Estrategias basadas en Machine Learning para la planificación de proyectos de diseño en ingeniería en la empresa Audubon, Sucursal Colombiana.

Oscar David Alpargatero Ulloa

Especialización en Machine Learning

Luz Anedy Cerdas Rodríguez

Especialización en Gerencia de Proyectos

Carlos Esteban Hernández Menco

Especialización en Gerencia de Proyectos

Jairo Alfonso Orozco Pastran

Especialización en Gerencia de Proyectos

Universidad Ean

Facultad de Ingeniería

Bogotá, Colombia

26/11/2025

Estrategias basadas en Machine Learning para la planificación de proyectos de diseño en ingeniería en la empresa Audubon, Sucursal Colombiana.

Oscar David Alpargatero Ulloa

Luz Anedy Cerdas Rodríguez

Carlos Esteban Hernández Menco

Jairo Alfonso Orozco Pastran

Trabajo de grado presentado como requisito para optar al título de:

Especialista en Machine Learning y Especialista en Gerencia de Proyectos

Director (a):

Luz Maribel Guevara Ortega

Modalidad:

Trabajo Dirigido

Universidad Ean

Facultad de Ingeniería

Bogotá, Colombia

26/11/2025

Nota de aceptación:

Firma del jurado

Firma del jurado

Firma del director del trabajo de grado

Bogotá, 24-Noviembre-2025

Agradecimientos

Agradecimientos a Luz Maribel Ortega, profesora asociada de la facultad de ingeniería y tutora de la asignatura Seminario de Investigación por la retroalimentación recibida durante el período académico para llevar a buen término este proyecto de investigación. Igualmente un saludo y agradecimiento para Julián Muñoz Ordoñez, profesor asociado y director del ecosistema de ciencia de datos de la facultad de ingeniería por tan valioso aporte y apoyo en el direccionamiento de este trabajo.

Resumen

Los proyectos de ingeniería en el sector Oil & Gas se caracterizan por una alta complejidad y exposición a riesgos financieros, consecuencia de la falta de planeación detallada y coordinación de los presupuestos, asignación de recursos y cronogramas. La presente investigación propone el diseño de un modelo de aprendizaje automático que identifique variables críticas y anticipe posibles impactos financieros negativos en la empresa Audubon. La metodología está basada en un enfoque descriptivo y aplicado que permita la caracterización, integración y uso de grandes volúmenes de datos en el diseño de algoritmos precisos para la anticipación de riesgos financieros y operativos. Se implementaron los algoritmos Regresión Logística, Árbol de Decisión, Random Forest y XGBoost, empleando métricas de desempeño como Recall, F1-Score y AUC-ROC para la validación de resultados.

Los hallazgos demuestran que los modelos de ensamblado, en particular Random Forest y XGBoost, presentan un desempeño superior respecto a los modelos tradicionales logísticos, con valores de recall y AUC-ROC superiores al 0.95 y 0.97, respectivamente. Estos modelos lograron capturar relaciones no lineales complejas y mantener una alta capacidad de generalización, lo que confirma su idoneidad para la predicción de viabilidad financiera en entornos con datos heterogéneos y de alta incertidumbre. Los resultados apoyarán la toma de decisiones estratégicas, optimizando la gestión de proyectos y contribuyendo a la mitigación de riesgos financieros en entornos de alta incertidumbre.

Palabras clave: Machine Learning, Visualización de Datos, Modelos predictivos, Procesamiento de datos, Diseño de ingeniería.

Abstract

Engineering projects in the Oil & Gas sector are characterized by high complexity and exposure to financial risks, resulting from a lack of detailed planning and coordination of budgets, resource allocation, and schedules. This research proposes the design of a machine learning model aimed at identifying critical variables and anticipating potential negative financial impacts within the company Audubon. The methodology is based on a descriptive and applied approach, enabling the characterization, integration, and use of large volumes of data for the development of accurate algorithms to predict financial and operational risks. The algorithms Logistic Regression, Decision Tree, Random Forest, and XGBoost were implemented, using performance metrics such as Recall, F1-Score, and AUC-ROC for result validation.

The findings demonstrate that ensemble models, particularly Random Forest and XGBoost, exhibit superior performance compared to logistic models, achieving recall and AUC-ROC values above 0.95 and 0.97, respectively. These models successfully captured complex nonlinear relationships and maintained a high generalization capacity, confirming their suitability for financial viability prediction in environments with heterogeneous data and high uncertainty. The results support strategic decision-making, optimizing project management, and contributing to the mitigation of financial risks in high-uncertainty contexts.

Keywords: Machine Learning, Data Visualization, Predictive Models, Data processing, Engineering Design.

Contenido

Pág.

Lista de Figuras

Pág.

Lista de Tablas

Pág.

Problema de Investigación

La empresa Audubon Companies LLC, con presencia en la industria Oil & Gas colombiana a través de su sucursal, enfrenta un reto recurrente en la gestión de sus proyectos de diseño de ingeniería: las desviaciones presupuestarias significativas y el sobrecosto que impactan directamente su rentabilidad y capacidad de respuesta ante clientes y aliados estratégicos. La causa subyacente radica en la ausencia de herramientas analíticas avanzadas durante la etapa de planificación y estimación, lo que limita la identificación temprana de factores críticos determinantes, tales como el alcance correctamente definido, la valoración de riesgos específicos del cliente y las condiciones iniciales restrictivas del proyecto (Oberlender et al., 2022) . Actualmente, la empresa basa la proyección financiera y de tiempo en metodologías tradicionales, como las establecidas en el Project Management Body of Knoweledge (Project Management Institute, 2021) y la revisión de lecciones aprendidas por parte del gerente de proyectos, apoyándose en la experiencia de equipos técnicos y proyectos realizados anteriormente. Este enfoque, aunque útil en la consolidación de buenas prácticas, carece de una capacidad predictiva que permita la anticipación real de desviaciones (Waqar et al., 2023).

Los síntomas y manifestaciones observadas evidencian la necesidad de innovación en el control de gestión. En la revisión de los reportes consolidados en cada cierre mensual se reportan diferencias importantes entre los costos estimados y los reales, con cifras que alcanzaron más de USD $4.4 millones en sobrecostos para el cierre de 2024 (Audubon, 2025b). La naturaleza técnica y la volatilidad del contexto económico del sector, sumadas a las presiones de cumplimiento de cronogramas y factores externos como la variabilidad del precio del crudo y exigencias regulatorias, intensifica la probabilidad de desviaciones financieras (Zhang et al., 2025). El monitoreo presupuestario actual, apoyado en reportes ex post y análisis Earned Value Management (EVM), ofrece diagnósticos tardíos, obligando a implementar acciones correctivas cuando el proyecto ya está avanzado en ejecución. Esto limita la

capacidad de respuesta preventiva, disminuye la competitividad y reduce la eficiencia operativa (Priyadarshy & Moonsammy, 2021) .

El pronóstico de seguir operando bajo este modelo tradicional señala un escenario poco alentador. La empresa mantendrá una exposición elevada a sobrecostos, afectará la satisfacción y confianza de sus clientes, y comprometerá su sostenibilidad financiera en entornos de alta incertidumbre y competencia global. La tendencia apunta a una reducción progresiva de la competitividad frente a empresas que adopten analítica avanzada y modelos robustos de Machine Learning en la gestión de proyectos (Priyadarshy & Moonsammy, 2021) . El riesgo se materializa tanto en términos financieros como en pérdidas de oportunidades y en el desgaste reputacional de la marca, resultados que en proyectos Oil & Gas se agravan por la falta de sistemas predictivos para anticipar y mitigar el riesgo de sobrecostos.

El control pronóstico y posible solución reside en la migración hacia un modelo de gestión presupuestaria basado en análisis de datos y Machine Learning, aprovechando el acervo de datos históricos que posee la organización. A través de la consolidación y análisis estructurado de datos de proyectos previos, es posible identificar variables críticas en la planificación, tales como cambios de alcance, duración de fases de ingeniería, complejidad disciplinar, indicadores macroeconómicos y comportamiento financiero del cliente, entre otros (Ma et al., 2025).

La problemática planteada se encuentra inmersa dentro de la gerencia de proyectos de ingeniería y Machine Learning, es una intervención directa en la organización y demanda el diseño aplicable de un modelo predictivo adaptado al contexto empresarial. En síntesis, la brecha entre la disponibilidad de datos históricos y su uso efectivo para predecir desviaciones presupuestarias limita la competitividad de Audubon y prolonga su exposición a impactos financieros negativos en un sector altamente exigente.

La siguiente pregunta surge a partir de este análisis y será abordada posteriormente: ¿Cómo puede un modelo de aprendizaje automático, alimentado con datos históricos de

proyectos de diseño de ingeniería en el sector Oil & Gas, identificar las variables con mayor incidencia en los sobrecostos futuros y entregar información anticipada que optimice la planificación presupuestaria en la empresa Audubon sucursal colombiana?

Este documento se estructura en cinco capítulos. El primer capítulo aborda y desarrolla el marco teórico, en el cual concentra una exploración detallada y teórica de los conceptos de Machine Learning y modelos predictivos en proyectos de sectores relacionados con ingeniería, diseño y construcción, para enseguida abordar la aplicación de AI y ML en el ciclo de vida de los proyectos, la predicción de sobrecostos y finalmente una revisión a teorías y modelos de referencia usados en proyectos. En el segundo capítulo se realiza el análisis bibliométrico entre los años 2019 y 2025 a artículos publicados relacionados con Predictive Modeling y Management para evaluar, cuantificar y analizar el impacto de las publicaciones que tienen relación cercana y/o directa al objeto de investigación. El tercer capítulo se centra en el diseño metodológico estableciendo cada una de las fases para el desarrollo de los objetivos, en las cuales se incluye el diagnóstico, procedimiento y modelo elegido para la intervención en la organización. Finalmente se presenta el plan de intervención que contiene estrategias y propuestas para la organización basados en los resultados obtenidos del análisis de datos y modelos predictivos.

Objetivos

Objetivo general

Diseñar una estrategia basada en Machine Learning para la planificación de proyectos de diseño en ingeniería en la empresa Audubon, Sucursal Colombiana.

Objetivos específicos

Proponer un modelo de aprendizaje automático que, a partir de datos históricos, identifique patrones y relaciones entre variables determinantes de los proyectos del sector Oil & Gas.

Evaluar el desempeño del modelo usando métricas de clasificación para validar su confiabilidad y capacidad predictiva.

Plantear estrategias para la empresa Audubon que optimicen la planificación de proyectos de ingeniería y contribuyan a la mitigación de riesgos financieros en el sector Oil & Gas.

Justificación

La planificación estratégica de proyectos guía a las organizaciones al cumplimiento de sus objetivos y permite afrontar de manera coherente los problemas que pueden presentarse en el camino de su ejecución (Parrales García et al., 2024). Considerando lo anterior, la empresa Audubon enfrenta el reto de predecir, coordinar y gestionar riesgos mediante la incorporación de herramientas que fortalecen la toma de decisiones y promueven la estabilidad en sus operaciones.

La finalidad de este estudio es plantear estrategias basadas en Machine Learning para la planificación de proyectos de diseño en ingeniería de la empresa Audubon, mediante modelos de aprendizaje automático que integren datos históricos, reconozcan variables críticas y anticipen la probabilidad de impactos financieros negativos (Caballero et al., 2023) al reducir la incertidumbre en la planificación de proyectos fortaleciendo la capacidad de respuesta ante escenarios adversos.

Adicionalmente, el uso de modelos predictivos fundamentados en datos históricos para reconocer variables críticas y predecir impactos financieros negativos permite mejorar la distribución de recursos, minimizar residuos y efectos no deseados sobre el entorno (Bodero Poveda et al., 2021), contribuyendo así a objetivos más amplios de desarrollo sostenible (Directorio de Sostenibilidad, 2025). En este contexto, el estudio adquiere relevancia social al mejorar la capacidad de respuesta de la empresa frente a escenarios complejos, generando beneficios notables tanto para sus colaboradores como para el entorno económico empresarial.

La relevancia práctica de esta investigación es notoria, porque el modelo propuesto puede desarrollarse como una herramienta predictiva en la empresa Audubon y replicarse en otros proyectos de ingeniería, generando valor en la evaluación inicial de riesgos financieros y en la definición de estrategias de mitigación. En este contexto Joyanes (2019) resalta que la inteligencia de negocios y la analítica de datos permiten transformar la información en

conocimiento aplicable, generando sistemas de apoyo que fortalecen la toma de decisiones estratégicas y operativas en las organizaciones.

Con respecto al valor teórico, este estudio enriquece la literatura en gestión de proyectos y analítica de datos al integrar fundamentos de estadística, Big data, minería de datos y aprendizaje automático en un entorno empresarial. Lo anterior genera un marco de referencia idóneo para futuras investigaciones interdisciplinarias (Lee & Lee, 2025).

Por último, la utilidad metodológica de este estudio se centra en el diseño de un modelo predictivo replicable y adaptable a distintos contextos de ingeniería de proyectos, fortaleciendo la investigación aplicada y promoviendo marcos innovadores en la intersección entre gestión de proyectos e inteligencia artificial. Como señala Lee & Lee (2025) la estructuración de la inteligencia artificial en el aprendizaje automático constituye tanto un recurso técnico, como una guía metodológica para construir modelos sólidos, escalables y aplicables a problemas complejos.

Este estudio se enmarca en el campo de emprendimiento y gerencia de la Universidad EAN, dentro del grupo de investigación en dirección y gestión de proyectos, específicamente en la línea de investigación sobre modelos, metodologías y sistemas de gestión para la gerencia de proyectos y gestión de proyectos, estrategia y competitividad. De esta manera, contribuye al fortalecimiento de prácticas innovadoras para la planificación, control y evaluación de proyectos en sectores estratégicos.

Marco Institucional

Estructura organizacional

Audubon, empresa norteamericana establecida en Colombia desde el año 2012, en donde la calidad es uno de sus valores fundamentales, presta servicios de ingeniería y diseño multidisciplinario, soporte operacional, gestoría, comisionamiento, terminación, gestión de adquisiciones, optimización de procesos y otros servicios en campo para la industria de los hidrocarburos. El alcance se logra a través de un capital humano capacitado y especializado en las disciplinas de Eléctrica, Instrumentación, Civil, Estructural, Procesos, Mecánica y Tubería que son fundamentales para el desarrollo de todas las ingenierías. (Audubon, 2025) Así mismo se cuenta con personal idóneo como apoyo a todo el proceso misional en las áreas de gerencia, financiera, control proyectos, control de documentos, automatización, calidad, recursos humanos, compras, seguridad y salud en el trabajo, que permiten a la organización ser reconocida en el mercado, por la alta convertibilidad de sus proyectos, lograda a través de la integridad, transparencia y competitividad de sus profesionales y sus políticas corporativas.

Valores corporativos

Los valores corporativos que definen a Audubon sucursal colombiana son el ingenio, la integridad, el entusiasmo, la eficiencia, la calidad y la confianza estableciendo la hoja de ruta para la organización, cuyo compromiso es entregar productos y servicios de calidad cumpliendo en su totalidad con la satisfacción del cliente. Estos valores nacen de la visión de llegar a ser la empresa líder del sector local llevando a cabo cada aspecto de las operaciones e interacciones a resultados exitosos como factor diferenciador, a través de un enfoque en las personas, la flexibilidad, las relaciones y la experiencia.

Historia de la Empresa

Audubon Engineering Company LLC fue fundada en Lousiana en el año 1997 por los ingenieros Ryan Hanemann, Denis Taylor y Bob Rosamond, con la visión de ser una de una compañía líder en ingeniería, consultoría, fabricación y servicios técnicos especializados, para

cubrir necesidades de los sectores de la energía, producción y mercados industriales. Por más de dos décadas, Audubon ha construido una reputación basada en la premisa de trabajar y desarrollar una cultura de mejora continua, cumpliendo con la entrega satisfactoria de proyectos a los diferentes clientes y abordando algunos de los retos más desafiantes a los que se enfrentan los sectores energéticos e industriales. (Audubon, 2022).

A través de los valores fundamentales, la experiencia y capacidades Audubon ha logrado convertirse en un líder global reconocido por asegurar fiabilidad, flexibilidad y récord en la ejecución de proyectos. Con el apoyo de las compañías afiliadas, Audubon Engineering, Audubon Field Solutions, Audubon Construction, Opero Energy y Audubon Carbon, se han entregado sucesivos proyectos exitosos, seguros, a tiempo y dentro del presupuesto (Audubon, 2022).

Trascendencia y posición en el mercado

En cuanto a la posición del mercado de acuerdo con la revista Engineering News Record (ENR), Audubon ocupa el puesto #111 entre el top 150 de firmas de diseño global para 2025. (Audubon, 2025c). Este reconocimiento sigue el camino de haber logrado el puesto #64 en el ranking Top 500 de las mejores firmas de diseño en abril del mismo año (Audubon, 2025b). Esto demuestra que los resultados se ven plasmados cuando se invierte en las personas correctas, la tecnología correcta y las mejores alianzas como input para impulsar la habilidad de la compañía en el desarrollo y entrega exitosa de proyectos de ingeniería.

Marco de Referencia

El siguiente marco de referencia plantea la revisión de antecedentes, modelos y marcos conceptuales que fundamentan la aplicación del aprendizaje automático en la gestión de proyectos de ingeniería. En primer lugar, se desarrolla un estado del arte que recopila los avances más significativos en el uso de Inteligencia Artificial y Machine Learning en la gestión de proyectos, así como los hallazgos relacionados con la predicción de sobrecostos y retrasos en contextos industriales y de construcción. En segundo lugar, se analizan los algoritmos de aprendizaje automático relacionados con la predicción de riesgos, identificando sus fortalezas y limitaciones. Se incluyen también los factores y variables críticas que influyen en la viabilidad de los proyectos, diferenciando entre aspectos internos y externos, además de las métricas financieras que permiten evaluar su factibilidad. Finalmente, se presenta un panorama de los marcos legales y normativos internacionales y nacionales, incluyendo estándares ISO, normativas de gobernanza de la inteligencia artificial, regulaciones de sostenibilidad y políticas públicas como el CONPES 4144 en Colombia, que orientan la implementación de modelos predictivos en sectores estratégicos.

El campo de la inteligencia artificial ha permitido en tiempos recientes un avance creciente en modelos y aplicaciones que dan soporte a la mayoría de las actividades en las organizaciones empresariales. Diversos estudios muestran que la AI está actualmente revolucionando industrias de la manufactura, ventas al por menor y telecomunicaciones. Subcampos como el Machine Learning vienen siendo ampliamente usados en industrias para lograr incrementos en productividad, eficiencia y seguridad (S. O. Abioye et al., 2021). Para el caso específico de estudio en la empresa de diseño en ingeniería para el sector Oil & Gas en el que el alcance cubre entregables de ingeniería que serán usados posteriormente en fases constructivas, hay numerosos retos y desafíos que deben ser revisados y cubiertos antes de su puesta en marcha con aplicaciones reales en la industria, por lo que este marco de referencia pretende hacer una revisión de antecedentes, teorías, modelos y otras investigaciones que

fundamenten la investigación y aplicabilidad del Machine Learning dentro del sector al cual pertenece la organización objeto de análisis.

De acuerdo con Abioye et al. (2021), en los últimos años, la incorporación de tecnologías de análisis de datos ha mostrado un gran potencial para mejorar la gestión de proyectos e incrementado la eficiencia y rentabilidad. La inteligencia Artificial (AI) y el Machine Learning han logrado transformar la forma en que las organizaciones analizan sus datos, permitiendo identificar patrones y anticiparse a tendencias o resultados que en años anteriores eran inimaginables mediante el procesamiento de datos, Datta et al. (2024a) argumenta que estas tecnologías son útiles para anticipar riesgos y estimar costos de manera más precisa, muchas empresas aún carecen de sistemas que integren información histórica para alimentar modelos predictivos y la falta de adopción de herramientas de análisis predictivo limita la capacidad de planificar de manera proactiva.

Tshidavhu y Khatleli (2020) argumentan que el desarrollo de proyectos de ingeniería en el sector Oil & Gas enfrenta retos significativos concernientes a la complejidad técnica, el tamaño de las inversiones y los altos niveles de incertidumbre asociados a los mercados de energía globales. Teniendo esto en cuenta, los problemas de sobrecostos y retrasos son comunes y comprometen la viabilidad financiera de las organizaciones. Como conclusión, después de la revisión de diversas fuentes, podemos afirmar que, no existen herramientas claras que permitan anticipar riesgos financieros (Flyvbjerg, 2014), y se deben aplicar diversos conocimientos para construir una metodología que por medio de información conocida permita predecir la viabilidad de proyectos futuros.

Diversos autores exponen que el boom de la inteligencia artificial (AI) y el aprendizaje automático (Machine Learning, ML) abre nuevas puertas para la gestión de proyectos, estas tecnologías permiten analizar grandes volúmenes de datos históricos para identificar patrones ocultos y predecir con mayor precisión el comportamiento futuro de variables importantes (Abioye et al., 2021; Datta et al., 2024b) . Por estas razones, aplicar el Machine Learning a

proyectos de ingeniería en el sector Oil and Gas se convierte en una necesidad innovadora para predecir la viabilidad financiera antes de su ejecución, permitiendo de esta manera mejorar las decisiones estratégicas en las organizaciones.

En el presente estudio se profundizará en la consulta de los estudios más significativos de Machine Learning en proyectos de ingeniería, predicción de sobrecostos y retrasos, fundamentación conceptual y teórica, teorías de gestión de proyectos, modelos predictivos, marcos conceptuales y normativos.

Aplicación de Inteligencia Artificial y Machine Learning en el ciclo de vida de proyectos

En industrias como la energía y la construcción, la incorporación de Inteligencia Artificial y Machine Learning en la gestión de proyectos han ganado relevancia en los últimos 10 años. Teniendo en cuenta lo expuesto por Datta et al. (2024b), la mayor parte de la aplicación de estas herramientas, se concentran en la fase de planificación y ejecución, etapas donde los modelos permiten estimar con mayor precisión tiempos, costos y riesgos. Su revisión sistemática muestra que los enfoques basados en ML superan los métodos tradicionales de estimación, al integrar múltiples variables de manera no lineal y proporcionar sistemas de alerta temprana.

Según un estudio comparativo realizado por Mali et al. (2025) y después de comparar diferentes modelos de ML aplicados a la gestión de proyectos de construcción, Artificial Neural Networks (ANN), Support Vector Machines (SVM) y Random Forest, concluyeron que, aunque las redes neuronales ofrecen un buen desempeño, el modelo Random Forest es el más robusto para predecir riesgos y optimizar la asignación de recursos, debido a su capacidad de manejar grandes volúmenes de datos con menor sobreajuste.

En el ámbito específico del sector Oil & Gas, no se encuentran estudios significativos, pero existen reportes de aplicación de ML en áreas como mantenimiento predictivo, análisis de integridad de activos y optimización de procesos de exploración. La aplicación directa de estas

técnicas a la evaluación de viabilidad financiera en proyectos de diseño de ingeniería constituye un vacío investigativo que justifica la presente investigación.

Predicción de sobrecostos y retrasos en proyectos de construcción

Los sobrecostos y retrasos han sido estudiados ampliamente en proyectos de construcción, que comparten características con los proyectos Oil & Gas por su escala, complejidad y dependencia de múltiples actores. En su estudio Al mnaseer et al. (2023a) emplearon redes neuronales artificiales (ANN) optimizadas con Tabu Search para predecir overruns en 191 proyectos en Jordania, alcanzando un R^2 de 0.93 para costos y 0.94 para tiempos, resultados que son mucho mejores comparado con los modelos tradicionales.

Por otra parte, Arabiat et al. (2023) aplicaron K-nearest neighbor (KNN) y ANN en un conjunto de proyectos completados, logrando precisiones del 83.7 % y 99.3 %, estos trabajos permiten ver lo útil del ML para anticipar desviaciones cuando se dispone de datos históricos consistentes.

En el espacio de los algoritmos Coffie y Cudjoe (2024) utilizaron el algoritmo Extreme Gradient Boosting (XGBoost) en proyectos realizados en Ghana, complementando el análisis con SHAP values para identificar las variables más influyentes en modelos. Como resultado resaltaron las variables de monto inicial del contrato, el número de cambios de alcance y la duración inicial como predictores claves de sobrecostos.

Los ya mencionados, Tshidavhu y Khatleli (2020) en su estudio de megaproyectos energéticos ubicados en Sudáfrica, determinaron que las principales causas de desviaciones no son exclusivamente técnicas, puesto que incluyen aspectos organizacionales, por ejemplo, lentitud en la toma de decisiones, falta de mano de obra calificada, estimaciones deficientes y cambios frecuentes en el alcance, con esto se concluye que se deben tener en cuenta tanto variables cuantitativas como cualitativas en la construcción de modelos predictivos.

Otros estudios refuerzan estas tendencias, Arabiat et al. (2023) combinaron ML con PSO, encontrando mejoras en precisión de costos; Salama (2025) integró GPR con AHP para

evaluar retrasos, logrando predicciones casi perfectas; Podder & Podder (2025) utilizaron lógica difusa y clustering para gestionar incertidumbre; Wang et al. (2025) diferenciaron entre sobrecostos regulares y grandes mediante clasificadores bayesianos; Maurya et al. (2025) destacaron que GBM logra un equilibrio entre precisión e interpretabilidad.

Modelos de aprendizaje automático aplicados

La revisión de la literatura evidencia que distintos algoritmos de Machine Learning han mostrado un potencial significativo para anticipar riesgos en proyectos de ingeniería. Como ejemplo claro se destacan las redes neuronales artificiales o ANN por sus siglas en inglés, cuya característica principal es su capacidad para describir relaciones complejas y no lineales entre variables, a pesar de sus limitaciones en la forma de interpretar los resultados (Al Mnaseer et al., 2023b; Arabiat et al., 2023). Otros enfoques, como los modelos Random Forest y las Support Vector Machines (SVM), han demostrado ser particularmente útiles y confiables para la clasificación y predicción de riesgos con conjuntos de datos grandes y complejos, mostrando un mejor control frente al sobreajuste (Mali et al., 2025).

De igual manera el algoritmo Extreme Gradient Boosting (XGBoost) ha cobrado relevancia en contextos de predicciones tabulares, ya que, además de su precisión, ofrece una ventaja en términos de interpretabilidad gracias a los valores SHAP, que permiten identificar las variables más influyentes en los modelos (Coffie & Cudjoe, 2024). Finalmente, el método K-Nearest Neighbor (KNN) se ha consolidado como una alternativa sencilla y efectiva para bases de datos de tamaño medio y bien estructuradas, proporcionando resultados competitivos con menor complejidad computacional (Arabiat et al., 2023).

De acuerdo con Salama (2025), nuevas contribuciones incluyen Voting Regression con PSO, GPR+AHP, de la misma manera Podder & Podder (2025) exponen la lógica difusa + clustering como herramienta, Wang et al. (2025) nombra a los clasificadores bayesianos para

sobrecostos severos y Maurya et al. (2025) GBM como balance entre precisión y explicabilidad (Maurya et al., 2025).

Es así como los modelos de clasificación son válidos para aplicar a este conjunto de datos porque se ha demostrado su aplicabilidad en contextos similares, con entornos cambiantes y complejos que requieren de modelos adaptativos a los diferentes rangos y etiquetas encontrados en la base de datos históricos de la compañía Audubon LLC sucursal Colombiana, XGBoost suena como el modelo más predominante que puede ofrecer mejores resultados de acuerdo al estado del arte considerado anteriormente.

Tabla 1

Resumen de modelos predictivos para el análisis de datos

Tipo de Modelo	Descripción	Propósito	Algoritmos
Modelos de regresión	Predicen una variable continua basada en otras variables.	Determinar cómo variables específicas afectan alcance, tiempo y costo.	Regresión lineal, regresión polinómica, regresión Ridge/Lasso
Modelos de clasificación	Identifican categorías o clases para datos de entrada.	Clasificar proyectos en categorías (alto impacto, bajo impacto, etc.).	Árboles de decisión, Random Forest, SVM.
Modelos de árboles de decisión y bosques aleatorios	Analizan relaciones no lineales y determinan variables importantes.	Identificar variables clave que impactan en los resultados del proyecto.	Árbol de decisión, Random Forest, Gradient Boosting
Modelos de análisis de contribución (Feature Importance)	Evaluar la influencia de cada variable en la predicción.	Priorizar variables que más afectan alcance, tiempo y costo.	Importancia de características en Random Forest, XGBoost, LightGBM
Modelos de regresión multivariable	Evaluar múltiples variables simultáneamente para predecir un resultado.	Comprender el impacto conjunto de variables en la ejecución del proyecto.	Regresión lineal múltiple, Ridge, Lasso
Modelos de clustering	Agrupar proyectos con características similares sin etiquetado previo.	Identificar perfiles de proyectos con impactos similares.	K-means, KNN, Hierarchical Clustering, DBSCAN

Modelos de análisis de sensibilidad	Analizan cómo cambios en variables afectan los resultados.	Evaluar la sensibilidad de KPI's respecto a variables específicas.	Análisis de sensibilidad, Modelo Sobol o técnicas de simulación

Nota. Elaboración propia.

La tabla 1 describe y categoriza algunos de los modelos de ML que más se han utilizado en investigaciones recientes y que han mostrado resultados favorables en la aplicación de proyectos de construcción y gestión de proyectos.

Variables críticas en proyectos de ingeniería

Las variables críticas que influyen en la viabilidad de un proyecto abarcan tantos factores internos (complejidad técnica, número de disciplinas, cambios de alcance, horas de ingeniería, retrabajo), los cuales han sido identificados en estudios previos (Al mnaseer et al., 2023c; Arabiat et al., 2023; Bohórquez Castellanos & Mejia-Aguilar, 2019; Coffie & Cudjoe, 2024). Así mismo, factores externos como el precio del petróleo, la inflación, las tasas de cambio y los retrasos en permisos influyen en la viabilidad de los proyectos, en línea con lo señalado por (Tshidavhu & Khatleli, 2020), (Dzhusupova et al., 2024), (Bruzzone et al., 2021) y (Liu et al., 2025)

Viabilidad financiera en proyectos

De acuerdo con Serrano-Gomez y Muñoz-Hernandez (2020), el Valor presente Neto (NPV), la Tasa interna de Retorno (IRR) y la relación Beneficio/Costo (B/C) son criterios financieros utilizados en construcción como puntos de partida para aprobar proyectos y toma de decisiones. El día a día en la gestión de proyectos relacionados con ingeniería, energías renovables, hidrocarburos permite comprobar la importancia y utilización de estos.

Complejidad, incertidumbre y riesgo

La literatura en gestión de proyectos sostiene que la complejidad organizacional y la incertidumbre del entorno amplifican los riesgos, lo que justifica la adopción de enfoques basados en datos y modelos predictivos (Cicmil et al., 2006).

Por otra parte, (S. O. Abioye et al., 2021) concluyen que, en el ámbito de la construcción, la adopción de la inteligencia artificial (IA) enfrenta importantes barreras culturales y de datos. En primer lugar, las barreras culturales se relacionan con la resistencia al cambio por parte de los profesionales y organizaciones del sector, quienes tienden a mostrar desconfianza hacia las nuevas tecnologías debido a la tradición de trabajar con métodos convencionales, al temor a la pérdida de empleos y a la falta de competencias digitales entre los trabajadores.

En segundo lugar, las barreras de datos hacen referencia a la carencia de información de calidad, estandarizada y accesible, necesaria para alimentar los algoritmos de AI. En el sector de la construcción, los datos suelen estar fragmentados, incompletos o dispersos en diferentes actores de la cadena de valor, lo que dificulta la creación de modelos confiables.

Normatividad legal en Colombia

En Colombia, los proyectos del sector Oil & Gas están regulados por la Agencia Nacional de Hidrocarburos (ANH), el Ministerio de Minas y Energía y la Unidad de Planeación Minero-Energética (UPME). Estos organismos exigen estudios de factibilidad técnica, ambiental y financiera antes de aprobar proyectos. Es evidente que la normatividad vigente es la guía que permite organizar los proyectos de manera técnica y financiera.

El diseño de un modelo de aprendizaje automático para la planificación de proyectos en el sector Oil & Gas debe corroborarse en un marco legal y normativo internacional que proporcione solidez a su implementación, puntualmente en todo lo relacionado a gestión de proyectos, riesgos financieros y gobernanza de la inteligencia artificial (AI).

Inicialmente en el área de gestión de proyectos, se resaltan las siguientes directrices ISO/TC 258. La ISO 21500:2021 las cuales establecen conceptos esenciales en la gestión de proyectos, programas y portafolios, siendo acogida por organizaciones públicas y privadas como modelo de gobernanza (Dawood & Ahmed, 2023). Además, la guía ISO 21502:2023 facilita los lineamientos fundamentales para la gestión de proyectos complejos, reforzando la

necesidad de integrar estándares internacionales en la planificación estratégica del sector Oil & Gas (Rumane, 2024).

Asimismo, el presente trabajo implementa modelos de aprendizaje automático, por lo tanto, es necesario mencionar la normativa internacional sobre gobernanza de AI. En este contexto, la ISO/IEC 42001:2023 determina los requisitos básicos para la creación de sistemas de gestión de inteligencia artificial, mientras que la ISO/IEC 23894:2023 destaca los lineamientos para la gestión de riesgos asociados a los modelos predictivos. Estos lineamientos ofrecen criterios claros para el control de riesgos técnicos, organizacionales y de datos en todas las fases del modelo (ISO, 2025). Simultáneamente, el NIST AI Risk Management Framework (2023) presenta una guía práctica para la identificación, evaluación y mitigación de riesgos de AI, reconocida al complementar los estándares ISO (Tabassi, 2023).

En el marco reglamentario, la Unión Europea ha divulgado el AI Act (2024),siendo la primera legislación internacional que controla los sistemas de inteligencia artificial bajo un enfoque basado en riesgos. Este marco normativo dispone de categorías de riesgo, obligaciones de documentación técnica y supervisión post-implementación, también exige transparencia y trazabilidad en el desarrollo y uso de modelos de AI (Ebers, 2024). Investigaciones recientes destacan que este enfoque será clave para las organizaciones que operen en sectores críticos como el Oil & Gas (Szadeczky & Bederna, 2025).

Según lo anterior, el uso de modelos predictivos que anticipan impactos financieros negativos en proyectos debe alinearse con las normativas internacionales de divulgación de riesgos y sostenibilidad. Los estándares IFRS S1 y S2 (2023), emitidos por el *International Sustainability Standards Board (ISSB)*, exigen a las empresas reportar información financiera relacionada con riesgos de sostenibilidad, incluyendo métricas e indicadores que fortalecen la transparencia en las organizaciones. En este contexto, es importante que los proyectos industriales integren prácticas de reporte financiero y no financiero como parte de la gestión de riesgos (Hummel & Jobst, 2024).

Por último, algunos autores han demostrado que los modelos de aprendizaje automático son eficaces en la predicción de sobrecostos y retrasos en proyectos industriales, respaldando su validez técnica y regulatoria. Por ejemplo, Turkyilmaz & Polat (2024) comprobó que los modelos de ML pueden estimar ratios de sobrecosto con alta precisión, mientras que Moussa et al. (2024) resalta su capacidad para anticipar interacciones de riesgos sistémicos en proyectos de gran escala. Estos aportes destacan la necesidad de integrar estándares normativos con la práctica de la ciencia de datos aplicada a la gestión de proyectos (págs. 5-9).

Para concluir, el marco legal y normativo que orienta este trabajo se apoya en: (i) los estándares internacionales de gestión de proyectos (ISO 21500 e ISO 21502), (ii) la regulación emergente de AI y marcos de gestión de riesgos (ISO/IEC 42001, ISO/IEC 23894 y NIST AI RMF), (iii) las regulaciones internacionales sobre transparencia y sostenibilidad financiera (IFRS S1 y S2), la evidencia científica reciente sobre el uso de aprendizaje automático en la predicción de sobrecostos y riesgos financieros. Esta normatividad vigente y fuentes bibliográficas permiten que la propuesta de modelo predictivo esté soportada con las mejores prácticas internacionales en gobernanza, sostenibilidad e innovación tecnológica en el sector Oil & Gas.

En Colombia, el Documento CONPES 4144 de 2025 constituye el principal lineamiento normativo para la adopción y regulación de la inteligencia artificial. Esta política nacional establece una hoja de ruta para el desarrollo, implementación y gobernanza ética de la AI, definiendo lineamientos en materia de transparencia, calidad de datos, gestión de riesgos y responsabilidad de los actores (CONPES 4144: Política nacional de inteligencia artificial, 2025).

En el contexto de proyectos de ingeniería en el sector Oil & Gas, el CONPES destaca la necesidad de aplicar la AI en procesos estratégicos, particularmente en la identificación de variables críticas y predicción de riesgos financieros, garantizando al mismo tiempo el cumplimiento de estándares internacionales y la protección de derechos fundamentales. Su enfoque en la trazabilidad de los algoritmos, la interoperabilidad de datos y la alineación con

principios de sostenibilidad convierte este documento en el eje central de la normativa colombiana aplicable a la investigación.

El marco teórico nos permite concluir que las herramientas disponibles de aprendizaje automático permiten obtener mejores resultados a las técnicas tradicionales en la predicción de sobrecostos y retrasos. Se encontró que no existe una aplicación directa a proyectos de diseño de ingeniería en Oil & Gas, específicamente en lo relativo a la viabilidad financiera**.**

El marco teórico aquí expuesto permite fundamentar el problema desde tres perspectivas: (i) la gestión de proyectos, que aporta metodologías estandarizadas; (ii) los modelos de predicción de riesgos y desempeño, que justifican el uso de sistemas de alerta temprana; y (iii) el aprendizaje automático, que provee herramientas poderosas para identificar variables críticas y anticipar la viabilidad de los proyectos.

En conclusión, la presente investigación se abordará teniendo en cuenta conocimientos de gestión de proyectos, análisis financiero y técnicas de ML (ANN, RF), con el propósito de diseñar un modelo que contribuya a reducir la incertidumbre y fortalecer la toma de decisiones estratégicas en el sector Oil & Gas. Durante el desarrollo del proyecto serán evaluados diferentes modelos para plantear y diseñar la estrategia basada en Machine Learning para la planificación de proyectos de diseño en ingeniería en la empresa Audubon, Sucursal Colombiana.

Revisión bibliométrica (2019 – 2025)

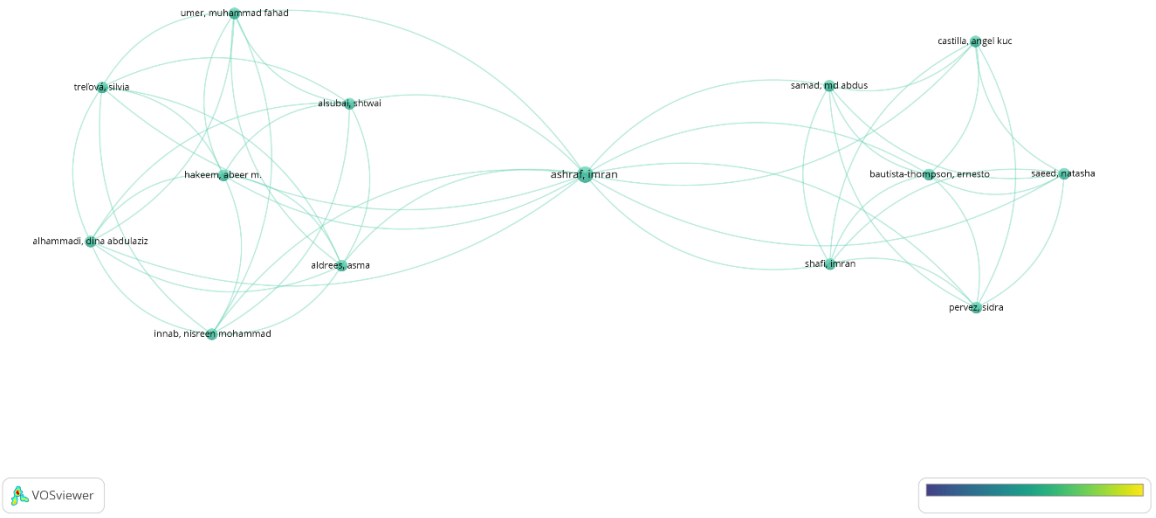
En este capítulo se desarrolla una revisión bibliométrica de la investigación en curso para descubrir correlaciones entre los distintos autores que han realizado investigaciones relacionadas a Predictive Modeling y Management usadas en títulos de artículos, resúmenes y palabras clave, a través una búsqueda de estos términos en la base de datos Scopus, entre los años 2019 y 2025, teniendo en cuenta filtros a solo artículos de investigación y limitados a áreas específicas de ingeniería, ciencia de la computación, Ciencias Sociales, energía,

negocios, gestión, contabilidad, matemáticas, economía, econométricas, finanzas y ciencias de decisión.

Publicaciones: Se identificaron 1,022 documentos publicados en la base de datos Scopus de acuerdo a los parámetros mencionados anteriormente.

Figura 1

Visualización por autores

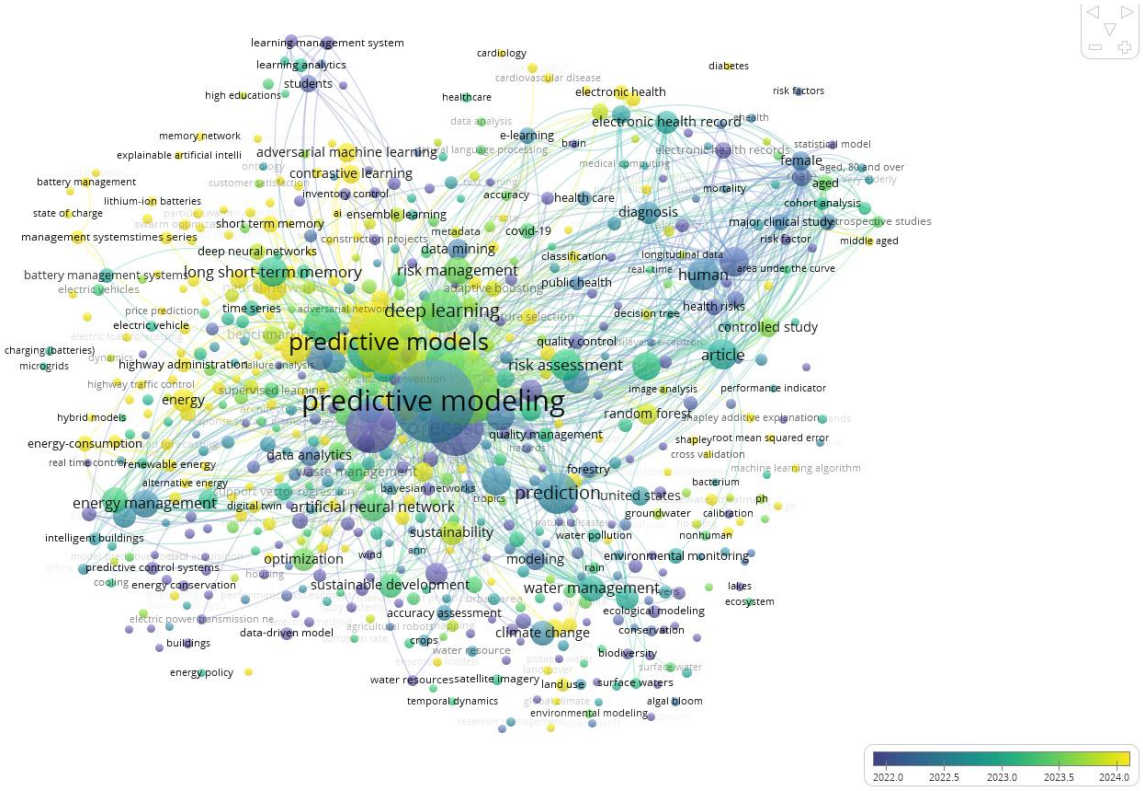


Nota. Elaborado a partir de la herramienta VOSViewer. (Scopus, 2025). https://www.scopus.com/home.uri

Se encontró una limitada interconexión entre los investigadores. Esto significa que, aunque existe una producción académica considerable sobre la temática, los autores trabajan de manera aislada y rara vez se citan entre sí, lo que evidencia la ausencia de una red consolidada de investigación. Esto sugiere que el campo aún se encuentra en una etapa de maduración, caracterizada por aportes fragmentados y enfoques diversos, sin que se haya consolidado un núcleo de autores que se dediquen a este campo especifico.

Figura 2

Visualización por palabras claves

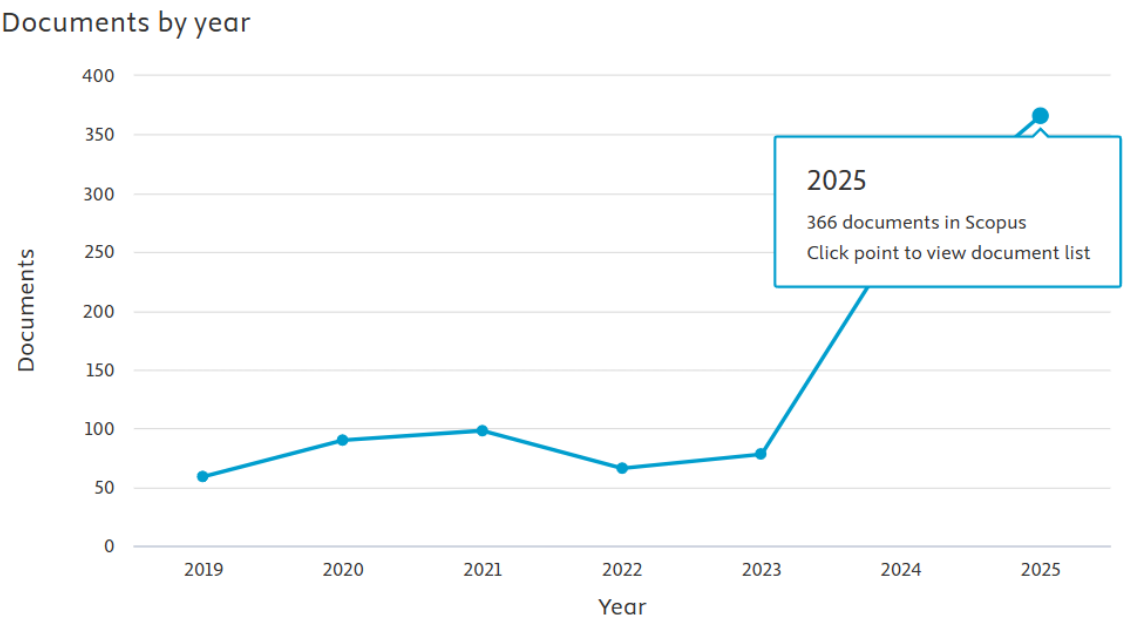


Nota. Elaborado a partir de la herramienta VOSViewer. (Scopus, 2025). https://www.scopus.com/home.uri

Las palabras clave asociadas a la revisión bibliográfica permitió identificar un conjunto de términos comunes que orientan las principales líneas de investigación. Las más destacadas predictive modeling, predictive models, Deep Learning, Artificial Neural Network, prediction, risk assessment, modeling y risk management. Este hallazgo evidencia el desarrollo de enfoques basados en la predicción y el uso de algoritmos avanzados para predecir resultados en proyectos. Encontrar los términos Deep Learning y Artificial Neural Networks evidencia el interés por técnicas capaces de procesar grandes volúmenes de datos. Las frases risk assessment y risk management evidencian el interés por aplicar estos modelos predictivos a la identificación y mitigación de riesgos, con el fin de fortalecer la toma de decisiones estratégicas en sectores de alta incertidumbre como el Oil & Gas.

Figura 3

Cantidad de documentos por año

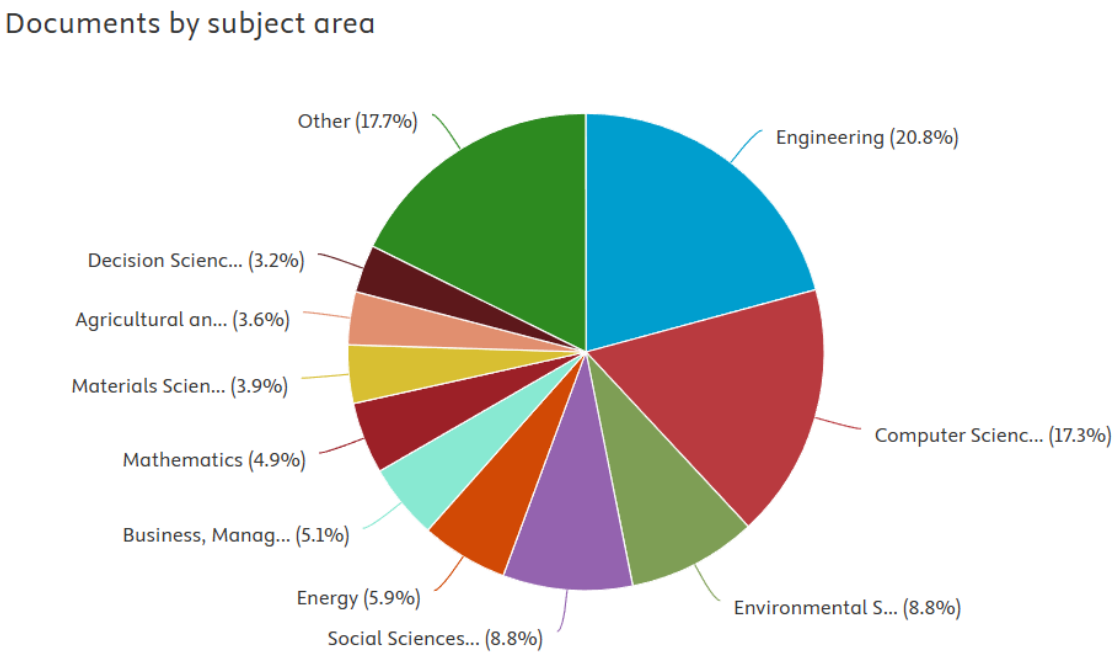


Nota. Elaborado a partir de la herramienta VOSViewer. (Scopus, 2025). https://www.scopus.com/home.uri

La figura 3 muestra el número de documentos emitidos por año donde se refleja que el año 2023 fue el año donde más crecimiento se percibió en el número de publicaciones después de un ajuste en el año 2022 producto de un crecimiento continuo con un pico de 100 publicaciones en el año 2021. La tendencia positiva en el número de publicaciones desde el año 2023 puede explicarse por el creciente interés y avance en las aplicaciones de la inteligencia artificial, más específicamente el Machine Learning en diferentes áreas del conocimiento.

Figura 4

Distribución por áreas de estudio



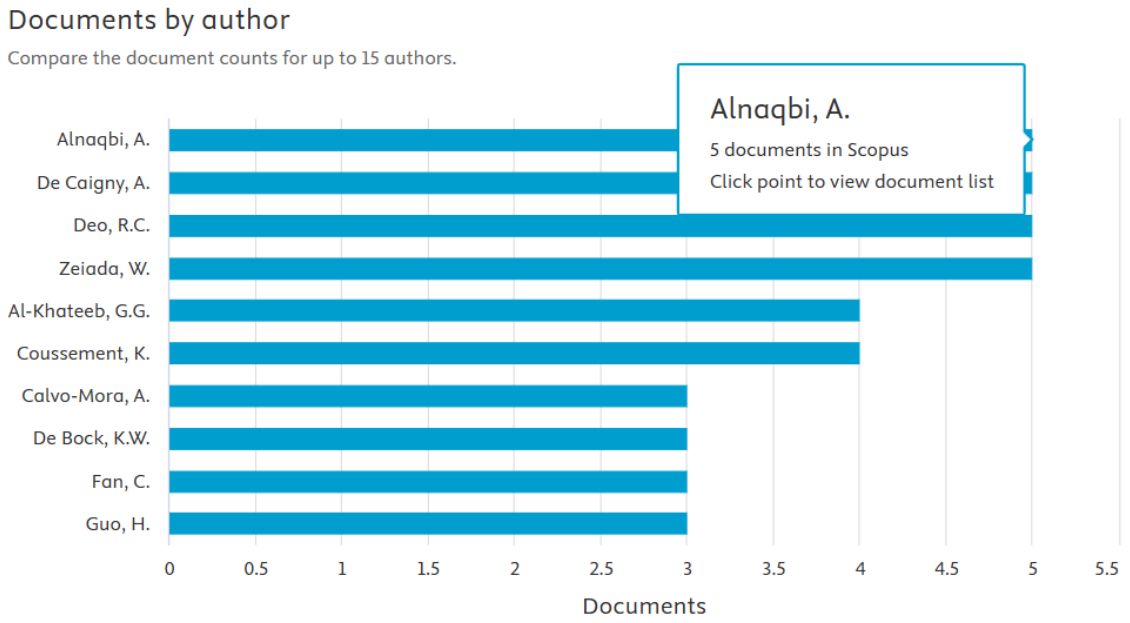
Nota. Elaborado a partir de la herramienta VOSViewer. (Scopus, 2025). https://www.scopus.com/home.uri

La figura 4. muestra la distribución porcentual de publicaciones científicas en Machine Learning (ML) y Management clasificadas por áreas de estudio. Se observa una concentración clara en disciplinas relacionadas con la ingeniería, ciencias computacionales, y otros, lo cual refleja tanto el carácter interdisciplinario del ML como su aplicación en sectores estratégicos.

El análisis bibliométrico evidencia que el Machine Learning se encuentra en la intersección entre la ciencia computacional y la gestión aplicada a diferentes áreas como Ciencias de decisión, agricultura, ciencia de los materiales, matemáticas, negocios, gestión, energía y ciencias sociales. Para el sector Oil & Gas, esta distribución confirma que el ML se desarrolla desde la perspectiva técnica y se proyecta hacia la optimización de la planificación de proyectos, gestión de riesgos financieros y cumplimiento regulatorio, integrando múltiples disciplinas.

Figura 5

Número de documentos por autor



Nota. Elaborado a partir de la herramienta VOSViewer. (Scopus, 2025). https://www.scopus.com/home.uri

El análisis de la figura 5 de documentos por autor**,** que compara la cantidad de documentos publicados por los autores más productivos en el área de aprendizaje entre 2019 y 2025, permite identificar a los principales referentes bibliográficos del campo. Destaca que Alnaqbi, A., De Caigny, A., Deo, R.C., Zeiada, W. lideran la producción con cinco artículos cada uno, estableciéndose como los autores con mayor capacidad de aporte y visibilidad científica.

El resto de los autores, como Al-Kathebb, G.G., Coussement, K., muestran una productividad sostenida de cuatro documentos en el periodo analizado, lo que evidencia una participación significativa y continua en la generación de conocimiento especializado. Esta distribución sugiere la existencia de una base de investigadores activos y recurrentes, aunque el liderazgo claramente recae en los autores más citados y prolíficos.

En conclusión, el campo está marcado por la consolidación de líderes académicos y la presencia de un grupo estable de expertos que contribuyen de forma relevante al desarrollo de la disciplina. La concentración de publicaciones en unos pocos autores también refleja su influencia como referencia obligada en futuras investigaciones y la potencial formación de colaboraciones estratégicas dentro del área de modelos predictivos, Machine Learning e inteligencia artificial aplicado a la gestión empresarial.

Diseño Metodológico

Alcance de la Investigación

El presente se plantea bajo un enfoque descriptivo, que permitirá recopilar, analizar y presentar el conjunto de datos, con el objetivo de identificar asociaciones entre las distintas variables que permitan proporcionar una perspectiva del escenario actual como un paso previo a la preparación y construcción de modelos predictivos. La investigación contempla la selección del modelo con mayor aplicación al estudio de caso y la posterior evaluación del rendimiento en el conjunto de prueba de los datos propio del estudio. El tratamiento de los datos partirá de un análisis transversal, dado que la recolección de datos se realiza en un momento único, a través de la observación y permitiendo describir a la población de estudio. Este enfoque ha sido utilizado en estudios de diagnóstico organizacional, así como en intervenciones orientadas a la transformación digital y en análisis predictivo. Diversos estudios han señalado que la integración de inteligencia artificial (AI) y Machine Learning en el sector de ingeniería permite anticipar riesgos financieros, respaldando la importancia del enfoque cuantitativo para evaluar el impacto de los modelos implementados (Bauskar et al., 2024; Mohseni & Mustafa Kamal, 2025).

Tipo de investigación

El estudio contempla ser desarrollado bajo una perspectiva descriptiva, ya que se busca caracterizar, explicar o describir las diferentes condiciones o relaciones entre las variables determinantes en los proyectos de ingeniería. A través de este diseño, la investigación busca

establecer la causalidad y efectividad de los nuevos métodos predictivos frente a la línea base de estimación, generando evidencia robusta para el proceso de transformación hacia un enfoque predictivo basado en modelos de inteligencia artificial dentro de Audubon.

Análisis Descriptivo

Actualmente la empresa Audubon cuenta con una base de datos de proyectos cerrados que contiene 15.861 registros acumulados desde el año 2014 hasta el año 2024 (Audubon, 2025b). Para propósitos de la investigación se contempla la utilización de la fuente de datos secundaria de Audubon para extraer una muestra probabilística estratificada, ya que como afirma Daraz et al. (2024), en cierta medida este tipo de muestreo aleatorio puede ayudar a lidiar con la variabilidad de los datos en las fases de planificación, diseño y estimación porque se generan grupos homogéneos conocidos como estratos a partir de características comunes de los datos. Esto garantizará que todos los grupos importantes se encuentren representados en la muestra, siendo útil en ML para mantener la distribución de variables claves en la muestra.

El modelo de Machine Learning deberá trabajar con un conjunto de datos relevantes que reflejen la variedad y características de los proyectos en un entorno real para asegurar la diversidad y variabilidad de la información, incluyendo diferentes casos y escenarios positivos y negativos en la ejecución de los proyectos de manera que el algoritmo aprenda a distinguir entre ellos.

El estrato que permitirá seleccionar la muestra probabilística será el tipo de contrato de cada proyecto dentro de los que se encuentra tiempo y materiales, costo y tarifa, suma global fija, entre otros (Audubon, 2025b); está variable es una de las consideraciones claves a la hora de estimar, planificar, desarrollar y entregar estos proyectos de ingeniería al cliente cumpliendo con la triple restricción de alcance, tiempo y costo. Para la determinación del tamaño de la muestra hay que tener en cuenta varios factores como la obtención de resultados más precisos y confiables, los recursos disponibles de tiempo, capacidad de procesamiento de máquina y la

variabilidad de los datos, razón por la cual se espera desarrollar el modelo con una muestra del 30% de los datos para obtener una representación suficiente de la población total en el modelo.

Antes de entrenar un modelo predictivo, es muy importante realizar una EDA (Exploratory Data Analysis) o análisis exploratorio de los datos para comprender y analizar mejor el conjunto de datos, con el fin de encontrar patrones, similitudes y detectar posibles errores en los datos, otorgando al analista indicios sobre que variables pueden llegar a ser adecuadas como predictores en los modelos (Hernández & Baquero, 2025).

A partir de los resultados del análisis exploratorio, se determinarán y segmentarán los datos y variables que harán parte del preprocesamiento de manera que se asegure la interpretabilidad de la información por parte del algoritmo de machine learning; esta etapa permitirá limpiar, transformar o imputar datos para corregir valores faltantes o segmentar las características o variables más significativas del set de datos. Así mismo siguiendo las técnicas de ML se dividirá la muestra total en conjuntos de entrenamiento, validación y prueba para de esta manera evaluar el rendimiento del modelo de manera adecuada.

Enfoque de la investigación

La presente investigación adoptará el enfoque cuantitativo como eje transversal para la evaluación y mejora de los procesos de planificación en los proyectos de diseño de ingeniería en Audubon, sucursal colombiana. La aplicación de la metodología cuantitativa permite la recolección y análisis de datos numéricos usando métodos estadísticos con el fin de encontrar relaciones lineales y no lineales entre las variables críticas que inciden en la gestión de los proyectos. Este tipo de enfoque, ampliamente respaldado en la literatura actual permite el estructurar y segmentar el conjunto de datos con el propósito de obtener modelos de aprendizaje automático capaces de generalizar y entender a partir de nuevos datos la realidad compleja de los proyectos de ingeniería.

Como señalan Bauskar et al. (2024), la metodología cuantitativa se ha convertido en el marco integro para trabajar con modelos de aprendizaje automático, ya que permite la

manipulación y comprensión de grandes volúmenes de datos usando técnicas estadísticas como base y de esta manera obtener algoritmos que interpreten de manera efectiva los datos y anticipen riesgos financieros y operativos. Mohseni y Mustafa Kamal (2025) subrayan que el rigor cuantitativo aporta solidez a los procesos de entrenamiento y validación, permitiendo comparar entre distintos modelos matemáticos para encontrar la fuente que describa mejor los datos, destacando la utilidad de la estadística descriptiva y correlacional en el diagnóstico organizacional y la mejora continua en entornos altamente competitivos como el sector Oil & Gas.

A través del enfoque cuantitativo se busca garantizar que la medición se realice con objetividad sin alterar ninguna de las variables con el fin de entender el origen medible de los datos y la relación que conservan con cada tipo de proyecto, esto asegura la aplicabilidad y validez estadística de los resultados. Es así como este enfoque facilita la sistematización, el análisis de grandes volúmenes de datos históricos y dota a la investigación de una capacidad interpretativa y predictiva de relevancia internacional, fundamental para la toma de decisiones estratégicas en el sector Oil & Gas (Bauskar et al., 2024; Mohseni & Mustafa Kamal, 2025).

Variables del estudio de investigación

El conjunto de variables que se presenta en la Tabla 2 son los parámetros recolectados a lo largo de los años de los proyectos cerrados en la empresa Audubon, sucursal colombiana (Audubon, 2025a).Este conjunto de datos está compuesto por 58 variables de tipo cualitativo y cuantitativo que describen y caracterizan a cada uno de los proyectos, reflejando características de interés que pueden ser tenidas en cuenta para el desarrollo del modelo, el diccionario de variables se presenta en la tabla 2.

Tabla 2

Diccionario de variables de estudio

Variable	Tipo	Definición
Project Health	Cualitativa	Salud del Proyecto de acuerdo a parámetros financieros como sobre ejecución, Cuentas por cobrar mayores a 90 días, Cuentas sin facturar mayores a 60 días, Margen de ganancia, etc. Fuente Audubon
Health Tags	Cualitativa	Etiquetas que identifican el estado del proyecto. Fuente Audubon
Sync Time	Cualitativa	última fecha de sincronización del proyecto. Fuente Audubon
Start Date	Cualitativa	Fecha de inicio del proyecto. Fuente Audubon
Duration	Cuantitativa	Duración estimada del proyecto. Fuente Audubon
Duration UOM	Cualitativa	Tipo de unidad estándar de medida para la medición de la duración. Fuente Audubon
PO	Cualitativa	Número de orden de contrato. Fuente Audubon
Company	Cualitativa	Nombre de la empresa filial. Fuente Audubon
Owning Company	Cualitativa	Nombre de la empresa dueña. Fuente Audubon
Execution Center	Cualitativa	Oficina de ejecución. Fuente Audubon
BU	Cualitativa	Unidad de negocio de la empresa. Fuente Audubon
Project Id	Cualitativa	Código único de asignación de proyecto. Fuente Audubon
Name	Cualitativa	Nombre descriptivo del proyecto. Fuente Audubon
Client Name	Cualitativa	Nombre del cliente o dueño del proyecto. Fuente Audubon
Manager	Cualitativa	Gerente del proyecto. Fuente Audubon
Type	Cualitativa	Tipo de contrato del proyecto de ingeniería. (CP = + Costo + tarifa, CPI= Costo + tarifa entre filiales, TM= Tiempo y Materiales, TMI= Tiempo y Materiales entre filiales, FFC = Suma Global Fija). Fuente Audubon
T/W Status	Cualitativa	Estado actual del proyecto (Activo/Inactivo para cargue de horas). Fuente Audubon
Contract Value	Cuantitativa	Valor en moneda corriente del contrato. Fuente Audubon
Labor Current Budget	Cuantitativa	Presupuesto directo del proyecto. Fuente Audubon
Non-Labor Current Budget	Cuantitativa	Presupuesto indirecto del proyecto. Fuente Audubon
Labor Cost PTD	Cuantitativa	Costo directo del proyecto. Fuente Audubon
Non-Labor Cost PTD	Cuantitativa	Costo indirecto del proyecto. Fuente Audubon

Hourly Budget	Cuantitativa	Cantidad de horas presupuestadas para la ejecución del proyecto. Fuente Audubon
PTD Actuals	Cuantitativa	Cantidad de horas utilizadas a la fecha. Fuente Audubon
Remaining Budget	Cuantitativa	Cantidad de horas restantes por ejecutar. Fuente Audubon
% Spent	Cuantitativa	Porcentaje de horas ejecutadas. Fuente Audubon
Revenue Current Budget	Cuantitativa	Ingresos presupuestados actuales. Fuente Audubon
Revenue PTD	Cuantitativa	Ingresos acumulados hasta la fecha. Fuente Audubon
Remaining Budget ($)	Cuantitativa	Presupuesto restante por ejecutar. Fuente Audubon
% Spent ($)	Cuantitativa	Porcentaje de ingresos ejecutados. Fuente Audubon
Billed To Date	Cuantitativa	Valor monetario facturado a la fecha. Fuente Audubon
Paid To Date	Cuantitativa	Valor monetario pagado a la fecha. Fuente Audubon
Unbilled	Cuantitativa	Valor monetario sin facturar a la fecha. Fuente Audubon
EAC	Cuantitativa	Métrica de gestión de proyectos que proyecta el costo total de un proyecto al finalizar. Fuente Audubon
Estimated Completion Date	Cualitativa	Fecha estima de completamiento o finalización del proyecto. Fuente Audubon
Target GM	Cuantitativa	Objetivo proyectado de ganancia para el proyecto. Fuente Audubon
Target HVEC	Cuantitativa	Objetivo de contribución o apoyo de oficinas aliadas. Fuente Audubon
Actual GM	Cuantitativa	Margen de ganancia actual. Fuente Audubon
Gross GM PTD	Cuantitativa	Margen de ganancia obtenido a la fecha. Fuente Audubon
Planned HVEC	Cuantitativa	Meta planeada de contribución o apoyo de oficinas aliadas. Fuente Audubon
HVEC to Date	Cuantitativa	Horas consumidas de trabajo de oficinas aliadas a la fecha. Fuente Audubon
HVEC % To Date	Cuantitativa	Porcentaje de contribución o apoyo de oficinas aliadas a la fecha. Fuente Audubon
HVEC FC	Cuantitativa	Distribución de horas localizadas en oficinas aliadas. Fuente Audubon
US to Date	Cuantitativa	Distribución de horas localizadas en oficinas de Estados Unidos. Fuente Audubon
US FC	Cuantitativa	Distribución de horas planeadas en oficinas de Estados Unidos. Fuente Audubon
Unbilled 0 to 30	Cuantitativa	Rango de 0 a 30 días de moneda corriente sin facturar. Fuente Audubon

Unbilled 31 to 45	Cuantitativa	Rango de 31 a 45 días de moneda corriente sin facturar. Fuente Audubon
Unbilled 46 to 60	Cuantitativa	Rango de 46 a 60 días de moneda corriente sin facturar. Fuente Audubon
Unbilled 61 to 90	Cuantitativa	Rango de 61 a 90 días de moneda corriente sin facturar. Fuente Audubon
Unbilled 90 to 120	Cuantitativa	Rango de 90 a 120 días de moneda corriente sin facturar. Fuente Audubon
Unbilled 120+	Cuantitativa	Rango mayor a 120 días de moneda corriente sin facturar. Fuente Audubon
Total Unbilled	Cuantitativa	Total de dinero sin facturar a la fecha. Fuente Audubon
AR Current	Cuantitativa	Valor monetario actual de cuentas por cobrar. Fuente Audubon
AR 1 to 30	Cuantitativa	Rango de 1 a 30 días de cuentas por cobrar. Fuente Audubon
AR 31 to 60	Cuantitativa	Rango de 31 a 60 días de cuentas por cobrar . Fuente Audubon
AR 61 to 90	Cuantitativa	Rango de 61 a 90 días de cuentas por cobrar. Fuente Audubon
AR 91+	Cuantitativa	Rango mayor a 91 días de cuentas por cobrar. Fuente Audubon
Total Aging	Cuantitativa	Total de dinero sin recibir a la fecha. Fuente Audubon
Last Charge Date	Cualitativa	Fecha ultima de cargue de horas. Fuente Audubon
Average Days To Pay	Cuantitativa	Días promedio para pagar. Fuente Audubon
SPI	Cuantitativa	Métrica que mide la eficiencia del cronograma. Fuente Audubon
CPI	Cuantitativa	Métrica que mide la eficiencia de los costos. Fuente Audubon
Complete Progress	Cuantitativa	Progreso del proyecto (1 para completado y 0 sin completar). Fuente Audubon
ST Multiplier	Cuantitativa	Multiplicador regular para la variable horas de ejecución. Fuente Audubon
OT Multiplier	Cuantitativa	Multiplicador de extras para la variable horas de ejecución. Fuente Audubon
% Progress	Cuantitativa	Porcentaje de progreso del proyecto. Fuente Audubon

Nota. Elaboración propia.

La tabla anterior describe y clasifica cada una de las variables del conjunto de datos de proyectos cerrados que hacen parte de la información secundaria de la empresa Audubon sucursal Colombiana, que serán utilizadas para el planteamiento del modelo de ML.

Ruta metodológica para el cumplimiento de los objetivos

El diseño metodológico de la presente investigación se estructura en cuatro fases que responden al objetivo general de formular una estrategia basada en *Machine Learning* para la planificación de proyectos de diseño en ingeniería en la empresa Audubon, Sucursal Colombiana. Cada fase se fundamenta en prácticas de gestión de proyectos y en los aportes recientes de la literatura académica sobre la aplicación de modelos de aprendizaje automático en el sector Oil & Gas y en industrias afines. La herramienta que se va a utilizar para el análisis y procesamiento de los datos será Python, reconocido como uno de los lenguajes de programación más robustos y versátiles en el ámbito de la ciencia de datos, el Machine Learning y la inteligencia artificial, esta herramienta cuenta con un amplio entorno de bibliotecas que permite llevar a cabo un análisis integral, comenzando con la exploración inicial hasta la validación de modelos predictivos. Como señala Caballero et al. (2023) este lenguaje permite el procesamiento de grandes volúmenes de información y la construcción de modelos analíticos. Hernández y Baquero (2025), señalan que Python permite aplicar de manera eficiente herramientas estadísticas y de aprendizaje automático en entornos empresariales.

Fase 1. Consolidación y preparación de datos históricos

Esta fase busca que la información base del estudio cumpla con criterios de calidad y sea coherente, esto debido a que los resultados de un modelo predictivo dependen de la integridad de los datos empleados (Datta et al., 2024b). Se tendrá en cuenta la recopilación de información de proyectos ejecutados entre 2014 y 2024, incluyendo variables categóricas referentes al tipo de proyecto, la unidad de negocio, el cliente, fecha de inicio del proyecto y variables numéricas como el valor del contrato, duración del proyecto, valor pagado a la fecha, etc. Posteriormente, se realizará la depuración, normalización, estandarización, codificación y selección de variables. La variable objetivo estará asociada al éxito financiero de los proyectos (proyectos financieramente sanos y proyectos con problemas), mientras que las variables predictoras incluirán variables categóricas y continuas de los proyectos. S. O. Abioye et al.

(2021) destaca que la estandarización de datos es una de las barreras más críticas en la implementación de inteligencia artificial en proyectos de ingeniería.

Fase 2. Análisis exploratorio y selección de variables críticas

En esta etapa se identificarán los factores más determinantes en el desempeño de los proyectos. Se aplicarán técnicas estadísticas y de visualización para reconocer patrones en los datos históricos, además de métodos de selección de variables como análisis de varianza y métricas de importancia generadas por algoritmos de aprendizaje automático. Estudios previos han demostrado que variables como la complejidad técnica, los cambios de alcance, las horas de ingeniería y los retrasos en permisos regulatorios influyen directamente en la viabilidad financiera (Bohórquez Castellanos & Mejia-Aguilar, 2019; Tshidavhu & Khatleli, 2020). Coffie & Cudjoe (2024) resaltan que la selección temprana de predictores relevantes fortalece la capacidad explicativa de los modelos.

Para el desarrollo de esta fase se emplearán librerías como Pandas y NumPy de Python para la manipulación, limpieza, transformación e imputación de datos faltantes. Por otra parte se usará Matplotlib y Seaborn como herramientas de visualización, para identificar patrones, tendencias y correlaciones iniciales en los proyectos analizados.

Fase 3. Modelado predictivo con algoritmos de Machine Learning

En esta fase se realizará el entrenamiento de los modelos seleccionados, iniciando con una división y segmentación de los datos en conjuntos de entrenamiento para validación y prueba, garantizando así un análisis confiable y controlado.

Una vez divididos los datos, se implementará en primer lugar, una regresión logística como modelo base, teniendo en cuenta su simplicidad y utilidad en la interpretación de relaciones entre variables como lo indica (Martínez Pérez & Pérez Martín, 2024). Posteriormente, se evaluarán los algoritmos Random Forest, D*ecision* Trees y *XGBoost*, útiles en la predicción de sobrecostos y en la clasificación de proyectos (Mali et al., 2025).

Como último paso de esta fase, se evaluará el ajuste de los modelos mediante validación cruzada, precisión, Recall, F1-score, AUC-ROC y matrices de confusión (López Ferreiro et al., 2025a).

Fase 4. Estrategias organizacionales y plan de intervención

La última fase se enfocará en la formulación de conclusiones y recomendaciones estratégicas, con el objetivo de mejorar y optimizar la planificación de proyectos de ingeniería en la empresa Audubon, mitigando de esta manera riesgos financieros a futuro. Adicionalmente los resultados de los modelos predictivos serán la base para plantear lineamientos para la gestión de datos, la implementación de herramientas analíticas y la necesidad de incluir nuevas y mejores variables medibles en los proyectos.

En línea con lo planteado por Dzhusupova et al. (2024), donde señalan que la adopción de inteligencia artificial en entornos empresariales depende en gran medida de la disposición organizacional y de la definición de políticas claras que faciliten su integración. La fase final no supone la ejecución práctica del modelo, en cambio, supone la construcción de un marco de recomendaciones estratégicas que permita a la gerencia de Audubon tomar decisiones fundamentadas y en el mediano plazo, avanzar hacia la integración responsable y sostenible de herramientas de *Machine Learning* en su proceso de planificación de proyectos.

Desarrollo y resultado de la investigación

Consolidación y preparación de la base de datos

Los modelos de clasificación son los modelos más predominantes cuando se habla de separar y segmentar los datos en dos o más conjuntos de resultados, siendo importante destacar que la base de datos contiene una etiqueta que determina el estado del proyecto; esta característica es la variable objetivo que se quiere llegar a predecir, por lo que teniendo en cuenta la base teórica e investigativa de varios autores, se opta por trabajar con el algoritmo XGBoost. Este algoritmo ha demostrado ser muy efectivo en la predicción de resultados y

además es ideal para trabajar con grandes sets de datos. Igualmente tal y como se describió anteriormente, se llevara a cabo el modelamiento con varios algoritmos de clasificación para evaluar su precisión con respecto a XGBoost.

La base de datos inicial estuvo conformada por 15.861 registros correspondientes a proyectos ejecutados bajo diferentes modalidades contractuales, entre las que se incluyen Tiempo y Materiales (TM), Costo más tarifa (CP), Suma Global Fija (FFC), Tiempo y Materiales entre compañías filiales (TMI), Costo más tarifa entre compañías filiales (CPI), Gastos Indirectos (OH) y Gastos Indirectos de Ofertas (OHP).

Como primer paso metodológico, se decidió excluir los registros asociados a gastos indirectos (OH y OHP), dado que no representan ingresos directos para la empresa. Estos conceptos corresponden a actividades de apoyo, gestión administrativa o preparación de ofertas, las cuales, aunque son necesarias para la operación general, no forman parte de los proyectos generadores de ingresos. Su función principal es permitir la trazabilidad de horas y costos internos asociados a labores de soporte y funcionamiento operativo, siendo contabilizados dentro de los gastos generales de la organización.

Una vez aplicada esta primera depuración, el conjunto de datos se redujo a 9.956 registros válidos, los cuales constituyen la base analítica para el desarrollo del modelo predictivo. Esta reducción garantiza que el análisis se centre exclusivamente en los proyectos con impacto financiero directo, eliminando posibles sesgos derivados de actividades no productivas o de carácter administrativo.

Finalizado el proceso de depuración, los datos fueron cargados y estructurados en el entorno de trabajo de Python, estableciendo la base inicial para el desarrollo de las etapas posteriores de análisis y modelado como se observa en la Figura 6.

Figura 6

Cargue de datos inicial en Python

```
Información del DataFrame actual:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 9956 entries, 0 to 9955
Data columns (total 67 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Project Health             9956 non-null   object
 1   Health Tags                5065 non-null   object
 2   Sync Time                  9956 non-null   object
 3   Start Date                 9956 non-null   object
 4   Duration                   9956 non-null   int64
 5   Duration UOM               8756 non-null   object
 6   PO                         9178 non-null   object
 7   Company                    9956 non-null   object
 8   Owning Company             9956 non-null   object
 9   Execution Center           9949 non-null   object
 10  BU                         9956 non-null   int64

 11  Project Id                 9956 non-null   object
 12  Name                       9956 non-null   object
 13  Client Name                9956 non-null   object
 14  Manager                    9956 non-null   object
 15  Type                       9956 non-null   object
 16  T/W Status                 9956 non-null   object
 17  Contract Value             9956 non-null   float64
 18  Labor Current Budget       9956 non-null   float64
 19  Non-Labor Current Budget   9956 non-null   float64
 20  Labor Cost PTD             9956 non-null   float64
 21  Non-Labor Cost PTD         9956 non-null   float64
 22  Hourly Budget              9956 non-null   float64
 23  PTD Actuals                9956 non-null   float64
 24  Remaining Budget           9956 non-null   float64
 25  % Spent                    9956 non-null   float64
 26  Revenue Current Budget     9956 non-null   float64
 27  Revenue PTD                9956 non-null   float64
 28  Remaining Budget ($)       9956 non-null   float64
```

```
 29  % Spent ($)                9956 non-null   object
 30  Billed To Date             9956 non-null   float64
 31  Paid To Date               9956 non-null   float64
 32  Unbilled                   9956 non-null   float64
 33  EAC                        9956 non-null   float64
 34  Estimated Completion Date  9666 non-null   object
 35  Target GM                  9956 non-null   float64
 36  Target HVEC                9956 non-null   float64
 37  Actual GM                  9956 non-null   object
 38  Gross GM PTD               9956 non-null   float64
 39  Planned HVEC               9956 non-null   float64
 40  HVEC to Date               9956 non-null   float64
 41  HVEC % To Date             9956 non-null   float64
 42  HVEC FC                    9956 non-null   int64
 43  US to Date                 9956 non-null   float64
 44  US FC                      9956 non-null   float64
 45  Unbilled 0 to 30           9956 non-null   int64
 46  Unbilled 31 to 45          9956 non-null   int64
 47  Unbilled 46 to 60          9956 non-null   int64
 48  Unbilled 61 to 90          9956 non-null   int64
 49  Unbilled 90 to 120         9956 non-null   int64
 50  Unbilled 120+              9956 non-null   float64
 51  Total Billed               9956 non-null   float64
 52  AR Aging Current           9956 non-null   int64
 53  AR Aging 1 to 30           9956 non-null   int64
 54  AR Aging 31 to 60          9956 non-null   int64
 55  AR Aging 61 to 90          9956 non-null   int64
 56  AR Aging 91+               9956 non-null   int64

 57  Total Aging                9956 non-null   int64
 58  Last Charge Date           9487 non-null   object
 59  Average Days To Pay        9956 non-null   float64
 60  SPI                        2 non-null      float64
 61  CPI                        2 non-null      float64
 62  % Complete Progress        9956 non-null   int64
 63  ST Multiplier              1219 non-null   float64
 64  OT Multiplier              4 non-null      float64
 65  % Progress                 9956 non-null   int64
 66  Status                     9956 non-null   int64
dtypes: float64(31), int64(17), object(19)
memory usage: 5.1+ MB
```

Nota. Elaboración propia.

Para la obtención de la muestra probabilística, se aplicó un proceso de estratificación de los datos con el propósito de conservar la representatividad proporcional de los distintos tipos de proyecto incluidos en la población. Este procedimiento permitió asegurar que la muestra

seleccionada mantuviera la misma distribución porcentual por categoría de proyecto observada en el conjunto total de datos.

La selección de la muestra se llevó a cabo en el entorno de Python, utilizando la librería sklearn y la función train_test_split, la cual permite dividir la base de datos de forma aleatoria. Como resultado se obtienen 3.982 registros tomando como base que se va a trabajar con el 40% de los datos que comprenden a la población total. Tal como se evidencia en la Figura 7, el resultado de la operación conserva la proporción porcentual de cada uno de los tipos de proyecto, garantizando que el muestreo estratificado es proporcional al tamaño de la población total.

Figura 7

Análisis probabilístico estratificado

```
Distribución de tipos de proyecto en el conjunto original:
Type
TM      0.625151
CP      0.145540
FFC     0.138309
TMI     0.056448
CPI     0.023905
NOPO    0.010647
Name: proportion, dtype: float64

Distribución de tipos de proyecto en la muestra estratificada:
Type
TM      0.625063
CP      0.145655
FFC     0.138373
TMI     0.056504
CPI     0.023857
NOPO    0.010547
Name: proportion, dtype: float64

Tamaño de la muestra estratificada: 3982
```

Nota. Elaboración propia.

Análisis exploratorio y selección de variables críticas

El análisis exploratorio de los datos inicia con las transformaciones necesarias a las variables que requerían ajustes, tales como la conversión de campos categóricos a formato numérico, la normalización de variables continuas y la corrección de inconsistencias

detectadas. Estos procesos se desarrollaron mediante el uso de las librerías pandas y NumPy en Python, garantizando la integridad estructural del conjunto de datos para su posterior modelado. Los resultados de estas transformaciones se ilustran en la Figura 8.

Figura 8

Transformación del tipo de datos

```
#Transformar el tipo de formato de algunas variables

df_sampled['BU'] = df_sampled['BU'].astype(str)

# Limpiar la columna 'Actual GM' antes de convertir a float
# Eliminar el '%' y convertir a numérico, convirtiendo errores a NaN
df_sampled['Actual GM'] = df_sampled['Actual GM'].astype(str).str.replace('%', '', regex=False)
df_sampled['Actual GM'] = pd.to_numeric(df_sampled['Actual GM'], errors='coerce')

# Limpiar la columna '% Spent ($)' antes de convertir a float
# Eliminar el '%' y convertir a numérico, convirtiendo errores a NaN
df_sampled['% Spent ($)'] = df_sampled['% Spent ($)'].astype(str).str.replace('%', '', regex=False)
df_sampled['% Spent ($)'] = pd.to_numeric(df_sampled['% Spent ($)'], errors='coerce')

# Convertir columnas a tipo fecha, coercing errors to NaT and specifying dayfirst
df_sampled['Estimated Completion Date'] = pd.to_datetime(df_sampled['Estimated Completion Date'], errors='coerce', dayfirst=True)
df_sampled['Start Date'] = pd.to_datetime(df_sampled['Start Date'], errors='coerce', dayfirst=True)
df_sampled['Sync Time'] = pd.to_datetime(df_sampled['Sync Time'], errors='coerce', dayfirst=True)
df_sampled['Last Charge Date'] = pd.to_datetime(df_sampled['Last Charge Date'], errors='coerce', dayfirst=True)
```

Nota. Elaboración propia.

Se procede a visualizar las primeras cinco filas con el fin de realizar una validación y comprobar que todas las variables se carguen correctamente en el entorno de trabajo. Posteriormente, se realiza una revisión del tipo de dato asignado a cada columna, con el objetivo de asegurar la coherencia entre la naturaleza de la variable y su formato en la base de datos como se evidencia en la Figura 9.

Figura 9

Primeras 5 filas del Dataframe

```
df_sampled.head()
```

	Project Health	Health Tags	Sync Time	Start Date	Duration	Duration UOM	PO	Company	Owning Company
0	OnTrack	NaN	2025-04-12	2020-08-21	13.0	weeks	MB20002-TO01	PCMLLC	PCMLLC
1	OnTrack	RevenueNearBudget	2025-04-12	2023-09-20	2.0	weeks	4504352105	AECLP	AECLP
2	OnTrack	NaN	2025-04-12	2019-03-08	12.0	weeks	SO 4526274889	AECLP	AECLP
3	OnTrack	RevenueNearBudget	2025-04-12	2017-07-31	2.0	weeks	Cost Center 46630061	PCMLLC	PCMLLC
4	OnTrack	NaN	2025-04-12	2023-01-09	8.0	weeks	4500037168	AECLP	AECLP

Nota. Elaboración propia.

Resumen Descriptivo

El resumen descriptivo de las variables cuantitativas, obtenido mediante la función .describe() de la librería pandas en Python, permite visualizar de manera general y resumida las principales características estadísticas de las variables numéricas. De acuerdo con Ekbote et al. (2023) este análisis exploratorio comprende varios componentes clave que son esenciales para obtener información y comprender la estructura subyacente de un conjunto de datos.

Previo al análisis descriptivo, se depuró el conjunto de datos con el fin de eliminar las columnas que no aportaban valor analítico directo al estudio. Entre ellas se encuentran aquellas que almacenaban fechas, identificadores numéricos de los proyectos y tiempos de sincronización de la base de datos, variables cuya naturaleza no contribuye al análisis estadístico ni a la etapa de modelación predictiva. Esta limpieza permitió enfocar el examen exclusivamente en los atributos con contenido cuantitativo relevante.

Posteriormente, se aplicó la función .describe() de la librería pandas en Python, la cual genera un resumen estadístico de las variables numéricas del conjunto de datos. Este procedimiento proporcionó información de los principales indicadores descriptivos. Los resultados de este análisis se ilustran en la Figura 10.

Figura 10

Exclusión de ciertas características iniciales

```
df_sampled = df_sampled.drop(columns=['Sync Time', 'Project Id', 'Start Date', 'Last Charge Date', 'Estimated Completion Date'])

df_sampled.describe()
```

Nota. Elaboración propia.

Al aplicar la función .describe() al conjunto de datos, se observa inicialmente que algunas columnas como Actual GM, SPI, CPI, ST Multiplier y OT Multiplier presentan una cantidad desigual de registros, lo que indica la existencia de valores faltantes. Por ello, en etapas posteriores se realizará un análisis exhaustivo de valores nulos con el fin de identificar, corregir o suprimir registros, columnas o valores ausentes que puedan comprometer la integridad del modelo.

De igual manera, se identificó que variables como Unbilled 0 to 30, Unbilled 31 to 45, Unbilled 46 to 60, Unbilled 61 to 90, Unbilled 90 to 120, Unbilled 120+, AR Aging 1 to 30, AR Aging 31 to 60, AR Aging 61 to 90, AR Aging 91+ y % Progress no contienen valores registrados. Esta ausencia sistemática de información sugiere que dichas métricas no eran recolectadas durante los primeros años de creación de la base de datos, lo que explica la nulidad observada y plantea la necesidad de evaluar su exclusión o imputación según su relevancia estadística.

Respecto a la duración de los proyectos, los resultados indican una media de 11,8 semanas, una desviación estándar de 16 semanas y una mediana de 6 semanas, lo que revela una asimetría positiva en la distribución. Este comportamiento sugiere que existen proyectos con duraciones que se alejan considerablemente del promedio y aumentan la variabilidad de los datos, generando una distribución sesgada hacia la derecha. La desviación estándar superior a

la media confirma la alta dispersión en la temporalidad de los proyectos, reflejando la presencia de valores atípicos y una distribución no normal de los datos.

Con respecto a las variables financieras como Contract Value, Labor Current Budget, Non-Labor Current Budget, Labor Cost PTD, Non-Labor Cost PTD, Revenue Current Budget, Remaining Budget, Billed to Date y Paid to Date presentan diferencias significativas entre la media y la mediana, siendo la primera sustancialmente mayor. Este comportamiento indica la existencia de valores extremos que desplazan la distribución hacia la derecha, una tendencia que se evidencia con mayor claridad en las gráficas de distribución para variables continuas.

Finalmente, se observa que las variables continuas no presentan patrones de distribución definidos que permitan establecer relaciones consistentes entre los distintos tipos de proyectos. Este alto grado de variabilidad e irregularidad sugiere que el conjunto de datos podría generar bajo sesgo (BIAS) y alta varianza durante el proceso de modelado, lo cual representa un desafío para la generalización del modelo predictivo. Los resultados detallados de esta etapa se ilustran en la Figura 11.

Figura 11

Resumen descriptivo para algunas variables cuantitativas

```
df_sampled.describe()
```

	Duration	Contract Value	Labor Current Budget	Non-Labor Current Budget	Labor Cost PTD	Non-Labor Cost PTD	Hourly Budget	PTD Actuals	Remaining Budget	% Spent	Revenue Current Budget	Revenue PTD	Remaining Budget ($)	3
count	3,716.0	3,716.0	3,716.0	3,716.0	3,716.0	3,716.0	3,716.0	3,716.0	3,716.0	3,716.0	3,716.0	3,716.0	3,716.0	
mean	11.8	145,509.6	24,321.3	2,010.9	64,956.3	18,514.5	786.4	1,139.1	-352.7	12.5	148,782.2	134,271.9	13,603.6	
std	16.0	1,193,700.6	184,373.0	26,078.0	522,050.2	171,014.3	6,307.5	9,570.6	5,953.8	211.7	1,209,139.5	1,212,318.6	188,355.1	
min	0.0	-121,625.0	-49,436.4	-1,000.0	0.0	-54,090.5	-260.0	0.0	-307,457.0	0.0	0.0	-42,300.0	-2,190,728.9	
25%	2.0	6,915.0	0.0	0.0	2,462.8	0.0	12.0	43.0	-88.1	0.2	7,650.9	5,444.8	0.0	
50%	6.0	22,000.0	0.0	0.0	8,702.3	237.6	105.0	153.0	-3.5	0.9	23,364.0	18,627.0	170.0	
75%	15.0	72,989.8	1,410.4	0.0	34,757.3	3,522.1	420.2	554.3	22.5	1.2	75,071.5	65,616.4	4,866.1	
max	383.0	46,588,533.0	8,432,627.0	1,055,628.0	20,470,095.0	8,820,635.3	246,802.0	322,069.0	39,955.2	8,188.6	46,588,537.0	48,779,265.9	8,551,497.0	

% Spent ($)	Billed To Date	Paid To Date	Unbilled	EAC	Target GM	Target HVEC	Actual GM	Gross GM PTD	Planned HVEC	HVEC to Date	HVEC % To Date	HVEC FC	US to Date	US FC
3,706.0	3,716.0	3,716.0	3,716.0	3,716.0	3,716.0	3,716.0	3,715.0	3,716.0	3,716.0	3,716.0	3,716.0	3,716.0	3,716.0	3,716.0
83.1	135,881.5	134,965.0	-43.8	143,353.1	0.1	0.0	32.3	50,801.1	0.0	161.6	0.1	0.4	1,329.2	1.0
53.8	1,249,594.8	1,212,843.4	2,057.3	1,204,193.9	0.2	0.1	21.1	638,849.8	0.0	2,263.0	0.2	14.9	9,761.3	25.3
0.0	0.0	0.0	-116,242.9	0.0	0.0	0.0	-307.3	-611,443.0	0.0	0.0	0.0	0.0	0.0	0.0
64.7	5,518.7	5,574.4	0.0	6,500.0	0.0	0.0	21.4	1,549.5	0.0	0.0	0.0	0.0	44.0	0.0
97.0	18,819.2	18,872.8	0.0	20,674.8	0.0	0.0	33.6	5,806.4	0.0	0.0	0.0	0.0	166.5	0.0
100.0	65,691.2	65,807.8	0.0	71,350.9	0.3	0.0	45.6	20,289.0	0.0	0.0	0.0	0.0	632.8	0.0
801.6	48,779,265.9	48,779,265.9	9,153.9	46,588,537.0	1.0	1.0	136.8	27,802,474.9	1.0	89,677.3	1.0	720.0	320,248.5	1,092.0

Nota. Elaboración propia.

Evaluación de valores Nulos

La revisión de valores nulos es una de las etapas más importantes de la exploración de datos porque una cantidad considerable de datos nulos puede desbalancear considerablemente el modelo afectando la precisión y rendimiento de este. Puede haber muchas razones que generen pérdida de datos o valores no encontrados en la base datos de Audubon siendo la principal la actualización y adición en los últimos años de nuevas columnas que no incluyen datos para los proyectos con mayor longevidad en la base de datos. Tal y como se observa en la figura 12, características como ‘Health Tags’ que representa una variable categórica dividida en cuatro clases para clasificar los proyectos (Overbudget, Budget near Contract, Unbilled, AR Aging), contiene 1814 valores nulos que representa un 48.8% del total de los registros de la muestra. Así mismo ‘Duration UOM’ con 13.1%, ‘PO’ con 8.5 % valores nulos, ‘% Spent’ con un 10% y finalmente dos variables ‘SPI’ y ‘CPI’ con un porcentaje de valores nulos del 100% demuestra que la organización en algún momento estableció estas dos nuevas características pero nunca se han recolectado datos en el período comprendido entre 2014 y 2024.

Figura 12

Proporción de valores nulos en el conjunto de datos

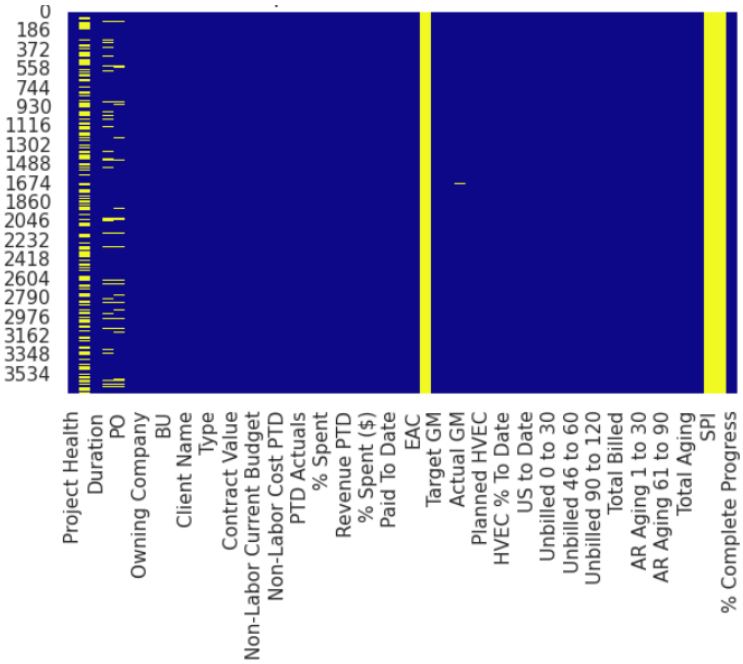
```
=== Valores nulos por columna ===
                        Missing Values  Percentage (%)
Project Health                       0            0.0
Health Tags                       1814           48.8
Duration                             0            0.0
Duration UOM                       485           13.1
PO                                 315            8.5
Company                              0            0.0
Owning Company                       0            0.0
Execution Center                     4            0.1
BU                                   0            0.0
Name                                 0            0.0
Client Name                          0            0.0
Manager                              0            0.0
Type                                 0            0.0
T/W Status                           0            0.0
Contract Value                       0            0.0
Labor Current Budget                 0            0.0
Non-Labor Current Budget             0            0.0
Labor Cost PTD                       0            0.0
Non-Labor Cost PTD                   0            0.0
Hourly Budget                        0            0.0
PTD Actuals                          0            0.0
Remaining Budget                     0            0.0
% Spent                              0            0.0
Gross GM PTD                         0            0.0
Planned HVEC                         0            0.0
HVEC to Date                         0            0.0
HVEC % To Date                       0            0.0
HVEC FC                              0            0.0
US to Date                           0            0.0
US FC                                0            0.0
Unbilled 0 to 30                     0            0.0
Unbilled 31 to 45                    0            0.0
Unbilled 46 to 60                    0            0.0
Unbilled 61 to 90                    0            0.0
Unbilled 90 to 120                   0            0.0
Unbilled 120+                        0            0.0
Total Billed                         0            0.0
AR Aging Current                     0            0.0
AR Aging 1 to 30                     0            0.0
AR Aging 31 to 60                    0            0.0
AR Aging 61 to 90                    0            0.0
AR Aging 91+                         0            0.0
Total Aging                          0            0.0
Average Days To Pay                  0            0.0
SPI                               3715          100.0
CPI                               3715          100.0
% Complete Progress                  0            0.0
```

Nota. Elaboración propia.

Para otorgar mayor claridad al investigador del conjunto de datos vacíos, en la figura 13 se visualiza el mapa de calor que representa esta nulidad de los datos en el Dataset de proyectos cerrados de la empresa Audubon.

Figura 13

Mapa de calor de valores nulos



Nota. Elaboración propia.

Como resultado del análisis y evaluación de valores nulos se excluyen las variables caracteristicas que contienen más del 40% de los datos nulos ya que no es conveniente imputar o borrar registros debido a la cantidad considerable de valores nulos en los datos. En total se eliminan 5 columnas y la nueva forma del Dataframe se compone de 3982 registros y 62 columnas. Ver figura 14.

Figura 14

Depuración de valores nulos

```
Se eliminaron 5 columnas:
- Estimated Completion Date: 82.67% de valores NaN
- SPI: 99.97% de valores NaN
- CPI: 99.97% de valores NaN
- ST Multiplier: 87.72% de valores NaN
- OT Multiplier: 99.92% de valores NaN

Forma original del DataFrame: (3982, 67)
Forma del DataFrame después de eliminar columnas: (3982, 62)
```

Nota. Elaboración propia.

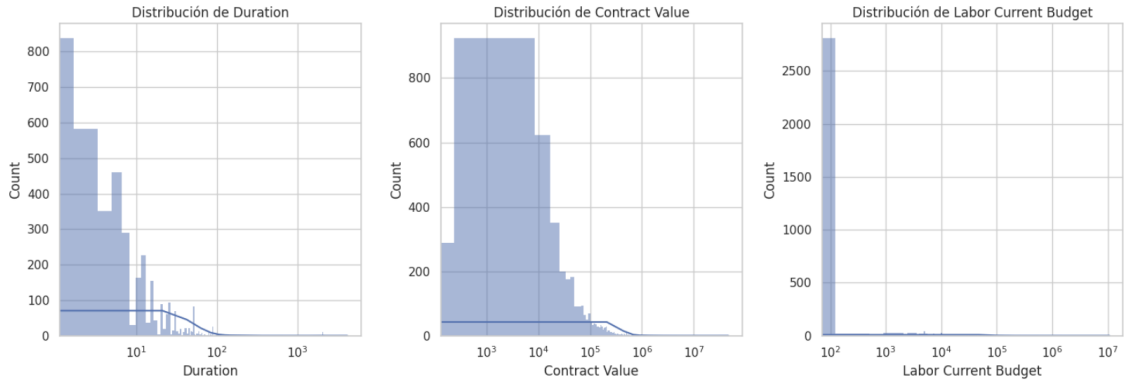
Histograma de frecuencias para variables continuas

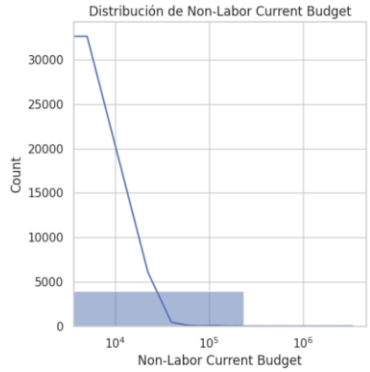
Los histogramas de las variables continuas correspondientes a datos financieros de los proyectos permiten observar visualmente el comportamiento de los números en las distintas características, confirmando lo descrito en el resumen descriptivo, variables como ‘Duration’ muestra que son muchos los proyectos de corta duración independientemente de que algunos se extiendan considerablemente. Las variables ‘Contract Value’, ‘Labor Cost PTD’, ‘Non Labor cost PTD’, ‘Hourly Budget’, ‘PTD Actuals’, ‘Remaining Budget’, ‘Revenue current Budget’, ‘Revenue PTD’, ‘Paid to Date’, ‘EAC’ entre otras características financieras describen el mismo patrón de desfase entre la media y la mediana, lo que genera que la distribución este sesgada hacia la derecha, claramente mostrando una alta variación entre proyectos pequeños y proyectos muy grandes debido a que hay proyectos que son significativamente más costosos.

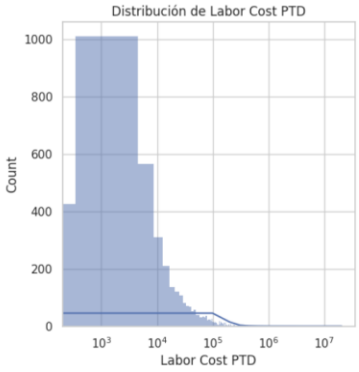
Por lo tanto se entiende que cada proyecto contiene su propia naturaleza de alcance, tiempo, y costo sin llegar a marcar tendencias o relaciones claras y semejantes entre las variables. Para el caso de variables como ‘Labor Current Budget’ se observa una cantidad considerable de datos atípicos que no reflejan ninguna distribución conocida. Por otro lado la variable ‘Actual GM’ que corresponde a los márgenes de ganancias obtenidos de los proyectos, conforma una única gráfica que parece aproximarse a una curva normal, centrada en un valor medio con una menor dispersión de los datos y sin sesgos notables, mostrando que la mayoría de los valores están dentro de un rango alrededor de la media, por lo que esta variable podría ser utilizada en algoritmos de ML sin transformaciones. Finalmente variables como relacionadas con periodos sin facturar y cuentas por cobrar (Unbilled y AR Aging) no describen ningún tipo de distribución y solo se observa una acumulación de los valores en un solo punto, sin variación visual en la gráfica; esto podría indicar que no hay registros suficientes o todos los registros tienen el mismo valor por lo que puede generar problemas a la hora de usar los datos como entrada para los modelos de ML. Variables que contienen valores constantes o muy reducidos determinan que sus datos no serán muy útiles en el análisis y se podrían omitir si no aportan información relevante al modelo.

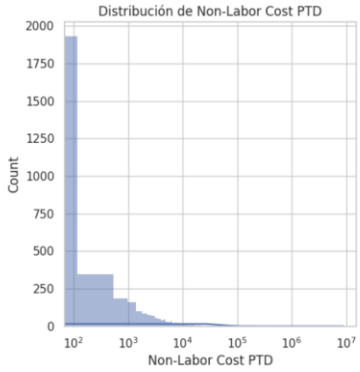
Figura 15

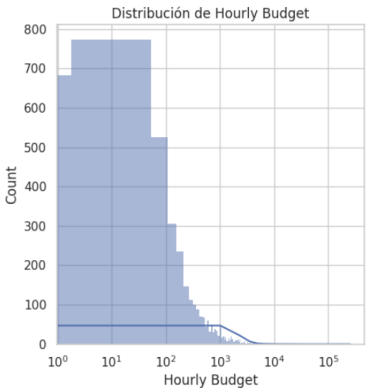
Histograma de frecuencias para variables continuas

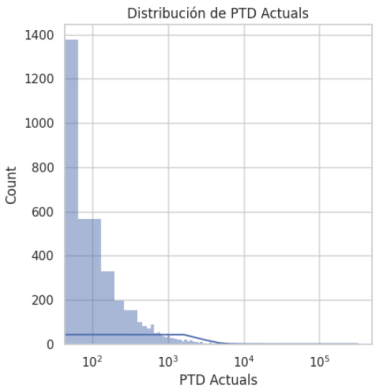


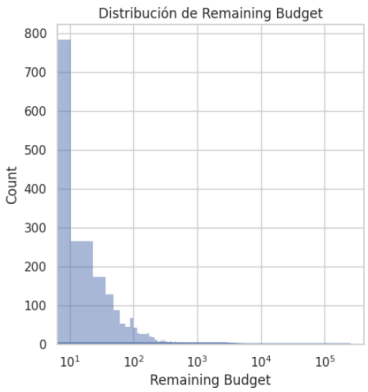

Distribución de Non-Labor Current Budget
Count
Non-Labor Current Budget

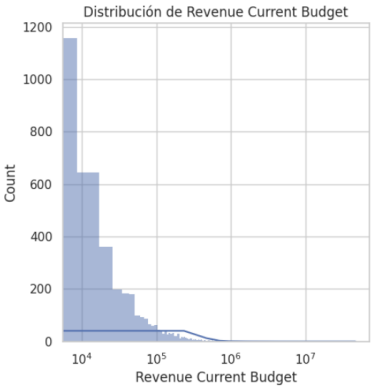

Distribución de Labor Cost PTD
Count
Labor Cost PTD

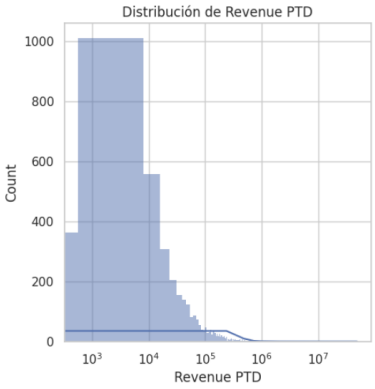

Distribución de Non-Labor Cost PTD
Count
Non-Labor Cost PTD

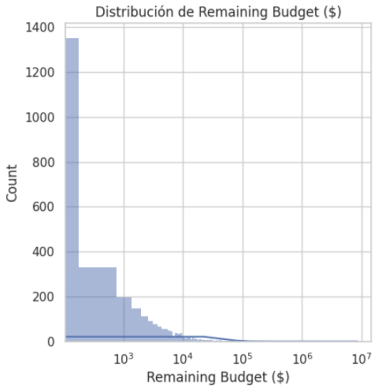

Distribución de Hourly Budget
Count
Hourly Budget

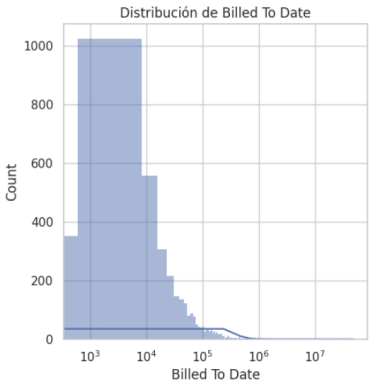

Distribución de PTD Actuals
Count
PTD Actuals

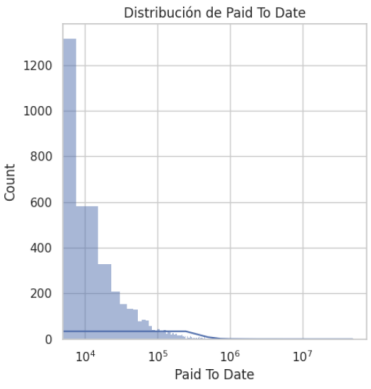

Distribución de Remaining Budget
Count
Remaining Budget

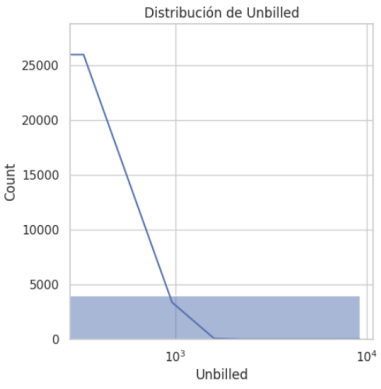

Distribución de Revenue Current Budget
Count
Revenue Current Budget

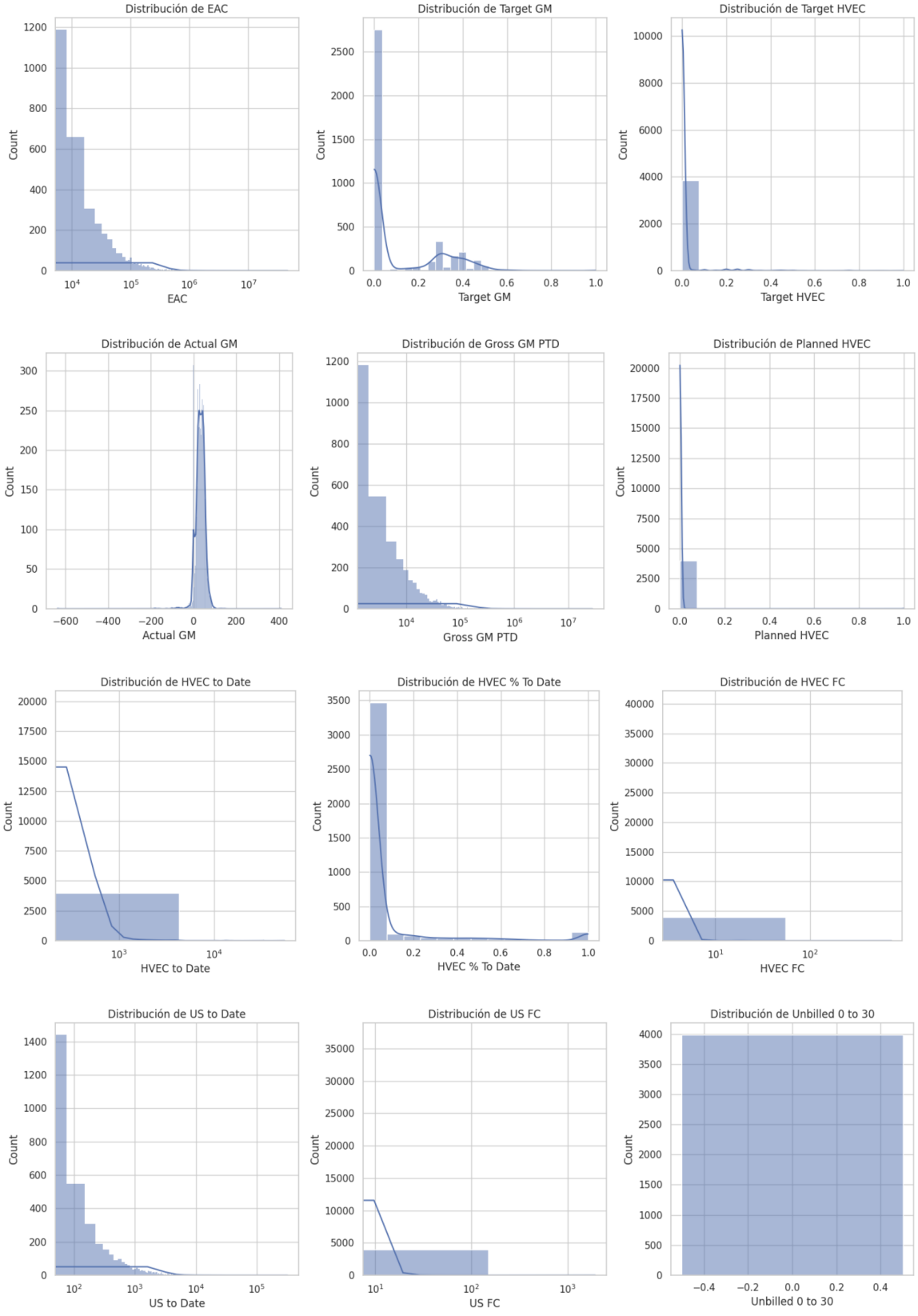

Distribución de Revenue PTD
Count
Revenue PTD

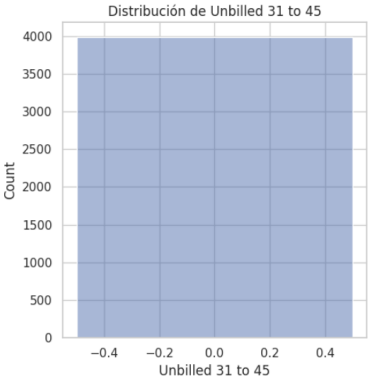

Distribución de Remaining Budget ($)
Count
Remaining Budget ($)

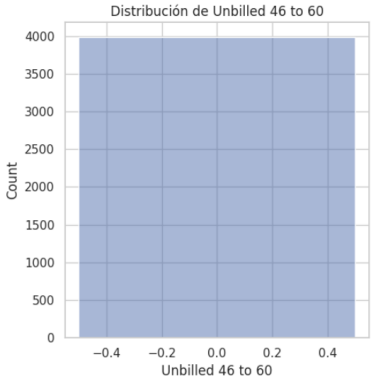

Distribución de Billed To Date
Count
Billed To Date

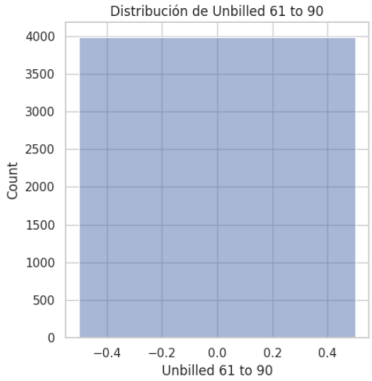

Distribución de Paid To Date
Count
Paid To Date

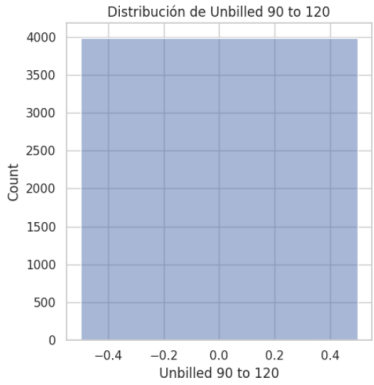

Distribución de Unbilled
Count
Unbilled

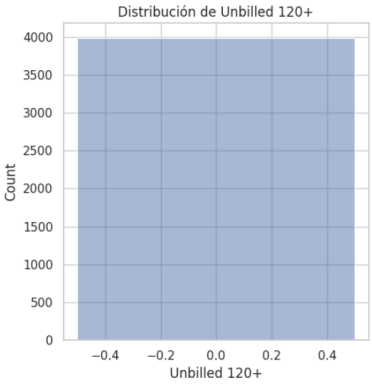

Distribución de EAC
Count
EAC
Distribución de Target GM
Count
Target GM
Distribución de Target HVEC
Count
Target HVEC
Distribución de Actual GM
Count
Actual GM
Distribución de Gross GM PTD
Count
Gross GM PTD
Distribución de Planned HVEC
Count
Planned HVEC
Distribución de HVEC to Date
Count
HVEC to Date
Distribución de HVEC % To Date
Count
HVEC % To Date
Distribución de HVEC FC
Count
HVEC FC
Distribución de US to Date
Count
US to Date
Distribución de US FC
Count
US FC
Distribución de Unbilled 0 to 30
Count
Unbilled 0 to 30

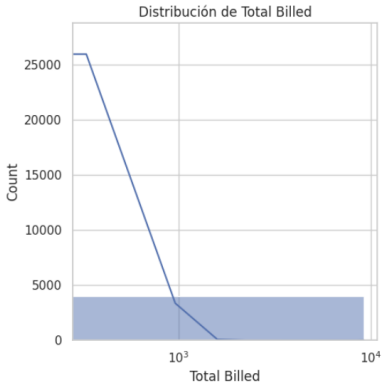

Distribución de Unbilled 31 to 45
Count
Unbilled 31 to 45

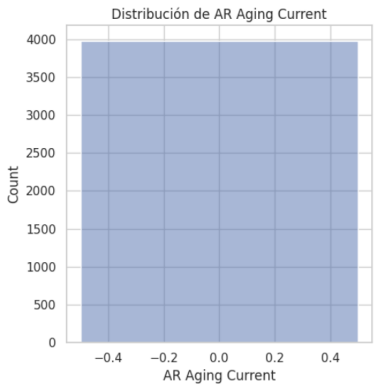

Distribución de Unbilled 46 to 60
Count
Unbilled 46 to 60

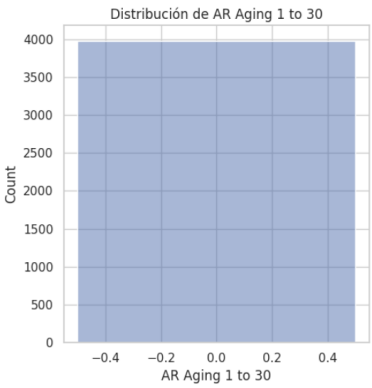

Distribución de Unbilled 61 to 90
Count
Unbilled 61 to 90

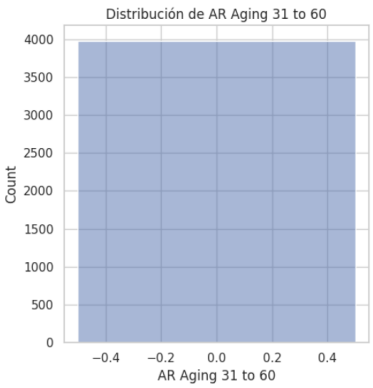

Distribución de Unbilled 90 to 120
Count
Unbilled 90 to 120

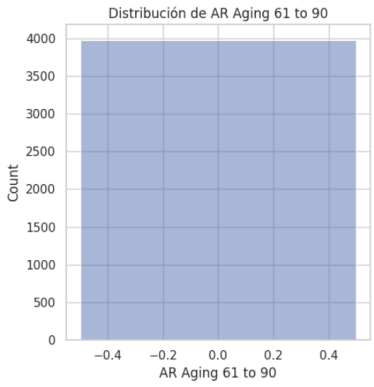

Distribución de Unbilled 120+
Count
Unbilled 120+

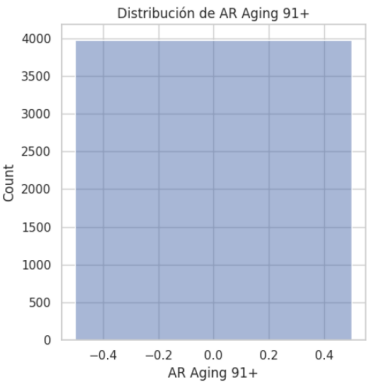

Distribución de Total Billed
Count
Total Billed

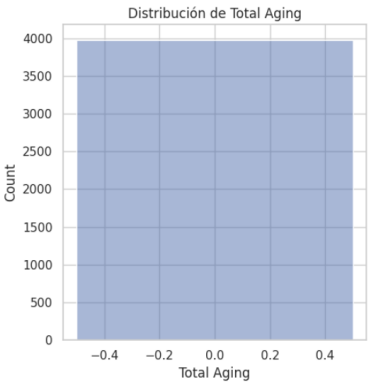

Distribución de AR Aging Current
Count
AR Aging Current

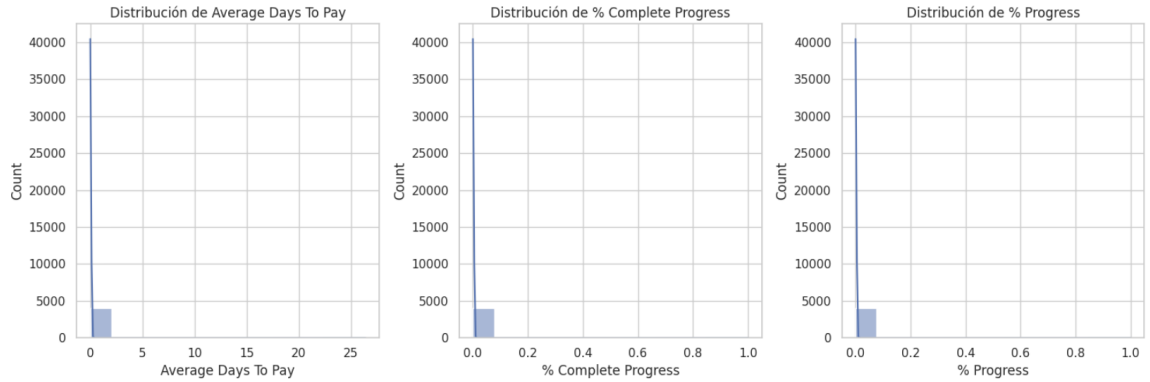

Distribución de AR Aging 1 to 30
Count
AR Aging 1 to 30


Distribución de AR Aging 31 to 60
Count
AR Aging 31 to 60


Distribución de AR Aging 61 to 90
Count
AR Aging 61 to 90


Distribución de AR Aging 91+
Count
AR Aging 91+


Distribución de Total Aging
Count
Total Aging



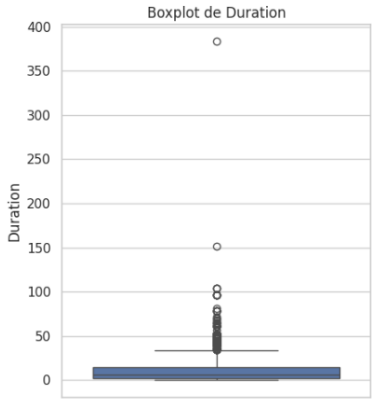
Nota. Elaboración propia.

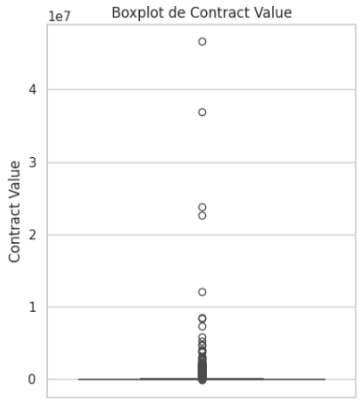
Visualización de diagramas de cajas

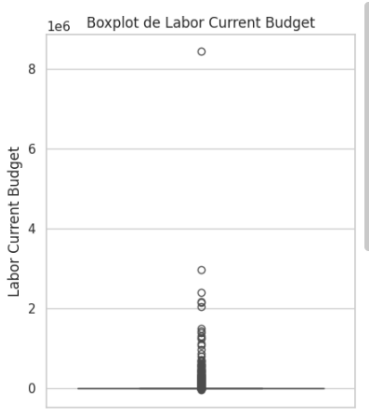
La visualización de diagrama de cajas es muy útil en estos casos para comprender la distribución de los datos numéricos y confirmar la variabilidad de estos, predominando fuertemente los outliers o valores atípicos ocasionados por una varianza muy alta en las variables predictoras así como una concentración de los datos en la parte baja de las gráficas. En cambio, variables como Duration, Target GM y Actual GM describen valores dentro de la caja que representan el 50% central de la distribución concentrada en un rango bajo pero sin outliers claros. Se debe considerar transformar o tratar estos valores anómalos y llevar una investigación más a nivel de ciencia de datos para buscar específicamente los proyectos que ocasionan estas desviaciones con el propósito de asimilar si los datos que se presentan son reales y acordes al contexto del proyecto.

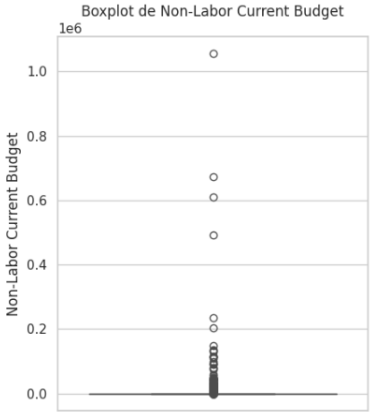
Figura 16

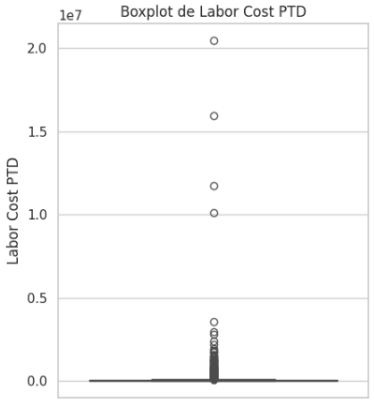
Diagrama de cajas para variables numéricas

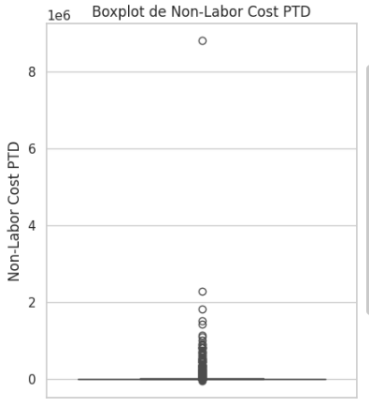

Boxplot de Duration
Duration
400
350
300
250
200
150
100
50
0

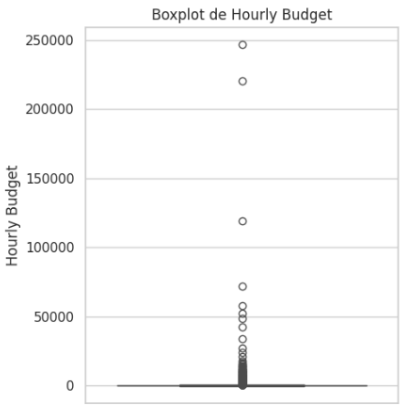

Boxplot de Contract Value
1e7
Contract Value
4
3
2
1
0

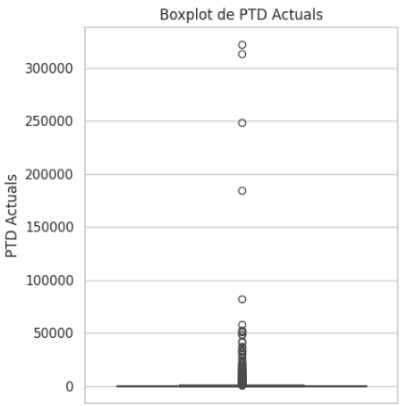

Boxplot de Labor Current Budget
1e6
Labor Current Budget
8
6
4
2
0

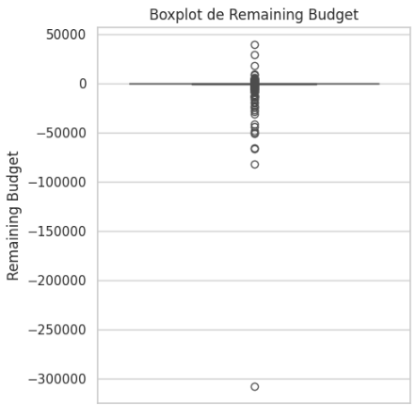

Boxplot de Non-Labor Current Budget
1e6
Non-Labor Current Budget
1.0
0.8
0.6
0.4
0.2
0.0

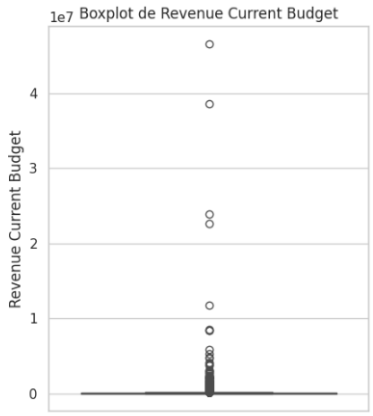

Boxplot de Labor Cost PTD
1e7
Labor Cost PTD
2.0
1.5
1.0
0.5
0.0

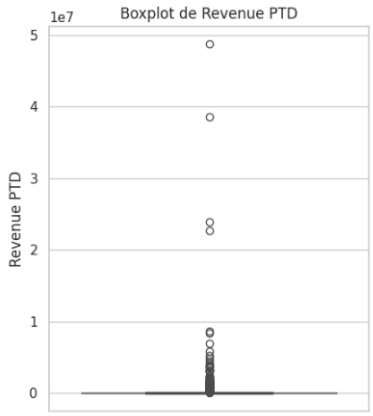

Boxplot de Non-Labor Cost PTD
1e6
Non-Labor Cost PTD
8
6
4
2
0

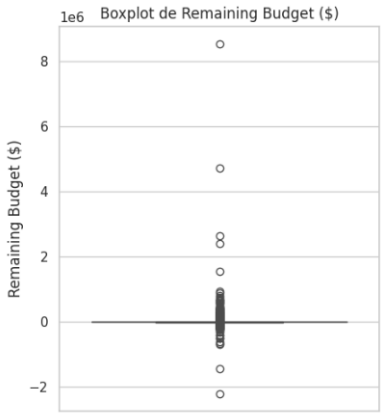

Boxplot de Hourly Budget
Hourly Budget
250000
200000
150000
100000
50000
0

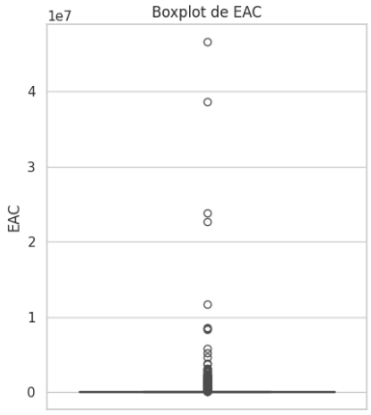

Boxplot de PTD Actuals
PTD Actuals
300000
250000
200000
150000
100000
50000
0

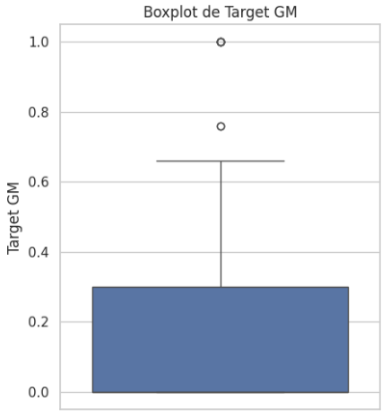

Boxplot de Remaining Budget
Remaining Budget
50000
0
−50000
−100000
−150000
−200000
−250000
−300000

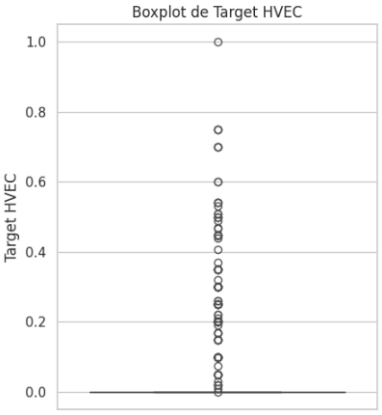

Boxplot de Revenue Current Budget
1e7
Revenue Current Budget
0
1
2
3
4

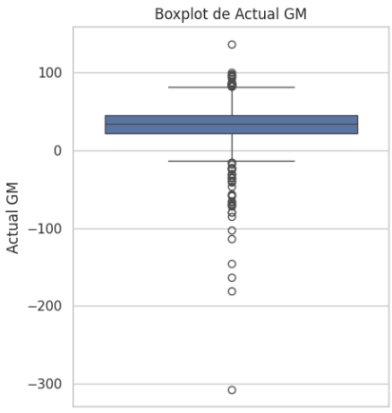

Boxplot de Revenue PTD
1e7
Revenue PTD
0
1
2
3
4
5

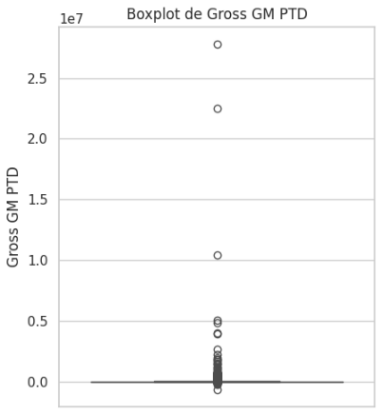

Boxplot de Remaining Budget ($)
1e6
Remaining Budget ($)
−2
0
2
4
6
8

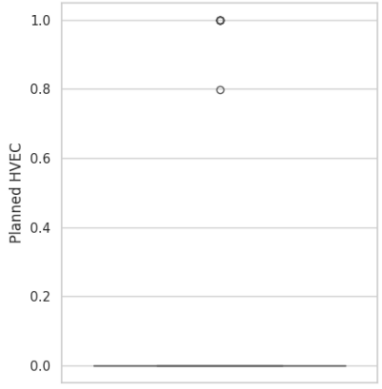

Boxplot de EAC
1e7
EAC
0
1
2
3
4

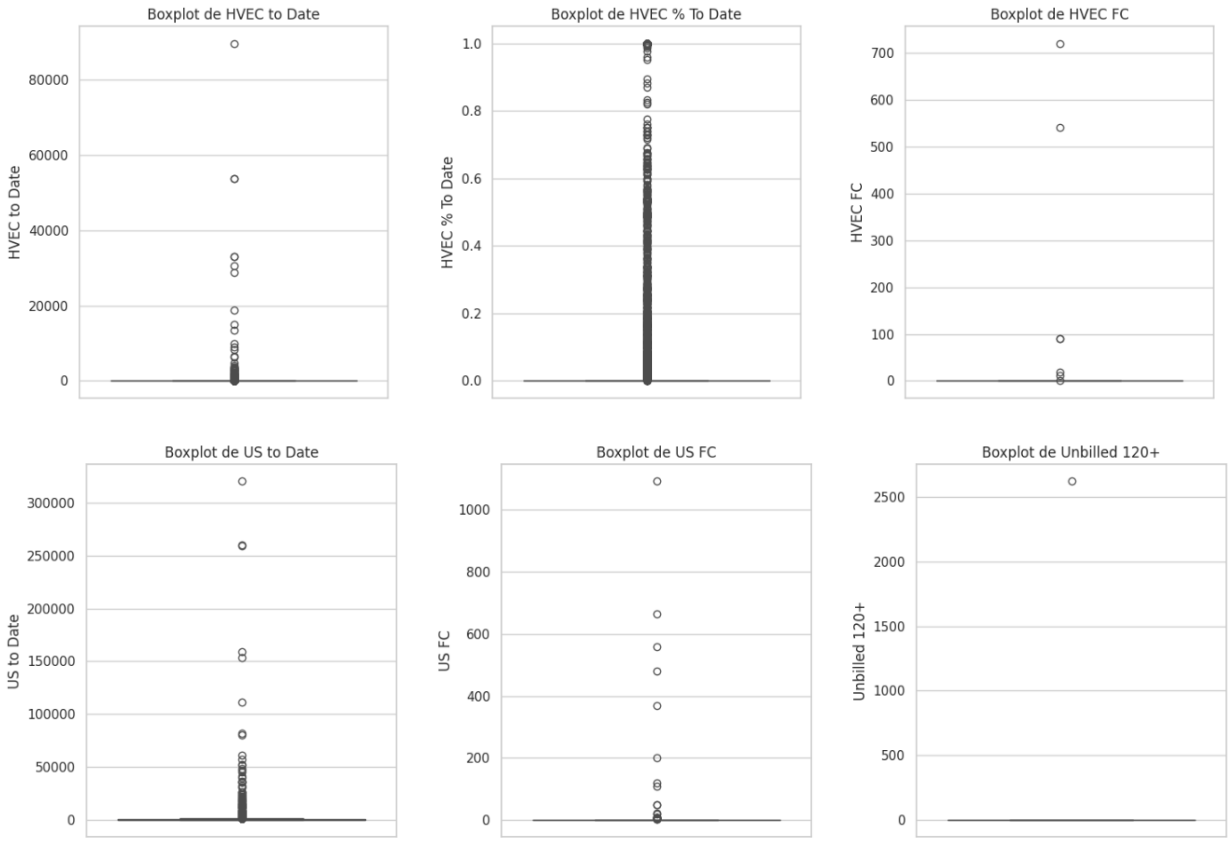

Boxplot de Target GM
Target GM
0.0
0.2
0.4
0.6
0.8
1.0


Boxplot de Target HVEC
Target HVEC
0.0
0.2
0.4
0.6
0.8
1.0


Boxplot de Actual GM
Actual GM
−300
−200
−100
0
100


Boxplot de Gross GM PTD
1e7
Gross GM PTD
0.0
0.5
1.0
1.5
2.0
2.5


Boxplot de Planned HVEC
Planned HVEC
0.0
0.2
0.4
0.6
0.8
1.0



Nota. Elaboración propia.

Visualización de distribuciones de frecuencias en variables categóricas

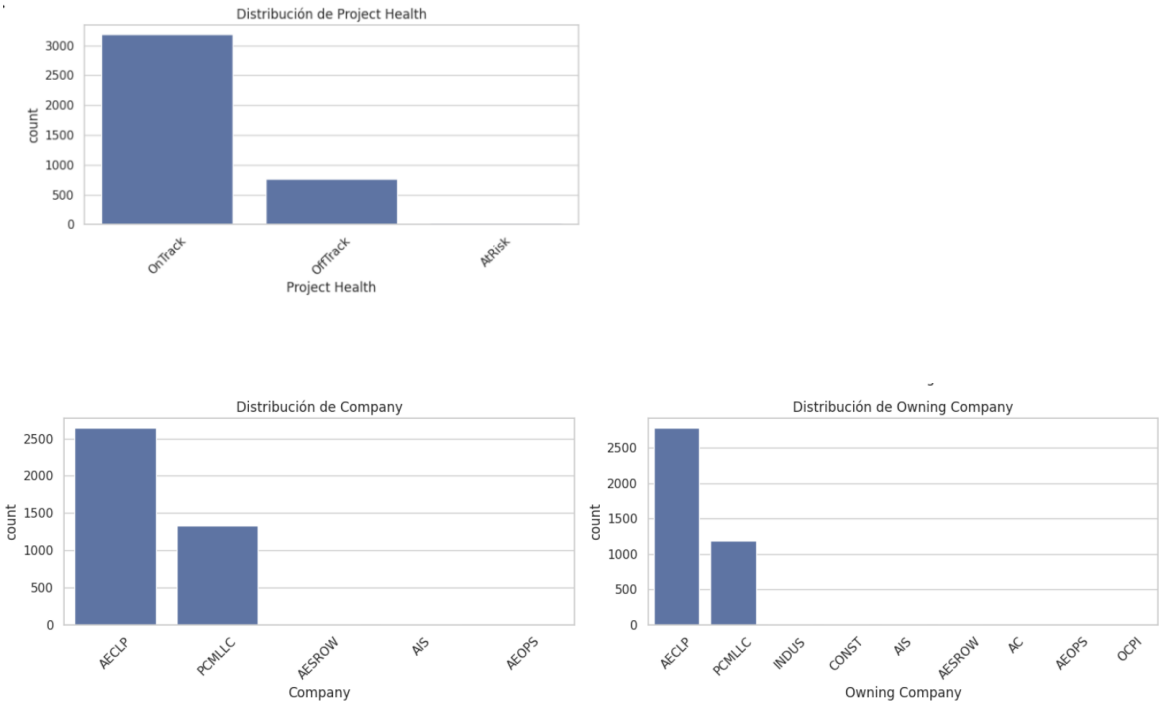
Las variables categóricas representan características que contienen información crucial de los proyectos que son determinantes para llevar a cabo análisis estadístico y conformación de modelos predictivos. Estas variables categóricas permiten al modelo aprender patrones específicos asociados a esas categorías, aumentando su capacidad predictiva. Las variables categóricas más importantes que contiene el Dataset de proyectos cerrados de diseño de ingeniería de la empresa Audubon son ‘Project health’, que indica la salud del proyecto y que para propósitos del modelo será la variable por predecir para nuevos proyectos. La muestra de estudio contiene 2497 proyectos ‘On Track’, 730 proyectos ‘Off Track’ y 39 proyectos ‘At Risk’. Este desbalanceo en las clases será determinante para el resultado del modelo ya que favorece tener una cantidad similar de datos para cada una de las clases a predecir dentro de

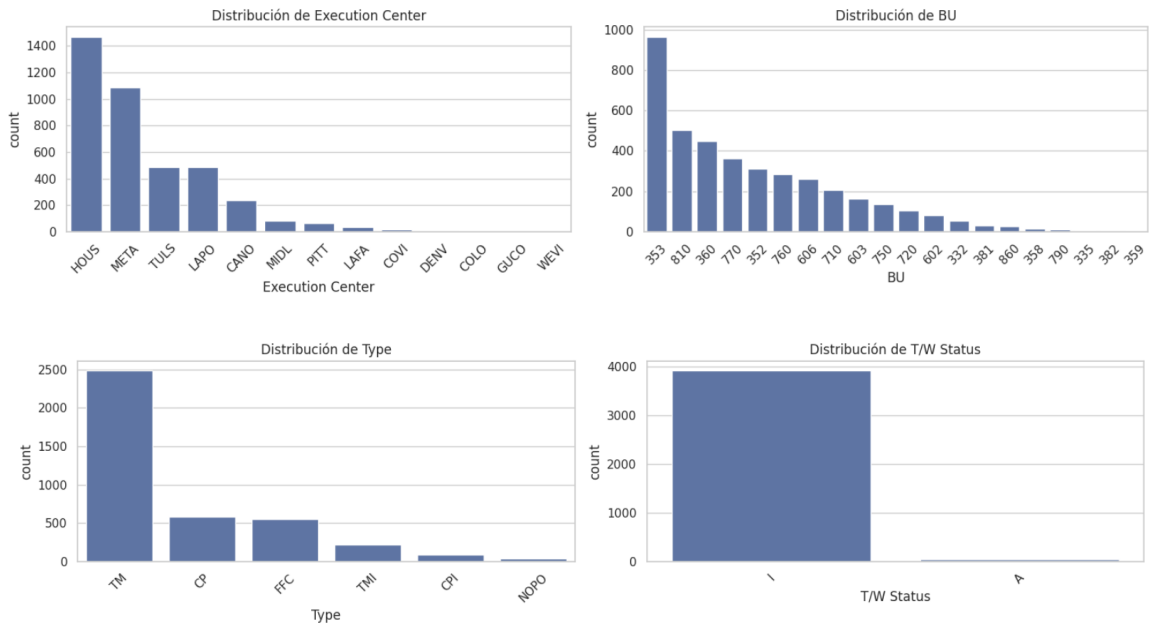
los modelos de clasificación, por lo que los resultados pueden señalar indicios vagos en la predicción debido a la poca cantidad de objetos para estudio.

Categorías como ‘Company’, ‘Owning company’, ‘Execution Center’, ‘BU’ y ‘Type’ describen correctamente la distribución de los distintos proyectos en filiales, oficinas, unidades de negocio y tipo de proyectos que sirven como base para el modelo en la identificación y valoración de proyectos exitosos, en riesgo o problemas de acuerdo a estos parámetros principales de los proyectos. Para este caso se considera que la variable ‘T/W Status’ no aporta mayor significancia al modelo porque esta variable claramente clasifica los proyectos inactivos en su totalidad, confirmando que se trata de proyectos cerrados a la fecha.

Figura 17

Diagrama de barras para variables categóricas





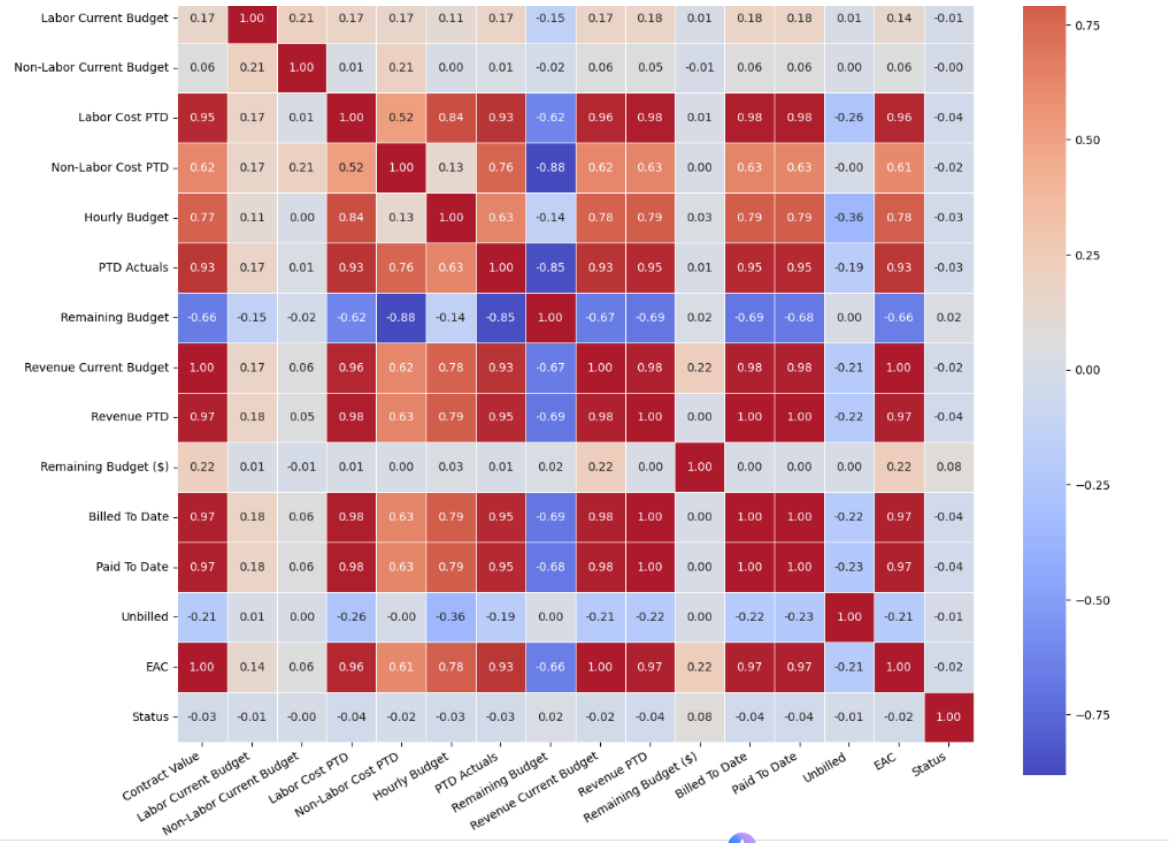
Nota. Elaboración propia.

Matriz de Correlación de variables

La matriz de correlación es una herramienta muy útil para identificar relaciones lineales de correspondencia entre las variables del conjunto de datos del cual se está llevando a cabo el análisis. Todas las variables relacionadas al desglose financiero de los proyectos presentan una alta correlación debido a su naturaleza, Revenue PTD con Billed to Date sugiere una alta correlación ya que el ingreso acumulado a la fecha está directamente relacionado con lo facturado, lo cual se espera que ocurra a medida que avanza un proyecto. En cuanto a las correlaciones negativas Unbilled y Billed to Date (-0.45) indica que cuanto más aumenta la facturación en una fecha, menos queda por facturar, reflejando el flujo de transacciones esperado en un ciclo de facturación. Este análisis visual de colinealidad entre las variables permite detectar que variables están altamente correlacionadas, lo cual puede ser indicar redundancia y así, facilitar la selección de variables para modelos predictivos, de manera que se puedan eliminar aquellas que aportan información adicional. Ver figura 18.

Figura 18

Mapa de calor de correlación de variables



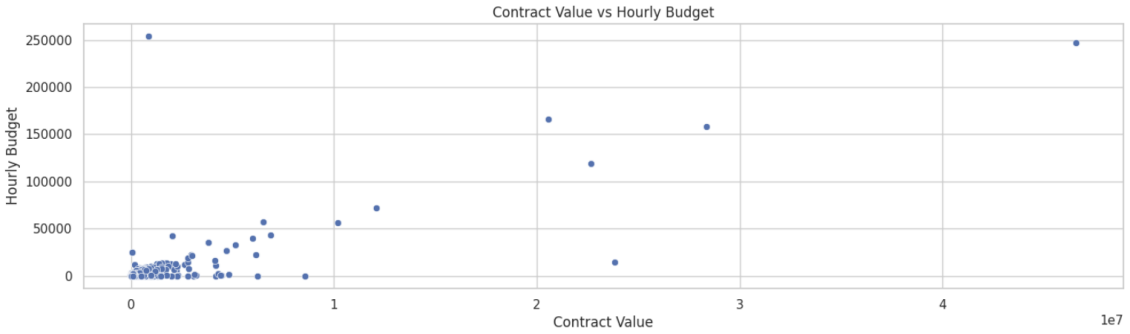
Nota. Elaboración propia.

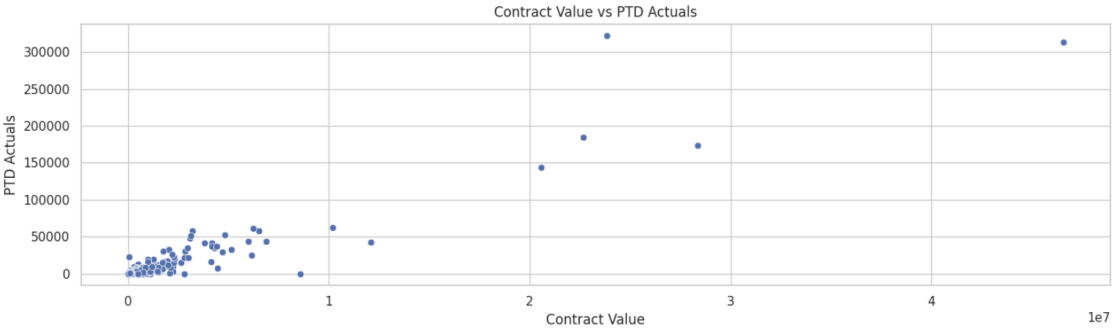
Conforme a este análisis se identifican que las variables cuantitativas que pueden tener mayor incidencia sobre el modelo son el remaining Budget de horas, Labor Cost PTD, Non-Labor cost PTD, Revenue PTD, Billed to Date y Unbilled. Para confirmar esta información se realizan gráficos de dispersión para pares de variables con correlación significativa, de manera que se visualice la alta colinealidad entre estas variables y se considere la eliminación de algunas para evitar el sobre ajuste del modelo. Ver Figura 19.

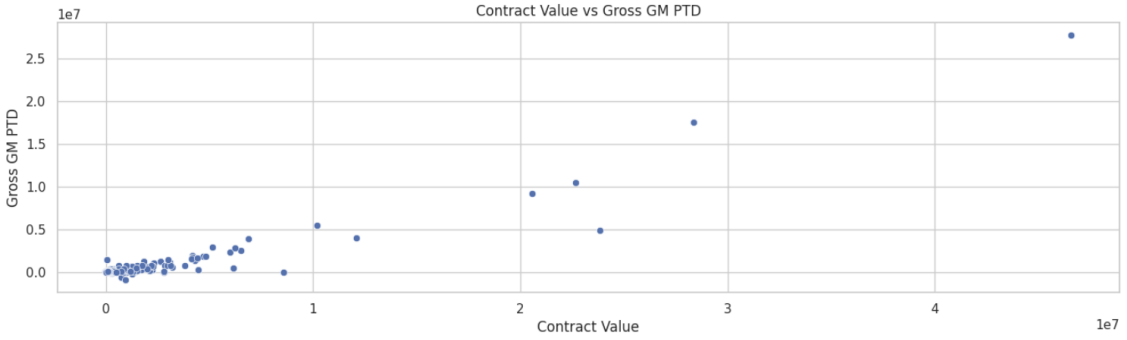
Figura 19

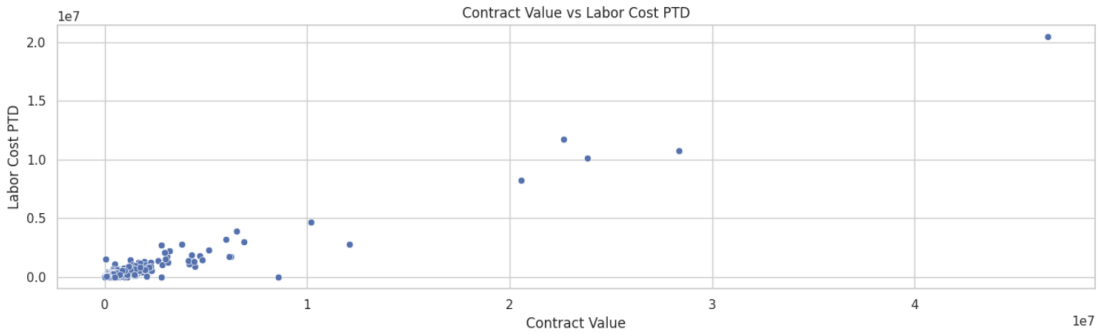
Gráficos de dispersión pares de variables con mayor correlación


Contract Value vs Remaining Budget ($)
1e6
Remaining Budget ($)
8
6
4
2
0
−2
0
1
2
3
4
Contract Value
1e7


Contract Value vs Hourly Budget
Hourly Budget
250000
200000
150000
100000
50000
0
0
1
2
3
4
Contract Value
1e7


Contract Value vs PTD Actuals
PTD Actuals
300000
250000
200000
150000
100000
50000
0
0
1
2
3
4
Contract Value
1e7


Contract Value vs Gross GM PTD
1e7
Gross GM PTD
2.5
2.0
1.5
1.0
0.5
0.0
0
1
2
3
4
Contract Value
1e7



Nota. Elaboración propia.

Modelamiento de los datos y resultados

Selección de variables críticas para el modelo

La selección de variables críticas considero todo el análisis realizado previamente en la exploración de los datos, teniendo en cuenta la dispersión, sesgo, distribución, colinealidad, nulidad y relevancia de cada uno de los valores numéricos y categóricos que pueden incidir en el resultado final de un proyecto de diseño de ingeniería en la empresa Audubon LLC sucursal colombiana, definidos por la variable objetivo 'Project Health' y que está compuesta por tres clases de estado de los proyectos (On Track, Off Track, At Risk). En la figura 29 se observan las variables que no aportan valor al modelo y se excluyen del Dataframe porque no van a ser tenidas en cuenta en el entrenamiento y prueba del modelo.

Figura 20

Eliminación de variables que no aportan valor al modelo

```
# Eliminar columnas que de acuerdo al análisis descriptivo no aportan valor al modelo

columns_not_useful = ['PO', 'Estimated Completion Date', 'Last Charge Date', 'Duration UOM', 'PO', 'Project Id', 'Name', 'Manager', 'Client Name', 'Company',
                      'Planned HVEC', 'Unbilled 0 to 30', 'Unbilled 31 to 45', 'Unbilled 46 to 60', 'Unbilled 61 to 90', 'Unbilled 90 to 120', 'AR Aging Current
                      'AR Aging 1 to 30', 'AR Aging 31 to 60', 'AR Aging 61 to 90', 'AR Aging 91+', 'Total Aging', 'Average Days To Pay', '% Complete Progress',
                      '% Progress', 'Owning Company', 'T/W Status', 'Unbilled', 'HVEC to Date', 'HVEC FC', 'US FC', 'Billed to Date', 'EAC']
```

Nota. Elaboración propia.

Imputación de variables en campos nulos

Se realiza la imputación de variables para aquellas que se consideran de relevancia para el modelo y que no conllevan una cantidad considerable de campos nulos que al ser transformados afecten de manera considerable la naturaleza de cada proyecto en específico.

Para las variables numéricas se realiza la imputación con la mediana de los datos y para las categóricas se completa con la moda de cada una de las variables. Al final se obtiene una muestra de 3716 registros para trabajar con los modelos de ML.

Figura 21

Imputación de variables

```
# Identificar columnas números y categóricas

numeric_features = df_data.select_dtypes(include=['int64', 'float64']).columns
categorical_features = df_data.select_dtypes(include=['object']).columns

# Se extrae la variable objetivo del df Categorical para darle otro tratamiento

categorical_features = categorical_features.drop('Project Health')

print("Características numéricas:", numeric_features)
print("Características categóricas:", categorical_features)
```

```
Características numéricas: Index(['Duration', 'Contract Value', 'Labor Current Budget',
       'Non-Labor Current Budget', 'Labor Cost PTD', 'Non-Labor Cost PTD',
       'Hourly Budget', 'PTD Actuals', 'Remaining Budget', '% Spent',
       'Revenue Current Budget', 'Revenue PTD', 'Remaining Budget ($)',
       '% Spent ($)', 'Billed To Date', 'Paid To Date', 'Target GM',
       'Target HVEC', 'Actual GM', 'Gross GM PTD', 'HVEC % To Date',
       'US to Date', 'Unbilled 120+', 'Total Billed'],
      dtype='object')
Características categóricas: Index(['Execution Center', 'BU', 'Type'], dtype='object')
```

```
# Función para imputar valores
def imputar_valores(df):
    for col in df.columns:
        if col in numeric_features:
            df[col] = pd.to_numeric(df[col], errors='coerce')
            df[col] = df[col].fillna(df[col].median())
        elif df[col].dtype == 'object':
            df[col] = df[col].fillna(df[col].mode()[0])
    return df

# Aplicar la función de imputación
df_data = imputar_valores(df_data)

print('\nDatos después de la imputación:')
print("Número de filas después de la imputación:", len(df_data))
df_data.info()
```

```
Número de filas antes de la imputación: 3716

Datos después de la imputación:
Número de filas después de la imputación: 3716
```

Nota. Elaboración propia.

Normalización de los datos

Conforme se cuenta con el total de los datos para análisis se procede a normalizar los datos. Los algoritmos de ML son muy sensibles a la magnitud y escala de las características, por lo que una falta de normalización puede afectar la velocidad, exactitud y precisión del algoritmo. Así mismo los valores anómalos u outliers pueden distorsionar los datos de entrenamiento y sesgar los modelos, generando overfitting y finalmente afectando la

generalización del modelo. Para este caso y teniendo en cuenta la alta dispersión de los datos se opta por trabajar con el objeto RobustScaler() y FunctionTransformer() de la librería scikit-learn. RobustScaler permite utilizar estadísticas robustas como la mediana y el rango Inter cuartil en lugar de la media y desviación estándar, lo que minimiza la sensibilidad a valores atípicos. Log_transform() se utiliza cuando las distribuciones están sesgadas a la derecha con el fin de normalizarlas y reducir la asimetría.

Para las variables categóricas se utiliza One Hot Encoding para convertir las clases en valores numéricos binarios en nuevas columnas, siendo esencial para el entendimiento del modelo de aprendizaje automático e incrementando la interpretabilidad y precisión. Ver figura 22.

Figura 22

Normalización de los datos

```
# Definir la función de transformación logarítmica
def log_transform(X):
    return np.log1p(X - X.min() + 1)

# Crear el pipeline de preprocesamiento para características numéricas
numeric_transformer = Pipeline(steps=[
    ('robust', RobustScaler()),
    ('log', FunctionTransformer(log_transform, validate=False))
])

# Aplicar la transformación a las columnas numéricas
df_data[numeric_features] = numeric_transformer.fit_transform(df_data[numeric_features])
```

```
# Aplicar One Hot Encoding a las características categóricas
encoder = OneHotEncoder(handle_unknown='ignore')
encoded_features = encoder.fit_transform(df_data[categorical_features])

# Crear un DataFrame con las características codificadas
encoded_df = pd.DataFrame(
    encoded_features.toarray(),
    columns=encoder.get_feature_names_out(categorical_features),
    index=df_data.index
)

# Concatenar las características codificadas con las numéricas
df_encoded = pd.concat([df_data[numeric_features], encoded_df, df_data['Project Health']], axis=1)


# Codificar variable objetivo con label encoder (No aumenta el número de columnas)
le = LabelEncoder() #Encoder ordena alfabeticamente las clases de la variable objetivo
```

Nota. Elaboración propia.

Selección de variables predictoras y objetivos

Como se observa en la figura 23 se separan las variables predictoras y objetivo en dos variables diferentes para llevar a cabo la división de los datos de entrenamiento y prueba en un 70% y 30% respectivamente.

Figura 23

Definición de variables predictoras y objetivo

```
X = df_encoded.drop('Project Health', axis=1)
y = le.fit_transform(df_encoded['Project Health'])
```

Nota. Fuente. Elaboración propia.

Inicialización de modelos

De acuerdo con Mali et al. (2025) los modelos de aprendizaje supervisado que utilizan un tipo de algoritmo de clasificación demostraron ser más capaces para predecir resultados a partir de datos de entrada que incluyen sobrecostos y estimación de presupuestos en proyectos. Se opta por trabajar con los algoritmos de regresión logística, arboles de decisión, bosques aleatorios y XGBoost ya que predominan como los algoritmos de clasificación más utilizados y resultaron ser efectivos en la clasificación de proyectos complejos.

Figura 24

Modelos de clasificación

```
# División del conjunto de datos
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42, stratify=y)

# Inicialización de modelos
models = {
    'Logistic Regression': LogisticRegression(max_iter=1000),
    'Decision Tree': DecisionTreeClassifier(random_state=42),
    'Random Forest': RandomForestClassifier(random_state=42),
    'XGBoost': XGBClassifier(random_state=42)
}
```

Nota. Elaboración propia.

Evaluación de modelos

El análisis comparativo de las métricas resultantes de los modelos de clasificación en la Figura 25, evidencia diferencias significativas en el rendimiento de los modelos evaluados. La Regresión Logística obtuvo un recall de 0.7901 y un AUC-ROC de 0.6959, lo que indica una

capacidad moderada de detección y discriminación de clases. Este comportamiento sugiere que el modelo presenta limitaciones para capturar relaciones no lineales y dependencias complejas entre las variables predictoras, aspecto esperado dada su naturaleza lineal.

El Decision Tree mostró un desempeño notablemente superior (recall de 0.9327 y AUC-ROC de 0.9014), reflejando una mejor capacidad para representar interacciones no lineales. Se comprobó que el modelo tiende al sobreajuste (overfitting) cuando se aplica a conjuntos de datos que contienen alta correlación entre sí, lo que puede afectar su capacidad de generalización.

Los modelos Random Forest y XGBoost presentan mejores resultados con métricas más cercanas a 1. En particular, Random Forest alcanzó un recall de 0.9659, F1-score de 0.9602 y AUC-ROC de 0.9579, demostrando alta precisión, estabilidad y robustez frente a ruido y variabilidad en los datos. Por su parte, XGBoost obtuvo el mejor desempeño global (AUC-ROC = 0.9795), teniendo en cuenta que es un modelo muy sensible a los parámetros inicialmente establecidos y su precisión aumenta si se ajustan debidamente las dimensiones y parámetros del modelo.

En conjunto, los resultados confirman que los modelos Random Forest y XGBoost superan ampliamente a los modelos de regresión logística y arboles de decisión, validando su posible uso para escenarios de planeación de proyectos de ingeniería.

Figura 25

Resultados de métricas de los modelos de clasificación

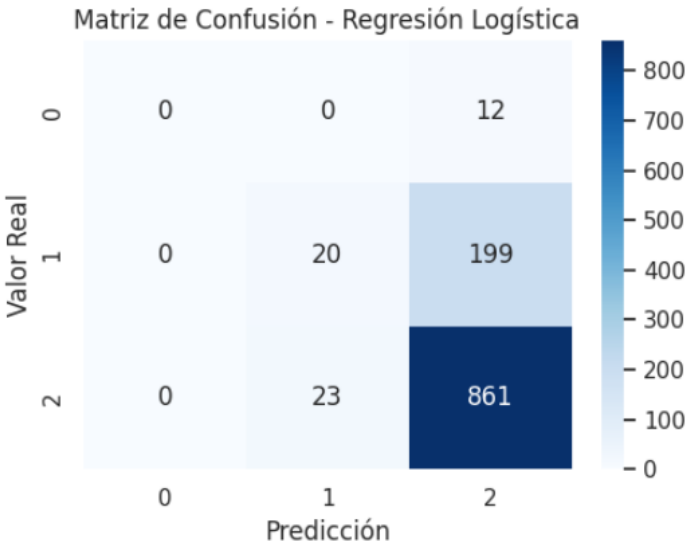
```
Training Logistic Regression...
Training Decision Tree...
Training Random Forest...
Training XGBoost...
Logistic Regression:
Recall: 0.7901
F1-Score: 0.7280
AUC-ROC: 0.6959442041309574

Decision Tree:
Recall: 0.9327
F1-Score: 0.9325
AUC-ROC: 0.9013617947951867

Random Forest:
Recall: 0.9659
F1-Score: 0.9602
AUC-ROC: 0.9579341044734739

XGBoost:
Recall: 0.9623
F1-Score: 0.9608
AUC-ROC: 0.9795185848812398
```

Nota. Elaboración propia.

Conforme a la evaluación de las métricas para estos modelos de clasificación, se procede a realizar un análisis de las matrices de confusión para cada modelo con el fin de observar y precisar la capacidad del modelo para predecir aciertos y errores de clasificación. En términos generales, las matrices de confusión expuestas en la Figura 26, 27, 28 y 29 permiten corroborar las tendencias observadas en las métricas globales, los modelos Random Forest y XGBoost logran una clasificación más precisa y equilibrada entre las tres categorías, minimizando falsos negativos y maximizando verdaderos positivos. En contraste, los modelos más simples Regresión Logística y Árbol de Decisión presentan mayores niveles de error en las clases intermedias y minoritarias, afectando la sensibilidad del modelo ante escenarios de mayor incertidumbre.

Figura 26

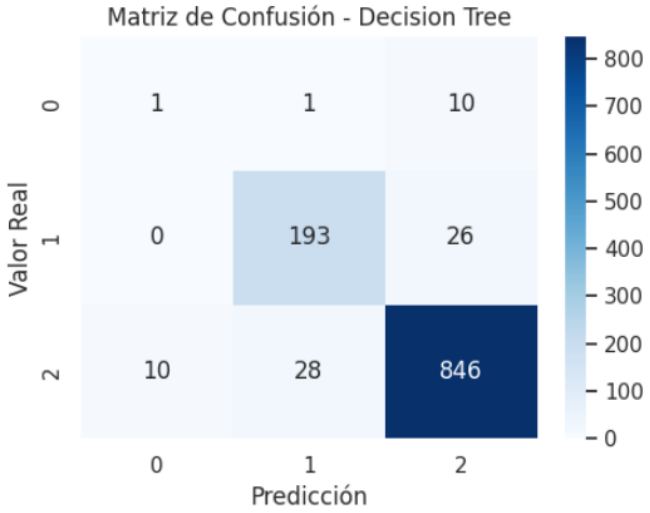
Matriz de Confusión para regresión logística



Nota. Elaboración propia.

Figura 27

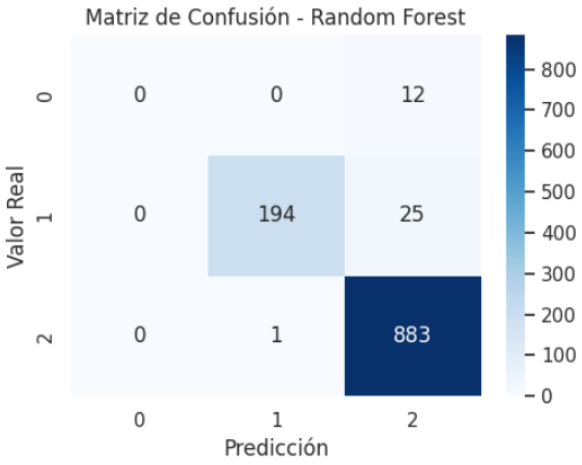
Matriz de Confusión para árbol de decisión



Nota. Elaboración propia.

Figura 28

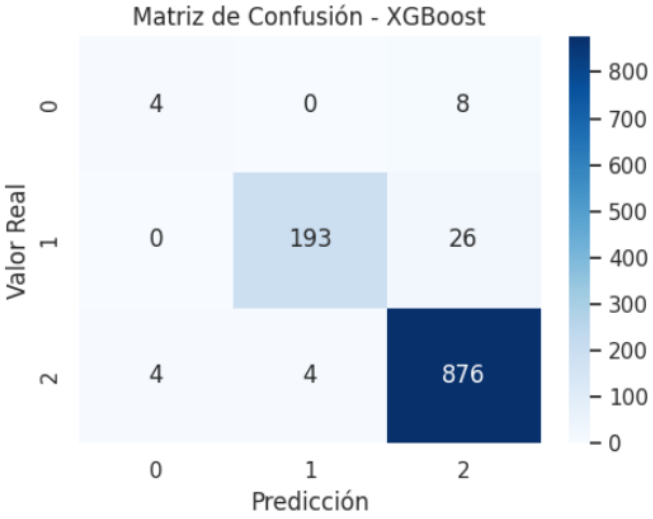
Matriz de Confusión para bosques aleatorios



Nota. Elaboración propia.

Figura 29

Matriz de Confusión para XGBoost



Nota. Elaboración propia.

Estos resultados confirman la hipótesis planteada, los modelos de ensamblado son más adecuados para problemas de predicción multiclase en los proyectos de la empresa, donde las relaciones entre variables son no lineales y los datos presentan heterogeneidad estructural.

Estrategia de mejora de la planeación basada en Project Management Ágil y analítica de datos

Introducción

Los resultados del modelo predictivo desarrollado en la investigación demostraron que los datos históricos disponibles en la empresa Audubon, Sucursal Colombiana, carecen de relevancia analítica suficiente para explicar o anticipar los resultados de éxito de los proyectos de ingeniería. El análisis identificó una dependencia excesiva de variables financieras y operativas, mientras que factores cualitativos y contextuales —como la clase de proyecto, la importancia estratégica del cliente, el desempeño de los vendors, la experiencia del Project Manager o el tipo de relación contractual— no estaban adecuadamente representados en el conjunto de datos.

Esta carencia limita el poder predictivo de los modelos de machine learning y, más aún, revela una debilidad estructural en la planeación organizacional. Tal como señalan Kerzner (2022) y el Project Management Institute (2021), los proyectos que carecen de una visión integral de sus variables no solo presentan mayor incertidumbre, sino que también desarrollan sistemas de planeación reactivas, incapaces de aprender del entorno y mejorar de manera iterativa.

En este contexto, se propone una estrategia integral de mejora de la planeación, fundamentada en los principios del Project Management moderno y el marco ágil adaptativo, que permita a Audubon evolucionar hacia un modelo de planeación predictiva iterativa, sostenida en datos cualitativos y ciclos de aprendizaje continuo.

Fundamentación teórica

La literatura reciente en dirección de proyectos ha transitado desde enfoques predictivos tradicionales hacia modelos híbridos de gestión, donde la agilidad organizacional y la capacidad analítica son factores determinantes del desempeño (Biesenthal & Wilden, 2022; Serrador & Pinto, 2019).

En entornos de ingeniería, donde la variabilidad de requerimientos, la dependencia técnica y la interacción cliente–proveedor son constantes, la planeación basada exclusivamente en indicadores financieros es insuficiente (Madiwale & Mahadik, 2023). Por el contrario, los marcos ágiles y los sistemas de data-driven project management promueven la incorporación de métricas contextuales, cualitativas y dinámicas que alimentan un proceso de planeación vivo y en evolución.

Según el Project Management Institute (2021), los dominios de desempeño del proyecto —Stakeholders, Planning, Delivery y Measurement— deben concebirse como sistemas adaptativos donde los datos fluyen de forma continua para fortalecer la toma de decisiones. Asimismo, Dzhusupova et al. (2024) sostienen que la integración entre machine learning y marcos ágiles potencia la capacidad de predicción organizacional, al combinar la exploración estadística con la validación empírica en ciclos iterativos.

Desde esta perspectiva, el reto no radica únicamente en “tener más datos”, sino en gestionar mejor los datos relevantes: contextualizados, categóricos y trazables. Por tanto, el propósito de la estrategia propuesta es redefinir la planeación como un sistema de aprendizaje y retroalimentación, apoyado por herramientas analíticas y metodologías ágiles.

Diagnóstico del problema estructural

El sistema actual de planeación de Audubon presenta tres deficiencias clave: Planeación lineal y poco adaptable: los planes de ejecución se elaboran de manera estática al inicio del proyecto, sin mecanismos ágiles de retroalimentación que permitan revisar supuestos o anticipar riesgos emergentes.

Datos históricos desbalanceados: las variables registradas se centran en costos, horas de ingeniería y rentabilidad, sin considerar atributos categóricos de contexto organizacional o relacional que influyen de forma significativa en la variabilidad del desempeño.

Gestión del conocimiento fragmentada: la información generada por los equipos de proyecto no se sistematiza ni retroalimenta los modelos analíticos. El conocimiento tácito se pierde entre fases, dificultando la madurez de la planeación predictiva.

Estas limitaciones no solo obstaculizan el uso efectivo de modelos de inteligencia artificial, sino que también reducen la capacidad de la organización para aprender de su propia experiencia (Nonaka & Takeuchi, 1995).

Estrategia propuesta: Marco de Planeación Predictiva Ágil (MPPA)

El diagnóstico realizado condujo a la propuesta del Marco de Planeación Predictiva Ágil (MPPA). Se trata de un modelo híbrido diseñado para integrar estratégicamente las metodologías base del Project Management Body of Knowledge con la agilidad de Scrum y las capacidades de datos de DataOps. Este marco busca instaurar una cultura de planeación continua, iterativa y basada en evidencia, donde la información de cada ciclo de proyecto retroalimente al siguiente.

Fase I. Rediseño estructural de la planeación

Implementar una Work Breakdown Structure orientada a datos (Data-Driven WBS), en la cual cada paquete de trabajo incorpore variables categóricas como tipo de cliente, relevancia estratégica, criticidad técnica y madurez del Project Manager.

Desarrollar una Matriz de Criticidad del Proyecto (PCM) que clasifique los proyectos según su nivel de complejidad, incertidumbre contractual y valor estratégico, generando una tipología que pueda alimentar los modelos predictivos.

Establecer una línea base ágil (Agile Baseline) revisable en intervalos de 4–6 semanas, que permita ajustar estimaciones y riesgos de manera incremental.

Fase II. Gobernanza de datos ágil

Para satisfacer la demanda de datos confiables, completos y consistentes que requiere la planeación predictiva, se propone el desarrollo de una Estructura de Gobernanza Ágil de

Datos (Agile Data Governance Framework, ADGF). Dicha estructura se compone de tres roles clave:

Data Product Owner: responsable de validar la relevancia y el valor de negocio de cada variable incorporada al modelo.

Data Steward: encargado de la calidad, limpieza y trazabilidad de los datos provenientes de distintas fuentes (ERP, CRM, Power BI, informes técnicos).

Scrum Master de datos: facilita los sprints de revisión y prioriza los data backlogs para mejorar el rendimiento analítico del sistema.

Gracias al enfoque DataOps, se automatizarán los procesos de ingesta, validación y publicación de datos. Esto permitirá reducir los ciclos de actualización y optimizar la sincronía entre la operación del proyecto y su respectiva analítica, una práctica respaldada por la literatura reciente (Erickson et al., 2021).

Fase III. Integración ágil-predictiva

El MPPA adopta un modelo híbrido en el que:

La planeación predictiva tradicional se utiliza para los componentes estructurales, como infraestructura, ingeniería y adquisiciones.

La planeación ágil iterativa se aplica a variables de alto dinamismo (relación cliente, desempeño de equipos, gestión de cambios, riesgo técnico).

Cada Sprint Planning incluirá una revisión de métricas de desempeño predictivo (Predictive Sprint Review), donde se evalúe la precisión de las estimaciones previas frente a los resultados reales y se ajusten las variables del modelo.

Este mecanismo convierte la planeación en un proceso de aprendizaje adaptativo, donde las desviaciones no se castigan, sino que se documentan como lecciones analíticas.

Fase IV. Sistema de aprendizaje organizacional

Se establecerá un Closed-Loop Learning System (CLLS), en el cual las lecciones aprendidas y métricas históricas se codifiquen en una base de conocimiento institucional. Este

sistema permitirá que cada proyecto concluido alimente el conjunto de datos del siguiente, consolidando una inteligencia organizacional evolutiva (Biesenthal & Wilden, 2022).

Mecanismos de evaluación

El impacto del MPPA será medido mediante indicadores de desempeño (KPIs) alineados con el estándar PMI (Project Management Institute, 2021) y con métricas de madurez analítica.

Indicador, Descripción, Fórmula

Forecast Accuracy Agile (FAA), Precisión del modelo predictivo tras cada iteración, (Predicción ajustada – Resultado real) / Resultado real.

Plan Predictivo Mejorado (PPM%), Porcentaje de desviaciones anticipadas gracias al modelo, (Desviaciones anticipadas / Total desviaciones) × 100.

Project Replanning Efficiency (PRE), Eficiencia del proceso de replanificación ágil, Tiempo medio de ajuste / Duración total del proyecto.

Analytical Maturity Index (AMI), Nivel de integración de datos cualitativos y cuantitativos en la planeación, Escala 1–5 (según ISO 8000 y DataOps maturity models).

Estos indicadores serán evaluados de forma iterativa, utilizando tableros de control en Power BI integrados con los resultados de los sprints de planeación.

Discusión

Es importante señalar que el MPPA no pretende invalidar la planeación tradicional, sino enriquecerla con un componente adaptativo y de aprendizaje. Solo a través de la evolución cultural de Audubon hacia una gestión centrada en datos será viable develar y aprovechar aquellos patrones que permanecen invisibles bajo los enfoques tradicionales.

Diversos estudios coinciden en que los enfoques híbridos aumentan la resiliencia organizacional y la precisión de la planificación, especialmente en industrias de alta complejidad técnica, (S. O. Abioye et al., 2021; Madiwale & Mahadik, 2023). Además, la incorporación de feedback loops analíticos transforma la planeación en un proceso de

inteligencia colectiva, donde el conocimiento de los equipos se convierte en un activo estratégico (Nonaka & Takeuchi, 1995).

Conclusiones

Los resultados que se han obtenido con la aplicación comparativa de los modelos de aprendizaje automático permiten concluir que el enfoque metodológico adoptado resultó idóneo para cumplir con el objetivo general, dirigido a establecer un marco predictivo para mejorar la eficiencia en la planificación de proyectos de ingeniería en la empresa Audubon. La secuencia metodológica que comprendió la consolidación y depuración de datos, la selección de variables críticas y la implementación de modelos supervisados permitió no solo validar la pertinencia de las técnicas utilizadas, sino también identificar las limitaciones y fortalezas de cada algoritmo en términos de desempeño, interpretabilidad y aplicabilidad empresarial.

Desde el punto de vista técnico, la Regresión Logística proporcionó una línea base de comparación representativa para establecer la capacidad predictiva del modelo conforme al análisis exploratorio inicial de los datos que permitió la depuración y selección de variables críticas conforme a la base de datos histórica. Sin embargo, su desempeño inferior frente a los modelos de ensamblado evidenció la necesidad de incorporar enfoques más flexibles y no paramétricos. Nuestros resultados confirman que el modelo XGBoost no solo supera a otros métodos en rendimiento, sino que también exhiben una mayor capacidad de generalización. Esta superioridad concuerda con las tendencias destacadas en la literatura reciente (almahameed & Bisharah, 2024; Datta et al., 2024a; López Ferreiro et al., 2025b) , donde las técnicas de ensemble learning se consolidan como herramientas idóneas para la predicción de sobrecostos, retrasos y viabilidad financiera en proyectos complejos.

La ruta metodológica propuesta, que combina la limpieza de datos, la estratificación, el análisis exploratorio y el modelado supervisado en Python, resulta replicable y escalable. Este marco puede ser adoptado por otras organizaciones del sector interesadas en integrar la analítica predictiva en su toma de decisiones. Además, la solidez del diseño metodológico queda

demostrada por el uso sistemático de técnicas de aprendizaje automático y métricas robustas (como F1-score, recall y AUC-ROC), lo que garantiza tanto la precisión de los modelos como su relevancia práctica en la gestión de proyectos reales.

El desarrollo del modelo de aprendizaje automático también permitió evidenciar que los datos históricos de Audubon, Sucursal Colombiana, presentan un enfoque tradicional hacia variables financieras y operativas, limitando su capacidad predictiva. No obstante, el proceso de modelado demostró que es posible identificar patrones de comportamiento relevantes siempre que se comprenda la naturaleza de los proyectos y las variables que inciden en él y de las que actualmente no se lleva ningún tipo de registro o control en la compañía: valoración de cliente, criticidad técnica, medición del desempeño del Project Manager en proyectos cerrados, etc. Como Kerzner (2022) y el Project Management Institute (PMI) (2021)han señalado, el éxito de un proyecto no es solo una cuestión de controlar costos y cronogramas. Requiere la integración de dimensiones contextuales que capten la complejidad de la realidad organizacional. Así, el modelo propuesto constituye un punto de partida para avanzar hacia una gestión basada en inteligencia analítica y aprendizaje continuo, donde los proyectos se convierten en fuentes de conocimiento predictivo y estratégico.

Finalmente, el bajo rendimiento predictivo del modelo actual no deriva de una debilidad algorítmica, sino de una insuficiente consolidación de datos relevantes para los proyectos, así como una inmadurez sistémica en la gestión del conocimiento y la planeación organizacional.

La implementación del Marco de Planeación Predictiva Ágil (MPPA) permitirá a Audubon integrar variables cualitativas, promover la gobernanza ágil de datos y consolidar un proceso de planeación adaptativo, cíclico y orientado a la mejora continua.

Al fortalecer el vínculo entre Project Management, Agile y analítica de datos, la organización no solo incrementará la precisión de sus modelos de predicción, sino que también elevará su madurez analítica y competitividad estratégica frente a los retos del sector Oil & Gas.

Referencias

Abioye, F. O. O. A. C. (2021). Integration of artificial intelligence and agile methodologies in construction project management. *Journal of Construction Project Management and Innovation*, *11*(1), 19–33.

Abioye, S. O., Oyedele, L. O., Akanbi, L., Ajayi, A., Davila Delgado, J. M., Bilal, M., Akinade, O. O., & Ahmed, A. (2021). Artificial intelligence in the construction industry: A review of present status, opportunities and future challenges. In *Journal of Building Engineering* (Vol. 44). Elsevier Ltd. https://doi.org/10.1016/j.jobe.2021.103299

Al mnaseer, R., Al-Smadi, S., & Al-Bdour, H. (2023a). Machine learning-aided time and cost overrun prediction in construction projects: application of artificial neural network. *Asian Journal of Civil Engineering*, *24*(7), 2583–2593. https://doi.org/10.1007/s42107-023-00665-7

Al mnaseer, R., Al-Smadi, S., & Al-Bdour, H. (2023b). Machine learning-aided time and cost overrun prediction in construction projects: application of artificial neural network. *Asian Journal of Civil Engineering*, *24*(7), 2583–2593. https://doi.org/10.1007/s42107-023-00665-7

Al mnaseer, R., Al-Smadi, S., & Al-Bdour, H. (2023c). Machine learning-aided time and cost overrun prediction in construction projects: application of artificial neural network. *Asian Journal of Civil Engineering*, *24*(7), 2583–2593. https://doi.org/10.1007/s42107-023-00665-7

almahameed, B. aldeen, & Bisharah, M. (2024). Applying Machine Learning and Particle Swarm Optimization for predictive modeling and cost optimization in construction project management. *Asian Journal of Civil Engineering*, *25*(2), 1281–1294. https://doi.org/10.1007/s42107-023-00843-7

Arabiat, A., Al-Bdour, H., & Bisharah, M. (2023). Predicting the construction projects time and cost overruns using K-nearest neighbor and artificial neural network: a case study from

Jordan. *Asian Journal of Civil Engineering*, *24*(7), 2405–2414. https://doi.org/10.1007/s42107-023-00649-7

Audubon. (n.d.). *Audubon Companies*. Retrieved August 30, 2025, from https://www.linkedin.com/company/audubon-companies?trk=nav_type_overview

Audubon. (2022). *Audubon Engineering Company Celebrates 25th Anniversary*. Article. https://audubooncompanies.com/news/audubon-engineering-company-celebrates-25th-anniversary/#:~:text=Milestone%20demonstrates%20a%20legacy%20of,client%20list%2C%20and%20talent%20pool.

Audubon. (2025a). *Audubon Climbs to #64 on 2025 ENR Top 500 Design Firms, #6 for Industrial Process/Oil & Gas Sector*. Article. https://audubooncompanies.com/news/audubon-climbs-to-64-on-2025-enr-top-500-design-firms/

Audubon. (2025b). *Audubon Companies*. Linkedin. https://www.linkedin.com/company/audubon-companies?trk=nav_type_overview

Audubon. (2025c). *ENR Ranks Audubon Companies a Top 150 Global Design Firm*. Audubon Publications. https://audubooncompanies.com/news/enr-ranks-audubon-companies-a-top-150-global-design-firm/

Bauskar, S. R., Madhavaram, C. R., Galla, E. P., Sunkara, J. R., Gollangi, H. K., & Rajaram, S. K. (2024). Predictive Analytics for Project Risk Management Using Machine Learning. *Journal of Data Analysis and Information Processing*, *12*(04), 566–580. https://doi.org/10.4236/jdaip.2024.124030

Biesenthal, C., & Wilden, R. (2022). Project learning and knowledge integration in dynamic environments. *International Journal of Project Management*, *40*(3), 267–281.

Bodero Poveda, E., De Giusti, M., & Morales Alarcón, C. (2021). La preservación digital a largo plazo y las bases de la planificación estratégica. *3C TIC: Cuadernos de Desarrollo Aplicados a Las TIC*, *10*(3), 17–39. https://doi.org/10.17993/3ctic.2021.103.17-39

Bohórquez Castellanos, J. J., & Mejia-Aguilar, G. (2019). *Relationship between cost overruns and complexity in engineering projects: a mixed approach*. https://doi.org/10.46421/sibragec.v11i00.59

Bruzzone, A. G., Chervisari, M. L., Faccio, F., Massei, M., & Cardelli, M. (2021). Models to apply Strategic Engineering at Digitalization Initiatives in Large Engineering Companies. *20th International Conference on Modeling and Applied Simulation, MAS 2021*, 194–198. https://doi.org/10.46354/i3m.2021.mas.025

Caballero, R., Martín, R. E., Adrián, M., & Rodríguez, R. (2023). *Análisis y minería de textos con PYTHON*.

Cicmil, S., Williams, T., Thomas, J., & Hodgson, D. (2006). Rethinking Project Management: Researching the actuality of projects. *International Journal of Project Management*, *24*(8), 675–686. https://doi.org/10.1016/j.ijproman.2006.08.006

Coffie, G. H., & Cudjoe, S. K. F. (2024). Using extreme gradient boosting (XGBoost) machine learning to predict construction cost overruns. *International Journal of Construction Management*, *24*(16). https://doi.org/10.1080/15623599.2023.2289754

CONPES 4144: Política nacional de inteligencia artificial, 0 (2025).

Daraz, U., Wu, J., Alomair, M. A., & Aldoghan, L. A. (2024). New classes of difference cum-ratio-type exponential estimators for a finite population variance in stratified random sampling. *Heliyon*, *10*(13), e33402. https://doi.org/10.1016/J.HELIYON.2024.E33402

Datta, S. D., Islam, M., Rahman Sobuz, Md. H., Ahmed, S., & Kar, M. (2024a). Artificial intelligence and machine learning applications in the project lifecycle of the construction industry: A comprehensive review. *Heliyon*, *10*(5). https://doi.org/10.1016/j.heliyon.2024.e26888

Datta, S. D., Islam, M., Rahman Sobuz, Md. H., Ahmed, S., & Kar, M. (2024b). Artificial intelligence and machine learning applications in the project lifecycle of the construction

industry: A comprehensive review. *Heliyon*, *10*(5). https://doi.org/10.1016/j.heliyon.2024.e26888

Dawood, F. S., & Ahmed, A. F. (2023). The applicability of the international standard (iso 21500:2021) managing projects, programs and portfolios at the saladin investment commission (case study). *International Journal of Professional Business Review*, *8*(4). https://doi.org/10.26668/businessreview/2023.v8i4.1293

Directorio de Sostenibilidad. (2025, March 1). *¿Cómo pueden los datos predictivos mejorar los resultados de los proyectos de sostenibilidad?*

Dzhusupova, R., Bosch, J., & Olsson, H. H. (2024). Choosing the right path for AI integration in engineering companies: A strategic guide. *Journal of Systems and Software*, *210*. https://doi.org/10.1016/j.jss.2023.111945

Ebers, M. (2024). *Stanford-Vienna Truly Risk-Based Regulation of Artificial Intelligence: How to Implement the EU's AI Act*. http://ttlf.stanford.edu

Ekbote, N., Dhanshetti, P., & Sakhrekar, S. (2023). Techniques of Exploratory Data Analysis. *Madhya Pradesh Journal of Social Sciences*, *28*. https://doi.org/10.13140/RG.2.2.13578.03522

Flyvbjerg, B. (2014). What you should know about megaprojects and why: An overview. In *Project Management Journal* (Vol. 45, Issue 2, pp. 6–19). https://doi.org/10.1002/pmj.21409

Hernández, Miguel., & Baquero, L. (2025). *Python con orientación a objetos y al análisis de datos* (A. Gutierrez, Ed.; Primera edición). Ediciones de la U. https://www-ebooks7-24-com.bdbiblioteca.universidadean.edu.co/?il=43453

Hummel, K., & Jobst, D. (2024). An Overview of Corporate Sustainability Reporting Legislation in the European Union. *Accounting in Europe*, *21*(3), 320–355. https://doi.org/10.1080/17449480.2024.2312145

ISO. (2025). *UNE-ISO/IEC 42001 Tecnología de la información Inteligencia artificial Sistema de gestión*. www.une.org

Joyanes, Luis. (2019). *Inteligencia de negocios y analítica de datos*.

Kerzner, H. (2022). *Project management: A systems approach to planning, scheduling, and controlling* (13th ed.). Wiley.

Lee, J. J., & Lee, M. (2025). Artificial Intelligence Structuration in Machine Learning. In *Journal of Strategic Innovation and Sustainability* (Vol. 20, Issue 2).

Liu, J., Gao, X., & Chen, X. (2025). Feasibility Analysis of Optimization Models for Natural Gas Distribution Networks Using Machine Learning. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, *29*(3), 614–622. https://doi.org/10.20965/jaciii.2025.p0614

López Ferreiro, M. Á., Ruiz, J. G., García, Ó., & De La Fuente Valentín, L. (2025a). Artificial Intelligent Application in Project Management: An Algorithm Comparison for Solar Plants Planning Construction. *Expert Systems*, *42*(9), 0–19. https://doi.org/10.1111/exsy.70105

López Ferreiro, M. Á., Ruiz, J. G., García, Ó., & De La Fuente Valentín, L. (2025b). Artificial Intelligent Application in Project Management: An Algorithm Comparison for Solar Plants Planning Construction. *Expert Systems*, *42*(9), 0–19. https://doi.org/10.1111/exsy.70105

Ma, F., Altalbawy, F. M. A., Patel, P., Manjunatha, R., Kalia, R., Formanova, S., Naveen, P. R., Joshi, K. K., Sinha, A., Kandahari, A. Y., Al-Rubaye, T. M. K., & Alam, M. M. (2025). Predictive modeling of oil rate for wells under gas lift using machine learning. *Scientific Reports*, *15*(1). https://doi.org/10.1038/s41598-025-12129-w

Madiwale, P., & Mahadik, R. (2023). Hybrid project management framework for the engineering sector. *Journal of Engineering Management Studies*, *5*(1), 45–58.

Mali, A. S., Kolhe, A., Gorde, P., Kolekar, A., Umbrajkar, A., Solepatil, S., & Zare, K. (2025). Application of artificial intelligence and machine learning in construction project

management: a comparative study of predictive models. *Asian Journal of Civil Engineering*, *26*(6), 2671–2686. https://doi.org/10.1007/s42107-025-01335-6

Martínez Pérez, J. A., & Pérez Martín, P. S. (2024). Regresión logística. *Medicina de Familia. SEMERGEN*, *50*(1), 102086. https://doi.org/https://doi.org/10.1016/j.semerg.2023.102086

Maurya, S., Lakkimsetty, N. R., Manjunath, T., Shukla, A., Sethy, B., & Behera, R. (2025). Balancing accuracy and interpretability: AI-driven predictive modeling of construction schedule performance in India. *Asian Journal of Civil Engineering*, *26*, 3083–3098. https://doi.org/10.1007/s42107-025-01363-2

Mohseni, M., & Mustafa Kamal, E. (2025). Evaluating Machine Learning Models for Predict Cost Overruns in Petrochemical Projects. *PaperASIA*, *41*(1b), 45–57. https://doi.org/10.59953/paperasia.v41i1b.301

Moussa, A., Ezzeldin, M., & El-Dakhakhni, W. (2024). Predicting and managing risk interactions and systemic risks in infrastructure projects using machine learning. *Automation in Construction*, *168*. https://doi.org/10.1016/j.autcon.2024.105836

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company*. Oxford University Press.

Oberlender, G. D., Spencer, G. R., & Lewis, R. M. (2022). Design Proposals. In *Project Management for Engineering and Construction: A Life-Cycle Approach* (4th Edition). McGraw-Hill Education. https://www.accessengineeringlibrary.com/content/book/9781264268443/chapter/chapter9

Parrales García, N. R., Baque Parrales, E. M., Baque Cantos, M. A., & Moreno Ponce, M. R. (2024). Integración de la Inteligencia artificial en la formulación de proyectos: Oportunidades, desafíos y perspectivas futuras. *RECIAMUC*, *8*(1), 463–477. https://doi.org/10.26820/reciamuc/8.(1).ene.2024.463-477

Podder, S., & Podder, S. (2025). *Cost Overrun Prediction in Road Construction: A Fuzzy Logic and Clustering Approach*. https://doi.org/10.1061/9780784486207.060

Priyadarshy, S., & Moonsammy, D. (2021, December 16). *A machine learning risk management framework for sustainable oil and gas solutions.* . Share Article. https://www.halliburton.com/en/energy-pulse/a-machine-learning-risk-management-framework-for-sustainable-oil-and-gas-solutions

Project Management Institute. (2021). *A guide to the project management body of knowledge (PMBOK® guide) (7th ed.)*. https://www.pmi.org/standards/pmbok

Project Management Institute (PMI). (2021). *A Guide to the Project Management Body of Knowledge (PMBOK® Guide) (7th ed.)*. Project Management Institute (PMI).

Rumane, A. R. (2024). *Quality Management in Oil and Gas Projects*.

salama, A. (2025). Evaluating the impact of construction delays on project duration using machine learning and multi-criteria decision analysis. *Asian Journal of Civil Engineering*, *26*(1), 389–399. https://doi.org/10.1007/s42107-024-01196-5

Serrador, P., & Pinto, J. K. (2019). Does Agile work?—A quantitative analysis of agile project success. *International Journal of Project Management*, *37*(5), 623–633.

Serrano-Gomez, L., & Muñoz-Hernandez, J. I. (2020). Risk influence analysis assessing the profitability of large photovoltaic plant construction projects. *Sustainability (Switzerland)*, *12*(21), 1–16. https://doi.org/10.3390/su12219127

Szadeczky, T., & Bederna, Z. (2025). Risk, regulation, and governance: evaluating artificial intelligence across diverse application scenarios. *Security Journal*, *38*(1). https://doi.org/10.1057/s41284-025-00495-z

Tabassi, E. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://doi.org/10.6028/NIST.AI.100-1

Tshidavhu, F., & Khatleli, N. (2020). An assessment of the causes of schedule and cost overruns in South African megaprojects: A case of the critical energy sector projects of Medupi and Kusile. *Acta Structilia*, *27*(1), 119–143. https://doi.org/10.18820/24150487/as27i1.5

Turkyilmaz, A. H., & Polat, G. (2024). Risk-Based Completion Cost Overrun Ratio Estimation in Construction Projects Using Machine Learning Classification Algorithms: A Case Study. *Buildings*, *14*(11). https://doi.org/10.3390/buildings14113541

Wang, P., Wang, K., Huang, Y., & Fenn, P. (2025). Probability, Formation, and Prediction of Large-Size Construction Cost Overruns Governed by a Power-Law Distribution. *Journal of Construction Engineering and Management*, *151*. https://doi.org/10.1061/JCEMD4.COENG-16445

Waqar, A., Othman, I., Shafiq, N., & Mansoor, M. S. (2023). Applications of AI in oil and gas projects towards sustainable development: a systematic literature review. *Artificial Intelligence Review*, *56*(11), 12771–12798. https://doi.org/10.1007/s10462-023-10467-7

Zhang, M., Lei, Z., Yan, C., Zeng, B., Huang, F., Qu, T., Wang, B., & Fu, L. (2025). Construction of Analogy Indicator System and Machine-Learning-Based Optimization of Analogy Methods for Oilfield Development Projects. *Energies*, *18*(15), 4076. https://doi.org/10.3390/en18154076