FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO

ESCUELA PROFESIONAL DE INGENIERÍA DE SISTEMAS

TESIS

Comparación de técnicas de estimación basadas en machine learning para predecir costos en los planes de adquisiciones de las entidades públicas del Perú

PARA OPTAR EL TÍTULO PROFESIONAL DE INGENIERO DE SISTEMAS

Autor

Bach. Chang Hidalgo, Haward Miguel

ORCID: https://orcid.org/0000-0002-7260-0983

Asesor

Mg. Chirinos Mundaca, Carlos Alberto

ORCID: https://orcid.org/0000-0002-6733-8992

Línea de Investigación

Infraestructura, Tecnología y Medio Ambiente

Pimentel – Perú

2023

COMPARACIÓN DE TÉCNICAS DE ESTIMACIÓN BASADAS DE MACHINE LEARNING PARA PREDECIR COSTOS EN LOS PLANES DE ADQUISICIONES DE LAS ENTIDADES PÚBLICAS DEL PERÚ

Aprobación del jurado

_____

MG. MEJIA CABRERA HEBER IVAN

Presidente del Jurado de Tesis

_____

DR. VASQUEZ LEYVA OLIVER

Secretario del Jurado de Tesis

_____

MG. BRAVO RUIZ JAIME ARTURO

Vocal del Jurado de Tesis

DECLARACIÓN JURADA DE ORIGINALIDAD

Quien(es) suscribe(imos) la **DECLARACIÓN JURADA**, soy(somos) Chang Hidalgo, Haward Miguel del Programa de Estudios de Ingeniería de Sistemas de la Universidad Señor de Sipán S.A.C, declaro (amos) bajo juramento que soy (somos) autor(es) del trabajo titulado:

COMPARACIÓN DE TÉCNICAS DE ESTIMACIÓN BASADAS DE MACHINE LEARNING PARA PREDECIR COSTOS EN LOS PLANES DE ADQUISICIONES DE LAS ENTIDADES PÚBLICAS DEL PERÚ

El texto de mi trabajo de investigación responde y respeta lo indicado en el Código de Ética del Comité Institucional de Ética en Investigación de la Universidad Señor de Sipán (CIEI USS) conforme a los principios y lineamientos detallados en dicho documento, en relación con las citas y referencias bibliográficas, respetando al derecho de propiedad intelectual, por lo cual informo que la investigación cumple con ser inédito, original y autentico.

En virtud de lo antes mencionado, firma:

Chang Hidalgo, Haward Miguel	DNI: 10868987	

Pimentel, 01 de febrero de 2023.

Dedicatoria

A “Yavé, Dios Padre de toda la Humanidad, por la dicha de contar con una excelente familia y de los mejores amigos.

A mis padres **Segundo Chang y Carmen Hidalgo,** por estar a mi lado en cada paso de mi vida y por hacer de mi persona un hombre de bien.

A mi hermano **Segundo** y mi cuñada **Cinthia**, por mantener la familia unida.

A mis sobrinos / hijos: **Kennie, Brian, Piero y Ángelo**, quienes me impulsan a ser mejor persona.

Agradecimientos

Agradezco a los maestros de la Escuela Profesional de Ingeniería de Sistemas de la Universidad Señor de Sipan – USS, por contribuir en mi desarrollo profesional y por transmitir en cada uno de sus alumnos la importancia de la práctica de valores.

De manera especial agradezco a mis autoridades y maestros Víctor Tuesta Monteza y Heber Mejía Cabrera, quienes nos inculcaron que “no basta con hacer un buen trabajo”, por el contrario “debemos hacer el mejor trabajo”.

ÍNDICE

Índice de ilustraciones

Índice de tablas

Resumen

Las Entidades Públicas, cada año tienen la obligación de estimar los presupuestos (en adelante costos), que permitirán la atención de las necesidades de bienes y servicios que deben estar consignados en los planes de adquisiciones. Para tal fin, el Ministerio de Economía y Finanzas (en adelante MEF) tiene la difícil tarea de revisar, evaluar, aprobar y asignar los recursos que estime conveniente, respecto de las propuestas presentadas por cada entidad pública. La problemática presentada, nos lleva a la necesidad de identificar una técnica de aprendizaje automático (Machine Learning), que permita facilitar el proceso de predicción de los costos, con la finalidad de financiar los planes de adquisiciones de las diversas Entidades Públicas del Perú, asimismo, a fin de resolver el problema expuesto, se determinó el siguiente método de trabajo, iniciando con la identificación del dataset de las contrataciones públicas, obtenido del portal web de datos abiertos de la Organismo Supervisor de Contrataciones Estatales – OSCE, en segundo lugar se procedió con la revisión de la literatura de artículos científicos que se relacionen con la presente investigación para identificar los algoritmos más utilizados y los resultados obtenidos para su implementación, en tercer lugar se priorizo la implementación de los siguientes algoritmos de Regresión: a) Lineal Múltiple, b) Arból de decisión, c) Bosque de aleatorio y d) Xgboots; y en cuarto lugar se llevaron a cabo las pruebas del desempeño de las técnicas implementadas, obteniéndose los siguientes resultados: en primer lugar, el modelo Regresión Lineal Múltiple con los siguientes índices de error MAE=4.03E+06, MAPE=0.30.MSE=4.04E+13, RMSE=6.36E+06 y un R2 = 0.79587, en segundo lugar, Random Forest con índices de MAE=5.43E+06, MAPE=0.33, MSE=6.20E+12, RMSE=7.88E+06 y un R2 = 0.68666, en tercer lugar, XGboost con índices de MAE=5.97E+06, MAPE=0.34, MSE=7.59E+13, RMSE=8.71E+06 y R2 = 0.61649 y en cuarto lugar, Árbol de Decisiones con índices de MAE=6.16E+06, MAPE=0.40, MSE=1.03E+14, RMSE=1.03E+14 y R2 = 0.53162, concluyéndose que el mejor desempeño lo obtuvo el algoritmo de regresión lineal múltiple.

Palabras Clave: Aprendizaje automático, predicción, costos, regresión lineal múltiple, árbol de decisiones, bosque de aleatorio y XGboost.

Abstract

The Public Entities, each year have the obligation to estimate the budgets (hereinafter costs), which will allow the attention of the needs of goods and services that must be consigned in the acquisition plans. For this, the Ministry of Economy and Finance (hereinafter MEF) has the difficult task of reviewing, evaluating, approving and allocating the resources it deems appropriate, regarding the proposals presented by each public entity. The problem presented, leads us to the need to identify an automatic learning technique (Machine Learning), which facilitates the cost prediction process, in order to finance the acquisition plans of the various Public Entities of Peru, as well as , in order to solve the exposed problem, the following work method was determined, starting with the identification of the dataset of public procurement, obtained from the open data web portal of the Supervisory Body for State Procurement - OSCE, secondly, With the review of the literature of scientific articles that are related to the present investigation to identify the most used algorithms and the results obtained for their implementation, in third place the implementation of the following Regression algorithms was prioritized: a) Linear Multiple, b) Decision Tree, c) Random Forest and d) Xgboots; and fourthly, the performance tests of the implemented techniques were carried out, obtaining the following results: firstly, the Multiple Linear Regression model with the following error rates MAE=4.03E+06, MAPE=0.30.MSE= 4.04E+13, RMSE=6.36E+06 and an R2 = 0.79587, in second place, Random Forest with indices of MAE=5.43E+06, MAPE=0.33, MSE=6.20E+12, RMSE=7.88E+06 and an R2 = 0.68666, in third place, XGboost with indexes of MAE=5.97E+06, MAPE=0.34, MSE=7.59E+13, RMSE=8.71E+06 and R2 = 0.61649 and in fourth place, Decision Tree with indexes of MAE=6.16E+06, MAPE=0.40, MSE=1.03E+14, RMSE=1.03E+14 and R2 = 0.53162, concluding that the best performance was obtained by the multiple linear regression algorithm.

Keywords: Machine learning, prediction, costs, multiple linear regression, decision tree, random forest and XGboost.

I. INTRODUCCIÓN

1.1. Realidad Problemática.

Cada año, el Gobierno Peruano, a través del Ministerio de Economía, tiene la tarea de elaborar la propuesta del presupuesto para atender diversas categorías de gastos como son: Adquisición de Activos No Financieros, Planillas de Personal, Bienes y Servicios, Servicio de la Deuda Pública, entre otros.

De las principales categorías de gasto, la que presenta mayor complejidad en su estimación es la de Bienes y Servicios, las cuales, se elaboran en el marco de un presupuesto asignado, estableciéndose en ocasiones la programación de valores estimados menores o mayores al promedio del mercado, esto en razón a no contar con un estudio previo de las necesitades a convocar, o de un procedimiento para pronosticar el valor estimado que se aproximen a los precios del mercado.

En el portal web de la Plataforma Nacional de Datos Abiertos de la Secretaría de Gobierno Digital de la PCM, se aprecia que, alrededor de 3080 entidades públicas periódicamente contratan bienes y servicios para cumplir con los fines institucionales, según el Organismo Supervisor de las Contrataciones del Estado [OSCE] (2021), al 31 de diciembre de 2021, se han realizado más 35,000 procesos de selección en el marco de los procedimientos normados por la vigente Ley de Contrataciones del Estado y su respectivo reglamento.

Según la nota periodística del diario La Gestión (2021), con el título: “Fiscalía investiga 15 denuncias a nivel nacional por compras sobrevaloradas para la PNP”, señala que la Fiscalía Anticorrupción viene investigando alrededor de 15 denuncias en diferentes regiones del ámbito nacional, respecto a presuntos hechos irregulares en la sobrevaloración en compras de diversos productos sanitarios por parte de la Policía Nacional del Perú (PNP).

Asimismo, en el sector salud, se presenta similar situación, con la nota periodística del diario El Comercio, (Medrano, 2021), con el título “Essalud: tomógrafos, camas UCI y otros equipos que se pudieron comprar con dinero sobrevalorado”, manifiesta que el Ministerio Público en colaboración con la División de investigación de delitos de Alta Complejidad (Diviac) vienen realizado la investigación por presuntas compras sobrevaloradas de equipos de tomografía, camas ucis, y otros, por un monto total de 28,9 millones de soles.

Los problemas descritos de manera general denotan deficiencia en el proceso de predicción para cálculo, asignación y control del uso de los presupuestos, situaciones que ya han venido siendo investigados por los siguientes autores:

Truong et al. (2020), mencionan que, para poder adquirir una vivienda, esta se condiciona a otros factores como la ubicación, el área, la población, número de habitaciones entre otra información, con lo cual se hace necesario identificar una forma de predecir el precio promedio de las viviendas. El autor menciona la existencia de artículos consultados que adoptan enfoques tradicionales de aprendizaje automático para predecir los precios de la vivienda con precisión, pero rara vez se preocupan por el rendimiento de modelos individuales y descuidan los modelos menos populares pero complejos.

Deepa et al. (2021), comentan que, la agricultura se enfrenta a un crecimiento descendente, debido a la menor cantidad de lluvia, no se encuentran obteniendo el rendimiento adecuado y tampoco el precio esperado para sus cultivos conllevándolos a menos ganancias y pérdidas. Requiriéndose la necesidad de la predicción del precio de la producción de algodón, lo cual facilitaría la decisión de seleccionar en que área geográfica llevar a cabo el cultivo.

Zhang et al. (2021), mencionan que, las complicadas fluctuaciones de los precios del cobre pueden afectar significativamente a otras industrias. Por lo

tanto, se debe proponer varios modelos de pronóstico para predecir los precios mensuales del cobre con algoritmos en aprendizaje automático.

Asimismo, los mencionados problemas de ingeniería fueron abordados con las siguientes soluciones:

Para el problema de ingeniería para la predicción del precio de las viviendas (Truong et al., 2020) se consideraron atributos específicos, al aplicarse los métodos de clasificación, presentando mejor desempeño la **Regresión Generalizada Apilada,** asimismo, para predecir el precio del algodón (Deepa et al., 2021) se utilizó la comparación de diversos algoritmos de regresión, presentando el mejor desempeño el de **Árbol de Decisiones potenciada o impulsada,** de igual manera, para pronosticar el precio mensual del cobre (Zhang et al., 2021), se compararon técnicas de aprendizaje automático, mostrándose **la red neuronal (con técnicas de aprendizaje profundo)** como el mejor método.

Por todo lo mencionado como realidad problemática, existe la necesidad de conocer los valores estimados de las contrataciones públicas, de manera que, al contarse con información histórica de las contrataciones, además del uso de las tecnologías de la información, resulta indispensable la comparación de técnicas de estimación de aprendizaje automático que permitan predecir los valores que se aproximen con mayor exactitud y que sirvan de soporte para establecer la demanda del presupuesto (recursos) necesarios al momento de programar los planes anuales de contrataciones de las entidades públicas del Estado Peruano, contribuyéndose de ese modo, con el aseguramiento de la calidad del gasto público.

1.2. Trabajos previos.

Como antecedente del presente trabajo se han citado investigaciones, mediante las cuales diversos autores comparten sus experiencias a través de los siguientes artículos científicos:

Deepa et al. (2021), en la investigación, Machine learning regression model for material synthesis prices prediction in agriculture, en India, plantea el problema ¿cómo predecir el precio del algodón producido en la India, mediante algoritmos de regresión de aprendizaje automático, para saber si los agricultores deben cultivar a no?, para ello, proponer como método de solución, diferentes tipos de algoritmos de regresión para la predicción de precios, tales como: Regresión Lineal, Regresión Lineal bayesiana, Regresión de Árbol de Decisiones Potenciada y Regresión del Bosque de Aleatorio. Teniendo como resultados que, al comparar los modelos de regresión, se identifica con mayor desempeño: En primer lugar al modelo Regresión de Árbol de Decisión potenciada - BDR - (con el mejor coeficiente de determinación – R2- por una valor 0.80 y menor RMSE de 242.33 y RSE de 0,194 respectivamente), en segundo lugar el modelo Regresión Lineal - LR (con el 2do mejor coeficiente de determinación -R2 - con 0.79966 y menor MAE de 120.01 y RAE de 0.353 respectivamente), en tercer lugar el modelo Regresión Lineal bayesiana - BLR - (con el 3er mejor coeficiente de determinación -R2- con 0.79965). concluyendo que, el uso de los modelos de predicción de aprendizaje automático fue de utilidad para la actividad de los agricultores.

Zhang et al. (2021), en la investigación, Forecasting monthly copper price: A comparative study of various machine learning-based methods, en Estados Unidos, plantea el problema ¿cómo predecir los precios mensuales del cobre?, para lo cual plantea como solución, el impulso de algoritmos de aprendizaje automático, como el perceptrón multicapa (MLP), k vecinos más próximos (KNN), máquinas de vector de soporte (SVM), árbol de aumento de gradiente (GBT) y bosque aleatorio (RF), aplicados a un conjunto de datos mensuales de precios del cobre, desde el año 1990 al 2019. Los resultados revelaron que en los países que tienen la producción más abundante de cobre en el mundo, presentan un efecto significativo en la volatilidad de los precios mensuales, asimismo al pronosticar los precios mensuales del cobre en el futuro, mediante comparación de varias técnicas de aprendizaje automático se muestra que la red neuronal MLP (con técnicas de aprendizaje profundo)

es el mejor método para pronosticar el precio mensual del cobre con un MAE de 228.617 y RMSE de 287.539. Considerando que, los otros modelos, como RF, SVM, GBT, KNN, proporcionaron errores más altos con un MAE en el rango de 308.691 a 453.147 y con una RMSE en el rango de 393.599 a 552.208. Concluyéndose que, la red neuronal MLP (con técnicas de aprendizaje profundo) es el mejor método para pronosticar el precio mensual del cobre, siendo la herramienta más confiable para pronosticar los precios del cobre en el futuro.

Chen et al. (2020), en la investigación, Bitcoin price prediction using machine learning: An approach to sample dimension engineering, en China, plantea el problema de, ¿cómo encontrar un método que pueda utilizar con precisión algoritmos de aprendizaje automático para predecir el precio de Bitcoin?; al respecto los autores proponen como método de solución el aprovechamiento de técnicas de aprendizaje automático como LR, LDA, XGB, SVM para diseñar dimensiones de muestra para la predicción de precios de Bitcoin, inspirados en el principio de la navaja de Cocan y las características de conjuntos de datos. De los resultados obtenidos, los dos métodos estadísticos son mejores en general, La precisión media de los métodos estadísticos (LR y LDA) es del 65,0%, superior a la precisión media de los modelos de aprendizaje automático (55,3%). El modelo LR obtuvo los mejores resultados, con una precisión del 66,0%. Entre los modelos de aprendizaje automático, XGB tuvo el peor desempeño, con una precisión del 48,3%, y SVM fue el mejor, con una precisión del 65,3%, competitivo con los métodos estadísticos. En general, LR y LDA superaron a los otros modelos de aprendizaje automático en el conjunto de datos de precios diarios, lo que indica que los conjuntos de características de alta dimensión correctamente seleccionados pueden compensar la simplicidad de los modelos en la predicción de precios diarios de Bitcoin. Finalmente, concluye que, si bien la mayoría de los trabajos anteriores simplemente aprovechan los algoritmos de aprendizaje automático en la predicción de precios de Bitcoin, en la presente investigación se muestra que se debe considerar la granularidad y las dimensiones de las características de la muestra, donde el precio diario agregado de Bitcoin,

adquirido de CoinMarketCap, facilita la inclusión de características de alta dimensión.

Rico & Taltavull. (2021), en la investigación, Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain, en España, plantea el problema de, ¿Qué modelo hedónico de aprendizaje automático tiene mejor precisión en el pronóstico de precios de viviendas?, proponiendo como método de solución, el estudio comparativo de dos modelos hedónicos y una selección de modelos de aprendizaje automático y algoritmos tales como: Vecinos más cercanos, Árbol de decisión, Bosque Aleatorio, XGBosst, AdaBoost, CatBoost, Red Neuronal (percepción multicapa), Regresión Lineal, Cresta Lineal, Lazo Lineal. De los resultados numéricos según la precisión, muestran al algoritmo de Bosque Aleatorio con un promedio de precisión de 0,970, seguido en segundo lugar por el algoritmo Vecino más cercanos con un valor de 0.924, Catboost con 0.880, XGB con 0.850, MLP con 0.845, Regresión y Cresta Lineal con 0.800; árbol de decisión con 0.766, Adaboost con 0.728 y Lazo Lineal con 0.510. Finalmente, concluye que la presente investigación muestra los resultados de varios modelos predictivos que se basan en dos metodologías diferentes: aprendizaje automático y regresiones hedónicas. Se probaron un total de siete modelos diferentes para determinar su capacidad predictiva, utilizando una gran base de datos micro de precios de vivienda (solicitados) en la ciudad de Alicante, España, para el período 1996-2012.

Raditya et al. (2021), en la investigación, Predicting Sneaker Resale Prices using Machine Learning, en Indonesia, plantea el problema de, ¿cómo predecir los precios de reventa de zapatillas?, proponiendo como método de solución, la regresión lineal y el bosque aleatorio para hacer las predicciones utilizando los datos del historial de ventas de zapatillas recopilados de Stock X, del concurso de datos Stock X de 2019 que consta de 99,956 ventas de zapatillas (con 52 zapatillas diferentes), también se agregaron 150 datos adicionales debido a la falta de variación de datos. De los resultados obtenidos se aprecia que, en la prueba inicial, la regresión lineal con un valor de 0.9286

de Prueba R funciona mejor que el bosque aleatorio con un valor de 0.9234, pero cuando se pone en juego la validación cruzada, el bosque aleatorio tiene puntuaciones más altas, con un valor de 0.9330 y la regresión lineal con un valor de 0.9060. Esto también afecta el MSE para ambos modelos, especialmente el bosque aleatorio, ya que podemos ver que el MSE para el bosque aleatorio disminuyó en 3000. Cuando se usa Landon Forest, el tren R2 es ligeramente más alto que la prueba R2, lo que indica un poco de sobreajuste. Después de realizar la validación cruzada, el puntaje R2 es más alto y el MSE es más bajo, por lo que podríamos corregir el sobreajuste mediante la validación cruzada. Sin embargo, con la regresión lineal, el R2 disminuyó. Finalmente, concluye que, el modelo Random Forest con 10 pliegues de validación cruzada funciona mejor en comparación con el modelo de regresión lineal en la predicción de precios de reventa de zapatillas.

Deina et al. (2021), en la investigación, A methodology for coffee price forecasting based on extreme learning machine, en Brasil, plantea el problema de, a través de que metodología de aprendizaje automático se podrá estimar los precios del café, proponiendo como método abordar los siguientes modelos: Suavizado exponencial (ES), Autorregresivo (AR) y modelos de media móvil e integrada autorregresiva (ARIMA), multicapa Redes neuronales Perceptron (MLP) y Extreme Learning Machines (ELM). Obteniendo como resultados, en cuanto al precio del café Arábica, es claro que nuestra propuesta de utilizar el ELM obtuvo mejores rendimientos para todos los horizontes y métricas, superando el MLP tradicional. Con respecto al tipo Robusta, hubo casi un empate entre ELM y MLP. El primero destaca por PAG = 1 y 3 (excepto MSE), mientras que el MLP destaca para los demás casos, En términos del general rendimiento de los modelos, el ELM superó a los otros modelos en 17 de 24 escenarios (71%). Finalmente, concluye que los resultados computacionales mostraron que la metodología que utilizaba ELM era capaz de superar todos los procedimientos lineales, y fue mejor que el MLP en el 71% de los casos

Díaz et al. (2019), en la investigación, Predicción and explanation of the formation of the Spanish day-ahead electricity price through machine learning regression, en España, plantea el problema de, ¿Cómo predecir el precio diario de la electricidad a través del aprendizaje automático?, aplicando como método de solución para la predicción, se especifica un modelo de regresión por cuantiles basado en árboles de regresión potenciados por gradientes, asimismo, se propone utilizar modelos de regresión, tales como: Regresión lineal múltiple (MLR), regresión sobre componentes principales (PCR) y Árboles de regresión de aumento de gradiente. De los resultados, respecto a la precisión de los modelos PCR y GBRT de percentil 50 en función de la hora que se realiza la predicción, se obtiene que las predicciones menos precisas fueron las realizadas a partir de las 18:00 horas, donde los valores del RMSE y MAE para el PCR fueron de 16.12 y 1.15, y para el GBRT de 2.10 y 1.15 respectivamente, donde el modelo de PCR es visiblemente peor que el modelo de GBRT. Asimismo, las métricas calculadas muestran que el modelo de la presente investigación produce errores de predicción notablemente bajos cuando se utiliza la mediana como método de predicción puntual, con valores para el RMSE con 2,78 € / MWh y MAE con un valor de 1,94 € / MWh, y por último MAPE con un valor de 0,059. Concluyendo que, el modelo GBRT del percentil 50 demuestra ser un método de predicción puntual.

Fang & Taylor. (2021), en la investigación, A machine Lear Ning based asset pricing factor model comparison on anomaly portfolios, en EE.UU, plantea el problema de, como determinar el mejor modelo de factores de precios de activos a través del aprendizaje automático, como método hace una comparación completa entre el rendimiento de los modelos de factores lineales y SVM, redes neuronales, modelos lineales regularizados y modelos basados en árboles. Obteniendo los resultados que las ligeras modificaciones de la regresión lineal, como una versión regularizada, elastic-net y su extensión impulsada, generalmente tienen un mayor rendimiento predictivo en las carteras de anomalías. Los modelos más complejos, como el bosque aleatorio, el árbol impulsado por gradiente y los predictores basados en redes neuronales, no logran ese resultado. En los modelos de baja dimensión f1 y

f2, destaca la Red Elástica con un valor promedio entre 0.7515 y 0.8239, seguido de la Regresión Lineal con valores entre 0.7414 y 0.8052. Finalmente concluye que, se debe ampliar el conjunto de variables predictoras más allá del de los ocho modelos de factores considerados en este estudio. Después de realizar una revisión más completa de las variables de mercado de base amplia utilizadas en la literatura sobre precios de activos, se pueden diseñar características a partir de estas variables y volver a ejecutar la comparación del modelo. Además, el conjunto de modelos podría ampliarse aún más y se pueden considerar pasos adicionales de optimización del modelo y ajuste de hiperparámetros.

Koo & Kim. (2021), en la investigación, Prediction of Bitcoin price based on manipulating distribution strategy, en Corea, plantea el problema de, como predecir el precio del bitcoin a través de la manipulación de la estrategia de distribución, aplicando como método de solución el procedimiento de entrenamiento en arquitecturas de redes neuronales artificiales (ANN), incluidas MLP, RNN y LSTM. Obteniendo como resultados que, las grandes desviaciones estándar de las métricas de error exhibidas en MLP y RNN, es razonable discutir el desempeño de FDS principalmente en términos de LSTM con un valor de precisión de 0.0128, MLP con 0.0118 y RNN con 0.0104. Finalmente, concluye que, para la predicción, consideramos el retorno de Bitcoin, que es intratable debido a una concentración cercana a cero y un movimiento ruidoso, y tres tipos de enfoques: MLP, RNN y LSTM. Estos enfoques son reconocidos como excelentes enfoques de machine learning en lo que se refiere a la predicción de series de tiempo. Mejoramos estos enfoques con una nueva estrategia. Específicamente, Introducimos el FDS basado en la teoría de la cópula y construimos nuestros algoritmos con FDS para predecir el retorno del precio de Bitcoin.

Gan et al. (2020), en la investigación, Machine learning solutions to challenges in finance: An application to the pricing of financial products, en China, plantea el problema de, como predecir con aprendizaje automático los precios de los instrumentos financieros de las opciones asiáticas geométricas y aritméticas,

aplicando como método de solución el aprendizaje automático basado en el aprendizaje profundo para fijar el precio de las opciones de promedios aritméticos y geométricos, a través de la realización de un experimento numérico completo con datos generados por computadora, y la verificación del nuevo método mediante una prueba empírica con datos reales. Obteniendo como resultados que, el método de aprendizaje planteado es sólido, tanto en conjuntos de entrenamiento como en conjuntos de pruebas en diferentes tamaños de datos de muestra generados por tres tipos de métodos. Estos sólidos resultados demuestran la efectividad del método de aprendizaje profundo para fijar el precio de las opciones asiáticas. El aprendizaje profundo es más preciso y más rápido que los métodos tradicionales de fijación de precios de opciones, el sesgo de predicción tiene una mediana del 0,8% y una media del 95% del 2% cuando se utilizan datos reales, el MSE está cerca de cero, R-cuadrado y la correlación entre los datos y la predicción está cerca de 1, y el modelo de aprendizaje profundo entrenado calcula 10,000 precios de opciones asiáticas en 1 s. Concluye que, en el experimento numérico, se investigó la efectividad utilizando tres conjuntos de datos que son generados por la computadora de acuerdo con tres tipos de métodos tradicionales: la fórmula exacta de opciones geométricas asiáticas, la simulación de opciones geométricas asiáticas y la simulación de opciones aritméticas asiáticas, los resultados numéricos y el análisis empírico muestran que no importa qué conjunto de datos se utilice para entrenar el modelo de aprendizaje profundo, puede predecir los precios de las opciones asiáticas con una precisión alta. Asimismo, en referencia a los tres métodos tradicionales, la velocidad del modelo de aprendizaje profundo entrenado es extremadamente rápida.

Truong et al. (2020), en la investigación, Housing Price Prediction via Improved Machine Learning Techniques, en Estados Unidos, plantea el problema de, como elegir el modelo de predicción adecuado para el precio de vivienda, aplica como método, la comparación y análisis de tres tipos diferentes de métodos de aprendizaje automático, Random Forest, XGBoost y LightGBM, y dos técnicas de aprendizaje automático, incluida la regresión

híbrida y la regresión de generalización apilada. Los resultados demuestran que, el método de bosque aleatorio tiene el error más bajo, con valores de RMSLE en el conjunto de entrenamiento de 0,12980 y en el equipo de prueba con 0.16568, seguido por el modelo extremo con valores de 0.14969 y 0.16372, y en tercer lugar el modelo de Regresión híbrida de la máquina de aumento con valores de 0.16118 y 0.16603. Concluye que, de la comparación y análisis de los tres tipos diferentes de métodos de aprendizaje automático, todos esos métodos lograron resultados deseables, los diferentes modelos tienen sus pros y sus contras. Se deben realizar más investigaciones sobre los siguientes temas para investigar más a fondo estos modelos, especialmente las combinaciones de diferentes modelos.

Xu et al. (2020), en la investigación, Carbon price forecasting with complex network and extreme learning machine, en China, plantea el problema de, como predecir el precio del carbono usando la tecnología de análisis de redes complejas, luego, como método de solución, realiza el mapeo de datos del precio del carbono en una red de precios del carbono (CPN), luego, extrae la información efectiva de las fluctuaciones del precio del carbono mediante el uso de la topología de la red, y el uso de la información efectiva extraída para reconstruir los datos de muestra del precio del carbono. Con los datos reconstruidos y el algoritmo de la máquina de aprendizaje extremo, se construye el modelo de máquina de aprendizaje extremo de red de precios de carbono (CPN-ELM). Los resultados obtenidos, desde el aspecto del índice de precisión MAPE, la precisión de predicción del modelo CPN-ELM en la ventana de tiempo del 63.38% es mejor que eso. de ELM. Al mismo tiempo, el valor medio del índice MAPE del modelo CPN-ELM en todas las ventanas de tiempo es 0.0206, que fue menor que el valor medio del modelo de predicción ELM de 0.0229. En términos de índice de precisión RMSE, El modelo CPN-ELM es mejor que el modelo ELM en la ventana de tiempo del 53,85%. Y mientras tanto, el valor medio del índice RMSE en todas las ventanas de tiempo del modelo CPN-ELM es 0.3102, que es menor que el valor medio del modelo ELM de 0.3417. Para el aspecto del índice de precisión de directividad, la precisión del modelo CPNELM en la ventana de tiempo del

69.23% es mejor que la del modelo de predicción ELM, y el valor medio del índice MAPE del modelo CPN-ELM en todas las ventanas de tiempo es 0.5292, que es mayor que el valor medio del modelo ELM de 0,4982, Los resultados muestran que CPN-ELM puede mejorar la precisión predictiva de ELM tanto en precisión de nivel como en precisión direccional. Mientras tanto, el modelo de predicción CPN-ELM tiene mejor robustez al enfrentarse a muestras aleatorias, datos de muestra con diferentes frecuencias o datos de muestra con cambios estructurales. Finalmente, concluye que, el modelo empleado, es un nuevo tipo de modelo de predicción combinado que combina la tecnología de análisis de redes complejas de series de tiempo con un algoritmo de inteligencia artificial, donde el modelo CPN-ELM es superior al modelo ELM puro en precisión horizontal y precisión direccional.

Chowdhury et al. (2020), en la investigación, An approach to predict and forecast the price of constituents and index of cryptocurrency using machine learning, en Bangladesh, plantea el problema de, como predecir y pronosticar el precio de los componentes y el índice de criptomonedas; como método de solución propone la utilización de algoritmos y modelos de aprendizaje automático, a través de técnicas y algoritmos, con la comparación de modelos entre sí para obtener el mejor resultado. Se da a conocer la comparación entre los resultados que se obtuvieron en el citado trabajo y otras investigaciones anteriores, lo que muestra que la precisión utilizando el método de aprendizaje por conjuntos es de 0,924, mientras que en la de anteriores investigaciones, donde la precisión más alta es 0,952 con el modelo LightGBM, asimismo, el RMSE que se ha obtenido por potenciación de árbol gradiente para Bitcoin, DogeCoin y NEM son los valores de 32,863, 0.000 y 0.001 respectivamente. Por otro lado, se ha demostrado que LSTM y RNN se han comportado mejor que el modelo ARIMA, LSTM proporciona precisión y RMSE de 52,78% y 6,87%, mientras que RNN proporciona precisión y RMSE de 50,25% y 5,45%; para el modelo ARIMA, los valores son 50,05% y 53,74% respectivamente. La mayor precisión y RMSE utilizando el modelo de árboles potenciados por gradiente son 0.900 y 0.001, y 0.924 y 0.002 usando el método de aprendizaje por conjuntos. Finalmente, concluye que, los modelos exhiben un muy buen

desempeño en la predicción general del cierre, sosteniendo que se ha demostrado que el rendimiento de sus modelos parece mejor y competitivo. Obteniendo un 92,4% de precisión utilizando el método de aprendizaje por conjuntos, lo que se considera el mejor entre todos los modelos utilizados en el desarrollo de la citada investigación.

Bonnet et al. (2021), en la investigación, Machine learning and oil price point and density forecasting, en Brasil, plantea el problema de, como encontrar técnicas de aprendizaje automático para pronosticar el precio del petróleo; proponiendo como método de solución, los pronósticos del precio del petróleo proveniente de 23 métodos de pronóstico, además de algunos enfoques tradicionales para pronosticar los precios del petróleo, como el paseo aleatorio y los modelos ARIMA que considera modelos factoriales, asimismo, un conjunto de métodos de pronóstico que incluye varios métodos de aprendizaje automático no lineales, basados en procedimientos de regularización (por ejemplo, LASSO y red elástica) o árboles de regresión. Los resultados indican un buen rendimiento de los métodos de aprendizaje automático a corto plazo. Hasta seis meses, los precios futuros del petróleo, VECM y el modelo Schwartz-Smith proporcionan los mejores pronósticos. En horizontes más largos, las combinaciones de pronósticos también se vuelven relevantes. En varios casos, las ganancias de precisión con respecto al pronóstico de caminata aleatoria son estadísticamente significativas y alcanzan cifras de dos dígitos, en términos porcentuales, usando la estadística fuera de muestra, Finalmente, concluye que, la presente investigación estudia la precisión del pronóstico de 23 métodos competidores, que se utilizan para construir pronósticos puntuales de la variación del precio del petróleo Brent. Los métodos de pronóstico aplicados pueden ser útiles para ayudar a mejorar el conjunto de herramientas que los académicos y agentes del mercado utilizan actualmente para construir pronósticos del precio del petróleo, ofreciendo así una valiosa contribución al campo de los pronósticos macroeconómicos.

Kim et al. (2021), en la investigación, Predicting Ethereum prices with machine learning based on Blockchain information, en Korea, plantea el problema de,

como predecir los precios de Ethereum, proponiendo como método de solución, la comparación de algoritmos de aprendizaje automático empleados en la literatura existente respecto a predicciones de precios de criptomonedas, empleando una serie de variables predictoras usadas comúnmente para la predicción del precio del Bitcoin. Los resultados obtenidos, el análisis muestra que la ANN funciona mejor que la SVM en todos los modelos. Entre ellos, los Modelos I-4 y II-4 con la ANN presentaron el mejor desempeño, como se muestra en Tablas 6 y 7 (RMSE = 0.068, MAPE = 0,048). Este estudio realizó un análisis paso a paso para los Modelos I-1 a I-6. El modelo I-1 incluye solo factores macroeconómicos (RMSE = 0.131, MAPE = 0,067). El Modelo I-2 agrega información genérica de Blockchain donde se encuentra que, RMSE y MAPE se mejoraron (RMSE = 0.086, MAPE = 0.054). Por el contrario, el Modelo I-3 con información Blockchain específica de Ethereum no mejoró significativamente los resultados del análisis (RMSE = 0.107 MAPE = 0.052). Sin embargo, descubrimos que RMSE y MAPE se mejoraron en el Modelo I-4, que incluía información de Blockchain de Bitcoin (RMSE = 0.068, MAPE = 0.048). Además, este estudio confirmó que en el Modelo I-5, al agregar la información de Blockchain de Litecoin no mejoró el resultado del análisis (RMSE = 0.107, MAPE = 0.053). También se encontró que el Modelo I-6, con la información Blockchain de Dashcoin, no mejoró el rendimiento (RMSE = 0.099, MAPE = 0.053). Finalmente, se concluye que, los resultados revelan que la información genérica de Blockchain incluye información que está directamente relacionada con los precios de Ethereum. Sin embargo, La información de Blockchain específica de Ethereum y la información de Blockchain de Litecoin y Dashcoin no contribuyeron significativamente a la predicción del precio de Ethereum. Por tanto, es posible que se incluyan variables innecesarias en los modelos de predicción, según el análisis escalonado de los Modelos I-1 a I-6.

1.3. Teorías relacionadas al tema.

De acuerdo con las variables definidas en la presente investigación, estas se fundamentan en las siguientes bases teóricas:

Variable Independente:

1.3.1. Machine Learning.

Bobadilla (2020), como también conocido como aprendizaje automático, que comprende el proceso de aprender a través de datos, generando conocimiento de un comportamiento que aplican las computadoras. para ello usa algoritmos que le permiten estudiar una realidad, identifica patrones y permite la predicción de nueva información.

Sandoval (2018), Describe como parte de la Inteligencia Artificial, encargada de la generación de algoritmos con capacidad de aprendizaje sin tener que realizar programaciones de manera explícita, evitando que el responsable demande horas de programación teniendo en consideración distintos escenarios y excepciones posibles; donde lo único que haría, es entrenar el algoritmo con un vasto volumen de data, con la finalidad que el algoritmo logre aprender y sepa responder en cada escenario que se le presente.

Franco y Ramos (2019), Mencionan que, algunos de los objetivos de Machine Learning es el aprendizaje de las computadoras a través de la utilización de distintas técnicas y métodos estadísticos, numéricos y lógicos, que permitan que las técnicas puedan controlar y comprender enormes cantidades de datos.

1.3.2. Clasificación automática

Cárdenas (2014), define la clasificación automática como la tarea que se ejecuta a través de un sistema artificial en base a un grupo de elementos, con la finalidad de ordenarlos, en categorías o clases, asimismo, (Scherz., 2018) comenta que, la clasificación automática utiliza técnicas de aprendizaje automático, donde, gracias a la construcción automática de un clasificador, aprende características de las categorías a raíz de un grupo preclasificado de documentos.

1.3.3. Técnicas de clasificación automática

Gil et al. (2019), comentan que la mayoría de los investigadores, acuden a técnicas de clasificación de tipo supervisada, las cuales poseen diferentes procesos de aprendizaje y entramiento previos a la clasificación de las imágenes; a su vez, se obtiene un cierto interés por las de tipo no supervisada.

1.3.4. Aprendizaje supervisado

Sandoval (2018), se refiere al entrenamiento del algoritmo otorgándole las preguntas (características) y respuestas (etiquetas) ya establecidas, de ese modo, el algoritmo en un futuro podrá predecir teniendo conocimiento las características del caso; para este tipo de aprendizaje, existen dos algoritmos, algoritmo de clasificación y regresión.

1.3.5. Aprendizaje no supervisado

Sandoval (2018), menciona que, solo se debe brindar al algoritmo las características, más no las etiquetas, queriendo que agrupe los datos que le proporcionamos teniendo como referencia sus características.

1.3.6. Algoritmo de clasificación

Sandoval (2018), describe que, se debe esperar que el algoritmo indique a cuál grupo pertenece el elemento de estudio, encontrando patrones en los datos proporcionados y clasificándolos en grupos; comparando posteriormente los nuevos datos y ubicándolos en unos de los grupos creados. La variable que se predecirá debe ser un conjunto de estados categóricos.

1.3.7. Algoritmo de regresión

Sandoval (2018), indica que, en este tipo de algoritmo, lo que se desea conseguir o determinar es un número, un valor específico, este algoritmo no se ubica en grupos ni se clasifica en categorías.

1.3.8. Modelos de predicción

Menoyo (2021), Define que los modelos de predicción pueden darse de tipo clasificación y de regresión. El modelo de clasificación se caracteriza por analizar una variable de entrada (x) y cuya salida se representa en una categoría o etiqueta (y) de tipo discreta, en el caso del modelo de regresión analiza una o más variables de entrada (x) y cuya salida es representada por un resultado (y) de tipo numérico entero (int) ó decimal (flotante).

1.3.9. Modelos lineales:

Rayón (2017), Se refiere a ubicar una línea que se “ajuste” a la atmósfera de puntos con los que se está trabajando, destacando modelos populares tales como: La regresión lineal o regresión de los mínimos cuadrados y la logística (forma de adaptación de la regresión lineal respecto a problemas de clasificación de variables discretas o categóricas); estos, presentan frecuentemente el problema del “overfit”, que tiene que ver con el ajuste exagerado a los datos que se encuentran disponibles; a su vez, al considerarse relativamente como modelos “simples”, no garantizan resultados muy eficientes en el caso de comportamientos de mayor complejidad.

1.3.10. Modelos de árbol:

Rayón (2017), Los considera como modelos de mayor precisión, estabilidad y simplicidad de interpretación, teniendo como base, su construcción a través de reglas de decisión representadas como un árbol, que pueden implementar relaciones no lineales para la resolución de problemas. Entre sus principales modelos, destacan los “árboles de decisión” y los “random forest”, que al caracterizarse por ser más precisos y elaborados, estos, tiene mayor capacidad de predicción, pero menor rendimiento.

1.3.11. Redes neuronales:

Rayón (2017), Comenta que, las redes artificiales neuronales tienen como característica principal, replicar el comportamiento del cerebro humano, que a través de los millones de neuronas que lo habitan, estas, se encuentran interconectadas para el envío de mensajes ente sí. Esta “réplica” es considerado actualmente como uno de los modelos de moda, debido a la gran cantidad de habilidades cognitivas de razonamiento que obtienen; ejemplo de ello, son el reconocimiento de imágenes, donde por su complejidad, aplicar una red neuronal sería lo ideal, teniendo como único punto débil, la lentitud para el entrenamiento y la necesidad de amplia capacidad de cómputo.

1.3.12. Características del modelo de regresión

Menoyo (2021), los modelos de regresión permiten, la predicción de cantidades, sus variables de entrada pueden ser múltiples, cuando los datos incluyen fechas permite aplicar series de tiempo, su evaluación se aplica con error en la predicción.

1.3.13. Regresión Lineal

Bosch et al.(2019), considera a la regresión lineal como la actividad de aprendizaje inductivo estudiado y utilizado ampliamente, se puede definir también como un problema de clasificación de clases continuas, donde se predice valores de tipo numérico en vez de etiquetas de clasificación discreta, en algunos casos, se puede inferir la regresión lineal como la predicción numérica que tiene que ver en la asignación de valores numéricos a las diferentes instancias de un determinado dominio, descrito por varios atributos de valor continuo, donde los puntos representan datos del aprendizaje y la línea la predicción de futuros eventos.

1.3.14. Regresión Lineal Múltiple

Amat (2016), Se desarrolla a través de un modelo lineal, cuyos datos de entrada son conjuntos de variables denominadas predictores (x1,

x2, x3,…) con la finalidad de obtener un resultado (y), su objetivo principal es determinar la influencia de los predictores en el resultado final, asimismo, para seleccionar los predictores se pueden aplicar los métodos: Jerárquico, entrada forzada y stepwise.

1.3.15. Árboles de decisión

Merayo (2020), los árboles de decisión representan uno de los algoritmos de mayor uso en la toma de decisiones de Machine Learning, definido como modelo de predicción que, a través de la división de espacio en los predictores agrupa observaciones de valores de igual similitud para la variable de respuesta. Asimismo, un árbol de decisión se entiende como un algoritmo de tipo supervisado, donde, para realizar el aprendizaje de modelo, se necesita de una variable dependiente en grupo de entrenamiento, su conformación se da a través de nodos, y su desarrollo es de arriba hacia abajo.

1.3.16. Regresión de Bosque Aleatorio-Ramdon Forest:

Martinez (2020), Es un grupo de árboles de decisión que al combinarse con el modelo de machine learning “bagging”, permitiendo que diferentes árboles vean diversos segmentos de los datos, permitiendo así que ningún árbol vea todos los datos de entrenamiento, haciendo que cada árbol entrene con distintas muestras de datos para un mismo problema.

1.3.17. Regresión XGboost:

Gonzalez (2018), Su implementación es en el marco de Gradient Boosting aplicados en algoritmos de aprendizaje automático, proporcionando un impulsando el árbol paralelo (GBDT / GBM) que atiende varios problemas de la ciencia de datos con rapidez y precisión. El código puede ejecutarse en los principales entornos distribuidos (SGE, Hadoop y MPI) y es aplicado para miles de millones de casos.

Variable Dependiente:

1.3.18. Predicción

Es la acción de estimar, en un ambiente de incertidumbre, valores futuros de variables temporales teniendo como referencia sus valores pasados; asimismo, se considera a la predicción como una de las herramientas importantes en la toma de decisiones dentro de los diferentes ámbitos de estudio (Garcia., 2016).

1.3.19. Plan Anual de Contrataciones

Se considera como una herramienta de gestión de uso obligatorio, permite la planificación, programación y seguimiento de las contrataciones que las Entidades Públicas realicen en un año determinado, el cual debe comprender: la relación de las contrataciones a realizarse, el tipo de proceso de selección, el valor estimado, y la fecha de convocatoria (Ministerio de Economía y Finanzas, 2018).

1.3.20. Valores Estimados

Es el valor que se le asignan a las futuras contrataciones, el cual se determina mediante la indagación del mercado de los bienes o servicios, asimismo, sirve de referencia para identificar el tipo de proceso de selección (Ministerio de Economía y Finanzas, 2018).

Herramientas de Machine Learning:

Franco y Ramos (2019), Para el desarrollo y ejecución de modelos de Machine Learning, se cuenta con diversas herramientas y lenguajes de programación, tales como “Python”, “GNU R”, “Weka” y “Rapidminder”, teniendo como característica común, su uso de forma gratuita “open-source”, y ser consideradas como multiplataformas, lo que les permite ser utilizadas en gran parte de los sistemas operativos actuales.

1.3.21. Python

Nolasco (2018), tipo de lenguaje de programación, diseñado por Guido Van Rosum en la década de 90s laboró en Google y en la actualidad en Dropboxm, su denominación se origina de la historieta comic Monty Python, posee una sintaxis limpia, entendible, de tipado dinámico, es decir, las variables pueden tener datos de diferentes tipos, de la mano de su interpretación, hacen de este lenguaje uno de los favoritos para empezar aprender a programar; a su vez, Phyton es uno de los lenguajes de interpretación que no requiere compilación del código fuente para su ejecución, ofreciendo ventajas respecto a otros lenguajes.

Franco y Ramos (2019), consideran a Phyton, como uno de los lenguajes de programación más usados en el rubro de los campos científicos y en la ciencia de datos, gracias a su versatilidad, simplicidad en su sintaxis, y por ser multiplataforma, puede ser ejecutado en distintos sistemas operativos tales como, “Linux”, “MacOS” y “Windows”; a su vez, permite la utilización de bibliotecas, que vienen hacer un conjunto de programas con un fin en específico y predeterminado.

1.3.22. Librerías de Python de aprendizaje automático

(Briega, s.f.) Python presenta una ventaja competitiva sobre los otros lenguajes de programación con respecto al uso de librerías especializadas para diversas aplicaciones en la materia de machine learning, entre las principales tenemos: Statsmodels (para modelos estadísticos), PyMC (para modelos estadísticos bayesianos), Scikit-Learn (implementa algoritmos de aprendizaje).

1.3.23. Colaboratory / Colab

Google (2018), Colaboratory, o reconocido también con el nombre de "Colab", es una aplicación de Google. Que permite al usuario programar y ejecutar en el navegador web, código de Python.

Usualmente, esta aplicación es utilizada para labores de machine learning, y su uso no depende de licencia con costo alguno, siendo una aplicación potente y que optimiza los recursos de los ordenadores.

1.4. Formulación del Problema.

¿Cuál es la técnica de estimación basada en Machine Learning, que predice con mayor precisión, los costos en los planes de las adquisiciones de las entidades públicas del Perú?

1.5. Justificación e importancia del estudio.

Permitirá contribuir con fomentar el conocimiento científico en la línea de Tecnología, a través de algoritmos supervisados de Machine Learning.

En el Perú, se realizan un promedio 35000 procesos de adquisiciones al año, siendo de vital importancia poder determinar costos referenciales cercanos a los reales, lo cual garantizara el desarrollo de Planes de Adquisiciones efectivos, disminución de los casos de sobrevaloraciones, y generando un significativo ahorro al Gobierno Peruano.

Finalmente, al contarse con el acceso a información de contrataciones del estado de los años 2018 - 2021, además de las herramientas para la realización del análisis de información, el presente trabajo de investigación resulta viable de desarrollo.

1.6. Hipótesis.

La técnica de estimación de Machine Learning Random Forest**,** es la que predice con mayor precisión los costos en los planes de las contrataciones de las entidades públicas del Perú.

1.7. Objetivos.

1.7.1. Objetivo general.

Comparar técnicas de estimación de machine learning para predecir los costos en los planes de adquisiciones de las entidades públicas del Perú.

1.7.2. Objetivos específicos.

a) Elaborar el dataset de las contrataciones por tipo de bien o servicio.
b) Seleccionar las técnicas de estimación automática de machine learning a implementar.
c) Implementar las técnicas de estimación automática de machine learning para predecir costos.
d) Realizar las pruebas del desempeño de las técnicas implementadas.

II. MATERIALES Y MÉTODO

2.1. Tipo y Diseño de Investigación.

2.1.1. El tipo de investigación es: Tecnológica – Aplicada

Cegarra (2004), define como investigación tecnológica, el invento de artefactos o procesos que permiten ofrecer un beneficio monetario o económico. Usualmente los casos son de tipo experimental.

La investigación aplicada tiene como objeto dar a conocer casos nuevos, al ser proyectada correctamente la investigación permitirá confiar en el descubrimiento y obtener información útil y significativa que contribuya a nuevas teorías (Baena., 2014).

2.1.2. El diseño es: Cuantitativa -- > cuasiexperimental

El Diseño Cuantitativa, según (Hernández et al., 2014), representan procesos secuenciales de tipo probatorios, donde a través de la utilización de métodos estadísticos se extraerán las conclusiones del estudio realizado.

Diseño Cuasi Experimental, tiene por objetivo contrastar la relación causa - efecto, sin fines de control pues este diseño no permite manipular las variables. Esta herramienta es utilizada en su mayoría de casos para contextos naturales, pues no se tiene control en sus factores (Muñoz., 2005).

2.2. Población y muestra.

2.2.1. Población.

La población comprende 27 técnicas o algoritmos de estimación de precios (ver Anexo N° 02), las que fueron identificadas luego de aplicar

la revisión sistemática de la literatura y que han sido mencionadas en la sección 1.2. Trabajos Previos.

2.2.2. Muestra 04 algoritmos.

La muestra se determinó, tomando en cuenta los 04 algoritmos con mayor presencia en las investigaciones mencionadas en la sección 1.2. Trabajos Previos:

Tabla 1 - Muestra de 04 algoritmos

N°	**Algoritmo**	**Siglas**
1	Regresión Lineal Múltiple	LR
2	Árbol de decisiones	TD
3	Random Forest	RF
4	Media móvil e integrada auto regresiva	XGboost

Fuente: Elaboración propia.

2.3. Variables, Operacionalización.

Los resultados esperados a obtenerse se lograron mediante los siguientes indicadores y de los instrumentos que permitieron la recolección de datos:

Tabla 2 - Variables, Operacionalización

Variables	Indicador	Ítem	Técnica / instrumentos de recolección de datos
Técnicas de estimación basados en machine learning	Consumo de **Memoria**	$\boldsymbol{cm = \sum_{j}^{n} \frac{cm_j}{n}}$	
	Tiempo de respuesta	$\boldsymbol{Tr = \sum_{j}^{n_f} \frac{tf_j - tf_i}{n}}$	
Predecir costos en los planes de adquisiciones de las entidades públicas del Perú.	**MAE** (Error absoluto medio)	$\boldsymbol{MAE = \frac{SAE}{N} = \frac{\sum_{i=1}^{N} \lvert x_i - \hat{x}_i \rvert}{N}}$	**Técnica:** Observación **Instrumento**: Ficha digital de observación
	MAPE (Error porcentual absoluto medio)	$MAPE = \frac{\sum_{t=1}^{n} \frac{\lvert X_t - Y_t \rvert}{\lvert X_t \rvert}}{N}$	
	MSE (Error cuadrático medio)	$MSE = \frac{1}{N}\sum_{i=1}^{n}(X_i - Y_i)^2$	
	RMSE ($\sqrt{}$del Error cuadrático medio)	$\boldsymbol{RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}}}$	
	R2 (Coeficiente de determinación)	$\boldsymbol{R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}}$ $\boldsymbol{SST = \sum (y - \bar{y})^2}$ $\boldsymbol{SSR = \sum (y' - \bar{y'})^2}$ $\boldsymbol{SSE = \sum (y - \overline{y'})^2}$	

Fuente: elaboración propia

2.4. Técnicas e instrumentos de recolección de datos, validez y confiabilidad.

2.4.1. Técnica: Observación.

Pérez et al. (2020), menciona que, la técnica de la observación permite recabar información tal cual sucede en la realidad, su utilización dependerá de lo que se está investigando, en el caso de su utilización, las variables que se obtengan ganarán en calidad de respuesta, a su vez, se obtendrá la certeza de que no se cuente una versión distorsionada a lo sucedido.

2.4.2. Instrumento: Como instrumentos se ha considerado la ficha digital de observación aplicado a:

Desempeño del algoritmo (Anexo N° 02).

Precisión del algoritmo (Anexo N° 03).

2.5. Procedimiento de análisis de datos.

2.5.1. Consumo de Memoria:

Indicador que muestra la memoria consumida durante las pruebas del modelo de aprendizaje, su fórmula es:

$$cm = \sum_{j}^{n} \frac{cm_j}{n}$$

Donde:

cm: Consumo de memoria

cm_j: Consumo de memoria en la prueba j

n: Total de pruebas.

2.5.2. Tiempo de respuesta:

Indicador es refleja el tiempo promedio de ejecución del modelo en la etapa de pruebas, su fórmula es:

$$Tr = \sum_{j}^{n_f} \frac{tf_j - tf_i}{n}$$

Donde:

Tr:Tiempo de respuesta

tf_j:Tiempo de respuesta final

tf_i:Tiempo de respuesta inicial

n: Total de pruebas.

2.5.3. MAE (Error absoluto medio):

Calcula la función MAE para el pronóstico y los resultados posibles, su fórmula es:

$$MAE = \frac{SAE}{N} = \frac{\sum_{i=1}^{N} \left| y_i - \hat{y}_i \right|}{N}$$

Donde:

y_i : Corresponde a la información actual de la serie de tiempo

$\hat{y}_i$: Corresponde a la serie de tiempo pronosticada.

SAE: Corresponde a la sumatoria de errores absolutos o desviaciones

N: Corresponde a los números de puntos de datos no faltantes

2.5.4. MAPE (Error porcentual absoluto medio):

Esta métrica permite calcular la dimensión del error de tipo absoluto expresado en porcentajes, siendo su fórmula:

$$MAPE = \frac{\sum_{t=1}^{n} \frac{|X_t - Y_t|}{|X_t|}}{N}$$

Donde:

X_t: $Valor\ real$

Y_t: $Valor\ de\ pronóstico$

N: $Número\ de\ puntos\ de\ datos\ no\ faltantes$

2.5.5. MSE (error cuadrático medio):

Esta métrica calcula el valor promedio de los errores elevados al cuadrado, su fórmula es:

$$MSE = \frac{1}{N} \sum_{i=1}^{n} (X_i - Y_i)^2$$

Donde:

X_i: $Valor\ real\ en\ el\ tiempo\ i$

Y_i: $Valor\ estimado\ en\ el\ tiempo\ i$

N: $Número\ de\ puntos\ de\ datos$

2.5.6. RMSE (Raíz - error cuadrático medio):

Calcula la función de raíz de la métrica MSE (conocida como desviación media cuadrática -RMSD), su fórmula es:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(p_i - a_i)^2}{n}}$$

Donde:

a: objetivo actual

p: objetivo previsto

2.5.7. R2 (Coeficiente de determinación):

Sumariza el poder de explicación del modelo de regresión y el computo de la suma de cuadrados, R2 describe la parte de varianza de la variable de tipo dependiente detallada del modelo de regresión. el modelo es “perfecto”, si el SSE es cero, y R2 es 1, asimismo, Si el modelo de regresión es un desastre, SSE es igual a SST, y no se puede explicar ninguna varianza por regresión, además R2 es cero, su fórmula es:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$$

$$SST = \sum (x - \bar{x})^2$$

$$SSR = \sum (x' - \bar{x'})^2$$

$$SSE = \sum (x - \overline{x'})^2$$

De lo cual:

$SST =$ Corresponde a la suma de cuadrados total

$SSR =$ Corresponde a la suma de cuadrados de regresión.

$SSE =$ Corresponde a la suma de cuadrados de error.

2.6. Criterios éticos.

Con la finalidad de conducir la investigación de forma correcta y respetando las reglas de éticas, se tomarán los siguientes:

2.6.1. Objetividad

El análisis de los procedimientos para la determinación de los costos (valores estimados) de las contrataciones públicas, se realizarán aplicando las técnicas establecidas para la estimación de aprendizaje automático.

2.6.2. Originalidad

El desarrollo de los métodos y la obtención de resultados de la presente investigación, se fundamentan en trabajos previos de investigadores, a quien se les referenciara en respeto y reconocimiento a contribución a la investigación.

2.6.3. Veracidad

La información (dataset) de las contrataciones del estado, son de acceso público y se analizarán conservando su autenticidad.

2.7. Criterios de Rigor Científico.

Para tener la seguridad de efectividad de la investigación, se eligieron los siguientes:

2.7.1. Confiabilidad

La información de las contrataciones a ser recolectada corresponderá a los comprendidos entre el año 2018 a 2021, para lo cual se realizarán las pruebas de consistencia durante el proceso de análisis.

2.7.2. Trabajo Metódico

En el desarrollo de la investigación, se aplicarán las normas, técnicas o procedimientos aplicables a cada etapa.

III. RESULTADOS.

3.1. Resultados en Tablas y Figuras.

Esta sección presenta una tabla resumen de las métricas alcanzadas de los 04 algoritmos, al ser evaluadas por los 07 indicadores (02 de la variable independiente y 05 de la variable dependiente) citados en la sección 2.3 de operacionalización de variables:

Tabla 3 - Medidas de rendimiento de los algoritmos implementados

Algoritmo de regresión	MAE	MAPE	MSE	RMSE	R2
Regresión Lineal Multiple	**4.03E+06**	**0.30**	**4.04E+13**	**6.36E+06**	**0.79587**
Árbol de decisión	**6.16E+06**	**0.40**	**1.03E+14**	**1.03E+14**	**0.53162**
Random Forest	**5.43E+06**	**0.33**	**6.20E+12**	**7.88E+06**	**0.68666**
Xgboost	**5.97E+06**	**0.34**	**7.59E+13**	**8.71E+06**	**0.61649**

Fuente: elaboración propia

La tabla 03. Muestra los resultados respecto a la precisión que ofrece cada algoritmo, observándose que el algoritmo que presenta mejores niveles de precisión es la Regresión Lineal Múltiple con un R2 de 0.79587 y los menores valores de error en la métrica de MAE, MAPE y RMSE.

Para una mayor comprensión de los resultados, es necesario conocer lo siguiente:

- Para las métricas de error: MAE, MAPE, MSE y RMSE (representa un mejor resultado el menor valor identificado al comparar los 04 algoritmos).
- Los valores MAE, MSE y RMSE, se representan con valores de muchos dígitos, debido a que los presupuestos de las partidas de adquisiciones pueden superar los millones de soles.

- Para la métrica de R2 (representa un mejor resultado el mayor valor identificado al comparar los 04 algoritmos).

Con la información previa, se detallan los resultados de las pruebas de precisión de los modelos de regresión aplicados a los presupuestos de las adjudicaciones publicas agrupados por nivel de gobierno fueron: en primer lugar, sitúa el modelo Regresión Lineal Múltiple con los siguientes índices de error MAE=4.03E+06, MAPE=0.30.MSE=4.04E+13, RMSE=6.36E+06 y R2 = 0.79587, en segundo lugar, se sitúa Random Forest con índices de MAE=5.43E+06, MAPE=0.33, MSE=6.20E+12, RMSE=7.88E+06 y R2 = 0.68666, en tercer lugar, se sitúa a XGboost con índices de MAE=5.97E+06, MAPE=0.34, MSE=7.59E+13, RMSE=8.71E+06 y R2 = 0.61649, y en cuarto lugar, Árbol de Decisiones con índices de MAE=6.16E+06, MAPE=0.40, MSE=1.03E+14, RMSE=1.03E+14 y R2 = 0.53162.

Como resultado de las pruebas de desempeño de los modelos de regresión aplicados a las adjudicaciones publicas agrupados por nivel de gobierno fueron: Lineal Múltiple con los siguientes índices: tiempo de respuesta (ejecución) = 0.008s y 1.21 M.RAM, Random Forest con los siguientes índices: 0.017s y 1.21 M.RAM, XGboost con los siguientes índices: 0.018s y 1.49 M.RAM, y Árbol de Decisiones con los siguientes índices: 016s y 1.53 M.RAM.

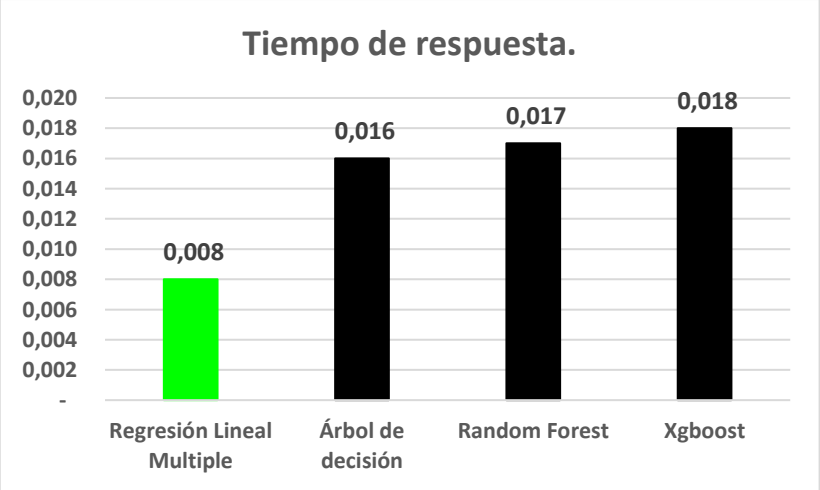

Ilustración 1 - Tiempo de respuesta de ejecución del algoritmo

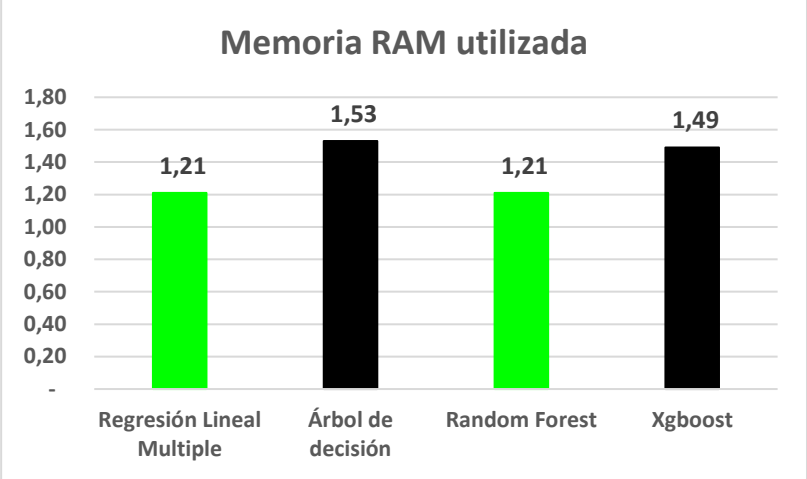

Ilustración 2 - Consumo de memoria RAM en la ejecución del algoritmo

3.2. Discusión de resultados

A continuación, se comentan los resultados de los trabajos previos del numeral 1.2 del presente documento, que por su problema de ingeniería se podría relacionar con la problemática identificadas en la presente investigación:

- **Deepa et al. (2021), en la investigación, predicción de precios en el ámbito de la agricultura con aprendizaje automático**, con el problema de ingeniería: determinar el precio del algodón en la India, se encuentra coincidencia en la implementación del algoritmo de regresión lineal múltiple, árbol de regresión de decisión y bosque aleatorio, sin embargo, el algoritmo que tuvo mejor desempeño fue árbol de decisiones.

- **Zhang et al. (2021), en la investigación, comparación de estudios para determinar el precio del cobre**, con el problema de ingeniería: determinar el precio del cobre en USA, se encuentra como coincidencia la implementación del algoritmo bosque de aleatorio, sin embargo, el algoritmo que tuvo mejor desempeño fue de redes neuronales percepton multicapa.

- **Rico & Taltavull. (2021), en la investigación, predicción del precio de productos inmobiliarios,** con el problema de ingeniería: determinar el precio de viviendas en el país de España, se encuentra coincidencia en la implementación de algoritmos de regresión lineal múltiple, árbol de regresión de decisión, bosque de aletarorio y Xgboost, sin embargo, el algoritmo que tuvo mejor desempeño fue el de bosque de aleatorio.

- **Chen et al. (2020), en la investigación, predicción del precio de bitcoin con aprendizaje automático,** con el problema de ingeniería: determinar el precio del bitcoin en China, se encuentra coincidencia en la implementación de los algoritmos de regresión lineal múltiple y Xgboost, asimismo, se coincide en que el algoritmo que tuvo mejor desempeño fue regresión lineal múltiple.

- **Raditya et al. (2021), en la investigación, predicción del precio de calzado deportivo mediante aprendizaje automático,** con el problema de ingeniería: determinar el precio de Zapatillas en Indonesia, se encuentra coincidencia en la implementación de los algoritmos de regresión lineal múltiple y bosque de decisiones, asimismo, se coincide en que el algoritmo que tuvo mejor desempeño fue regresión lineal múltiple.

Los trabajos previos, han servido de guía para el desarrollo de la presente investigación, esto gracias a que en las diversas investigaciones se plantean diversas formas de abordar el problema de ingeniería orientado a la estimación de precios o costos.

3.3. Aporte práctico.

El desarrollo del aporte practico comprende las siguientes etapas:

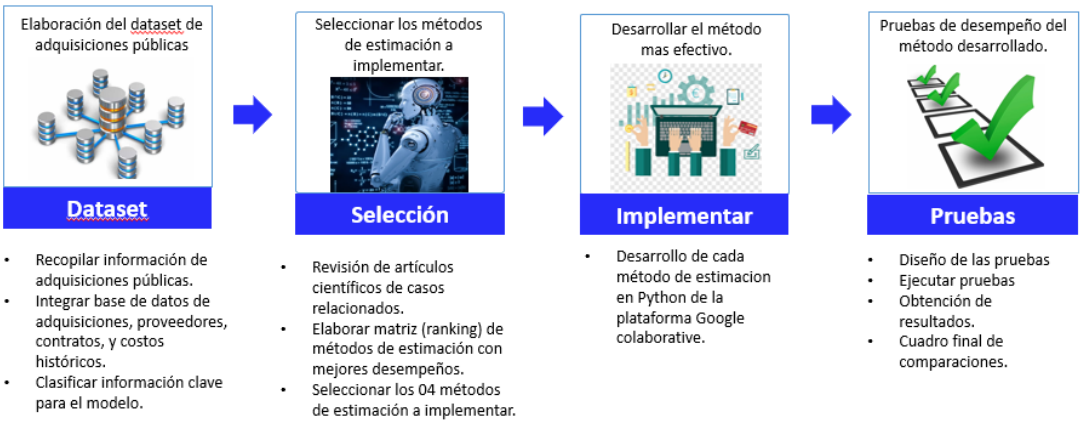

Ilustración 3 - Metodología del modelo propuesto

a. Elaborar el dataset de las contrataciones por tipo de bien o servicio.

La información a recabar, para la construcción del dataset utilizado en el presente estudio, se ubicó en el portal de datos abiertos del OSCE (https://portal.osce.gob.pe), asimismo, se revisaron los diccionarios de datos de los dataset publicados, eligiéndose el diccionario de **Datos de la Convocatoria o Invitación** (anexo 05), por tratarse de la información que se relacionaría mejor con los objetivos definidos.

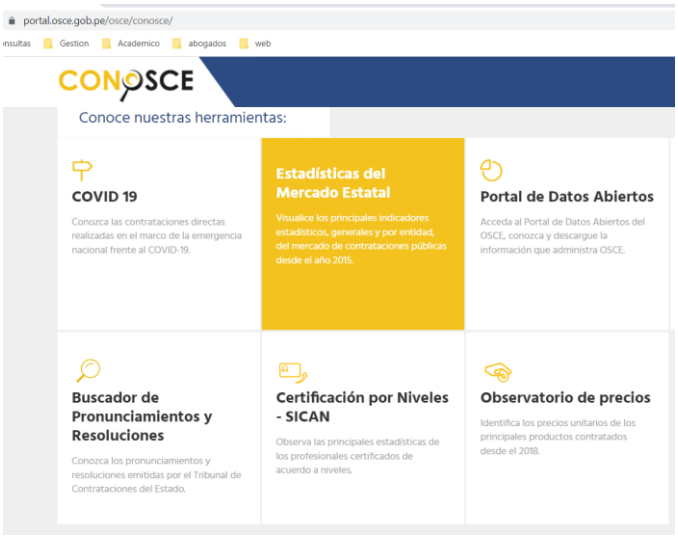

Ilustración 4 - Vista del portal de datos abiertos del OSCE

De igual modo, al acceder a la sección de Datos de la Convocatoria o Invitación, se visualizaron las publicaciones en formato Excel de los procesos de selección por años, desde enero 2018 - abril 2022, descargándose los 5 (cinco) archivos en formato Excel, denominados “descargar todos los procesos”, para su consolidación en un solo dataset preliminar, obteniéndose 28 campos (columnas) y 248,399 registros (filas), con lo cual se habilita el análisis y limpieza de información.

De igual modo, luego del análisis y limpieza de 248,399 registros, se obtuvo 85,578 registros aptos para dar inicio al proceso de normalización, según el siguiente detalle:

Tabla 4 - Datos a utilizarse en el Dataset

DATASET / CAMPOS	REGISTROS ELIMINADOS	SUBTOTAL	TOTAL
DATASET PRELIMINAR			**248,399**
Menos (-)			
OBJETOCONTRACTUAL	Consultoria, Obra y Servicio	121,578	
ESTADOITEM	Adjudicado y consentido	6,998	
PAQUETE	Si	30,255	
FECHAPRESENTACIONPROPUESTA	2022	3,990	
Subtotal de registros eliminados			**162,821**
DATASET PARA NORMALIZACION			**85,578**

Fuente: elaboración propia

Asimismo, con la finalidad de eliminar los datos atípicos, en la web de colab de Google, se identificó la composición de la información del dataset de los 85,578 registros, con el comando describe:

```
df.describe()
```

	MONTOREFERENCIAL	MES_FECHAPRESENTACIONPROPUESTA	mes
count	8.557800e+04	85578.000000	85578.00000
mean	5.245248e+06	2019.528570	7.20954
std	3.912835e+07	1.097216	3.35767
min	0.000000e+00	2018.000000	1.00000
25%	7.376819e+04	2019.000000	4.00000
50%	1.638929e+05	2020.000000	7.00000
75%	4.295156e+05	2020.000000	10.00000
max	6.157510e+08	2021.000000	12.00000

Ilustración 5 - Descripción de los datos del dataset

Una vez conocida el número de valores, mediana, la desviación estándar, el valor mínimo, máximo y los cuartiles, se desea conocer que segmento de la información tiene mayor representación, con la finalidad de evitar los datos atípicos, con la siguiente instrucción:.

```
plt.hist(df.MONTOREFERENCIAL)
plt.show()

plt.boxplot(df.MONTOREFERENCIAL)
plt.show()
```

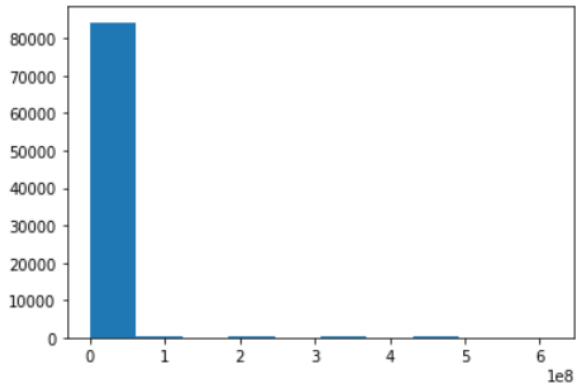

Ilustración 6 - Concentración de información

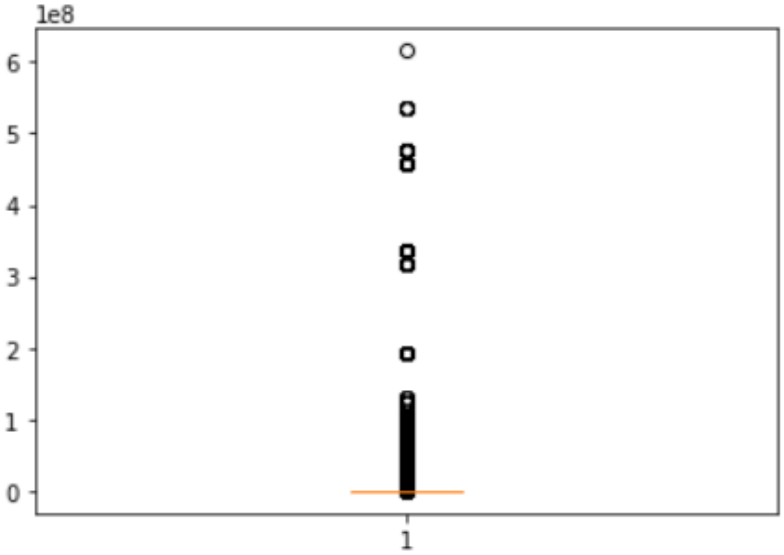

Ilustración 7 - Concentración de mayores datos

Seguidamente, se identificaron las columnas **Monto, año, mes, tipo, Entidad** (de la base de datos), siendo la columna monto la variable que nos permitirá realizar el análisis de las predicciones.

```
df.columns=['Monto','año','mes','Tipo','Entidad']
df.head()
```

Una vez, habiéndose determinado que la mayor concentración de los datos se encuentra en el umbral menor o igual a un presupuesto en adquisiciones de 1.638929e+05 soles, se procede con trabajar en adelante con 42,789 registros.

```
df=df[df.Monto<1.638929e+05]
```

Finalmente, en el anexo 06, se presenta la versión final de la estructura del dataset, como insumo para el desarrollo de los objetivos específicos c y d.

b. Seleccionar las técnicas de regresión automática de machine learning a implementar.

Para el desarrollo del presente objetivo se procedió con la revisión de literatura sistemática con la definición del objetivo: identificar las mejores técnicas computacionales para la estimación de precios, luego se determinaron las variables del PICOC (P=población, I=intervención, C=comparación, O=outcome “salida” y C=contexto); asimismo, se determinaron las preguntas que se pretenden absolver con la revisión de la literatura, la forma para acopiar, seleccionar y priorizar los artículos más representativos y relacionados a la industria, a continuación se presentan las etapas que comprendió dicho proceso:

Ilustración 8 - Proceso de planificación para la revisión sistemática

b.1 Definición de preguntas de investigación

En esta sección se definen las preguntas que nos permitirán ampliar nuestro conocimiento, respecto a las técnicas computacionales para la estimación de precios y los principales resultados que podamos obtener:

b.1.1 ¿Cuáles son las técnicas de aprendizaje automático que se desarrollaron para estimar precios?

La formulación de la presente pregunta permitirá conocer cuál es la población de las técnicas, métodos o algoritmos mencionados en los artículos priorizados.

b.1.2 ¿Cuál es la técnica de aprendizaje automático más utilizada para estimar precios?

Esta pregunta permitirá identificar la técnica, método o algoritmo que se utiliza con mayor frecuencia de la población.

b.1.3 ¿Cuáles son las métricas que se utilizan para evaluar los resultados?

Con los resultados obtenidos al aplicar las técnicas, métodos o algoritmos, esta pregunta busca conocer cuáles son los indicadores que se han utilizado en la determinación de la efectividad de la predicción o estimación de los precios de los casos de uso.

b.2 Realización de la investigación

Las búsquedas de los artículos de investigación, fueron realizadas en los siguientes 04 portales web: a) ACM Digital Library http://portal.acm.org, b) IEEE Digital Library http://ieeexplore.ieee.org, c) Science@Direct http://www.sciencedirect.com y d) Scopus http://www.scopus.com; a través de búsquedas avanzadas con las palabras: “predicción de precios”, “cálculo de precios”, o “estimación de precios”; y “técnicas computacionales”, “algoritmos”, o “machine learning”; asimismo, las búsquedas fueron acotadas con los filtros de publicaciones entre los años 2019 y 2021.

b.3 Proyección de artículos relevantes

El proceso de proyección de artículos relevantes inicia con la definición previa de los criterios de exclusión e inclusión y; entre los criterios de inclusión de artículos tenemos: relacionado con técnicas de predicción de precios de productos finales, relacionado con machine learning, publicaciones finales; entre los criterios de exclusión tenemos: No relacionado con técnica de predicción o estimación de precios, sin publicar, cuyos temas estén relacionados con mercado de valores, energía y explotación de minerales.

En segundo lugar, se definieron 05 preguntas de evaluación de la calidad (¿Es una literatura sobre técnicas de estimación de precios computacionales?, ¿Está identificado el problema de ingeniería?, ¿El método desarrollado está claramente desarrollado?, ¿Están

identificadas las técnicas de estimación utilizadas?) que permitieron evaluar si el artículo contribuirá con la finalidad de la revisión de la literatura sistemática, teniendo como posibles respuestas “Si, Parcialmente y No” con valores de “1, 0.5 y 0”, respectivamente; de tal manera que al evaluarse la admisibilidad de los artículos científicos, la puntuación mínima será de 3.5 y la máxima de 5.

b.4 Palabras clave sobre la base del resumen

Teniendo en cuenta las palabras clave como: predicción, cálculo o estimación de precios; y por otro lado las palabras de machine learning, técnicas ó algoritmos, se procede con la lectura de los resúmenes y palabras claves de cada artículo, con la finalidad de ir clasificándolos como: aceptados, duplicados o rechazados.

b.5 Proceso de extracción y mapeo de datos

Etapa que consiste en establecer los elementos o campos, de la información que se desea extraer de los 24 artículos, con la finalidad de construir nueva información a partir del acopio, consolidación y comparación de los datos registrados a continuación, se detallan los siguientes elementos:

Tabla 5 - Información de los datos a extraer de la publicación

N°	Elemento de datos	Descripción
1	Año	Fecha
2	Autor	Autor o autores
3	Titulo	Título
4	País	Lugar
5	Fuente	Revista, Simposio, Conferencia.
6	Procedencia	Empresa o Academia

Fuente: Elaboración propia.

b.6. Resultados de la búsqueda

Como producto de la recopilación inicial de artículos científicos, se logró el acopio de 949 de ellas, de las cuales al realizar una revisión a los resúmenes y palabras claves de cada documento, se descartaron 802

artículos, quedando en calidad de aceptados la cantidad de 147 artículos, los mismo que al descartar 53 artículos con la calidad de duplicados, restan 94 artículos priorizados por resumen.

Finalmente, al aplicar la evaluación de calidad mencionada en el número 3.3 del presente documento, se procedió con la extracción y mapeo de los datos de 24 artículos científicos.

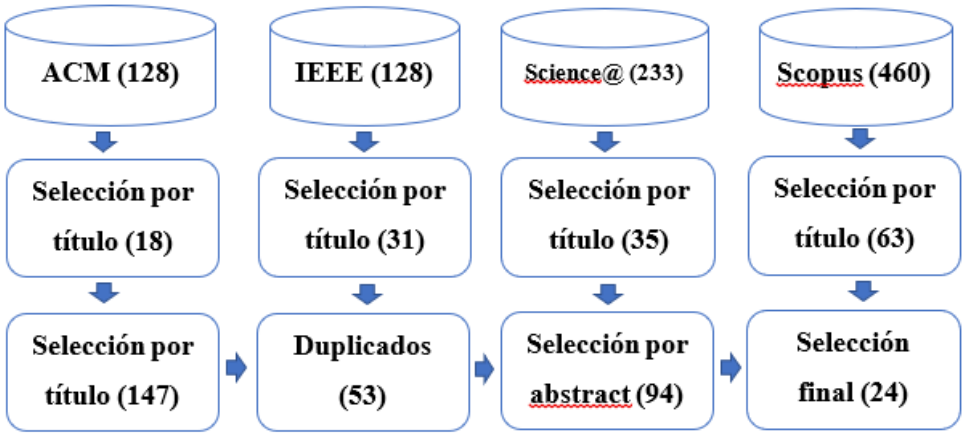

Ilustración 9 - Proceso de selección de trabajos relevantes

Tabla 6 - Artículos seleccionados

#	Autor	Año	Tipo de publicación	Caso de uso
1	**Wang, Tianyi and Pouyanfar, Samira and Tian, Haiman and Tao, Yudong and Alonso, Miguel and Luis, Steven and Chen, Shu-Ching** [1]	**2019**	**Conferencia**	**Sector Transporte Aéreo**
2	**Joshi, N. and Singh, G. and Kumar, S. and Jain, R. and Nagrath, P.** [2]	**2020**	**Conferencia**	**Sector Transporte Aéreo**
3	**Ma, Wei and Nowocin, Kendall and Marathe, Niraj and Chen, George H.** [3]	**2019**	**Conferencia**	**Sector agricultura**
4	**Hossain, S.M.S. and Rawat, J. and Logofatu, D.** [4]	**2021**	**Conferencia**	**Sector Inmobiliario**
5	**Jue Wang and Zhen Wang and Xiang Li and Hao Zhou** [5]	**2022**	**Revista**	**Sector agricultura**
6	**Hasan, M.M. and Zahara, M.T. and Sykot, M.M. and Nur, A.U. and Saifuzzaman, M. and Hafiz, R.** [6]	**2020**	**Conferencia**	**Sector agricultura**

7	**Shiliang Su and Shenjing He and Chenxi Sun and Hui Zhang and Lirong Hu and Mengjun Kang** [7]	**2021**	**Revista**	**Sector Inmobiliario**
8	**Madhuri, CH. Raga and Anuradha, G. and Pujitha, M. Vani** [8]	**2019**	**Conferencia**	**Sector Inmobiliario**
9	**Peng, Ningxin and Li, Kangcheng and Qin, Yiyuan** [9]	**2020**	**Conferencia**	**Sector Inmobiliario**
10	**Yuan, Chen Zhi and Ling, Sin Kai**[10]	**2020**	**Revista**	**Sector agricultura**
11	**Manasa, J. and Gupta, R. and Narahari, N.S.** [11]	**2020**	**Conferencia**	**Sector Inmobiliario**
12	**Zhu, A. and Li, R. and Xie, Z.** [12]	**2020**	**Conferencia**	**Sector Inmobiliario**
13	**Lirong Hu and Shenjing He and Zixuan Han and He Xiao and Shiliang Su and Min Weng and Zhongliang Cai** [13]	**2019**	**Revista**	**Sector Inmobiliario**
14	**Fan, M. and Huang, J. and Zhuo, A. and Li, Y. and Li, P. and Wang, H.** [14]	**2019**	**Conferencia**	**Sector Inmobiliario**
15	**Mostafa Mir and H.M. Dipu Kabir and Farnad Nasirzadeh and Abbas Khosravi** [15]	**2021**	**Revista**	**Sector Inmobiliario**
16	**Sameh, Ahmed and Abunadi, Ibrahim** [16]	**2019**	**Revista**	**Sector Inmobiliario**
17	**Wang, X. and Gao, S. and Zhou, S. and Guo, Y. and Duan, Y. and Wu, D.** [17]	**2021**	**Revista**	**Sector Inmobiliario**
18	**Jain, M. and Rajput, H. and Garg, N. and Chawla, P.** [18]	**2020**	**Conferencia**	**Sector Inmobiliario**
19	**Steven Davenport** [19]	**2021**	**Revista**	**Sector agricultura**
20	**DAI, Muyun and WANG, Wan and MIAO, Lixin** [20]	**2020**	**Revista**	**Sector Comercio**
21	**Mrsic, L. and Jerkovic, H. and Balkovic, M.** [21]	**2020**	**Conferencia**	**Sector Inmobiliario**
22	**Katai, Y. and Hasuike, T.** [22]	**2019**	**Conferencia**	**Sector Textil**
23	**Mehedi Hasan, M. and Zahara, M.T. and Mahamudunnobi Sykot, M. and Hafiz, R. and Saifuzzaman, M.** [23]	**2020**	**Conferencia**	**Sector agricultura**
24	**Zhang, Y. and Zhang, D. and Miller, E.J., Prof.** [24]	**2021**	**Revista**	**Sector Inmobiliario**

Fuente: Elaboración propia.

La tabla 06, presenta la relación de los 24 artículos científicos aceptados y validados con la evaluación de calidad, el cual comprende la información de los autores, año de publicación, tipo de publicación y el sector del caso de uso.

b.7. Determinación de las técnicas y métricas más utilizadas

En esta sección, se desarrollaron las siguientes interrogantes:

b.7.1 ¿Cuáles son las técnicas de aprendizaje automático que se desarrollaron para estimar precios?

Según la revisión de artículos seleccionados, las técnicas empleadas para la estimación de precios son las de tipo aprendizaje supervisado, en la siguiente tabla, se lista las diferentes técnicas de aprendizaje encontradas y su presencia en los artículos revisados.

Tabla 7 - Técnicas de estimación de precios

Técnicas de aprendizaje	Presencia en investigaciones
Random Forest	8
Regresión Lineal	5
Árbol de decisión	4
XGBoost	4
SVR - Regresión de Vectores de Soporte	4
GBR - Regresión de aumento de Gradiente	4
K-NN vecinos más cercanos	4
SVM - Máquina de Vectores de Soporte	3
Red Neuronal Profunda (DNN), Regresión de crestas, Regresión de Lazo, Regresión Lineal Múltiple MLR, Perceptrón multicapa MLP, Naive Bayyes	2

Regresión de red elástica, Regresión de AdaBoost, y Red Neuronal Artificial ANN	1
Red Neuronal	1

Fuente: Elaboración propia.

b.7.2 ¿Cuál es la técnica de aprendizaje automático más utilizada para estimar precios?

En la siguiente tabla, se aprecia la técnica más utilizada en lo que se refiere la estimación de precios, resaltando para esta interrogante, la técnica de Random Forest (RF) con presencia en ocho (8) investigaciones y con un porcentaje de uso del 33.33% en comparación con las demás técnicas, y seguida por la técnica de Regresión Lineal (5) con 20.83%.

Tabla 8 - Técnicas más utilizadas para predecir precios

Técnicas de aprendizaje	Presencia en investigaciones	% de uso
Random Forest	**8**	33.33 %
Regresión Lineal	**5**	20.83 %
Arbol de decisión	**4**	16.66 %
XGBoost	**4**	16.66 %
SVR - Regresión de Vectores de Soporte	4	16.66 %
GBR - Regresión de aumento de Gradiente	4	16.66 %
K-NN vecinos más cercanos	4	16.66 %
SVM - Máquina de Vectores de Soporte	3	12.5 %
Red Neuronal Profunda (DNN), Regresión de crestas, Regresión de Lazo, Regresión Lineal Múltiple MLR, Perceptrón multicapa MLP, Naive Bayyes	2	8.33 %

Técnicas de aprendizaje	Presencia en investigaciones	% de uso
Regresión de red elástica, Regresión de AdaBoost, y Red Neuronal Artificial ANN	1	4.16 %
Red Neuronal	1	4.16 %

Fuente: Elaboración propia.

b.7.3 ¿Cuáles son las métricas para evaluar el resultado?

Los resultados de los artículos revisados, se ha verificado que, ocho (8) artículos no precisan las métricas para la evaluación de resultados, de las otras restantes, en la siguiente tabla, se enlista las métricas identificadas y su porcentaje de uso, siendo la métrica de Error cuadrático medio (RMSE) con presencia en alrededor de nueve (9) investigaciones, con un porcentaje de uso del 37.5%, seguido por las métricas de coeficiente de determinación (R2) y error absoluto medio (MAE), ambas presentes en cuatro (4) investigaciones con un porcentaje de uso del 16.66%,

Tabla 9 - Métricas más utilizadas para evaluar resultados de predicción de precios

Métricas de evaluación	Presencia en investigaciones	% de uso
Error cuadrático medio (RMSE)	9	37.75 %
Coeficiente de determinación (R2)	4	16.66 %
Error absoluto medio (MAE)	4	16.66 %
Error porcentual absoluto (MAPE)	2	8.33 %
Uso de datos	2	8.33 %
Validación cruzada generalizada mínima (GCV), Validación cruzada K-Fold, Error cuadrático medio de la raíz (RMLSE), R2 adj, Error porcentual absoluto (MAPE), Error cuadrático relativo y Error absoluto relativo.	1	4.16 %

Fuente: Elaboración propia.

c. Implementar las técnicas de regresión automática de machine learning para predecir costos.

- **Actividades previas al modelado**

  El registro del código inicia, con la importación de las siguientes librerías de Python, que se utilizaran para desarrollo del proyecto: a) Panda (pd), b) Numpy(np), c) Matplotlib.pyplot (plt) y d) Seaborn (sns).

  A continuación, se establece el enlace donde se encuentra la base de datos, se lee el archivo excel y luego se visualizan los primeros 05 registros para constatar que la conexión se haya realizado correctamente.

```
link = 'https://docs.google.com/spreadsheets/d/e/2PACX-1vSRbtBtT5d1lDr1JPNtCkREvPaCRYTCBlQ6TAP3iQCBdxFBPnwX6VK6xSVpzQL-AQ/pub?output=xlsx'

df = pd.read_excel(link)
df.head()
```

	MONTOREFERENCIAL	MES_FECHAPRESENTACIONPROPUESTA	mes	TIPOENTIDAD	ENTIDAD
0	84370.00	2018.0	2.0	GOBIERNO LOCAL	MUNICIPALIDAD DISTRITAL DE ALTO SELVA ALEGRE
1	140785.81	2018.0	4.0	GOBIERNO REGIONAL	GOBIERNO REGIONAL DE PIURA-SALUD
2	124082.00	2018.0	6.0	GOBIERNO REGIONAL	GOBIERNO REGIONAL DE CUSCO - PROYECTO ESPECIAL...
3	6201899.52	2018.0	11.0	GOBIERNO NACIONAL	IAFAS DEL EJERCITO DEL PERU (FOSPEME)
4	201630.00	2018.0	11.0	GOBIERNO LOCAL	MUNICIPALIDAD DISTRITAL DE CHALLHUAHUACHO

Ilustración 10 - Prueba de acceso al dataset

Con la finalidad de establecer la serie de tiempo, se agrupan los valores en la base de datos grupo_fecha, de acuerdo a los criterios de las columnas año y mes, y que muestre la cantidad de registros.

```
grupo_fecha = df.groupby(['año','mes']).agg({'Tipo':'count'}).reset_index()
grupo_fecha.head()
```

Se inserta un índice correlativo (row num), que permita identificar los 48 periodos o meses a ser analizados, con la finalidad de preparar la data para el entrenamiento y prueba.

```
grupo_fecha['row_num'] = np.arange(1,len(grupo_fecha)+1)
print (grupo_fecha)
```

Seguidamente, se agrupa la información de los campos: año, mes y tipo (tipo de entidad), con la condición, que se acumulen los montos de las diversas operaciones de adjudicaciones.

```
grupo_df = df.groupby(['año','mes','Tipo']).agg({'Mo
nto':'sum'}).reset_index()
grupo_df.head()
```

Luego se hace un cruce de información de la base de datos grupo_fecha con grupo_df, obteniéndose una tabla denominada df_total que contiene los siguientes campos año, mes, row_num, tipo de entidad y los montos.

```
df_total = pd.merge(grupo_fecha,df,how = 'left', on
=['año','mes'])
df_total.head()
```

Con la información trabajada, se proyecta un plot que permita visualizar 01 grafico por cada tipo de Entidad, el cual permita observar la tendencia de las adquisiciones de 48 meses correspondientes a los años 2018, 2019, 2020 y 2021.

```
g = sns.FacetGrid(melt, col='Tipo'
                      # ,row = 2
                       ,sharey=False
                  )
g.map(sns.lineplot, "row_num", "Monto")
```

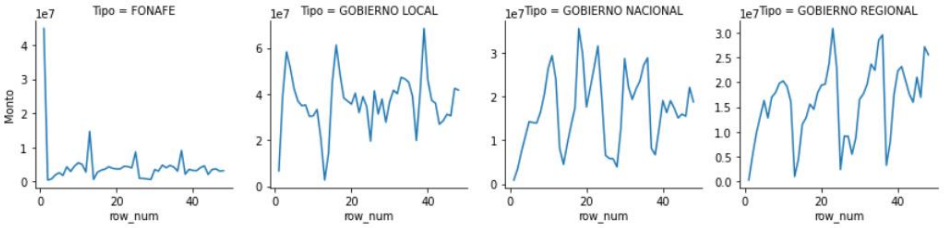

Ilustración 11 - Comparativo de presupuesto para compra de bienes por tipo de Entidad

Al conocerse la tendencia de las adquisiciones por tipo de entidad se procede con la creación de la variable que nos va a representar el valor futuro **(Variable Y),** considerado como **monto_tmas_1,** usando la agrupación de los campos Tipo y Monto.

```
melt['monto_tmas_1'] = melt.groupby('Tipo')['Monto']
.shift(-1)
```

De igual modo se procede con la creación de 03 predictores adicionales para la variable X (features): a) periodo anterior, b) diferencia de periodo anterior versus periodo actual, y c) media móvil

```
melt['monto_tmenos_1'] = melt.groupby('Tipo')['Monto
'].shift(1)


melt['diff_monto_1'] = melt.groupby('Tipo')['Monto']
.diff(1)


melt['MA_3']=melt.groupby('Tipo')['Monto'].rolling(3
).mean().reset_index(level=0,drop=True)
```

Seguidamente, se determina el porcentaje de partición de la data para determinar los datos que se asignen al entrenamiento 80% y prueba 20%.

```
split_point = int(len(grupo_fecha)*.80)
print(split_point)


melt_train = melt[melt['row_num'] < split_point].copy()
melt_valid = melt[melt['row_num'] >= split_point].copy()
```

Determinación de la matriz de caracterización, en esta etapa se identifican los campos que se relacionen con la variable X ,Y, siendo en el caso del X, los features son: Monto, monto_tmenos_1, diff_monto_1 y MA_3, y para el caso de la variable Y es: monto_tmas_1 y recuperando valores perdidos

```
from sklearn.impute import SimpleImputer


features =['Monto','monto_tmenos_1','diff_monto_1','MA_3']
melt_train = melt_train.dropna()


X_train = (melt_train[features])
```

Recuperar valores perdidos

```
print(len(X_train),len(y_train))


np.median(y_train)
```

- **Modelo 1: Árbol de Decisión de Regresión**

  Se crea el modelo de Árbol de Decisión con Regresión con la importación del modelo sklearn.tree y el algoritmo

DecisionTreeRegressor, y luego se asigna al modelo un coeficiente de regulación de 0.1, con la finalidad que no llegue al overfiting.

```
from sklearn.tree import DecisionTreeRegressor


mod1 = DecisionTreeRegressor(ccp_alpha = 0.1)
mod1.fit(X_train,y_train)
```

Se procede con el inicio de las predicciones con la data de validación o testing, considerando las variables (X,Y).

```
Xval = melt_valid[features]
yval = melt_valid['monto_tmas_1']
p = mod1.predict(Xval)
print(len(p))
```

A continuación, se evalúan las métricas del modelo R2,

```
#R2
mod1.score(X_train,y_train)
```

Para determinar las métricas que permitan medir el error del modelo y el coeficiente de determinación, se importa el modelo sklearn.metric, eligiendo las métricas MAE, MAPE, MSE, RMSE y R2, luego comparamos los valores reales con la data de validación, aplicando la eliminación de los elementos vacíos.

```
from sklearn.metrics import r2_score,
mean_squared_error,mean_absolute_error, mean_absolut
e_percentage_error


# comparamos el valor real vs prediccion con la data
 de validacion
comp = pd.DataFrame(np.concatenate([yval.values.resh
ape(-1,1),p.reshape(-1,1)],axis=1))


comp = comp.dropna()
```

- **Modelo 2: Regresión Lineal Múltiple**

Se crea el modelo de Regresión Lineal Múltiple con la importación del modelo sklearn.linear_model y el algoritmo LinearRegression.

```
from sklearn.linear_model import LinearRegression
mod1 = LinearRegression()
mod1.fit(X_train, y_train)
```

Se procede con el inicio de las predicciones con la data de validación o testing, considerando las variables (X,Y).

```
Xval = melt_valid[features]
yval = melt_valid['monto_tmas_1']
p = mod1.predict(Xval)
print(len(p))
```

Para determinar las métricas que nos permiten medir el error del modelo y el coeficiente de determinación, se importa el modelo sklearn.metric, eligiendo las métricas MAE, MAPE, MSE, RMSE y R2, luego comparamos los valores reales con la data de validación, aplicando la eliminación de los elementos vacíos.

```
from sklearn.metrics import r2_score,
mean_squared_error,mean_absolute_error, mean_absolut
e_percentage_error


# comparamos el valor real vs prediccion con la data
 de validacion
comp = pd.DataFrame(np.concatenate([yval.values.resh
ape(-1,1),p.reshape(-1,1)],axis=1))


comp = comp.dropna()
```

- **Modelo 3: Random Forest de regresión**

Se crea el modelo de Random Forest con Regresión con la importación del modelo sklearn.ensemble y el algoritmo RandomForestRegressor, y luego se asigna al modelo una estimación con 300 árboles.

```
from sklearn.ensemble import RandomForestRegressor
regression = RandomForestRegressor(n_estimators = 300, random
_state = 0)
regression.fit(X_train, y_train)
```

Se procede con el inicio de las predicciones con la data de validación o testing, considerando las variables (X,Y).

```
Xval = melt_valid[features]
yval = melt_valid['monto_tmas_1']
p = mod1.predict(Xval)
print(len(p))
```

Para determinar las métricas que nos permiten medir el error del modelo y el coeficiente de determinación, se importa el modelo sklearn.metric, eligiendo las métricas MAE, MAPE, MSE, RMSE y

R2, luego comparamos los valores reales con la data de validación, aplicando la eliminación de los elementos vacíos.

```
from sklearn.metrics import r2_score,
mean_squared_error,mean_absolute_error, mean_absolut
e_percentage_error


# comparamos el valor real vs prediccion con la data
 de validacion
comp = pd.DataFrame(np.concatenate([yval.values.resh
ape(-1,1),p.reshape(-1,1)],axis=1))


comp = comp.dropna()
```

- **Modelo 4: XGBoost Regresor**

  Se crea el modelo de Xgboost con regresión con la importación del modelo xgboost y el algoritmo XGBRegressor.

```
from xgboost import XGBRegressor
mod1 = XGBRegressor()
mod1.fit(X_train, y_train)
```

Se procede con el inicio de las predicciones con la data de validación o testing, considerando las variables (X,Y).

```
Xval = melt_valid[features]
yval = melt_valid['monto_tmas_1']
Xval.head()
p = mod1.predict(Xval)
```

Para determinar las métricas que nos permiten medir el error del modelo y del coeficiente de determinación, se importa el modelo sklearn.metric, eligiendo las métricas MAE, MAPE, MSE, RMSE y R2, luego comparamos los valores reales con la data de validación, aplicando la eliminación de los elementos vacíos.

```
from sklearn.metrics import r2_score,
mean_squared_error,mean_absolute_error, mean_absolut
e_percentage_error


# comparamos el valor real vs prediccion con la data
 de validacion
comp = pd.DataFrame(np.concatenate([yval.values.resh
ape(-1,1),p.reshape(-1,1)],axis=1))


comp = comp.dropna()
```

d. Realizar las pruebas del desempeño de las técnicas implementadas.

Al realizarse las pruebas de desempeño se obtuvieron los siguientes resultados:

d.1. **Modelo 1: Árbol de Regresión**

MAE = 6.16E+06 R2 = 0.53162

MAPE = 0.40 RAM = 1.53

MSE = 1.03E+14 Tiempo = 0.016

RMSE = 1.03E+14

d.2 **Modelo 2: Regresión Lineal Múltiple**

MAE = 4.03E+06 R2 = 0.79587

MAPE = 0.30 RAM = 1.21

MSE = 4.040E+13 Tiempo = 0.008

RMSE = 6.36E+06

d.3. **Modelo 3: Random Forest de regresión**

MAE = 5.43E+06 R2 = 0.68666

MAPE = 0.33 RAM = 1.21

MSE = 6.20E+12 Tiempo = 0.017

RMSE = 7.88E+06

d.4. **Modelo 4: XGBoost Regresor**

MAE = 5.97E+06 R2 = 0.61649

MAPE = 0.34 RAM = 1.49

MSE = 7.59E+13 Tiempo = 0.018

RMSE = 8.71E+06

IV. CONCLUSIONES Y RECOMENDACIONES

4.1. Conclusiones.

- Se elaboro el dataset de las adjudicaciones públicas 2018 – 2021, obteniéndose, luego del proceso de análisis y limpieza de datos, la cantidad de 42,789 adquisiciones (registros), que representa un presupuesto en adquisiciones de 1.638929e+05 soles, siendo el insumo para el despliegue de las labores de entrenamiento y prueba de los modelos de regresión a ser implementados.

- De la revisión sistemática de la literatura, se obtuvieron 24 artículos científicos (investigaciones) de los cuales, por su frecuencia en su aplicación, fueron seleccionados para su implementación los siguientes modelos de regresión de machine learning: Regresión Lineal Múltiple, Árbol de Decisiones, Bosque de aleatorio (Random Forest) y XGboost.

- La implementación de los modelos de regresión de machine learning fueron desarrollados en el lenguaje de programación Python de la aplicación colaboratory de Google.

- Como resultado de las pruebas de precisión de los modelos de regresión aplicados a las adjudicaciones publicas agrupados por nivel de gobierno fueron: en primer lugar, sitúa el modelo Regresión Lineal Múltiple con los siguientes índices de error MAE=4.03E+06, MAPE=0.30.MSE=4.04E+13, RMSE=6.36E+06 y R2 = 0.79587, en segundo lugar, se sitúa Random Forest con índices de MAE=5.43E+06, MAPE=0.33, MSE=6.20E+12, RMSE=7.88E+06 y R2 = 0.68666, en tercer lugar, se sitúa a XGboost con índices de MAE=5.97E+06, MAPE=0.34, MSE=7.59E+13, RMSE=8.71E+06 y R2 = 0.61649, y en cuarto lugar, Árbol de Decisiones con índices de MAE=6.16E+06, MAPE=0.40, MSE=1.03E+14, RMSE=1.03E+14 y R2 = 0.53162.

- Como resultado de las pruebas de desempeño de los modelos de regresión aplicados a las adjudicaciones publicas agrupados por nivel de gobierno fueron: Lineal Múltiple con los siguientes índices: tiempo de ejecución = 0.008s y 1.21 M.RAM, Random Forest con los siguientes índices: tiempo de ejecución = 0.017s y 1.21 M.RAM, XGboost con los siguientes índices: tiempo de ejecución = 0.018s y 1.49 M.RAM, y Árbol de Decisiones con los siguientes índices: tiempo de ejecución = 0.016s y 1.53 M.RAM.

4.2. Recomendaciones.

- Se recomienda continuar con la implementación de algoritmos de regresión a nivel de Entidad Pública, con la finalidad de identificar la demanda de recursos para el proceso de formulación del presupuesto institucional.

- Se recomienda utilizar el método propuesto en el presente trabajo, con fines de estimar los presupuestos en adquisiciones por tipo gobierno (Nacional, Regional, Local y Empresarial) y a nivel de cada Entidad Pública.

REFERENCIAS.

Bosch, A., Casas, J., & Lozano, T. (2019). Deep learning, Principios y fundamentos. Barcelona: Editorial UOC.

Cardenas, J. (2014). Clasificación automática de textos usando redes de palabras. *Scielo, 47*(86), 346-364. doi:http://dx.doi.org/10.4067/S0718-09342014000300001

Chen, Z., Li, C., & Sun, W. (2020). Bitcoin price prediction using machine learning: An approach to sample dimension engineering. Journal of Computational and Applied Mathematics, 365(112395), 1-13. doi:https://doi.org/10.1016/j.cam.2019.112395

Chowdhury, R., Rahman, M. A., Rahman, M. S., & Mahdy, M. (2020). An approach to predict and forecast the price of constituents and index of cryptocurrency using machine learning. Physica A: Statistical Mechanics and its Applications, 551(124569), 1-17. doi: https://doi.org/10.1016/j.physa.2020.124569

Costa, A., Ferreira, P., Gaglianone, W., Guillén, O., Issler, J., & Lin, Y. (2021). Machine learning and oil price point and density forecasting. Energy Economics, 102(105494), 1-21. doi: https://doi.org/10.1016/j.eneco.2021.105494Carolina, D., Matheus, D. A.,

Deepa, S., Alli, A., Sheetac, & Gokila, S. (2021). Machine learning regression model for material synthesis prices prediction in agriculture. Materials Today: Proceedings, 1-5. doi:https://doi.org/10.1016/j.matpr.2021.04.327

Deina, C., Do Amaral, M., Rodrigues, C., Ribeiro, M., Trojan, F., Stevan, S., & Valadares, H. (2021). A methodology for coffee price forecasting based on extreme learning machines. Information Processing in Agriculture, 1-29. doi: https://doi.org/10.1016/j.inpa.2021.07.003Dita, R., Nicholas, E. P., S, F. A., & Novita, H. (2021). Predicting Sneaker Resale Prices using Machine Learning. *Procedia Computer Science, 179*, 533-540. doi:https://doi.org/10.1016/j.procs.2021.01.037

Díaz, G., Coto, J., & Gómez, J. (2019). Prediction and explanation of the formation of the Spanish day-ahead electricity price through machine learning

regression. Applied Energy, 239, 610-625. doi: https://doi.org/10.1016/j.apenergy.2019.01.213

Fang, M., & Taylor, S. (2021). A machine learning based asset pricing factor model comparison on anomaly portfolios. Economics Letters, 204, 1-7. doi: https://doi.org/10.1016/j.econlet.2021.109919

F. Franco, E., & J. Ramos, R. (2019). Aprendizaje de máquina y aprendizaje profundo en biotecnología: aplicaciones, impactos y desafíos. Ciencia, Ambiente y Clima, 2(2), 7-26. Doi: https://doi.org/10.22206/cac.2019.v2i2.pp7-26

Gan, L., Wang, H., & Yang, Z. (2020). Machine learning solutions to challenges in finance: An application to the pricing of financial products. Technological Forecasting and Social Change, 153(119928), 1-11. doi: https://doi.org/10.1016/j.techfore.2020.119928Luciano, P., Rúben, P., & María, S. (2020). *Metodología de la investigación científica.* Buenos aires: Editoria Mauipe.

García, J. (2016). *Predicción en el dominio del tiempo. Analisis de series temporales para ingenieros.* Valencia: Universitat politécnica de Valéncia.

Gestión, D. L. (03 de Mayo de 2020). *Gestión*. Obtenido de https://gestion.pe/peru/politica/coronavirus-peru-fiscalia-investiga-15-denuncias-a-nivel-nacional-por-compras-sobrevaloradas-para-policia-nacional-pnp-covid-19-nndc-noticia/

Gil, I., Diaz, P., & Rodriguez, J. (2019). Técnicas y usos en la clasificación automática de imágenes. Universidad de Murcia, Murcia.

Hernandez, R., Fernández, C., & Baptista, P. (2014). Metodología de la investigación (Sexta ed.). México: Interamericana Editores.

Koo, E., & Kim, G. (2021). Prediction of Bitcoin price based on manipulating distribution strategy. Applied Soft Computing, 110, 1-10. doi: https://doi.org/10.1016/j.asoc.2021.107738

Kim, H.-M., Bock, G.-W., & Lee, G. (2021). Predicting Ethereum prices with machine learning based on Blockchain information. *Expert Systems with Applications, 184*(115480), 1-8. doi:https://doi.org/10.1016/j.eswa.2021.115480

Martínez, J. (28 de Mayo de 2019). *IArtificial.net*. Obtenido de https://www.iartificial.net/maquinas-de-vectores-de-soporte-svm/

Medrano, H. (7 de agosto de 2021). *El Comercio*. Obtenido de https://elcomercio.pe/lima/sucesos/essalud-tomografos-camas-uci-y-otros-equipos-que-se-pudieron-comprar-con-dinero-sobrevalorado-noticia/

Merayo, P. (mayo de 2020). *Máxima formación*. Obtenido de https://www.maximaformacion.es/blog-dat/que-son-los-arboles-de-decision-y-para-que-sirven/

Ministerio de Economía y Finanzas. (2018). Reglamento de la Ley N° 30225, Ley de Contrataciones del Estado.

Nolasco, J. (2018). *Python, Aplicaciones prácticas.* Madrid: RA-MA Editorial.

OSCE. (31 de Agosto de 2021). *CONOSCE - Portal de Datos Abiertos del OSCE*. Obtenido de CONOSCE - Portal de Datos Abiertos del OSCE: http://bi.seace.gob.pe/pentaho/api/repos/%3Apublic%3Aportal%3Adatosabiertosconvocatorias.html/content?userid=public&password=key

Rayon, A. (25 de Abril de 2017). *Deusto Data*. Obtenido de https://blogs.deusto.es/bigdata/guia-para-comenzar-con-algoritmos-de-machine-learning/

Rico, J., & Taltavull, P. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in Alicante, Spain. Expert Systems with Applications, 171, 1-14. doi: https://doi.org/10.1016/j.eswa.2021.114590Ligdi, G. (03 de Mayo de 2018). *Aprende IA*. Obtenido de https://aprendeia.com/ide-para-machine-learning-con-python/

Sandoval, L. (2018). Algortimos de aprendizaje automático para análisis y predicción de datos. *MACHINE LEARNING ALGORITHMS FOR DATA ANALYSIS AND PREDICTION.* ITCA - FEPADE, Santa Tecla, El Salvador. Obtenido de http://www.redicces.org.sv/jspui/bitstream/10972/3626/1/Art6_RT2018.pdf

Scherz, A. (2018). Clasificación automática de papers de Ciencias de la Computación. *(Tesis de licenciatura).* Universidad de Buenos Aires, Buenos Aires.

Truong, Q., Nguyen, M., Dang, H., & Mei, B. (2020). Housing Price Prediction via Improved Machine Learning Techniques. Procedia Computer Science, 174, 433-442. doi:https://doi.org/10.1016/j.procs.2020.06.111

Xu, H., Wang, M., Jiang, S., & Yang, W. (2020). Carbon price forecasting with complex network and extreme learning machine. *Physica A: Statistical mode*https://doi.org/10.1016/j.physa.2019.122830

Zhang, H., Nguyen, H., Vu, D.-A., Bui, X.-N., & Pradhan, B. (2021). Forecasting monthly copper price: A comparative study of various machine learning-based methods. *Resources Policy, 73*(102189), 1-6. doi: https://doi.org/10.1016/j.resourpol.2021.102189

Menoyo Ros, D. García López, E. y García Cabot, A. (2021). Fundamentos de la ciencia de datos. Alcalá de Henares, España, Editorial Universidad de Alcalá. Recuperado de https://elibro.net/es/ereader/bibsipan/177631?page=177.

ANEXOS.

Anexo 01: Resolución de aprobación del proyecto de investigación

UNIVERSIDAD SEÑOR DE SIPÁN

FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO

RESOLUCIÓN N°1179--2021/FIAU-USS

Pimentel, 10 de diciembre de 2021

VISTO:
El Acta de reunión N°1611-2021 del Comité de investigación de la Escuela profesional de INGENIERÍA DE SISTEMAS remitida mediante Oficio N°0382-2021/FIAU-IS-USS de fecha 24 de noviembre de 2021, y;

CONSIDERANDO:
Que, de conformidad con la Ley Universitaria N° 30220 en su artículo 48° que a letra dice: "La investigación constituye una función esencial y obligatoria de la universidad, que la fomenta y realiza, respondiendo a través de la producción de conocimiento y desarrollo de tecnologías a las necesidades de la sociedad, con especial énfasis en la realidad nacional. Los docentes, estudiantes y graduados participan en la actividad investigadora en su propia institución o en redes de investigación nacional o internacional, creadas por las instituciones universitarias públicas o privadas.";

Que, de conformidad con el Reglamento de grados y títulos en su artículo 21° señala: "Los temas de trabajo de investigación, trabajo académico y tesis son aprobados por el Comité de Investigación y derivados a la Facultad o Escuela de Posgrado, según corresponda, para la emisión de la resolución respectiva. El periodo de vigencia de los mismos será de dos años, a partir de su aprobación. En caso un tema perdiera vigencia, el Comité de Investigación evaluará la ampliación de la misma.

Que, de conformidad con el Reglamento de grados y títulos en su artículo 24° señala: La tesis es un estudio que debe denotar rigurosidad metodológica, originalidad, relevancia social, utilidad teórica y/o práctica en el ámbito de la escuela profesional. Para el grado de doctor se requiere una tesis de máxima rigurosidad académica y de carácter original. Es individual para la obtención de un grado; es individual o en pares para obtener un título profesional. Asimismo, en su artículo 25° señala: "El tema debe responder a alguna de las líneas de investigación institucionales de la USS S.A.C.".

Que, según documentos de Vistos el Comité de investigación de la Escuela profesional de INGENIERÍA DE SISTEMAS acuerdan aprobar los temas de las Tesis a cargo de los estudiantes que se detallan en el anexo de la presente Resolución.

Estando a lo expuesto, y en uso de las atribuciones conferidas y de conformidad con las normas y reglamentos vigentes;

SE RESUELVE:

ARTÍCULO 1°: APROBAR, el tema de la Tesis perteneciente a la línea de investigación de INFRAESTRUCTURA, TECNOLOGÍA Y MEDIO AMBIENTE, a cargo de los estudiantes del Programa de estudios de INGENIERÍA DE SISTEMAS según se detalla en el anexo de la presente Resolución.

ARTÍCULO 2°: ESTABLECER, que la inscripción del Tema de la Tesis se realice a partir de emitida la presente resolución y tendrá una vigencia de dos (02) años.

ARTÍCULO 3°: DEJAR SIN EFECTO, toda Resolución emitida por la Facultad que se oponga a la presente Resolución.

REGÍSTRESE, COMUNÍQUESE Y ARCHÍVESE

Mg. Victor Alexei Tuesta Montero
Decano (e) / Facultad De Ingeniería, Arquitectura Y Urbanismo
UNIVERSIDAD SEÑOR DE SIPÁN SAC.

MBA. María Noelia Sioler Rivera
Secretaria Académica / Facultad de Ingeniería, Arquitectura y Urbanismo
UNIVERSIDAD SEÑOR DE SIPÁN SAC.

Cc: Interesado, Archivo

UNIVERSIDAD SEÑOR DE SIPÁN

FACULTAD DE INGENIERÍA, ARQUITECTURA Y URBANISMO

RESOLUCIÓN N°1179--2021/FIAU-USS

Pimentel, 10 de diciembre de 2021

ANEXO

N°	AUTOR(ES)	TEMA DE TESIS
1	CABRERA SANCHEZ KEVIN ALONSO MENDOZA FERRE ESPERANZA NATALY	DESARROLLO DE UNA METODOLOGÍA DE GESTIÓN DE RIESGOS PARA MEJORAR LA DISPONIBILIDAD DE SERVICIO DE TI DE UN MUNICIPIO DISTRITAL
2	ROJAS ARRUNATEGUI JOEL ENRIQUE YAFAC LAU CESAR LEONIDAS	DESARROLLO DE UN MODELO DE PROCESOS PARA LA ADQUISICIÓN DE SOFTWARE BASADO EN LA NTP-ISO/IEC 12207 PARA MEJORAR LA GESTIÓN DE LAS ADQUISICIONES DE SOFTWARE EN MICROEMPRESAS PERUANAS
3	FERNANDEZ MALUQUIS JOSE EFRAIN	ANALISIS DE ALGORITMOS BALANCEADORES DE CARGA PARA UN CLÚSTER DE SERVIDORES PARA MEJORAR LA DISPONIBILIDAD DE UN SERVIDOR
4	RAMOS SANDOVAL FABIOLA ARACELY CANTORAL MONTEJO CESAR ENRIQUE	DESARROLLO DE UN MÉTODO DE CLASIFICACIÓN AUTOMÁTICA PARA LA DETECCIÓN EFICIENTE DEL RIESGO DE ANEMIA INFANTIL A PARTIR DE HÁBITOS DE ALIMENTACIÓN Y CUIDADOS
5	BOCANEGRA GUERRERO YERSON HUAMAN HUANCAS DERBIS	ANÁLISIS COMPARATIVO DE ARQUITECTURAS DE APRENDIZAJE PROFUNDO PARA LA CLASIFICACIÓN DE ROYA AMARILLA EN HOJAS DE CAFÉ
6	SANDOVAL CHERO CESAR ARTURO	MODELO DE LA GESTIÓN DE LA SEGURIDAD DE LA INFORMACIÓN ALINEADA A LA NORMA ISO/IEC 27001 ORIENTADO A LAS MICROEMPRESAS
7	DENNIS MAURICIO AVILES ODAR	APLICACIÓN DE BUENAS PRÁCTICAS PARA ENTORNOS DE DESARROLLO DE SOFTWARE BASADOS EN DEVOPS PARA MEJORAR LA INTEGRACIÓN Y DESPLIEGUE DE PROYECTOS EN UNA EMPRESA CONSULTORA DE LA CIUDAD DE CHICLAYO
8	RIVAS PLATA CASAS CARLOS GUALBERTO	DETECCIÓN DE CÁNCER DE PULMÓN EN IMÁGENES DE TOMOGRAFÍAS MEDIANTE PROCESAMIENTO DE IMÁGENES Y APRENDIZAJE AUTOMÁTICO
9	PECHE SANCHEZ CHRISTIAN WILFREDO	DISEÑO DE ARQUITECTURA DE MICROSERVICIOS PARA OPTIMIZAR PROCESOS EN LA GESTIÓN DE VENTAS ONLINE
10	SEVERINO HERNÁNDEZ YAMPIER GILBERTO	EVALUACIÓN DEL RENDIMIENTO DE UNA APLICACIÓN WEB CON ARQUITECTURA DE MICROSERVICIOS SOPORTADOS EN LA NUBE EN UN AMBIENTE DE ALTA CONCURRENCIA
11	CHANG HIDALGO HAWARD MIGUEL	COMPARACIÓN DE TÉCNICAS DE ESTIMACIÓN BASADAS EN MACHINE LEARNING PARA PREDECIR COSTOS EN LOS PLANES DE ADQUISICIONES DE LAS ENTIDADES PÚBLICAS DEL PERÚ
12	PUICON PISFIL MIRIAN ALICIA VILCHEZ CHANGANAQUI RICHARD ALEXIS	DESARROLLO DE UN MODELO DE PROCESOS BASADO EN ESTÁNDARES PARA LA EVALUACIÓN DE LA USABILIDAD WEB PARA MICROEMPRESAS PERUANAS
13	LOPEZ ABANTO GUILLERMO ANTONIO	EVALUACIÓN DE LA SEGURIDAD DE UN SISTEMA DE VOTACIÓN ELECTRÓNICA CON BLOCKCHAIN
14	CALDERON ZUÑIGA JESUS TELLO TANTARICO DILSON GUZMAN	DESARROLLO DE UN MODELO DE GOBERNANZA DE TI BASADO EN MARCOS DE GOBIERNO Y GESTIÓN DE TECNOLOGÍAS DE LA INFORMACIÓN PARA INSTITUCIONES PÚBLICAS PERUANAS

Anexo 02: Población de algoritmos

N°	Algoritmo
1	Árbol de decisión
2	Autoregresivo (AR)
3	CatBoost
4	Cresta Lineal
5	Lazo Lineal
6	Multicapa Redes Neuronales Perceptron (MLP)
7	AdaBoost
8	Análisis Discriminante Cuadrático (QDA)
9	Análisis Discriminante Lineal (LDA)
10	Árboles de aumento de gradiente (GBT)
11	Árboles de regresión de aumento de gradiente (GBRT)
12	Bosque Aleatorio (RF)
13	Extreme Learning Machine (ELM)
14	K vecinos más cercanos (KNN)
15	Maquina de Vectores de Soporte (SVM)
16	Red Neuronal de Memoria a Largo Plazo (LSTM)
17	Modelos de media móvil e integrada autoregresiva (ARIMA)
18	Red Elástiva
19	Red Neuronal Artificial (RNN)
20	Regresión de árbol de decisiones potenciada (BDR)
21	Regresión de componentes principales (PCR)
22	Regresión del Bosque de Decisión (DFR)
23	Regresión Lineal (LR)
24	Regresión Lineal Bayesiana (BLR)
25	Regresión Logística
26	Suavisado exponencial (ES)
27	XGBoost (XGB)

Fuente: elaboración propia

Anexo 03: Ficha digital de observación – Desempeño del algoritmo

N°	Nombre	Consumo de Memoria	Tiempo de respuesta
1			
2			
3			
4			

Fuente: elaboración propia

Anexo 04: Ficha digital de observación – Precisión del algoritmo

N°	Nombre	MAE (Error absoluto medio)	MAPE (Error porcentual absoluto medio)	MSE (Error cuadrático medio)	RMSE (Raíz del error cuadrático medio)	R2 (Coeficiente de determinación)
1						
2						
3						
4						

Fuente: elaboración propia

Anexo 05: Estructura de datos del dataset preliminar

DICCIONARIO DE DATOS - DATASET CONVOCATORIA O INVITACION		
Nombre del camp	**Descripción del campo**	**Tipo de dat**
CODIGOENTIDAD	Código de la entidad	Texto
ENTIDAD_RUC	RUC de la entidad convocante	Texto
ENTIDAD	Nombre de la entidad convocante	Texto
TIPOENTIDAD	Tipo de la entidad, puede ser gobierno local, regional, entre otros	Texto
CODIGOCONVOCATORIA	Código de la convocatoria	Número
DESCRIPCION_PROCESO	Descripción del proceso de la selección. Esta información es registrada por la Entidad	Texto
PROCESO	Nomenclatura del proceso de selección	Texto
TIPOCOMPRA	Tipo de compra. Puede ser: por la Entidad. Encargo o Compra Corporativa	Texto
OBJETOCONTRACTUAL	Objeto Contractual. Puede ser Bien, Servicio, Obra, Consultoria de Obras	Texto
SECTOR / NIVEL DE GOBIERNO	Nombre del sector de gobierno al cual esta adscrito la entidad. Puede ser Salud, Educación, Economía, entre otros	Texto
SISTEMA CONTRATACION	Sistema de contratación del proceso de selección. Puede tener los siguientes valores: Tarifas, Precios Unitarios, Costo Reembolsable, Suma Alzada, Honorario Fijo y comisión de Éxito, Precios Unitarios, Precios Unitarios, tarifas o porcentajes.	Texto
TIPOPROCESOSELECCION	Tipo de proceso de selección. Pueden los correspondientes a la Ley de Contrataciones del Estado (como subasta inversa electrónica, adjudicación simplificada, licitación pública, entre otros) como a otros regímenes (como Proceso Especial de Contratación, Contratación Internacional, etc)	Texto
MONTOREFERENCIAL	Valor referencial (o valor estimado) total del proceso de selección en moneda original. Debe recordarse que, para el caso de procedimientos de bienes y servicios convocados bajo el ámbito de la Ley de Contrataciones, se utiliza el "valor estimado", y este no es público mientras el procedimiento aún se encuentre en marcha.	Número
N_ITEM	Número del ítem	Número
DESCRIPCION_ITEM	Descripción del ítem del proceso de selección. Un proceso de selección puede convocarse según relación de ítems.	Texto
UNIDAD_MEDIDA	Unidad de medida del ítem. Es elegido por la Entidad	Texto
ESTADOITEM	Estado del ítem del proceso de selección. Puede contener los siguientes valores: Adjudicado, Consentido o Contratado	Texto
PAQUETE	Un ítem puede ser un paquete o combo. Es habitual en Alimentos, como por ejemplo, adquisición de canastas. Este campo es sólo Indicador de paquete del ítem: SI o NO.	Número
CODIGOITEM	Código del quinto nivel del Catálogo Único de Bienes, Servicios y Obras (CUBSO). Este código es elegido por la Entidad. Para el caso de Ítem Paquete, este campo estará vacío.	Texto
ITEMCUBSO	Descripción del quinto nivel del Catálogo Único de Bienes, Servicios y Obras (CUBSO), el cual ha sido elegido por la Entidad. Para el caso de Ítem paquete, este campo está vacío	Texto
DISTRITO_ITEM	Nombre del distrito del Ítem	Texto
PROVINCIA_ITEM	Nombre de la provincia del ítem	Texto
DEPARTAMENTO_ITEM	Nombre del departamento del ítem	Texto
MONTO_REFERENCIAL_ITEM	Monto referencial del ítem en moneda original	Número
MONEDA	Es la moneda en la que se ha definido los montos para el proceso de selección.	Número
FECHA_CONVOCATORIA	Fecha de la convocatoria o fecha de invitación	Fecha
FECHAINTEGRACIONBASES	Fecha de la integración de las bases. Para los procedimientos en los que no es aplicable esta acción (por ejemplo Contrataciones Directas o procesos de regímenes distintos a los regulados por la Ley de Contrataciones), este campo está vacío y por defecto el reporte genera la fecha 01/01/1900.	Fecha
FECHAPRESENTACIONPROPUESTA	Fecha de presentación de propuestas. Para los procedimientos en los que no es aplicable esta acción, este campo está vacío y por defecto el reporte genera la fecha 01/01/1900.	Fecha

Fuente: Portal de datos abiertos del OSCE.

Anexo 06: Estructura de datos del dataset final

N°	Nombre del campo	Descripción del campo	Tipo de dato
1	COD_ENTIDAD	Código de la entidad	Texto
2	ENTIDAD_RUC	RUC de la entidad convocante	Texto
3	NOM_ENTIDAD	Nombre de la entidad convocante	Texto
4	TIPO_ENTIDAD	Tipo de la entidad, puede ser gobierno local, regional, entre otros	Texto
5	COD_CONVOCATORIA	Código de la convocatoria	Número
6	DESC_PROCESO	Descripción del proceso de la selección. Esta información es registrada por la Entidad	Texto
7	NOM_PROCESO	Nomenclatura del proceso de selección	Texto
8	TIPO_COMPRA	Tipo de compra. Puede ser: por la Entidad. Encargo o Compra Corporativa	Texto
9	SECTOR	Nombre del sector de gobierno al cual esta adscrito la entidad. Puede ser Salud, Educación, Economía, entre otros	Texto
10	SIST_CONTRATACION	Sistema de contratación del proceso de selección. Puede tener los siguientes valores: Tarifas, Precios Unitarios, Costo Reembolsable, Suma Alzada, Honorario Fijo y comisión de Éxito, Precios Unitarios, Precios Unitarios, tarifas o porcentajes.	Texto
11	TIPO_PROCESO_SELECCION	Tipo de proceso de selección. Pueden los correspondientes a la Ley de Contrataciones del Estado (como subasta inversa electrónica, adjudicación simplificada, licitación pública, entre otros) como a otros regímenes (como Proceso Especial de Contratación, Contratación Internacional, etc)	Texto
12	MONTO_REFERENCIAL 1	Valor referencial (o valor estimado) total del proceso de selección en moneda original. Debe recordarse que, para el caso de procedimientos de bienes y servicios convocados bajo el ámbito de la Ley de Contrataciones, se utiliza el "valor estimado", y este no es público mientras el procedimiento aún se encuentre en marcha.	Número
13	N_ITEM	Número del ítem	Número
14	DESC_ITEM	Descripción del ítem del proceso de selección. Un proceso de selección puede convocarse según relación de ítems.	Texto
15	UNIDAD_MEDIDA	Unidad de medida del ítem. Es elegido por la Entidad	Texto
16	COD_ITEM	Código del quinto nivel del Catálogo Único de Bienes, Servicios y Obras (CUBSO). Este código es elegido por la Entidad. Para el caso de Ítem Paquete, este campo estará vacío.	Texto
17	ITEM_CUBSO	Descripción del quinto nivel del Catálogo Único de Bienes, Servicios y Obras (CUBSO), el cual ha sido elegido por la Entidad. Para el caso de Ítem paquete, este campo está vacío	Texto
18	DISTRITO_ITEM	Nombre del distrito del Ítem	Texto
19	PROVINCIA_ITEM	Nombre de la provincia del ítem	Texto
20	DEPARTAMENTO_ITEM	Nombre del departamento del ítem	Texto
21	MONTO_REF_ITEM	Monto referencial del ítem en moneda original	Número
22	MONEDA	Es la moneda en la que se ha definido los montos para el proceso de selección.	Número
23	FECHA_PROPUESTA	Fecha de la convocatoria o fecha de invitación	Fecha
24	AÑO_FECHA_PROPUESTA	Año de la fecha de la convocatoria o fecha de invitación	Fecha
25	MES_FECHA_PROPUESTA	Mes de la fecha de la convocatoria o fecha de invitación	Fecha

Fuente: Elaboración propia.