



## ANEXO 1

### FORMATO PARA LA PRESENTACIÓN DE PROYECTOS DE INVESTIGACIÓN CON EL FINANCIAMIENTO DEL FEDU

#### 1. Título del proyecto

**Análisis de sentimientos para determinar el comportamiento de criptodivisas usando modelos predictivos**

#### 2. Área de Investigación

Área de investigación	Línea de Investigación	Disciplina OCDE
Ingeniería y tecnología	Sistemas, computación e Informática	Inteligencia Artificial

#### 3. Duración del proyecto (meses)

12

#### 4. Tipo de proyecto

<b>Individual</b>	<input type="radio"/>
<b>Multidisciplinario</b>	<input checked="" type="radio"/>
<b>Director de tesis pregrado</b>	<input type="radio"/>

#### 4. Datos de los integrantes del proyecto

Apellidos y Nombres	<b>SOTOMAYOR ALZAMORA GUINA GUADALUPE</b>
Escuela Profesional	<b>INGENIERÍA DE SISTEMAS</b>
Código Docente	<b>2170502</b>
Celular	<b>953620511</b>
Correo Electrónico	<a href="mailto:gsotomayor@unap.edu.pe">gsotomayor@unap.edu.pe</a>

Apellidos y Nombres	<b>CONDORI ALEJO HENRY IVAN</b>
Escuela Profesional	<b>INGENIERÍA DE SISTEMAS</b>
Código Docente	<b>970478</b>
Celular	<b>958315508</b>
Correo Electrónico	<a href="mailto:hcondori@unap.edu.pe">hcondori@unap.edu.pe</a>

Apellidos y Nombres	<b>ACEITUNO ROJO MIGUEL ROMILIO</b>
Escuela Profesional	<b>INGENIERÍA DE SISTEMAS</b>
Código Docente	<b>2170845</b>
Celular	<b>952010600</b>
Correo Electrónico	<a href="mailto:maceituno@unap.edu.pe">maceituno@unap.edu.pe</a>



- I. **Título** (El proyecto de tesis debe llevar un título que exprese en forma sintética su contenido, haciendo referencia en lo posible, al resultado final que se pretende lograr. Máx. palabras 25)

Análisis de sentimientos aplicando modelos predictivos e IA para determinar el comportamiento del BITCOIN

- II. **Resumen del Proyecto de Tesis** (Debe ser suficientemente informativo, presentando -igual que un trabajo científico- una descripción de los principales puntos que se abordarán, objetivos, metodología y resultados que se esperan)

El presente proyecto tiene como objetivo proponer un modelo para predecir el comportamiento del precio de la criptomoneda Bitcoin, realizada a partir del análisis de sentimientos en publicaciones de Twitter, considerando también sus interacciones y variables de alcance. Se considerará la dificultad existente en la predicción de precio, fuertemente afectado con el sentimiento del inversor. La metodología abarca el realizar la recopilación de datos financieros históricos de BTC/USD y de la data histórica de Twitter relacionada al Bitcoin, especialmente publicaciones en inglés, esta data será pre-procesada para realizar la eliminación del ruido, estructuración y análisis de la data que se obtenga. Se aplicará el análisis de sentimientos de las publicaciones obtenidas usando un clasificador de sentimientos adecuado, para realizar el análisis de sentimientos en redes sociales, adaptando este al contexto del Bitcoin, con la finalidad de obtener clasificaciones más precisas, según las necesidades del proyecto. Con la data resultante se seleccionarán las variables más representativas, utilizando el índice de correlación de Spearman. Se implementará un modelo por corte de horas, considerando variables propias de Bitcoin y las que muestran expectativas e interacciones del entorno, obtenidas de las publicaciones de Twitter, además de variables de alcance y en combinación con algoritmos de aprendizaje supervisado de Deep Learning. Los modelos de predicción usarán redes neuronales recurrentes (RNN), y su validación servirá para medir su capacidad de predicción.

- III. **Palabras claves (Keywords)** (Colocadas en orden de importancia. Máx. palabras: cinco)

Análisis de sentimientos, Bitcoin, Deep Learning, Predicción, Twitter

- IV. **Justificación del proyecto** (Describa el problema y su relevancia como objeto de investigación. Es importante una clara definición y delimitación del problema que abordará la investigación, ya que temas cuya definición es difusa o amplísima son difíciles de evaluar y desarrollar)

Bitcoin es la criptomoneda con mayor capitalización en el mercado, aproximadamente 790,00 Billones de dólares (investing.com, 2021) y con mayor cantidad de comerciantes activos en el mundo, entre 51,2 y 52,4 millones en lo que respecta a criptomonedas (CH&Co., 2019), todo esto a partir del denominado "Boom del Bitcoin" del año 2017 y su posterior revalorización en el año 2021. Con esto se ve el crecimiento día a día de la comunidad de Bitcoin, la cual es extremadamente grande, a pesar de la caída del mercado el 2022. Sin embargo, el problema principal es la incertidumbre generada por la alta volatilidad del Bitcoin lo que afecta en gran medida la toma de decisiones financieras de sus inversionistas. La importancia de realizar este trabajo de investigación radica en 2 puntos importantes: Primero, se han realizado distintos trabajos de investigación relacionados con la influencia de los sentimientos en la toma de



decisiones por parte de los actores del mercado (Badiola Ramos, 2019) pero la gran mayoría se enfoca más sobre el cómo afecta la información de medios tradicionales o redes sociales a los mercados bursátiles tradicionales y aquellos pocos trabajos que se centran en la influencia de los sentimientos en el criptomercado difieren en sus resultados ya que algunos aseguran que sí existe una relación de influencia entre el sentimiento y la variación de las criptomonedas mientras que otros niegan la existencia de esta relación. Además, la mayoría de los estudios previos relacionados a este tema tuvieron muchas limitaciones como por ejemplo en la extracción de datos históricos, lo que implica la realización del análisis en franjas de tiempo muy pequeñas lo cual no es óptimo al momento de utilizar modelos predictivos con franjas de tiempo definidas. Segundo, el modelo predictivo a ser propuesto servirá de apoyo en la toma de decisiones financieras para los inversionistas de Bitcoin y permitirá la evaluación de la eficacia de las técnicas de IA aplicadas en este segmento.

**V. Antecedentes del proyecto (Incluya el estado actual del conocimiento en el ámbito nacional e internacional. La revisión bibliográfica debe incluir en lo posible artículos científicos actuales, para evidenciar el conocimiento existente y el aporte de la Tesis propuesta. Esto es importante para el futuro artículo que resultará como producto de este trabajo)**

Abraham et al. (2018), presentaron un método para predecir cambios en los precios de Bitcoin y Ethereum utilizando data de Twitter y Google Trends, para ello investigaron la relación entre el volumen y el sentimiento de los tweets que hablan sobre las criptomonedas Bitcoin y Ethereum, y el volumen de búsquedas en Google de ambas criptomonedas con la fluctuación del precio por tres meses en 2018. Descartaron la variabilidad del sentimiento al descubrir que no tenía una gran correlación con el precio. Los autores atribuyeron la causa a que el estudio fue realizado durante la caída del mercado, y otros estudios que sí encontraron relación lo hicieron en un momento de subida del mercado. Aun así, con el modelo lineal utilizado, que tomó como entrada tweets y data de Google Trends, fueron capaces de predecir acertadamente la dirección de los cambios de precio.

Awoke et al. (2020), compararon 2 modelos de predicción basados en Deep Learning: Long Short-Term Memory (LSTM) y Gated Recurrent Unit (GRU), con ello manejaron la volatilidad del precio del Bitcoin y obtuvieron una precisión alta.

Badiola (2019), estudió la posibilidad de predecir el precio del Bitcoin aplicando análisis de sentimientos en más de 17 millones de tweets sobre Bitcoin entre agosto de 2017 y enero de 2019. Partió de la hipótesis de que el sentimiento de los actores del mercado si tiene relación con los movimientos del precio de Bitcoin y que el volumen de los datos sería irrelevante. Tras el análisis completo de los datos obtenidos se observó que el sentimiento medio expresado en los mensajes se mantuvo positivo incluso cuando el valor del Bitcoin cayó. Esto indica que el sentimiento no es una buena variable para predecir el valor de Bitcoin. También observó que el volumen de los mensajes tiene una correlación alta y sólida durante todo el periodo, incluso cuando el precio cae. El autor creó tres modelos para predicciones de 1h, 4h y 24h en el futuro, la investigación consideró dos tipos de variables para crear los modelos de predicción, volumen de tweets y sentimiento medio de los Tweets. Los modelos presentaron un error cuadrático medio de 387, 483 y 818 respectivamente, pero algo interesante a resaltar es que la primera parte de la predicción resultó bastante certera, pero tras una gran caída el modelo tiende a predecir por encima de los valores reales, esto debido a que la caída fue un evento único y no estuvo presente en los datos de entrenamiento del modelo. Sin embargo, al incluir el entrenar el modelo con el evento único los resultados de predicción son mucho mejores con unos errores



cuadráticos medios de 244, 320, 600 para los modelos de 4h, 24h y 48h. Jiang (2020) analizó diferentes redes de aprendizaje profundo y diversos métodos para mejorar la precisión, incluida la normalización mínima y máxima, el optimizador Adam y la normalización mínima y máxima de Windows. Recopiló datos sobre el precio del Bitcoin por minuto y los reorganizó para reflejar el precio del Bitcoin en horas, haciendo un total de 56832 puntos. Además, consideró 24 horas de datos como entrada y salida del precio del Bitcoin de la siguiente hora, luego comparó los diferentes modelos y descubrió que la falta de memoria significa que el perceptrón multicapa (MLP) no es adecuado para el caso de predecir el precio según la tendencia actual, así mismo descubrió que LSTM proporciona relativamente la mejor predicción cuando la memoria pasada y la red recurrente cerrada (GRU) se incluyen en el modelo.

Li & Dai (2019) propusieron un modelo de red neuronal híbrida basada en una red neuronal convolucional (CNN) y una red neuronal LSTM. Consideraron como datos de entrada los datos de las transacciones de Bitcoin en sí, así como la información externa, las variables macroeconómicas y la atención de los inversores. Primero utilizaron CNN para la extracción de características. Luego, los vectores de características se ingresan en LSTM para entrenar y pronosticar el precio a corto plazo del Bitcoin. El resultado muestra que la red neuronal híbrida CNN-LSTM puede mejorar efectivamente la precisión de la predicción del valor y la predicción de dirección en comparación con la red neuronal de estructura única.

Pagolu et al. (2017), observaron los cambios en los precios de las acciones de una empresa, entre subidas y caídas, además están correlacionadas con las opiniones públicas expresadas en tweets sobre una empresa. El objetivo del análisis de sentimientos es comprender la opinión del autor a partir de un texto, por lo que se emplearon dos representaciones textuales diferentes, Word2vec y Ngram, para analizar los sentimientos del público en los tweets. Los autores aplicaron el análisis de sentimiento y principios de machine learning supervisado a los tweets extraídos de Twitter y analizaron la correlación entre los movimientos del mercado de valores de una empresa y sentimientos en tweets. Mostraron que, de forma elaborada, las noticias y tweets positivos sobre una empresa en las redes sociales definitivamente animan a la gente a invertir en las acciones de esa empresa y como resultado el precio de la acción de esa empresa aumentaría. Por lo que concluyeron que existe una fuerte correlación entre los aumentos y caídas de los precios de las acciones con los sentimientos del público en Twitter.

Regal et al. (2019) planteó la medida en que las publicaciones de redes sociales capturan las expectativas colectivas de los inversores para afectar el valor futuro de la moneda. Su objetivo fue pronosticar el desempeño diario de un mercado basándose en sus componentes: Definir el comportamiento de la criptomoneda (volumen, valores de: apertura, cierre, máximo y mínimo) y las expectativas e interacciones del entorno, obtenidas de los tweets recolectados. Propusieron el uso del tipo de red neuronal recurrente denominado "Long Short Term Memory" (LSTM). Usaron como metodología el preprocesamiento de los datos y la aplicación de esta técnica de pronóstico de series temporales les permite obtener una predicción con un Error Porcentual Absoluto Medio de 34.9%; implicando que la representación de la variable de percepción en redes sociales no ha sido la pertinente, y motiva nuevos trabajos con la finalidad de modelar dichas variables mediante el uso de otras técnicas de NLP.

Sin & Wang (2017), exploran la relación entre las características del Bitcoin y el cambio de precio en el Bitcoin al día siguiente utilizando un enfoque conjunto de RNA llamado AG basado en una red neuronal selectiva conjunta, construida usando un perceptrón multicapa como modelo base para cada RN del conjunto.

Stenqvist & Lönnö (2017) analizaron 2,27 millones de tweets relacionados con Bitcoin para las fluctuaciones de sentimiento que podrían indicar un cambio de precio en un futuro próximo. Usaron el método Naive de atribuir únicamente el



incremento o caída, según la gravedad de cambio de sentimiento de Twitter agregado durante periodos que oscilan entre 5 minutos y 4 horas, incrementando estas predicciones en el tiempo 1,2,3 o 4 periodos de tiempo para indicar el tiempo de intervalo BTC correspondiente. Evaluaron el modelo de predicción agregando sentimientos en tweets durante un periodo de 30 minutos con 4 cambios hacia adelante y un umbral de cambio de sentimiento de 2,2% con una precisión del 79%.

#### **VI. Hipótesis del trabajo** (Es el aporte proyectado de la investigación en la solución del problema)

El modelo predice asertivamente el comportamiento del precio de Bitcoin aplicando análisis de sentimientos en Twitter.

#### **VII. Objetivo general**

Proponer un Modelo predictivo basado en Análisis de sentimientos y machine learning para determinar el comportamiento del precio de BITCOIN

#### **VIII. Objetivos específicos**

- Recopilar y preprocesar la data histórica referida al Bitcoin procedente de Twitter.
- Analizar la data extraída de Twitter y aplicar el análisis de sentimientos.
- Implementar el modelo de predicción
- Evaluar el performance del modelo predictivo

#### **IX. Metodología de investigación** (Describir el(los) método(s) científico(s) que se empleará(n) para alcanzar los objetivos específicos, en forma coherente a la hipótesis de la investigación. Sustentar, con base bibliográfica, la pertinencia del(los) método(s) en términos de la representatividad de la muestra y de los resultados que se esperan alcanzar. Incluir los análisis estadísticos a utilizar)

El presente proyecto corresponde a data science, por lo que se aplicará la metodología CRISP-DM (Cross Industry Standard Process for Data Mining) ya que proporciona una descripción normalizada del ciclo de vida de un proyecto estándar de análisis de datos, tal metodología consta de seis fases, que son:

##### Fase 1. Business Understanding

La fase inicial se enfoca en la comprensión de los objetivos del proyecto. Con el conocimiento de los datos se define el problema de Minería de Datos y se realiza un plan preliminar para alcanzar los objetivos.

##### Fase 2. Data Understanding

La fase de entendimiento de datos inicia con la colección de datos inicial, continúa con las actividades de familiarización con los datos, identificación de los problemas de calidad, descubrimiento del conocimiento preliminar sobre los datos y el descubrimiento de subconjuntos interesantes para formar hipótesis mientras la información está oculta.

##### Fase 3. Data Preparation

La fase de preparación de datos cubre las actividades necesarias para construir un conjunto final de datos que serán utilizados en las herramientas de modelado, a partir de los datos en bruto iniciales. Las tareas incluyen la selección de tablas, registros y atributos, así como la transformación y la limpieza de datos para las herramientas de modelado

##### Fase 4. Modeling

La fase de modelado incluye la selección y aplicación de técnicas de modelado



pertinentes al problema (se buscarán las más adecuadas), y los parámetros serán calibrados a valores óptimos. Se evaluarán las técnicas para el mismo tipo de problema de minería de datos. Algunas técnicas tienen requerimientos específicos sobre la forma de los datos. Se espera que esta fase se convierta en la preparación de datos.

#### Fase 5. Evaluation

En esa fase los modelos a ser construidos serán evaluados para ver si alcanzan la calidad suficiente desde la perspectiva de análisis de datos. Al final de esta fase se debería obtener una decisión sobre la aplicación de los resultados del proceso de análisis de datos.

#### Fase 6. Deployment

Luego de la creación del modelo, se verificará si se ha incrementado el conocimiento de datos, que será organizado y presentado de tal forma que el cliente pueda usarlo. Esto puede darse desde la generación de un informe o tan compleja como la realización periódica y quizás automatizada de un proceso de análisis de datos en la organización.

## X. Referencias (Listar las citas bibliográficas con el estilo adecuado a su especialidad)

Abraham J., Higdon D., Nelson J. et al. (2018). Cryptocurrency Price Prediction Using Tweet Volumes and Sentiment Analysis. SMU Data Science Review: Vol. 1: No. 3, Article 1.

Awoke, T., Rout T., Mohanty L. et al. (2020). Bitcoin Price Prediction and Analysis Using Deep Learning Models. Lecture Notes in Networks and Systems. School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar, Odisha, India.

Badiola R., Jaime (2019). ¿Podemos comerciar Bitcoin usando análisis de sentimientos sobre Twitter? Universidad Pontificia de Comillas. Madrid – España.

Jiang, X.X. (2020) Bitcoin Price Prediction Based on Deep Learning Methods. Journal of Mathematical Finance, 10, 132-139.

Li Y. y Dai W. (2019). Bitcoin price forecasting method based on CNN-LSTM hybrid neural network model. The 3rd Asian Conference on Artificial Intelligence Technology (ACAIT). Department of Financial Engineering, Central University of Finance and Economics, Beijing, People's Republic of China, pp. 344-347

Pagolu V., Reddy K., Panda G. et al. (2017). Sentiment analysis of Twitter data for predicting stock market movements. International Conference on Signal Processing, Communication, Power and Embedded System, SCOPES 2016 – Proceedings, India.

Regal A., Morzán J., Fabbri C. et al. (2019). Proyección del precio de criptomonedas basado en Tweets empleando LSTM. Ingeniare. Revista Chilena de Ingeniería, vol. 27 N° 4, pp. 696-706.

Sin E. y Wang L. (2018). Bitcoin Price prediction using ensembles of neural networks. ICNC-FSKD 2017 – 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. Nanyang Technological University Singapore.

Stenqvist, E., & Lönnö, J. (2017). Predicting Bitcoin price fluctuation with Twitter sentiment analysis. KTH Royal Institute of Technology – School of Computer Science and Communication, Stockholm, Sweden.

CH&Co. (2019). How many active crypto traders are there across the globe? investing.com. (20 de septiembre de 2021). investing.com. Obtenido de <https://es.investing.com/crypto/bitcoin/historical-data>.



**XI. Uso de los resultados y contribuciones del proyecto (Señalar el posible uso de los resultados y la contribución de los mismos)**

Permitirá contar una definición de eficiencia del modelo predictivo a partir de tweets, se pueda predecir el comportamiento del precio de BITCOIN.

**XII. Impactos esperados**

**i. Impactos en Ciencia y Tecnología**

Plantea una propuesta teórica de los modelos aplicables a predicción de precios de BITCOIN basado en machine learning.

**ii. Impactos económicos**

La evaluación de modelos permitirá determinar el modelo de machine learning aplicable a predicción del precio de BITCOIN con la correspondiente reducción de incertidumbre.

**iii. Impactos sociales**

Mejorar el proceso de uso de alternativas de dinero hacia los criptoactivos y su posible inversión.

**iv. Impactos ambientales**

No genera impacto o costo ambiental

**XIII. Recursos necesarios (Infraestructura, equipos y principales tecnologías en uso relacionadas con la temática del proyecto, señale medios y recursos para realizar el proyecto)**

Laptop  
Internet  
Google Colab  
Libreta de apuntes  
API de Twitter  
Material bibliográfico  
Machine Learning  
Natural Language Processing (NLP)  
Sentiment Analysis  
Teoría de Grafos  
Kaggle

**XIV. Localización del proyecto (indicar donde se llevará a cabo el proyecto)**

A principio los tweets a analizar serán extraídos de la plataforma gratuita Kaggle que dispone a los usuarios un conjunto de datasets sobre distintos temas, en este caso se utilizará el dataset "Bitcoin tweets - 16M tweets" que contiene tweets desde 01-01-2016 hasta 29-03-2019.

**XV. Cronograma de actividades**



Actividad	Meses											
	1	2	3	4	5	6	7	8	9	10	11	12
Selección de antecedentes Documentarios	X	X										
Identificación del Problema y Formulación de los Objetivos			X									
Antecedentes Bibliográficos			X	X	X							
Análisis de los modelos			X	X								
Caracterización general de las variables			X									
Construcción de modelos y análisis				X	X	X	X	X	X	X		
Evaluación de modelos										X	X	
Desarrollo del Informe											X	X
Presentación												X

## XVI. Presupuesto

Descripción	Unidad de medida	Costo Unitario (S/.)	Cantidad	Costo total (S/.)
Bibliografía e información	Unidad	300	10	3000
Almacenamiento	Unidad	30	6	180
Útiles de escritorio	Unidad	50	3	150
Capacitación	Unidad	1000	3	3000
Impresiones	Hojas	0.20	250	50
Digitación	Hoja	1.00	100	100
Servicio de Internet	Unidad	140	3	420
Viáticos	Unidad	500	3	1500
Movilidad	Unidad	200	3	600
Análisis, Construcción, Implementación y ejecución de bases y Pruebas	Unidad	5000	1	5000
			Sub Total	14000
			Imprevistos (10%)	140
			Total	14140