UNIVERSIDAD PERUANA DE CIENCIAS APLICADAS
FACULTAD DE INGENIERÍA

CARRERA PROFESIONAL
DE CIENCIAS DE LA COMPUTACIÓN

Asignatura:

CC57- MACHINE LEARNING

Sección: CC72

TRABAJO FINAL

Modelo de selección y predicción del mercado de las criptomonedas

Autor

Samuel Esteban Cano Chocce – U202116508

Eduardo Elías Puglisevich Vergara – U20201e850

Nicolás Miguel Guerrero Icochea – U202115535

Profesor

Patricia Daniela Reyes Silva

Lima, abril de 2024

TABLA DE CONTENIDOS

I. SITUACIÓN DE CONTEXTO REAL

Con el avance tecnológico de las últimas décadas, ha aparecido en el mercado otro tipo de dinero distinto al billete o las monedas convencionales: las criptomonedas. También conocidas como cripto activos o criptodivisas, son un tipo de activo digital completamente descentralizado que utiliza protocolos criptográficos para asegurar las transacciones, verificar su autenticidad y preservar el anonimato de los usuarios (Martin, 2022). A diferencia del dinero tradicional, las criptomonedas no tienen una forma física y operan fuera del control de instituciones centralizadas como los bancos. Estos activos se basan en una red peer-to-peer, un sistema completamente descentralizado llamado Blockchain. Este sistema se puede describir como una red de computadoras que se basa en el consenso, es decir, busca alcanzar un acuerdo entre todos los usuarios respecto a la información contenida en cada registro digital. Estos registros son ampliamente conocidos como bloques y se validan mediante la resolución de problemas criptográficos por agentes llamados mineros.

Por otro lado, según Egaña (2018), las criptomonedas ofrecen beneficios tanto a la economía como para el individuo. En términos económicos, ayudan a reducir los costos de las transacciones y a aumentar su velocidad; ya que, al eliminar intermediarios, los costos se reducen y el procesamiento se acelera considerablemente. Además, su impacto en el mercado económico y tecnológico es innegable, ya que el uso de criptomonedas impulsa la creación de nuevos negocios, plataformas y hardware especializado, proporcionando una alternativa para el comercio y las transacciones. En cuanto a los beneficios individuales, las criptomonedas ofrecen seguridad, transparencia y anonimato que las monedas tradicionales no pueden garantizar, además de ser una gran opción de inversión en la actualidad.

Por este último punto, las criptomonedas han aumentado su popularidad significativamente en los últimos años, ya que representan un nuevo tipo de inversión con el potencial de ofrecer retornos muy lucrativos. Sin embargo, al igual que con otros métodos de inversión, las criptomonedas conllevan riesgos. Por ejemplo, la volatilidad extrema inherente a este tipo de activo puede llevar a pérdidas sustanciales en un corto lapso de tiempo y, al ser activos no regulados por ninguna institución, los inversores pueden enfrentarse a la falta de recursos para buscar reclamaciones o responsabilidades si pierden el dinero por errores humanos. A pesar de estos riesgos, según Gordon (2023), la gestión del riesgo desempeña un papel crucial en la inversión en criptomonedas. Un análisis detallado y cuidadoso de la gestión del riesgo puede aumentar significativamente las posibilidades de éxito en este tipo

de activos. Estos análisis se suelen dividir en dos tipos, análisis fundamental y análisis técnico. Por un lado, el análisis fundamental se centra en profundizar en los fundamentos del proyecto que la respalda, como conocer las metas del proyecto, conocer quiénes son los implicados del proyecto, tanto desarrolladores como sponsors, así como investigar cual es el respaldo financiero con el que cuentan (Kriptomat, 2023). Además, de ser posible, la búsqueda de competidores que ofrezcan servicios similares dentro del mercado cripto es otro método que puede ayudar a validar si la criptomoneda elegida es o no un activo con potencial de crecimiento. El análisis técnico abarca el análisis de gráficas, datos o estadísticas para encontrar patrones a través de estas. Evidentemente, profundizar en este tema puede resultar complejo, pero es importante al menos comprender los conceptos básicos: por ejemplo, es fundamental conocer los tipos de gráficos que se pueden encontrar en la web y su significado (Kriptomat, 2023). Entre los gráficos más comunes se encuentran los gráficos de velas japonesas, los gráficos de líneas y los gráficos de barras, siendo los gráficos de velas japonesas los más utilizados y recomendados para aprender. Estos gráficos detallan el precio de un criptoactivo en un período de tiempo específico, lo que puede ayudar a los inversores a tomar decisiones informadas sobre sus inversiones en criptomonedas.

En este tipo de análisis nos adentramos en lo que se conoce como análisis on-chain, el cual consiste en extraer, analizar y comprender la gran cantidad de información y métricas que la misma blockchain proporciona de forma pública. Dado que se trata de un sistema descentralizado, la información de la cadena de bloques es totalmente transparente, inmutable, gratuita y accesible para cualquier persona en el mundo (Clarke, 2024). Comprender y analizar estos datos on-chain brinda a los inversores una ventaja significativa sobre aquellos que no están familiarizados con este enfoque, lo que puede marcar la diferencia entre una buena o mala inversión. Sin embargo, lograr este entendimiento puede ser una tarea ardua y extensa que requiere horas de dedicación y esfuerzo. Entre los datos on-chain más relevantes se encuentran el precio, el recuento de transacciones, la frecuencia de transacciones, el movimiento de tokens y el número de direcciones de billeteras.

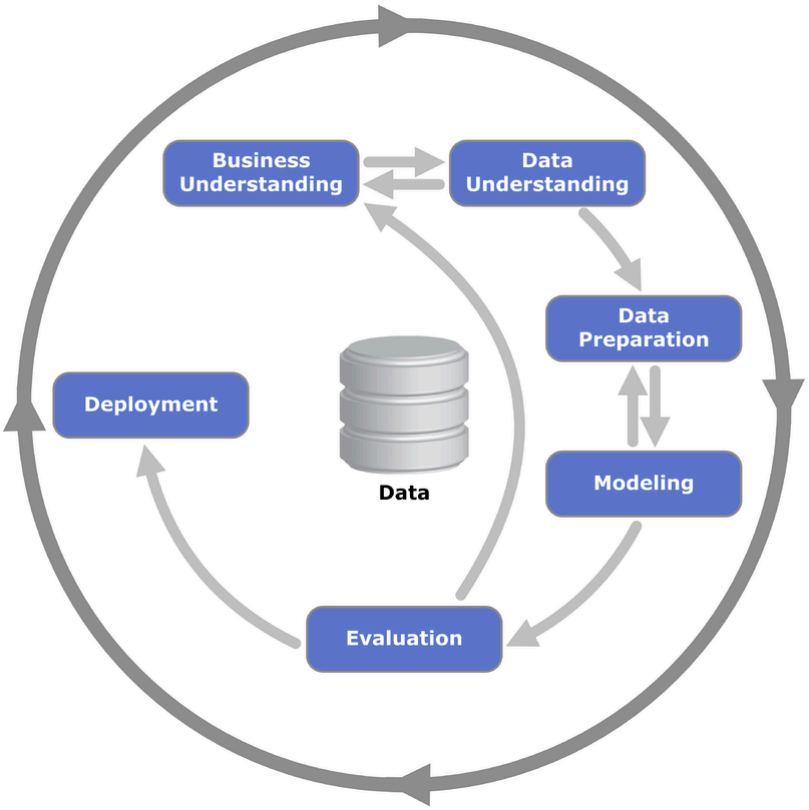
El objetivo del presente trabajo es extraer datos relevantes del mercado cripto para comprender mejor la correlación entre los datos y desarrollar un modelo de aprendizaje automático para la predicción del aumento de precios de las criptomonedas. Esto ayudará significativamente a las empresas que desean invertir en criptomonedas nuevas, pues conocer y predecir el comportamiento de las monedas permite anticiparse frente a otros inversores y obtener más ganancias.

II. PROPUESTA

La propuesta del presente informe se alinea con la metodología de trabajo CRISP-DM, la cual consiste en las fases de comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue (IBM, 2021). Para el primer hito, solo se han contemplado las primeras tres fases, tal como se muestra en el cronograma de la Figura 1.

Figura 1

Cronograma de fases de la metodología CRISP-DM según hitos



- **Comprensión del negocio:** En esta fase se busca comprender a fondo el mercado de las criptomonedas. Para ello, se debe plantear preguntas clave que puedan ayudar a comprender mejor qué factores influyen en el comportamiento de las criptomonedas y qué cualidades aumentan la probabilidad de un alza en su precio. Las preguntas planteadas fueron: ¿Qué factores influyen en el crecimiento de las criptomonedas? ¿Qué factores en común tienen las criptomonedas que presentaron un alza significativa? ¿De qué formas se puede medir el crecimiento de una criptomoneda? Al

término de esta fase, se logró obtener un mayor conocimiento en cuanto al mercado de criptomonedas y qué factores influyen en el crecimiento y el precio de una criptomoneda.

- **Comprensión de los datos:** En esta fase se realiza la extracción, exploración y análisis de los datos. Durante esta fase, se plantearon preguntas que puedan ayudar a identificar qué fuentes y qué tipos de datos ayudarán a abordar el tema planteado. Se plantearon las siguientes preguntas: ¿Qué fuentes de datos disponibles se tienen? ¿Qué características relevantes se pueden extraer de estas fuentes? ¿Se deben realizar actividades de extracción adicionales? ¿Qué tan buena es la calidad de los datos? Al término de esta fase, se optó por utilizar a Crypto Compare y su API como fuente de extracción de datos. De esta fuente de datos se pudo extraer características relevantes como la capitalización de mercado, suministro actual, volatilidad, entre otros. Las características extraídas se detallarán a profundidad en el punto IV del informe.
- **Preparación de los datos:** Una vez obtenidos los datos a analizar, en esta fase se realizó las actividades necesarias para que los datos tengan una buena calidad. Se realizaron actividades de limpieza de datos y creación de nuevas características que puedan ayudar a abordar la problemática. Como es usual, en la fase de extracción de datos se presentaron desafíos que se tuvieron que solucionar mediante la extracción y/o limpieza manual de datos. Por ejemplo, entre los principales problemas al intentar buscar características específicas de vital importancia para nuestro proyecto como la capitalización del mercado o la cantidad de suministro de la moneda no estaban disponibles en las APIs usadas para anteriores trabajos tales como CoinMarketCap, y si estaban eran de pago como es el caso de CoinGecko. Después de buscar extensivamente, encontramos la API de Crypto Compare que como se mencionó anteriormente fue usada como fuente principal de extracción.

Técnicas de aprendizaje utilizadas

Para el trabajo final, se utilizaron tanto técnicas de aprendizaje no supervisado como supervisado. A continuación, se explica cómo fueron usadas las técnicas respectivas.

- **Análisis de Componentes Principales (PCA):** Este algoritmo de aprendizaje no supervisado consiste en una técnica de reducción de dimensionalidad que puede ayudar a identificar patrones y relaciones entre las características de los datos. Al

aplicar PCA a características relacionadas con criptomonedas, se podría descubrir qué variables o características tienen más peso en la determinación de la popularidad.

- **Análisis de Clúster (K-Means):** Los algoritmos de clúster, como lo es K-Means en el presente trabajo, pueden agrupar criptomonedas similares en grupos o clústeres en función de características compartidas. De esta manera se identifican grupos de criptomonedas con estadísticas similares y se comprende qué factores las distinguen de otras independientemente de la clase a la que pertenezcan.
- **Isolation Forest:** es un algoritmo de aprendizaje no supervisado diseñado principalmente para la detección de anomalías. Este método de aislamiento crea un conjunto de árboles de decisión de forma aleatoria y mide el número de particiones necesarias para aislar una observación. Las observaciones que requieren menos particiones son consideradas anomalías. Al aplicar Isolation Forest a datos relacionados con criptomonedas, se puede identificar qué transacciones, patrones de trading, o características de criptomonedas son atípicas y podrían indicar comportamientos inusuales o fraudulentos. Esto permite una mayor seguridad y vigilancia en el mercado de criptomonedas.
- **VSM One-Class**: Es un enfoque de aprendizaje no supervisado utilizado para la detección de anomalías y la clasificación de datos en un solo grupo de referencia. Este método busca encontrar un hiperplano que maximice la separación entre los datos y el origen, encapsulando la mayor cantidad de datos de una clase particular. Al aplicar VSM One-Class a características de criptomonedas, se puede determinar cuáles transacciones o comportamientos se desvían significativamente de la norma establecida, permitiendo la detección de irregularidades y la identificación de criptomonedas que no siguen los patrones usuales del mercado.
- **Regresión Lineal:** La regresión lineal es otra técnica de aprendizaje supervisado que se utiliza para modelar la relación entre una variable dependiente continua y una o más variables independientes. En el análisis de criptomonedas, la regresión lineal puede ser útil para predecir el valor futuro de una criptomoneda en función de sus características actuales, de esta manera determinaremos si duplicó o triplicó el valor de ésta con respecto al anterior halving.

III. ADQUISICIÓN Y PRE PROCESAMIENTOS DE LOS DATOS

Origen de los datos

Como se mencionó anteriormente, los datos de las criptomonedas que se extrajeron provienen de la API de Crypto Compare con una licencia gratuita. En la Tabla 1 se lista las características extraídas y su descripción.

Tabla 1

Características extraídas con la API Crypto Compare

Característica	Descripción
time	Fecha del día de las características en segundos
Symbol	Código de la criptomoneda.
new_addresses	Cantidad de nuevas direcciones de billeteras.
active_addresses	Cantidad de direcciones de billeteras activas.
transaction_count	Cantidad de transacciones
Large_transaction_count	Cantidad de transacciones grandes
average_transaction_value	Valor promedio de transacciones
current_supply	Cantidad actual de monedas en circulación.
high	Valor más alto de la criptomoneda en las últimas 24 horas.
low	Valor más bajo de la criptomoneda en las últimas 24 horas.
open	Precio de la criptomoneda a las 00:00 horas del día.
volumefrom	Volumen total de la criptomoneda de origen
volumeto	Volumen total de la criptomoneda de destino
close	Precio de cierre de la criptomoneda a las 23:59 horas del día.
date	Fecha del día de las características en formato yyyy-mm-dd
volatility	Medida de magnitud de los cambios de los precios de las monedas.

Nota. Elaboración propia.

Análisis Exploratorio de los Datos (EDA)

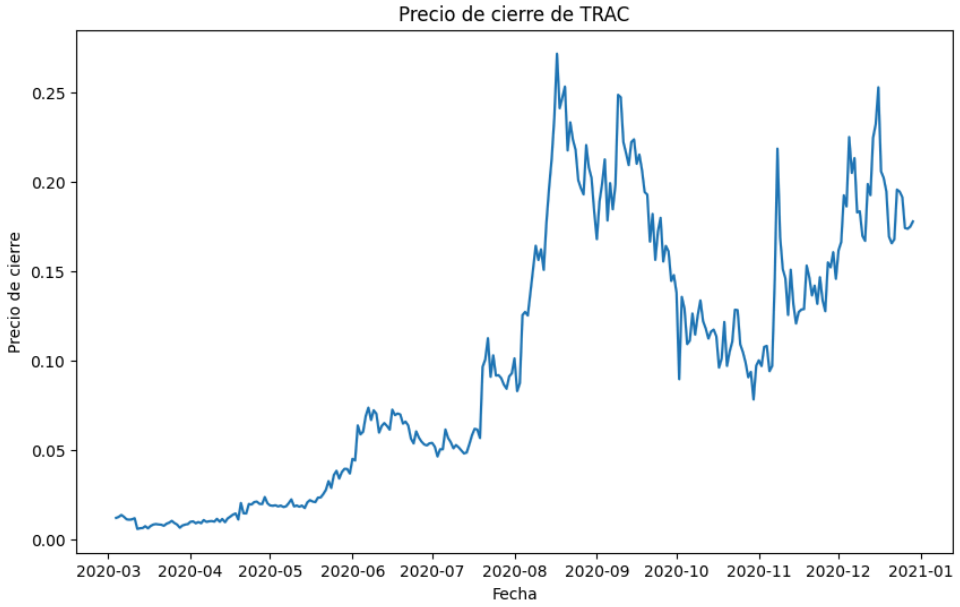
El Análisis Exploratorio de Datos (EDA) es el primer paso crítico en el análisis de datos. El objetivo fue resumir las principales características de los datos y detectar patrones,

anomalías y relaciones preliminares entre variables, lo cual es crucial para formular hipótesis y diseñar modelos predictivos efectivos. A continuación, se detalla el procedimiento llevado a cabo para realizar el análisis EDA.

1. **Carga de datos e inspección inicial:** Luego de haber recolectado los datos mediante la API de Crypto Compare, inicialmente el conjunto de datos estaba compuesto por muchas columnas y registros con datos nulos. Específicamente, se tenían 36120 registros de 45 columnas, en su mayoría compuestos por datos numéricos.
2. **Análisis de valores faltantes:** Se continuó con un análisis de datos nulos. En esta fase, existen distintas técnicas para corregir la data faltante como la eliminación de registro o columnas, o el rellenado de datos. En esta ocasión, se decidió eliminar las columnas que presentaban un porcentaje de datos nulos mayor al 80% y, posteriormente, se rellenó los datos faltantes con 3 métodos de interpolación *bfill, ffill* y *linear interpolate.* Se asignó una puntuación a cada método por cada columna del conjunto de datos y se realizó el rellenado de datos siguiendo el método de interpolación con la puntuación más alta de las 3. Luego de esto, el conjunto de datos final fue de 22876 registros y 14 columnas.
3. **Creación de columnas:** Luego de la limpieza de datos, se crearon dos nuevas columnas en el conjunto de datos. Estas son las columnas de *date,* la cual es simplemente la fecha en formato *yyyy-mm-dd,* la columna *volatility,* la cual se calculó mediante el retorno logarítmico. Asimismo, se creó la columna *market_cap* de ese día que simplemente es la multiplicación del suministro actual por el precio de cierre de ese día. Por último, se creó una columna *diff_high_low* que representa la diferencia entre el precio máximo y mínimo de aquel día. Luego de un nuevo renombramiento y eliminación de algunas columnas, el conjunto de datos resultante t
4. **Visualización de los datos:** Luego del procesamiento, se realizaron algunas visualizaciones para obtener un mejor concepto de los datos que tenemos. En la siguiente Figura 2, se ha realizado la gráfica del precio de cierre de la moneda TRAC a través del tiempo.

Figura 2

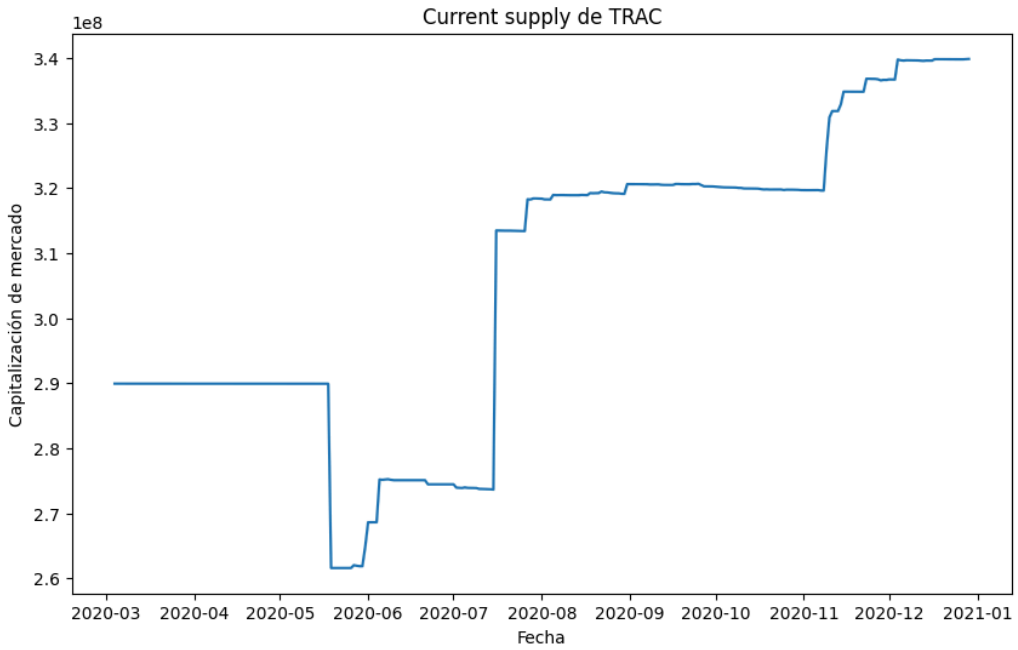
Precio de cierre de TRAC a través del tiempo



Nota. Elaboración propia.

Figura 3

Suministro diario de la moneda TRAC

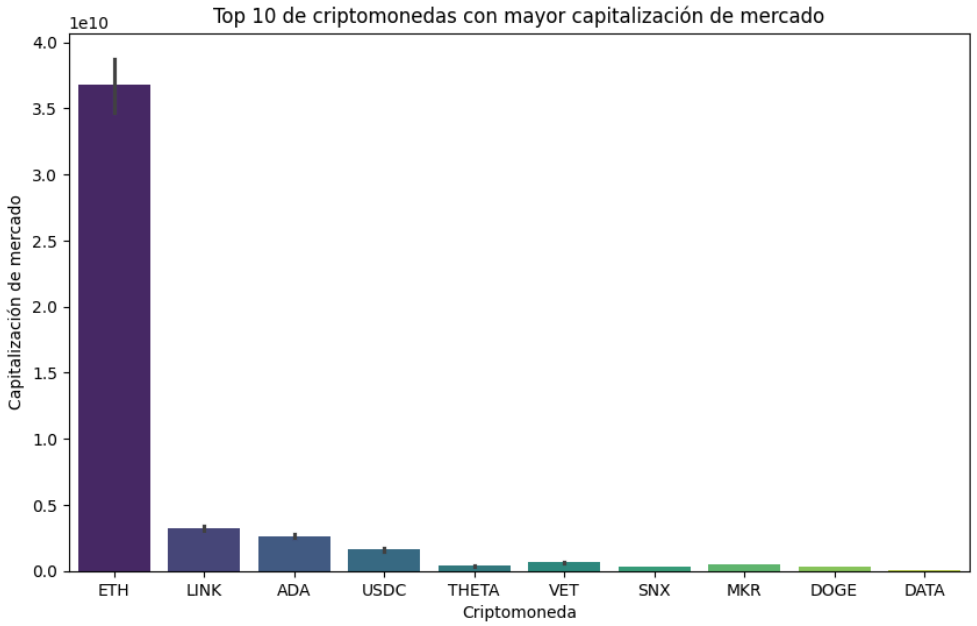


Nota. Elaboración propia.

En la siguiente Figura 4, se graficó la distribución de monedas según la mayor capitalización de mercado registrada.

Figura 4

Top 10 monedas con mayor capitalización de mercado

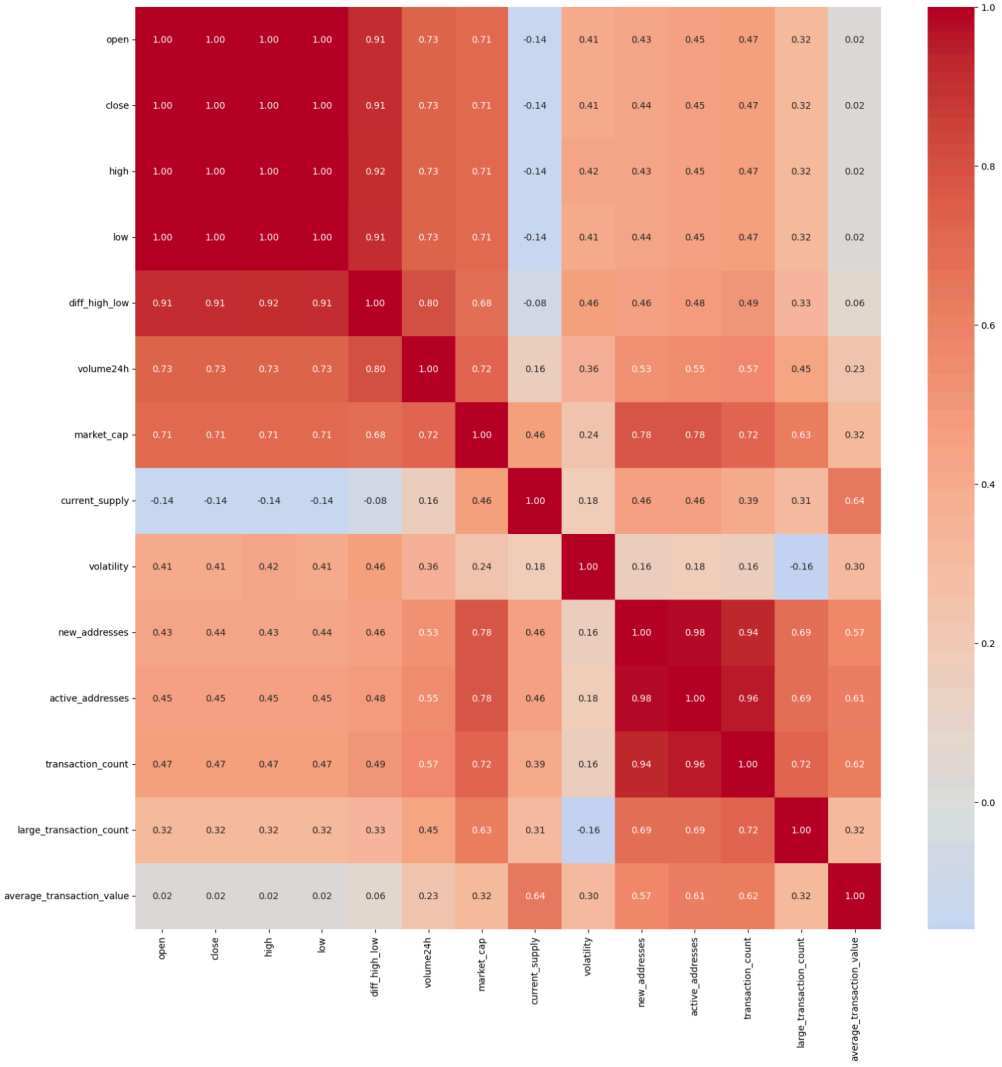


Nota. Elaboración propia.

Por último, se graficó el mapa correlacional de las variables del conjunto de datos en la Figura 5.

Figura 5

Mapa de calor



Nota. Elaboración propia.

IV. INGENIERÍA DE CARACTERÍSTICAS

Luego del análisis EDA, se procede a la selección final de características a utilizar en los modelos. En el caso del clustering, se decidió agrupar las variables mencionadas anteriormente en base a métricas estadísticas como el máximo, la media, la suma y la desviación estándar. De esta manera, el clustering determinará los patrones intrínsecos de las criptomonedas analizadas y los grupos formados pasarán al segundo modelo de predicción. Dado el rango de fechas requeridas, las características obtenidas tuvieron que recolectarse de una fuente externa al repositorio de datos de las entregas anteriores porque ningún grupo había recolectado métricas importantes como la capitalización del mercado, el punto más alto y bajo de la criptomoneda, etc.

Tabla 2

Características destinadas a la clusterización

Característica	Descripción
symbol	Símbolo del token
max_market_cap	Máxima capitalización de mercado por moneda.
std_dev_market_cap	Desviación Estándar de la Capitalización de Mercado
avg_transaction_count	Promedio del Conteo de Transacciones
total_large_transaction	Total de Grandes Transacciones
total_new_addresses	Total de Nuevas Direcciones
avg_volatility	Promedio de Volatilidad
avg_volume24h	Promedio del Volumen de 24 Horas
std_dev_volume24h	Desviación Estándar del Volumen de 24 Horas
avg_diff_high_low	Promedio de la Diferencia entre Alto y Bajo

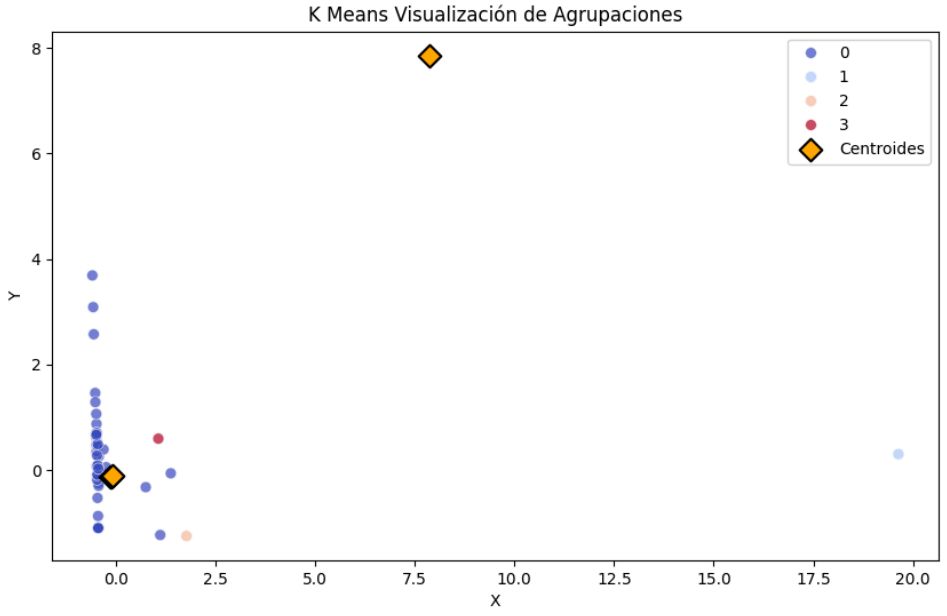
Nota. Elaboración propia.

V. EXPERIMENTOS

Experimentación con algoritmos de clustering: Para la fase de clasificación inicial se experimentó a usar los algoritmos de KNN y Mean Shift. En el caso de KNN se probó con diferentes valores de K, en el rango de 1 al 11. Mediante las técnicas de evaluación de la Silueta y del Codo, se determina el mejor valor K, que para la primera técnica es 2 y para la segunda es 4. Debido a que buscamos diversificar los grupos generados, optamos por quedarnos con K=4. Finalmente utilizamos PCA para la reducción de dimensionalidad de los datos y se crea la visualización de los agrupamientos generados, la cual se puede ver en la Figura 6.

Figura 6

Clustering con KNN

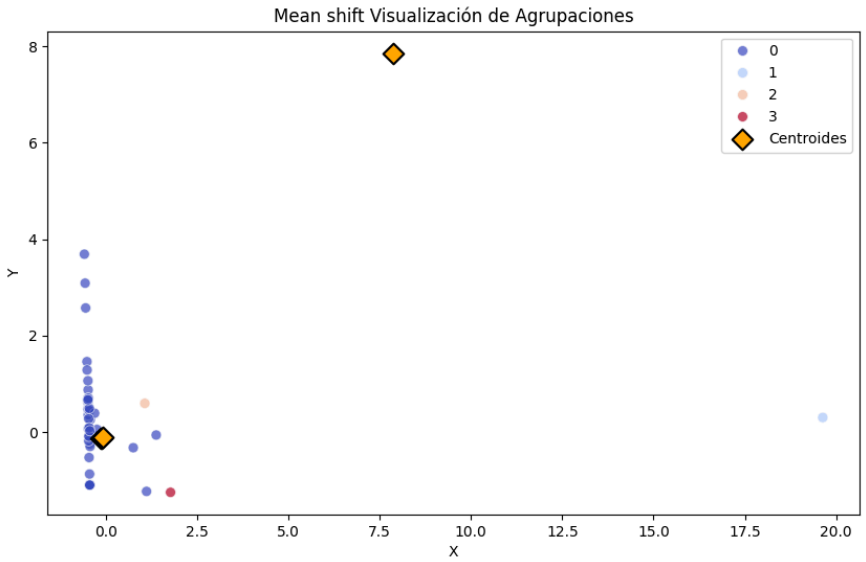


Nota. Elaboración propia.

En el caso de Mean Shift, se probó con varios anchos de banda para determinar el más adecuado. El método de la silueta indicaba que este valor rondaba alrededor de 2.5 a 3, por lo que se decidió tomar un valor de 2.5 como mejor candidato. Después del clustering, se puede visualizar en la Figura 7 cómo Mean Shift agrupó las criptomonedas.

Figura 7

Clustering con Mean Shift



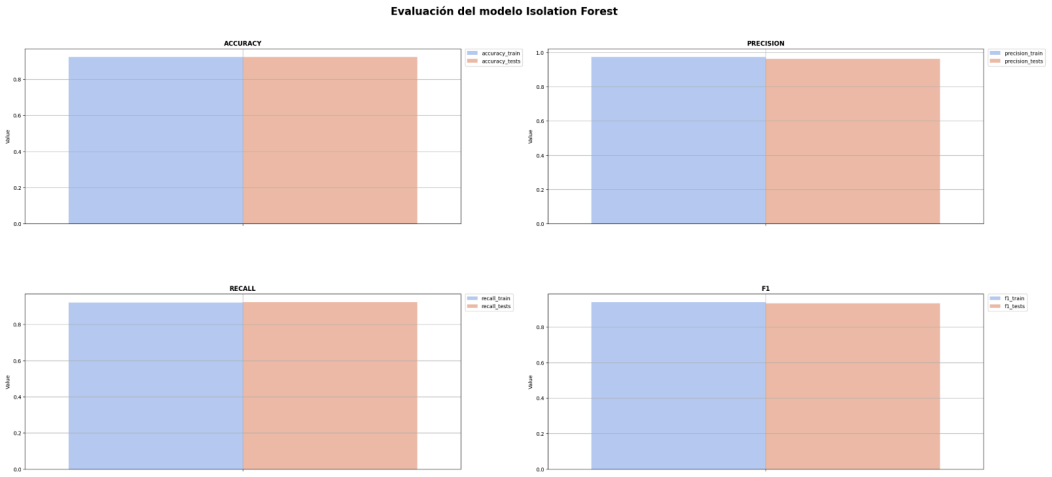
Nota. Elaboración propia.

Para ambos modelos la evaluación de la silueta fue de 0.79 aproximadamente, pero se decidió tomar Mean Shift como modelo definitivo debido a ser capaz de descubrir automáticamente el número óptimo de clusters y adaptarse a la estructura de los datos.

Detección de anomalías: Después de la fase de agrupamiento, se determinó buscó el mejor algoritmo para la detección de anomalías dentro de los grupos generados, para lo cual se probó con los algoritmos de Isolation Forest y One Class SVM. En el caso de Isolation Forest, se puede visualizar en la Figura 8 el rendimiento del modelo mediante las métricas Accuracy, Precision, Recall y F1.

Figura 8

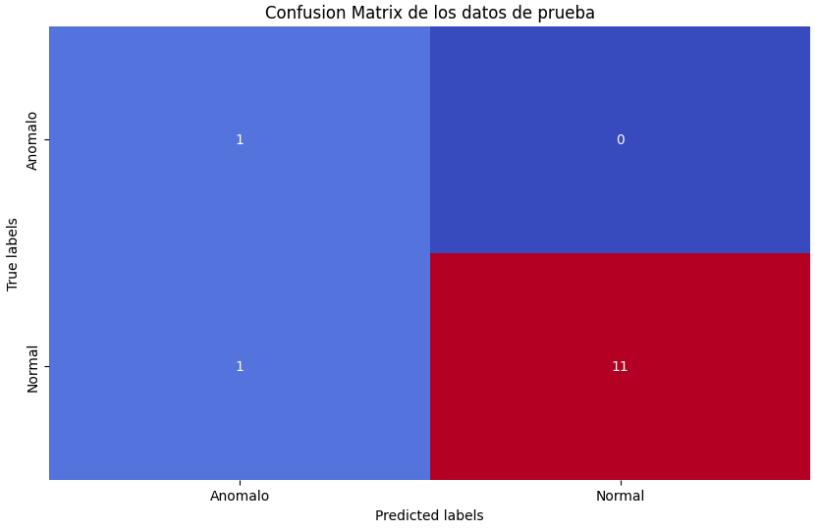
Evaluación del modelo Isolation Forest



Asimismo, se muestra en la Figura 9 cómo clasificó el modelo las anomalías mediante una matriz de confusión.

Figura 9

Matriz de confusión de datos anómalos con Isolation Forest

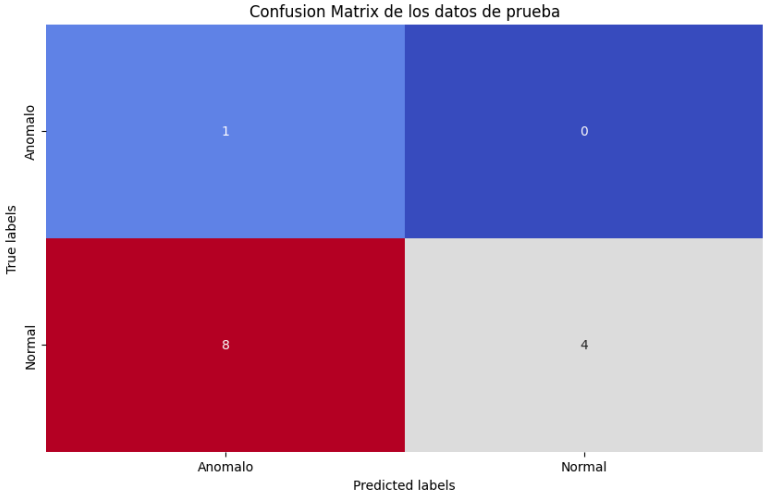


Nota. Elaboración propia.

Para el caso de One SVM, se determinó de la misma forma la matriz de confusión para detectar datos anómalos, como se puede ver en la Figura 10.

Figura 10

Matriz de confusión de datos anómalos con One SVM



IX. REFERENCIAS BIBLIOGRÁFICAS

Clarke, A. (15 de enero de 2024). *Datos on-chain: Cómo los traders se adelantan a los acontecimientos.* Cointelegraph. Recuperado el 6 de abril de 2024 de https://es.cointelegraph.com/news/on-chain-data-market-traders

Egaña, J. (2018). *Criptomonedas: Pasado, Presente y ¿Futuro?* [Trabajo de fin de grado, Universidad de Sevilla]. Depósito de Investigación Universidad de Sevilla. https://hdl.handle.net/11441/88306

Gordon, A. (2023). Evaluación del rendimiento de inversiones en criptomonedas: Riesgos y oportunidades. *Revista Ingenio global*, *2*(1), 13–24. https://editorialinnova.com/index.php/rig/article/view/58

Kriptomat. (27 de diciembre de 2023). *¿Cuáles son los elementos clave del análisis fundamental en el comercio de criptomonedas?*. Recuperado el 5 de abril de 2024 de *https://kriptomat.io/es/finanzas-e-inversion/cuales-son-los-elementos-clave-del-analisis-fundamental-en-el-comercio-de-criptomonedas/*

Martin, C. (2022). *Criptomonedas* [Trabajo de fin de grado, Universidad de Valladolid]. Universidad de Valladolid Repositorio Documental. https://uvadoc.uva.es/handle/10324/54475

IBM. (2017). *Conceptos básicos de ayuda de CRISP-DM - Documentación de IBM*. Recuperado el 1 de mayo de 2024 de https://www.ibm.com/docs/es/spss-modeler/saas?topic=dm-crisp-help-overview.