



UNIVERSIDAD
NACIONAL
DE COLOMBIA

Predicción de sobrecostos en proyectos de construcción de edificación (uso no residencial) empleando una técnica de Machine Learning. Caso de estudio: Capital Project Schedules and Budgets

Andrés Felipe Ospina Castañeda

Universidad Nacional de Colombia

Facultad de Minas, Departamento de Ciencias de la Computación y la decisión

Medellín, Colombia

2025

Predicción de sobrecostos en proyectos de construcción de edificación (uso no residencial) empleando una técnica de Machine Learning. Caso de estudio: Capital Project Schedules and Budgets

Andrés Felipe Ospina Castañeda

Trabajo de investigación presentado como requisito parcial para optar al título de:
Magister en Ingeniería - Analítica

Director:

Ph.D. Albeiro Espinosa Bedoya

Codirector (a):

Ph.D. Miguel David Rojas López

Grupo de Investigación:

Centro de Investigación y Consultoría Organizacional - CINCO

Universidad Nacional de Colombia

Facultad de Minas, Departamento de Ciencias de la Computación y la decisión

Medellín, Colombia

2025

*A mis padres,
Fernando y Marina
Y mis hermanos,
Melisa y Daniel*

Agradecimientos

Agradezco profundamente a Dios por acompañarme con sabiduría y fortaleza en cada etapa de este proceso. A mi familia, en especial a mis padres y hermanos, por haber sido siempre un pilar firme de apoyo, confianza y afecto desde los primeros años de mi formación. Su ejemplo y dedicación han sido fundamentales para alcanzar este logro.

Expreso mi gratitud a la Universidad Nacional de Colombia, por abrirme las puertas a una formación rigurosa y con sentido social. Agradezco especialmente al Instituto de Educación en Ingeniería (IEI) de la Facultad de Minas, por permitirme formar parte de su equipo de trabajo y por el respaldo económico que me ofrecieron a través de los servicios prestados, lo cual me permitió cumplir con las obligaciones universitarias y avanzar con estabilidad en esta etapa académica.

A todos los docentes que, directa o indirectamente, contribuyeron con sus enseñanzas, reflexiones y exigencia académica a la construcción de este trabajo. De manera especial, al profesor Albeiro Espinosa Bedoya, director de esta tesis, y al profesor Miguel David Rojas López, codirector, por su compromiso, orientación constante y valiosos aportes. Su acompañamiento hizo de este proceso una experiencia de profundo aprendizaje, rigurosidad y sentido.

Resumen

Predicción de sobrecostos en proyectos de construcción de edificación (uso no residencial) empleando una técnica de Machine Learning. Caso de estudio: Capital Project Schedules and Budgets

El presente estudio propone un modelo predictivo basado en técnicas de *Machine Learning* para anticipar sobrecostos en proyectos de construcción de edificación de uso no residencial, tomando como caso de estudio la base de datos “*Capital Project Schedules and Budgets*” de la *School Construction Authority* (SCA) de Nueva York. La investigación surge ante la necesidad de superar las limitaciones de los enfoques tradicionales en la gestión de riesgos en costos. A partir del enfoque CRISP-DM, se llevó a cabo un proceso estructurado que incluyó la comprensión del negocio, análisis exploratorio de datos, selección de variables relevantes, transformación de datos y entrenamiento de modelos predictivos. Se evaluaron cuatro algoritmos: *Linear Regression*, *Random Forest Regressor*, *Multi Layer Perceptron Regressor*, y *Gradient Boosting Regressor*, siendo este último el de mejor desempeño, alcanzando un coeficiente de determinación (R^2) de 0.9824, con un error cuadrático medio (MSE) de 309.699.558 y un error absoluto medio (MAE) de 3.887. El análisis identificó que las variables más influyentes en los sobrecostos fueron de tipo financiero, destacándose el presupuesto total del proyecto, el gasto real estimado a la fecha y el presupuesto final estimado. En contraste, variables categóricas como el tipo de proyecto o la fase constructiva mostraron baja significancia estadística. Asimismo, la validación del modelo mediante *K-Fold Cross Validation* confirmó su capacidad de generalización, sin indicios de sobreajuste.

Palabras clave: sobrecostos, Machine Learning, construcción, predicción, gestión de riesgos, Gradient Boosting.

Abstract

Prediction of Cost Overruns in Building Construction Projects (Non-Residential Use) Using a Machine Learning Technique. Case Study: Capital Project Schedules and Budgets

This study proposes a predictive model based on Machine Learning techniques to anticipate cost overruns in non-residential building construction projects, using the “Capital Project Schedules and Budgets” dataset from the New York School Construction Authority (SCA) as a case study. The research emerges from the need to overcome the limitations of traditional approaches to cost risk management. Following the CRISP-DM framework, a structured process was conducted, including business understanding, exploratory data analysis, selection of relevant variables, data transformation, and training of predictive models. Four algorithms were evaluated: Linear Regression, Random Forest Regressor, Multi-Layer Perceptron Regressor, and Gradient Boosting Regressor. The latter showed the best performance, achieving a coefficient of determination (R^2) of 0.9824, a mean squared error (MSE) of 309,699,558, and a mean absolute error (MAE) of 3,887. The analysis identified financial variables as the most influential in cost overruns, with total project budget, estimated actual expenditure to date, and final estimated budget standing out. In contrast, categorical variables such as project type or construction phase showed low statistical significance. Moreover, the model's validation through K-Fold Cross Validation confirmed its generalization capability, with no signs of overfitting.

Keywords: cost overruns, Machine Learning, construction, prediction, risk management, Gradient Boosting.

Contenido

	Pág.
Agradecimientos	VI
Resumen.....	VII
Abstract.....	VIII
Lista de figuras.....	XI
Lista de tablas	XIII
Introducción	1
1. Presentación del proyecto.....	5
1.1 Planteamiento del problema.....	5
1.2 Justificación.....	6
1.3 Objetivos	9
1.3.1 Objetivo general	9
1.3.2 Objetivos específicos.....	9
1.4 Alcance.....	9
2. Marco teórico.....	11
2.1 Aplicación de IA en construcción	11
2.1.1 Linear Regression.....	14
2.1.2 Random Forest.....	15
2.1.3 Gradient Boosting Regressor.....	16
2.1.4 Neural Networks.....	16
2.1.5 Evaluación del desempeño de modelos	17
2.2 Gestión de riesgos en proyectos de construcción de edificación	19
2.2.1 Definición y clasificación de riesgos en construcción.....	19
2.2.2 Metodologías para la identificación, análisis y mitigación de riesgos	22
2.2.3 Factores que contribuyen a sobrecostos en proyectos de construcción.....	26
2.2.4 Métodos determinísticos.....	27
2.2.5 Métodos probabilísticos.....	29
3. Estado del arte	31
3.1 Revisión sistemática de literatura.....	31
3.2 Análisis bibliométrico	36

3.3	Análisis estadístico.....	39
3.4	Hallazgos y discusión.....	42
4.	Metodología.....	47
4.1	Comprensión del negocio.....	48
4.1.1	Problema y su contexto.....	49
4.1.2	Valor del análisis de datos.....	49
4.2	Comprensión de los datos.....	50
4.2.1	Identificación de los datos.....	51
4.2.2	Análisis exploratorio de los datos.....	53
4.3	Preparación de los datos.....	63
4.3.1	Selección de variables relevantes.....	63
4.3.2	Transformación de datos.....	68
4.3.3	División del conjunto de datos.....	69
4.4	Modelado.....	69
4.4.1	Linear Regression.....	70
4.4.2	Random Forest Regressor.....	72
4.4.3	Gradient Boosting Regressor.....	73
4.4.4	Multi Layer Perceptron Regressor.....	74
4.4.5	Evaluación de modelos.....	75
4.5	Validación.....	79
4.5.1	Esquema de validación.....	79
4.5.2	Hiperparámetros evaluados.....	80
4.5.3	Iteraciones y resultados.....	80
4.6	Despliegue.....	83
5.	Conclusiones y recomendaciones.....	85
5.1	Conclusiones.....	85
5.2	Recomendaciones.....	87
A.	Anexo: Factores de riesgo que implican sobrecostos en proyectos.....	89
B.	Anexo: Relación de los factores de riesgo con variables de los proyectos.....	90
C.	Anexo: Modelo de bosques aleatorios.....	92
D.	Instrucciones y uso del repositorio.....	93
	Bibliografía.....	96

Lista de figuras

	Pág.
Figura 1-1: Tendencias de publicación en implementación de IA.	8
Figura 2-1: Componentes, tipos y subcampos de IA.....	11
Figura 2-2: Aplicación de los subcampos de IA en la construcción.	13
Figura 2-3: Estructura de desglose de riesgos para proyectos de construcción.....	21
Figura 3-1: Dominios para ecuación de búsqueda.....	32
Figura 3-2: Flujo basado en PRISMA para revisión sistemática de literatura.	35
Figura 3-3: Diagrama de Co-ocurrencia entre palabras clave.	37
Figura 3-4: Diagrama de Co-ocurrencia de palabras clave y evolución en el tiempo.....	38
Figura 3-5: Frecuencia de algoritmos usados en la gestión de riesgos en construcción. ..	39
Figura 3-6: Fases de gestión de riesgos cubierta en las fuentes analizadas.....	40
Figura 3-7: Evolución de los términos clave en la fuente analizada.	41
Figura 4-1: Marco conceptual para cálculo de sobrecostos basado en CRISP-DM.....	48
Figura 4-2: Criterios de evaluación del modelo propuesto.....	50
Figura 4-3: Renombramiento de variables.	53
Figura 4-4: Estado de presupuesto de los proyectos.	55
Figura 4-5: Distribución de presupuestos y gastos.....	56
Figura 4-6: Distribución de Duraciones y Retrasos.	57
Figura 4-7: Distribución de desviación presupuestaria.	59
Figura 4-8: Matriz de dispersión entre variables numéricas.....	60
Figura 4-9: Sobrecosto por fase del proyecto.....	61
Figura 4-10: Desviación presupuestaria según categoría temática.....	62
Figura 4-11: P-Values de variables categóricas frente a la cantidad real gastada.....	66
Figura 4-12: Coeficiente de correlación de Pearson.....	68

Figura 4-13: Predicciones Vs. Valores reales en el modelado.	77
Figura 4-14: Importancia de variables en los modelos.....	78
Figura 4-15: Comportamiento de métricas por iteración.....	82

Lista de tablas

	Pág.
Tabla 2-1: Comparación entre PMBOK, PRINCE 2, y ICB.	22
Tabla 2-2: Descripción general de la gestión de los riesgos del proyecto.	23
Tabla 2-3: Especificaciones de los riesgos identificados.	26
Tabla 3-1: Técnicas de referencia reportadas en la fuente analizada.	44
Tabla 4-1: Características de la base de datos.	51
Tabla 4-2: Cantidad y tipo de registros - SCA Capital Projects.	52
Tabla 4-3: Resumen estadístico de presupuesto, gastos y duración de proyectos.	58
Tabla 4-4: Categorías generales de la descripción del proyecto.	61
Tabla 4-5: Métricas de evaluación en el conjunto de prueba.	76
Tabla 4-6: Registro de iteraciones.	81

Introducción

El desarrollo de proyectos de construcción de edificación está orientado al cumplimiento de objetivos relacionados con el costo, el tiempo, la calidad y la seguridad. No obstante, estos objetivos están expuestos a una amplia gama de riesgos e incertidumbres, derivados de la naturaleza compleja y singular de cada proyecto (Arthur, 2021). La interacción de múltiples actores, la alta variabilidad de condiciones técnicas y contextuales, y la fuerte dependencia del juicio experto en la toma de decisiones hacen que los proyectos de construcción de edificación estén expuestos a desviaciones no planificadas (Flanagan & Norman, 1993).

Uno de los riesgos más frecuentes y con mayor impacto es el sobrecosto, entendido como el incremento del costo real de ejecución respecto al presupuesto inicialmente aprobado. Esta situación no solo compromete la viabilidad financiera del proyecto, sino que puede generar consecuencias en la calidad del producto final, en el cumplimiento de plazos contractuales y en la reputación de las organizaciones responsables.

La incorporación de tecnologías emergentes, como la inteligencia artificial (IA), ofrece nuevas oportunidades para transformar el proceso de gestión de riesgos. En particular, el aprendizaje automático (*Machine Learning*, ML) ha demostrado ser eficaz para automatizar tareas complejas, identificar patrones ocultos en grandes volúmenes de datos y generar modelos predictivos aplicables a problemas reales de la industria (Khodabakhshian et al., 2024). En el campo de la construcción de edificaciones, esto se traduce en una capacidad creciente para anticipar riesgos como los sobrecostos, y tomar decisiones más informadas durante el ciclo de vida del proyecto.

No obstante, el uso efectivo de esta técnica requiere conjuntos de datos suficientemente robustos, estructurados y representativos. Como advierten Tayefeh Hashemi et al. (2020) y Wahab & Wang (2022), la estimación de sobrecostos en proyectos de construcción de

edificación está condicionada por la disponibilidad y calidad de los datos históricos. La carencia de información adecuada puede conducir a estimaciones poco precisas, lo cual desencadena una serie de decisiones erradas que afectan el cumplimiento de los objetivos a nivel financiero del proyecto (Arabzadeh et al., 2018).

En este contexto, la presente investigación propone un modelo predictivo basado en técnicas de *Machine Learning* para anticipar sobrecostos en proyectos de construcción de edificación de uso no residencial. Como caso de estudio se emplea la base de datos “*Capital Project Schedules and Budgets*”, desarrollada por la *School Construction Authority* (SCA) de Nueva York, la cual contiene información sobre fechas, presupuestos, fases constructivas y ejecución financiera de proyectos de construcción escolares.

La metodología utilizada se estructura bajo el enfoque CRISP-DM (*Cross Industry Standard Process for Data Mining*), reconocido en la industria analítica por su capacidad para guiar procesos de modelado de datos de manera ordenada y replicable. Este enfoque comprende seis fases: comprensión del negocio, comprensión de los datos, preparación de los datos, modelado, evaluación y despliegue. A lo largo del estudio, se aplicaron técnicas de análisis exploratorio, selección y transformación de variables, y entrenamiento de modelos de regresión para el desarrollo de la propuesta.

El alcance de esta investigación se ajusta al desarrollo y validación técnica de un modelo predictivo a partir de datos históricos. No contempla su implementación operativa en plataformas de gestión ni la adaptación organizacional de sus resultados. Adicionalmente, debido a la carencia de bases de datos públicas con nivel de detalle comparable en el contexto colombiano, no se considera su aplicación inmediata a nivel nacional. No obstante, el estudio busca sentar bases conceptuales y metodológicas que puedan ser utilizadas en futuras investigaciones o desarrollos institucionales que aspiren a incorporar inteligencia artificial en la gestión del riesgo de sobrecostos.

Este trabajo se enmarca en un campo de estudio emergente que vincula la ingeniería de datos con la gestión de proyectos de construcción. Su aporte se orienta a demostrar el potencial de

las técnicas de *Machine Learning* como herramienta de soporte a la toma de decisiones, especialmente en un entorno como el de la construcción de edificaciones, caracterizado por su alta exposición al riesgo, la presión por cumplir objetivos financieros y la creciente demanda de eficiencia y transparencia.

1. Presentación del proyecto

1.1 Planteamiento del problema

El desarrollo de proyectos de construcción de edificación enfrenta riesgos que pueden influir en sus resultados. A lo largo de cada fase, desde la etapa inicial hasta la entrega final, estos proyectos son susceptibles a factores impredecibles, como condiciones climáticas adversas, fluctuaciones en los costos, retrasos en el cronograma y otras variables que afectan la calidad y la seguridad (Aggabou et al., 2024). En este escenario, algunos de los objetivos principales en proyectos de construcción de edificación incluyen preservar el cronograma, el costo, la calidad, la seguridad y los objetivos ambientales, independientemente de la naturaleza y las fuentes de los riesgos (Zou et al., 2007).

Ante esta realidad, la gestión de riesgos se convierte en un componente clave de la dirección de proyectos. El Project Management Institute (2021) define la gestión de riesgos como el conjunto de acciones destinadas a maximizar la probabilidad y el impacto de los riesgos positivos, y a minimizar la probabilidad e impacto de los riesgos negativos. Una adecuada gestión de riesgos es importante, ya que los riesgos no controlados pueden desviar los proyectos de sus objetivos.

En las prácticas tradicionales de gestión de riesgos, los expertos en el área generan un inventario de riesgos potenciales que podrían afectar los objetivos del proyecto. Posteriormente, se realizan análisis cualitativos y cuantitativos para priorizar estos riesgos en función de factores como la probabilidad, el impacto y las consecuencias potenciales. Las respuestas a los riesgos se desarrollan para abordar los riesgos priorizados, asegurando que el proyecto permanezca encaminado y dentro de sus limitaciones. Adicionalmente, se monitorea y controla continuamente los riesgos a lo largo del ciclo de vida del proyecto, con

revisiones y reevaluaciones periódicas para garantizar que el perfil de riesgo esté actualizado y que se tomen las acciones necesarias (Weng, 2023).

No obstante, estas prácticas convencionales de identificación y evaluación de riesgos, basadas en juicios individuales y experiencias previas, son personalizadas y dependientes del contexto. Esto plantea problemas en la transferencia de conocimiento y la generalización de modelos para futuros proyectos (Li et al., 2018).

Con base en lo anterior, este trabajo involucra uno de los desafíos en la gestión de proyectos de construcción de edificación: la gestión del riesgo de sobrecostos. Los sobrecostos, también conocidos como desviaciones presupuestarias, ocurren cuando el costo real de un proyecto de construcción en general excede el presupuesto inicialmente aprobado. Este fenómeno es común en la construcción a nivel global y plantea retos para el desempeño y la viabilidad financiera de los proyectos (Ammar et al., 2022). Aproximadamente el 86% de los proyectos de construcción experimentan sobrecostos (Hammad et al., 2014). Controlar el presupuesto del proyecto a lo largo de su ciclo de vida es un desafío constante para las empresas de construcción, debido a la interdependencia y naturaleza dinámica de los factores de riesgo asociados al sobrecosto (Ashtari et al., 2022).

1.2 Justificación

El sector de la construcción se enfrenta a desafíos que han obstaculizado su crecimiento y han dado lugar a niveles de productividad bajos en comparación con otros sectores como el de la manufactura (Abioye et al., 2021). La ausencia de conocimientos digitales y la adopción de tecnología dentro del sector de la construcción se ha relacionado con ineficiencias de costos, retrasos en los proyectos, bajo desempeño, toma de decisiones desinformada y baja productividad (Nikas et al., 2007).

Dentro de las opciones para la aplicación de tecnologías en el sector de la construcción, se presenta la inteligencia artificial (IA), la cual ha ayudado a lograr mejoras en las operaciones comerciales, procesos de servicio y productividad de varios sectores en los últimos años

(Abioye et al., 2021). Los subcampos de la IA, como el *Machine Learning* (ML) y el *Natural Language Processing* (NLP), se han aplicado para abordar problemas complejos y respaldar la toma de decisiones (Rao et al., 2021). Por ejemplo, en la industria manufacturera, la aparición de la cuarta revolución industrial, comúnmente conocida como Industria 4.0, ha llevado a importantes mejoras de procesos, rentabilidad, reducción de tiempos de producción, mejora de la seguridad y ha ayudado a alcanzar los objetivos de sostenibilidad en las empresas (Chien et al., 2020).

El sector de la construcción está experimentando una revolución digital, impulsada por avances en tecnologías emergentes como los gemelos digitales y el Internet de las Cosas (IoT). Sin embargo, la adopción de estas herramientas en gestión de riesgos en proyectos de construcción de edificación ha sido limitada (Chenya et al., 2022). Las razones de esto se deben principalmente a la falta de datos estructurados, la dependencia de juicios de expertos y la dificultad para realizar un análisis adecuado de las interdependencias causales entre los riesgos (Khodabakhshian et al., 2023). Los métodos tradicionales de gestión de riesgos siguen siendo manuales y están basados en la experiencia y juicio individual. Esta forma de trabajo ha demostrado ser ineficaz y tiende a no adaptarse a los rápidos cambios y desafíos que surgen a lo largo de los proyectos de construcción de edificación (Khodabakhshian et al., 2023).

La motivación de este estudio radica en abordar uno de los desafíos en cuanto a riesgo, que comúnmente se presenta en proyectos de construcción de edificación: *los sobrecostos*. A través del uso de un subcampo de IA como el *Machine Learning*. Se busca desarrollar un modelo predictivo que integre las interdependencias entre los factores de riesgo y permita anticipar fluctuaciones desviaciones presupuestarias. Este enfoque no solo contribuirá a la reducción de desviaciones, sino que también establecerá bases de conocimiento para futuras aplicaciones. La implementación de este modelo permitirá mejorar la toma de decisiones.

Con el propósito de explorar el estado del arte sobre la aplicación de herramientas de IA en la gestión de riesgos en proyectos de construcción, se realizó una búsqueda bibliográfica en

la base de datos *Scopus*. Para ello, se utilizó la **Ecuación 3-1**, diseñada para combinar términos relacionados con IA, gestión de riesgos y proyectos de construcción de edificación.

El análisis se centró en publicaciones académicas correspondientes a los últimos 10 años, con el objetivo de identificar tendencias recientes y evaluar el grado de desarrollo del tema en la literatura científica. Los resultados iniciales arrojaron un conjunto amplio de documentos. Sin embargo, tras aplicar criterios de depuración que incluyeron la eliminación de duplicados, la revisión del enfoque temático y la exclusión de trabajos no pertinentes el número de registros relevantes se redujo a 169.

Esta limitada cantidad de publicaciones sugiere que la integración de inteligencia artificial en la gestión de riesgos en proyectos de construcción de edificaciones es aún un campo emergente en la literatura científica. La **Figura 1-1** presenta la evolución temporal de las publicaciones identificadas, lo que permite visualizar la dinámica del interés investigativo en este tema a lo largo del periodo analizado. No obstante, se observa una tendencia de investigación. Esto refleja interés en la integración de IA en gestión de riesgos, es decir, el campo de estudio se encuentra en fases iniciales. Este incremento en la investigación puede indicar que el sector de la construcción está empezando a reconocer el potencial de la IA para transformar la forma en que se gestionan los riesgos.

Figura 1-1: Tendencias de publicación en implementación de IA.



Fuente: Propia.

1.3 Objetivos

1.3.1 Objetivo general

Proponer un modelo basado en *Machine Learning* para predecir sobrecostos en proyectos de construcción de edificación de uso no residencial. Caso de estudio: Capital Project Schedules and Budgets of School Construction Authority.

1.3.2 Objetivos específicos

- Estudiar las variables relevantes que influyen en los sobrecostos de proyectos de construcción de edificación de uso no residencial, considerando datos históricos.
- Proponer un modelo basado en *Machine Learning* teniendo en cuenta las variables identificadas en el objetivo anterior.
- Validar el modelo predictivo con base en la información suministrada en la base de datos del caso de estudio.

1.4 Alcance

Este trabajo tiene como alcance la formulación y desarrollo, a partir de código, de un modelo predictivo con capacidad para anticipar desviaciones presupuestarias en proyectos de construcción de edificación. El modelo se construye con base en técnicas analíticas y algoritmos de *Machine Learning*, empleando conjuntos de datos históricos que permitan identificar patrones y factores asociados al sobrecosto.

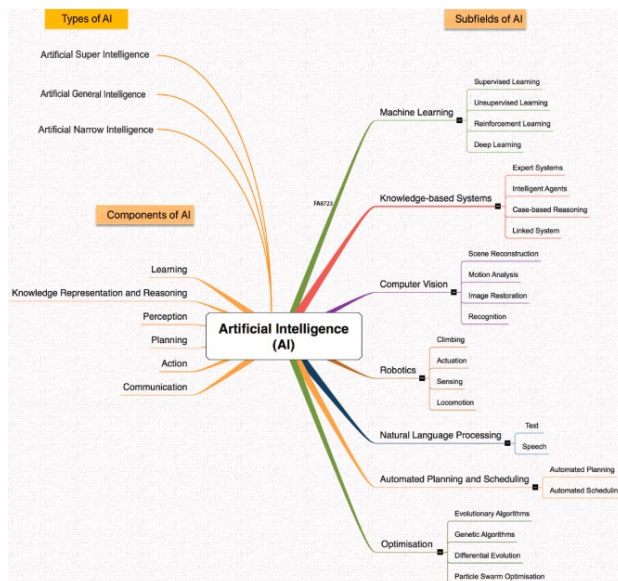
El enfoque del estudio se limita al diseño, entrenamiento, validación y evaluación del desempeño del modelo desde una perspectiva técnica, sin contemplar su implementación en entornos organizacionales ni su integración en sistemas productivos o plataformas de gestión empresarial.

2. Marco teórico

2.1 Aplicación de IA en construcción

La idea de desarrollar máquinas con inteligencia similar a la humana tiene sus raíces en disciplinas como la filosofía, la ficción, la imaginación, la informática, la electrónica y la ingeniería. Un hito clave en el campo de la inteligencia artificial (IA) fue la prueba de *Turing*, propuesta por Alan Turing, la cual superó las perspectivas teológicas tradicionales y las conclusiones matemáticas sobre la posibilidad de máquinas inteligentes. Años después, las máquinas inteligentes han superado a los humanos en múltiples ámbitos, como el aprendizaje. Según Abioye et al., (2021) la IA se puede describir como "*el estudio de cómo hacer que las máquinas realicen tareas que, por el momento, los humanos hacen mejor*". La **Figura 2-1** muestra una descripción general de los tipos, componentes y subcampos de la IA.

Figura 2-1: Componentes, tipos y subcampos de IA.



Fuente: (Abioye et al., 2021)

Para analizar el estado de la inteligencia artificial en la construcción, se requiere identificar sus áreas de desarrollo. A lo largo del tiempo, la aplicación de la IA ha dado lugar a diferentes ramas, entre ellas: (a) aprendizaje automático, (b) visión por computadora, (c) procesamiento del lenguaje natural, (d) optimización, (e) robótica y (f) planificación y programación automatizadas. A continuación, se explica cada una de estas áreas según Abioye et al., (2021):

(A) Aprendizaje automático: El aprendizaje automático se enfoca en desarrollar programas capaces de aprender de datos previos para modelar, controlar o predecir sin necesidad de una programación explícita. Dentro de sus enfoques está el aprendizaje supervisado, donde las máquinas toman decisiones basadas en datos etiquetados; el aprendizaje no supervisado, que permite identificar patrones en datos no etiquetados mediante técnicas de agrupamiento y reducción de dimensión; y el aprendizaje profundo (*Deep Learning*), una técnica más avanzada, capaz de ofrecer predicciones más precisas que los métodos tradicionales.

(B) Visión por computadora: La visión artificial es un área multidisciplinaria enfocada en replicar el sistema visual humano de manera artificial. Su objetivo es permitir que las máquinas interpreten imágenes digitales y multidimensionales a un nivel avanzado, capturándolas con dispositivos especializados, procesándolas mediante algoritmos avanzados y analizándolas para optimizar la toma de decisiones.

(C) Procesamiento del lenguaje natural: El procesamiento del lenguaje natural es una rama de la inteligencia artificial que desarrolla modelos computacionales para imitar las habilidades lingüísticas humanas. Se aplica en traducción automática, resumen de textos, interfaces de usuario, recuperación de información, reconocimiento de voz y sistemas expertos.

(D) Optimización: La optimización consiste en tomar decisiones que ofrezcan los mejores resultados dentro de ciertas restricciones o el proceso de elegir la mejor opción entre varias disponibles.

(E) Robótica: Los robots son dispositivos automatizados diseñados para ejecutar tareas físicas en el mundo real. La robótica, como disciplina interdisciplinaria, abarca su diseño, fabricación, operación y mantenimiento, permitiendo que imiten acciones humanas mediante sensores. Estos dispositivos suelen especializarse en tareas específicas y no siempre tienen forma humanoide.

(F) Planificación y programación automatizadas: La planificación en IA ayuda a los sistemas inteligentes a alcanzar objetivos ordenando acciones según sus resultados esperados, mientras que la programación distribuye el tiempo y los recursos necesarios para ejecutar esos planes. Estas técnicas se usan en problemas complejos donde sus beneficios superan los costos.

En **Figura 2-2** se presentan las aplicaciones de los subcampos de la IA previamente definidos, dentro de actividades y operaciones de construcción.

Figura 2-2: Aplicación de los subcampos de IA en la construcción.

Salud y seguridad	A	B	C			
Programación	A	D	F			
Estimación de costos	A		D			
Aspectos legales	A	B	C	D		
Cadena de suministro y logística	A		D			
Monitoreo del sitio y evaluación del rendimiento	A	B	C	E		
Gestión de materiales	A	B	C	D	E	
Montaje fuera del sitio	A		E			
Gestión de plantas y equipos	A	B	D	E		
Planificación de proyectos	A	B	C	D	E	F
Gestión del conocimiento	A	B	C			
Diseño	A	B	D			
Gestión de riesgos	A	C	D	E		
Estructuras temporales	A		D			
Ofertas/Licitaciones	A		C			
Gestión de la energía/Sostenibilidad	D					

Fuente: Adaptado de Abioye et al. (2021)

En la **Figura 2-2** se puede notar que el aprendizaje automático es el subcampo con mayor aplicabilidad, extendiéndose a casi todas las áreas, su capacidad para analizar volúmenes de datos y generar predicciones lo hacen apropiado para la toma de decisiones.

Otro subcampo es la visión por computador, la cual se utiliza en la supervisión del sitio de construcción y en la evaluación del rendimiento; su capacidad para procesar imágenes y videos la hace apropiada en actividades como la detección de riesgos laborales, el control de calidad y la automatización de inspecciones en obra. De manera similar, la planificación y programación automatizadas son relevantes en la optimización de la ejecución de proyectos, asegurando la asignación eficiente de recursos y la reducción de ineficiencias en la programación de actividades. Además, el procesamiento de lenguaje natural se centra en la gestión de contratos, licitaciones y documentación, permitiendo la extracción automatizada de información para optimizar la administración de documentos legales y reducir conflictos contractuales.

Por otro lado, los algoritmos de optimización tienen un papel determinante en la mejora de la eficiencia operativa y la sostenibilidad, al minimizar costos, mejorar tiempos de ejecución y optimizar la asignación de recursos en proyectos de gran escala.

En relación con el aprendizaje automático, a continuación, se presentan conceptos sobre algoritmos y métricas de desempeño.

2.1.1 Linear Regression

Es una técnica estadística utilizada para modelar y analizar la relación entre una variable dependiente y una o más variables independientes. Su objetivo es encontrar una línea recta (o un hiperplano en dimensiones superiores) que minimice la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos (Huang, 2020).

Desde una perspectiva matemática, la regresión lineal se expresa mediante la ecuación:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots \dots \dots + \beta_n x_n + \epsilon$$

Donde y es la variable dependiente, x_1, x_2, \dots, x_n representan las variables independientes, $\beta_0, \beta_1, \dots, \beta_n$ son los coeficientes que indican la relación entre las variables y ϵ es el término de error.

El método más común para estimar los coeficientes β es el de Mínimos Cuadrados Ordinarios (OLS, por sus siglas en inglés), el cual minimiza la suma de los residuos al cuadrado, es decir, las diferencias entre los valores observados y los valores estimados por el modelo.

La regresión lineal se aplica en campos como las ciencias sociales, los negocios, la ingeniería y la salud, ayudando a predecir resultados y comprender la relación entre variables. Sin embargo, para que este modelo sea válido, es necesario cumplir con ciertos supuestos estadísticos, entre los que se incluyen: la relación entre las variables debe ser lineal, los errores deben ser independientes entre sí, la varianza del error debe ser constante, los errores deben seguir una distribución normal (Hazelton, 2009).

2.1.2 *Random Forest*

Es un algoritmo de aprendizaje supervisado utilizado tanto para tareas de clasificación como de regresión. Su funcionamiento se basa en la construcción de múltiples árboles de decisión durante la fase de entrenamiento y la combinación de sus resultados para realizar una predicción final. Para garantizar diversidad entre los árboles, el algoritmo emplea técnicas como el *bagging* (*Bootstrap Aggregating*) y la selección aleatoria de características. Al ser un método de ensamble, *Random Forest* combina las predicciones de varios clasificadores base (árboles de decisión) con el fin de mejorar la precisión y la robustez del modelo (Kulkarni et al., 2014). En el caso de las tareas de clasificación, la decisión final se toma mediante votación mayoritaria entre los distintos árboles, mientras que, en las tareas de regresión, el resultado se obtiene al calcular el promedio de las predicciones generadas por cada árbol.

2.1.3 Gradient Boosting Regressor

Es un modelo de aprendizaje automático para regresión que construye predicciones de forma secuencial, combinando múltiples árboles de decisión simples para minimizar una función de pérdida mediante descenso por gradiente. Cada nuevo árbol corrige los errores del modelo anterior, mejorando gradualmente la precisión (Guelman, 2012). Este método combina aprendices débiles para formar un modelo robusto, ajustando la función de pérdida al problema específico. Se destaca por su alta precisión, capacidad para manejar datos complejos y valores faltantes, y por realizar selección automática de variables (Biau & Cadre, 2021).

Se ha aplicado con éxito en áreas como seguros (predicción de costos por accidentes), energía (pronóstico de consumo residencial) y salud (evaluación física con dispositivos portátiles) (Guelman, 2012).

Su rendimiento se evalúa con métricas como R^2 , MSE, MAE y RMSE, y suele superar a modelos como regresión lineal, árboles simples o máquinas de soporte vectorial. Variantes como *XGBoost* y *CatBoost* han sido desarrolladas para mejorar aún más su eficiencia y precisión (Konstantinov et al., 2021).

2.1.4 Neural Networks

Son modelos computacionales inspirados en la estructura y función del cerebro humano. Están compuestas por unidades interconectadas llamadas neuronas, que procesan información de manera similar a las neuronas biológicas. Su diseño busca replicar la organización del cerebro, donde las neuronas establecen conexiones y se comunican a través de sinapsis. Estas redes tienen la capacidad de aprender a partir de los datos, ajustando sus parámetros mediante procesos de entrenamiento para mejorar su desempeño en tareas específicas (Katal & Singh, 2022).

Existen distintos tipos de redes neuronales, entre las que destacan las redes neuronales artificiales, utilizadas en disciplinas como la informática y la ingeniería. Las redes

neuronales convolucionales, especializadas en el procesamiento de datos en forma de grilla, como imágenes, y son efectivas en clasificación de imágenes y detección de objetos. Por otro lado, las redes neuronales de picos imitan la dinámica temporal de las neuronas biológicas y se emplean en aplicaciones de bajo consumo energético (Dampfoffer et al., 2023).

Entre sus principales ventajas se encuentra su capacidad de adaptación, ya que pueden ajustarse a nuevos datos y mejorar su rendimiento con el tiempo. También destacan por su capacidad de procesamiento en paralelo, lo que las hace eficientes para manejar grandes volúmenes de información, además de su robustez, ya que pueden operar con datos ruidosos o incompletos, lo que las hace ideales para aplicaciones del mundo real (Katal & Singh, 2022).

2.1.5 Evaluación del desempeño de modelos

La evaluación del rendimiento de modelos es un proceso que consiste en analizar el desempeño de un modelo con datos nuevos y no vistos, facilitando así la selección del modelo predictivo óptimo. Dentro de este proceso, la selección de modelos se basa en puntuaciones de evaluación que reflejan su calidad, garantizando además la consistencia y fiabilidad de los resultados en diferentes conjuntos de datos.

Existen varias métricas para medir el rendimiento de los modelos, dependiendo de su naturaleza. En modelos de clasificación, es común el uso de métricas como el área bajo la curva ROC y el puntaje F1, mientras que, en modelos de regresión, el caso del presente estudio, se emplean medidas como el coeficiente de determinación (R^2), el error cuadrático medio (MSE) y el error absoluto medio (MAE). Para evaluar el rendimiento de manera efectiva, se aplican diversas técnicas, entre ellas la validación cruzada, donde métodos como la validación cruzada *k-fold* permiten probar el modelo en datos diferentes a los usados en su entrenamiento (Pretnar Žagar & Demšar, 2022).

- **Coeficiente de determinación (R^2).** Representa el porcentaje de la varianza de la variable dependiente que es explicado por las variables independientes incluidas en el modelo. Este valor, se utiliza como un indicador de la bondad de ajuste, permitiendo evaluar qué tan bien el modelo representa los datos observados.

- **Error cuadrático medio (MSE).** Es una métrica común en el análisis de regresión, un método estadístico para modelar la relación entre una variable dependiente y una o más variables independientes. Para calcular el MSE, se toma la diferencia entre cada valor observado y su valor predicho correspondiente, se elevan al cuadrado estas diferencias y, a continuación, se promedian en todo el conjunto de datos. Este proceso resalta grandes errores debidos al paso de cuadratura, lo que hace que MSE sea sensible a los valores atípicos.

- **Error medio absoluto (MAE).** Es una métrica utilizada en estadística y análisis de los datos que cuantifica la magnitud promedio de los errores en un conjunto de predicciones, sin considerar su dirección. Se define como el promedio de las diferencias absolutas entre los valores predichos y los valores reales.

- **Validación cruzada.** Es una técnica utilizada para evaluar el desempeño de modelos de aprendizaje automático, mediante el entrenamiento del modelo en diferentes subconjuntos de los datos disponibles y su posterior evaluación en los subconjuntos complementarios. Este enfoque permite estimar la capacidad de generalización del modelo a nuevos datos no vistos. Una de las principales utilidades de la validación cruzada es la detección del sobreajuste (*overfitting*), que ocurre cuando el modelo presenta un alto rendimiento en los datos de entrenamiento, pero un desempeño deficiente en los datos de prueba.

2.2 Gestión de riesgos en proyectos de construcción de edificación

La gestión de riesgos en proyectos de construcción de edificación se emplea con el fin de minimizar incertidumbres y mejorar la toma de decisiones estratégicas. Un enfoque estructurado permite identificar, analizar y mitigar riesgos que pueden generar sobrecostos, retrasos y problemas de calidad. Para ello, se emplea la Estructura de Desglose de Riesgos (RBS), que clasifica factores internos y externos, facilitando su control y seguimiento.

Metodologías como PMBOK, PRINCE2 e ICB 4.0 guían el proceso de gestión, desde la planificación hasta el monitoreo y respuesta ante riesgos. Herramientas como matrices de probabilidad e impacto, mapas de calor y simulaciones de Monte Carlo permiten evaluar la incertidumbre y cuantificar desviaciones en costos y plazos. El sobrecosto, uno de los problemas en construcción de edificaciones, surge de factores interdependientes como planificación deficiente, cambios en el alcance y retrasos contractuales, lo que exige una evaluación constante de costos.

2.2.1 Definición y clasificación de riesgos en construcción

El riesgo en la gestión de proyectos se entiende como un evento o condición incierta que, si ocurre, puede afectar positiva o negativamente los objetivos del proyecto. Los riesgos negativos, conocidos como amenazas, pueden generar impactos desfavorables, mientras que los riesgos positivos, llamados oportunidades, pueden aportar beneficios. Dado que todo proyecto implica cierto grado de incertidumbre, es fundamental que el equipo de trabajo identifique de manera proactiva los riesgos a lo largo del desarrollo del proyecto. Para gestionar eficazmente estos riesgos, es necesario establecer umbrales que definan el nivel de exposición aceptable, estos umbrales permiten orientar las estrategias de respuesta y establecer acciones que minimicen las amenazas y potencien las oportunidades (Project Management Institute, 2021).

En los proyectos de construcción de edificación, el riesgo está presente en cada fase, debido a la complejidad de las actividades, la inversión requerida y la interacción de múltiples

factores. Se considera cualquier evento imprevisto que pueda generar efectos adversos en el presupuesto, los plazos, la calidad o la seguridad. Con el crecimiento del sector de la construcción y la adopción de nuevas tecnologías, los riesgos han aumentado en áreas como la seguridad laboral, la disponibilidad y eficiencia de los equipos, la gestión financiera y la contratación de personal. Para reducir estos riesgos, se desarrollan planes de gestión que incluyen su identificación y evaluación, así como la implementación de medidas de control destinadas a minimizar su impacto y garantizar el cumplimiento de los objetivos del proyecto (Teja & Ch, 2017).

Si bien cada proyecto de construcción es único, se puede crear una Estructura de Desglose de Riesgos (RBS) estándar según los tipos de proyectos que una empresa suele desarrollar, en la **Figura 2-3** se presenta una RBS típica basada en estudios previos. El gráfico representa una clasificación estructurada de los riesgos en proyectos de construcción, agrupa los factores de riesgo en diversas categorías y subcategorías. Se identifican riesgos relacionados con aspectos ambientales y sociales, legales, económicos, políticos, diseño, ejecución, gestión del proyecto, planificación, recursos humanos, equipos y materiales, y relaciones con socios. Cada categoría desglosa amenazas específicas que pueden afectar el desarrollo del proyecto.

Este tipo de clasificación es útil para la gestión de riesgos en proyectos de construcción de edificación, ya que permite identificar y evaluar amenazas potenciales. Al proporcionar una visión organizada de los riesgos, sirve como una herramienta para la toma de decisiones (Sohrabinejad & Rahimi, 2015).

Figura 2-3: Estructura de desglose de riesgos para proyectos de construcción.

Riesgo de los proyectos de construcción	Ambiental y social	Huelga	Cambios en las normas	Ambigüedad en las normas	Riesgos ambientales	Situación inestable en materia de seguridad	Desastres naturales
	Factores jurídicos	Retrasos en la resolución de litigios	Cambio en la tendencia de negociación	Retrasos en los pagos	Falta de moderador	Disputas legales	
	Factores económicos	Problemas de inversión	Norma de cambio de divisa	Inestabilidad económica	Retrasos en los pagos	Quiebras financieras de contratistas	Inflación
	Política	Sanción	Guerra	Cambios gubernamentales	Cambios en las normas gubernamentales		
	Diseño	Diseñadores no cualificados	Cambios en el diseño	Prisas en el diseño	Diseño erróneo	Retrasos en el diseño	
	Ejecución	Inestabilidad medioambiental	Desviaciones en la implementación	Materiales e instrumentos inadecuados	Cierre del Proyecto	Permisos y licencias	Revisiones retrasadas, incompletas e incorrectas
		Acontecimientos durante la ejecución	Retrasos en la confirmación del plan ejecutivo	Daños materiales y personales	Desprendimiento de tierra		
	Gestión de proyectos	Pérdida de personal experimentado	Cambios en los métodos de gestión	No asignar tareas a personal experimentado	Equipo de Proyecto inadecuado	Definiciones inadecuadas de metas y objetivos	Prisa en la subasta
	Planificación y programación	Planificación ineficaz	Errores de estimación/ planificación	Previsión errónea del plan temporal			
	Recursos humanos	Rotación de recursos humanos	Falta de especialistas				
	Recursos y equipos	Rendimiento de los equipos	Avería de las máquinas	Problemas de suministro de piezas de repuesto	Falta de máquinas y equipos	Cambio en los materiales durante la construcción	Calidad de los materiales
		Falta de materiales	Retrasos en el suministro de materiales				
	Socios	Interferencia del propietario	Retrasos de los consultores	Retraso en la toma de decisiones (propietario)	Retraso en el pago de los contratistas	Desafíos entre las partes interesadas	Falta de conocimiento de lo consultores sobre términos y condiciones
Método de licitación y selección de proveedores		Comunicación deficiente entre las partes del proyecto	Falta de mediadores para resolver problemas	Retraso de los contratistas	Mala gestión de los contratistas	Gestión deficiente del taller	
Falta de experiencia de los contratistas		Poder financiero de los contratistas					

Fuente: Adaptado de Sohrabinejad & Rahimi (2015)

2.2.2 Metodologías para la identificación, análisis y mitigación de riesgos

Según el Project Management Institute (2021) la gestión de riesgos en un proyecto abarca una serie de procesos que incluyen la planificación, identificación, análisis, diseño e implementación de respuestas, así como el seguimiento de los riesgos a lo largo del proyecto. Su propósito es incrementar la probabilidad e impacto de los riesgos favorables y reducir la probabilidad e impacto de los riesgos negativos, con el objetivo de mejorar las posibilidades de éxito del proyecto. Ahora bien, la gestión de riesgos en proyectos se aborda de manera diferente según el estándar utilizado. En la **Tabla 2-1** se compara los enfoques de PMBOK 6, PRINCE2 e ICB 4.0 en términos de estructura, definición y fases a seguir.

Tabla 2-1: Comparación entre PMBOK, PRINCE 2, y ICB.

Estándar	Estructura	Definición	Fases
PMBOK 6	Dentro de las diez áreas de conocimiento principales.	Aumentar la probabilidad y/o el impacto de los riesgos positivos y reducir la probabilidad y/o el impacto de los riesgos negativos para optimizar las posibilidades de éxito del proyecto.	1) Planificar la gestión de riesgos. 2) Identificar los riesgos. 3) Realizar el análisis cualitativo de riesgos. 4) Realizar el análisis cuantitativo de riesgos. 5) Planificar las respuestas a los riesgos. 6) Implementar las respuestas a los riesgos. 7) Monitorear los riesgos. (Project Management Institute, 2017)
PRINCE2	Uno de los siete temas principales de PRINCE2.	Aplicación sistemática de principios, enfoques y procesos para identificar, evaluar y planificar los riesgos, además de comunicar la gestión de riesgos con los interesados.	1) Identificación de riesgos. 2) Evaluación de probabilidad, impacto y tiempo (evaluación de proximidad). 3) Planificación e implementación de respuestas. 4) Monitoreo y control. (Axelos, 2017)
IBC 4.0	Clasificado dentro de la competencia práctica (relacionada con el proyecto), abordada en los niveles de gestión de proyectos, programas y portafolios.	Identificación, evaluación, planificación de respuestas, implementación y control de riesgos y oportunidades en los proyectos.	No se establecen pasos específicos. (International Project Management Association, 2015)

Fuente: (Khodabakhshian, 2023)

El estándar PMBOK es uno de los más conocidos y aplicados en relación con proyectos de construcción. La **Tabla 2-2**, basada en este estándar, muestra una descripción general de los

procesos de gestión de los riesgos del proyecto. Se presentan como procesos diferenciados con interfaces definidas, sin embargo, en la práctica se superponen e interactúan.

Tabla 2-2: Descripción general de la gestión de los riesgos del proyecto.

Descripción general de la gestión de los riesgos del proyecto			
11.1 Planificar la Gestión de los Riesgos	11.2 Identificar los Riesgos	11.3 Realizar el Análisis cualitativo de Riesgos	11.4 Realizar el Análisis Cuantitativo de Riesgos
<p>Entradas</p> <ul style="list-style-type: none"> - Acta de constitución del proyecto. - Plan para la dirección del proyecto. - Documentos del proyecto. - Factores ambientales de la empresa. - Activos de los procesos de la organización. <p>Herramientas y Técnicas</p> <ul style="list-style-type: none"> - Juicio de expertos. - Análisis de datos. <p>Salidas</p> <ul style="list-style-type: none"> - Plan de gestión de los riesgos. 	<p>Entradas</p> <ul style="list-style-type: none"> - Plan para la dirección del proyecto. - Documentos del proyecto. - Acuerdos. - Documentación de las adquisiciones. - Factores ambientales de la empresa. - Activos de los procesos de la organización. <p>Herramientas y técnicas</p> <ul style="list-style-type: none"> - Juicio de expertos. - Recopilación de datos. - Análisis de datos. - Habilidades interpersonales y de equipo. - Listas rápidas. - Reuniones. <p>Salidas</p> <ul style="list-style-type: none"> - Registro de riesgos. - Informe de riesgos. - Actualizaciones a los documentos del proyecto. 	<p>Entradas</p> <ul style="list-style-type: none"> - Plan para la dirección del proyecto. - Documentos del proyecto. - Factores ambientales de la empresa. - Activos de los procesos de la organización. <p>Herramientas y técnicas</p> <ul style="list-style-type: none"> - Juicio de expertos. - Recopilación de datos. - Análisis de datos. - Habilidades interpersonales y de equipo. - Categorización de riesgos. - Representación de datos. - Reuniones. <p>Salidas</p> <ul style="list-style-type: none"> - Actualizaciones a los documentos del proyecto. 	<p>Entradas</p> <ul style="list-style-type: none"> - Plan para la dirección del proyecto. - Documentos del proyecto. - Factores ambientales de la empresa. - Activos de los procesos de la organización. <p>Herramientas y técnicas</p> <ul style="list-style-type: none"> - Juicio de expertos. - Recopilación de datos. - Habilidades interpersonales y de equipo. - Representaciones de la incertidumbre. - Análisis de datos. <p>Salidas</p> <ul style="list-style-type: none"> - Actualizaciones a los documentos del proyecto.
11.5 Planificar la Respuesta a los Riesgos	11.6 Implementar la Respuesta a los Riesgos	11.7 Monitorear los Riesgos	
<p>Entradas</p> <ul style="list-style-type: none"> - Plan para la dirección del proyecto. - Documentos del proyecto. - Factores ambientales de la empresa. - Activos de los procesos de la organización. <p>Herramientas y técnicas</p> <ul style="list-style-type: none"> - Juicio de expertos. - Recopilación de datos. - Habilidades interpersonales y de equipo. - Estrategias para amenazas. - Estrategias para oportunidades. - Estrategias de respuesta a contingencias. - Estrategias para el riesgo general del proyecto. - Análisis de datos. - Toma de decisiones. <p>Salidas</p> <ul style="list-style-type: none"> - Solicitudes de cambio. - Actualizaciones al plan para la dirección del proyecto. - Actualizaciones a los documentos del proyecto. 	<p>Entradas</p> <ul style="list-style-type: none"> - Plan para la dirección del proyecto. - Documentos del proyecto. - Activos de los procesos de la organización. <p>Herramientas y técnicas</p> <ul style="list-style-type: none"> - Juicio de expertos. - Habilidades interpersonales y de equipo. - Sistema de información para la dirección de proyectos. <p>Salidas</p> <ul style="list-style-type: none"> - Solicitudes de cambio. - Actualizaciones a los documentos del proyecto. 	<p>Entradas</p> <ul style="list-style-type: none"> - Plan para la dirección del proyecto. - Documentos del proyecto. - Datos de desempeño del trabajo. - Informes de desempeño del trabajo. <p>Herramientas y técnicas</p> <ul style="list-style-type: none"> - Análisis de datos. - Auditorías. - Reuniones. <p>Salidas</p> <ul style="list-style-type: none"> - Información de desempeño del trabajo. - Solicitudes de cambio. - Actualizaciones al plan para la dirección del proyecto. - Actualizaciones a los documentos del proyecto. - Actualizaciones a los activos de los procesos de la organización. 	

Fuente: Adaptado de Project Management Institute (2017)

La gestión de riesgos en proyectos de construcción aporta ventajas, entre las que se destaca la disminución de sobrecostos y demoras, al permitir una identificación anticipada de problemas potenciales y su tratamiento oportuno (Kadume & Naji, 2021). Asimismo, favorece el aseguramiento de la calidad del proyecto, al posibilitar la detección temprana de riesgos que podrían comprometer el cumplimiento de los estándares establecidos (Banaitiene et al., 2011). Adicionalmente, promueve una comunicación más efectiva entre las partes interesadas, al garantizar que todos estén al tanto de los riesgos identificados y de las medidas propuestas para su control.

A continuación, se presentan las definiciones de cada uno de los procesos mencionados en la **Tabla 2-2**, basadas en la guía PMBOK 6th del *Project Management Institute*. Esta referencia es reconocida a nivel mundial como un estándar en la gestión de proyectos, siendo adoptada por organizaciones y profesionales para mejorar la planificación, ejecución y control de proyectos.

Planificar la gestión de los riesgos: La planificación de la gestión de riesgos define el enfoque y alcance de las actividades relacionadas con los riesgos dentro del proyecto. Su principal beneficio es asegurar que los esfuerzos de gestión sean proporcionales a la magnitud de los riesgos y a la importancia del proyecto para la organización y sus interesados. Este proceso se realiza una vez o en momentos puntuales durante el ciclo del proyecto.

Identificar los riesgos: La identificación de riesgos implica reconocer y documentar tanto los riesgos individuales como las fuentes de riesgo que puedan impactar el proyecto en general. Su principal beneficio es generar un registro detallado que sirva de base para una respuesta efectiva por parte del equipo. Este proceso se lleva a cabo de forma continua a lo largo del proyecto.

Realizar el análisis cualitativo de riesgos: El análisis cualitativo de riesgos consiste en priorizar los riesgos del proyecto con base en la evaluación de su probabilidad de ocurrencia, impacto y otras características. Su principal ventaja es permitir que los esfuerzos se

concentren en los riesgos más críticos. Este proceso se realiza de manera continua a lo largo del ciclo del proyecto.

Realizar el análisis cuantitativo de riesgos: El análisis cuantitativo de riesgos evalúa numéricamente el efecto combinado de los riesgos identificados y otras incertidumbres sobre los objetivos globales del proyecto. Su principal aporte es ofrecer una estimación objetiva de la exposición total al riesgo, lo que facilita una mejor planificación de las respuestas. Aunque no es necesario en todos los proyectos, cuando se implementa puede aplicarse en distintas etapas a lo largo del ciclo del proyecto.

Planificar la respuesta a los riesgos: La planificación de la respuesta a los riesgos consiste en definir estrategias y acciones específicas para gestionar tanto los riesgos individuales como la exposición global del proyecto. Su principal beneficio es establecer enfoques eficaces para mitigar los riesgos, asignar recursos de forma adecuada y actualizar los documentos del proyecto según se requiera. Este proceso se desarrolla de manera continua a lo largo del proyecto.

Implementar la respuesta a los riesgos: La implementación de la respuesta a los riesgos implica poner en marcha las acciones definidas para gestionar los riesgos del proyecto. Su principal beneficio es asegurar la ejecución efectiva de las estrategias previstas, lo que permite reducir amenazas, aprovechar oportunidades y controlar la exposición global al riesgo. Este proceso se lleva a cabo de forma continua durante todo el ciclo del proyecto.

Monitorear los riesgos: El monitoreo de los riesgos consiste en supervisar la ejecución de las respuestas planificadas, hacer seguimiento a los riesgos existentes, identificar nuevos riesgos y evaluar la eficacia de las acciones implementadas. Su principal beneficio es ofrecer información actualizada que respalde la toma de decisiones en función de la exposición al riesgo, tanto a nivel individual como global del proyecto. Este proceso se mantiene activo durante toda la ejecución del proyecto.

2.2.3 Factores que contribuyen a sobrecostos en proyectos de construcción

La estimación del costo de finalización en proyectos de construcción es un proceso dinámico que requiere actualizaciones continuas a lo largo de la ejecución. Sin embargo, los sobrecostos suelen ser inevitables. Estos riesgos, además de ser variables en el tiempo, presentan interdependencias entre sí, por lo que su evaluación inadecuada constituye una de las principales causas de desviaciones presupuestarias. Por lo tanto, una evaluación precisa y periódica de los costos de finalización es esencial para garantizar un control financiero efectivo y mitigar el impacto de los riesgos asociados.

El estudio realizado por Ashtari et al. (2022) cuyo objetivo fue predecir los sobrecostos y analizar los factores de riesgo asociados mediante un modelo de aprendizaje automático interpretable, capaz de considerar las relaciones entre los distintos riesgos, se basó en las respuestas de expertos y profesionales involucrados en proyectos de construcción gubernamentales en la provincia de Zanján, Irán. A partir de esta información, se identificaron los factores de riesgo vinculados a los sobrecostos, los cuales se resumen en la **Tabla 2-3**.

Tabla 2-3: Especificaciones de los riesgos identificados.

Gestión	
Estudio de viabilidad deficiente	Conflicto entre las partes del proyecto
Debilidad en la gestión del contratista	Debilidad en la gestión del consultor
Comunicación deficiente entre las partes	Propietario incapaz de gestionar el proyecto
Materiales y equipos	
Aumento del precio de los materiales	Escasez de los materiales
Escasez de los equipos	Nuevos equipos/problemas tecnológicos
Retraso de los proveedores en la entrega de equipos en la obra	
Mano de obra	
Falta de conocimientos y experiencia	Escasez de mano de obra
Falta de personal cualificado (personal técnico) in situ	
Financieros	
Tipo de cambio	Múltiples fuentes de fondos

Inflación	Escasez de fondos del contratista
Escasez de fondos del propietario y retrasos en los pagos	
Proyecto	
Cambio adverso de las condiciones geológicas	Complejidad del proyecto
Limitaciones del emplazamiento	
Propietario	
Disponibilidad del emplazamiento	Retrasos en la adquisición de terrenos
Ordenes de cambio durante la construcción	Suministro de servicios públicos
Retrasos en la toma de decisiones	Selección del mejor licitador
Política aduanera del propietario y complejidad (retraso en la contratación)	
Contratista	
Falta de conocimientos y experiencia	Seguridad en las obras
Retrasos en las adquisiciones	Calidad de la construcción (defectos)
Retrasos de los subcontratistas en las obras anteriores	Planificación y programación deficientes
Gestión financiera inadecuada	
Consultor	
Falta de conocimientos y experiencia	Retrasos en la entrega del diseño
Diseño inadecuado/errores de diseño	Cambio de equipo, o especificación del equipo, durante la construcción
Medio ambiente	
Mal tiempo o situación de emergencia	Ley de preservación del medio ambiente
Siniestros/lesiones inesperadas	

Fuente: Adaptado de Ashtari et al. (2022)

2.2.4 Métodos determinísticos

Una lista de técnicas de *Machine Learning* (ML) aplicadas en disciplinas relacionadas con la construcción incluye *Artificial Neural Networks* (ANNs), *Decision Trees* (DT), *Logistic Regression* (LR), *Naive Bayes Models* (NB) y *Support Vector Machines* (SVM). ML combina métodos de estadística, análisis de bases de datos, minería de datos, reconocimiento de patrones e inteligencia artificial para extraer tendencias, interrelaciones, patrones de interés y conocimientos útiles a partir de conjuntos de datos complejos (Khodabakhshian et al., 2023).

Un enfoque determinista es utilizado por gran parte de los algoritmos de ML. Estos algoritmos pueden emplearse en una de las siguientes aplicaciones en la gestión de riesgos:

- Regresión para predecir resultados numéricos continuos, como el retraso causado por un riesgo, incluyendo LR, DT, SVM y *Neural Network* (NN) *techniques*.
- Clasificación para determinar la clase de salida basada en algunas características de entrada, como la identificación de riesgos, incluyendo NN, *Random Forest* (RF), SVM y *Genetic Algorithm* (GA).
- Agrupamiento para explorar datos en busca de agrupaciones naturales, como la identificación de eventos relacionados que causan un riesgo, incluyendo *K-means* y SVM.
- Evaluación de atributos para clasificarlos según su relación con la variable objetivo, como la identificación de las causas más significativas de accidentes, incluyendo DT, RF.
- Detección de anomalías para identificar casos inusuales basados en desviaciones, como la identificación de riesgos de accidentes, incluyendo SVM y *Deep Neural Networks* (DNNs).

A diferencia de otros ámbitos en la construcción, las aplicaciones de ML han sido limitadas y se han centrado principalmente en predecir riesgos de retraso, el impacto de cambios contractuales en el tiempo y calidad del desempeño, y el análisis y modelado de bases de datos de incidentes para la predicción de riesgos en salud y seguridad ocupacional (Khodabakhshian et al., 2023).

Las redes *Artificial Neural Networks* (ANNs) son el método de ML más aplicado en la evaluación de riesgos en ingeniería, seguidas por SVM, DT, RF, NB, *K-means* y LR (Hegde & Rokseth, 2020).

2.2.5 Métodos probabilísticos

El enfoque probabilístico se emplea en *Structural Equation Modeling* (SEM), *Bayesian Networks* (BNs), *Fuzzy Logic* y *Fuzzy Cognitive Maps* (FCMs), que pueden integrarse con otros métodos (Khodabakhshian et al., 2023). El SEM es una técnica estadística multivariante versátil que representa relaciones estructurales causales entre múltiples variables mediante un diagrama esquemático. Tiene aplicación en el análisis de riesgos en seguridad en la construcción, con el *Exploratory Factor Analysis* (EFA), que permite descubrir la estructura subyacente de un gran conjunto de variables cuando no existen hipótesis previas sobre su organización (Liu et al., 2018).

Por otro lado, las BNs son el modelo gráfico probabilístico más aplicado en la industria de la construcción (Hon et al., 2022). Se basan en la teoría de la probabilidad y la teoría de grafos para representar las relaciones causales entre variables y sus probabilidades en redes de riesgo. Estas redes se presentan como grafos compuestos por nodos (variables aleatorias) y arcos dirigidos (relaciones causales entre variables), lo que se conoce como modelo de *Directed Acyclic Graph* (DAG). Además, incluyen una *Conditional Probability Distribution* (CPD) para variables continuas o una *Conditional Probability Table* (CPT) para variables categóricas, representando las influencias entre los nodos. Los parámetros de la CPD o CPT pueden aprenderse mediante algoritmos a partir de datos históricos extensivos, opinión de expertos, o ambos (Borujeni et al., 2021).

Las BNs son aplicadas además en la modelación, identificación y análisis de riesgos en proyectos de construcción, como reclamaciones y riesgos contractuales, salud estructural, calidad de operaciones, sobrecostos y retrasos en cronogramas, así como riesgos de seguridad (Borujeni et al., 2021).

En cuanto a la *Fuzzy Logic* tiene aplicación en la modelación de datos cualitativos y subjetivos obtenidos de expertos, permitiendo razonar con información ambigua. Expresiones verbales de probabilidad se transforman en números difusos, con grados de veracidad o falsedad representados por un rango de valores entre 1 (verdadero) y 0 (falso),

utilizando funciones de membresía triangulares, trapezoidales o gaussianas. Un mapa cognitivo difuso combina lógica difusa y mapa cognitivo, utilizando variables lingüísticas subjetivas y vagas de expertos del dominio para realizar un análisis de causa raíz. Este modelo representa sistemas complejos y dinámicos con numerosos indicadores, dependencias causales y pesos. Además, permite realizar análisis de escenarios "what-if" para la predicción y evaluación de riesgos en un modelo gráfico ponderado difuso, con tolerancia a la imprecisión y la incertidumbre (Fayek, 2020).

3. Estado del arte

3.1 Revisión sistemática de literatura

El presente estudio utiliza un enfoque de revisión sistemática de literatura (SLR, por sus siglas en inglés) sobre bases de datos científicas, con varios métodos de análisis bibliométricos para encontrar interrelaciones entre la gestión de riesgos, la inteligencia artificial, con énfasis en *Machine Learning*, y la gestión de proyectos de construcción de edificación.

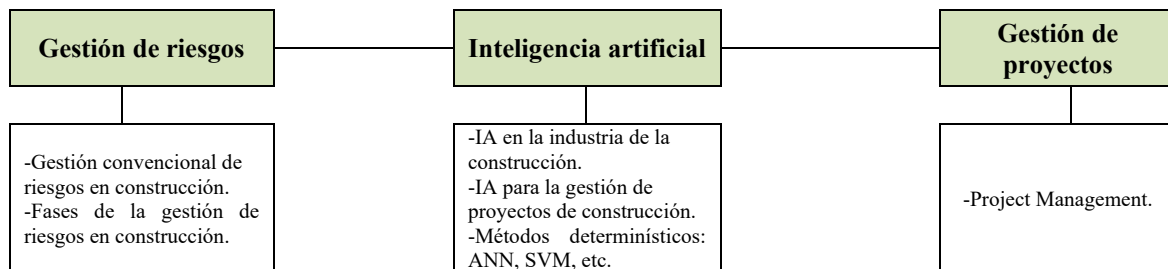
Una SLR es un enfoque riguroso y metódico para revisar la investigación existente sobre un tema específico, diseñado para proporcionar un resumen integral, transparente y reproducible de la evidencia disponible. Višić (2022) sugiere que sus componentes clave incluyen:

- Formulación de una pregunta de investigación clara y concreta, que guía todo el proceso de revisión.
- Desarrollo de un protocolo detallado que describa los métodos y criterios para la revisión, garantizando transparencia y reproducibilidad.
- Búsqueda bibliográfica exhaustiva en bases de datos relevantes para identificar todos los estudios potenciales, incluyendo trabajos publicados y no publicados.
- Selección y cribado de estudios aplicando criterios de inclusión y exclusión para asegurar su relevancia con la pregunta de investigación.

- Evaluación de la calidad de los estudios seleccionados, para garantizar la fiabilidad de los hallazgos.
- Extracción y síntesis de datos relevantes de los estudios seleccionados, empleando métodos cuantitativos (metaanálisis) o cualitativos (metasíntesis).
- Elaboración de un informe detallado que documente la metodología, los hallazgos y las conclusiones de la revisión, asegurando su transparencia y reproducibilidad.

En la **Figura 3-1** se presentan los dominios de investigación definidos para la generación de la ecuación de búsqueda, seleccionados por su relevancia para abordar el tema de manera integral. Este enfoque no solo considera los sobrecostos, sino que también incorpora herramientas y metodologías que permiten una comprensión más amplia y estratégica del problema en el contexto de la industria de la construcción .

Figura 3-1: Dominios para ecuación de búsqueda.



Fuente: Propia.

En esta revisión, se seleccionó la base de datos *Scopus* debido a sus avanzadas capacidades de búsqueda, su carácter interdisciplinario y su reconocimiento como fuente confiable de publicaciones. Para llevar a cabo la SLR se emplearon las directrices PRISMA (*Preferred Reporting Items for Systematic Reviews and Meta-Analysis*), estas directrices se publicaron por primera vez en 2009 para mejorar la transparencia y la integridad de los informes en revisiones sistemáticas y metaanálisis. Estas pautas ayudan a los autores a informar por qué se realizó la revisión , que se hizo y que se encontró, mejorando así la calidad y confiabilidad de las revisiones sistemáticas y de los ensayos clínicos (Page et al., 2021). A continuación,

se emplea un diagrama de flujo de 4 fases que consiste en (a) identificación, (b) selección, (c) elegibilidad, y (d) inclusión. Además, se presenta de forma esquemática el proceso en la **Figura 3-2**.

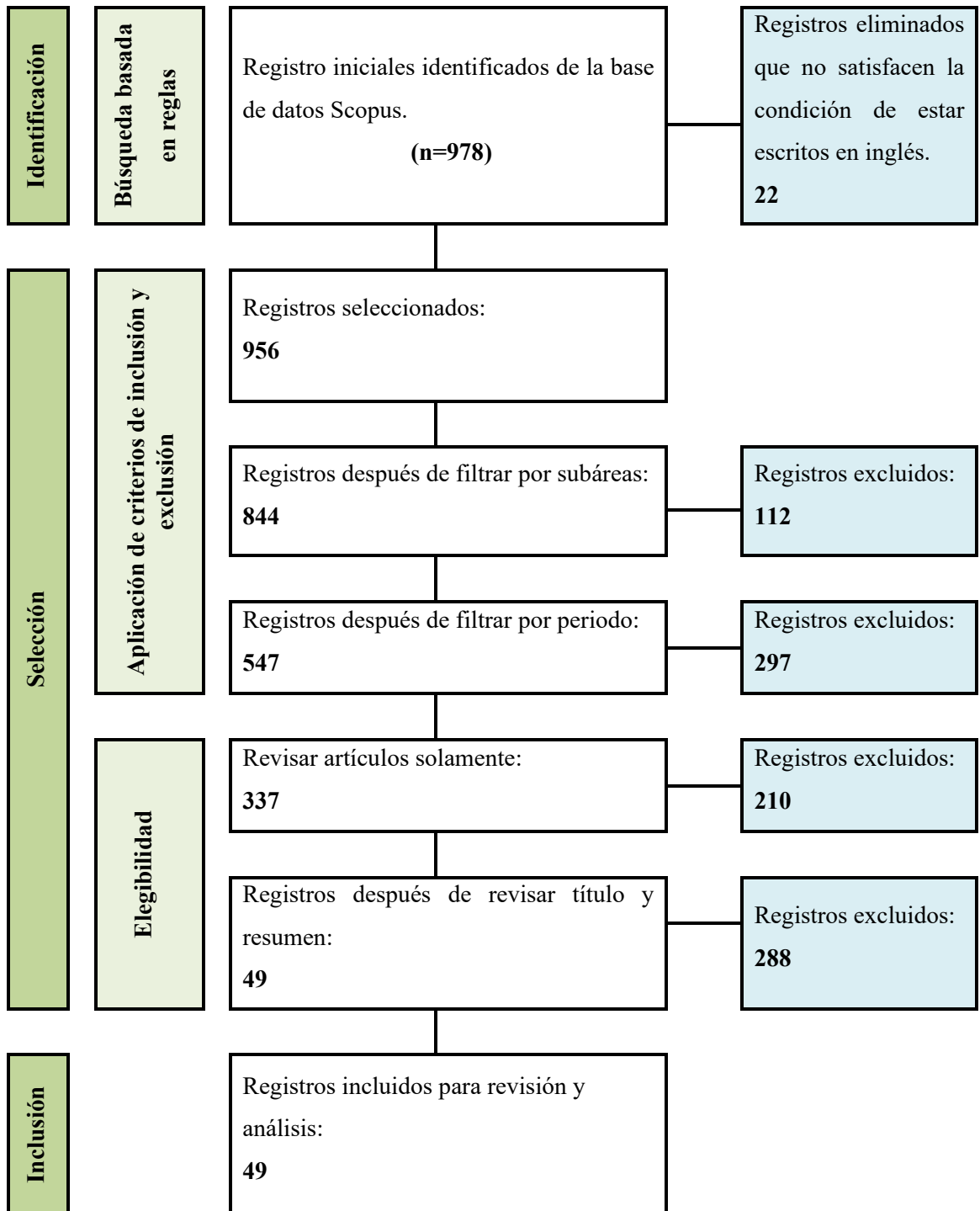
- a) En la fase de **identificación**, se emplea la ecuación de búsqueda, **Ecuación 3-1**, obteniendo un número significativo de registros que cumplen con los criterios establecidos. Se consideran únicamente los registros con escritura en inglés, por lo que se excluye un total de 22.
- b) En la fase de **selección**, se aplican los siguientes criterios: en primer lugar, la búsqueda se limita a subáreas de ingeniería, ciencias de la computación y negocios y administración, dado el interés en desarrollar un modelo que facilite la toma de decisiones a nivel organizacional. En segundo lugar, se considera un periodo de cinco años debido a la creciente relevancia de la inteligencia artificial en los últimos años, especialmente en su integración con la digitalización en la industria de la construcción, un campo en constante avance. Tras aplicar estos criterios, se obtienen 547 registros, de los cuales los tres tipos más comunes son artículos científicos (337), artículos de conferencias (140) y revisiones (35). Dado que el objetivo es analizar soluciones a problemas de investigación previamente planteados, el enfoque se centrará en los artículos, los cuales servirán como referencia para la evaluación y validación del modelo propuesto.
- c) En la fase de **elegibilidad**, como se mencionó anteriormente, se seleccionan únicamente los registros correspondientes a artículos. Posteriormente, se revisan manualmente los títulos y resúmenes con el fin de identificar aquellos que aborden específicamente la gestión de riesgos en proyectos de construcción, con énfasis en sobrecostos, retrasos y disputas contractuales.
- d) En la fase de **inclusión**, estrechamente ligada a la de elegibilidad, se seleccionan un total de 49 artículos para revisión detallada. Cabe destacar que una parte significativa

de los artículos evaluados en la fase anterior se enfoca en riesgos cibernéticos, seguridad y salud ocupacional, mientras que otros abordan aspectos técnicos, como la resistencia del concreto y el acero, lo cual no es interés para este trabajo.

Ecuación 3-1: ecuación de búsqueda

("risks" OR "Risk management" OR "risk analysis" OR "risk assessment") AND ("Artificial intelligence" OR intelligence OR "AI" OR "machine learning" OR "fuzzy logic" OR "expert system" OR "data mining" OR "Deep learning" OR "Natural language processing" OR "NLP") AND ("Construction industry" OR "construction project" OR "construction project management" OR "construction risk management" OR "architectural, engineering, and construction industry" OR "AEC").

Figura 3-2: Flujo basado en PRISMA para revisión sistemática de literatura.



Fuente: Propia.

3.2 Análisis bibliométrico

El análisis bibliométrico es un método cuantitativo utilizado para evaluar y analizar la literatura científica mediante técnicas estadísticas y matemáticas aplicadas a datos bibliográficos, como libros y artículos. Su objetivo es identificar patrones, tendencias e influencias dentro de un campo de estudio específico (Donthu et al., 2021). A través de este enfoque, se exploran grandes volúmenes de datos científicos para determinar el estado del arte en un área de investigación. Además, permite la representación gráfica de datos para identificar autores, instituciones y países influyentes, utilizando herramientas como *VosViewer* (Obregón et al., 2019). También posibilita el rastreo de la evolución de un campo temático, analizando cómo ha cambiado la producción de investigación a lo largo del tiempo y en distintas regiones (González Alcaide et al., 2011).

Otro aspecto es que el análisis bibliométrico permite identificar vacíos en el conocimiento y sugerir oportunidades para futuras investigaciones mediante el análisis de citas y tendencias de publicación. Asimismo, juega un papel clave en la planificación estratégica y la toma de decisiones en políticas de financiamiento e investigación, facilitando la gestión y asignación de recursos (Hasan & Singh, 2015). Para su implementación, se utilizan técnicas como el análisis de citas, que examina la frecuencia y patrones de citación para medir la relevancia e impacto de los trabajos científicos, el estudio de redes de coautoría y colaboración, que analiza las relaciones entre autores, instituciones y países para comprender la estructura de las comunidades científicas, y la coocurrencia de palabras clave, que permite identificar temas y tendencias mediante el análisis de frecuencia y coocurrencia de términos clave (Ravindran & Deepak, 2023).

En la **Figura 3-3** se presenta una red de co-ocurrencia de palabras clave en los 49 artículos seleccionados, donde los nodos más grandes corresponden a términos con mayor frecuencia de aparición y conexión con otros conceptos.

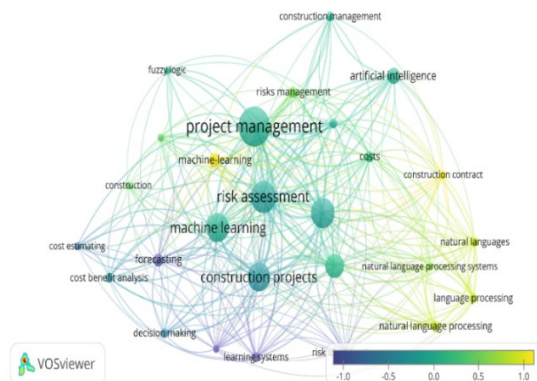
Los colores indican diferentes agrupaciones temáticas dentro del análisis. Se identifican tres grupos: el primero, en rojo, se centra en la gestión de proyectos y riesgos en construcción,

según la escala en la parte inferior. Los valores cercanos a -1 (color azul oscuro y morado) representan términos que fueron más frecuentes alrededor de 2020, mientras que los valores cercanos a 1 (color amarillo) indican términos emergentes y tendencias hacia 2025.

Se observa que términos como *Forecasting*, *Decision Making* y *Learning Systems* aparecen en la región azul y morada, esto indica que han sido áreas de enfoque en años anteriores, posiblemente relacionadas con la aplicación temprana de modelos de predicción y toma de decisiones en la gestión de proyectos. En contraste, términos como *Construction Contract*, *Natural Languages*, *Natural Language Processing* y *Language Processing* aparecen en la región amarilla, esto sugiere un creciente interés en la aplicación de procesamiento de lenguaje natural (PLN) en el sector de la construcción en los próximos años. Los términos centrales como *Project Management*, *Risk Assessment*, *Machine Learning* y *Construction Projects* aparecen en una tonalidad intermedia (verde), esto sugiere que han mantenido relevancia a lo largo del período analizado.

Sin embargo, el hecho de que *Machine Learning* y *Artificial Intelligence* estén en una zona ligeramente más verde puede indicar una evolución continua en su aplicación dentro de la gestión de riesgos y proyectos de construcción.

Figura 3-4: Diagrama de Co-ocurrencia de palabras clave y evolución en el tiempo.

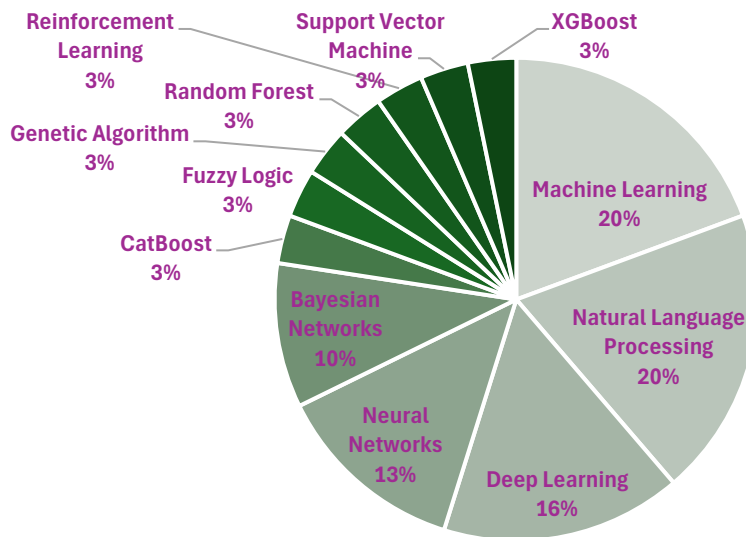


En general, la evolución muestra una transición desde modelos más tradicionales de análisis y predicción (2020) hacia la integración de inteligencia artificial y procesamiento de lenguaje natural en la construcción (2025).

3.3 Análisis estadístico

La **Figura 3-5** muestra la distribución de diferentes enfoques y técnicas dentro del aprendizaje automático y la inteligencia artificial, revelando patrones sobre la popularidad y aplicabilidad de cada metodología.

Figura 3-5: Frecuencia de algoritmos usados en la gestión de riesgos en construcción.



Fuente: Propia.

Las categorías más dominantes son *Machine Learning* y *Natural Language Processing*, con un 20% cada una, dando a entender que estos son los campos más explorados o utilizados en el contexto del análisis. La presencia significativa de NLP sugiere una creciente importancia en la interpretación, generación y análisis de texto en tareas como minería de datos, automatización de procesos y toma de decisiones basada en lenguaje natural.

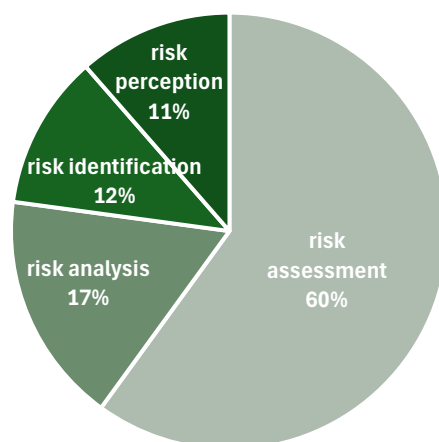
El *Deep Learning* (16%) y las *Neural Networks* (13%) tienen una presencia importante. La diferencia entre *Deep Learning* y *Neural Networks* puede interpretarse como que el primero incluye técnicas más avanzadas y profundas, mientras que las redes neuronales abarcan un espectro más amplio, incluyendo modelos menos profundos o especializados en tareas particulares. Las Redes Bayesianas (10%) tienen una representación notoria, esto puede indicar que los métodos probabilísticos siguen siendo empleados para la modelación de

incertidumbre, inferencia causal y toma de decisiones en entornos complejos. Por otro lado, las técnicas con menor representación (todas con un 3%) incluyen *CatBoost*, *Fuzzy Logic*, *Genetic Algorithm*, *Random Forest*, *Reinforcement Learning*, *Support Vector Machine* y *XGBoost*.

Del análisis anterior se puede concluir que las redes neuronales profundas y el procesamiento de lenguaje natural son dominantes, además, las metodologías probabilísticas y técnicas clásicas siguen estando vigentes. La menor representación de ciertos enfoques podría deberse a su aplicabilidad en nichos más especializados o a que han sido reemplazados progresivamente por modelos más avanzados.

Continuando con el análisis estadístico, en la **Figura 3-6** se presenta la distribución de la frecuencia con la que se mencionan distintas fases de la gestión de riesgos en los 49 artículos estudiados. La etapa predominante es *Risk Assessment*, que representa el 60% de las menciones, la interpretación que podría darse es que la mayoría de los estudios revisados se centran en la evaluación general del riesgo, probablemente abordando su impacto y probabilidad de ocurrencia dentro de distintos contextos.

Figura 3-6: Fases de gestión de riesgos cubierta en las fuentes analizadas.



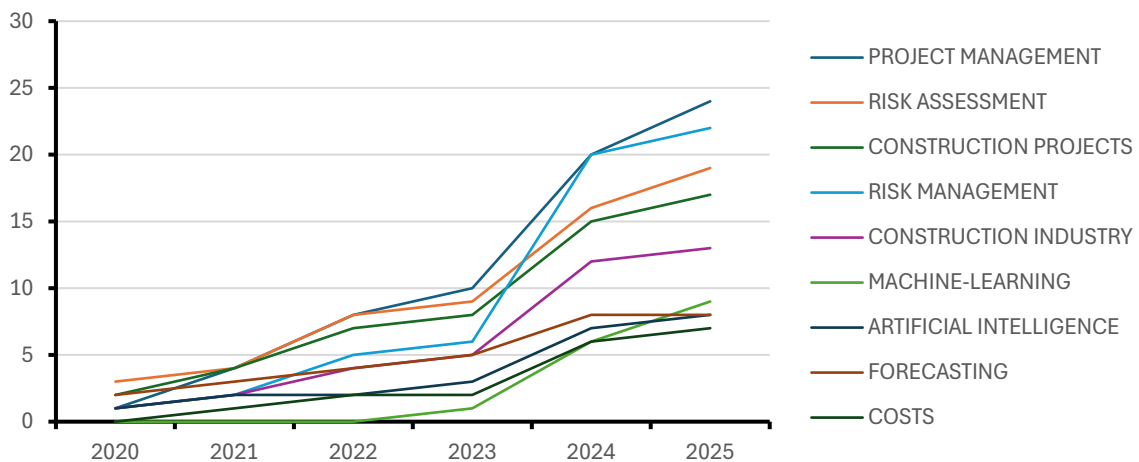
Risk Analysis ocupa el segundo lugar con un 17%, lo que indica que una parte de la literatura se enfoca en técnicas analíticas para cuantificar y modelar los riesgos, utilizando enfoques como simulaciones, modelos matemáticos o inteligencia artificial.

Risk Identification aparece con un 12%, aunque es una fase importante dentro de la gestión de riesgos, recibe menos atención en comparación con la evaluación y el análisis. Esto podría deberse a que la identificación es vista como una etapa preliminar, mientras que la evaluación y el análisis permiten profundizar en el impacto y la toma de decisiones.

Risk Perception tiene la menor representación con un 11%, lo que sugiere que la subjetividad y los factores humanos en la gestión de riesgos son menos explorados en la literatura revisada. Sin embargo, en ciertos sectores, la percepción del riesgo juega un papel clave en la toma de decisiones, por lo que este menor enfoque podría ser una oportunidad de investigación en el futuro.

Finalmente, en la **Figura 3-7** se presenta un gráfico de líneas, el cual representa la evolución en la frecuencia de ciertas palabras clave a lo largo del tiempo en la revisión de 49 artículos. Permite identificar tendencias en la investigación y los temas que han ganado más relevancia en los últimos años.

Figura 3-7: Evolución de los términos clave en la fuente analizada.



Fuente: Propia.

"*Project Management*" y "*Risk Management*" son los términos con mayor crecimiento, especialmente después de 2023, dando a entender que la literatura ha incrementado su enfoque en la gestión de proyectos y riesgos, y que existe un interés creciente en metodologías de administración eficiente. "*Risk Assessment*" y "*Construction Projects*" han

mostrado un crecimiento sostenido, lo que sugiere que la evaluación de riesgos y su aplicación en proyectos de construcción han sido de interés constante. *"Construction Industry"* ha tenido un crecimiento más moderado en comparación con los otros términos, pero aún muestra una tendencia positiva.

"Machine Learning" y *"Artificial Intelligence"* muestran un crecimiento paulatino desde 2021 y un aumento más notable en 2024-2025. Esto indica que los enfoques basados en IA están ganando relevancia en la gestión de proyectos y riesgos, posiblemente para mejorar la toma de decisiones y la predicción de problemas.

"Forecasting" y *"Costs"* han aumentado su frecuencia, aunque a un ritmo más lento. Esto sugiere que la investigación ha explorado métodos de pronóstico y análisis de costos, pero sin el mismo crecimiento expandido que la gestión de proyectos y riesgos.

3.4 Hallazgos y discusión

Los hallazgos encontrados en las publicaciones en cuanto a la implementación de técnicas de IA para la gestión de riesgos en proyectos de construcción muestran lo siguiente:

El estudio de Turkyilmaz & Polat (2024) propone un modelo de estimación de costos utilizando ML, que predice sobrecostos según las fluctuaciones en las puntuaciones de riesgo. Se validaron seis algoritmos de clasificación, destacando el rendimiento superior del clasificador de árboles de decisión. Esta investigación ayuda a mejorar las capacidades de planificación en la fase de ejecución al proporcionar un sistema de estimación rápida.

Yi & Luo (2024) proponen un modelo de estimación y control de costos impulsado por IA (AI-CCECA) que emplea redes neuronales profundas para analizar el costo preliminar y la gestión dinámica del sistema de control. El uso de IA y ML mejora la eficiencia operativa y la competitividad en la industria de la construcción, permitiendo optimizar la estimación de costos en fases tempranas de los proyectos.

Ashtari et al. (2022) propusieron un modelo de red bayesiana (BN) que considera las interacciones entre los riesgos de sobrecosto para mejorar las predicciones de costos en proyectos de construcción. El modelo BN superó a otros enfoques tradicionales (Naive Bayes y árbol de decisiones) con una precisión del 78,86%, destacando factores críticos como el costo de materiales y la falta de experiencia del personal.

Nyqvist et al. (2024) comparan el modelo GPT-4 de ChatGPT con expertos humanos en la gestión de riesgos de proyectos de construcción. Aunque la IA superó a los humanos en la generación de planes integrales de gestión de riesgos, se identificaron limitaciones en la aplicabilidad práctica de las estrategias. Se resalta el potencial sinérgico de la IA y la experiencia humana, sugiriendo un modelo colaborativo.

George et al. (2022) desarrollaron un modelo predictivo utilizando técnicas de ML para evaluar el estado de seguridad en los sitios de construcción, mejorando la prevención de accidentes. El análisis de un conjunto de datos de la Administración de Seguridad y Salud Ocupacional mostró que el modelo “*Gradient Boosting*”, entrenado con un conjunto de atributos, logró los mejores resultados en términos de predicción.

Moon et al. (2022) utilizaron BERT, una técnica de procesamiento de lenguaje natural, para clasificar cláusulas contractuales en proyectos de construcción. El modelo alcanzó una precisión de 0.889, mejorando la gestión de riesgos contractuales y facilitando la revisión de especificaciones en proyectos de construcción.

Cheng & Darsa (2021) desarrollaron el modelo CSRAM para predecir retrasos en los cronogramas de proyectos de construcción en Etiopía. Utilizando redes neuronales artificiales (ANN), identificaron factores de riesgo como la mala gestión de recursos y la corrupción, proponiendo estrategias para mitigar los retrasos significativos en los proyectos.

Sanni-Anibire et al. (2021) propusieron un marco basado en IA para mitigar los retrasos en proyectos de construcción, utilizando técnicas de ML como redes neuronales y máquinas de

soporte vectorial. El marco se validó mediante entrevistas con expertos y mostró ser útil para abordar problemas de retraso en proyectos a través de la estimación de costos, duración y evaluación de riesgos.

Yaseen et al. (2020) desarrollaron un modelo híbrido RF-GA para predecir problemas de retraso en proyectos de construcción. El modelo alcanzó una precisión del 91,67%, demostrando ser eficaz en el monitoreo y la sostenibilidad de la gestión de proyectos.

Okudan et al. (2021) desarrollaron una herramienta basada en razonamiento de casos (CBRisk) para la gestión de riesgos en la construcción. Esta herramienta, validada mediante pruebas de caja negra y revisión de expertos, mejora la efectividad de la gestión de riesgos al integrar diversas fases del proceso y utilizar un enfoque basado en el conocimiento.

Chattapadhyay et al. (2021) crearon un sistema de predicción de riesgos para megaproyectos utilizando ML y un algoritmo de agrupamiento GA-K-means. El modelo identificó factores de alto riesgo relacionados con costos, tiempo, calidad y alcance, mejorando la gestión de riesgos en megaproyectos de construcción.

Ahora bien, de los estudios analizados, aquellos que se centran específicamente en el sobrecosto en proyectos de construcción de edificación se presentan en la **Tabla 3-1**. Ahí se detallan los autores, las técnicas de modelado utilizadas y las métricas de evaluación aplicadas en cada investigación. Este resumen proporciona una visión estructurada y comparativa de los enfoques existentes. Además, permite identificar tendencias, fortalezas y oportunidades de mejora en las metodologías actuales.

Tabla 3-1: Técnicas de referencia reportadas en la fuente analizada.

Referencia	ForouzeshNejad et al., 2024
Técnica Usada	Técnica híbrida que combina el algoritmo <i>eXtreme Gradient Boosting</i> (XGBoost) con el algoritmo de <i>recocido simulado</i> (Simulated Annealing, SA) para predecir el tiempo y costo en proyectos de construcción. Esta combinación busca mejorar la precisión de las

	predicciones al considerar la complejidad de las redes de actividades y las incertidumbres inherentes en los entornos de proyectos.
Métricas de evaluación	El modelo híbrido <i>XGBoost-SA</i> logró una precisión de predicción del 92%. Además, los resultados mostraron que este modelo redujo el error de predicción de costos en casi un 50% y el error de predicción de tiempo en aproximadamente un 80% en comparación con métodos tradicionales como la Gestión del Valor Ganado (EVM) y el Método de Programación Ganada (ESM).

Referencia	(Turkyilmaz & Polat, 2024)
Técnica Usada	Se implementaron y evaluaron seis algoritmos de clasificación para predecir las clases de tasa de sobrecostos en función de las puntuaciones de riesgo totales en cualquier momento del proyecto. Los seis algoritmos de clasificación empleados en el estudio fueron: <ul style="list-style-type: none"> • Árbol de Decisión (Decision Tree) • Bosque Aleatorio (Random Forest) • Máquina de Soporte Vectorial (Support Vector Machine, SVM) • k-Vecinos Más Cercanos (k-Nearest Neighbors, k-NN) • Regresión Logística • Naive Bayes
Métricas de evaluación	Para evaluar el rendimiento de estos algoritmos, se utilizaron las siguientes métricas: <ul style="list-style-type: none"> • Precisión (Accuracy) • Precisión (Precision) • Sensibilidad o Recall (Recall) • Puntuación F1 (F1-Score) Entre los algoritmos evaluados, el Árbol de Decisión (<i>Decision Tree</i>) demostró el mejor desempeño en términos de las métricas mencionadas.

Referencia	Al-Nahhas et al., 2024
Técnica Usada	Sistema de Inferencia Difusa de Mamdani: Este enfoque utiliza lógica difusa para manejar la incertidumbre y la imprecisión en la evaluación de factores que contribuyen a los sobrecostos.
Métricas de evaluación	El estudio no proporciona métricas de evaluación específicas como precisión, exactitud o puntuación F1. En su lugar, la validación del modelo se llevó a cabo mediante un análisis de sensibilidad, comparando los sobrecostos predichos con los reales en dos proyectos completados en Arabia Saudita.

Referencia	(Ashtari et al., 2022)
-------------------	------------------------

Técnica Usada	<p>Redes Bayesianas (BN): Se implementó un modelo de clasificación que considera las posibles interacciones entre los factores de riesgo de sobrecostos.</p> <p>Naive Bayes (NB): Se utilizó este modelo para comparar su desempeño con el de la red bayesiana, asumiendo independencia entre los factores de riesgo.</p> <p>Árbol de Decisión (DT): Este algoritmo también se empleó para comparar su rendimiento con los modelos anteriores.</p>
Métricas de evaluación	<p>Precisión (Accuracy): Se evaluó la capacidad de cada modelo para predecir correctamente los riesgos de sobrecostos. Los modelos de redes bayesianas alcanzaron una precisión promedio del 78.86%, superando significativamente a los modelos NB y DT.</p>

Referencia	Tayefeh Hashemi et al., 2020
Técnica Usada	<p>El estudio identifica y analiza diversas técnicas de aprendizaje automático empleadas en la estimación de costos, entre las que destacan:</p> <ul style="list-style-type: none"> • Redes Neuronales Artificiales (ANN). • Máquinas de Soporte Vectorial (SVM). • Algoritmos de Árboles de Decisión y Bosques Aleatorios. • Redes Bayesianas.
Métricas de evaluación	<p>Para evaluar el rendimiento de los modelos de aprendizaje automático en la estimación de costos, el artículo destaca el uso de diversas métricas, entre las que se incluyen:</p> <ul style="list-style-type: none"> • Error Absoluto Medio (MAE) • Error Cuadrático Medio (MSE) • Raíz del Error Cuadrático Medio (RMSE) • Coeficiente de Determinación (R^2)

Fuente: Propia.

Los estudios revisados demuestran que no existe un único algoritmo óptimo para la predicción de desviaciones presupuestarias o sobrecostos en proyectos de construcción de edificación, sino que la elección del modelo depende de la naturaleza de los datos y del objetivo del análisis.

En cuanto a las métricas de evaluación, los estudios emplean un enfoque integral que incluye métricas de clasificación y regresión, permitiendo una evaluación más completa del desempeño de los modelos.

4. Metodología

El desarrollo de este estudio se realiza a partir de CRISP-DM (Cross-Industry Standard Process for Data Mining), una metodología estructurada para proyectos de minería de datos que garantiza un desarrollo coherente desde la definición del problema hasta la implementación de la solución.

Según Goh et al. (2017) las fases de la metodología pueden describirse de la siguiente manera:

Comprensión del negocio: en esta etapa se identifican los objetivos y requisitos desde una perspectiva empresarial, traduciéndolos en un problema de minería de datos y definiendo un plan preliminar de trabajo.

Comprensión de los datos: consiste en recopilar, explorar y evaluar la calidad de la información, con el fin de detectar patrones iniciales y posibles inconsistencias.

Preparación de los datos: implica seleccionar, limpiar y transformar los datos para garantizar su formato e idoneidad antes del modelado.

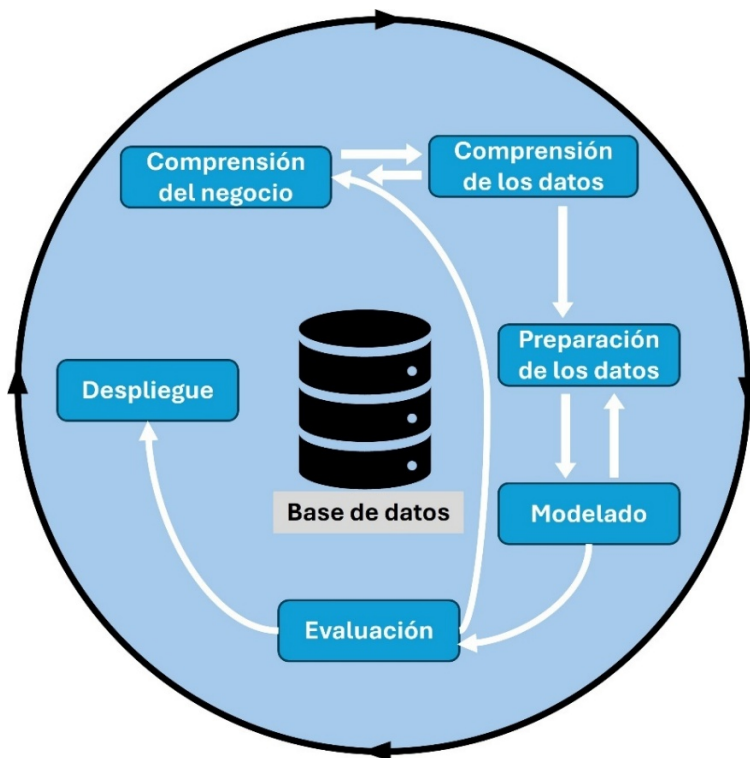
Modelado: se aplican diversas técnicas analíticas y se ajustan sus parámetros para encontrar la solución más adecuada al problema planteado.

Evaluación: se validan los modelos generados, verificando que satisfagan los objetivos del negocio y sean suficientemente confiables para su uso.

Despliegue: en esta fase se implementan los resultados en el entorno empresarial, ya sea mediante informes, integración en sistemas operativos o automatización de procesos, asegurando que el conocimiento obtenido aporte valor real a la organización.

La **Figura 4-1** presenta las fases de CRISP-DM, en el centro se observa una base de datos, la cual simboliza el papel central que juegan los datos en este proceso. Toda la metodología gira en torno a la recopilación, análisis y transformación de la información para extraer conocimientos útiles y aplicables a la toma de decisiones. Alrededor de esta base de datos se organizan las seis fases del modelo en una estructura circular con flechas interconectadas. CRISP-DM no es un proceso estrictamente lineal, sino más bien iterativo y adaptable. Esto significa que, a medida que se avanza en el análisis, pueden surgir nuevos hallazgos que requieran volver a fases anteriores para refinar los datos, ajustar modelos o redefinir objetivos del negocio.

Figura 4-1: Marco conceptual para cálculo de sobrecostos basado en CRISP-DM.



Fuente: Adaptado de *Cross-Industry Standard Process for Data Mining*, (IBM) (2021)

4.1 Comprensión del negocio

La primera etapa de la metodología CRISP-DM, tiene como objetivo entender el problema desde una perspectiva organizacional, identificar los objetivos clave del proyecto y

establecerlos en una formulación clara para su análisis mediante técnicas de ciencia de datos. Esta etapa permite alinear los enfoques analíticos con las necesidades reales del negocio (Goh et al., 2017).

En este trabajo, se presenta la necesidad de abordar un desafío del sector de la construcción: la gestión de riesgos asociados a los sobrecostos en proyectos de construcción de edificación. Tradicionalmente, y como se ha reiterado en secciones anteriores, su gestión se basa en la experiencia y el juicio de los profesionales, lo que limita la capacidad de anticiparse a los riesgos.

Dado lo anterior, este trabajo busca traducir esa necesidad del negocio en un problema de análisis de datos, aprovechando el potencial del aprendizaje automático.

4.1.1 Problema y su contexto

Como se expuso en la sección **1.1.Planteamiento del problema**, las desviaciones presupuestarias en proyectos de construcción de edificación representan un desafío dentro de la gestión de riesgos, especialmente debido a la complejidad e incertidumbre inherente al sector. Esta sección profundiza en dicho problema.

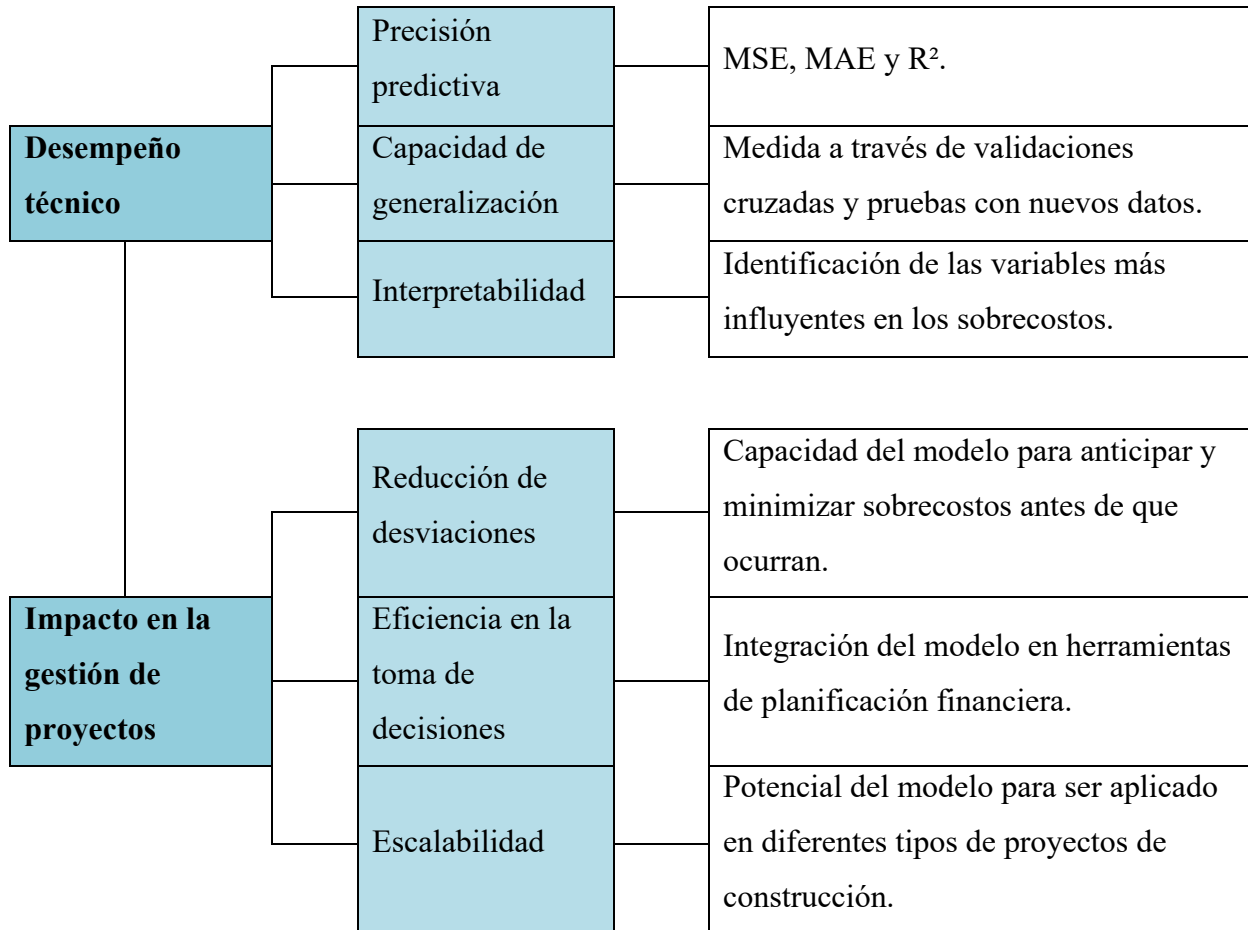
4.1.2 Valor del análisis de datos

El análisis de datos es una herramienta para mejorar la toma de decisiones, la eficiencia operativa y la planificación estratégica en diferentes sectores. Su aplicación permite a las organizaciones descubrir patrones ocultos y tomar decisiones precisas y oportunas (Sharma et al., 2024), optimizando operaciones y reduciendo costos, como en el sector energético, donde se predice la demanda de gas con mayor exactitud (Yabsley & Coleman, 2019).

En sectores específicos, el análisis de datos impulsa mejoras, como en la atención médica, donde optimiza la calidad del servicio, los diagnósticos y la prevención de enfermedades (Saini & Kanna, 2023); en telecomunicaciones, al identificar factores clave y optimizar la adopción tecnológica; y en la construcción y gestión de instalaciones, al integrar tendencias como BIM para aumentar la productividad y eficiencia.

El análisis de datos, además de optimizar procesos y mejorar la toma de decisiones, es útil para evaluar el rendimiento de modelos predictivos. En este sentido, la eficacia del modelo propuesto para la predicción de sobrecostos en proyectos de construcción de edificación se medirá con base en los indicadores de la **Figura 4-2**.

Figura 4-2: Criterios de evaluación del modelo propuesto.



Fuente: Propia.

4.2 Comprensión de los datos

La comprensión de los datos es un paso que ayuda a familiarizarse con los atributos involucrados y, por lo tanto, comprender su importancia. La eliminación de aquellos atributos que no brindan suficiente información para el análisis es importante durante el

modelado de datos, ya que esto afecta la eficiencia del modelo (George et al., 2022), por lo que es relevante una reducción en el tamaño del conjunto de datos.

4.2.1 Identificación de los datos

La base de datos "*Capital Project Schedules and Budgets*" es un recurso público proporcionado por la *School Construction Authority* (SCA) de la ciudad de Nueva York. Esta, ofrece información sobre los proyectos de construcción de capital, incluyendo presupuestos y tiempos. La **Tabla 4-1** resume algunas de sus características.

Tabla 4-1: Características de la base de datos.

Origen	School Construction Authority (SCA) de Nueva York
Nombre	Capital Project Schedules and Budgets
Año de creación	2013 (vigente hasta la actualidad)
Cantidad de registros	Más de 15,000
Número de variables	14

Fuente: Propia.

A continuación, se presenta la descripción de variables, proporcionada por la SCA:

Project Geographic District: Distrito donde se ubica el edificio.

Project Building Identifier: Identificación del edificio.

Project School Name: Nombre de la escuela.

Project Type: Tipo de proyecto en función de la financiación.

Project Description: Componente(s) del trabajo a realizar.

Project Phase Name: Ejemplos de Alcance, Diseño y Construcción.

Project Status Name: Dentro de la fase más detallada, es decir, In-Progress, Hold.

Project Phase Actual Start Date: Fecha de inicio real de la fase.

Project Phase Planned End Date: Fecha de finalización de la fase originalmente programada.

Project Phase Actual End Date: Fecha de finalización efectiva de la fase.

Project Budget Amount: Adjudicación del presupuesto base por fases.

Final Estimate of Actual Costs Through End of Phase Amount: Estimación final proyectada actual a la finalización del proyecto por fase.

Total Phase Actual Spending Amount: Gastos acumulados efectivos por etapa.

DSF Number(s): Número utilizado para identificar el proyecto en el Plan de Cinco Años.

La **Tabla 4-2** proporciona la cantidad y tipo de registro, mostrando el número de valores no nulos en cada variable (*Non-Null Count*) y el tipo de datos almacenados en ellas (*Dtype*). La cantidad de valores no nulos es similar en la mayoría de las variables, excepto en unas cuantas, donde se tienen datos incompletos, es decir, no se cuenta con los 15303. Por otro lado, el tipo de datos predominante es *Object*, indicando que la mayoría de las columnas contienen información textual o categórica, mientras que las variables numéricas, como costos y presupuestos, están representadas en *Float64* o *Int64*.

Tabla 4-2: Cantidad y tipo de registros - SCA Capital Projects.

Column	Non-Null Count	Dtype
Project geographic district	15303 non-null	Int 64
Project building Identifier	15303 non-null	Object
Project school name	15303 non-null	Object
Project type	15303 non-null	Object
Project description	15303 non-null	Object
Project phase name	15303 non-null	Object
Project status name	15303 non-null	Object
Project phase actual start date	15303 non-null	Object
Project phase planned end date	15269 non-null	Object
Project phase actual end date	10128 non-null	Object
Project budget amount	15303 non-null	Object
Final estimate of actual costs through end phase amount	15279 non-null	Float64
Total phase actual spending amount	15023 non-null	Float64
DSF number (s)	15262 non-null	Object

Fuente: Propia.

De lo anterior, se concluye que el conjunto de datos requiere procesamiento para análisis cuantitativos, especialmente en lo relacionado con estimaciones de costos.

4.2.2 Análisis exploratorio de los datos

La **Tabla 4-2**, presentada en la sección anterior, permite establecer cuáles son las variables que contienen valores nulos: “*Project Phase Planned End Date*” (34), “*Project Phase Actual End Date*” (5175), “*Final Estimate of Actual Costs Through End of Phase Amount*” (24), “*Total Phase Actual Spending Amount*” (280), y “*DSF Number(s)*” (41). La presencia de valores nulos representa los datos que no fueron registrados o ciertos registros/proyectos que aún están en desarrollo. En particular, los datos relacionados con fechas de finalización y costos finales presentan vacíos importantes, esto puede deberse a retrasos en la ejecución, falta de actualización en los registros o inconvenientes en la disponibilidad de información según la fase del proyecto. Dado lo anterior, se requiere tratamiento para manejar las ausencias y evitar sesgos en cualquier análisis posterior.

Por facilidad y para un mayor control y entendimiento, las variables son renombradas de acuerdo con la nomenclatura de la **Figura 4-3**.

Figura 4-3: Renombramiento de variables.

Project Geographic District	Distrito Geográfico
Project Building Identifier	ID Edificio
Project School Name	Nombre Escuela
Project Type	Tipo de Proyecto
Project Description	Descripción del Proyecto
Project Phase Name	Fase del Proyecto
Project Status Name	Estado del Proyecto
Project Phase Actual Start Date	Fecha de Inicio
Project Phase Planned End Date	Fecha de Fin Proyectada
Project Phase Actual End Date	Fecha de Fin Real
Project Budget Amount	Presupuesto del Proyecto
Final Estimate of Actual Costs Through End of Phase Amount	Presupuesto Final Estimado
Total Phase Actual Spending Amount	Cantidad Real Gastada
DSF Number(s)	Identificador DSF

Fuente: Propia.

Antes de eliminar los valores nulos, es necesario abordar un caso particular en la variable “presupuesto del proyecto”, en esta variable, se esperaría que todos sus registros fueran numéricos y mayores que cero, no obstante, algunos contienen valores como 'DOER', '0', 'DIIR', 'DOES', 'IEH', 'FTK', 'EMER', 'DIIT', 'DOEL', 'TPL', 'DOEP', esto indica errores en la captura o el procesamiento de datos. Por lo que, el primer paso en la limpieza consiste en identificar y eliminar estos registros para garantizar la calidad y coherencia de la información.

Además, la variable categórica “estado del proyecto” contiene tres posibles valores: “PNS”, “In-Progress” y “Complete”. Por efectos prácticos, el análisis estará enfocado sólo en proyectos finalizados, la justificación de esta decisión se menciona más adelante.

Una vez realizados estos dos pasos, se verifica nuevamente la presencia de valores nulos en el conjunto de datos, confirmando que ya no existen. A partir de este punto, se procede a ajustar el formato/tipo de las variables “Presupuesto del Proyecto”, “Fecha de Inicio”, “Fecha de Fin Proyectada” y “Fecha de Fin Real” para asegurar su correcta interpretación. Además, se eliminan las variables “Identificador DSF” e “ID Edificio”, ya que no aportan valor al análisis.

Luego de la limpieza inicial, se generan cuatro nuevas variables a partir de las existentes, cuyo objeto es facilitar el análisis de correlaciones. A continuación, se describen.

Las tres primeras variables están relacionadas con la duración del proyecto, la primera consiste en la diferencia entre “fecha de fin proyectada” y “fecha de inicio”, esta variable se denomina “duración proyectada”; luego se calcula la “duración real” como la diferencia entre “fecha de fin real” y “fecha de inicio”; la tercera variable es el “retraso” y se calcula como la diferencia entre “fecha de fin real” y “fecha de fin proyectada”.

Adicionalmente, se incorpora la variable “desviación presupuestaria” en porcentaje, definida como una fracción, donde el numerador es la diferencia entre “cantidad real gastada

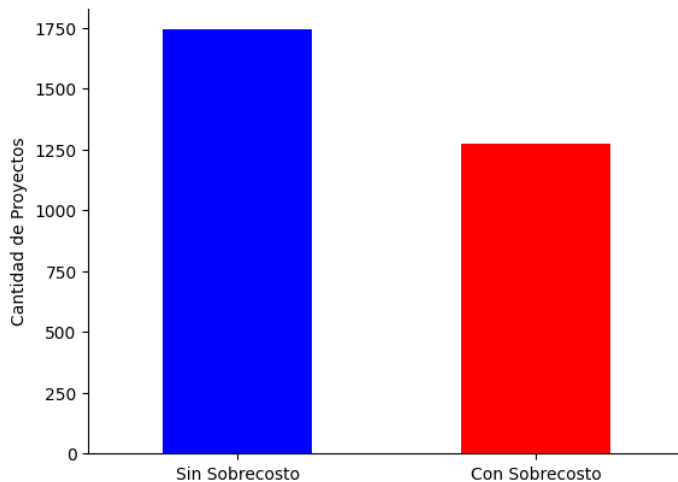
y el “presupuesto del proyecto”, y el denominador es “presupuesto del proyecto”, todo multiplicado por 100.

4.2.2.1. Variables numéricas

El análisis de variables numéricas consta de dos partes, en primer lugar, lo relacionado al presupuesto y en segundo lugar a la duración de los proyectos.

La **Figura 4-4** muestra la distribución de los proyectos en relación con la desviación presupuestaria. Se evidencia que una parte de los proyectos no presenta sobrecostos, lo que indica que el gasto real al finalizar fue igual o inferior al presupuesto estimado. No obstante, también se observa una proporción considerable de proyectos con sobrecostos, esto resalta la necesidad de realizar un análisis detallado y justifica la pertinencia del presente estudio. Si bien el análisis se basa en proyectos ya finalizados, pues solo en ellos se conoce la cantidad real gastada, sus resultados están orientados a apoyar la toma de decisiones en proyectos en ejecución, donde aún se pueden aplicar medidas de control y mitigación.

Figura 4-4: Estado de presupuesto de los proyectos.

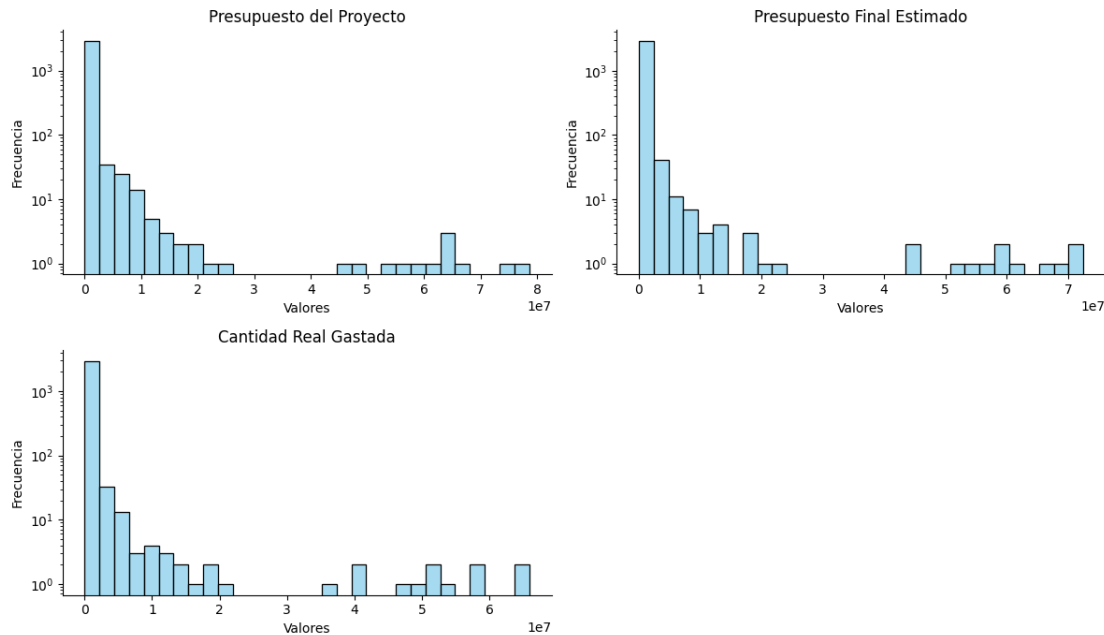


Fuente: Propia.

Continuando con el análisis de las variables numéricas, en la **Figura 4-5** se presenta el histograma de presupuestos y gastos, ahí se puede evidenciar que la mayoría de los proyectos presentan valores bajos tanto en el “presupuesto del proyecto”, el “presupuesto final

estimado” y la “cantidad real gastada”, es decir, los proyectos de menor costo son más frecuentes. El “presupuesto del proyecto” tiende a ser más conservador, con una distribución más concentrada en valores bajos y una disminución rápida en frecuencias a medida que los montos aumentan. En contraste, el “presupuesto final estimado” y la “cantidad real gastada” muestran una distribución más amplia, con una presencia en valores más altos, esto da indicios de ajustes y sobrecostos en algunos proyectos. Se hace particularmente evidente en el rango de 40 a 80 millones de dólares, donde se identifican valores extremos con presupuestos finales y gastos reales mucho mayores que los iniciales.

Figura 4-5: Distribución de presupuestos y gastos.

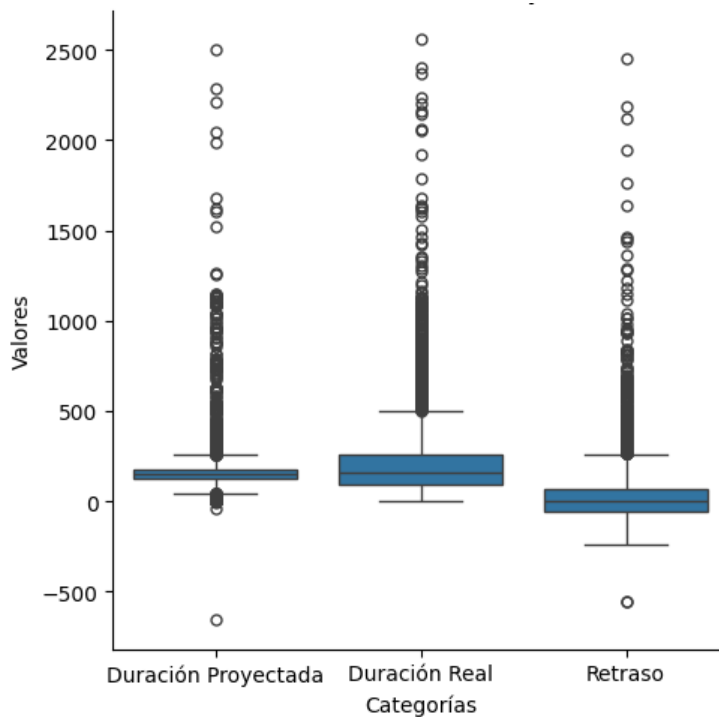


Fuente: Propia.

Ahora bien, en relación con la duración de los proyectos, en **Figura 4-6**, las “duraciones proyectadas” presentan una variabilidad limitada, con la mayoría de los valores agrupados cerca de la mediana. Sin embargo, existen casos extremos con estimaciones inusualmente largas o cortas. Las “duraciones reales”, en contraste, muestran una mayor dispersión y una mediana superior a la de las proyecciones iniciales, esto puede indicar que los proyectos suelen extenderse más de lo estimado. El “retraso”, medido como la diferencia entre la

duración real y la proyectada, tiene una dispersión con valores tanto positivos (retrasos) como negativos (adelantos). Aunque la mediana está cerca de cero, la alta variabilidad sugiere que, si bien algunos proyectos se adelantan, otros enfrentan retrasos importantes.

Figura 4-6: Distribución de Duraciones y Retrasos.



Fuente: Propia.

A partir de las figuras anteriores y la información presentada en la **Tabla 4-3**, se puede notar que las variables analizadas presentan una alta dispersión y presencia de valores atípicos. El presupuesto y gasto real, en particular, muestran diferencias amplias entre sus valores máximos y la mediana, el resultado de esto puede ser proyectos que enfrentaron sobrecostos significativos o ajustes presupuestarios. En cuanto a la duración, se detectan valores negativos en la duración proyectada, lo cual podría reflejar errores en la base de datos. Por su parte, los retrasos presentan una alta variabilidad, con casos de proyectos finalizados con una amplia antelación y otros con demoras considerables.

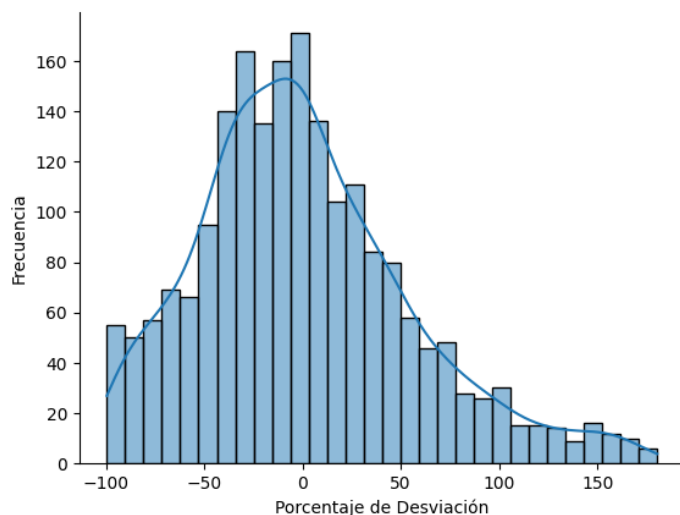
Tabla 4-3: Resumen estadístico de presupuesto, gastos y duración de proyectos.

	Presupuesto del Proyecto	Presupuesto Final Estimado	Cantidad Real Gastada	Duración Proyectada	Duración Real	Retraso	Desviación Presupuestaria
count	3013.00	3013.00	3013.00	3013.00	3013.00	3013.00	3013.00
mean	681242.10	591573.10	516797.80	185.51	231.62	46.10	12.39
std	4161808.00	3926412.00	3448535.00	171.11	255.51	199.30	136.85
min	201.00	28.00	6.00	-657.00	1.00	-551.00	-99.90
25%	21827.00	25329.00	19323.00	127.00	99.00	-57.00	-40.93
50%	83250.00	104251.00	92400.00	153.00	159.00	0.00	-9.23
75%	300370.00	310017.00	274885.00	178.00	261.00	75.00	32.89
Max	78613500.00	72497380.00	65905830.00	2498.00	2557.00	2451.00	4406.41

Fuente: Propia.

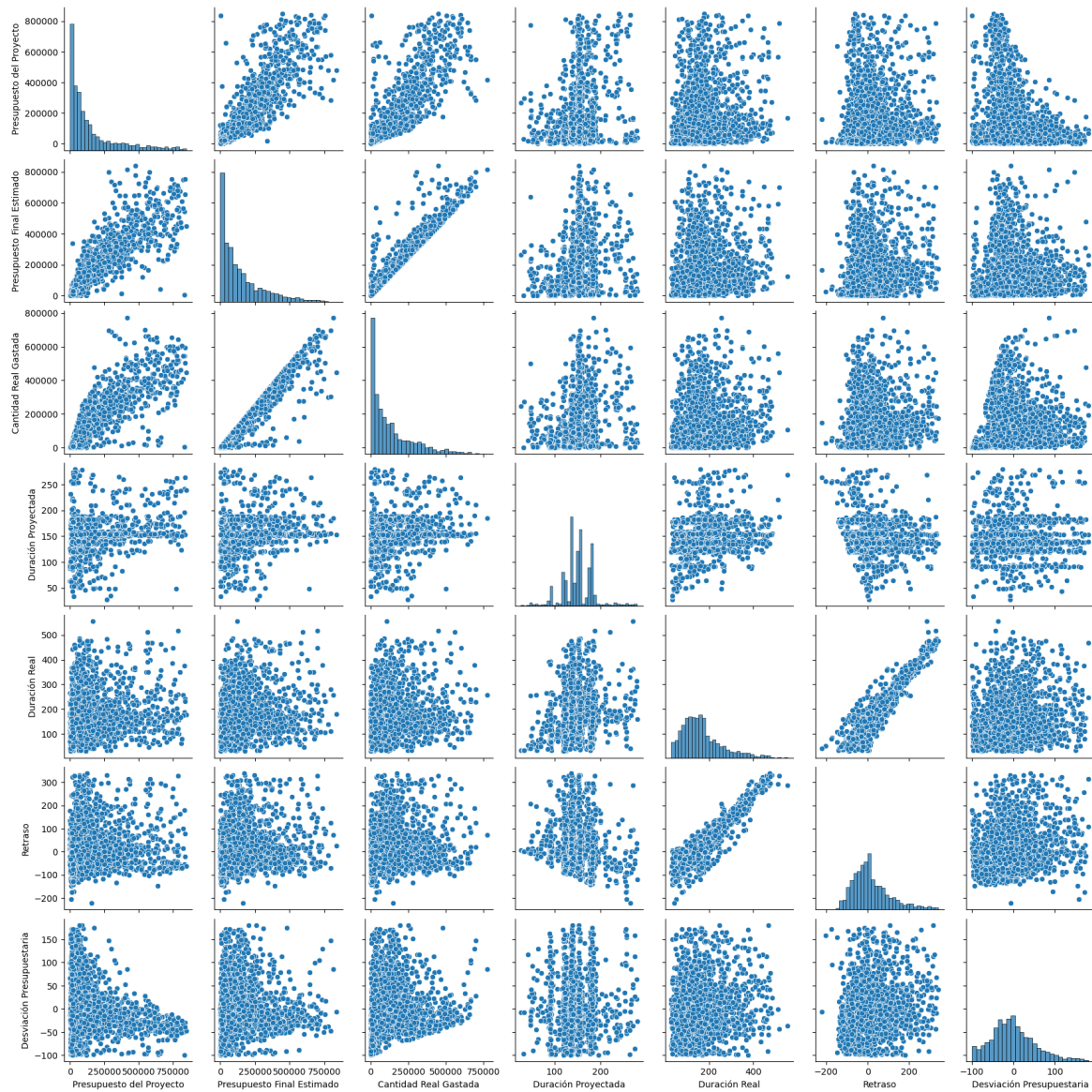
Dado este comportamiento, y con el objetivo de mejorar la calidad de los datos para los modelos que se desarrollarán posteriormente, se optó por aplicar un método de eliminación de valores atípicos basado en el rango intercuartílico (IQR). Para ello, se empleó una función que calcula los cuartiles Q1 y Q3 para cada variable, determina el IQR ($Q3-Q1$) y elimina aquellos registros cuyos valores estén por fuera de los límites definidos como $Q1-2 \times IQR$ y $Q3+2 \times IQR$. Esta técnica permite reducir el efecto de valores extremos sin eliminar datos representativos.

Una vez aplicado el ajuste por IQR para eliminar los valores atípicos en todas las variables numéricas incluida la “desviación presupuestaria”, la distribución resultante de esta última variable revela particularidades sobre el comportamiento de los proyectos analizados. La **Figura 4-7** muestra una asimetría hacia la derecha, con una mayor concentración de frecuencia en torno al 0%, además, una proporción de proyectos se ejecutó sin desviaciones respecto al presupuesto (desviaciones negativas). Aún se conserva una dispersión, con una cola prolongada hacia los valores positivos. Esto evidencia que, aunque los valores extremos fueron removidos, persisten casos con desviaciones elevadas que reflejan variabilidad en los datos y no necesariamente anomalías. La forma de la distribución sugiere que las desviaciones positivas tienden a ser más frecuentes que las negativas.

Figura 4-7: Distribución de desviación presupuestaria.

Fuente: Propia.

Por otro lado, se analizan las relaciones entre las variables numéricas, utilizando un gráfico de dispersión múltiple, **Figura 4-8**. A partir de este, se observa que no hay una relación lineal fuerte entre la desviación presupuestaria y la mayoría de las variables, es decir, su comportamiento no está determinado de forma directa por una única variable cuantitativa. Sin embargo, se identifican ciertos patrones: la desviación presupuestaria tiende a disminuir a medida que aumentan tanto el presupuesto del proyecto como la cantidad real gastada, esto puede indicar que los proyectos de gran magnitud presentan mayor control o capacidad de absorción ante desviaciones. También se perciben asociaciones más difusas con las variables de tiempo. Aunque no hay una tendencia clara, hay indicios de que los retrasos pueden contribuir al aumento de la desviación presupuestaria en algunos casos, aunque esta relación no es uniforme.

Figura 4-8: Matriz de dispersión entre variables numéricas.

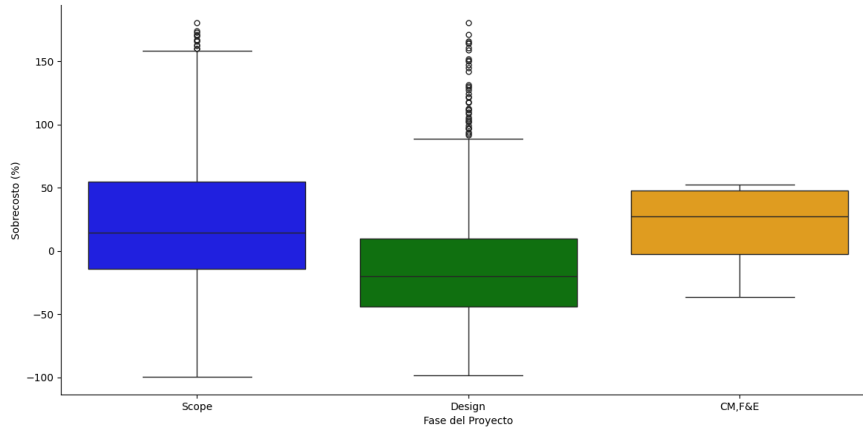
Fuente: Propia.

4.2.2.1. Variables categóricas

Inicialmente, se revisa el comportamiento de una de las variables categóricas, “fase del proyecto”, con respecto a la “desviación presupuestaria”. La **Figura 4-9** muestra diferencias entre las fases: en *Design*, los proyectos tienden a registrar valores negativos, es decir, se gasta menos de lo estimado. En contraste, las fases *Scope* y *CM,F&E* presentan una mayor concentración de valores positivos. Esto sugiere que los sobrecostos se manifiestan con

mayor fuerza en etapas avanzadas del proyecto, mientras que en la fase de diseño es más común observar ahorros o presupuestos conservadores.

Figura 4-9: Sobrecosto por fase del proyecto.



Fuente: Propia.

Continuando con el análisis de variables categóricas, se realizó un proceso de categorización temática sobre la columna "descripción del proyecto", que originalmente presentaba una alta definición conceptual con más de 600 categorías distintas. Esta variedad dificultaba su análisis e interpretación, especialmente al tratar de relacionar el tipo de proyecto con variables como la desviación presupuestaria.

Con el fin de reducir esta complejidad y facilitar el análisis, se implementó una estrategia de clasificación basada en palabras clave, agrupando las descripciones en un conjunto más manejable de categorías generales. Estas categorías se presentan en la **Tabla 4-4**.

Tabla 4-4: Categorías generales de la descripción del proyecto.

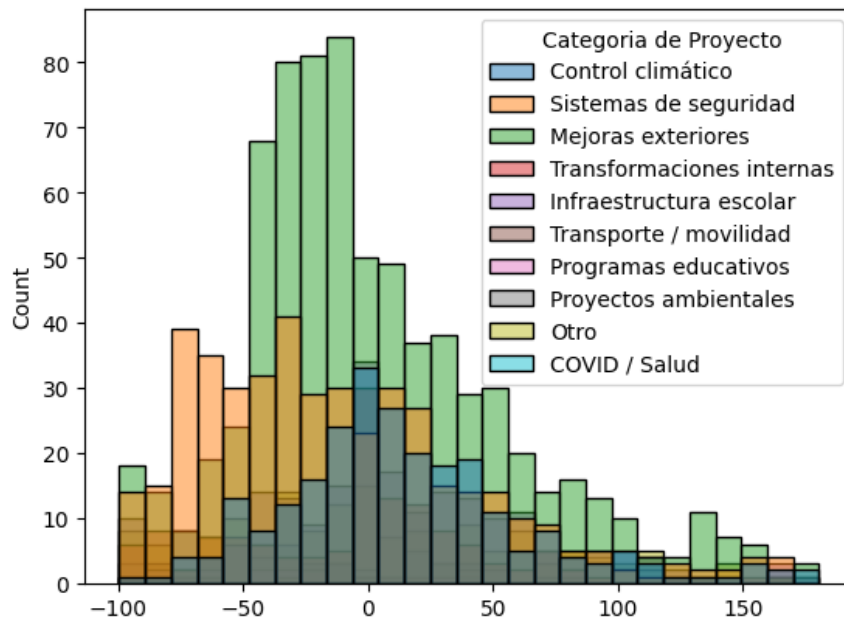
Categoría de Proyecto	Palabras Clave Asociadas
Mejoras exteriores	exterior, masonry, roofs, parapets, windows, replacement
Sistemas de seguridad	fire, alarm, camera, system, ipdvs, installations
Control climático	climate, ventilation, boiler, freezers, electrical, low, voltage
Infraestructura escolar	auditorium, playground, redevelopment, walk, accessibility

Categoría de Proyecto	Palabras Clave Asociadas
Proyectos ambientales	flood, elimination, PLANYC
Transformaciones internas	conversion, upgrade, installation, systems
COVID / Salud	covid
Transporte / movilidad	path, travel
Programas educativos	program, full, reso

Fuente: propia.

La **Figura 4-10** muestra la distribución de la desviación presupuestaria según la categoría temática del proyecto. En general, la mayoría de las categorías concentran valores alrededor de cero o negativos, indicando ejecución dentro del presupuesto o con ahorro. No obstante, en todas se observan casos con sobrecostos (desviaciones positivas), indicando que este riesgo es transversal, independientemente del tipo de intervención. Las categorías más frecuentes, como mejoras exteriores y control climático, destacan por su mayor concentración de registros.

Figura 4-10: Desviación presupuestaria según categoría temática.



Fuente: propia.

Como etapa final del análisis exploratorio, se construyeron dos nuevas variables: “probabilidad” e “impacto”, utilizando una escala tipo Likert. Para su definición, se tomó como referencia el estudio de Ashtari et al. (2022), el cual identifica factores de riesgo con potencial incidencia en sobrecostos en proyectos de construcción. Estos factores, junto con sus respectivas escalas de probabilidad e impacto, se detallan en el Anexo A.

El procedimiento consistió en asociar los factores de riesgo a tres variables disponibles del conjunto de datos: “fase del proyecto”, “retraso” y “presupuesto del proyecto”. Se identificaron los riesgos más representativos para cada una, como aquellos vinculados al alcance o diseño, los que implican demoras superiores al 30 %, y los relacionados con desviaciones presupuestales, conforme se describe en el Anexo B.

Dado que un mismo proyecto puede presentar uno o varios de estos escenarios, se asignaron los valores correspondientes de probabilidad e impacto a cada caso y finalmente, se calculó el promedio de estos valores para cada proyecto.

4.3 Preparación de los datos

En esta fase siguió una serie de procesos, a saber, se seleccionaron variables relevantes a partir de una base de datos limpia con 2010 registros y 15 características/variables. Se descartaron variables sin valor explicativo o sin disponibilidad en proyectos en progreso, como “nombre escuela”, “estado del proyecto” o “retraso”. Adicionalmente, se incorporó una nueva variable estimada de “gasto real a la fecha” para mejorar la aplicabilidad del modelo. Se analizaron correlaciones mediante ANOVA y el coeficiente de Pearson. Finalmente, se prepararon los datos mediante estandarización y codificación, y se dividió el conjunto en datos de entrenamiento y prueba.

4.3.1 Selección de variables relevantes

Luego del análisis exploratorio realizado en la sección 4.2, se cuenta con un total de 2010 proyectos y 15 variables, entre las cuales se incluyen: “distrito geográfico”, “nombre escuela”, “tipo de proyecto”, “fase del proyecto”, “estado del proyecto”, “presupuesto del

proyecto”, “presupuesto final estimado”, “cantidad real gastada”, “duración proyectada”, “duración real”, “retraso”, “categoría del proyecto”, “probabilidad”, “impacto” y “desviación presupuestaria”.

Antes de proceder con la selección de variables relevantes para el modelo, es necesario identificar y justificar aquellas que serán descartadas del análisis. En primer lugar, la variable “distrito geográfico” se excluye dado que representa únicamente un identificador numérico entre 1 y 32, sin un significado contextual claro que contribuya a explicar el sobrecosto. De manera similar, el “nombre escuela” corresponde a un dato nominal, que no aporta información generalizable ni patrones útiles para el modelo. Por otro lado, la variable “estado del proyecto” también se descarta, ya que en este conjunto de datos todos los registros se encuentran en estado "completado", lo cual no ofrece variabilidad ni valor predictivo.

Adicionalmente, se excluyen las variables “duración real” y “retraso”. Aunque son variables relacionadas con el sobrecosto, su uso se limita por razones prácticas: si bien están disponibles para los proyectos completados, no están presentes ni pueden calcularse para los proyectos en ejecución. Esto se debe a que en los proyectos en progreso no se cuenta con datos de seguimiento en tiempo real ni con una fecha de corte que permita inferir el avance, la duración acumulada o un posible retraso. Este punto es clave, ya que el modelo se entrenará con datos históricos de proyectos completados, pero su aplicación práctica se espera que sea en proyectos en ejecución.

Para garantizar que el modelo sea funcional en ambos escenarios, se requiere que las variables utilizadas durante el entrenamiento y validación también estén disponibles en su puesta en marcha. Utilizar variables que solo existen en proyectos finalizados comprometería la capacidad del modelo para generalizar y hacer predicciones durante el curso de la ejecución del proyecto. Esta alineación garantiza que el modelo pueda ser aplicado efectivamente a proyectos en curso, cumpliendo así con el propósito práctico de anticipar sobrecostos en etapas tempranas o intermedias de la ejecución.

En los proyectos en ejecución es común realizar controles periódicos para evaluar su estado, en términos presupuestales. Para ello, se suele revisar el gasto real acumulado hasta una fecha determinada.

En la base de datos utilizada, los proyectos en curso incluyen una variable denominada "gasto real", la cual se interpreta como el valor registrado en un momento específico del avance del proyecto. Con el fin de incorporar una nueva variable explicativa que mejore la precisión de los modelos predictivos, se construyó la variable "gasto real a la fecha estimado". Para su obtención, se formuló y entrenó un modelo de bosques aleatorios utilizando los datos disponibles de proyectos en ejecución.

Tras verificar que su capacidad de generalización era adecuada, se aplicó el modelo para estimar el gasto real a la fecha en proyectos completados. Los detalles sobre la formulación y validación del modelo se presentan en el Anexo C.

Posterior a la exclusión e inclusión de variables, el paso siguiente es el análisis de correlaciones, para ello se emplean dos herramientas estadísticas según el tipo de variables: el Análisis de Varianza (*ANOVA*) para variables categóricas y el coeficiente de correlación de Pearson para variables numéricas.

El *ANOVA* es un método que permite comparar las medias de tres o más grupos con el fin de identificar si existen diferencias estadísticamente significativas entre ellos. Además, ayuda a controlar el riesgo de cometer errores tipo I (falsos positivos) que pueden surgir al realizar múltiples comparaciones (Kim, 2017).

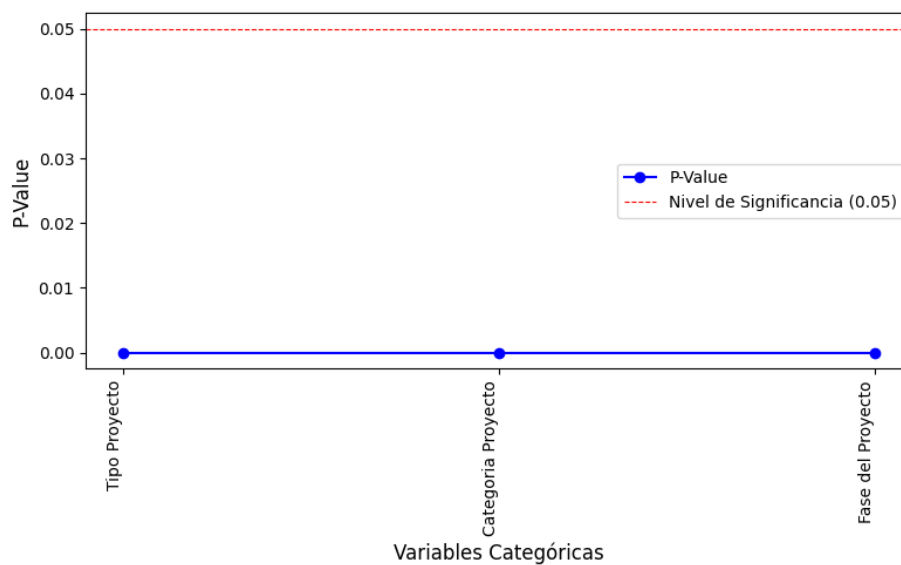
Por su parte, el coeficiente de correlación de Pearson evalúa la fuerza y dirección de la relación lineal entre dos variables continuas (Akoglu, 2018). Su valor oscila entre -1 y 1, donde: 1 indica una correlación positiva perfecta; -1 una correlación negativa perfecta; y 0 la ausencia de correlación lineal.

La **Figura 4-11** presenta los resultados del análisis *ANOVA* aplicado a las variables categóricas del conjunto de datos con la variable objetivo “*cantidad real gastada*”. En el gráfico se observa, en el eje X, cada una de las variables categóricas analizadas, mientras que en el eje Y se representan sus respectivos valores p. Estos valores p indican la probabilidad de que las diferencias observadas entre las medias de gasto real en las distintas categorías se deban al azar.

Para facilitar su interpretación, se incluye una línea roja horizontal que representa el umbral comúnmente utilizado como nivel de significancia ($\alpha = 0.05$). Cuando el p-value de una variable se encuentra por debajo de esta línea, se considera que existe suficiente evidencia estadística para rechazar la hipótesis nula, lo que implica que dicha variable tiene un efecto significativo sobre la variable objetivo.

En este caso, los resultados muestran que todas las variables categóricas analizadas presentan valores p muy cercanos a cero y por debajo del umbral de 0.05. En otras palabras, las diferencias observadas en el gasto real entre las distintas categorías no son producto del azar, sino que están asociadas al tipo, fase o categoría del proyecto.

Figura 4-11: P-Values de variables categóricas frente a la cantidad real gastada.



Fuente: Propia.

Ahora bien, la **Figura 4-12** presenta los coeficientes de correlación de Pearson entre distintas variables numéricas y la variable objetivo “*cantidad real gastada*”. De esta se puede inferir lo siguiente:

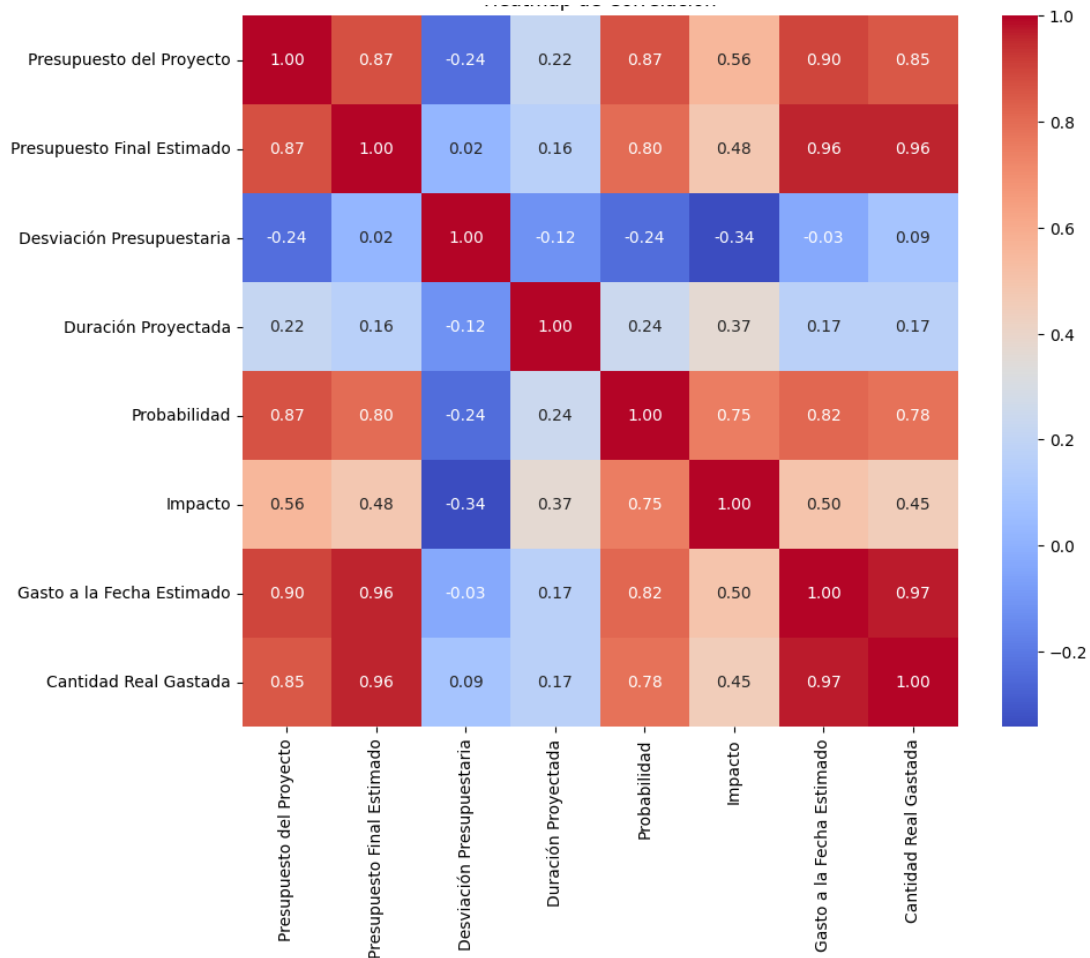
El gasto a la fecha estimado tiene una correlación cercana a 1, lo cual es esperable, ya que esta variable fue generada específicamente para aproximarse al comportamiento del gasto real en proyectos en ejecución. Su fuerte asociación valida su inclusión como una variable explicativa relevante en un modelo predictivo.

Le siguen en importancia presupuesto final estimado y presupuesto del proyecto, ambas con correlaciones superiores a 0.85, lo que indica que a mayor presupuesto proyectado (inicial o final), mayor tiende a ser el gasto real observado. Estas variables reflejan el alcance y magnitud del proyecto, por lo que es lógico que se relacionen directamente con el nivel de gasto.

La variable probabilidad (asociada al riesgo) también presenta una correlación significativa, aunque ligeramente menor (0.78), sugiriendo que a medida que aumenta la probabilidad de ocurrencia de riesgos, el gasto real tiende a incrementarse, posiblemente por la materialización de sobrecostos.

En contraste, variables como impacto, duración proyectada y desviación presupuestaria muestran correlaciones más moderadas o bajas. La variable impacto presenta una correlación media (0.45), lo que sugiere que, si bien la severidad potencial de los riesgos influye en el gasto, no lo hace de forma tan directa como la probabilidad.

La duración proyectada y desviación presupuestaria tienen coeficientes bajos (inferiores a 0.2), esto indica una relación más débil.

Figura 4-12: Coeficiente de correlación de Pearson.

Fuente: Propia.

Como conclusión, se tiene que la incorporación de variables categóricas y numéricas en modelos predictivos resulta pertinente, ya que aportan información sobre el contexto, la naturaleza y el estado de ejecución de los proyectos. Los resultados del análisis exploratorio respaldan la viabilidad de desarrollar un modelo capaz de estimar con precisión la cantidad real que se gastará en los proyectos. Esto permitiría anticipar posibles sobrecostos y facilitar la toma de decisiones oportunas para implementar medidas correctivas o preventivas.

4.3.2 Transformación de datos

Como parte del proceso de preparación, se implementa una estrategia de transformación de datos que permite adaptar las variables a los requerimientos de los algoritmos de aprendizaje

automático. Esta transformación considera la naturaleza de las variables disponibles, diferenciando entre variables numéricas y categóricas.

Para las variables numéricas, se aplica una estandarización *StandardScaler()*, con el fin de llevarlas a una escala común. Esta técnica ajusta los valores para que tengan una media igual a cero y una desviación estándar igual a uno, evitando que diferencias en magnitudes influyan de manera desproporcionada en el modelo.

En cuanto a las variables categóricas, se realiza una codificación mediante el método de *one-hot encoding*, que convierte cada categoría en una columna binaria. Esto permite representar información cualitativa de forma que los algoritmos puedan interpretarla sin introducir un orden artificial entre categorías.

4.3.3 División del conjunto de datos

Con el propósito de desarrollar modelos predictivos y evaluar sus capacidades para generalizar a nuevos datos, se realizó una partición del conjunto de datos. En esta etapa, se separa la variable objetivo, “*cantidad real gastada*”, de las variables independientes o explicativas. Posteriormente, se empleó una técnica de muestreo aleatorio para dividir los datos en dos subconjuntos: un conjunto de entrenamiento y un conjunto de prueba.

El 80% de los datos se asignó al conjunto de entrenamiento (x_{train} , y_{train}), el 20% restante se destinó al conjunto de prueba (x_{test} , y_{test}). Este conjunto se emplea para evaluar el rendimiento del modelo sobre datos no vistos, permitiendo así estimar su capacidad de generalización. La división se realiza de forma aleatoria, pero controlada mediante una semilla fija ($\text{random_state}=42$), con el fin de garantizar la reproducibilidad del experimento.

4.4 Modelado

El proceso de modelado incluyó la formulación de cuatro modelos: *linear regression*, *random forest regressor*, *gradient boosting regressor* y *multilayer perceptron regressor*.

Cada modelo se desarrolló siguiendo un flujo de trabajo común, con diferencias específicas en la construcción del pipeline y en la estrategia de búsqueda en malla (*grid search*), según las particularidades de cada algoritmo. Los detalles completos del desarrollo y parametrización de los modelos están disponibles en el siguiente repositorio público: <https://github.com/Anfospina/cost-overflow-prediction>

4.4.1 Linear Regression

A partir de características relacionadas con el presupuesto, la duración, los riesgos y el tipo de proyecto, se desarrolló un modelo de regresión lineal para predecir la cantidad real gastada en proyectos. Este flujo de trabajo se estructuró en cuatro fases: preprocesamiento de datos, construcción de pipeline, búsqueda de hiperparámetros mediante validación cruzada, y evaluación del modelo.

- Preprocesamiento de datos

El preprocesamiento se diseñó para preparar adecuadamente los datos antes del entrenamiento del modelo. Se identificaron dos tipos de variables:

Variables numéricas, como: gasto a la fecha estimado, presupuesto final estimado, presupuesto del proyecto, probabilidad e impacto de riesgos, duración proyectada, desviación presupuestaria.

Variables categóricas, como: tipo de proyecto, fase del proyecto.

Para el tratamiento de estas variables se utilizó un *ColumnTransformer*, que aplica transformaciones específicas a cada tipo. Las variables numéricas fueron estandarizadas con *StandardScaler*, para asegurar que todas tengan media cero y desviación estándar uno, favoreciendo el desempeño del modelo. Las variables categóricas fueron transformadas mediante *OneHotEncoder*, lo que permite convertirlas en variables binarias sin introducir ordenaciones arbitrarias.

- Construcción del pipeline de modelado

Se creó un Pipeline de *scikit-learn* para encapsular todo el flujo de trabajo. El pipeline incluía los siguientes pasos: (1) Preprocesamiento, aplica las transformaciones descritas anteriormente. (2) Selección de características, se utilizó *SelectKBest* con la prueba estadística *f_regression*, lo que permite seleccionar automáticamente las variables con mayor poder explicativo. (3) Modelo de regresión, se utilizó *LinearRegression* como algoritmo base, con posibilidad de ajustar si se incluye o no un término de intercepto.

- Búsqueda en cuadrícula

Para ajustar los hiperparámetros del pipeline, se implementó una búsqueda exhaustiva (*GridSearchCV*) con validación cruzada de 10 pliegues. Esto permite identificar la mejor combinación de parámetros, minimizando el error cuadrático medio en distintos subconjuntos de datos.

Los hiperparámetros evaluados fueron: (1) *feature_selection__k*, número de características a seleccionar (5, 10, 15, 17). (2) *regressor__fit_intercept*, inclusión o no del término independiente en la regresión (True o False).

El uso de *GridSearchCV* automatizó la evaluación sistemática de estas combinaciones y seleccionó la que presentó el mejor desempeño promedio.

- Entrenamiento, evaluación y almacenamiento del modelo

El conjunto de datos se dividió en entrenamiento (80%) y prueba (20%) para validar el rendimiento del modelo. Una vez finalizado el ajuste con *GridSearchCV*, se evaluaron métricas de desempeño tanto en el conjunto de entrenamiento como en el de prueba, incluyendo: R^2 (coeficiente de determinación), Error cuadrático medio (MSE), Error absoluto mediano (MAD). Estas métricas se almacenaron en un archivo JSON para su posterior análisis. Finalmente, el modelo optimizado fue almacenado en formato comprimido (*pkl.gz*) usando *pickle* y *gzip*, permitiendo su reutilización en futuros procesos sin necesidad de reentrenamiento.

4.4.2 *Random Forest Regressor*

Como segundo modelo se implementó *Random Forest Regressor*, con el propósito de mejorar la capacidad predictiva respecto a la *Cantidad Real Gastada* en proyectos, utilizando variables que caracterizan su planificación, presupuesto, riesgo y tipología. El enfoque adoptado permitió capturar relaciones no lineales y reducir la sensibilidad a valores atípicos o a la redundancia entre predictores.

Al igual que en el modelo anterior (basado en regresión lineal), se mantuvo la misma lógica para el tratamiento de datos. El conjunto de variables se dividió en: variables numéricas, que fueron normalizadas mediante *StandardScaler* y variables categóricas, transformadas con *OneHotEncoder* para codificación binaria.

Estas transformaciones se integraron en un *ColumnTransformer*, que posteriormente fue incorporado en un *Pipeline* junto al estimador principal. En este caso, el algoritmo elegido fue *RandomForestRegressor*, un modelo de tipo *ensemble* que combina múltiples árboles de decisión para generar predicciones más estables y precisas.

- **Búsqueda en cuadrícula**

Para mejorar el desempeño del modelo y ajustarlo adecuadamente al problema, se implementó una búsqueda de hiperparámetros mediante *GridSearchCV*, que evalúa múltiples combinaciones de configuraciones utilizando validación cruzada de 10 pliegues.

Los hiperparámetros evaluados incluyeron:

- *n_estimators*: número de árboles en el bosque (100, 200, 300).
- *max_depth*: profundidad máxima de cada árbol (10, 20, 30, sin límite).
- *min_samples_split*: número mínimo de muestras necesarias para dividir un nodo (2, 5, 10).
- *min_samples_leaf*: número mínimo de muestras en una hoja (1, 2, 4).
- *max_features*: número de características consideradas al dividir un nodo (auto, sqrt, log2).

- *bootstrap*: si se usa muestreo con reemplazo (*True, False*).

El resultado fue la selección automática de la mejor combinación de estos hiperparámetros, con base en su capacidad para reducir el error en distintas divisiones de los datos.

Finalmente, tal como sucedió con la última fase del proceso anterior (regresión lineal), se almacenan las métricas y el modelo en los formatos establecidos.

4.4.3 Gradient Boosting Regressor

En esta tercera iteración del proceso de modelado se implementó un modelo basado en *Gradient Boosting Regressor*, una técnica de aprendizaje supervisado que construye modelos de forma secuencial, donde cada nuevo árbol busca corregir los errores cometidos por el conjunto anterior. Al igual que en los modelos anteriores, se empleó una estructura de pipeline para mantener la trazabilidad del flujo de trabajo, aunque el foco principal en este caso estuvo en ajustar adecuadamente los hiperparámetros del algoritmo para maximizar su rendimiento.

- Búsqueda en cuadrícula

Para identificar la configuración adecuada del modelo, se utilizó el método de búsqueda en cuadrícula (*GridSearchCV*) con validación cruzada de 6 pliegues.

Los hiperparámetros explorados incluyeron:

- *n_estimators*: número de árboles que conforman el modelo.
- *learning_rate*: controla la contribución de cada nuevo árbol en el ajuste.
- *max_depth*: profundidad de cada árbol, clave para capturar interacciones entre variables.
- *min_samples_split* y *min_samples_leaf*: definen criterios mínimos para la creación de nodos y hojas.
- *subsample*: proporción de datos utilizada para entrenar cada árbol, lo cual introduce aleatoriedad y mejora la generalización.

- *max_features*: cantidad de variables consideradas en cada división, lo que influye en la diversidad de los árboles.

El resultado fue un modelo configurado con los parámetros más efectivos según el rendimiento observado, el cual se entrenó con los datos de entrenamiento y se evaluó posteriormente sobre el conjunto de prueba.

4.4.4 Multi Layer Perceptron Regressor

El cuarto modelo implementado para predecir la variable de interés fue una red neuronal artificial (*MLPRegressor*), una técnica de aprendizaje profundo que simula el funcionamiento de las neuronas humanas mediante capas conectadas entre sí. Aunque la fase de preprocesamiento y la construcción del pipeline se mantuvieron consistentes con los enfoques anteriores, en este caso el foco principal estuvo en diseñar y ajustar adecuadamente la arquitectura de la red neuronal.

- Búsqueda en cuadrícula

Para encontrar la estructura más adecuada de la red y maximizar su rendimiento, se utilizó nuevamente una búsqueda en cuadrícula (*GridSearchCV*) con validación cruzada de 5 pliegues. Se exploraron diferentes combinaciones de hiperparámetros:

- *hidden_layer_sizes*: define el número de capas ocultas y la cantidad de neuronas en cada una (por ejemplo, una sola capa con 100 neuronas o dos capas de 50).
- *activation*: función de activación utilizada por las neuronas (relu o tanh), que influye en la capacidad del modelo para aprender relaciones no lineales.
- *solver*: algoritmo de optimización empleado para actualizar los pesos (adam o sgd).
- *alpha*: coeficiente de regularización L2, que ayuda a prevenir el sobreajuste.
- *learning_rate*: estrategia de actualización de la tasa de aprendizaje durante el entrenamiento (constant o adaptive).

4.4.5 Evaluación de modelos

Los resultados presentados en la **Tabla 4-5**, correspondientes a las predicciones realizadas sobre el conjunto de prueba, evidencian diferencias en el desempeño de los modelos.

El *Gradient Boosting Regressor* se destaca como el modelo más preciso, alcanzando un coeficiente de determinación R^2 de 0.9824, el más alto del grupo. Además, presenta el menor error cuadrático medio (MSE) con 309,699,558.66, lo que indica una capacidad sobresaliente para capturar las variaciones del costo real. Su error absoluto medio (MAE) también es bajo (3887.30), esto refuerza su consistencia en predicciones individuales.

Por otro lado, el *Random Forest Regressor* también muestra un rendimiento sólido con un R^2 de 0.9764, un MSE de 414,936,490.77 y el menor MAE de todos los modelos (3357.15), esto indica precisión puntual, especialmente en escenarios con menor sensibilidad a valores extremos.

El modelo de Regresión Lineal obtiene un R^2 de 0.9616 y un MSE elevado de 676,569,986.56, con un MAE de 9018.84, el mayor entre todos. Si bien logra capturar una parte considerable de la varianza, sus predicciones individuales tienen mayor dispersión, reflejando limitaciones al no considerar relaciones no lineales presentes en los datos.

Finalmente, la Red Neuronal (*Multi-Layer Perceptron*) presenta un R^2 similar que la regresión lineal (0.9616), pero con un leve aumento en MSE (676,999,465.96) y una reducción del MAE (7541.29). A pesar de ser un modelo con potencial para capturar relaciones complejas, en este caso no logró superar a los modelos de ensamble, probablemente debido a la falta de más datos para generalizar mejor.

Tabla 4-5: Métricas de evaluación en el conjunto de prueba.

	R²	MSE	MAE
Linear Regression	0.9616	676569986.56	9018.84
Random Forest Regressor	0.9764	414936490.77	3357.15
Gradient Boosting Regressor	0.9824	309699558.66	3887.30
Multi Layer Perceptron Regressor	0.9616	676999465.96	7541.29

Fuente: Propia.

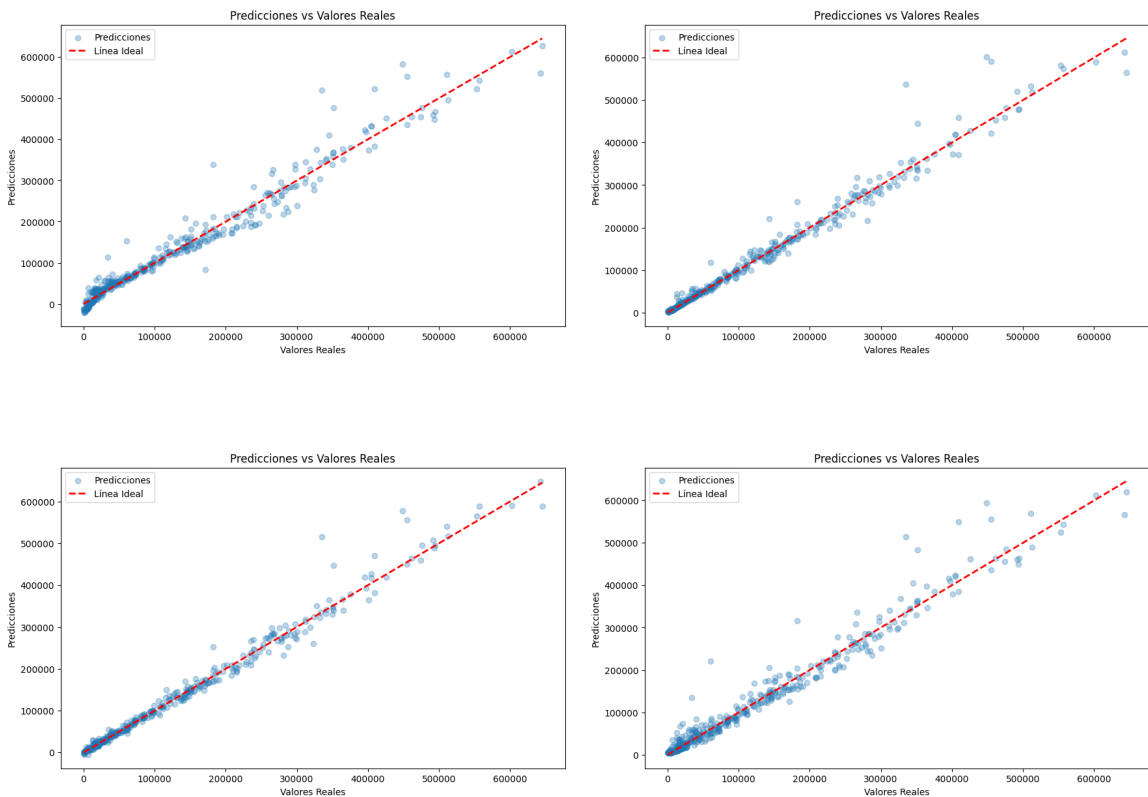
En términos visuales como se muestra en **Figura 4-13**, se confirma que los modelos de ensamble *Random Forest* y *Gradient Boosting* generan predicciones más ajustadas a la línea ideal (roja punteada), que representa una predicción perfecta. En ambos casos, los puntos están densamente agrupados alrededor de esta línea, lo cual refuerza su capacidad para capturar patrones no lineales y manejar la complejidad de la relación entre las variables predictoras y la variable objetivo.

El *Gradient Boosting*, en la esquina inferior izquierda, presenta la mejor alineación visual: los puntos siguen casi perfectamente la línea diagonal, con muy pocos valores atípicos o dispersos. Esto es coherente con su superior desempeño en las métricas cuantitativas.

El *Random Forest*, en la parte superior derecha, también muestra un patrón de dispersión contenido y alineado, aunque se observan ligeras desviaciones en los extremos, que explican su mayor MSE frente al *boosting*. Por su parte, la Regresión Lineal (esquina superior izquierda) muestra mayor dispersión, especialmente en los extremos del rango de valores. Aunque la tendencia general es adecuada, la recta de predicción no se ajusta con la misma precisión, lo cual se traduce en errores más altos y menor robustez frente a relaciones no lineales.

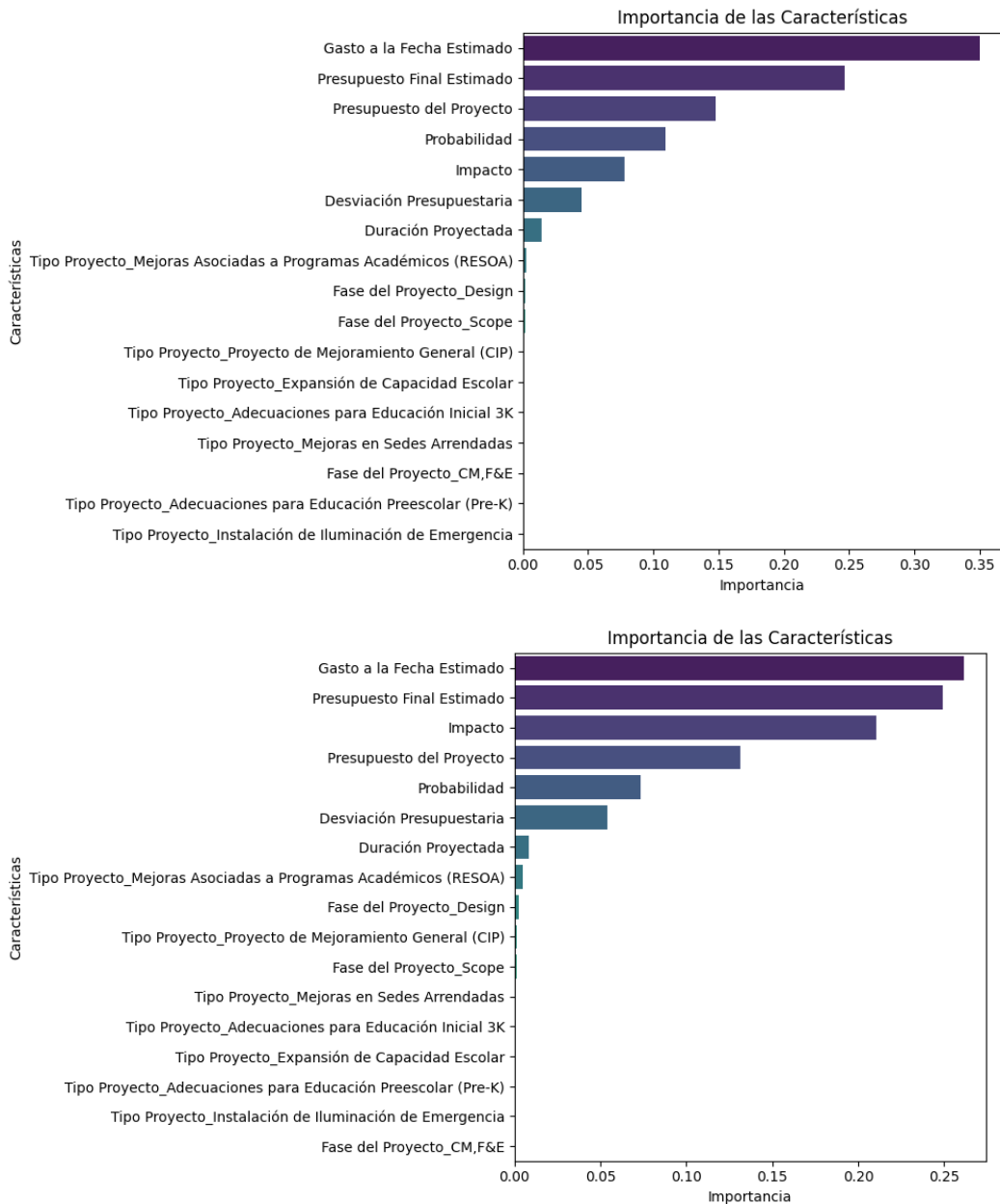
Finalmente, la Red Neuronal (MLP), en la esquina inferior derecha, exhibe un comportamiento intermedio. Aunque la nube de puntos sigue una tendencia general positiva y próxima a la línea ideal, se perciben errores más pronunciados en predicciones elevadas, lo que podría indicar una limitación en la arquitectura seleccionada o en la capacidad del modelo para generalizar.

Figura 4-13: Predicciones Vs. Valores reales en el modelado.



Fuente: Propia.

La **Figura 4-14** muestra la importancia de las características en los modelos de regresión *Random Forest* (superior) y *Gradient Boosting* (inferior), y ofrecen información para la interpretación y la toma de decisiones sobre los factores que más influyen en la predicción del costo real final de los proyectos.

Figura 4-14: Importancia de variables en los modelos.

Fuente: Propia.

Ambos modelos destacan como variables dominantes al gasto a la fecha estimado, el presupuesto final estimado y el presupuesto del proyecto. Este resultado es coherente desde una perspectiva técnica, ya que estas variables cuantitativas reflejan el avance financiero y la estructura presupuestal del proyecto, dos aspectos importantes al momento de anticipar

sobrecostos. La fuerte influencia que ejercen sobre la variable objetivo indica que los modelos están captando correctamente la lógica financiera subyacente en la evolución del presupuesto.

Sin embargo, se observan diferencias en la ponderación de otras características, especialmente en variables asociadas al riesgo. En el modelo de *Gradient Boosting*, la variable impacto aparece como la tercera más importante, superando al presupuesto del proyecto. Por su parte, en el modelo *Random Forest*, impacto tiene un peso menor. Asimismo, la probabilidad de ocurrencia del riesgo tiene una relevancia moderada en ambos modelos, aunque nuevamente con mayor peso en *Gradient Boosting*.

Respecto a las variables de menor peso, ambos modelos coinciden en asignar poca relevancia a las características categóricas relacionadas con el tipo de proyecto o la fase de este (por ejemplo, “Fase del Proyecto_Scope”, “Tipo Proyecto_Expansión de Capacidad Escolar”, etc.). Estas variables pueden contener información contextual útil, pero no parecen ser determinantes para predecir el sobrecosto en comparación con las variables numéricas directas. Este resultado podría deberse a una codificación categórica con demasiadas clases dispersas, baja variabilidad o una limitada relación directa con la variable objetivo.

4.5 Validación

Con base en los resultados de la sección anterior, se establece el modelo *gradient boosting regressor* como el más relevante para el objetivo general del trabajo, se toma como base esta estructura y se realizan una serie de iteraciones cuya finalidad es obtener un modelo capaz de generalizar ante datos nunca vistos y con resultados superiores a los del modelo inicial.

4.5.1 Esquema de validación

Con el objetivo de evaluar la generalización del modelo entrenado y evitar el sobreajuste, se aplica validación cruzada en diferentes configuraciones:

- Validación cruzada con $cv=8$ y $cv=10$.

- División previa del conjunto de datos culminados en train/test (80/20) para evaluación final fuera de muestra.

4.5.2 Hiperparámetros evaluados

Los siguientes hiperparámetros fueron ajustados:

- *n_estimators*: Número de árboles
- *learning_rate*: Tasa de aprendizaje
- *max_depth*: Profundidad máxima del árbol
- *min_samples_split*: Mínimo de muestras para dividir un nodo
- *min_samples_leaf*: Muestras mínimas en una hoja
- *subsample*: Proporción de muestras utilizadas por árbol
- *max_features*: Máximo número de variables a considerar por división

Se implementó una búsqueda en malla (*GridSearchCV*) sobre múltiples combinaciones de hiperparámetros, a través de cinco iteraciones independientes, con variaciones tanto en el tamaño como en la granularidad de los valores de los hiperparámetros. En cada iteración se evaluaron los modelos con las siguientes métricas sobre el conjunto de prueba (test):

- **R²** (Coeficiente de determinación): mide el grado de ajuste del modelo.
- **MSE** (Error cuadrático medio): penaliza fuertemente los errores grandes.
- **MAE** (Desviación absoluta media): proporciona una medida robusta del error medio.

4.5.3 Iteraciones y resultados

En la **Tabla 4-6** se presentan los detalles de cada iteración, con enfoque en el tamaño de la cuadrícula y el número de pliegues para validación cruzada, además, en la **Figura 4-15** se muestra cómo se comporta cada métrica ante las variaciones de cada iteración, evidenciando que el mejor modelo se obtiene en la iteración número 4.

Tabla 4-6: Registro de iteraciones.

Iteración	CV	Tamaño del param_grid	Hiperparámetros modificados	Observaciones
1	8	576 combinaciones (4×4×3×3×3×2×2)	Exploración inicial. Se variaron todos los hiperparámetros.	Buen desempeño inicial. Sirvió como base para entender la sensibilidad del modelo ante variaciones amplias.
2	10	576 combinaciones (igual a Iteración 1)	Misma grilla que Iteración 1. Cambio en la validación cruzada (cv=10) para mejorar robustez de evaluación.	Validación más estricta. No se modificó la grilla, pero se buscó estabilidad del desempeño.
3	8	144 combinaciones (3×3×2×2×2×2×1)	Reducción de complejidad. Se limitaron los valores de: n_estimators, learning_rate, max_depth, min_samples_split, min_samples_leaf, y se dejó solo 'sqrt' para max_features.	Menor complejidad y tiempo de cómputo. Se sacrificó precisión para explorar modelos más simples.
4	8	135 combinaciones (3×5×3×3×3×3×3)	Mayor granularidad en learning_rate (de 0.01 a 0.1 con incrementos pequeños). Se incorporó max_features=0.5 además de 'sqrt' y 'log2'.	Mejor configuración encontrada. Ajuste sobresaliente y baja desviación. Balance óptimo entre complejidad y precisión.
5	8	180 combinaciones (4×5×3×3×3×3×3)	Misma estrategia que la Iteración 4, pero se añadió un valor más alto en n_estimators (hasta 400) para probar mejoras por profundidad del ensamble.	Resultados muy buenos, pero sin mejoras sobre Iteración 4. Aumento de estimadores no aportó beneficios adicionales significativos.

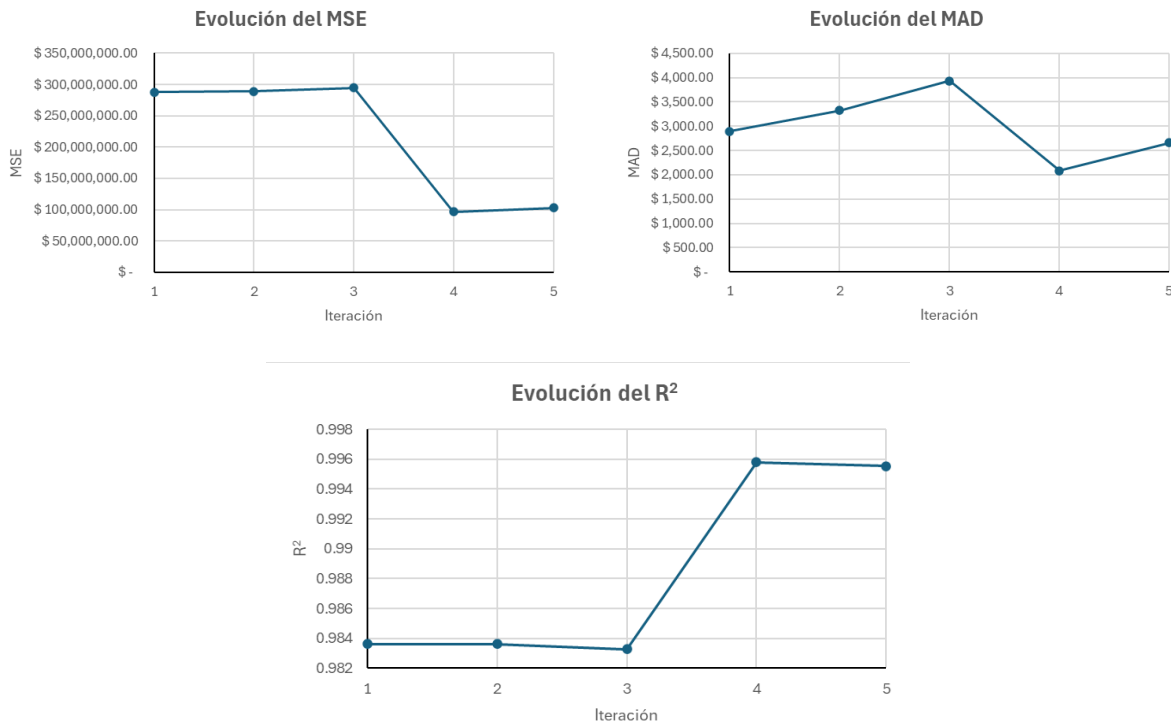
Fuente: Propia.

La estrategia de variación se puede resumir como sigue:

- **Iteración 1 → Iteración 2:** Misma grilla, variación en el método de validación (cv=10 en vez de cv=8).

- **Iteración 2 → Iteración 3:** Reducción drástica de la grilla, buscando un modelo más simple con menor carga computacional.
- **Iteración 3 → Iteración 4:** Se ajustó la grilla para permitir más precisión en `learning_rate` y se incorporaron valores mixtos en `max_features` para capturar interacciones adicionales.
- **Iteración 4 → Iteración 5:** Se mantuvo la configuración ganadora, pero se aumentó `n_estimators` para verificar si mayor profundidad del modelo aportaba mejoras.

Figura 4-15: Comportamiento de métricas por iteración.



Fuente: Propia.

Finalmente, se concluye que, a través de la estrategia de ajuste progresiva, basada en validación cruzada y experimentación controlada sobre múltiples combinaciones de hiperparámetros, se logró identificar un modelo de *Gradient Boosting Regressor* con alto poder predictivo. Este modelo servirá como herramienta para la estimación del gasto total en proyectos de construcción en progreso, aportando valor en la gestión financiera y la anticipación riesgos.

4.6 Despliegue

La fase de despliegue marca el punto en el que el modelo de *Machine Learning*, ya desarrollado y validado, se implementa en un entorno de producción para comenzar a generar predicciones con datos reales, y, por lo tanto, a generar impacto directo en la toma de decisiones del negocio.

Desplegar un modelo en producción presenta varios retos técnicos, especialmente porque el entorno de desarrollo y el de producción suelen ser diferentes. Estos entornos, conocidos como entornos de ejecución (*runtime environments*), son el conjunto de herramientas, versiones de software y configuraciones que permiten ejecutar el modelo (Stijnman, 2024).

Según Stijnman (2024), durante el desarrollo, el modelo se entrena en un entorno controlado, utilizando versiones específicas de Python, bibliotecas (como scikit-learn, pandas, etc.) y configuraciones locales. Sin embargo, cuando se traslada a producción, este entorno puede ser distinto: versiones desactualizadas, bibliotecas faltantes o sistemas operativos diferentes pueden causar que el modelo no funcione correctamente o que los resultados sean inconsistentes.

Para entenderlo mejor, se puede usar la analogía de una cocina: desarrollar un modelo en un entorno es como crear una receta en una cocina específica, con sus propios utensilios y condiciones. Si se lleva esa misma receta a otra cocina que tiene diferentes ingredientes o herramientas, el resultado puede variar. Lo mismo ocurre al mover un modelo de desarrollo a producción.

No obstante, es importante señalar que esta fase se encuentra fuera del objetivo del trabajo, dado que no se contempla la aplicación práctica del modelo en una organización específica, ni se cuenta con los recursos técnicos e infraestructurales necesarios (como ambientes de producción, herramientas de integración continua o plataformas de despliegue) para su implementación real.

5. Conclusiones y recomendaciones

5.1 Conclusiones

El modelo propuesto, construido con base en técnicas de *Machine Learning*, logró predecir desviaciones presupuestarias con un desempeño satisfactorio. El algoritmo *Gradient Boosting Regressor* fue el que presentó mejor rendimiento, alcanzado un coeficiente de determinación (R^2) de 0.99. Este resultado evidencia que es posible anticipar sobrecostos en proyectos de construcción de uso no residencial con un buen nivel de precisión cuando se dispone de datos históricos consistentes y bien estructurados.

Del análisis exploratorio y la evaluación estadística de las variables se concluyó que las variables numéricas, especialmente aquellas relacionadas con el presupuesto y ejecución financiera, fueron las más relevantes en la predicción de sobrecostos. Entre ellas, el gasto real estimado a la fecha, el presupuesto final estimado y el presupuesto total del proyecto mostraron una fuerte correlación con la desviación presupuestaria, mientras que las variables categóricas como la fase del proyecto o el tipo de edificación no aportaron valor explicativo significativo en los modelos predictivos.

Entre los modelos propuestos, el *Gradient Boosting Regressor* fue el que presentó el mejor desempeño al ser evaluado con el conjunto de prueba. Alcanzó un coeficiente de determinación (R^2) de 0.9824, un error cuadrático medio (MSE) de 309,699,558.66 y un error absoluto medio (MAE) de 3,887.30, superando a modelos como regresión lineal y redes neuronales. Estos resultados demuestran que el modelo fue capaz de capturar de manera precisa las variaciones en el gasto real, consolidándolo como la opción más robusta y precisa entre las alternativas evaluadas. Este hallazgo es coherente con estudios comparativos donde se muestra que, para datos tabulares con número moderado de

observaciones y variables heterogéneas, los modelos de ensamble basados en árboles como *Gradient Boosting* tienden a superar a las redes neuronales multicapa, salvo que estas últimas reciban un ajuste muy exhaustivo en arquitecturas, funciones de activación y regularización (Fernández-Delgado et al., 2014; Shwartz-Ziv & Armon, 2022). En particular, se ha documentado que los *Gradient Boosting Decision Trees* (GBDT) manejan de forma más eficiente interacciones y no linealidades con menor sensibilidad a hiperparámetros, esto los hace más competitivos cuando no se dispone de una gran cantidad de datos ni de una búsqueda extensiva de configuraciones para redes profundas. Así, el mejor rendimiento observado en este estudio se sustenta tanto en la evidencia empírica de la literatura como en las características propias del conjunto de datos utilizado.

El modelo seleccionado fue sometido a un proceso de validación cruzada con esquemas de K-fold con $cv=8$ y $cv=10$, además de una partición del conjunto de datos en train/test (80/20). Esta validación permitió verificar la capacidad de generalización del modelo ante datos no vistos, evidenciando estabilidad en las métricas de desempeño y ausencia de sobreajuste. La consistencia del R^2 y el MAE entre diferentes pliegues y conjuntos confirma que el modelo tiene una buena fiabilidad y puede ser aplicado.

La preparación de datos, incluyendo estimación de variables faltantes y exclusión de atributos sin valor explicativo, fue clave para mejorar la precisión del modelo. Esto resalta la necesidad de fortalecer la recolección y estandarización de datos en proyectos de construcción.

El análisis bibliométrico reveló que, el área de estudio aún se encuentra en una etapa inicial. Esto representa una oportunidad para que investigaciones futuras profundicen en la combinación de IA y análisis de riesgos específicos en la construcción.

Aunque este trabajo se centró en proyectos de edificación no residencial, la estructura metodológica puede adaptarse a otras áreas de infraestructura. Esto permitiría ampliar el impacto de la investigación y facilitar la toma de decisiones informada en diferentes contextos del sector de la construcción.

5.2 Recomendaciones

A diferencia de países como Estados Unidos, donde bases de datos abiertas como la de *School Construction Authority* permiten el desarrollo de modelos analíticos, en el contexto colombiano no se tienen repositorios abiertos con información estandarizada sobre presupuestos, cronogramas y ejecución de proyectos de construcción. Esta carencia restringe la posibilidad de replicar estudios similares, lo que evidencia la necesidad de promover políticas públicas orientadas a la creación de sistemas de información abiertos, interoperables y confiables, que fortalezcan la toma de decisiones basadas en evidencia en el sector de infraestructura.

En Colombia, la gestión y disponibilidad de información sobre proyectos de construcción presenta un grado de fragmentación. Las entidades públicas como el Departamento Nacional de Planeación (DNP), el Instituto Nacional de Vías (INVIAS) y la Agencia Nacional de Infraestructura (ANI) publican información sobre contratos y avances físicos-financieros en portales como SECOP II, pero los datos no siempre cuentan con el nivel de desagregación, estandarización y completitud necesario para alimentar modelos predictivos. Esta situación coincide con lo planteado por (Wahab & Wang, 2022), quienes señalan que la precisión de las estimaciones de costos está directamente condicionada por la calidad y consistencia de los registros históricos.

En el sector público, una acción realista y de impacto sería consolidar, a través del DNP, un repositorio único que unifique la información ya existente en distintas plataformas (SECOP, MÍO, SINERGIA, Observatorios de Infraestructura), incorporando campos como presupuesto inicial, costos ejecutados por fases, modificaciones contractuales y causas de variaciones. En lugar de crear un sistema desde cero, esta estrategia se apoyaría en bases ya disponibles, pero las enriquecería con un formato homogéneo y criterios claros de reporte, como lo sugiere (Tayefeh Hashemi et al., 2020) en su discusión sobre estandarización de datos para *Machine Learning* en construcción.

En el sector privado, las empresas constructoras y firmas interventoras poseen registros detallados que, por motivos de confidencialidad, rara vez se comparten. Sin embargo, siguiendo el modelo de *data trusts* utilizado en Reino Unido (ODI, 2019), se podría promover la creación de acuerdos de intercambio de datos anonimizados gestionados por gremios como Camacol. Estos acuerdos permitirían a las empresas aportar información de costos, cronogramas y variaciones, eliminando datos sensibles, pero manteniendo su utilidad estadística para investigación y desarrollo.

La combinación de estas dos acciones, integración y estandarización de datos públicos ya existentes, junto con mecanismos de colaboración privada segura, podría representar un camino factible para alimentar modelos predictivos como el desarrollado en esta investigación.

En el contexto colombiano, un modelo de este tipo podría aplicarse inicialmente en proyectos de infraestructura pública, por ejemplo, edificaciones educativas, hospitales o sedes administrativas, donde las entidades contratantes ya cuentan con registros de presupuesto y ejecución, aunque dispersos en plataformas como SECOP II y el Observatorio de Infraestructura. Con información más estructurada, sería posible anticipar desviaciones presupuestales en etapas tempranas, orientar la asignación de recursos y priorizar intervenciones preventivas en los proyectos con mayor riesgo de sobrecosto.

En el sector privado, su implementación sería factible en empresas con procesos maduros de control de costos, que registren de manera sistemática presupuestos iniciales, avances físicos y financieros, y modificaciones contractuales. En este entorno, el modelo podría apoyar la estimación de costos de obras similares basadas en experiencias previas.

De manera progresiva, y conforme se fortalezca la disponibilidad y calidad de los datos, este tipo de modelo podría convertirse en una herramienta de apoyo para la toma de decisiones estratégicas en proyectos de construcción, no como un sustituto del criterio profesional, sino como un complemento que reduce la incertidumbre y aumenta la objetividad en la gestión de riesgos.

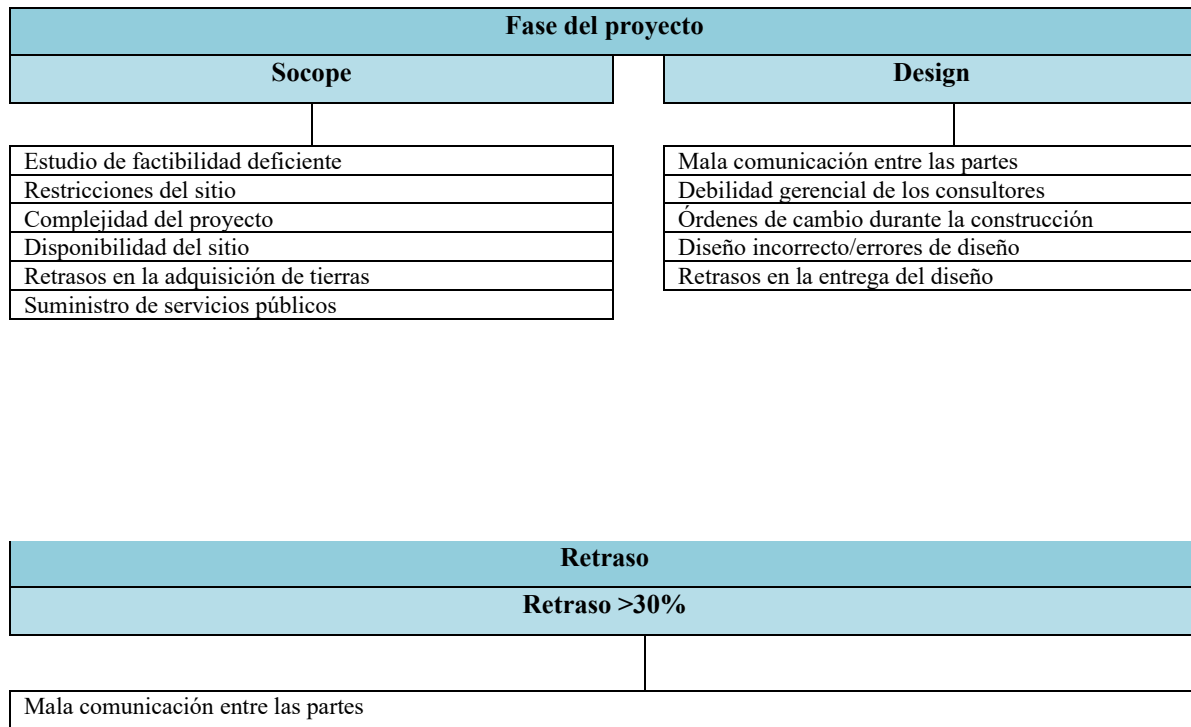
A. Anexo: Factores de riesgo que implican sobrecostos en proyectos

Factor de riesgo	Fuente	Media de Probabilidad	Std P	Impacto Medio	Std I
Estudio de factibilidad deficiente	Directivo	3.28	1.21	3.95	1.02
Debilidad de la gestión de los contratistas	Directivo	3.05	1.05	4.28	0.76
Mala comunicación entre las partes	Directivo	2.59	0.68	3.28	1.02
Conflicto entre las partes del proyecto	Directivo	2.67	1.03	3.33	0.98
Debilidad gerencial de los consultores	Directivo	2.90	0.97	3.82	0.82
Aumento del precio de los materiales	Materiales y equipos	4.59	0.55	4.74	0.59
Escasez de equipos	Materiales y equipos	2.79	1.20	3.18	1.27
Retraso de los proveedores en la entrega de los equipos a la obra	Materiales y equipos	2.74	1.04	3.23	1.13
Escasez de materiales	Materiales y equipos	2.56	1.19	3.18	1.23
Nuevos problemas de equipos/tecnología	Materiales y equipos	2.23	1.20	2.69	1.19
Falta de conocimiento y experiencia	Personal	2.74	0.88	3.28	1.02
Escasez de mano de obra	Personal	2.18	1.10	3.26	1.07
Falta de personal calificado (personal técnico) en el sitio	Personal	2.61	1.14	3.56	1.05
Tipo de cambio de divisas	Financiero	4.20	1.13	4.49	0.91
Inflación	Financiero	4.69	0.52	4.85	0.36
Escasez de fondos del propietario y retrasos en los pagos	Financiero	4.05	1.10	4.36	0.84
Múltiples fuentes de fondos	Financiero	2.54	1.00	3.13	1.15
Escasez de fondos para contratistas	Financiero	3.38	0.78	3.92	1.06
Cambios adversos en las condiciones geológicas	Proyecto	2.10	1.14	3.08	1.26
Restricciones del sitio	Proyecto	2.23	1.01	2.69	1.05
Complejidad del proyecto	Proyecto	2.51	1.05	3.20	1.15
Disponibilidad del sitio	Dueño	2.26	0.97	3.00	1.32
Órdenes de cambio durante la construcción	Dueño	3.44	1.16	3.77	0.96
Retrasos en la toma de decisiones	Dueño	3.38	1.02	3.77	0.96
Política aduanera del propietario y complejidad (demora en la contratación)	Dueño	2.95	1.34	3.48	1.33
Retrasos en la adquisición de tierras	Dueño	2.54	1.21	3.28	1.39
Suministro de servicios públicos	Dueño	4.08	0.84	2.13	1.00
Selección del postor más bajo	Dueño	3.67	1.11	3.74	1.12
Falta de conocimiento y experiencia	Contratista	3.20	0.92	3.85	0.93
Retrasos en las adquisiciones	Contratista	2.85	0.84	3.54	0.91
Retrasos de los subcontratistas en trabajos anteriores	Contratista	3.05	0.97	3.33	1.06
Gestión financiera inadecuada	Contratista	3.15	1.09	3.77	0.90
Seguridad en el sitio	Contratista	2.98	1.22	3.38	1.39

Factor de riesgo	Fuente	Media de Probabilidad	Std P	Impacto Medio	Std I
Calidad de la construcción (defectos)	Contratista	3.10	1.12	3.74	1.19
Mala planificación y programación	Contratista	3.28	1.19	3.97	0.90
Falta de conocimiento y experiencia	Consultor	2.74	1.09	3.67	1.08
Diseño incorrecto/errores de diseño	Consultor	2.95	1.02	3.79	1.13
Retrasos en la entrega del diseño	Consultor	2.70	1.03	3.49	0.97
Cambio de equipo, o especificación de equipo, durante la construcción	Consultor	2.64	0.99	3.26	1.07
Mal tiempo o condición de emergencia	Medio ambiente	2.64	0.93	3.28	0.94
Bajas/lesiones inesperadas	Medio ambiente	1.77	1.01	2.36	1.33
Ley de preservación del medio ambiente	Medio ambiente	1.46	0.82	1.92	1.18

Fuente: (Ashtari et al., 2022)

B. Anexo: Relación de los factores de riesgo con variables de los proyectos



Aumento del precio de los materiales			
Escasez de equipos			
Retraso de los proveedores en la entrega de los equipos a la obra			
Escasez de materiales			
Escasez de mano de obra			
Restricciones del sitio			
Complejidad del proyecto			
Órdenes de cambio durante la construcción			
Suministro de servicios públicos			
Retrasos en las adquisiciones			
Retrasos de los subcontratistas en trabajos anteriores			
Calidad de la construcción (defectos)			
Mala planificación y programación			
Diseño incorrecto/errores de diseño			
Retrasos en la entrega del diseño			
Mal tiempo o condición de emergencia			
Presupuesto del proyecto			
Rango 1 Presupuesto ≤ q(0.25)	Rango 2 q(0.25) > presupuesto ≤ q(0.75)	Rango 3 q(0.75) < presupuesto ≤ q(0.9)	Rango 4 Presupuesto > q(0.9)
Falta de personal calificado (personal técnico) en el sitio	Órdenes de cambio durante la construcción	Inflación	Tipo de cambio de divisas
Escasez de fondos del propietario y retrasos en los pagos	Retrasos en la toma de decisiones	Restricciones del sitio	Escasez de fondos para contratistas
Restricciones del sitio	Retrasos en la adquisición de tierras	Mala planificación y programación	Política aduanera del propietario y complejidad (demora en la contratación)

Fuente: Propia.

C. Anexo: Modelo de bosques aleatorios

```
from sklearn.ensemble import RandomForestRegressor
from sklearn.preprocessing import OneHotEncoder
from sklearn.compose import ColumnTransformer
from sklearn.pipeline import Pipeline
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_absolute_error, r2_score
from sklearn.model_selection import train_test_split

df_train = project_2.copy()

# Entrenamiento del modelo con datos de proyectos en progreso
X = df_train[['Tipo Proyecto',
              'Categoria Proyecto', 'Fase del Proyecto',
              'Presupuesto del Proyecto', 'Presupuesto Final Estimado',
              'Desviación Presupuestaria', 'Duración Proyectada']]
y = df_train['Cantidad Gastada a la Fecha']

# División de datos para prueba/entrenamiento
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
                                                    random_state=42)

# Codificación de variables categóricas
categorical_cols = ['Tipo Proyecto', 'Categoria Proyecto', 'Fase del Proyecto']
numeric_cols = ['Presupuesto del Proyecto', 'Presupuesto Final Estimado',
                'Desviación Presupuestaria', 'Duración Proyectada']

preprocessor = ColumnTransformer([
    ('cat', OneHotEncoder(handle_unknown='ignore'), categorical_cols)
], remainder='passthrough')

modelo_gasto_a_fecha = Pipeline(steps=[
    ('preprocessor', preprocessor),
    ('regressor', RandomForestRegressor(n_estimators=100, random_state=42))
])

# entrenamiento del modelo con datos de entrenamiento
modelo_gasto_a_fecha.fit(X_train, y_train)

# Predecir en test
y_pred = modelo_gasto_a_fecha.predict(X_test)
```

D. Instrucciones y uso del repositorio

El repositorio disponible públicamente en GitHub: <https://github.com/Anfospina/cost-overflow-prediction> documenta el desarrollo de un proyecto orientado a predecir sobrecostos en proyectos de construcción de edificación, utilizando técnicas de análisis de datos y algoritmos de aprendizaje automático (*Machine Learning*). Su estructura está organizada para facilitar la comprensión, reutilización y mejora del proceso. A continuación, se describe cada una de sus partes:

1. Carpeta *_InternalFunctions*

Contiene funciones auxiliares que se utilizan en las primeras etapas del análisis, especialmente para:

- **Limpieza de datos:** Eliminar o corregir valores erróneos o ausentes.
- **Formateo:** Asegurar que los datos estén en el formato adecuado para ser procesados.
- **Generación de características:** Crear nuevas variables (o "*features*") que capturan mejor la información contenida en los datos originales.

Estas funciones están diseñadas para ser reutilizables con otras bases de datos que tengan estructuras similares, lo que las hace adaptables y escalables. Se recomienda no modificar su contenido directamente para mantener la compatibilidad del flujo de trabajo.

2. Carpeta *_ModelFunctions*

Incluye funciones que permiten construir, entrenar y evaluar diferentes modelos de predicción. Estas funciones se enfocan específicamente en la lógica de modelado y se han formulado para ser compatibles con varios algoritmos de *Machine Learning*.

3. Carpeta *Input_output*

En esta carpeta se encuentran tanto los datos originales utilizados en el análisis como los resultados intermedios que se generan a lo largo del proceso:

- **Datos de entrada:** La base de datos inicial con información sobre proyectos de construcción.
- **Datos de salida:** Conjuntos de datos ya procesados, con variables limpias y nuevas características listas para ser usadas por los modelos predictivos.

Esto permite reproducir fácilmente los resultados sin necesidad de rehacer todo el proceso desde cero.

4. *Archivo development.ipynb*

Es un cuaderno interactivo de Jupyter Notebook que actúa como núcleo del proyecto. Presenta de manera estructurada el flujo completo de trabajo, desde el inicio hasta la generación de modelos. En él se puede seguir paso a paso:

- La limpieza y transformación de los datos.
- La creación de variables derivadas.
- **La selección y entrenamiento de cuatro modelos de predicción:** *Linear Regression, Random Forest Regressor, Gradient Boosting Regressor, MLP regressor*

Este archivo es ideal para comprender el enfoque metodológico y verificar cómo se implementaron las funciones desarrolladas en las carpetas anteriores.

5. *Archivo model_optimization.py*

Este script en formato .py se enfoca en mejorar el rendimiento del modelo con mejores resultados obtenidos previamente, que en este caso fue el *Gradient Boosting Regressor*.

Durante este proceso de optimización se:

- Prueban distintas combinaciones de parámetros (técnica conocida como ajuste de hiperparámetros).
- Calculan métricas de desempeño para cada variante del modelo.
- Guardan automáticamente los resultados de cada iteración.

El modelo que muestra el mejor desempeño final es almacenado en la carpeta *optimization*, con el nombre *GradientBoosting_4.pkl.gz*, listo para ser usado en nuevas predicciones.

El archivo *GradientBoosting_4.pkl.gz* corresponde al modelo predictivo con mejor desempeño obtenido durante el desarrollo del trabajo. Este modelo fue entrenado para estimar el valor real de *costos en proyectos de construcción*, y se encuentra listo para ser reutilizado en nuevos conjuntos de datos. A continuación, se explica cómo utilizarlo y qué resultados puede ofrecer.

El uso del modelo implica seguir tres pasos: primero, cargar el modelo en memoria utilizando una función adecuada que lea el archivo *.pkl.gz*; segundo, preparar los datos de entrada, asegurándose de que estén preprocesados de forma idéntica a los datos usados en el entrenamiento (mismos nombres y tipos de columnas, sin datos faltantes o errores de formato); y tercero, aplicar el modelo sobre los nuevos datos para obtener las predicciones del valor objetivo.

Al aplicarse sobre un conjunto de datos, el modelo genera una estimación numérica específica, por ejemplo, el costo real esperado de ejecución para cada proyecto teniendo en cuenta las características particulares de cada uno. Estas predicciones se devuelven como una lista o columna de valores, en el mismo orden en que se presentaron los datos de entrada. Cada valor representa una proyección basada en el conocimiento aprendido por el modelo durante su entrenamiento, el cual integra relaciones complejas entre múltiples variables, tanto numéricas como categóricas.

El modelo opera de manera automática y consistente: por cada nuevo caso, entrega una estimación puntual que puede ser utilizada directamente para análisis, comparación con presupuestos iniciales o decisiones operativas en el contexto de control de costos y riesgos en proyectos. En resumen, el modelo devuelve un valor predictivo por cada fila de datos que se le proporcione, permitiendo así anticipar posibles desviaciones presupuestales antes de que ocurran.

Bibliografía

- Abioye, S. O., Oyedele, L. O., Akanbi, L., Ajayi, A., Dávila 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. *Journal of Building Engineering*, *44*, 103299. <https://doi.org/10.1016/J.JOBE.2021.103299>
- Aggabou, L. K., Lakehal, B., & Mouda, M. (2024). An Artificial Neural Network Approach for Construction Project Risk Management. *International Journal of Safety and Security Engineering*, *14*(2), 553–561. <https://doi.org/10.18280/ijssse.140222>
- Akoglu, H. (2018). User's guide to correlation coefficients. In *Turkish Journal of Emergency Medicine* (Vol. 18, pp. 91–93). Emergency Medicine Association of Turkey. <https://doi.org/10.1016/j.tjem.2018.08.001>
- Al-Nahas, Y. S., Hadidi, L. A., Islam, M. S., Skitmore, M., & Abunada, Z. (2024). Modified Mamdani-fuzzy inference system for predicting the cost overrun of construction projects. *Applied Soft Computing*, *151*. <https://doi.org/10.1016/j.asoc.2023.111152>
- Ammar, T., Abdel-Monem, M., & El-Dash, K. (2022). Risk factors causing cost overruns in road networks. *Ain Shams Engineering Journal*, *13*. <https://doi.org/10.1016/j.asej.2022.101720>
- Arabzadeh, V., Niaki, S. T. A., & Arabzadeh, V. (2018). Construction cost estimation of spherical storage tanks: artificial neural networks and hybrid regression—GA algorithms. *Journal of Industrial Engineering International*, *14*, 747–756. <https://doi.org/10.1007/s40092-017-0240-8>
- Arthur, A. (2021). *Construction Risk Management Decision Making*.
- Ashtari, M. A., Ansari, R., Hassannayebi, E., & Jeong, J. (2022). Cost Overrun Risk Assessment and Prediction in Construction Projects: A Bayesian Network Classifier Approach. *Buildings*, *12*(10). <https://doi.org/10.3390/buildings12101660>

- Axelos. (2017). *Managing Successful Projects with PRINCE2®*. TSO (The Stationery Office).
- Banaitiene, N., Banaitis, A., & Norkus, A. (2011). Risk management in projects: Peculiarities of Lithuanian construction companies. *International Journal of Strategic Property Management*, *15*, 60–73. <https://doi.org/10.3846/1648715X.2011.568675>
- Biau, G., & Cadre, B. (2021). Optimization by Gradient Boosting. In *Advances in Contemporary Statistics and Econometrics: Festschrift in Honor of Christine Thomas-Agnan* (pp. 23–44). Springer International Publishing. https://doi.org/10.1007/978-3-030-73249-3_2
- Borujeni, S. E., Nannapaneni, S., Nguyen, N. H., Behrman, E. C., & Steck, J. E. (2021). Quantum circuit representation of Bayesian networks. *Expert Systems with Applications*, *176*. <https://doi.org/10.1016/j.eswa.2021.114768>
- Chattapadhyay, D. B., Putta, J., & Rama Mohan Rao, P. (2021). Risk identification, assessments, and prediction for mega construction projects: A risk prediction paradigm based on cross analytical-machine learning model. *Buildings*, *11*(4). <https://doi.org/10.3390/buildings11040172>
- Cheng, M.-Y., & Darsa, M. H. (2021). Construction schedule risk assessment and management strategy for foreign general contractors working in the Ethiopian construction industry. *Sustainability (Switzerland)*, *13*(14). <https://doi.org/10.3390/su13147830>
- Chenya, L., Aminudin, E., Mohd, S., & Yap, L. S. (2022). Intelligent Risk Management in Construction Projects: Systematic Literature Review. *IEEE Access*, *10*, 72936–72954. <https://doi.org/10.1109/ACCESS.2022.3189157>
- Chien, C. F., Dauzère-Pérès, S., Huh, W. T., Jang, Y. J., & Morrison, J. R. (2020). Artificial intelligence in manufacturing and logistics systems: algorithms, applications, and case studies. In *International Journal of Production Research* (Vol. 58, pp. 2730–2731). Taylor and Francis Ltd. <https://doi.org/10.1080/00207543.2020.1752488>
- Dampfhofer, M., Mesquida, T., Valentian, A., & Anghel, L. (2023). Backpropagation-Based Learning Techniques for Deep Spiking Neural Networks: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*. <https://doi.org/10.1109/TNNLS.2023.3263008>

- Donthu, N., Kumar, S., Mukherjee, D., Pandey, N., & Lim, W. M. (2021). How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, *133*, 285–296. <https://doi.org/10.1016/j.jbusres.2021.04.070>
- Fayek, A. R. (2020). Fuzzy Logic and Fuzzy Hybrid Techniques for Construction Engineering and Management. *Journal of Construction Engineering and Management*, *146*. [https://doi.org/10.1061/\(asce\)co.1943-7862.0001854](https://doi.org/10.1061/(asce)co.1943-7862.0001854)
- Fernández-Delgado, M., Cernadas, E., Barro, S., Amorim, D., & Fernández-Delgado, A. (2014). Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? *Journal of Machine Learning Research*, *15*, 3133–3181. <http://www.mathworks.es/products/neural-network>.
- Flanagan, R., & Norman, G. (1993). Developing Guideline for Risk Management of Tunnel Construction in Ethiopia. *Open Journal of Safety Science and Technology*, *11*, 171–183. <https://doi.org/10.4236/ojsst.2021.114012>
- ForouzeshNejad, A. A., Arabikhan, F., & Aheleroff, S. (2024). Optimizing Project Time and Cost Prediction Using a Hybrid XGBoost and Simulated Annealing Algorithm. *Machines*, *12*. <https://doi.org/10.3390/machines12120867>
- George, M. R., Nalluri, M. R., & Anand, K. B. (2022). Application of Ensemble Machine Learning for Construction Safety Risk Assessment. *Journal of The Institution of Engineers (India): Series A*, *103*(4), 989–1003. <https://doi.org/10.1007/s40030-022-00690-w>
- Goh, S. C., Elliott, C., & Richards, G. (2017). Setting the context for analytics: Performance management in Canadian public organizations: Findings of a multi-case study. In *Big Data and Analytics Applications in Government: Current Practices and Future Opportunities* (pp. 29–55). CRC Press. <https://doi.org/10.4324/9781315153582>
- González Alcaide, G., Valderrama Zurián, J. C., Aleixandre Benavent, R., & González De Dios, J. (2011). La investigación pediátrica Española en Anales de Pediatría: Grupos y ámbitos temáticos (2003-2009). *Anales de Pediatría*, *74*, 239–254. <https://doi.org/10.1016/j.anpedi.2010.10.023>
- Guelman, L. (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction. *Expert Systems with Applications*, *39*, 3659–3667. <https://doi.org/10.1016/j.eswa.2011.09.058>

- Hammad, A., AbouRizk, S., & Mohamed, Y. (2014). Application of KDD Techniques to Extract Useful Knowledge from Labor Resources Data in Industrial Construction Projects. *Journal of Management in Engineering*, 30. [https://doi.org/10.1061/\(asce\)me.1943-5479.0000280](https://doi.org/10.1061/(asce)me.1943-5479.0000280)
- Hasan, N., & Singh, M. (2015). Library and Information Science Research Output: A study based on Web of Science. *Collnet Journal of Scientometrics and Information Management*, 9, 47–64. <https://doi.org/10.1080/09737766.2015.1027089>
- Hazelton, M. L. (2009). Univariate Linear Regression. In *International Encyclopedia of Education, Third Edition* (pp. 482–488). Elsevier. <https://doi.org/10.1016/B978-0-08-044894-7.01373-7>
- Hegde, J., & Rokseth, B. (2020). Applications of machine learning methods for engineering risk assessment – A review. In *Safety Science* (Vol. 122). Elsevier B.V. <https://doi.org/10.1016/j.ssci.2019.09.015>
- Hon, C. K. H., Sun, C., Xia, B., Jimmieson, N. L., Way, K. A., & Wu, P. P. Y. (2022). Applications of Bayesian approaches in construction management research: a systematic review. *Engineering, Construction and Architectural Management*, 29, 2153–2182. <https://doi.org/10.1108/ECAM-10-2020-0817>
- Huang, M. (2020). Theory and Implementation of linear regression. *Proceedings - 2020 International Conference on Computer Vision, Image and Deep Learning, CVIDL 2020*, 210–217. <https://doi.org/10.1109/CVIDL51233.2020.00-99>
- Kadume, N. H., & Naji, H. I. (2021). Building Schedule Risks Simulation by Using BIM with Monte Carlo Technique. *IOP Conference Series: Earth and Environmental Science*, 856. <https://doi.org/10.1088/1755-1315/856/1/012059>
- Katal, A., & Singh, N. (2022). Artificial Neural Network: Models, Applications, and Challenges. In *EAI/Springer Innovations in Communication and Computing* (pp. 235–257). Springer Science and Business Media Deutschland GmbH. https://doi.org/10.1007/978-3-030-78284-9_11
- Khodabakhshian, A. (2023). *MACHINE LEARNING FOR RISK MANAGEMENT IN CONSTRUCTION PROJECTS*. Politecnico Milano.
- Khodabakhshian, A., Malsagov, U., & Re Cecconi, F. (2024). Machine Learning Application in Construction Delay and Cost Overrun Risks Assessment. *Lecture Notes*

- in Networks and Systems*, 921 LNNS, 222–240. https://doi.org/10.1007/978-3-031-54053-0_17
- Khodabakhshian, A., Puolitaival, T., & Kestle, L. (2023). Deterministic and Probabilistic Risk Management Approaches in Construction Projects: A Systematic Literature Review and Comparative Analysis. *Buildings*. <https://doi.org/10.3390/buildings13051312>
- Kim, T. K. (2017). Understanding one-way anova using conceptual figures. *Korean Journal of Anesthesiology*, 70, 22–26. <https://doi.org/10.4097/kjae.2017.70.1.22>
- Konstantinov, A., Utkin, L., & Muliukha, V. (2021). Gradient boosting machine with partially randomized decision trees. *Conference of Open Innovation Association, FRUCT, 2021-January*. <https://doi.org/10.23919/FRUCT50888.2021.9347631>
- Kulkarni, V. Y., Petare, M., & Sinha, P. K. (2014). Analyzing random forest classifier with different split measures. *Advances in Intelligent Systems and Computing*, 236, 691–699. https://doi.org/10.1007/978-81-322-1602-5_74
- Li, J., Wang, J., Xu, N., Hu, Y., & Cui, C. (2018). Importance degree research of safety risk management processes of urban rail transit based on text mining method. *Information (Switzerland)*, 9. <https://doi.org/10.3390/info9020026>
- Liu, W., Zhao, T., Zhou, W., & Tang, J. (2018). Safety risk factors of metro tunnel construction in China: An integrated study with EFA and SEM. *Safety Science*, 105, 98–113. <https://doi.org/10.1016/j.ssci.2018.01.009>
- Moon, S., Chi, S., & Im, S.-B. (2022). Automated detection of contractual risk clauses from construction specifications using bidirectional encoder representations from transformers (BERT). *Automation in Construction*, 142. <https://doi.org/10.1016/j.autcon.2022.104465>
- Nikas, A., Poulymenakou, A., & Kriaris, P. (2007). Investigating antecedents and drivers affecting the adoption of collaboration technologies in the construction industry. *Automation in Construction*, 16, 632–641. <https://doi.org/10.1016/j.autcon.2006.10.003>
- Nyqvist, R., Peltokorpi, A., & Seppänen, O. (2024). Can ChatGPT exceed humans in construction project risk management? *Engineering, Construction and Architectural Management*, 31(13), 223–243. <https://doi.org/10.1108/ECAM-08-2023-0819>

- Obregón, L., Orozco, C., Camargo, J., Duarte, J., & Valencia, G. (2019). Research trend on nuclear energy from 2008 to 2018: A bibliometric analysis. *International Journal of Energy Economics and Policy*, 9, 542–551. <https://doi.org/10.32479/ijee.8515>
- ODI. (2019). *Data trusts: lessons from three pilots (report)*. <https://theodi.org/insights/reports/odi-data-trusts-report/>
- Okudan, O., Budayan, C., & Dikmen, I. (2021). A knowledge-based risk management tool for construction projects using case-based reasoning. *Expert Systems with Applications*, 173. <https://doi.org/10.1016/j.eswa.2021.114776>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *Journal of Clinical Epidemiology*, 134, 178–189. <https://doi.org/10.1016/j.jclinepi.2021.03.001>
- Pretnar Žagar, A., & Demšar, J. (2022). Model Evaluation: How to Accurately Evaluate Predictive Models. In *Tourism on the Verge: Vol. Part F1051* (pp. 253–274). Springer Nature. https://doi.org/10.1007/978-3-030-88389-8_13
- Project Management Institute. (2017). *A guide to the project management body of knowledge (PMBOK® guide) – Sixth edition*. Project Management Institute.
- Project Management Institute. (2021). *A guide to the project management body of knowledge (PMBOK® guide) – Seventh edition*. Project Management Institute.
- Rao, T. V. N., Gaddam, A., Kurni, M., & Saritha, K. (2021). Reliance on artificial intelligence, machine learning and deep learning in the era of industry 4.0. In *Smart Healthcare System Design: Security and Privacy Aspects* (pp. 281–300). Wiley. <https://doi.org/10.1002/9781119792253.ch12>
- Ravindran, D., & Deepak, S. (2023). Bibliometric Analysis of Network Marketing for Business Sustainability Using Co-citation Method. *Studies in Computational Intelligence*, 1113, 299–309. https://doi.org/10.1007/978-3-031-43300-9_25
- Sanni-Anibire, M. O., Zin, R. M., & Olatunji, S. O. (2021). Machine learning - Based framework for construction delay mitigation. *Journal of Information Technology in Construction*, 26, 303–318. <https://doi.org/10.36680/j.itcon.2021.017>

- Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, *81*, 84–90. <https://doi.org/10.1016/J.INFFUS.2021.11.011>
- Sohrabinejad, A., & Rahimi, M. (2015). Risk Determination, Prioritization, and Classifying in Construction Project Case Study: Gharb Tehran Commercial-Administrative Complex. *Journal of Construction Engineering*, *2015*, 1–10. <https://doi.org/10.1155/2015/203468>
- Stijnman, F. (2024). MLOps Concepts. *Deploying Machine Learning into Production [DataCamp]*.
- Tayefeh Hashemi, S., Ebadati, O. M., & Kaur, H. (2020). Cost estimation and prediction in construction projects: a systematic review on machine learning techniques. In *SN Applied Sciences* (Vol. 2). Springer Nature. <https://doi.org/10.1007/s42452-020-03497-1>
- Teja, S., & Ch, S. A. (2017). Risk Management in Construction Equipment. *International Journal of Civil Engineering and Technology*, *8*(5), 160–167. <http://iaeme.com/Home/issue/IJCIET?Volume=8&Issue=5><http://iaeme.com>
- 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>
- Višić, M. (2022). CONNECTING PUZZLE PIECES: SYSTEMATIC LITERATURE REVIEW METHOD IN THE SOCIAL SCIENCES. *Sociologija*, *64*, 543. <https://doi.org/10.2298/SOC2204543V>
- Wahab, A., & Wang, J. (2022). Factors-driven comparison between BIM-based and traditional 2D quantity takeoff in construction cost estimation. *Engineering, Construction and Architectural Management*, *29*, 702–715. <https://doi.org/10.1108/ECAM-10-2020-0823>
- Weng, J. (Connor). (2023). *Putting Intellectual Robots to Work: Implementing Generative AI Tools in Project Management*. <https://Archive.Nyu.Edu/Handle/2451/69531>.
- Yaseen, Z. M., Ali, Z. H., Salih, S. Q., & Al-Ansari, N. (2020). Prediction of risk delay in construction projects using a hybrid artificial intelligence model. *Sustainability (Switzerland)*, *12*(4). <https://doi.org/10.3390/su12041514>

-
- Yi, Z., & Luo, X. (2024). Construction Cost Estimation Model and Dynamic Management Control Analysis Based on Artificial Intelligence. *Iranian Journal of Science and Technology - Transactions of Civil Engineering*, 48(1), 577–588.
<https://doi.org/10.1007/s40996-023-01173-z>
- Zou, P. X. W., Zhang, G., & Wang, J. (2007). Understanding the key risks in construction projects in China. *International Journal of Project Management*, 25, 601–614.
<https://doi.org/10.1016/j.ijproman.2007.03.001>