

Universidad Católica de Santa María

Facultad de Arquitectura e Ingenierías Civil y del Ambiente

Escuela Profesional de Ingeniería Civil

Desarrollo de modelos de machine learning para la predicción de sobrecostos y retrasos en obras públicas: estudio de caso de la región Arequipa, Perú

Tesis presentada por:

Retamozo Garcia, Jafet Amir

ORCID: 0009-0001-0603-766X

para optar el Título Profesional de Ingeniero Civil

Asesora:

Mgter. Torres Almiron, Jeniffer Carla

ORCID: 0000-0002-3417-4581

Arequipa - Perú

2026

UCSM-ERP

UNIVERSIDAD CATÓLICA DE SANTA MARÍA

INGENIERIA CIVIL

TITULACIÓN CON TESIS

DICTAMEN APROBACIÓN DE BORRADOR

Arequipa, 01 de Octubre del 2026

Dictamen: 016793-C-EPIC-2026

Visto el borrador del expediente 016793, presentado por:

2018802051 - RETAMOZO GARCIA JAFET AMIR

Titulado:

DESARROLLO DE MODELOS DE MACHINE LEARNING PARA LA PREDICCIÓN DE SOBRECOSTOS Y RETRASOS EN OBRAS PÚBLICAS: ESTUDIO DE CASO DE LA REGIÓN AREQUIPA, PERÚ

Nuestro dictamen es:

APROBADO

Título Profesional/Titulo de Segunda Especialidad/Grado Académico a optar:

INGENIERO CIVIL

29245617 - DIAZ GALDOS MIGUEL RENATO
DICTAMINADOR

46735424 - LOAYZA RODRIGUEZ JOSE GERMAN
DICTAMINADOR

46309156 - MONTOYA VILLANUEVA FILIBERTO RODY
DICTAMINADOR

DESARROLLO DE MODELOS DE MACHINE LEARNING PARA LA PREDICCIÓN DE SOBRECOSTOS Y RETRASOS EN OBRAS PÚBLICAS: ESTUDIO DE CASO DE LA REGIÓN AREQUIPA, PERÚ

INFORME DE ORIGINALIDAD

9%	8%	3%	4%
INDICE DE SIMILITUD	FUENTES DE INTERNET	PUBLICACIONES	TRABAJOS DEL ESTUDIANTE

FUENTES PRIMARIAS

1	Submitted to Universidad Católica de Santa María Trabajo del estudiante	1%
2	repositorio.ucsm.edu.pe Fuente de Internet	1%
3	hdl.handle.net Fuente de Internet	1%
4	docplayer.fi Fuente de Internet	<1%
5	repositorio.upt.edu.pe Fuente de Internet	<1%
6	repositorio.upeu.edu.pe Fuente de Internet	<1%
7	www.coursehero.com Fuente de Internet	<1%
8	Submitted to Universidad Europea de Madrid Trabajo del estudiante	<1%
9	repositorio.continental.edu.pe Fuente de Internet	<1%
10	Submitted to Universidad ESAN -- Escuela de Administración de Negocios para Graduados Trabajo del estudiante	<1%
11	dspace.unach.edu.ec Fuente de Internet	<1%
12	repositorioacademico.upc.edu.pe Fuente de Internet	<1%
13	up-rid.up.ac.pa Fuente de Internet	<1%
14	Submitted to Universitat Oberta de Catalunya Trabajo del estudiante	<1%

Dedicatoria

Dedico esta tesis a mis padres, en sincero reconocimiento a su entrega constante y a sus sacrificios para brindarme siempre lo mejor.

Agradecimiento

A mis padres y a mi familia, por ser mi mayor sustento y por brindarme los medios necesarios para culminar mi formación profesional.

A mi novia, por su compañía y respaldo constante a lo largo de este camino.

De manera especial, a los ingenieros Jeniffer Torres Almiron, Filiberto Rody Montoya Villanueva, Milagros Guillén Málaga y Carlos Sánchez Carigga, por su valiosa participación en el equipo de investigación, así como por su compromiso para hacer posible la ejecución de este estudio.

Reconocimiento

La presente investigación se desarrolló con el respaldo financiero de la Universidad Católica de Santa María, habiendo resultado seleccionada en la convocatoria "Fondo Concursable para la Investigación Científica y Desarrollo Tecnológico UCSM 2025". El trabajo contó con una asignación presupuestal de S/ 40,000.00 (cuarenta mil con 00/100 soles), formalizada mediante la Resolución N.° 31951-R-2025.

Los resultados de este estudio están orientados a su publicación en una revista científica indexada de cuartil Q1.

RESUMEN

La construcción concentra gran parte de la inversión pública peruana, pero enfrenta problemas crónicos de sobrecostos y retrasos que comprometen los recursos del Estado y postergan el cierre de la brecha de infraestructura. Esta investigación desarrolla modelos de *machine learning* para anticipar sobrecostos y retrasos en obras públicas, a partir de 98 obras públicas liquidadas en Arequipa (S/ 4.5 a S/ 30 millones) con variables predictoras previas a la ejecución. Las variables independientes surgieron de una revisión bibliográfica de los factores más influyentes, y las dependientes se establecieron según el triángulo de hierro de la gestión de proyectos (costo, plazo y alcance). Tras construir el conjunto de datos mediante un análisis forense y documental de los registros contractuales, se realizó un análisis exploratorio y una estratificación no supervisada que separó tres perfiles de obra: exitosas, moderadas y colapsadas. Después se entrenaron modelos supervisados con variables seleccionadas por objetivo, comparando modelos lineales (LASSO lineal y cuadrático) con *random forest*, evaluados por validación cruzada estratificada repetida; el modelo final se validó externamente con obras de Cusco, Puno y Tacna. Predecir el valor exacto del desvío resultó poco viable ($R^2 \leq 0.17$), pero su estructuración como un modelo de clasificación de riesgo por terciles demostró ser una alternativa efectiva; los tres clasificadores obtuvieron un AUC interno de 0.69 a 0.76. En la validación externa, el orden de riesgo se mantuvo para el retraso y el avance, pero no para el sobrecosto (AUC 0.58, cercano al azar), y el umbral debió recalibrarse por región. Las variables más influyentes cambian según el objetivo: el sobrecosto depende del plazo inicial, la modalidad de licitación y el tipo de contratista; el retraso, del sistema de contratación, la tipología de obra y el contexto de la pandemia. Finalmente, se propone una herramienta de gestión, el Semáforo de Alerta Temprana, que clasifica las obras y prioriza la revisión preventiva del expediente antes de la licitación.

Palabras clave: *Machine learning*, sobrecostos, retrasos.

ABSTRACT

Construction concentrates much of Peru's public investment, yet it faces chronic *cost overruns* and *delays* that compromise State resources and postpone the closing of the infrastructure gap. This research develops *machine learning* models to anticipate *cost overruns* and *delays* in public works, based on 98 completed public works in Arequipa (S/ 4.5 to S/ 30 million) with pre-execution predictor variables. The independent variables emerged from a literature review of the most influential factors, and the dependent ones were established according to the iron triangle of project management (cost, time, and scope). After building the *dataset* through a forensic and documentary analysis of the contractual records, an exploratory data analysis and an unsupervised stratification were carried out, separating three work profiles: successful, moderate, and collapsed. Supervised models were then trained with variables selected per objective, comparing linear models (linear and quadratic LASSO) with *random forest*, evaluated through repeated stratified cross-validation; the final model was externally validated with works from Cusco, Puno, and Tacna. Predicting the exact value of the deviation proved barely feasible ($R^2 \leq 0.17$), but framing it as a tercile-based risk classification proved an effective alternative: the three classifiers achieved an internal AUC of 0.69 to 0.76. In external validation, the risk ordering held for delay and progress, but not for *cost overrun* (AUC 0.58, close to chance), and the cutoff threshold had to be recalibrated by region. The most influential variables change by objective: *cost overrun* depends on the initial schedule, the bidding modality, and the contractor type; delay, on the contracting system, the work typology, and the pandemic context. Finally, a management tool is proposed, the Early Warning Traffic Light, which classifies works and prioritizes the preventive review of the technical file before the tender.

Keywords: Machine learning, cost overruns, schedule delays.

ÍNDICE GENERAL

Dedicatoria
Agradecimiento
Reconocimiento
RESUMEN
ABSTRACT

INTRODUCCIÓN 1
1. PLANTEAMIENTO DEL PROBLEMA 3
1.1. Problemática de la Investigación 3
1.2. Justificación de la Investigación 5
1.3. Hipótesis 7
1.4. Objetivos de la Investigación 7
1.4.1. *Objetivo General* 7
1.4.2. *Objetivos Específicos* 7
1.5. Cuadro de Consistencia 8
1.6. Alcances 9
1.7. Limitaciones 12
2. FUNDAMENTO TEÓRICO 13
2.1. Estado del Arte 13
2.2. Marco Legal 17
2.3. Marco Teórico 18
3. METODOLOGÍA 24
3.1. Tipo y Nivel de Investigación 24
3.2. Diseño de Investigación 25
3.3. Métodos de Investigación 25
4. TRATAMIENTO DE INFORMACIÓN 29
4.1. Definición de Variables 29
4.1.1. *Selección de Variables Independientes* 29
4.1.2. *Selección de Variables Dependientes* 34
4.2. Recolección de Datos 38
4.2.1. *Diagnóstico de Viabilidad* 38
4.2.2. *Técnicas de Recolección de Datos* 39
4.2.3. *Criterios de Filtrado de la Muestra* 41

4.2.4. ***Diseño y Validación del Instrumento de Extracción*** 42
4.3. Gestión y Preparación del *Dataset* 43
4.3.1. ***Preparación y Estructuración de Datos*** 43
4.3.2. ***Limpieza de Datos*** 45
4.4. Análisis Exploratorio de Datos (EDA) 49
4.4.1. ***Datos y Alcance del Análisis*** 49
4.4.2. ***Estadística Descriptiva Univariada*** 50
4.4.3. ***Tratamiento de Valores Extremos: Winsorización*** 53
4.4.4. ***Colinealidad y Correlación Bivariada*** 55
4.4.5. ***Contraste de Variables Dependientes según Variables Categóricas*** 57
4.4.6. ***Información Mutua*** 62
4.4.7. ***Robustez Frente al Nivel de Evidencia Documental*** 66
4.4.8. ***Interpretación y Decisiones en Base al EDA*** 67
4.5. Ingeniería de Características 68
4.5.1. ***Construcción de Indicadores Relativos y Transformación Logarítmica*** 68
4.5.2. ***Winsorización Local*** 69
4.5.3. ***Codificación Categórica*** 69
4.5.4. ***Puntaje de Relevancia*** 70
4.5.5. ***Preparación para el Análisis No Supervisado y Estado Final del Dataset*** 71
5. MODELADO APRENDIZAJE NO SUPERVISADO 73
5.1. Objetivo de *Clustering* 73
5.2. Selección de Variables para *Clustering* 74
5.3. Definición de Hiperparámetros y Procedimiento 75
5.4. Resultados del Agrupamiento 79
5.5. Interpretación de los Clústeres 81
6. MODELADO APRENDIZAJE SUPERVISADO 84
6.1. Selección Multivariada de Características 84
6.1.1. ***Objetivo de Selección de Características*** 84
6.1.2. ***Definición de Hiperparámetros y Procedimiento*** 85
6.1.3. ***Consenso de Selección*** 89
6.2. Regresión Supervisada de Modelos Línea Base 91
6.2.1. ***Definición de Hiperparámetros y Procedimiento*** 91
6.2.2. ***Resultados de Modelos Línea Base*** 93

6.2.3. ***Análisis Diagnóstico por Objetivo*** 95
6.2.4. ***Comparación de Resultados según Modelos*** 98
6.2.5. ***Interpretación*** 99
6.3. Modelos de Clasificación de Riesgo y Alerta Temprana 100
6.3.1. ***Objetivo de Modelos de Clasificación*** 100
6.3.2. ***Definición de las Clases de Riesgo*** 100
6.3.3. ***Definición de Hiperparámetros y Procedimiento*** 101
6.3.4. ***Resultados de Clasificación*** 103
6.3.5. ***Calibración del Umbral de Corte (Índice de* Youden*)*** 105
6.4. Validación Externa 106
6.4.1. ***Objetivo y Procedimiento*** 106
6.4.2. ***Resultados de validación externa*** 108
7. DISCUSIÓN DE RESULTADOS 112
8. CONCLUSIONES Y RECOMENDACIONES 125
8.1. Conclusiones 125
8.2. Recomendaciones 127
9. REFERENCIAS 129

ÍNDICE DE TABLAS

Tabla 1 Cuadro de Consistencia 8
Tabla 2 Estado del Arte 13
Tabla 3 Marco Legal 17
Tabla 4 Codificación de Metodología 25
Tabla 5 Resumen de Variables Independientes Utilizadas 29
Tabla 6 Banco de Variables Independientes Adoptado para Modelo Predictivo 32
Tabla 7 Variables Dependientes del Modelo Predictivo 35
Tabla 8 Detalle de Recolección de Datos No Estructurados 44
Tabla 9 Estadística Descriptiva Univariada 50
Tabla 10 Distribución de Frecuencias de las Variables Categóricas 50
Tabla 11 Efecto de la Winsorización 54
Tabla 12 Correlación de Pearson (Predictoras Numéricas × Dependientes) 55
Tabla 13 Correlación de Spearman (Predictoras Numéricas × Dependientes) 55
Tabla 14 Factor de Inflación de la Varianza (VIF) 57
Tabla 15 Pruebas de Kruskal-Wallis con Significación Global ($P < 0.05$) 60
Tabla 16 Pruebas U de Mann-Whitney con Significación Global ($P < 0.05$) 61
Tabla 17 Prueba *Post-hoc* de Dunn con Ajuste de Bonferroni 61
Tabla 18 Información Mutua 63
Tabla 19 Pruebas de Robustez de las Variables Endógenas y EMS 66
Tabla 20 Efecto de Log-transformación en Asimetría y Curtosis 68
Tabla 21 Umbrales de Winsorización (Referenciales) 69
Tabla 22 Puntaje de Relevancia (*Rank_score*) 70
Tabla 23 Transformación de Variables Independientes 72
Tabla 24 Hiperparámetros de *K-means* 77
Tabla 25 Centroides de los Clústeres 79
Tabla 26 Validación de la Separación entre Clústeres (Kruskal-Wallis) 81
Tabla 27 Hiperparámetros Modelo *random forest* 86
Tabla 28 Predictores Seleccionados por Objetivo 90
Tabla 29 Desempeño de Regresión Modelos Línea Base 93
Tabla 30 Umbrales de Definición de la Clase de Riesgo 101
Tabla 31 Hiperparámetro de Regresión Logística 102
Tabla 32 Desempeño de Clasificación de Riesgo 104

Tabla 33 Métricas Operativas en el Umbral de Corte 105
Tabla 34 Desempeño de los Clasificadores en la Validación Externa 108
Tabla 35 Validación Externa por Obra 110
Tabla 36 Perfil de los Tres Arquetipos de Obra 113
Tabla 37 Asociación entre las Variables Categóricas y el Arquetipo de Desempeño de Obras 114
Tabla 38 Factores Críticos por Variable Objetivo 118
Tabla 39 Semáforo de Alerta Temprana por Objetivo con Umbral de Corte 121

ÍNDICE DE FIGURAS

Figura 1 Esquema Metodológico 28
Figura 2 Distribución de BA (Sobrecosto): Histograma y Diagrama de Caja 52
Figura 3 Distribución de BB (Incidencia Económica): Histograma y Diagrama de Caja .. 52
Figura 4 Distribución de BC (Avance Físico): Histograma y Diagrama de Caja 52
Figura 5 Distribución de BD (Retraso): Histograma y Diagrama de Caja 53
Figura 6 Distribución de BE (Incidencia Temporal): Histograma y Diagrama de Caja 53
Figura 7 Matriz de Correlación de Spearman (Predictoras Continuas y Dependientes) 56
Figura 8 Dependientes según Naturaleza de Intervención (Kruskal-Wallis; * = p<0.05) .. 58
Figura 9 Dependientes según Tipología de Obra (Kruskal-Wallis; * = p<0.05) 58
Figura 10 Dependientes según Sistema de Contratación (Kruskal-Wallis; * = p<0.05) 58
Figura 11 Dependientes según Modalidad de Selección (Kruskal-Wallis; * = p<0.05) 58
Figura 12 Dependientes según Nivel de Gobierno (Kruskal-Wallis; * = p<0.05) 59
Figura 13 Dependientes según Escenario COVID (Kruskal-Wallis; * = p<0.05) 59
Figura 14 Dependientes según Ubicación Metropolitana (Kruskal-Wallis; * = p<0.05) ... 59
Figura 15 Dependientes según Familia SUCS (Kruskal-Wallis; * = p<0.05) 59
Figura 16 Dependientes según Calidad de Fundación (Kruskal-Wallis; * = p<0.05) 60
Figura 17 Dependientes según Tipo de Contratista (Kruskal-Wallis; * = p<0.05) 60
Figura 18 Información Mutua con el Sobrecosto (BA) 63
Figura 19 Información Mutua con la Incidencia Económica (BB) 64
Figura 20 Información Mutua con el Avance Físico (BC) 64
Figura 21 Información Mutua con el Retraso (BD) 65
Figura 22 Información Mutua con la Incidencia Temporal (BE) 65
Figura 23 Funcionamiento de Algoritmo de *Clustering* 73
Figura 24 Método del Codo: Inercia Frente al Número de Clústeres 77
Figura 25 Construcción del Espacio de Agrupamiento en JupyterLab 78
Figura 26 Aplicación de Algoritmo *K-means* en JupyterLab 79
Figura 27 Obras en el Espacio de Desempeño BA–BD–BC 80
Figura 28 Proyección PCA 2D de los Clústeres 81
Figura 29 Incidencia Endógena Económica (BB) por Clúster 83
Figura 30 Incidencia Endógena de Plazo (BE) por Clúster 83
Figura 31 Algoritmo *Sequential Forward Selection* 85
Figura 32 Winsorización y Validación Cruzada para *random forest* 88

Figura 33 Aplicación e Iteración de Algoritmo *Forward Selection* 88
Figura 34 Frecuencia de Selección por Variable 89
Figura 35 Curva de Selección *Forward*: R^2 Frente al Número de Variables K 90
Figura 36 Validación Cruzada Repetida y Comparación de Modelos en JupyterLab 93
Figura 37 R^2 por Objetivo y Modelo 94
Figura 38 BD: Predicho vs. Real y Residuos (*random forest*) 95
Figura 39 BC: Predicho vs. Real y Residuos (*random forest*) 96
Figura 40 BA: Predicho vs. Real y Residuos (*random forest*) 97
Figura 41 BB: Predicho vs. Real y Residuos (*random forest*) 97
Figura 42 BE: Predicho vs. Real y Residuos (*random forest*) 98
Figura 43 Cálculo del AUC para *random forest* Frente a Regresión Logística en JupyterLab 103
Figura 44 AUC por Objetivo y Modelo de Clasificación 104
Figura 45 Curvas ROC de la Clasificación de Riesgo 104
Figura 46 Matrices de Confusión en Umbral de Corte (Predicciones Fuera de Pliegue) 106
Figura 47 AUC por Variable Objetivo 108
Figura 48 Retraso Predicho Frente al Real por Obra 109
Figura 49 Matrices de Confusión de la Validación Externa por Variable Objetivo 110
Figura 50 Distribución de Arquetipos por Sistema de Contratación 114
Figura 51 Distribución de Arquetipos por Tipo de Contratista 115
Figura 52 Distribución de Arquetipos por Nivel de Gobierno 115
Figura 53 Distribución de Arquetipos por Tipología de Obra 116
Figura 54 Distribución de Arquetipos por Escenario COVID 117
Figura 55 R^2 por Objetivo y Modelo 119
Figura 56 AUC por Objetivo y Modelo 120
Figura 57 Semáforo de Alerta Temprana 122

ÍNDICE DE ANEXOS

Anexo 1 Base de datos sin transformar empleada en la validación externa: variables predictoras 137
Anexo 2 Base de datos sin transformar empleada en la validación externa: variables objetivo 138

INTRODUCCIÓN

La ejecución de obras de infraestructura pública constituye uno de los principales motores del desarrollo socioeconómico y del cierre de brechas de competitividad. Sin embargo, convive con una dificultad persistente, pocos proyectos se ejecutan dentro del costo y el plazo previstos. La evidencia internacional es contundente al respecto, nueve de cada diez proyectos de infraestructura experimentan sobrecostos (Afana et al., 2024); a marzo de 2025 se contabilizaban más de 2500 obras paralizadas en los tres niveles de gobierno, con Arequipa a la cabeza del saldo de inversión por ejecutar (Contraloría General de la República del Perú, 2025). Detrás de estas cifras subyace un problema de gestión, los sobrecostos y los retrasos rara vez se anticipan y, cuando se detectan, suele ser demasiado tarde para corregirlos.

Los métodos tradicionales de control de proyectos son mayoritariamente reactivos y se apoyan en la experiencia y en enfoques lineales, lo que limita su capacidad para prever desviaciones en escenarios complejos. Frente a ello, el aprendizaje automático (*machine learning*) ofrece una alternativa, aprender de datos históricos de obras ya ejecutadas para reconocer, con anticipación, los patrones asociados a un mal desempeño (Khursheed et al., 2024). Pese a su potencial, en el Perú estas técnicas casi no se han aplicado a la predicción del desempeño de la obra pública, lo que deja un vacío tanto metodológico como práctico.

La presente investigación aborda ese vacío y se integra formalmente a la línea de investigación en Construcción de Ingeniería Civil, enmarcándose en el área de gestión de la construcción. El objetivo general es desarrollar modelos predictivos basados en *machine learning* para anticipar sobrecostos y retrasos en proyectos de infraestructura pública en el Perú, identificando y analizando las variables críticas que los originan. Como caso de estudio se emplean obras civiles liquidadas de la región Arequipa, de mediana envergadura, seleccionadas por la completitud y trazabilidad de su información en los portales oficiales del Estado. Asimismo, como parte de su contribución práctica, el modelo final de clasificación de

riesgo se acopla a un protocolo de Semáforo de Alerta Temprana. De este modo, a partir de las variables previas a la ejecución del expediente técnico, cada obra es clasificada en niveles de criticidad (verde, ámbar o rojo), permitiendo que se detecten los proyectos de alto riesgo y se priorice el control preventivo antes de la licitación y del movimiento de tierras.

El documento se organiza en ocho capítulos. El Capítulo I plantea el problema, la justificación, las hipótesis y los objetivos. El Capítulo II desarrolla el marco teórico y el estado del arte. El Capítulo III describe la metodología. El Capítulo IV presenta el proceso de recolección de datos, análisis exploratorio y la ingeniería de características. El Capítulo V expone la estratificación no supervisada por *clustering*. El Capítulo VI abarca la selección de variables, la regresión de los modelos línea base y la clasificación de riesgo con su validación externa. Para concluir, El Capítulo VII discute e integra los resultados, y el Capítulo VIII presenta las conclusiones y recomendaciones.

1. PLANTEAMIENTO DEL PROBLEMA

1.1. Problemática de la Investigación

A nivel mundial, la industria de la construcción representa aproximadamente el 10 % del Producto Bruto Interno (PIB), siendo un motor clave para el crecimiento económico (Abel et al., 2024). Sin embargo, esta misma industria enfrenta altos niveles de incertidumbre debido a su complejidad inherente y a la exposición constante a riesgos impredecibles durante la ejecución de los proyectos. Según el estudio de Flyvbjerg et al. (2002), que analizó 258 proyectos de infraestructura en 20 países durante un período de 70 años, el 90% de los proyectos experimentan sobrecostos, con un promedio de 28% de sobrecosto respecto al presupuesto inicial. Esta problemática se agrava considerablemente en proyectos de mayor envergadura, donde los sobrecostos promedio alcanzan el 44.7% para proyectos ferroviarios y 33.7% para enlaces fijos como túneles y puentes (Flyvbjerg et al., 2003). Estudios más recientes confirman que esta tendencia persiste: Afana et al. (2024) reportan en su revisión sistemática que el 90% de proyectos experimentan sobrecostos a nivel global, mientras que investigaciones específicas documentan sobrecostos promedio del 60% en proyectos de energía y salud, 40% en transporte y agua, y 30% en educación (Allahaim, 2015). Esta problemática es universal y predominante, los factores críticos incluyen deficiencias en la planificación inicial, subestimación sistemática de costos, condiciones imprevistas del sitio, cambios en el alcance del proyecto y gestión inadecuada de riesgos (Afana et al., 2024).

En el contexto peruano, esta situación se agrava debido a deficiencias institucionales y limitaciones en capacidades técnicas, particularmente en el sector de infraestructura pública. Esto afecta de forma directa la eficiencia del gasto público, así como la credibilidad de las entidades ejecutoras (Avila Chumpisuca, 2019). Según un análisis del Consejo Privado de Competitividad (CPC) en junio de 2025, más de 12,000 proyectos públicos en el país presentan una variación de costo respecto al presupuesto viable original, alcanzando una variación

promedio de 118% (Consejo Privado de Competitividad, 2025a). Paralelamente, 8,805 proyectos registran retrasos en su ejecución con un promedio de 671 días de demora, cifra que supera significativamente los 595 días registrados el año anterior, siendo el Gobierno nacional el que presenta el mayor retraso promedio con 836 días por proyecto (Consejo Privado de Competitividad, 2025b). Esta crisis se ve reflejada en el Reporte de Obras Paralizadas de Marzo del 2025, existen 2,572 obras detenidas en los tres niveles de gobierno. Siendo Arequipa la que encabeza la lista con un saldo de inversión por ejecutar de S/ 5.197 millones. Las principales causas incluyen incumplimientos contractuales (23%), falta de recursos financieros y liquidez (21%), y deficiencias en expedientes técnicos (10%) (Contraloría General de la República del Perú, 2025). Por su parte, estudios específicos en el sector de agua y saneamiento peruano demuestran que el 55.7% de los proyectos experimentan retrasos y el 44.7% enfrentan sobrecostos (Julca-Varas et al., 2025). Esta situación no solo refleja limitaciones sistemáticas en la planificación y gestión que perpetúan ciclos de ineficiencia, sino también una carencia de sistemas capaces de anticipar y prevenir desviaciones significativas en el desempeño de las obras.

Los métodos tradicionales de gestión de proyectos han demostrado limitaciones estructurales para anticipar y mitigar eficazmente estos riesgos. La dependencia excesiva de evaluaciones subjetivas, procesos manuales de documentación, sistemas reactivos de control y enfoques lineales limita la objetividad y precisión en la toma de decisiones críticas por parte de la Supervisión (Gado, 2024). Esta brecha evidente entre las limitaciones de los métodos tradicionales y las capacidades avanzadas del *machine learning* plantea una oportunidad para modernizar integralmente la gestión pública de obras en Perú. La implementación estratégica de técnicas de *machine learning* podría transformar los procesos de planificación, monitoreo y control de proyectos, permitiendo una gestión considerablemente más eficiente de los recursos públicos y reduciendo sustancialmente la incertidumbre inherente a la ejecución de obras de

infraestructura. En este contexto crítico, surge la siguiente pregunta fundamental de investigación: ¿Es posible desarrollar y validar modelos predictivos robustos basados en datos históricos de proyectos liquidados con técnicas de *machine learning* que permitan anticipar los sobrecostos y retrasos en proyectos de infraestructura pública en Perú?

1.2. Justificación de la Investigación

La gestión de obras públicas en el Perú atraviesa una crisis estructural reflejada en el alto índice de obras paralizadas a nivel nacional, una problemática que compromete una inversión superior a S/ 43 mil 163 millones, lo que representa aproximadamente el 2% del PBI nacional (Contraloría General de la República del Perú, 2025). A ello se suma que, según la Escuela de Gestión Pública de la Universidad del Pacífico (2025), por encargo de la Asociación para el Fomento de la Infraestructura Nacional (AFIN), la brecha de infraestructura para el periodo 2016-2025 asciende a US$ 159 mil 549 millones, lo que exige una inversión anual del 8.27% del PBI para su cierre. Este escenario evidencia la necesidad de herramientas más precisas para anticipar y gestionar los riesgos de sobrecostos y retrasos en proyectos en ejecución, a fin de orientar una gestión pública más eficiente (Traverso-Frisancho & Romero Alva, 2023).

Esta investigación representa un aporte científico y tecnológico significativo para modernizar la gestión pública con la implementación de herramientas de la ingeniería 4.0, mediante *machine learning*. El desarrollo de modelos computacionales permite anticipar sobrecostos y retrasos con mayor precisión que los métodos tradicionales, facilitando una gestión más eficiente de los recursos públicos limitados. La identificación de arquetipos de obra mediante técnicas de análisis no supervisado y de patrones ocultos mediante análisis supervisado, proporciona conocimiento estratégico para gestionar los riesgos de forma anticipada (Nwosu, 2023). Si bien la literatura internacional ha demostrado que los modelos de *machine learning* mejoran significativamente la estimación de costos, plazos y riesgos (Singh,

2024), esta investigación aplica de manera sistemática estas técnicas para predecir desviaciones en obras públicas peruanas, cubriendo así un vacío relevante en la literatura científica nacional. Los estudios previos en el país se han centrado principalmente en análisis descriptivos o en simulaciones probabilísticas, sin explorar el potencial predictivo de los algoritmos de *machine learning* sobre datos históricos de proyectos reales.

Como principal aporte aplicativo, se propone un instrumento de gestión de riesgos: el Semáforo de Alerta Temprana, una herramienta diseñada para que los modelos desarrollados sirvan como apoyo efectivo en la planificación y control de proyectos públicos. Sus recomendaciones, basadas en evidencia, ayudan a formular políticas más efectivas para prevenir desviaciones de presupuesto y de plazo. La capacidad predictiva también permite asignar de forma más eficiente los recursos de supervisión, porque concentra los esfuerzos en los proyectos de mayor riesgo. Además, al identificar las causas raíz de los sobrecostos y los retrasos, el trabajo facilita el diseño de estrategias preventivas y reduce la incertidumbre propia de la ejecución de obras de infraestructura. Estos objetivos son coherentes con los cambios que introduce la Ley General de Contrataciones Públicas (Ley N.° 32069) frente a la norma anterior, que prioriza la calidad, la innovación y la sostenibilidad por encima de la oferta más económica y refuerza la gestión de riesgos desde la etapa inicial del proyecto. Por último, el estudio abre nuevas líneas de investigación sobre el uso de inteligencia artificial en la gestión pública, con aportes a la transparencia, al uso eficiente de los recursos y a una ejecución de infraestructura guiada por criterios de eficiencia y sostenibilidad.

La justificación de este estudio se apoya en varios ámbitos. En lo social, anticipar los sobrecostos y los retrasos ayuda a sostener la continuidad de los servicios públicos, porque reduce las obras paralizadas que golpean a sectores como educación, salud y transporte, con un efecto directo sobre las poblaciones más vulnerables. En lo económico, permite aprovechar mejor unos recursos públicos que ya son limitados: cuando bajan las pérdidas por sobrecostos,

se liberan fondos para nuevas inversiones. Y en lo tecnológico, el trabajo lleva técnicas de inteligencia artificial a la gestión pública para el manejo de riesgos, un terreno que hasta ahora se ha atendido casi siempre de forma reactiva.

En el plano político e institucional, los modelos aportan criterios objetivos para decidir, lo que puede mejorar la eficiencia de la gestión y reforzar la confianza de la ciudadanía en las instituciones. En lo ambiental, planificar mejor las obras públicas reduce el consumo de recursos y las emisiones ligadas a la construcción, en línea con criterios de sostenibilidad. Tomados en conjunto, estos ámbitos justifican la aplicación de modelos predictivos para el seguimiento de la obra pública en el país.

1.3. Hipótesis

Los modelos predictivos basados en técnicas de *machine learning*, entrenados con datos históricos de obras públicas liquidadas en Arequipa según factores de riesgo identificados previo al inicio de la ejecución, permiten anticipar de forma estadísticamente significativa el riesgo de sobrecostos y retrasos en proyectos públicos en el sur del Perú.

1.4. Objetivos de la Investigación

1.4.1. *Objetivo General*

Desarrollar y validar modelos predictivos basados en *machine learning* para anticipar el riesgo de sobrecostos y retrasos en obras públicas, mediante la identificación de factores críticos previos a la ejecución contractual, tomando como caso de estudio la región de Arequipa, Perú.

1.4.2. *Objetivos Específicos*

- Identificar y categorizar las variables críticas asociadas a sobrecostos y retrasos en obras de infraestructura, mediante una revisión sistemática de literatura científica y un proceso de selección de variables.
- Desarrollar y entrenar modelos de *machine learning* para predecir riesgo de sobrecostos y

retrasos en proyectos de infraestructura pública en la etapa previa al inicio de la ejecución física.

- Identificar patrones y arquetipos de desempeño en costo, plazo y alcance mediante algoritmos no supervisados de agrupamiento (*clustering*) con el fin de tipificar los perfiles de colapso en la ejecución de la infraestructura pública.
- Evaluar el desempeño de los modelos desarrollados en contextos reales y formular una herramienta de gestión preventiva para su aplicación en la gestión pública.

1.5. Cuadro de Consistencia

Tabla 1

Cuadro de Consistencia

Problema General	**Objetivo General**	**Hipótesis General**
P.G.1: ¿Es posible desarrollar modelos predictivos robustos basados en técnicas de *machine learning* que permitan anticipar el riesgo de sobrecostos y retrasos en obras públicas en Perú, utilizando factores críticos previos al inicio de ejecución y datos históricos de obras liquidadas?	**O.G.1:** Desarrollar y validar modelos predictivos basados en *machine learning* para anticipar el riesgo de sobrecostos y retrasos en obras públicas, mediante la identificación de factores críticos previos a la ejecución contractual, tomando como caso de estudio la región de Arequipa, Perú.	**H.G.1:** Los modelos predictivos basados en técnicas de *machine learning*, entrenados con datos históricos de obras públicas liquidadas en Arequipa según factores de riesgo identificados previos al inicio de la ejecución, permiten clasificar y anticipar de forma estadísticamente significativa el riesgo de sobrecostos y retrasos en proyectos públicos en el sur del Perú, sirviendo de base para la estructuración de una herramienta de gestión preventiva de alerta temprana.
Problemas Específicos	**Objetivos Específicos**	**Hipótesis Específicas**
P.E.1: ¿Cuáles son las variables críticas más frecuentes y significativas asociadas a sobrecostos y retrasos en obras de infraestructura según la literatura científica y técnica especializada?	**O.E.1:** Identificar y categorizar las variables críticas asociadas a sobrecostos y retrasos en obras de infraestructura, mediante una revisión sistemática de literatura científica y un proceso de selección de variables.	**H.E.1**: Existe un conjunto finito y categorizable de variables críticas técnicas, administrativas y contextuales que se asocian significativamente con la ocurrencia de sobrecostos y retrasos en obras públicas, según la evidencia científica internacional.
P.E.2: ¿Qué algoritmos de *machine learning* demuestran	**O.E.2:** Desarrollar y entrenar modelos de *machine learning*	**H.E.2:** Los algoritmos de *machine learning* supervisado

mayor capacidad predictiva para anticipar sobrecostos y retrasos en obras públicas utilizando datos históricos reales previos al inicio de la ejecución física?	para predecir el riesgo de sobrecostos y retrasos en proyectos de infraestructura pública en la etapa previa al inicio de la ejecución física.	(regresión lineal, cuadrática y *random forest*) entrenados con registros contractuales históricos de obras públicas permiten predecir sobrecostos y retrasos en una etapa previa al inicio de la ejecución física con mayor precisión predictiva que los métodos tradicionales de estimación.
P.E.3: ¿Qué patrones recurrentes y arquetipos de desempeño pueden identificarse a partir de las dimensiones de costo, plazo y alcance en los proyectos liquidados mediante técnicas de análisis no supervisado?	**O.E.3:** Identificar patrones y arquetipos de desempeño en costo, plazo y alcance mediante algoritmos no supervisados de agrupamiento (*clustering*) con el fin de tipificar los perfiles de colapso en la ejecución de la infraestructura pública.	**H.E.3:** El análisis de agrupamiento no supervisado (*clustering*) aplicado a las dimensiones de desempeño revela patrones recurrentes y agrupaciones temáticas que tipifican de manera estadísticamente significativa los perfiles de desvío contractual.
P.E.4: ¿Cuál es el nivel de generalización y aplicabilidad práctica de los modelos desarrollados cuando se validan en contextos externos a los datos de entrenamiento?	**O.E.4:** Evaluar el desempeño de los modelos desarrollados en contextos reales y formular una herramienta de gestión preventiva para su aplicación en la gestión pública.	**H.E.4:** Los modelos predictivos desarrollados mantienen un desempeño estadísticamente aceptable cuando se validan externamente en proyectos no incluidos en el conjunto de entrenamiento, una herramienta de gestión de riesgos permite su aplicación práctica en la gestión pública de obras.

1.6. Alcances

Los modelos de *machine learning* aplicados a la predicción de sobrecostos y retrasos han mostrado buena precisión, sobre todo por su capacidad de procesar de forma automática grandes volúmenes de datos históricos. A diferencia de los enfoques convencionales, estos algoritmos detectan patrones ocultos, relaciones no lineales y dependencias complejas entre variables técnicas, administrativas y contextuales que suelen escapar a los análisis basados solo en la experiencia o en modelos lineales. Funcionan ajustando de forma iterativa los parámetros del modelo para minimizar el error y mejorar la generalización ante nuevos casos, incluso en condiciones de incertidumbre y riesgo. En esa línea, la investigación se centra en desarrollar

modelos predictivos que anticipen los sobrecostos y las ampliaciones de plazo en proyectos de infraestructura pública en el Perú, con la región de Arequipa como caso de estudio.

De acuerdo con el análisis de los antecedentes presentados en la Tabla 2 se revela que los tamaños de muestra en investigaciones similares de *machine learning* aplicado a construcción varían significativamente según el alcance, tipo de algoritmo y disponibilidad de datos. La investigación se desarrolló con una muestra de al menos 98 proyectos públicos correspondientes al departamento de Arequipa. Se considera todo tipo de obras civiles, liquidadas entre 2020 y 2026. Estos proyectos, de mediana envergadura con montos de inversión entre S/ 4.5 y S/ 30 millones, fueron seleccionados por la limitada disponibilidad, completitud y trazabilidad de la información en portales oficiales del Estado (Consulta Amigable y Sistema de Seguimiento de Inversiones del Ministerio de Economía y Finanzas (MEF), Sistema Electrónico de Contrataciones del Estado (SEACE) del Organismo Supervisor de las Contrataciones del Estado (OSCE) e Infobras de la Contraloría General de la República). Este criterio de selección prioriza la calidad y completitud de los datos sobre la cantidad absoluta de proyectos, enfocándose en obras de mediana y gran envergadura que típicamente cuentan con mejor documentación y seguimiento institucional.

La elección de este tamaño muestral se sustenta en la evidencia empírica y en antecedentes internacionales que utilizaron un rango similar de proyectos (Elbashbishy et al., 2022; Sahu et al., 2024; Timalsina et al., 2025). Además, cumple con la regla técnica de modelado en *machine learning*, la cual sugiere que el número de muestras sea al menos diez veces superior al número de variables independientes (Graphite Note, 2024). Considerando que el estudio realizó considerando entre siete y nueve variables críticas, la muestra de 98 proyectos resulta adecuada para el entrenamiento, validación y prueba de los modelos predictivos, garantizando solidez estadística y capacidad de generalización. Adicionalmente, para la validación externa se considera el 10% del total de la muestra, la escala del 10% se encuentra

en paridad con antecedentes nacionales e internacionales en ingeniería civil que emplearon entre 4 y 10 casos de control (Avila Chumpisuca, 2019; Elbashbishy et al., 2022; Sahu et al., 2024)

El análisis se basa exclusivamente en datos estructurados, es decir, variables numéricas y categóricas relacionadas con características del proyecto asociadas a sobrecostos y retrasos, siguiendo criterios establecidos en la literatura científica y técnica internacional. Una vez recopilados, los datos fueron depurados, normalizados y codificados según corresponde.

Los modelos predictivos desarrollados se validan en dos niveles. En primer lugar, se realiza una evaluación interna mediante validación cruzada (*k-fold cross-validation*) sobre el conjunto de entrenamiento, lo que permite estimar la estabilidad del modelo y reducir el riesgo de sobreajuste, empleando métricas como el coeficiente de determinación (R^2), la raíz del error cuadrático medio (RMSE) y el error absoluto medio (MAE) para medir la precisión y el error de predicción. Finalmente, se realiza una validación externa utilizando ≈10 proyectos liquidados de otros departamentos, distintos a los usados en el entrenamiento, con el propósito de comprobar la capacidad de generalización del modelo en contextos geográficos y administrativos diferentes. Este proceso permite verificar su desempeño real y robustez ante condiciones diversas.

Como producto aplicativo de la investigación, el modelo predictivo se acopla al protocolo del Semáforo de Alerta Temprana, constituyendo una herramienta de gestión aplicable y dirigida a las entidades públicas para la gestión y mitigación de riesgos en una etapa previa al inicio de la ejecución física. Todo el procesamiento y análisis de datos se realiza en entornos de programación basados en Python, haciendo uso de bibliotecas especializadas como pandas, scikit-learn, matplotlib y seaborn. Este entorno permite implementar algoritmos de aprendizaje supervisado (regresión lineal, regresión cuadrática y bosques aleatorios) y no

supervisado (*k-means*). Los procesos de validación y visualización también se realizan en este entorno de *software*, asegurando reproducibilidad y transparencia metodológica.

1.7. Limitaciones

Esta investigación presenta limitaciones principalmente relacionadas con la disponibilidad y calidad de los datos necesarios para la construcción del *dataset*. Si bien se recurrirá a fuentes oficiales del Estado Peruano, como los portales de Infobras, Sistema de Seguimiento de Inversiones (SSI) del MEF, SEACE y la Contraloría General de la República, estos sistemas no siempre cuentan con registros completos, actualizados o estandarizados. Esta situación restringe la cantidad de casos útiles para el análisis, además de requerir procedimientos adicionales de depuración y validación.

En caso de ser necesario, se contempla la necesidad de complementar la información mediante solicitudes directas a gobiernos locales y regionales. No obstante, la respuesta y colaboración de estas entidades dependerán de su disponibilidad y nivel de transparencia, lo cual podría limitar el acceso a expedientes técnicos o datos de obras culminadas que no estén disponibles en línea. No se trabajará con información no estructurada, como informes técnicos narrativos completos o minutas de gestión

Otro aspecto a considerar es el posible sesgo en los resultados debido a que muchos de los proyectos incluidos fueron finalizados durante el período 2020–2022, coincidente con la pandemia por COVID-19. Esta coyuntura generó alteraciones significativas en los procesos de ejecución, financiamiento y supervisión de obras públicas, distorsionando patrones que no necesariamente corresponden al comportamiento típico de los proyectos en condiciones normales.

2. FUNDAMENTO TEÓRICO

2.1. Estado del Arte

La inteligencia artificial y el *machine learning* están transformando la ingeniería civil al optimizar procesos. Van Tam y Toan (2021) identifican cinco áreas clave de aplicación: gestión de riesgos, seguridad, costos, cronogramas y demanda energética. En esa línea, para conocer los métodos y resultados de otros autores a nivel global, se revisó literatura científica en Scopus y Web of Science, cuyos hallazgos se presentan en la Tabla 2.

Tabla 2

Estado del Arte

Ítem	Autor	Relevancia	Métodos	Código
1	Andrić et al. (2024)	Predice sobrecostos en proyectos de infraestructura en Asia del Sur.	- Revisión literaria para definir variables críticas. - Base de datos: informes de 138 proyectos recuperados del Banco Asiático de Desarrollo. - Modelos de Regresión lineal, cuadrática y *random forest*. - Evaluación y Validación Interna.	M1.1 M1.2 M1.3 M1.5 M1.6
2	Al-Bataineh et al. (2024)	Identifica factores críticos de retraso en construcción en países en desarrollo.	- Revisión literaria para definir factores críticos. - Base de datos: cuestionarios administrados a 68 empresas. - Método de *K-means clustering* y PCA. - Aplicaron PCA para reducir la dimensionalidad del datase	M1.1 M1.2 M1.4
3	Andrić et al. (2025)	Modela retrasos en infraestructura en Asia del Sur.	- Revisión de literatura para definir variables críticas. - Base de datos: informes de 138 proyectos. Recuperados del Banco Asiático de Desarrollo. - Modelos de Regresión lineal, cuadrática y *random forest*. - Evaluación y Validación Interna.	M1.1 M1.2 M1.3 M1.5 M1.6
4	Avila Chumpisuca (2019)	Simulación de probabilidades de ampliaciones y adicionales en obras públicas.	- Revisión literaria para definir variables críticas. - Base de datos recopilada en Municipalidad Local. - Simulación Monte Carlo mediante @Risk. - Validación Externa mediante aplicación a 4 casos reales.	M1.1 M1.2 M1.5 M1.7
5	Boamah et al. (2025)	Creación de modelo IA para identificar riesgos en infraestructura.	- Revisión literaria para construir vocabulario técnico, procesamiento textual, agrupación de riesgos. - Base de datos: informes de proyectos, entrevistas y encuestas a expertos en el campo. - Modelo IA basado en NLP, embeddings, similitud semántica y *clustering*. - Validación externa con precisión.	M1.1 M1.2 M1.4 M1.5 M1.7

6	Andrić et al. (2019)	Identificación y evaluación de patrones y variables que influyen en el rendimiento de costos.	- Revisión literaria para formular hipótesis sobre causas principales. - Base de datos: información de 102 proyectos de infraestructura recopilada del Asian Development Bank. - Regresión Lineal Múltiple	M1.1 M1.2 M1.3 M1.5
7	Timalsina et al. (2025)	Predicción de ocurrencia de disputas y mecanismos de resolución en proyectos de tuberías.	- Revisión literaria para identificar variables críticas. - Base de datos: 60 proyectos de "*Pipeline*" recolectada de Global Energy Monitor. - Modelos de predicción: *Random forest*, *Logistic Regression,* Naïve Bayes*, K-Nearest Neighbors, AdaBoost.* - Evaluación y Validación Interna	M1.1 M1.2 M1.3 M1.6
8	Elbashbishy et al. (2022)	Evalúa el impacto acumulativo de riesgos en sobrecostos.	- Revisión literaria para categorizar riesgos significativos. - La base de datos: información de 53 proyectos de construcción recopilada mediante encuestas. Se modelan Algoritmos Genéticos y Red Neuronal Artificial. - Verificación y Validación Interna.	M1.1 M1.2 M1.3 M1.6
9	Oh et al. (2025)	Analiza recurrencia de órdenes de cambio en proyectos de carreteras y su impacto en costos y plazos.	- Revisión literaria para definir variables claves. - Base de datos: 1182 órdenes de cambio en 68 proyectos de construcción de carreteras por el Departamento de Transporte de Carolina del Sur. - Modelado de Eventos Recurrentes: Modelo Andersen Gill, PWP y de Fragilidad de Cox. Modelos: *Random forest*, *gradient boosting* Machine. - Evaluación y Validación Interna	M1.1 M1.2 M1.3 M1.5 M1.6
10	Sahu et al. (2024)	Identifica causas de retraso y predicción de riesgos para obras en India.	- Revisión literaria para establecer causas de disputa. - La base de datos: 50 proyectos mediante cuestionarios hechos a profesionales de la construcción. - Modelo de *random forest* optimizado con Algoritmo Genético (RFGA). - Evaluación y Validación Interna	M1.1 M1.2 M1.3 M1.6
11	Gómez-Cabrera et al. (2025)	Analiza y predice desviaciones de tiempo y costo en infraestructura pública escolar en Colombia.	- Revisión de literatura para la conceptualización de variables de entrada - Base de datos: 175 contratos de colegios públicos liquidados extraídos de la plataforma de datos abiertos del Estado (SECOP II). - Selección de variables mediante Disminución Media de la Precisión con *random forest*. - Algoritmo supervisado *random forest regression*. - Validación cruzada interna.	M1.1 M1.2 M1.3 M1.5 M1.6
12	Tochaiwat & Budda (2026)	Predice el precio final del contrato adjudicado para	- Revisión literaria y entrevistas a 8 expertos para la selección de 14 variables físicas y contractuales *ex-ante*.	M1.1 M1.2 M1.3

		edificaciones del gobierno de Tailandia.	- Base de datos: 300 proyectos de construcción de edificios gubernamentales del portal electrónico de compras públicas. - *Deep Learning* (Multilayer Factorization Net - MLFN) y función de pérdida cuadrática. - Evaluación y Validación Interna	M1.5 M1.6
13	Restrepo Ramírez & Rúa Machado (2025)	Predice retrasos y sobrecostos finales utilizando flujos de caja y avance físico histórico.	- Revisión literaria sobre marcos de trabajo analíticos de proyectos, Gestión de Valor Ganado y Cronograma Ganado. - Base de datos: Datos de proyectos de vivienda residencial en Colombia (102 proyectos para tiempos y 81 para flujos de caja). - Seis modelos de regresión de *machine learning*: OLS, regresión robusta, K-Vecinos Cercanos y *random forest* regression, comparados contra modelos probabilísticos de cadenas de Márkov. - Métricas de evaluación de rendimiento calculados en diferentes intervalos de tiempo constructivos.	M1.1 M1.2 M1.3 M1.5 M1.6
14	Nassar & Elbisy (2024)	Predice porcentajes de retraso en el cronograma para proyectos de construcción de infraestructura marina.	- Revisión de literatura para la identificación de 43 causas críticas de retraso evaluadas mediante el Índice de Importancia Relativa (RII) - Base de datos: Cuestionarios estructurados aplicados a profesionales y datos históricos. - *General Regression Neural Networks, Tree Boost Regression* y *Support Vector Machines.* - Validación interna mediante diagramas de Taylor cruzando métricas de evaluación de rendimiento.	M1.1 M1.2 M1.3 M1.5 M1.6
15	Ashtari et al. (2022)	Evalúa e identifica interrelaciones de riesgo para predecir niveles categóricos de sobrecosto.	- Revisión literaria para identificar 43 riesgos de sobrecosto, evaluados mediante encuestas estructuradas a expertos. - Base de datos: Registros de opinión de expertos recopilados bajo la escala ISO 31000:2018. - Reducción de variables: Método *Correlation-based Feature Selection* (CFS) para filtrar a 8 características clave. - *Network Classifier* (BN), Naive Bayes (NB) y Árboles de Decisión (DT) - Evaluación y Validación Interna	M1.1 M1.2 M1.3 M1.5 M1.6
16	Coffie & Cudjoe (2023)	Desarrolla modelos de predicción de la magnitud monetaria del costo de finalización y sobrecosto en edificaciones públicas.	- Revisión de literatura para la conceptualización de variables de diseño arquitectónico y de licitación. - Base de datos: Registros históricos contractuales y físicos de 911 proyectos de edificaciones escolares públicas. - Análisis de correlación de Pearson y *Support Vector Machine Regression* (SVR) optimizado mediante múltiples *kernels*. - Evaluación y Validación Interna	M1.1 M1.2 M1.3 M1.5 M1.6

17	Zhong & Zhang (2025)	Evalúa la probabilidad de retrasos y realiza diagnósticos inversos de riesgo en el cronograma de edificaciones prefabricadas.	-	Revisión de literatura para identificar 14 factores críticos de retraso en 5 dimensiones operativas y logísticas. - Base de datos: Cuestionarios y juicio de expertos. - Enfoque híbrido que integra el análisis de toma de decisiones conceptual DEMATEL-ISM mapeado directamente a Redes Bayesianas (BN). - Validación: Inferencia diagnóstica posterior y análisis de sensibilidad estructural.	M1.1 M1.2 M1.3 M1.4 M1.5 M1.6

Los antecedentes muestran avances recientes en el uso de *machine learning* para gestionar riesgos y optimizar la gestión en infraestructura a nivel mundial, constituyendo una base sólida para adaptar estos enfoques al contexto peruano. Aunque existen estudios aislados sobre análisis de datos y modelos predictivos en construcción, su desarrollo es aún limitado, especialmente en el sector público. Este estudio busca cubrir esta brecha, pues hasta ahora no se han propuesto modelos de *machine learning* para predecir costos y retrasos en obras públicas en Perú.

Las técnicas de aprendizaje automático seleccionadas (regresión lineal, polinómica y *random forest*) ofrecen ventajas prácticas para modelar las desviaciones de costo y plazo en proyectos de infraestructura, sobre todo por su capacidad de aprender relaciones complejas y no lineales entre muchas variables. Aung et al. (2023) mostraron que los algoritmos de *machine learning* superan de forma consistente a los métodos convencionales en precisión predictiva, con tasas superiores al 85 % al clasificar proyectos con riesgo de desviaciones de presupuesto y plazo. Chang Hidalgo (2023), al usar regresión lineal múltiple para estimar costos de obras públicas peruanas, obtuvo un coeficiente de determinación $R^2 \approx 0.80$, lo que indica que el modelo explica buena parte de la variabilidad de los datos. Aun así, los métodos no paramétricos avanzados como *random forest* resultan especialmente útiles para manejar muchas variables e interacciones complejas, y a menudo superan en precisión a las regresiones tradicionales.

2.2. Marco Legal

En la Tabla 3 se presenta el conjunto de leyes, normas y documentos que ofrece el Estado Peruano que serán de utilidad para la investigación.

Tabla 3

Marco Legal

Marco Legal	**Relevancia**
Ley General de Contrataciones Públicas - Ley N° 32069	Establece el marco normativo que regula todas las contrataciones de obras públicas, proporcionando el contexto normativo para el análisis de sobrecostos y retrasos. Entra en vigencia a partir del 22 de abril de 2025.
Ley General de Contrataciones Públicas - Ley N° 30225	Marco normativo anterior para regular las contrataciones de obras públicas. Las obras analizadas en la muestra estuvieron regidas por este marco debido a la temporalidad. Estuvo en vigencia hasta el 21 de abril de 2025.
Sistema Electrónico de Contrataciones del Estado (SEACE)	Sistema informático que permite el intercambio de información y difusión sobre las contrataciones del Estado. Será Fuente principal de datos sobre procedimientos de selección, contratos y su ejecución, además de modificaciones contractuales, ampliaciones de plazo y sobrecostos.
Sistema de Información de Obras Públicas (Infobras)	Sistema implementado por la Contraloría General que registra información sobre la ejecución de obras públicas. Fuente de información sobre avances físicos, presupuestales y estado de las obras públicas, permitiendo el análisis de desviaciones en plazos y costos.
Ley de Transparencia y Acceso a la Información Pública - Ley N° 27806	Promueve la transparencia de los actos del Estado y regula el derecho fundamental de acceso a la información pública. Es el Marco legal que garantiza el acceso a la información pública necesaria para la investigación
Código Civil del Perú	Proporciona el marco jurídico para el análisis de las causas de sobrecostos y retrasos relacionadas con incumplimientos contractuales y modificaciones.
Ley General del Sistema Nacional de Presupuesto - Ley N° 28411	Marco normativo para entender las limitaciones presupuestales y los procesos de modificación presupuestal que pueden influir en sobrecostos de obras públicas.
Consulta Amigable - MEF	Herramienta de acceso público para obtener información sobre ejecución presupuestal de obras públicas, complementando los datos del SEACE e Infobras.
Ley Orgánica del Sistema Nacional de Control - Ley N° 27785	Establece el marco legal para el control gubernamental de obras públicas y define las responsabilidades de las entidades en el registro de información, fundamentando la obligatoriedad del uso de Infobras.

2.3. Marco Teórico

a. Conceptos de Gestión de Obras

- Contrataciones del Estado: Marco normativo e institucional que regula de forma integral los procesos de gestión contractual y ejecución de obras públicas bajo los lineamientos de la Ley N.° 32069 y su reglamento. Comprende las modalidades de selección, los adicionales de obra y los deductivos vinculantes para garantizar la eficiencia en el uso de los recursos del Estado.
- Gestión de proyectos: La dirección de proyectos aplica conocimientos y técnicas a las actividades de un proyecto para cumplir con sus requisitos y generar valor. Su marco de referencia es el triángulo de hierro, que equilibra tres restricciones interdependientes: el costo, el plazo y el alcance (Project Management Institute, 2021).
- Gestión de riesgos: Es el proceso de identificar, analizar y responder a los eventos inciertos que pueden afectar los objetivos de la obra, priorizándolos según su probabilidad e impacto (Project Management Institute, 2021).
- Gestión de costos: Comprende la estimación, el presupuesto y el control del costo de la obra. Su principal desviación es el sobrecosto (*cost overrun*), definido como la cantidad en que los gastos reales exceden el costo planificado del proyecto (Asiedu & Ameyaw, 2020) o, según Flyvbjerg et al. (2002), la diferencia entre los costos pronosticados y los reales de construcción.
- Gestión del plazo: Comprende la programación y el control del cronograma de la obra. Su principal desviación es la ampliación de plazo (*delay*) definida como la extensión del cronograma contractual por causas atribuibles al cliente, al contratista o a factores externos (Larsen et al., 2016).

- Gestión del alcance: Comprende la definición y el control de las metas físicas de la obra respecto a lo previsto en el expediente técnico, delimitando qué trabajo está incluido en el proyecto y qué queda estrictamente excluido. Su objetivo principal es asegurar que se edifique exactamente lo contratado, previniendo desviaciones contractuales como sobrecostos, retrasos o el fenómeno conocido como *scope creep* (crecimiento descontrolado del alcance sin presupuesto adicional).
- Adicionales de Obra: Son trabajos no previstos requeridos para el cumplimiento del alcance del contrato, los cuales pueden deberse a fallas del contratista, falta de documentación o problemas de diseño.
- Paralización de Obra: Se refiere a la detención temporal de un proyecto, servicio u obligaciones contractuales debido a condiciones específicas delineadas en un acuerdo (Cobrief, 2025).
- Variables *ex-ante*: Son las características de la obra conocidas antes de iniciar la construcción, tomadas del expediente técnico y de las condiciones del proyecto. Comprenden factores técnicos (tipología de la obra y condiciones del terreno y de la fundación), contractuales (monto contratado, plazo inicial, sistema de contratación, modalidad de selección y tipo de contratista) y contextuales.

b. Conceptos Estadísticos

- Correlación de Pearson y Spearman: Miden la fuerza de asociación entre dos variables: Pearson capta relaciones lineales y Spearman, al operar sobre rangos, capta relaciones monótonas y es robusta ante la asimetría y los valores atípicos.
- Información mutua: Cuantifica la reducción de incertidumbre sobre una variable al conocer otra, sin suponer una forma funcional, por lo que detecta dependencias no lineales que la correlación lineal no revela.

- Pruebas estadísticas de comparación de grupos: Son pruebas basadas en rangos para comparar una variable entre grupos: Kruskal-Wallis contrasta tres o más grupos, la U de Mann-Whitney es su equivalente para dos grupos, y la prueba *post-hoc* de Dunn con ajuste de Bonferroni identifica qué pares de grupos difieren cuando Kruskal-Wallis resulta significativa.
- Winsorización: Técnica que acota los valores extremos de una variable a un percentil determinado, atenuando el efecto de las colas largas sin eliminar observaciones.
- Transformación logarítmica y estandarización: El logaritmo estabiliza la varianza de variables monetarias muy asimétricas, mientras que la estandarización *z-score* expresa cada variable con media cero y desviación unitaria, condición necesaria para algoritmos basados en distancias como *K-means*.

c. Analítica de datos aplicada a la gestión de la construcción

- Construcción 4.0: Corresponde a la incorporación de tecnologías digitales y de analítica de datos en la ingeniería civil; en este marco, el *machine learning* se emplea como herramienta para gestionar el riesgo de la obra y no como un fin en sí mismo.
- *Machine learning*: Es el campo de estudio de la inteligencia artificial que se ocupa del desarrollo y estudio de algoritmos estadísticos que pueden aprender de los datos y generalizar a datos no vistos, realizando así predicciones y detectando patrones significativos (Xu et al., 2021).
- Modelos Predictivos: Los modelos predictivos son algoritmos de *machine learning* que buscan crear un modelo descubriendo relaciones entre las características y los datos de salida y luego hacen predicciones de los valores de respuesta para un nuevo conjunto de datos (MathWorks, 2025).

- Datos estructurados y no estructurados: Los datos estructurados se organizan en formato tabular (filas y columnas), mientras que los no estructurados no siguen un esquema fijo (documentos, actas y expedientes); esta investigación integró ambos tipos a partir de los registros de obra.

d. Algoritmos de *machine learning*

- Modelos Supervisados: Son un tipo común de algoritmo que utiliza un conjunto de datos previamente etiquetado para entrenar un modelo capaz de establecer una relación entre los datos de entrada y una salida específica. Su objetivo principal es predecir resultados o clasificar nuevos datos (IBM, 2023) Entre los algoritmos de aprendizaje supervisado se encuentran la regresión lineal y la regresión cuadrática, los cuales se utilizan para predecir un valor de salida cuya relación con los datos de entrada puede modelarse de forma continua, ya sea mediante una función lineal o un polinomio de grado n. (Maldonado de la Torre, 2025)
- Modelos No Supervisados: Son algoritmos de *machine learning* para analizar y agrupar conjuntos de datos no etiquetados de datos. Estos algoritmos descubren patrones ocultos o agrupaciones de datos sin necesidad de intervención humana (MathWorks, 2025).
- *Random forest*: es una técnica de aprendizaje automático basada en el uso conjunto de múltiples árboles de decisión. Esta metodología combina los resultados individuales de varios árboles, los cuales analizan diferentes subconjuntos aleatorios de los datos. En el caso de la clasificación, los resultados se integran mediante votación, mientras que en tareas de regresión se utiliza el promedio, lo que permite obtener predicciones más precisas y robustas (IBM, 2021).
- Agrupamiento (*Clustering*): Es una técnica de minería de datos que consiste en agrupar datos no etiquetados en función de sus similitudes o diferencias. Este

proceso organiza objetos en conjuntos, llamados clústeres, cuyos elementos comparten características similares entre sí y se distinguen claramente de los pertenecientes a otros grupos (MathWorks, 2025).

- *Least Absolute Shrinkage and Selection Operator* (LASSO): Es una técnica de regresión lineal regularizada que añade una penalización matemática (normal L_1) a la función de pérdida del modelo. Esta metodología ajusta los coeficientes de las variables independientes castigando su magnitud, lo que obliga a los predictores menos relevantes o redundantes a reducir su peso a cero exacto ($\beta = 0$).

e. Métodos de Validación

- Validación Interna: Consiste en la aplicación de la técnica de validación cruzada (por ejemplo, *K-fold*), la cual divide la muestra de datos en k subconjuntos de igual tamaño para entrenar y evaluar el modelo de manera rotativa. Este procedimiento garantiza que todas las observaciones sean utilizadas como prueba, permitiendo estimar la estabilidad del algoritmo, minimizar el riesgo de sobreajuste (*overfitting*) y proporcionar una evaluación insesgada de la capacidad predictiva del modelo ante datos no observados.
- Validación externa: Consiste en evaluar el modelo entrenado sobre un conjunto distinto de proyectos (no usados en el entrenamiento) para verificar su capacidad de generalizar sobre nuevos entornos, midiendo métricas de desempeño y comparando sus predicciones con los resultados reales disponibles.
- Métricas de desempeño predictivo: Son los indicadores matemáticos utilizados para cuantificar la precisión, el error y la capacidad de discriminación de los modelos desarrollados. En los problemas de regresión continua, el coeficiente de determinación (R^2) mide la proporción de la varianza total de la variable dependiente que es explicada exitosamente por el modelo respecto a la media

histórica, su valor óptimo es 1.0. Para dimensionar el error en las mismas unidades de medida físicas de la variable objetivo se emplean el Error Absoluto Medio (MAE), que refleja la magnitud promedio de las equivocaciones del modelo expresándose de manera lineal; y la Raíz del Error Cuadrático Medio (RMSE), que calcula la desviación estándar de los residuos; al elevar las diferencias al cuadrado penaliza severamente los errores de gran magnitud. Por otro lado, para los modelos de clasificación, se utiliza el área bajo la curva de Característica Operativa del Receptor (ROC), la cual evalúa la capacidad global del modelo para discriminar y ordenar correctamente los proyectos críticos frente a los no críticos, independientemente del umbral de corte elegido.

3. METODOLOGÍA

3.1. Tipo y Nivel de Investigación

La presente investigación es de enfoque cuantitativo, dado que aborda el problema mediante la recolección objetiva y la estructuración numérica de los registros contractuales de obras públicas liquidadas, y contrasta sus hipótesis a través de procedimientos estadísticos y de modelamiento matemático y computacional (Hernández-Sampieri & Mendoza, 2020). La totalidad de las variables del estudio, tanto los predictores previos a la ejecución como los objetivos de desempeño contractual (sobrecosto, retraso y avance físico), se expresaron como magnitudes numéricas o categorías codificadas, y su tratamiento se sustentó en estadística descriptiva e inferencial, pruebas de asociación no paramétricas (Kruskal-Wallis y U de Mann-Whitney), información mutua, aprendizaje no supervisado (*k-means*) y modelos supervisados de *machine learning*.

Por su nivel o alcance, la investigación es de carácter explicativo y predictivo (Hernández-Sampieri & Mendoza, 2020). En su dimensión explicativa, analiza las relaciones no lineales y los mecanismos de interacción multivariada entre las condiciones previas a la ejecución contractual (factores técnicos, administrativos y contextuales del expediente) y la generación de contingencias de costo y plazo. En su dimensión predictiva, desarrolla, entrena y valida algoritmos de *machine learning* supervisado orientados a jerarquizar y clasificar el nivel de riesgo de sobrecostos y retrasos antes del inicio de las obras.

De este modo, el estudio no se limita a describir las obras públicas, sino que busca explicar los factores que condicionan el desvío contractual y anticipar su ocurrencia en proyectos futuros de infraestructura pública, en coherencia con el objetivo general de la investigación. Asimismo, se propone una herramienta práctica de alerta temprana orientada a optimizar la gestión de riesgos en el ámbito público

3.2. Diseño de Investigación

La investigación adoptó un diseño no experimental, de corte transversal y de tipo correlacional-predictivo. No es experimental porque las variables no se manipularon, sino que se observaron y registraron tal como se presentaron en obras públicas ya ejecutadas y liquidadas (Hernández-Sampieri & Mendoza, 2020). Es transeccional porque los datos se recolectaron en un único momento, correspondiente a proyectos liquidados entre 2020 y 2026 en la región de Arequipa. Y es correlacional-predictivo porque el estudio no solo identificó las asociaciones entre las condiciones de preinversión y los desvíos de costo y plazo, sino que además construyó y validó modelos capaces de anticipar y clasificar dichos desvíos en proyectos de infraestructura pública con características similares.

3.3. Métodos de Investigación

La presente investigación se enfoca en desarrollar modelos predictivos que permitan anticipar la ocurrencia de sobrecostos y retrasos en proyectos de infraestructura pública en Perú. El proceso se articula en torno al triángulo de hierro de la gestión de proyectos (costo, plazo y alcance) y la gestión de riesgos, tomando como referencia metodologías analizadas en los artículos científicos revisados las cuales fueron clasificadas como se muestra en la Tabla 4.

Tabla 4

Codificación de Metodología

Código	Descripción
M1.1	Revisión de literatura para definición de variables críticas
M1.2	Recopilación y procesamiento de datos
M1.3	Modelado y análisis supervisado
M1.4	Análisis no supervisado
M1.5	Análisis estadístico
M1.6	Evaluación y validación interna
M1.7	Evaluación y validación externa

Durante la primera etapa, se realizó una revisión bibliográfica de literatura científica y técnica (M1.1), con el propósito de identificar y consolidar las variables más frecuentes asociadas a sobrecostos y ampliaciones de plazo en proyectos de infraestructura. Las fuentes

incluyeron artículos científicos de revistas indexadas en las bases de datos más rigurosas de la actualidad: Scopus y Web of Science. Las ecuaciones de búsqueda utilizadas fueron conformadas por la combinación de las palabras clave "*delay", "cost overruns", "infrastructure", "project management", "risk management"*. A partir de esta revisión, se identificaron y categorizaron las principales variables de influencia.

En la segunda etapa, de acuerdo con las variables identificadas en el punto anterior, se realizó la recolección de información para la construcción del *dataset* (M1.2). Se recopilaron datos reales de proyectos de mediana envergadura, liquidados entre 2020 y marzo de 2026 de la región de Arequipa. Esta información fue obtenida de portales oficiales como SEACE del Organismo Supervisor de las Contrataciones del Estado (OSCE), Infobras de la Contraloría General de la República y SSI del Ministerio de Economía y Finanzas mediante una auditoría directa a los documentos contractuales y de liquidación. Una vez recopilada, la información fue sometida a un proceso de limpieza, filtrado y selección de parámetros relevantes. Tras la depuración y la verificación de consistencia sobre la muestra, se realizó un análisis exploratorio de datos (M1.5) que incluye estadística descriptiva, correlaciones de Pearson y Spearman, factor de inflación de la varianza (VIF), información mutua y contraste de dependientes según variables categóricas; estos resultados fundamentan las decisiones de transformación de datos. La lectura e interpretación de expedientes, liquidaciones, adicionales y ampliaciones se realizó bajo el marco de la normativa de contrataciones del Estado.

Posteriormente se desarrolló el aprendizaje no supervisado (M1.4) mediante *k-means clustering* sobre las variables de desempeño, con esto se clasificaron los arquetipos de obra. Más que un fin descriptivo, la etiqueta de clúster cumple una función instrumental, esta se emplea como criterio de estratificación de los pliegues en la validación cruzada del modelamiento supervisado, garantizando que cada pliegue represente de forma equilibrada los tres perfiles de desempeño.

Después se desarrolló el modelamiento supervisado (M1.3) con un enfoque por variable objetivo. Primero, mediante algoritmos como Selección Secuencial Hacia Adelante, guiada por el desempeño (R^2), se identificó para cada objetivo el subconjunto de predictores más informativo. Segundo, sobre ese subconjunto se entrenaron los modelos de regresión de línea base, LASSO lineal, LASSO cuadrático y *random forest*. La validación interna (M1.6) se realizó con validación cruzada repetida y estratificada por el clúster de desempeño conseguido en la etapa anterior (M1.4), reportando R^2, MAE y RMSE como media y desviación estándar, con el fin de estabilizar las métricas dada la muestra reducida. Finalmente, en función de los resultados, se reformuló una clasificación binaria de riesgo orientada a una alerta temprana. Se utilizaron los modelos de regresión logística y *random forest*, evaluados por el área bajo la curva (AUC), y se fijó un punto de corte que permitiera traducir el modelo en una regla operativa expresada en términos de sensibilidad y especificidad. Todos los resultados numéricos se generaron con semilla fija (*random_state* = 42).

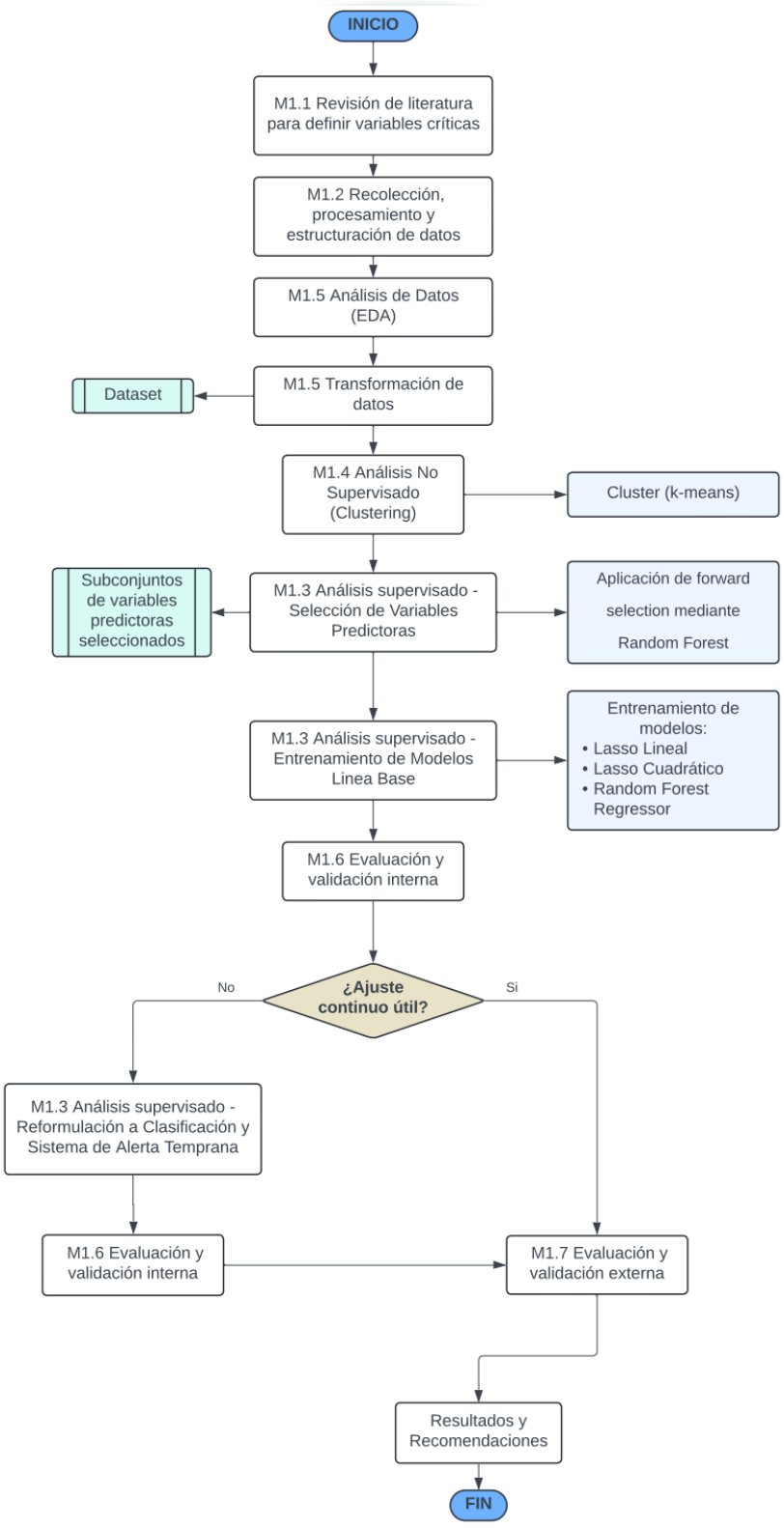
En la cuarta etapa, se realizó la validación externa de los modelos de aprendizaje supervisado desarrollados, aplicándolos a los 10 proyectos liquidados de los departamentos de Cusco, Puno y Tacna (M1.7). Esto permite evaluar el desempeño del modelo en casos reales de obras civiles, verificando su capacidad de generalización en contextos no observados previamente. Como producto final de este proceso analítico, los modelos predictivos se materializan en una herramienta de gestión preventiva denominada Semáforo de Alerta Temprana.

Finalmente, en la quinta etapa, con base en los modelos desarrollados y las variables identificadas como más influyentes, se identificaron los causales principales y su correlación con la generación de sobrecostos y ampliaciones de plazo en obras públicas. Adicionalmente, se presenta un set de recomendaciones basadas en análisis de datos reales, orientadas a mejorar

la toma de decisiones en el control y planeamiento de la gestión de obras públicas. En la Figura 1 se presenta el esquema metodológico planteado.

Figura 1

Esquema Metodológico



4. TRATAMIENTO DE INFORMACIÓN

4.1. Definición de Variables

4.1.1. *Selección de Variables Independientes*

La identificación y selección de las variables independientes se fundamenta en una revisión sistemática de la literatura científica reciente orientada a modelos de *machine learning* aplicados a la estimación y predicción de sobrecostos y retrasos en proyectos de construcción, cuya síntesis metodológica y conceptual se consolida en la Tabla 5.

Tabla 5

Resumen de Variables Independientes Utilizadas

Autor	Ámbito	Método Usado	Variables Independientes	Variables Dependientes
Gómez-Cabrera et al. (2025)	Proyectos de infraestructura educativa (colegios públicos) en Colombia	*Random forest regression* (RFR).	Valor del contrato adjudicado, Duración inicial (plazo contractual), Ajuste de oferta (*Award Growth*), Intensidad de proyecto (*Project Intensity*), Tiempo suspendido (%), Número de postores, Tipo de proceso de contratación, Tipo de proyecto (Obra nueva / Mejoramiento), Año de licitación, Región geográfica.	Desviación de Costo (%). Desviación de Tiempo (%).
Andrić et al. (2024)	Proyectos de infraestructura de transporte y energía (carreteras, puentes, ferrocarriles, etc.) en Asia del Sur	Regresión Lineal y Cuadrática, *random forest regression* (RFR).	Tipo de proyecto, Ubicación geográfica (país), Tamaño del proyecto (Costo estimado), Período de implementación (duración planificada), Desvío de tiempo (retraso acumulado).	Sobrecosto económico (%).
Andrić et al. (2025)	Proyectos de infraestructura civil, vial y energética en la región de Asia del Sur.	Regresión Lineal y Regresión Cuadrática, *random forest regression* (RFR).	Tipo de proyecto, Ubicación geográfica (país), Costo estimado del proyecto (tamaño), Duración planificada de la obra, Porcentaje de sobrecosto (económico).	Retraso temporal (%).
Restrepo Ramírez & Rúa Machado (2025)	Proyectos de construcción residencial (cooperativas de	*k-nearest neighbors regression* (KNNR) y	Avance físico e histórico programado y real, Flujos de caja reales gastados en intervalos de tiempo, Inversión económica inicial pre-construcción,	Retraso final acumulado (%).

	vivienda) en Colombia.	*random forest* r*egression* (RFR).	Cronograma Ganado (ES), Planned Value (PV), Earned Value (EV).	Sobrecosto económico total (%).
Nassar & Elbisy (2024)	Proyectos de construcción marina (puertos, estructuras de gas/petróleo onshore y offshore, rompeolas, etc.).	Máquinas de Vectores de Soporte (SVM), *Tree boost regression*.	43 factores que influyen en los retrasos, agrupados en 9 categorías de riesgo (Contratista, Externos, Consultor, Propietario, Diseñador, Contrato, Recursos, Fuerza Mayor, Todas las partes).	Porcentaje de retraso total (*%*).
Ashtari et al. (2022)	Proyectos de construcción en general (viviendas, edificaciones, obras públicas, etc.).	Bayesian Network Classifier (BN), Naive Bayes (NB) y Árboles de Decisión (DT).	43 riesgos de sobrecosto (reducidos a 8 variables de alta importancia mediante elección de Características Basada en Correlación (CFS): incremento de materiales, inflación, falta de conocimiento/experiencia del recurso humano, políticas aduaneras, escasez de equipos, etc.).	Nivel categórico final del sobrecosto (*Bajo, Moderado, Alto*).
Al-Hazim et al. (2017)	Proyectos de infraestructura vial y redes de agua/saneamiento en Ammán, Jordania.	Procesamiento matemático de desvíos, Clasificación y *Ranking* de causas.	20 causas y factores reales extraídos de la correspondencia oficial de obra (condiciones del terreno y topografía, condiciones climáticas, nuevas órdenes de variación/adicionales de obra, errores en el diseño o expediente técnico, deficiente planificación de tiempo y costo).	Porcentaje de sobrecosto final (%). Porcentaje de retraso de tiempo (%).
Coffie & Cudjoe (2023)	Proyectos de edificación de infraestructura educativa gubernamental en Ghana.	*Support Vector Machine* (SVM) con *kernels* Lineal, Función de Base Radial y Polinomial.	Costo contractual inicial, Plazo contractual inicial, Área construida bruta, Número de pisos, Tipo de contrato, Suma provisional, Método de contratación, Modificaciones de alcance.	Magnitud de costo real de finalización de la obra (moneda local).
Sahu et al. (2024)	Proyectos de construcción de obras públicas diversas (viviendas, plantas de energía, carreteras, puentes, redes de saneamiento, etc.) en India (50 proyectos).	Relative Importance Index (RII), *random forest, Genetic Algorithm* (RFGA).	40 causas de reclamo organizadas por etapas del proyecto (Las principales resultaron: Retraso en el pago de valorizaciones mensuales, desastres naturales, Falta de conocimiento del contrato, Retraso en pagos, Condiciones deficientes del sitio de trabajo, Retrasos en la aprobación de cronogramas y órdenes de cambio / adicionales de obra).	Retraso del proyecto (%). Ocurrencia de conflictos y disputas entre partes.

Los antecedentes convergen en un conjunto de factores estructurales del proyecto en varias etapas y su entorno contractual, las variables con la relevancia predictiva más representativa son: 1) Plazo inicial, 2) Monto inicial, 3) Órdenes de cambio / Variaciones de alcance, 4) Tipo de obra, 5) Retrasos de pago / Dificultades financieras del cliente, 6) Condiciones de sitio / Características geotécnicas, 7) Tipo de contrato / Método de contratación, 8) Ubicación, 9) Ajuste de oferta / Número de postores, 10) Factores del cliente.

Las investigaciones de Gómez-Cabrera (2025), Andrić (2024, 2025) o Coffie & Cudjoe (2023), incurren en fuga de datos (*data leakage*) al alimentar sus modelos con variables *ex-post* conocidas únicamente al cierre de la construcción. Utilizar variables tales como el sobrecosto económico, retraso de tiempo, tiempo de paralización y órdenes de cambio aprobadas, mejoran artificialmente el desempeño de los modelos entrenados porque la máquina deduce relaciones obvias, por ejemplo, a más adicionales aprobados, mayor costo final. Otras investigaciones, como la realizada por Restrepo & Rúa (2025), utilizan curvas de avance físico real e histórico, flujos de caja reales gastados en intervalos de tiempo para ajustar predicciones en plena ejecución; estas tienen valor como herramientas de monitoreo y control de obra en campo durante la ejecución de un proyecto.

En concordancia con Larsen et al. (2016), la literatura científica demuestra que el esfuerzo invertido en las fases iniciales del proyecto es determinante para asegurar el cumplimiento del costo y plazo final. Por ello, la presente investigación se restringe de manera estricta al uso de variables predictoras *ex-ante* disponibles en el "Día 0" (a la firma contractual), orientándose a clasificar proyectos con riesgo de desviación e identificar sus factores más influyentes. Este enfoque garantiza el desarrollo de una herramienta preventiva de alerta temprana con capacidad de generalización ante nuevos proyectos públicos, cumpliendo con el valor estratégico de la inteligencia artificial como un mecanismo preliminar de "*benchmarking*" en etapas tempranas de planeamiento (Tochaiwat & Budda, 2026). En la

Tabla 6 se presenta el detalle de las variables independientes planteadas y los criterios de clasificación para estas.

Tabla 6

Banco de Variables Independientes Adoptado para Modelo Predictivo

Variable Independiente	**Tipo**	**Fuente**	**Descripción**
Monto Adjudicado	Numérica	Infobras / SEACE. Contrato	Costo base del contrato original con IGV. Presente en la mayoría de antecedentes revisados como predictor estructural.
Plazo de Ejecución Inicial	Numérica	Infobras / SEACE. Contrato	Días calendario pactados originalmente. Predictor estándar en la literatura de retrasos.
Naturaleza de Intervención	Categórica binaria	Infobras	Obra nueva vs. intervención preexistente (mejoramiento, ampliación, rehabilitación). Diferencia el riesgo de hallazgos en sitio.
Tipología de Obra	Categórica (4 clases)	Infobras / Clasificación propia	La naturaleza de la infraestructura. Se ha establecido la siguiente clasificación: • Infraestructura Vial y Transporte: Proyectos lineales puros que conectan dos puntos geográficos distintos (pueblos, provincias). Carreteras, trochas carrozables, vías interurbanas de evitamiento, grandes puentes de conexión. • Infraestructura Urbana y Espacios Públicos: Corresponde a una intervención intraurbana, como el mejoramiento de la habitabilidad dentro de la ciudad (incluye el asfalto local y ornato). Mejoramiento de calles (pistas urbanas y veredas), parques, plazas, bermas, ciclovías, losas. • Infraestructura Pública: Estructuras arquitectónicas cerradas para albergar servicios del Estado. Colegios, pabellones universitarios, hospitales, postas, comisarías. • Saneamiento e Hidráulica: Manejo del recurso hídrico (consumo, agricultura, residuos) y protección fluvial. Redes de agua/desagüe, PTAR, canales de riego.
Sistema de Contratación	Categórica	SEACE	Variable estructural crítica por su distinto perfil de riesgo ante cambios de metrados. Suma Alzada, Precios Unitarios, Esquema Mixto.
Modalidad de Selección	Categórica	SEACE	*Proxy* de competencia y transparencia del proceso. Licitación Pública, Adjudicación Simplificada, Procedimiento Especial de Contratación.

Tasa de Pobreza Distrital	Numérica	INEI, Mapa de Pobreza 2018	*Proxy* de ubicación según distrito, representa la capacidad institucional local y la densidad de servicios técnicos disponibles en la ubicación de la obra.
Ubicación Metropolitano	Binaria	INEI (2018), Censos Nacionales 2017	*Proxy* binario de ubicación que distingue las obras del área metropolitana de Arequipa de las situadas fuera de ella según la agregación censal "Arequipa Metropolitana"(INEI, 2018).
Nivel de Gobierno	Categórica ordinal	Infobras	Local (1), Regional (2), Nacional (3). *Proxy* de capacidad técnica e institucional de la entidad ejecutora.
Número de postores efectivos	Numérica	SEACE. Acta de Otorgamiento de Buena Pro.	*Proxy* de la presión competitiva que mide la cantidad de ofertas admitidas en el proceso de selección.
Competitividad / Ajuste de Oferta	Numérica	Infobras / SEACE	Captura el comportamiento estratégico del postor ganador y los efectos del sobreajuste de precios. Según Ley de Contrataciones del Estado del Perú, Ley N° 30225 y su Reglamento, las ofertas económicas en la ejecución de obras públicas están sujetas a un rango de variación legal estricto: no pueden ser menores al 90% del valor referencial; en consecuencia, el rango de variación de esta variable queda acotado entre 0% y 10%. (Valor Referencial − Monto Adjudicado) / Valor Referencial.
Tipo de contratista	Binaria	SEACE	Distingue la naturaleza del adjudicatario. (Consorcio=1 / Individual=0).
Familia SUCS	Categórica	SEACE / INGEMMET. Estudio de Mecánica de Suelos.	Macro-categoría del suelo representativo según el Sistema Unificado de Clasificación de Suelos, extraído del Estudio Mecánico de Suelos (EMS). Se diferencian las categorías de: Grave, Arena, Fino, Suelo Especial.
Calidad de fundación	Ordinal	SEACE. Estudio de Mecánica de Suelos.	Clasifica a capacidad portante admisible del suelo en baja/media/alta. Se consideran los umbrales para clasificación de: Para q_adm: baja < 1.5 kg/cm², media 1.5–3.0, alta > 3.0. Para CBR de subrasante (categorías MTC/AASHTO): baja < 6%, media 6–20%, alta > 20%.
Material inadecuado superficial	Binaria	SEACE. Estudio de Mecánica de Suelos.	Toma el valor 1 cuando el EMS menciona relleno no controlado o antrópico, o suelo colapsable/deleznable (pómez/puzolánico), o suelo orgánico (turba, materia orgánica) significativo; toma 0 en caso contrario

Escenario COVID	Categórica nominal	Derivada de fecha de inicio	Clasifica cada obra según su ventana de ejecución en tres regímenes. Pre COVID, Pandemia, Post COVID. Variable de control exógeno bajo el marco de la Directiva N° 005-2020-OSCE/CD que permite atribuir la fracción de demoras provocada por restricciones sanitarias a un régimen causal exógeno común, impidiendo que dicho efecto distorsione la estimación de las variables endógenas del proyecto.

4.1.2. *Selección de Variables Dependientes*

La definición de las variables dependientes se estructuró en torno al "triángulo de hierro" de la gestión de proyectos (Project Management Institute, 2021), el cual reconoce tres dimensiones interdependientes del desempeño: costo, plazo y alcance. Este marco conceptual, adoptado asimismo por Larsen et al. (2016), permite descomponer el desempeño global en indicadores medibles e independientes, evitando la formulación de métricas compuestas sujetas a una ponderación arbitraria de pesos.

Para las dimensiones de costo y plazo se separó una variable bruta, que captura el impacto financiero y temporal agregado sobre la Entidad, de una variable endógena referida a órdenes de cambio (prestaciones adicionales de obra o ampliaciones de plazo). Esta segunda variable aísla la señal predecible a partir de las características estructurales del proyecto y deja fuera los *shocks* exógenos que no dependen de la calidad del expediente técnico. Esta distinción endógeno-exógeno es central en la investigación y responde al criterio metodológico establecido por el protocolo AACE International para el análisis de retrasos en construcción RP 29R-03, según el cual las causales de desviación deben separarse por naturaleza (excusable compensable, excusable no compensable, no excusable) para permitir inferencias causales limpias (AACE International, 2011).

La necesidad de esta separación radica en la naturaleza de las fuentes generadoras de desvío, mientras que las modificaciones endógenas derivan de deficiencias en los estudios de ingeniería o contingencias técnicas de la ejecución, los factores exógenos corresponden a

causas externas no imputables a la calidad del diseño (tales como paralizaciones por fuerza mayor o disposiciones normativas). De este modo, aislar el componente endógeno permite modelar con precisión la variabilidad vinculada a la planificación y gestión técnica del proyecto, a la vez que el análisis conjunto de ambos tipos de variables brinda una visión integral y multidimensional sobre los mecanismos que detonan los sobrecostos y retrasos en la infraestructura pública. La Tabla 7 presenta las cinco variables dependientes y su rol conceptual en el marco.

Tabla 7

Variables Dependientes del Modelo Predictivo

Cod.	Variable Dimensión	Rol	Fórmula
BA	% Vulnerabilidad Financiera Total	Costo (bruta)	(Costo Final Liquidado − Monto Adjudicado) / Monto Adjudicado
BB	% Incidencia Económica por Adicionales Endógenos	Costo (endógena)	Σ (Adicionales Aprobados − Deductivos Vinculantes) / Monto Adjudicado
BC	% Avance Físico	Alcance	Monto Ejecutado del Alcance Original / Monto Adjudicado
BD	% Vulnerabilidad del Plazo Total	Plazo (bruta)	(Plazo Cronológico Real − Plazo Contractual Inicial) / Plazo Contractual Inicial
BE	% Incidencia Temporal por Ampliaciones Endógenas	Plazo (endógena)	Σ Días de Ampliación de Plazo Aprobados / Plazo Contractual Inicial

La información empleada para la conformación de la muestra procede de fuentes oficiales (Infobras y SEACE), cuyos registros se asumen formalmente válidos y verídicos; por consiguiente, los análisis e interpretaciones se fundamentan estrictamente en lo consignado en dichos documentos, sin formular suposiciones sobre la existencia de datos no registrados en las plataformas.

a. Porcentaje de Vulnerabilidad Financiera Total (BA)

Mide la desviación presupuestal total acumulada por la entidad pública al cierre de la liquidación final de la obra respecto al monto contratado originalmente.

$$BA = \frac{(Costo\ Final\ Liquidado\ -\ Monto\ Adjudicado)}{Monto\ Adjudicado}$$

Representa el impacto financiero integral del proyecto. Agrupa todas las fuentes de incremento monetario: adicionales de obra, reajustes de precios por fórmulas polinómicas, mayores metrados, laudos arbitrales, compensaciones, mayores gastos generales variables.

b. Porcentaje de Incidencia Económica por Adicionales Netos (BB)

Cuantifica la proporción del presupuesto original que fue modificada estrictamente mediante presupuestos adicionales aprobados (deduciendo los deductivos vinculantes) debido principalmente a variaciones técnicas en el expediente.

$$BB = \frac{\sum Adicionales\ Aprobados\ -\ \sum Deductivos\ Vinculantes}{Monto\ Adjudicado}$$

Esta variable aísla el impacto financiero directo derivado de la calidad del diseño, como errores en el metrado, omisiones de partidas o vicios ocultos del terreno. En los casos donde la liquidación, el acta de recepción o las resoluciones no registran conceptos por prestaciones adicionales, se descarta cualquier estimación presuntiva y se asume de forma estricta que el proyecto no requirió modificaciones técnicas derivadas del expediente original.

c. Porcentaje de Cumplimiento de Avance Físico respecto al Alcance Inicial (BC)

Mide la fracción del alcance y presupuesto originales del expediente técnico que efectivamente se mantuvo, ejecutó y liquidó al cierre de la obra.

$$BC = \frac{Monto\ Ejecutado\ del\ Alcance\ Original}{Monto\ Adjudicado}$$

Debido a que la totalidad de los proyectos estudiados serán concluidos y liquidados en condiciones de plena operatividad, una disminución en el indicador de cumplimiento del alcance original no denota que la obra se encuentre inconclusa o en estado de abandono; expresa que una fracción de las metas y partidas contempladas en el diseño inicial fue suprimida, sustituida o reformulada mediante deductivos para asegurar la viabilidad técnica y operativa del proyecto. Esta variable cuantifica la fidelidad de la ejecución material respecto a la concepción original formulada en el expediente técnico.

d. Porcentaje de Vulnerabilidad del Plazo Contractual Total (BD)

Cuantifica el incremento total de tiempo transcurrido en el calendario desde la fecha del acta de inicio de obra hasta la fecha del acta de recepción física final, respecto al plazo contractual inicial.

$$BD = \frac{Plazo\ Cronológico\ Real - Plazo\ Contractual\ Inicial}{Plazo\ Contractual\ Inicial}$$

Representa el desvío temporal bruto que afectó a la obra. Incluye tanto el tiempo concedido por ampliaciones formales como los días de suspensión de plazo y paralizaciones de carácter exógeno como cuarentenas por COVID-19, falta de disponibilidad de terreno, huelgas o lluvias, en los cuales el cómputo del plazo contractual estuvo legalmente detenido.

e. Porcentaje de Incidencia Temporal por Ampliaciones Aprobadas (BE)

Mide la proporción de tiempo adicional que fue formalmente aprobada mediante resoluciones de ampliación de plazo contractual derivadas de modificaciones técnicas o mayores trabajos.

$$BE = \frac{\sum Días\ de\ Ampliación\ de\ Plazo\ Aprobados}{Plazo\ Contractual\ Inicial}$$

Aísla el retraso contractual originado por la necesidad de ejecutar partidas adicionales o subsanar modificaciones de ingeniería. Un valor nulo en este indicador evidencia que el contrato no requirió la aprobación de prórrogas por causas técnicas. Por consiguiente, cuando un proyecto registra una dilatación considerable en su plazo global, en conjunto con un valor nulo en esta variable de prórrogas técnicas, el desfase temporal se explica íntegramente por suspensiones de plazo o paralizaciones de carácter exógeno, descartando la necesidad de ampliaciones originadas en deficiencias del expediente técnico.

4.2. Recolección de Datos

4.2.1. *Diagnóstico de Viabilidad*

La factibilidad de la investigación se apoyó en un diagnóstico previo de la disponibilidad y el acceso a información pública sobre obras liquidadas en el Perú. Aunque la Ley N.° 27806 de Transparencia y Acceso a la Información Pública obliga a publicar los actos del Estado, en la práctica los portales del sector construcción tienen niveles desiguales de completitud y trazabilidad, sobre todo en los gobiernos subnacionales (Traverso-Frisancho & Romero Alva, 2023). Esto coincide con lo que reporta la Contraloría General de la República (2025), que señala deficiencias sistemáticas en los sistemas de información como uno de los factores estructurales que dificultan el monitoreo de la inversión pública.

El diagnóstico se realizó sobre tres portales oficiales, cada uno con un rol funcional distinto dentro del ciclo de la contratación pública peruana:

- El Sistema Electrónico de Contrataciones del Estado (SEACE), administrado por el Organismo Supervisor de las Contrataciones del Estado (OSCE), constituye la fuente transaccional primaria del proceso contractual: publica las Bases Integradas, la nomenclatura del procedimiento de selección, el sistema de contratación y el contrato original firmado, así como sus modificaciones formalmente registradas.

- El Sistema de Información de Obras Públicas (Infobras), operado por la Contraloría General de la República bajo la Ley N.° 27785, registra el avance físico y financiero de las obras y almacena las actas de recepción y liquidación en formato PDF. Constituye la fuente autorizada para el cierre físico-financiero del contrato, e incluye cronogramas y valorizaciones según la completitud documental de cada obra. Esta fue la vía para obtener el volcado íntegro de los datos estructurados que sirvió como base para la construcción del *dataset*.
- El Sistema de Seguimiento de Inversiones (SSI) del Ministerio de Economía y Finanzas registra la ejecución presupuestal a nivel de Código Único de Inversión (CUI), agregando componentes que pueden incluir múltiples obras físicas bajo una misma inversión.

Complementariamente, se revisaron las páginas web institucionales de las entidades ejecutoras, en las cuales, en ocasiones, se encuentran publicadas las resoluciones de ampliaciones de plazo, modificaciones contractuales y paralizaciones emitidas durante la ejecución de la obra.

Ante la existencia de datos parcialmente redundantes y con niveles distintos de detalle y confiabilidad, se estableció una jerarquía de prelación de fuentes documentales para resolver eventuales discrepancias. Esta jerarquía prioriza el documento contractual formalmente cerrado (acta de recepción y liquidación de obra en Infobras) sobre los registros transaccionales intermedios (SEACE) y sobre los sistemas presupuestales agregados (SSI del MEF), dado que el SSI opera a nivel de CUI y frecuentemente agrupa varios contratos bajo una misma inversión, lo que impide atribuir de forma limpia el devengado presupuestal a un contrato específico.

4.2.2. *Técnicas de Recolección de Datos*

La construcción de la base de datos combinó dos técnicas complementarias, seleccionadas en función de la naturaleza y el formato de la información disponible en cada portal.

a. Extracción de bases de datos estructuradas

El portal Infobras ofrece la exportación de sus registros en formato tabular (CSV, XLSX), lo que permitió la extracción masiva de datos estructurados. Estos datos constituyeron la columna vertebral del *dataset*. Asimismo, se realizó las descargas de bases de datos provenientes de fuentes abiertas, tales como el Portal de Datos Abiertos del OSCE, el cual provee información sobre convocatorias, adjudicaciones y contratos de obras públicas.

b. Descarga manual de documentos oficiales y fuentes no estructuradas

La mayoría de la información relativa al cierre operativo de los contratos carece de sistematización en las plataformas oficiales. Aquellos registros disponibles en los sistemas, tales como el costo final de la obra y los montos correspondientes a prestaciones adicionales, presentaban inconsistencias y datos incompletos. Asimismo, los documentos de sustento (actas de recepción, resoluciones de liquidación, resoluciones individuales de adicionales, ampliaciones de plazo y actas de paralización) no se encuentran en formato estructurado, por lo que requirieron la descarga manual de archivos PDF desde las fichas individuales de cada proyecto en Infobras, SEACE y el SSI – MEF

De forma similar, para recopilar las variables independientes del número de postores admitidos y las características geotécnicas de la obra, fue necesaria la auditoría directa de las Actas de Otorgamiento de la Buena Pro, así como la evaluación y calificación de los Estudios de Mecánica de Suelos (EMS) contenidos en los expedientes técnicos. Toda la documentación legalmente accesible fue recopilada, verificada y organizada en repositorios locales estructurados por número de ítem, asegurando la trazabilidad y la integridad de los datos primarios que alimentan la matriz de investigación.

4.2.3. *Criterios de Filtrado de la Muestra*

La delimitación de la muestra respondió a un balance entre representatividad estadística y trazabilidad documental. Se aplicaron los siguientes criterios de inclusión y exclusión sobre el universo de obras públicas registradas en Infobras:

- Ámbito territorial: obras localizadas en el departamento de Arequipa. La delimitación regional controla heterogeneidad institucional y geográfica que introduciría ruido en un modelo entrenado a nivel nacional.
- Ámbito temporal: obras liquidadas entre 2020 y marzo de 2026. Este rango captura el ciclo pandémico completo y su recuperación posterior, permitiendo modelar el efecto del contexto COVID-19 mediante variables de control.
- Rango de inversión: valor referencial entre S/ 4.5 millones y S/ 30 millones, umbral que concentra la muestra en obras de mediana envergadura con documentación más completa y trazable, excluyendo contratos menores de baja densidad informativa y megaproyectos con dinámicas atípicas.
- Modalidad de ejecución: exclusivamente obras por contrata. Se excluyen administración directa, obras por impuestos y otras modalidades no contractuales, cuyos regímenes de modificación, control y liquidación son estructuralmente distintos y no comparables bajo el mismo marco de variables dependientes.
- Tipología: se excluyen obras principalmente eléctricas, electromecánicas y de telecomunicaciones (subestaciones, líneas de transmisión, instalación de fibra óptica), cuya estructura de costos está dominada por la adquisición de equipos, montajes electromecánicos, protocolos de prueba y no por trabajos civiles.
- Cierre contractual: se excluyen las obras cuyo contrato fue terminado por resolución, pues su costo y plazo reflejan el corte administrativo y no la ejecución completa,

- Omisión Documentaria: se excluyen las obras sin registro de cierre documental (acta o liquidación) sustentadas en un archivo oficial descargable en las fuentes oficiales.

4.2.4. *Diseño y Validación del Instrumento de Extracción*

La transformación de los documentos no estructurados en variables numéricas y categóricas requiere el diseño de una matriz de extracción estandarizada y un protocolo de fusión de datos con reglas explícitas, deterministas y reproducibles. La matriz define, para cada obra, un conjunto de columnas correspondientes a las variables, y el protocolo establece las reglas para poblarlas a partir de las fuentes, incluyendo las jerarquías de prelación en caso de discrepancia y los criterios de clasificación categórica.

Los aspectos centrales del protocolo incluyen: (i) la jerarquía de fuentes, que sitúa al acta de recepción como fuente primaria, la liquidación como complemento y las resoluciones individuales como fuente autoritativa condicional para la clasificación causal; (ii) reglas de aritmética de fechas que evitan errores de conteo en días de paralización y ampliación; (iii) los criterios de clasificación endógeno/exógeno de cada modificación; y (iv) verificaciones cruzadas de consistencia entre montos y plazos declarados en distintas fuentes.

Este procedimiento de validación por trazabilidad, además de garantizar la cobertura conceptual del instrumento respecto al fenómeno estudiado, asegura que la estructura de las variables (numéricas continuas, numéricas discretas, categóricas ordinales, categóricas nominales, binarias) sea apta para su procesamiento por algoritmos supervisados de *machine learning*, tanto en los modelos paramétricos de referencia (LASSO) como en los modelos basados en árboles de decisión (*random forest*). La especificación de los campos que conforman la matriz de extracción y sus definiciones operacionales se encuentran consolidadas en las Tablas 6 y 7, ya presentadas anteriormente.

4.3. Gestión y Preparación del *Dataset*

4.3.1. *Preparación y Estructuración de Datos*

La preparación de la base de datos consolidó un archivo tabular único a partir de la información procedente de dos flujos paralelos de recolección: los datos masivos estructurados extraídos de Infobras y del Portal de Datos Abiertos del OSCE, y las variables construidas manualmente mediante la extracción documental sobre archivos PDF conforme al protocolo descrito en la sección previa. Este procedimiento, denominado marco de fusión de datos, se ejecutó en dos etapas secuenciales.

En la primera etapa se construyó la base preliminar a partir del volcado tabular de Infobras, filtrada conforme a los criterios de selección establecidos en la sección de criterios de filtrado de muestra. La base resultante comprendió 112 obras con sus variables descriptivas generales: identificadores únicos (CUI, Código Infobras y Código SNIP), entidad ejecutora, denominación del proyecto, clasificadores tipológicos, ubicación geográfica, montos aprobados en el expediente técnico y montos adjudicados, plazos contractuales iniciales, fechas de aprobación del expediente, fechas de inicio y término real, así como el avance físico y financiero registrado al momento de la liquidación. Esta base constituyó el esquema inicial de la matriz de datos. A partir de este repositorio y empleando los identificadores únicos, se realizó el cruce de información para incorporar la nomenclatura del proceso en el SEACE, la modalidad de selección, el sistema de contratación y los montos contractuales, datos provenientes de los registros estructurados descargados directamente del Portal de Datos Abiertos del OSCE.

En la segunda etapa metodológica se llevó a cabo el acople horizontal de los datos no estructurados a partir de la auditoría manual de los archivos PDF (actas de recepción, liquidaciones técnico-financieras, resoluciones contractuales, actas de buena pro y estudios geotécnicos). Para cada proyecto de la muestra es necesaria la extracción de datos no

estructurados de forma manual; con el objetivo de completar las variables independientes y dependientes. Estos datos se detallan en la Tabla 8 según su rol funcional en el estudio.

Tabla 8

Detalle de Recolección de Datos No Estructurados

Categoría de Variable	Dato	Tipo de Dato	Documento de Sustento
Datos no estructurados – Variables Dependientes (Cierre Contractual)	Monto Neto de Adicionales Endógenos	Numérica continua	Acta de Entrega de Obra / Resoluciones de Adicionales y Deductivos Vinculantes / Liquidación de Obra
	Días de Ampliación Endógena	Numérica discreta	Acta de Entrega de Obra / Resoluciones de Ampliación de Plazo
	Días de Ampliación Exógena – Referencial	Numérica discreta	Acta de Entrega de Obra / Resoluciones de Ampliación de Plazo
	Días de Paralización / Suspensión Total – Referencial	Numérica discreta	Acta de Entrega de Obra / Cronogramas Actualizados de Obra / Actas de Suspensión y Reinicio de Plazo
	Cantidad Total de Adicionales y Ampliaciones – Referencial	Numérica discreta	Acta de Entrega de Obra / Resoluciones Contractuales / Liquidación
	Nivel de Evidencia / Confiabilidad Documental – Referencial	Categórica ordinal	Acervo documental en PDFs auditados.
Datos no estructurados – Variables Independientes Predictoras	Número de Postores Admitidos	Numérica discreta	Acta de Otorgamiento de la Buena Pro
	Macrofamilia Clasificación SUCS	Categórica	Estudio de Mecánica de Suelos (EMS)
	Material Inadecuado Superficial	Binaria	Estudio de Mecánica de Suelos (EMS)
	Calidad de Fundación del Suelo	Categórica ordinal	Estudio de Mecánica de Suelos (EMS)
	Parámetros Geotécnicos de Calidad de Fundación del Suelo (qadm / CBR) – Referencial	Numérica / Categórica	Estudio de Mecánica de Suelos (EMS)
	Nivel de Evidencia del EMS – Referencial	Categórica ordinal	Disponibilidad de Mecánica de Suelos / Datos de entorno

El acople se realizó por medio del Código Infobras como clave primaria en Infobras, CUI y Nomenclatura del proceso en SEACE garantizando la unicidad e integridad referencial

entre ambos flujos. La base de datos resultante constituye la matriz final sobre la cual se aplican las etapas posteriores de limpieza e ingeniería de características.

4.3.2. *Limpieza de Datos*

a. Valores faltantes e imputación

La disponibilidad heterogénea de documentos oficiales en los portales públicos peruanos genera situaciones en las que una obra puede carecer de acta de recepción digitalizada, de resoluciones individuales de aprobación de adicionales, o de otros documentos requeridos para la extracción de variables. Se adoptó un tratamiento diferenciado en función del tipo de dato faltante y de su impacto en el modelo.

Para las variables dependientes globales (BA y BD), la ausencia total de documentación operativa determinó la exclusión de la obra bajo la condición de "No procesable". En cambio, para las variables auxiliares descriptivas asociadas a modificaciones contractuales (BB y BE), se admitió la presencia de datos faltantes sin descartar la obra de la muestra, debido a que estos campos son complementarios.

Se descartó la imputación estadística (media, mediana o modelos de regresión) sobre las variables dependientes para evitar sesgos algorítmicos y distorsiones en las métricas de evaluación, impidiendo que los modelos predictivos aprendan a partir de valores sintéticos. Como consecuencia, el único proyecto con prestación adicional cuyo monto neto no se encuentra documentado se mantiene como valor faltante en la variable BB, analizándose esta sobre un subconjunto de n-1 observaciones respecto al total de la muestra.

Respecto a las variables predictoras, se identificaron 11 proyectos sin Estudio de Mecánica de Suelos (EMS) digitalizado en el expediente original; estas obras fueron categorizadas con nivel de evidencia documental B y sus parámetros se completaron mediante imputación espacial por proximidad geográfica o cartografía geológica

regional de INGEMMET. Los 87 proyectos restantes disponen de un EMS de laboratorio directo, catalogados con nivel de evidencia A.

b. Piso lógico de las variables objetivo

Las variables BA (sobrecosto) y BD (retraso) tienen un piso lógico en cero: una obra que costó o duró menos de lo contratado, pero cuyo avance físico no alcanzó el 100 %, constituye un falso éxito y no un ahorro genuino. La verificación confirma 11 obras en el piso 0 para BA y 5 obras para BD.

c. Valores atípicos y winsorización

Los valores atípicos detectados corresponden a tres fuentes: errores manifiestos de registro en los portales (montos con dígitos duplicados, plazos negativos, fechas invertidas); obras que combinan modificaciones excepcionales concentradas en un periodo corto; y obras cuya magnitud de desvío, aun siendo empíricamente real, distorsiona la escala. Los errores del primer tipo se corrigieron por consulta cruzada con la fuente primaria (acta o liquidación). Los del segundo se conservaron por representar comportamientos válidos del fenómeno. Para los del tercero se aplica winsorización, que sustituye los valores más allá de un umbral por el propio umbral, acotando el apalancamiento de los extremos sin descartar observaciones (Andrić et al., 2024).

La winsorización se aplica de forma local por pliegue, el umbral se calcula únicamente con las observaciones de entrenamiento dentro de cada partición de la validación cruzada, y no con la muestra completa. Esta decisión evita la fuga de información (*data leakage*) dado que estimar los límites de corte de manera global incorporaría propiedades del conjunto de prueba y generaría métricas de desempeño artificialmente optimistas (Krstajic et al., 2014; Cawley & Talbot, 2010). Se fija el percentil 95 superior para BD y BE, cuyas colas derechas concentran la señal de retraso,

y el percentil 5 inferior para BC, cuya cola relevante es la de avances anormalmente bajos. En todos los casos, la evaluación y el cálculo de métricas de error de los modelos se efectúan sobre los valores empíricos reales y no sobre las series winsorizadas.

d. Reclasificación operativa y auditoría de modificaciones contractuales

Este paso constituye la contribución metodológica más específica de la investigación al procesamiento de *datasets* sobre obra pública peruana. La normativa contractual bajo la Ley N.° 30225 y su antecedente en el Decreto Supremo N.° 350-2015-EF reconoce mecanismos legalmente distintos para las modificaciones al plazo (ampliaciones y suspensiones / paralizaciones). La diferencia operativa establece que las ampliaciones extienden el plazo con obra en ejecución, mientras las paralizaciones lo congelan sin ejecución. Sin embargo, la práctica administrativa durante la emergencia sanitaria generó, bajo la Directiva N.° 005-2020-OSCE/CD, la figura de la "Ampliación Excepcional de Plazo", que permitió reconocer bajo esa denominación tanto días de paralización efectiva como días de ampliación real. El resultado documental es que numerosas resoluciones presentan bajo la etiqueta de "ampliación" bloques de días que operacionalmente fueron paralizaciones.

La captura mecánica de estas modificaciones bajo su etiqueta administrativa produce un *dataset* contaminado, en el que la variable de ampliaciones agrega días de naturaleza heterogénea y la de paralizaciones subestima su magnitud. Para modelos que buscan aislar la señal endógena atribuible a la calidad del expediente técnico, esta contaminación introduce ruido no interpretable y compromete la validez del *ranking* de importancia de variables.

El protocolo de depuración adoptado establece la reclasificación de las modificaciones temporales según la actividad constructiva real en campo, prescindiendo de la denominación formal consignada en el acto resolutivo: cada

periodo se computa como paralización si las labores estuvieron totalmente detenidas o como ampliación real si los trabajos continuaron en ejecución, incluso de forma parcial. Bajo este criterio, en resoluciones con causales mixtas se desglosa y clasifica cada componente de manera independiente, asignando términos como "cese de actividades", "cuarentena obligatoria" o "adecuación de ambientes con obra cerrada" al rubro de paralización total, mientras que conceptos como "afectación al rendimiento", "ejecución con cuadrillas reducidas" o "menor productividad por protocolos" se registran como ampliaciones de plazo por causal exógena.

Complementariamente, la clasificación endógeno-exógeno para las variables dependientes aplica una cascada de tres niveles de evidencia. El Nivel A se aplica cuando la resolución individual está disponible y declara la causal explícitamente; el Nivel B, cuando la clasificación se realiza por indicios convergentes en el acta o la liquidación; y el Nivel C, cuando solo se dispone del dato agregado, en cuyo caso se aplica presunción endógena por defecto. Esta presunción se sustenta en el diagnóstico de la Contraloría General de la República (2025), según el cual las causas dominantes de modificaciones en obra pública peruana son endógenas al sistema contractual. El nivel de evidencia alcanzado por la obra queda documentado como columna auxiliar y se emplea en el análisis de robustez del EDA.

Un ejemplo empírico ilustra la importancia del procedimiento. En una obra de la Universidad Nacional de San Agustín con plazo inicial de 360 días, el acta registró una "Ampliación Excepcional COVID" de 141 días, dividida en tres subcomponentes: 81 días de paralización efectiva (marzo a junio de 2020), 15 días de adecuación de ambientes con obra cerrada y 45 días por afectación al saldo bajo protocolos sanitarios. Una lectura mecánica habría clasificado los 141 días como ampliación exógena; el protocolo reclasificó los 96 días de los dos primeros subcomponentes como

paralización total y conservó los 45 días del tercero como ampliación exógena real. Sumadas dos ampliaciones ordinarias (20 y 43 días) y 73 días de paralización confirmados como endógenos, el resultado final fue de 63 días de ampliación endógena, 45 de ampliación exógena y 169 de paralización total, cifra consistente con el plazo real (621 días) y con la fecha de plazo ampliado del acta.

4.4. Análisis Exploratorio de Datos (EDA)

4.4.1. *Datos y Alcance del Análisis*

El análisis exploratorio de datos (EDA) constituye la etapa diagnóstica que articula la preparación del *dataset* con el modelado predictivo. Su finalidad es caracterizar la estructura de las variables predictoras y objetivo mediante estadística descriptiva, correlación y pruebas no paramétricas, además de fundamentar con evidencia cuantitativa las decisiones de transformación y selección. Coherente con su naturaleza exploratoria y con el tamaño muestral, los resultados se interpretan como indicios orientadores del preprocesamiento y generadores de hipótesis, no como pruebas confirmatorias. Todas las asociaciones se describen en términos de correlación, nunca de causalidad.

De los 112 proyectos del registro consolidado se retiraron 14 obras debido a los siguientes criterios de depuración: ausencia de documentación técnica esencial para el cálculo de las variables dependientes, tipología no representativa de infraestructura civil convencional, resoluciones contractuales con ejecución física inconclusa, financiamiento mediante convenios de endeudamiento externo regidos por normativas ajenas a la Ley de Contrataciones del Estado, y la ausencia de información competente en los archivos oficiales de la obra. La muestra analítica resultante es de n = 98 obras. Además, una obra adicional (Código Infobras 131548) registra adicionales cuyo monto no consta en la fuente documental. En consecuencia, la variable BB se analiza sobre n = 97 observaciones y el resto del análisis sobre n = 98.

4.4.2. *Estadística Descriptiva Univariada*

La caracterización univariada resume tendencia central, dispersión, asimetría y curtosis en exceso (normal = 0). La asimetría y la curtosis anticipan la necesidad de transformaciones. La Tabla 9 reporta los descriptivos de las variables numéricas y de las cinco dependientes; la Tabla 10, la distribución de frecuencias de las variables categóricas.

Tabla 9

Estadística Descriptiva Univariada

Variable	n	Media	Mediana	DE	Mín	Máx	Asimetría	Curtosis
Monto adjudicado (S/)	98	7 717 254.8	6 721 118.5	3 452 730.2	4 003 106.0	28 460 193.0	2.70	12.70
Plazo inicial (días)	98	206.9	180.0	61.8	105.0	450.0	1.27	2.03
N.º de postores	98	3.99	2.00	4.23	1.00	19.00	2.16	4.24
Tasa de pobreza	98	0.093	0.083	0.076	0.007	0.522	2.66	11.81
Competitividad	98	0.076	0.100	0.042	0.000	0.131	−1.17	−0.52
BA (sobrecosto)	98	0.129	0.117	0.095	0.000	0.441	0.73	0.73
BB (incid. económica)	97	0.070	0.066	0.048	0.000	0.154	0.20	−1.04
BC (avance físico)	98	0.934	0.960	0.072	0.648	1.000	−1.55	2.64
BD (retraso)	98	0.917	0.723	0.853	0.000	4.467	2.08	5.80
BE (incid. temporal)	98	0.335	0.274	0.279	0.000	1.387	1.24	1.73

Tabla 10

Distribución de Frecuencias de las Variables Categóricas

Variable	Categoría	Frecuencia	%
Naturaleza de intervención	Intervención preexistente	86	87.8
	Obra nueva	12	12.2
Tipología de obra	Infraestructura urbana y espacios públicos	37	37.8
	Infraestructura pública	26	26.5
	Infraestructura vial y transporte	23	23.5
	Saneamiento e hidráulica	12	12.2
Sistema de contratación	Precios unitarios	69	70.4
	Suma alzada	29	29.6
Modalidad de selección	Licitación pública	80	81.6
	Adj. simplificada y proc. especiales	18	18.4
Nivel de gobierno	Gobierno local	81	82.7

	Gobierno nacional	9	9.2
	Gobierno regional	8	8.2
Escenario COVID	Post-COVID	43	43.9
	Pandemia	37	37.8
	Pre-COVID	18	18.4
Ubicación metropolitana	Metropolitano	74	75.5
	No metropolitano	24	24.5
Familia SUCS	Arena	79	80.6
	Grava	19	19.4
Calidad de fundación	Media	50	51.0
	Baja	28	28.6
	Alta	20	20.4
Tipo de contratista	Consorcio	56	57.1
	Individual	42	42.9
Material inadecuado superficial	Presente	53	54.1
	Ausente	45	45.9

Salvo el % de Avance Físico, acotado entre 0 y 100 %, las demás variables objetivo son proporciones o razones; en particular, la Vulnerabilidad del Plazo Contractual Total (BD) es un porcentaje que puede superar el 100 % en obras con retrasos muy marcados. La visualización de las cinco variables dependientes confirma lo que anticipan la asimetría y la curtosis (Figura 2 a la 6). El sobrecosto (BA) y la incidencia económica (BB) presentan distribuciones de forma moderada, mientras que el retraso (BD) y la incidencia temporal (BE) muestran colas largas hacia la derecha, con un caso extremo de BD cercano al 447 % del plazo; el avance físico (BC), en cambio, exhibe una cola hacia la izquierda (asimetría negativa), con la masa concentrada cerca del 100 % y unas pocas obras de avance bajo. Este diagnóstico sugiere acotar las colas de BD, BE y BC mediante winsorización antes del modelado, para que los valores extremos no dominen el ajuste.

Figura 2

Distribución de BA (Sobrecosto): Histograma y Diagrama de Caja

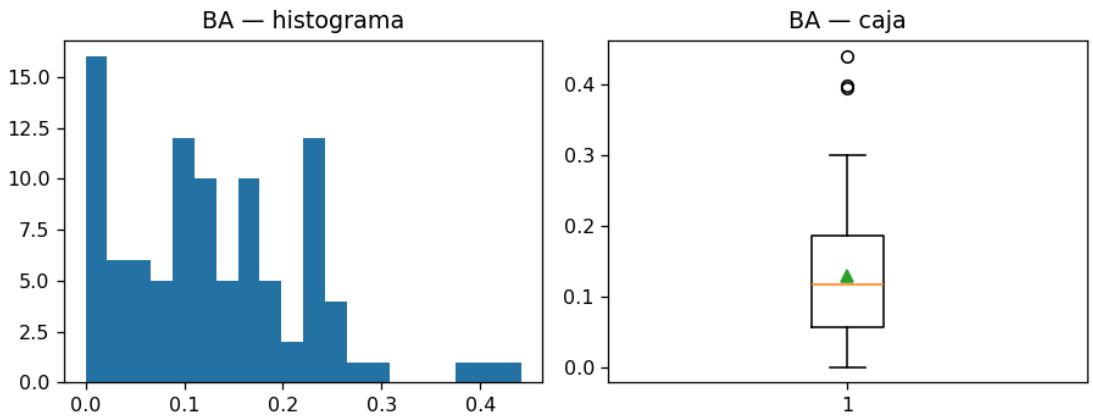


Figura 3

Distribución de BB (Incidencia Económica): Histograma y Diagrama de Caja

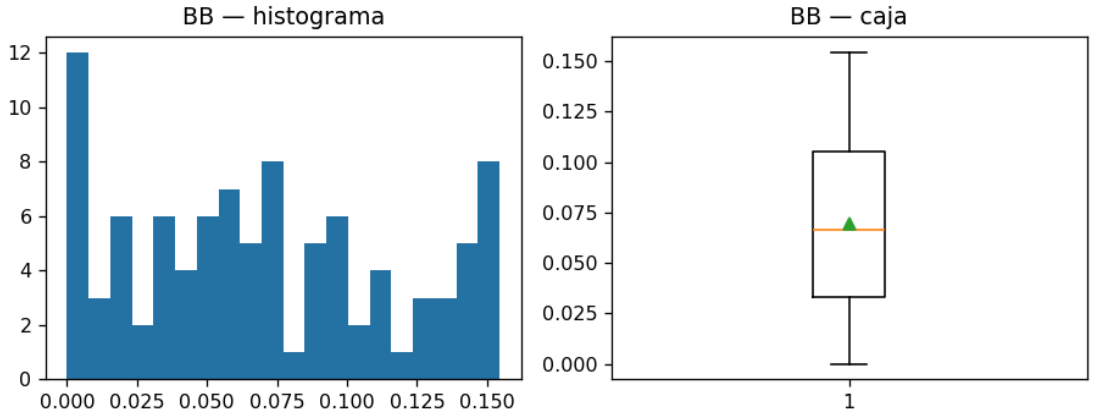


Figura 4

Distribución de BC (Avance Físico): Histograma y Diagrama de Caja

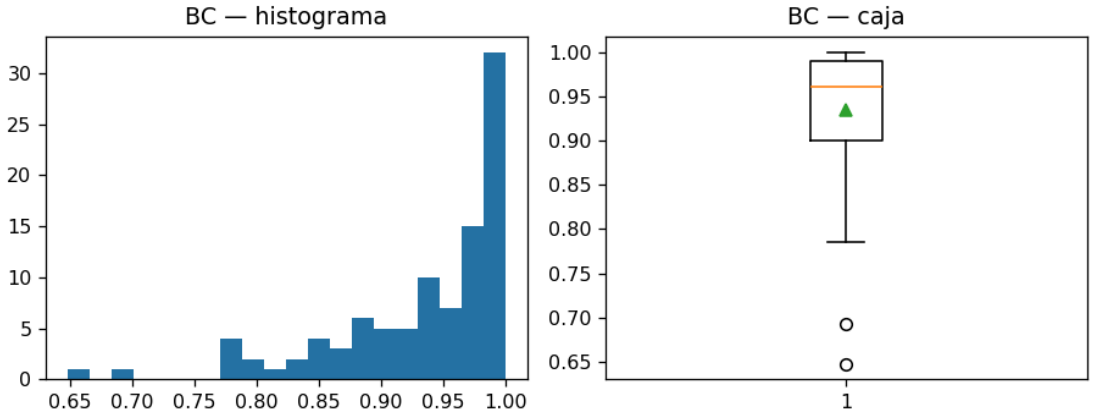


Figura 5

Distribución de BD (Retraso): Histograma y Diagrama de Caja

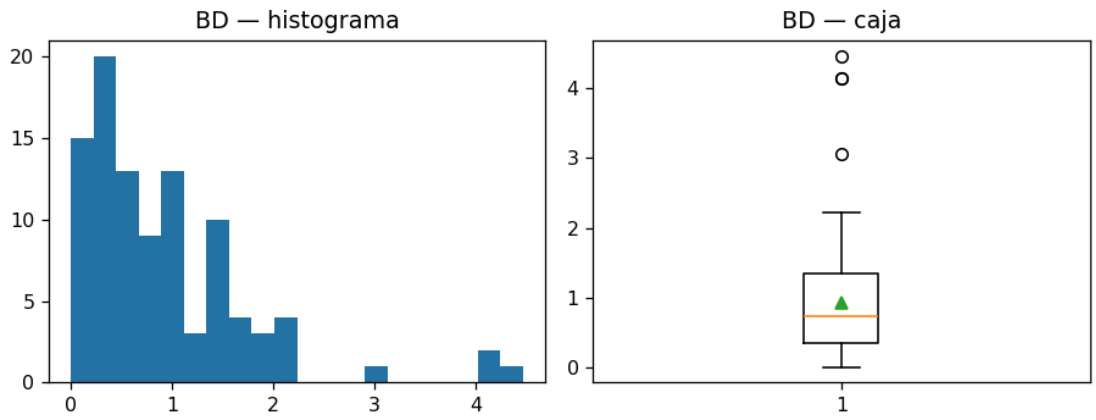
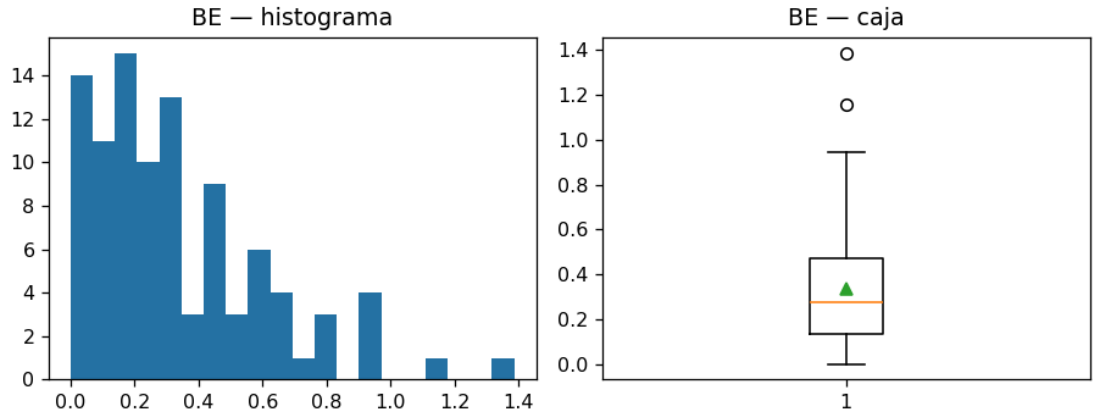


Figura 6

Distribución de BE (Incidencia Temporal): Histograma y Diagrama de Caja



4.4.3. *Tratamiento de Valores Extremos: Winsorización*

La winsorización se aplica únicamente a las variables dependientes de cola larga (BD y BE, con recorte superior; BC, con recorte inferior) y no a las variables independientes. El objetivo del recorte es evitar que unos pocos valores extremos de la variable a predecir dominen el ajuste y la métrica de error; ese riesgo se concentra en las dependientes, que son las que el modelo optimiza. Adicionalmente, las variables independientes numéricas se tratan por otras vías más adecuadas a su rol, como el Monto, muy asimétrico, que se propone estabilizar con logaritmo.

Esta subsección documenta, con fines ilustrativos, el efecto de la winsorización calculada sobre la muestra completa; en el modelado el umbral se recalcula por pliegue. La Tabla 11 muestra el efecto del recorte, contrastando además una cola frente a dos colas para justificar la elección de recorte de una sola cola.

Tabla 11

Efecto de la Winsorización

Variable	Winsorización	Media	DE	Asimetría	Curtosis
BD (retraso)	Original	0.917	0.848	2.077	5.796
	Una cola (p95 superior)	0.841	0.619	0.612	−0.618
	Dos colas (p2.5–p97.5)	0.899	0.782	1.646	3.456
BE (incid. temporal)	Original	0.335	0.278	1.239	1.729
	Una cola (p95 superior)	0.327	0.254	0.801	−0.138
	Dos colas (p2.5–p97.5)	0.329	0.258	0.850	0.000
BC (avance)	Original	0.934	0.072	−1.549	2.637
	Una cola (p5 inferior)	0.937	0.064	−1.020	0.003
	Dos colas (p2.5–p97.5)	0.937	0.065	−1.027	0.023

El recorte de una sola cola reduce de manera apreciable la asimetría y la curtosis de las variables de cola larga sin eliminar observaciones y sin tocar el extremo opuesto de la distribución, que es información válida. En BD la asimetría cae de 2.08 a 0.61 y la curtosis de 5.80 a −0.62 una vez acotado el extremo de 4.47 (447 % de retraso), correspondiente a una obra con paralización prolongada que se conserva acotada, no se descarta; el recorte de dos colas resulta claramente menos efectivo (asimetría 1.65). En BE la asimetría desciende de 1.24 a 0.80, y en BC el recorte inferior atenúa la cola de avances anormalmente bajos (asimetría de −1.55 a −1.02). Se adopta, por tanto, la winsorización de una cola en el sentido de la cola larga de cada variable.

4.4.4. *Colinealidad y Correlación Bivariada*

a. Correlación Pearson y Spearman

Se calcularon los coeficientes de Pearson (asociación lineal) y de Spearman (asociación monotónica sobre rangos, capaz de captar relaciones no lineales) entre las predictoras numéricas y las cinco variables objetivo. Divergencias marcadas entre ambos coeficientes revelarían no linealidad o influencia de atípicos. Las correlaciones se presentan en las Tablas 12 a 13.

Tabla 12

Correlación de Pearson (Predictoras Numéricas × Dependientes)

Predictora	BA	BB	BC	BD	BE
Monto adjudicado (S/)	0.122	−0.044	0.072	−0.131	0.063
Plazo inicial (días)	0.374	0.002	0.168	−0.151	−0.076
N.º de postores	−0.211	−0.177	0.016	−0.048	−0.131
Tasa de pobreza	0.021	−0.230	0.136	0.095	−0.125
Competitividad	0.118	0.039	−0.062	0.027	0.035

Tabla 13

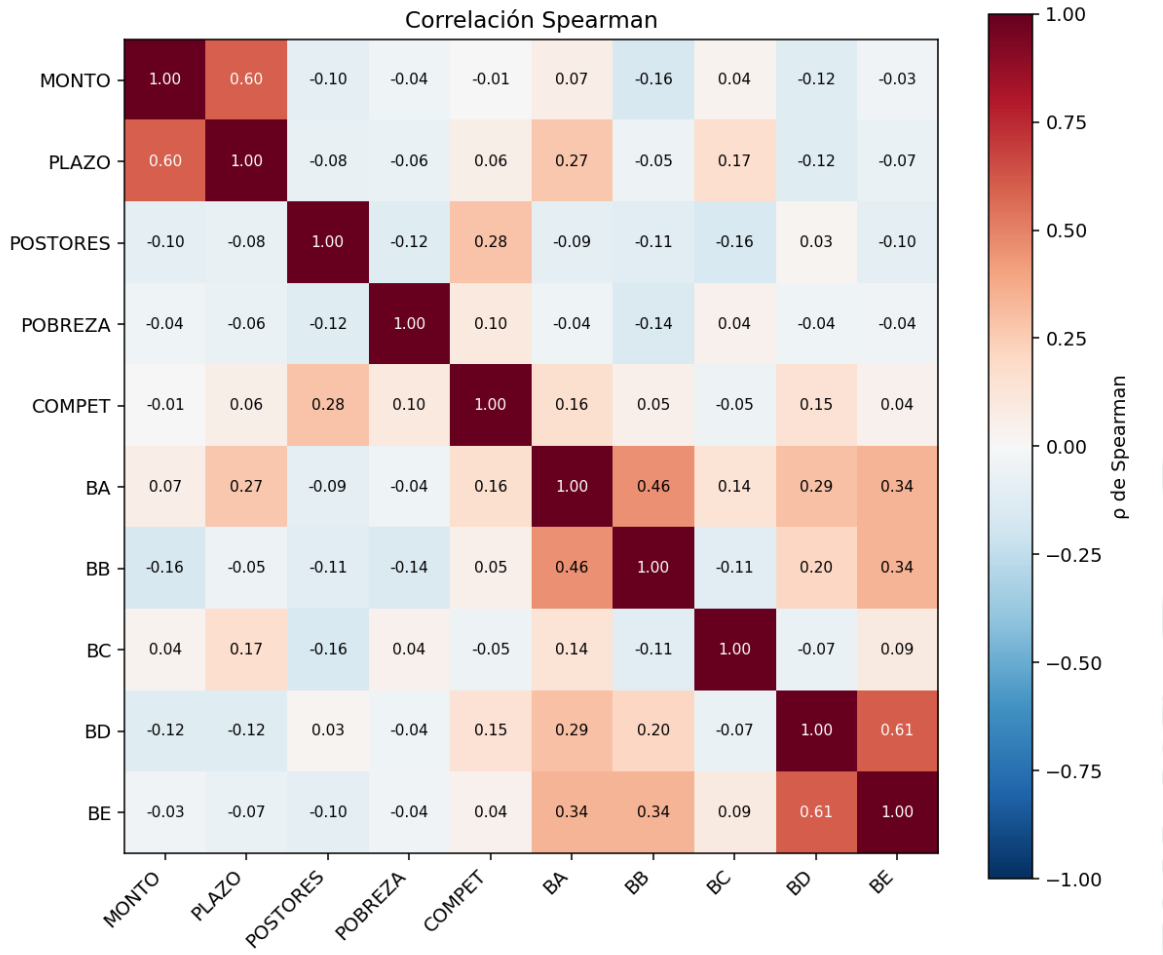
Correlación de Spearman (Predictoras Numéricas × Dependientes)

Predictora	BA	BB	BC	BD	BE
Monto adjudicado (S/)	0.070	−0.163	0.035	−0.123	−0.032
Plazo inicial (días)	0.267	−0.048	0.170	−0.120	−0.066
N.º de postores	−0.090	−0.112	−0.155	0.028	−0.098
Tasa de pobreza	−0.040	−0.144	0.042	−0.039	−0.036
Competitividad	0.163	0.048	−0.047	0.147	0.045

En la Figura 7 se presenta el mapa de calor de Spearman, al operar sobre rangos reconoce asociaciones monótonas no lineales que el coeficiente de Pearson puede subestimar. La matriz se presenta de forma completa, de modo que puedan leerse todas las relaciones entre pares, incluidas las que se dan entre las propias variables objetivo.

Figura 7

Matriz de Correlación de Spearman (Predictoras Continuas y Dependientes)



Las asociaciones lineales y monótonas son en conjunto débiles, congruente con la naturaleza multicausal del fenómeno. La correlación máxima entre una predictora numérica y una variable objetivo es $|r| \approx 0.37$ (Pearson, Plazo inicial × BA) y $|\rho| \approx 0.27$ en Spearman (Plazo inicial × BA).

b. Diagnóstico de multicolinealidad (VIF)

El factor de inflación de la varianza (VIF) se calcula sobre las predictoras numéricas en su escala original (sin transformar), incluyendo el bloque continuo completo del dataset. Los resultados se muestran en la Tabla 14.

Tabla 14

Factor de Inflación de la Varianza (VIF)

Variable	VIF
Monto adjudicado (S/)	1.598
Plazo inicial (días)	1.586
N.º de postores	1.067
Tasa de pobreza	1.016
Competitividad	1.085

Todos los VIF se sitúan muy por debajo del umbral de atención de 5, por lo que no existe multicolinealidad relevante entre las predictoras continuas y no se justifica excluir ninguna por este criterio. La ausencia de multicolinealidad respalda, además, el uso de la penalización L1 (LASSO) frente a L2 (Ridge) en el modelado posterior.

4.4.5. *Contraste de Variables Dependientes según Variables Categóricas*

La prueba de Kruskal-Wallis es la alternativa no paramétrica al Análisis de Varianza (ANOVA), esta compara si la distribución de una variable dependiente difiere entre los niveles de una categórica ordenando todas las observaciones por rango y contrastando los rangos medios de cada grupo; es apropiada aquí porque las dependientes son sesgadas. Su tamaño de efecto se resume con Eta² a partir del estadístico H. Cuando la categórica tiene solo dos niveles, la prueba equivalente es la prueba U de Mann-Whitney, que contrasta los rangos de los dos grupos. Finalmente, cuando Kruskal-Wallis resulta significativa en una categórica de tres o más niveles, la prueba *post-hoc* de Dunn con ajuste de Bonferroni identifica qué pares de niveles concretos son responsables de la diferencia, corrigiendo por el número de comparaciones. Primero, para cada variable categórica se grafican los diagramas de caja de las cinco dependientes con el valor p de Kruskal-Wallis en cada panel; los paneles significativos ($p < 0.05$) se marcan con asterisco, estos se presentan en las Figuras 8 a 17.

Figura 8

*Dependientes según Naturaleza de Intervención (Kruskal-Wallis; * = p<0.05)*

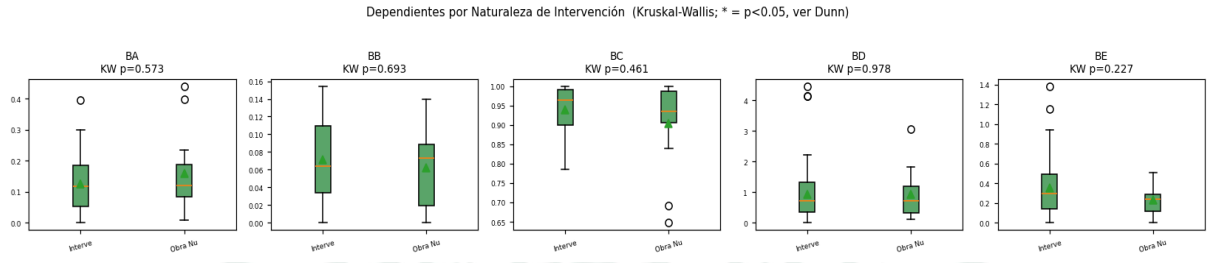


Figura 9

*Dependientes según Tipología de Obra (Kruskal-Wallis; * = p<0.05)*

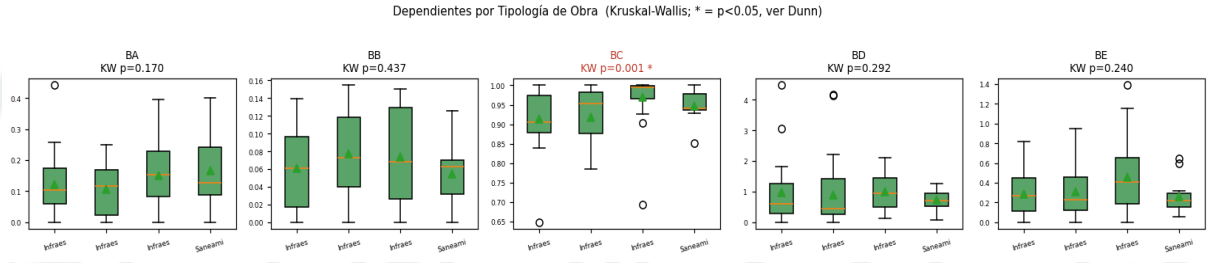


Figura 10

*Dependientes según Sistema de Contratación (Kruskal-Wallis; * = p<0.05)*

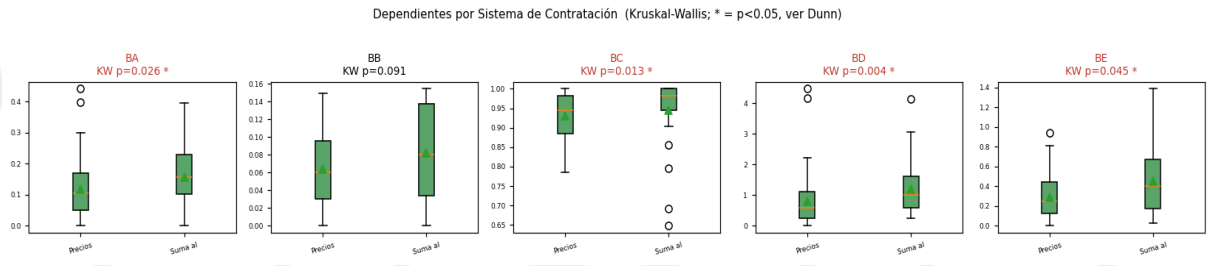


Figura 11

*Dependientes según Modalidad de Selección (Kruskal-Wallis; * = p<0.05)*

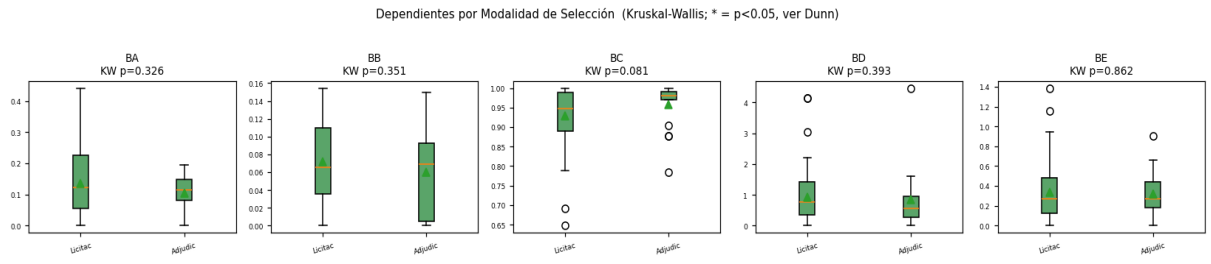


Figura 12

*Dependientes según Nivel de Gobierno (Kruskal-Wallis; * = p<0.05)*

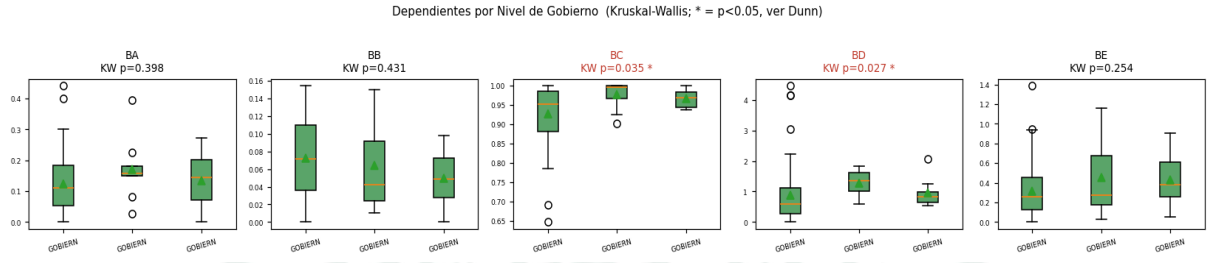


Figura 13

*Dependientes según Escenario COVID (Kruskal-Wallis; * = p<0.05)*

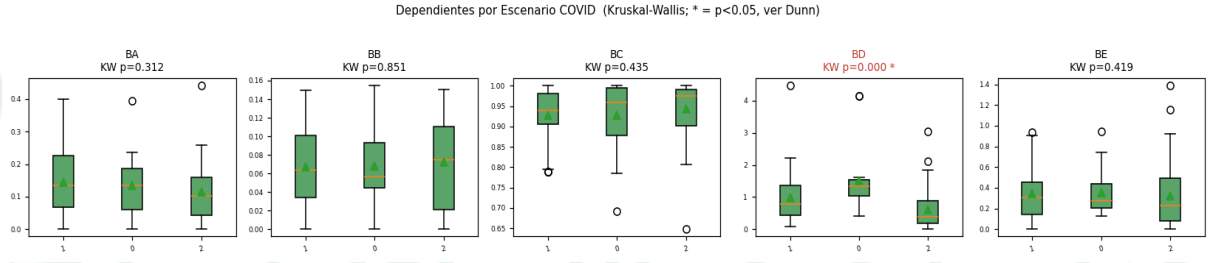


Figura 14

*Dependientes según Ubicación Metropolitana (Kruskal-Wallis; * = p<0.05)*

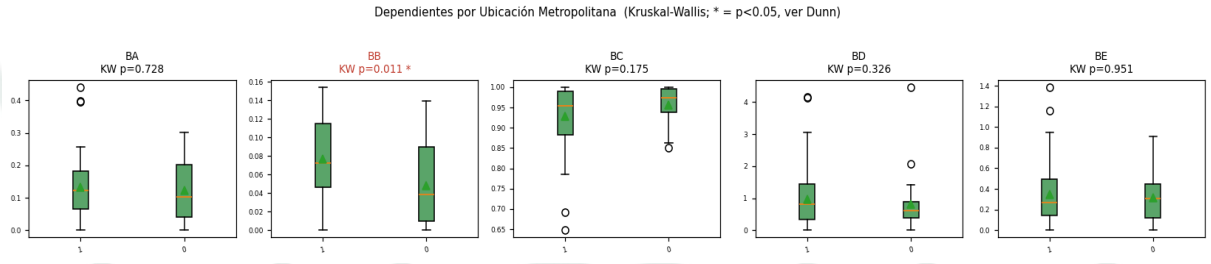


Figura 15

*Dependientes según Familia SUCS (Kruskal-Wallis; * = p<0.05)*

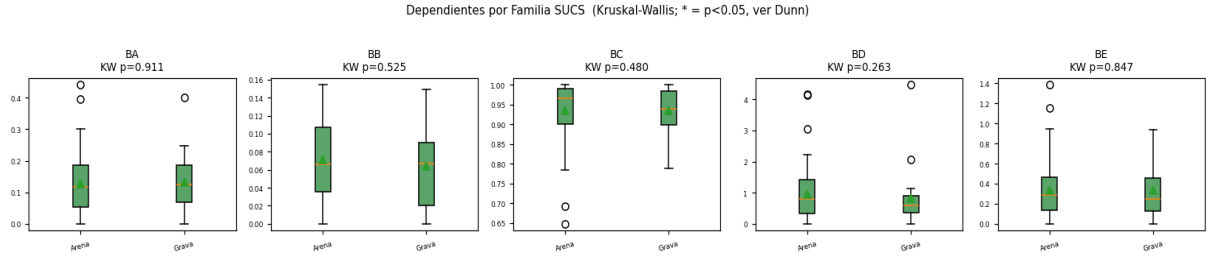


Figura 16

*Dependientes según Calidad de Fundación (Kruskal-Wallis; * = p<0.05)*

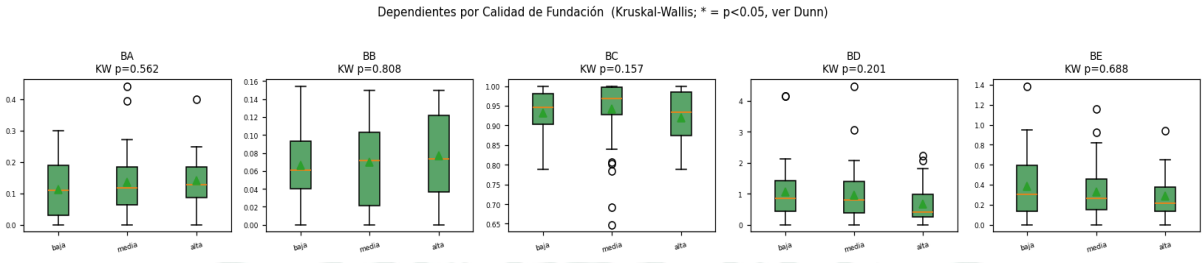
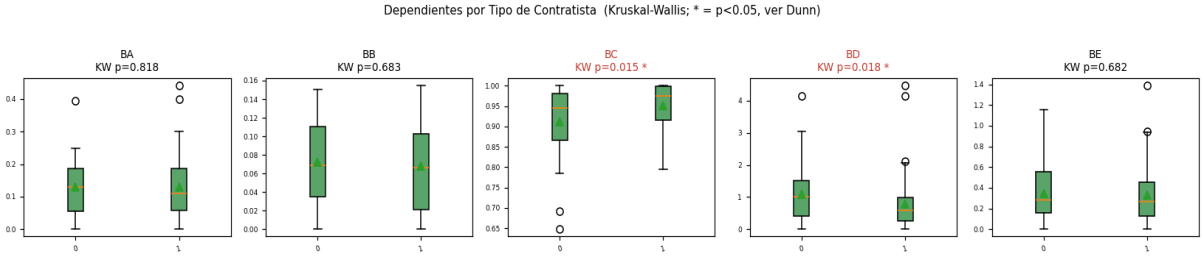


Figura 17

*Dependientes según Tipo de Contratista (Kruskal-Wallis; * = p<0.05)*



Después, se consolidan en tablas los pares con significación global de Kruskal-Wallis para las categóricas de tres o más niveles en la Tabla 15, y la prueba U de Mann-Whitney para las binarias en la Tabla 16.

Tabla 15

Pruebas de Kruskal-Wallis con Significación Global (P < 0.05)

Predictor	Objetivo	H	p	Eta²	k
Escenario COVID	BD	20.550	0.0000	0.195	3
Tipología	BC	17.663	0.0005	0.156	4
Nivel de gobierno	BD	7.192	0.0274	0.055	3
Nivel de gobierno	BC	6.707	0.0350	0.050	3

Tabla 16

Pruebas U de Mann-Whitney con Significación Global (P < 0.05)

Predictor	Objetivo	U	z	p
Sistema de contratación	BD	1369.0	2.868	0.0041
Ubicación metropolitana	BB	572.5	−2.539	0.0111
Sistema de contratación	BC	1321.0	2.497	0.0125
Tipo de contratista	BC	1515.0	2.436	0.0148
Tipo de contratista	BD	847.0	−2.362	0.0182
Sistema de contratación	BA	1285.5	2.219	0.0265
Sistema de contratación	BE	1257.5	2.001	0.0454

Sobre los pares significativos según Kruskal-Wallis de tres o más niveles se aplica la prueba de Dunn, se presentan resultados en la Tabla 17.

Tabla 17

Prueba Post-hoc de Dunn con Ajuste de Bonferroni

Predictor	Objetivo	Par de niveles	z	p (Bonferroni)	Sign.
Tipología	BC	Infraestructura pública vs. Urbana	3.759	0.0010	Sí
Tipología	BC	Infraestructura pública vs. Vial	3.557	0.0022	Sí
Tipología	BC	Infraestructura pública vs. Saneamiento	2.196	0.1684	No
Tipología	BC	Urbana vs. Vial	0.212	1.0000	No
Tipología	BC	Urbana vs. Saneamiento	−0.589	1.0000	No
Tipología	BC	Vial vs. Saneamiento	−0.707	1.0000	No
Nivel de gobierno	BD	Local vs. Nacional	−2.546	0.0327	Sí
Nivel de gobierno	BD	Local vs. Regional	−1.081	0.8391	No
Nivel de gobierno	BD	Nacional vs. Regional	1.016	0.9284	No
Nivel de gobierno	BC	Local vs. Nacional	−2.430	0.0453	Sí
Nivel de gobierno	BC	Local vs. Regional	−1.123	0.7848	No
Nivel de gobierno	BC	Nacional vs. Regional	0.901	1.0000	No

Escenario COVID	BD	Pre-COVID vs. Post-COVID	4.432	0.0000	Sí
Escenario COVID	BD	Pandemia vs. Post-COVID	2.523	0.0349	Sí
Escenario COVID	BD	Pre-COVID vs. Pandemia	2.361	0.0547	No

El efecto más fuerte es el del Escenario COVID sobre el retraso (BD): el *post-hoc* de Dunn confirma que el régimen post-COVID difiere significativamente del pre-COVID ($p < 0.001$) y del de pandemia ($p = 0.035$), con el retraso decreciendo hacia el post-COVID, coherente con las secuelas de paralizaciones y ampliaciones del periodo; el contraste pre-COVID frente a pandemia queda en el límite y no alcanza significancia ($p = 0.055$). El segundo efecto claro es el de la Tipología sobre el avance físico (BC): la categoría "Infraestructura pública" difiere de la urbana ($p = 0.001$) y de la vial ($p = 0.002$) y concentra los avances más bajos respecto del alcance original, mientras que el resto de pares no resulta significativo, de modo que el efecto se debe a que esa tipología se aparta de las demás y no a diferencias entre las otras. El Nivel de gobierno muestra diferencias moderadas sobre BC y BD, localizadas en el contraste Local vs. Nacional ($p = 0.045$ y $p = 0.033$); los pares que involucran al nivel Regional no son significativos, lo que es consistente con su reducido tamaño muestral (8 obras) y con el ajuste conservador de Bonferroni. Entre las binarias, el Sistema de contratación es la variable transversal, se asocia significativamente con BD, BC, BA y BE, indicio de que el esquema contractual (precios unitarios vs. suma alzada) discrimina el desempeño; el Tipo de contratista se asocia con BC y BD, y la condición metropolitana con BB.

4.4.6. *Información Mutua*

Para identificar dependencias no lineales que los coeficientes de Pearson y Spearman no logran capturar, se estimó la información mutua entre cada predictora y las cinco variables dependientes. Basada en la teoría de la información de Shannon, la MI mide la reducción de incertidumbre de la variable objetivo al conocer el predictor sin asumir una forma funcional previa, lo que permite detectar patrones complejos o no monótonos donde las correlaciones

clásicas suelen dar falsos valores cercanos a cero. Los resultados principales se detallan en la Tabla 18 y el ordenamiento jerárquico de cada variable objetivo se muestra en las Figuras 18 a 22.

Tabla 18

Información Mutua

Objetivo	Predictoras con mayor información mutua (valor)
BA (sobrecosto)	Plazo inicial (0.173); Tasa de pobreza (0.126); Competitividad (0.113); N.º de postores (0.093)
BB (incid. económica)	Competitividad (0.153); N.º de postores (0.088); Tasa de pobreza (0.078); Monto adjudicado (0.052)
BC (avance)	Tipología (0.174); Plazo inicial (0.173); Nivel de gobierno (0.137); N.º de postores (0.087)
BD (retraso)	Escenario COVID (0.176); Monto adjudicado (0.147); Plazo inicial (0.117); Competitividad (0.090)
BE (incid. temporal)	Tasa de pobreza (0.099); Tipología (0.090); Plazo inicial (0.081); N.º de postores (0.077)

Figura 18

Información Mutua con el Sobrecosto (BA)

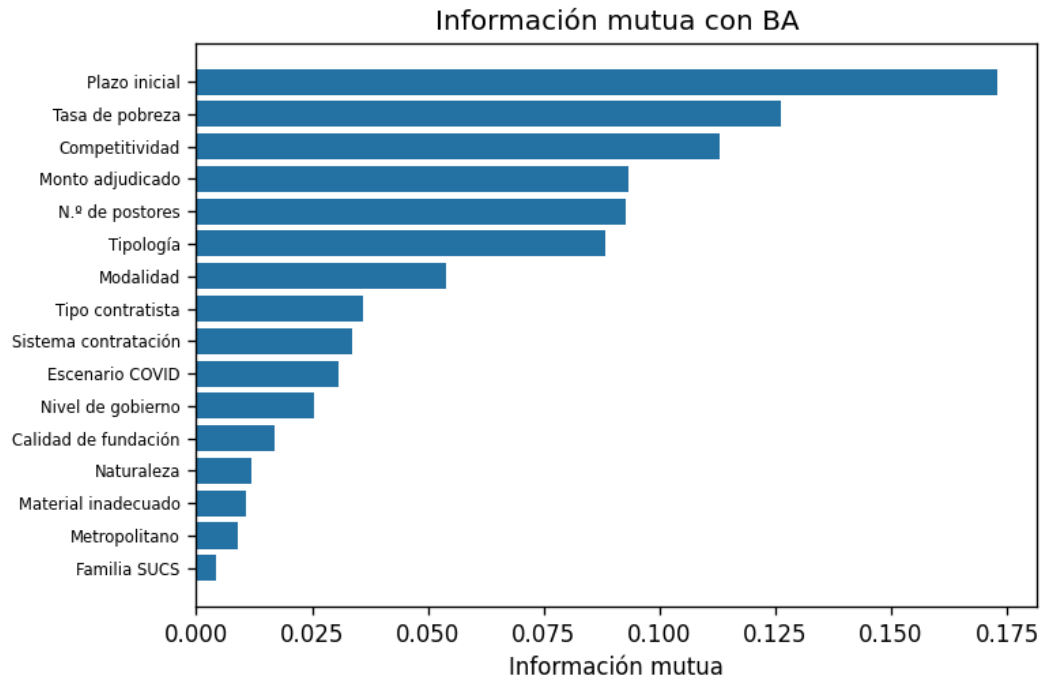


Figura 19

Información Mutua con la Incidencia Económica (BB).

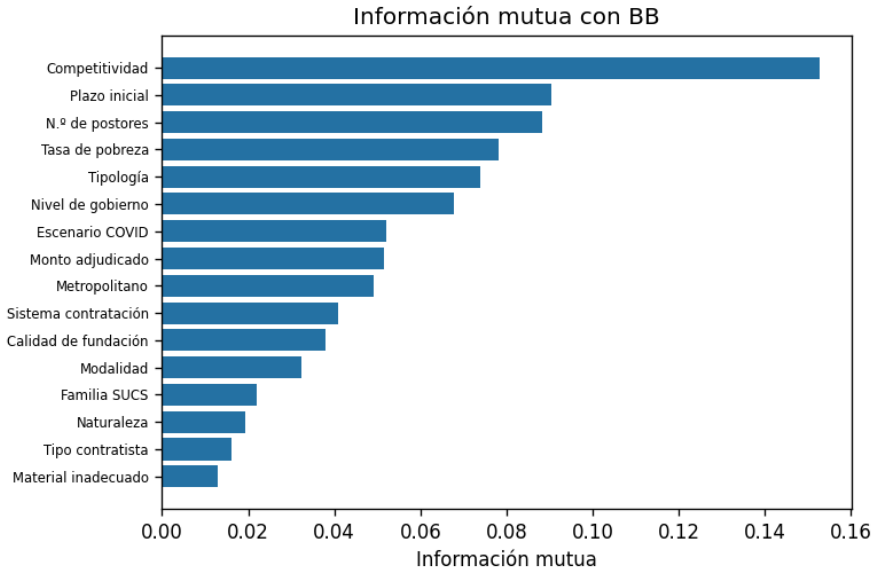


Figura 20

Información Mutua con el Avance Físico (BC).

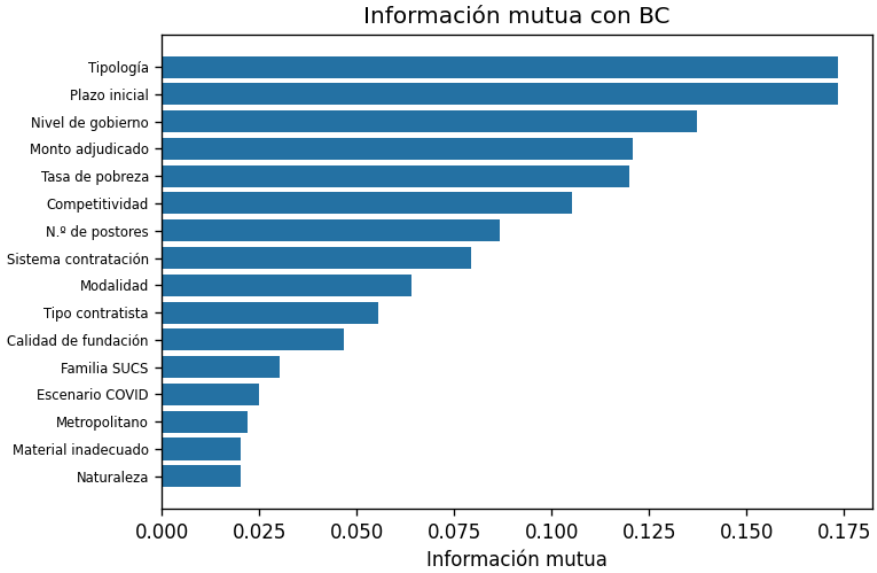


Figura 21

Información Mutua con el Retraso (BD)

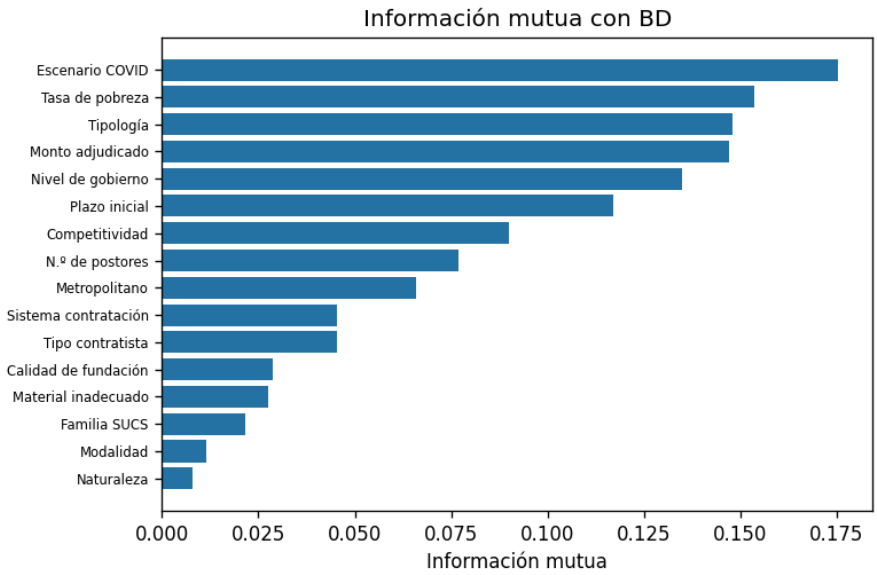
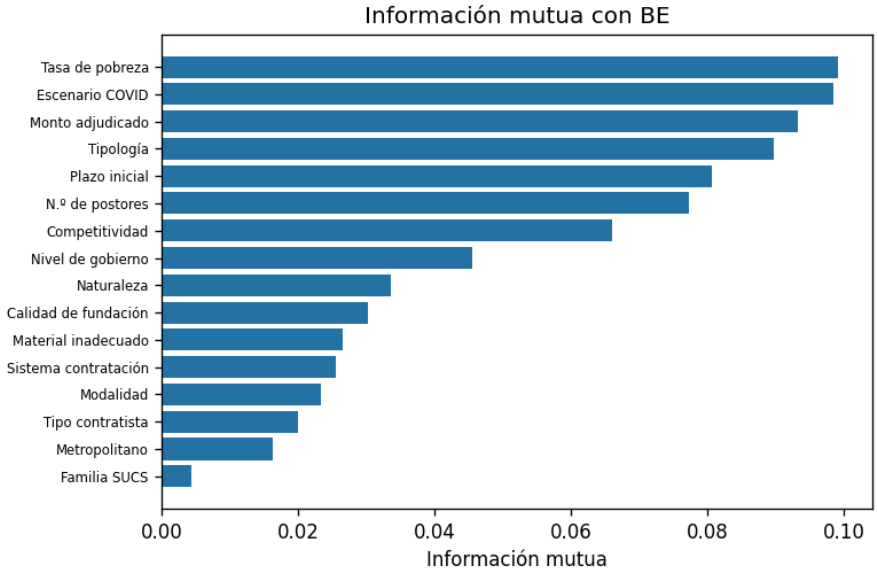


Figura 22

Información Mutua con la Incidencia Temporal (BE)



La información mutua revela que variables sin correlación lineal aparente, como el Número de Postores y el Nivel de Gobierno sobre el retraso, o la Tipología sobre el avance, sí aportan dependencia no lineal, lo que justifica preservarlas para los modelos basados en árboles

aun cuando su correlación de Pearson sea baja. Esta constatación es la base del criterio de descarte de la ingeniería de características, no se elimina una predictora por baja correlación lineal.

4.4.7. *Robustez Frente al Nivel de Evidencia Documental*

Las variables endógenas (BB y BE) se construyen a partir de la clasificación de adicionales y ampliaciones, apoyada en la evidencia documental disponible por obra (niveles A, B y C). Para confirmar que esta construcción es estable y no queda condicionada por el detalle documental, se contrastó la distribución de cada endógena a lo largo de los tres niveles mediante Kruskal-Wallis. Del mismo modo, se contrastó cada objetivo entre las obras con EMS real (A) y las imputadas (B) mediante la prueba U de Mann-Whitney. Ambos análisis se presentan en Tabla 19.

Tabla 19

Pruebas de Robustez de las Variables Endógenas y EMS

Prueba	**Estadístico**	**p**	**Lectura**
BB × validez del adicional (A/B/C)	H = 2.21	0.331	Sin diferencias: estable
BE × validez de la ampliación (A/B/C)	H = 2.36	0.307	Sin diferencias: estable
BA × EMS (A vs. B)	U = 573	0.996	Sin diferencia
BD × EMS (A vs. B)	U = 421	0.128	Sin diferencia significativa
BC × EMS (A vs. B)	U = 547	0.804	Sin diferencia

Las medianas por nivel resultan próximas y las diferencias no alcanzan significancia, por lo que las variables endógenas se comportan de forma consistente con independencia del nivel de evidencia. Igualmente, las obras con EMS real e imputado no difieren significativamente en ninguna variable objetivo, lo que respalda la validez de la imputación geotécnica. La lectura mantiene la cautela por el reducido subconjunto del nivel C.

4.4.8. *Interpretación y Decisiones en Base al EDA*

a. Interpretación

Existe un techo estructural de predictibilidad. Un patrón atraviesa todo el análisis: la asociación máxima entre cualquier predictora individual y una variable objetivo es baja ($|r| \approx 0.37$ para el sobrecosto con el plazo, y $Eta^2 \approx 0.20$ para el retraso con el escenario COVID). Esto no es una deficiencia del *dataset*, sino una característica documentada del fenómeno, el marco del sesgo optimista sostiene que los sobrecostos y retrasos de la obra pública responden en buena medida a factores conductuales y de estimación que no quedan capturados por las variables técnicas del expediente. En consecuencia, ninguna variable aislada predice con fuerza el desvío, lo que fundamenta las dos decisiones estratégicas del resto de la tesis, priorizar el *ranking* de factores por objetivo antes que la maximización del R^2, y reformular el problema hacia un sistema de clasificación y alerta temprana por niveles de riesgo.

La ausencia de multicolinealidad severa respalda la elección de la regularización L1 (LASSO) frente a L2 (Ridge) para los modelos de regresión lineal y cuadrática. Dado que la señal reside primordialmente en efectos categóricos y no lineales, como el escenario COVID frente al retraso (BD) o la tipología constructiva frente al avance (BC), se justifica el empleo complementario de modelos lineales regularizados y ensambles de árboles de decisión. Por encima de todo, esta heterogeneidad fundamenta la adopción de un enfoque individualizado por objetivo, puesto que los factores dominantes difieren entre dimensiones, el plazo inicial y la competencia condicionan el sobrecosto, el contexto coyuntural rige el retraso y la tipología define el alcance físico. Además, el techo de predictibilidad recomienda evaluar con validación cruzada repetida en lugar de una partición única.

b. Decisiones para Transformación de datos

A partir de las conclusiones anteriores se adoptan las siguientes decisiones, que definen el estado del *dataset* previo a los modelados de *machine learning*:

- Binarización del Nivel de Gobierno a (Local, No local), por su fuerte desbalance (81 / 9 / 8 en las categorías originales).
- Log-transformación del Monto Adjudicado y log1p del Número de Postores, por su fuerte asimetría.
- Winsorización local (percentil 95 superior para BD y BE; percentil 5 inferior para BC), aplicada por pliegue durante el modelado.
- Conservación del bloque geotécnico completo por su aporte de dependencia no lineal (información mutua), pese a su baja correlación lineal.

4.5. Ingeniería de Características

4.5.1. *Construcción de Indicadores Relativos y Transformación Logarítmica*

Las variables dependientes BA, BB, BD y BE se expresan como porcentajes del monto adjudicado o del plazo inicial, en lugar de valores absolutos, lo que permite comparar obras de distinta magnitud bajo una escala común. El Monto Adjudicado, de fuerte asimetría en el rango de S/ 4.5 a S/ 30 millones, se corrige mediante logaritmo natural, procedimiento estándar en el modelado de variables monetarias de cola pesada (Uddin et al., 2022). El número de postores, conteo sesgado con posibles valores mínimos, se transforma mediante *log1p*, seguro ante ceros. Los resultados se resumen en la Tabla 20.

Tabla 20

Efecto de Log-transformación en Asimetría y Curtosis

Variable	Transformación	Asim. antes	Curt. antes	Asim. desp.	Curt. desp.
Monto adjudicado	*log natural*	2.70	12.70	0.83	0.86
N.º de postores	*log1p*	2.16	4.24	0.83	−0.13

4.5.2. *Winsorización Local*

El recorte al percentil 95 superior de BD y BE, y al percentil 5 inferior de BC, reduce de manera apreciable la asimetría y la curtosis sin descartar observaciones. Con fines de trazabilidad, la Tabla 21 reporta los umbrales calculados sobre la muestra completa (referenciales; en el modelado se recalculan por pliegue) y el número aproximado de obras afectadas.

Tabla 21

Umbrales de Winsorización (Referenciales)

Variable	Cola	Umbral	Obras afectadas
BD Vulnerabilidad del plazo	Superior (p95)	2.129	5
BE Incidencia temporal	Superior (p95)	0.907	5
BC % avance físico	Inferior (p5)	0.788	5

4.5.3. *Codificación Categórica*

Las categóricas se codifican según su naturaleza. Se binarizan las de dos niveles o reagrupables: Nivel de Gobierno (Local / No local), Familia SUCS (Familia_Grava), condición metropolitana y presencia de material inadecuado. La Calidad de Fundación se codifica de forma ordinal (baja/media/alta como 0/1/2), respetando su orden natural.

Las nominales con más de dos niveles se codifican con *one-hot encoding* y eliminación de una categoría de referencia (*drop-first*) para evitar colinealidad perfecta, por ejemplo, Tipología de Obra (cuatro categorías, se transforma a tres indicadores, tomando como referencia “Infraestructura Urbana”) y Escenario COVID (tres regímenes, se transforma a dos indicadores, tomando como referencia “pre-COVID”). Sin embargo, vale mencionar que esta adopción de la categoría de referencia (*drop-first*) solo aplica para modelos LASSO (Regresión Lineal y Cuadrática), ya que los modelos no paramétricos, como *random forest*, son robustos y no presentan colinealidad significativa frente a codificación *One-hot* estándar. Respecto a la variable que explica el Contexto COVID, se adopta codificación como *dummies* y no ordinal,

dado que el régimen post-COVID no representa un extremo de mayor severidad respecto al pre-COVID, sino un régimen cualitativamente distinto caracterizado por secuelas inflacionarias y ajustes de precios. Los algoritmos basados en árboles manejan indistintamente ambas codificaciones.

4.5.4. *Puntaje de Relevancia*

Para orientar la selección posterior, se calculó un puntaje de relevancia por objetivo que combina, en partes iguales y normalizadas, la fuerza de asociación (|Spearman| para las numéricas, Eta² para las categóricas) y la información mutua de cada predictora. El *rank_Score* no sustituye a la selección multivariada del Capítulo VI, sino que ofrece un ordenamiento univariado preliminar, útil para contrastar el consenso posterior. La Tabla 22 resume las cuatro predictoras de mayor puntaje por objetivo.

Tabla 22

Puntaje de Relevancia (Rank_score)

Objetivo	Predictora	Asociación	Info. mutua	*rank_Score*
BA	Plazo inicial	0.267	0.173	1.000
BA	Competitividad	0.163	0.113	0.628
BA	Tasa de pobreza	0.040	0.126	0.436
BA	N.º de postores	0.090	0.093	0.430
BB	Tasa de pobreza	0.144	0.078	0.675
BB	Competitividad	0.048	0.153	0.646
BB	Monto adjudicado	0.163	0.052	0.638
BB	N.º de postores	0.112	0.088	0.614
BC	Plazo inicial	0.170	0.173	1.000
BC	Tipología	0.111	0.174	0.827
BC	N.º de postores	0.155	0.087	0.672
BC	Nivel de gobierno	0.058	0.137	0.552
BD	Escenario COVID	0.149	0.176	1.000
BD	Monto adjudicado	0.123	0.147	0.827
BD	Competitividad	0.147	0.090	0.739

BD	Plazo inicial	0.120	0.117	0.727
BE	N.º de postores	0.098	0.077	0.884
BE	Tipología	0.071	0.090	0.809
BE	Plazo inicial	0.066	0.081	0.737
BE	Tasa de pobreza	0.036	0.099	0.681

4.5.5. *Preparación para el Análisis No Supervisado y Estado Final del Dataset*

El tratamiento de esta fase resulta crítico, pues como advierten Uddin et al. (2022), en aplicaciones de *machine learning* a la gestión de proyectos, la calidad del *dataset* condiciona directamente la validez del modelo predictivo, más que la sofisticación algorítmica adoptada. El criterio adoptado no descarta predictoras por baja correlación lineal, precisamente para preservar los efectos no lineales que los árboles aprovechan (evidenciados por la información mutua). Solo se descarta por colinealidad directa (VIF > 5 o |r| > 0.85), que aquí no se encontró. No se retiran variables predictoras en esta sección; solo se limita a anticipar, con la evidencia de correlación e información mutua, las variables que presentan asociación débil con las variables objetivo.

Las decisiones tomadas para la transformación de datos responden a tres objetivos: estabilizar la varianza de variables asimétricas, transformar variables cualitativas en representaciones numéricas procesables y construir indicadores normalizados que faciliten la interpretación comparativa entre obras de distinta escala. Todas las transformaciones sensibles a fuga se ejecutan dentro del *pipeline* de validación cruzada, ajustándose solo con el pliegue de entrenamiento.

Las 16 variables independientes se transforman en 21 columnas de entrada al aplicar codificación *one-hot* a las dos variables categóricas con más de dos niveles: la tipología de obra (4 columnas) y el escenario de pandemia (3 columnas). En los modelos paramétricos (LASSO lineal y cuadrático) se omite la categoría de referencia de cada una, Tip_Publica y Covid_PreCovid, para evitar colinealidad perfecta. Para los modelos no paramétricos (*random*

forest) se conservan todas las columnas. Tras aplicar las decisiones anteriores, el *dataset* queda conformado por n = 98 obras (n = 97 para BB), con 21 predictoras efectivas tras el *encoding* (Tabla 23) y las cinco variables objetivo. La razón de observaciones por predictor con el total de factores es 4.67, por debajo del umbral recomendado de 10, lo que hace indispensable la selección rigurosa de características.

Tabla 23

Transformación de Variables Independientes

Variable independiente	**Variables independientes tras *encoding***	**Tipo de dato**
Monto contratado	Monto_log	Decimal (logaritmo natural)
Plazo de ejecución inicial	Plazo_ejecucion_inicial	Entero (días)
Número de postores	Num_Postores_log1p	Decimal (log1p)
Ubicación metropolitana	Metropolitano	Binario (0/1)
Nivel de gobierno local	Nivel_local	Binario (0/1)
Tipo de contratista	Tipo_Contratista	Binario (consorcio = 1)
Presencia de material inadecuado	Material_inadecuado	Binario (0/1)
Familia de suelo (SUCS)	Familia_Grava	Binario (grava = 1)
Calidad de fundación	Calidad_fundacion_ord	Entero ordinal (baja/media/alta = 0/1/2)
Sistema de contratación	Sist_precios_unitarios	Binario (precios unitarios = 1)
Modalidad de licitación	Mod_licitacion_publica	Binario (licitación pública = 1)
Naturaleza de la obra	Naturaleza_obra_nueva	Binario (obra nueva = 1)
Tasa de pobreza del distrito	Site_tasa_pobreza	Decimal
Índice de competitividad regional	Competitividad	Decimal
Tipología de obra	Tip_Infraestructura Urbana y Espacios Publicos;	Categórica: 4 columnas binarias (*one-hot*)
	Tip_Infraestructura Vial y Transporte;	
	Tip_Saneamiento e Hidraulica;	
	Tip_Publica (Referencia)	
Escenario de pandemia	Covid_Pandemia;	Categórica: 3 columnas binarias (*one-hot*)
	Covid_PostCovid;	
	Covid_PreCovid (Referencia)	

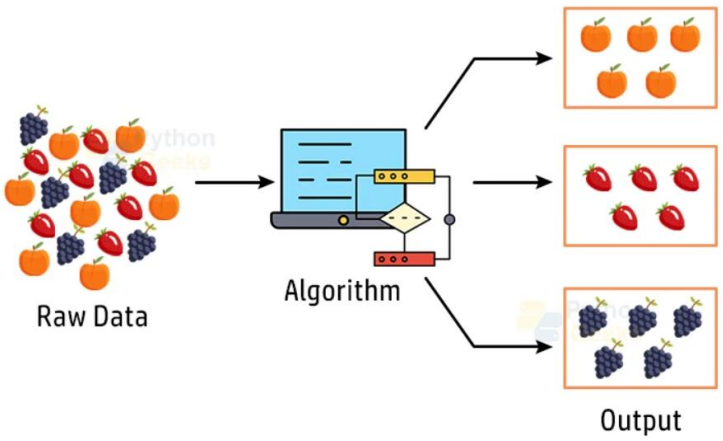
5. MODELADO APRENDIZAJE NO SUPERVISADO

5.1. Objetivo de *Clustering*

A diferencia de los modelos supervisados, donde se conocen las variables objetivo y se busca predecirlas, el análisis no supervisado no dispone de una etiqueta al predecir. Se particiona un conjunto de observaciones en grupos internamente homogéneos y externamente heterogéneos, de modo que las obras de un mismo grupo se parezcan entre sí más de lo que se parecen a las de otros grupos. En la literatura de gestión de obra pública, el *clustering* se ha empleado para identificar factores de riesgo y arquetipos de proyecto: Al-Bataineh et al. (2024) lo aplican para reconocer los factores clave de retraso, y Abdelalim et al. (2025) combinan *K-means* con decisión grupal difusa para agrupar proyectos por riesgo de sobrecosto. La presente investigación adopta esta lógica pero con un objetivo doble y explícito: describir los perfiles de desempeño y, sobre todo, construir estratos de forma equilibrada para la validación cruzada supervisada, evitando particiones degeneradas por el reducido tamaño muestral. La Figura 23 muestra el esquema de funcionamiento del algoritmo de *clustering*.

Figura 23

Funcionamiento de Algoritmo de Clustering



Nota: Recuperado de PythonGeeks (s. f.).

5.2. Selección de Variables para *Clustering*

El agrupamiento se realiza sobre las tres variables objetivo continuas que representan de forma completa las tres dimensiones del triángulo de hierro: la Vulnerabilidad Financiera Total (BA, costo), la Vulnerabilidad del Plazo Contractual Total (BD, plazo) y el % de Avance Físico (BC, alcance). Estas tres variables definen el espacio de desempeño de una obra: cuánto se desvió en costo, cuánto en plazo y cuánto conservó de su alcance físico.

Las variables endógenas (BB y BE) se excluyen deliberadamente del cálculo de los clústeres. Estas variables no son dimensiones primarias del desempeño, sino variables que descomponen las causas del desvío capturado por BA y BD; funcionan como instrumentos de diagnóstico causal *post-hoc*. Su papel es explicar, una vez formados los grupos, qué parte de la severidad de un clúster proviene de causas endógenas (deficiencias del expediente técnico) y qué parte de causas exógenas (paralizaciones o ampliaciones ajenas al proyecto).

Precisamente por eso su incorporación como coordenadas del espacio de agrupamiento sería contraproducente, puesto que introduciría en la distancia la información de diagnóstico que se quiere reservar para la interpretación posterior. Esta separación es la que evita falsos positivos al leer los índices globales. Una obra puede exhibir un sobrecosto elevado (por ejemplo, BA cercano a 200 %) con una incidencia de adicionales baja (BB cercano a 20 %); ello indica que el sobrecosto no proviene de adicionales surgidos en obra, sino de otros factores como reajustes de precios, mayores gastos generales o ineficiencias de gestión. De modo simétrico, una obra puede mostrar un retraso bruto grande (por ejemplo, BD cercano a 280 %) con BE = 0 %, lo que revela que no hubo ampliaciones de plazo por causas propias del proyecto, sino paralizaciones o suspensiones de origen mayormente exógeno. Mantener BB y BE fuera del cálculo de distancias y emplearlas después como lente causal es lo que habilita este tipo de lectura; por ello se reservan para el análisis posterior (sección 5.6).

Como nota estadística complementaria, ambas variables concentran además una fracción de ceros legítimos, cercana al 10 % en BB y al 8 % en BE, correspondiente a obras que efectivamente no requirieron adicionales o ampliaciones endógenas; en un algoritmo basado en distancia euclidiana como *K-means*, esa acumulación tendería a atraer los centroides y a reflejar la presencia o ausencia de una modificación antes que el gradiente de severidad del desempeño. Este efecto refuerza, en el plano operativo, la decisión ya justificada en el plano conceptual. En consecuencia, el espacio de agrupamiento queda definido por el vector tridimensional (BA, BD, BC) de cada obra, y BB y BE se incorporan únicamente como variables descriptivas para caracterizar los clústeres ya formados.

5.3. Definición de Hiperparámetros y Procedimiento

a. Estandarización de variables *z-score*

Antes de aplicar *K-means*, las tres variables se estandarizan mediante la transformación *z-score*, que a cada valor le resta la media de su variable y lo divide por su desviación estándar, de modo que cada variable queda con media 0 y desviación estándar 1. Este paso es una condición necesaria para *K-means*, cuyo criterio de agrupamiento es la distancia euclidiana entre observaciones (Al-Bataineh et al., 2024). La distancia euclidiana suma las diferencias al cuadrado de cada coordenada; si las variables tienen escalas distintas, la variable de mayor magnitud domina la distancia y anula el aporte de las demás.

En este caso la asimetría de escalas es evidente: la Vulnerabilidad del Plazo (BD) alcanza valores cercanos a 4.5 (447 % del plazo), el % de Avance Físico (BC) se mueve en el rango estrecho de 0.65 a 1.0, y la Vulnerabilidad Financiera (BA) entre 0 y 0.44. Sin estandarizar, la distancia estaría gobernada casi exclusivamente por BD, y las diferencias en avance físico o sobrecosto serían prácticamente invisibles para el

algoritmo. La estandarización coloca a las tres dimensiones en pie de igualdad, permitiendo que cada una contribuya de forma equilibrada a la formación de los grupos.

Adicionalmente, y solo para el cálculo de distancias, la variable BD se winsoriza al percentil 98 antes de estandarizar. El motivo es que la obra con retraso extremo (447 %) actuaría como un valor influyente que estiraría la escala de BD y desplazaría su centroide; al acotar su cola superior se estabiliza el reparto sin eliminar la observación. Esta winsorización es exclusivamente un recurso de cálculo de la partición, los centroides y todas las estadísticas descriptivas se reportan sobre los valores reales (no winsorizados).

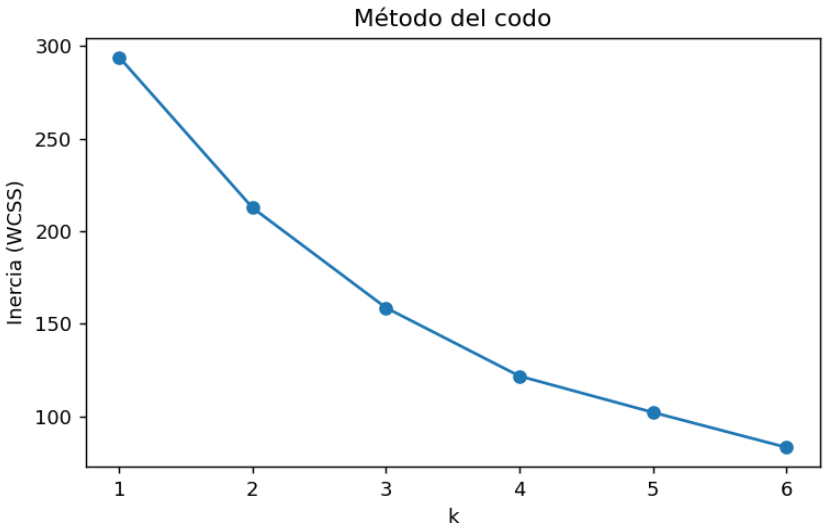
b. Definición de Clústeres

La elección de k *clusters* combina dos criterios: el método del codo y el criterio de dominio. El método del codo grafica la inercia (suma de distancias al cuadrado de cada obra a su centroide, WCSS) frente a k. La inercia siempre decrece al aumentar k, pero el "codo", el punto donde la ganancia marginal se aplana, señala el k a partir del cual añadir grupos deja de compensar. En este caso la inercia cae de 294.0 (k = 1) a 212.6 (k = 2) y a 158.6 (k = 3), y a partir de allí las reducciones son progresivamente menores (121.7 en k = 4), configurando un codo en k = 3. Se presenta gráfico en Figura 24.

Adicionalmente, según el criterio de dominio, la decisión final se fundamenta en el marco conceptual: el desempeño de una obra pública admite una estratificación natural en tres estados (exitosa, moderada y colapsada), coherente con la práctica de clasificación de riesgo en la literatura (Abdelalim et al., 2025) y con el objetivo de traducir los resultados en categorías accionables para la gestión. La convergencia de codo y dominio sustenta k = 3.

Figura 24

Método del Codo: Inercia Frente al Número de Clústeres



c. Parámetros de ejecución

El resto de hiperparámetros se fija para garantizar una solución estable y reproducible, resumidos en la Tabla 24.

Tabla 24

Hiperparámetros de K-means

Parámetro	Valor	Justificación
init	*K-means++*	Siembra inteligente de centroides iniciales que dispersa las semillas y reduce el riesgo de óptimos locales.
n_init	50	Número de reinicios con semillas distintas; se conserva la mejor solución (menor inercia), lo que estabiliza el resultado ante el azar de la inicialización.
max_iter	500	Tope de iteraciones por corrida; holgado para asegurar la convergencia del algoritmo con n pequeño.
random_state	42	Semilla fija para reproducibilidad exacta de la partición.
n_clusters	3	Definido por codo + dominio.

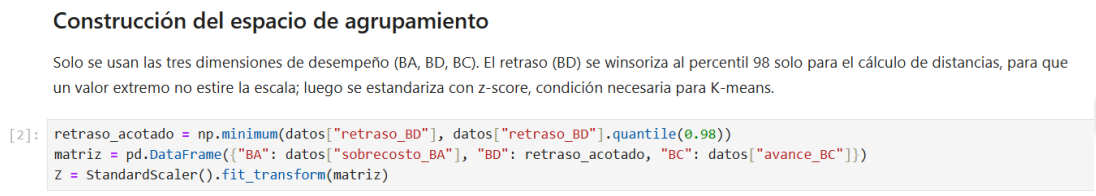
d. Código

El procesamiento y modelamiento de la fase de agrupamiento no supervisado se desarrollaron en Python utilizando la biblioteca scikit-learn, complementada con pandas, numpy y matplotlib para la estructuración y visualización de datos. El algoritmo

seleccionado fue *K-means*, aplicado sobre las dimensiones de desempeño continuo previamente estandarizadas a una escala común mediante *Z-score* (StandardScaler). Como paso crítico en la preparación del espacio vectorial, se aplicó una winsorización local al percentil 98 exclusivamente sobre la variable de retraso (BD) para el cálculo de distancias euclidianas, evitando que valores atípicos extremos distorsionen la segmentación (Figura 25).

Figura 25

Construcción del Espacio de Agrupamiento en JupyterLab



La configuración del modelo *K-means* definió la formación de tres agrupaciones (n_clusters=3), utilizando el método *K-means*++ para la siembra inteligente de centroides iniciales. Para asegurar la estabilidad y reproducibilidad de la partición, se configuraron 50 inicializaciones independientes (n_init=50), un máximo de 500 iteraciones (max_iter=500) y una semilla algorítmica fija (random_state=42). Finalmente, la nomenclatura de los arquetipos ("Exitosa", "Moderada" y "Colapsada") se automatizó calculando la severidad de cada grupo mediante la suma de sus centroides estandarizados de sobrecosto y retraso (Figura 26).

Figura 26

Aplicación de Algoritmo K-means en JupyterLab

Ajuste final con k = 3 y nombrado por severidad

El grupo con mayor suma de centroides estandarizados de sobrecosto y retraso (zBA + zBD) es el más severo (Colapsada); el de menor valor, (Exitosa).

```
[4]: kmeans = KMeans(n_clusters=3, init="k-means++", n_init=50, max_iter=500, random_state=SEMILLA).fit(Z)
etiquetas = kmeans.labels_
centros = kmeans.cluster_centers_              # columnas: BA, BD, BC

severidad = pd.Series(centros[:, 0] + centros[:, 1], index=range(3))   # zBA + zBD
orden = severidad.sort_values().index.tolist()
nombres = {orden[0]: "Exitosa", orden[1]: "Moderada", orden[2]: "Colapsada"}
datos["Cluster"] = pd.Series(etiquetas).map(nombres)

silueta_global = silhouette_score(Z, etiquetas)
reparto = datos["Cluster"].value_counts()[["Exitosa", "Moderada", "Colapsada"]]
print("Reparto:", reparto.to_dict(), "| Silueta global:", round(silueta_global, 3))

Reparto: {'Exitosa': 47, 'Moderada': 16, 'Colapsada': 35} | Silueta global: 0.309
```

5.4. Resultados del Agrupamiento

Con los parámetros anteriores, *K-means* asigna las 98 obras a tres grupos con un reparto de 47 obras exitosas, 16 moderadas y 35 colapsadas, y un coeficiente de silueta global de 0.309. La etiqueta de cada clúster se asigna de forma objetiva por su severidad de colapso costo-plazo, definida como la suma de los centroides estandarizados de sobrecosto y retraso, el grupo de menor valor es “Exitosa” y el de mayor valor, “Colapsada”. La Tabla 25 presenta los centroides en escala real.

Tabla 25

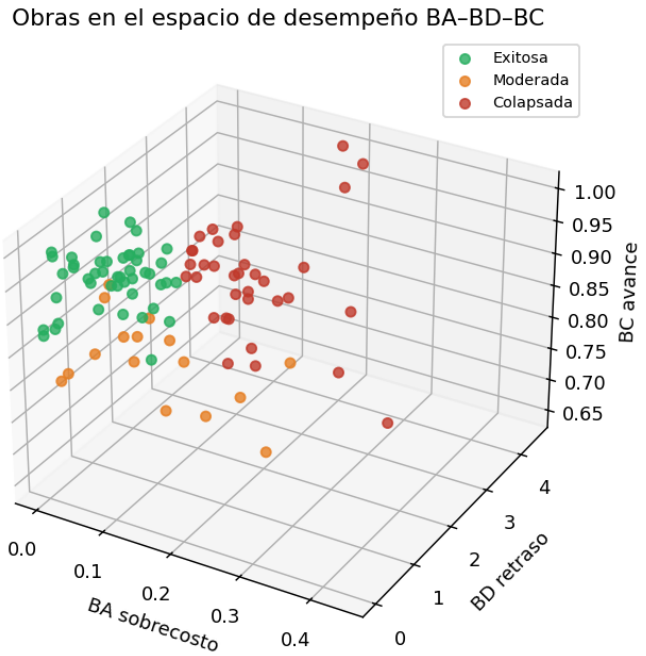
Centroides de los Clústeres

Clúster	n	BA sobrecosto	BD retraso	BC avance	BB endóg. costo	BE endóg. plazo
Exitosa	47	0.072	0.503	0.959	0.051	0.202
Moderada	16	0.095	1.121	0.807	0.088	0.311
Colapsada	35	0.222	1.379	0.959	0.087	0.526

La lectura del reparto (47/16/35) es coherente con el marco de sesgo optimista, las obras se distribuyen entre un grupo mayoritario de desempeño sano y dos grupos de desvío, sin que exista un único arquetipo dominante. La Figura 27 muestra las obras en el espacio de desempeño según las tres variables objetivo como ejes.

Figura 27

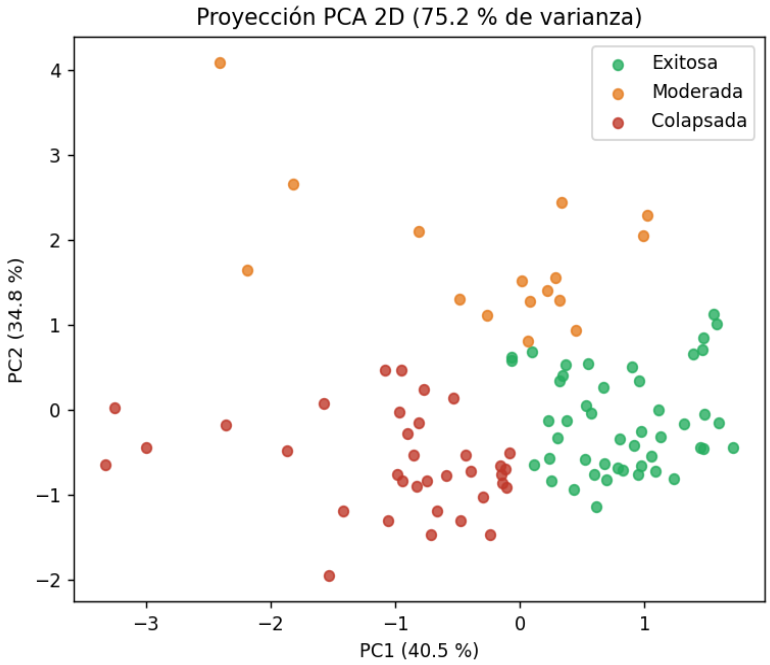
Obras en el Espacio de Desempeño BA–BD–BC



La proyección por Análisis de Componentes Principales (PCA) reduce las tres dimensiones de desempeño (sobrecosto, retraso y avance físico) a dos componentes que resumen el 75.3 % de la varianza total. La primera componente (PC1, 40.5 %) refleja sobre todo la magnitud de los desvíos de plazo y de costo: el retraso y el sobrecosto aportan casi todo su peso, cerca del 53 % y 44 % respectivamente, y el avance físico es marginal. Este eje resume la severidad conjunta de las ampliaciones de plazo y los sobrecostos, y ordena las obras desde las exitosas hasta las colapsadas sobre el eje horizontal. La segunda componente (PC2, 34.8 %) está dominada por el avance físico, que concentra cerca del 80 % de su peso, con un aporte menor del sobrecosto. Este eje mide el cumplimiento de metas físicas y es en buena medida independiente del primero. En conjunto, el retraso y el sobrecosto varían juntos, porque comparten el primer eje, mientras que el avance físico aporta una variación propia sobre el segundo. La Figura 28 presenta esta proyección.

Figura 28

Proyección PCA 2D de los Clústeres



5.5. Interpretación de los Clústeres

Los centroides y las pruebas *post-hoc* permiten caracterizar cada arquetipo en términos de gestión de obra. Para validar que las diferencias entre grupos son estadísticamente reales, se aplicó la prueba de Kruskal-Wallis a cada variable entre los tres clústeres, con la prueba *post-hoc* de Dunn para localizar los pares de grupos que difieren (Tabla 26).

Tabla 26

Validación de la Separación entre Clústeres (Kruskal-Wallis)

Variable	H	p	Eta²	Rol
BA sobrecosto	55.47	<0.001	0.563	*Clustering* — separa Colapsada del resto
BD retraso	30.41	<0.001	0.299	*Clustering* — separa Exitosa del resto
BC avance	38.20	<0.001	0.381	*Clustering* — aísla a Moderada (avance bajo)
BB endóg. costo	12.17	0.002	0.108	*Post-hoc* — Exitosa difiere de los desviados
BE endóg. plazo	27.08	<0.001	0.256	*Post-hoc* — aumento monótono hacia Colapsada

Las cinco variables difieren significativamente entre clústeres (todas con $p \leq 0.003$), y los tamaños de efecto son grandes en las tres dimensiones del agrupamiento (Eta^2 de 0.31 a 0.57), lo que confirma que la partición captura estructura real. La prueba de Dunn precisa la firma de cada grupo:

- Clúster Exitoso (n = 47). Bajo desvío integral: sobrecosto de 7.2 %, retraso de 50.3 % y avance físico de 95.9 %. La prueba de Dunn lo señala como el único grupo con retraso significativamente bajo y con la menor incidencia endógena de costo.
- Clúster Colapsado (n = 35). El colapso costo-plazo: concentra el mayor sobrecosto (22.2 %, cerca del triple del resto; difiere de ambos grupos con $p < 0.001$) y el mayor retraso (137.9 %). Es relevante que sus obras sí alcanzan el 96 % de avance físico: no son obras inconclusas, sino obras que se terminaron a costa de fuertes sobrecostos y ampliaciones. Su incidencia endógena de plazo (BE = 0.526) es la más alta y difiere significativamente de la del grupo exitoso ($p < 0.001$), lo que indica que su colapso obedece a causas endógenas, como adicionales y ampliaciones vinculadas a deficiencias del expediente técnico, más que a paralizaciones exógenas.
- Clúster Moderado (n = 16). Un perfil intermedio con una firma distintiva: es el único grupo con avance físico significativamente bajo (80.7 %; difiere de Exitosa y Colapsada con $p < 0.001$), pese a un sobrecosto contenido (9.5 %, cercano al exitoso) y un retraso intermedio (112.1 %). Corresponde a obras formalmente liquidadas cuyo alcance ejecutado quedó por debajo del previsto en el expediente original, un patrón cualitativamente distinto del colapso financiero del grupo anterior.

El análisis *post-hoc* de las variables endógenas aporta la lectura causal que el agrupamiento por sí solo no da. La incidencia endógena de plazo (BE) crece monótonamente de Exitosa (0.202) a Moderada (0.311) y Colapsada (0.526), y su diferencia entre extremos es significativa. Esto sugiere que, a medida que una obra se acerca al arquetipo colapsado, una

fracción creciente de su retraso proviene de ampliaciones endógenas, es decir, atribuibles principalmente a la calidad del expediente técnico. La misma dirección se observa, con menor magnitud, en la incidencia endógena de costo (BB). Este hallazgo es coherente con el diagnóstico de la Contraloría General de la República (2025), que sitúa las deficiencias de los expedientes técnicos entre las causas dominantes del desvío en la obra pública peruana. La distribución de los clústeres según las variables BB y BE se presentan en la Figura 29 y 30.

Figura 29

Incidencia Endógena Económica (BB) por Clúster

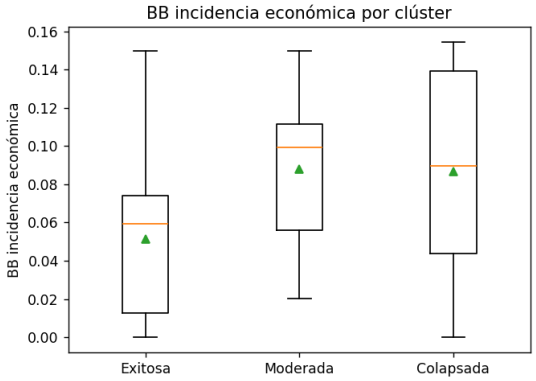
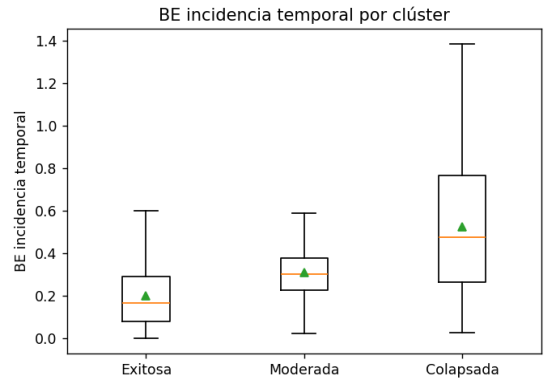


Figura 30

Incidencia Endógena de Plazo (BE) por Clúster



6. MODELADO APRENDIZAJE SUPERVISADO

6.1. Selección Multivariada de Características

6.1.1. *Objetivo de Selección de Características*

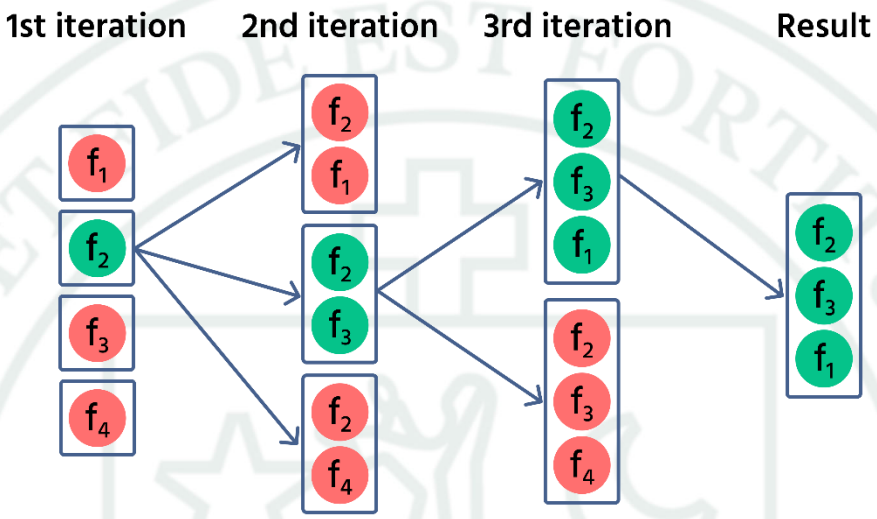
En *machine learning*, se recomienda que la cantidad de datos supere al menos por un factor de diez al número de predictores (Graphite Note, 2024). Al contar con 98 observaciones y 21 variables, esta relación queda por debajo del umbral recomendado. Por ello, se hace necesario un procedimiento sistemático de selección de variables para preservar la validez del análisis. Esta sección se organiza en torno a dos principios. El primero es un enfoque diferenciado por objetivo: como los determinantes del sobrecosto financiero no coinciden necesariamente con los del retraso o los del avance físico, cada variable dependiente pasa por su propio proceso de selección, que reduce el espacio de predictores a un subconjunto pequeño y reproducible con la mayor capacidad predictiva posible. El segundo es el uso de *random forest* como modelo no lineal de referencia, por su robustez ante interacciones complejas, su tolerancia a variables de distinta naturaleza y su baja dependencia de supuestos distribucionales.

Seleccionar características consiste en elegir, del banco de predictores disponible, el subconjunto que mejor explica la variable objetivo. Los métodos envolventes (*wrappers*) evalúan subconjuntos completos de variables por su desempeño predictivo estimado. En este estudio, el método principal de selección es la selección secuencial hacia adelante (*forward selection*), un *wrapper* que parte de un conjunto vacío y, en cada paso, incorpora la variable cuya inclusión produce el mayor incremento del coeficiente de determinación (R^2) estimado por validación cruzada; el proceso se detiene en el tamaño de subconjunto que maximiza dicho R^2. La ventaja de este enfoque frente a un simple *ranking* de importancia es que detecta variables cuyo aporte solo aparece en combinación con otras, un efecto que un *ranking*

univariado o un criterio de relevancia marginal puede pasar por alto. En la Figura 31 se esquematiza el funcionamiento del algoritmo.

Figura 31

Algoritmo Sequential Forward Selection



Nota: Recuperado de Codefinity (s. f.).

6.1.2. *Definición de Hiperparámetros y Procedimiento*

a. Validación cruzada estratificada y repetida

Con una muestra pequeña (n = 98), estimar el desempeño con una única partición de entrenamiento y prueba resulta inestable. Por eso se emplea validación cruzada *K-fold* repetida: los datos se dividen en cinco pliegues; cada pliegue actúa una vez como prueba mientras los otros cuatro entrenan; y todo el proceso se repite varias veces con particiones distintas, promediando los resultados y reportando media y desviación estándar. Cinco pliegues equilibran el compromiso sesgo–varianza del estimador y las repeticiones atenúan el azar de la partición (Krstajic et al., 2014).

Además, la validación se estratifica por la etiqueta de clúster obtenida en el Capítulo V. Como la muestra es heterogénea, conviven obras exitosas y colapsadas, estratificar asegura que los tres regímenes queden representados en cada pliegue, lo que

reduce la varianza del estimador; es la extensión natural de la validación cruzada estratificada, estándar cuando existen subgrupos. Este mismo criterio de estratificación se aplica por igual a las cinco variables objetivo y en ambas fases, de modo que las comparaciones sean consistentes. La etiqueta de clúster solo balancea la composición de los pliegues; no interviene en el ajuste de los modelos.

b. Transformaciones: winsorización local y escalado

Algunas variables objetivo presentan colas largas. Para que unos pocos valores extremos no dominen el entrenamiento, se aplica winsorización local, el valor se acota al percentil 95 superior (BD, BE) o al percentil 5 inferior (BC). Que sea local significa que el umbral se calcula únicamente con los datos de entrenamiento de cada pliegue, lo que evita la fuga de información. Además, la evaluación se realiza siempre contra los valores reales, no winsorizados.

c. Hiperparámetros modelo *random forest*

Para la ejecución del método de selección secuencial hacia adelante, se utilizará un modelo *random forest* que por su naturaleza de ensamble con profundidad y hojas acotadas, tolera bien varios predictores débiles sin sobreajustar. Este modelo representa un conjunto de árboles de decisión entrenados sobre remuestras *bootstrap*, donde cada división considera un subconjunto aleatorio de variables; la predicción promedia los árboles y reduce la varianza. La selección de hiperparámetros se orienta por el compromiso sesgo–varianza dado el tamaño muestral, de acuerdo con la Tabla 27.

Tabla 27

Hiperparámetros Modelo random forest

Parámetro	Valor	Justificación
n_estimators	300	Número de árboles; suficientes para estabilizar el promedio.
max_depth	5	Profundidad máxima; limitarla es esencial con n = 98, pues los árboles a profundidad libre memorizan la muestra (varianza excesiva) y no generalizan.

min_samples_leaf	5	Mínimo de observaciones por hoja; evita hojas basadas en una o dos obras.
max_features	4	Valor por debajo de las 7 a 9 variables de cada objetivo, de modo que cada árbol trabaja con un subconjunto distinto, lo que decorrelaciona los árboles y mejora la generalización del bosque.
random_state	42	Semilla fija para asegurar la reproducibilidad exacta del remuestreo *bootstrap* y las particiones en cada ejecución.

Se fijan alrededor de ocho variables por objetivo. El criterio adoptado es una selección por estabilidad; se repite la selección forward K = 10 veces, cada una con una partición de validación cruzada distinta, tomando en cada corrida la trayectoria completa de la selección; se registra la frecuencia con que cada variable es seleccionada. Finalmente se retienen las que superan un umbral de consenso del 50 % (mayoría de las corridas), con un piso mínimo de siete variables por objetivo.

d. Código

El modelamiento para la selección de características se desarrolló en Python utilizando la biblioteca scikit-learn (específicamente RandomForestRegressor y StratifiedKFold), con el apoyo de pandas y numpy para el procesamiento matricial. El enfoque algorítmico implementado fue una Selección Secuencial hacia Adelante (*Forward Selection Wrapper*) reforzada metodológicamente con un criterio de Selección por Estabilidad (*Stability Selection*). El código se estructura en dos procedimientos críticos diseñados para evitar la fuga de información (*data leakage*) y aislar el ruido estadístico. En primer lugar, se definió una función de evaluación mediante validación cruzada estratificada de 5 pliegues; esta función incorpora un filtro dinámico (notna()) que excluye automáticamente las observaciones con valores nulos en la variable objetivo (permitiendo evaluar incidencias como BB con n=97 sin afectar el resto del *dataset*), y aplica una winsorización estrictamente local sobre los datos de entrenamiento para calcular el coeficiente de determinación (R^2) frente a los valores reales de prueba (Figura 32).

Figura 32

Winsorización y Validación Cruzada para random forest

Winsorización y evaluación por validación cruzada

```
def winsor(y,code):
    p=WIN.get(code)
    if p is None: return y
    return np.minimum(y,np.percentile(y,p)) if p>0 else np.maximum(y,np.percentile(y,-p))
RF=dict(n_estimators=300,max_depth=5,min_samples_leaf=5,random_state=RNG)
def cv_r2(cols,code,seed):
    ok=df[TGT[code]].notna().values                # descarta filas sin dato del objetivo (p.ej. BB)
    X=df.loc[ok,cols].values; y=df.loc[ok,TGT[code]].values; cl=clab[ok]
    skf=StratifiedKFold(5,shuffle=True,random_state=seed); oof=np.zeros(len(y))
    for tr,va in skf.split(X,cl):
        rf=RandomForestRegressor(max_features=min(4,len(cols)),**{**RF,'random_state':seed}); rf.fit(X[tr],winsor(y[tr],code)); oof[
    return r2_score(y,oof)
```

El núcleo algorítmico se encuentra en las funciones forward_path y *consenso*. La primera evalúa iterativamente el banco de 21 columnas, incorporando paso a paso la variable que mayor incremento genere en el R^2 hasta conformar subconjuntos de aproximadamente 8 predictores (kmax=8). Debido a que la muestra es moderada, la función de consenso repite este proceso iterativo 10 veces (K=10) con distintas particiones aleatorias, calculando la frecuencia de selección de cada variable. Finalmente, el código dictamina como predictores definitivos únicamente a aquellos que superan un umbral de consenso del 50% a lo largo de las corridas (Figura 33).

Figura 33

Aplicación e Iteración de Algoritmo Forward Selection

Forward de trayectoria completa y consenso ≥ 50 %

Cada corrida construye la trayectoria forward hasta `kmax=8` (sin truncar en el máximo, para poder rankear alrededor de 8 variables). Se repite K = 10 veces y se cuenta la frecuencia de selección.

```
def forward_path(code,seed,kmax=8):
    rest=list(BANK); orden=[]; path=[]
    while rest and len(orden)<kmax:
        best=max(((cv_r2(orden+[cc],code,seed),cc) for cc in rest),key=lambda t:t[0])
        orden.append(best[1]); rest.remove(best[1]); path.append(best[0])
    return orden, path

def consenso(code,K=10,umbral=0.5,minimo=7):
    frec=Counter(); paths=[]
    for r in range(K):
        orden,path=forward_path(code,seed=100+r*13); paths.append(path)
        for v in set(orden): frec[v]+=1
    ranked=[v for v,_ in frec.most_common()]
    sel=[v for v in ranked if frec[v]/K>=umbral]
    if len(sel)<minimo: sel=ranked[:minimo]
    return sel, frec, K, paths

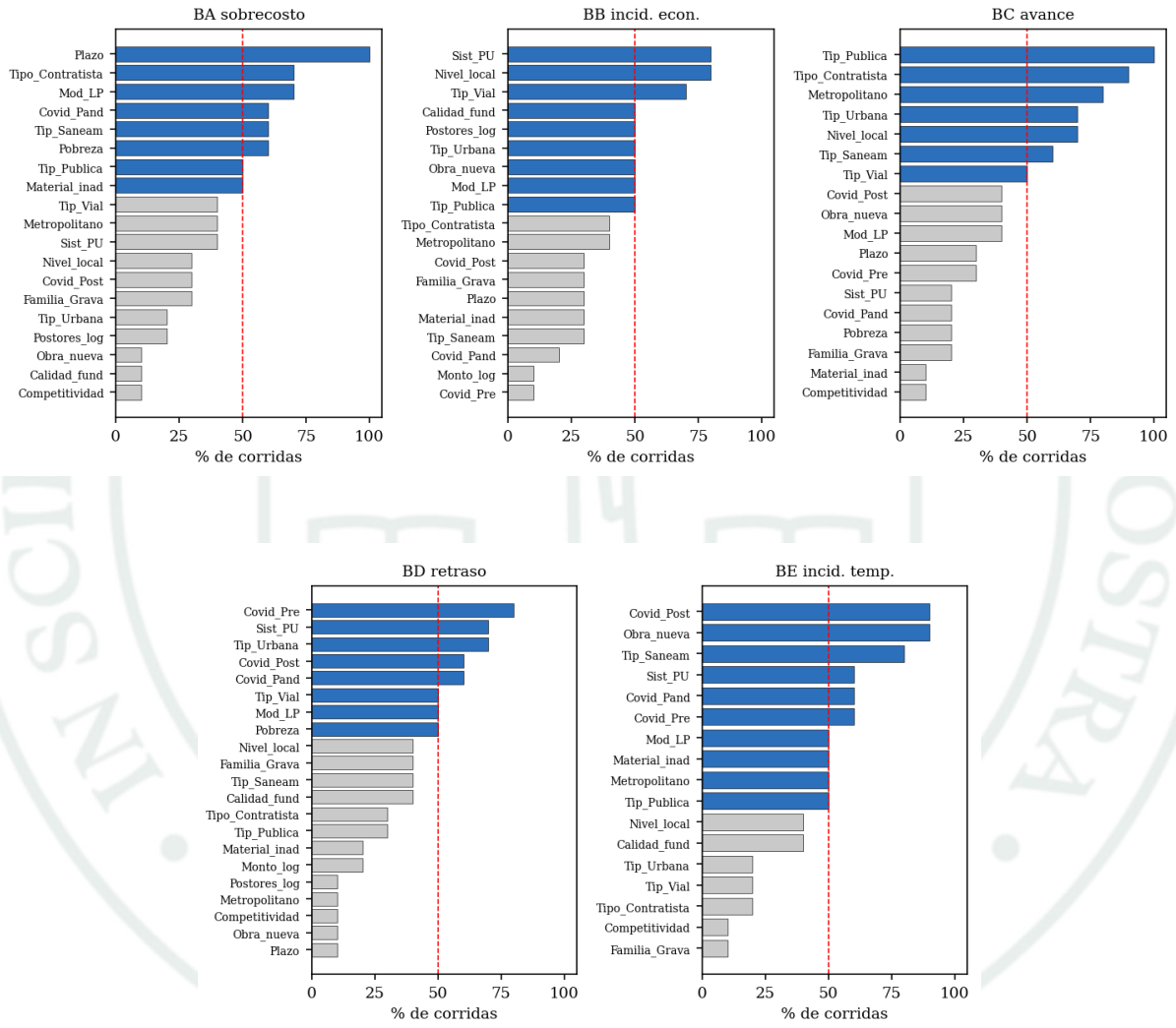
RESULT={c:consenso(c) for c in ['BA','BB','BC','BD','BE']}
for c,(s,f,K,p) in RESULT.items(): print(c,len(s),'->',s)
```

6.1.3. *Consenso de Selección*

La Figura 34 muestra, para las cinco variables objetivo, la frecuencia de selección de cada variable en las diez corridas. En azul, las variables que superan el umbral de consenso del 50 % (línea roja).

Figura 34

Frecuencia de Selección por Variable

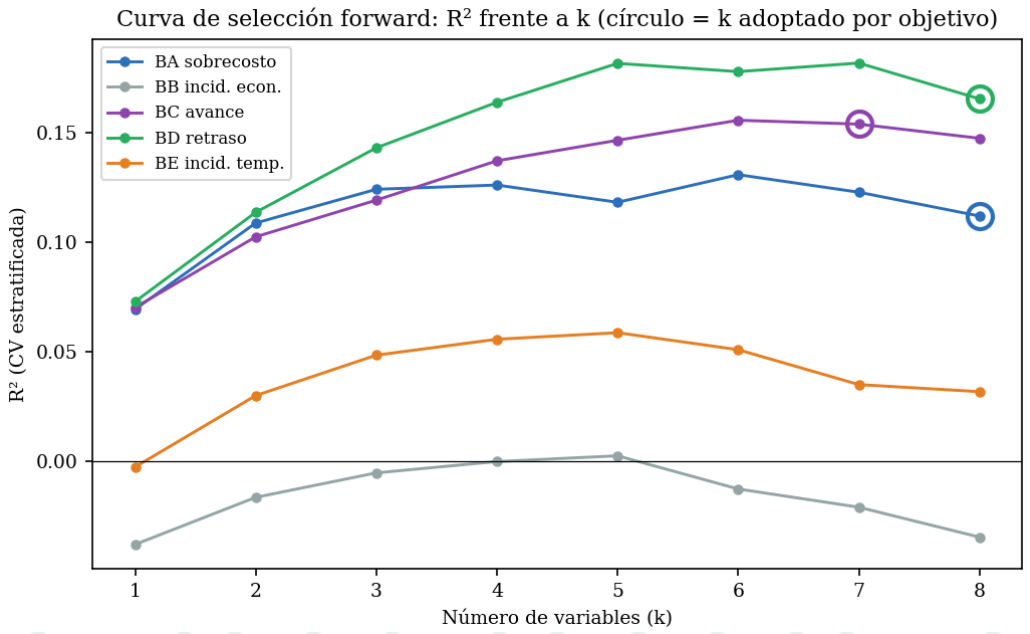


La selección hacia adelante genera, para cada objetivo, una curva del R^2 de validación cruzada en función del número de variables incorporadas (k). Para el sobrecosto (BA) y el avance (BC), el R^2 de validación cruzada sigue creciendo hasta que se tienen 7-8 variables predictoras; para el retraso (BD) alcanza su mejor nivel y se sostiene en ese entorno; los

objetivos tienden a mejorar conforme aumenta la cantidad de predictoras hasta que llegan alrededor de 7. Por su parte, las variables BB y BE muestran un R^2 bajo, lo que demuestra la baja capacidad predictiva de estas variables. La información se muestra en la Figura 35.

Figura 35

Curva de Selección Forward: R^2 Frente al Número de Variables K



Como el forward incorpora en cada paso la variable que más incrementa el R^2, las variables de mayor frecuencia de selección son, por construcción, las que más elevan el desempeño; frecuencia y contribución al ajuste coinciden. Finalmente, en la Tabla 28 se recopilan los subconjuntos seleccionados.

Tabla 28

Predictores Seleccionados por Objetivo

Objetivo	N.º	Predictores seleccionados
BA (sobrecosto)	8	Plazo inicial (100 %), Modalidad licitación pública (70 %), Tipo de contratista (70 %), Contexto pandemia (60 %), Tasa de pobreza (60 %), Tipología saneamiento (60 %), Material inadecuado (50 %), Infraestructura pública, ref. (50 %)
BB (incid. económica)	9	Sistema precios unitarios (80 %), Nivel de gobierno local (80 %), Tipología vial (70 %), N.º postores (50 %), Tipología urbana (50 %), Calidad de fundación (50

		%), Modalidad licitación pública (50 %), Naturaleza obra nueva (50 %), Infraestructura pública, ref. (50 %)
BC (avance)	7	Infraestructura pública, ref. (100 %), Tipo de contratista (90 %), Ámbito metropolitano (80 %), Tipología urbana (70 %), Nivel de gobierno local (70 %), Tipología saneamiento (60 %), Tipología vial (50 %)
BD (retraso)	8	Contexto pre-COVID, ref. (80 %), Sistema de contratación (70 %), Tipología urbana (70 %), Contexto post-COVID (60 %), Contexto pandemia (60 %), Tipología vial (50 %), Modalidad licitación pública (50 %), Tasa de pobreza (50 %)
BE (incid. temporal)	9	Contexto post-COVID (90 %), Naturaleza obra nueva (90 %), Tipología saneamiento (80 %), Sistema precios unitarios (60 %), Contexto pandemia (60 %), Contexto pre-COVID (60 %), Modalidad licitación pública (50 %), Material inadecuado (50 %), Ámbito metropolitano (50 %)

La composición de cada subconjunto integra dimensiones técnicas, administrativas y contractuales de la obra junto con factores contextuales, garantizando un espacio de características suficiente para la diferenciación empírica de proyectos. Ante la marcada correlación observada entre el fenómeno del COVID-19 y el retraso contractual (BD), y en vista de que la totalidad de los proyectos prospectivos pertenecen al régimen post-COVID, se integraron factores estructurales del expediente como el sistema de contratación, la tipología, la modalidad y el contexto social (tasa de pobreza). De este modo, la capacidad predictiva del modelo se sostiene en factores técnicos permanentes y no en una contingencia temporal transitoria.

6.2. Regresión Supervisada de Modelos Línea Base

6.2.1. *Definición de Hiperparámetros y Procedimiento*

Se evaluaron dos variantes de regresión lineal regularizada mediante penalización (LASSO): lineal y cuadrática. En la versión cuadrática, el subconjunto de variables continuas se expande a polinomios de segundo grado para capturar curvaturas e interacciones no lineales simples. El hiperparámetro de penalización (alpha) se ajusta por validación cruzada interna al

conjunto de entrenamiento de cada pliegue, de modo que su elección no contamina la estimación de desempeño.

El objetivo es estimar, con el subconjunto de la Selección de variables, el desempeño predictivo continuo de cada objetivo, comparando la referencia lineal regularizada frente al modelo no lineal de ensamble. El procedimiento implementado evalúa cada par bajo el mismo esquema de validación de la sección 6.1., ajustando dentro de cada pliegue de entrenamiento la winsorización, el escalado y el alpha de LASSO, y midiendo sobre los valores reales.

La regresión de línea base se implementó en Python con scikit-learn, empleando tres estimadores: *LassoCV* para la regresión lineal regularizada, la misma clase combinada con *PolynomialFeatures* mediante un *make_pipeline* con StandardScaler para la variante cuadrática, y *RandomForestRegressor* como modelo no lineal de ensamble. La evaluación se apoya en *RepeatedStratifiedKFold* y en las métricas de desempeño, pandas y numpy dan soporte al manejo de datos.

La parte central del código es la función oof (out-of-*fold*), que produce las predicciones fuera de pliegue de cada modelo. A diferencia de una partición única, se emplea una validación cruzada repetida *RepeatedStratifiedKFold* con 5 pliegues y 8 repeticiones, estratificada por la etiqueta de clúster. Así, las predicciones de cada observación se promedian entre repeticiones, lo que estabiliza la estimación del R^2 con una muestra moderada. Similar a la selección de variables, la función descarta con *notna()* las filas sin dato del objetivo y winsoriza únicamente el pliegue de entrenamiento. El bucle final recorre los cinco objetivos y los tres modelos, y consolida los resultados (R^2, MAE y RMSE) en la tabla comparativa (Figura 36).

Figura 36

Validación Cruzada Repetida y Comparación de Modelos en JupyterLab

Evaluación por validación cruzada (R^2, MAE, RMSE)

Se generan predicciones fuera de pliegue para cada modelo y objetivo.

```
[ ]: RF=dict(n_estimators=300,max_depth=5,min_samples_leaf=5,random_state=RNG)
def oof(code, model):
    cols=SETS[code]; ok=df[TGT[code]].notna().values   # descarta filas sin dato del objetivo (p.ej. BB)
    X=df.loc[ok,cols].values; y=df.loc[ok,TGT[code]].values; cl=clab[ok]
    rk=RepeatedStratifiedKFold(n_splits=5,n_repeats=8,random_state=RNG)  # CV repetida
    acc=np.zeros(len(y)); cnt=np.zeros(len(y))
    for tr,va in rk.split(X,cl):
        yt=winsor(y[tr],code)
        if model=='rf':
            m=RandomForestRegressor(max_features=min(4,len(cols)),**RF)
        elif model=='lin':
            m=make_pipeline(StandardScaler(), LassoCV(cv=3,random_state=RNG,max_iter=5000))
        else:
            m=make_pipeline(StandardScaler(), PolynomialFeatures(2,include_bias=False), LassoCV(cv=3,random_state=RNG,max_iter=5000))
        m.fit(X[tr],yt); acc[va]+=m.predict(X[va]); cnt[va]+=1
    return y, acc/np.maximum(cnt,1)
```

6.2.2. *Resultados de Modelos Línea Base*

La Tabla 29 reporta las métricas de evaluación y la Figura 37 resume el R^2 por objetivo.

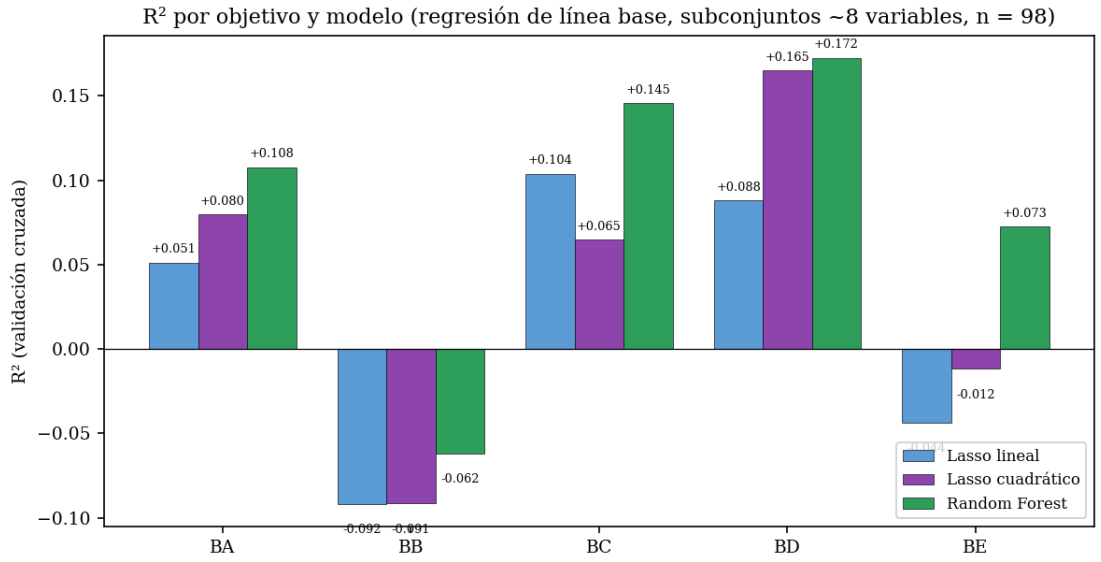
Tabla 29

Desempeño de Regresión Modelos Línea Base

Objetivo	Modelo	R^2	MAE	RMSE
BA	LASSO lineal	+0.051	0.0725	0.0898
BA	LASSO cuadrático	+0.080	0.0719	0.0887
BA	*Random forest*	+0.108	0.0702	0.0894
BB	LASSO lineal	−0.092	0.0412	0.0487
BB	LASSO cuadrático	−0.091	0.0412	0.0486
BB	*Random forest*	−0.062	0.0413	0.0492
BC	LASSO lineal	+0.104	0.0516	0.0662
BC	LASSO cuadrático	+0.065	0.0519	0.0676
BC	*Random forest*	+0.145	0.0481	0.0664
BD	LASSO lineal	+0.088	0.5387	0.7808
BD	LASSO cuadrático	+0.165	0.5118	0.7513
BD	*Random forest*	+0.172	0.5024	0.7718
BE	LASSO lineal	−0.044	0.2180	0.2779
BE	LASSO cuadrático	−0.012	0.2132	0.2737
BE	*Random forest*	+0.073	0.2082	0.2673

Figura 37

R^2 por Objetivo y Modelo



El retraso contractual (BD) se consolidó como el objetivo de mayor capacidad predictiva en la modelación continua, alcanzando un R^2=+0.172 mediante *random forest* y +0.165 con LASSO cuadrático, seguido en desempeño por el avance físico (BC, R^2 = +0.145), el sobrecosto financiero (BA, R^2=+0.108) y la incidencia temporal (BE, R^2 = +0.073). La retención de un espacio dimensional parsimonioso en torno a 7 u 8 predictores incrementó sustancialmente la precisión en BD y BC al incorporar factores complementarios como la tipología de obra, el nivel de gobierno y el tipo de contratista, permitiendo a *random forest* igualar o superar a las formulaciones LASSO en los objetivos con señal empírica válida. La baja modelabilidad continua de BB y BE responde a dos factores principales. Primero, la concentración de ceros legítimos (obras sin adicionales o sin ampliaciones endógenas) deja distribuciones difíciles de ajustar como variable continua; por eso BB no supera a la media histórica y BE no alcanza una discriminación operativa en la fase de clasificación. Segundo, ambas variables dependen de eventos contingentes que surgen durante la ejecución, como adicionales, ampliaciones y cambios de alcance ligados a circunstancias particulares de cada

obra; su componente aleatorio es alto y no queda determinado por las variables *ex-ante* disponibles antes de la licitación.

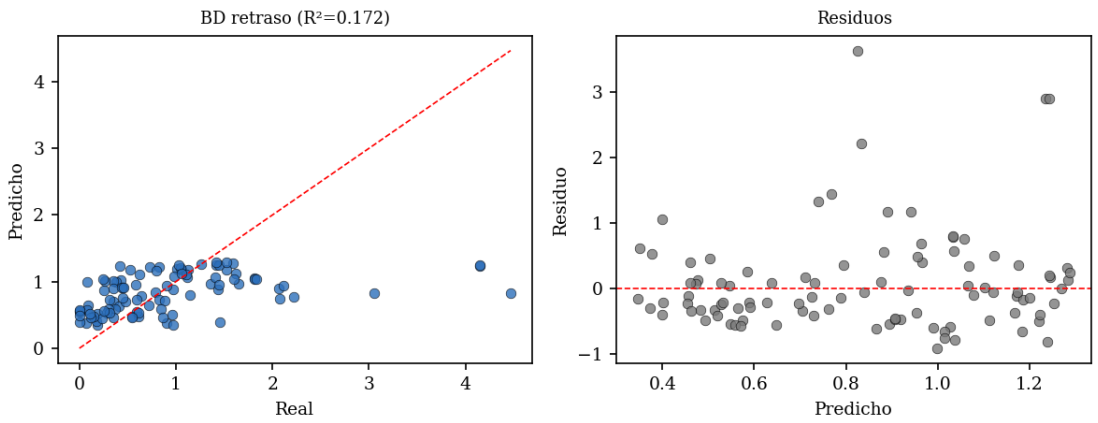
6.2.3. *Análisis Diagnóstico por Objetivo*

a. Retraso de plazo (BD)

El retraso contractual (BD) alcanzó el mejor desempeño predictivo de la modelación continua (R^2 = +0.172) con *random forest*. El subconjunto seleccionado de ocho variables articula condiciones estructurales del expediente técnico, como el esquema de contratación, las tipologías vial y urbana, la modalidad del procedimiento de selección y el nivel de pobreza de la zona, con las variables de control del período de emergencia sanitaria. Esta distribución equilibrada entre atributos propios del proyecto y factores contextuales resulta metodológicamente valiosa, pues garantiza que el modelo no quede anclado a la coyuntura de la pandemia y pueda transferirse con solvencia a la evaluación de obras futuras. El gráfico de residuos y de predicho vs real se presenta en la Figura 38.

Figura 38

BD: Predicho vs. Real y Residuos (random forest)



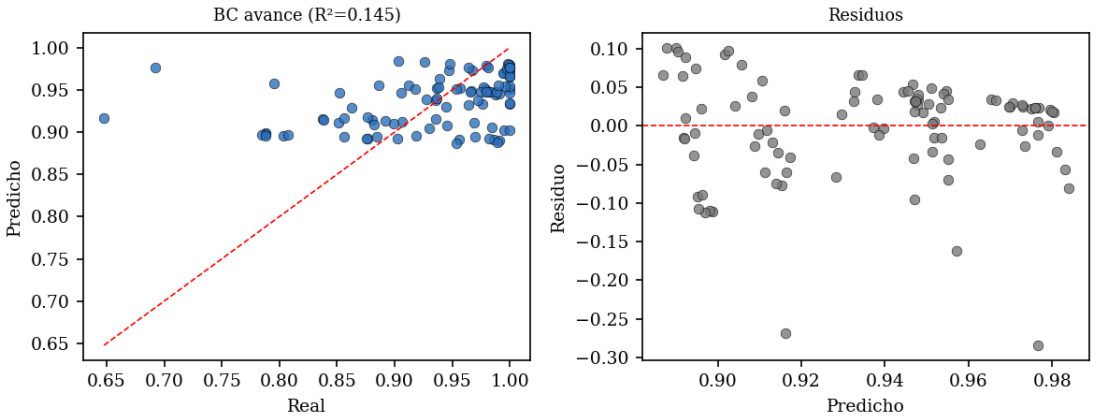
b. Avance físico de alcance inicial (BC)

La variable BC mide el porcentaje del alcance original del expediente técnico que se conservó al momento de la liquidación, no un grado de terminación de la obra. Todas las obras del conjunto están formalmente concluidas y liquidadas. Este fue el segundo

objetivo con mejor desempeño (R^2 = +0.145), encabezado por la tipología Infraestructura pública, el tipo de contratista y el ámbito metropolitano. La ventaja de *random forest* sobre las alternativas lineales sugiere interacciones entre estos factores que el modelo lineal no representa. El gráfico de residuos y de predicho vs real se presenta en la Figura 39.

Figura 39

BC: Predicho vs. Real y Residuos (random forest)

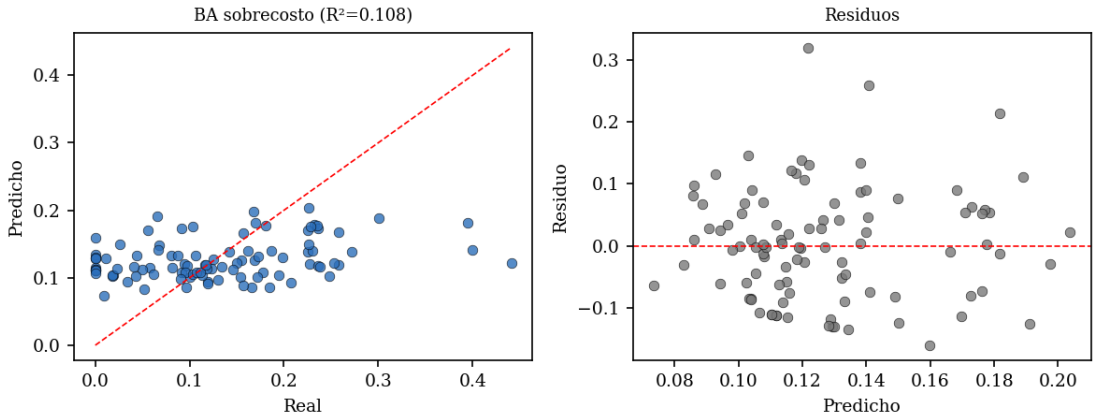


c. Sobrecosto financiero (BA)

El sobrecosto financiero (BA) alcanzó un ajuste continuo de R^2 = +0.108 mediante *random forest*, sustentado principalmente en el plazo inicial como el factor dominante y de mayor consenso, seguido por la modalidad del procedimiento de selección y el tipo de contratista. Este núcleo predictivo se complementa con variables de entorno y escala como la tasa de pobreza, la tipología sectorial y el contexto de pandemia. La selección sistemática y prioritaria de la duración programada no solo ratifica su peso empírico dentro de los algoritmos de ensamble, sino que coincide plenamente con los diagnósticos de asociación exploratoria identificados en el EDA. El gráfico de residuos y de predicho vs real se presenta en la Figura 40.

Figura 40

BA: Predicho vs. Real y Residuos (random forest)



d. Incidencia por adicionales (BB y BE)

Ambos objetivos concentran una masa de ceros legítimos (obras sin adicionales o sin ampliaciones endógenas), lo que dificulta la regresión continua. A esto se suma el componente aleatorio de las variables endógenas, el cual es alto y no queda determinado por las variables *ex-ante* disponibles antes de la licitación. En BB, ninguno de los algoritmos superó al modelo nulo trivial, arrojando valores de R^2 negativos (*random forest* -0.062); mientras que en BE, si bien *random forest* extrajo un ajuste positivo marginal (R^2 = +0.073, MAE = 0.2082) frente a las pendientes negativas de LASSO, dicho rendimiento resultó insuficiente para estructurar una frontera de decisión analítica confiable.

Figura 41

BB: Predicho vs. Real y Residuos (random forest)

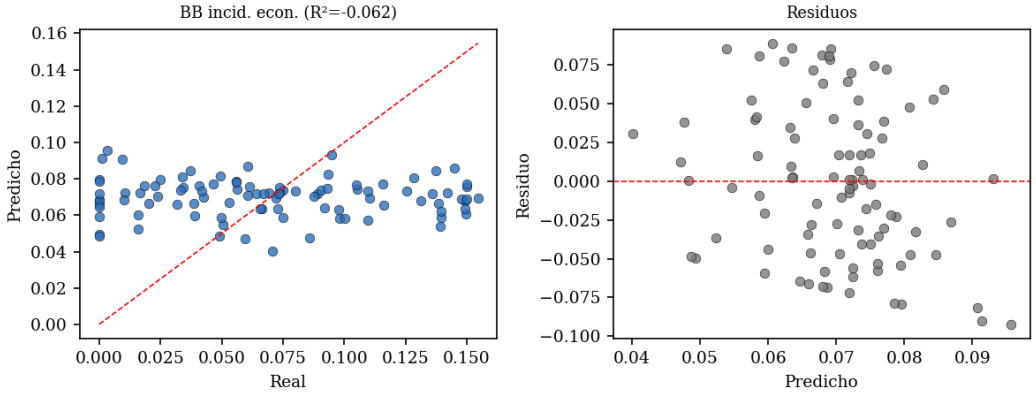
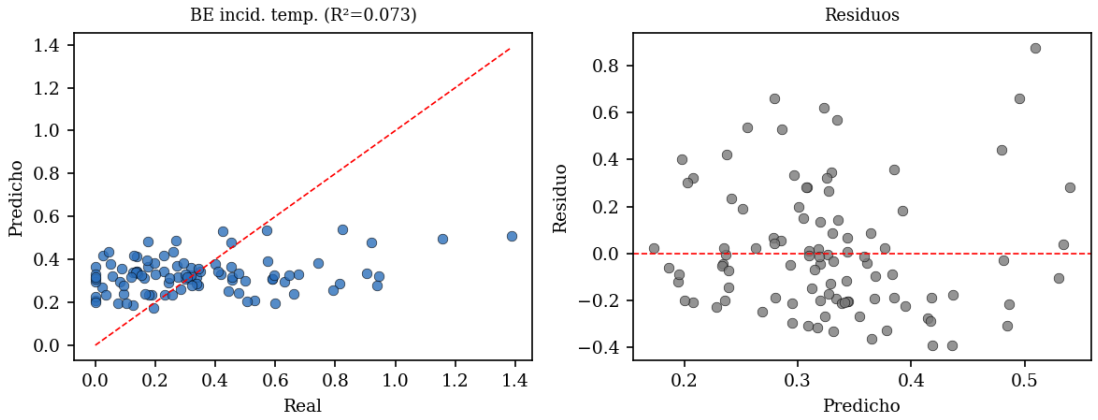


Figura 42

BE: Predicho vs. Real y Residuos (random forest)



6.2.4. *Comparación de Resultados según Modelos*

La evaluación continua de línea base evidenció que las formulaciones no lineales superan de forma consistente a la regularización paramétrica en los objetivos con señal predictiva válida, destacando el retraso contractual (BD) como el indicador de mejor ajuste global (R^2 = +0.172, MAE = 0.5024 con *random forest* y +0.165 con LASSO cuadrático). En el avance físico (BC) y el sobrecosto financiero (BA), *random forest* consolidó también el mayor desempeño con un R^2 de +0.145 (MAE = 0.0481) y +0.108 (MAE = 0.0702), respectivamente, aventajando a las variantes LASSO lineal y cuadrática al capturar de manera más eficiente las interacciones complejas entre el plazo inicial, la intensidad competitiva, la tipología sectorial y la gobernanza del proyecto.

Por el contrario, las variables endógenas ligadas a trámites de adicionales y prórrogas (BB y BE) colapsaron en la modelación continua debido a su marcada concentración de ceros legítimos y a la alta dispersión administrativa de su registro. A esa dificultad distribucional se suma una razón más de fondo: los adicionales y las ampliaciones son eventos contingentes que aparecen durante la ejecución, ligados a circunstancias particulares de cada obra con un alto componente de aleatoriedad.

6.2.5. *Interpretación*

El análisis conjunto de los predictores seleccionados permite caracterizar los mecanismos mediante los cuales las condiciones *ex-ante* del proyecto inducen desviaciones contractuales, destacando los siguientes factores determinantes:

- Plazo contractual, modalidad de licitación y tipo de contratista (BA): La magnitud de la duración programada actúa como el inductor dominante, donde plazos más extensos incrementan la exposición a reajustes y contingencias, mientras que el procedimiento de selección y la conformación del ejecutor modulan la capacidad operativa y financiera para absorber desvíos presupuestales.
- Sistema de contratación y tipología de obra (BD, BC, BB): El esquema contractual (precios unitarios vs. suma alzada) estructura la asignación de riesgos de metrados e imprevistos, al tiempo que la naturaleza constructiva, particularmente en proyectos viales y urbanos, concentra mayores interferencias físicas y de replanteo que impactan en el cumplimiento de cronogramas y la preservación del alcance original.
- Nivel de gobierno y localización territorial (BC, BB): Las disparidades de madurez técnica y capacidad administrativa entre niveles de gobierno condicionan la culminación de metas físicas, mientras que el entorno metropolitano introduce restricciones de interferencias de servicios que inciden en las modificaciones de obra.
- Tasa de pobreza distrital (BA, BD): Refleja un condicionante socioeconómico e institucional estructural donde distritos con menor capacidad de gestión suelen presentar mayor propensión a retrasos y sobrecostos, proveyendo una señal contextual estable y transferible a nuevas intervenciones.
- Contexto COVID-19 (BA, BD, BE): Representa un choque exógeno extraordinario de disrupciones en el suministro y suspensiones obligatorias; aunque explica un alto

porcentaje de la dispersión de plazos pasados, se instrumentaliza como una covariable de control temporal para evitar distorsiones en los factores permanentes del expediente.

6.3. Modelos de Clasificación de Riesgo y Alerta Temprana

6.3.1. *Objetivo de Modelos de Clasificación*

La sección anterior mostró que el valor exacto de la desviación es difícil de predecir con variables *ex-ante*, los R^2 son positivos pero modestos. Sin embargo, en la práctica, para la gestión pública lo accionable no es tanto anticipar el número preciso del sobrecosto o del retraso, sino identificar con antelación qué obras tienen alta probabilidad de terminar en un desempeño problemático. La finalidad de esta sección es reformular cada objetivo continuo como un problema de clasificación binaria de riesgo, que sirva de mecanismo de alerta temprana. El planteamiento es coherente con la evidencia de la literatura, que reconoce mayor viabilidad en clasificar el nivel de riesgo que en predecir el valor puntual de la desviación. Solo se llevan a esta fase los tres objetivos con señal de regresión aprovechable BA, BC y BD; en cambio, BB y BE quedan fuera desde el inicio por la debilidad de su señal, evidenciada en la sección anterior, ya que reportar un clasificador sin capacidad discriminante no aporta valor y podría inducir a error.

6.3.2. *Definición de las Clases de Riesgo*

Para cada objetivo se define una clase binaria de "alto riesgo" a partir de la distribución real de la variable. El criterio adoptado es el tercil de la propia distribución muestral, elegido por tres razones: es una regla objetiva y reproducible, no depende de umbrales normativos externos que varían entre entidades, y fija una prevalencia de riesgo del 34 % (33 de 98 obras) que evita el desbalance severo y hace comparables las métricas entre objetivos.

La dirección del recorte depende del significado de cada variable. En el sobrecosto (BA) y el retraso (BD), un valor alto es desfavorable, de modo que la clase de riesgo corresponde al tercil superior (las obras con mayor desviación). En el avance físico (BC) la

lógica se invierte, lo problemático es conservar un menor porcentaje del alcance original, por lo que la clase de riesgo es el tercil inferior. Solo se modelan BA, BC y BD.

La Tabla 30 traduce estos terciles a umbrales concretos sobre los valores reales, que constituyen la definición operativa de "obra en riesgo" para cada objetivo.

Tabla 30

Umbrales de Definición de la Clase de Riesgo

Objetivo	Regla	Umbral (valor real)	Interpretación del umbral	Obras en riesgo
BA (sobrecosto)	≥ tercil superior (P66.7)	0.1675	Sobrecosto ≥ 16.75 % del monto adjudicado	33 / 98 (34 %)
BC (avance)	≤ tercil inferior (P33.3)	0.9265	Avance ≤ 92.65 % del alcance original	33 / 98 (34 %)
BD (retraso)	≥ tercil superior (P66.7)	1.0444	Retraso ≥ 104.4 % del plazo inicial	33 / 98 (34 %)

6.3.3. *Definición de Hiperparámetros y Procedimiento*

a. Hiperparámetros

Se entrenan dos clasificadores por objetivo, una regresión logística regularizada, como referencia lineal interpretable, y un *random forest*, como modelo no lineal. La validación es la misma, validación cruzada con 8 repeticiones de *StratifiedKFold* de 5 pliegues, esta vez por la clase de riesgo para preservar su proporción en cada pliegue. La métrica principal es el área bajo la curva (AUC), que mide la capacidad de ordenar las obras por riesgo con independencia del umbral de corte, un AUC de 0.5 equivale al azar y 1.0 a la discriminación perfecta.

Respecto a la parametrización de los clasificadores, *random forest* mantiene los mismos hiperparámetros definidos en la regresión de línea base, mientras que para la regresión logística se consideraron los parámetros según la Tabla 31.

Tabla 31

Hiperparámetro de Regresión Logística

Parámetro	Valor	Sustento
penalty / C	L2 / 1.0	Regularización *Ridge* moderada (valores por defecto); controla la magnitud de los coeficientes con pocas observaciones y evita la inestabilidad numérica.
solver	lbfgs	Optimizador estándar y eficiente para la penalización L2 en problemas de tamaño pequeño.
max_iter	1000	Número máximo de iteraciones; garantiza la convergencia del optimizador.
Preprocesamiento	StandardScaler	Estandariza las variables (media 0, varianza 1); necesario para que la penalización L2 trate a todas las variables en la misma escala.

b. Código

La reformulación como clasificación de riesgo se desarrolló con scikit-learn, comparando dos algoritmos: *RandomForestRegressor* y *LogisticRegression* sobre las variables estandarizadas con StandardScaler. La evaluación usa *RepeatedStratifiedKFold* y *StratifiedKFold* para las particiones, y las métricas de desempeño *roc_auc_score*, *roc_curve* y *confusion_matrix*; pandas y numpy dan soporte.

El bloque central es la función cv_scores, que unifica la evaluación bajo un mismo esquema: repite ocho veces una validación cruzada estratificada de cinco pliegues (estratificada por la clase de riesgo, definida por el tercil de cada objetivo) y devuelve, por un lado, el AUC de cada repetición y, por otro, la puntuación fuera de pliegue (OOF) promediada por obra. Ese vector promediado alimenta las curvas ROC y el punto de operación. Para el avance (BC), cuyo riesgo es un avance bajo, se invierte el signo de la puntuación del *random forest* (que predice el avance); la probabilidad de la regresión logística no se invierte, porque ya modela directamente la clase de riesgo (Figura 43).

Figura 43

Cálculo del AUC para random forest Frente a Regresión Logística en JupyterLab

AUC por validación cruzada: Random Forest vs. regresión logística

Se comparan dos clasificadores. El Random Forest usa su predicción continua del objetivo como puntuación de riesgo; la regresión logística entrega la probabilidad de la clase. La función `cv_scores` ejecuta las 8 repeticiones y devuelve: el AUC de cada repetición (para la media y sd) y la puntuación OOF promediada (para la ROC).

```
def cv_scores(code, modelo):
    cols=SETS[code]; X=df[cols].values; y=df[TGT[code]].values; yb=clase(code)
    acc=np.zeros(len(y)); aucs=[]
    for rep in range(N_REP):
        skf=StratifiedKFold(5,shuffle=True,random_state=RNG+rep); oof=np.zeros(len(y))
        for tr,va in skf.split(X,yb):
            if modelo=='rf':
                m=RandomForestRegressor(max_features=min(4,len(cols)),**RF); m.fit(X[tr],winsor(y[tr],code))
                p=m.predict(X[va]); p=-p if code=='BC' else p
            else:
                sc=StandardScaler().fit(X[tr]); lg=LogisticRegression(max_iter=1000).fit(sc.transform(X[tr]),yb[tr])
                p=lg.predict_proba(sc.transform(X[va]))[:,1]   # la logistica ya modela la clase; NO se invierte
            oof[va]=p
        aucs.append(roc_auc_score(yb,oof)); acc+=oof
    return acc/N_REP, float(np.mean(aucs)), float(np.std(aucs))   # (oof_promediado, AUC_medio, sd)

STAT={cc:{m:cv_scores(cc,m) for m in ['rf','log']} for cc in ['BD','BC','BA']}
rows=[[cc, round(STAT[cc]['rf'][1],3), round(STAT[cc]['rf'][2],3), round(STAT[cc]['log'][1],3), round(STAT[cc]['log'][2],3)] for cc in ['BD','BC','BA']]
tab=pd.DataFrame(rows,columns=['Objetivo','AUC_RF','sd_RF','AUC_LOG','sd_LOG']); tab
```

El *random forest* se adopta como clasificador final; a partir de su puntuación OOF promediada se traza la curva ROC y se fija el punto de corte mediante el índice de Youden *(np.argmax(tpr − fpr))*, insumo con el que se construyen las matrices de confusión y las métricas operativas (sensibilidad y especificidad).

6.3.4. *Resultados de Clasificación*

La Tabla 32 y la Figura 44 resumen el área bajo la curva (AUC) de los modelos de clasificación. Adicionalmente, en la Figura 45 se muestran las curvas de característica operativa del receptor (ROC), las cuales grafican la sensibilidad frente a la tasa de falsos positivos a lo largo de todos los posibles umbrales de decisión continua. Esta trayectoria permite evaluar la capacidad del modelo para discriminar las observaciones entre perfiles de riesgo sin depender de un punto de corte arbitrario, donde la proximidad de la curva al vértice superior izquierdo denota una separabilidad diagnóstica superior frente a la diagonal de clasificación aleatoria.

Tabla 32

Desempeño de Clasificación de Riesgo

Objetivo	Clase de riesgo	AUC — *random forest*	AUC — Regresión logística
BD (retraso)	Tercil superior	0.761 ± 0.018	0.731 ± 0.027
BC (avance)	Tercil inferior	0.706 ± 0.013	0.670 ± 0.023
BA (sobrecosto)	Tercil superior	0.694 ± 0.024	0.651 ± 0.027

Figura 44

AUC por Objetivo y Modelo de Clasificación

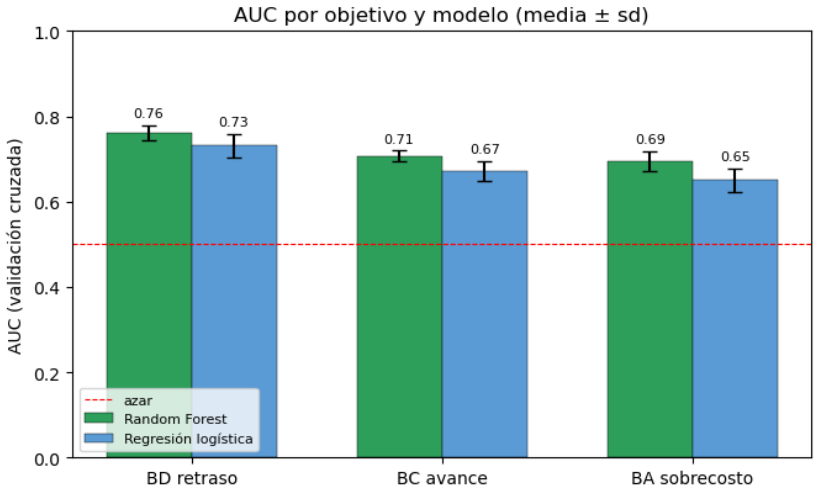
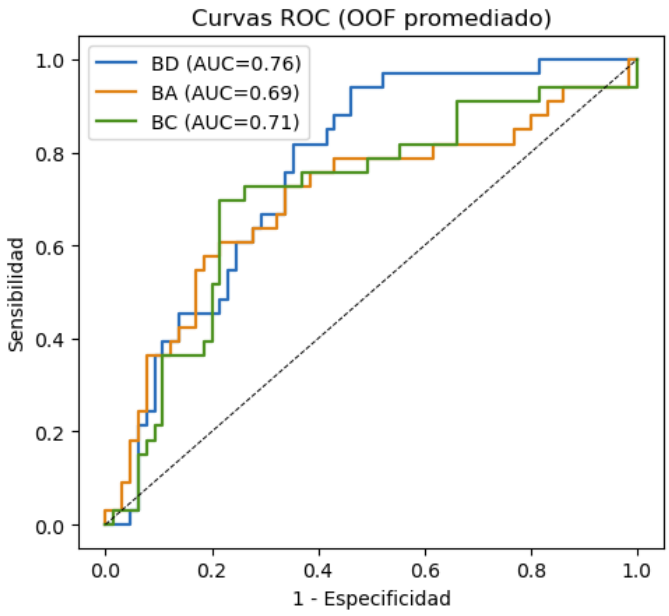


Figura 45

Curvas ROC de la Clasificación de Riesgo



El clasificador de retraso alcanza un AUC de 0.76 (bueno), y los de avance y sobrecosto 0.71 y 0.69 (moderados). Las desviaciones estándar son pequeñas (≤ 0.01–0.03), reflejo de la robustez del consenso. *Random forest* y la regresión logística rinden de forma comparable en los tres objetivos (BD 0.76 frente a 0.73; BC 0.71 frente a 0.67; BA 0.69 frente a 0.65), lo que indica que la frontera de riesgo es en buena medida lineal y que un modelo simple e interpretable resulta competitivo. Se adopta *random forest* como clasificador final por su desempeño ligeramente superior y consistente, con la regresión logística como alternativa transparente de respaldo.

6.3.5. *Calibración del Umbral de Corte (Índice de* Youden*)*

Para transformar las puntuaciones continuas de los modelos en decisiones binarias de alerta (riesgo vs. no riesgo), se calibra el umbral mediante el Índice de Youden (J):

$$J = Sensibilidad + Especificidad\text{-}1$$

Este índice mide la calidad de cada punto de corte, fija el umbral de corte c* como el valor que maximiza J. Se identifica el punto de corte sobre la curva ROC que maximiza la tasa de verdaderos positivos minimizando los falsos positivos, superando la arbitrariedad del umbral clásico de 0.50. De esta manera se establece una regla operativa, fijando el punto de corte por el índice de Youden; la Tabla 33 y la Figura 46 muestran el comportamiento de cada clasificador como regla de decisión.

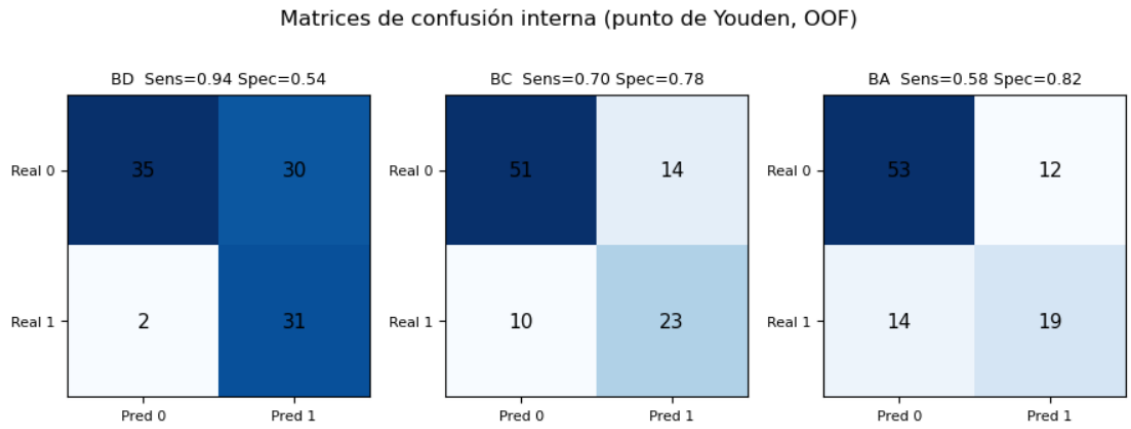
Tabla 33

Métricas Operativas en el Umbral de Corte

Objetivo	AUC	Sensibilidad	Especificidad	Índice de Youden	Umbral de Corte (ĉ *)
BD (retraso)	0.761	0.939	0.538	0.477	Retraso predicho ≥ 0.789
BC (avance)	0.706	0.697	0.785	0.482	Avance predicho ≤ 0.922
BA (sobrecosto)	0.694	0.576	0.815	0.391	Sobrecosto predicho ≥ 0.133

Figura 46

Matrices de Confusión en Umbral de Corte (Predicciones Fuera de Pliegue)



El clasificador de retraso (BD) prioriza la sensibilidad (0.94), detecta la gran mayoría de las obras que terminarán con alto retraso, a costa de algunas falsas alarmas (especificidad 0.54), un balance adecuado para una alerta temprana donde omitir un caso de riesgo es más costoso que revisarlo de más. El clasificador de sobrecosto (BA) también es específico (0.82), pero poco sensible (0.58), lo que significa que es un clasificador conservador que rara vez genera falsas alarmas, pero pueden pasar desapercibidos proyectos críticos. El de objetivo de avance (BC) equilibra mejor (sensibilidad 0.70, especificidad 0.79). Para una alerta temprana, priorizar la sensibilidad es razonable, ya que es preferible revisar de más que dejar pasar una obra de riesgo.

6.4. Validación Externa

6.4.1. *Objetivo y Procedimiento*

La validación externa no busca reentrenar el algoritmo, sino evaluar su capacidad de generalización fuera de la muestra (*out-of-sample*). Por eso, diez obras de otros departamentos bastan para comprobar si su poder discriminante se transfiere a nuevos contextos geográficos y administrativos. Esa escala, además, es consistente con la literatura especializada (por ejemplo, Avila Chumpisuca, 2019), donde es habitual validar modelos de riesgo sobre conjuntos de control acotados pero auditados con rigor.

Para evaluar la generalización del instrumento a otros contextos, se aplicaron los clasificadores finales a una muestra externa de diez obras públicas liquidadas de otros departamentos del sur del país provenientes de Cusco (4), Tacna (4) y Puno (2), las cuales no fueron utilizadas en el entrenamiento. La muestra registra todas las variables del expediente (sistema de contratación, tipo de contratista, número de postores, clasificación de suelo, calidad de fundación, material inadecuado, familia de suelo, tipología, nivel de gobierno y escenario) y sus tres objetivos (sobrecosto, avance y retraso). La muestra registra todas las variables del expediente técnico y contractual requeridas por el modelo, cuyos registros tabulares completos y valores originales no transformados se detallan en los Anexos 1 y 2.

Cada clasificador final (*random forest*) se entrenó sobre las 98 obras de Arequipa y se aplicó a las diez obras externas, codificadas de forma idéntica a la muestra de entrenamiento (monto en logaritmo natural, número de postores en log1p, calidad de fundación ordinal baja/media/alta = 0/1/2, tipologías y escenario en *one-hot*). Todos los campos provienen directamente del expediente, sin imputación. Conviene distinguir dos umbrales. El umbral de tercil opera sobre el valor real del resultado y define la clase de riesgo verdadera; se mantiene el corte calibrado en Arequipa (Tabla 30) para que las obras externas sean comparables con la muestra de entrenamiento. El umbral de corte de Youden, en cambio, opera sobre la puntuación predicha por el modelo y decide cuándo se emite la alerta. Como este segundo umbral depende de la escala de las predicciones, la validación externa se reporta bajo dos reglas: (A) reutilizando el umbral de corte (por índice de Youden) de Arequipa y (B) recalibrándolo localmente con las diez obras. Esto permite separar la capacidad de discriminación del modelo (AUC, que no depende del umbral) de su calibración (el punto de corte, que sí debe ajustarse por zona). Se reporta el área bajo la curva ROC (AUC) y, en el punto de corte, la sensibilidad y especificidad.

6.4.2. *Resultados de validación externa*

La Tabla 34 resume el desempeño. Se hace la distinción entre los resultados obtenidos utilizando (A) el umbral de corte de Arequipa y (B) el umbral de corte recalibrado localmente con las obras de la validación externa; la Figura 47 lo representa. Adicionalmente, se presenta en la Figura 48 el retraso predicho frente al real por obra y objetivo.

Tabla 34

Desempeño de los Clasificadores en la Validación Externa

Objetivo	Obras Críticas	AUC	Sens (A)	Spec (A)	Aciertos (A)	Sens (B)	Spec (B)	Aciertos (B)
BD (retraso)	2 / 10	0.969	1.00	0.12	3/10	1.00	0.88	9/10
BC (avance)	2 / 10	1.000	0.00	1.00	8/10	1.00	1.00	10/10
BA (sobrecosto)	4 / 10	0.583	0.50	0.50	5/10	0.50	0.83	7/10

Figura 47

AUC por Variable Objetivo

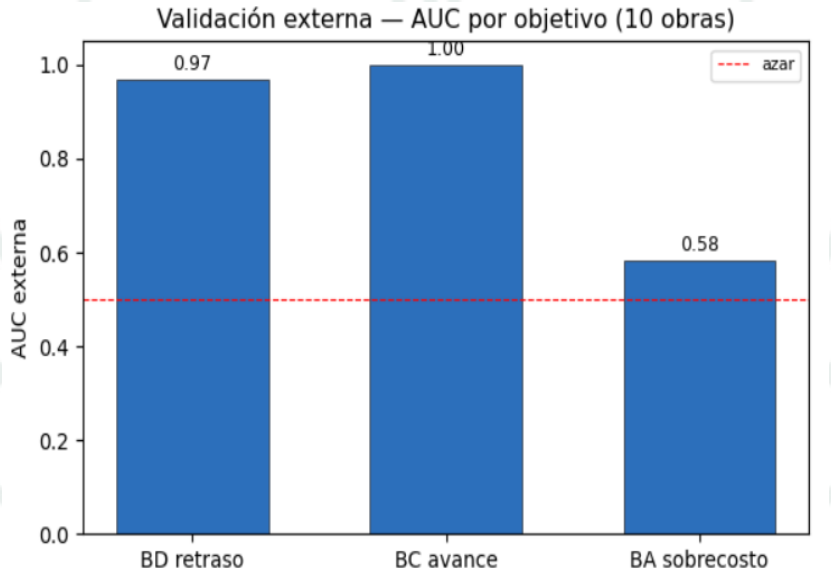
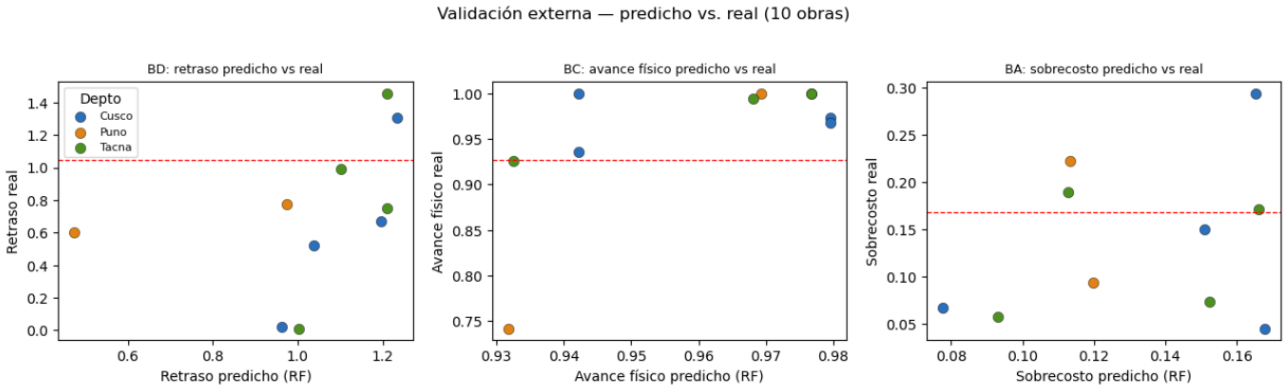


Figura 48

Retraso Predicho Frente al Real por Obra



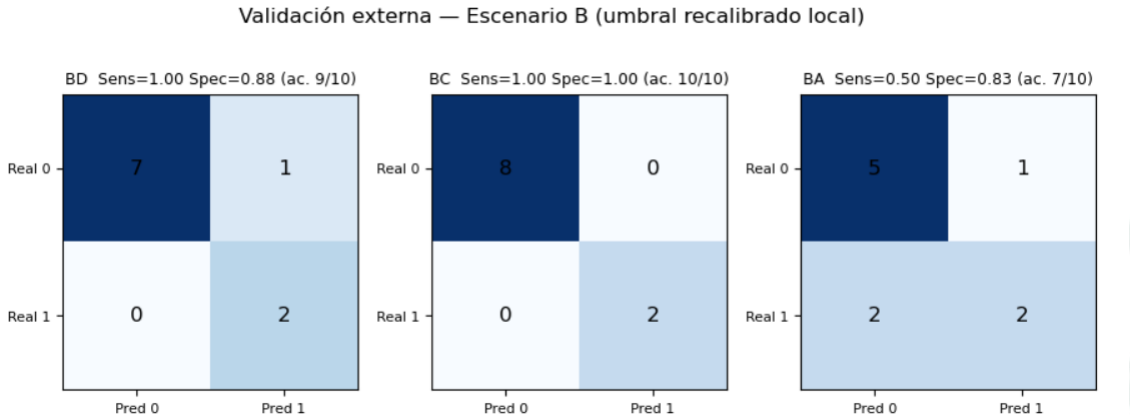
Con el umbral de Arequipa aplicado (A), la regla no transfiere. El retraso (BD) satura de falsas alarmas (especificidad 0.12) y en el avance (BC) no dispara ninguna alerta (sensibilidad 0.00). Al recalibrar el umbral localmente (B), el desempeño se recupera (retraso 9/10, avance 10/10), confirmando que lo que generaliza es la capacidad de ordenar por riesgo y que el punto de corte debe fijarse por zona. El clasificador de retraso (BD) discrimina de forma sobresaliente fuera de Arequipa (AUC 0.97), asignando las puntuaciones más altas a las dos obras con alto retraso real (obras con ítems 1 y 11). Sus predictores, como sistema de contratación, tipologías, modalidad y contexto, se transfieren a otros departamentos del sur. El de avance (BC) logra separación perfecta (AUC 1.00), resultado que debe leerse con la cautela propia de una muestra reducida con pocos casos positivos (dos de diez), pero que confirma su capacidad de ordenamiento. El de sobrecosto (BA) es el más débil (AUC 0.58), no se reconocieron dos obras críticas, este indicador es más cercano al azar y no transfiere de forma fiable, coherente con que el sobrecosto depende de factores de ejecución no observables *ex-ante*.

Para leer el desempeño con claridad conviene precisar que un "positivo" es una obra que sí presentó el problema (retraso alto, avance bajo o sobrecosto alto, según los umbrales). De las diez obras externas, dos tuvieron retraso alto real, dos tuvieron avance bajo real y cuatro

tuvieron sobrecosto alto real; el resto son negativos reales. El acierto del modelo se mide comparando la clase real con la clase predicha en el punto de corte según el índice de Youden, y se resume en las matrices de confusión de la Figura 49.

Figura 49

Matrices de Confusión de la Validación Externa por Variable Objetivo



La lectura es concreta, el clasificador de retraso (BD) acierta en 9 de 10 obras, detecta las dos con retraso real (ningún falso negativo) y solo marca una falsa alarma; el de avance (BC) acierta en las 10; y el de sobrecosto (BA) acierta en 7 de 10, captura dos de los cuatro casos reales, y con solamente una falsa alarma, coherente con su menor sensibilidad.

La Tabla 35 detalla obra por obra la clase real, la clase predicha por el modelo y si coincidieron. Se presenta con clases binarias para que el resultado de cada obra sea inequívoco.

Tabla 35

Validación Externa por Obra

Ítem	Depto.	BD (real → pred)	Res. BD	BA (real → pred)	Res. BA	BC (real → pred)	Res. BC
1	Cusco	1 → 1	✓	0 → 1	✗	0 → 0	✓
2	Cusco	0 → 0	✓	0 → 0	✓	0 → 0	✓
4	Cusco	0 → 0	✓	1 → 1	✓	0 → 0	✓
5	Cusco	0 → 0	✓	0 → 0	✓	0 → 0	✓
6	Puno	0 → 0	✓	0 → 0	✓	1 → 1	✓

8	Puno	0 → 0	✓	1 → 0	✗	0 → 0	✓
9	Tacna	0 → 1	✗	0 → 0	✓	0 → 0	✓
11	Tacna	1 → 1	✓	1 → 0	✗	0 → 0	✓
13	Tacna	0 → 0	✓	0 → 0	✓	1 → 1	✓
14	Tacna	0 → 0	✓	1 → 1	✓	0 → 0	✓

7. DISCUSIÓN DE RESULTADOS

- **Estructura de las variables y hallazgos del EDA**

El análisis exploratorio de las 98 obras dejó un resultado transversal que condiciona toda la interpretación posterior: la asociación entre cualquier predictora individual y una variable objetivo es débil. La correlación máxima observada fue de $|r| \approx 0.37$ (Pearson, entre el plazo inicial y el sobrecosto) y de $|\rho| \approx 0.27$ en Spearman. Este patrón no refleja una deficiencia de los datos, sino un rasgo ya documentado del fenómeno: los sobrecostos y los retrasos dependen en buena medida de factores de ejecución y de comportamiento (sesgo optimista, cambios de alcance, gestión contractual) que las variables técnicas del expediente no capturan.

Donde sí apareció señal fue en los efectos categóricos y no lineales. Las pruebas no paramétricas identificaron diferencias significativas del retraso (BD) según el contexto del efecto COVID-19 ($p < 0.001$) y del avance físico (BC) según la tipología de obra ($p = 0.0005$), además de asociaciones del sistema de contratación con el retraso, el avance, el sobrecosto y la incidencia temporal (prueba U de Mann-Whitney), y del tipo de contratista con el avance y el retraso. La información mutua confirmó que variables sin correlación lineal aparente, como número de postores, nivel de gobierno, tipología, aportan dependencia no lineal, lo que justificó conservarlas para los modelos basados en árboles. Estos hallazgos anticiparon los factores que la selección de variables retendría más adelante.

- **Arquetipos de desempeño de la obra pública**

El *clustering* mediante *K-means* sobre las tres dimensiones del desempeño (sobrecosto, retraso y avance físico) identificó tres arquetipos claramente diferenciados, con un coeficiente de silueta de 0.309 y diferencias significativas entre grupos en todas las variables (Kruskal-Wallis $p \leq 0.003$; tamaños de efecto Eta² de 0.31 a 0.57). La Tabla 36 resume su perfil.

Tabla 36

Perfil de los Tres Arquetipos de Obra

Arquetipo	n	Sobrecosto (BA)	Retraso (BD)	Avance (BC)	Incid. económica endógena (BB)	Incid. temporal endógena (BE)
Exitosa	47	7.2 %	50.3 %	95.9 %	0.051	0.202
Moderada	16	9.5 %	112.1 %	80.7 %	0.088	0.311
Colapsada	35	22.2 %	137.9 %	96.0 %	0.087	0.526

El grupo Exitoso (48 % de la muestra) reúne obras de bajo desvío integral, ejecutadas dentro de márgenes controlados. El grupo Colapsado (36 %) concentra el mayor sobrecosto, 22.2 % y el mayor retraso; es relevante que sus obras sí alcanzan el 96 % del alcance original y con la incidencia endógena de plazo más alta (BE = 0.526). El grupo Moderado (16 %) tiene características distintas, es el único con avance físico bajo (80.7 %), es decir, obras formalmente liquidadas cuyo alcance ejecutado quedó por debajo del previsto en el expediente original, con recorte o sustitución de metas durante la ejecución. La incidencia endógena de plazo crece de forma monótona de Exitosa a Colapsada, lo que indica que, a medida que una obra se acerca al arquetipo colapsado, una fracción creciente de su desvío proviene de causas endógenas y no de paralizaciones exógenas. En contraste, la incidencia endógena de los adicionales (BB) diferencia de manera sustancial al grupo de obras exitosas frente a los grupos de moderado y colapsado, los cuales registran magnitudes comparables entre sí.

- **Perfil de riesgo según las variables categóricas del proyecto**

Un aporte central de la aplicación de la técnica de *clustering* es cruzar los arquetipos de desempeño con las características categóricas del proyecto, para responder a qué tipos de obra concentran el riesgo. Para cada variable se calculó la tabla de contingencia con la etiqueta de clúster y se cuantificó la fuerza de la asociación con la V de Cramér y su significación con la prueba Chi-cuadrado (Tabla 37). Dado el tamaño muestral (n = 98), la potencia es limitada, por lo que se interpretan tanto la significación como la magnitud del efecto.

Tabla 37

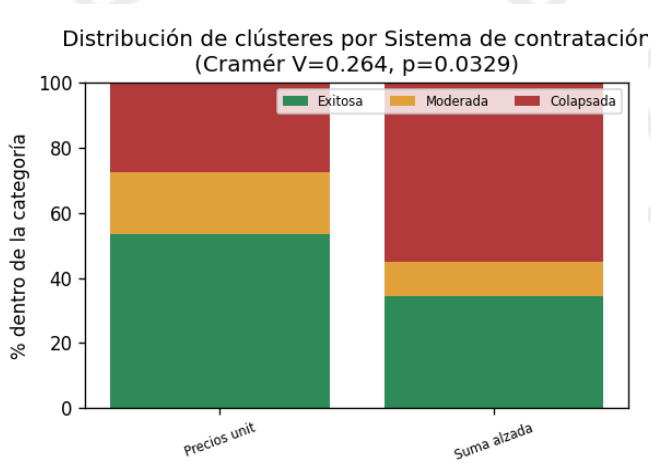
Asociación entre las Variables Categóricas y el Arquetipo de Desempeño de Obras

Variable	Chi²	gl	p	V de Cramér	Significativa
Sistema de contratación	6.83	2	0.033	0.264	Sí
Tipo de contratista	6.80	2	0.034	0.263	Sí
Nivel de gobierno	10.48	4	0.033	0.231	Sí
Tipología de obra	9.95	6	0.127	0.225	No
Escenario COVID	8.18	4	0.085	0.204	No
Calidad de fundación	2.14	4	0.711	0.104	No
Material inadecuado superficial	1.03	2	0.599	0.102	No
Naturaleza de intervención	0.04	2	0.983	0.019	No

Tres asociaciones resultan significativas y comparten una magnitud moderada (V de Cramér ≈ 0.26). El sistema de contratación es la más nítida: bajo suma alzada, el 55 % de las obras cae en el arquetipo Colapsado, frente a apenas el 28 % bajo precios unitarios (Figura 50). Esto es coherente con la teoría de riesgo contractual: en la suma alzada el contratista asume el riesgo de cantidades, y ante deficiencias del expediente el conflicto se traslada a adicionales y ampliaciones que terminan en colapso costo-plazo.

Figura 50

Distribución de Arquetipos por Sistema de Contratación



El tipo de contratista muestra el efecto esperado de la capacidad de gestión, las obras ejecutadas por consorcios se ubican en el arquetipo Exitoso en el 57 % de los casos, frente al 36 % de las ejecutadas por empresas individuales, que además concentran más obras en el grupo Moderado (Figura 51). El nivel de gobierno también discrimina, aunque con la cautela de los tamaños reducidos: las obras del gobierno nacional caen en el arquetipo Colapsado en el 78 % de los casos (9 obras), mientras que las de gobierno local, el 83 % de la muestra, se reparten de forma más equilibrada (30 % colapsadas). La distribución según el Nivel de Gobierno se muestra en la Figura 52.

Figura 51

Distribución de Arquetipos por Tipo de Contratista

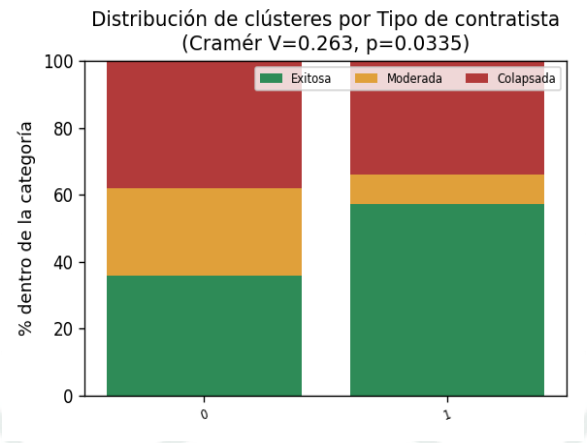
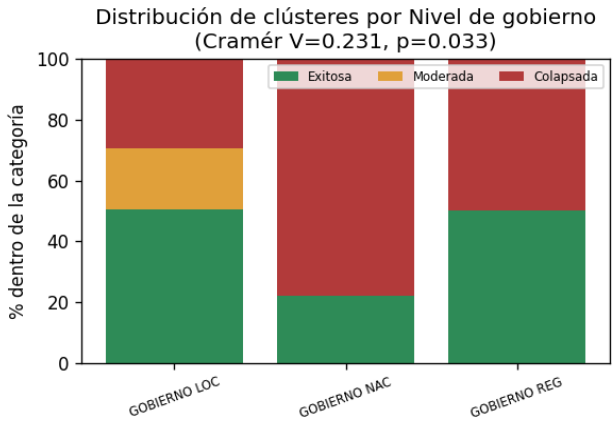


Figura 52

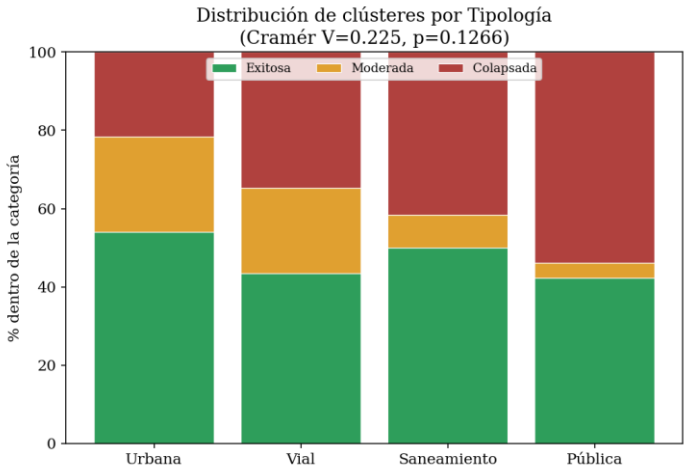
Distribución de Arquetipos por Nivel de Gobierno



Entre las asociaciones no significativas, pero de magnitud apreciable, la tipología de obra ordena el riesgo de forma nítida: la Infraestructura Pública concentra la mayor proporción de obras colapsadas (53.8 %) y la menor de moderadas (3.8 %); le siguen Saneamiento (41.7 % colapsada) y Vial (34.8 % colapsada); mientras que la Infraestructura Urbana muestra el mejor perfil, con 54.1 % de obras exitosas y solo 21.6 % colapsadas (Figura 53). El gradiente que va entre los distintos tipos de obra en la proporción de colapso es coherente con la mayor complejidad y variabilidad de metrados de las edificaciones institucionales frente a las obras urbanas.

Figura 53

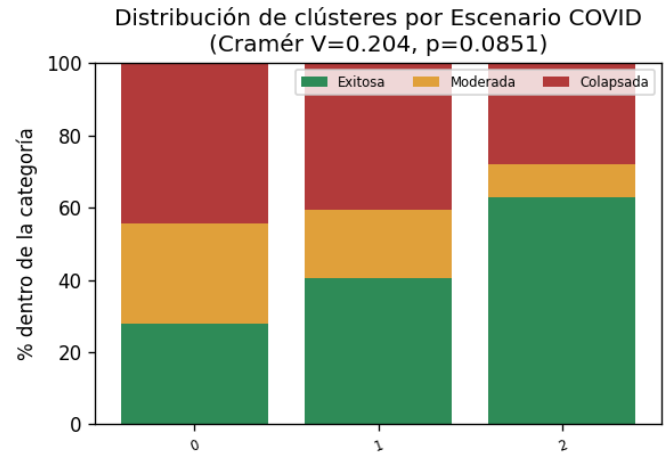
Distribución de Arquetipos por Tipología de Obra



El análisis por contexto del fenómeno COVID-19 muestra que las obras ejecutadas por completo en el periodo post-COVID tienen el mejor desempeño (63 % exitosas), mientras que las que empezaron antes de la pandemia y quedaron atrapadas en ella concentran la mayor proporción de colapso (44 %), lo que es coherente con el efecto de las paralizaciones sanitarias sobre el plazo (Figura 54). En cambio, las variables geotécnicas (calidad de fundación y material inadecuado) y la naturaleza de la intervención no mostraron asociación con el arquetipo; su influencia, si existe, actúa de forma más sutil y no lineal de lo que una tabla de contingencia puede captar.

Figura 54

Distribución de Arquetipos por Escenario COVID



De estos cruces se desprende una lectura de riesgo accionable: una obra de edificación institucional, contratada bajo suma alzada, ejecutada por una empresa individual y en el ámbito del gobierno nacional, reúne el perfil de mayor probabilidad de colapso costo-plazo. El perfil opuesto sería una obra urbana, a precios unitarios, ejecutada por consorcio, la cual tiende al desempeño exitoso.

- **Selección de variables y capacidad predictiva continua**

La selección de características no se hizo en un solo pase, sino de forma robusta. A partir del banco completo de 21 columnas se aplicó una selección por estabilidad: el procedimiento forward se repitió sobre múltiples particiones de validación cruzada, y por cada objetivo se retuvieron las variables elegidas en la mayoría de las corridas (umbral de consenso del 50 %), hasta llegar a un subconjunto de siete a ocho variables. Con una muestra pequeña y señales individuales débiles, este consenso sustituye un subconjunto inestable por uno reproducible.

Una decisión de codificación resultó determinante para la interpretación, en los modelos basados en árboles se conservaron todas las categorías, sin eliminar la de referencia. Debido a que los árboles no sufren colinealidad en el *one-hot encoding*. Mantener todas las

indicadoras permitió que la selección reconociera como factores de primer orden a categorías que un esquema con referencia habría dejado ocultas. Los subconjuntos de consenso confirman que los determinantes difieren entre objetivos, lo que valida el enfoque por variable dependiente:

Tabla 38

Factores Críticos por Variable Objetivo

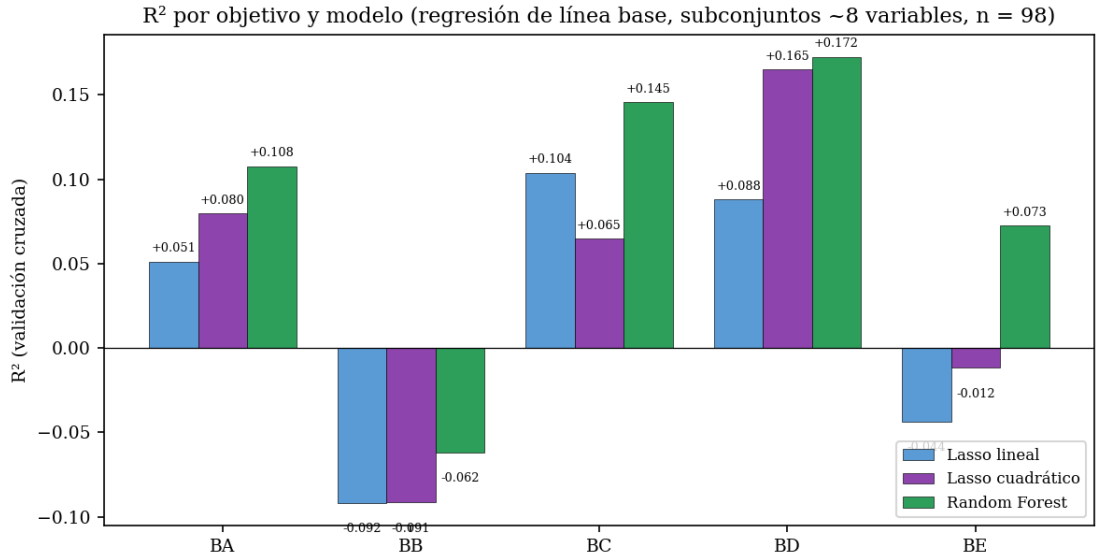
Variable objetivo	N.º	Factores críticos
Sobrecosto (BA)	8	Plazo inicial (seleccionado en el 100% de las corridas), modalidad de licitación, tipo de contratista, contexto de pandemia, tasa de pobreza distrital, tipología de saneamiento, material inadecuado e infraestructura pública.
Avance físico (BC)	7	Infraestructura pública, tipo de contratista, ámbito metropolitano, tipología urbana, tipología de saneamiento, tipología vial y nivel de gobierno local.
Retraso (BD)	8	Régimen pre-COVID, sistema de contratación, tipología urbana, tipología vial, contexto post-COVID, contexto de pandemia, modalidad de licitación y tasa de pobreza distrital.

El plazo inicial encabeza el sobrecosto de forma sistemática, coherente con la correlación observada en el EDA; el sistema de contratación y las tipologías estructuran el retraso y el avance; y el contexto COVID (interpretado como un control temporal). La convergencia entre estos subconjuntos y los cruces categóricos de la sección anterior refuerza la validez de los factores identificados.

En los objetivos con señal, el *random forest* igualó o superó al LASSO lineal y cuadrático (Figura 55), lo que confirma que la relación entre los predictores y los desvíos es predominantemente no lineal y de interacciones complejas. El techo de predictibilidad continua es consistente con la naturaleza del fenómeno y con la literatura, predecir el valor puntual del desvío con variables *ex-ante* tiene un límite estructural. Este resultado motivó la reformulación del problema.

Figura 55

R^2 por Objetivo y Modelo

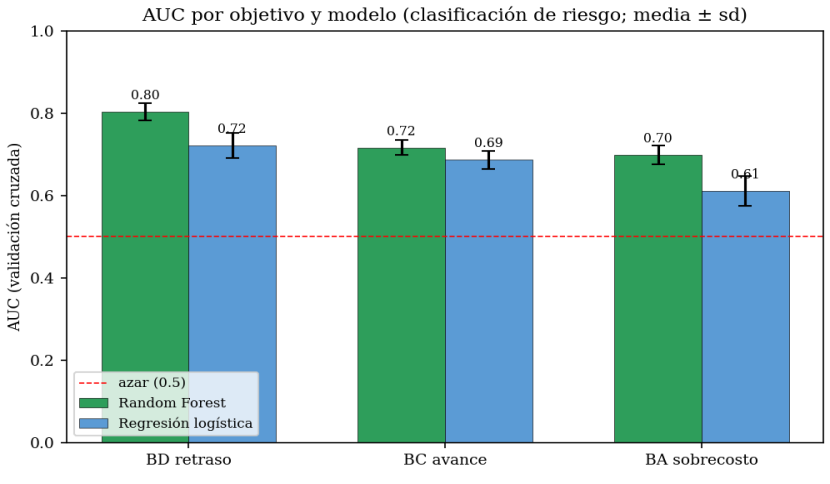


- **Clasificación de riesgo**

Al reformular cada objetivo como una clasificación binaria de riesgo con terciles, el desempeño recupera valor práctico. El clasificador de retraso alcanza un AUC de 0.76, el de avance 0.71 y el de sobrecosto 0.69, con desviaciones estándar pequeñas gracias al consenso (Figura 56). *Random forest* y la regresión logística rinden de forma comparable en los tres objetivos (BD 0.76 frente a 0.73; BC 0.71 frente a 0.67; BA 0.69 frente a 0.65), lo que indica que la frontera de riesgo es en buena medida lineal y que un modelo simple e interpretable es competitivo. Se adopta *random forest* como clasificador final por su desempeño ligeramente superior y consistente, con la regresión logística como alternativa transparente. La incidencia temporal endógena (BE) y la económica (BB) quedaron fuera del sistema por su débil señal.

Figura 56

AUC por Objetivo y Modelo



- **Validación externa**

El resultado principal es que la capacidad discriminante se transfiere al sur en el retraso (BD, AUC 0.97) y el avance físico (BC, AUC 1.00), además, permite distinguir las propiedades de discriminación y calibración. Para esos dos objetivos el modelo ordena el riesgo de forma efectiva, pero el umbral calibrado en Arequipa no se transfiere directamente, porque el punto de corte óptimo depende de la prevalencia y de la distribución de puntuaciones, que difieren entre regiones. Al recalibrar el punto de corte localmente, el desempeño se recupera (retraso 9/10, avance 10/10). El sobrecosto (BA) es la excepción, su AUC externo (0.58) es cercano al azar, de modo que ni su ordenamiento se transfiere de forma fiable, y ajustar el umbral solo mejora el punto de operación sin restaurar la discriminación. De este modo, lo que realmente generaliza es el poder del algoritmo para jerarquizar los proyectos según su nivel de vulnerabilidad, mientras que la frontera de decisión sugiere un ajuste contextual previo a su implementación en nuevas regiones.

En consecuencia, se recomienda usar el modelo como herramienta de ordenamiento de riesgo (priorización) más que como una regla de corte fija, y recalibrar el umbral de corte con una pequeña muestra local de cada región. Con esa recalibración, la evidencia externa respalda

el uso del clasificador de retraso como instrumento de alerta temprana transferible para el avance como apoyo sólido, y del de sobrecosto como referencia complementaria.

- **Herramienta recomendada**

Se recomienda la herramienta de "Semáforo de alerta temprana". El clasificador entrega una puntuación de riesgo que, comparada con el umbral de corte, sitúa a la obra en un semáforo de tres niveles: verde (trámite ordinario), ámbar (revisión reforzada) y rojo (revisión obligatoria y plan de mitigación). La tabla 39 resume la regla por objetivo con su sensibilidad, especificidad y con su umbral de corte.

Tabla 39

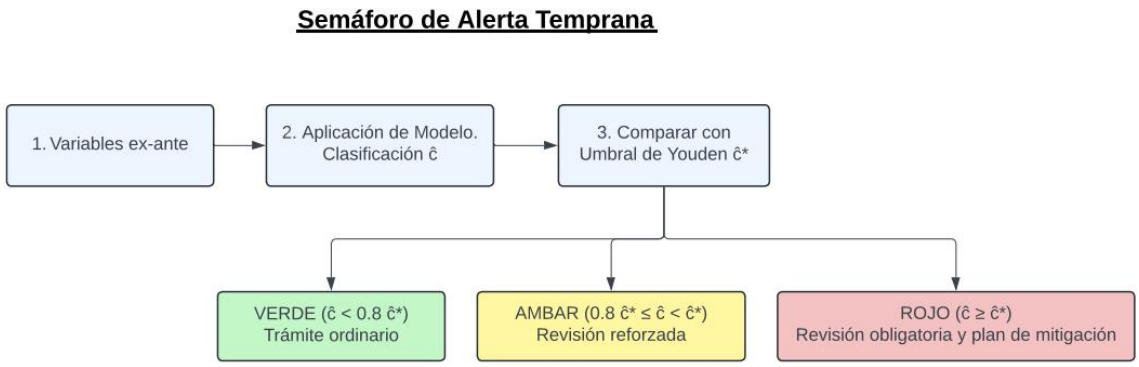
Semáforo de Alerta Temprana por Objetivo con Umbral de Corte

Objetivo	Umbral de corte ($\hat{c}^*$)	Verde ($\hat{c} < 0.8 \cdot \hat{c}^*$)	Ámbar ($0.8 \cdot \hat{c}^* \leq \hat{c} < \hat{c}^*$)	Rojo ($\hat{c} \geq \hat{c}^*$)	Sensibilidad	Especificidad
BD (retraso)	0.789	Riesgo bajo	Vigilancia	Alerta de alto retraso	0.94	0.54
BC (avance)	0.922	Riesgo bajo	Vigilancia	Alerta de recorte de alcance	0.70	0.78
BA (sobrecosto)	0.133	Riesgo bajo	Vigilancia	Alerta de sobrecosto	0.58	0.82

El flujo, previo a la ejecución consta de cuatro pasos, reunir las variables *ex-ante* del expediente, calcular la puntuación de riesgo $\hat{c}$; compararla con el umbral de corte $\hat{c}^*$, asignar el color y actuar según el color. Es importante mencionar que para el objetivo de alcance BC (donde menor es peor), los signos del semáforo se invertirían. Como el retraso se apoya en factores estructurales del expediente además del contexto, su regla es aplicable a obras futuras con menor dependencia de la coyuntura. Este proceso se grafica en la Figura 57.

Figura 57

Semáforo de Alerta Temprana



- **Discusión**

Los hallazgos muestran que los tres vértices del triángulo de hierro no operan por separado, sino que comparten determinantes administrativos. El sistema de contratación se asoció de forma significativa con el retraso, el avance físico, el sobrecosto y la incidencia temporal, y es el único factor que cruza los tres vértices a la vez; el tipo de contratista, en cambio, se vinculó con el avance y el retraso. La conexión más directa entre tiempo y costo aparece en la selección de variables: el plazo de ejecución inicial es el principal predictor del sobrecosto, de modo que la forma en que se dimensiona el cronograma en el expediente condiciona el resultado económico de la obra. El análisis de componentes principales lo confirma de forma cuantitativa: el retraso y el sobrecosto concentran casi todo el peso de la primera componente (juntos cerca del 97 %), lo que indica que ambos desvíos varían de manera acoplada, mientras que el avance físico define una segunda componente casi independiente. Esa misma estructura reaparece en la tipificación no supervisada: las obras del arquetipo colapsado reúnen a la vez más retraso, menos avance y más sobrecosto, mientras que las exitosas mantienen los tres vértices bajo control. El deterioro, entonces, es sobre todo un fenómeno conjunto de plazo y costo, al que el avance físico añade una dimensión propia de cumplimiento de metas.

El análisis supervisado evidenció un límite estructural fundamental: la predicción exacta de la magnitud del desvío resulta inviable utilizando exclusivamente información *ex-ante* ($R^2 \leq 0.17$). A diferencia de estudios previos como los de Gómez-Cabrera et al. (2025) o Restrepo y Rúa (2025), que reportan ajustes superiores al alimentar sus modelos con datos de ejecución en tiempo real (fuga de información), el presente estudio demuestra que el desvío en la obra pública peruana posee una alta carga de aleatoriedad ligada a contingencias de campo y gestión administrativa. Esto valida teóricamente la presencia del sesgo optimista descrito por Flyvbjerg et al. (2002) y justifica metodológicamente por qué el enfoque predictivo en la ingeniería civil debe transitar desde la regresión continua hacia la clasificación preventiva del riesgo.

La validación externa aportó el hallazgo de mayor impacto para la escalabilidad institucional de la herramienta: el ordenamiento del riesgo de retrasos y de alcance posee una naturaleza estructural compartida en el sur del país. Sin embargo, la limitada transferibilidad del modelo de sobrecosto evidencia que el comportamiento financiero depende de factores distintos a los considerados en Arequipa, como idiosincrasias administrativas locales, variaciones en los mercados de materiales regionales y la liquidez particular de los gobiernos de turno. En consecuencia, la discusión metodológica establece que, si bien el algoritmo generaliza eficazmente el diagnóstico de riesgo, el umbral operativo de decisión requiere obligatoriamente una recalibración del umbral de corte (índice de Youden) antes de su implementación como política pública.

En cuanto a la arquitectura de los modelos, el desempeño comparable entre la regresión logística y los ensambles no lineales (*random forest*) al momento de clasificar el riesgo demuestra una implicancia práctica relevante: la frontera de criticidad en las obras públicas es en gran medida lineal. Esto difiere de estudios internacionales enfocados en megaproyectos donde predominan interacciones algorítmicas de alta complejidad (Ashtari et al., 2022). Para

el contexto peruano de mediana envergadura, esto indica que herramientas parsimoniosas e interpretables son suficientes para capturar el riesgo, lo cual facilita la adopción tecnológica y la auditoría de estos modelos por parte de las entidades estatales.

El principal condicionante es el tamaño muestral (n = 98), que produce desviaciones estándar amplias en las métricas de validación cruzada y limita la potencia de las pruebas de asociación; por ello se privilegiaron modelos parsimoniosos, selección por estabilidad y validación repetida. La validación externa, aunque confirma la transferencia del ordenamiento de riesgo, se apoya en una muestra reducida (diez obras) con pocos casos positivos, por lo que sus métricas deben leerse como evidencia preliminar.

En comparación con los antecedentes internacionales, los R^2 alcanzados son inferiores a los reportados por estudios similares. La diferencia se explica por la naturaleza de las variables, diversos estudios modelan el costo final a partir de variables de obra ejecutada, mientras que este estudio se restringe deliberadamente a variables *ex-ante* disponibles antes de la ejecución, precisamente las útiles para una alerta temprana. El aporte de esta tesis, entonces, no reside en maximizar el R^2 sino en identificar de forma sistemática los factores de riesgo por objetivo y construir un sistema de clasificación aplicable a la gestión preventiva sobre datos reales de obra pública peruana.

8. CONCLUSIONES Y RECOMENDACIONES

8.1. Conclusiones

Primera, el objetivo general se cumplió: se desarrollaron y validaron modelos de *machine learning*, entrenados con datos reales de obras públicas peruanas, capaces de anticipar el riesgo de sobrecostos y retrasos a partir de factores previos a la ejecución, con la región de Arequipa como caso de estudio. La investigación caracterizó cómo y por qué se producen los sobrecostos y los retrasos, y demostró que su riesgo puede anticiparse antes de iniciar la construcción. La contribución central es un cambio de enfoque de ingeniería: el monto exacto del desvío no puede estimarse con la información del expediente ($R^2 \leq 0.17$), porque nace en gran medida de decisiones y eventos de la ejecución, como los adicionales y las ampliaciones; en cambio, el nivel de riesgo sí resulta anticipable con utilidad práctica (AUC de 0.69 a 0.76). Este resultado, coherente con antecedentes que reportan mayor viabilidad al clasificar el riesgo que al estimarlo de forma exacta (Aung et al., 2023; Timalsina et al., 2025), traslada el aporte desde la predicción numérica hacia una herramienta de control preventivo aplicable en la etapa previa a la licitación. Con ello, la hipótesis general se confirma en su versión operativamente relevante, la que se refiere a la clasificación del riesgo y no a la predicción del valor puntual.

Segunda, los desvíos no son aleatorios: se explican por factores de gestión y contratación identificables y, en su mayoría, controlables. El determinante más nítido y accionable es el esquema contractual: bajo suma alzada cerca del 55 % de las obras colapsa en costo y plazo, frente al 28 % bajo precios unitarios; a ello se suman el tipo de contratista, pues los consorcios tienden al desempeño exitoso y las empresas individuales al colapso, y el nivel de gobierno. Los factores determinantes difieren según el objetivo: el sobrecosto depende del plazo inicial, la modalidad de licitación y el tipo de contratista; el retraso, del sistema de contratación, la tipología y el contexto de la pandemia; y el avance, de la tipología, el tipo de contratista y el nivel de gobierno. La causa raíz es mayoritariamente endógena, ya que la

incidencia de adicionales y ampliaciones crece hacia el arquetipo colapsado; esto apunta a deficiencias del expediente técnico más que a factores externos, en línea con el diagnóstico de la Contraloría General de la República. Además, el sobrecosto y el retraso se comportan de forma acoplada, mientras que el avance físico constituye una dimensión propia, y la señal de estos factores no reside en correlaciones lineales fuertes ($|r| \leq 0.37$) sino en efectos categóricos y no lineales.

Tercera, se desarrollaron y entrenaron los modelos supervisados. En la predicción continua, el *random forest* igualó o superó a las alternativas lineales, con su mejor ajuste en el retraso ($R^2 = 0.172$), seguido del avance (0.145) y el sobrecosto (0.108). Las incidencias por adicionales (BB y BE) no alcanzaron señal modelable y se conservaron solo con carácter diagnóstico, por su fuerte concentración de ceros y porque dependen de eventos contingentes de la ejecución, como adicionales y ampliaciones, que las variables *ex-ante* no logran anticipar. En la clasificación de riesgo, los tres objetivos resultaron clasificables con un AUC de utilidad práctica (0.69 a 0.76). El *random forest* fue ligeramente superior y más consistente que la regresión logística en los tres objetivos (por ejemplo, AUC de BD 0.76 frente a 0.73), por lo que se adoptó como modelo final. Que la regresión logística rinda casi igual indica que la frontera de riesgo es en buena medida lineal; por eso la hipótesis sobre la superioridad de los algoritmos de *machine learning* se confirma de forma parcial: el *random forest* se impone por un margen estrecho y el modelo lineal queda como alternativa transparente de respaldo.

Cuarta, el aprendizaje no supervisado identificó tres arquetipos de desempeño bien diferenciados: Exitosa (48 % de la muestra, con bajo desvío integral), Colapsada (36 %, con sobrecosto y retraso elevados pese a completar el alcance) y Moderada (16 %, con recorte del alcance y desvíos intermedios). El cruce de estos arquetipos con las características del proyecto reveló asociaciones significativas: el sistema de contratación por suma alzada, la ejecución por empresa individual y el nivel de gobierno nacional concentran el mayor riesgo de colapso

costo-plazo, mientras que las obras urbanas, a precios unitarios y ejecutadas por consorcios, tienden al desempeño exitoso. Esto confirma la hipótesis sobre el valor del análisis no supervisado, tanto por la interpretabilidad de los arquetipos como por su función instrumental al estratificar la validación cruzada.

Quinta, se evaluó el desempeño en un contexto real con una validación externa sobre diez obras de Cusco, Puno y Tacna. Lo que se transfiere es la capacidad de ordenar el riesgo, es decir, la discriminación, no el umbral de decisión. El retraso (AUC 0.97) y el avance (1.00) conservaron un ordenamiento sólido, mientras que el sobrecosto (0.58) quedó cerca del azar y no llegó a transferirse de manera fiable. En todos los casos, el umbral de corte calibrado en Arequipa debe reajustarse por zona antes de emitir alertas; al recalibrarlo localmente, el retraso acertó en 9 de 10 obras y el avance en 10 de 10. Sobre esta base se formuló una herramienta de gestión preventiva, el Semáforo de Alerta Temprana, para su aplicación en la gestión pública orientada a anticipar sobrecostos y retrasos en la obra pública.

8.2. Recomendaciones

Para la gestión pública se recomienda incorporar un tamizaje de riesgo previo a los procesos de licitación. Se recomienda aplicar la herramienta "Semáforo de alerta temprana" sobre las variables *ex-ante* del expediente técnico. Las obras señaladas en rojo deberían pasar por una revisión reforzada del expediente y un plan de mitigación documentado antes de la adjudicación, priorizando el control preventivo donde el modelo anticipa mayor riesgo.

Atender los factores de riesgo controlables. Dado que el sistema de contratación por suma alzada, la baja competencia y la ejecución por empresas individuales se asocian a peor desempeño, se recomienda reforzar la calidad del expediente técnico en obras bajo suma alzada, promover la concurrencia efectiva de postores y ponderar la capacidad de gestión del contratista en la evaluación. La fuerte incidencia endógena de plazo en las obras colapsadas confirma que buena parte del desvío es prevenible con mejores expedientes.

Fortalecer la trazabilidad de la información. La principal restricción del estudio fue la disponibilidad y completitud de la información oficial. Se recomienda a las entidades mantener actualizados los portales abiertos de Infobras, el SEACE y los repositorios de actas y liquidaciones, condición necesaria para escalar este tipo de herramientas predictivas.

Para una investigación futura, se recomienda ampliar la muestra y consolidar la validación externa iniciada en este estudio, lo que mejoraría la potencia estadística y permitiría calibrar umbrales de corte locales por región; este punto es especialmente crítico para el sobrecosto, cuyo ordenamiento no llegó a transferirse y requiere más datos antes de usarse como regla fuera de Arequipa. Además, la inclusión de variables de ejecución, como cambios de diseño, desempeño real del contratista e hitos de avance, las cuales solo se conocen una vez iniciada la obra, podría elevar el techo de la predicción continua. Finalmente, se sugiere explorar la integración del sistema de alerta temprana en las plataformas de gestión pública, evaluando su factibilidad operativa y su impacto en la toma de decisiones.

9. REFERENCIAS

AACE International. (2011). Recommended Practice No. 29R-03: Forensic schedule analysis (Rev. 25 de abril de 2011). Association for the Advancement of Cost Engineering International.

Abdelalim, A. M., Salem, M., Salem, M., Al-Adwani, M., & Tantawy, M. (2025). Analyzing cost overrun risks in construction projects: A multi-stakeholder perspective using fuzzy group decision-making and K-means *clustering*. Buildings, 15(3), 447. https://doi.org/10.3390/buildings15030447

Abel, G. B. J. J., Benito, P. L. L., Rocio, V. B. L., & Jaime, R. C. Y. (2024). Challenges and opportunities of digital transformation and artificial intelligence in sustainable construction in Peru. Proceedings of the LACCEI International Multi-Conference for Engineering, Education and Technology. https://doi.org/10.18687/LEIRD2024.1.1.462

Abramov, M. (2024). Building with brains: How AI and *machine learning* are transforming construction. Keymakr. https://keymakr.com/blog/building-with-brains-how-ai-and-machine-learning-are-transforming-construction/

Afana, O., Al Zubaidi, R., Abu Dabous, S., & Ibrahim, F. (2024). Categories and factors of cost overrun in construction projects: A systematic review. Engineering, Technology & Applied Science Research, 14(6), 18330–18347. https://doi.org/10.48084/etasr.9006

Al-Bataineh, F., Khatatbeh, A. A., & Alzubi, Y. (2024). Unsupervised *machine learning* for identifying key risk factors contributing to construction delays. Organization, Technology and Management in Construction, 16(1), 170–185. https://doi.org/10.2478/otmcj-2024-0014

Al-Hazim, N., Salem, Z. A., & Ahmad, H. (2017). Delay and cost overrun in infrastructure projects in Jordan. Procedia Engineering, 182, 18–24. https://doi.org/10.1016/j.proeng.2017.03.059

Allahaim, F. S. A. (2015). An empirical classification of cost overrun in infrastructure projects by using cluster analysis. Proceedings of the Utzon Symposium, UNSW Australia.

Andrić, J. M., Lin, S., Cheng, Y., & Sun, B. (2024). Determining cost and causes of overruns in infrastructure projects in South Asia. Sustainability, 16(24), 11159. https://doi.org/10.3390/su162411159

Andrić, J. M., Lin, S., Cheng, Y., & Sun, B. (2025). Investigating the possible regression functions for modelling delays in infrastructure projects in South Asia. KSCE Journal of Civil Engineering, 29(9), 100209. https://doi.org/10.1016/j.kscej.2025.100209

Andrić, J. M., Mahamadu, A.-M., Wang, J., Zou, P. X. W., & Zhong, R. (2019). The cost performance and causes of overruns in infrastructure development projects in Asia. Journal of Civil Engineering and Management, 25(3), 203–214. https://doi.org/10.3846/jcem.2019.8646

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), 1660. https://doi.org/10.3390/buildings12101660

Asiedu, R., & Ameyaw, C. (2020). A system dynamics approach to conceptualise causes of cost overrun of construction projects in developing countries. International Journal of Building Pathology and Adaptation, 39(2), 1–22. https://doi.org/10.1108/IJBPA-05-2020-0043

Aung, T., Liana, S. R., Htet, A., & Bhaumik, A. (2023). Using *machine learning* to predict cost overruns in construction projects. Journal of Technology Innovations and Energy, 2(2), 1–7.

Avila Chumpisuca, J. E. (2019). Aplicación de la simulación Monte Carlo para el control de ampliaciones de plazo y adicionales en obras públicas [Tesis de licenciatura, Pontificia Universidad Católica del Perú].

AWS. (2025). What's the difference between structured data and unstructured data? Amazon Web Services. https://aws.amazon.com/es/compare/the-difference-between-structured-data-and-unstructured-data/

Boamah, F. A., Jin, X., Senaratne, S., & Perera, S. (2025). AI-driven risk identification model for infrastructure project: Utilising past project data. Expert Systems with Applications, 283, 127891. https://doi.org/10.1016/j.eswa.2025.127891

Cawley, G. C., & Talbot, N. L. C. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. Journal of *Machine learning* Research, 11, 2079–2107.

Chang Hidalgo, H. M. (2023). Comparación de técnicas de estimación basadas en *machine learning* para predecir costos en los planes de adquisiciones de las entidades públicas del Perú [Tesis de maestría]. https://hdl.handle.net/20.500.12802/10566

Cobrief. (2025). Suspension of work: Overview, definition, and example. https://www.cobrief.app/resources/legal-glossary/suspension-of-work-overview-definition-and-example/

Codefinity. (s. f.). Sequential backward and forward selection [Ilustración]. Codefinity. https://codefinity.com/blog/Sequential-Backward-and-Forward-Selection

Coffie, G. H., & Cudjoe, S. K. F. (2023). Toward predictive modelling of construction cost overruns using support vector machine techniques. Cogent Engineering, 10(2), 2269656. https://doi.org/10.1080/23311916.2023.2269656

Consejo Privado de Competitividad. (2025a). Más de 12 mil obras en Perú son más caras que lo previsto al inicio. Gestión. https://gestion.pe/economia/

Consejo Privado de Competitividad. (2025b). Proyectos cuestan en promedio más del doble que lo previsto al inicio. https://www.compite.pe/noticia/

Contraloría General de la República del Perú. (2025). Informe de obras públicas paralizadas en el territorio nacional a marzo 2025 (N.° 002-2025-CG/SESNC). https://www.gob.pe/institucion/contraloria/

Elbashbishy, T. S., Hosny, O. A., Waly, A. F., & Dorra, E. M. (2022). Assessing the impact of construction risks on cost overruns: A risk path simulation–driven approach. Journal of Management in Engineering, 38(6), 04022058. https://doi.org/10.1061/(ASCE)ME.1943-5479.0001090

Escuela de Gestión Pública de la Universidad del Pacífico. (2025). Brecha de infraestructura suma US$ 159,5 mil millones. https://www.up.edu.pe/egp/noticias/brecha-infraestructura-afin-egp/

Flyvbjerg, B., Holm, M. S., & Buhl, S. (2002). Underestimating costs in public works projects: Error or lie? Journal of the American Planning Association, 68(3), 279–295. https://doi.org/10.1080/01944360208976273

Flyvbjerg, B., Skamris Holm, M. K., & Buhl, S. L. (2003). How common and how large are cost overruns in transport infrastructure projects? Transport Reviews, 23(1), 71–88. https://doi.org/10.1080/01441640309904

Gado, N. G. (2024). AI revolutionizes construction management: Building smarter, safer, and efficiently addressing industry challenges. Engineering Research Journal, 183(3), 330–344. https://doi.org/10.21608/erj.2024.299920.1062

Gómez-Cabrera, A., León, L. C., Lopez, M. L., & Torres, A. (2025). Analyzing time and cost deviations in educational infrastructure projects: A data-driven approach using Colombia's public data platform. Buildings, 15(19), 3505. https://doi.org/10.3390/buildings15193505

Graphite Note. (2024). How much data do you need for *machine learning*. https://graphite-note.com/how-much-data-is-needed-for-machine-learning/

Hernández-Sampieri, R., & Mendoza, C. (2020). Metodología de la investigación: Las rutas cuantitativa, cualitativa y mixta. McGraw-Hill.

IBM. (2021). What is random forest? https://www.ibm.com/think/topics/random-forest

IBM. (2023). What is a *machine learning* algorithm? https://www.ibm.com/es-es/think/topics/machine-learning-algorithms

Instituto Nacional de Estadística e Informática. (2018). Censos Nacionales 2017: XII de Población, VII de Vivienda y III de Comunidades Indígenas. Arequipa: Resultados definitivos (Tomo I). https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1551/04TOMO_01.pdf

Instituto Nacional de Estadística e Informática. (2018). Mapa de pobreza monetaria provincial y distrital 2018. https://www.inei.gob.pe/media/MenuRecursivo/publicaciones_digitales/Est/Lib1718/Libro.pdf

Julca-Varas, C., García-Segura, T., Torres-Machi, C., & Montalbán-Domingo, L. (2025). Primary risks for delays and cost overruns in the construction of drinking water

supply and sanitation projects in Peru. Journal of Management in Engineering, 41(3), 04025008. https://doi.org/10.1061/JMENEA.MEENG-6496

Khursheed, S., Sharma, S., Paul, V. K., Al-Farouni, M., & Yoldosheva, C. (2024). Forecasting delays and budget overruns resulting from material-related factors. E3S Web of Conferences, 563, 02045. https://doi.org/10.1051/e3sconf/202456302045

Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. Journal of Cheminformatics, 6, 10. https://doi.org/10.1186/1758-2946-6-10

Larsen, J. K., Shen, G. Q., Lindhard, S. M., & Brunoe, T. D. (2016). Factors affecting schedule delay, cost overrun, and quality level in public construction projects. Journal of Management in Engineering, 32(1), 04015032. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000391

Ley N.° 27785, Ley Orgánica del Sistema Nacional de Control y de la Contraloría General de la República. (2002, 23 de julio). Diario Oficial El Peruano.

Ley N.° 27806, Ley de Transparencia y Acceso a la Información Pública. (2002, 2 de agosto). Diario Oficial El Peruano.

Ley N.° 30225, Ley de Contrataciones del Estado. (2014). Diario Oficial El Peruano.

Ley N.° 32069, Ley de Contrataciones del Estado. (2024). Diario Oficial El Peruano.

Maldonado de la Torre, L. (2025). Introducción a *Machine learning* en Ingeniería Estructural. INESA Tech. https://www.inesa-tech.com/blog/introduccion-machine-learning/

MathWorks. (2025). What is supervised learning? https://www.mathworks.com/discovery/supervised-learning.html

Nassar, K., & Elbisy, M. S. (2024). A *machine learning* approach to predict time delays in marine construction projects. Engineering, Technology & Applied Science Research, 14(5), 16125–16134. https://doi.org/10.48084/etasr.8173

Nwosu, O. C. V. (2023). How artificial intelligence influences project management. https://doi.org/10.51483/IJAIML.3.1.2023.19-34

Oh, J., Touran, A., D'Angelo, D., Clark, T., Fisher, C., Gaskins, C., & Ashuri, B. (2025). Machine learning–enhanced recurrent event modeling for change order recurrence in highway construction. Journal of Management in Engineering, 41(5), 04025032. https://doi.org/10.1061/JMENEA.MEENG-6583

Project Management Institute. (2021). A guide to the project management body of knowledge (PMBOK guide) (7th ed.). Project Management Institute.

PythonGeeks. (s. f.). Clustering in machine learning [Ilustración]. PythonGeeks. https://pythongeeks.org/clustering-in-machine-learning/

Restrepo Ramírez, A. F., & Rúa Machado, C. A. (2025). Predicting delays and cost overruns in construction projects: A *machine learning* approach. Revista EIA, 22(44), Reia4415. https://doi.org/10.24050/reia.v22i44.1861

Sahu, P., Bera, D. K., & Parhi, P. K. (2024). Analyzing the impact of construction delays on disputes in India: A statistical and *machine learning* approach. Journal of Mechanics of Continua and Mathematical Sciences, 19(9), 24–34. https://doi.org/10.26782/jmcms.2024.09.00004

Singh, D. (2024). Application of *machine learning* in civil engineering: Review. Advancements in Civil Engineering & Technology, 6. https://doi.org/10.31031/ACET.2024.06.000639

Timalsina, S., Zhang, C., Quapp, U., Demirkesen, S., & Olorunnishola, A. (2025). Machine learning–based predictive model for dispute occurrence and resolution strategies in pipeline projects. Journal of Legal Affairs and Dispute Resolution in Engineering and Construction, 17(3), 04525036. https://doi.org/10.1061/JLADAH.LADR-1334

Tochaiwat, K., & Budda, A. (2026). Construction contract price prediction model for government buildings using a deep learning technique: A study from Thailand. Buildings, 16(2), 539. https://doi.org/10.3390/buildings16020539

Traverso-Frisancho, A., & Romero Alva, V. (2023). Application of BIM methodology in public and private electricity and telecommunications projects in Peru. International Journal of Engineering Trends and Technology, 71(7), 67–74. https://doi.org/10.14445/22315381/IJETT-V71I7P207

Uddin, S., Ong, S., & Lu, H. (2022). Machine learning in project analytics: A data-driven framework and case study. Scientific Reports, 12, 15252. https://doi.org/10.1038/s41598-022-19728-x

Van Tam, N., & Toan, N. (2021). Research trends on *machine learning* in construction management: A scientometric analysis. Journal of Applied Science and Technology Trends, 2(3), 96–104. https://doi.org/10.38094/jastt203105

Xu, Y., Zhou, Y., Sekula, P., & Ding, L. (2021). Machine learning in construction: From shallow to deep learning. Developments in the Built Environment, 6, 100045. https://doi.org/10.1016/j.dibe.2021.100045

Zhong, C., & Zhang, S. (2025). Schedule risk analysis of prefabricated building projects based on DEMATEL-ISM and Bayesian networks. Buildings, 15(3), 508. https://doi.org/10.3390/buildings15030508

ANEXOS

Anexo 1

Base de datos sin transformar empleada en la validación externa: variables predictoras

Item	cod_Infobras	Nomenclatura	Monto_adjudicado	Plazo_ejecucion inicial	Naturaleza_de_intervencion	Tipologia_de_obra	Sistema_contratacion	Modalidad_seleccion	Ubicación_metropolitano	Nivel_gobierno	Escenario_macro_covid	Site_tasa_pobreza	Competitividad_ajuste_oferta	Num_Postores_Efectivos	Tipo_Contratista	Familia_SUCS	Calidad_fundacion	Material_inadecuado_superficial
1	82727	LP-SM-5-2018-MDK/CS-1	10542732	360	Intervención Preexistente	Infraestructura Vial y Transporte	Precios Unitarios	LP	0	Local	0	25%	0%	1	1	Fino	baja	0
2	83604	AS-SM-40-2018-MDP/CS-2	8698210.48	300	Intervención Preexistente	Infraestructura Pública	Suma Alzada	AS y Otros	0	Local	0	31%	0%	3	1	Fino	baja	0
4	143200	LP-SM-5-2020-MDP/CS-1	7127009.12	300	Obra Nueva	Infraestructura Pública	Suma Alzada	LP	0	Local	1	31%	0%	2	1	Fino	baja	0
5	172992	LP-SM-14-2022-MDUA/CS-1	7128855.22	210	Intervención Preexistente	Infraestructura Vial y Transporte	Suma Alzada	LP	0	Local	2	31%	0%	1	1	Fino	media	0
6	78607	LP-SM-1-2021-MDU/CS-1	6676959.12	180	Intervención Preexistente	Saneamiento e Hidraulica	Precios Unitarios	LP	0	Local	2	24%	0%	1	0	Arena	media	1
8	168117	LP-SM-3-2022-UNAJ/CS-1	5896676.42	180	Intervención Preexistente	Infraestructura Pública	Suma Alzada	LP	0	Nacional	2	22%	10%	1	0	Arena	baja	0
9	68501	LP-SM-5-2019-MIGRACIONES-1	5171968.23	210	Intervención Preexistente	Infraestructura Pública	Suma Alzada	LP	1	Nacional	0	34%	10%	36	1	Grava	alta	0
11	112852	LP-SM-5-2019-GOB.REG.TACNA-1	10416307.86	240	Intervención Preexistente	Infraestructura Pública	Suma Alzada	LP	1	Regional	0	20%	10%	4	0	Arena	baja	0
13	130102	DIRECTA-PROC-15-2020-EPS TACNA S.A.-1	12664571.92	180	Intervención Preexistente	Saneamiento e Hidraulica	Precios Unitarios	AS y Otros	1	Local	1	8%	0%	1	1	Arena	media	1
14	503599	LP-SM-7-2022-GOB.REG.TACNA-1	9651051.63	180	Intervención Preexistente	Infraestructura Pública	Suma Alzada	LP	1	Regional	2	8%	10%	2	1	Grava	media	0

Anexo 2

Base de datos sin transformar empleada en la validación externa: variables objetivo

Item	cod_Infobras	Nomenclatura	BA_Vulnerabilidad_financiera_total	BB_Incidencia_economica_adicionales_endogenos	BC_%_avance_fisico	BD_Vulnerabilidad_plazo_total	BE_Incidencia_temporal_adicionales_endogenos
1	82727	LP-SM-5-2018-MDK/CS-1	4.49%	0.25%	93.57%	130.56%	27.50%
2	83604	AS-SM-40-2018-MDP/CS-2	14.95%	10.54%	97.38%	66.67%	18.33%
4	143200	LP-SM-5-2020-MDP/CS-1	29.34%	8.30%	96.74%	52.33%	15.00%
5	172992	LP-SM-14-2022-MDUA/CS-1	6.62%	4.10%	100.00%	1.90%	0.00%
6	78607	LP-SM-1-2021-MDU/CS-1	9.29%	0.00%	74.18%	60.00%	22.22%
8	168117	LP-SM-3-2022-UNAJ/CS-1	22.20%	14.52%	100.00%	77.22%	57.78%
9	68501	LP-SM-5-2019-MIGRACIONES-1	5.77%	0.49%	99.96%	74.76%	6.67%
11	112852	LP-SM-5-2019-GOB.REG.TACNA-1	18.86%	9.85%	99.40%	145.42%	63.33%
13	130102	DIRECTA-PROC-15-2020-EPS TACNA S.A.-1	7.27%	0.35%	92.60%	1.11%	1.11%
14	503599	LP-SM-7-2022-GOB.REG.TACNA-1	17.10%	9.93%	100.00%	98.89%	98.89%