Aplicación de un algoritmo de aprendizaje automático para mejorar la estimación a la terminación de un proyecto (EAC) en una empresa del sector Oil & Gas con base en métricas de gestión de valor ganado (EVM)

Fredy David Infante Gonzalez

Asesor

Rafael Roberto Ruiz Escorcia

Universidad Nacional Abierta y a Distancia UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería ECBTI

Ciencias de Datos y Analítica

2025

Resumen

Este trabajo propone aplicar un modelo predictivo de aprendizaje automático para mejorar la estimación del costo final de proyectos en el sector Oil & Gas, utilizando métricas de gestión del valor ganado (EVM). La gestión del valor ganado es una metodología tradicionalmente aceptada para estimar costos finales, pero presenta diferencias significativas con los costos reales debido a diversos factores. El escenario esperado es implementar un algoritmo de aprendizaje automático que, basado en datos de seguimiento y métricas de EVM, genere estimaciones más precisas, optimizando la toma de decisiones de los Project Management Professional (PMP). La metodología incluye un diseño cuantitativo de enfoque correlacional, utilizando datos de fuentes privadas organizacionales (Project Online y Excel) y librerías (Pandas, Numpy, Matplotlib, Scikit-learn, etc) de Python para análisis el descriptivo y el desarrollo de los modelos. Se espera validar el modelo propuesto con métricas como el Error cuadrático medio para reconocer la precisión de las estimaciones respecto a los datos reales.

***Palabras clave*:** Aprendizaje automático, EVM, Costos, Proyecto, Estimación, Métricas

Abstract

This work proposes to develop a machine learning predictive model to improve the final cost estimation of projects in the Oil & Gas sector using earned value management (EVM) metrics. Earned value management is a traditionally accepted methodology for estimating final costs, but it shows significant differences from actual costs due to various factors. The expected scenario is to implement a machine learning algorithm that, based on tracking data and EVM metrics, generates more accurate estimations, optimizing decision-making for Project Management Professionals (PMP). The methodology includes a quantitative design with a correlational approach, using data from organizational private sources (Project Online and Excel) and Python libraries (Pandas, Numpy, Matplotlib, Scikit-learn, etc.) for descriptive analysis and model development. The proposed model is expected to be validated with metrics such as Mean Squared Error to recognize the accuracy of the estimations compared to real data.

Keywords**:** Machine learning, EVM, Cost, Project, Estimation, Metrics

Tabla de Contenido

Lista de Tablas

Lista de Figuras

Introducción

La precisión en la estimación del costo final de los proyectos es un desafío recurrente en el sector Oil & Gas. La gestión del valor ganado (EVM, por sus siglas en inglés) es una metodología reconocida para esta tarea, pero presenta variaciones significativas cuando se compara con los costos reales debido a diferentes factores.

La idea central de este trabajo de grado es desarrollar un modelo predictivo de aprendizaje automático que ayude a mejorar la estimación del costo final de un proyecto en el sector de Oil & Gas y se integre con el análisis de métricas de EVM. Se espera optimizar la toma de decisiones para los Project Management Professional (PMP), bajo la premisa de que un algoritmo de aprendizaje automático puede ayudar a generar mejores estimaciones de los costos finales que la metodología clásica de EVM por sí sola.

El modelo se entrenará utilizando una base de datos que contiene características relacionadas con el proyecto, como el tiempo de ejecución y métricas previamente calculadas de EVM (PV y AC). Con estas estimaciones se pretende calcular el Estimate at Completion (EAC) del proyecto.

Por razones de confidencialidad, los nombres de los proyectos y de la empresa propietaria de la información se mantendrán en anonimato, ya que las bases de datos utilizadas son corporativas y privadas.

Una vez finalizado este proyecto, se pretende proponer el modelo para su implementación como un esfuerzo de investigación para la mejora continua, sin embargo, esta fase no está dentro del alcance del presente trabajo de grado y será la organización la que decida la viabilidad de su uso interno. Esta propuesta busca cerrar la brecha existente entre las

estimaciones tradicionales y los costos reales, habilitando una gestión de proyectos más precisa y eficiente en el sector Oil & Gas con el apoyo de nuevos saberes y nuevas tecnologías.

Descripción del Problema

Planteamiento del Problema

La organización es una importante empresa del sector y para la gestión de sus proyectos sigue minuciosamente los lineamientos del estándar del PMI, por lo tanto, la metodología para el seguimiento de los proyectos se hace con métricas de análisis de valor ganado, con base en estos análisis se realizan estimaciones futuras de los costos asociados. Aunque no existen predicciones cien por ciento precisas de eventos futuros para proyectos únicos, las métricas de valor ganado suelen generar estimaciones aceptables, pero con un grado considerable de imprecisión e incertidumbre.

La imprecisión en la predicción de costos al finalizar un proyecto es un desafío crítico que puede tener consecuencias significativas, como sobrecostos, falsas promesas de entrega, ineficiencia en la asignación de recursos y daño a la reputación de la empresa y los gerentes de proyectos. Esta imprecisión a menudo se debe al uso inadecuado de estimaciones, la dificultad para estimar actividades únicas, el empleo de datos irrelevantes y la falta de consideración del contexto específico del proyecto. Sin embargo, los avances en inteligencia artificial y algoritmos de aprendizaje automático ofrecen una solución prometedora. Estos modelos predictivos pueden analizar grandes volúmenes de datos históricos y métricas de gestión del valor ganado (EVM), proporcionando estimaciones de costos más precisas. La implementación de estas tecnologías puede optimizar la toma de decisiones, mejorar la eficiencia operativa y aumentar la competitividad, mitigando las consecuencias negativas de las estimaciones inexactas y fortaleciendo la posición de la organización en el competitivo sector Oil & Gas.

Sistematización del Problema

Este apartado tiene como objetivo identificar y organizar las diversas dimensiones que constituyen la problemática central a través del desglose en preguntas específicas que servirán para orientar el desarrollo de este proyecto. De acuerdo con lo anterior, se plantean las siguientes interrogantes:

¿Cuáles son las limitaciones actuales de la metodología EVM en la estimación de costos finales?

¿De qué manera un algoritmo de aprendizaje automático puede integrarse con las métricas EVM para mejorar la precisión del EAC?

¿Qué variables adicionales (además de PV y AC) son relevantes para la predicción del costo final en proyectos del sector Oil & Gas?

¿Cómo se puede validar la efectividad del modelo propuesto en un entorno corporativo?

Justificación

Según el Project Management Institute (2022) la gestión del valor ganado (EVM) es una metodología de gestión que integra el alcance, el cronograma y los recursos para medir objetivamente el desempeño y el progreso del proyecto, y para pronosticar los resultados de este. Esta metodología es una herramienta tradicionalmente aceptada para la estimación del costo al finalizar un proyecto.

Sin embargo, a menudo existen diferencias significativas entre las estimaciones de costos iniciales y los costos reales debido a diversos factores como el uso inadecuado de estimaciones, la dificultad para estimar los costos de actividades únicas, el uso de datos irrelevantes y las estimaciones fuera de contexto (Cezar-Petre e Irinel, 2018). Estas incorrectas estimaciones pueden resultar en graves consecuencias como sobrecostos y falsas promesas de entrega, afectando la eficiencia del proyecto, la asignación adecuada de recursos y la reputación de la compañía y de los gerentes de proyectos ante los grupos de interés.

Gracias al avance de nuevas tecnologías como la inteligencia artificial y algoritmos computaciones para la estimación de costos, se desea aplicar algunas estas herramientas para generar modelos predictivos que proporcionen pronósticos más precisos, ayudando a mitigar las potenciales consecuencias de estimaciones inexactas, especialmente en cuanto a los costos finales de un proyecto.

Objetivos

Objetivo General

Implementar un algoritmo de aprendizaje automático que optimice la predicción de costos futuros de un proyecto con base en la técnica de gestión del valor ganado (EVM).

Objetivos Específicos

Realizar el análisis exploratorio de datos mediante técnicas estadísticas.

Transformar los datos utilizando técnicas de programación y herramientas propias de las ciencias de datos para su posterior procesamiento.

Desarrollar un modelo de aprendizaje automático para la predicción de variables cuantitativas.

Evaluar el modelo y la precisión de sus predicciones respecto a los valores esperados.

Marco de Referencia

Estado del Arte

Antecedentes Relevantes

Evaluation of EVM/ES forecasting methods in hospital construction projects. El objetivo de este trabajo fue evaluar 22 métodos de pronóstico del tiempo y 12 del costo, basados en el Valor Ganado (EVM/ES), bajo tres supuestos de desempeño futuro y en términos de la oportunidad, precisión y estabilidad de los pronósticos (Durón et al., 2022).

Comparison of neural network, Gaussian regression, support vector machine, long short-term memory, multi-gene genetic programming, and M5 Trees methods for solving civil engineering problems. En esta investigación, se examinó cómo los métodos de aprendizaje automático (ML) rinden en distintos problemas con características diversas. Se emplearon seis enfoques de ML que incluyen redes neuronales artificiales (ANN), regresión de procesos gaussianos (GPR), regresión con máquinas de vectores de soporte (SVMR), memoria a corto plazo (LSTM), programación genética multigénica (MGGP) y el árbol de modelo M5 (M5Tree) para analizar tres problemas independientes en ingeniería civil (Uncuoglu et al., 2022).

A machine learning study to improve the reliability of project cost estimates. En esta investigación, se utiliza el aprendizaje automático (ML) para aumentar la confiabilidad en la predicción de costos de proyectos. Se desarrolla un modelo de pronóstico basado en XGBoost y se llevan a cabo experimentos computacionales utilizando datos reales de 110 proyectos, que comprenden 1268 puntos de datos de costos (Narbaev et al., 2024). Es impresionante cómo la integración de técnicas avanzadas como XGBoost puede mejorar significativamente la precisión en la estimación de costos, aspecto crucial para una gestión de proyectos más efectiva y eficiente. Al aprovechar datos reales de una amplia muestra de proyectos, este estudio

proporciona insights valiosos que pueden transformar las prácticas tradicionales de gestión de costos en la industria.

A method for project completion cost predicting using LSTM in earned value management technique. En este artículo, se presenta un método de cuatro pasos que utiliza redes LSTM para predecir el costo final de un proyecto. El enfoque propuesto emplea conjuntos de datos precalculados, utilizados en las técnicas de Gestión del Valor Ganado (EVM), como datos de entrada. El modelo de red LSTM procesa estos valores en su capa de entrada para estimar el costo de finalización del proyecto (Le et al., 2020).

Hybrid machine learning model performance in IT project cost and duration prediction. Este estudio tiene como objetivo desarrollar un modelo híbrido de aprendizaje automático que sea altamente confiable y eficiente, con el fin de mejorar la precisión en la predicción de costos y duración. Para lograr esto, se comparó el rendimiento de cinco modelos de aprendizaje automático en tres conjuntos de datos distintos, evaluándolos mediante seis indicadores de rendimiento (Pang, 2023).

Development of machine learning models for prediction of IT project cost and duration. Este estudio busca una solución para mejorar la capacidad de predicción mediante el uso de un modelo de aprendizaje automático (ML). Se llevó a cabo un experimento que comparó el rendimiento de cada modelo de ML, utilizando tres conjuntos de datos distintos y catorce modelos diferentes, evaluados con seis indicadores de desempeño (Pang, 2022).

Integrated earned value method for IT project cost and duration estimation. Este estudio tiene como objetivo desarrollar un Método Integrado de Valor Ganado (IEVM) que sea altamente confiable y eficiente para mejorar la precisión en la predicción de costos y duración. Para lograr esto, se realizó un experimento que comparó el rendimiento de cinco modelos de

aprendizaje automático en tres conjuntos de datos distintos y seis indicadores de rendimiento, y se verificaron los modelos con otros tres tipos de datos de proyectos en vivo (Pang, 2023)

Impacto de la estandarización y escalado: factor para predicción de costos en proyectos a través de una red neuronal artificial. Este estudio presenta una comparación entre los métodos de estandarización y escalado utilizados en la predicción de costos. Se aplicaron cuatro técnicas de estandarización y escalado para el preprocesamiento de datos; posteriormente, los datos fueron procesados mediante una red neuronal artificial (RNA) (Rodríguez González et al., 2021).

Reference Class Forecasting and Machine Learning for Improved Offshore Oil and Gas Megaproject Planning: Methods and Application. Este estudio presenta el desarrollo y descripción de métodos rigurosos para el pronóstico en proyectos de petróleo y gas. Utiliza un enfoque predictivo basado en aprendizaje automático (ML) que considera factores específicos del proyecto para estimar con mayor precisión los costos y los posibles excesos en el cronograma (Natarajan, 2022).

Marco Contextual

Este proyecto se enmarca en una empresa líder del transporte de hidrocarburos en el sector Oil & Gas, que sigue rigurosamente los lineamientos del PMI y utiliza el análisis del valor ganado para la gestión de proyectos. El entorno corporativo, caracterizado por la alta inversión y la complejidad de las operaciones, demanda soluciones innovadoras para optimizar la gestión de costos y recursos.

Además, la empresa está comprometida con la implementación de tecnologías de vanguardia y la adopción de metodologías ágiles que le permitan responder eficientemente a las fluctuaciones del mercado energético global. La integración de herramientas digitales avanzadas, como sistemas de información geográfica y análisis predictivo, potencia la capacidad de

anticipar riesgos y optimizar la asignación de recursos. Este enfoque estratégico no solo mejora la toma de decisiones, sino que también impulsa la sostenibilidad y la responsabilidad ambiental, aspectos cruciales en el sector Oil & Gas actual.

Marco Teórico

Gestión de Valor Ganado (EVM): La gestión del valor ganado (EVM) es una técnica de gestión de proyectos que se utiliza para encontrar desviaciones en los proyectos en función de la comparación del trabajo realizado y el trabajo planificado. EVM es una técnica útil en la previsión de proyectos. Se considera como una técnica para controlar la ejecución del proyecto, combinando tanto el control de costos como el progreso del proyecto (Le et al. 2020).

Además, la gestión del valor ganado es una metodología de gestión para integrar el alcance, el cronograma y los recursos; para medir objetivamente el desempeño y el progreso del proyecto; y para pronosticar los resultados del proyecto (Project Management Institute, 2011).

Además, según el Project Management Institute (2011) indica que una vez establecidas las líneas base del proyecto, estas se convierten en la mejor fuente para comprender el rendimiento del proyecto durante la ejecución. Una comparación del desempeño real (tanto del costo como del cronograma) con estas líneas base se proporciona información sobre el estado y los datos del proyecto, no solo para proyectar resultados probables, sino también para que la gerencia tome decisiones oportunas y útiles utilizando datos objetivos.

Como metodología de gestión del desempeño, EVM agrega algunas prácticas críticas al proceso de gestión de proyectos. Estas prácticas se llevan a cabo principalmente en las áreas de planificación, ejecución y control de proyectos, y están relacionadas con el objetivo de medir, analizar, pronosticar e informar sobre los datos de rendimiento de costos y cronogramas para su evaluación y acción por parte del equipo del proyecto y otras partes interesadas clave.

Métricas EVM

Costo real (AC): El costo realizado incurrido por el trabajo realizado en una actividad durante un período de tiempo específico. Esto se puede reportar para el acumulado hasta la fecha o para un período de reporte específico. También puede conocerse como el costo real del trabajo realizado (ACWP) (Project Management Institute, 2011).

Presupuesto al finalizar (BAC): La suma de todos los presupuestos establecidos para el trabajo que se va a realizar en un proyecto, componente de estructura de desglose del trabajo, cuenta de control o paquete de trabajo. El BAC del proyecto es la suma de todos los BAC del paquete de trabajo. (Project Management Institute, 2011).

Valor ganado (EV): La medida de la obra realizada, expresada en términos del presupuesto autorizado para dicha obra. El valor acumulado se puede reportar para el acumulado hasta la fecha o para un período de informe específico. También puede conocerse como el costo presupuestado por el trabajo realizado (CPTR) (Project Management Institute, 2011).

Estimación al finalizar (EAC): El costo total esperado de completar todo el trabajo expresado como la suma del costo real hasta la fecha y la estimación para completar (ETC) (Project Management Institute, 2011).

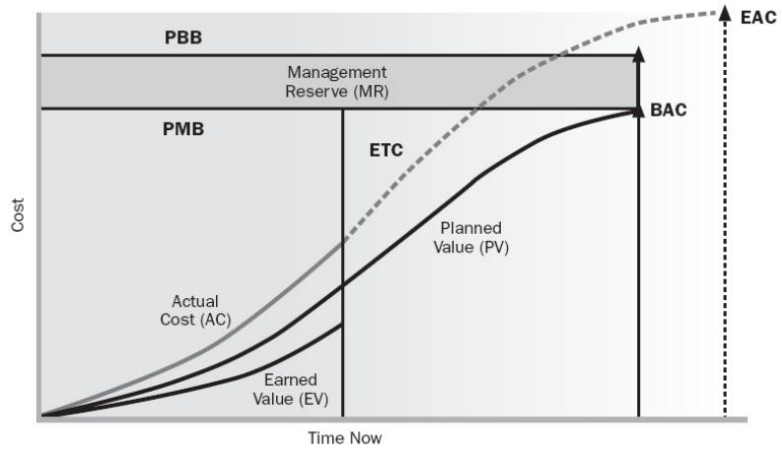
Estimación para completar (ETC): El costo estimado para terminar todo el trabajo restante. Agregar el ETC al costo real (AC) dará como resultado la estimación al finalizar (EAC) en cualquier nivel del proyecto (Project Management Institute, 2011).

Valor planificado (PV): El presupuesto autorizado asignado al trabajo programado a partir de una fecha de informe determinada. En cualquier momento, el valor planificado define el trabajo que debería haberse realizado. El valor planificado se puede notificar para el trabajo

acumulado hasta la fecha o para un período de informe específico. También se conoce como el costo presupuestado para el trabajo programado (Project Management Institute, 2011).

Figura 1

Representación de Relación de Términos



Nota. Tomada de (Project Management Institute, 2011).

Uso del Aprendizaje Automático en la Estimación de Costos

En años recientes se ha visto como se ha incrementado considerablemente el uso de herramientas basadas en la inteligencia artificial para utilizar las máquinas para facilitar la toma de decisiones en diversos sectores e industrias. El pronóstico de costos y otras variables han comenzado a ser objetivo de estas nuevas herramientas para mejorar la precisión al utilizarse grandes conjuntos de datos. En el campo de gestión de proyectos, donde es importante generar proyecciones más precisas se ha venido utilizando en buena medida.

En la gestión de proyectos, hay muchas áreas de aplicación y beneficios potenciales del uso de la analítica de datos y la IA (Munir 2019; Ong y Uddin 2020, como se citó en Narbaev et al , 2024). Sin embargo, como indica Narvaev et al. (2024) la bibliografía sobre las aplicaciones es escasa. Esta escasez puede estar relacionada con el hecho de que cada proyecto es único, y la

predicción de los resultados del proyecto utilizando datos históricos limitados es ineficaz. Sin embargo, en cualquier industria, muchas actividades de diferentes proyectos son comunes.

Existen una gran cantidad de algoritmos de inteligencia artificial y del aprendizaje automático que son utilizados para la estimación de variables en la gestión de proyectos. Varias técnicas de regresión, redes neuronales, árboles de decisión, clustering y algoritmos de reconocimiento de patrones podrían considerarse técnicas de aprendizaje automático (Narvaev et al., 2024).

Es por tal motivo que otros autores han investigado y trabajado en el tema de la aplicación del aprendizaje automático, como Elmousalami (2021) que también implementa ML para apoyar la toma de decisiones en las primeras etapas de los proyectos, específicamente en las estimaciones conceptuales de costos. Integran factores de incertidumbre en los modelos utilizando la teoría difusa (Narvaev et al., 2024).

Wauters y Vanhoucke (2016) utilizan árboles de decisión, bagging, bosque aleatorio y técnicas de impulso para predecir el tiempo de finalización del proyecto y comparar el rendimiento de estas técnicas. Generan sus datos para las pruebas y, utilizando estos datos, muestran que todos estos métodos de ML tienen mejores capacidades de predicción que los métodos de pronóstico tradicionales (Narvaev et al., 2024).

Algunos estudios recientes sugieren el uso de ML en la previsión de duración y costes (por ejemplo, Pellerin y Perrier 2018; Willems y Vanhoucke 2015). Willems y Vanhoucke (2015) afirman que las técnicas de ML tienen como objetivo aprender de la experiencia (por ejemplo, los patrones de gasto de costos hasta la fecha actual) y aplicar este conocimiento en nuevas situaciones (por ejemplo, para pronosticar el costo final del proyecto) (Narvaev et al., 2024).

Narvaev et al. (2024), cómo los demás autores citados en su propia investigación utilizaron el aprendizaje automático para estimar los costos de proyectos.

Marco Normativo

Se incluirá la normativa y estándares relevantes para la gestión de proyectos, tales como:

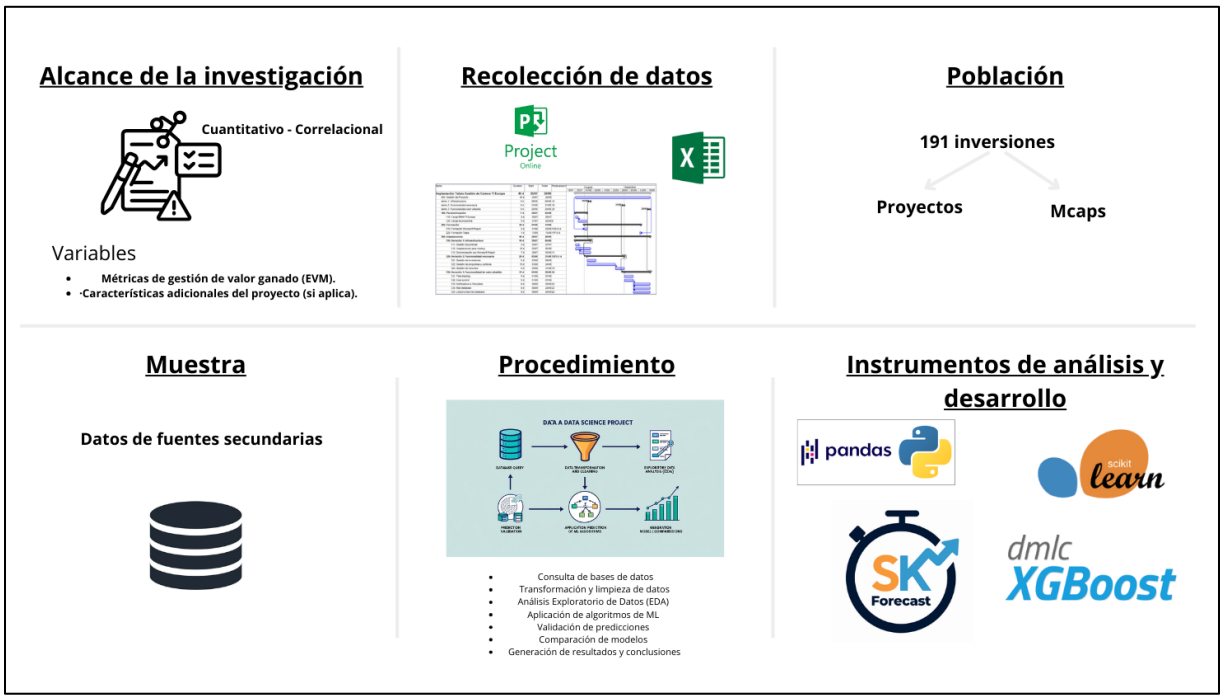
- Guias del Project Management Institute (PMI).
- Normativas internas de la empresa.
- Regulaciones específicas del sector Oil & Gas que afectan la estimación y control de costos.
- Ley 1581 de 2012 (2012). Ley 1581 de 2012. Congreso de la República de Colombia. [https://www.alcaldiabogota.gov.co/sisjur/normas/Norma1.jsp?i=49981]: esta ley busca fortalecer el derecho constitucional de cada persona a acceder, actualizar y corregir la información que se ha recopilado sobre ella en bases de datos o archivos. Además, reafirma los demás derechos, libertades y garantías constitucionales mencionados en el artículo 15 de la Constitución Política, y refuerza el derecho a la información consignado en el artículo 20 de la misma.
- Consideraciones éticas y tratamiento de información

Se garantizará la confidencialidad y el anonimato de los participantes. Esto incluye a la organización, nombres de proyectos y de procesos, para garantizar que la difusión de datos sensibles no perjudique a los sujetos de la investigación. Todos los datos se almacenarán de manera segura y se utilizarán exclusivamente para fines de investigación.

Metodología

Figura 2

Resumen de Principales Aspectos de la Metodología



Alcance de la Investigación

Esta investigación utiliza un diseño cuantitativo de tipo correlacional, ya que la variable o característica de interés es de carácter cuantitativo o numérica. Se estudia y compara el desempeño de un modelo de naturaleza matemática o estadística y como este influye en el mejoramiento de la estimación de la variable de interés. El alcance correlacional se aplica al analizar la relación entre variables independientes o predictoras y las variables dependientes.

Variables

- Métricas de gestión de valor ganado (EVM).
- Características adicionales del proyecto (si aplica).

Recolección de Datos

Los datos de los proyectos con las variables de interés son de carácter privado que se pueden extraer de dos opciones de fuentes de datos: Project Online y Excel. Ambas fuentes tienen falencias en la calidad de los datos, que deben ser analizadas al detalle, para definir la fuente apropiada para realizar el análisis de resultados luego de implementar las herramientas de análisis y modelado. Los datos están disponibles para usuarios con permisos exclusivos. En el caso de la fuente de Excel, el equipo de seguimiento y control de proyectos se encarga de consolidar y actualizar los datos de las métricas de EVM y de almacenarlos de manera periódica en un repositorio en la nube.

Para el caso de la fuente de Project Online los profesionales de seguimiento y control se encargan de actualizar el avance del cronograma en el software de gestión de proyectos y las métricas de EMV se calculan automáticamente. Esta fuente, está en estado de implementación inicial, por lo que no se tiene completitud en los datos, sin embargo, es la que está mejor parametrizada y con el tiempo la organización esperaría utilizar, ya que esta herramienta reduciría la probabilidad de error humano y por ende mejorar la calidad de la información capturada.

Población

La población objetivo son 191 inversiones las cuáles se dividen en proyectos y mantenimientos capitalizables. Ambos tipos de inversión se analizan bajo las mismas métricas de valor ganado.

Muestra

Se extrajeron los datos de fuentes secundarias, los datos son generados al recopilarse los datos de seguimiento de las métricas EVM de los proyectos en Excel y Project Online.

Procedimiento

Comenzar consultando las bases de datos disponibles, transformando y limpiando los datos para garantizar su integridad. Luego, realizar un Análisis Exploratorio de Datos (EDA) para descubrir patrones y tendencias clave. Aplicar algoritmos de aprendizaje automático para construir modelos predictivos, definiendo el más adecuado según su desempeño. Validar las predicciones con técnicas estadísticas, comparando los modelos para seleccionar el óptimo. Finalmente, generar resultados y conclusiones que aporten insights valiosos y fortalezcan la toma de decisiones en el proyecto.

Instrumentos de Análisis y Desarrollo

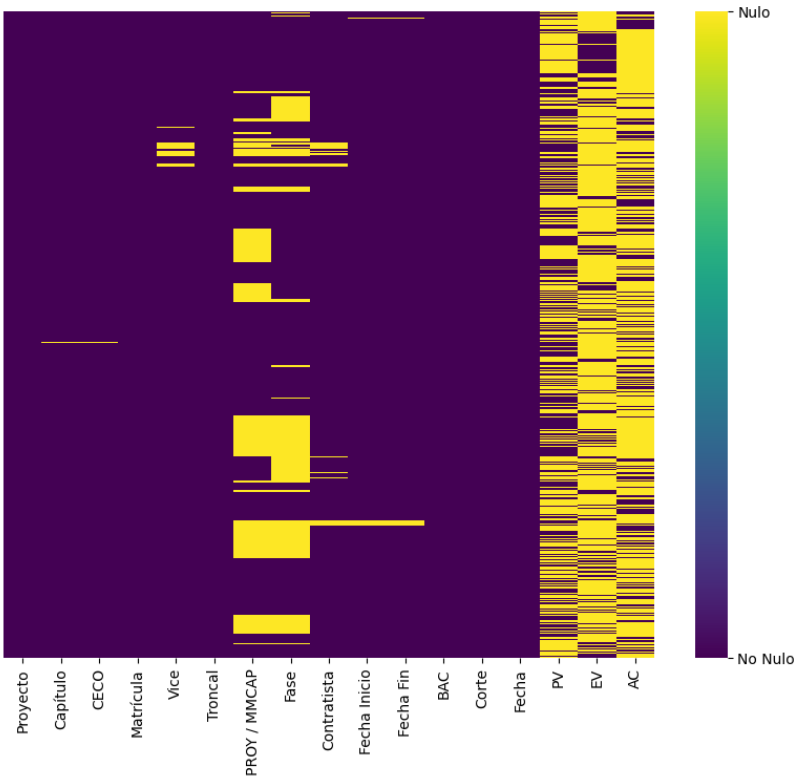
Se realizará un análisis exploratorio de datos (EDA) exhaustivo sobre las bases de datos de los proyectos utilizando Python o Power BI, con el objetivo de encontrar patrones, tendencias, datos atípicos y errores, y proceder a una limpieza y depuración detallada de los datos. Posteriormente, se aplicarán modelos de aprendizaje automático apropiados para el conjunto de datos analizado. Para validar la precisión de los resultados obtenidos, se emplearán métricas específicas de validación de algoritmos de regresión, como el error cuadrático medio y el coeficiente de determinación.

Se realizó un análisis exploratorio de datos con el fin de cumplir con los objetivos de este proyecto, ya que este preprocesamiento de los datos es necesario para conocer como están compuestos los datos de estudio y encontrar patrones, distribuciones, tendencias y relaciones. Para este análisis se utilizaron las librerías disponibles en Python y los conocimientos adquiridos durante la especialización en ciencias de datos, librerías como Pandas, Numpy, Matplotlib, Seaborn, plotly, etc y técnicas estadísticas. El análisis fue principalmente descriptivo con enfoque correlacional.

Se hizo un análisis de valores faltantes para la base de datos original, el cual se puede observar en la figura 3.

Figura 3

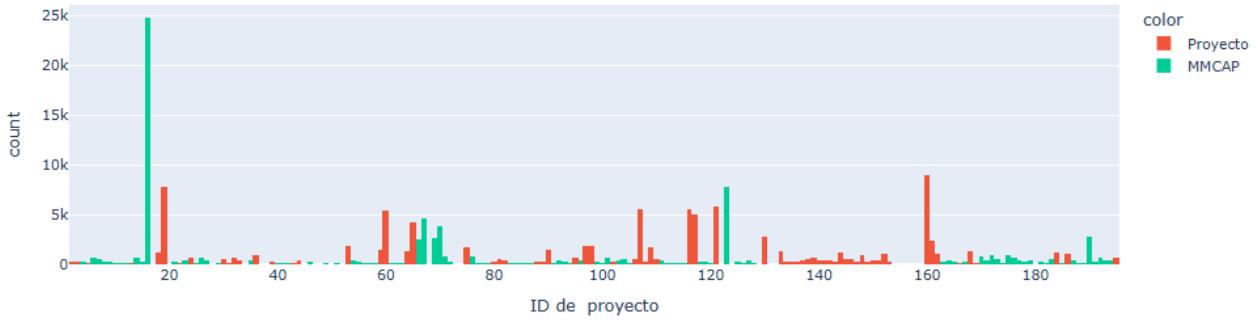
Validación de Valores Faltantes-Datos Originales



Se corrigió el nombramiento de las etiquetas de la variable PROY / MCAP, las cuál nos indica el tipo de inversión. Existen únicamente dos tipos de inversión: Proyecto y MMCAP (Mantenimientos capitalizables). Se observa que luego de realizar el ajuste, se aumenta el número de registros de tipo Proyecto, que se encontraba divido en otras etiquetas mal escritas o redundantes figura 4.

Figura 4

Distribución Corregida de las Inversiones

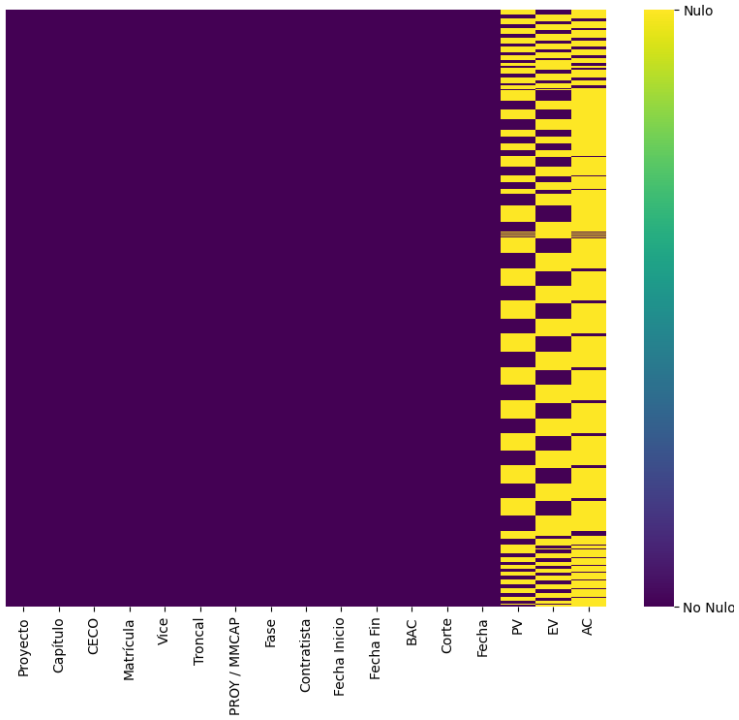


En la figura 4 se observa cómo se distribuyen los datos para los dos tipos de inversión (Proyecto y MMCAP), Los MMCAP cuentan con una inversión que contiene una gran cantidad de registros, sin embargo, este tipo de inversiones tiene unas características financieras diferentes a las de tipo Proyecto, debido a que este trabajo tiene como propósito el análisis y optimización de estimación de costos de un proyecto, definimos el proyecto que cuenta con la mayor cantidad de datos. Por lo que se acota el conjunto de datos y se hace el análisis para el proyecto Numero 160.

Al conjunto de datos acotados únicamente para el proyecto 160, se le realizó su respectivo análisis de valores faltantes, en donde se observa que la cantidad de valores faltantes se redujo considerablemente, sin embargo, aún se observa que hay una gran cantidad de valores faltantes para las variables PV, AC y EV, según la figura 5.

Figura 5

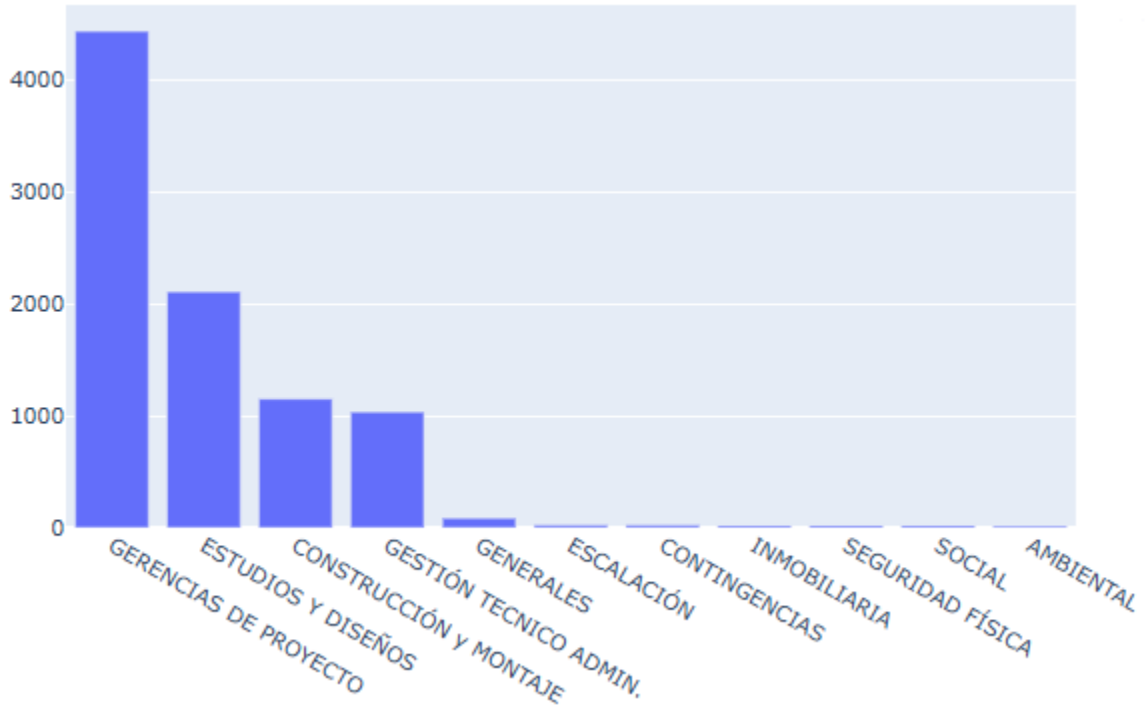
Validación de Valores Faltantes Proyecto 160



Se analizaron las distribuciones de las variables categóricas por medio de gráficos de barras, en donde se observa que la mayoría de los datos para la variable Capitulo, están asociados a la “Gerencia proyecto”. Por otro lado, vemos que no se logra una diferencia significativa para la distribución de los datos para la variable CECO. Para la variable Vice, se observa que la mayoría de los datos pertenecen a la clase Poliductos, y que las Troncales con mayores registros son “Magdalena Medio” y “Central.”

Figura 6

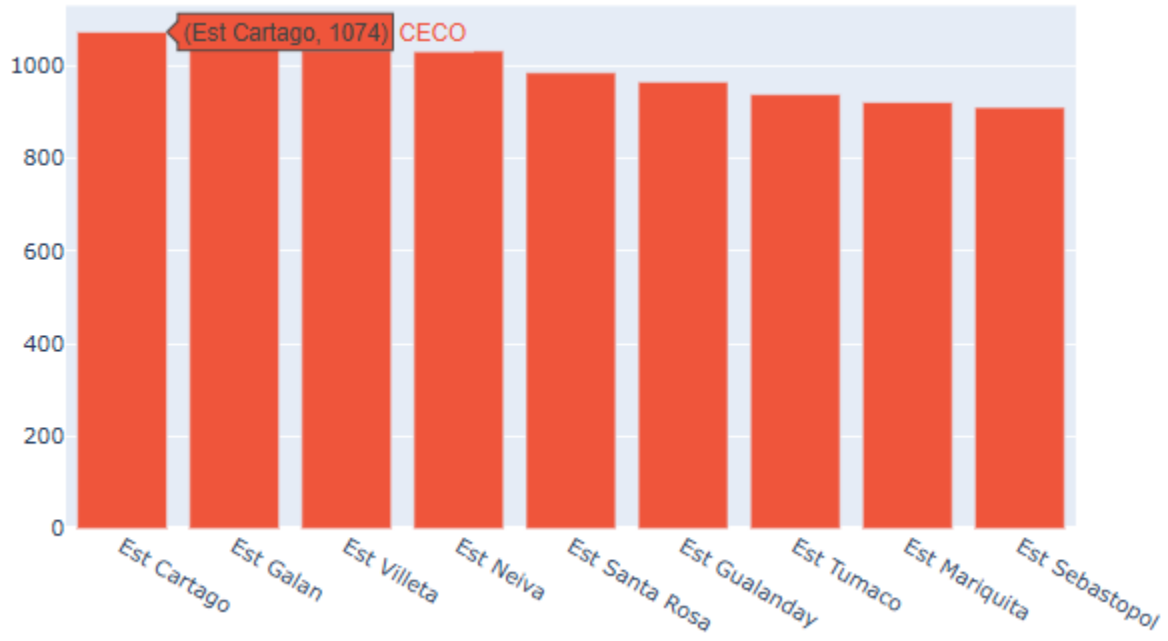
Distribución por Variable "Capitulo"



En la figura 6, se puede observar la distribución por la variable capítulo, donde se puede notar que la gerencia de proyectos es el capítulo que contiene más información o costos asociados, seguido por Estudios y diseños y construcción y montaje.

Figura 7

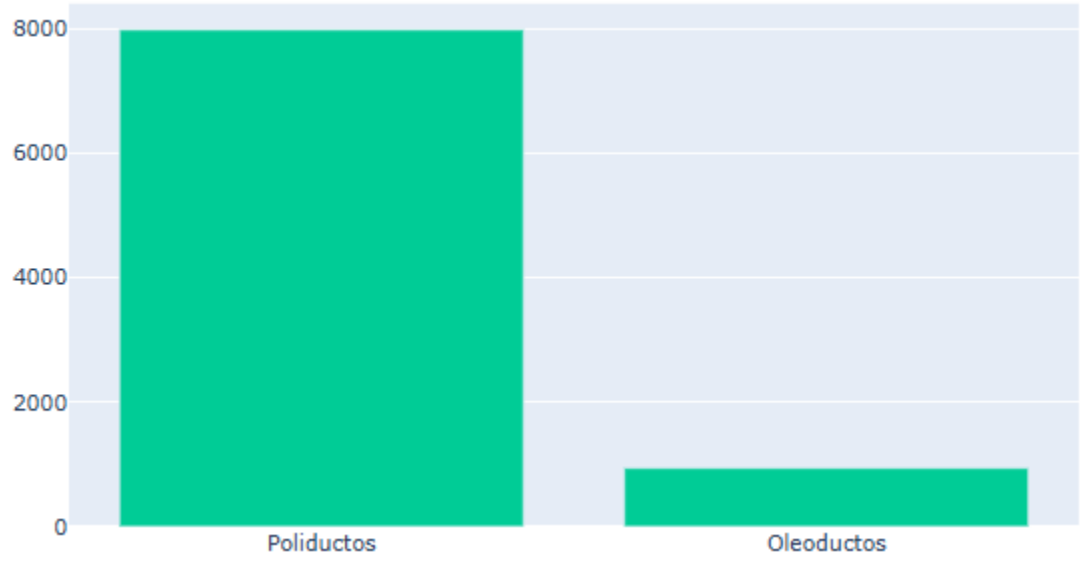
Distribución por Variable "CECO"



La figura 7, muestra la distribución por CECO o centros de costos, también llamadas estaciones, Se observa que para el proyecto 160, la estación con más información o costos asociados es la Estación Cartago, además se observa que no hay diferencian importantes entre las estaciones.

Figura 8

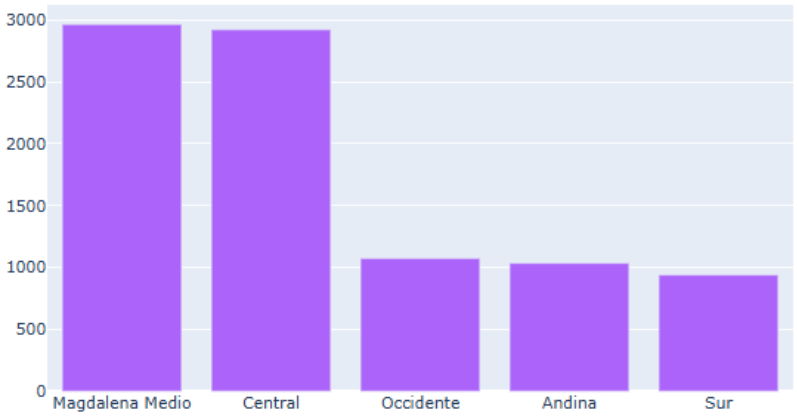
Distribución por Variable "Vicepresidencia"



En la Figura 8 se observa que la vicepresidencia con más información de manera significativa es la vicepresidencia de Poliductos.

Figura 9

Distribución por Variable "Troncal"



En la figura 9, podemos notar que las troncales que son más afectadas por el proyecto son la troncal Magdalena Medio y Central, aproximadamente contienen tres veces más información que la troncal Occidente, Andina y Sur.

Una vez realizado el análisis exploratorio de datos y las transformaciones respectivas, se identificaron las variables que serían utilizadas para la serie de tiempo y que serviría para entrenar los modelos de predicción, los resultados de la serie de tiempo se observan en la figura 10. Se desarrollaron tres modelos, un modelo tradicional ARIMA el cual sus resultados se observan en la figura 11 y dos modelos basados en XGboost para la predicción de una serie de tiempo univariada que se presenta en la figura 12 y figura 13.

Se utilizó el modelo autoARIMA para encontrar los parámetros de cada uno de los componentes AR, I, MA, los resultados de este modelo se observan en la figura 11.

Una vez desarrollado el modelo ARIMA como modelo de referencia para comparar con los modelos de aprendizaje automático, se desarrollaron dos modelos XGboost, uno basado en la librería SKforecast, el cuál optimiza el tiempo requerido en el preprocesamiento manual de la

base de datos y además permite trabajar fácilmente con otras variables exógenas el cual se muestra en la figura 8. También se desarrolló el modelo XGboost de la librería scikit-learn, sin embargo, este necesitó un preprocesamiento, al que se le aplicó el suavizamiento de los datos con la técnica de suavizado exponencial (SE) para reducir el ruido y la transformación de la serie de tiempo para que el modelo pudiera ser entrenado correctamente, se observa en la figura 13.

Una vez desarrollados los tres modelos, se compararon y se evaluaron, para determinar el modelo con el mayor rendimiento, utilizando como métricas de evaluación el MSE y el MAE, los resultados se muestran en la tabla 1.

Resultados

Primer Resultado

La base de datos inicial abarca 227,267 registros y 17 variables, reflejando inversiones en proyectos y mantenimientos capitalizables. El proyecto con mayor número de registros es el Proy ID - 160, con 8,932 entradas, seleccionado para un análisis detallado debido a su relevancia y magnitud; este proyecto se desarrolla en las troncales Magdalena Medio y Central. La variable con más datos faltantes es AC (*Actual Cost: Costo real*), de tipo Float, con solo 987 datos existentes, lo que podría impactar en la precisión del análisis. Entre las variables numéricas se incluyen Proyecto, Matrícula, Fase, BAC, Corte, PV, EV y AC; las categóricas son Capítulo, CECO, Vice, Troncal, PROY/MMCAP y Contratista; y las variables de tipo Datetime comprenden Fecha Inicio, Fecha Fin y Fecha. Es notable que no existe correlación lineal entre las variables numéricas, lo que sugiere relaciones más complejas que podrían explorarse mediante análisis multivariado. Esta ausencia de correlación lineal plantea interrogantes sobre la influencia de factores no cuantificados y abre la oportunidad para profundizar en técnicas estadísticas avanzadas, potenciando así la optimización en la gestión de costos y recursos en futuros proyectos.

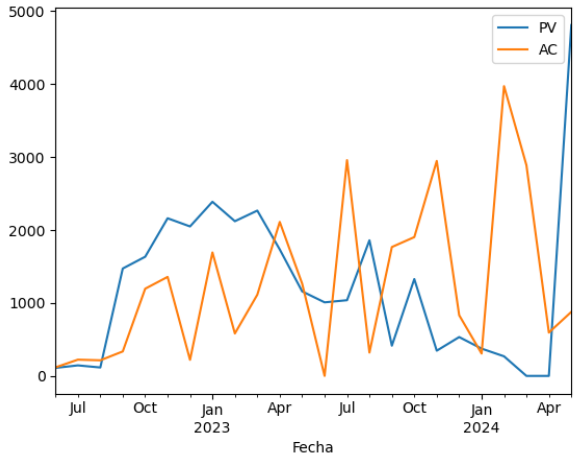
Segundo Resultado

Utilizando Python y las librerías Pandas y NumPy para la transformación y limpieza de datos, se corrigieron 495 datos en la variable "PROY / MMCAP", depurando y seleccionando tres variables clave: 'Fecha', 'PV' y 'AC'. Se agrupó por Fecha y se realizó la agregación por suma de los valores AC y PV; al completar esta agrupación, se eliminaron los valores faltantes en AC debido a que la frecuencia original era semanal. Posteriormente, se efectuó un remuestreo y ajuste de frecuencias de fecha, cambiando de semanas a meses, lo cual permitió definir una serie

temporal univariada con frecuencia mensual. Se eliminaron los valores en cero por no contribuir al modelo a desarrollar, y tras las transformaciones realizadas, la serie de tiempo consta de 24 períodos, iniciando desde el 30-06-2022 hasta el 31-05-2024.

Figura 10

Series de Tiempo para la Variable PV y AC



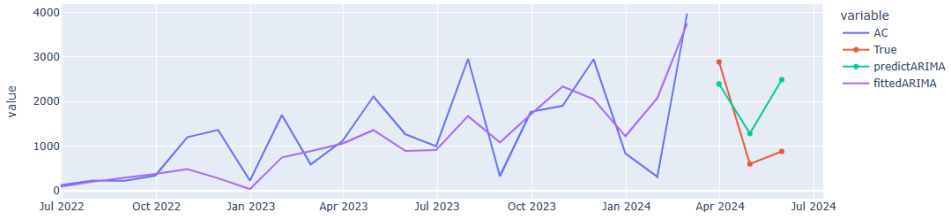
Tercer Resultado

Los resultados del análisis de serie de tiempo revelan que la prueba de Dickey-Fuller arrojó un p-valor de 0.1, superior al umbral de 0.05, lo que indica que la serie no es estacionaria; el análisis de descomposición confirmó la existencia de una tendencia, pero no se identificó un patrón de estacionalidad. Al aplicar el modelo ARIMA tradicional utilizando la librería autoarima en Python, se determinaron los parámetros óptimos (3,1,1): el componente AR considera tres valores pasados de la serie, I aplica una diferenciación de primer orden, y MA toma un valor pasado de los errores. El modelo presentó un Criterio de Información de Akaike (AIC) de 333.974, y la prueba de Ljung-Box reportó un p-valor de 0.76, sugiriendo que, para un rezago de 1, el modelo captura eficazmente las dependencias temporales; además, se obtuvo una

métrica MSE (Error Cuadrático Medio) de 1,103,604 y MAE (Error Medio Absoluto) de 929.99, reflejando la precisión del modelo en la predicción de los datos.

Figura 11

Serie de Tiempo y Predicción del Modelo ARIMA



Cuarto Resultado

Se empleó Skforecast, una biblioteca de Python para pronóstico de series temporales que integra modelos de aprendizaje automático, utilizando el algoritmo XGBoost para regresión, el cual aplica árboles de decisión y estimaciones iterativas para refinar las predicciones. Se incorporó la variable exógena "PV" al modelo para mejorar la estimación, logrando una métrica MSE (Error Cuadrático Medio) de 535,781 y un MAE (Error Medio Absoluto) de 611.02. Por otro lado, se implementó XGBoost con Suavización Exponencial Doble, dado que la serie presenta tendencia sin estacionalidad, realizando transformaciones adicionales a los datos para validar posibles mejoras; sin embargo, este enfoque resultó en una métrica MSE de 1,605,589 y MAE de 1118.54, indicando un desempeño menor en comparación con el modelo que incluyó la variable exógena.

Figura 12

Xgboost SKforecaste con Variable Exógena

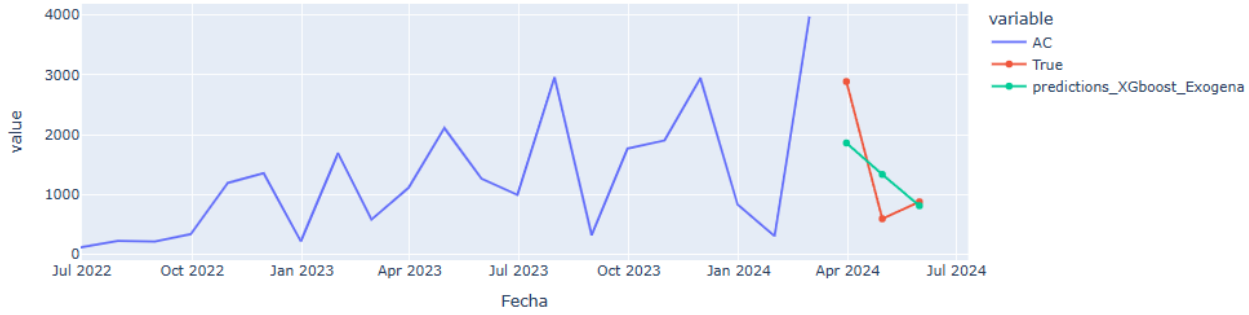
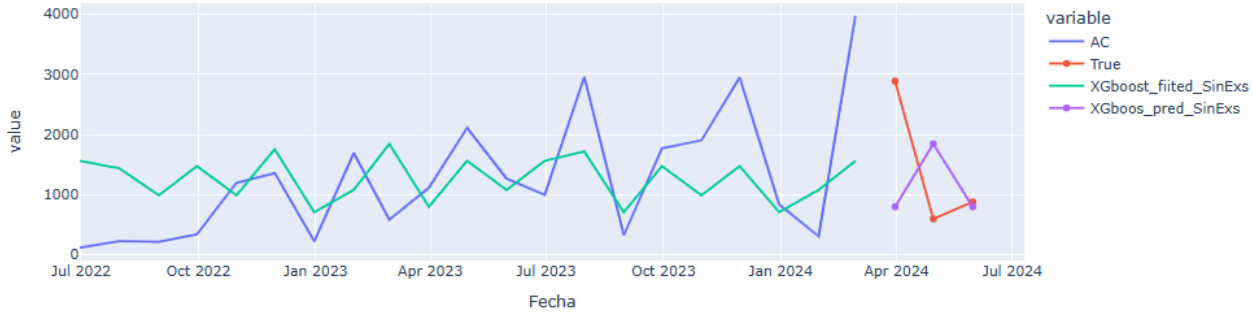


Figura 13

Xgboost Scikit-Learn con Suavizado Exponencial



Quinto Resultado

Comparando los resultados de los tres modelos, el XGBoost con variable exógena "PV" logró el mejor desempeño con un MSE de 535,781 y MAE de 611.02, indicando mayor precisión en las predicciones. El modelo ARIMA obtuvo un MSE de 1,103,604 y MAE de 929.99, siendo menos preciso que el anterior. Por último, el XGBoost con Suavización Exponencial Doble presentó el MSE más alto de 1,605,589 y MAE de 1118.54, reflejando el desempeño más bajo entre los tres.

Tabla 1

Comparativa de los Tres Modelos Desarrollados por Métricas MSE y MAE

Modelo	MSE	MAE
Modelo ARIMA	1.10e+06	929.9
Modelo skforescast xgboost	5.36e+05	611.02
Xgboost + SE	1.60e+06	1118.5

Nota. Comparación de los tres modelos, ARIMA y de aprendizaje automático, tomando como métricas de evaluación el MSE y MAE, se muestra que el modelo skforescast xgboost genero el mejor resultado.

Conclusiones

El análisis exploratorio permitió identificar que la base de datos inicial contenía 227,267 registros y 17 variables, de las cuales la variable con mayor cantidad de datos faltantes fue **AC (Costo real)**, la cual es la variable de interés, con solo 987 datos existentes. Este hallazgo resalta la importancia de abordar problemas de datos incompletos en etapas tempranas para garantizar la calidad del análisis posterior. Además, se determinó que no existía correlación lineal entre las variables numéricas, lo que indica que no se podían asumir relaciones directas entre las mismas.

El conocimiento adquirido sobre la estructura de los datos, como la categorización en variables numéricas, categóricas y tipo Datetime, fue fundamental para comprender la naturaleza de la información y su potencial uso en modelos predictivos. En particular, la agrupación y visualización de datos por tipo de inversión (proyectos y mantenimientos capitalizables) permitió identificar patrones iniciales en las troncales Magdalena Medio y Central, aportando contexto valioso al problema abordado.

Este análisis exploratorio fue esencial para enmarcar el proyecto 160 dentro de un problema de predicción en series de tiempo, ya que permitió seleccionar y limpiar las variables más relevantes. La identificación de datos como la serie temporal univariada del AC, que abarca 24 periodos entre junio de 2022 y mayo de 2024, estableció la base para el desarrollo de modelos que permitan la optimización de costos de este y otros proyectos.

Las transformaciones realizadas en la base de datos, como el agrupamiento por fecha, la eliminación de valores faltantes y el remuestreo de frecuencias semanales a mensuales, facilitaron la definición de una serie temporal univariada más manejable y consistente para el análisis. Esto demuestra que técnicas de programación (Pandas, Numpy) son esenciales para preparar datos en problemas complejos.

El uso de herramientas de programación permitió no solo limpiar los datos, sino también optimizar el conjunto de variables. De las 17 variables originales, se seleccionaron tres claves para el análisis: Fecha, PV (Planned Value) y AC (Actual Cost). Este nuevo conocimiento sobre las variables críticas aporta al problema al reducir la complejidad del modelo y enfocar los recursos computacionales en los datos más relevantes.

El proceso de transformación y limpieza de datos destacó la importancia de eliminar valores que no aportan al modelo, como los ceros en las variables numéricas. Esto, junto con la definición de un periodo específico para la serie de tiempo, sentó las bases para construir modelos más precisos, asegurando que las predicciones estén alineadas con las necesidades prácticas del problema.

El modelo ARIMA tradicional, con parámetros (3, 1, 1) obtenidos mediante autoarima, demostró ser efectivo para capturar las dependencias temporales de la serie, como lo evidencia el p-valor de 0.76 en la prueba de Ljung-Box. Sin embargo, el MSE obtenido (1,103,604) refleja que este enfoque podría beneficiarse de ajustes adicionales o de la inclusión de variables exógenas.

El modelo XGBoost, al incorporar la variable exógena PV (Planned Value), mostró un desempeño significativamente mejor que ARIMA, con un MSE de 535,781. Esto resalta la ventaja de utilizar algoritmos de aprendizaje automático para problemas de series de tiempo.

El uso de suavización exponencial en combinación con XGBoost no produjo mejores resultados, con un MSE de 1,605,589. Este hallazgo sugiere que, aunque las técnicas combinadas pueden ser útiles, es crucial evaluar cuidadosamente su impacto en cada problema. La experimentación con diferentes enfoques confirmó que la incorporación de variables adicionales al modelo base tiene un impacto significativo en la mejora de las predicciones.

La comparación de los tres modelos implementados mostró que el modelo XGBoost con variable exógena es el más adecuado para este problema, con el menor error cuadrático medio (MSE). Este resultado valida la hipótesis de que incluir información adicional relevante mejora la precisión de los modelos predictivos.

El análisis estadístico realizado en los datos originales, junto con la evaluación de los modelos, aportó una solución práctica al problema abordado, al proporcionar un modelo que no solo predice de manera más precisa, sino que también tiene la flexibilidad de incorporar nuevas variables si se dispone de más información.

El enfoque metodológico basado en pruebas estadísticas (Dickey-Fuller, Ljung-Box) y métricas de error permite respaldar con solidez los resultados obtenidos. Esto contribuye directamente a optimizar los procesos de inversión, demostrando cómo el análisis de datos puede aportar soluciones prácticas y basadas en evidencia.

Recomendaciones

Se recomienda para futuras investigaciones, probar otros algoritmos de aprendizaje automático de última generación como LightGBM, el cuál puede generar resultados similares o mejores, sin embargo, este algoritmo obtiene mayor funcionalidad cuando se trabajan conjuntos de datos de gran tamaño.

Se espera que, en próximas investigaciones, agregar más variables exógenas que puedan influir en un mejor rendimiento del modelo a utilizar, por ejemplo, el uso de una variable como las fases del proyecto y el tipo de actividad en la que se incurre el costo sería interesante conocer como estas variables podrían mejorar las estimaciones.

Es importante resaltar que debido a su característica estocástica, En algunas series de tiempo los valores pasados no garantizan la precisión de predicciones futuras (paseo aleatorio), por lo que se debe tener cuidado al realizar estimaciones en horizontes temporales demasiado largos, debido al componente de ruido blanco, sin embargo estos modelos nos ayudan a tener un punto de referencia para realizar estimaciones que en combinación con el juicio de un experto permitan construir conjuntos de datos que incorporen todas las variables que puedan explicar y mejorar la predicción de los costos durante el ciclo de vida de los proyecto cuando se trabaje con modelos de aprendizaje automático o aprendizaje profundo.

En la construcción de los modelos vistos en este trabajo, se utilizaron técnicas de optimización y búsqueda de hiperparámetros , como la optimización bayesiana, esta actividad depende adicional de los rangos o la selección de las variables que el modelo de optimización debe buscar, lo cual no sustituye en su totalidad la actividad manual de ajustar los hiperparametros, sino que facilita su búsqueda, por lo que se espera en próximas investigaciones, probar otras combinaciones de hiperparametros y analizar sus resultados.

Se espera, en próximas investigaciones mostrar los resultados del EAC al considerarse que ETC = Predicción del modelo *SKforest-XGboost con variable exógena*, en lugar de únicamente aproximaciones subjetivas. Posteriormente realizar su comparación con el BAC o línea base, para conocer el desempeño del proyecto.

Referencias Bibliográficas

Cezar-Petre, S., & Irinel, M.C. (2018). *Project Cost Estimate at Completion: Earned Value Management versus Earned Schedule-Based Regression Models. A Comparative Analysis of the Models Application in the Construction Projects in Romania*. ECONOMIC COMPUTATION AND ECONOMIC CYBERNETICS STUDIES AND RESEARCH.

Dagnino, J. (2014). *Regresión lineal*. Rev. chil. anest, 43(2).

Durón González, Flavio R., Rivas Tovar, Luis Arturo, & Cárdenas Tapia, Magali. (2022). *Evaluation of EVM/ES forecasting methods in hospital construction projects*. Revista ingeniería de construcción, 37(3), 405-416. https://dx.doi.org/10.7764/ric.00043.21

González Casimiro, P. (2009). *Análisis de series temporales: Modelos ARIMA*. Departamento de Economía Aplicada III (Econometría y Estadística), Facultad de Ciencias Económicas y Empresariales, Universidad del País Vasco (UPV-EHU).

Inter-American Development Bank. (2019). *Guía práctica PM4R Agile (PDF).* https://connectamericas.com/sites/default/files/articles_files/Guia%20practica%20PM4R%20Agile%20web_0.pdf

Le, T.-A., Huynh, Q.-T., Nguyen, T.-H., Nguyen, N.-H., & Cao, P.-N. (2020). *A method for project completion cost predicting using LSTM in earned value management technique*. En 2020 4th International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom) (pp. 87-92). Hanoi, Vietnam. https://doi.org/10.1109/SigTelCom49868.2020.9199064

Montoya Zapata, C., & Corredor Roa, C. A. (2022). *Aplicación de modelos de deep learning y machine learning para el pronóstico de la serie de tiempo del par de divisas Euro-Dólar.*

Narbaev, T., Hazir, Ö., Khamitova, B., & Talgat, S. (2024). *A machine learning study to improve the reliability of project cost estimates.* International Journal of Production Research, 62(12), 4372-4388. https://doi.org/10.1080/00207543.2023.2262051

Natarajan, A. (2022). *Reference Class Forecasting and Machine Learning for Improved Offshore Oil and Gas Megaproject Planning: Methods and Application.* Project Management Journal, 53(5), 456–484. https://doi.utadeoproxy.elogim.com/10.1177/87569728211045889

Navarrete Barrenechea, J. L. (2020). *Gestión de la planificación basado en la guía del PMBOK para el cumplimiento de la triple restricción del proyecto: suministro e instalación del sistema de climatización y extracción para pacientes infectocontagiosos e inmunodepresivos ejecutado por la empresa Cova Contratistas SAC en el año 2020.*

Pang, D. (2023). *Hybrid machine learning model performance in IT project cost and duration prediction.* Advances in Science, Technology and Engineering Systems Journal, 8(2), 108-115.

Pang, D.-J., Shavarebi, K., & Ng, S. (2022). *Development of machine learning models for prediction of IT project cost and duration*. En 2022 IEEE 12th Symposium on Computer Applications & Industrial Electronics (ISCAIE) (pp. 228-232). IEEE. https://doi.org/10.1109/ISCAIE54458.2022.9794529

Project Management Institute (2021). *Guía de los Fundamentos para la Dirección de Proyectos (Guía del PMBOK) – Séptima edición y El Estándar para la Dirección de Proyectos.* Project Management Institute.

Project Management Institute. (2011). *Practice standard for earned value management (2nd ed.).* Project Management Institute.

Project Management Institute. (2022). *Earned Value Management (2nd ed.).* Project Management Institute.

Raschka, S., & Mirjalili, V. (2017). *Python Machine Learning: Second Edition: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow*. Packt Publishing.

Rivas, G., López, L. A., & Velasco, A. (1993). *Regresión no lineal*. Revista colombiana de estadística, 14(27).

Rivera Martínez, F. (2010). *Administración de proyectos. Guía para el aprendizaje*. Pearson Educación.

Rodríguez González, J., & Ugalde Saborio, E. (2021). *Impacto de la estandarización y escalado: factor para predicción de costos en proyectos a través de una red neuronal artificial*. INGENIARE - Revista Chilena de Ingeniería, 29(2), 265–275.

Sánchez Torres, A. (2020). *La triple restricción en gestión de proyectos: Marco documental.*

Uncuoglu, E., Citakoglu, H., Latifoglu, L., Bayram, S., Laman, M., Ilkentapar, M., & Oner, A. A. (2022). *Comparison of neural network, Gaussian regression, support vector machine, long short-term memory, multi-gene genetic programming, and M5 Trees methods for solving civil engineering problems.* Applied Soft Computing, 129, 109623. https://doi.org/10.1016/j.asoc.2022.109623