicos.

La comparación incluye modelos de referencia, modelos clásicos, modelos con regresores exógenos, modelos en espacio de estados, algoritmos de machine learning, arquitecturas de deep

learning, modelos aditivos y modelos híbridos. Esta organización permite evaluar si la incorporación de información multivariada y el uso de métodos más flexibles generan mejoras predictivas frente a modelos simples y parsimoniosos.

Todos los modelos fueron evaluados bajo esquemas de validación temporal, respetando el orden cronológico de la serie y evitando fuga de información. El desempeño se analizó mediante métricas de error, métricas direccionales y pruebas estadísticas de comparación predictiva, con el fin de valorar tanto la precisión en magnitud como la capacidad para anticipar la dirección del retorno.

La Tabla 5 resume las principales familias de modelos implementadas y diferencia su aplicación en los enfoques univariado y multivariado. En el enfoque univariado, los modelos utilizan únicamente información histórica del retorno logarítmico, mientras que en el enfoque multivariado se incorporan variables exógenas asociadas al volumen, la volatilidad, el momentum y el flujo de mercado. Esta estructura permite contrastar si la información adicional mejora el desempeño predictivo frente a modelos basados solo en la dinámica pasada del retorno.

Tabla 5

Familias de modelos implementadas

Familia	Enfoque univariado	Enfoque multivariado
Referencia	Naive, SNaive	Benchmark comparativo
Clásicos	AutoETS, SARIMA, AutoARIMA, Theta	ARIMAX, AutoARIMAX, SARIMAX, ETSX
Machine learning	MLP, Random Forest, XGBoost	MLP, Random Forest, XGBoost con exógenas

Familia	Enfoque univariado	Enfoque multivariado
Deep learning	LSTM, GRU	LSTM, GRU con exógenas
Aditivos	Prophet, NeuralProphet, GAM	Prophet, NeuralProphet, GAM con regresores
Híbridos	AutoARIMA + MLP, AutoARIMA + XGBoost	Según implementación final

Nota: elaboración propia con base en los modelos implementados en Python.

Como se observa en la Tabla 5, el diseño metodológico permite comparar modelos simples, modelos estadísticos interpretables y modelos flexibles de aprendizaje automático y profundo. Esta estructura facilita evaluar si la mayor complejidad metodológica se traduce en mejoras reales frente a los modelos de referencia, principio fundamental en la evaluación de pronósticos financieros (Hyndman & Athanasopoulos, 2021; Hyndman & Koehler, 2006).

8.1. Modelos de referencia

Los modelos de referencia, también conocidos como modelos baseline, se implementaron con el propósito de establecer una línea base mínima para la comparación predictiva. Su función principal no es capturar estructuras complejas de la serie, sino servir como punto de partida para evaluar si los modelos más sofisticados logran generar mejoras reales en el pronóstico del retorno logarítmico de BTC-USD.

En este proyecto se consideraron dos modelos de referencia: Naive y Seasonal Naive SNaive(7). El modelo Naive asume que el mejor pronóstico para el siguiente periodo corresponde al último valor observado de la serie. En términos prácticos, este modelo proyecta hacia adelante

el retorno logarítmico más reciente, sin incorporar estructuras adicionales de tendencia, estacionalidad o variables explicativas.

Por su parte, el modelo SNaive(7) utiliza como pronóstico el valor observado siete periodos atrás. Dado que la serie analizada tiene frecuencia diaria, el rezago de siete días permite incorporar una referencia semanal simple, útil para evaluar si existe algún patrón de repetición asociado al comportamiento semanal del mercado.

La inclusión de estos modelos resulta importante porque los retornos financieros suelen presentar un componente aleatorio elevado y baja autocorrelación lineal. En este contexto, superar consistentemente a modelos simples como Naive o SNaive no siempre es trivial. Por esta razón, estos modelos permiten aplicar el principio de parsimonia, según el cual un modelo más complejo solo se justifica si ofrece una mejora clara frente a alternativas simples y fácilmente interpretables (Hyndman & Athanasopoulos, 2021; Hyndman & Koehler, 2006).

En el presente estudio, los modelos Naive y SNaive(7) se emplean como benchmarks para contrastar el desempeño de las demás familias de modelos. En el enfoque univariado, estos modelos se aplican directamente sobre la serie de retornos logarítmicos; mientras que, en la comparación general, sirven como referencia para evaluar si los modelos clásicos, multivariados, de machine learning, deep learning, aditivos e híbridos logran capturar información predictiva adicional.

Los resultados cuantitativos y gráficos asociados a estos modelos se presentan posteriormente en el capítulo de resultados, donde se comparan sus métricas de error y desempeño direccional frente a los demás enfoques implementados.

8.2. Modelos clásicos univariados

Los modelos clásicos univariados se implementaron con el propósito de evaluar la capacidad predictiva contenida únicamente en la historia del retorno logarítmico de BTC-USD. A diferencia de los modelos multivariados, estos enfoques no incorporan variables exógenas como volumen, volatilidad o indicadores técnicos, sino que utilizan la estructura temporal pasada de la propia variable objetivo.

En este grupo se consideraron modelos ampliamente utilizados en el análisis de series de tiempo, como AutoETS, SARIMA, AutoARIMA y Theta. Estos modelos permiten capturar distintos tipos de dependencia temporal, componentes de error, posibles patrones estacionales y estructuras autorregresivas presentes en la serie. Su inclusión resulta relevante porque constituyen una base estadística interpretable frente a la cual se pueden comparar modelos más flexibles y complejos (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025).

El modelo AutoETS se basa en la descomposición de la serie en componentes de error, tendencia y estacionalidad, seleccionando automáticamente la especificación más adecuada. Por su parte, los modelos SARIMA y AutoARIMA permiten representar relaciones autorregresivas, diferenciación y componentes de medias móviles, incluyendo posibles patrones estacionales. El modelo Theta, por su parte, se considera una alternativa clásica de pronóstico que ha mostrado buen desempeño empírico en distintos contextos de series temporales.

En el presente proyecto, estos modelos se aplicaron sobre la serie de retornos logarítmicos de Bitcoin con el fin de establecer un punto de comparación estadístico más robusto que los modelos de referencia Naive y SNaive(7). Su evaluación permite determinar si las estructuras clásicas de series de tiempo logran capturar información predictiva relevante en los retornos de BTC-USD antes de incorporar variables exógenas.

Dado que los retornos financieros suelen presentar baja autocorrelación lineal, alta volatilidad y presencia de valores extremos, los modelos clásicos univariados pueden enfrentar limitaciones para capturar dinámicas no lineales o cambios abruptos del mercado. Por esta razón, sus resultados se interpretan como una línea base estadística frente a la cual se comparan posteriormente los modelos multivariados, de machine learning, deep learning, aditivos e híbridos.

8.3. Modelos clásicos multivariados

Los modelos clásicos multivariados se implementaron con el propósito de evaluar si la incorporación de variables exógenas mejora el pronóstico del retorno logarítmico de BTC-USD frente a los modelos univariados. A diferencia de los modelos clásicos basados únicamente en la historia del retorno, estos enfoques permiten incluir información adicional asociada al volumen, la volatilidad y los indicadores técnicos.

En este grupo se consideraron modelos como ARIMAX, AutoARIMAX, SARIMAX y un modelo estructural tipo ETSX con variables exógenas. Estos modelos amplían la lógica de los modelos clásicos de series de tiempo al incorporar regresores externos, manteniendo una estructura estadística interpretable. Su utilidad radica en que permiten analizar si variables como `log_volume`, `volume_ratio`, `rolling_vol_7`, `rolling_vol_30`, `RSI_14`, `MACD`, `ATR_14`, `OBV` y `CMF_20` aportan información predictiva adicional al retorno histórico.

Los modelos ARIMAX y AutoARIMAX incorporan variables exógenas dentro de una estructura autorregresiva y de medias móviles. El modelo SARIMAX extiende esta formulación al incluir componentes estacionales, lo cual resulta pertinente en series diarias cuando se desea evaluar posibles patrones semanales. Por su parte, el modelo estructural tipo ETSX se implementó mediante un enfoque de espacio de estados con regresores externos, permitiendo combinar componentes dinámicos de la serie con información exógena del mercado.

En el presente proyecto, estos modelos cumplen una función metodológica central: sirven como puente entre los modelos clásicos univariados y los modelos más flexibles de machine learning y deep learning. Su implementación permite evaluar si la información multivariada mejora el desempeño predictivo dentro de una estructura estadística interpretable antes de recurrir a algoritmos más complejos.

Dado que los retornos de Bitcoin presentan alta volatilidad, bajo nivel de dependencia lineal y posibles relaciones no lineales, los modelos clásicos multivariados pueden presentar limitaciones frente a enfoques más flexibles. No obstante, su inclusión es importante porque permite establecer una comparación rigurosa entre modelos con y sin variables exógenas, manteniendo criterios de parsimonia e interpretabilidad dentro del proceso de pronóstico (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025).

8.4. Modelos en espacio de estados

Los modelos en espacio de estados se implementaron como una alternativa estructural e interpretable para el pronóstico del retorno logarítmico de BTC-USD. Este tipo de modelos permite representar la serie mediante componentes dinámicos no observables, como nivel, tendencia o estacionalidad, los cuales se actualizan a medida que se incorpora nueva información temporal.

En el presente proyecto, este enfoque se aplicó mediante un modelo estructural con variables exógenas, implementado en Python a través de la clase `UnobservedComponents` de `statsmodels`. Esta formulación permitió combinar un componente dinámico de la serie con regresores externos asociados al volumen, la volatilidad y los indicadores técnicos. De esta manera, el modelo funcionó como una aproximación tipo ETSX estructural, dentro del conjunto de modelos clásicos multivariados.

La utilidad de estos modelos radica en que permiten incorporar información externa sin perder interpretabilidad estadística. A diferencia de los modelos puramente autorregresivos, el enfoque de espacio de estados permite actualizar los componentes latentes de la serie mediante procedimientos recursivos, usualmente asociados al filtro de Kalman. Esto resulta pertinente en series financieras como Bitcoin, donde la dinámica del mercado puede cambiar con rapidez y presentar episodios de alta volatilidad (Shumway & Stoffer, 2025).

Dentro del estudio, el modelo en espacio de estados cumple una función comparativa frente a los modelos ARIMAX, SARIMAX y los modelos de aprendizaje automático. Su inclusión permite evaluar si una estructura estadística dinámica, con regresores exógenos, logra capturar información relevante del volumen, la volatilidad y los indicadores técnicos para mejorar el pronóstico del retorno logarítmico.

No obstante, al igual que otros modelos estadísticos clásicos, este enfoque puede presentar limitaciones frente a relaciones altamente no lineales o interacciones complejas entre predictores. Por esta razón, sus resultados se interpretan como parte del bloque de modelos estructurales e interpretables, y se comparan posteriormente con modelos más flexibles, como Random Forest, XGBoost, LSTM, GRU y modelos aditivos.

8.5. Modelos de machine learning

Los modelos de machine learning se implementaron con el propósito de evaluar enfoques más flexibles para el pronóstico del retorno logarítmico de BTC-USD. A diferencia de los modelos clásicos, estos métodos no requieren especificar de forma estricta una estructura lineal o paramétrica, sino que aprenden patrones a partir de los datos y pueden capturar relaciones no lineales e interacciones entre variables predictoras (Hastie, Tibshirani, & Friedman, 2017).

En este proyecto se consideraron tres modelos principales: MLP, Random Forest y XGBoost. El modelo MLP corresponde a una red neuronal multicapa utilizada para aproximar relaciones no lineales entre los rezagos del retorno y la variable objetivo. Random Forest combina múltiples árboles de decisión, reduciendo la variabilidad individual de cada árbol y mejorando la estabilidad del pronóstico. Por su parte, XGBoost utiliza un esquema de gradient boosting, en el cual los árboles se construyen de forma secuencial para corregir errores de modelos anteriores, incorporando además mecanismos de regularización para reducir el sobreajuste (Chen & Guestrin, 2016).

Estos modelos fueron aplicados tanto en el enfoque univariado como en el multivariado. En el caso univariado, los predictores corresponden principalmente a rezagos del retorno logarítmico. En el enfoque multivariado, además de los rezagos de la variable objetivo, se incorporan variables exógenas asociadas al volumen, la volatilidad, el momentum y el flujo de mercado. Esta estructura permite evaluar si los algoritmos de aprendizaje automático logran aprovechar información adicional frente a los modelos basados únicamente en la historia del retorno.

La inclusión de modelos de machine learning resulta pertinente en el contexto de Bitcoin, debido a que sus retornos pueden presentar baja autocorrelación lineal, alta volatilidad, valores extremos y relaciones complejas entre variables de mercado. En este sentido, estos modelos ofrecen una alternativa flexible para capturar patrones que pueden no ser adecuadamente representados por modelos estadísticos tradicionales.

No obstante, su implementación requiere controlar el riesgo de sobreajuste mediante validación temporal, selección adecuada de variables, escalamiento cuando corresponde y ajuste de hiperparámetros. Por esta razón, en el presente estudio los modelos de machine learning fueron

evaluados bajo los mismos esquemas de validación que las demás familias de modelos, garantizando una comparación homogénea y respetando la estructura cronológica de los datos.

8.6. Modelos de deep learning

Los modelos de deep learning se implementaron con el propósito de evaluar arquitecturas secuenciales capaces de aprender patrones temporales complejos en el retorno logarítmico de BTC-USD. A diferencia de los modelos clásicos y de algunos algoritmos de machine learning, estas arquitecturas trabajan con ventanas de observaciones pasadas, lo que permite representar la dependencia temporal de la serie dentro del proceso de entrenamiento.

En este proyecto se consideraron dos arquitecturas principales: LSTM (Long Short-Term Memory) y GRU (Gated Recurrent Unit). Las redes LSTM están diseñadas para conservar información relevante a través del tiempo mediante mecanismos de memoria y compuertas, lo que ayuda a mitigar el problema del desvanecimiento del gradiente en secuencias largas (Hochreiter & Schmidhuber, 1997). Las redes GRU tienen una lógica similar, pero con una estructura más compacta, lo que puede reducir el costo computacional y facilitar el entrenamiento.

Estos modelos fueron aplicados tanto en el enfoque univariado como en el multivariado. En el enfoque univariado, las secuencias de entrada se construyeron a partir de rezagos del retorno logarítmico. En el enfoque multivariado, las ventanas temporales incorporaron además variables exógenas asociadas al volumen, la volatilidad y los indicadores técnicos, permitiendo evaluar si esta información adicional mejora la capacidad predictiva del modelo.

La inclusión de LSTM y GRU resulta pertinente en el contexto de Bitcoin debido a que sus retornos pueden presentar patrones no lineales, episodios de volatilidad agrupada y cambios abruptos en la dinámica del mercado. Estudios recientes han explorado el uso de modelos de aprendizaje profundo para el pronóstico de criptomonedas, destacando su potencial para capturar

relaciones complejas en series financieras altamente volátiles (McNally, Roche, & Caton, 2018; Lahmiri & Bekiros, 2020).

Desde el punto de vista metodológico, estos modelos requieren procesos específicos de preparación de datos, como escalamiento de variables, construcción de ventanas temporales y transformación de los datos en estructuras tridimensionales. Asimismo, se emplearon estrategias de control de sobreajuste, como early stopping y regularización, con el fin de mejorar la capacidad de generalización. Sus resultados se comparan posteriormente con los modelos clásicos, de machine learning, aditivos e híbridos bajo los mismos esquemas de validación temporal.

8.7. Modelos aditivos

Los modelos aditivos se implementaron como una familia flexible e interpretable para el pronóstico del retorno logarítmico de BTC-USD. Estos modelos permiten representar la serie como una combinación de componentes, tales como tendencia, estacionalidad, efectos de calendario, relaciones no lineales y regresores externos. Su principal ventaja es que ofrecen una estructura relativamente interpretable, manteniendo capacidad para adaptarse a patrones complejos de la serie.

En este proyecto se consideraron tres enfoques principales: Prophet, NeuralProphet y GAM. El modelo Prophet permite descomponer la serie en componentes de tendencia, estacionalidad y efectos adicionales, además de incorporar regresores externos cuando se trabaja bajo un enfoque multivariado (Taylor & Letham, 2017). NeuralProphet amplía esta estructura mediante componentes autorregresivos y redes neuronales, conservando parte de la interpretabilidad del enfoque aditivo, pero con mayor flexibilidad para capturar patrones temporales complejos (Triebe et al., 2021).

Por su parte, los modelos GAM permiten representar relaciones no lineales entre la variable objetivo y los predictores mediante funciones suaves. Esta característica resulta útil cuando el efecto de variables como volumen, volatilidad o indicadores técnicos sobre el retorno logarítmico no puede describirse adecuadamente mediante una relación lineal simple.

Estos modelos fueron aplicados tanto en el enfoque univariado como en el multivariado. En el enfoque univariado, se emplearon principalmente componentes temporales y rezagos del retorno. En el enfoque multivariado, se incorporaron regresores asociados al volumen, la volatilidad, el momentum y el flujo de mercado, con el fin de evaluar si estas variables aportan información adicional al proceso de pronóstico.

La inclusión de modelos aditivos permite complementar la comparación entre modelos clásicos, modelos de machine learning y arquitecturas de deep learning. Su valor metodológico radica en que ofrecen un equilibrio entre flexibilidad e interpretabilidad, lo cual resulta relevante en un proyecto donde no solo interesa obtener buenos pronósticos, sino también analizar el posible aporte de las variables exógenas en el comportamiento predictivo del retorno logarítmico de Bitcoin.

8.8. Modelos híbridos

Los modelos híbridos se implementaron con el propósito de combinar la capacidad interpretativa de los modelos clásicos de series de tiempo con la flexibilidad de los algoritmos de machine learning. Este enfoque parte de la idea de que un modelo estadístico puede capturar la estructura temporal principal de la serie, mientras que un modelo de aprendizaje automático puede aprovechar patrones remanentes presentes en los residuos.

En el presente proyecto se consideraron modelos híbridos basados en la combinación de AutoARIMA + MLP y AutoARIMA + XGBoost. En una primera etapa, el modelo AutoARIMA

se ajusta sobre la serie del retorno logarítmico de BTC-USD para capturar dependencias lineales, componentes autorregresivos y estructuras de medias móviles. Posteriormente, los residuos generados por este modelo se utilizan como insumo para un segundo modelo, orientado a identificar posibles patrones no lineales no capturados por la estructura clásica.

El modelo AutoARIMA + MLP busca complementar la estructura estadística inicial mediante una red neuronal multicapa, capaz de aproximar relaciones no lineales en los residuos. Por su parte, el modelo AutoARIMA + XGBoost utiliza árboles secuenciales de gradient boosting para corregir errores residuales y capturar interacciones más complejas entre rezagos o variables complementarias.

La utilidad metodológica de los modelos híbridos radica en que permiten evaluar si la combinación de enfoques mejora el desempeño predictivo frente a modelos clásicos o de machine learning aplicados de forma independiente. En el contexto de Bitcoin, donde los retornos pueden presentar alta volatilidad, bajo nivel de autocorrelación lineal y patrones no lineales, esta estrategia permite explorar si los residuos de un modelo clásico aún contienen información aprovechable para el pronóstico.

Al igual que las demás familias de modelos, los modelos híbridos fueron evaluados mediante esquemas de validación temporal, evitando fuga de información y respetando el orden cronológico de la serie. Sus resultados se interpretan como parte de la comparación general entre enfoques univariados, multivariados y combinados, con el fin de determinar si la integración de modelos estadísticos y algoritmos flexibles aporta mejoras en la predicción del retorno logarítmico de BTC-USD.

9. Resultados

En este capítulo se presentan los resultados de los modelos utilizados para pronosticar el retorno logarítmico de BTC-USD. La evaluación sigue los esquemas definidos en la metodología: Hold-Out y Rolling-Origin, con horizontes de uno y siete días. Ambos procedimientos respetan el orden cronológico de la serie y permiten medir el desempeño fuera de muestra, condición necesaria en ejercicios de pronóstico financiero (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025).

Se incluyen modelos univariados y multivariados de distintas familias: modelos de referencia, modelos clásicos, modelos con regresores exógenos, modelos en espacio de estados, algoritmos de machine learning, arquitecturas de deep learning, modelos aditivos y modelos híbridos. El interés principal es examinar si las variables de volumen, volatilidad e indicadores técnicos aportan información útil frente a modelos contruidos únicamente con los rezagos del retorno.

La precisión en magnitud se evalúa con MAE, RMSE, sMAPE y MASE. La capacidad para anticipar el signo del retorno se revisa con Accuracy, Precision, Recall, F1-score, MCC y AUC. MASE se adopta como métrica principal porque permite comparar el error de cada modelo frente a un referente ingenuo (Hyndman & Koehler, 2006; Hyndman & Athanasopoulos, 2021).

La prueba de Diebold-Mariano complementa las métricas descriptivas. Se aplica para contrastar formalmente si los errores de determinados modelos difieren de los del benchmark Naive, bajo la hipótesis de igualdad de precisión predictiva (Diebold & Mariano, 1995).

El capítulo no se limita a ordenar modelos por error. También revisa estabilidad entre esquemas de validación, desempeño direccional y comportamiento del enfoque multivariado. Las

tablas y figuras muestran los resultados principales; las salidas completas se incluyen en los anexos..

9.1. Resultados Hold-Out

La validación Hold-Out corresponde a la primera evaluación fuera de muestra de los modelos implementados. Cada modelo se entrenó con la información disponible antes del periodo de validación y luego se evaluó sobre una ventana final de datos que no participó en el ajuste. De esta forma, se evita mezclar información pasada y futura y se reproduce una situación básica de pronóstico (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025).

El procedimiento se aplicó a los modelos univariados y multivariados. Los modelos univariados utilizaron únicamente los rezagos del retorno logarítmico de BTC-USD. Los modelos multivariados incorporaron, además, variables de volumen, volatilidad, momentum, tendencia y flujo de mercado. La comparación permite observar si esa información adicional se asocia con una reducción del error frente a modelos basados solo en la historia del retorno.

Para facilitar la lectura, el capítulo presenta los mejores modelos de cada familia, ordenados según MASE. Esta métrica permite comparar el error relativo frente a un modelo ingenuo de referencia, mientras que MAE y RMSE complementan la evaluación de la magnitud del error (Hyndman & Koehler, 2006). Los resultados completos se incluyen en los anexos: el Anexo 2 reúne los modelos univariados y el Anexo 3 los modelos multivariados.

Tabla 6*Mejores modelos univariados Hold-Out por familia ordenados por MASE*

Modelo	Familia	MAE	RMSE	sMAPE (%)	MASE	Hit Ratio	Precision	Recall	F1	DM Stat.	DM p-value
AutoARIMA + XGBoost	Híbrido	0.017	0.022	175.989	0.459	0.533	0.500	0.071	0.125	-3.156	0.002
MLP óptimo	Machine Learning	0.017	0.022	158.563	0.460	0.550	1.000	0.036	0.069	-2.810	0.005
NeuralProphet	Aditivo	0.017	0.022	170.644	0.461	0.550	0.516	0.571	0.542	-3.273	0.001
GAM	Aditivo	0.017	0.022	188.529	0.466	0.433	0.429	0.643	0.514	-3.384	0.001
GRU	Deep Learning	0.017	0.022	169.808	0.468	0.467	0.467	1.000	0.636	-3.583	0.000
AutoARIMA	Clásico	0.017	0.022	172.150	0.468	0.467	0.467	1.000	0.636	-3.565	0.000

Nota: elaboración propia con base en los resultados obtenidos en Python.

La Tabla 6 muestra que el modelo híbrido AutoARIMA + XGBoost obtuvo el menor MASE dentro del enfoque univariado. Le siguen MLP óptimo, NeuralProphet, GAM, GRU y AutoARIMA. Las diferencias entre sus valores de error son reducidas, por lo que ninguna familia presenta una ventaja amplia en esta primera evaluación Hold-Out.

Los resultados direccionales no siguen exactamente el mismo orden. GRU y AutoARIMA alcanzan valores altos de Recall, pero su Precision y F1-score son más bajos. Por esta razón, un mayor Recall no basta para considerar que un modelo predice mejor la dirección del retorno. La evaluación debe combinar las métricas de error con los indicadores direccionales, ya que precisión en magnitud y acierto en el signo del retorno pueden conducir a conclusiones distintas.

Tabla 7

Mejores modelos multivariados Hold-Out por familia ordenados por MASE

Modelo	Familia	MAE	RMSE	sMAPE (%)	MASE	Accuracy	F1-score	MCC	AUC	DM p-value
RandomForest_multivariado	Machine Learning	0.015	0.020	122.857	0.410	0.600	0.500	0.189	0.674	0.014
AutoARIMAX_MLP_Híbrido	Híbrido	0.015	0.020	132.449	0.428	0.617	0.610	0.236	0.633	0.007
GAM_multivariado	Aditivo	0.015	0.020	121.125	0.431	0.733	0.714	0.464	0.786	0.008
AutoARIMAX	Clásico	0.016	0.021	153.010	0.457	0.633	0.560	0.259	0.631	0.001
LSTM_multi	Deep Learning	0.017	0.022	152.285	0.467	0.533	0.000	0.000	0.595	0.011
SNaive	Baseline	0.020	0.027	134.656	0.577	0.567	0.519	0.126	0.564	0.930

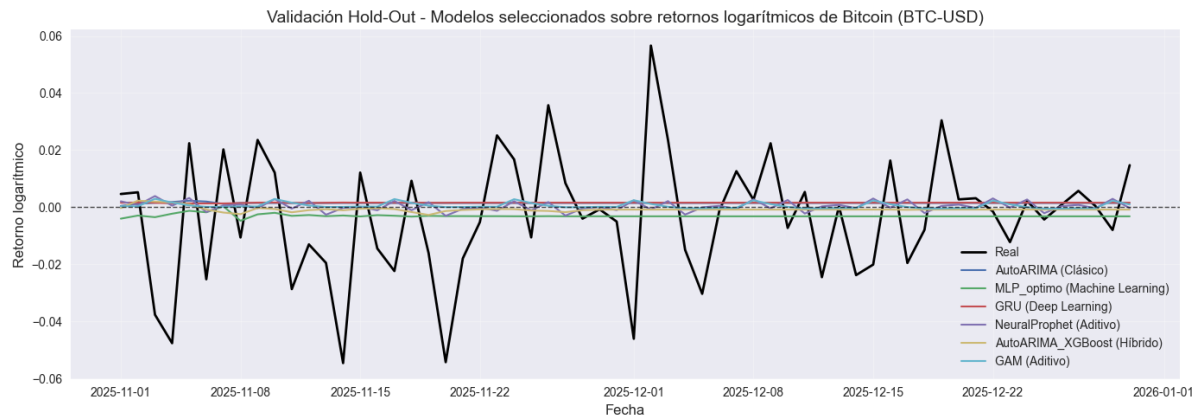
Nota: elaboración propia con base en los resultados obtenidos en Python; La prueba Diebold-Mariano se calcula frente al modelo Naive como referencia.

La Tabla 7 muestra que, en la validación Hold-Out, los modelos multivariados obtienen los menores errores entre las alternativas seleccionadas. RandomForest multivariado registra el menor MASE y el menor RMSE, seguido por el modelo híbrido AutoARIMAX + MLP y por GAM multivariado. En esta ventana de validación, las variables de volumen, volatilidad e indicadores técnicos aportan información que ayuda a reducir el error frente a los modelos univariados.

El resultado cambia al revisar la dirección del retorno. GAM multivariado presenta los valores más altos de Accuracy, F1-score, MCC y AUC dentro de la tabla. Aunque no tiene el menor error de magnitud, clasifica mejor el sentido del movimiento del retorno. Por eso, Random Forest resulta más conveniente cuando se prioriza MASE, mientras que GAM es la alternativa más destacada cuando el interés principal es anticipar si el retorno será positivo o negativo.

Figura 8

Pronósticos Hold-Out de modelos univariados seleccionados frente al retorno observado de BTC-USD



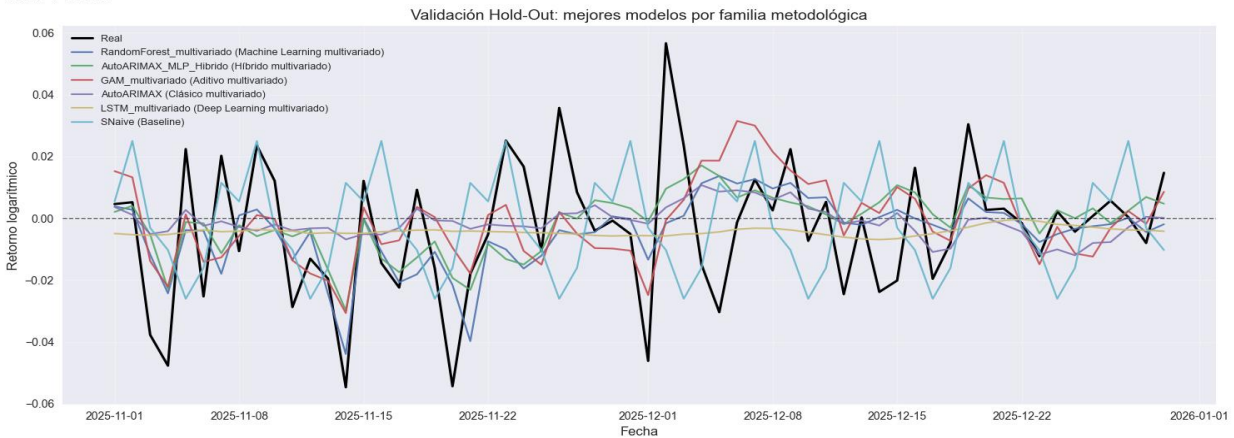
Nota: elaboración propia con base en datos OHLCV de BTC-USD obtenidos de Yahoo Finance; modelos estimados en Python.

La Figura 8 permite observar que los modelos univariados tienden a suavizar la dinámica del retorno observado y presentan dificultad para capturar movimientos extremos. Esto es esperable, dado que estos modelos utilizan únicamente la historia del retorno logarítmico y no incorporan información adicional sobre volumen, volatilidad o indicadores técnicos. En consecuencia, aunque algunos modelos univariados mejoran frente al benchmark, su capacidad para reproducir cambios abruptos del mercado resulta limitada.

La validación Hold-Out ofrece una primera comparación favorable al enfoque multivariado. Los mejores modelos multivariados registran menor error que los principales modelos univariados y, en algunos casos, mejores resultados direccionales. Sin embargo, esta conclusión corresponde a una sola ventana temporal de validación. El análisis debe complementarse con Rolling-Origin, que permite revisar si el desempeño se mantiene al cambiar el origen del pronóstico.

Figura 9

Pronósticos Hold-Out de mejores modelos multivariados por familia frente al retorno observado de BTC-USD



Nota: elaboración propia con base en datos OHLCV de BTC-USD obtenidos de Yahoo Finance; modelos multivariados estimados en Python.

La prueba de Diebold-Mariano presenta valores-p inferiores a 0.05 para varios modelos seleccionados frente al benchmark Naive. Esto indica diferencias estadísticamente significativas de precisión en esas comparaciones puntuales. La lectura final combina esta evidencia con MASE, las métricas direccionales y los resultados Rolling-Origin, evitando basar la evaluación en un único corte Hold-Out.

9.2. Resultados Rolling-Origin $h=1$

La validación Rolling-Origin con horizonte ($h = 1$) evalúa la capacidad de los modelos para pronosticar el retorno logarítmico de BTC-USD del día siguiente. A diferencia del esquema Hold-Out, el origen del pronóstico se actualiza de forma progresiva y genera varios cortes temporales de evaluación. Esto permite revisar el desempeño de los modelos en distintos momentos del periodo de validación, en lugar de depender de una sola ventana fija (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025).

El pronóstico diario representa un escenario exigente. Los retornos de Bitcoin presentan alta variabilidad, poca autocorrelación lineal y movimientos abruptos, por lo que resulta necesario revisar tanto el error de magnitud como la capacidad para anticipar la dirección del retorno.

Los resultados se presentan por separado para los enfoques univariado y multivariado. En ambos casos, los modelos se ordenan según MASE, que permite comparar su error relativo frente a un modelo ingenuo de referencia. MAE, RMSE y sMAPE complementan la evaluación de la magnitud del error. También se reportan Accuracy, Hit Ratio, F1-score, MCC y AUC cuando estas métricas están disponibles en las salidas consolidadas (Hyndman & Koehler, 2006)..

Tabla 8

Resultados Rolling-Origin $h = 1$ de modelos univariados ordenados por MASE

Modelo	Familia	MAE	RMSE	sMAPE (%)	MASE	Hit Ratio	Precisión	Recall	F1	DM Stat.	DM p-value
MLP_RO	Machine Learning	0.017	0.021	166.577	0.463	0.483	0.429	0.321	0.367	-3.208	0.001
AutoARIMA_MLP_RO	Híbrido	0.017	0.022	168.086	0.464	0.483	0.429	0.321	0.367	-3.221	0.001
AutoARIMA_RO	Clásico	0.017	0.022	181.080	0.462	0.483	0.455	0.536	0.492	-3.340	0.001
NeuralProphet_RO	Aditivo	0.017	0.022	161.646	0.465	0.533	0.500	0.536	0.517	-3.263	0.001
GAM_RO	Aditivo	0.017	0.022	173.881	0.466	0.483	0.385	0.179	0.244	-2.911	0.004
GRU_RO	Deep Learning	0.019	0.025	158.805	0.535	0.483	0.440	0.393	0.415	-0.919	0.358

Nota. La prueba Diebold-Mariano se calcula frente al modelo Naive como referencia, elaboración propia con base en los resultados obtenidos en Python.

La Tabla 8 muestra que MLP_RO obtuvo el menor MASE dentro del enfoque univariado bajo Rolling-Origin ($h = 1$). Le siguen AutoARIMA_MLP_RO, AutoARIMA_RO, NeuralProphet_RO y GAM_RO. El orden cambia frente a la validación Hold-Out, donde

AutoARIMA + XGBoost ocupó el primer lugar. Por tanto, el ranking de los modelos univariados depende del esquema de validación utilizado.

Aun con ese cambio, los valores de MAE, RMSE y MASE son cercanos entre los principales modelos. Ninguna familia presenta una ventaja amplia dentro del enfoque univariado. Rolling-Origin permite observar este comportamiento con mayor detalle porque evalúa los modelos en varios cortes temporales y no en una sola ventana final.

La lectura direccional también cambia. En Hold-Out, GRU y AutoARIMA registraron los valores más altos de Recall. En Rolling-Origin, NeuralProphet_RO ofrece el mejor equilibrio entre Hit Ratio y F1-score, mientras que AutoARIMA_RO mantiene un F1 competitivo. MLP_RO y AutoARIMA_MLP_RO reducen más el error de magnitud, pero obtienen menor Recall. El modelo que minimiza MASE no necesariamente es el que mejor identifica el signo del retorno.

La comparación entre las Tablas 6 y 8 muestra que los errores generales se mantienen en rangos similares, aunque las posiciones cambian al modificar el procedimiento de validación. Esto explica por qué una sola partición Hold-Out no es suficiente para seleccionar el mejor modelo univariado.

La prueba de Diebold-Mariano arroja valores-p inferiores a 0.05 para la mayoría de los modelos de la Tabla 8, excepto GRU_RO, frente al benchmark Naive. En esas comparaciones se rechaza la igualdad de precisión predictiva. Esta evidencia debe leerse junto con las métricas de error, los indicadores direccionales y los resultados Hold-Out; la significancia estadística no garantiza una ventaja práctica en todas las dimensiones del pronóstico.

La Tabla 9 muestra cambios en el orden de los modelos multivariados al pasar de Hold-Out a Rolling-Origin ($h = 1$). En Hold-Out, RandomForest_multivariado obtuvo el menor MASE, seguido por AutoARIMAX_MLP_Híbrido y GAM_multivariado. En Rolling-Origin, el primer

lugar corresponde a XGBoost_multivariado_RO, seguido por RandomForest_multivariado_RO, AutoARIMAX_MLP_Híbrido_RO y AutoARIMAX_RO.

Tabla 9

Resultados Rolling-Origin $h = 1$ de modelos multivariados ordenados por MASE

Modelo	Familia	MAE	RMSE	sMAPE (%)	MASE	Accuracy	F1-score	MCC	AUC
XGBoost_multivariado_RO	Machine Learning multivariado	0.014	0.014	129.440	0.398	0.633	0.267	0.000	—
RandomForest_multivariado_RO	Machine Learning multivariado	0.015	0.015	130.700	0.423	0.550	0.167	0.000	—
AutoARIMAX_MLP_Híbrido_RO	Híbrido multivariado	0.016	0.016	143.863	0.455	0.583	0.200	0.000	—
AutoARIMAX_RO	Clásico multivariado	0.016	0.016	145.272	0.457	0.600	0.217	0.000	—
GAM_multivariado_RO	Aditivo multivariado	0.017	0.017	120.402	0.468	0.667	0.300	0.000	—
LSTM_multivariado_RO	Deep Learning multivariado	0.018	0.018	165.506	0.712	0.467	0.050	0.000	—

Nota. Los valores se presentan con tres decimales. El símbolo — indica que la métrica AUC no fue calculada para ese modelo en la salida consolidada, elaboración propia con base en los resultados obtenidos en Python.

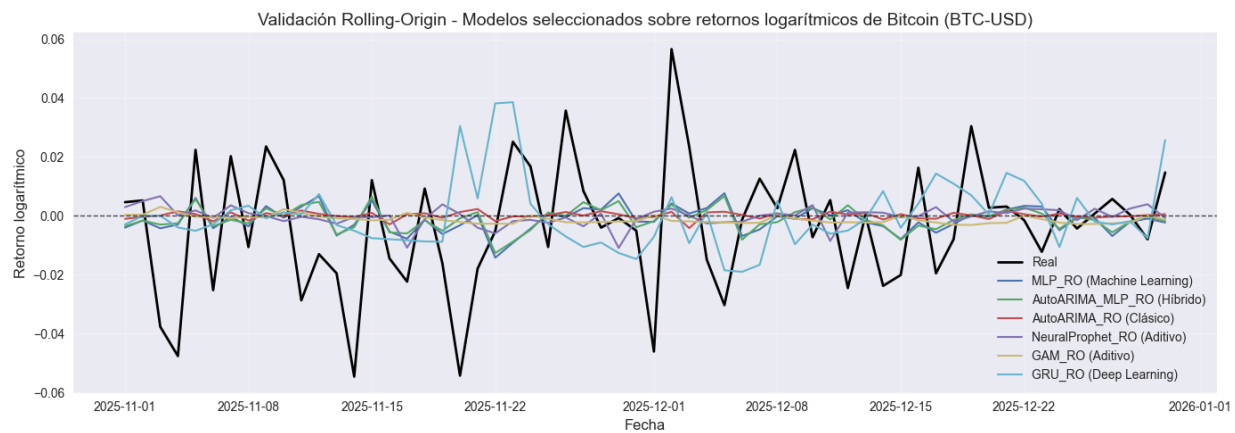
Aunque el ranking cambia, los modelos multivariados de machine learning permanecen en las primeras posiciones en ambos esquemas. Random Forest lidera la evaluación Hold-Out y XGBoost lidera Rolling-Origin. Los modelos basados en árboles aprovechan de manera consistente la información adicional de volumen, volatilidad e indicadores técnicos.

La diferencia frente a los modelos univariados es clara en el horizonte diario. Los valores de MASE de la Tabla 9 son menores que los obtenidos por los modelos univariados en la Tabla 8. En los cortes Rolling-Origin evaluados, las variables exógenas se asocian con una reducción del error de pronóstico. Sin embargo, los cambios de posición entre la Tabla 7 y la Tabla 9 muestran que una sola ventana Hold-Out no basta para valorar completamente el desempeño de un modelo.

XGBoost_multivariado_RO registra el menor MASE, MAE y RMSE de la Tabla 9. Este resultado lo convierte en la mejor alternativa para el pronóstico diario cuando el criterio principal es la precisión en magnitud. RandomForest_multivariado_RO también mantiene un desempeño competitivo, en línea con su primer lugar en la evaluación Hold-Out.

La clasificación direccional presenta una lectura distinta. En Hold-Out, GAM_multivariado alcanzó los mejores valores de Accuracy, F1-score, MCC y AUC. En Rolling-Origin ($h = 1$), GAM_multivariado_RO conserva el mayor Accuracy, pero no se encuentra entre los modelos con menor MASE. Por tanto, el modelo más preciso en magnitud no necesariamente es el que mejor anticipa el signo del retorno.

La comparación entre las Tablas 7 y 9 respalda el uso de Rolling-Origin como complemento del Hold-Out. El enfoque multivariado mantiene mejores resultados de error en el horizonte diario, especialmente con XGBoost y Random Forest. A la vez, la variación del ranking entre esquemas confirma que la selección final debe apoyarse en varios cortes temporales y no únicamente en una partición fija..

Figura 10*Comparación Rolling-Origin $h=1$ de modelos univariados seleccionados*

Nota: elaboración propia con base en datos OHLCV de BTC-USD obtenidos de Yahoo Finance; modelos estimados en Python.

La Figura 10 muestra los pronósticos univariados obtenidos con Rolling-Origin ($h = 1$). Frente a los resultados Hold-Out de la Figura 8, las predicciones presentan mayor variación porque el origen del pronóstico se actualiza en cada corte y los modelos incorporan la información más reciente disponible.

En la Figura 8, los pronósticos Hold-Out tienden a mantenerse alrededor de cero y responden poco a los cambios del retorno observado. En Rolling-Origin, algunos modelos reaccionan mejor ante variaciones locales de la serie. Esta diferencia no implica que anticipen todos los movimientos, pero permite evaluarlos bajo un proceso de actualización secuencial más cercano al uso real de un modelo de pronóstico.

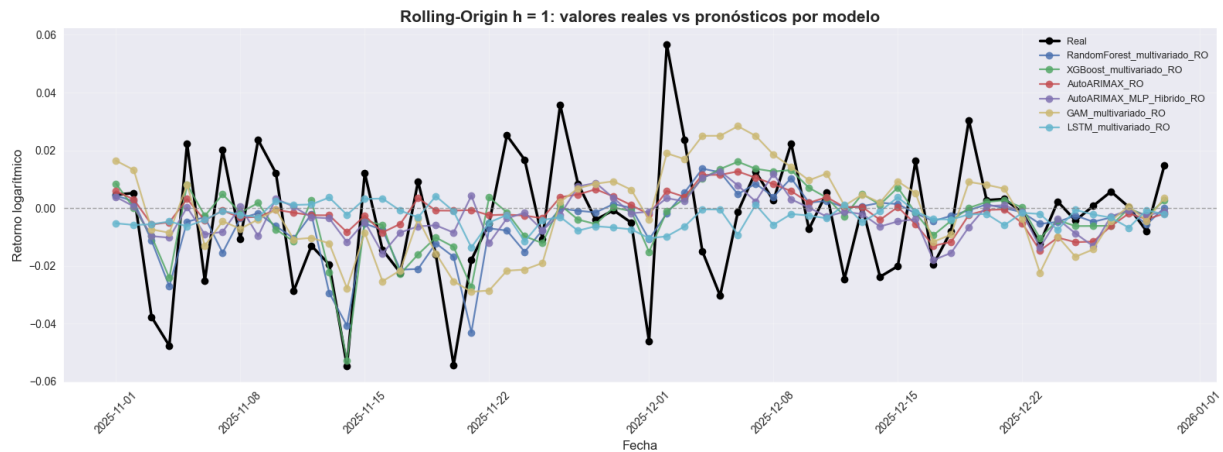
Los picos positivos y negativos más fuertes siguen sin ser capturados completamente. Esa limitación es esperable en modelos que utilizan solo los rezagos del retorno logarítmico, sin información adicional sobre volumen, volatilidad o indicadores técnicos.

La comparación entre las Figuras 8 y 10 muestra la utilidad de Rolling-Origin para examinar la respuesta de los modelos en distintos momentos del periodo de validación. También

deja claro que la actualización temporal, por sí sola, no resuelve la dificultad de anticipar movimientos abruptos de Bitcoin. Por ello, estos resultados se contrastan posteriormente con modelos multivariados.

Figura 11

Comparación Rolling-Origin $h=1$ de modelos multivariados seleccionados



Nota: elaboración propia con base en datos OHLCV de BTC-USD obtenidos de Yahoo Finance; modelos estimados en Python.

La Figura 11 presenta los pronósticos multivariados obtenidos con Rolling-Origin ($h = 1$), comparados con el retorno observado de BTC-USD. Frente a la Figura 9, correspondiente al esquema Hold-Out, los modelos responden de manera más dinámica a los cambios locales de la serie. Esto ocurre porque el origen del pronóstico se actualiza en cada corte y los modelos utilizan la información disponible hasta ese momento.

Los modelos multivariados ya mostraban una mayor variación que los univariados en la validación Hold-Out. En Rolling-Origin, esa respuesta se observa durante todo el periodo de evaluación, debido a la actualización secuencial de los modelos y de las variables de volumen, volatilidad e indicadores técnicos.

Algunos pronósticos se acercan mejor a los movimientos de menor amplitud del retorno observado. Los cambios extremos, positivos o negativos, continúan siendo difíciles de anticipar. La inclusión de variables exógenas mejora la información disponible, pero no elimina la incertidumbre propia del pronóstico diario de Bitcoin.

La comparación entre las Figuras 9 y 11 muestra por qué Rolling-Origin complementa la evaluación Hold-Out. Mientras Hold-Out resume el desempeño en una sola ventana, Rolling-Origin permite revisar cómo responden los modelos en varios cortes temporales. La figura respalda los resultados de la Tabla 9: los modelos multivariados se adaptan mejor a la información reciente, aunque mantienen limitaciones frente a los movimientos más abruptos del mercado.

9.3. Resultados Rolling-Origin $h=7$ multivariado

La validación Rolling-Origin para horizonte $h=7$ se desarrolló únicamente para el enfoque multivariado, con el propósito de evaluar la capacidad de los modelos para pronosticar el retorno logarítmico de BTC-USD en un horizonte semanal de corto plazo. A diferencia del horizonte diario $h=1$, este escenario incrementa la dificultad del pronóstico, debido a que el modelo debe anticipar el comportamiento futuro del retorno en un periodo más amplio, donde la incertidumbre acumulada tiende a ser mayor.

En este caso, los modelos multivariados incorporan información exógena asociada al volumen, la volatilidad, el momentum, la tendencia y el flujo de mercado. Esta estructura permite evaluar si las variables técnicas mantienen utilidad predictiva cuando el horizonte de pronóstico se amplía de un día a siete días. La evaluación se realizó bajo el esquema Rolling-Origin, respetando el orden temporal de la serie y actualizando progresivamente el origen del pronóstico, con el fin de simular un proceso de predicción más cercano a condiciones reales de mercado (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025).

Para conservar coherencia con los análisis anteriores, los modelos se ordenan principalmente por MASE, métrica que permite comparar el desempeño relativo frente a un modelo ingenuo de referencia. De manera complementaria, se consideran métricas como MAE, RMSE y sMAPE para evaluar la magnitud del error, así como métricas direccionales cuando se encuentran disponibles.

Tabla 10

Resultados Rolling-Origin $h=7$ de modelos multivariados ordenados por MASE

Modelo	Familia	MAE	RMSE	sMAPE (%)	MASE	Accuracy	F1-score	MCC	AUC
XGBoost RO $h=7$	Machine Learning multivariado	0.015	0.018	131.549	0.411	0.608	0.442	0.140	0.655
Random Forest RO $h=7$	Machine Learning multivariado	0.016	0.019	132.878	0.440	0.537	0.264	- 0.026	0.576
AutoARIMAX + MLP RO $h=7$	Híbrido multivariado	0.016	0.020	148.981	0.464	0.545	0.283	0.076	0.681
AutoARIMAX RO $h=7$	Clásico multivariado	0.017	0.020	148.990	0.467	0.624	0.488	0.272	0.698
GAM RO $h=7$	Aditivo multivariado	0.018	0.021	127.907	0.503	0.622	0.438	0.197	0.653
LSTM RO $h=7$	Deep Learning multivariado	0.018	0.022	162.978	0.715	0.508	0.213	- 0.017	0.560

Nota. Los valores se presentan con tres decimales. El horizonte $h = 7$ corresponde a la validación Rolling-Origin semanal del enfoque multivariado, elaboración propia con base en los resultados obtenidos en Python.

La Tabla 10 muestra el desempeño de los modelos multivariados en el horizonte semanal ($h = 7$). XGBoost multivariado mantiene el primer lugar por MASE, al igual que en Rolling-Origin ($h = 1$). En el horizonte diario registró un MASE de 0.398 y, en el semanal, obtuvo 0.411. Es el modelo que conserva el mejor desempeño relativo al ampliar el horizonte de uno a siete días.

El error aumenta en varios modelos al pasar al horizonte semanal. El RMSE de XGBoost pasa de 0.014 a 0.018, mientras que el de Random Forest aumenta de 0.015 a 0.019. Este resultado es esperable: anticipar el retorno de una semana implica mayor incertidumbre que pronosticar el día siguiente.

A pesar de ese incremento, el orden general se mantiene. XGBoost y Random Forest ocupan las primeras posiciones en las Tablas 9 y 10. Los modelos basados en árboles conservan buen desempeño cuando se incorporan volumen, volatilidad e indicadores técnicos, incluso en el horizonte semanal.

El modelo híbrido AutoARIMAX + MLP también se mantiene competitivo. Ocupa el tercer lugar en ambos horizontes, con MASE de 0.455 en ($h = 1$) y 0.464 en ($h = 7$). AutoARIMAX presenta un comportamiento parecido: su MASE pasa de 0.457 a 0.467 y permanece cerca del modelo híbrido. Estos resultados muestran que los modelos con regresores exógenos siguen siendo alternativas válidas, aunque no alcanzan el nivel de error de XGBoost y Random Forest.

GAM multivariado presenta una lectura diferente. En ($h = 1$) obtuvo un MASE de 0.468 y el mayor Accuracy de la Tabla 9. En ($h = 7$), su MASE aumenta a 0.503, pero conserva un Accuracy de 0.622. El modelo pierde posiciones al medir error de magnitud, pero mantiene una capacidad direccional competitiva. Error y dirección no siempre conducen al mismo ranking.

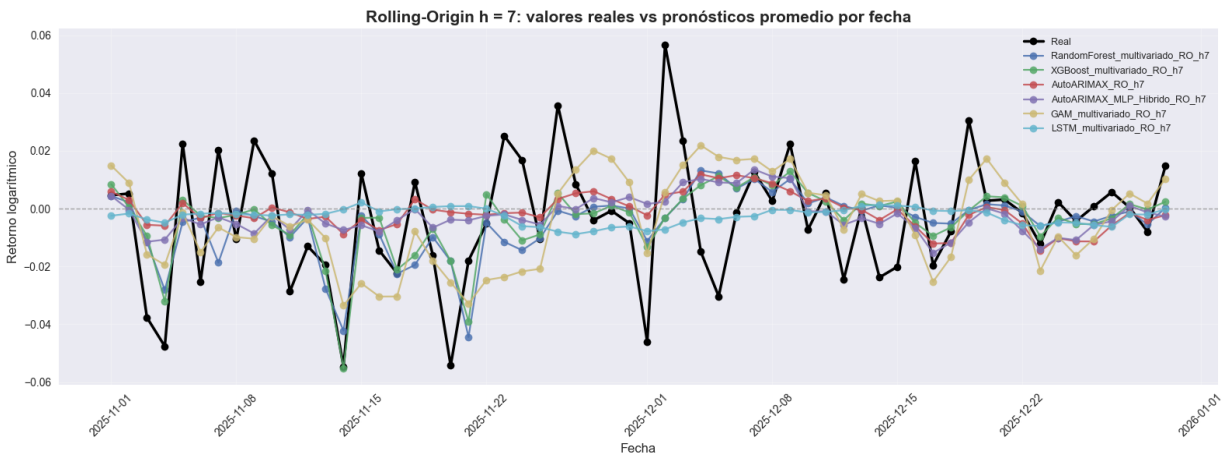
LSTM multivariado registra el desempeño más débil en ambos horizontes. Su MASE pasa de 0.712 en ($h = 1$) a 0.715 en ($h = 7$). Bajo la configuración utilizada, la red LSTM no aprovecha la información multivariada con la misma eficiencia que los modelos basados en árboles, los híbridos o los modelos clásicos con regresores.

La comparación entre las Tablas 9 y 10 muestra que el enfoque multivariado mantiene resultados competitivos al pasar del horizonte diario al semanal, especialmente con XGBoost y

Random Forest. La pérdida de precisión al ampliar el horizonte es normal; por ello, el criterio relevante es identificar cuáles modelos conservan mejor desempeño relativo. Bajo Rolling-Origin, XGBoost multivariado es la alternativa más estable entre los modelos evaluados.

Figura 12

Pronósticos Rolling-Origin $h=7$ de modelos multivariados seleccionados frente al retorno observado de BTC-USD



Nota: elaboración propia con base en datos OHLCV de BTC-USD obtenidos de Yahoo Finance; modelos estimados en Python.

La Figura 12 presenta los pronósticos multivariados bajo Rolling-Origin para el horizonte semanal ($h = 7$). Frente a la Figura 11, correspondiente al horizonte diario, las trayectorias son más suavizadas y reaccionan menos a los cambios puntuales del retorno observado. Al ampliar el horizonte de pronóstico aumenta la incertidumbre y resulta más difícil anticipar choques abruptos del mercado.

En el horizonte diario ($h = 1$), las variables recientes de volumen, volatilidad e indicadores técnicos tienen una relación más inmediata con el pronóstico. Por eso, en la Figura 11 algunos modelos responden mejor a movimientos locales de la serie. En cambio, los pronósticos semanales de la Figura 12 capturan una dinámica más general, pero no reproducen completamente los picos positivos y negativos del retorno.

XGBoost y Random Forest mantienen trayectorias relativamente cercanas al comportamiento general de la serie en ambos horizontes. Esta lectura visual coincide con las Tablas 9 y 10, donde los dos modelos se ubican entre los menores valores de MASE tanto para ($h = 1$) como para ($h = 7$).

La gráfica debe interpretarse junto con los resultados cuantitativos. Un modelo puede reaccionar poco ante un movimiento puntual y aun así obtener un desempeño competitivo en métricas agregadas. La Tabla 10 permite complementar esa lectura al comparar el error promedio de cada modelo.

La comparación entre las Figuras 11 y 12 muestra que el enfoque multivariado conserva capacidad predictiva al pasar del horizonte diario al semanal, aunque pierde sensibilidad ante los movimientos extremos. El horizonte ($h = 1$) mide la respuesta inmediata del modelo; el horizonte ($h = 7$) permite evaluar qué alternativas mantienen mejor desempeño relativo en un periodo más amplio.

9.4. Comparación por familias de modelos

La comparación por familias de modelos permite integrar los resultados obtenidos en los distintos esquemas de validación y evaluar el desempeño de los enfoques metodológicos implementados de manera conjunta. En esta sección no se busca analizar cada modelo de forma aislada, sino identificar qué familias presentan mayor capacidad predictiva, mayor estabilidad temporal y mejor desempeño relativo frente a los modelos de referencia.

El análisis se desarrolla desde dos perspectivas complementarias. En primer lugar, se comparan los modelos según métricas de error, utilizando MASE como criterio principal, con el fin de evaluar qué enfoques reducen en mayor medida la magnitud del error de pronóstico. En segundo lugar, se analiza el desempeño direccional mediante métricas como Accuracy, F1-score,

MCC y AUC, con el propósito de determinar qué modelos logran anticipar mejor el signo del retorno logarítmico de BTC-USD.

Esta doble evaluación es necesaria porque el modelo con menor error de magnitud no necesariamente coincide con el modelo que mejor predice la dirección del movimiento. Por tanto, la comparación por familias permite valorar de forma más completa el desempeño predictivo, diferenciando entre precisión numérica, capacidad direccional y estabilidad bajo distintos esquemas de validación temporal.

9.4.1. Comparación por métricas de error

La comparación por métricas de error permite identificar qué familias de modelos presentan mejor desempeño predictivo en términos de magnitud del error. Para este análisis se utiliza MASE como criterio principal, debido a que permite comparar el desempeño relativo de cada modelo frente a un benchmark ingenuo y facilita la evaluación entre modelos de distinta naturaleza (Hyndman & Koehler, 2006; Hyndman & Athanasopoulos, 2021). En el contexto de series financieras, esta comparación resulta especialmente relevante, ya que los retornos de Bitcoin suelen presentar alta volatilidad, baja autocorrelación lineal y movimientos extremos, por lo que la evaluación fuera de muestra debe realizarse bajo esquemas temporales rigurosos (Shumway & Stoffer, 2025).

Tabla 11

Comparación de mejores modelos por familia y esquema de validación

Esquema	Familia destacada	Mejor modelo	MASE	Comentario principal
Rolling-Origin $h = 1$ multivariado	Machine Learning	XGBoost multivariado_RO	0.398	Mejor resultado diario bajo validación dinámica.

Esquema	Familia destacada	Mejor modelo	MASE	Comentario principal
Hold-Out multivariado	Machine Learning	RandomForest multivariado	0.410	Mejor resultado general en Hold-Out.
Rolling-Origin $h = 7$ multivariado	Machine Learning	XGBoost multivariado_RO $h=7$	0.411	Mantiene estabilidad en horizonte semanal.
Hold-Out univariado	Híbrido	AutoARIMA XGBoost	0.459	Mejor desempeño dentro del enfoque univariado.
Rolling-Origin $h = 1$ univariado	Machine Learning	MLP_RO	0.463	Mejor desempeño diario dentro del enfoque univariado.

Nota: Los modelos se ordenan de mejor a peor desempeño según MASE; menor MASE indica mejor desempeño relativo frente al benchmark ingenuo.

La Tabla 11 reúne los mejores resultados por familia y esquema de validación. Los menores valores de MASE corresponden a modelos multivariados de machine learning. XGBoost multivariado_RO obtuvo un MASE de 0.398 en Rolling-Origin ($h = 1$) y de 0.411 en Rolling-Origin ($h = 7$). En la evaluación Hold-Out, RandomForest multivariado alcanzó un MASE de 0.410.

Estos resultados ubican a los modelos basados en árboles entre las alternativas más consistentes del estudio. Random Forest y XGBoost trabajan con múltiples predictores y pueden capturar interacciones no lineales entre los rezagos del retorno, el volumen, la volatilidad y los indicadores técnicos. Esa flexibilidad resulta útil en una serie como BTC-USD, donde las relaciones entre las variables no son necesariamente lineales (Hastie, Tibshirani, & Friedman, 2017; Chen & Guestrin, 2016).

La diferencia entre los enfoques univariado y multivariado también es visible en Rolling-Origin ($h = 1$). MLP_RO fue el mejor modelo univariado, con un MASE de 0.463, mientras que

XGBoost multivariado_RO obtuvo 0.398. En los cortes temporales evaluados, la incorporación de volumen, volatilidad, momentum y flujo de mercado se asocia con una reducción observada del error de pronóstico. Esta comparación es descriptiva y corresponde a los escenarios de validación analizados.

Los modelos híbridos también presentan resultados competitivos. En Hold-Out univariado, AutoARIMA + XGBoost obtuvo un MASE de 0.459, el mejor valor dentro de ese enfoque. Combinar una estructura autorregresiva con un algoritmo de aprendizaje automático puede aprovechar parte de la dinámica lineal del retorno y posibles patrones residuales. Aun así, los menores errores de la tabla pertenecen a los modelos multivariados basados en árboles.

El horizonte semanal mantiene el mismo patrón. XGBoost multivariado_RO ($h = 7$) conserva el primer lugar con un MASE de 0.411, cercano al resultado diario de 0.398. Aunque el pronóstico a siete días introduce mayor incertidumbre, el modelo sostiene un desempeño relativo favorable frente a las demás alternativas multivariadas.

Los modelos clásicos e híbridos cumplen un papel importante como referencias interpretables y como alternativas de comparación. Las arquitecturas de deep learning no aparecen entre los mejores resultados consolidados. Bajo la configuración empleada, LSTM y GRU no alcanzaron los niveles de error obtenidos por XGBoost y Random Forest. Esta lectura debe hacerse con cautela, porque el desempeño de las redes depende del tamaño de muestra, la longitud de las ventanas, el escalamiento y la selección de hiperparámetros (Hochreiter & Schmidhuber, 1997; McNally, Roche, & Caton, 2018).

9.4.2. Comparación por métricas direccionales

La comparación por métricas direccionales permite evaluar la capacidad de los modelos para anticipar correctamente el signo del retorno logarítmico de BTC-USD, es decir, si el

movimiento futuro del activo será positivo o negativo. A diferencia de las métricas de error, que miden la distancia entre el valor observado y el valor pronosticado, las métricas direccionales permiten analizar si el modelo acierta la dirección del movimiento, aspecto relevante en contextos financieros y de toma de decisiones.

Para este análisis se consideran métricas como Accuracy/Hit Ratio, F1-score, MCC y AUC. Aunque Accuracy permite medir el porcentaje global de aciertos direccionales, puede resultar insuficiente cuando existe desbalance entre movimientos positivos y negativos. Por esta razón, el F1-score se utiliza como criterio principal de ordenamiento, ya que resume el equilibrio entre Precision y Recall. De manera complementaria, el MCC permite evaluar la calidad global de la clasificación considerando aciertos y errores en ambas direcciones, mientras que el AUC mide la capacidad del modelo para discriminar entre retornos positivos y negativos cuando dicha métrica se encuentra disponible (Hastie, Tibshirani, & Friedman, 2017).

Tabla 12

Comparación de mejores modelos por desempeño direccional y esquema de validación.

Esquema	Familia destacada	Mejor modelo direccional	Accuracy / Hit Ratio	F1-score	MCC	AUC	Comentario principal
Hold-Out multivariado	Aditivo	GAM multivariado	0.733	0.714	0.464	0.786	Mejor desempeño direccional general en Hold-Out.
Hold-Out univariado	Deep Learning Clásico	GRU / AutoARIMA	0.467	0.636	—	—	Mejor desempeño direccional dentro del enfoque univariado Hold-Out.
Rolling-Origin h = 1 univariado	Aditivo	NeuralProphet-RO	0.533	0.517	—	—	Mejor equilibrio direccional dentro

Esquema	Familia destacada	Mejor modelo direccional	Accuracy / Hit Ratio	F1-score	MCC	AUC	Comentario principal
							del enfoque univariado diario.
Rolling-Origin h = 7 multivariado	Clásico multivariado	AutoARIMAX_ RO h=7	0.624	0.488	0.272	0.698	Mejor desempeño direccional multivariado en horizonte semanal.
Rolling-Origin h = 1 multivariado	Aditivo	GAM multivariado_R O	0.667	0.300	0.000	—	Mayor Accuracy dentro del enfoque multivariado diario.

Nota: Los modelos se ordenan de mejor a peor desempeño según F1-score; Accuracy/Hit Ratio, MCC y AUC se presentan como métricas complementarias. El símbolo — indica que la métrica no fue calculada o no se encuentra disponible en la salida consolidada, elaboración propia con base en los resultados obtenidos en Python.

La Tabla 12 presenta una lectura distinta a la obtenida con las métricas de error. El mejor resultado direccional general aparece en la validación Hold-Out multivariada: GAM multivariado alcanza un Accuracy de 0.733, un F1-score de 0.714, un MCC de 0.464 y un AUC de 0.786. Random Forest y XGBoost reducen más el error de magnitud, pero GAM identifica mejor el signo del retorno en esta ventana de evaluación.

En Hold-Out univariado, GRU y AutoARIMA obtienen el mayor F1-score de su grupo, con 0.636. Sin embargo, el Accuracy es de 0.467. Ese contraste obliga a revisar más de una métrica, porque un F1 alto puede estar influido por la distribución de clases y el comportamiento del Recall.

Para Rolling-Origin (h = 1) univariado, NeuralProphet_RO presenta el mejor equilibrio direccional, con Hit Ratio de 0.533 y F1-score de 0.517. Dentro de los modelos univariados

evaluados en distintos cortes temporales, es la alternativa que mejor balancea aciertos y precisión en las señales. Su desempeño, sin embargo, no alcanza el resultado direccional de GAM multivariado en Hold-Out.

En el horizonte semanal, AutoARIMAX_RO ($h = 7$) obtiene el mejor resultado direccional multivariado: Accuracy de 0.624, F1-score de 0.488, MCC de 0.272 y AUC de 0.698. Aunque no registra el menor MASE, conserva una capacidad competitiva para anticipar el signo del retorno a siete días.

GAM_multivariado_RO muestra una situación diferente en Rolling-Origin ($h = 1$). Tiene el mayor Accuracy del grupo, con 0.667, pero su F1-score es 0.300 y el MCC es 0.000. El resultado indica que el porcentaje global de aciertos no describe por sí solo una clasificación balanceada entre retornos positivos y negativos.

Los rankings por error y por dirección no coinciden. XGBoost y Random Forest destacan al ordenar por MASE; GAM multivariado y AutoARIMAX_RO ($h = 7$) sobresalen en la evaluación direccional. La selección final no puede depender de una sola métrica: debe considerar el objetivo del pronóstico, el error de magnitud y la capacidad para anticipar el signo del retorno.

Esta diferencia es esperable en series financieras. Un modelo puede aproximar bien el valor promedio del retorno, pero equivocarse con frecuencia en su dirección. Otro puede acertar el signo en varios casos, aunque presente un error de magnitud mayor. Por eso, la evaluación combina métricas de error, indicadores direccionales y validación temporal fuera de muestra (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025).

9.5. Comparación univariado vs. Multivariado

La comparación entre los enfoques univariado y multivariado permite evaluar de manera directa si la incorporación de variables exógenas aporta mejoras reales frente a modelos basados

únicamente en la historia del retorno logarítmico de BTC-USD. Mientras el enfoque univariado utiliza rezagos de la propia variable objetivo, el enfoque multivariado incorpora información adicional del mercado, como volumen, volatilidad, momentum, tendencia y flujo de mercado.

Este análisis se desarrolla desde dos perspectivas complementarias: primero, mediante métricas de error, con énfasis en el MASE como medida principal de desempeño relativo; y segundo, mediante métricas direccionales, orientadas a evaluar la capacidad de los modelos para anticipar el signo del retorno. Esta separación permite determinar si el enfoque multivariado mejora la precisión en magnitud, la direccionalidad del mercado o ambos criterios de forma simultánea.

9.5.1. Comparación por métricas de error

La comparación entre los enfoques univariado y multivariado constituye uno de los análisis centrales del proyecto, ya que permite evaluar si la incorporación de variables exógenas mejora el pronóstico del retorno logarítmico de BTC-USD frente a modelos basados únicamente en la historia del retorno. En el enfoque univariado, los modelos utilizan rezagos de la variable objetivo; en contraste, el enfoque multivariado incorpora información adicional asociada al volumen, la volatilidad, el momentum, la tendencia y el flujo de mercado.

Desde el punto de vista metodológico, esta comparación es relevante porque permite contrastar si el aumento en la complejidad del conjunto de variables se traduce en una mejora real del desempeño fuera de muestra. En problemas de pronóstico financiero, los modelos más complejos solo se justifican si logran superar de manera consistente a modelos más simples o parsimoniosos bajo esquemas de validación temporal rigurosos (Hyndman & Athanasopoulos, 2021; Hyndman & Koehler, 2006). En este caso, la comparación se realiza utilizando MASE como métrica principal, dado que permite evaluar el error relativo frente a un modelo ingenuo de referencia.

Tabla 13

Comparación entre mejores modelos univariados y multivariados según esquema de validación.

Esquema	Mejor univariado	MASE univariado	Mejor multivariado	MASE multivariado	Mejor enfoque
Hold-Out	AutoARIMA + XGBoost	0.459	RandomForest multivariado	0.410	Multivariado
Rolling-Origin $h = 1$	MLP_RO	0.463	XGBoost multivariado_RO	0.398	Multivariado
Rolling-Origin $h = 7$	No evaluado	—	XGBoost multivariado_RO $h = 7$	0.411	Multivariado

Nota. El horizonte Rolling-Origin $h = 7$ fue evaluado únicamente para el enfoque multivariado; por tanto, no se reporta comparación directa con modelos univariados para dicho horizonte. Menor MASE indica mejor desempeño relativo frente al benchmark ingenuo, elaboración propia con base en los resultados obtenidos en Python.

La Tabla 13 presenta la comparación directa entre los mejores modelos univariados y multivariados en los esquemas donde ambos enfoques fueron evaluados. En la validación Hold-Out, AutoARIMA + XGBoost fue el mejor modelo univariado, con un MASE de 0.459. RandomForest multivariado obtuvo el menor MASE del enfoque multivariado, con 0.410. En esta ventana de validación, la incorporación de variables exógenas se asocia con una reducción observada del error.

El mismo patrón aparece en Rolling-Origin ($h = 1$). MLP_RO fue el mejor modelo univariado, con un MASE de 0.463, mientras que XGBoost_multivariado_RO alcanzó un MASE de 0.398. Esta diferencia se obtiene a partir de múltiples cortes temporales, por lo que complementa la lectura de la evaluación Hold-Out y permite revisar el desempeño bajo actualizaciones sucesivas del origen del pronóstico.

Para Rolling-Origin ($h = 7$) no se efectuó una comparación directa con modelos univariados, dado que este horizonte se evaluó únicamente con modelos multivariados. En este escenario, XGBoost_multivariado_RO ($h = 7$) obtuvo un MASE de 0.411. El resultado muestra que el modelo mantiene un nivel de error competitivo al ampliar el horizonte de uno a siete días, aunque no permite establecer una comparación directa entre enfoques.

Los resultados descritos constituyen evidencia empírica fuera de muestra, limitada al periodo, los modelos y los esquemas de validación considerados. En las comparaciones directas realizadas, los modelos multivariados presentaron menores valores de MASE que los mejores modelos univariados. El volumen, la volatilidad y los indicadores técnicos aportan información adicional que los modelos flexibles pueden aprovechar en el proceso de pronóstico. Esta lectura es consistente con estudios que reconocen relaciones dinámicas entre precio, volumen, liquidez y volatilidad en mercados de criptomonedas (Bouri et al., 2020; Vidal-Tomás & Ibáñez, 2023; Anastasiou, Ballis, & Drakos, 2023).

La reducción del error no implica automáticamente un mejor desempeño direccional. Como se observó en la Tabla 12, el modelo con menor MASE no siempre coincide con el que mejor anticipa el signo del retorno. La selección final debe considerar conjuntamente el error de pronóstico, las métricas direccionales y la estabilidad observada bajo los distintos esquemas de validación.

En síntesis, la Tabla 13 muestra que el enfoque multivariado obtuvo menores errores de magnitud en los escenarios con comparación directa, especialmente con Random Forest y XGBoost. Este resultado respalda la inclusión del feature set de volumen, volatilidad e indicadores técnicos dentro del diseño predictivo del proyecto.

9.5.2. Comparación por métricas direccionales

La comparación por métricas direccionales permite evaluar si los modelos logran anticipar correctamente el signo del retorno logarítmico de BTC-USD, es decir, si el movimiento futuro del mercado será positivo o negativo. Este análisis complementa la comparación por métricas de error, ya que un modelo puede presentar bajo error de magnitud, pero no necesariamente acertar la dirección del movimiento.

Para este análisis se utiliza principalmente el F1-score, debido a que resume el equilibrio entre Precision y Recall. Esta métrica resulta más informativa que observar únicamente el Accuracy, porque permite evaluar si el modelo no solo acierta una proporción general de casos, sino si identifica adecuadamente los movimientos relevantes del mercado. De manera complementaria, se consideran Accuracy/Hit Ratio, MCC y AUC, cuando estas métricas se encuentran disponibles.

La Tabla 14 compara el desempeño direccional de los enfoques univariado y multivariado en los esquemas donde existe comparación directa. En la validación Hold-Out, GAM multivariado obtiene el mejor resultado: Accuracy de 0.733, F1-score de 0.714, MCC de 0.464 y AUC de 0.786. El mejor resultado univariado corresponde a GRU y AutoARIMA, con F1-score de 0.636. En esta ventana de validación, el enfoque multivariado muestra una mejor capacidad para anticipar el signo del retorno.

El resultado cambia en Rolling-Origin ($h = 1$). NeuralProphet_RO es el mejor modelo univariado según F1-score, con Hit Ratio de 0.533 y F1-score de 0.517. GAM_multivariado_RO alcanza un Accuracy superior, de 0.667, pero su F1-score es de 0.300. En otras palabras, el modelo multivariado acierta una mayor proporción total de casos, mientras que NeuralProphet_RO logra

un mejor balance entre Precision y Recall. Bajo el criterio principal de F1-score, el mejor resultado direccional en Rolling-Origin ($h = 1$) corresponde al enfoque univariado.

Tabla 14

Comparación entre mejores modelos univariados y multivariados según desempeño direccional

Esquema	Mejor univariado	Acc./Hit uni.	F1 uni.	Mejor multivariado	Acc. mult.	F1 mult.	MCC mult.	AUC mult.	Mejor enfoque
Hold-Out	GRU / AutoARIMA	0.467	0.63	GAM multivariado	0.73	0.71	0.46	0.78	Multivariado
Rolling-Origin $h = 1$	NeuralProphet_RO	0.533	0.51	GAM multivariado_RO	0.66	0.30	0.00	—	Univariado
Rolling-Origin $h = 7$	No evaluado	—	—	AutoARIMAX_RO $h = 7$	0.62	0.48	0.27	0.69	Multivariado

Nota. La comparación direccional se ordena principalmente por F1-score, debido a que esta métrica resume el equilibrio entre Precision y Recall. Accuracy/Hit Ratio, MCC y AUC se presentan como métricas complementarias. El horizonte Rolling-Origin $h = 7$ fue evaluado únicamente para el enfoque multivariado; por tanto, no existe comparación directa con modelos univariados en ese horizonte. El símbolo — indica que la métrica no fue calculada o no se encuentra disponible en la salida consolidada, elaboración propia con base en los resultados obtenidos en Python.

Este contraste muestra por qué Accuracy no debe analizarse de forma aislada. En el caso de GAM_multivariado_RO, el Accuracy es alto, pero el F1-score y el MCC indican una clasificación menos equilibrada entre retornos positivos y negativos. La interpretación direccional requiere revisar varias métricas en conjunto.

En Rolling-Origin ($h = 7$) no fue posible comparar directamente los dos enfoques, porque solo se evaluaron modelos multivariados. AutoARIMAX_RO ($h = 7$) obtuvo el mejor resultado direccional en ese horizonte, con Accuracy de 0.624, F1-score de 0.488, MCC de 0.272 y AUC de 0.698. Aunque no tuvo el menor MASE, logró una capacidad competitiva para anticipar el signo del retorno semanal.

La direccionalidad no presenta el mismo patrón estable observado en las métricas de error. En Hold-Out, GAM multivariado lidera la clasificación direccional. En Rolling-Origin ($h = 1$), NeuralProphet_RO obtiene el mejor equilibrio de métricas. El desempeño depende del esquema de validación, del horizonte y del indicador utilizado.

Por esta razón, los resultados direccionales deben complementar la comparación por MASE, no reemplazarla. Para decisiones que dependan de anticipar el signo diario del retorno, Rolling-Origin ofrece una referencia más exigente porque evalúa varios cortes temporales. La selección final debe equilibrar precisión en magnitud, capacidad direccional y estabilidad temporal.

9.5.3. Contraste empírico descriptivo entre enfoques univariado y multivariado

Con el propósito de comparar el desempeño predictivo de los enfoques univariado y multivariado, se realizó un contraste empírico descriptivo entre el mejor modelo univariado y el mejor modelo multivariado, utilizando el MASE como métrica principal de evaluación. Esta comparación permite examinar si la incorporación de variables exógenas asociadas al volumen, la volatilidad y los indicadores técnicos se relaciona con una reducción del error predictivo frente a modelos basados únicamente en la historia del retorno logarítmico de Bitcoin.

La Tabla 15 presenta la comparación entre los mejores modelos univariados y multivariados en tres escenarios: comparación global, validación Hold-Out y validación Rolling-Origin. La diferencia se calcula como $(MASE_U - MASE_M)$ de modo que valores positivos indican menor error del modelo multivariado frente al modelo univariado. La mejora porcentual corresponde a la reducción relativa del MASE del modelo multivariado respecto al modelo univariado.

Tabla 15

Contraste empírico descriptivo entre el mejor modelo univariado y el mejor modelo multivariado

Comparación	Mejor modelo univariado	MASE univ.	Mejor modelo multivariado	MASE multiv.	Dif. MASE U - M	Mejora multiv. (%)	Mejor enfoque
Global	AutoARIMA_XGBoost	0.4592	XGBoost_multivariado_RO	0.3981	0.0611	13.31	Multivariado
Hold-Out	AutoARIMA_XGBoost	0.4592	RandomForest_multivariado	0.4096	0.0496	10.80	Multivariado
Rolling- -Origin RO	AutoARIMA_RO	0.4625	XGBoost_multivariado_RO	0.3981	0.0644	13.92	Multivariado

Nota. La tabla presenta una comparación empírica descriptiva entre el mejor modelo univariado y el mejor modelo multivariado, utilizando el MASE como métrica principal. La diferencia se calcula como $(MASE_U - MASE_M)$ por lo que valores positivos indican menor error del modelo multivariado frente al univariado. La mejora porcentual corresponde a la reducción relativa del MASE del modelo multivariado respecto al modelo univariado. Este análisis corresponde a una comparación descriptiva basada en los resultados obtenidos bajo los esquemas de validación y no permite inferir significancia estadística agregada entre enfoques, sino una contrastación descriptiva basada en los resultados obtenidos en los esquemas de validación.

Los resultados muestran que, en los escenarios evaluados, los modelos multivariados presentaron menor MASE que los modelos univariados de comparación. A nivel global, XGBoost_multivariado_RO registró menor error que AutoARIMA_XGBoost. De manera similar, en Hold-Out, RandomForest_multivariado presentó menor MASE que el mejor modelo univariado, mientras que en Rolling-Origin XGBoost_multivariado_RO mantuvo un menor error frente a AutoARIMA_RO.

Estos resultados sugieren que la incorporación de variables exógenas asociadas al volumen, la volatilidad y los indicadores técnicos aportó información útil para el pronóstico del retorno logarítmico de BTC-USD en los escenarios analizados. La reducción observada del MASE se

mantiene en las comparaciones global, Hold-Out y Rolling-Origin, lo cual respalda la pertinencia metodológica del enfoque multivariado dentro del diseño de evaluación utilizado.

No obstante, los resultados deben interpretarse como evidencia empírica descriptiva fuera de muestra, limitada al periodo, los modelos y los esquemas de validación considerados. La comparación de los valores de MASE permite identificar diferencias observadas en el error predictivo, pero no constituye una prueba estadística formal de significancia agregada entre los enfoques univariado y multivariado.

9.6. Análisis de métricas direccionales

El análisis direccional evalúa si los modelos anticipan correctamente el signo del retorno logarítmico de BTC-USD: positivo o negativo. Esta revisión complementa las métricas de error, porque un modelo puede aproximar bien la magnitud del retorno y, al mismo tiempo, equivocarse en la dirección del movimiento.

Para este propósito se utilizaron Accuracy, Hit Ratio, Precision, Recall, F1-score, MCC y AUC. Accuracy y Hit Ratio indican la proporción total de aciertos. Sin embargo, pueden dar una visión incompleta cuando existe desbalance entre retornos positivos y negativos. El F1-score combina Precision y Recall; el MCC evalúa la calidad global de la clasificación; y el AUC mide la capacidad de distinguir entre ambas direcciones cuando la salida del modelo permite calcularla (Hastie, Tibshirani, & Friedman, 2017).

Los resultados direccionales no coinciden siempre con el ranking por MASE. XGBoost y Random Forest multivariados se ubicaron entre los mejores modelos por error de magnitud, mientras que algunos modelos aditivos y clásicos lograron mejores resultados al anticipar el signo del retorno. Se trata de dos dimensiones distintas del desempeño predictivo.

En la validación Hold-Out, GAM multivariado alcanzó el mejor resultado direccional general. Registró los valores más altos de Accuracy, F1-score, MCC y AUC dentro de la comparación. Este resultado corresponde a una sola ventana final de validación y debe interpretarse en ese contexto.

En Rolling-Origin ($h = 1$), NeuralProphet_RO presentó el mejor equilibrio direccional dentro de los modelos evaluados. GAM_multivariado_RO alcanzó un Accuracy relativamente alto, pero sus valores de F1-score y MCC fueron inferiores. El porcentaje global de aciertos, por sí solo, no garantiza una clasificación equilibrada entre retornos positivos y negativos.

Para Rolling-Origin ($h = 7$), evaluado solo en el enfoque multivariado, AutoARIMAX_RO ($h = 7$) obtuvo el mejor comportamiento direccional. No fue el modelo con menor MASE, pero sus métricas de Accuracy, F1-score, MCC y AUC muestran que los modelos clásicos con regresores exógenos pueden conservar información útil para anticipar el signo del retorno semanal.

La selección de un modelo no debe depender de una única métrica. Un modelo con bajo error puede ser apropiado cuando interesa aproximar la magnitud del retorno. En cambio, un modelo con mejor desempeño direccional puede ser más útil cuando la prioridad es anticipar el sentido del movimiento. También es necesario revisar si ese comportamiento se mantiene bajo Rolling-Origin, donde los modelos se evalúan en varios cortes temporales.

El análisis direccional matiza los resultados obtenidos mediante MASE. Los modelos multivariados presentan menores errores de magnitud en los escenarios evaluados, pero la predicción del signo es más sensible al horizonte, al esquema de validación y a la métrica utilizada. Por eso, la evaluación final combina precisión en magnitud, desempeño direccional y estabilidad temporal.

9.7. Prueba Diebold-Mariano

La prueba de Diebold-Mariano se utilizó como contraste estadístico complementario para evaluar si las diferencias observadas en la precisión predictiva entre los modelos implementados y el modelo de referencia Naive son estadísticamente significativas. A diferencia de las métricas descriptivas, como MAE, RMSE, sMAPE o MASE, esta prueba permite determinar si la mejora de un modelo frente al benchmark puede considerarse estadísticamente relevante o si podría atribuirse a variaciones aleatorias del periodo evaluado (Diebold & Mariano, 1995).

En este proyecto, el modelo Naive se tomó como referencia porque representa una línea base simple y parsimoniosa dentro del análisis de pronóstico. Bajo esta lógica, cada modelo evaluado se contrastó frente a Naive, con el fin de identificar si su desempeño predictivo ofrecía una mejora significativa. Esta comparación resulta importante porque, en problemas de series financieras, los modelos más complejos solo se justifican si logran superar de manera consistente a modelos simples bajo condiciones fuera de muestra (Hyndman & Athanasopoulos, 2021; Hyndman & Koehler, 2006).

La prueba se formula mediante las siguientes hipótesis:

Hipótesis Nula:

$$H_0: E(d_t) = 0$$

- No hay diferencia promedio en el error de pronóstico entre el modelo evaluado y el modelo Naive.

Hipótesis Alternativa:

$$H_1: E(d_t) \neq 0$$

- Sí existe una diferencia promedio en el error de pronóstico entre el modelo evaluado y el modelo Naive.

donde (d_t) representa la diferencia entre las pérdidas de pronóstico de dos modelos competidores. En este caso, la hipótesis nula (H_0) plantea que el modelo evaluado y el modelo

Naive tienen igual capacidad predictiva. Por su parte, la hipótesis alternativa (H_1) indica que existen diferencias estadísticamente significativas entre ambos modelos.

El criterio de decisión se estableció con un nivel de significancia del 5 %. Por tanto, cuando el valor-p de la prueba Diebold-Mariano es inferior a 0.05, se rechaza la hipótesis nula de igualdad de precisión predictiva. En consecuencia, se considera que el modelo evaluado presenta una diferencia estadísticamente significativa frente al modelo Naive. Si el valor-p es mayor o igual a 0.05, no se rechaza la hipótesis nula, por lo que no existe evidencia suficiente para afirmar que el modelo evaluado supera estadísticamente al benchmark.

Tabla 16

Síntesis de resultados de la prueba Diebold-Mariano frente al modelo Naive

Esquema de validación	Enfoque	Modelos con evidencia significativa frente a Naive	Modelos sin evidencia significativa	Interpretación
Hold-Out	Univariado	AutoARIMA + XGBoost; MLP óptimo; NeuralProphet; GAM; GRU; AutoARIMA	SNaive	La mayoría de modelos univariados supera estadísticamente al benchmark Naive.
Hold-Out	Multivariado	RandomForest; XGBoost; AutoARIMAX + MLP; AutoARIMAX + XGBoost; GAM; AutoARIMAX; SARIMAX; ARIMAX; Prophet con regresores; ETSX; LSTM; GRU; NeuralProphet	SNaive	Los modelos multivariados muestran evidencia amplia de mejora frente a Naive.

Esquema de validación	Enfoque	Modelos con evidencia significativa frente a Naive	Modelos sin evidencia significativa	Interpretación
Rolling-Origin $h = 1$	Univariado	MLP_RO; AutoARIMA_MLP_RO; AutoARIMA_RO; NeuralProphet_RO; GAM_RO	GRU_RO	La mayoría mantiene diferencias significativas bajo validación dinámica, excepto GRU_RO.
Rolling-Origin $h = 1$	Multivariado	No reportado en tabla consolidada con DM p-value	No reportado	La comparación se evaluó principalmente mediante métricas de error y direccionales.
Rolling-Origin $h = 7$	Multivariado	No reportado en tabla consolidada con DM p-value	No reportado	No se dispone de contraste DM consolidado para este horizonte.

Nota. Se considera evidencia estadísticamente significativa cuando el valor- p de la prueba Diebold-Mariano es inferior a 0.05. La comparación se realiza frente al modelo Naive como referencia; En los esquemas Rolling-Origin $h=1$ y $h=7$ multivariado no se reportó el contraste Diebold-Mariano en la salida consolidada, porque la significancia estadística frente al modelo Naive se evaluó principalmente en Hold-Out. En Rolling-Origin, el análisis se enfocó en la estabilidad temporal del desempeño mediante métricas de error y direccionales. Por tanto, la ausencia del valor- p en estos esquemas responde a una delimitación metodológica y no a una falta de evaluación del desempeño predictivo; Elaboración propia con base en los resultados obtenidos en Python.

La Tabla 16 resume los resultados de la prueba de Diebold-Mariano frente al modelo Naive.

En los esquemas donde se reportó la prueba, varios modelos presentan valores- p inferiores a 0.05, lo que permite rechazar la hipótesis nula de igualdad de precisión predictiva frente al benchmark.

En Hold-Out univariado, AutoARIMA + XGBoost, MLP óptimo, NeuralProphet, GAM, GRU y AutoARIMA muestran diferencias estadísticamente significativas frente a Naive. SNaive no presenta evidencia suficiente para rechazar la igualdad de precisión bajo este esquema.

En Hold-Out multivariado, la mayoría de los modelos evaluados registra diferencias significativas frente a Naive. Entre ellos se encuentran Random Forest, XGBoost, AutoARIMAX + MLP, AutoARIMAX + XGBoost, GAM, AutoARIMAX, SARIMAX, ARIMAX, Prophet con regresores, ETSX, LSTM, GRU y NeuralProphet. Esto indica que, en la ventana Hold-Out analizada, sus errores difieren de manera estadísticamente significativa de los obtenidos por el benchmark.

En Rolling-Origin ($h = 1$) univariado, MLP_RO, AutoARIMA_MLP_RO, AutoARIMA_RO, NeuralProphet_RO y GAM_RO también presentan valores-p inferiores a 0.05. GRU_RO no muestra evidencia estadística de diferencia frente a Naive en esta validación dinámica.

La prueba de Diebold-Mariano no se reportó en las salidas consolidadas de Rolling-Origin ($h = 1$) multivariado ni de Rolling-Origin ($h = 7$) multivariado. En esos esquemas, el análisis se concentró en las métricas de error, las métricas direccionales y la estabilidad observada entre los distintos cortes temporales.

Los resultados de Diebold-Mariano deben interpretarse junto con las demás métricas. Una diferencia significativa frente a Naive no implica que el modelo sea el mejor en todos los criterios. Algunos modelos univariados superan estadísticamente al benchmark, pero registran MASE menos favorables que los mejores modelos multivariados. Del mismo modo, un modelo puede reducir el error de magnitud sin ser la mejor alternativa para anticipar el signo del retorno.

La prueba confirma que varios modelos presentan diferencias de precisión frente a Naive, especialmente en Hold-Out y Rolling-Origin ($h = 1$) univariado. La selección final se basa en una lectura conjunta de MASE, desempeño direccional, estabilidad temporal y evidencia estadística frente al modelo de referencia.

9.8. Selección de mejores modelos

La selección de los mejores modelos se realizó a partir de una evaluación integral que consideró cuatro criterios principales: error de pronóstico, desempeño direccional, estabilidad temporal y evidencia estadística frente al modelo de referencia Naive. Esta aproximación permite evitar una selección basada en una única métrica, ya que en problemas de pronóstico financiero un modelo puede presentar bajo error de magnitud, pero no necesariamente ser el más adecuado para anticipar la dirección del mercado.

Desde el punto de vista metodológico, la selección final debe considerar el desempeño fuera de muestra y la capacidad de generalización de los modelos. En series financieras altamente volátiles, como el retorno logarítmico de BTC-USD, los modelos más complejos solo se justifican si muestran mejoras consistentes frente a modelos simples de referencia y si mantienen estabilidad bajo esquemas de validación temporal, como Hold-Out y Rolling-Origin (Hyndman & Athanasopoulos, 2021; Hyndman & Koehler, 2006).

La Tabla 17 reúne los modelos que sobresalen según el criterio de evaluación utilizado. Para el pronóstico diario bajo Rolling-Origin ($h = 1$), XGBoost_multivariado_RO obtuvo el menor MASE. Es el mejor resultado de error en el esquema que actualiza el origen del pronóstico de forma sucesiva.

Tabla 17*Selección final de modelos según criterio de evaluación*

Criterio de selección	Modelo recomendado	Enfoque	Justificación
Menor error diario	XGBoost multivariado_RO	Multivariado	Presentó el menor MASE en Rolling-Origin $h = 1$.
Menor error Hold-Out	RandomForest multivariado	Multivariado	Obtuvo el mejor resultado general en Hold-Out multivariado.
Mejor estabilidad semanal	XGBoost multivariado_RO $h = 7$	Multivariado	Mantuvo desempeño competitivo en horizonte $h = 7$.
Mejor desempeño direccional Hold-Out	GAM multivariado	Multivariado	Presentó los mejores valores de Accuracy, F1-score, MCC y AUC.
Mejor equilibrio direccional diario	NeuralProphet_RO	Univariado	Obtuvo el mejor F1-score en Rolling-Origin $h = 1$ univariado.
Mejor evidencia estadística frente a Naive	Modelos multivariados Hold-Out	Multivariado	La mayoría presentó significancia en la prueba Diebold-Mariano.
Modelo final recomendado	XGBoost multivariado_RO	Multivariado	Combina menor error, estabilidad temporal y capacidad de aprovechar variables exógenas.

Nota. La selección final diferencia entre criterios de error, desempeño direccional, estabilidad temporal y significancia estadística. El modelo final recomendado corresponde al que presenta mejor equilibrio general para el objetivo predictivo del estudio; Elaboración propia con base en los resultados obtenidos en Python.

En la validación Hold-Out, RandomForest multivariado registró el menor MASE. Ambos resultados ubican a los modelos multivariados basados en árboles entre las alternativas con mejor desempeño de magnitud. Su ventaja está en la capacidad para trabajar con relaciones no lineales e interacciones entre los rezagos del retorno, el volumen, la volatilidad y los indicadores técnicos (Hastie, Tibshirani, & Friedman, 2017; Chen & Guestrin, 2016).

XGBoost_multivariado_RO ($h = 7$) también mantuvo un resultado competitivo en el horizonte semanal. Aunque el error aumenta al ampliar el horizonte, el modelo conservó el menor MASE entre las alternativas multivariadas evaluadas para siete días.

Los resultados direccionales llevan a una selección distinta. GAM multivariado fue el modelo más destacado en Hold-Out por Accuracy, F1-score, MCC y AUC. En Rolling-Origin ($h = 1$), NeuralProphet_RO alcanzó el mejor F1-score dentro del enfoque univariado. Por tanto, el modelo que reduce más el error de magnitud no necesariamente es el que mejor anticipa el signo del retorno.

La prueba de Diebold-Mariano aporta evidencia adicional frente al benchmark Naive en los casos donde fue aplicada. Los modelos multivariados evaluados en Hold-Out concentraron varios resultados con diferencias estadísticamente significativas. Esta evidencia se interpreta junto con MASE y las métricas direccionales; por sí sola no define la selección final (Diebold & Mariano, 1995).

Con base en el criterio multicriterio definido en el estudio, el modelo final recomendado es XGBoost_multivariado_RO. Su selección responde principalmente al menor error diario en Rolling-Origin, al buen desempeño en el horizonte semanal y a la capacidad de utilizar información exógena. GAM multivariado y NeuralProphet_RO se conservan como referencias cuando el interés principal es la predicción direccional.

9.8.1. Criterio multicriterio para la selección del modelo final

La selección del modelo final se realizó con un criterio multicriterio. No se eligió el modelo únicamente por una métrica o por el resultado de una sola ventana de validación. Se revisaron cuatro aspectos: error de pronóstico, estabilidad temporal, desempeño direccional e interpretabilidad financiera.

El criterio principal fue la precisión en magnitud, medida principalmente con MASE. Esta métrica tiene mayor peso dentro del estudio porque permite comparar el error relativo frente a un modelo ingenuo. Se priorizaron los modelos con menor MASE fuera de muestra, especialmente en Rolling-Origin.

También se revisó la estabilidad temporal. Un modelo fue valorado de forma favorable cuando mantuvo un desempeño competitivo al cambiar el esquema de validación o al ampliar el horizonte de pronóstico. En particular, se consideró importante conservar bajo error tanto en Rolling-Origin ($h = 1$) como en Rolling-Origin ($h = 7$).

El desempeño direccional se evaluó mediante Accuracy, F1-score, MCC y AUC, según la disponibilidad de cada modelo. Este criterio complementa la evaluación del error: en aplicaciones financieras también interesa anticipar si el retorno será positivo o negativo. Sin embargo, la dirección no tuvo prioridad sobre MASE, dado que el objetivo principal del trabajo es pronosticar el retorno logarítmico.

Finalmente, se consideró la interpretabilidad financiera. Se valoró que el modelo permitiera relacionar su resultado con variables de volumen, volatilidad, momentum, tendencia y flujo de mercado. En XGBoost y Random Forest, esta revisión se apoyó en el análisis de importancia de variables.

Con base en estos criterios, XGBoost_multivariado_RO fue seleccionado como modelo final recomendado. Obtuvo el menor MASE en Rolling-Origin ($h = 1$), mantuvo el mejor desempeño relativo en Rolling-Origin ($h = 7$) y permitió aprovechar el conjunto de variables exógenas del enfoque multivariado. Además, su interpretación se complementó con la importancia relativa de los predictores.

Otros modelos también mostraron fortalezas puntuales. GAM_multivariado obtuvo mejores resultados direccionales en Hold-Out y RandomForest_multivariado registró el menor error en esa misma validación fija. Sin embargo, XGBoost_multivariado_RO presentó el balance más consistente entre precisión, estabilidad temporal e interpretabilidad dentro del diseño del estudio.

La selección no significa que XGBoost_multivariado_RO sea el mejor modelo en todas las métricas ni en todos los escenarios. Representa la alternativa recomendada a partir de los resultados observados en los esquemas de validación analizados.

10. Discusión de resultados

Este capítulo interpreta los resultados obtenidos en la evaluación de los modelos de pronóstico del retorno logarítmico de BTC-USD. El capítulo anterior presentó las métricas, los rankings y las pruebas realizadas; aquí se analiza qué significan esos resultados desde una perspectiva estadística, predictiva y financiera.

La discusión se concentra en el comportamiento de los modelos multivariados de machine learning, especialmente en su desempeño por error y estabilidad bajo distintos esquemas de validación. También se revisa la diferencia entre precisión en magnitud y capacidad direccional. Un modelo puede registrar bajo error promedio sin ser el más acertado al anticipar el signo del retorno.

Además, se examina el papel de las variables incorporadas al enfoque multivariado: volumen, volatilidad, momentum, tendencia y flujo de mercado. El propósito es valorar si estas señales complementan la información contenida en los rezagos del retorno y ayudan a describir la dinámica de corto plazo de Bitcoin.

Finalmente, el capítulo revisa el desempeño de los modelos bajo condiciones de alta y baja volatilidad, así como de alto y bajo volumen. Este análisis permite observar cómo cambia el error según el contexto de mercado y discutir las implicaciones de los resultados sin perder de vista las limitaciones propias del pronóstico de retornos financieros volátiles.

10.1. Interpretación general de resultados

La lectura conjunta de los resultados muestra que el enfoque multivariado obtuvo un desempeño favorable frente al univariado en los esquemas donde fue posible compararlos directamente. La diferencia fue más visible en los modelos de machine learning, que trabajaron con los rezagos del retorno y con variables de volumen, volatilidad e indicadores técnicos.

Los menores valores de MASE se concentraron en modelos multivariados como XGBoost y Random Forest. En los periodos evaluados, la información adicional del mercado ayudó a reducir el error de pronóstico frente a los modelos contruidos únicamente con la historia del retorno. Este resultado es razonable en Bitcoin, donde los cambios de liquidez, la actividad de negociación, la volatilidad y la presión compradora o vendedora pueden aportar señales que no están contenidas por completo en los rezagos de la serie.

Sin embargo, el menor error de magnitud no siempre coincidió con el mejor desempeño direccional. Algunos modelos redujeron más el MASE, mientras que otros mostraron mejores resultados al anticipar si el retorno sería positivo o negativo. La magnitud y el signo del retorno son dimensiones relacionadas, pero no equivalentes. Por esta razón, MASE debe analizarse junto con las métricas direccionales.

El esquema de validación también cambia la interpretación. Hold-Out permite una primera evaluación fuera de muestra en una ventana fija, mientras que Rolling-Origin revisa el desempeño en varios momentos del periodo de validación. Esta segunda alternativa es particularmente útil en

series financieras, porque permite observar si un modelo conserva su desempeño cuando cambian las condiciones de mercado.

En conjunto, los modelos multivariados de machine learning representan una alternativa sólida cuando el objetivo principal es reducir el error de pronóstico y aprovechar información adicional del mercado. La elección final, no obstante, depende del propósito del análisis: minimizar el error de magnitud, anticipar la dirección del retorno o buscar un equilibrio entre precisión, estabilidad e interpretación..

10.2. Aporte del enfoque multivariado

El enfoque multivariado es uno de los aportes metodológicos centrales del proyecto. Su propósito es evaluar si las variables de volumen, volatilidad e indicadores técnicos mejoran el pronóstico del retorno logarítmico de BTC-USD frente a modelos contruidos solo con la historia de la variable objetivo.

Los modelos univariados utilizan principalmente los rezagos del retorno. En cambio, el enfoque multivariado incorpora señales relacionadas con intensidad de negociación, liquidez relativa, momentum, tendencia, volatilidad reciente y flujo de mercado. Entre las variables incluidas se encuentran volumen relativo, volatilidades móviles, ATR, RSI, MACD, OBV y CMF.

Los resultados muestran que los modelos multivariados obtuvieron menores valores de MASE en los esquemas donde se realizó comparación directa. Esto indica que la historia del retorno no recoge toda la información útil para el pronóstico. Las variables exógenas aportan elementos adicionales para representar la dinámica de Bitcoin y aproximar mejor la magnitud de sus variaciones.

Este resultado es relevante porque BTC-USD presenta baja autocorrelación lineal, alta volatilidad, valores extremos y cambios en las condiciones del mercado. En este contexto, un

modelo basado únicamente en la memoria de la serie puede tener limitaciones. La información de volumen, volatilidad y señales técnicas amplía el conjunto de predictores disponible para el proceso de pronóstico.

El aporte fue más evidente en Random Forest y XGBoost. Los modelos basados en árboles pueden incorporar relaciones no lineales e interacciones entre predictores sin imponer una estructura funcional rígida. Esta flexibilidad resulta útil cuando los movimientos del retorno se relacionan simultáneamente con cambios en el volumen, la volatilidad y los indicadores técnicos.

La mejora por error de magnitud no se trasladó de manera uniforme a la predicción direccional. El modelo con menor MASE no siempre fue el que mejor anticipó si el retorno sería positivo o negativo. Por ello, la magnitud y la dirección deben analizarse como dimensiones complementarias: MASE permite evaluar el tamaño del error, mientras que las métricas direccionales revisan la clasificación del signo del retorno.

Desde una perspectiva financiera, las variables incorporadas contienen señales útiles para describir el entorno de mercado de corto plazo. Esto no implica una relación causal. Que una variable mejore el pronóstico significa que aporta información predictiva dentro del periodo, los modelos y los esquemas de validación utilizados; no significa que determine directamente el movimiento futuro del precio.

En conjunto, el enfoque multivariado fortaleció el análisis al incorporar información que no está contenida únicamente en los rezagos del retorno. Su principal aporte se observa en la reducción empírica del error de pronóstico y en la posibilidad de examinar el papel de las variables exógenas en modelos avanzados. Los resultados deben entenderse como evidencia descriptiva fuera de muestra, limitada al periodo analizado y al diseño metodológico aplicado.

10.3. Relación entre error, direccionalidad e interpretabilidad

La evaluación de los modelos no puede limitarse a una sola métrica. En este estudio, algunos modelos reducen mejor el error de magnitud del retorno, mientras que otros obtienen mejores resultados al anticipar si el movimiento será positivo o negativo. Por ello, la interpretación combina tres dimensiones: precisión numérica, desempeño direccional e interpretabilidad de las variables utilizadas.

Los modelos multivariados de machine learning, en particular XGBoost y Random Forest, obtuvieron los menores errores de magnitud en los principales esquemas de validación. Su ventaja está en la capacidad de incorporar relaciones no lineales e interacciones entre el retorno.

10.4. Discusión por condiciones de mercado

La discusión por condiciones de mercado permite revisar cómo cambia el desempeño de los modelos cuando varían la volatilidad y el volumen de negociación de BTC-USD. En Bitcoin, un modelo puede funcionar bien en promedio y perder precisión durante periodos de alta incertidumbre, movimientos extremos o cambios fuertes en la actividad transaccional.

Por esta razón, el análisis no se limita al error global. Se comparan los resultados bajo condiciones de alta y baja volatilidad, así como de alto y bajo volumen. Esta separación permite identificar qué modelos conservan un comportamiento más estable y cuáles son más sensibles a cambios en el entorno de mercado.

Los modelos multivariados tienen una ventaja potencial en este análisis porque incorporan variables que describen distintas dimensiones del mercado. El volumen, ATR, volatilidades móviles, RSI, MACD, OBV y CMF aportan información sobre actividad de negociación, volatilidad, momentum, tendencia y flujo de capital. Estas señales pueden ser especialmente útiles

cuando la dinámica reciente del mercado no queda reflejada por completo en los rezagos del retorno.

El análisis por condiciones también permite distinguir entre un modelo con buen desempeño promedio y otro que mantiene resultados competitivos cuando cambian las condiciones de mercado. En los apartados siguientes se presentan los hallazgos relacionados con los niveles de volatilidad y volumen, así como la estabilidad relativa de los modelos en cada escenario. La interpretación es descriptiva y corresponde al periodo de validación analizado.

10.4.1. Construcción de regímenes de mercado

La clasificación por condiciones de mercado se construyó para revisar si el desempeño de los modelos cambia según la volatilidad y el volumen de negociación de BTC-USD. En Bitcoin, el comportamiento del retorno no es uniforme en el tiempo. Existen periodos de mayor incertidumbre, cambios en la liquidez y variaciones importantes en la actividad transaccional que pueden afectar la precisión de los pronósticos.

En este estudio, el término “régimen” se utiliza en un sentido operativo y descriptivo. No se busca identificar estados económicos estructurales ni estimar cambios de estado latentes del mercado. La clasificación divide las observaciones de la muestra de validación según condiciones relativas de mayor o menor volatilidad y volumen, con el fin de comparar el error de los modelos en cada escenario.

El análisis se aplicó a los pronósticos generados en la validación Hold-Out. Cada fecha de la muestra de validación se clasificó de acuerdo con los valores observados de `rolling\_vol\_30` y `volume\_ratio`. Por tanto, las tablas posteriores muestran una evaluación condicional fuera de muestra: describen cómo se comportaron los modelos en observaciones asociadas con distintos niveles de volatilidad y actividad de negociación.

Para la volatilidad se utilizó `rolling\_vol\_30`, una medida calculada sobre una ventana móvil de 30 días del retorno logarítmico. El umbral fue la mediana de esta variable dentro de la muestra de validación. Las observaciones con valores iguales o inferiores a la mediana se clasificaron como de baja volatilidad; las observaciones con valores superiores se clasificaron como de alta volatilidad.

La mediana se empleó como punto de corte porque es una medida robusta frente a valores extremos y permite separar la muestra en grupos comparables. No representa un nivel teórico de volatilidad ni una frontera económica permanente. La clasificación tiene carácter exploratorio y no reemplaza métodos formales de identificación de regímenes, como modelos de cambio de régimen de Markov, clustering, umbrales endógenos o modelos de volatilidad condicional.

Para el volumen se utilizó `volume\_ratio`, que compara el volumen diario con su promedio móvil reciente. El punto de referencia fue 1.00. Los valores iguales o inferiores a 1.00 se clasificaron como bajo volumen, mientras que los valores superiores se clasificaron como alto volumen. Esta separación distingue fechas con menor actividad transaccional de aquellas con una negociación relativamente más intensa.

El análisis no permite afirmar que un nivel de volatilidad o volumen cause el desempeño de un modelo. Su propósito es describir cómo cambió el error de pronóstico bajo las condiciones observadas durante la validación Hold-Out.

La Tabla 18 presenta los criterios utilizados para clasificar las observaciones según volatilidad y volumen. La clasificación no pretende identificar regímenes financieros estructurales ni establecer relaciones causales entre estas variables y el retorno. Su función es organizar la evaluación predictiva bajo condiciones observables del mercado y revisar si el error de los modelos cambia entre escenarios de mayor o menor volatilidad y actividad transaccional.

Tabla 18*Criterios de clasificación de regímenes de mercado*

Dimensión	Variable utilizada	Régimen	Umbral de referencia	Criterio de clasificación	Interpretación
Volatilidad	rolling_vol_30	Baja volatilidad	≤ 0.023185 (retorno logarítmico diario $\approx 2.3185\%$)	Valores iguales o inferiores a la mediana de rolling_vol_30 calculada en la muestra de validación.	Mercado relativamente estable, con menor amplitud en las variaciones del retorno.
Volatilidad	rolling_vol_30	Alta volatilidad	> 0.023185 (retorno logarítmico diario $\approx 2.3185\%$)	Valores superiores a la mediana de rolling_vol_30 calculada en la muestra de validación.	Mercado con mayor incertidumbre, oscilaciones amplias y mayor dificultad predictiva.
Volumen	volume_ratio	Bajo volumen	≤ 1.00 (adimensional)	Volumen diario igual o inferior a su promedio móvil reciente.	Menor intensidad de negociación y posible menor participación del mercado.
Volumen	volume_ratio	Alto volumen	> 1.000 (adimensional)	Volumen diario superior a su promedio móvil reciente.	Mayor intensidad de negociación, aumento de liquidez y posible presión compradora o vendedora.

Nota: El umbral de volatilidad corresponde a la mediana de rolling_vol_30 calculada sobre la muestra de validación; no representa una constante teórica del mercado, sino un punto de corte empírico para clasificar cada observación diaria. El umbral de volumen se define en volume_ratio = 1, donde valores superiores indican volumen por encima del promedio móvil reciente y valores iguales o inferiores indican volumen igual o menor a dicho promedio; Elaboración propia con base en variables calculadas a partir de datos OHLCV de BTC-USD en Python.

El análisis se realizó únicamente con los pronósticos de la validación Hold-Out. Cada fecha del periodo de validación se clasificó con base en los valores observados de `rolling\_vol\_30` y `volume\_ratio`. Por ello, las tablas de desempeño por condición representan una evaluación condicional fuera de muestra dentro de esa ventana Hold-Out.

Rolling-Origin se utiliza en otros apartados para examinar la estabilidad temporal de los modelos, pero no hace parte de esta clasificación. En consecuencia, los resultados deben interpretarse como una comparación descriptiva entre segmentos de la muestra Hold-Out, no como una modelación formal de cambios de régimen.

Esta segmentación permite observar si un modelo conserva resultados similares entre condiciones de mercado o si su desempeño es más sensible a la volatilidad o al volumen. Complementa la evaluación global, pero no sustituye métodos formales de identificación de regímenes. Sus conclusiones se limitan a la muestra de validación analizada y a los umbrales operativos definidos en la Tabla 18.

10.4.2. Desempeño predictivo por régimen de volatilidad

El análisis por volatilidad permite revisar si los modelos conservan su desempeño cuando cambia el nivel de inestabilidad del mercado. En Bitcoin, los periodos de alta volatilidad suelen incluir movimientos más abruptos, mayor dispersión de los retornos y valores extremos. En esas condiciones, el pronóstico tiende a ser más difícil.

La clasificación se construyó con la variable ``rolling_vol_30`` y divide la muestra Hold-Out en observaciones de baja y alta volatilidad. Con esta segmentación se puede observar si un modelo funciona principalmente en periodos relativamente estables o si mantiene resultados competitivos cuando aumenta la incertidumbre.

La comparación incluye dos dimensiones. La primera corresponde a las métricas de error, con MASE como indicador principal. La segunda revisa la capacidad para anticipar el signo del retorno mediante métricas direccionales. Esta distinción es necesaria porque un modelo puede reducir el error promedio y, aun así, no mejorar la clasificación entre movimientos positivos y negativos.

Los resultados por volatilidad permiten, entonces, complementar la evaluación global y revisar la estabilidad práctica de los modelos bajo las condiciones observadas en la muestra de validación.

10.4.2.1. Desempeño por métricas de error

El análisis del desempeño por métricas de error permite evaluar si la precisión de los modelos cambia bajo escenarios de alta y baja volatilidad del mercado de Bitcoin. Para ello, se utilizó la clasificación construida previamente a partir de la variable `rolling_vol_30`, donde los periodos de baja volatilidad corresponden a observaciones con valores iguales o inferiores al umbral empírico de referencia, mientras que los periodos de alta volatilidad corresponden a observaciones que superan dicho umbral.

Desde el punto de vista metodológico, este análisis no implica reentrenar los modelos ni modificar los pronósticos generados. Cada predicción se asocia con el régimen de volatilidad observado en la fecha correspondiente, permitiendo evaluar el comportamiento del error dentro de cada subconjunto de validación. En este apartado se prioriza el MASE, dado que constituye la métrica principal de comparación del estudio.

Tabla 19

Comparación del MASE por modelo según régimen de volatilidad

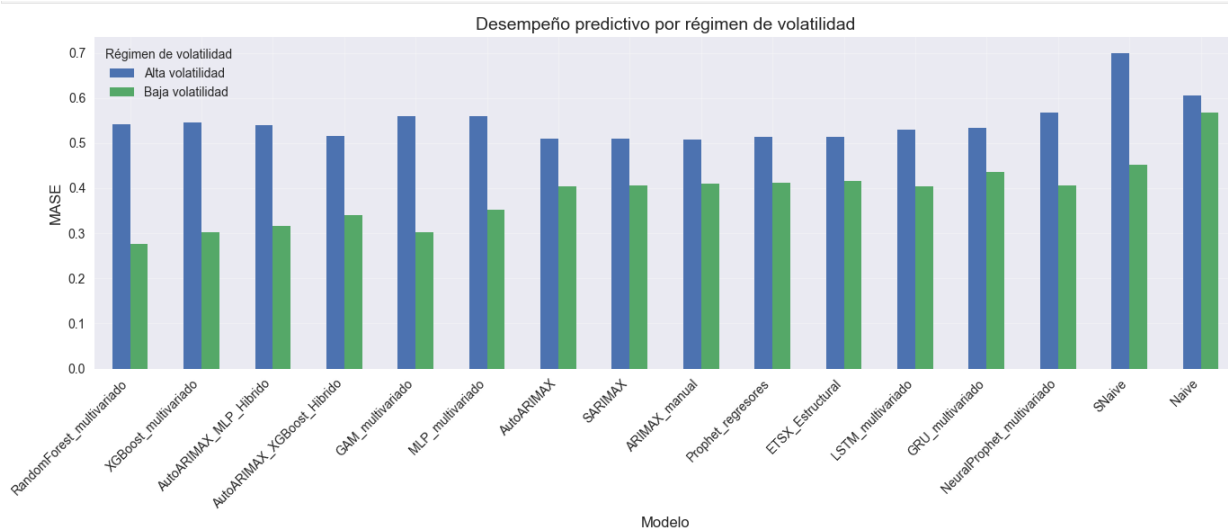
Modelo	Alta volatilidad	Baja volatilidad	Diferencia _abs	MASE_pr omedio
RandomForest_multivariado	0.5422	0.2770	0.2652	0.4096
XGBoost_multivariado	0.5447	0.3029	0.2418	0.4238
AutoARIMAX_MLP_Hibrido	0.5392	0.3159	0.2232	0.4276
AutoARIMAX_XGBoost_Hibrido	0.5160	0.3396	0.1764	0.4278
GAM_multivariado	0.5598	0.3022	0.2576	0.4310
MLP_multivariado	0.5603	0.3512	0.2091	0.4557
AutoARIMAX	0.5099	0.4049	0.1050	0.4574

Modelo	Alta volatilidad	Baja volatilidad	Diferencia _abs	MASE_pr omedio
SARIMAX	0.5096	0.4059	0.1037	0.4578
ARIMAX_manual	0.5072	0.4102	0.0971	0.4587
Prophet_regresores	0.5142	0.4112	0.1030	0.4627
ETSX_Estructural	0.5127	0.4163	0.0963	0.4645
LSTM_multivariado	0.5303	0.4043	0.1260	0.4673
GRU_multivariado	0.5332	0.4368	0.0964	0.4850
NeuralProphet_multivariado	0.5668	0.4052	0.1616	0.4860
SNaive	0.6994	0.4511	0.2483	0.5753
Naive	0.6051	0.5670	0.0381	0.5861

Nota: elaboración propia con base en los modelos implementados en Python.

La Tabla 19 y la Figura 13 muestran que el error de pronóstico fue mayor en las observaciones clasificadas como de alta volatilidad y menor en las de baja volatilidad. Bajo condiciones de mayor inestabilidad, los retornos presentan más dispersión y movimientos más abruptos, por lo que anticipar su magnitud resulta más difícil.

Los modelos multivariados de machine learning registran los menores errores promedio en ambos grupos. RandomForest_multivariado obtiene el menor MASE promedio y XGBoost_multivariado se ubica en la siguiente posición. En la muestra Hold-Out analizada, los modelos basados en árboles aprovechan mejor la información de volumen, volatilidad e indicadores técnicos que las demás alternativas evaluadas.

Figura 13*Comparación del MASE por modelo según régimen de volatilidad*

Nota: Elaboración propia con base en datos OHLCV de BTC-USD obtenidos de Yahoo Finance; modelos estimados en Python.

La diferencia entre alta y baja volatilidad no debe leerse de manera aislada. Algunos modelos clásicos muestran una variación menor entre ambos grupos, pero mantienen un MASE promedio superior al de Random Forest y XGBoost. Un modelo puede cambiar poco entre condiciones de mercado y, aun así, tener un error elevado en ambos casos. La robustez requiere revisar, al mismo tiempo, el nivel promedio de error y la diferencia observada entre segmentos.

Los modelos Naive y SNaive ilustran este punto. Naive presenta una diferencia reducida entre los dos grupos de volatilidad, pero su MASE promedio se encuentra entre los más altos. Esa estabilidad aparente no representa una ventaja predictiva; refleja una respuesta limitada frente a los cambios de la serie.

En conjunto, la baja volatilidad se asocia con menores errores de pronóstico, mientras que la alta volatilidad eleva la dificultad del problema. Dentro de las observaciones Hold-Out evaluadas, RandomForest_multivariado y XGBoost_multivariado ofrecen la combinación más favorable entre error promedio y comportamiento bajo ambos niveles de volatilidad.

10.4.2.2. Desempeño por métricas direccionales

Además del análisis por métricas de error, resulta necesario evaluar la capacidad de los modelos para anticipar correctamente la dirección del retorno logarítmico de Bitcoin. Esta dimensión es relevante porque un modelo puede presentar bajo error promedio en magnitud y, aun así, no ser el más adecuado para identificar si el retorno futuro será positivo o negativo.

En este apartado se analizan las métricas direccionales bajo los mismos regímenes de volatilidad definidos previamente mediante `rolling_vol_30`. La evaluación se centra principalmente en Accuracy y F1-score, mientras que MCC y AUC se utilizan como métricas complementarias para valorar la consistencia global de la clasificación direccional. El detalle completo de las métricas por modelo y régimen se presenta en el Anexo A4.1. Métricas completas de desempeño predictivo por régimen de volatilidad.

La Tabla 20 muestra que la capacidad para anticipar el signo del retorno fue, en general, más favorable en las observaciones de baja volatilidad. Cuando la variación del mercado es menor, los movimientos tienden a ser menos abruptos y las señales de volumen, momentum y flujo de mercado pueden resultar más aprovechables. En alta volatilidad aumenta el ruido de la serie y varios modelos pierden capacidad para clasificar correctamente la dirección.

GAM_multivariado presenta el mejor desempeño direccional promedio, especialmente en F1-score, MCC y AUC. Esto indica que el modelo logra un buen equilibrio entre aciertos y errores de clasificación dentro de la muestra analizada. Su resultado no coincide necesariamente con el menor MASE, lo que vuelve a mostrar que el error de magnitud y la predicción del signo son criterios distintos.

Tabla 20

Comparación de la capacidad direccional por modelo según régimen de volatilidad.

Modelo	Accurac y alta volatilid ad	Accuracy baja volatilidad	Accurac y promedi o	F1 alta volatilid ad	F1 baja volatili dad	F1 prome dio	MCC prome dio	AUC promedi o
RandomForest_multivariado	0.4667	0.7333	0.6000	0.4286	0.6000	0.5143	0.2315	0.6920
XGBoost_multivariado	0.6667	0.7667	0.7167	0.6429	0.6957	0.6693	0.4422	0.7232
AutoARIMAX_MLP_Hibrido	0.5333	0.7000	0.6167	0.5882	0.6400	0.6141	0.2460	0.6540
AutoARIMAX_XGBoost_Hibrido	0.5000	0.6000	0.5500	0.4000	0.3333	0.3667	0.1021	0.6763
GAM_multivariado	0.6667	0.8000	0.7333	0.6667	0.7692	0.7179	0.4697	0.7835
MLP_multivariado	0.4000	0.6333	0.5167	0.4706	0.6857	0.5782	0.0659	0.6071
AutoARIMAX	0.6000	0.6667	0.6333	0.6000	0.5000	0.5500	0.2864	0.6250
SARIMAX	0.6000	0.6667	0.6333	0.6000	0.5000	0.5500	0.2864	0.6317
ARIMAX_manual	0.6000	0.6000	0.6000	0.6000	0.4000	0.5000	0.2029	0.6228
Prophet_regresores	0.6000	0.5333	0.5667	0.6250	0.4167	0.5208	0.1327	0.5960
ETSX_Estructural	0.5333	0.5667	0.5500	0.5333	0.4348	0.4841	0.0940	0.5938
LSTM_multivariado	0.5333	0.5333	0.5333	0.0000	0.0000	0.0000	0.0000	0.6049
GRU_multivariado	0.5333	0.4333	0.4833	0.1250	0.1053	0.1151	0.1106	0.3348
NeuralProphet_multivariado	0.5667	0.6667	0.6167	0.6486	0.6667	0.6577	0.2697	0.5915
SNaive	0.4000	0.7333	0.5667	0.3077	0.7143	0.5110	0.1230	0.5703
Naive	0.4667	0.4667	0.4667	0.6364	0.6364	0.6364	0.0000	0.5000

Nota. La tabla resume la capacidad direccional de los modelos bajo regímenes de alta y baja volatilidad. El F1-score promedio se utiliza como criterio principal de lectura, ya que combina Precision y Recall, mientras que Accuracy, MCC y AUC complementan la evaluación del signo del retorn; Elaboración propia con base en los modelos implementados en Python.

XGBoost_multivariado combina resultados direccionales competitivos con bajo error de magnitud. Por esta razón, ofrece un balance favorable entre las dos dimensiones de evaluación. RandomForest_multivariado, aunque registra el menor error promedio, presenta una menor

capacidad direccional en alta volatilidad. El mejor ajuste en magnitud no garantiza, por sí solo, una buena clasificación del signo del retorno.

Los modelos clásicos y estructurales muestran resultados direccionales intermedios. Pueden recoger parte de las señales asociadas al movimiento del mercado, pero no alcanzan el desempeño global de los modelos mejor ubicados. El caso de Naive también requiere cautela: un F1-score relativamente estable no compensa su bajo Accuracy promedio y un MCC cercano a cero, indicadores de escasa capacidad para diferenciar entre retornos positivos y negativos.

En conjunto, las observaciones de baja volatilidad registran una mejor clasificación direccional que las de alta volatilidad. La comparación refuerza la necesidad de seleccionar los modelos con una lectura conjunta de MASE, métricas direccionales y estabilidad bajo las condiciones de mercado observadas.

10.4.3. Desempeño predictivo por régimen de volumen

El análisis por régimen de volumen permite evaluar si el desempeño predictivo de los modelos cambia cuando el mercado de Bitcoin opera bajo distintos niveles de actividad transaccional. Esta dimensión es relevante porque el volumen de negociación puede reflejar cambios en liquidez, participación de mercado, presión compradora o vendedora y episodios de mayor actividad especulativa.

A partir de la clasificación construida mediante la variable `volume_ratio`, se comparan los resultados bajo escenarios de bajo y alto volumen. Esta segmentación permite analizar si las variables exógenas aportan mayor información predictiva cuando el mercado presenta mayor intensidad de negociación o si, por el contrario, los modelos se comportan mejor en contextos de menor actividad y menor ruido de mercado.

El análisis se desarrolla en dos dimensiones complementarias: métricas de error, con énfasis en el MASE como criterio principal, y métricas direccionales, orientadas a evaluar la capacidad de anticipar el signo del retorno. Esta separación permite distinguir entre precisión en la magnitud del pronóstico y capacidad clasificatoria, dos aspectos que no necesariamente coinciden en series financieras. Los resultados detallados por modelo y régimen se conservan en el Anexo 5, mientras que en el cuerpo del capítulo se presentan tablas resumidas para facilitar la interpretación crítica de los hallazgos.

10.4.3.1. Desempeño por métricas de error

El desempeño por métricas de error según régimen de volumen permite evaluar si la precisión de los modelos cambia cuando el mercado de Bitcoin presenta alta o baja actividad transaccional. En este análisis, el régimen de volumen se definió a partir de la variable `volume_ratio`, clasificando como alto volumen los periodos en los que el volumen diario superó su promedio móvil reciente y como bajo volumen aquellos en los que fue igual o inferior a dicho promedio.

Desde el punto de vista metodológico, esta evaluación no implica reentrenar los modelos, sino calcular las métricas de error sobre los subconjuntos de validación definidos por el régimen de volumen. En este apartado se prioriza el MASE, dado que es la métrica principal del estudio y permite comparar el desempeño relativo de los modelos frente a un referente de escala común.

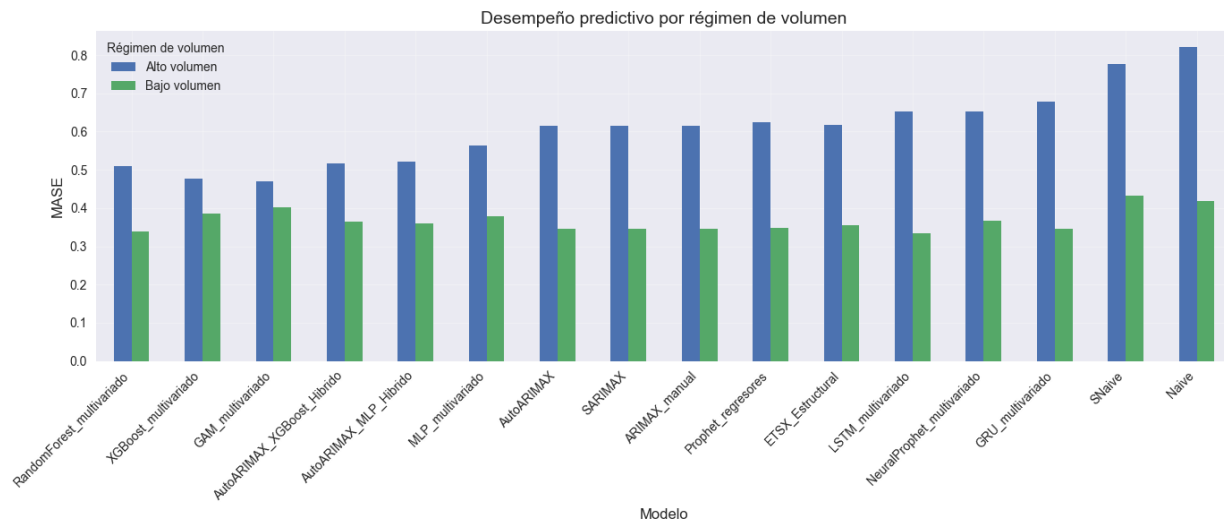
Tabla 21

Comparación del MASE por modelo según régimen de volumen

Modelo	Alto volumen	Bajo volumen	Diferencia_abs	MASE_promedio
RandomForest_multivariado	0.5087	0.3388	0.1699	0.4238
XGBoost_multivariado	0.4775	0.3854	0.0921	0.4314

Modelo	Alto volumen	Bajo volumen	Diferencia_abs	MASE_promedio
GAM_multivariado	0.4709	0.4025	0.0684	0.4367
AutoARIMAX_XGBoost_Hibrido	0.5167	0.3643	0.1524	0.4405
AutoARIMAX_MLP_Hibrido	0.5223	0.3599	0.1624	0.4411
MLP_multivariado	0.5635	0.3788	0.1847	0.4711
AutoARIMAX	0.6145	0.3452	0.2693	0.4799
SARIMAX	0.6160	0.3447	0.2713	0.4804
ARIMAX_manual	0.6154	0.3468	0.2685	0.4811
Prophet_regresores	0.6236	0.3477	0.2759	0.4857
ETSX_Estructural	0.6181	0.3548	0.2633	0.4865
LSTM_multivariado	0.6528	0.3348	0.3180	0.4938
NeuralProphet_multivariado	0.6537	0.3661	0.2876	0.5099
GRU_multivariado	0.6795	0.3461	0.3334	0.5128
SNaive	0.7762	0.4318	0.3444	0.6040
Naive	0.8218	0.4177	0.4041	0.6197

Nota. La tabla resume el MASE por régimen de volumen. Un menor MASE indica mejor desempeño predictivo en términos de error; Eboración propia con base en los modelos implementados en Python.

Figura 14*Comparación del MASE por modelo según régimen de volumen*

Nota: elaboración propia con base en datos OHLCV de BTC-USD obtenidos de Yahoo Finance; modelos estimados en Python.

La Tabla 21 y la Figura 14 muestran que el error de pronóstico fue mayor en las observaciones clasificadas como de alto volumen y menor en las de bajo volumen. Una mayor actividad transaccional no facilita necesariamente la predicción del retorno de Bitcoin. En la muestra analizada, los episodios de alto volumen estuvieron asociados con movimientos más bruscos, cambios rápidos en la liquidez y mayor dificultad para anticipar la magnitud del retorno.

Los modelos multivariados mantienen los mejores resultados promedio. RandomForest_multivariado, XGBoost_multivariado y GAM_multivariado se ubican entre los menores valores de MASE. Esto indica que las variables de volumen, volatilidad, momentum y flujo de mercado aportan información útil en ambos niveles de actividad transaccional.

El modelo con mejor resultado cambia según la condición de volumen. En alto volumen, GAM_multivariado y XGBoost_multivariado presentan resultados competitivos. En bajo volumen, RandomForest_multivariado obtiene el menor error. Los modelos no responden de la

misma forma a los cambios en la intensidad de negociación, por lo que el ranking depende parcialmente del contexto de mercado.

La diferencia entre alto y bajo volumen debe analizarse junto con el error promedio. Algunos modelos presentan brechas reducidas entre los dos grupos, pero conservan un MASE alto. Una menor sensibilidad al volumen no equivale, por sí sola, a una mejor capacidad predictiva. La evaluación de robustez requiere considerar el nivel de error y su variación entre condiciones.

Naive y SNaive registran los mayores errores, especialmente en alto volumen. Los modelos de referencia tienen mayores limitaciones cuando el mercado presenta actividad transaccional elevada. Los modelos multivariados más flexibles aprovechan mejor la información adicional disponible, aunque el alto volumen continúa siendo una condición más exigente para el pronóstico.

En conjunto, las observaciones de bajo volumen presentaron menores errores que las de alto volumen. RandomForest_multivariado, XGBoost_multivariado y GAM_multivariado ofrecen los resultados más favorables dentro de la muestra Hold-Out analizada, lo que respalda el uso de variables de actividad de mercado, volatilidad, momentum y flujo financiero en el enfoque multivariado.

10.4.3.2. Desempeño por métricas direccionales

El análisis por métricas direccionales según régimen de volumen permite evaluar si la capacidad de los modelos para anticipar el signo del retorno logarítmico de BTC-USD cambia cuando el mercado opera bajo distintos niveles de actividad transaccional. Esta dimensión resulta relevante porque el volumen de negociación puede reflejar cambios en liquidez, presión compradora o vendedora y participación de mercado, factores que pueden afectar no solo la magnitud del error, sino también la dirección del movimiento pronosticado.

A diferencia de las métricas de error, que evalúan la distancia entre el valor observado y el valor pronosticado, las métricas direccionales permiten analizar si el modelo identifica correctamente si el retorno será positivo o negativo. Para este propósito, se consideran principalmente Accuracy y F1-score, mientras que MCC y AUC se utilizan como métricas complementarias para valorar la consistencia global de la clasificación.

En este apartado, la comparación se realiza entre escenarios de alto y bajo volumen definidos a partir de la variable `volume_ratio`. Esta segmentación permite analizar si la mayor actividad transaccional aporta señales útiles para anticipar el signo del retorno o si, por el contrario, incrementa el ruido de mercado y dificulta la clasificación direccional.

Tabla 22

Comparación de la capacidad direccional por modelo según régimen de volumen

Modelo	Acc. alto	Acc. bajo	Acc. prom.	F1 alto	F1 bajo	F1 prom.	MCC prom.	AUC prom.
RandomForest_multivariado	0.7200	0.5143	0.6171	0.4615	0.5143	0.4879	0.1954	0.6674
XGBoost_multivariado	0.8400	0.6286	0.7343	0.7778	0.6061	0.6919	0.4669	0.7184
GAM_multivariado	0.8000	0.6857	0.7429	0.7059	0.7179	0.7119	0.4608	0.7884
AutoARIMAX_XGBoost_Hibrido	0.6800	0.4571	0.5686	0.3333	0.3871	0.3602	0.0869	0.6548
AutoARIMAX_MLP_Hibrido	0.7600	0.5143	0.6371	0.6250	0.6047	0.6148	0.2284	0.6241
MLP_multivariado	0.5200	0.5143	0.5171	0.4000	0.6531	0.5265	-0.0201	0.5623
AutoARIMAX	0.8000	0.5143	0.6571	0.7059	0.4848	0.5954	0.3021	0.6977
SARIMAX	0.8000	0.5143	0.6571	0.7059	0.4848	0.5954	0.3021	0.7061
ARIMAX_manual	0.7600	0.4857	0.6229	0.6667	0.4375	0.5521	0.2362	0.7114
Prophet_regresores	0.6800	0.4857	0.5829	0.6364	0.4706	0.5535	0.1852	0.6817

Modelo	Acc. alto	Acc. bajo	Acc. prom.	F1 alto	F1 bajo	F1 prom.	MCC prom.	AUC prom.
ETSX_Estructural	0.6800	0.4571	0.5686	0.6000	0.4242	0.5121	0.1361	0.6667
LSTM_multivariado	0.6400	0.4571	0.5486	0.0000	0.0000	0.0000	0.0000	0.5431
NeuralProphet_multivariado	0.6400	0.6000	0.6200	0.6400	0.6667	0.6533	0.2859	0.6140
GRU_multivariado	0.6400	0.3714	0.5057	0.1818	0.0833	0.1326	-0.0975	0.3633
SNaive	0.5600	0.5714	0.5657	0.2667	0.6154	0.4410	0.0506	0.5240
Naive	0.3600	0.5429	0.4514	0.5294	0.7037	0.6166	0.0000	0.5000

Nota. La tabla conserva el mismo orden de modelos de la tabla de error. El F1-score promedio se utiliza como apoyo para valorar la capacidad direccional, complementado por Accuracy, MCC y AUC; Elaboración propia con base en los modelos implementados en Python.

La Tabla 22 muestra que el desempeño direccional no cambia de manera uniforme entre las observaciones de alto y bajo volumen. Mientras las métricas de error fueron más favorables en bajo volumen, la capacidad para anticipar el signo del retorno varía según el modelo. Algunos modelos clasifican mejor bajo alta actividad transaccional y otros bajo bajo volumen. El efecto del volumen depende, por tanto, de la forma en que cada modelo utiliza la información exógena.

GAM_multivariado presenta el mejor resultado direccional promedio, con un balance favorable entre Accuracy, F1-score, MCC y AUC. El modelo aprovecha las relaciones entre volumen, volatilidad y flujo de mercado para clasificar el signo del retorno. Sin embargo, este resultado debe leerse junto con MASE, porque el mejor modelo direccional no coincide necesariamente con el de menor error de magnitud.

XGBoost_multivariado también ofrece un resultado equilibrado. Mantiene métricas direccionales competitivas y, al mismo tiempo, se ubica entre los modelos con menor error. RandomForest_multivariado, aunque lidera varias comparaciones por MASE, presenta una

capacidad direccional más moderada. Esto muestra nuevamente que aproximar la magnitud del retorno y acertar su signo son tareas distintas.

Los modelos clásicos y estructurales presentan resultados intermedios. En algunos escenarios capturan señales direccionales, pero su desempeño cambia más al pasar de alto a bajo volumen. Naive y SNaive deben interpretarse con cautela: algunos valores de F1-score pueden parecer aceptables, pero MCC y AUC reflejan una capacidad limitada para diferenciar entre retornos positivos y negativos.

En conjunto, ningún modelo domina todos los criterios. GAM_multivariado destaca en dirección, RandomForest_multivariado en error de magnitud y XGBoost_multivariado ofrece uno de los balances más favorables entre precisión, direccionalidad y estabilidad bajo las condiciones de volumen observadas.

10.4.4. Estabilidad de los modelos entre regímenes

La estabilidad entre condiciones de mercado permite revisar si el desempeño de un modelo cambia cuando varían la volatilidad y el volumen de BTC-USD. En este estudio se consideran dos clasificaciones descriptivas: volatilidad, definida a partir de rolling_vol_30, y volumen, definida con volume_ratio. El propósito no es identificar estados financieros estructurales, sino comparar el comportamiento de los modelos bajo observaciones de alta y baja volatilidad, así como de alto y bajo volumen.

La estabilidad se define como la variación del error entre dos condiciones de mercado. Para medirla se utiliza la diferencia absoluta del MASE:

$$\Delta MASE = |MASE_{\text{régimen alto}} - MASE_{\text{régimen bajo}}| \quad (\text{Ecuación 16})$$

En volatilidad, se compara el MASE de las observaciones de alta volatilidad con el de baja volatilidad. En volumen, se compara el MASE de alto volumen con el de bajo volumen. Un valor

menor de ($\Delta MASE$) indica que el error cambia menos entre los dos grupos y, por tanto, que el modelo es más estable bajo ese criterio.

Sin embargo, una baja diferencia no basta para considerar que un modelo sea conveniente. Un modelo puede registrar errores similares entre regímenes y, aun así, tener un MASE alto en ambos grupos. Por esta razón, la estabilidad debe revisarse junto con el error promedio.

En este trabajo, la robustez se entiende como la combinación entre bajo error promedio y baja sensibilidad a los cambios de volatilidad o volumen. Un modelo robusto no es solo aquel que varía poco entre condiciones, sino aquel que mantiene un error competitivo cuando el entorno de mercado cambia.

Esta distinción evita dos interpretaciones incompletas. Un modelo con bajo MASE promedio, pero con grandes diferencias entre regímenes, puede ser sensible a determinadas condiciones de mercado. A su vez, un modelo muy estable, pero impreciso en todos los escenarios, tampoco constituye una alternativa adecuada.

Los apartados siguientes revisan la variación del error entre regímenes, la respuesta de los modelos ante cambios conjuntos de volatilidad y volumen, y las alternativas que combinan mejor precisión promedio y estabilidad dentro de la muestra Hold-Out analizada.

10.4.4.1. Variación del error entre regímenes

La variación del error entre regímenes permite evaluar qué tan sensible es cada modelo frente a los cambios en las condiciones del mercado. Para ello, se comparan las diferencias absolutas del MASE entre alta y baja volatilidad, así como entre alto y bajo volumen. Esta lectura permite identificar no solo los modelos con menor error promedio, sino también aquellos cuyo desempeño se mantiene relativamente estable cuando cambia el entorno de mercado.

Tabla 23*Estabilidad del MASE entre regímenes de volatilidad y volumen*

Modelo	Dif. volatilidad	MASE prom. volatilidad	Dif. volumen	MASE prom. volumen	Dif. promedio	MASE prom. global
GAM_multivariado	0.2576	0.4310	0.0684	0.4367	0.1630	0.4338
AutoARIMAX_XGBoost_Hibrido	0.1764	0.4278	0.1524	0.4405	0.1644	0.4342
XGBoost_multivariado	0.2418	0.4238	0.0921	0.4314	0.1670	0.4276
ETSX_Estructural	0.0963	0.4645	0.2633	0.4865	0.1798	0.4755
ARIMAX_manual	0.0971	0.4587	0.2685	0.4811	0.1828	0.4699
AutoARIMAX	0.1050	0.4574	0.2693	0.4799	0.1872	0.4686
SARIMAX	0.1037	0.4578	0.2713	0.4804	0.1875	0.4691
Prophet_regresores	0.1030	0.4627	0.2759	0.4857	0.1895	0.4742
AutoARIMAX_MLP_Hibrido	0.2232	0.4276	0.1624	0.4411	0.1928	0.4343
MLP_multivariado	0.2091	0.4557	0.1847	0.4711	0.1969	0.4634
GRU_multivariado	0.0964	0.4850	0.3334	0.5128	0.2149	0.4989
RandomForest_multivariado	0.2652	0.4096	0.1699	0.4238	0.2175	0.4167
Naive	0.0381	0.5861	0.4041	0.6197	0.2211	0.6029
LSTM_multivariado	0.1260	0.4673	0.3180	0.4938	0.2220	0.4806
NeuralProphet_multivariado	0.1616	0.4860	0.2876	0.5099	0.2246	0.4979
SNaive	0.2483	0.5753	0.3444	0.6040	0.2964	0.5896

Nota: La tabla resume la variación absoluta del MASE y el MASE promedio de los modelos entre regímenes de volatilidad y volumen; Elaboración propia con base en los modelos implementados en Python.

La Tabla 23 permite distinguir entre dos criterios diferentes: el nivel promedio de error y la variación de ese error entre condiciones de mercado. Una diferencia pequeña entre alta y baja volatilidad, o entre alto y bajo volumen, indica que el modelo cambia menos ante esas condiciones. Sin embargo, esa estabilidad solo es útil cuando el error promedio también es competitivo.

GAM_multivariado, AutoARIMAX_XGBoost_Híbrido y XGBoost_multivariado presentan las menores diferencias promedio entre regímenes. Dentro de este grupo, XGBoost_multivariado ofrece el balance más favorable, porque combina una variación moderada del MASE con un error promedio menor que el de los otros modelos más estables.

RandomForest_multivariado registra el menor MASE promedio global. Su principal fortaleza es, por tanto, la precisión en magnitud. No obstante, presenta una diferencia mayor entre regímenes, lo que indica una sensibilidad más alta a los cambios de volatilidad y volumen.

Los modelos clásicos y estructurales muestran diferencias intermedias entre regímenes, pero sus valores de MASE global permanecen por encima de los modelos líderes. Naive y SNaive también deben interpretarse con cautela: aunque puedan mostrar cierta estabilidad, sus errores promedio son elevados.

La Tabla 23 no permite elegir un modelo solo por estabilidad o solo por precisión. XGBoost_multivariado se destaca por combinar bajo error promedio y estabilidad relativa. RandomForest_multivariado conserva la mejor precisión global, mientras que GAM_multivariado presenta menor variación entre las condiciones de mercado, aunque con un error promedio ligeramente mayor..

10.4.4.2. Robustez frente a cambios de volatilidad y volumen

La robustez frente a cambios de volatilidad y volumen permite evaluar qué modelos mantienen un desempeño relativamente estable cuando cambian simultáneamente las condiciones

de incertidumbre y actividad transaccional del mercado. Para ello, la Tabla 22 resume la diferencia del MASE entre regímenes de volatilidad, la diferencia del MASE entre regímenes de volumen, la diferencia promedio y el MASE promedio global de cada model.

Tabla 24

Robustez del MASE frente a cambios de volatilidad y volumen

Modelo	Dif. volatilidad	MASE prom. volatilidad	Dif. volumen	MASE prom. volumen	Dif. promedio	MASE prom. global
GAM_multivariado	0.2576	0.4310	0.0684	0.4367	0.1630	0.4338
AutoARIMAX_XGB oost_Hibrido	0.1764	0.4278	0.1524	0.4405	0.1644	0.4342
XGBoost_multivariado	0.2418	0.4238	0.0921	0.4314	0.1670	0.4276
ETSX_Estructural	0.0963	0.4645	0.2633	0.4865	0.1798	0.4755
ARIMAX_manual	0.0971	0.4587	0.2685	0.4811	0.1828	0.4699
AutoARIMAX	0.1050	0.4574	0.2693	0.4799	0.1872	0.4686
SARIMAX	0.1037	0.4578	0.2713	0.4804	0.1875	0.4691
Prophet_regresores	0.1030	0.4627	0.2759	0.4857	0.1895	0.4742
AutoARIMAX_MLP _Hibrido	0.2232	0.4276	0.1624	0.4411	0.1928	0.4343
MLP_multivariado	0.2091	0.4557	0.1847	0.4711	0.1969	0.4634
GRU_multivariado	0.0964	0.4850	0.3334	0.5128	0.2149	0.4989
RandomForest_multivariado	0.2652	0.4096	0.1699	0.4238	0.2175	0.4167
Naive	0.0381	0.5861	0.4041	0.6197	0.2211	0.6029
LSTM_multivariado	0.1260	0.4673	0.3180	0.4938	0.2220	0.4806

Modelo	Dif. volatilidad	MASE prom. volatilidad	Dif. volumen	MASE prom. volumen	Dif. promedio	MASE prom. global
NeuralProphet_multivariado	0.1616	0.4860	0.2876	0.5099	0.2246	0.4979
SNaive	0.2483	0.5753	0.3444	0.6040	0.2964	0.5896

***Nota.** La tabla resume la robustez del error predictivo frente a cambios de volatilidad y volumen. La diferencia promedio mide la sensibilidad del MASE entre regímenes, mientras que el MASE promedio global permite evaluar el nivel general de error. Un modelo robusto debe combinar baja diferencia promedio con bajo MASE promedio global; Elaboración propia con base en los modelos implementados en Python.*

La Tabla 24 permite evaluar la robustez de los modelos a partir de dos elementos: el error promedio global y la variación del error entre condiciones de volatilidad y volumen. Bajo este criterio, no basta con que un modelo cambie poco entre regímenes; también debe conservar un nivel de MASE competitivo.

GAM_multivariado presenta la menor diferencia promedio entre regímenes. Su error varía poco cuando cambian simultáneamente la volatilidad y el volumen, por lo que se destaca por estabilidad. Sin embargo, su MASE promedio global es superior al de los modelos con mejor precisión.

XGBoost_multivariado ofrece un resultado más equilibrado. Mantiene una diferencia moderada entre regímenes y, al mismo tiempo, registra uno de los menores valores de MASE global. Dentro de la muestra evaluada, logra adaptarse a distintos escenarios de mercado sin perder de forma importante su precisión.

RandomForest_multivariado obtiene el menor MASE promedio global. Su fortaleza principal es la reducción del error agregado. No obstante, la diferencia entre regímenes es mayor,

lo que indica que su desempeño es más sensible a cambios en volatilidad y volumen que el de GAM o XGBoost.

Los modelos clásicos, estructurales y de referencia presentan una posición menos favorable. Algunos registran diferencias moderadas entre regímenes, pero sus errores promedio permanecen por encima de los principales modelos multivariados. En Naive y SNaive, una aparente estabilidad no compensa sus valores altos de MASE.

La lectura conjunta de la tabla distingue tres fortalezas: GAM_multivariado es el modelo más estable, RandomForest_multivariado alcanza la mejor precisión promedio y XGBoost_multivariado ofrece el balance más favorable entre ambos criterios. Por esta razón, XGBoost se mantiene como la alternativa recomendada cuando se busca buen desempeño bajo cambios simultáneos de volatilidad y volumen.

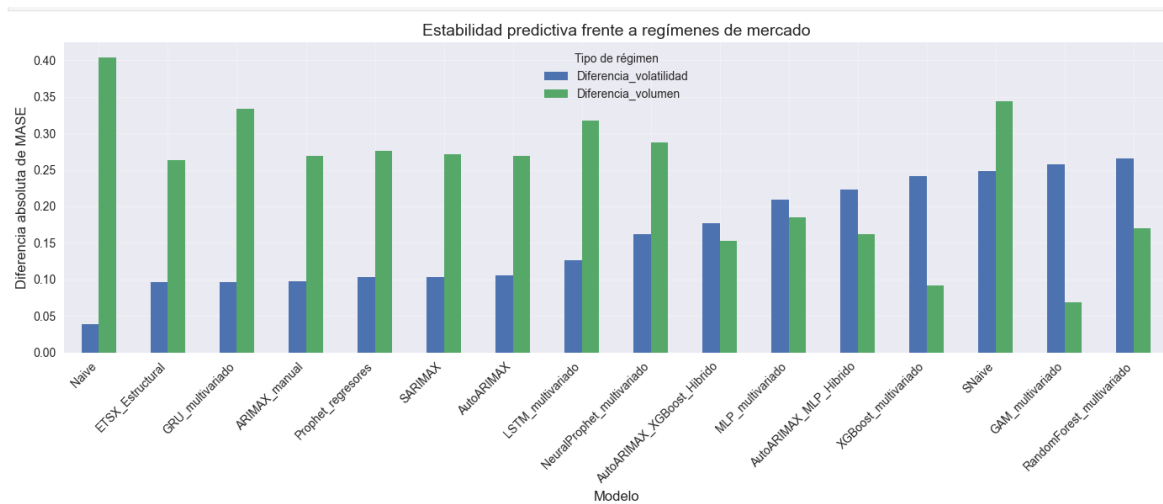
10.4.4.3. Identificación de modelos más estables

La identificación de los modelos más estables se realizó considerando dos criterios complementarios: la diferencia promedio del MASE entre regímenes y el MASE promedio global. La diferencia promedio permite evaluar qué tan sensible es cada modelo frente a cambios de volatilidad y volumen, mientras que el MASE promedio global permite valorar el nivel general de error. Por tanto, un modelo estable no debe entenderse únicamente como aquel que presenta menor variación entre regímenes, sino como aquel que combina estabilidad relativa con buen desempeño predictivo.

Tabla 25*Modelos más estables según MASE frente a cambios de volatilidad y volumen*

Modelo	Dif. volatilidad	MASE prom. volatilidad	Dif. volumen	MASE prom. volumen	Dif. promedio	MASE prom. global
GAM_multivariado	0.2576	0.4310	0.0684	0.4367	0.1630	0.4338
AutoARIMAX_XGB oost_Hibrido	0.1764	0.4278	0.1524	0.4405	0.1644	0.4342
XGBoost_multivaria do	0.2418	0.4238	0.0921	0.4314	0.1670	0.4276
ETX_Estructural	0.0963	0.4645	0.2633	0.4865	0.1798	0.4755
ARIMAX_manual	0.0971	0.4587	0.2685	0.4811	0.1828	0.4699
AutoARIMAX	0.1050	0.4574	0.2693	0.4799	0.1872	0.4686
SARIMAX	0.1037	0.4578	0.2713	0.4804	0.1875	0.4691
Prophet_regresores	0.1030	0.4627	0.2759	0.4857	0.1895	0.4742
AutoARIMAX_MLP _Hibrido	0.2232	0.4276	0.1624	0.4411	0.1928	0.4343
MLP_multivariado	0.2091	0.4557	0.1847	0.4711	0.1969	0.4634

Nota: La tabla resume los modelos más estables, considerando la diferencia promedio del MASE entre regímenes y el MASE promedio global. Los valores se presentan con cuatro decimales; Elaboración propia con base en los modelos implementados en Python.

Figura 15*Estabilidad predictiva del MASE frente a cambios de volatilidad y volumen*

Nota: elaboración propia con base en datos OHLCV de BTC-USD obtenidos de Yahoo Finance; modelos estimados en Python.

La Tabla 25 y la Figura 15 permiten comparar la sensibilidad del MASE frente a cambios de volatilidad y volumen. La estabilidad no depende únicamente de la menor diferencia entre regímenes. Un modelo puede variar poco entre condiciones de mercado y, aun así, mantener un error promedio elevado. Por eso, la lectura combina la diferencia promedio entre regímenes con el MASE global.

GAM_multivariado presenta la menor variación promedio. Su error cambia poco cuando varían conjuntamente la volatilidad y el volumen, por lo que se destaca por estabilidad relativa. Sin embargo, su MASE global es superior al de los modelos más precisos. Su principal fortaleza está en la consistencia del error, no en alcanzar el menor nivel de error promedio.

XGBoost_multivariado combina una variación moderada entre regímenes con uno de los menores valores de MASE global. Este resultado lo ubica entre los modelos más equilibrados, ya que conserva una precisión competitiva sin mostrar una sensibilidad excesiva a los cambios de mercado.

RandomForest_multivariado obtiene el menor MASE promedio global, pero presenta una variación mayor entre regímenes. Su mejor resultado corresponde a la precisión agregada, aunque su desempeño cambia más cuando se modifican las condiciones de volatilidad y volumen.

La Figura 15 muestra además que los modelos no responden de igual forma a ambas dimensiones. Algunos son más sensibles a los cambios de volatilidad y otros a los cambios de volumen. Por ello, la robustez debe evaluarse de manera conjunta y no a partir de un solo indicador.

La comparación deja tres resultados claros: GAM_multivariado es el más estable, RandomForest_multivariado es el más preciso en promedio y XGBoost_multivariado ofrece el balance más favorable entre estabilidad y precisión. Este último se mantiene como una de las alternativas más consistentes para la recomendación final del estudio.

10.4.5. Síntesis del análisis por regímenes

El análisis por regímenes permitió evaluar el desempeño de los modelos bajo condiciones diferenciadas de volatilidad y volumen, con el propósito de identificar si la precisión predictiva y la capacidad direccional se mantienen estables cuando cambian las condiciones del mercado. Este análisis se realizó sobre los pronósticos generados en la validación Hold-Out, clasificando cada observación del periodo de validación según el régimen observado de volatilidad y volumen.

Tabla 26

Mejores modelos por régimen de mercado según MASE

Tipo de régimen	Régimen	Modelo	N_obs	MAE	RMSE	sMAPE (%)	MASE	Accuracy	F1-score	MC C	AU C
Volatilidad	Baja volatilidad	RandomForest_multivariado	30	0.009	0.014	91.265	0.277	0.733	0.600	0.53	0.888

Tipo de régimen	Régimen	Modelo	N_obs	MAE	RMSE	sMAPE (%)	MASE	Accuracy	F1-score	MC C	AUC
Volatilidad	Alta volatilidad	ARIMAX_manual	30	0.018	0.022	154.876	0.507	0.600	0.600	0.20	0.64
Volumen	Bajo volumen	LSTM_multivariado	35	0.011	0.015	151.050	0.334	0.457	0.000	0.00	0.63
Volumen	Alto volumen	GAM_multivariado	25	0.016	0.021	106.840	0.470	0.800	0.705	0.55	0.85

Nota. La tabla resume el mejor modelo por régimen de mercado según MASE. Los valores se presentan con cuatro decimales; Elaboración propia con base en los modelos implementados en Python.

Tabla 27

Modelos líderes según error global y estabilidad entre regímenes

Criterio	Modelo	Dif. volatilidad	MASE prom. volatilidad	Dif. volumen	MASE prom. volumen	Dif. promedio	MASE prom. global
Menor MASE global	RandomForest_multivariado	0.2652	0.4096	0.1699	0.4238	0.2175	0.4167
Menor MASE global	XGBoost_multivariado	0.2418	0.4238	0.0921	0.4314	0.1670	0.4276
Menor MASE global	GAM_multivariado	0.2576	0.4310	0.0684	0.4367	0.1630	0.4338
Menor MASE global	AutoARIMAX_XGBoost_Hibrido	0.1764	0.4278	0.1524	0.4405	0.1644	0.4342
Menor MASE global	AutoARIMAX_MLP_Hibrido	0.2232	0.4276	0.1624	0.4411	0.1928	0.4343

Criterio	Modelo	Dif. volatilidad	MASE prom. volatilidad	Dif. volumen	MASE prom. volumen	Dif. promedio	MASE prom. global
Mayor estabilidad	GAM_multivariado	0.2576	0.4310	0.0684	0.4367	0.1630	0.4338
Mayor estabilidad	AutoARIMAX_ XGBoost_Hibrido	0.1764	0.4278	0.1524	0.4405	0.1644	0.4342
Mayor estabilidad	XGBoost_ multivariado	0.2418	0.4238	0.0921	0.4314	0.1670	0.4276
Mayor estabilidad	ETSX_Estructural	0.0963	0.4645	0.2633	0.4865	0.1798	0.4755
Mayor estabilidad	ARIMAX_manual	0.0971	0.4587	0.2685	0.4811	0.1828	0.4699

Nota. La tabla contrasta los modelos líderes por menor MASE promedio global y por mayor estabilidad entre regímenes; Elaboración propia con base en los modelos implementados en Python.

La Tabla 26 muestra que el modelo con menor MASE cambia según la condición de mercado. En baja volatilidad, RandomForest_multivariado obtiene el mejor resultado; en alta volatilidad, el menor error corresponde a ARIMAX_manual. Por volumen, LSTM_multivariado lidera en bajo volumen y GAM_multivariado en alto volumen. Estos resultados puntuales indican que ningún modelo domina todos los escenarios.

La comparación de la Tabla 27 ayuda a poner estos resultados en contexto. RandomForest_multivariado presenta el menor MASE promedio global, seguido por XGBoost_multivariado y GAM_multivariado. Sin embargo, GAM_multivariado y AutoARIMAX_XGBoost_Híbrido registran las menores diferencias promedio entre regímenes. Tienen mayor estabilidad relativa, aunque su error global es superior al de los modelos más precisos.

Por tanto, precisión y estabilidad no son el mismo criterio. Un modelo puede obtener el menor error promedio y ser más sensible a los cambios de volatilidad o volumen. También puede variar poco entre regímenes, pero mantener un error elevado. La selección final debe considerar ambos aspectos, además del comportamiento observado en cada escenario particular.

XGBoost_multivariado presenta un resultado favorable en esta comparación. No tiene el menor MASE global ni la menor variación entre regímenes, pero combina un error promedio bajo con una diferencia moderada entre condiciones de mercado. Este balance resulta relevante para Bitcoin, donde la volatilidad y la actividad de negociación pueden cambiar rápidamente.

El análisis por condiciones de mercado deja tres hallazgos. Primero, los modelos multivariados obtienen mejores resultados promedio que los modelos de referencia. Segundo, el ranking cambia según volatilidad y volumen, por lo que no conviene elegir un modelo solo por un resultado puntual. Finalmente, XGBoost_multivariado mantiene uno de los balances más favorables entre precisión, estabilidad y uso de información exógena dentro de la muestra Hold-Out analizada.

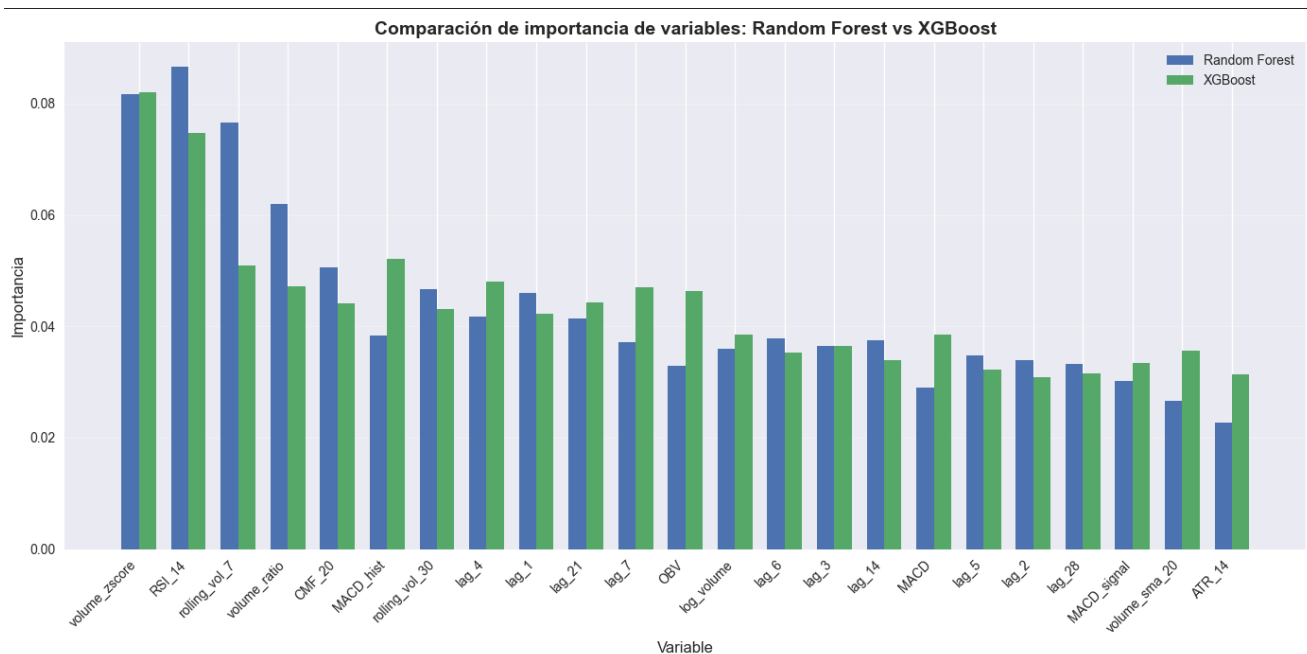
10.5. Implicaciones estadísticas y financieras

Los resultados obtenidos tienen implicaciones relevantes desde el punto de vista estadístico y financiero. En términos estadísticos, el estudio muestra que la incorporación de variables exógenas asociadas al volumen, la volatilidad, el momentum, la tendencia y el flujo de mercado permitió mejorar descriptivamente el desempeño predictivo frente a modelos basados únicamente en la historia del retorno. Esta conclusión respalda el enfoque multivariado propuesto, especialmente en un activo como Bitcoin, caracterizado por alta volatilidad, cambios de régimen y relaciones no lineales entre variables de mercado.

La comparación de importancia de variables entre Random Forest y XGBoost, presentada en la Figura 16, permite complementar la evaluación predictiva con una lectura de interpretabilidad. Variables como `volume_zscore`, `volume_ratio`, `rolling_vol_7`, `rolling_vol_30`, `RSI_14`, `MACD_hist`, `OBV` y `CMF_20` aparecen entre los predictores más relevantes, lo que sugiere que el desempeño de los modelos no depende únicamente de la memoria histórica del retorno, sino también de señales asociadas con actividad transaccional, volatilidad reciente, momentum técnico y flujo de capital.

Figura 16

Comparación de importancia de variables: Random Forest vs XGBoost



Nota: elaboración propia con base en datos OHLCV de BTC-USD obtenidos de Yahoo Finance; modelos estimados en Python.

Tabla 28*Lectura financiera de variables relevantes del enfoque multivariado*

Bloque financiero	Variable	Lectura financiera	Aporte al modelo multivariado
Volumen	Volume _zscore	Captura desviaciones anormales del volumen frente a su comportamiento reciente.	Permite detectar episodios de actividad extraordinaria del mercado.
Volumen	Volume _ratio	Mide si el volumen actual está por encima o por debajo de su media móvil reciente.	Ayuda a identificar cambios relativos en la liquidez y actividad del mercado.
Volatilidad intradía	ATR_1 4	Representa la amplitud intradía y la intensidad de los movimientos del mercado.	Aporta información sobre riesgo, expansión de rangos y presión de volatilidad.
Momentum	RSI_14	Identifica fuerza relativa, sobrecompra, sobreventa y posibles señales de reversión.	Contribuye a capturar señales técnicas de corto plazo.
Momentum / tendencia	MACD _hist	Resume aceleración o desaceleración del momentum de corto plazo.	Aporta señales asociadas a cambios en la fuerza de la tendencia.
Flujo de capital	OBV	Aproxima acumulación o distribución de volumen según la dirección del precio.	Incorpora información acumulada de flujo de mercado.
Presión compradora / vendedora	CMF_2 0	Mide presión compradora o vendedora combinando precio, rango intradía y volumen.	Permite aproximar acumulación, distribución y presión de mercado.

Nota. La tabla resume la lectura financiera de las variables relevantes del enfoque multivariado y su aporte conceptual al pronóstico del retorno logarítmico de BTC-USD; Elaboración propia con base en los modelos implementados en Python.

Esta lectura se complementa con el Anexo 6, ****Interpretabilidad e importancia de variables en modelos multivariados****. Allí se presentan los resultados de importancia de variables para MLP, Random Forest y XGBoost, el ranking global de predictores y la lectura de los regresores

exógenos utilizados en Prophet. El anexo conserva el detalle técnico del análisis sin recargar el capítulo principal con tablas extensas.

Los resultados muestran que el volumen y los indicadores técnicos aportan información útil para describir la dinámica de corto plazo de Bitcoin. Esta utilidad debe entenderse en términos predictivos, no causales. Que una variable tenga mayor importancia dentro de un modelo no significa que cause directamente el movimiento futuro del precio; indica que ayudó a mejorar la representación estadística del mercado dentro del periodo y los esquemas de validación evaluados.

La Tabla 27 traduce los predictores más relevantes a una lectura financiera. Variables como `volume\_zscore` y `volume\_ratio` representan cambios en la actividad de negociación; `ATR\_14` recoge volatilidad reciente; `RSI\_14` y `MACD\_hist` se relacionan con momentum y tendencia; mientras que `OBV` y `CMF\_20` aportan información sobre flujo de capital y presión compradora o vendedora. Esta agrupación facilita relacionar la importancia estadística de las variables con dimensiones observables del mercado.

La evaluación tampoco debe reducirse a una única métrica. El modelo con menor MASE no siempre fue el mejor en dirección ni el más estable entre condiciones de volatilidad y volumen. Por ello, la comparación incluyó MASE, métricas direccionales, prueba de Diebold-Mariano, análisis por condiciones de mercado e interpretabilidad de variables.

Los modelos pueden servir como apoyo al análisis cuantitativo, pero no como reglas automáticas de inversión. Bitcoin mantiene alta volatilidad y movimientos abruptos, y las diferencias entre precisión en magnitud y desempeño direccional exigen cautela. Un modelo con bajo MASE ayuda a aproximar el tamaño esperado del retorno; otro con mejores métricas direccionales puede ser más útil cuando el interés se centra en anticipar su signo.

La recomendación de XGBoost_multivariado_RO responde a esa lectura conjunta. RandomForest_multivariado obtuvo resultados destacados por error promedio y GAM_multivariado mostró fortalezas en estabilidad y dirección. XGBoost, sin embargo, ofreció el balance más favorable entre precisión, estabilidad relativa y uso de información exógena dentro del diseño del estudio.

En conjunto, el enfoque multivariado aporta valor al pronóstico del retorno logarítmico de BTC-USD. Permite comparar modelos frente a referencias simples, revisar su comportamiento bajo distintos esquemas de validación y analizar el papel de variables de volumen, volatilidad, momentum y flujo de mercado. Sus resultados deben utilizarse como apoyo analítico, no como una señal directa de compra o venta.

10.6. Contraste con la literatura previa

Los resultados del estudio se ubican dentro de la literatura reciente sobre predicción de Bitcoin y criptomonedas. En conjunto, coinciden con la idea de que pronosticar retornos financieros en estos mercados es difícil debido a la alta volatilidad, la baja persistencia lineal, los cambios frecuentes en las condiciones de mercado y la posible presencia de relaciones no lineales. Kristoufek (2015), Catania, Grassi y Ravazzolo (2019), y Bouri, Gil-Alana, Gupta y Roubaud (2020) describen precisamente un mercado con alta variabilidad e inestabilidad temporal, donde los modelos contruidos solo con el comportamiento pasado del precio o del retorno pueden resultar insuficientes.

Los resultados también son consistentes con los estudios que incorporan información distinta al precio. Vidal-Tomás e Ibáñez (2023) analizan la relación dinámica entre precio y volumen de Bitcoin; Anastasiou, Ballis y Drakos (2023) estudian señales de tendencia en el pronóstico de criptomonedas; y Guo, Sang, Tu y Wang (2024) encuentran relaciones predictivas

entre retornos de distintos criptoactivos. En este trabajo, la comparación descriptiva fuera de muestra muestra menores errores para los modelos multivariados en los escenarios donde se contrastaron directamente con los univariados. Las variables de volumen, volatilidad histórica, RSI, MACD, OBV y CMF aportaron información útil dentro del diseño aplicado. Este resultado no representa una prueba formal de superioridad agregada entre enfoques.

La literatura también ha explorado el uso de machine learning y deep learning para representar patrones no lineales en mercados de criptomonedas. McNally, Roche y Caton (2018), Lahmiri y Bekiros (2020) y Lahmiri y Bekiros (2023) reflejan ese interés. En el presente estudio, Random Forest y XGBoost mostraron resultados competitivos frente a los modelos clásicos y frente a las arquitecturas de deep learning evaluadas. Esta lectura es coherente con la capacidad de los métodos basados en árboles y boosting para capturar interacciones complejas entre predictores, como señalan Chen y Guestrin (2016) y Hastie, Tibshirani y Friedman (2017).

La comparación con estudios previos tiene límites claros. Los resultados publicados no son directamente equivalentes porque cambian el periodo analizado, la frecuencia de los datos, la variable objetivo, el horizonte de pronóstico, las variables explicativas, las métricas y los esquemas de validación. Por ello, el propósito no es afirmar que los modelos de este proyecto superen numéricamente a los reportados en otros trabajos, sino ubicar los hallazgos dentro de una línea de investigación que reconoce la complejidad de Bitcoin y la utilidad potencial de los enfoques multivariados.

El aporte del proyecto consiste en evaluar, dentro de un mismo pipeline reproducible, modelos clásicos, modelos con regresores exógenos, algoritmos de machine learning, arquitecturas de deep learning, modelos aditivos y modelos híbridos para pronosticar el retorno logarítmico de BTC-USD. La comparación integra Hold-Out y Rolling-Origin, métricas de error y de dirección,

análisis por condiciones de volatilidad y volumen, e interpretabilidad de variables. De este modo, el estudio aporta evidencia propia sobre el papel del volumen, la volatilidad y los indicadores técnicos, limitada al periodo, los datos y las decisiones metodológicas consideradas.

10.7. Limitaciones del estudio

Aunque el estudio permitió comparar un conjunto amplio de modelos univariados y multivariados para pronosticar el retorno logarítmico de Bitcoin, los resultados deben interpretarse dentro de varios límites metodológicos y prácticos.

El análisis se realizó únicamente sobre BTC-USD. Bitcoin es el criptoactivo con mayor relevancia de mercado, pero sus resultados no pueden extenderse automáticamente a otras criptomonedas. Cada activo puede presentar condiciones distintas de liquidez, volumen, volatilidad, capitalización y comportamiento especulativo (Catania, Grassi, & Ravazzolo, 2019; Bouri, Gil-Alana, Gupta, & Roubaud, 2020).

Las variables exógenas se construyeron a partir de información de mercado e indicadores técnicos: volumen, volatilidad histórica, momentum, tendencia y flujo de capital. No se incluyeron variables macroeconómicas, noticias, sentimiento de mercado, actividad en redes sociales, flujos institucionales ni información on-chain. El conjunto de predictores representa, por tanto, una aproximación basada en datos OHLCV, pero no cubre todas las fuentes que podrían contener información útil para el pronóstico.

Bitcoin también puede reaccionar a eventos extremos, cambios regulatorios, crisis de liquidez o episodios especulativos que no siempre quedan reflejados en la información histórica disponible. Incluso un buen desempeño fuera de muestra puede deteriorarse ante una ruptura estructural o un cambio abrupto en las condiciones del mercado (Shumway & Stoffer, 2025; Catania, Grassi, & Ravazzolo, 2019).

El análisis por condiciones de volatilidad y volumen se aplicó solo a los pronósticos de la validación Hold-Out. La clasificación permitió comparar el desempeño condicional en una misma ventana fuera de muestra, pero tiene carácter descriptivo. Los umbrales utilizados —la mediana de `rolling\_vol\_30` y `volume\_ratio = 1`— son operativos y reproducibles, aunque no sustituyen métodos formales como cambios de régimen de Markov, clustering dinámico, umbrales endógenos o modelos de volatilidad condicional.

Algunos resultados también pueden depender de las decisiones de modelación. MLP, LSTM, GRU, Random Forest y XGBoost son sensibles a hiperparámetros, rezagos, escalamiento, arquitectura, regularización y criterios de entrenamiento. Una optimización más extensa podría modificar el desempeño relativo de algunas familias de modelos (Hastie, Tibshirani, & Friedman, 2017; Chen & Guestrin, 2016; Hochreiter & Schmidhuber, 1997).

La interpretabilidad se desarrolló con mayor detalle para los modelos multivariados de machine learning, principalmente MLP, Random Forest y XGBoost. En los modelos clásicos, estructurales, aditivos y de deep learning, el análisis de predictores requeriría herramientas específicas, como coeficientes, funciones parciales, análisis de sensibilidad o valores SHAP. Estas alternativas no se abordaron en profundidad.

El estudio evalúa desempeño estadístico y direccional, pero no incluye un backtesting financiero formal. Los resultados no representan evidencia directa de rentabilidad ni una regla de inversión. Para evaluar una estrategia sería necesario incorporar costos de transacción, deslizamiento, liquidez, tamaño de posición, gestión del riesgo y reglas explícitas de entrada y salida.

Por último, la prueba de Diebold-Mariano permitió contrastar diferencias de precisión frente al modelo Naive en los casos donde fue aplicada. Sus resultados dependen del periodo y de la muestra de validación utilizada. La comparación entre los enfoques multivariado y univariado se presenta como un contraste empírico descriptivo basado en MASE, no como una prueba formal de superioridad agregada (Diebold & Mariano, 1995; Hyndman & Koehler, 2006; Hyndman & Athanasopoulos, 2021).

Estas consideraciones delimitan el alcance del estudio y orientan futuras extensiones: incorporar otros criptoactivos y fuentes de información, aplicar métodos formales de identificación de condiciones de mercado, profundizar la optimización de modelos y evaluar estrategias de decisión bajo criterios financieros reales.

11. Conclusiones

El proyecto evaluó el aporte del volumen de negociación, la volatilidad y los indicadores técnicos al pronóstico del retorno logarítmico diario de Bitcoin (BTC-USD). Se compararon modelos univariados y multivariados mediante validación Hold-Out y Rolling-Origin, con horizontes de uno y siete días.

La variable objetivo fue el retorno logarítmico diario de BTC-USD. Esta transformación permitió trabajar con variaciones relativas del precio, en lugar de precios en niveles, lo cual resulta adecuado ante los cambios de escala y la alta volatilidad observada en Bitcoin. A partir de esta variable se construyó un conjunto de predictores con rezagos del retorno, variables de volumen, medidas de volatilidad, indicadores de momentum, tendencia y flujo de mercado.

En los esquemas donde se realizó una comparación directa, los modelos multivariados obtuvieron menores valores de MASE que los mejores modelos univariados. Random Forest y XGBoost fueron las alternativas más destacadas por error de magnitud. Dentro de la muestra y los

esquemas evaluados, las variables de volumen, volatilidad, momentum y flujo de mercado aportaron información adicional para aproximar el retorno logarítmico de Bitcoin.

Esta lectura debe entenderse como evidencia empírica descriptiva fuera de muestra. La comparación entre los enfoques univariado y multivariado no se interpretó como una prueba estadística formal de superioridad agregada. Los resultados se limitan al periodo, los modelos, las variables y los esquemas de validación utilizados.

La comparación entre Hold-Out y Rolling-Origin mostró que el ranking de los modelos puede cambiar según el procedimiento de evaluación. Algunos modelos que obtuvieron buenos resultados en una ventana fija no conservaron la misma posición al actualizar progresivamente el origen del pronóstico. Rolling-Origin permitió observar esa variación temporal y aportó una evaluación más exigente del desempeño diario.

El análisis direccional mostró que el modelo con menor error de magnitud no siempre fue el más adecuado para anticipar el signo del retorno. Los modelos multivariados de machine learning destacaron por MASE, mientras que GAM multivariado y otros modelos mostraron fortalezas direccionales en escenarios específicos. Por ello, la selección no se basó solo en error, sino también en métricas de dirección y estabilidad temporal.

El análisis por condiciones de volatilidad y volumen mostró que los errores fueron menores en observaciones de baja volatilidad y bajo volumen. Los escenarios de alta volatilidad y alto volumen resultaron más exigentes. Además, el mejor modelo puntual cambió según la condición de mercado, lo que muestra que el desempeño de Bitcoin no puede evaluarse con una única condición promedio.

RandomForest_multivariado obtuvo el menor MASE global, pero mostró mayor variación entre regímenes. GAM_multivariado presentó una estabilidad relativa más alta, aunque con un

error promedio mayor. XGBoost_multivariado ofreció el balance más favorable entre bajo error, estabilidad razonable y uso de información exógena. Por esta razón, XGBoost_multivariado_RO se mantiene como el modelo final recomendado.

El análisis de importancia de variables reforzó el valor del enfoque multivariado. Predictores como `volume\_zscore`, `volume\_ratio`, `rolling\_vol\_7`, `rolling\_vol\_30`, `RSI\_14`, `MACD\_hist`, `OBV` y `CMF\_20` aportaron información relacionada con actividad anormal de negociación, volatilidad reciente, momentum, tendencia y presión compradora o vendedora. Estas relaciones deben interpretarse en términos predictivos, no causales.

La prueba de Diebold-Mariano permitió contrastar algunos modelos frente al benchmark Naive. En varias comparaciones se encontraron diferencias estadísticamente significativas. Esta evidencia complementa las métricas descriptivas, pero no reemplaza la evaluación integral ni demuestra una superioridad práctica en todos los criterios.

En conjunto, el estudio muestra que el pronóstico del retorno logarítmico de Bitcoin requiere combinar métricas de error, métricas direccionales, validación temporal, análisis por condiciones de mercado e interpretabilidad de variables. La volatilidad y los cambios frecuentes del mercado impiden asumir que un único modelo será el mejor en todos los escenarios.

El enfoque multivariado constituye una alternativa metodológica útil para el análisis predictivo de BTC-USD. No representa una regla directa de inversión ni evidencia causal, pero permite ampliar la información disponible para el pronóstico. Dentro del diseño aplicado, XGBoost_multivariado_RO ofrece el balance más consistente entre precisión, estabilidad relativa y aprovechamiento de variables exógenas.

12. Recomendaciones y trabajos futuros

A partir de los resultados, una primera línea de trabajo consiste en ampliar las fuentes de información utilizadas en el pronóstico. El estudio se concentró en variables de precio, volumen, volatilidad e indicadores técnicos. Futuras investigaciones podrían incluir variables macroeconómicas y financieras, como tasas de interés, inflación, índices bursátiles, liquidez global o indicadores del entorno financiero. Esto permitiría revisar si Bitcoin responde únicamente a señales internas del mercado cripto o también a condiciones externas (Kristoufek, 2015; Bouri, Gil-Alana, Gupta, & Roubaud, 2020).

También sería útil incorporar variables de atención y sentimiento de mercado. Noticias financieras, búsquedas en internet, redes sociales e indicadores de percepción del riesgo pueden aportar señales sobre expectativas de los inversionistas. Estas fuentes podrían ser especialmente relevantes en episodios de alta volatilidad o actividad transaccional intensa (Kristoufek, 2015; Anastasiou, Ballis, & Drakos, 2023).

En relación con las condiciones de mercado, una extensión natural es utilizar procedimientos formales de identificación de cambios de régimen. En este proyecto, las condiciones de volatilidad y volumen se definieron con `'rolling_vol_30'` y `'volume_ratio'`, mediante umbrales operativos. Trabajos posteriores podrían emplear modelos de volatilidad condicional, correlación dinámica, cambio de régimen o métodos de clasificación más flexibles. Estas alternativas permitirían estudiar con mayor detalle la agrupación temporal de la volatilidad y las variaciones en la relación entre predictores (Engle, 1982; Bollerslev, 1986; Engle, 2002; Bouri et al., 2020).

También conviene ampliar la validación temporal. Además de Hold-Out y Rolling-Origin para uno y siete días, se podrían evaluar ventanas móviles, ventanas expansivas, varios periodos

de prueba y horizontes de 3, 14 o 30 días. Esto ayudaría a examinar el desempeño de los modelos bajo distintos plazos y condiciones de mercado, siempre respetando el orden temporal de los datos (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025).

Los modelos de machine learning y deep learning podrían beneficiarse de una búsqueda más extensa de hiperparámetros. XGBoost, Random Forest, MLP, LSTM y GRU dependen de decisiones sobre arquitectura, regularización, número de rezagos, tamaño de ventana y criterios de entrenamiento. Una optimización temporalmente validada permitiría revisar si el ranking entre las familias de modelos se mantiene o cambia bajo configuraciones alternativas (Hastie, Tibshirani, & Friedman, 2017; Chen & Guestrin, 2016; Hochreiter & Schmidhuber, 1997; Lahmiri & Bekiros, 2023).

Otra línea de trabajo es profundizar la interpretabilidad. En los modelos basados en árboles podrían utilizarse análisis de contribución de variables; en Prophet, NeuralProphet y GAM, revisar de forma detallada los componentes, regresores y funciones suaves; y en modelos clásicos con variables exógenas, analizar coeficientes y efectos dinámicos. Esto permitiría entender con mayor detalle el papel de `volume\_zscore`, `volume\_ratio`, `RSI\_14`, `MACD\_hist`, `OBV` y `CMF\_20` en cada familia de modelos (Chen & Guestrin, 2016; Taylor & Letham, 2017; Triebe et al., 2021; Lahmiri & Bekiros, 2023).

La extensión a otros criptoactivos también resulta relevante. Ethereum, Binance Coin, Solana y otros activos presentan condiciones distintas de liquidez, capitalización y volatilidad. Un análisis multiactivo permitiría establecer si los resultados obtenidos para BTC-USD son particulares de Bitcoin o se mantienen en otros mercados, e incluso revisar posibles relaciones predictivas entre criptomonedas (Catania, Grassi, & Ravazzolo, 2019; Guo, Sang, Tu, & Wang, 2024).

Desde una perspectiva aplicada, sería necesario complementar las métricas estadísticas con un backtesting financiero formal. Este estudio no evaluó una estrategia de inversión. Trabajos posteriores podrían incorporar reglas de entrada y salida, costos de transacción, deslizamiento, liquidez, tamaño de posición y medidas de riesgo. Así sería posible determinar si las mejoras predictivas se traducen en utilidad económica.

Finalmente, se recomienda avanzar hacia modelos probabilísticos y estimaciones de incertidumbre. En un activo tan volátil como Bitcoin, intervalos de predicción, escenarios de riesgo o distribuciones predictivas pueden ser más informativos que una estimación puntual del retorno (Hyndman & Athanasopoulos, 2021; Shumway & Stoffer, 2025).

Estas extensiones deberían conservar la lógica de evaluación integral aplicada en el proyecto: comparación frente a modelos de referencia, validación temporal rigurosa, métricas de error y dirección, análisis de estabilidad e interpretación de variables. La prueba de Diebold-Mariano puede seguir utilizándose para contrastes puntuales de precisión frente a benchmarks, sin sustituir la lectura conjunta del desempeño predictivo (Diebold & Mariano, 1995; Hyndman & Koehler, 2006; Makridakis, Spiliotis, & Assimakopoulos, 2020, 2022).

Referencias Bibliográficas

- Anastasiou, D., Ballis, A., & Drakos, K. (2023). Trend-based forecast of cryptocurrency returns. *Economic Modelling*, 118, 106151. <https://doi.org/10.1016/j.econmod.2023.106151>
- Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), 307–327. [https://doi.org/10.1016/0304-4076\(86\)90063-1](https://doi.org/10.1016/0304-4076(86)90063-1)
- Bouri, E., Gil-Alana, L. A., Gupta, R., & Roubaud, D. (2020). Predictability of Bitcoin returns: Evidence from robust tests. *The European Journal of Finance*, 26(14), 1320–1339. <https://doi.org/10.1080/1351847X.2020.1835685>
- Bouri, E., Shahzad, S. J. H., Roubaud, D., Kristoufek, L., & Lucey, B. (2019). RETRACTED: Is Bitcoin a better safe-haven investment than gold and commodities? *International Review of Financial Analysis*, 63, 322–330. <https://doi.org/10.1016/j.irfa.2019.01.002>
- Catania, L., Grassi, S., & Ravazzolo, F. (2019). Forecasting cryptocurrencies under model and parameter instability. *International Journal of Forecasting*, 35(2), 485–501. <https://doi.org/10.1016/j.ijforecast.2018.09.005>
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). <https://doi.org/10.1145/2939672.2939785>
- Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263. <https://doi.org/10.1080/07350015.1995.10524599>
- Engle, R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50(4), 987–1007. <https://doi.org/10.2307/1912773>

- Engle, R. F. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3), 339–350. <https://doi.org/10.1198/073500102288618487>
- Guo, L., Sang, B., Tu, J., & Wang, Y. (2024). Cross-cryptocurrency return predictability. *Journal of Economic Dynamics and Control*, 163, 104863. <https://doi.org/10.1016/j.jedc.2024.104863>
- Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. <https://doi.org/10.1007/978-0-387-84858-7>
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. <https://otexts.com/fpp3/>
- Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), 679–688. <https://doi.org/10.1016/j.ijforecast.2006.03.001>
- Kristoufek, L. (2015). What are the main drivers of the Bitcoin price? Evidence from wavelet coherence analysis. *PLOS ONE*, 10(4), e0123923. <https://doi.org/10.1371/journal.pone.0123923>
- Lahmiri, S., & Bekiros, S. (2020). Cryptocurrency forecasting with deep learning chaotic neural networks. *Chaos, Solitons & Fractals*, 134, 109721. <https://doi.org/10.1016/j.chaos.2020.109721>

- Lahmiri, S., & Bekiros, S. (2023). Forecasting cryptocurrency returns with machine learning models. *Expert Systems with Applications*, 213, 118849. <https://doi.org/10.1016/j.eswa.2023.118849>
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 36(1), 54–74. <https://doi.org/10.1016/j.ijforecast.2019.04.014>
- Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2022). The M5 accuracy competition: Results, findings, and conclusions. *International Journal of Forecasting*, 38(4), 1346–1364. <https://doi.org/10.1016/j.ijforecast.2021.11.013>
- McNally, S., Roche, J., & Caton, S. J. (2018). Predicting the price of Bitcoin using machine learning. In *Proceedings of the 26th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP)*. <https://doi.org/10.1109/PDP2018.2018.00060>
- Mudassir, M., Bennbaia, S., Unal, D., & Hammoudeh, M. (2020). Time-series forecasting of Bitcoin prices using high-dimensional features: A machine learning approach. *Neural Computing and Applications*, 37, 22979–22993. <https://doi.org/10.1007/s00521-020-05129-6>
- OpenAI. (2026). ChatGPT (GPT-5.5 Thinking) [Modelo de lenguaje de inteligencia artificial]. <https://chat.openai.com/>
- Shmueli, G. (2010). To explain or to predict? *Statistical Science*, 25(3), 289–310. <https://doi.org/10.1214/10-STS330>
- Shumway, R. H., & Stoffer, D. S. (2025). *Time series analysis and its applications: With R examples* (4th ed.). Springer. <https://doi.org/10.1007/978-3-319-52452-8>

Taylor, S. J., & Letham, B. (2017). Forecasting at scale. *The American Statistician*, 72(1), 37–45.

https://facebook.github.io/prophet/static/prophet_paper_20170113.pdf

Triebe, O., Hewamalage, H., Pilyugina, P., Laptev, N., Bergmeir, C., & Rajagopal, R. (2021).

NeuralProphet: Explainable forecasting at scale. *arXiv preprint arXiv:2111.15397*.

<https://arxiv.org/abs/2111.15397>

Vidal-Tomás, D., & Ibáñez, A. M. (2023). Re-examining Bitcoin's price–volume relationship: A time-varying approach. *Journal of Risk and Financial Management*, 16(7), 324.

<https://doi.org/10.3390/jrfm16070324>

Apéndices

Apéndice A. Comparación entre correlaciones Pearson y Spearman

Tabla A 1

Comparación entre correlaciones de Pearson y Spearman para variables explicativas seleccionadas

Índice	Variable_1	Variable_2	Pearson	Spearman	Diferencia
5	log_volume	volume_ratio	0.149	0.266	0.117
16	ATR_14	log_volume	0.600	0.717	0.117
6	log_volume	volume_zscore	0.169	0.277	0.108
39	MACD	RSI_14	0.552	0.658	0.106
57	MACD_hist	RSI_14	0.569	0.655	0.086
41	MACD	MACD_hist	0.312	0.226	0.085
49	MACD_signal	RSI_14	0.416	0.501	0.085
74	CMF_20	MACD	0.505	0.585	0.080
75	CMF_20	MACD_signal	0.440	0.516	0.076
25	ATR_14	OBV	0.760	0.832	0.072
65	OBV	rolling_vol_30	0.283	0.221	0.062
64	OBV	rolling_vol_7	0.188	0.126	0.061
54	MACD_hist	volume_zscore	0.105	0.158	0.053
24	ATR_14	MACD_hist	0.150	0.098	0.052
36	MACD	volume_zscore	0.131	0.180	0.049

Nota: elaboración propia con base en los resultados obtenidos en Python.

Apéndice B. Resultados Hold-Out univariados ordenados por MASE**Tabla B 1***Métricas de error de modelos univariados Hold-Out*

Modelo	Familia	MAE	RMSE	sMAPE (%)	MASE
AutoARIMA_XGBoost	Híbrido	0.017	0.022	175.989	0.459
AutoARIMA_MLP	Híbrido	0.017	0.022	158.683	0.459
MLP_optimo	Machine Learning	0.017	0.022	158.563	0.460
NeuralProphet	Aditivo	0.017	0.022	170.644	0.461
Theta	Clásico	0.017	0.022	194.274	0.463
SARIMA_manual	Clásico	0.017	0.022	199.510	0.463
RandomForest	Machine Learning	0.017	0.022	176.577	0.465
GAM	Aditivo	0.017	0.022	188.529	0.466
AutoETS	Clásico	0.017	0.022	171.119	0.468
XGBoost	Machine Learning	0.017	0.022	170.956	0.468
GRU	Deep Learning	0.017	0.022	169.808	0.468
LSTM	Deep Learning	0.017	0.022	169.556	0.468
AutoARIMA	Clásico	0.017	0.022	172.150	0.468
Prophet	Aditivo	0.017	0.022	166.412	0.470
SNaive	Referencia	0.020	0.027	134.656	0.563
Naive	Referencia	0.021	0.026	144.027	0.574

Nota: elaboración propia con base en los resultados obtenidos en Python.

Tabla B 2*Métricas direccionales y contraste Diebold-Mariano de modelos univariados Hold-Out*

Modelo	Familia	Accuracy	Precision	Recall	F1-score	DM Stat.	DM value	p-value
AutoARIMA_XGBoost	Híbrido	0.533	0.500	0.071	0.125	-3.156	0.002	
AutoARIMA_MLP	Híbrido	0.550	1.000	0.036	0.069	-2.816	0.005	
MLP_optimo	Machine Learning	0.550	1.000	0.036	0.069	-2.810	0.005	
NeuralProphet	Aditivo	0.550	0.516	0.571	0.542	-3.273	0.001	
Theta	Clásico	0.467	0.467	1.000	0.636	-3.322	0.001	
SARIMA_manual	Clásico	0.433	0.400	0.429	0.414	-3.308	0.001	
RandomForest	Machine Learning	0.467	0.466	0.964	0.628	-3.461	0.001	
GAM	Aditivo	0.433	0.429	0.643	0.514	-3.384	0.001	
AutoETS	Clásico	0.467	0.467	1.000	0.636	-3.561	0.000	
XGBoost	Machine Learning	0.467	0.467	1.000	0.636	-3.564	0.000	
GRU	Deep Learning	0.467	0.467	1.000	0.636	-3.583	0.000	
LSTM	Deep Learning	0.467	0.467	1.000	0.636	-3.588	0.000	
AutoARIMA	Clásico	0.467	0.467	1.000	0.636	-3.565	0.000	
Prophet	Aditivo	0.533	0.500	0.821	0.622	-3.526	0.000	

Modelo	Familia	Accur acy	Precision	Recall	F1- score	DM Stat.	DM value	p-
SNaive	Referen cia	0.567	0.538	0.500	0.519	0.088	0.930	
Naive	Referen cia	0.467	0.467	1.000	0.636	NaN	NaN	

Nota: elaboración propia con base en los resultados obtenidos en Python.

Apéndice C. Resultados Hold-Out de modelos multivariados

Tabla C 1

Métricas de error de modelos multivariados Hold-Out

Modelo	Ran king	Familia	MAE	RMSE	sMAPE (%)	MASE
RandomForest_multivariado	1	Machine Learning multivariado	0.015	0.020	122.857	0.410
XGBoost_multivariado	2	Machine Learning multivariado	0.015	0.019	125.847	0.424
AutoARIMAX_MLP_Hibrido	3	Híbrido multivariado	0.015	0.020	132.449	0.428
AutoARIMAX_XGBoost_Hibrido	4	Híbrido multivariado	0.015	0.019	134.056	0.428
GAM_multivariado	5	Aditivo multivariado	0.015	0.020	121.125	0.431
MLP_multivariado	6	Machine Learning multivariado	0.016	0.021	146.527	0.456
AutoARIMAX	7	Clásico multivariado	0.016	0.021	153.010	0.457
SARIMAX	8	Clásico multivariado	0.016	0.021	153.136	0.458
ARIMAX_manual	9	Clásico multivariado	0.016	0.021	154.163	0.459
Prophet_regresores	10	Aditivo multivariado	0.016	0.021	153.020	0.463
ETSX_Estructural	11	Clásico multivariado	0.016	0.021	154.940	0.465

Modelo	Ran king	Familia	MAE	RMSE	sMAPE (%)	MASE
LSTM_multivariado	12	Deep Learning multivariado	0.017	0.022	152.285	0.467
GRU_multivariado	13	Deep Learning multivariado	0.017	0.022	163.959	0.485
NeuralProphet_multivariado	14	Aditivo multivariado	0.017	0.022	140.253	0.486
SNaive	15	Baseline	0.020	0.027	134.656	0.577
Naive	16	Baseline	0.021	0.026	144.027	0.587

Nota: elaboración propia con base en los resultados obtenidos en Python.

Tabla C 2

Métricas direccionales y contraste Diebold-Mariano de modelos multivariados Hold-Out

Modelo	Accur acy	F1- score	MCC	AUC	DM Stat.	DM p- value	DM Comparación
RandomForest_multivariado	0.600	0.500	0.189	0.674	-2.446	0.014	RandomForest_multivariado vs Naive
XGBoost_multivariado	0.717	0.667	0.431	0.713	-2.528	0.011	XGBoost_multivariado vs Naive
AutoARIMAX_MLP_Hibrido	0.617	0.610	0.236	0.633	-2.714	0.007	AutoARIMAX_MLP_Hibrido vs Naive
AutoARIMAX_XGBoost_Hibrido	0.550	0.372	0.077	0.654	-2.527	0.012	AutoARIMAX_XGBoost_Hibrido vs Naive
GAM_multivariado	0.733	0.714	0.464	0.786	-2.658	0.008	GAM_multivariado vs Naive
MLP_multivariado	0.517	0.580	0.062	0.586	-2.264	0.024	MLP_multivariado vs Naive
AutoARIMAX	0.633	0.560	0.259	0.631	-3.397	0.001	AutoARIMAX vs Naive

Modelo	Accur acy	F1- score	MCC	AUC	DM Stat.	DM p- value	DM Comparación
SARIMAX	0.633	0.560	0.259	0.635	-3.440	0.001	SARIMAX vs Naive
ARIMAX_manual	0.600	0.520	0.189	0.635	-3.450	0.001	ARIMAX_manual vs Naive
Prophet_regresores	0.567	0.536	0.129	0.614	-3.521	0.000	Prophet_regresores vs Naive
ETSX_Estructural	0.550	0.491	0.090	0.604	-3.389	0.001	ETSX_Estructural vs Naive
LSTM_multivariado	0.533	0.000	0.000	0.595	-2.536	0.011	LSTM_multivariado vs Naive
GRU_multivariado	0.483	0.114	-0.132	0.359	-2.374	0.018	GRU_multivariado vs Naive
NeuralProphet_multivariado	0.617	0.657	0.266	0.586	-2.921	0.003	NeuralProphet_multivariado vs Naive
SNaive	0.567	0.519	0.126	0.564	0.088	0.930	SNaive vs Naive
Naive	0.467	0.636	0.000	0.500	—	—	Modelo de referencia

Nota: elaboración propia con base en los resultados obtenidos en Python. Nota: DM corresponde a la comparación frente al modelo Naive.

Apéndice D. Métricas completas de desempeño por régimen de volatilidad

Tabla D 1

Métricas completas de desempeño predictivo por régimen de volatilidad

Régimen - Volatilidad	Modelo	N_obs	MAE	RMSE	sMAPE (%)	MASE	Accuracy	Precision	Recall	F1-score	MCC	AUC
Alta volatilidad	ARIMAX_manual	30	0.02	0.02	154.88	0.51	0.60	0.56	0.64	0.60	0.21	0.64

Régimen - Volatilidad	Modelo	N_obs	MAE	RMSE	sMAPE (%)	MASE	Accuracy	Precision	Recall	F1- score	MCC	AUC
Alta volatilidad	SARIMAX	30	0.02	0.02	155.94	0.51	0.60	0.56	0.64	0.60	0.21	0.63
Alta volatilidad	AutoARIMAX	30	0.02	0.02	155.97	0.51	0.60	0.56	0.64	0.60	0.21	0.62
Alta volatilidad	ETSX_Estructural	30	0.02	0.02	157.31	0.51	0.53	0.50	0.57	0.53	0.07	0.62
Alta volatilidad	Prophet_regresores	30	0.02	0.02	157.45	0.51	0.60	0.56	0.71	0.62	0.22	0.63
Alta volatilidad	AutoARIMAX_XGBoost_Hibrido	30	0.02	0.02	145.22	0.52	0.50	0.45	0.36	0.40	- 0.02	0.56
Alta volatilidad	LSTM_multivariado	30	0.02	0.02	148.53	0.53	0.53	0.00	0.00	0.00	0.00	0.61
Alta volatilidad	GRU_multivariado	30	0.02	0.02	145.60	0.53	0.53	0.50	0.07	0.12	0.02	0.44
Alta volatilidad	AutoARIMAX_MLP_Hibrido	30	0.02	0.02	155.77	0.54	0.53	0.50	0.71	0.59	0.09	0.54
Alta volatilidad	RandomForest_multivariado	30	0.02	0.02	154.45	0.54	0.47	0.43	0.43	0.43	- 0.07	0.50
Alta volatilidad	XGBoost_multivariado	30	0.02	0.02	142.72	0.54	0.67	0.64	0.64	0.64	0.33	0.60
Alta volatilidad	GAM_multivariado	30	0.02	0.02	134.51	0.56	0.67	0.62	0.71	0.67	0.34	0.65
Alta volatilidad	MLP_multivariado	30	0.02	0.02	160.32	0.56	0.40	0.40	0.57	0.47	- 0.19	0.48
Alta volatilidad	NeuralProphet_multivariado	30	0.02	0.02	149.47	0.57	0.57	0.52	0.86	0.65	0.20	0.56
Alta volatilidad	Naive	30	0.02	0.03	142.67	0.61	0.47	0.47	1.00	0.64	0.00	0.50
Alta volatilidad	SNaive	30	0.02	0.03	156.70	0.70	0.40	0.33	0.29	0.31	- 0.22	0.40

Régimen - Volatilidad	Modelo	N_obs	MAE	RMSE	sMAPE (%)	MASE	Accuracy	Precision	Recall	F1- score	MCC	AUC
Baja volatilidad	RandomForest_multivariado	30	0.01	0.01	91.27	0.28	0.73	1.00	0.43	0.60	0.53	0.89
Baja volatilidad	GAM_multivariado	30	0.01	0.01	107.74	0.30	0.80	0.83	0.71	0.77	0.60	0.92
Baja volatilidad	XGBoost_multivariado	30	0.01	0.01	108.97	0.30	0.77	0.89	0.57	0.70	0.55	0.85
Baja volatilidad	AutoARIMAX_MLP_Hibrido	30	0.01	0.02	109.12	0.32	0.70	0.73	0.57	0.64	0.40	0.77
Baja volatilidad	AutoARIMAX_XGBoost_Hibrido	30	0.01	0.02	122.89	0.34	0.60	0.75	0.21	0.33	0.22	0.79
Baja volatilidad	MLP_multivariado	30	0.01	0.02	132.73	0.35	0.63	0.57	0.86	0.69	0.32	0.74
Baja volatilidad	LSTM_multivariado	30	0.01	0.02	156.04	0.40	0.53	0.00	0.00	0.00	0.00	0.60
Baja volatilidad	AutoARIMAX	30	0.01	0.02	150.05	0.40	0.67	0.83	0.36	0.50	0.37	0.63
Baja volatilidad	NeuralProphet_multivariado	30	0.01	0.02	131.03	0.41	0.67	0.62	0.71	0.67	0.34	0.62
Baja volatilidad	SARIMAX	30	0.01	0.02	150.34	0.41	0.67	0.83	0.36	0.50	0.37	0.63
Baja volatilidad	ARIMAX_manual	30	0.01	0.02	153.45	0.41	0.60	0.67	0.29	0.40	0.20	0.60
Baja volatilidad	Prophet_regresores	30	0.01	0.02	148.59	0.41	0.53	0.50	0.36	0.42	0.05	0.56
Baja volatilidad	ETSX_Estructural	30	0.01	0.02	152.57	0.42	0.57	0.56	0.36	0.43	0.12	0.56
Baja volatilidad	GRU_multivariado	30	0.02	0.02	182.32	0.44	0.43	0.20	0.07	0.11	- 0.24	0.23
Baja volatilidad	SNaive	30	0.02	0.02	112.61	0.45	0.73	0.71	0.71	0.71	0.46	0.74

Régimen - Volatilidad	Modelo	N_obs	MAE	RMSE	sMAPE (%)	MASE	Accuracy	Precision	Recall	F1-score	MCC	AUC
Baja volatilidad	Naive	30	0.02	0.03	145.38	0.57	0.47	0.47	1.00	0.64	0.00	0.50

Nota. Esta tabla conserva el detalle técnico completo de métricas de error y métricas direccionales por régimen.

Apéndice E. Métricas completas de desempeño por régimen de volumen

Tabla E 1

Métricas completas de desempeño predictivo por régimen de volumen

Régimen - Volumen	Modelo	N_obs	MAE	RMSE	sMAPE (%)	MASE	Accuracy	Precision	Recall	F1-score	MCC	AUC
Alto volumen	GAM_multivariado	25	0.02	0.02	106.84	0.47	0.80	0.75	0.67	0.71	0.56	0.90
Alto volumen	XGBoost_multivariado	25	0.02	0.02	110.42	0.48	0.84	0.78	0.78	0.78	0.65	0.86
Alto volumen	RandomForest_multivariado	25	0.02	0.02	111.62	0.51	0.72	0.75	0.33	0.46	0.35	0.80
Alto volumen	AutoARIMAX_XGBoost_Hibrido	25	0.02	0.02	120.11	0.52	0.68	0.67	0.22	0.33	0.24	0.82
Alto volumen	AutoARIMAX_MLP_Hibrido	25	0.02	0.02	123.83	0.52	0.76	0.71	0.56	0.62	0.46	0.78
Alto volumen	MLP_multivariado	25	0.02	0.02	144.40	0.56	0.52	0.36	0.44	0.40	0.01	0.68
Alto volumen	AutoARIMAX	25	0.02	0.03	156.17	0.61	0.80	0.75	0.67	0.71	0.56	0.89
Alto volumen	ARIMAX_manual	25	0.02	0.03	157.27	0.62	0.76	0.67	0.67	0.67	0.48	0.91
Alto volumen	SARIMAX	25	0.02	0.03	157.22	0.62	0.80	0.75	0.67	0.71	0.56	0.90
Alto volumen	ETSX_Estructural	25	0.02	0.03	159.56	0.62	0.68	0.55	0.67	0.60	0.34	0.83
Alto volumen	Prophet_regresores	25	0.02	0.03	159.47	0.62	0.68	0.54	0.78	0.64	0.39	0.84

Régimen - Volumen	Modelo	N_obs	MAE	RMSE	sMAPE (%)	MASE	Accuracy	Precision	Recall	F1-score	MCC	AUC
Alto volumen	LSTM_multivariado	25	0.02	0.03	154.01	0.65	0.64	0.00	0.00	0.00	0.00	0.45
Alto volumen	NeuralProphet_multivariado	25	0.02	0.03	153.36	0.65	0.64	0.50	0.89	0.64	0.39	0.77
Alto volumen	GRU_multivariado	25	0.02	0.03	168.74	0.68	0.64	0.50	0.11	0.18	0.09	0.31
Alto volumen	SNaive	25	0.03	0.03	145.66	0.78	0.56	0.33	0.22	0.27	-0.03	0.44
Alto volumen	Naive	25	0.03	0.04	155.02	0.82	0.36	0.36	1.00	0.53	0.00	0.50
Bajo volumen	LSTM_multivariado	35	0.01	0.02	151.05	0.33	0.46	0.00	0.00	0.00	0.00	0.63
Bajo volumen	RandomForest_multivariado	35	0.01	0.02	130.88	0.34	0.51	0.56	0.47	0.51	0.04	0.54
Bajo volumen	SARIMAX	35	0.01	0.02	150.22	0.34	0.51	0.57	0.42	0.48	0.05	0.52
Bajo volumen	AutoARIMAX	35	0.01	0.02	150.75	0.35	0.51	0.57	0.42	0.48	0.05	0.51
Bajo volumen	GRU_multivariado	35	0.01	0.02	160.54	0.35	0.37	0.20	0.05	0.08	-0.28	0.42
Bajo volumen	ARIMAX_manual	35	0.01	0.02	151.95	0.35	0.49	0.54	0.37	0.44	-0.01	0.51
Bajo volumen	Prophet_regresores	35	0.01	0.02	148.42	0.35	0.49	0.53	0.42	0.47	-0.02	0.52
Bajo volumen	ETSX_Estructural	35	0.01	0.02	151.64	0.35	0.46	0.50	0.37	0.42	-0.07	0.50
Bajo volumen	AutoARIMAX_MLP_Hibrido	35	0.01	0.02	138.61	0.36	0.51	0.54	0.68	0.60	-0.00	0.47
Bajo volumen	AutoARIMAX_XGBoost_Hibrido	35	0.01	0.02	144.01	0.36	0.46	0.50	0.32	0.39	-0.06	0.49
Bajo volumen	NeuralProphet_multivariado	35	0.01	0.02	130.89	0.37	0.60	0.61	0.74	0.67	0.18	0.46

Régimen - Volumen	Modelo	N_obs	MAE	RMSE	sMAPE (%)	MASE	Accuracy	Precision	Recall	F1-score	MCC	AUC
Bajo volumen	MLP_multivariado	35	0.01	0.02	148.05	0.38	0.51	0.53	0.84	0.65	-0.05	0.44
Bajo volumen	XGBoost_multivariado	35	0.01	0.02	136.87	0.39	0.63	0.71	0.53	0.61	0.28	0.58
Bajo volumen	GAM_multivariado	35	0.01	0.02	131.33	0.40	0.69	0.70	0.74	0.72	0.36	0.68
Bajo volumen	Naive	35	0.01	0.02	136.18	0.42	0.54	0.54	1.00	0.70	0.00	0.50
Bajo volumen	SNaive	35	0.02	0.02	126.80	0.43	0.57	0.60	0.63	0.62	0.13	0.60

Nota: elaboración propia con base en los resultados obtenidos en Python.

Apéndice F. Interpretabilidad e importancia de variables en modelos multivariados

Este anexo presenta los resultados complementarios de interpretabilidad de los modelos multivariados. Se incluyen las principales variables identificadas por los modelos de machine learning, los regresores exógenos del modelo Prophet y el ranking global de importancia de variables en Random Forest y XGBoost.

La importancia de variables se calculó principalmente para los modelos multivariados de machine learning, dado que estos permiten obtener medidas de relevancia de predictores de forma más directa y comparable. En el caso del MLP se utilizó importancia por permutación, mientras que en Random Forest y XGBoost se emplearon las medidas internas de importancia asociadas a modelos basados en árboles. Para modelos clásicos, estructurales, aditivos y de deep learning, la interpretación de variables requiere enfoques específicos —coeficientes, funciones suaves, contribuciones parciales o métodos de sensibilidad— que no son directamente comparables con las medidas de importancia utilizadas en los modelos de machine learning. Por esta razón, el análisis de importancia se concentró en los modelos donde la medida resulta metodológicamente más estable e interpretable.

Tabla F 1*Ranking global de importancia de variables en Random Forest y XGBoost*

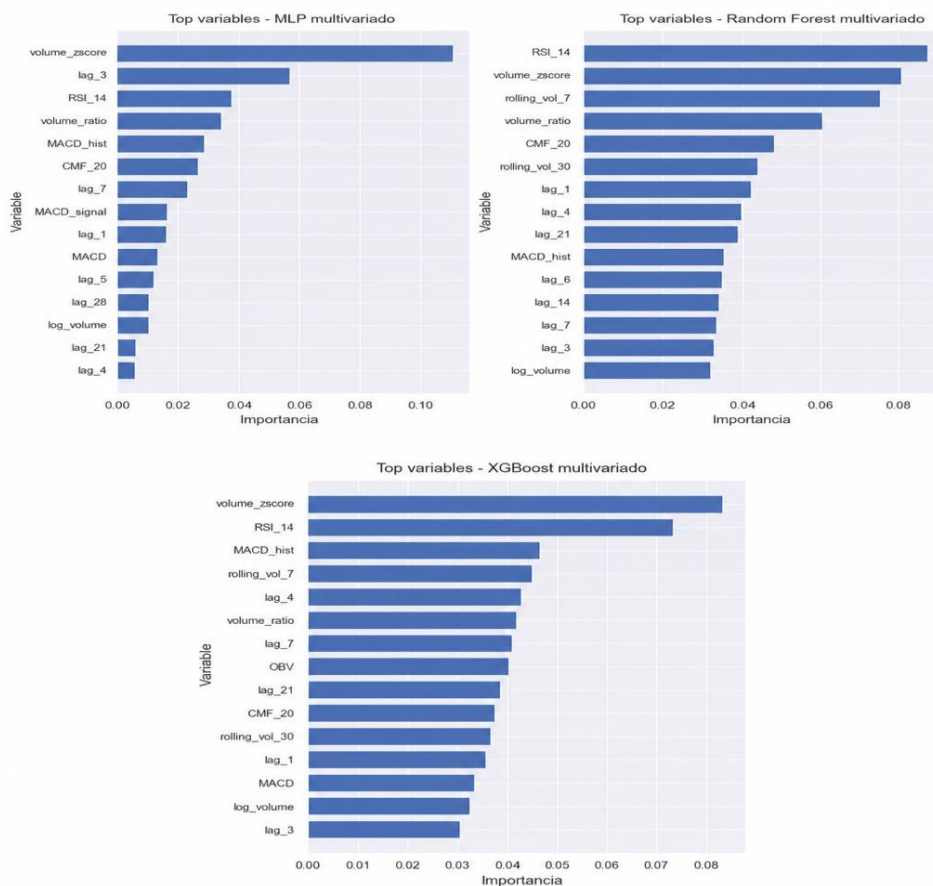
Ranking global de importancia	Variable	Importancia RF	Importancia XGB	Importancia promedio
1	volume_zscore	0.081641	0.082092	0.081867
2	RSI_14	0.086635	0.074662	0.080649
3	rolling_vol_7	0.076610	0.050983	0.063797
4	volume_ratio	0.062019	0.047171	0.054595
5	CMF_20	0.050657	0.044065	0.047361
6	MACD_hist	0.038274	0.052029	0.045152
7	rolling_vol_30	0.046691	0.043118	0.044905
8	lag_4	0.041728	0.047994	0.044861
9	lag_1	0.046060	0.042328	0.044194
10	lag_21	0.041386	0.044348	0.042867
11	lag_7	0.037178	0.046954	0.042066
12	OBV	0.032853	0.046265	0.039559
13	log_volume	0.035925	0.038427	0.037176
14	lag_6	0.037836	0.035291	0.036563
15	lag_3	0.036539	0.036484	0.036511
16	lag_14	0.037456	0.033987	0.035722
17	MACD	0.029071	0.038534	0.033802
18	lag_5	0.034826	0.032239	0.033533
19	lag_2	0.033881	0.030925	0.032403

Ranking global de importancia	Variable	Importancia RF	Importancia XGB	Importancia promedio
20	lag_28	0.033223	0.031565	0.032394
21	MACD_signal	0.030141	0.033471	0.031806
22	volume_sma_20	0.026685	0.035643	0.031164
23	ATR_14	0.022685	0.031422	0.027053

Nota. La tabla presenta el ranking global de importancia de variables calculado a partir de los modelos Random Forest y XGBoost. Los valores se reportan con seis decimales; Elaboración propia con base en los resultados obtenidos en Python.

Figura F 1

Importancia de variables en modelos MLP, Random Forest y XGBoost multivariados



Nota: elaboración propia con base en los resultados obtenidos en Python.

Tabla F 2*Regresores exógenos del modelo Prophet y coeficientes beta aproximados*

#	Regresor	Coeficiente_beta_aprox	Abs_beta
0	RSI_14	0.025589	0.025589
1	OBV	0.008829	0.008829
2	MACD	-0.006353	0.006353
3	MACD_signal	-0.005417	0.005417
4	CMF_20	0.005408	0.005408
5	MACD_hist	-0.004314	0.004314
6	volume_zscore	-0.004060	0.004060
7	volume_ratio	0.003794	0.003794
8	ATR_14	-0.003787	0.003787
9	rolling_vol_30	0.002289	0.002289
10	volume_sma_20	-0.002172	0.002172
11	log_volume	-0.001135	0.001135
12	rolling_vol_7	0.000623	0.000623

Nota: La tabla presenta los regresores exógenos del modelo Prophet, el coeficiente beta aproximado y su valor absoluto; Elaboración propia con base en los resultados obtenidos en Python.