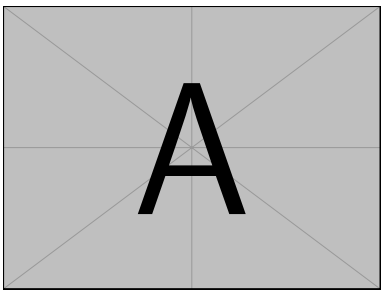

Propuesta de modelo para pronóstico del precio de Bitcoin y su clasificación para la toma de decisiones de inversión

Presentado en cumplimiento parcial de los requisitos
para el grado de

Licenciado en Matemáticas Aplicadas

por

Sacbe García García

Bajo la supervición de:

Dr. Álvaro Castañeda Mendoza
Mtr. Carlos Abraham Carballo Monsivais

Matemáticas Aplicadas
Universidad del Istmo campus Tehuantepec
Oaxaca, México - 70760
24 de marzo de 2022

Agradecimientos

Nunc sed pede. Praesent vitae lectus. Praesent neque justo, vehicula eget, interdum id, facilisis et, nibh. Phasellus at purus et libero lacinia dictum. Fusce aliquet. Nulla eu ante placerat leo semper dictum. Mauris metus. Curabitur lobortis. Curabitur sollicitudin hendrerit nunc. Donec ultrices lacus id ipsum.

Lugar: Universidad del Istmo
Fecha: 24 de marzo de 2022

Sacbe García García

Resumen

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus

enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Abstract

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus

enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Keywords: Keywords 1, Keywords 2, Keywords 3

Abreviaturas

OHLC	Open-High-Low-Close
SLR	Systematic Literature Review
AI	Artificial Intelligence
DL	Deep Learning
SES	Simple Exponential Smoothing
ETS	Exponential Smoothing
ARIMA	Autoregressive Integrated Moving Average
RTS	Regression Time Series
SVM	Support Vector Machine
RF	Random Forest
LSTM	Long Short Term Memory

Índice general

Capítulo 1

Introducción

El mercado de las criptomonedas se ha disparado en los últimos años y a comparación de las monedas tradicionales las mayores innovaciones de los activos criptográficos son establecer un nuevo sistema de pago con sistemas criptográficos sofisticados, transacciones de persona a persona, con un sistema descentralizado y que asegura la privacidad del usuario.
El Bitcoin destaca exitosamente en todos los puntos anteriores consolidándose como la criptodivisa líder del mercado. Debido a su naturaleza altamente volátil se necesitan buenas predicciones en las que basar las decisiones de inversión. En los mercados financieros tradicionales puede interesar la aplicación de estrategias comerciales automatizadas que ayudan a obtener un rendimiento razonable de los fondos invertidos.

1.1. Introducción

Bitcoin es la primera implementación de un concepto conocido como moneda criptográfica, la cual fue descrita por primera vez en 1998 por Wei Dai. La especificación inicial del protocolo Bitcoin y la prueba del concepto la publicó Satoshi Nakamoto en el 2009 [1] la cual abandonó a finales de 2010 sin revelar nada de su persona, desde entonces la comunidad ha crecido de forma exponencial.

En noviembre de 2020 el bitcoin alcanzó máximos históricos [2] consolidándose así como una moneda legitima que desde finales de 2019 se ha disparado más del 175 %, incluso gigantes de pago como Paypal y Square permiten la compra y venta de bitcoins. Países como Estados Unidos, Japón y El Salvador actualmente permiten transacciones con esta moneda, destacando este último por ser el primer país en legalizar el bitcoin como moneda legal a partir del 7 de septiembre de 2021. En contraste países como La República Dominicana o China no cuentan con el respaldo de sus bancos centrales, inclusive prohibiendo éste último la compra y venta de esta criptomoneda.

La tecnología Bitcoin es popular ya que permite evitar pasar por un intermediario para validar una transacción, genera valor al realizar las validaciones, y este valor se transforma en un elemento de intercambio. Como se ha visto en los últimos años el bitcoin puede recuperar su valor después de caídas significativas, incluso cuando la incertidumbre es alta en el mercado como durante la pandemia de la COVID-19 [3], esto es importante ya que consolida a esta moneda como un activo de inversión fiable.

Se define una moneda electrónica como una cadena de firmas digitales donde cada propietario, al transferir una monto a otra persona, firma digitalmente un *hash* de la transacción previa y la clave pública del siguiente propietario, añadiendo ambos al final de la moneda [1]. Este conjunto de transacciones es público y propagado a todos los usuarios en la red. Esto también es conocido como blockchain o cadena de bloques, donde cada bloque contiene una transacción con los hashes correspondientes, eliminando así la centralización y la confianza en un solo punto.

Se puede asegurar que alguien no puede agregar una transacción maliciosa en la blockchain ya que se utiliza una prueba de trabajo o proof of work. Para llevar a cabo esta prueba primero se tiene que convertir un bloque de la cadena de bloques en un hash y ser verificado, esto es, comprobar que el hash cumpla con un número de ceros requeridos en un orden especifico o dicho de otra forma resolver un problema criptográfico que requiere poder computacional. Al

agregarse otro bloque el atacante tiene que asegurarse que este esté conectado a su bloque malicioso por lo que tiene que competir contra el poder computacional de los demás usuarios de la red y así indefinidamente, lo que resulta inviable. La red blockchain selecciona la cadena de bloques más larga, esto quiere decir que selecciona la cadena de bloques que resuelve los problemas criptográficos más rápido, siendo la más confiable por ser la soportada por la comunidad.

Por todo lo anterior y debido a su naturaleza altamente volátil, se necesitan buenas predicciones en las que basar las decisiones de inversión. En los mercados financieros tradicionales puede interesar la aplicación de estrategias comerciales automatizadas que ayudan a obtener un rendimiento razonable de los fondos invertidos. Los métodos tradicionales (Naive, SES, Holt, ETS y ARIMA) dependen en gran medida de una hipótesis que requiere datos con estacionalidad, tendencia y ruido para que sean eficaces. Al no cumplirse estas características los mecanismos de inteligencia artificial entran en acción, considerando especialmente la naturaleza volátil y temporal de los datos como los del bitcoin [4].

1.2. Motivación

En la literatura disponible se muestran resultados limitados sobre aplicaciones y desarrollo de métodos con el uso de aprendizaje automático y trasfondo matemático para la toma de decisiones de inversión en criptomonedas [5, 4, 6]. Este campo tiene dos oportunidades de desarrollo potencial: la primera aplicando los métodos existentes en la creación de parámetros de comparación y su capacidad para caracterizar el entorno real, la segunda buscar alternativas de modelación para fines de predicción. Esta propuesta de investigación permitirá contribuir al cuerpo de conocimiento de la ciencia de datos con nuevos métodos para la caracterización de oportunidades de inversión financiera. Se espera que al final de la investigación se proponga al menos un método eficiente y robusto para facilitar a consultores de inversiones, analistas de riesgos e incluso inversionistas no calificados en la mejora de toma de decisiones para crecer sus inversiones a mediano y largo plazo.

1.3. Planteamiento del problema

Actualmente los métodos y herramientas para realizar modelación y caracterización de tendencias de comportamiento financiero están limitados al uso de técnicas simples en la

mayoría de los casos prácticos [7]. El problema a resolver es encontrar un modelo de predicción para el precio del bitcoin en dólares usando las métricas de la blockchain que más influyen en su comportamiento utilizando reducción de dimensionalidad y análisis topológico de datos, como también, proponer un modelo que ayude a la toma de decisiones de inversión.

1.4. Objetivo

1.4.1. Objetivo general

Proponer un modelo de aprendizaje maquina para predecir el precio del bitcoin usando métricas financieras y de la blockchain, averiguar si estas influyen en su volatilidad y plantear una metodología para la clasificación de estos precios para toma de decisiones de inversión a mediano y largo plazo.

1.4.2. Objetivos específicos

Los objetivos de la investigación son:

- Identificar si hay métricas que influyen en la variabilidad del precio del bitcoin y si es el caso determinar las de mayor peso usando análisis de regresión, análisis de componentes principales y agrupamiento jerárquico.
- Aplicar modelos estadísticos y de aprendizaje maquina sobre datos transformados para predecir el precio del bitcoin.
- Seleccionar el mejor modelo con base al índice de error RMSE y el índice de eficiencia Theil's U.
- Describir el panorama persistente de los datos usando análisis topológico de datos.
- Caracterizar diferentes herramientas de inteligencia artificial que permitan mejorar la toma de decisiones en un mercado financiero.
- Clasificar la decisión de inversión a mediano y largo plazo en comprar, vender y riesgo.

1.5. Contribuciones

Se espera que la contribución de esta investigación sea un modelo que minimice el error mínimo cuadrático en las predicciones y que estas sean fiables conforme al índice Theil's U, esto se reflejará en que los pronósticos realizados no diferirán mucho con respecto al precio original. Por otro lado se pretende aportar una metodología que ayude a clasificar el precio predicho para toma de decisiones de inversión a mediano y largo plazo a diferencia de los modelos para trading actuales.

1.6. Alcances y limitaciones

La investigación del estado de arte se limita a las publicaciones de herramientas y algoritmos que se han publicado recientemente. La comparación sera exclusiva entre los métodos de predicción y clasificación seleccionados en la investigación (Naive, SES, Holt, ETS, ARIMA, RTS, SVM, RF, LSTM y CNN). Se usan datos experimentales (series de tiempo del precio del bitcoin) para estimar la eficiencia y robustez de los métodos, los resultados tienen validez en el mismo contexto de los datos experimentales y las limitaciones de los métodos.

1.7. Contenido general

A continuación se mostrará de manera breve y concreta lo que el lector encontrará en cada capitulo.

1. **Introducción**: Presentación breve sobre orígenes del bitcoin, objetivos, justificación y contribuciones de la investigación.
2. **Estado del arte**: Aquí se encuentra la revisión sistemática de la literatura en la cual se presentan las preguntas de investigación, palabras clave y selección de estudios primarios.
3. **Metodología propuesta**: Se explica tomando como punto de partida un diagrama de flujo que muestra todo el proceso de la tesis y luego se explica a detalle el uso de cada método y herramienta usada.
4. **Resultados en Bitcoin**: Este capitulo describe los resultados obtenidos siguiendo la metodología propuesta.
5. **Conclusiones y trabajo futuro**: Conclusiones generales.

Capítulo 2

Estado del arte

Los modelos de inteligencia artificial ya sean de aprendizaje maquina o aprendizaje profundo han mostrado ser los mejores para predecir el precio volátil de algunas criptomonedas, entre ellas el bitcoin. Esto se debe a que los datos del precio no cumplen con los supuestos de estacionalidad, tendencia y ruido de algunos modelos estadísticos requieren.
Otros estudios muestran la importancia de agregar métricas tecnológicas como la de la blockchain como variables explicativas para mejorar la predicción utilizando métodos basados en correlación o análisis de sensibilidad.
Se encuentra también que los modelos de clasificación más comunes para inversión en criptomonedas están basados en el trading, donde la clasificación toma en cuenta las alzas o bajas del precio con respecto al actual en intervalos pequeños de tiempo, y con una exactitud un poco mayor al del tiro de una moneda.

2.1. Metodología

Para el seguimiento del estado del arte se utilizó la revisión sistemática de la literatura (SLR) expuesta en el Apéndice B donde se detalla cada paso de la misma. Esta metodología consiste en interpretar y sintetizar de forma adecuada la información obtenida de un tema de investigación. Se deben cumplir tres fases principales en una SLR: (1) Planificación de revisión, (2) Realización de la revisión, (3) Reporte de resultados. Los resultados para cada apartado de la investigación realizada se muestran a continuación utilizando los estudios principales obtenidos en la SLR de forma cronológica.

2.2. Predicción del precio

En un artículo publicado en marzo del 2018 McNally et al. [8] compararon modelos que utilizan redes neuronales recurrentes (RNN), modelos long-short term memory (LSTM) y ARIMA para realizar una clasificación binaria que determine las subidas y bajadas del precio con respecto al valor anterior. Para la clasificación se evaluó el rendimiento de la regresión utilizada para predicción con el indicador de error RMSE. En este trabajo se mostró que el modelo LSTM genera una mejor exactitud para la clasificación, sin embargo, basados en el RMSE el mejor modelo fue RNN con datos que van desde el 19 de agosto de 2013 a 19 de julio de 2016.

En julio de 2018 Phaladisailoed et al. [6] publicaron un articulo utilizando datos de Bitstamp que abarcan desde el 1 de enero de 2012 hasta 8 de enero de 2018 en intervalos de un minuto para realizar trading con Bitcoin. Se utilizaron todas características financieras del formato OHLC y se alcanzó un mejor rendimiento con el modelo GRU (Gated Recurrent Unit) entre LSMT, Theil-Sen Regression y Huber Regression.

Por otro lado en el artículo publicado en marzo de 2019 Tandon et al. [4], utilizando datos de CoinMarket sobre modelos como RNN, LSTM, regresión lineal y random forest (RF) con 10-fold cross validation, se mostró que RNN con LSTM supera a random forest y regresión lineal usando como predictores el precio de cierre y volumen del mismo con datos desde 2013 a 2019.

En un estudio comparativo de octubre de 2019 realizado por Felizardo et al. [5] nos muestran los resultados de modelos como ARIMA, RF, SVM, LSTM y Wavenets concluyendo que el modelo SVM es el mejor a un día de predicción y a una ventana de cinco a diez días

SVM y ARIMA son superiores a los demás modelos con datos de 2012 hasta inicios del 2017.

En julio de 2020 Mudassir et al. [3] obtienen resultados utilizando características de la blockchain, para ello se redujo la dimensionalidad de los datos filtrando las características con análisis de componentes principales y random forest, del análisis se seleccionaron aquellas características con mayor importancia, resultando que la arquitectura LSTM fue la mejor entre los modelos ANN, SVM y SANN.

De los estudios disponibles se concluye que no hay una metodología establecida para la predicción del precio del bitcoin, ya sea por las fechas de los datos de entrenamiento, los indices de comparación o los horizontes de predicción, esto puede deberse a que el Bitcoin es una tecnología en desarrollo y que los distintos modelos requieren de supuestos estadísticos distintos. Por otra parte, de los modelos utilizados más populares, el LSTM es que el que más destaca ya que fue el mejor en dos de cinco artículos primarios.

2.3. Análisis de métricas

En abril de 2018 tenemos un articulo de Saad et al. [9] sobre la caracterización de la blockchain donde se trata de identificar las métricas que explican los altos precios alcanzados. Usando la correlación de Pearson entre criptomonedas y características de la blockchain se encontró que el número de carteras, hashrate y UTXO determinan el número de nuevos usuarios, mineros y el balance que pueden gastar todos estos.

En septiembre de 2019 Ji et al. [10] encontraron que usando la correlación de Spearman con veintinueve características de la blockchain dieciocho de ellas fueron seleccionadas incluyendo el precio. De la selección se excluyeron aquellas características que tuviesen un coeficiente de Spearman menor a 0.75 y mayores a 0.95, esto debido a que las características con un coeficiente cercano a uno eran propiedades que dependían explícitamente del precio como la capitalización del mercado. De las características seleccionadas las que obtuvieron un mayor coeficiente fueron difficulty, est-trans-vol-usd, hash-rate, my-wallets, trade-vol y trans-fees-usd.

En julio de 2020 Mudassir et al. [3] obtienen características de la blockchain filtrando estas con análisis de componentes principales y random forest, seleccionando aquellas con mayor importancia. En el estudio no se muestran las características obtenidas por el PCA pero este logra capturar un 95 % de la varianza de los datos, sin embargo, el rendimiento en los modelos de regresión con estas características fue menor de lo esperado. Se utilizó random forest

como método principal de selección de características donde la importancia de las mismas dependía de la ventana de tiempo utilizada para el análisis. Las características más importantes que se mantuvieron a lo largo de las distintas ventanas de tiempo fueron las siguientes: median_transaction_fee30trxUSD, size90trx, price3wmaUSD, difficulty30rsi, mining_profitability y transactionvalueUSD.

En enero de 2021 Chen et al. [11] partiendo de predictores económicos y tecnológicos, y seleccionando aquellas con un factor de importancia utilizando random forest y ANNs se encontró que entre los modelos propuestos (ARIMA, SVR, ANFIS y LSTM) LSTM fue el mejor usando como datos de entrada los predictores obtenidos con el análisis de factor de importancia. Se encontró también que los predictores tecnológicos de alto impacto cambian constantemente con respecto al tiempo. En adición se muestra que la tarifa promedio por transacción es un fuerte predictor con alta importancia. Se concluye que las características que se mantienen en el tiempo (del 1 de enero de 2011 a 31 de julio de 2018) con alto impacto son la capitalización del mercado y la tarifa promedio por transacción.

Con esto, el estudio concluyó que la información obtenida de determinantes económicos y tecnológicos son más importantes que únicamente la tasa de cambio de la moneda (precio en USD) aun en periodos largos de tiempo como los usados en el estudio (1 de julio de 2015 a 31 de julio de 2018).

El panorama del estado del arte en lo que respecta al análisis de métricas de la blockchain para mejorar la predicción del precio parece no muy concluyente. En la mayoría de los estudios primarios se realiza un análisis de correlación, ya sea usando correlación de Pearson o de Spearman, estas nos indican el grado de relación lineal que hay entre las variables. Chen y Mudassir et al. [11, 3] utilizan el método de análisis de sensibilidad para medir la importancia de los factores económicos y tecnológicos usando random forest y artificial neural networks. Los resultados generales de los estudios no nos muestran características en común concluyentes. Eliminando las variables como la capitalización del mercado que son dependientes directamente del precio tenemos que la tarifa promedio por transacción, el hashrate, y el número de carteras son algunas características en común en al menos dos de los estudios.

2.4. Clasificación del precio para inversión

En febrero de 2020 Chen et al. [12] usando clasificación binaria, características de la blockchain y financieras clasificaron los precios del bitcoin con modelos estadísticos como

regresión lineal y análisis lineal de discriminante en intervalos de 5 minutos alcanzando una exactitud del 53 %. Por otro lado métodos de machine learning como random forest, XG-Bosst, discriminante cuadrático, SVM, y LSTM alcanzaron en promedio una exactitud del 62.2 %. La clasificación considero las subidas y bajadas del precio con respecto al actual y las características de la blockchain para predicción fueron seleccionadas usando Boruta.

En mayo de 2020 Pintelas et al. [13] mostraron que la combinación de algoritmos de deep learning (LSTM, Bi-directional LSTM, Convolutional Layers) con estrategias de ensamble (everaging, baggin, stacking) se pueden complementar mostrando buena auto-correlación en los residuos. En Bitcoin se alcanzó una exactiud de 54.62 % con el modelo CNN-BiLSTM y ensamble Bagging. La clasificación tomó en cuenta la dirección del precio con respecto al actual.

En julio de 2020 Mudassir et al. [3] obtuvieron resultados utilizando características de la blockchain, filtrando estas con análisis de componentes principales o random forest y seleccionando aquellas con mayor importancia. En la clasificación se alcanzó un promedio por arriba de un 65 % de exactitud para el día siguiente. Para 7 y 9 días se obtuvo de un 62 % a 64 % de exatitud con los modelos propuestos ANN,SVM,SANN y LSTM.

En Septiembre de 2020 Jaquart et al. [14] realizaron clasificación con datos de movimiento a corto plazo (1-60 minutos). El conjunto de características usadas incluyen características técnicas (retorno del bitcoin), de la blockchain (número de bitcoins, transacciones), de sentimiento (Sentimiento de Twitter y Sentimiento de Twitter ponderado por la fuerza de la emoción), y características base del activo (S&P500, VIX returns, Gold returns, MSCI World returns). Se usaron los modelos GRU, LSTM, FNN (Feedforward neural networks), regresión logística, gradiend boosting clasifier y random forest. Se concluyó que aparentemente incrementar el intervalo de tiempo trading aumentaba la exactitud del mismo. Se obtuvo el mejor resultado con LSTM a 60 minutos de predicción (56 %) con datos de 4 de marzo de 2019 a 10 diciembre de 2019.

En enero de 2021 Ibrahim et al. [15] realizaron clasificación del precio haciendo predicción hacia arriba o hacia abajo en intervalos de tiempo de 5 minutos. Se comparan modelos como ARIMA, Prophet (Facebook), RF, RF lagged-Auto-regression y multi-layer perceptron. El estudio concluyó que MLP alcanzó la mejor exacitud con 54 %.

En febrero de 2021 Akyildirim et al. [16] realizaron clasificación de precios del bitcoin para trading de un día, 15, 30, 60 min. utilizando el criterio -1 si OPEN<CLOSE y 1 si OPEN>CLOSE refiriéndose a el precio de apertura y cierre en un periodo de tiempo abarca

desde 10 agosto de 2017 a 23 junio de 2018. Se alcanzó una exactitud de 55-65 % en promedio resultando ser SVM el mejor modelo entre regresión logística, ANN y random forest como modelos candidatos.

En general podemos observar que los métodos de clasificación se basan en el trading y la dirección del precio, esto es, se utilizan datos de movimiento a corto plazo para realizar una predicción de compra o venta. Los resultados obtenidos en el estado del arte tienen una exactitud un poco mayor al del tiro de una moneda, sin embargo la metodología propuesta por Mudassir [3] alcanza una exactitud máxima de 64 % con ventana de tiempo de 7 a 9 días, haciendo notar que tal vez el trading no sea la mejor estrategia de inversión, sino las inversiones a mediano y largo plazo.

Capítulo 3

Metodología Propuesta

La metodología se basa en la recolección de datos, el pre-procesamiento de los mismos, en fases de entrenamiento para los distintos modelos usados y su respectiva validación con sus indicadores de error y eficiencia.
Para mejorar la calidad de la toma de decisiones de inversión se agregó un análisis de métricas antes de la fase de entrenamiento para saber cuales son las características de la blockchain que más variabilidad agregan al precio. Por ultimo se hizo una inspección a los datos usando análisis topológico para poder predecir posibles tiempos de crisis.
La salida de esta metodología es un posible precio en un intervalo de tiempo futuro, una recomendación de inversión y una posible advertencia de una crash en la serie de tiempo financiera.

3.1. Metodología

La metodología propuesta se divide en cuatro apartados: La predicción del precio utilizando series de tiempo del bitcoin en dólares, el análisis de las métricas de la blockchain, la clasificación de las alzas y bajas del precio y el análisis topológico de datos.

En la predicción del precio se darán a conocer rendimientos futuros utilizando un conjunto de datos de prueba sobre el cual se realizará la predicción usando modelos estadísticos y de machine learning. Para el análisis de métricas de la blockchain se encontrarán las variables que más influyen en el precio del bitcoin utilizando análisis de regresión, análisis de componentes principales y clustering jerárquico, por otra parte la clasificación para toma de decisiones de inversión a mediano y largo plazo convertirá la serie de tiempo en imágenes y las clasificará en comprar, vender o incertidumbre utilizando redes neuronales convolucionales, por último, el análisis topológico de datos analizará posibles crashes en el precio utilizando la norma topológica C1 y k-means.

La metodología general del trabajo se muestra en la Figura 3.1. Esta se divide en el pre-procesamiento de los datos, la fase de entrenamiento, análisis de métricas, análisis topológico de datos y la fase de validación.

3.1.1. Datos y pre-procesamiento

Los datos del precio del bitcoin están disponibles en linea de forma gratuita. Los datos para este estudio fueron recolectados de tres fuentes distintas, Yahoo! Finance [17], Kraken [18] e Investing [19], para la descarga de estos datos se usó la librería *quandl* en Python para descargarlos de forma automática en formato OHLC (Open-High-Low-Close). El intervalo de los datos es considerado desde el 2 de febrero del 2012 hasta el 11 de octubre de 2021.

Las tres fuentes de datos mostraban diferencias al compararlas, no todas estaban en formato OHLC y habían gaps entre los precios. Para solucionar lo anterior se creó la función *merge_dfs_on_column* en Python, que tiene por entrada una lista de dataframes con columnas en común, una lista con los nombres de exchanges usados asociados a los dataframes y la columna en común sobre la cual se trabajará. Esta función devuelve un dataframe que combina las columnas en común especificadas. Las columnas de dataframe obtenido anteriormente son sumadas y dividas entre el número de columnas, creándose así una única columna con la media.

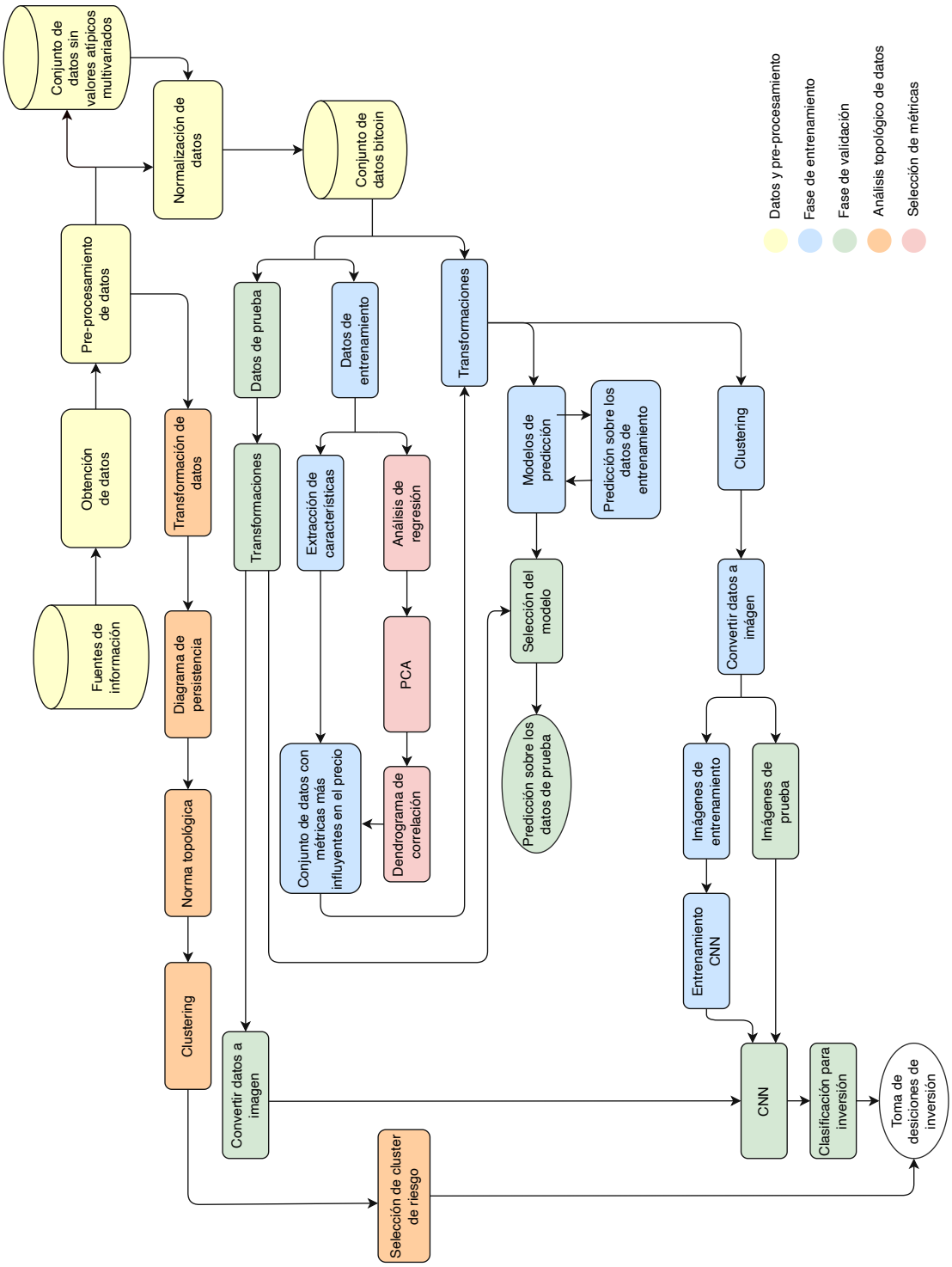

Figura 3.1: Metodología general propuesta

Al final se crea un conjunto de datos usando como columnas las medias obtenidas utilizando la función antes especificada. Ya que algunos exchanges no registran los precios con precisión o no trabajan en días festivos o fines de semana, entonces se eliminan estas fechas al dataframe creado con la media de los precios y se inputan los datos faltantes utilizando interpolación splines cúbico en ambas direcciones.
Los datos para realizar el análisis de métricas fueron recolectados utilizando la API de la comunidad de Coin Metrics [20]. Se obtuvo un archivo CSV con un total de 140 métricas como se observa en la Tabla A.1. Se consideró un intervalo de los datos desde el 2 de febrero del 2012 hasta el 11 de agosto de 2021 donde se utilizó interpolación de splines de orden tres para igualmente inputar datos faltantes.

Se eliminaron los valores atípicos multivariados del conjunto de datos de métricas utilizando la distancia de Mahalanovis definida como sigue:

$$D_M(x) = \sqrt{(x-\mu)\mathbf{S}^{-1}(x-\mu)}$$

donde $x = (x_1, x_2, x_3, ..., x_N)^T$ es una observación, $\mu = (\mu_1, \mu_2, \mu_3, ..., \mu_N)^T$ es la media y $\mathbf{S}$ es la matriz de covarianza. Se considera un valor atípico si la distancia del dato (la métrica en este caso) encontrada es mayor que dos veces la media de las distancias.

Se escalaron los datos de las métricas como del precio al intervalo [0,1] con la formula $(x - min(x))/(max(x) - min(x))$ del método MinMax para un mejor funcionamiento de los algoritmos de machine learning.

3.1.2. Fase de entrenamiento

Las bases de datos creadas se dividieron con un 80 % para los datos de entrenamiento y 20 % para los datos de prueba. Para los datos de predicción el conjunto de entrenamiento contiene 2782 elementos. Para el conjunto de entrenamiento de métricas de la blockchain se utilizó la función *train_test_split* de la librería sklearn en Python creando de forma aleatoria 3604 datos con semilla 123. Por último, después de las transformaciones correspondientes el conjunto de entrenamiento para clasificación y toma de decisiones de inversión resulto en 616 imágenes.

Predicción del precio: Para la extracción de características de los modelos de machine learning se usó el método de correlación [4], donde se realizó la estimación del coeficiente de

correlación de Pearson para examinar la fuerza y la dirección de la relación lineal entre las variables de la base de datos, en este caso, apertura, precio máximo, mínimo y cierre. Se determinó que todas las características están fuertemente relacionadas entre sí y fueron utilizadas como variables explicativas en los modelos no univariados (SVM, RF, RTS). A las características extraídas se les agregó las métricas de la blockchain que más influyeron en la variabilidad del precio como se detalla en la Subsección 3.1.3. Para las redes LSTM se utilizó únicamente el precio de cierre con retraso de un día para su ajuste.

Antes de aplicar los modelos a los datos de entrenamiento se realizaron distintas transformaciones como BoxCox [3.1], log [3.2], $diff$ [3.3] y $diff(\log)$ [3.4]. Para la transformación BoxCox el parámetro λ se estimó con la función BoxCox.lambda() de R para encontrar el parámetro que mejor estabiliza la varianza, en el caso del la transformación log se uso la base natural e.

Se ajustaron los precios del bitcoin usando diferentes modelos estadísticos y de machine learning basados en el método Naive [3.5], SES [3.6], Holt [3.7], ETS [1], ARIMA [3.8], RTS [3.9], SVM [3.10], RF [3.12], y LSTM [3.11].

Para el método Naive, SES, Holt, ETS y ARIMA se usaron las funciones estándar de R del paquete *fpp2*, únicamente cambiando la entrada de los datos de entrenamiento por la respectiva transformación, por otro lado, para el método RTS, SVM y RF se realizaron las inferencias sobre las características dichas anteriormente, manteniendo los hiperparámetros por defecto. La configuración usada para el modelos LSTM es la mostrada en la Tabla 3.1.

Hiperparámetro	**Configuración**
Optimizer	Adam
Hidden layers	2 (2,1)
Learning rate	0.02
Epochs	100
Batch size	1
Activation	tanh
Loss function	MSE

Tabla 3.1: Configuración de los hiperparámetros LSTM

[1]Encuentra el modelo que minimiza mejor el AIC (Akaike's Information Criterion) o BIC (Schwarz's Bayesian Information Criterion) usando SES de primer, segundo o tercer orden.

[2]Generalización del modelo SES que toma en cuenta el error, la tendencia y la estacionalidad.

Transformación	Fórmula	
BoxCox	$y_t(\lambda) = \begin{cases} \frac{y_t^\lambda - 1}{\lambda} & si \quad \lambda \neq 0 \\ \log y_y & si \quad \lambda = 0 \end{cases}$	(3.1)
log	$y_t = \log y_t$	(3.2)
diff	$\nabla y_t = y_t - y_{t-1}$	(3.3)
diff (log)	$\nabla y_t = \log y_t - \log y_{t-1}$	(3.4)

Tabla 3.2: Transformaciones utilizadas en el modelo de entrenamiento

Clasificación para toma de decisiones de inversión: Después del pre-procesamiento de los datos se aplica una transformación logarítmica al precio de cierre en dólares para estabilizar la varianza. El objetivo de esta transformación es tratar de identificar una tendencia monótona. En este estudio nos referiremos a este ajuste como la característica estadística tradicional (TSF)[21].

La entropía de Shannon es fundamental en la teoría de la información y también es conocida por medir la incertidumbre. La ecuación original es como sigue

$$H(x) = -\sum_{i=1}^{n} p(x_i) \log p(x_i)$$

donde x es una variable aleatoria discreta con posibles salidas $x_1, x_2, x_3, ..., x_n$. Entonces como en [21] utilizaremos una modificación de esta ecuación que tiene la siguiente forma

$$H(x) = \frac{1}{n} \sum_{i=1}^{n} -TFS(x_i) \log_2 TSF(x_i)$$

aquí n es el tamaño de la ventana deslizante. Esta última ecuación nos permite destacar las características obtenidas de la TSF. De este modo podemos elegir el TSF junto con la entropía de Shannon para observar un aumento o disminución bien definidos de los datos a lo largo del tiempo.

Modelo	Notación		Parámetros
Naive	$\hat{y}_{t+h\|t} = y_t$	(3.5)	
SES	$\hat{y}_{t+h\|t} = l_t$ $l_t = \alpha y_t + (1-\alpha) l_{t-1}$	(3.6)	$0 \leq \alpha \leq 1$
Holt	$\hat{y}_{t+h\|t} = l_t + hb_t$ $l_t = \alpha y_t + (1-\alpha)(l_{t-1} + b_{t-1})$ $b_t = \beta(l_t - l_{t-1}) + (1-\beta) b_{t-1}$	(3.7)	$0 \leq \alpha \leq 1$ $0 \leq \beta \leq 1$
ETS[2]			
ARIMA	$\nabla y_t = c + \phi \nabla_1 y_{t-1} + ... + \phi \nabla_p y_{t-p}$ $+\theta_1 \epsilon_{t-1} + ... + \theta_q \epsilon_{t-q} + \epsilon_t$	(3.8)	p: orden de parte autorregresiva d: grado de las diferencias q: orden de promedios móviles
RTS	$y_t = \beta_0 + \beta_1 x_{1,t} + ... + \beta_k x_{k,t} + \epsilon_t$	(3.9)	
SVM	$min \frac{1}{2}\|w\|^2 + \nu \sum_{i=1}^{m} \xi_i$ $s.to\ y_i(wx_i - b) \geq 1 - \xi_i,\ \xi \geq 0$	(3.10)	$w \in R^n$ $b \in R$ ν parámetro de regularización ξ_i variables de holgura
RF	$g_c(x) = \frac{1}{t} \sum_{j=1}^{t} \hat{P}(c \mid v_j(x))$ $P(c \mid v_j(x)) = \frac{P(c, v_j(x))}{\sum_{l=1}^{n} P(c_l, v_j(x))}$	(3.11)	t: número de arboles creados en un subespacio aleatorio c: clase 1,2,...,n $v_j(x)$: nodo terminal del punto x en el árbol T_j $(j = 1, 2, ..., t)$
LSTM	$x = \begin{bmatrix} x_t \\ h_{t-1} \end{bmatrix}$ $f_t = \delta(W_f X + b_f)$ $i_t = \delta(W_i X + b_i)$ $o_t = \delta(W_o X + b_o)$ $\tilde{C}_t = tanh(W_c[h_{t-1}, x_t] + b_c)$ $C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t$ $h_t = o_t \cdot tanh(C_t)$	(3.12)	x_t: input al tiempo t h_t: hiden state al tiempo t W_f, W_i, W_o, W_c: matrices de peso b_f, b_i, b_o, b_c: parámetros de entrenamiento δ: función de activación

Tabla 3.3: Descripción de modelos usados en la comparación

El siguiente paso es agrupar los datos después de las transformaciones correspondientes para así tener las etiquetas que clasifiquen los precios del bitcoin. Se utilizó el algoritmo k-means para agrupar los datos con características similares. Este algoritmo consiste en minimizar la suma de las distancias euclidianas de cada uno de los puntos con respecto al centroide del cluster. Se utilizó con $k = 10$ para una mayor precisión a la hora de escoger los grupos para la toma de decisiones de inversión.

El proceso para convertir los datos a imágenes con datos originales R y N puntos de muestreo se observan en la Figura 3.2. Si las imágenes a generar son de una resolución de $M \times M$, entonces el primer paso es tomar una sub-muestra L de tamaño M^2 de los datos originales R, esto es, la i-ésima muestra esta dada por $L_i = \{R(i \cdot s + 1), R(i \cdot s + 2), R(i \cdot s + 3), ..., R(i \cdot s + M^2)\}$ donde $s \in \mathbb{Z}^+$ es el paso entre muestras y el índice $i = \{1, 2, 3, ..., \lfloor \frac{N}{s} \rfloor\}$, con $\lfloor \cdot \rfloor$ la función piso. Cada punto de la sub-muestra L llena una matriz de $M \times M$ de izquierda a derecha y de arriba a abajo. Cada punto es normalizado de 0 a 255 y representa el valor del píxel en una escala de grises. Para la i-ésima imagen $P(x, y)$ el proceso se define como sigue

$$P_i(x, y) = round\left(\frac{L_i((x-1) \cdot M + y) - min(L_i)}{max(L_i) - min(L_i)} \cdot 255\right)$$

Para el propósito de este estudio se ha escogido una resolución de 32×32 píxeles con un paso $s = 32$.

La configuración de la arquitectura utilizada esta basada en AlexNet y se muestra en la Tabla 3.4. Esta mostró ser la optima en el estudio del desgaste de baleros con una precisión por arriba del 98,64 % [21].

3.1.3. Selección de métricas

Para la selección de métricas que más influyen en el precio del bitcoin primero se seleccionaron aquellas que tuvieran significación estadística usando mínimos cuadrados ordinarios. De las variables seleccionadas se escogieron aquellas que tuvieran la misma dirección y el mismo o mayor peso que la variable precio usando análisis de componentes principales. Al final se corroboró la información usando un dendrograma de correlación para ver la relación entre las variables.

Capa	Característica
1	Conv($5 \times 5 \times 96$)
2	Maxpool(2×2)
3	Conv($3 \times 3 \times 256$)
4	Maxpool(2×2)
5	Conv($3 \times 3 \times 384$)
6	Maxpool(2×2)
7	Conv($3 \times 3 \times 384$)
8	Conv($3 \times 3 \times 256$)
9	Maxpool(2×2)
10	FC (n)
11	FC (n)
12	FC (x)

Tabla 3.4: Configuración de la arquitectura propuesta.

Análisis de regresión: Las variables endógenas que se utilizaron son las mostradas en el Tabla A.1, la variable exógena es el precio en dólares obtenido de Coin Metrics. A los datos de entrenamiento de las métricas se ajustó un modelo de mínimos cuadrados ordinarios de la librería statsmodels en Python y se seleccionaron las variables con un valor p menor que 0,05 con base en el estadístico t.

Análisis de componentes principales: De las métricas obtenidas utilizando análisis de regresión se seleccionaron aquellas que tuvieran la misma dirección y, mayor o igual magnitud que la variable precio, para ello se utilizó la función PCA de la librería *sklearn.decomposition* en Python. Se utilizó como parámetro la cantidad total de métricas obtenidas anteriormente y se utilizó el siguiente criterio para determinar las métricas con mayor influencia en el precio

$$F' = \{y : \|y\|_2 \geq \|x\|_2, \, sign(y) = sign(x), \, y \in F\}$$

donde F es el conjunto de características obtenidas en ek análisis de regresión, x es el vector precio, $\|\cdot\|_2$ es la norma euclidiana usual y $sign$ es la función signo.

Dendrograma de correlación: Se utilizó la función heatmap de la librería biokit en Python utilizando todas la características obtenidas en el paso anterior empleando como distancia la correlación entre las métricas y se seleccionaron aquellas que estuvieran en el mismo cluster que el precio o cercanas al mismo.

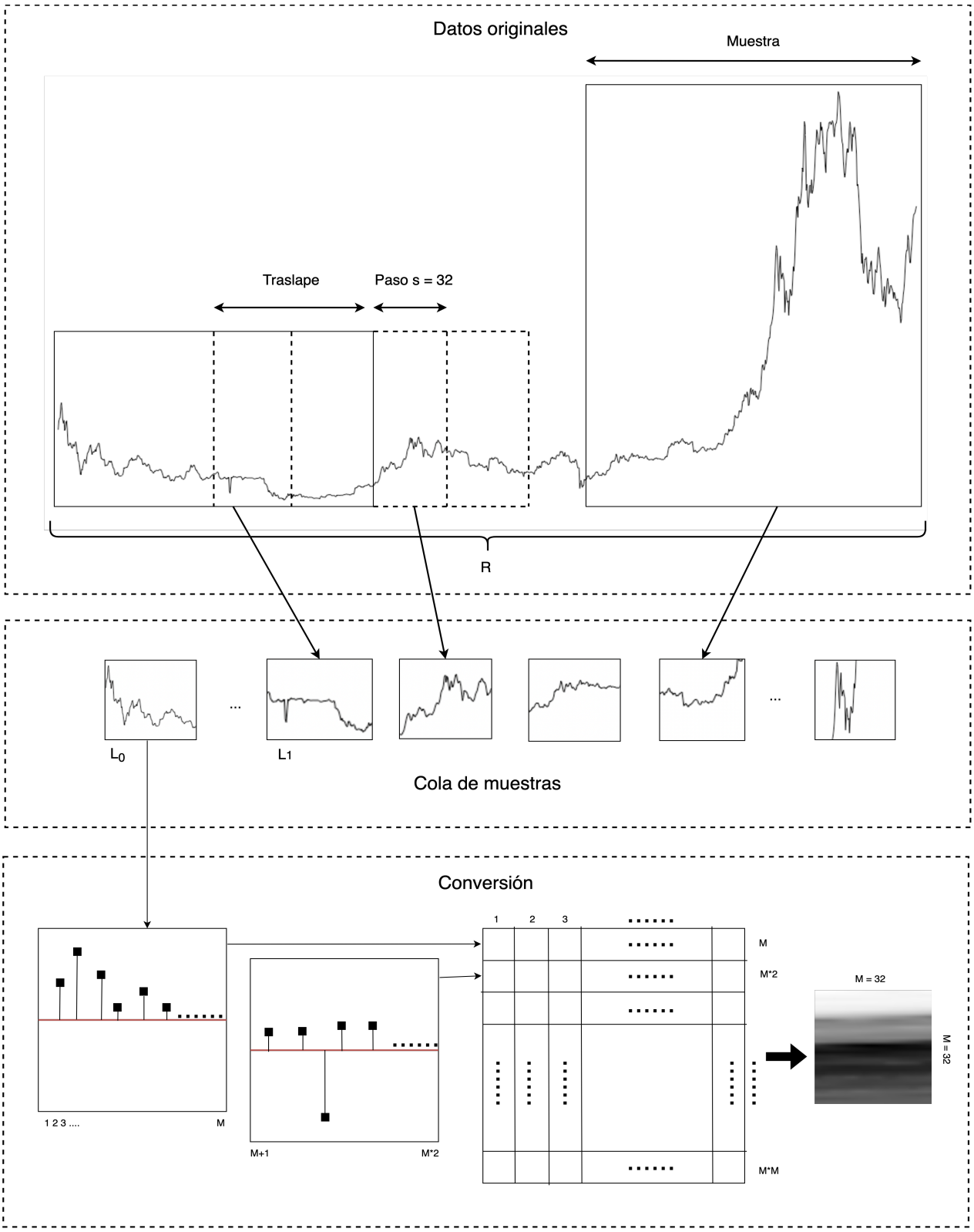

Figura 3.2: Método para convertir datos a imágenes.

3.1.4. Análisis topológico de datos

Se siguieron los siguientes pasos para determinar los posibles crashes en el precio:

Incrustar los datos en un espacio de dimensión superior: La serie de tiempo del precio del bitcoin es incrustada en un espacio de dimensión 4 usando el teorema de Takens.

Creación del panorama de persistencia: Con los datos en un espacio de dimensión superior se calcula la homología persistente de los datos para crear el respectivo panorama de persistencia. Se utiliza la librería TDA de R con la función $ripsDiag()$.

Cálculo de norma topológica: El panorama de persistencia creado en R al ser un elemento vectorial se calcula la norma L1 para después crear la norma C1 respectiva.

Creación de clusters: Se aplica el algoritmo k-means a un dataframe que tiene por columnas el logaritmo del precio del bitcoin y la norma C1 para visualizar los posibles clusters de los datos.

Selección de cluster de riesgo: En base a los siguientes criterios se selecciona el cluster que nos advierte de una posible burbuja.

1. La mayoría de los elementos del cluster cumplen que $\|x_t\|_{C1} > 0{,}5$
2. Las fechas en el cluster deben ser consecutivas.

3.1.5. Fase de validación

Predicción del precio: En cada modelo se realizo una predicción sobre el 80 % de los datos, en estas predicciones se consideró el indicador de error RMSE (Root Mean Square Error) y la segunda versión del indice de eficiencia Theil's U [22]. Hubo un total de 49 modelos tomando en cuenta cada transformación propuesta, de estos, el modelo seleccionado en este estudio es el que tiene menor RMSE e índice Theil's U más cercano a cero. Con el modelo seleccionado se hace la predicción sobre el conjunto de validación como se muestra en la Sección 5.1.

Indicador	Fórmula
RMSE	$\sqrt{\frac{\sum_{i=1}^{n}(\hat{y_n}-y_n)^2}{n}}$
Theil's U	$\sqrt{\frac{\sum_{i=1}^{n}(\hat{y_n}-y_n)^2}{\sum_{i=1}^{n} y_n^2}}$

Tabla 3.5: Fórmulas de los índices RMSE y Theil's U

Clasificación del precio: Se crea un dataframe con las imágenes generadas. Una columna contiene los archivos de imagen y la otra las etiquetas de su cluster respectivo. Los clusters son renombrados a comprar, vender o incertidumbre.

Se crea el conjunto de entrenamiento del dataframe creado, se configura la CNN como en la Tabla 3.4 y se entrena con los datos anteriores.

La precisión del modelo es calculada utilizando el método accuracy de tensorflow la cual crea dos variables locales, *total* y *count*, que se utilizan para calcular la frecuencia con la que las predicciones coinciden con las etiquetas. Esta frecuencia se devuelve finalmente como una operación que simplemente divide el total entre el recuento.

Capítulo 4

Fundamento teórico

Hola

4.1. Fundamento teórico

Los datos obtenidos de observaciones del precio del bitcoin son recolectados de forma secuencial y de forma cronológica a lo largo del tiempo. Como ejemplo de lo anterior se puede observar diariamente el precio de apertura y cierre de este activo modelando así lo que comúnmente conocemos como serie de tiempo. En este apartado se hará una revisión breve de los modelos de predicción más utilizados en la literatura como también de análisis multivariante y análisis topológico de datos ya que son fundamentales para el estudio de series temporales, modelar los altibajos del precio y clasificar tendencias.

4.2. Medidas de error

4.2.1. RMSE

4.2.2. MSE

4.2.3. MAE

4.2.4. Theils's U

4.3. Modelos estadísticos

Hola

4.3.1. Naive

Para el método de predicción Naive, simplemente hacemos que el valor predicho sea igual a la última observación, esto es,

$$\hat{y}_{T+h|T} = y_T$$

Aquí, la notación $\hat{y}_{T+h|T}$ es una manera corta de estimar $y_{T+h|T}$ basado en el los datos $y_1, \ldots, y_T$. Este método funciona bastante bien para series de tiempo económicas y financieras [23]. Se asume que las observaciones más recientes son las más importantes y que las observaciones previas no proveen de información para el futuro.

4.3.2. SES

El método simple exponential smothing (SES) es apropiado para la predicción de datos sin una tendencia clara o sin un patrón estacional [23]. Su concepto principal es dar mayor peso a las observaciones más recientes que a las del pasado lejano. Las predicciones se realizan mediante medias ponderadas, donde los pesos disminuyen de forma exponencial a medida que las observaciones son más antiguas.

$$\hat{y}_{T+1|T} = \alpha y_T + \alpha(1-\alpha)y_{T-1} + \alpha(1-\alpha)^2 y_{T-2} + \cdots \tag{4.1}$$

donde $0 \leq \alpha \leq 1$ es el parámetro de suavizado. La forma de media ponderada es la siguiente:

$$\hat{y}_{T+1|T} = \sum_{j=0}^{T-1} \alpha(1-\alpha)^j y_{T-j} + (1-\alpha)^T \ell_0$$

dónde ℓ_0 es el primer valor ajustado al tiempo $t = 1$, el cual tenemos que estimar y se obtiene dependiendo del tipo de error a maximizar. El valor $(1-\alpha)^T \ell_0$ tiende a 0 si T es muy grande, así que esta forma es similar a la Ecuación 4.1.

Una representación alternativa es la forma en componentes que está dada por:

$$\begin{aligned} &\text{Ecuación de predicción} && \hat{y}_{t+h|t} = \ell_t \\ &\text{Ecuación de nivel} && \ell_t = \alpha y_t + (1-\alpha)\ell_{t-1}, \end{aligned}$$

donde ℓ_t es el nivel o valor de suavizado de la serie al tiempo t.

4.3.3. Holt

Holt extendió el método SES para permitir la predicción de datos con una tendencia. Este método involucra una ecuación de predicción y dos de suavizado (una para el nivel y otra para la tendencia).

$$\begin{aligned} &\text{Ecuación de predicción} && \hat{y}_{t+h|t} = \ell_t + hb_t \\ &\text{Ecuación de nivel} && \ell_t = \alpha y_t + (1-\alpha)\left(\ell_{t-1} + b_{t-1}\right) \\ &\text{Ecuación de tendencia} && b_t = \beta^*\left(\ell_t - \ell_{t-1}\right) + (1-\beta^*)\, b_{t-1}, \end{aligned}$$

donde ℓ_t denota un estimador de nivel de la serie al tiempo t, b_t denota un estimador de la tendencia (pendiente) de la serie al tiempo t, α es el parámetro de suavizado para el nivel,

$0 \leq \alpha \leq 1$ y β^* es el parámetro de suavizado para la tendencia., $0 \leq \beta^* \leq 1$.

Como en el método SES, ℓ_t es una media ponderada de las observaciones y_t, en este caso b_t, la ecuación de tendencia es también una media ponderada sobre $\ell_t - \ell_{t-1}$ y b_t el estimado previo de la tendencia.

El método de Holt muestra una tendencia constante indefinida hacia el futuro tendiendo a sobre pronosticar para horizontes grandes de predicción. Dada la observación anterior Gardner y McKenzie introdujeron un parámetro de amortización que hace plana la tendencia después de un tiempo en el futuro. Se ha probado que los métodos que incluyen una tendencia amortiguada han sido más exitosos y más populares cuando las predicciones son automatizadas [23].

El método de Holt amortizado tiene la siguiente forma

$$\begin{aligned}
\hat{y}_{t+h|t} &= \ell_t + \left(\phi + \phi^2 + \cdots + \phi^h\right) b_t \\
\ell_t &= \alpha y_t + (1-\alpha)\left(\ell_{t-1} + \phi b_{t-1}\right) \\
b_t &= \beta^*\left(\ell_t - \ell_{t-1}\right) + \left(1-\beta^*\right)\phi b_{t-1}.
\end{aligned}$$

donde α y β son como el método Holt y ϕ es el parámetro de amortización $0 \leq \phi \leq 1$. Notemos que la ecuación de predicción converge a $\ell_T + \phi b_T/(1-\phi)$ conforme $h \to \infty$. Esto significa que las predicciones a corto plazo tienen tendencia mientras que las predicciones a largo plazo son constantes.

4.3.4. ETS

explicar que se uso ETS con todos los parametros en automatico para ver la mejor predicción que mejora la verosimilitud.

4.3.5. ARIMA

Modelos autorregresivos

Un modelo autorregresivo es aquel en el que predecimos la variable de interés usando una combinación lineal de valores pasados de la variable. El termino autorregresión indica que es una regresión de la variable contra sí misma [23].

Por tanto, un modelo autorregresivo de orden p puede ser escrito como

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \varepsilon_t$$

donde ε_t es ruido blanco y y_{t-p} son valores retardados de y_t como predictores. Esto es comúnmente conocido como $\mathbf{AR}(p)$.

Modelo de medias móviles

Un modelo de medias móviles usa errores de predicción pasados en una regresión, a diferencia de un modelo autorregresivo que usa valores retardados del valor a predecir [23]. Esto se puede expresar como

$$y_t = c + \varepsilon_t + \theta_1\varepsilon_{t-1} + \theta_2\varepsilon_{t-2} + \cdots + \theta_q\varepsilon_{t-q}$$

donde ε_t es ruido blanco. Esto es comúnmente conocido como $\mathbf{MA}(q)$, un modelo de medias móviles de orden q.

Modelos ARIMA

ARIMA es el acrónimo para modelo autorregresivo integrado de promedio móvil (del inglés AutoRegressive Integrated Moving Average), que es un modelo que combina la diferenciación con autorregresión y medias móviles [23]. En este caso la parte integrada del modelo hace referencia al converso de la diferenciación.

El modelo puede ser escrito

$$y'_t = c + \phi_1 y'_{t-1} + \cdots + \phi_p y'_{t-p} + \theta_1\varepsilon_{t-1} + \cdots + \theta_q\varepsilon_{t-q} + \varepsilon_t,$$

donde y'_t es la serie diferenciada y los predictores de la parte izquierda de la ecuación son son valores retardados de y_t y ε_t. Llamamos a esto un modelo $\mathbf{ARIMA}(p, d, q)$ donde

- p = orden de la parte autorregresiva.
- d = grado de la primer diferenciación involucrada.
- q = orden de la parte de medias móviles.

El operador B (backward shift) es una notación usada para trabajar con series de tiempo retardadas

$$By_t = y_{t-1}.$$

Si aplicamos dos veces B a y_t entonces desplaza los datos hacia atrás dos períodos:

$$B\left(By_t\right) = B^2 y_t = y_{t-2}.$$

En general este operador es particularmente útil ya que puede ser tratado usando reglas algebraicas ordinarias. También tiene la propiedad de que una diferencia de d-ésimo orden puede ser escrita como

$$(1-B)^d y_t.$$

Usando el operador B podemos escribir la ecuación del modelo ARIMA de la siguiente forma

$$\begin{array}{cccc} (1-\phi_1 B - \cdots - \phi_p B^p) & (1-B)^d y_t & = c + (1+\theta_1 B + \cdots + \theta_q B^q)\,\varepsilon_t \\ \uparrow & \uparrow & \uparrow \\ \mathrm{AR}(p) & d \text{ diferencias} & \mathrm{MA}(q) \end{array}$$

4.4. Transformaciones matemáticas

Ajustar los datos en algunos casos puede mejorar la predicción de los mismos ya que permiten disminuir la variabilidad o simplificar los patrones o tendencias de los registros históricos. En este estudio se aplicaron trasformaciones matemáticas para generar series de tiempo que refuerzan la tendencia o hacerla estacionaria.

4.4.1. Transformación *log*

Si denotamos las observaciones originales como $y_1, \ldots, y_T$ y las observaciones transformadas como $w_1, \ldots, w_T$, entonces $w_t = \log\left(y_t\right)$. Los logaritmos son útiles dado que son interpretables. Por ejemplo, si *log* base 10 es usada entonces un incremento de 1 en la escala logarítmica significa una multiplicación de 10 en la escala original. Otra propiedad útil es que restringe la predicción a permanecer positiva en la escala original.

4.4.2. Transformación Box-Cox

La familia de transformaciones Box-Cox que dependen del parámetro λ se define como sigue

$$w_t = \begin{cases} \log(y_t) & \text{if } \lambda = 0 \\ \left(y_t^{\lambda} - 1\right)/\lambda & \text{otro caso} \end{cases}$$

Esta transformación abarca la transformación logarítmica y de potencias, si $\lambda = 0$ entonces el logaritmo natural es usado si no se usa una trasformación de potencias con un escalamiento.

La ventaja de esta transformación es que tomando un buen valor de λ podemos estabilizar la varianza en toda la serie de tiempo, haciendo el modelo de predicción más simple [23].

4.4.3. Transformación $diff$

La diferencia en series es el cambio entre observaciones consecutivas en la serie original y puede ser escrita de la siguiente manera

$$y'_t = y_t - y_{t-1}$$

Una de las ventajas de esta transformación es que puede ayudar a estabilizar la media de una serie de tiempo al eliminar los cambios de nivel de la misma y, por tanto, eliminar (o reducir) la tendencia y la estacionalidad [23].

4.4.4. Transformación $diff(log)$

La transformación $diff(log)$ calcula cambios relativos, esto es, la diferencia fraccional entre dos números que, multiplicado por cien es la diferencia porcentual entre los mismos. Estos cambios son simétricos y útiles para explorar las relaciones con datos continuos y de valor positivo [24].

$$y'_t = \log y_t - \log y_{t-1}$$

En finanzas esta transformación puede ser usada para analizar el cambio porcentual de la tendencia.

4.5. Modelos de machine learning

Hola

4.5.1. RTS

El modelo de regresión sobre series de tiempo (RTS) es una variación de los modelos de regresión lineal que permite agregar variables para la tendencia y temporada [25].

La forma general de un modelo de regresión lineal multiple es la siguiente:

$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t$$

donde y es la variable a predecir y $x_1, \ldots, x_k$ son las k variables predictoras. Los coeficientes $\beta_1, \ldots, \beta_k$ miden el efecto de cada predictor después de tener en cuenta los efectos de todos los demás predictores del modelo.

Las suposiciones implícitas que se hacen al aplicar un modelo de regresión lineal son las siguientes:

- La relación entre la variable de pronóstico y las variables predictoras satisfacen una ecuación lineal.
- Los errores tienen media cero.
- Los errores no están autocorrelacionados.
- Los errores no están correlacionados con las variables predictoras.

Otra suposición importante es que cada predictor no sea una variable aleatoria, tienen que ser datos reproducibles. En datos financieros o económicos no es posible controlar sus valores, por tanto suponemos que no son una variable aleatoria.

La variable tendencia del modelo RTS puede ser modelada usando $x_{1,t} = t$ como predictor.

$$y_t = \beta_0 + \beta_1 t + \varepsilon_t$$

donde $t = 1, ..., T$. Por otra parte la temporada se modela agregando variables ficticias para codificar categorías. Cada categoría corresponde a una ventana de tiempo. Estas variables ficticias miden el efecto de la categoría relativa a la categoría omitida [23].

4.5.2. SVM

Support vector machines (SVM) es un algoritmo de machine learning supervisado que se basa en separar los puntos de los datos usando hiperplanos tal que la distancia de separación

es máxima. Los vectores de soporte son los puntos más cercanos al hiperplano para calcular su posición [3]. Para el cálculo de los SVM la función objetivo 4.2 debe ser minimizada sujeto a la condición 4.3.

$$\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\zeta_i \tag{4.2}$$

En la Ecuación 4.2 la variable de holgura es ζ_i, la penalización es C y $\mathbf{w}$ es la normal al hiperplano.

$$y_i\left(\mathbf{w}\cdot\phi\left(x_i\right)+b\right) \geq 1-\zeta_i, \qquad \text{con } \zeta_i \geq 0. \tag{4.3}$$

En la restricción 4.3 x_i y y_i son puntos en los datos y $\phi(x_i)$ son los datos transformados.

Kernel? En Mudassir

Análogamente existe una versión de regresión del algoritmo SVM que consiste en resolver el siguiente problema

Minimizar $\frac{1}{2}\|w\|^2$

Sujeto a $|y_i - \langle w, x_i\rangle - b| \leq \varepsilon$

donde x_i es una muestra de entrenamiento con valor objetivo y_i. Aquí $\langle w, x_i\rangle + b$ es la predicción y ϵ es un parámetro libre que sirve como umbral. Todas las predicciones tienen que estar dentro de un rango ϵ de predicciones verdaderas.

4.5.3. Random Forest

El algoritmo de bosque aleatorio o random forest combina la salida de múltiples arboles de decisión tal que cada árbol depende de los valores de un vector aleatorio probado independientemente y con la misma distribución para cada uno de estos y así alcanzar un solo resultado. Es usado para problemas de regresión y clasificación [26].

Árboles de decisión

Un árbol de decisión divide una pregunta inicial en subpreguntas para llegar a una decisión final. Observaciones que se ajusten a un criterio seguirán una rama "SÍ", y aquellas que no,

seguirán un camino distinto.

Entonces un árbol de decisión busca encontrar la mejor partición de un subconjunto de los datos y estos generalmente son entrenados con el algoritmo Classification and Regression Tree (CART). Las métricas usadas para evaluar la calidad de las particiones son Gini impurity, information gain, o MSE.

Métodos de conjunto (Ensamble methods)

Los métodos de conjunto son técnicas que buscan mejorar la precisión de resultados en modelos combinando múltiples modelos en vez de usar uno solo. Los modelos de conjunto más populares son: bagging, boosting y staking. En el método bagging se selecciona una muestra aleatoria de datos en un conjunto de entrenamiento con remplazo, significando esto que los datos se pueden elegir más de una vez. Después de generar varias muestras de los datos estos se entrenan de forma independiente ya sea regresión o clasificación. El promedio o mayoría de de esas predicciones nos levan a una mejor precisión.

Algoritmo random forest

El algoritmo random forest es una exención del método baggin ya que utiliza tanto el baggig como la selección aleatoria de características para crear un bosque no correlacionado de arboles de decisión. La aleatoriedad de características, también conocida como bagging de características o "el método del subespacio aleatorio" es la diferencia clave entre los bosques aleatorios y los arboles de decisión ya que mientras los arboles de decisión consideran todas las posibles divisiones de características, los bosques aleatorios solo seleccionan un subconjunto de ellas.

4.5.4. LSTM

Long Short-term memory (LSTM) constituye un caso especial de una red neuronal convolucional, la cual fue propuesta para modelar dependencias a corto y largo plazo. Este modelo de deep learnig es especialmente útil para modelado y predicción de datos de series de tiempo [3]. Una unidad LSTM consiste en una memoria de celda que almacena información y es actualizada por tres puertas principales: la puerta de entrada, la puerta de olvido y la puerta de salida [12]. En cada paso t la puerta de entrada i_t determina que información es agregada a la celda de estado S_t (memoria), la puerta de olvido f_t determina que información es desechada

de la celda de estado mediante la decisión de una función de transformación en la capa de la puerta de olvido, mientras que la puerta de salida o_t determina qué información del estado de la celda se utilizará como salida [27]. En en bloque LSTM, C_{t-1} es la memoria o celda de estado del bloque anterior, h_{t-1} es la salida del bloque anterior, X_t es el vector de entrada, C_t es la memoria o celda de estado del bloque presente y h_t es la salida del bloque actual. Las puertas y celdas de estado LSTM están dadas por las ecuaciones 4.4 a 4.9.

$$f_t = \sigma_g \left(W_f x_t + U_f h_{t-1} + b_f \right) \tag{4.4}$$

donde f_t es el vector de activación de la puerta de olvido, W y U son matrices ponderadas, b es el vector de sesgo y σ_g es la función sigmoide.

$$i_t = \sigma_g \left(W_i x_t + U_i h_{t-1} + b_i \right) \tag{4.5}$$

donde i_t es el vector de activación de la puerta de entrada o activación.

$$o_t = \sigma_g \left(W_o x_t + U_o h_{t-1} + b_o \right) \tag{4.6}$$

donde o_t es el vector de activación de la puerta de salida.

$$\tilde{c}_t = \sigma_h \left(W_c x_t + U_c h_{t-1} + b_c \right) \tag{4.7}$$

donde el vector de activación de la celda de entrada está dada por c_t y σ_h es la función tangente hiperbólica.

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tilde{c}_t \tag{4.8}$$

donde c_t es el estado de la celda o vector memoria.

$$h_t = o_t \otimes \sigma_h \left(c_t \right) \tag{4.9}$$

donde h_t es el vector de salida del bloque LSTM o el vector de estado oculto.

4.5.5. CNN

VER LIEBERIS ET AL, EN ESTADO DEL ARTE CLASIFICACION Y NOTAS IOS

4.5.6. K-Means

El algoritmo K-means es un algoritmo de partición con la meta de asignar a cada punto de dato un único agrupamiento. Divide un conjunto de n muestras X dentro de k agrupamientos disjuntos c_i, $i = 1, ..., k$, cada uno descrito por la media μ_i de las muestras en el agrupamiento. Las medias son comúnmente llamados centroides del agrupamiento. El algoritmo K-means asume que todos los k grupos tienen la misma varianza [28].

El agrupamiento K-means resuelve el siguiente problema de minimización.

$$\arg\min_{c} \sum_{j=1}^{k} \sum_{x \in c_j} d\left(x, \mu_j\right) = \arg\min_{c} \sum_{j=1}^{k} \sum_{x \in c_j} \|x - \mu_j\|_2^2 \tag{4.10}$$

donde c_i es el conjunto de puntos que pertenecen al agrupamiento i y μ_i es el centro de la clase c_i. La función objetivo del agrupamiento K-means usa el cuadrado de la distancia Euclidiana $d\left(x, \mu_j\right) = \|x - \mu_j\|^2$ que también es conocida como inercia.

4.6. Análisis de regresión

La regresión esta relacionada a como hacemos predicciones de cantidades del mundo real, las predicciones hacen referencia a preguntas que tienen una estructura en común: se preguntan por una respuesta que puede ser expresada como una combinación de una o más variables (independientes) que también son llamadas covariables o predictores. El papel de la predicción es construir un modelo para predecir la respuesta de las variables y que puede ser útil en diferentes tareas, (1) analizar el comportamiento de los datos, (2) predecir los valores de los datos y (3) encontrar variables importantes para el modelo [28].

4.6.1. Mínimos cuadrados ordinarios

Mínimos cuadrados ordinarios es un método estadístico para estimar parámetros desconocidos en un modelo de regresión lineal [29]. Entonces sea x una variable independiente y sea $y(x)$ una función desconocida de x la cual queremos aproximar. Suponiendo que tenemos m observaciones

$$\left(x_1, y_1\right), \left(x_2, y_2\right), \ldots, \left(x_m, y_m\right)$$

donde $y_i \sim y(x_i)$, $i = 1, ..., m$, la idea es modelar $y(x)$ por medio de una combinación de

n funciones base $\phi_1(x), \phi_2(x), ...\phi_n(x)$. En el caso lineal suponemos que la función se ajusta a los datos en una combinación lineal de la forma

$$y(x) = c_1\phi_1(x) + c_2\phi_2(x) + \ldots + c_n\phi_n(x)$$

Entonces, los datos deben satisfacer de manera aproximada

$$y_i = c_1\phi_1(x_i) + c_2\phi_2(x_i) + \ldots + c_n\phi_n(x_i), \quad i = 1, 2, \ldots, m$$

La ecuación anterior puede expresarse de forma matricial como sigue

$$\begin{bmatrix} \phi_1(x_1) & \phi_2(x_1) & \ldots & \phi_n(x_1) \\ \phi_1(x_2) & \phi_2(x_2) & \ldots & \phi_n(x_2) \\ \vdots & & \ddots & \vdots \\ \phi_1(x_m) & \phi_2(x_m) & \ldots & \phi_n(x_m) \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_n \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}$$

El enfoque de mínimos cuadrados consiste en buscar aquel vector de coeficientes c que minimice el residual $r = y - Ac$. Entonces, el problema consiste en resolver

$$\min_{c \in \mathbb{R}^n} \|Ac - y\|^2$$

Es decir, para encontrar el ajuste de mínimos cuadrados debemos encontrar el vector de coeficientes $c = (c_1, ..., c_n)^T$ que minimiza la suma de los cuadrados:

$$\min_{c \in \mathbb{R}^n} \sum_{i=1}^{m} \left(c_1\phi_1(x_i) + c_2\phi_2(x_i) + \cdots + c_n\phi_n(x_i) - y_i\right)^2.$$

4.7. Análisis de componentes principales

La idea principal del análisis de componentes principales (PCA) es reducir la dimensionalidad de un conjunto de datos con variables que están interrelacionados entre sí, mientras retienen la mayor variación posible. Esto se logra transformándolas a un nuevo conjunto de variables, los componentes principales (PCs), las cuales no están relacionadas y están ordenadas de tal forma que las primeras tienen la mayor variación presente de todas las variables originales [30].

Supongamos que x es un vector de p variables aleatorias y que la varianza de las p

variables aleatorias y la estructura de la covarianza o correlación entre las p variables es de interés. El primer paso del PCA es buscar una función lineal $\boldsymbol{\alpha}_1'\boldsymbol{x}$ de los elementos de $\boldsymbol{x}$ que tenga máxima varianza, donde $\boldsymbol{\alpha}_1$ es un vector de p constantes $\alpha_{11}, \alpha_{12}, ..., \alpha_{1p}$ y $'$ denota la traspuesta, así que

$$\boldsymbol{\alpha}_1'\boldsymbol{x} = \alpha_{11}x_1 + \alpha_{12}x_2 + \cdots + \alpha_{1p}x_p = \sum_{j=1}^{p} \alpha_{1j}x_j$$

Consideremos de momento que el vector de variables aleatorias $\boldsymbol{x}$ tiene una matriz de covarianza conocida $\boldsymbol{\Sigma}$, esta es la matriz cuya (i, j)-ésimos elementos es la covarianza conocida entre el i-ésimo y j-ésimo elemento de $\boldsymbol{x}$ cuando $i \neq j$ y la varianza del j-ésimo elemento de $\boldsymbol{x}$ cuando $i = j$. Resulta que para $k = 1, 2, .., p$ el k-ésimo PC está dado por $z_k = \boldsymbol{\alpha_k'}\boldsymbol{x}$ donde $\boldsymbol{\alpha_k}$ es un eigenvector de $\boldsymbol{\Sigma}$ correspondiente al k-ésimo eigenvalor más grande λ_k. Aún más, si $\boldsymbol{\alpha_k'}$ es escogido de tal forma que tiene longitud unidad ($\boldsymbol{\alpha_k'}\boldsymbol{\alpha_k} = 1$), entonces la $var(z_k) = \lambda_k$, donde $var(z_k)$ denota la varianza de z_k.

Para derivar la forma de los PC, consideremos primero $\boldsymbol{\alpha_k'}$ donde $\boldsymbol{\alpha_k}$ maximiza $var(\boldsymbol{\alpha_k'}) = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1$. Entonces el problema a resolver es el siguiente:

Maximizar $\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1$

Sujeto a $\boldsymbol{\alpha_k'}\boldsymbol{\alpha_k} = 1$

Usando la técnica de los multiplicadores de Lagrange obtenemos que el eigenvalor λ de $\boldsymbol{\Sigma}$ con $\boldsymbol{\alpha}_1$ su correspondiente eigenvector. Para decidir cual de los p eigenvectores da $\boldsymbol{\alpha}_1'\boldsymbol{x}$ con máxima varianza, notemos que la cantidad a ser maximizada es la siguiente

$$\boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \boldsymbol{\alpha}_1'\lambda\boldsymbol{\alpha}_1 = \lambda\boldsymbol{\alpha}_1'\boldsymbol{\alpha}_1 = \lambda,$$

así λ debe ser tan grande como sea posible. Por tanto, $\boldsymbol{\alpha}_1$ es el eigenvector correspondiente al eigenvalor más grande de $\boldsymbol{\Sigma}$ y $\text{var}\,(\boldsymbol{\alpha}_1'\mathbf{x}) = \boldsymbol{\alpha}_1'\boldsymbol{\Sigma}\boldsymbol{\alpha}_1 = \lambda_1$, es el eigenvalor más grande. En general el k-ésimo PC de $\boldsymbol{x}$ es $\boldsymbol{\alpha_k'}\boldsymbol{x}$ y $var(\boldsymbol{\alpha_k'}\boldsymbol{x}) = \lambda_k$ donde λ_k es el k-ésimo eigenvalor más grande de $\boldsymbol{\Sigma}$, y $\boldsymbol{\alpha_k}$ es el correspondiente eigenvector.

Los vectores de los coeficientes $\boldsymbol{\alpha}_3, \boldsymbol{\alpha}_4, \ldots, \boldsymbol{\alpha}_p$, son los eigenvectores de $\boldsymbol{\Sigma}$ correspondiente a $\lambda_1, \lambda_2, ..., \lambda_p$.

$$\text{var}\left[\boldsymbol{\alpha}_k'\mathbf{x}\right] = \lambda_k \quad \text{para } k = 1, 2, \ldots, p$$

La derivación de los coeficientes PC y varianzas como eigenvectores y eigenvalores de una matriz de covarianza es estándar.

4.8. Agrupamiento jerárquico

El agrupamiento jerárquico es un método de análisis de grupos puntuales el cual busca construir una jerarquía de grupos. Hay dos estrategías para el agrupamiento jerárquico:

- Aglomerativas: Este es un acercamiento ascendente: cada observación comienza en su propio grupo, y los pares de grupos son mezclados mientras uno sube en la jerarquía.
- Divisivas: Este es un acercamiento descendente: todas las observaciones comienzan en un grupo, y se realizan divisiones mientras uno baja en la jerarquía.

Los resultados de un agrupamiento jerárquico son usualmente presentados en un dendrograma.

En orden de decidir qué grupos deberían ser combinados, o cuando un grupo debería ser dividido, una medida de disimilitud entre conjuntos de observaciones es requerida y un criterio de enlace el cual especifica la disimilitud de conjuntos como una función de las distancias dos a dos entre observaciones en los conjuntos.

Algunas métricas usualmente usadas para el agrupamiento jerárquico son las mostradas en la Sección 4.8.

Nombres	Fórmula
Distancia euclidiana	$\|a-b\|_2 = \sqrt{\sum_i (a_i - b_i)^2}$
Distancia euclidiana al cuadrado	$\|a-b\|_2^2 = \sum_i (a_i - b_i)^2$
Distancia Manhattan	$\|a-b\|_1 = \sum_i \|a_i - b_i\|$
distancia máxima	$\|a-b\|_\infty = \max_i \|a_i - b_i\|$
Distancia de Mahalanobis	$\sqrt{(a-b)^\top S^{-1}(a-b)}$ donde S es la matriz de covarianza
Similitud coseno	$\frac{a \cdot b}{\|a\|\|b\|}$

El criterio de enlace determina la distancia entre conjuntos de observaciones como una función de las distancias entre observaciones dos a dos. Algunos criterios de enlace entre dos conjuntos de observaciones A y B frecuentemente usados son las mostradas en Sección 4.8

Nombres	Fórmula
Agrupamiento de máximo o completo enlace	$\max\{d(a,b) : a \in A, b \in B\}$.
Agrupamiento de mínimo o simple enlace	$\min\{d(a,b) : a \in A, b \in B\}$.
Agrupamiento de enlace media o promedio, o UPGMA	$\frac{1}{\lvert A\rvert\lvert B\rvert}\sum_{a\in A}\sum_{b\in B} d(a,b)$.
Agrupamiento de mínima energía	$\frac{2}{nm}\sum_{i,j=1}^{n,m}\lVert a_i - b_j\rVert_2 - \frac{1}{n^2}\sum_{i,j=1}^{n}\lVert a_i - a_j\rVert_2 - \frac{1}{m^2}\sum$

donde d es la métrica escogida. Otros criterios de enlace incluyen pueden ser: la suma de todas las varianzas del intragrupo, el decrecimiento en la varianza para los grupos que están siendo mezclados (criterio de Ward) o la probabilidad de que grupos candidatos se produzcan desde la misma función de distribución (V-enlace).

4.9. Análisis topológico de datos

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

4.9.1. Teorema de Takens

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci

eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

4.9.2. Diagrama de persistencia

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

4.9.3. Panorama de persistencia

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

4.9.4. Normas topológicas

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque ha-

bitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Capítulo 5

Resultados en Bitcoin

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

5.1. Resultados

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper,

leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Capítulo 6

Conclusiones y trabajo futuro

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum. Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

6.1. Conclusión

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper,

leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Apéndice A

Métricas de la Blockchain

Métrica	**Definición**
AdrActCnt	El recuento de sumas de direcciones únicas que estaban activas en la red (ya sea como destino o fuente de un cambio de libro mayor) ese día.
AdrBal1in100KCnt	La suma de direcciones únicas que poseen al menos una en Xth del suministro actual de unidades nativas al final de ese día
AdrBal1in100MCnt	La suma de direcciones únicas que poseen al menos una en Xth del suministro actual de unidades nativas al final de ese día.
AdrBal1in10BCnt	La suma de direcciones únicas que poseen al menos una en Xth del suministro actual de unidades nativas al final de ese día.
AdrBal1in10KCnt	La suma de direcciones únicas que poseen al menos una en Xth del suministro actual de unidades nativas al final de ese día.
AdrBal1in10MCnt	La suma de direcciones únicas que poseen al menos una en Xth del suministro actual de unidades nativas al final de ese día.
AdrBal1in1BCnt	La suma de direcciones únicas que poseen al menos una en Xth del suministro actual de unidades nativas al final de ese día.
AdrBal1in1KCnt	La suma de direcciones únicas que poseen al menos una en Xth del suministro actual de unidades nativas al final de ese día.
AdrBal1in1MCnt	La suma de direcciones únicas que poseen al menos una en Xth del suministro actual de unidades nativas al final de ese día.
AdrBalCnt	El recuento de sumas de direcciones únicas que tengan cualquier cantidad de unidades nativas al final de ese día
AdrBalNtv0.001Cnt	El recuento de sumas de direcciones únicas que tengan al menos X unidades nativas al final de ese día.
AdrBalNtv0.01Cnt	El recuento de sumas de direcciones únicas que tengan al menos X unidades nativas al final de ese día.
AdrBalNtv0.1Cnt	El recuento de sumas de direcciones únicas que tengan al menos X unidades nativas al final de ese día.
AdrBalNtv100Cnt	El recuento de sumas de direcciones únicas que tengan al menos X unidades nativas al final de ese día.

Métrica	Definición
AdrActCnt	El recuento de sumas de direcciones únicas que estaban activas en la red (ya sea como destino o fuente de un cambio de libro mayor) ese día.
AdrBal1in100KCnt	La suma de direcciones únicas que poseen al menos una en Xth del suministro actual de unidades nativas al final de ese día
AdrBal1in100MCnt	La suma de direcciones únicas que poseen al menos una en Xth del suministro actual de unidades nativas al final de ese día.
AdrBal1in10BCnt	La suma de direcciones únicas que poseen al menos una en Xth del suministro actual de unidades nativas al final de ese día.
AdrBal1in10KCnt	La suma de direcciones únicas que poseen al menos una en Xth del suministro actual de unidades nativas al final de ese día.
AdrBal1in10MCnt	La suma de direcciones únicas que poseen al menos una en Xth del suministro actual de unidades nativas al final de ese día.
AdrBal1in1BCnt	La suma de direcciones únicas que poseen al menos una en Xth del suministro actual de unidades nativas al final de ese día.
AdrBal1in1KCnt	La suma de direcciones únicas que poseen al menos una en Xth del suministro actual de unidades nativas al final de ese día.
AdrBal1in1MCnt	La suma de direcciones únicas que poseen al menos una en Xth del suministro actual de unidades nativas al final de ese día.
AdrBalCnt	El recuento de sumas de direcciones únicas que tengan cualquier cantidad de unidades nativas al final de ese día
AdrBalNtv0.001Cnt	El recuento de sumas de direcciones únicas que tengan al menos X unidades nativas al final de ese día.
AdrBalNtv0.01Cnt	El recuento de sumas de direcciones únicas que tengan al menos X unidades nativas al final de ese día.
AdrBalNtv0.1Cnt	El recuento de sumas de direcciones únicas que tengan al menos X unidades nativas al final de ese día.
AdrBalNtv100Cnt	El recuento de sumas de direcciones únicas que tengan al menos X unidades nativas al final de ese día.

Cuadro A.1: Métricas de la blockchain y su definición.

Apéndice B

Metodología del estado del arte

B.1. Metodología SLR

La revisión sistemática de literatura (SLR) es un método que nos permite interpretar y sintetizar de forma adecuada la información obtenida de un tema de investigación. Dado que se ejecuta una manera sistemática sus ventajas son un incremento en la posibilidad de tener mejores resultados en la búsqueda de información [31]. Hay tres fases principales en una SLR: (1) Planificación de revisión, (2) Realización de la revisión, (3) Reporte de resultados.

Planificación de revisión Esta fase se centra en: (1) La identificación de la necesidad de realizar el SLR, (2) formular las preguntas de investigación que guían la ejecución del SLR, (3) generar la búsqueda de cadenas, y (4) la selección de las fuentes de información para la extracción de los estudios primarios.

Realización de la revisión La segunda fase del SLR corresponde a la definición de los criterios de inclusión y exclusión, la ejecución de un proceso de selección de estudios y la evaluación de la calidad de los mismos, lo que da lugar a la selección de los estudios primarios.

La metodología se aplicará a tres factores clave de la tesis: el pronóstico del precio, análisis de métricas de la blockchain y la clasificación de precios de la criptomoneda.

B.2. Pronóstico del precio del bitcoin

B.2.1. Planificación de revisión

Identificar la necesidad de realizar el SLR

Tener un enfoque sistemático a la hora de comparar y escoger un algoritmo para la predicción del bitcoin es crucial por la cantidad de datos e información que se pueden obtener hoy día con actualización constante. Una buena comprensión de las variables explicativas, formateo de datos y métodos para comparar los modelos es fundamental a la hora de comprender el modelo propuesto.

Definiendo las preguntas de investigación

Las preguntas de investigación para este estudio son las siguientes:

- ¿Cuál es el estado del arte de modelos estadísticos para pronósticos del bitcoin?
- ¿Cómo implementar trading con bitcoin?
- ¿Cuales son los modelos de pronósticos más utilizados?
- ¿Qué modelos de aprendizaje máquina se están utilizando?
- ¿Cuál es el mejor modelo para realizar pronóstico del bitcoin?

Generar las cadenas de búsqueda

Para facilitar la búsqueda de los estudios primarios se identificaron las palabras clave resultantes de las preguntas de investigación. Resultaron en las siguientes:

- Bitcoin
- Machine learning
- Trading
- Forecasting

Combinando las palabras clave con el uso de conectores lógicos "AND" y "OR", la siguiente cadena de búsqueda es obtenida:

Bitcoin **AND** (machine learning **OR** trading **OR** forecasting)

Selección de fuentes de información

Las siguientes fuentes de información fueron seleccionadas para la extracción de los estudios:

- Google Scholar
- IEEE Xplorer
- ELSEVIER Science
- Springer Link

B.2.2. Realización de la revisión

Criterio de inclusión y exclusión

Para filtrar los estudios no relevantes para esta investigación se definen los criterios de inclusión y exclusión, que da lugar a los presentados en la Tabla B.1.

Criterio de inclusión	Criterio de exclusión
Los primeros estudios mas relevantes según los filtros de búsqueda de las fuentes de información seleccionadas.	Estudios no relevantes según los filtros de búsqueda de las fuentes de información.
El título contiene la palabra clave "Bitcoin" y al menos otra palabra clave.	El título no contiene ninguna palabra clave
Abstract está relacionado con predicción de precios del bitcoin usando machine learning o métodos estadísticos.	Estudios que no están relacionados con la predicción del bitcoin.
Estudios que contienen enfoques de metodologías, modelos, métodos, técnicas de predicción del bitcoin.	Estudios relacionados con el bitcoin pero no tienen un enfoque en ciencia de datos.
Estudios que contengan resultados sobre la predicción del precio del bitcoin con sus respectivos indicadores de error y comparación con otros métodos.	Estudios con resultados sobre la predicción del bitcoin pero sin comparación de modelos.
Estudios publicados entre 2016 y 2020.	Estudios publicados antes del 2016.

Tabla B.1: Criterios de inclusión y exclusión

Selección de estudios primarios

El protocolo SLR [31] sugiere definir un proceso de selección de estudios para obtener los estudios primarios, estructurado de la siguiente manera: (1) aplicar y adaptar la cadena de búsqueda a cada fuente de datos, (2) filtrar los estudios aplicando los primeros criterios de inclusión a cada fuente de información, (3) aplicar el resto de los criterios de inclusión y exclusión, y (4) seleccionar los estudios primarios. Después de aplicar los criterios de inclusión y exclusión como parte del proceso de selección de estudios, se seleccionaron 8 estudios primarios para esta investigación como se ve en la Figura B.1.

Evaluación de la calidad del estudio

Al evaluar la calidad del estudio, se garantiza que la información contenida en cada uno de los estudios primarios sea pertinente y valiosa para la investigación. En la Tabla B.2 se presenta la evaluación de la calidad de los estudios que se aplicarán.

Después de evaluar los estudios primarios utilizando las evaluaciones de calidad antes mencionadas quedaron **6 estudios primarios** [4, 12, 3, 5, 8, 6].

B.3. Análisis de métricas de la blockchain

B.3.1. Planificación de revisión

Identificar la necesidad de realizar el SLR

Tener un enfoque sistemático de las métricas que más influyen en el precio del bitcoin es crucial por la cantidad de datos e información que se pueden obtener hoy día con actua-

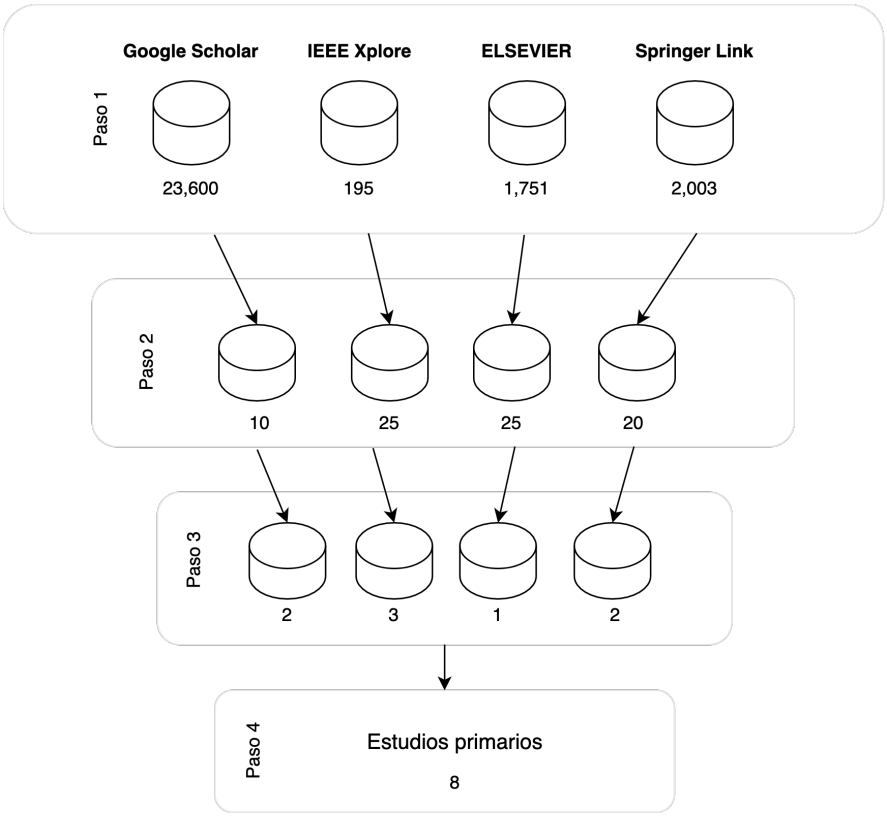

Figura B.1: Selección de estudios primarios y resultados

ID	Evaluación de la calidad del estudio
PC1	¿El estudio cuenta con una base teórica que explique los métodos utilizados?
PC2	¿El estudio detalla el método con el cuál pronostica y que variables explicativas utiliza para hacer la predicción?
PC3	¿El estudio compara modelos de predicción para pronosticar el precio del bitcoin?
PC4	¿El estudio utiliza algoritmos de machine learning para realizar la predicción?
PC5	¿El estudio muestra un modelo con mejor rendimiento comparado con los demás?

Tabla B.2: Evaluación de la calidad del estudio

lización constante, más tratándose de un campo tecnológico con evolución persistente. Una buena comprensión de las variables explicativas ayuda a mejorar y ahorrar esfuerzos en el entendimiento del asenso y descenso del precio de esta criptomoneda.

Definiendo las preguntas de investigación

Las preguntas de investigación son las siguientes:

- ¿Con qué métodos se están seleccionando las mejores métricas de la blockchain para

predicción del precio?

- Actualmente, ¿cuales son las métricas de la blockchain que más influyen en el precio del bitcoin?
- ¿Cuánto mejora la predicción con las métricas de la blockchain?

Generar las cadenas de búsqueda

Para facilitar la búsqueda de los estudios primarios se identificaron las palabras clave resultantes de las preguntas de investigación. Resultaron en las siguientes:

- Bitcoin
- Blockchain
- Metrics
- Prediction
- Features

Combinando las palabras clave con el uso de conectores lógicos "AND" y "OR", la siguiente cadena de búsqueda es obtenida:

(Bitcoin **AND** blockchain **AND** metrics **AND** prediction)
OR (Bitcoin **AND** features **AND** prediction)

Selección de fuentes de información

Las siguientes fuentes de información fueron seleccionadas para la extracción de los estudios:

- IEEE Xplore
- ELSEVIER Science Direct
- Springer Link

B.3.2. Realización de la revisión

Criterio de inclusión y exclusión

Para filtrar los estudios no relevantes para esta investigación se definen los criterios de inclusión y exclusión, que da lugar a los presentados en la Tabla B.3.

Criterio de inclusión	Criterio de exclusión
Los primeros estudios más relevantes según los filtros de búsqueda de las fuentes de información seleccionadas.	Estudios no relevantes según los filtros de búsqueda de las fuentes de información.
El titulo contiene la palabra Bitcoin o Blockchain y al menos otra palabra clave.	El titulo no contiene ninguna palabra clave.
El estudio muestra el o los método utilizados para la selección de métricas.	El estudio no muestra el o los métodos utilizados para la selección de métricas.
Estudios publicados entre 2017-2021	Estudios publicados antes del 2017.

Tabla B.3: Criterios de inclusión y exclusión

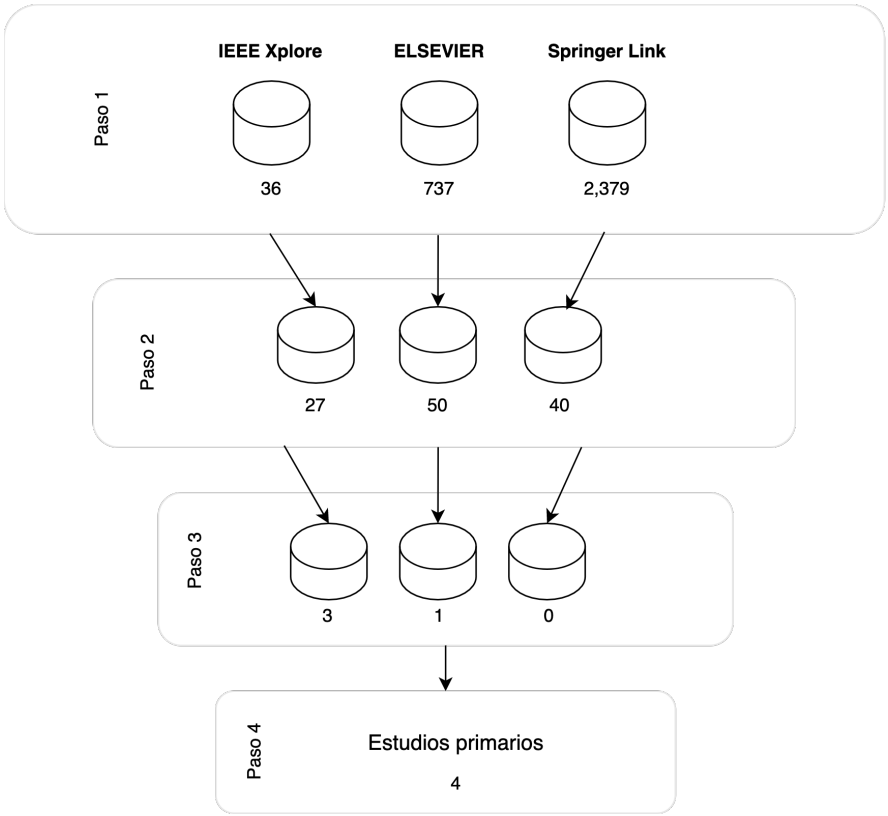

Figura B.2: Selección de estudios primarios y resultados

Selección de estudios primarios

Después de aplicar los criterios de inclusión y exclusión como parte del proceso de selección de estudios, se seleccionaron 4 estudios primarios para esta investigación como se ve en la Figura B.2.

Evaluación de la calidad del estudio

Al evaluar la calidad del estudio, se garantiza que la información contenida en cada uno de los estudios primarios sea pertinente y valiosa para la investigación. En la Tabla B.4 se presenta la evaluación de la calidad de los estudios que se aplicarán.

ID	Evaluación de la calidad del estudio
PC1	¿El estudio detalla y explica la metodología utilizada para encontrar las variables explicativas que mejoran la predicción?
PC2	¿El estudio concluye si agregar variables de la blockchain mejora o no la predicción del bitcoin?

Tabla B.4: Evaluación de la calidad del estudio

Después de evaluar los estudios primarios utilizando las evaluaciones de calidad antes mencionadas quedaron 3 estudios primarios [11, 32, 9].

B.4. Clasificación del precio para inversión

B.4.1. Planificación de revisión

Identificar la necesidad de realizar el SLR

Un enfoque sistemático de los métodos de clasificación más utilizados es de suma importancia ya que por la existencia de una gran variedad de métodos y cambios constantes en los mismos con un ritmo acelerado pude ser abrumador, por lo anterior el entendimiento de los modelos de clasificación con un enfoque SLR puede impulsar la creación de nuevos modelos más robustos y con mejor precisión.

Definiendo las preguntas de investigación

Las preguntas de investigación para este estudio son las siguientes:

- ¿Qué modelos de clasificación se están utilizando?
- ¿Cómo se clasifican los precios del bitcoin para la toma de decisiones de inversión?
- ¿Cuál es la mejor metodología para clasificación de precios del bitcoin?

Generar las cadenas de búsqueda

Para facilitar la búsqueda de los estudios primarios se identificaron las palabras clave resultantes de las preguntas de investigación. Resultaron en las siguientes:

- Bitcoin
- Blockchain
- Investment
- Classification
- Deep Learning

Combinando las palabras clave con el uso de conectores lógicos “AND” y “OR”, la siguiente cadena de búsqueda es obtenida:

(Bitcoin **AND** classification **AND** deep learning)
OR (Bitcoin **AND** classification **AND** deep learning **AND** investment)

Selección de fuentes de información

Las siguientes fuentes de información fueron seleccionadas para la extracción de los estudios:

- IEEE Xplore
- ELSEVIER Science Direct
- Springer Link

B.4.2. Realización de la revisión

Criterio de inclusión y exclusión

Para filtrar los estudios no relevantes para esta investigación se definen los criterios de inclusión y exclusión, que da lugar a los presentados en la Tabla B.5.

Criterio de inclusión	**Criterio de exclusión**
Los primeros estudios más relevantes según los filtros de búsqueda de las fuentes de información seleccionadas.	Estudios no relevantes según los filtros de búsqueda de las fuentes de información.
El titulo contiene la palabra Bitcoin o Blockchain y al menos otra palabra clave.	El titulo no contiene ninguna palabra clave.
El abstract está relacionado con la clasificación del precios del Bitcoin.	Estudios que no estaban relacionados con la clasificación del precio del bitcoin.
Estudios que contienen resultados sobre la clasificación de precios del Bitcoin con sus respectivos indicadores de precisión.	Estudios que no muestran indicadores de precisión de la clasificación.
Estudios publicados entre 2017-2021	Estudios publicados antes del 2017.

Tabla B.5: Criterios de inclusión y exclusión

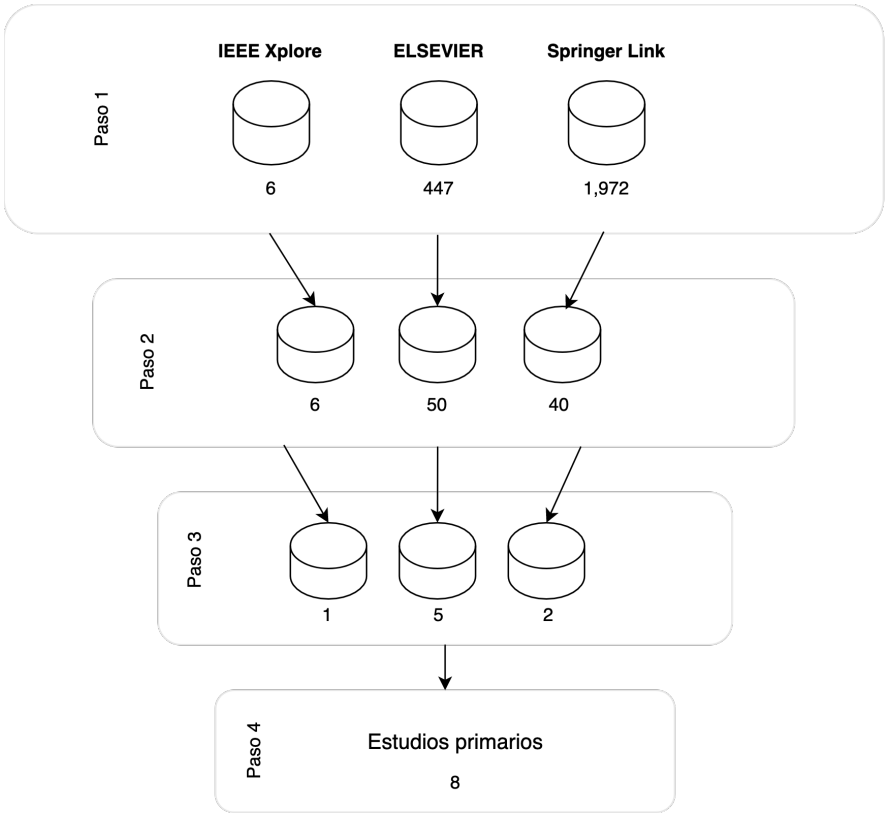

Figura B.3: Selección de estudios primarios y resultados

Selección de estudios primarios

Después de aplicar los criterios de inclusión y exclusión como parte del proceso de selección de estudios, se seleccionaron 8 estudios primarios para esta investigación como se ve en la Figura B.3.

Evaluación de la calidad del estudio

Al evaluar la calidad del estudio, se garantiza que la información contenida en cada uno de los estudios primarios sea pertinente y valiosa para la investigación. En la Tabla B.6 se presenta la evaluación de la calidad de los estudios que se aplicarán.

ID	Evaluación de la calidad del estudio
PC1	¿El estudio detalla la metodología utilizada para realizar la clasificación de los precios?
PC2	¿El estudio da una descripción de los modelos utilizados para realizar la clasificación?
PC3	¿El estudio muestra un modelo de clasificación con un mejor rendimiento comparado con los demás?

Tabla B.6: Evaluación de la calidad del estudio

Después de evaluar los estudios primarios utilizando las evaluaciones de calidad antes mencionadas quedaron 5 estudios primarios [15, 14, 12, 16, 13].

B.5. Discusión de la metodología SLR

En esta sección daremos respuesta a nuestras preguntas de investigación con base en los artículos primarios.

B.5.1. Pronóstico del precio del bitcoin

1.- ¿Cuál es el estado del arte de modelos estadísticos para pronósticos de bitcoin?
Bitcoin es una criptomoneda altamente usada hoy día, por ello hay algunos modelos para la predicción de su precio, ya sea con modelos estadísticos más tradicionales como regresión lineal o métodos de deep learning como LSTM [4]. Sin embargo, algunos inversores no tratan al bitcoin como una moneda de acuerdo al criterio usado por los economistas y hacen una inversión especulativa [12]. Una pregunta natural sería que características tomar en cuenta a la hora de hacer la predicción. Ya que bitcoin no cuenta una tendencia clara y tienen una alta volatilidad entonces los métodos de deep learning son una solución efectiva [4]. Más aun, se puede demostrar que es una serie de tiempo no estacionaria [3] y por consiguiente mas conveniente los métodos de aprendizaje maquina, aunque, igualmente, dado que los datos financieros están en el formato OHLC se muestra que todas las variables están altamente correlacionadas entre si y los métodos estadísticos mas tradicionales pueden ser efectivos [6].

2. ¿Cómo implementar trading con bitcoin?
Hay muchos enfoques a la hora de realizar la compra y venta de este activo, una cantidad numerosa de estudios han llegado a la conclusión de que usando indicadores técnicos del bitcoin se puede predecir con precisión los retornos generados [3].
Por otro lado se tienen estudios que indican que una manera de hacer trading es calcular el cambio del precio en intervalos de una hora [6], al final del día o cada treinta o noventa día basándose en el volumen del movimientos de los precios del bitcoin [3]. También se ha concluido que hacer trading tomando en cuenta las variables asociadas al precio como la apertura, valor máximo, mínimo y cierre, dan buenos resultado [5, 6]. Por ultimo se puede hacer ingeniería de características y tomar características de alta dimensionalidad para realizar trading en intervalos de cinco minutos [12] .

3. ¿Cuáles son los modelos de pronósticos más utilizados?
Dada la naturaleza del precio del bitcoin los métodos de aprendizaje profundo son los favoritos en este caso RNN (Recurrent Neural Network) y LSTM [8].

4. ¿Qué modelos de aprendizaje máquina se estan utilizando?
En los estudios comparativos es frecuente utilizar los siguientes modelos: ARIMA, RTS, RF, SVM y LSTM [5, 6].

5. ¿Cuál es el mejor modelo para realizar pronóstico del bitcoin?

En la mitad de los estudios primarios seleccionados [3, 12, 8] el mejor modelo de predicción son las redes LSTM basados en la exactitud (accuracy) con una puntuación que va desde un 52.78 % [8], hasta un 67.2 % [12]. En los demás estudios se tiene incertidumbre y especifican mas investigación.

B.5.2. Análisis de métricas de la blockchain

1. ¿Con que métodos se están seleccionando las mejores métricas de la blockchain para predicción del precio?
En la literatura disponible se están utilizando diversas aproximaciones para la selección de métricas con mayor influencia. En los estudios primarios los métodos de correlación [32, 9] fueron los mas predominantes. En estos se calcula la correlación que existe entre el precio y las características de la blockchain, se seleccionan las que tienen el coeficiente más alto y a su vez mejoran los modelos de predicción que no incluyen las métricas propuestas. Cambien existen métodos que utilizan análisis de sensibilidad y arboles aleatorios para reducir el subconjunto de predictores midiendo la importancia del factor tecnológico [11]. En este se seleccionaron las métricas con el índice de importancia más alto.

2. Actualmente, ¿cuales son las métricas de la blockchain que más influyen en el precio del bitcoin?
En [32] se encontró que entre 84 métricas obtenidas las mejores fueron aquellas relacionas con el indice de transacciones nVout que cuenta con cinco variables. Por otro lado en [9] utilizando igualmente análisis de correlación se obtuvo que las métricas con mayor influencia fueron las relacionadas con el número de carteras, la dificultad de minado, el hash rate y UTX's que es el conjunto de salidas de transacciones no gastadas. Por último en [11] las mejores características de la blockchain fueron la capitalización del mercado, el valor promedio de transacciones, la tasa promedio de transacciones, la dificultad de minado y el tamaño del bloque.

3. ¿Cuánto mejora la predicción con las métricas de la blockchain?
En [9, 11] se utilizo el modelo LSTM para la comparación en las mejoras, en [11] se logró una reducción de hasta 73 dólares en términos del RMSE promedio, mientras que en el error medio absoluto (MAE) se logro una reducción de 606,69 hasta 548,15. Por otro lado en [9] se alcanzó un MAE mínimo 0,0889 sobre el conjunto de validación. En [32] se utilizo regresión lineal sobre series de tiempo y se alcanzo una precisión del 95,04 %, mostrando una mejora en comparación del estado del arte [8] donde se logró una precisión del 52,78 %.

B.5.3. Clasificación del precio para inversión

1. ¿Que modelos de clasificación se están utilizando?
El estados del arte de los estudios primarios nos muestran que los modelos de machine learning y deep learning son los más populares entre ellos destacando Random Forest (RF), Linear Regresion (LR), Support Machine Vector (SVM), XGBoost y Long short-term memory (LSTM) [15, 14, 12, 16, 13].

2. ¿Cómo se clasifican los precios del bitcoin para la toma de decisiones de inversión?
En la mayoría de estudios la clasificación se realiza para trading, esto es, se clasifica el precio con base en si este sube o baja tomando en cuenta alguna característica que involucre pequeños periodos de tiempo. Se puede encontrar clasificación de precios que toman en cuenta si el precio incrementó en periodos de tiempo de 5, 15, 30, 60 minutos o diariamente [15, 14, 12, 13]. También hay criterios que toman en cuenta si el precio de apertura del día es mayor o menor que el precio de cierre para realizar la clasificación [16].

3. ¿Cuál es la mejor metodología para clasificación del precio del bitcoin?
De todos los estudios primarios estudiados el modelo que mejor precisión alcanzó fue LSTM alcanzando un 67,2 % de *accuracy* sobre un intervalo de actualización de 5 minutos [12]. El método propuesto en este estudio se caracteriza por incluir características de alta dimensionalidad como métricas de la blockchain y del mercado.

Referencias

[1] S. Nakamoto, "Bitcoin: A Peer-to-Peer Electronic Cash System," p. 9.

[2] Bitcoin USD (BTC-USD) Price, News, Quote & History - Yahoo Finance. [Online]. Available: https://finance.yahoo.com/quote/BTC-USD/chart/

[3] M. Mudassir, S. Bennbaia, D. Unal, and M. Hammoudeh, "Time-series forecasting of Bitcoin prices using high-dimensional features: A machine learning approach." [Online]. Available: https://link.springer.com/10.1007/s00521-020-05129-6

[4] S. Tandon, S. Tripathi, P. Saraswat, and C. Dabas, "Bitcoin Price Forecasting using LSTM and 10-Fold Cross validation," in *2019 International Conference on Signal Processing and Communication (ICSC)*. IEEE, pp. 323–328. [Online]. Available: https://ieeexplore.ieee.org/document/8938251/

[5] L. Felizardo, R. Oliveira, E. Del-Moral-Hernandez, and F. Cozman, "Comparative study of Bitcoin price prediction using WaveNets, Recurrent Neural Networks and other Machine Learning Methods," in *2019 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC)*. IEEE, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8963009/

[6] T. Phaladisailoed and T. Numnonda, "Machine Learning Models Comparison for Bitcoin Price Prediction," in *2018 10th International Conference on Information Technology and Electrical Engineering (ICITEE)*. IEEE, pp. 506–511. [Online]. Available: https://ieeexplore.ieee.org/document/8534911/

[7] K. Żbikowski, "Application of Machine Learning Algorithms for Bitcoin Automated Trading," in *Machine Intelligence and Big Data in Industry*, ser. Studies in Big Data, D. Ryżko, P. Gawrysiak, M. Kryszkiewicz, and H. Rybiński, Eds. Springer International Publishing, vol. 19, pp. 161–168. [Online]. Available: http://link.springer.com/10.1007/978-3-319-30315-4_14

[8] S. McNally, J. Roche, and S. Caton, "Predicting the Price of Bitcoin Using Machine Learning," in *2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP)*. IEEE, pp. 339–343. [Online]. Available: https://ieeexplore.ieee.org/document/8374483/

[9] M. Saad and A. Mohaisen, "Towards characterizing blockchain-based cryptocurrencies for highly-accurate predictions," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*. IEEE, pp. 704–709. [Online]. Available: https://ieeexplore.ieee.org/document/8406859/

[10] S. Ji, J. Kim, and H. Im, "A Comparative Study of Bitcoin Price Prediction Using Deep Learning," vol. 7, no. 10, p. 898. [Online]. Available: https://www.mdpi.com/2227-7390/7/10/898

[11] W. Chen, H. Xu, L. Jia, and Y. Gao, "Machine learning model for Bitcoin exchange rate prediction using economic and technology determinants," vol. 37, no. 1, pp. 28–43. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0169207020300431

[12] Z. Chen, C. Li, and W. Sun, "Bitcoin price prediction using machine learning: An approach to sample dimension engineering," vol. 365, p. 112395. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S037704271930398X

[13] E. Pintelas, I. E. Livieris, S. Stavroyiannis, T. Kotsilieris, and P. Pintelas, "Investigating the Problem of Cryptocurrency Price Prediction: A Deep Learning Approach," in *Artificial Intelligence Applications and Innovations*, ser. IFIP Advances in Information and Communication Technology, I. Maglogiannis,

L. Iliadis, and E. Pimenidis, Eds. Springer International Publishing, vol. 584, pp. 99–110. [Online]. Available: http://link.springer.com/10.1007/978-3-030-49186-4_9

[14] P. Jaquart, D. Dann, and C. Weinhardt, "Short-term bitcoin market prediction via machine learning," vol. 7, pp. 45–66. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2405918821000027

[15] A. Ibrahim, R. Kashef, and L. Corrigan, "Predicting market movement direction for bitcoin: A comparison of time series modeling methods," vol. 89, p. 106905. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0045790620307576

[16] E. Akyildirim, A. Goncu, and A. Sensoy, "Prediction of cryptocurrency returns using machine learning," vol. 297, no. 1-2, pp. 3–36. [Online]. Available: http://link.springer.com/10.1007/s10479-020-03575-y

[17] Yahoo Finance - Stock Market Live, Quotes, Business & Finance News. [Online]. Available: https://finance.yahoo.com/

[18] Bitcoin & Cryptocurrency Exchange — Bitcoin Trading Platform — Kraken. [Online]. Available: https://www.kraken.com/

[19] Investing.com - Stock Market Quotes & Financial News. Investing.com. [Online]. Available: https://www.investing.com/

[20] API Basics. [Online]. Available: https://docs.coinmetrics.io/api

[21] L. A. Pinedo-Sánchez, D. A. Mercado-Ravell, and C. A. Carballo-Monsivais, "Vibration analysis in bearings for failure prevention using CNN," vol. 42, no. 12, p. 628. [Online]. Available: http://link.springer.com/10.1007/s40430-020-02711-w

[22] F. Bliemel, "Theil's Forecast Accuracy Coefficient: A Clarification," vol. 10, no. 4, pp. 444–446. [Online]. Available: https://doi.org/10.1177/002224377301000413

[23] R. J. Hyndman and G. Athanasopoulos, "Forecasting: Principles and Practice," p. 504.

[24] T. J. Cole and D. G. Altman, "Statistics Notes: Percentage differences, symmetry, and natural logarithms," p. j3683. [Online]. Available: https://www.bmj.com/lookup/doi/10.1136/bmj.j3683

[25] Tslm: Fit a linear model with time series components in forecast: Forecasting Functions for Time Series and Linear Models. [Online]. Available: https://rdrr.io/cran/forecast/man/tslm.html

[26] What is Random Forest? [Online]. Available: https://www.ibm.com/cloud/learn/random-forest

[27] I. E. Livieris, E. Pintelas, S. Stavroyiannis, and P. Pintelas, "Ensemble Deep Learning Models for Forecasting Cryptocurrency Time-Series," vol. 13, no. 5, p. 121. [Online]. Available: https://www.mdpi.com/1999-4893/13/5/121

[28] L. Igual and S. Seguí, *Introduction to Data Science: A Python Approach to Concepts, Techniques and Applications*, 1st ed., ser. Undergraduate Topics in Computer Science. Springer International Publishing : Imprint: Springer.

[29] D. L. H. Juárez and D. A. Léon, "4to coloquio del departamento de matemáticas," p. 83.

[30] I. T. Jolliffe, *Principal Component Analysis*, 2nd ed., ser. Springer Series in Statistics. Springer.

[31] B. Kitchenham, O. Pearl Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic literature reviews in software engineering – A systematic literature review," vol. 51, no. 1, pp. 7–15. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0950584908001390

[32] S.-H. Ji, U.-J. Baek, M.-G. Shin, Y.-H. Goo, J.-S. Park, and M.-S. Kim, "Best Feature Selection using Correlation Analysis for Prediction of Bitcoin Transaction Count," in *2019 20th Asia-Pacific Network Operations and Management Symposium (APNOMS)*. IEEE, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8892896/