DOCENTES DE ADMINISTRACIÓN FINANCIERA

45 Jornadas Nacionales de Administración Financiera

Septiembre 18 y 19, 2025

De datos a decisiones estratégicas

Machine learning para el análisis de riesgos e identificación de patrones en finanzas tradicionales y descentralizadas

Rita Beatriz Morrone

Universidad de Buenos Aires

SUMARIO

Para comentarios:
ritamorrone@economicas.uba.ar

Resumen

Se aborda la aplicabilidad de *machine learning* (ML) como un paradigma transformador cuyo objetivo principal es extraer información de los datos para generar herramientas estratégicas de soporte orientadas a reducir la incertidumbre en los mercados financieros. Se enfoca en la extracción de patrones complejos a partir de grandes volúmenes de datos históricos y en tiempo real, tanto en mercados tradicionales como en los de criptoactivos, resaltando la necesidad de una adaptación metodológica debido a sus diferencias estructurales, operativas y de comportamiento. Estas diferencias se extienden al tipo de datos disponibles, habiendo en los mercados tradicionales predominio de datos estructurados de mayor calidad, mientras que en los mercados de criptoactivos abundan datos heterogéneos, no lineales y no estructurados, lo que condiciona las técnicas analíticas aplicables.

Se explora un marco comprensivo de las principales técnicas de *machine learning*, diferenciando entre el aprendizaje supervisado y no supervisado. Dentro del aprendizaje supervisado, se examinan métodos como la regresión lineal, *random forest*, regresión logística y redes neuronales recurrentes (RNN), incluyendo las LSTM, destacando aplicaciones para la predicción de precios del día siguiente y la clasificación de activos por perfil de riesgo. También se abordan métodos de análisis de sentimiento de noticias financieras mediante procesamiento del lenguaje natural (NLP), explorando cómo modelos avanzados como *transformers* permiten extraer información valiosa de datos no estructurados, para anticipar movimientos en mercados. Para el aprendizaje no supervisado se abordan el análisis de componentes principales (pca) para la reducción de dimensionalidad y la detección de anomalías, y las técnicas de *clustering* para el agrupamiento de activos según su perfil de riesgo.

Los resultados prácticos revelan que cada problema financiero requiere un enfoque específico según el objetivo de análisis, el tipo y volumen de datos disponibles, considerando las características intrínsecas de cada mercado. Para la predicción de precios, las LSTM superan a la regresión lineal y *random forest* en entornos volátiles como las criptomonedas, gracias a su capacidad para manejar patrones no lineales y dependencias temporales largas. La regresión lineal es útil en mercados estables y *random forest* captura no linealidades con menor interpretabilidad. La regresión logística es adecuada para predecir categorías de riesgo, identificando predictores clave y dinámicas diferenciadas por activo. Finalmente, el análisis de componentes principales y *clustering* son adecuados para reducir dimensionalidad, filtrar ruido de mercado, detectar anomalías y segmentar activos en grupos de riesgo homogéneos, revelando estructuras ocultas en los datos. La elección del método depende primordialmente de los objetivos y las características específicas de los datos.

1. Introducción

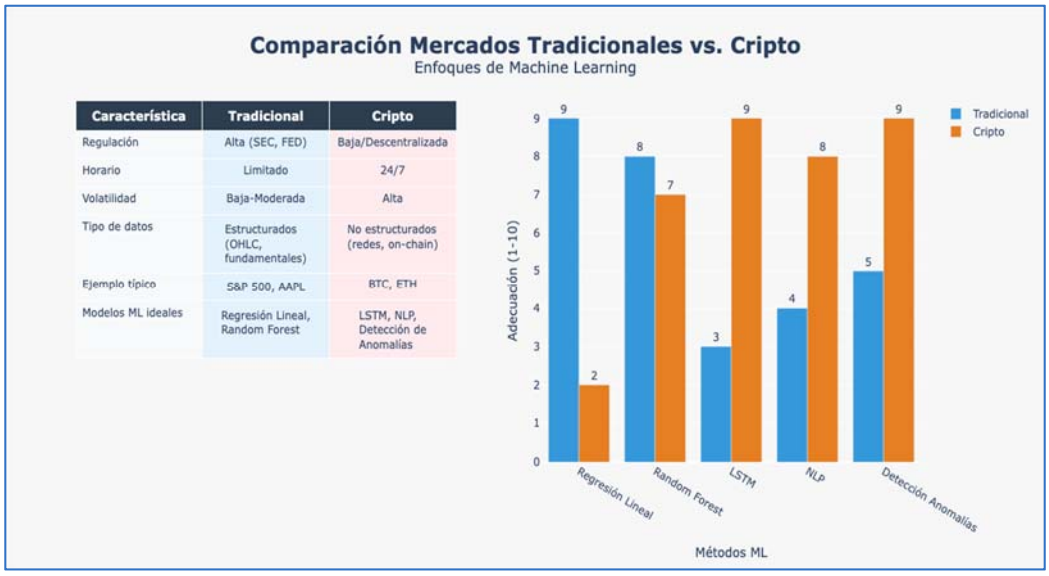
En la actualidad, el análisis de datos aplicado a los mercados financieros mediante técnicas de *machine learning* (ML) constituye una diciplina clave cuyo objetivo primordial es transformar la complejidad de los datos en herramientas estratégicas de soporte orientadas a reducir la incertidumbre. *Machine learning* o aprendizaje automático corresponde a una rama de la inteligencia artificial que permite a un sistema computacional aprender patrones complejos a partir de datos históricos, sin ser programados explícitamente para cada tarea. A través de algoritmos matemáticos y estadísticos, estas técnicas se basan en identificar relaciones ocultas, generalizar comportamientos y realizar predicciones o clasificaciones con precisión creciente a medida que se alimentan de mayor volumen de información. Permiten diseñar un experimento con el fin de probar hipótesis y entrenarlo, evaluando su precisión.

En el ámbito financiero, el ML emerge como un paradigma transformador para el análisis de los mercados tanto tradicionales como de criptoactivos, permitiendo obtener patrones complejos a partir de grandes volúmenes de datos históricos y en tiempo real. Estas técnicas permiten predecir precios de activos del día siguiente, clasificar activos según su perfil de riesgo y optimizar estrategias de inversión, transformando grandes volúmenes de datos. A su vez, facilita el procesamiento e integración de datos estructurados (series temporales) y no estructurados (noticias) con el propósito de generar *insights* accionables. Sin embargo, su aplicación efectiva requiere comprender tanto las capacidades de los modelos como las particularidades estructurales de cada mercado, donde factores como la volatilidad, la liquidez y el tipo de datos disponibles determinan el éxito de las estrategias implementadas.

La divergencia entre los mercados tradicionales y los mercados de criptoactivos evidencia la necesidad de una adaptación metodológica. Estos presentan diferencias estructurales, operativas y de comportamiento, que impactan directamente en la gestión de riesgos, estrategias de inversión y marcos regulatorios. A grandes rasgos, los mercados tradicionales están altamente regulados por entidades centralizadas, con horarios fijos de negociación, alta liquidez en activos de referencia y volatilidad moderada vinculada a indicadores macroeconómicos. Por el contrario, los mercados de criptoactivos se caracterizan por su naturaleza descentralizada, operatividad 24/7 y volatilidad extrema influenciada por factores tecnológicos y narrativas especulativas.

Estas diferencias que caracterizan a ambos mercados se extienden al tipo de datos disponibles, lo que condiciona las metodologías analíticas aplicables. En los mercados tradicionales, predominan datos estructurados de mayor calidad como series temporales, indicadores macroeconómicos y flujos institucionales, que permiten aplicar modelos como regresión lineal y *random forest*. Por el contrario, los mercados de criptoactivos operan con datos heterogéneos, no lineales y no estructurados como series temporales de alta frecuencia, métricas on-chain, contenido de redes sociales y eventos tecnológicos, requiriendo técnicas avanzadas como procesamiento del lenguaje natural (NLP) para análisis de sentimiento, redes neuronales recurrentes (LSTM) para predecir precios volátiles y algoritmos de análisis de componentes principales (PCA) y *clustering* para identificar patrones en datos multidimensionales (ilustración 1).

Ilustración 1: Comparativa de mercados financieros con enfoque en análisis de datos y ML



A diferencia de los modelos estadísticos tradicionales, que requieren supuestos rígidos (como linealidad o normalidad), *machine learning* se adapta a la complejidad y especificidad de los mercados, donde la volatilidad diferenciada, la interdependencia de variables y la tipología específica de datos disponibles exigen enfoques flexibles y escalables. Su implementación depende críticamente de tres pilares: la calidad de los datos, la selección del algoritmo adecuado al problema y la validación rigurosa de resultados para evitar sobreajustes o sesgos.

Con el fin de proporcionar un marco comprensivo sobre la utilidad y oportunidades de los métodos de aprendizaje automático aplicados al análisis de datos de los mercados financieros, se realizará un recorrido por las principales técnicas de *machine learning* (ML), destacando su aplicabilidad específica y sus ventajas en cada contexto de mercado. En el apartado 2 se establecen las generalidades de ML, delineando las distinciones fundamentales entre el aprendizaje supervisado y no supervisado, y se propone un árbol de decisión para la selección metodológica óptima, considerando el objetivo del análisis que se desea realizar, el tipo de datos disponibles y los alcances de cada técnica.

Avanzando en la aplicación práctica, el apartado 3 se centra en el análisis de sentimiento mediante procesamiento del lenguaje natural (NLP), explorando cómo modelos avanzados como *transformers* permiten extraer información valiosa de datos no estructurados, para anticipar movimientos en mercados. El apartado 4 aborda la predicción de precios para el día siguiente de activos específicos, contrastando la efectividad de las redes neuronales recurrentes, la regresión lineal y *random forest*, en función de los mercados específicos y características distintivas. El apartado 5, aborda el método de regresión logística, una técnica adecuada para predecir categorías. Se muestra un ejemplo de predicción de categorías de alto/bajo riesgo sobre activos específicos. El apartado 6, explora la detección de anomalías financieras y el agrupamiento de activos por perfiles de riesgo a través del análisis de componentes principales (PCA) para la reducción de dimensionalidad y el filtrado de ruido, y diversas técnicas de *clustering* como K-Means y el *clustering* jerárquico, incluyendo el uso de dendrogramas para visualizar

relaciones. Estos ejemplos se han implementado utilizando código Python en el entorno de Google Colaboratory, basándose en la extracción de datos a través de APIs específicas como Yahoo Finance y CoinGecko.

2. *Machine learning* y análisis de datos

En el ámbito de *machine learning* (ML), las metodologías se categorizan principalmente en dos paradigmas fundamentales, los modelos de aprendizaje supervisado y los modelos de aprendizaje no supervisado. Estos enfoques, difieren en su propósito y en la naturaleza de los datos que emplean. El aprendizaje supervisado requiere un conjunto de datos etiquetados para entrenar los algoritmos. Se caracteriza por la presencia de variables de entrada (predictoras) y una variable de respuesta asociada (output), cuyo valor se busca predecir o estimar. El algoritmo aprende a generar la salida deseada a medida que se introducen mayor cantidad de datos y se ajustan sus ponderaciones. El objetivo es obtener un modelo que relacione la respuesta con los predictores, ya sea para predecir resultados futuros con precisión o para comprender mejor la relación entre las variables (Hair *et al*, 2019; James *et al*, 2023).

Estos modelos de aprendizaje supervisado requieren ciertos niveles de experiencia para poder estructurar el análisis con precisión. Implica proporcionar al modelo un conjunto de datos de entrenamiento que incluya entradas y salidas correctas. Dentro de esta categoría, se encuentran diversos métodos. La regresión lineal es un método clásico y útil para predecir una respuesta cuantitativa. Asume una relación lineal entre las variables predictoras y la variable predecida. Por su naturaleza, es comúnmente aplicada para la predicción de valores de variables correspondientes a mercados financieros tradicionales. La regresión logística es un método de clasificación utilizado para predecir respuestas cualitativas o categóricas, estimando la probabilidad de que una observación pertenezca a una clase específica. Es una herramienta valiosa para predecir, si el precio de un activo aumentará o disminuirá. *Random forest* es un método basado en árboles de decisión que se distingue por su capacidad para manejar datos no lineales y mejorar la precisión predictiva. Opera combinando un gran número de árboles de decisión individuales, lo que los convierte en un método de ensamble robusto (Hair *et al*, 2019; James *et al*, 2023).

Por otro lado, las redes neuronales recurrentes (RNN) y los modelos *transformers* (LLM), aunque también forman parte del aprendizaje supervisado, merecen una mención especial por su aplicabilidad a datos secuenciales y no estructurados. Las RNN, incluyendo las LSTM (*long short-term memory*), son ideales para trabajar con series temporales altamente volátiles, como las que caracterizan a los mercados de criptoactivos, ya que pueden capturar patrones no lineales y dependencias a largo plazo para predecir el precio de un activo. Los *transformers* (LLM) como BERT, por su parte, son cruciales para el procesamiento del lenguaje natural (NLP), permitiendo el análisis de sentimiento a partir de noticias financieras o comentarios de redes sociales y anticipando movimientos del mercado basados en datos no estructurados (Vaswani *et al*, 2017; Kong *et al*, 2024).

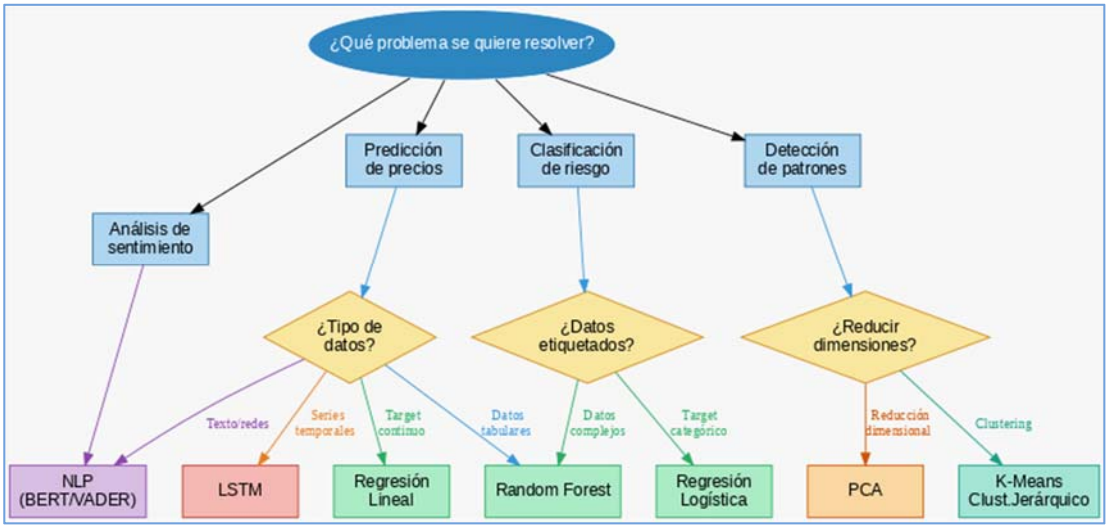
El aprendizaje no supervisado, por su parte, se aplica en escenarios donde solo se dispone de variables de entrada, sin una variable de respuesta asociada. Busca patrones y estructuras no visibles en datos sin etiquetas predefinidas. El modelo se ajusta a las observaciones pues no se

dispone de un conocimiento a priori. El conjunto de datos de entrada será tratado y los algoritmos se basan en reducir la dimensionalidad de los datos, evidenciar patrones y generar agrupamiento sin necesidad de intervención humana. El propósito aquí no es la predicción, sino descubrir relaciones, estructuras o patrones ocultos dentro de los datos. Este tipo de análisis es a menudo exploratorio y más subjetivo en su evaluación. Entre las técnicas más destacadas se encuentran el análisis de componentes principales (PCA) y el *clustering*.

El análisis de componentes principales (PCA) es una técnica poderosa para la reducción de dimensionalidad en conjuntos de datos multivariados, lo que simplifica la interpretación y ayuda a filtrar el "ruido del mercado". Al identificar componentes clave que explican la mayor variabilidad de los datos, permite la detección de anomalías mediante el error de reconstrucción y es un paso previo para el agrupamiento de activos por perfil de riesgo en mercados complejos. Los métodos de *clustering*, como K-Means y el *clustering* jerárquico, se utilizan para encontrar subgrupos o "clústeres" homogéneos dentro de los datos. Estas técnicas organizan automáticamente los activos, revelando aquellos que comparten perfiles de riesgo similares. El dendrograma, una representación visual del *clustering* jerárquico, actúa como un "mapa de relaciones" que facilita la exploración de mercados complejos como el cripto y la identificación de valores atípicos (Hair *et al*, 2019; James *et al*, 2023; Jolliffe & Cadima, 2016; Kaufman & Rousseeuw, 2009).

En resumen, la elección del método de ML, ya sea supervisado o no supervisado, y la especificidad de las técnicas (como regresión lineal, regresión logística, *random forest*, RNN, PCA o *clustering*), dependen críticamente del tipo de problema a resolver y de las características intrínsecas de los datos de cada mercado financiero. En la ilustración 2 se propone un árbol de decisión para la elección de la metodología a aplicar. Su objetivo principal es facilitar la selección del modelo más adecuado para una tarea específica, basándose en dos criterios clave, el tipo de problema que se busca resolver (ya sea predecir precios de activos o clasificar riesgos financieros) y la naturaleza de los datos disponibles (como tablas, texto o series de tiempo), guiando al lector en la elección de la herramienta analítica más idónea para cada escenario.

Ilustración 2: Árbol de decisión para la elección del modelo según el tipo de problema y datos disponibles



El árbol de decisión propuesto funciona como un marco orientativo para la selección de algoritmos de *machine learning*. Si el objetivo es anticipar movimientos del mercado de activos basándose en el tono de noticias o el sentimiento de redes sociales, y los datos disponibles son no estructurados, como textos, el árbol dirige hacia el uso de modelos de procesamiento del lenguaje natural (NLP) que logran extraer el tono positivo o negativo de la información textual, lo que permite inferir posibles impactos en el mercado financiero. Si el objetivo es una predicción numérica, como la del precio de un activo para el día siguiente, y los datos se presentan en formato tabular, muy común en mercados tradicionales, se sugiere el empleo de técnicas de regresión lineal o *random forest*. Por otro lado, si se dispone de series de tiempo altamente volátiles, características del mercado de criptomonedas, el camino se dirige hacia redes neuronales recurrentes (RNN) y sus variantes como las LSTM, debido a su habilidad para capturar patrones no lineales y dependencias a largo plazo esenciales en entornos tan dinámicos.

Cuando la necesidad es una respuesta cualitativa, por ejemplo, la clasificación del riesgo inherente a un activo (alto/bajo riesgo) o si el precio de un activo (aumentará/disminuirá), y se dispone de datos etiquetados con una variable target categórica, el árbol de decisión se orienta hacia el método de regresión logística. Una ventaja clave de esta es su robustez y que no requiere cumplir con supuestos estrictos como la normalidad multivariada, lo que la hace adecuada para una amplia gama de situaciones en mercados financieros En el caso de datos más complejos o relaciones altamente no lineales se orienta a técnicas de *random forest*, ideales para la clasificación en mercados de activos, donde las relaciones entre las variables pueden ser intrincadas y no lineales, como en la predicción de movimientos de precios o la identificación de perfiles de riesgo complejos.

Asimismo, el árbol de decisión aborda escenarios donde la meta no es la predicción directa de una variable, sino la exploración de la estructura inherente de los datos o la identificación de patrones ocultos sin etiquetas predefinidas. En este punto, se destacan técnicas como el análisis de componentes principales (PCA), utilizada para la reducción de dimensionalidad en datos financieros multivariados, lo que permite simplificar la información sin perder elementos clave y detectar anomalías al filtrar el "ruido" del mercado. Complementariamente, los métodos de *clustering* (como K-Means o *clustering* jerárquico) son propuestos para agrupar activos con perfiles de riesgo similares. Estas herramientas, al revelar *outliers* y correlaciones no evidentes a través de visualizaciones como el dendrograma.

En los siguientes apartados, se realizará un recorrido práctico mostrando posibles aplicaciones de las metodologías propuestas, lo que permitirá comprender los alcances de cada enfoque y visualizar ejemplos concretos de cómo los modelos seleccionados se podrían aplicar para la generación de insights predictivos y analíticos que fundamenten la toma de decisiones estratégicas y la gestión de riesgos en los mercados tradicionales y de criptoactivos.

3. Análisis de sentimiento

Una herramienta clave, a la hora de analizar los mercados es el análisis de sentimiento de noticias. Su importancia radica en poder anticipar movimientos o tendencias, basados en el "estado de ánimo" del mercado. Esta capacidad es vital, dado que el sentimiento de los participantes del mismo puede influir de manera significativa en las decisiones de inversión y, por

ende, en la dinámica de los precios. Tal como se mencionó anteriormente, para realizar este análisis se debe disponer de datos no estructurados como ser noticias sobre activos financieros, audios de conferencias, opiniones en redes sociales como Twitter (ahora X), que pueden ser valiosos y aportar visibilidad a la hora de disminuir la incertidumbre.

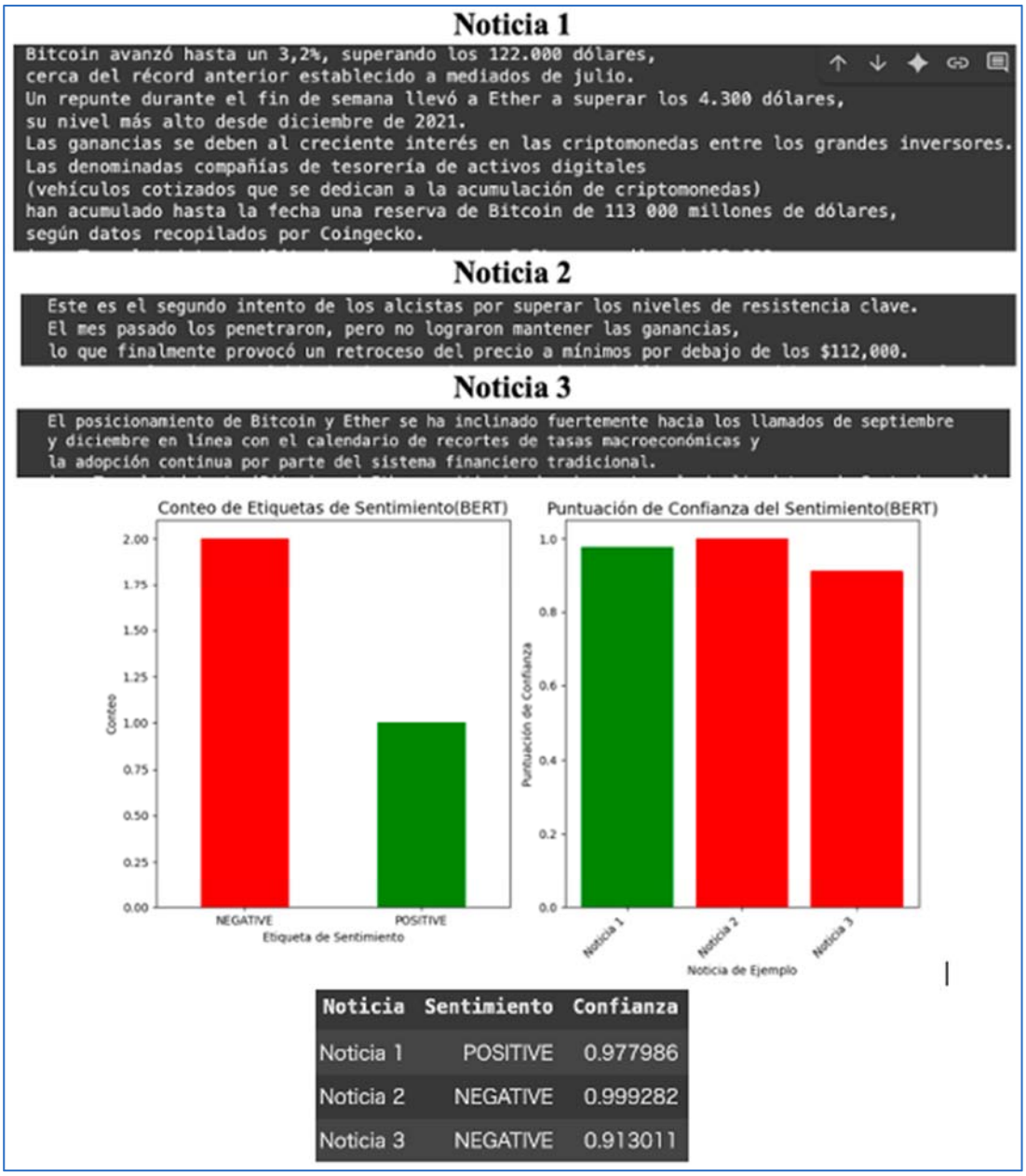
Este tipo de análisis es parte de una técnica de procesamiento del lenguaje natural (NPL) basada en inteligencia artificial que permite a las computadoras entender, interpretar y generar lenguaje humano. Desde el punto de vista técnico, el análisis de sentimiento implica varias etapas clave. La primera consta de la extracción de datos a partir de diversas fuentes como noticias financieras, opiniones de redes sociales, o incluso mediante scraping de tweets con hashtags específicos. La segunda corresponde al preprocesamiento de texto (tokenización, eliminación de stopwords, embeddings para convertir palabras a vectores numéricos). La tercera se refiere al proceso de clasificación de sentimiento mediante modelado del lenguaje a través de distintos modelos preentrenados para lograr la clasificación del tono emocional y agregación de resultados para generar métricas accionables (Vaswani *et al*, 2017; Kong *et al*, 2024; Liang *et al*, 2025).

En los mercados tradicionales, donde la información fluye a través de medios formales como reportes financieros y declaraciones corporativas, existen modelos avanzados como BERT son ideales para analizar el tono y matices en documentos complejos, ayudando a predecir reacciones institucionales. Por otro lado, en los mercados de criptoactivos, caracterizados por su alta volatilidad y dependencia de redes sociales, herramientas como VADER brindan rapidez para procesar el flujo constante de opiniones en redes sociales como Twitter, donde términos como "fomo" (miedo a perderse oportunidades) o "hodl" (mantener posiciones) pueden indicar cambios abruptos en la demanda.

Desde una perspectiva técnica, BERT (*bidirectional encoder representations from transformers*) es un modelo de *deep learning* de la familia de los *transformers* (*large language models*) muy eficaz para capturar dependencias de largo plazo. Juega un papel clave al ser capaz de comprender el significado contextual del lenguaje, incluso en el ámbito financiero, extrayendo el tono positivo o negativo de grandes volúmenes de texto, empleando mecanismos de atención para capturar relaciones semánticas profundas entre palabras. A través de su entrenamiento, puede discernir matices como ironía, sarcasmo o dobles sentidos, lo que lo hace superior para analizar documentos o noticias más complejas. También, tiene capacidad de fine-tuning lo que permite adaptarlo a dominios específicos, como el financiero, ajustando sus parámetros con datos etiquetados para mejorar la precisión en tareas como la clasificación de riesgo o la detección de tendencias en mercados (Vaswani *et al*, 2017; Kong *et al*, 2024; Liang *et al*, 2025). En la ilustración 3 se muestra un ejemplo del análisis de tres noticias por esta metodología.

En contraste, VADER (Valence Aware Dictionary and sEntiment Reasoner) es un algoritmo basado en reglas léxicas predefinidas y heurísticas simples, diseñado para evaluar el tono emocional en textos cortos e informales. A diferencia de los modelos de inteligencia artificial avanzada, no utiliza aprendizaje automático ni redes neuronales, sino que opera mediante un diccionario de palabras con puntuaciones de polaridad asignadas, combinado con reglas contextuales básicas como el manejo de negaciones o intensificadores. Este enfoque lo hace eficiente para análisis en tiempo real, pero limitado en precisión al interpretar contextos complejos o lenguaje especializado. En la ilustración 4 se muestra un ejemplo del análisis de tres comentarios empleando este método.

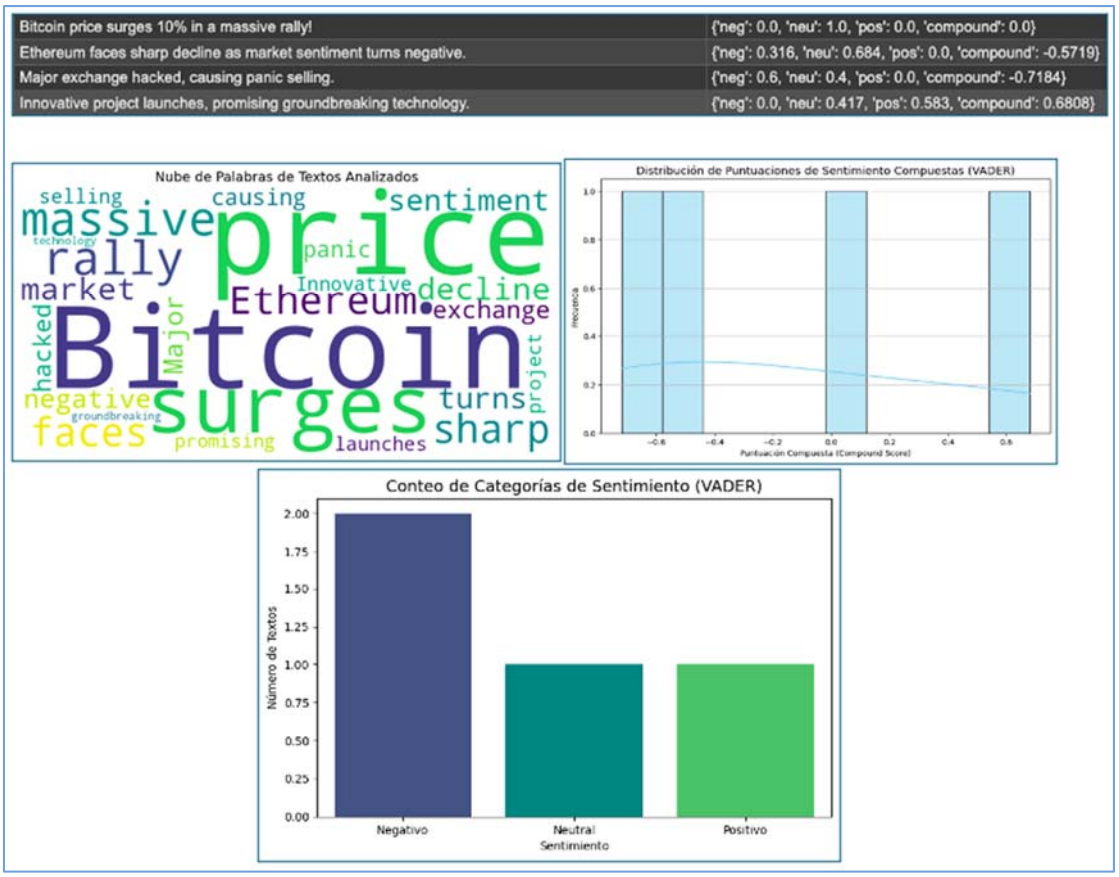
Ilustración 3: Análisis de sentimiento con BERT



Noticia	Sentimiento	Confianza
Noticia 1	POSITIVE	0.977986
Noticia 2	NEGATIVE	0.999282
Noticia 3	NEGATIVE	0.913011

La combinación estratégica de ambos modelos (VADER para velocidad en datos informales y BERT para profundidad en análisis contextual) permite a los actores financieros detectar señales tempranas, ya sea para evitar riesgos en acciones tradicionales o capitalizar tendencias emergentes en criptoactivos. Esta dualidad refleja cómo el análisis de sentimiento moderno debe adaptarse a las particularidades de cada mercado. La utilización conjunta y combinación de estos modelos dependerá del equilibrio requerido entre velocidad y profundidad analítica.

Ilustración 4: Análisis de sentimiento con VADER



4. Análisis de predicción de precios

En el ámbito de los mercados financieros, la predicción de precios de un activo para el día siguiente es una tarea crítica, la cual requiere la utilización de modelos que sean capaces de adaptarse a los datos disponibles según las características específicas de cada mercado. Para tal fin, se presenta primero un modelo de red neuronal recurrente para el análisis predictivo a partir de una serie de tiempo de bitcoin. Luego, se examina la aplicación de modelos de regresión lineal y *random forest* para predecir los precios tanto de bitcoin como del índice S&P500. El objetivo final es contrastar la efectividad de estas técnicas en función de las características distintivas de cada tipo de mercado.

4.1 Redes neuronales

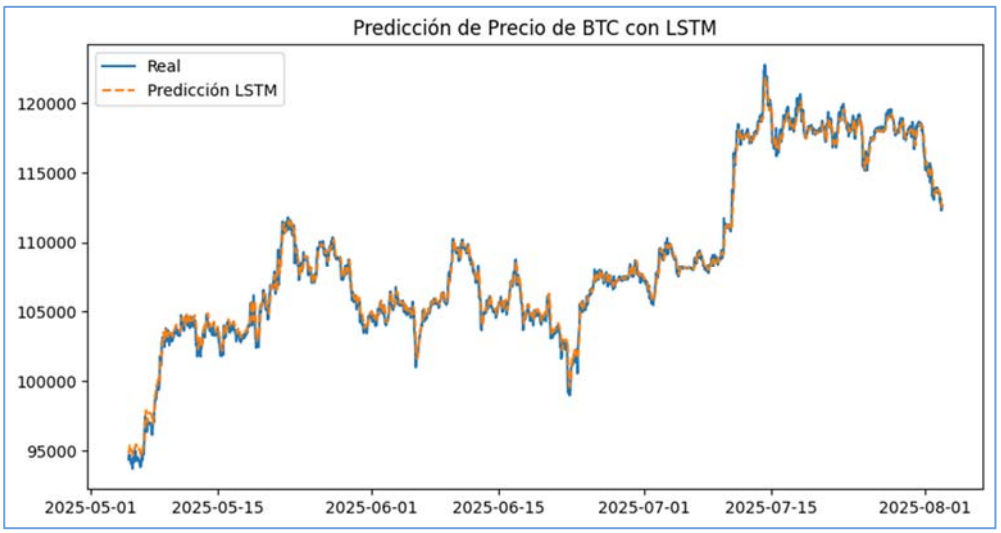
Las redes neuronales recurrentes (RNN), en particular la arquitectura *long short-term memory* (LSTM), son modelos de aprendizaje automático especialmente diseñados para trabajar con datos secuenciales y series temporales. En los mercados financieros, esto las hace ideales

para predecir el precio de un activo para el día siguiente, ya que pueden analizar patrones históricos y tendencias recientes simultáneamente. Su fortaleza radica en su capacidad para capturar patrones no lineales y dependencias a largo plazo en datos volátiles, como los de los mercados de criptoactivos, procesando la información de manera secuencial y considerando la historia de los datos anteriores para realizar la predicción. Son modelos altamente flexibles (James *et al*, 2023).

Con la finalidad de comprender su aplicabilidad, se presenta un ejemplo en donde se busca predecir el precio de bitcoin del día siguiente a partir de una ventana establecida de precios de siete días previos. Inicialmente se extrajeron datos del precio de BTC de los últimos 90 días mediante la CoinGeckoAPI. Se realizó el preprocesamiento, escalado de los datos y se crearon secuencias de precios con lapsos temporales de siete días para lograr predecir el precio del octavo día. Luego se define el modelo LSTM utilizando la librería tensorflow.keras. Esta red neuronal constituye un modelo secuencial con dos capas LSTM y una capa densa de salida. Al apilar dos capas LSTM, la salida de la primera se convierte en la entrada de la segunda permitiendo que el modelo aprenda representaciones más complejas y de mayor nivel a partir de los datos secuenciales, lo que es beneficioso para tareas de predicción de series de tiempo complejas como las disponibles del mercado de criptomonedas.

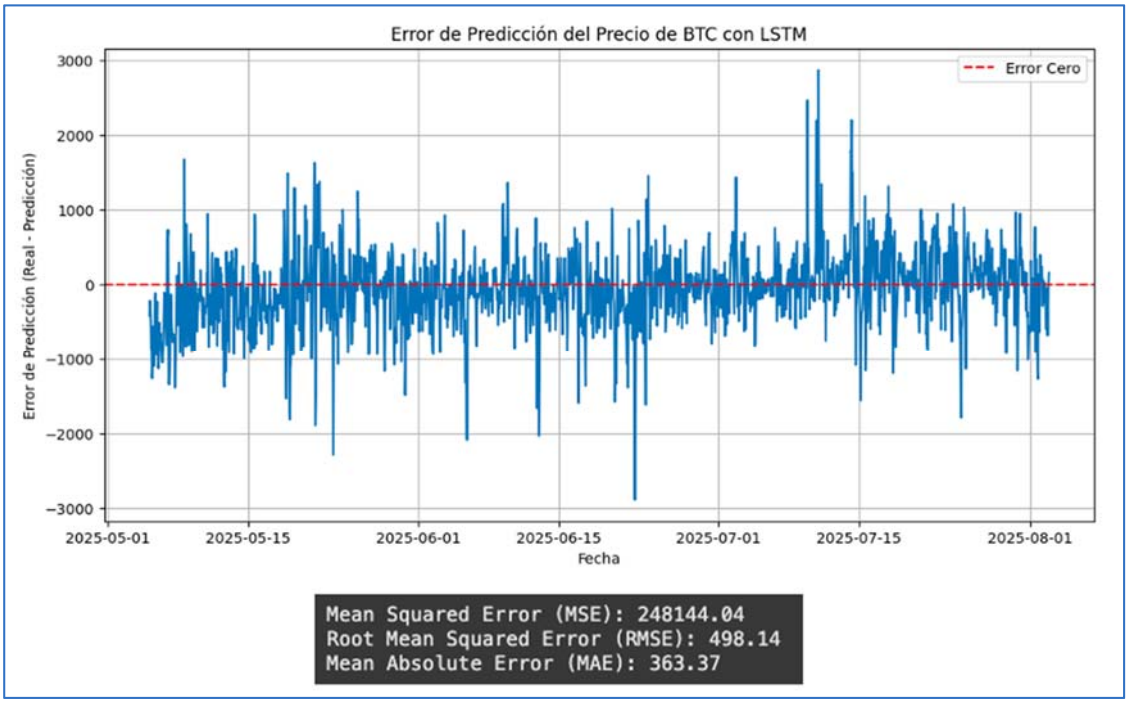
El modelo se entrena con datos preparados durante 20 épocas con un tamaño de lote de 32, o sea, se procesará todo el conjunto de datos de entrenamiento 20 veces durante el proceso de aprendizaje. El tamaño de lote significa que el modelo procesa 32 secuencias de datos de bitcoin a la vez, calcula el error promedio para ese lote y luego actualiza sus pesos basándose en ese error. En cada época, el modelo ajusta sus pesos internos para tratar de minimizar el error de predicción. Más épocas permiten que el modelo aprenda mejor, pero demasiadas pueden llevar a un problema llamado "sobreajuste" (*overfitting*), donde el modelo se vuelve demasiado bueno para predecir los datos de entrenamiento pero no los datos nuevos. Posteriormente al entrenamiento se realizan las predicciones sobre las secuencias de entrada y se realiza la transformación inversa para volver a la escala de precio original. En la ilustración 5 se muestra una gráfica que compara los precios reales de bitcoin con los precios predichos por el modelo.

Ilustración 5: Comparativa de precios reales versus precios predichos de BTC



A partir de las predicciones realizadas se puede cuantificar el error de predicción, que es la diferencia entre los valores reales observados y los valores predichos por el modelo. En la ilustración 6 se muestra la gráfica del error de predicción a lo largo del tiempo. La serie de tiempo exhibe la diferencia entre el precio real y el precio predicho para cada punto de datos en el conjunto donde hay predicciones disponibles. Un error positivo significa que el modelo predijo un precio más bajo de lo real, y un error negativo significa que predijo un precio más alto. También se exponen métricas relevantes como MSE, RMSE, MAE resultantes de este modelo. El error cuadrático medio (MSE, *mean squared error*) es una medida que calcula el promedio de las diferencias al cuadrado entre los valores predichos y los valores reales. Al realizarse sobre el conjunto de datos de prueba sirve para medir la capacidad de generalización del modelo y el objetivo es minimizarlo. La raíz del error cuadrático medio (RMSE, *root mean squared error*) es la raíz cuadrada del MSE que mide la falta de ajuste del modelo a los datos y se interpreta en las mismas unidades que la variable dependiente, lo que facilita la comprensión del error promedio. El error absoluto medio (MAE, *mean absolute error*) es el promedio de los valores absolutos de las diferencias entre las predicciones y los valores reales. Está en las mismas unidades que la variable que se predice, pero es menos sensible a los valores atípicos que el MSE o RMSE, ya que no eleva al cuadrado los errores grandes (James *et al*, 2023).

Ilustración 6: Error de predicción con LSTM y métricas relevantes



Acorde con las métricas exhibidas y para una muestra de datos de precios de 90 días, el modelo presentó un rendimiento prometedor ya que mostró una capacidad razonable para seguir los movimientos del precio de bitcoin en el período analizado. Las predicciones se desviaron alrededor de 498 usd del precio real. De manera similar, la magnitud promedio del error de predicción fue de unos 363 usd. Estos valores, en el contexto de la volatilidad inherente a la dinámica de bitcoin, sugieren que el modelo pudo capturar patrones y dependencias en las series

de tiempo de precios con un nivel de error relativamente contenido. La arquitectura de estas redes neuronales, caracterizada por su habilidad para interpretar secuencias de datos y aprender de la información pasada, parece ser adecuada para abordar la naturaleza dinámica y a menudo no lineal de los criptoactivos. Las métricas obtenidas validan la utilidad de este enfoque para la predicción de precios a corto plazo.

4.2 Regresión lineal versus *random forest*

Con la finalidad de comparar métodos de predicción y su aplicación, ya sea para predecir el precio del día siguiente de un activo, tanto del mercado tradicional como del mercado de criptoactivos, se realiza una comparativa entre regresión lineal y *random forest*. El rendimiento de cada técnica se evaluó utilizando series de precios de los últimos dos años del índice S&P 500, representando el mercado tradicional, y de bitcoin, representando el mercado de criptoactivos. Se entrenaron estos dos modelos para predecir el precio de cierre del día siguiente utilizando precios de cierre de cinco días anteriores. Se busca evaluar tanto ventajas como limitaciones de cada uno y su aplicabilidad en función de que tipo de activo se desea analizar.

La regresión lineal es un modelo de aprendizaje supervisado que busca establecer una relación lineal entre una variable dependiente (precio de cierre del día siguiente de un activo) y una o más variables independientes (precio de cierre de los 5 días anteriores como predictores). El objetivo es predecir una respuesta cuantitativa, ajustando una función lineal que minimice la suma de los errores al cuadrado por el método de mínimos cuadrados. Este enfoque facilita entender la dirección y magnitud de la relación de cada predictor con la variable de interés. En mercados financieros tradicionales, la regresión lineal puede ser una herramienta útil. Sin embargo, su principal limitación radica en la suposición de linealidad, lo cual a menudo no se corresponde con la compleja realidad de los mercados, donde las relaciones son inherentemente no lineales (Hair *et al*, 2019; Tabachnick & Fidell, 2019; James *et al*, 2023).

Además, es susceptible a problemas de multicolinealidad, donde la alta correlación entre las variables independientes puede dificultar la estimación precisa de los coeficientes individuales y enmascarar los verdaderos efectos de las variables, afectando la interpretabilidad y la estabilidad del modelo. Se debe tener en cuenta que la generalización a partir de este tipo de modelo puede verse comprometida por un tamaño de muestra inadecuado, especialmente cuando el número de parámetros a estimar es cercano al tamaño de la muestra, lo que puede llevar al sobreajuste (*overfitting*). Para mitigar esto, se recomienda una proporción mínima de 5 a 1 observaciones por variable independiente. La validación de sus resultados es clave y se puede realizar mediante la división de la muestra o el examen de los errores (James *et al*, 2023).

En contraste, los *random forests* son un método de conjunto (ensemble method) que construye múltiples árboles de decisión a partir de muestras bootstrap (con reemplazo) del conjunto de datos de entrenamiento. La predicción final se obtiene promediando las predicciones de todos los árboles individuales. La clave de su efectividad reside en la decorrelación de los árboles, que se logra seleccionando aleatoriamente un subconjunto de predictores en cada división del árbol (raíz cuadrada del total de predictores). Esto permite que el modelo explore más a fondo el espacio de soluciones y reduce la varianza de las predicciones, mejorando significativamente la precisión en comparación con un solo árbol.

Los *random forests* son ideales para manejar datos no lineales y relaciones complejas, lo que los hace adecuados para mercados de criptoactivos, conocidos por su alta volatilidad y

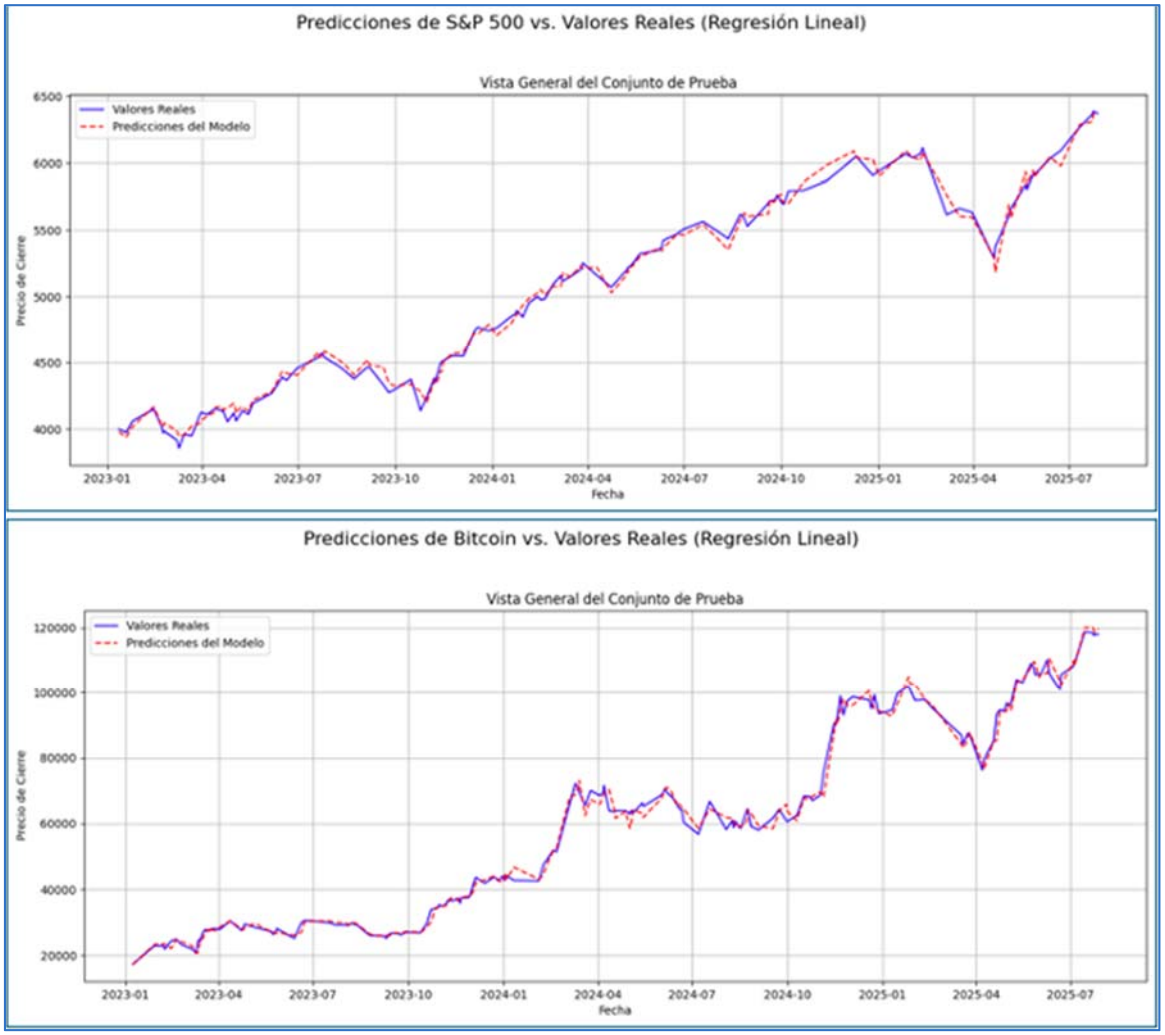
patrones intrincados. A diferencia de la regresión lineal, son más robustos al sobreajuste y a las pequeñas perturbaciones en los datos. También pueden manejar naturalmente una mezcla de tipos de variables (cuantitativas y cualitativas) y son menos sensibles a la multicolinealidad entre predictores debido a la selección aleatoria de características. Sin embargo, la principal desventaja es su menor interpretabilidad en comparación con los modelos lineales simples, ya que operan como una "caja negra" al combinar las predicciones de numerosos árboles. Aunque se pueden obtener medidas de importancia de las variables, el mecanismo exacto de cómo cada predictor influye en la predicción final es menos transparente. Al igual que con otros modelos, la selección adecuada de los parámetros de ajuste es importante para el rendimiento óptimo (Hair *et al*, 2019; Tabachnick & Fidell, 2019; James *et al*, 2023).

Continuando con la evaluación comparativa de estos dos modelos se utilizaron datos históricos de precios de los últimos dos años del índice S&P500, representando al mercado tradicional y de bitcoin, perteneciente al mercado de criptoactivos. Se entrenaron estos dos modelos para predecir el precio de cierre del día siguiente utilizando precios de cierre de 5 días anteriores. En las ilustraciones 7 y 8 se pueden ver las gráficas obtenidas para cada activo que comparan las predicciones versus los valores reales y en la ilustración 9 se muestran las métricas resultantes para cada activo de cada modelo.

Al interpretar los resultados, se observan diferencias sutiles pero significativas en el rendimiento de los modelos. Para el S&P 500, la regresión lineal muestra un R^2 de 0.9941, MSE de 3014.28 y RMSE de 54.90, mientras que el *random forest* tiene un R^2 de 0.9937, MSE de 3236.27 y RMSE de 56.89. Esto indica que la regresión lineal tuvo un rendimiento ligeramente superior en términos de R^2y MSE para el S&P 500. Sin embargo, el MAE del *random forest* (42.51) es inferior al de la regresión lineal (43.68), sugiriendo que, en promedio, sus predicciones tienen un error absoluto ligeramente menor. Para bitcoin, la regresión lineal presenta un R^2 de 0.9948, MSE de 4.49e+06 y RMSE de 2118.46; el *random forest* logra un R^2 de 0.9946, MSE de 4.61e+06 y RMSE de 2148,06. De nuevo, la regresión lineal exhibe métricas de MSE y R^2 ligeramente mejores. No obstante, al igual que con el S&P 500, el *random forest* obtiene un MAE más bajo (1482,50) en comparación con la regresión lineal (1522,25), lo que podría indicar una mayor robustez frente a valores atípicos y una mejor aproximación del error promedio en un mercado volátil como el de criptoactivos. Estas valoraciones resaltan que, si bien la regresión lineal puede capturar las tendencias dominantes de manera muy efectiva incluso en mercados complejos, los *random forests* pueden ofrecer una ventaja en la minimización del error absoluto, un factor relevante en aplicaciones financieras.

Además, si se compara los resultados de estos modelos para la predicción del precio de bitcoin con los resultados del modelo de redes neuronales LSTM aplicado al mismo conjunto de datos, se puede apreciar que las métricas obtenidas fueron notablemente inferiores en comparación estos dos enfoques (MSE de 212752, RMSE de 461, un MAE de 335, R^2de 0,99), logrando una reducción sustancial en todas las métricas de error. Esta superioridad en el rendimiento de LSTM para bitcoin es coherente con su capacidad inherente para manejar la alta volatilidad y los patrones no lineales que caracterizan el mercado de criptoactivos, a diferencia de los modelos de regresión lineal que asumen relaciones más simples o los *random forest* que, aunque robustos para no linealidades, podrían no capturar las dependencias temporales complejas tan eficientemente como una red LSTM. Esto subraya la importancia de seleccionar modelos cuya arquitectura se alinee con las características intrínsecas del tipo de datos y mercado que se desea analizar.

Ilustración 7: Predicciones versus valores reales del precio de S&P500 y BTC con regresión lineal



En resumen, la regresión lineal ofrece un entendimiento claro de las relaciones lineales, siendo un punto de partida para mercados con dinámicas más estables como los tradicionales. Los *random forests*, se adaptan mejor a la complejidad y volatilidad de los mercados modernos, como el de criptoactivos, al capturar patrones no lineales y reducir la varianza, aunque con un costo en la interpretabilidad directa. La elección óptima dependerá siempre de los objetivos específicos del problema y las características intrínsecas de los datos financieros a analizar. A su vez, mientras que la regresión lineal y *random forest* pueden ofrecer una primera aproximación rápida y capturar la tendencia general utilizando características simples, las redes neuronales LSTM se perfilan como un modelo más adecuado para la predicción del precio de criptoactivos debido a su capacidad para modelar la complejidad y la volatilidad de las series temporales financieras. Su menor error absoluto en las predicciones puntuales la hace una opción más prometedora para tareas donde la precisión en los valores es relevante. Sin embargo, es importante recordar que la predicción de precios en mercados volátiles sigue siendo un desafío, y el rendimiento de cualquier modelo debe ser validado rigurosamente en datos fuera de muestra y en condiciones de mercado cambiantes.

Ilustración 8: Predicciones versus valores reales del precio de S&P500 y BTC con random forest

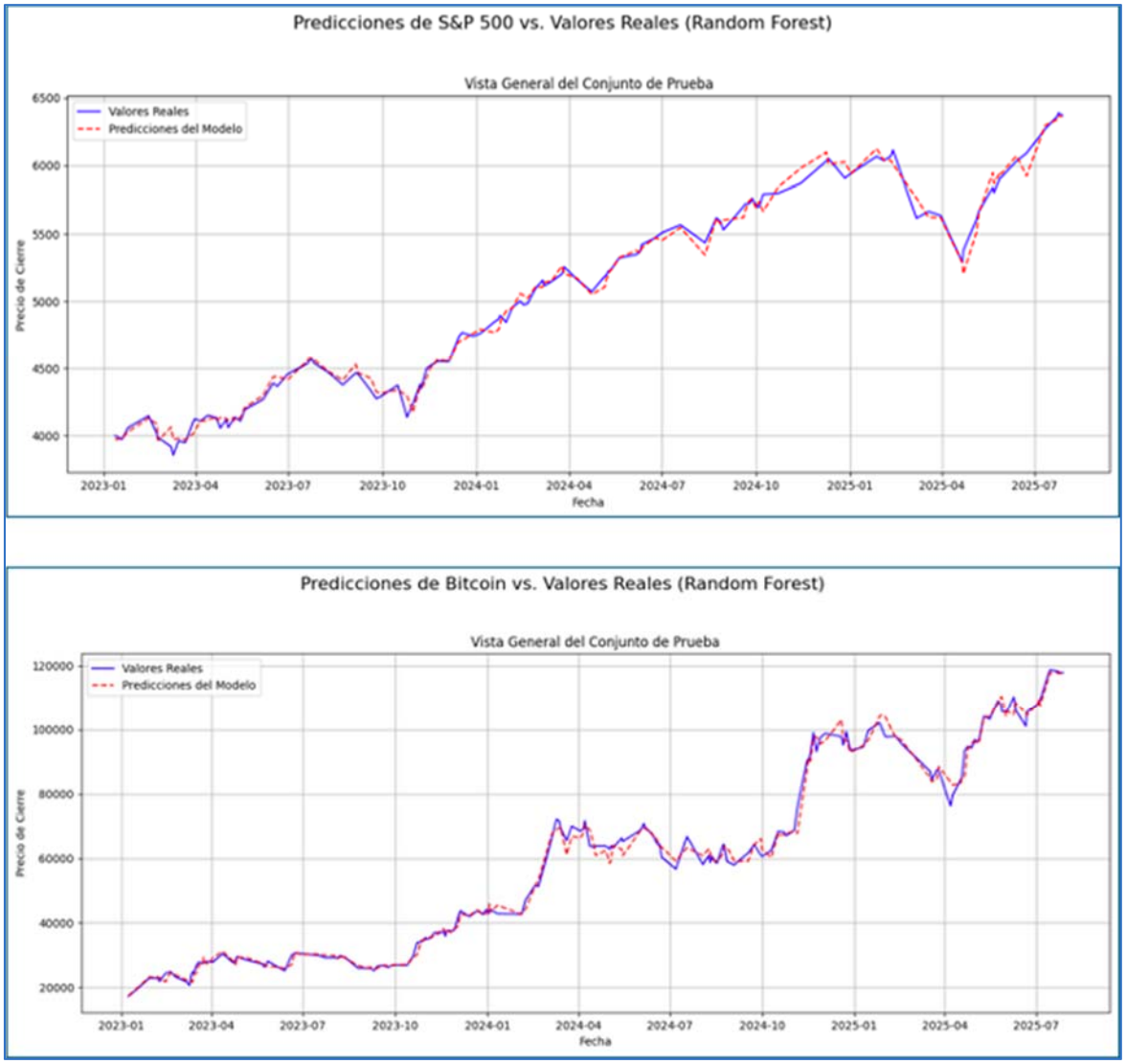


Ilustración 9: Regresión lineal versus random forest para S&P500 Y BTC

```
--- Resumen General de Resultados ---
             Modelo  Mercado          MSE       RMSE        MAE  R2 Score
0  Regresión Lineal  S&P 500  3.014277e+03    54.9024    43.6761    0.9941
1     Random Forest  S&P 500  3.236266e+03    56.8882    42.5063    0.9937
2  Regresión Lineal  Bitcoin  4.487881e+06  2118.4620  1522.2466    0.9948
3     Random Forest  Bitcoin  4.614176e+06  2148.0634  1482.5000    0.9946
```

5. Clasificación de activos por perfil de riesgo

Para abordar la necesidad de analizar datos desde la perspectiva de la clasificación del riesgo inherente a un activo (como alto/bajo riesgo) o la predicción de movimientos de precios (aumentará/disminuirá), donde se dispone de datos etiquetados con una variable objetivo categórica, la regresión logística se presenta como un método apropiado. Una de sus principales ventajas es su robustez, ya que no impone supuestos estrictos como la normalidad multivariada o la igualdad de matrices de varianza-covarianza entre grupos, siendo ideal para el análisis de situaciones en mercados financieros donde estas condiciones rara vez se cumplen. Además, permite manejar eficazmente variables independientes tanto métricas como no métricas. Se diferencia de la regresión lineal en que, en lugar de modelar directamente la respuesta de la variable predecida, modela la probabilidad de pertenecer a una categoría específica, garantizando que los valores predichos siempre estén entre 0 y 1. Sus principales objetivos son identificar el grupo al que pertenece un objeto y explicar las bases de la pertenencia a ese grupo a través de un conjunto de variables independientes (Hair *et al*, 2019; Tabachnick & Fidell, 2019; James *et al*, 2023).

El proceso de estimación se basa en el uso del método de máxima verosimilitud (MLE), a diferencia del método de mínimos cuadrados utilizado en la regresión lineal. Dada la naturaleza no lineal de la transformación logit aplicada a la variable dependiente (que convierte probabilidades en "odds"), el MLE se emplea de manera iterativa para encontrar los coeficientes que maximizan la probabilidad de que los eventos observados ocurran. La interpretación de los coeficientes logísticos, tanto en su dirección como en su magnitud, permite entender la relación entre las variables independientes y el cambio en la probabilidad predicha. Una relación positiva indica que un aumento en la variable independiente se asocia con un aumento en la probabilidad predicha, y viceversa (Hair *et al*, 2019; Tabachnick & Fidell, 2019; James *et al*, 2023).

Para mostrar su aplicabilidad práctica, se presenta un análisis a partir de datos históricos de bitcoin y ether obtenidos mediante la API de Yahoo Finance, para un período de 2 años. Se calcularon métricas de volatilidad anualizada móvil de 30 días, ratio de Sharpe móvil de 30 días y variación porcentual semanal de volumen. Se desarrollo y entrenó un modelo de regresión logística para predecir categorías de alto/bajo riesgo, definido como los días de mayor volatilidad y menor rendimiento. Para generar la variable target se consideró alto riesgo (=1) y bajo riesgo (=0) estableciendo un criterio basado en volatilidad superior al percentil 65 y ratio de Sharpe inferior al percentil 35, combinando alta volatilidad con bajo rendimiento. El modelo fue optimizado mediante técnicas de balanceo de clases (SMOTE) y ajuste automático de umbrales basado en la curva Precision-Recall. Este enfoque busca capturar períodos de estrés en los mercados cripto mientras se logra un equilibrio entre sensibilidad en la detección de riesgos reales y especificidad evitando falsas alarmas. En la ilustración 10 se muestran la volatilidad anualizada y el ratio de Sharpe para el período de análisis.

Luego de entrenar el modelo se obtuvieron las métricas clave para el modelo. En la ilustración 11 se muestran los resultados obtenidos para cada uno de los activos. El AUC-ROC mide la capacidad para distinguir entre clases, mientras que el AUC-PR es indicador en conjuntos de datos desbalanceados, evaluando el compromiso entre precisión y sensibilidad. El umbral óptimo representa el punto de equilibrio donde se maximiza la efectividad global del modelo se-

Ilustración 10: Volatilidad anualizada y ratio de Sharpe de BTC y ETH para el período 2023-2025

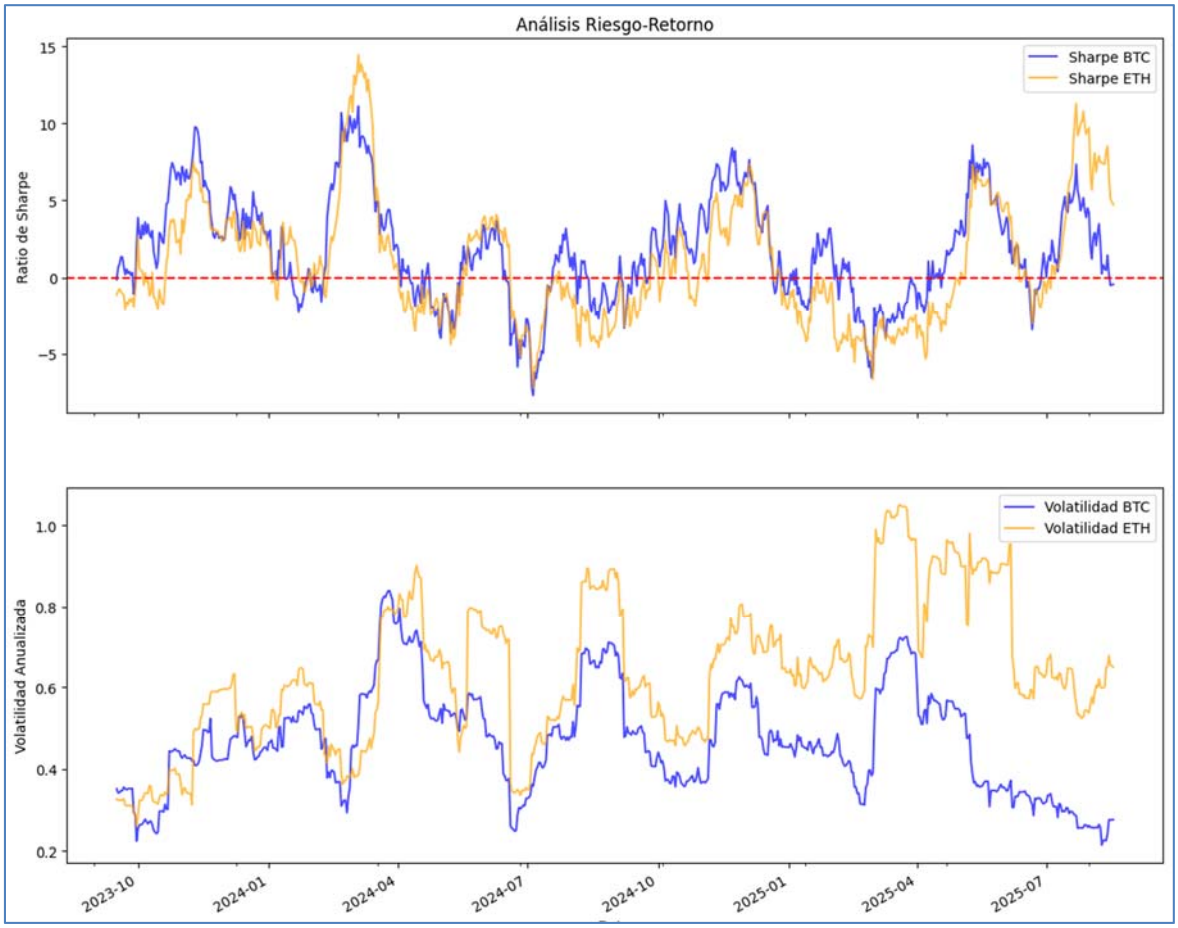


Ilustración 11: Resultados del modelo de regresión logística para BTC y ETH

🔍 Resultados para BTC:
Umbral óptimo: 0.5442
AUC-ROC: 0.9666
AUC-PR: 0.8741

Reporte de clasificación:

	precision	recall	f1-score	support
0	0.98	0.91	0.94	168
1	0.73	0.93	0.82	43
accuracy			0.91	211
macro avg	0.85	0.92	0.88	211
weighted avg	0.93	0.91	0.92	211

🔍 Resultados para ETH:
Umbral óptimo: 0.6619
AUC-ROC: 0.9999
AUC-PR: 0.9994

Reporte de clasificación:

	precision	recall	f1-score	support
0	1.00	0.99	1.00	171
1	0.98	1.00	0.99	40
accuracy			1.00	211
macro avg	0.99	1.00	0.99	211
weighted avg	1.00	1.00	1.00	211

gún el objetivo específico de la aplicación. La precisión del modelo se puede evaluar a partir de la proporción de predicciones correctas sobre el total de casos (Accuracy). También es importante considerar de todas las instancias predichas como positivas, cuántas fueron realmente

positivas (Precision) y de todas las instancias realmente positivas, cuántas fueron predichas correctamente como positivas (Recall–sensibilidad). El F1-Score es la media armónica de Precision y Recall.

Los resultados obtenidos demuestran un buen desempeño del modelo en ambos criptoactivos. Para bitcoin, alcanzó un área bajo la curva ROC (AUC-ROC) de 0.9666, indicando una alta capacidad discriminativa para distinguir entre períodos de alto y bajo riesgo. El área bajo la curva Precision-Recall (AUC-PR) de 0.8741 refleja equilibrio entre precisión y exhaustividad, dado el carácter desbalanceado de los datos. El umbral óptimo de clasificación se estableció en 0.5442, logrando un recall de 93 %, lo que implica que solo 7 % de los casos de alto riesgo no fueron detectados y una precisión de 73 % en la identificación de períodos de alto riesgo. En el caso de ether, el desempeño del modelo resultó mostrando un AUC-ROC de 0.9999 y un AUC-PR de 0.9994. El umbral óptimo se situó en 0.6619, alcanzando un recall de 100 % con detección de todos los casos de alto riesgo y una precisión de 98 %. Estas métricas pueden sugerir que los patrones de riesgo de ETH sean más predecibles que BTC o también puede indicar un posible sobreajuste del modelo. En las ilustraciones 12 y 13 se muestran las matrices de confusión, distribución de probabilidades y peso de las variables predictoras en el modelo.

La matriz de confusión es una herramienta importante a la hora de evaluar el rendimiento del modelo. Presenta cuatro categorías de clasificación, verdaderos positivos (casos de riesgo correctamente identificados), verdaderos negativos (períodos de calma correctamente clasificados), falsos positivos (falsas alarmas) y falsos negativos (casos de riesgo no detectados). Para BTC se observaron 3 falsos negativos y 15 falsos positivos. Esta distribución sugiere un modelo conservador que prioriza evitar omisiones de riesgo a costa de algunas falsas alarmas. Para ETH se evidenció un falso positivo. Esto podría atribuirse a patrones más predecibles o sobreajustes del modelo.

Los coeficientes estandarizados del modelo revelan patrones diferenciados entre bitcoin y ether. Para BTC, la volatilidad (2.63) muestra una fuerte relación positiva con el riesgo, confirmando que los períodos de alta fluctuación en los precios incrementan significativamente la probabilidad de eventos adversos. En contraste, el ratio de Sharpe (-3.46) evidencia una asociación inversa aún más pronunciada, donde un menor rendimiento ajustado al riesgo eleva la probabilidad de clasificación como alto riesgo. El volumen (-0.08), aunque marginal, sugiere que aumentos en la actividad transaccional podrían moderar levemente el riesgo. Para ETH, estos efectos son más intensos, la volatilidad (3.74) tiene un impacto aún mayor que en BTC, mientras que el ratio de Sharpe (-4.58) consolida su papel como principal indicador contrario al riesgo. El volumen en ETH (0.39) presenta una relación positiva débil con el riesgo, lo que podría reflejar que, a diferencia de BTC, los picos de volumen se asocian con mayor estrés en el mercado.

En resumen, la volatilidad es un predictor consistente de riesgo en ambos activos, pero con mayor peso en ETH. El Sharpe Ratio opera como un termómetro de estabilidad más preciso en ETH que en BTC. El volumen muestra comportamientos opuestos, efecto calmante en BTC versus amplificador en ETH, lo que sugiere dinámicas de liquidez distintas entre ambas criptomonedas. Estos hallazgos marcan la necesidad de enfoques diferenciados para la gestión de riesgo en cada activo, donde ETH exhibe relaciones más marcadas entre los predictores y el riesgo, mientras que BTC requiere mayor atención al contexto del mercado.

Ilustración 12: Resultados y gráficas para BTC

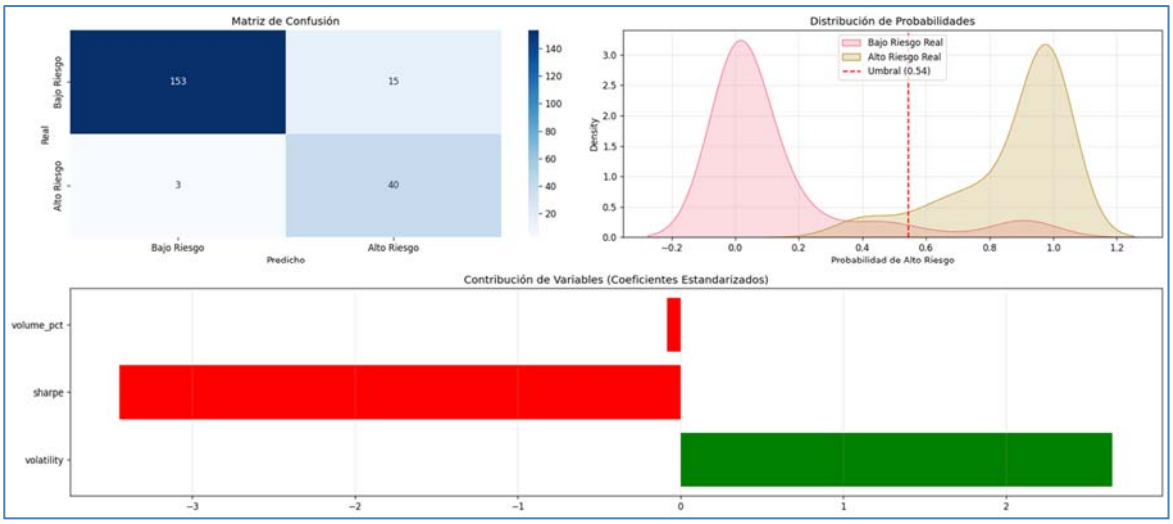
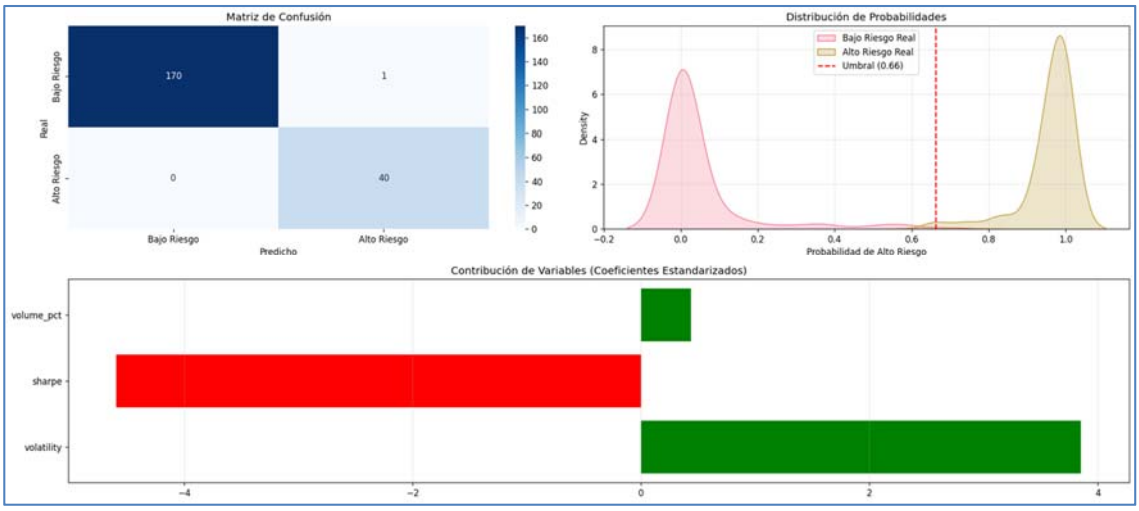


Ilustración 13: Resultados y gráficas para ETH



6. Detección de patrones

Las técnicas de aprendizaje no supervisado constituyen una herramienta metodológica ideal para el análisis exploratorio de datos no etiquetados de activos financieros, donde la ausencia de variables de respuesta exige algoritmos capaces de revelar patrones intrínsecos. En este contexto, el análisis de componentes principales (PCA) actúa como mecanismo de reducción de dimensionalidad, aislando factores latentes que maximizan la varianza explicativa y facilitan la identificación de anomalías mediante métricas de reconstrucción. Paralelamente, los algoritmos de *clustering* como K-Means o métodos jerárquicos permiten la segmentación automática de activos en grupos homogéneos basados en medidas de similitud multivariada, generando taxo-

nomías objetivas que trascienden categorizaciones subjetivas. Estas técnicas combinadas ofrecen un marco analítico apropiado para descomponer la complejidad de mercados altamente no lineales. Para exponer su aplicación, se proponen dos ejemplos prácticos. El primero, se basa en la detección de anomalías financieras sobre un conjunto de datos de diversos activos. El segundo, se basa en el agrupamiento de criptoactivos por su perfil de riesgo.

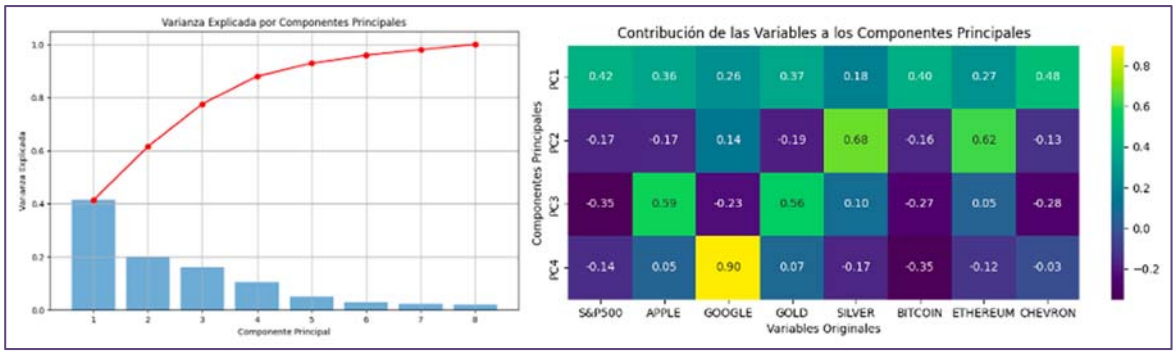
6.1 Detección de anomalías

El análisis de componentes principales (PCA) es una técnica muy útil para reducir la dimensionalidad en grandes conjuntos de datos e identificar los factores que explican la mayor parte de la variabilidad contenida en ellos. Su objetivo principal es simplificar datos sin perder información clave. Este enfoque multivariado facilita la transformación de variables originales potencialmente correlacionadas, en un nuevo conjunto de variables no correlacionadas llamadas componentes principales. Estos, constituyen un sistema de coordenadas ortogonales que maximiza la variabilidad explicada de los datos. El algoritmo realiza una descomposición espectral de la matriz de covarianza, generando componentes principales que representan direcciones de máxima variabilidad en el espacio multidimensional original. Desde el punto de vista operativo, cada componente se construye como una combinación lineal óptima de las variables originales. Este procedimiento garantiza que el primer componente (PC1) capture la máxima varianza posible, el segundo componente (PC2), ortogonal a PC1, explique la siguiente mayor porción de varianza, y así sucesivamente (Jolliffe & Cadima, 2016; Hair *et al*, 2019; Tabachnick & Fidell, 2019).

En el contexto de los mercados financieros, esta técnica ayuda a identificar los factores subyacentes más influyentes que explican los movimientos o características de los activos. Para llevar a cabo esta aplicación se seleccionaron 8 activos de diversas categorías (S&P500, Apple, Google, oro, plata, bitcoin, ether, Chevron) y se extrajeron datos para un período de 3 años (2022-2025), se calcularon los retornos diarios y se realizó la estandarización de estos datos para luego aplicar el algoritmo de PCA. La estandarización previa asegura que todas las variables estén en la misma escala, evitando que aquellas con mayor varianza sesguen el modelo. La selección de los componentes a considerar se determinó a través de dos criterios relevantes, el de retener los componentes cuyos valores propios son mayores a uno, indicando que aportan más varianza que una sola variable original, y el criterio que implica examinar el gráfico de proporción de varianza explicada y buscar un "codo" donde la proporción por cada componente subsiguiente disminuya bruscamente, sugiriendo el número óptimo de componentes. En la ilustración 14 se puede observar, a la izquierda que los primeros cuatro componentes contienen 85% de la varianza explicada. Luego de realizar esta selección, a la derecha se muestra las cargas factoriales (*loadings*) que señalan la contribución (peso y dirección) de las variables originales, a cada componente principal elegido.

El criterio de haber seleccionado los primeros 4 componentes principales radica en que éstos capturan la mayor parte de la señal o estructura principal de los datos, mientras que los componentes posteriores a menudo contienen más "ruido" o variabilidad atípica. Para continuar con el análisis, se aplica la función de transformación inversa y se calcula el error de reconstrucción (MSE), que constituye la diferencia entre los datos originales y los reconstruidos utilizando solo los componentes principales retenidos. Se establece un umbral para identificar los datos anómalos y se seleccionan aquellos puntos cuyo error de reconstrucción es mayor que el umbral.

Ilustración 14: Varianza explicada por cada componente

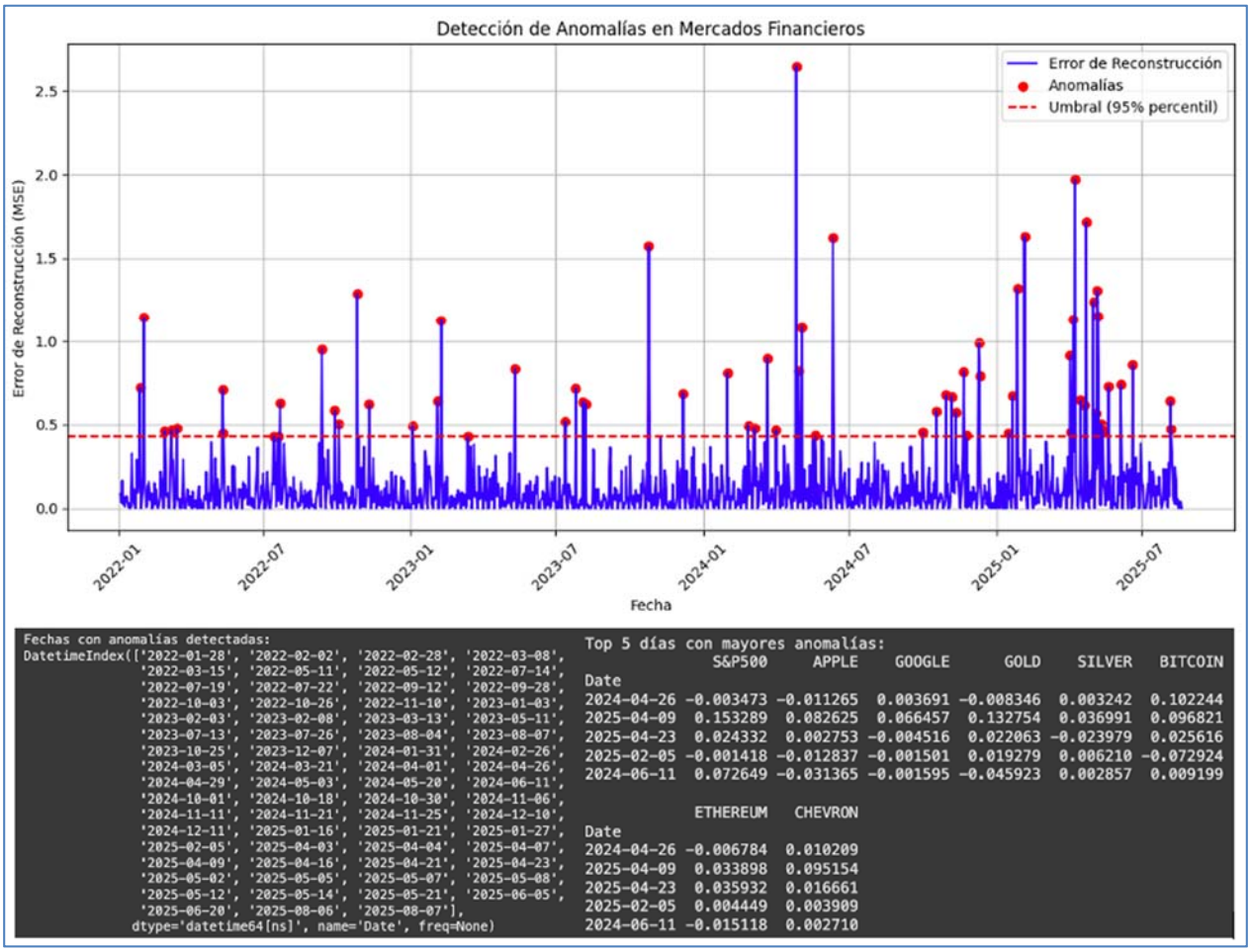


Un error alto sugiere que una observación se desvía significativamente del patrón principal capturado por los componentes dominantes, lo que permite detectar anomalías o eventos atípicos en los activos financieros. En la ilustración 15 se muestran las anomalías detectadas para el período y se detallan los días que coinciden con estas. Si bien este análisis se realiza con datos históricos, puede contribuir a obtener mayor comprensión de patrones y tendencias si se relaciona esta información con eventos relevantes acontecidos.

Luego de obtener las anomalías a partir de establecer un umbral para determinar los errores de reconstrucción significativos e identificar los días con mayores atipias, se puede confirmar visualmente las diferencias en la distribución de los retornos estandarizados de días anómalos versus días sin anomalías (ilustración 16). Es importante para completar este análisis contextualizar las anomalías detectadas con eventos relevantes para el mercado a través de la exploración de noticias, cambios regulatorios, factores macroeconómicos u otros eventos para la comprensión de patrones y tendencias.

Si bien el análisis de componentes principales (PCA) constituye una técnica valiosa para la reducción de dimensionalidad, filtrado de ruido y la detección de anomalías financieras mediante la evaluación del error de reconstrucción, se debe tener en cuenta algunas precauciones para asegurar la validez y utilidad de los resultados. La calidad de los datos es un punto crítico en mercados con bajo ratio señal/ruido como el financiero, lo que exige una curación de estos para filtrar elementos superfluos y ruido inesperado. Un paso indispensable es la estandarización previa de los datos, ya que asegura que todas las variables estén en la misma escala, evitando que aquellas con mayor varianza sesguen el modelo. En cuanto a la presencia de outliers, es importante interpretarlos con cautela, ya que no son inherentemente problemáticos, sino que podrían ser errores de recopilación. Se recomienda no eliminarlos a menos que se compruebe que son no representativos. También se debe considerar la eliminación de variables no informativas ya que, pueden introducir ruido aleatorio y comprometer el proceso de análisis. Finalmente, la validación de los resultados con datos fuera de muestra es relevante para prevenir el sobreajuste y siempre se debe complementar el análisis con el conocimiento del mercado para una interpretación contextual de las anomalías (James *et al*, 2023).

Ilustración 15: Días con anomalías detectadas mediante valoración del error de reconstrucción

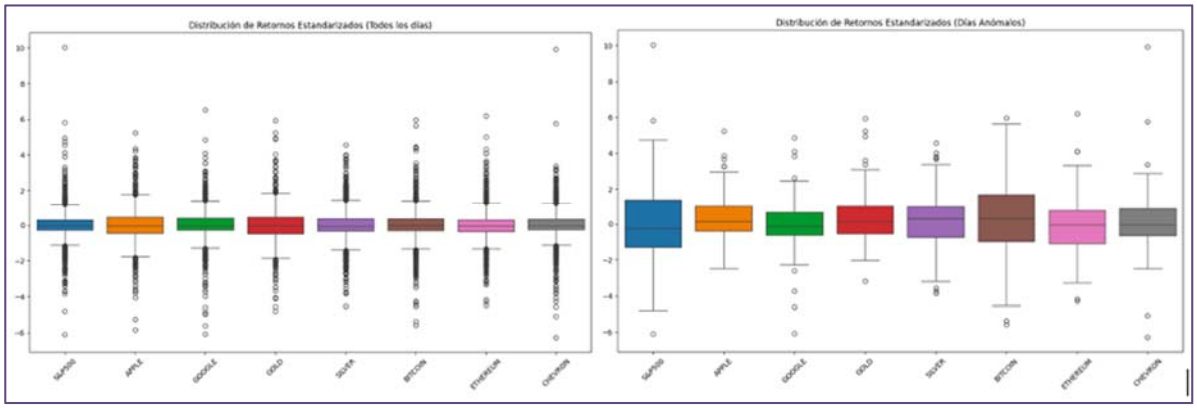



```
Fechas con anomalías detectadas:
DatetimeIndex(['2022-01-28', '2022-02-02', '2022-02-28', '2022-03-08',
               '2022-03-15', '2022-05-11', '2022-05-12', '2022-07-14',
               '2022-07-19', '2022-07-22', '2022-09-12', '2022-09-28',
               '2022-10-03', '2022-10-26', '2022-11-10', '2023-01-03',
               '2023-02-03', '2023-02-08', '2023-03-13', '2023-05-11',
               '2023-07-13', '2023-07-26', '2023-08-04', '2023-08-07',
               '2023-10-25', '2023-12-07', '2024-01-31', '2024-02-26',
               '2024-03-05', '2024-03-21', '2024-04-01', '2024-04-26',
               '2024-04-29', '2024-05-03', '2024-05-20', '2024-06-11',
               '2024-10-01', '2024-10-18', '2024-10-30', '2024-11-06',
               '2024-11-11', '2024-11-21', '2024-11-25', '2024-12-10',
               '2024-12-11', '2025-01-16', '2025-01-21', '2025-01-27',
               '2025-02-05', '2025-04-03', '2025-04-04', '2025-04-07',
               '2025-04-09', '2025-04-16', '2025-04-21', '2025-04-23',
               '2025-05-02', '2025-05-05', '2025-05-07', '2025-05-08',
               '2025-05-12', '2025-05-14', '2025-05-21', '2025-06-05',
               '2025-06-20', '2025-08-06', '2025-08-07'],
              dtype='datetime64[ns]', name='Date', freq=None)
```

```
Top 5 días con mayores anomalías:
              S&P500     APPLE    GOOGLE      GOLD    SILVER   BITCOIN
Date
2024-04-26 -0.003473 -0.011265  0.003691 -0.008346  0.003242  0.102244
2025-04-09  0.153289  0.082625  0.066457  0.132754  0.036991  0.096821
2025-04-23  0.024332  0.002753 -0.004516  0.022063 -0.023979  0.025616
2025-02-05 -0.001418 -0.012837 -0.001501  0.019279  0.006210 -0.072924
2024-06-11  0.072649 -0.031365 -0.001595 -0.045923  0.002857  0.009199

              ETHEREUM   CHEVRON
Date
2024-04-26 -0.006784  0.010209
2025-04-09  0.033898  0.095154
2025-04-23  0.035932  0.016661
2025-02-05  0.004449  0.003909
2024-06-11 -0.015118  0.002710
```

Ilustración 16: Comparativa de retornos estandarizados entre días típicos y días anómalos detectados

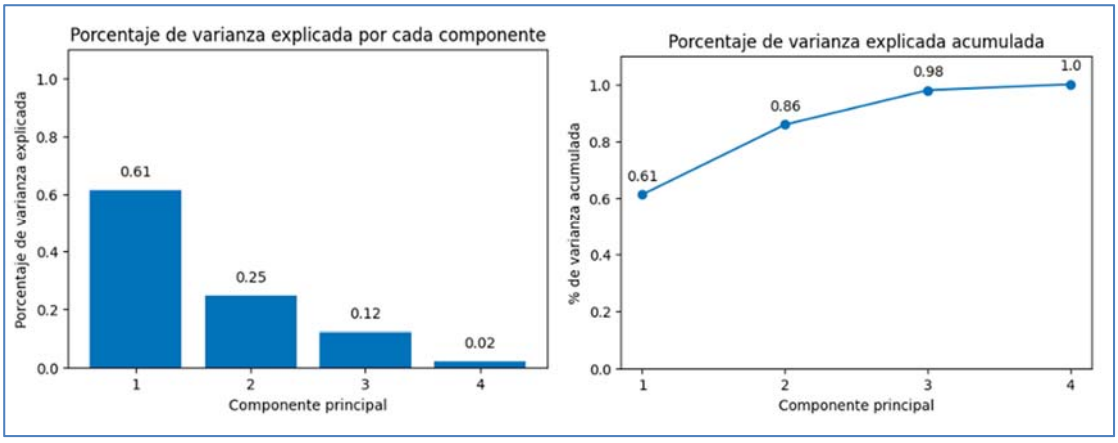


6.2 Agrupamiento de activos

La combinación de análisis de componentes principales con técnicas de *clustering* ofrece una herramienta muy útil para la interpretación de mercados financieros complejos a través del agrupamiento de activos según su perfil de riesgo. Como ya se mencionó, PCA cumple un rol clave en la reducción de la dimensionalidad de grandes conjuntos de datos, disminuyendo su complejidad, eliminando componentes con baja varianza (ruido de mercado) y descorrelacionando las variables. Por su parte, los algoritmos de *clustering* como K-Means o el *clustering* jerárquico son útiles para organizar los activos en grupos homogéneos de forma automática. Esta sinergia es ideal para detectar patrones ocultos en activos, como aquellos con volatilidad similar, correlación o perfiles de riesgo similares (Kaufman & Rousseeuw, 2009; James *et al*, 2023).

Para realizar esta aplicación se extrajeron datos históricos de un período de 90 días de veinte criptomonedas utilizando la API CoinGecko y se procesó una base de datos obteniendo métricas de volatilidad, liquidez, max drawdown y antigüedad. Luego, se escalaron los datos con el objetivo de aplicar el algoritmo de PCA. Acorde a los resultados, se seleccionaron los dos primeros componentes principales para este análisis, cuyos valores propios son mayores a uno y contienen la mayor proporción de varianza explicada como se muestra en la ilustración 17. Reteniendo estos dos primeros componentes principales, se logra conservar 86 % de la varianza total de los datos originales. Esto permite trabajar con un espacio de menor dimensión mientras se mantiene una cantidad significativa de la información.

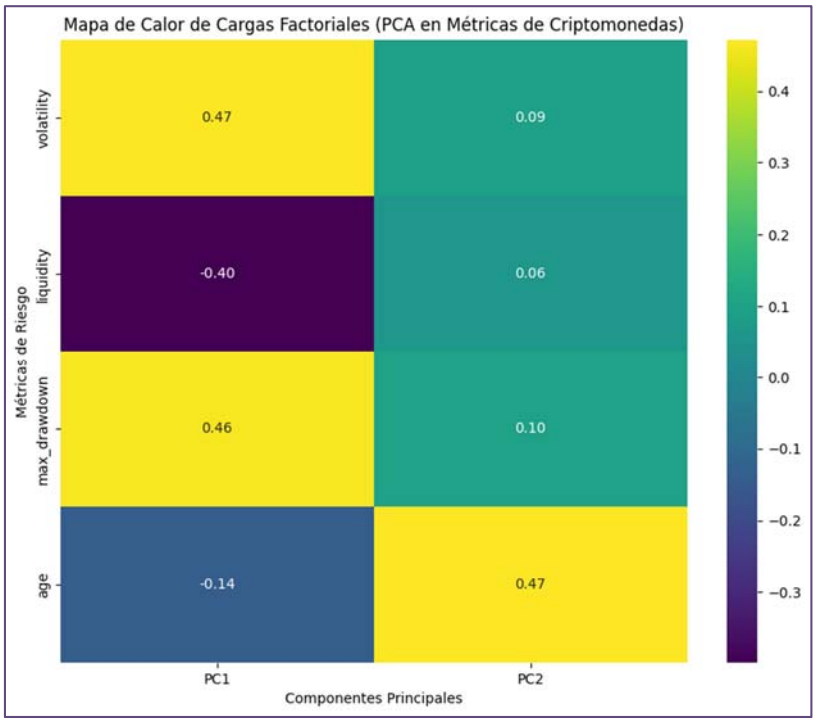
Ilustración 17: Varianza explicada por cada componente



Luego se analizaron las cargas factoriales o loadings, las cuales señalan la contribución (peso y dirección) de las variables originales a cada componente principal elegido, interpretando su significado en términos de dimensiones clave que influyen en la caracterización del perfil de riesgo. En la ilustración 18 se puede apreciar que el CP1 tiene cargas positivas altas para volatilidad (0.47) y max drawdown (0.46), y una carga negativa alta para liquidez (-0.40). Esto sugiere que captura una dimensión relacionada con el riesgo y la liquidez. Las criptomonedas con alta volatilidad, alto máximo drawdown y baja liquidez tendrán valores altos en CP1, mientras que aquellas con baja volatilidad, bajo máximo drawdown y alta liquidez tendrán valores bajos. Por otra parte, el CP2 tiene una carga positiva muy alta para antigüedad (0.47) y

cargas bajas en las otras variables. Esto indica que está fuertemente asociado con la antigüedad de la criptomoneda. Las más antiguas tendrán valores más altos en CP2, mientras que las más nuevas tendrán valores más bajos. En síntesis, el primer componente representa un factor de riesgo/liquidez, y el segundo es un factor de antigüedad y tolerancia. Ambos capturan las dimensiones más importantes de variabilidad en las métricas consideradas.

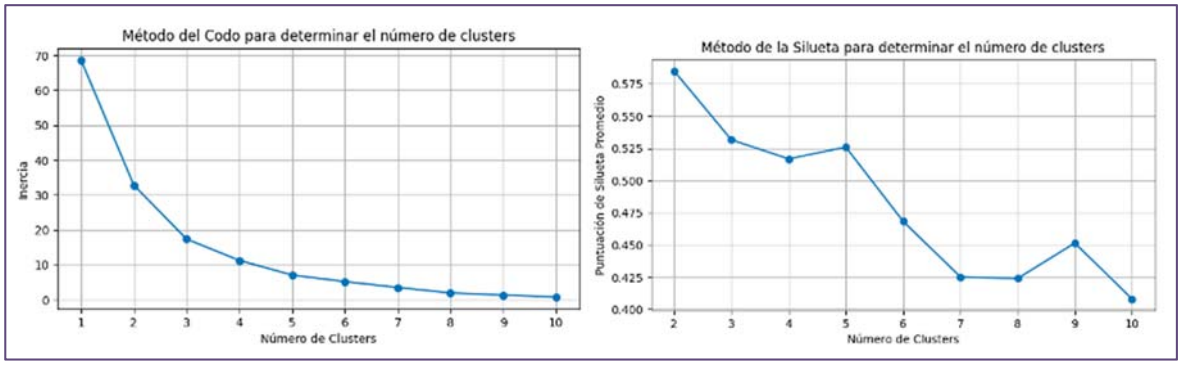
Ilustración 18: Contribución de variables originales a cada componente



Previamente a aplicar el algoritmo de *clustering*, se determinó el número óptimo de clústeres utilizando dos criterios de elección. Por una parte se consideró el método del Codo, en el cual su gráfica muestra la inercia para diferentes números de clústeres. Se calcula como la suma de las distancias cuadradas de cada punto al centro de su clúster. Se busca un codo en la curva, que indica un punto donde la disminución de la inercia comienza a ralentizarse. En la ilustración 19, a la izquierda, se observa un codo alrededor de 3 o 4 clústeres. Por otra parte, se valoró el método de la Silueta. Su gráfica muestra la puntuación de silueta promedio para diferentes números de clústeres. Una puntuación alta indica que los activos están bien agrupados dentro de su propio clúster y bien separados de otros. En la ilustración 19, a la derecha, se observa la puntuación más alta para 2 clústeres, pero 3 y 4 también tienen puntuaciones razonables. A partir de la combinación de ambos métodos y considerando la interpretabilidad de los clústeres, se decidió realizar un agrupamiento considerando tres clústeres.

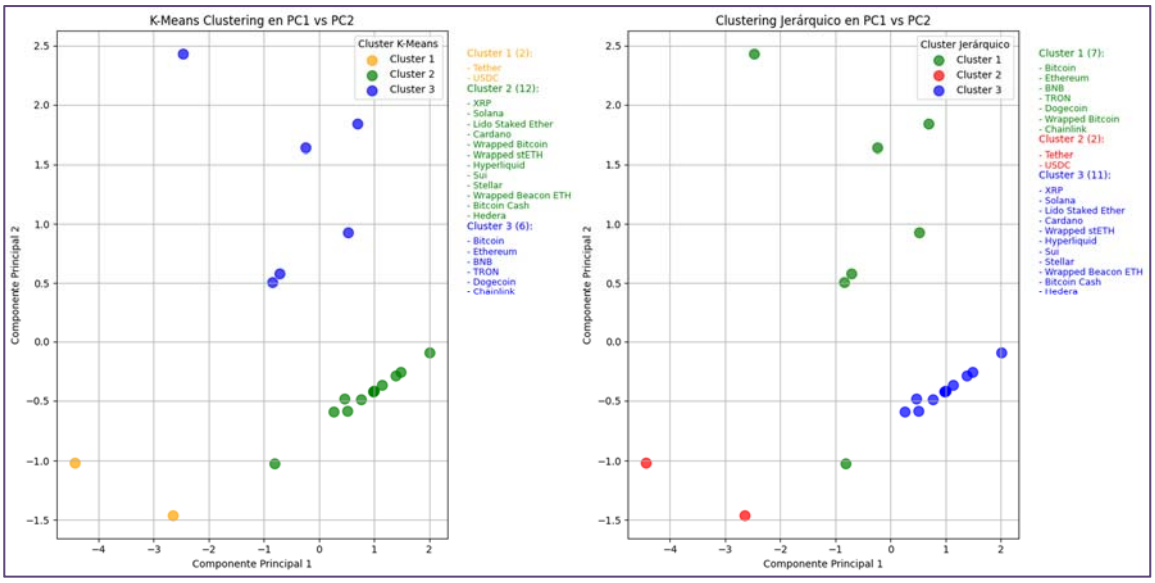
A continuación, se aplica el algoritmo de *clustering* por el método de K-Means. Paralelamente se realizó un *clustering* jerárquico para visualizar la estructura de los datos y comparar los resultados. El K-Means busca particionar el conjunto de datos en k clústeres distintos y no superpuestos. El algoritmo asigna iterativamente cada observación al centroide del clúster más

Ilustración 19: Selección del número óptimo de clústeres a considerar



cercano, utilizando la distancia euclidiana cuadrada para minimizar la variación dentro del grupo. Por otro lado, el *clustering* jerárquico presenta mayor capacidad para revelar estructuras anidadas y las relaciones entre grupos. A partir de este, se construye una representación visual en forma de dendrograma, que muestra las fusiones de los objetos. En este caso se aplicó un método aglomerativo, es decir, comienzan con cada objeto como su propio clúster y los fusionan progresivamente hasta formar uno solo. Ambas técnicas se complementan para descubrir patrones ocultos o realizar el agrupamiento de activos por perfiles de riesgo similares (Kaufman & Rousseeuw, 2009; James *et al*, 2023). En la ilustración 20 se muestra los resultados de ambos métodos y la distribución de los agrupamientos realizados en el espacio definido por los dos primeros componentes principales.

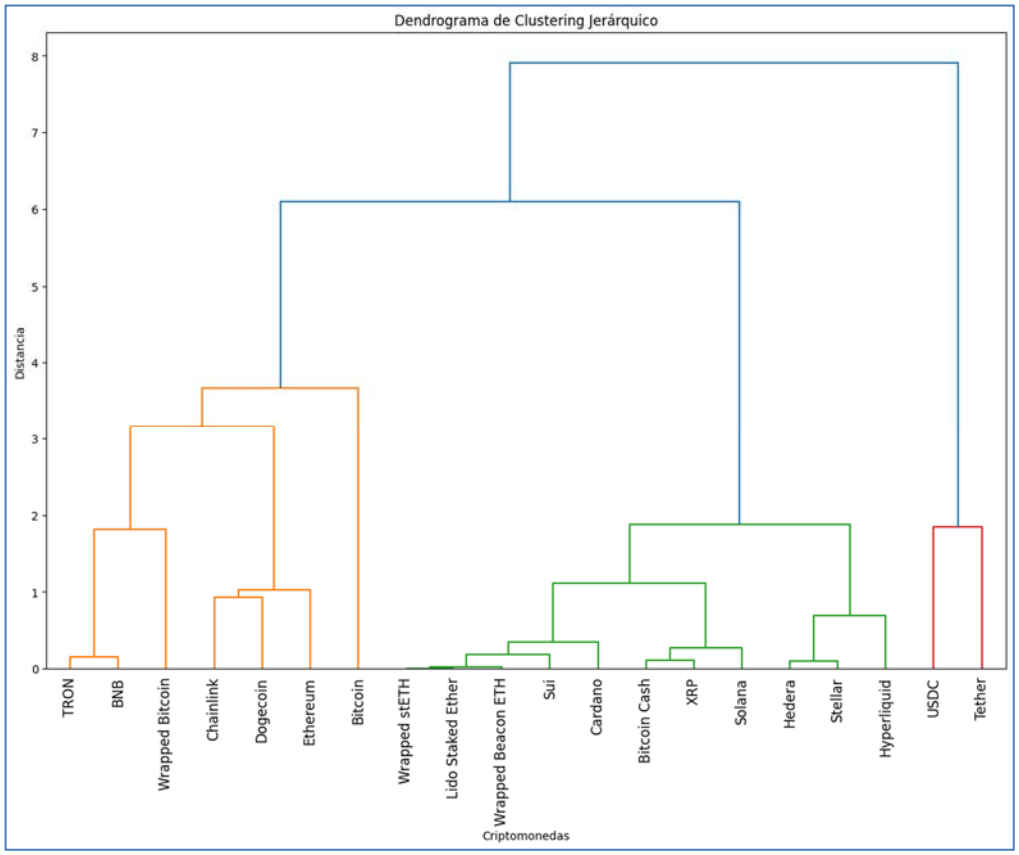
Ilustración 20: Clustering por método K-Means y método jerárquico



Se puede observar una que ambos métodos tienden a identificar agrupaciones similares en los activos. Las pequeñas diferencias podrían deberse a la naturaleza algorítmica distinta de K-Means, basado en centroides y el *clustering* jerárquico, basado en la unión progresiva de puntos

o clústeres. En la ilustración 21 se muestra el dendrograma resultante. Constituye una representación visual resultante del *clustering* jerárquico y actúa como un mapa de relaciones muy útil en el análisis de estos activos financieros. Muestra cómo se agrupan los activos de forma natural antes de definir clústeres rígidos. La altura de fusión en el dendrograma indica la disimilitud entre los grupos, con fusiones más bajas señalando mayor similitud y fusiones más altas, mayor diferencia. Esto facilita la exploración de mercados complejos como el de criptomonedas, donde se pueden develar outliers o activos con comportamientos atípicos. Además, permite a los analistas financieros elegir un número apropiado de clústeres simplemente haciendo un corte horizontal a la altura deseada, lo que es relevante para la diversificación de carteras al comprender mejor las relaciones entre los activos (Kaufman & Rousseeuw, 2009).

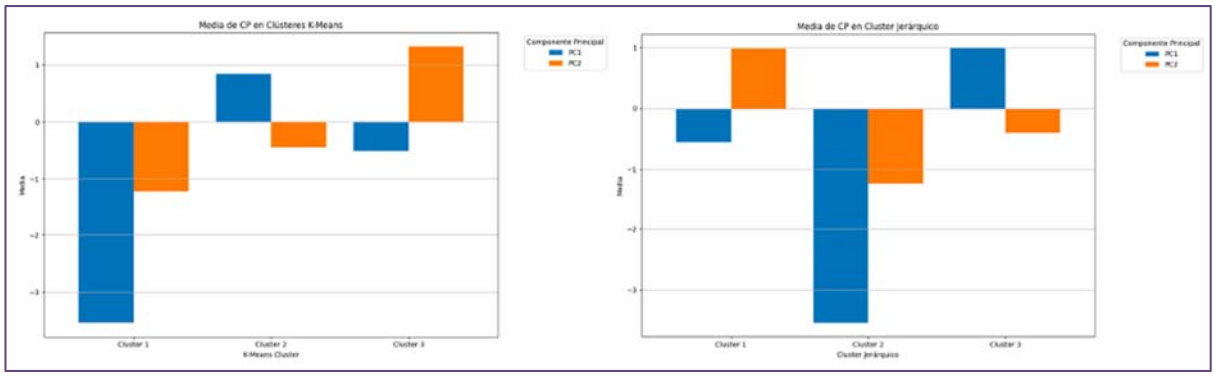
Ilustración 21: Dendrograma de clustering jerárquico



Para continuar con el análisis, se detalla la composición media de cada componente principal por cada uno de los tres clústeres hallados (ilustración 21). Esta representación facilita la comprensión de cómo cada clúster se posiciona en relación con los factores subyacentes identificados por PCA, permitiendo visualizar qué dimensiones son más prominentes o distintivas dentro de cada grupo. Al examinar los valores promedio de los CP para cada clúster, se puede inferir las características latentes que definen a cada segmento y cómo estas características contribuyen a sus perfiles de riesgo.

Se puede observar que el clúster 1 K-Means se corresponde completamente con el clúster 2 jerárquico y contiene dos criptomonedas, Tether y USDC, tiende a tener baja volatilidad y bajo

Ilustración 22: Composición media de CP por cada clúster



max drawdown, con una mayor liquidez. También tienden a ser más antiguas. El modelo logró identificarlas como stablecoins. Su función no es especulativa, sino preservar el valor y facilitar transacciones. Actúan como el efectivo digital que permite operar sin la volatilidad característica del mercado.

El Clúster 2 K-Means se corresponde con el clúster 3 Jerárquico y contienen 11 criptomonedas en común, XRP, Solana, Lido Staked Ether, Cardano, Wrapped Bitcoin, Wrapped stETH, Hyperliquid, Sui, Stellar, Wrapped Beacon ETH, Bitcoin Cash, Hedera. Se caracteriza por alta volatilidad y alto max drawdown. La liquidez es baja en este grupo, y la antigüedad es menor en comparación con el clúster 1 K-Means. El modelo logró identificarlas como altcoins. Son proyectos más jóvenes, con mayor potencial de crecimiento pero con alto riesgo y gran volatilidad. Muchas de estas suelen amplificar los movimientos de bitcoin, tienen menos adopción probada y son más susceptibles a noticias y fallos técnicos.

El Clúster 3 (K-Means) se corresponde con el clúster 1 (jerárquico) y contiene 6 criptomonedas, bitcoin, ether, BNB, TRON, dogecoin, Chainlink. Muestra una volatilidad y max drawdown moderados. La liquidez es moderada a alta, y la antigüedad es intermedia entre los otros dos clústeres. Este grupo representa los blue chips del espacio digital. Ofrecen un equilibrio entre riesgo y estabilidad. No son tan volátiles como las altcoins más nuevas, pero tienen un potencial de apreciación mucho mayor que las stablecoins. Su antigüedad intermedia y su liquidez moderada a alta atraen tanto a inversores institucionales como a minoristas que buscan exposición al mercado sin adentrarse en la frontera más especulativa.

7. Conclusiones

A lo largo de este artículo se abordó la aplicabilidad de diferentes técnicas de *machine learning* para el análisis de datos de activos financieros tanto del mercado tradicional como de criptoactivos. El objetivo primordial fue analizar la utilidad particular de diferentes modelos como herramientas estratégicas de soporte orientadas a reducir la incertidumbre en los mercados, generando insights predictivos y analíticos que fundamentan la toma de decisiones estratégicas. Dada la divergencia estructural, operativa y de comportamiento entre los mercados

financieros, se destacó la necesidad de una adaptación metodológica en la aplicación de estas técnicas, ya que los mercados tradicionales se caracterizan por datos estructurados de mayor calidad, mientras que los criptoactivos operan con datos heterogéneos, no lineales y no estructurados.

Los métodos de procesamiento del lenguaje natural (NLP), mediante modelos como *transformers* (BERT) y VADER, constituyen una herramienta clave para el análisis de sentimiento de noticias y redes sociales. El modelo BERT es eficaz para analizar el tono y matices en documentos complejos de mercados tradicionales, mientras que VADER es rápido y eficiente para el flujo constante de opiniones informales en criptoactivos. La combinación estratégica de ambos modelos permite detectar señales tempranas y anticipar movimientos.

Para la predicción de precios del día siguiente, en mercados volátiles como el de criptoactivos, las redes neuronales recurrentes (RNN), en particular las arquitecturas LSTM, demostraron un rendimiento superior respecto a los modelos de regresión lineal y *random forest*. Las LSTM lograron una reducción sustancial en todas las métricas de error, lo que se atribuye a su capacidad inherente para manejar la alta volatilidad, patrones no lineales y dependencias a largo plazo que caracterizan a las series de tiempo de criptoactivos.

Por su parte, la regresión lineal demostró un entendimiento claro de las relaciones lineales, siendo un punto de partida para la predicción de precios del día siguiente de activos en mercados con dinámicas más estables como los tradicionales. Los *random forests*, se adaptaron mejor a la complejidad y volatilidad de los mercados modernos, como el de criptoactivos, al capturar patrones no lineales y reducir la varianza, aunque con un costo en la interpretabilidad directa.

La elección óptima de los métodos predictivos dependerá siempre de los objetivos específicos del problema y las características intrínsecas de los datos financieros a analizar. Mientras que la regresión lineal y *random forest* pueden ofrecer una primera aproximación rápida y capturar la tendencia general utilizando características simples, las redes neuronales LSTM se perfilan como un modelo más adecuado para la predicción del precio de criptoactivos debido a su capacidad para modelar la complejidad y la volatilidad de las series temporales financieras.

La regresión logística demostró ser un método adecuado para predecir categorías de riesgo (alto/bajo riesgo) en activos financieros, sin requerir supuestos estrictos como la normalidad multivariada. Su aplicación para criptomonedas como bitcoin y ether reveló patrones diferenciados de riesgo, donde la volatilidad y el ratio de sharpe fueron predictores clave, pero con pesos e impactos distintos en cada criptoactivo. El volumen mostró un efecto calmante en BTC y amplificador en ETH, sugiriendo dinámicas de liquidez distintas y la necesidad de enfoques diferenciados en la gestión de riesgo.

Las técnicas de aprendizaje no supervisado, como el análisis de componentes principales y *clustering*, fueron relevantes para la exploración de la estructura inherente y la identificación de patrones ocultos en datos financieros no etiquetados. El método de PCA es ideal para reducir la dimensionalidad, filtrar el ruido de mercado y detectar anomalías mediante el error de reconstrucción, lo que ayuda a identificar eventos atípicos en activos. El *clustering*, en combinación con PCA, permitió segmentar criptoactivos en grupos homogéneos según su perfil de riesgo, logrando identificar categorías como stablecoins, altcoins y blue chips digitales.

El éxito de la implementación de *machine learning* en finanzas depende críticamente de la calidad de los datos, la selección del algoritmo adecuado al problema y la validación rigurosa de resultados para evitar sobreajustes o sesgos. Además, la interpretación contextual de los re-

sultados y la complementación con el conocimiento del mercado son esenciales para una comprensión profunda y la toma de decisiones informadas. Reconociendo que no existe un modelo único que se adapte a todas las complejidades de los mercados financieros, la clave es combinar distintas metodologías, ajustando parámetros, validando resultados y priorizando la interpretabilidad. Por ello, a futuros trabajos se indagará en modelos de *machine learning* híbridos y métodos de simulación avanzados con la finalidad de construir un marco analítico más comprensivo y adaptable, en la búsqueda de herramientas más efectivas y precisas para el análisis de los mercados financieros.

REFERENCIAS

Hair, J. F., Black, W. C., Babin, B. J. & Anderson, R. E. (2019). *Multivariate data analysis,* 8th Ed. Cengage Learning.

James, G., Witten, D., Hastie, T., Tibsshirani, R. & Taylor, J. (2023). *An introduction to statistical learning with applications in Python.* Springer

Jolliffe, I. T. & Cadima, J. (2016). *Principal component analysis: A review and recent developments*. Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 374 (2065): 20150202.

Kaufman, L. & Rousseeuw, P. J. (2009). *Finding groups in data: An introduction to cluster analysis*. Wiley.

Kong, Y., Nie, Y., Dong, X., Mulvey, J. M., Poor, H. V., Wen, Q. & Zohren, S. (2024). *Large language models for financial and investment management: Models, opportunities, and challenges*. Journal of Portfolio Management, 51 (2): 211-231.

Liang, Y., Liu, Y., Wang, N., Yang, H. Zhang, B. & Dan Wang, C. (2025). *FinGPT: Enhancing sentiment-based stock movement prediction with dissemination-aware and context-enriched LLMs*. arXiv preprint arXiv: 2412.10823v2.

Tabachnick, B. G. & Fidell, L. S. (2019). *Using multivariate statistics,* 7th Ed. Pearson.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. & Polosukhin, I. (2017). *Attention is all you need*. 31st Conference on Neural Information Processing Systems (NIPS 2017): 6000-6010.