

Redes Neuronales LTSM En El Análisis De Tendencias De Criptomonedas

Autor(es)

[IVER JOHAN HINCAPIE BETANCUR](mailto:iver.hincapie@udea.edu.co)

[ELIANA HERRAN LOGREIRA](mailto:eliana.herran@udea.edu.co)

Universidad de Antioquia

Facultad de Ingeniería

Ciudad, Colombia

2022

INSTRUCCIONES

* Copia este documento en tu drive
* Usa Heading 1,2,3, etc para estructurar las secciones en tu texto y que así el índice se actualice automáticamente
* Usa [[Cross Reference](https://gsuite.google.com/marketplace/app/cross_reference/269114033347)] para referenciar tablas, figuras, bibliografía, etc. en tu texto (es un add-on de Google docs). Lee las [instrucciones](https://github.com/davidrthorn/cross_reference/blob/master/README.md) detenidamente
* Elimina esta página de tu documento.
* No hay un requisito máximo ni mínimo de longitud de este documento. A modo de guía, entre 15 y 30 páginas es una longitud promedio.
* Respeta las secciones de primer nivel. Modifica las subsecciones si lo consideras necesario.
* Sé breve y preciso en tus descripciones. El objetivo no es que este documento sea largo, sino que describa adecuadamente el trabajo que has hecho.

NOTEBOOKS

* Has de acompañar este documento con un repositorio Github con los notebooks que implementen los procesos e iteraciones de tu proyecto.
* Incluye un notebook llamado ‘**00-INSTRUCCIONES.ipynb’** en el que describes cómo ejecutar tus notebooks.
* Los notebooks del curso han de estar publicados en un repositorio Github y, si posible, han de ser ejecutables desde Colab.
  + Si desarrollaste tu proyecto en otro entorno (por ejemplo Kaggle) incluye en el notebook de instrucciones cómo ejecutar tus notebooks en Kaggle
* Si puedes, incluye en tu Github los datos de tu proyecto. Si no puedes (porque son muy grandes, porque ya están en Kaggle, o porque son privados) incluye en el notebook con instrucciones sobre cómo obtenerlos.
  + Si tus notebooks son privados (p.ej. por restricciones de la empresa para la cual se hizo la monografía) haz el repositorio privado.
  + Si tus datos son privados incluye en las instrucciones tu dirección de correo como contacto para solicitar los datos.
* Adjunta un zip del repositorio (usa *download zip* desde Github) en tu entrega en la biblioteca.

TABLA DE CONTENIDOS

[**1. RESUMEN EJECUTIVO**](#_heading=h.30j0zll) **4**

[**2. DESCRIPCIÓN DEL PROBLEMA**](#_heading=h.1fob9te) **5**

[2.1 PROBLEMA DE NEGOCIO](#_heading=h.3znysh7) 5

[2.2 APROXIMACIÓN DESDE LA ANALÍTICA DE DATOS](#_heading=h.2et92p0) 5

[2.3 ORIGEN DE LOS DATOS](#_heading=h.tyjcwt) 5

[2.4 MÉTRICAS DE DESEMPEÑO](#_heading=h.3dy6vkm) 5

[**3. DATOS**](#_heading=h.1t3h5sf) **6**

[3.1 DATOS ORIGINALES](#_heading=h.4d34og8) 6

[3.2 DATASETS](#_heading=h.2s8eyo1) 6

[3.3 DESCRIPTIVA](#_heading=h.17dp8vu) 6

[**4. PROCESO DE ANALÍTICA**](#_heading=h.3rdcrjn) **7**

[4.1 PIPELINE PRINCIPAL](#_heading=h.26in1rg) 7

[4.2 PREPROCESAMIENTO](#_heading=h.lnxbz9) 7

[4.3 MODELOS](#_heading=h.35nkun2) 7

[4.4 MÉTRICAS](#_heading=h.1ksv4uv) 7

[**5. METODOLOGÍA**](#_heading=h.44sinio) **8**

[5.1 BASELINE](#_heading=h.2jxsxqh) 8

[5.2 VALIDACIÓN](#_heading=h.z337ya) 8

[5.3 ITERACIONES y EVOLUCIÓN](#_heading=h.3j2qqm3) 8

[5.4 HERRAMIENTAS](#_heading=h.1y810tw) 8

[**6. RESULTADOS**](#_heading=h.4i7ojhp) **9**

[6.1 MÉTRICAS](#_heading=h.2xcytpi) 9

[6.2 EVALUACIÓN CUALITATIVA](#_heading=h.1ci93xb) 9

[6.3 CONSIDERACIONES DE PRODUCCIÓN](#_heading=h.3whwml4) 9

# 1. RESUMEN EJECUTIVO

Esto es un abstract. Realiza una descripción general de tu proyecto: qué se persigue, qué datos se tiene, qué estrategia se siguió para las iteraciones, que obstáculos hubo, qué resultados se obtuvieron, etc. **La longitud máxima es de una página**

Incluye al final de dicha página la dirección de tu repositorio github

# 2. DESCRIPCIÓN DEL PROBLEMA

# A partir del año 2009, en el mercado financiero entró una nueva divisa llamada bitcoin, una moneda digital creada con técnicas criptográficas para asegurar que las transacciones que se realicen con ella sean lo más seguras posibles, permitiendo la compra y venta de objetos en el mercado internacional como cualquier otra moneda, siempre y cuando esta haya sido aceptada como divisa de cambio. Este tipo de dinero, transformaría por completo el paradigma de transacciones que se tenía hasta ese entonces de forma electrónica, al posicionarse como una nueva manera de abrir mercado sin la necesidad de tener la validación de la banca mundial, ni organizaciones gubernamentales. Con esta divisa se avecinaba una nueva etapa de descentralización de la moneda. Y aunque años anteriores en los Estados Unidos ya funcionaba esta idea de criptografía, con la denominada eCash y posterior digiCash, en la década de los 90's, pero no fue hasta la crisis económica del 2008 que esta idea cobró una mayor fuerza, generando un rápido crecimiento del Bitcoin que impulsó la creación de más criptomonedas en años posteriores y con esto los analistas financieros empezaron a incluir en sus análisis el comportamiento de este activo y su interacción con el mercado existente.

Por lo tanto, y dado el interés alrededor de las variaciones que se presentan en los mercados de criptomonedas por su relación con el mundo moderno y por sus altos indicadores de ganancias y fluctuaciones, hemos sentido la motivación de desarrollar un modelo de machine learning que nos permita acercarnos a una predicción sobre el comportamiento de estas monedas en el mercado bursátil, de tal manera que esto ayude a tomar decisiones de inversión con relación al cierre del día anterior de la moneda de interés.

## 2.1 PROBLEMA DE NEGOCIO

Uno de los problemas más comunes dentro de la economía financiera, es la de poder

determinar con una gran precisión el comportamiento futuro de las diferentes monedas en el

mercado bursátil.

Las herramientas tecnológicas actuales disponen de nuevos recursos para proponer un

modelo que prediga la tendencia de las criptomonedas, apoyándose en los conceptos

clásicos de probabilidad y estadística en especial de series de tiempo como ARIMA, ARCH,

GARCH; esto es clave para el entendimiento, descripción y modelamiento del comportamiento de los datos.

Igualmente, se busca aprovechar el concepto de machine learning y Redes neuronales aplicadas en algunos modelos como, redes autorregresivas y Long Short-Term Memory (LSTM) que serán las técnicas utilizadas en esta monografía.

## 2.2 APROXIMACIÓN DESDE LA ANALÍTICA DE DATOS

Explica para qué servirían los modelos predictivos desarrollados (p.ej. Se desarrollarán modelos para recomendar productos a clientes)

Se desarrollará un modelo de serie de tiempo para predecir el comportamiento de las criptomonedas de interés, realizando un análisis exploratorio de los datos para determinar las características más relevantes que permitan modelar las tendencias en la variación de los precios de las monedas.

Lo anterior se hace aplicando la prueba de Dickey-Fuller para determinar la estacionariedad; además de descomponer la serie de tiempo en los componentes principales para hallar su temporalidad, aleatoriedad y tendencia.

Una vez determinadas las características de la serie, se busca aplicar un modelo de redes neuronales Long Short-Term Memory (LSTM), dado que las tendencias analizadas varían en cortos periodos de tiempo, en el caso de los movimientos de las criptomonedas, las variaciones son diarias. Finalmente se busca predecir el valor de las monedas al día siguiente, con el fin de determinar si para ese momento es beneficioso o no realizar una compra o venta de la divisa.

## 2.3 ORIGEN DE LOS DATOS

Explica qué representan los datos y cómo se obtuvieron. P.ej. tenemos datos clínicos e imágenes de pacientes de X patologías, obtenidos en tal hospital, en tales condiciones y durante tal periodo de tiempo. No describas aquí los datos, eso lo harás más adelante.

Los datos son extraídos de la página web <https://coinmarketcap.com/>. Por medio de la técnica de web Scraping se realiza una conexión para acceder a los datos en tiempo real y con ayuda de la extensión proporcionada por google chrome llamada Web Scraper se obtienen los datos históricos hasta la fecha de entrega de este documento. Los datos son almacenados en archivos .csv para cada una de las criptomonedas y están disponibles en un repositorio creado para el proyecto en la plataforma de github[[1]](#footnote-0)

## 2.4 MÉTRICAS DE DESEMPEÑO

Incluye

* las métricas de machine learning (desempeño de los modelos)
* las métricas de negocio (desempeño del proceso de negocio que usaría tus modelos)
* un juicio sobre qué valor mínimo de las métricas son necesarias en el contexto de aplicación
  + P.ej. a partir de x% de acierto (métrica de ML) , se estima un ROI de USD y (métrica de negocio) que es lo mínimo para cubrir la operación, despliegue, etc. de los modelos.

Las métricas de desempeño utilizadas para el tipo de problema a solucionar, que consiste en una serie de tiempo son:

● Error cuadrático medio (MSE, RMSE), mide la aproximación de los puntos de la curva de predicción y la curva de test.

● Error absoluto medio (MAE), es el promedio del valor absoluto de la diferencia entre los datos observados y los datos predichos, Al ser un valor absoluto todas las diferencias individuales se ponderan por igual en el promedio

El modelo debería asegurar un porcentaje de error de por lo menos un 20% , con lo cuál se pueda dar por aceptado el valor predicho.

# 3. DATOS

## 3.1 DATOS ORIGINALES

Describe los datos crudos de los que se dispone, incluyendo formatos, cómo se distribuyen los datos en ficheros, descripción de las columnas, formatos y resoluciones de imágenes, etiquetado de los mismos (para aprendizaje supervisado), tamaño de los datos (número de registros, tamaño total en MB o GB), etc. etc.

Describe también el modo de acceso a los datos y restricciones que pudiera haber sobre los mismos.

Para el análisis de tendencias de los valores de aperturas de las criptomonedas, que vamos

a realizar nos basamos en el movimiento actual del mercado que está contenido en la

Página web <https://coinmarketcap.com/>, cuya información se encuentra libre para cualquier usuario.

En este sitio vemos que las 5 criptomonedas más transadas son: Bitcoin, Etherium, Binance Coin y Tether y Cardano, que serán el objetivo de estudio de la presente monografía, para lo cual usaremos el registro histórico de estas monedas desde el inicio de su creación hasta noviembre de 2021.

Acontinuación se describen los campos asociados a cada dataset:

* Date: Dia de observación
* Open: Precio de apertura en el día de la observación en dólares
* High: El precio más alto obtenido en el día de la observación en dólares
* Low: El precio más bajo obtenido en el día de la observación en dólares
* Close: Precio de cierre en el día de la observación en dólares
* Volume: Volumen de transacciones realizadas en el día de la observación
* Market Cap: Valorización en dólares en el mercado, de la moneda el día la observación.

| **Moneda** | **Fecha Inicial** | **Cantidad Registros** | **Tamaño (KB)** |
| --- | --- | --- | --- |
| Bitcoin | 2013-04-29 | 3254 | 382 |
| Etherium | 2015-08-08 | 2423 | 280 |
| BinanceCoin | 2017-07-26 | 1705 | 186 |
| Tether | 2015-02-26 | 2586 | 265 |
| Cardano | 2017-10-02 | 1637 | 175 |

## Tabla 1 Características archivos Históricos para entrenamiento del modelo

## 3.2 DATASETS

Describe cómo, a partir de los datos, se construyen los datasets de entrenamiento y validación

Dentro de nuestro desarrollo se implementaron dos técnicas para la predicción del comportamiento de las tendencias, para lo cual se aplicaron diferentes datasets.

# AutoArima : Para este caso, debido a qué el modelo requería que se compararan entre sí las monedas para conocer su relación e incidencia entre ellas, se ajustó el dataset para que cada uno de ellos tuviese la misma fecha de inicio. Para este caso no se partieron los datos en entrenamiento ni testing

# Red Neuronal LSTM: Para este caso se usó un dataset con un 65% de los datos como entrenamiento y un 35% de los datos como test.

## 3.3 DESCRIPTIVA

Muestra algunos ejemplos de tu dataset (imágenes, registros), incluyendo una breve analítica descriptiva (distribuciones de datos, etc.)

| **SNo** | **Name** | **Symbol** | **Date** | **High** | **Low** | **Open** | **Close** | **Volume** | **Marketcap** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 | Bitcoin | BTC | 2013-04-29 23:59:59 | 147.488007 | 134.000000 | 134.444000 | 144.539993 | 0.0 | 1.603769e+09 |
| 2 | Bitcoin | BTC | 2013-04-30 23:59:59 | 146.929993 | 134.050003 | 144.000000 | 139.000000 | 0.0 | 1.542813e+09 |
| 3 | Bitcoin | BTC | 2013-05-01 23:59:59 | 139.889999 | 107.720001 | 139.000000 | 116.989998 | 0.0 | 1.298955e+09 |
| 4 | Bitcoin | BTC | 2013-05-02 23:59:59 | 125.599998 | 92.281898 | 116.379997 | 105.209999 | 0.0 | 1.168517e+09 |
| 5 | Bitcoin | BTC | 2013-05-03 23:59:59 | 108.127998 | 79.099998 | 106.250000 | 97.750000 | 0.0 | 1.085995e+09 |
| 6 | Bitcoin | BTC | 2013-05-04 23:59:59 | 115.000000 | 92.500000 | 98.099998 | 112.500000 | 0.0 | 1.250317e+09 |
| 7 | Bitcoin | BTC | 2013-05-05 23:59:59 | 118.800003 | 107.142998 | 112.900002 | 115.910004 | 0.0 | 1.288693e+09 |
| 8 | Bitcoin | BTC | 2013-05-06 23:59:59 | 124.663002 | 106.639999 | 115.980003 | 112.300003 | 0.0 | 1.249023e+09 |
| 9 | Bitcoin | BTC | 2013-05-07 23:59:59 | 113.444000 | 97.699997 | 112.250000 | 111.500000 | 0.0 | 1.240594e+09 |
| 10 | Bitcoin | BTC | 2013-05-08 23:59:59 | 115.779999 | 109.599998 | 109.599998 | 113.566002 | 0.0 | 1.264049e+09 |

Tabla 2: Muestra de los primeros 10 registros del dataset de bitcoin

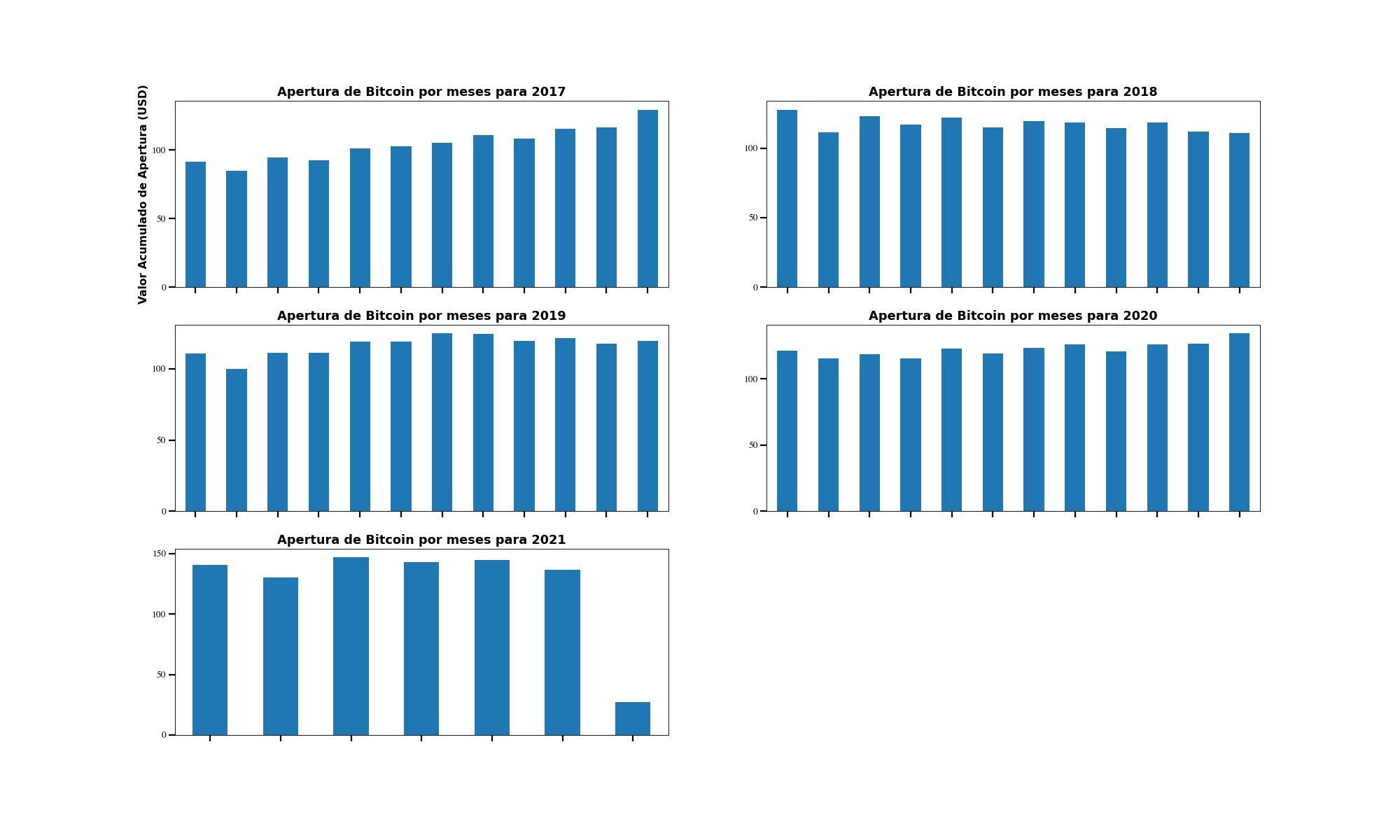


Figura 1: Histograma de la distribución del valor acumulado de apertura del bitcoin por mes, entre los años 2017-2021

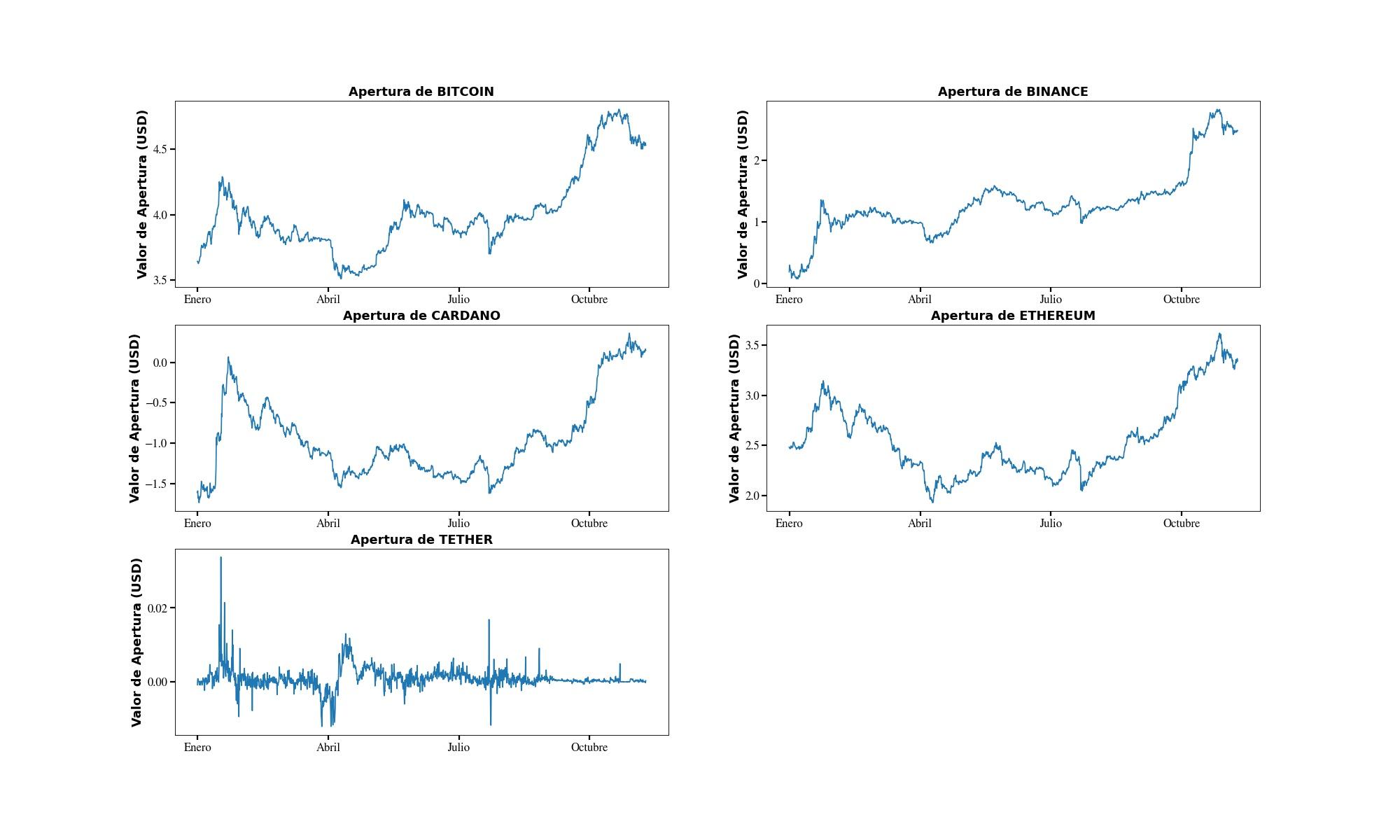


Figura 2 : Gráfica de tendencia de crecimiento de las 5 monedas más transadas en el portal sobre el valor acumulado de apertura para cada uno de los meses entre los años de creación y Julio 2021

# 

# 

# 

# 

# 

# 4. PROCESO DE ANALÍTICA

## 4.1 PIPELINE PRINCIPAL

Describe con un gráfico el flujo de trabajo general de los datos en tu proyecto. Añade o modifica secciones si lo consideras necesario

El proceso de los datos en el proyecto se lleva a cabo en dos partes, la primera consiste en crear y entrenar el modelo a partir de datos históricos, que han sido descargados en un repositorio de github, desde esa ubicación se leen los archivos y se cargan en un dataframe para hacer los procesos de transformación y limpieza de los datos pertinentes. Luego se analizan para determinar las características que se van a usar en el modelo, posteriormente se entrena y se crea el modelo que va a ser consumido por medio de herramientas de visualización, para mostrar la tendencia que se predice. La Figura a continuación esboza la arquitectura del proceso de construcción del modelo.

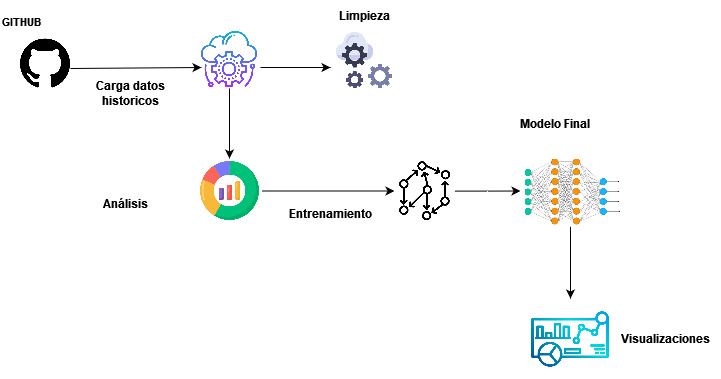


Figura 3: Arquitectura del procesamiento de datos para el modelo analítico.

## 4.2 PREPROCESAMIENTO

Describe las alternativas de preprocesamiento de datos que consideraste, aumentación de datos, etc.

Para el procesamiento de los datos se siguieron las siguientes estrategías:

1. Se homogeneizaron los datasets para que estuvieran en el mismo rango de fechas para hacer las comparaciones respectivas entre las diferentes monedas.
2. Se calcularon los logaritmos para cada una de tendencias de las monedas, con el fin de conocer las pendientes de crecimiento y decrecimiento para cada una de ellas.
3. Adicionalmente se calcula la primera derivada del logaritmo, con el fin de conocer si existe un comportamiento periodico.
4. Por último se calcula la estacionalidad del valor de apertura.

## 4.3 MODELOS

Describe los distintos modelos que consideraste, sus configuraciones, etc.

En un primer acercamiento de modelo, se utilizó un ARIMA, que es muy utilizado en Estadística y Econometría para la predicción de series temporales. Este modelo consiste en pronosticar basándose en el comportamiento de la variable en periodos anteriores dejando de lado las variables exógenas relacionadas al evento que se quiere predecir.

Para el modelado se usó la Función Autarima de la libreria pmdarima de Python con los siguientes parámetros:

auto\_arima(df\_Coins\_open\_copy['ret\_bitcoinOpen'][1:],

exogenous = df\_Coins\_open\_copy[['ret\_binanceOpen', 'ret\_cardanoOpen','ret\_ethereumOpen','ret\_tetherOpen']][1:],

m = 5,

max\_order = None,

max\_p = 7,

max\_q = 7,

max\_d = 1,

max\_P = 4,

max\_Q = 4,

max\_D = 1,

maxiter = 50,

alpha = 0.05,

n\_jobs = -1,

trend = 'ct',

information\_criterion = 'oob',

out\_of\_sample\_size = int(len(df\_Coins\_open\_copy)\*0.2))

# Modelo De Red Neuronal LSTM

model = Sequential()

model.add(LSTM(10, input\_shape=(1, look\_back)))

## 4.4 MÉTRICAS

Describe cómo calculas las métricas de desempeño ML y de negocio (p.ej. Con qué funciones de sklearn, tensorflow, etc.)

Para las métricas se usó la siguiente función de sklearn:

Error medio cuadrático:

model.compile(loss='mean\_squared\_error', optimizer='adam')

Graficas de análisis, resultados y comparaciones

# 5. METODOLOGÍA

## 5.1 BASELINE

Describe tu primera iteración, sus resultados y los problemas técnicos que tuviste que resolver

## 5.2 VALIDACIÓN

Describe tu proceso de validación, particiones de train/test/validación, etc.

## 5.3 ITERACIONES y EVOLUCIÓN

Describe en qué te enfocaste en las iteraciones subsiguientes. No es necesario hacer un listado exhaustivo de las iteraciones, pero sí has de describir qué perseguías y qué obtuviste con cada grupo de iteraciones. P.ej. primero hicimos un conjunto de iteraciones para definir las opciones de preprocesamiento, luego hicimos unas iteraciones probando distintos modelos de tal tipo, etc. etc.

## 5.4 HERRAMIENTAS

Menciona las herramientas que usaste para tu proyecto

# 6. RESULTADOS

## 6.1 MÉTRICAS

Describe los resultados numéricos de las métricas de las iteraciones que consideres más relevantes, junto con las configuraciones. Usa tablas o gráficas según consideres.

## 6.2 EVALUACIÓN CUALITATIVA

Realiza una evaluación cualitativa de los resultados, indicando casos de overfitting, underfitting, utilidad de los resultados, relación obtenida entre la métrica de ML y la de negocio, etc. etc.

## 6.3 CONSIDERACIONES DE PRODUCCIÓN

Menciona consideraciones técnicas de una posible puesta en producción. P.ej. condiciones de monitoreo del desempeño de los modelos, integración con streams de datos, servicios en la nube, etc.

Manuel API de las monedas: <https://coinmarketcap.com/api/documentation/v1/#section/Endpoint-Overview>

El consumo de los datos en el modelo entrenado para predecir la tendencia de las criptomonedas, se realizará con la técnica de web scraping para tener información en tiempo real del comportamiento de las divisas, así como se ilustra en la figura a continuación.

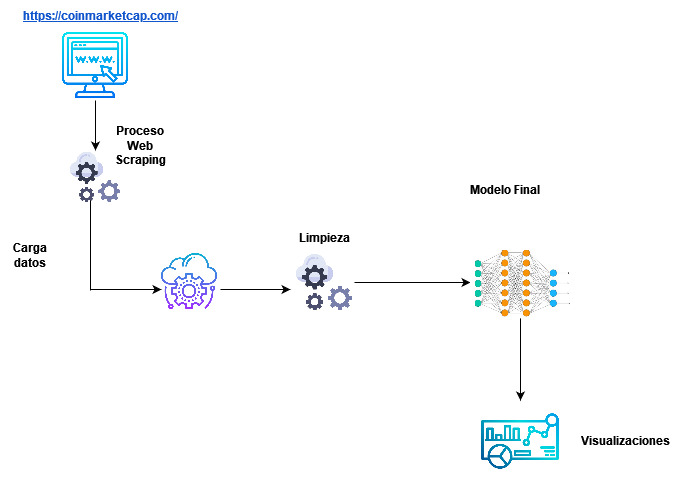


Figura 4: Arquitectura del proceso en tiempo real para predicción de las tendencias de las criptomonedas

# 7. CONCLUSIONES

Haz un breve juicio crítico de tu proyecto y del posible trabajo futuro que podría surgir a partir de tus resultados.

# 8. REFERENCIAS BIBLIOGRÁFICAS

Chávez Quisbert, Nicolás. (1997). MODELOS ARIMA. Revista Ciencia y Cultura, (1), 23-30. Recuperado de http://www.scielo.org.bo/scielo.php?script=sci\_arttext&pid=S2077-33231997000100005&lng=es&tlng=es.

Alcocer, U. M. R., Tello-Leal, E., & Alvarado, A. B. R. (2018). Modelo basado en redes neuronales recurrentes lstm para la predicción de la siguiente actividad en procesos de negocio (lstm recurrent neural network based-model for the prediction of the next activity in business processes). Pistas Educativas, 40(130).

Larios Meoño, J. F., & Alvarez, V. J. (2014). Análisis econométrico de series de tiempo. Universidad San Ignacio de Loyola.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press.

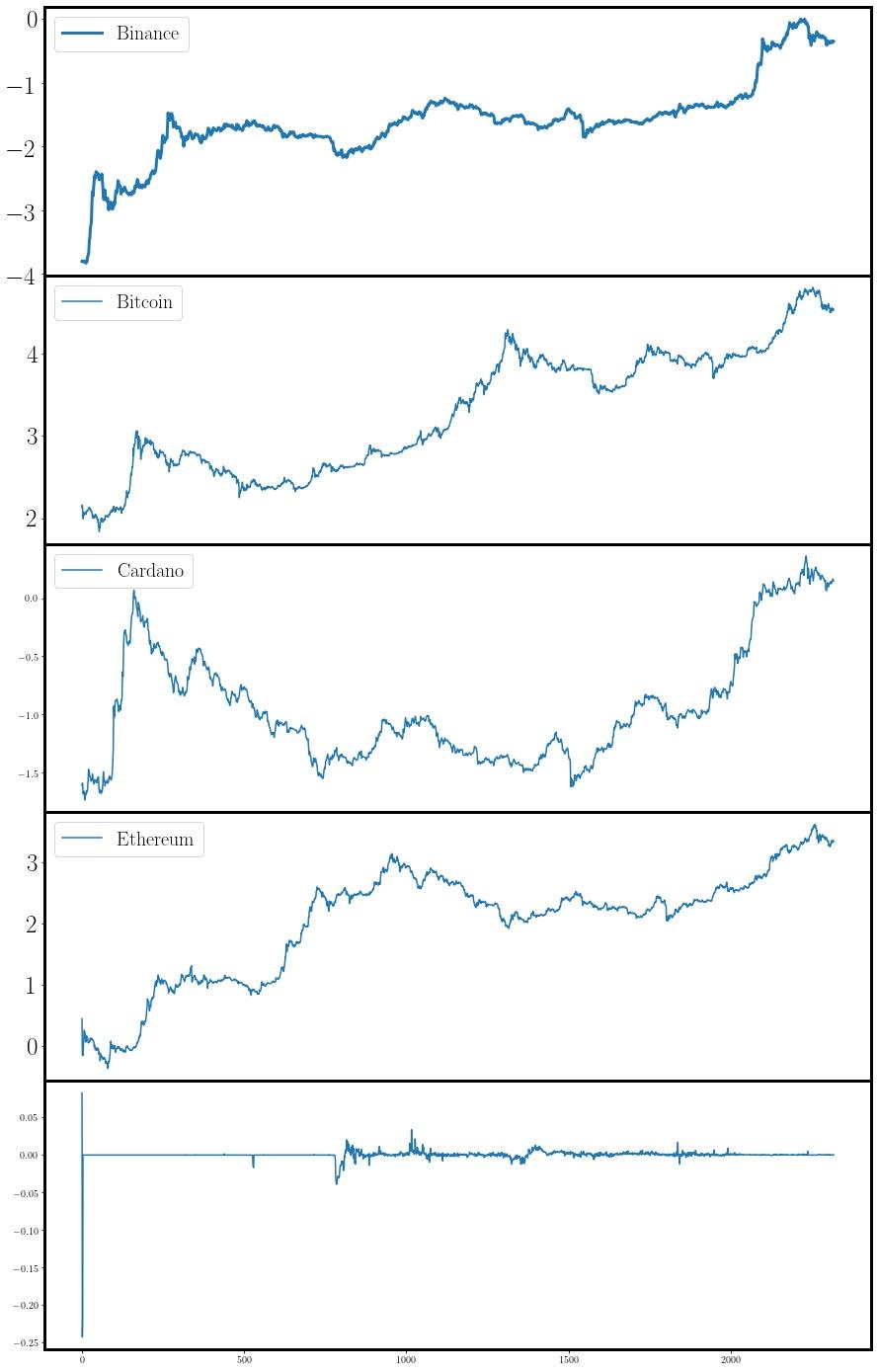
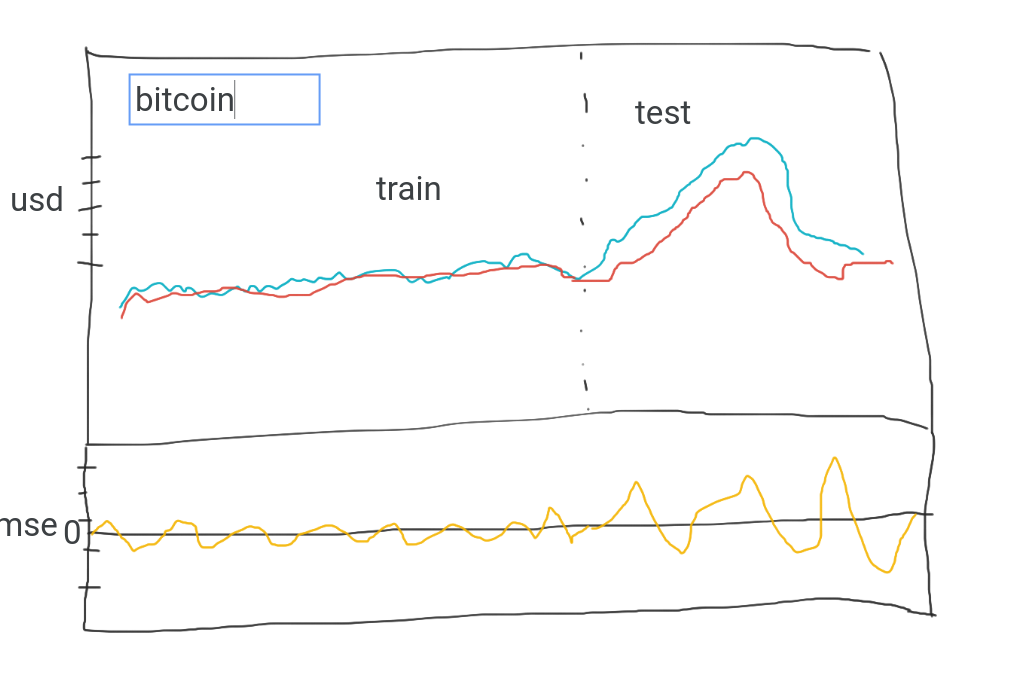


Figura:

Tareas:

* Organizar el repo con los datos, los codigos
* Guardar los modelos entrenados en un pkl

Hacer una grafica donde esten los datos, la prediccion del entrenamiento y el test. Y Separa la etapa de entrenamiento y el test con una linea vertical. Hacerlo para cada moneda.



1. <https://github.com/eherranudea/monografia_UDEA.git> [↑](#footnote-ref-0)