REVISTA BOLETÍN REDIPE: 14 (9) SEPTIEMBRE 2025 ISSN 2256-1536
RECIBIDO EL 15 DE MAYO DE 2025 - ACEPTADO EL 17 DE AGOSTO DE 2025

DESARROLLO DE UN PROTOTIPO DE HERRAMIENTA ANALÍTICA PARA EL CONTROL DE PROYECTOS DE INFRAESTRUCTURA VIAL EN COLOMBIA INCORPORANDO MACHINE LEARNING

ANALYTICAL TOOL PROTOTYPE FOR MONITORING ROAD INFRASTRUCTURE PROJECTS IN COLOMBIA INCORPORATING MACHINE LEARNING

Daniel David Bonilla Bonilla*

Nicolás Alejandro Castellanos Roncancio**

Lina María Gómez Montenegro***

César Augusto Leal Coronado****

Universidad Escuela Colombiana de Ingeniería Julio Garavito, Bogotá D.C, Colombia.

Resumen

Este artículo presenta el desarrollo de un prototipo de herramienta analítica orientada al control de proyectos de infraestructura vial en Colombia, incorporando técnicas de machine learning como recurso pedagógico y de innovación en la gestión de proyectos. Su propósito es fortalecer los procesos de formación y toma de decisiones, tanto en entornos académicos como profesionales, mediante modelos predictivos que permiten estimar duración y costos, optimizar recursos y generar recomendaciones técnicas automáticas. La metodología adoptó el enfoque CRISP-DM, partiendo de la recopilación y depuración de datos históricos del Gestor de Proyectos de Infraestructura (GPI), y el desarrollo de cuatro modelos: LightGBM para predicción

de duración, K-Means para optimización de recursos, regresión lineal para estimación de costos y Random Forest para recomendaciones. Estos modelos se integraron en una interfaz interactiva que posibilita su uso en tiempo real, favoreciendo el aprendizaje aplicado y el análisis reflexivo. Los resultados evidencian alta precisión y clasificación efectiva en tres niveles de eficiencia. Se concluye que la incorporación de herramientas analíticas basadas en datos, en contextos formativos y profesionales, no solo mejora el control y la planificación de proyectos, sino que también fomenta competencias críticas en análisis de información, mitigación de riesgos y toma de decisiones estratégicas en escenarios complejos.

Abstract

This article presents the development of an analytical tool prototype designed for the control of road infrastructure projects in Colombia, incorporating machine learning techniques as both a pedagogical resource and an innovation in project management. Its purpose is to strengthen decision-making and training processes in both academic and professional settings through predictive models that estimate project duration and costs, optimize resources, and generate automatic technical recommendations. The methodology followed the CRISP-DM approach, starting with the collection and cleaning of historical data from the Infrastructure Project Manager (GPI), and the development of four models: LightGBM for duration prediction, K-Means for resource optimization, linear regression for cost estimation, and Random Forest for recommendations. These models were integrated into an interactive interface that enables real-time use, fostering applied learning and reflective analysis. The results show high predictive accuracy and effective classification into three efficiency levels. It is concluded that incorporating data-driven analytical tools in educational and professional contexts not only improves project control and planning but also fosters critical competencies in data analysis, risk mitigation, and strategic decision-making in complex scenarios.

Palabras clave

Aprendizaje automático, control de proyectos, pronóstico de duración, optimización de recursos, pronóstico de costos, sugerencias.

Keywords

Machine learning, project monitoring, duration forecasting, resource optimization, cost forecasting, recommendations

Introducción

La industria de la construcción es fundamental para el desarrollo económico ya que impulsa la conectividad, la competitividad y genera empleo en múltiples sectores. No obstante, ha enfrentado un persistente estancamiento a nivel global, con mejoras apenas del 10% entre 2000 y 2022, lo que representa solo una quinta parte de avance observado en la economía general (Vsimple, 2024).

En Colombia, el desempeño del sector construcción ha enfrentado desafíos significativos en los últimos años evidenciado por su bajo crecimiento y baja productividad. En el 2023, el PIB del sector registró una disminución del 4,1%, posicionándose como uno de los sectores con peor comportamiento económico del país. En particular, el subsector de carreteras y vías de ferrocarril experimentó una caída del 12,3%, reflejando una preocupante desaceleración en la inversión y ejecución de proyectos (CAMACOL, 2024).

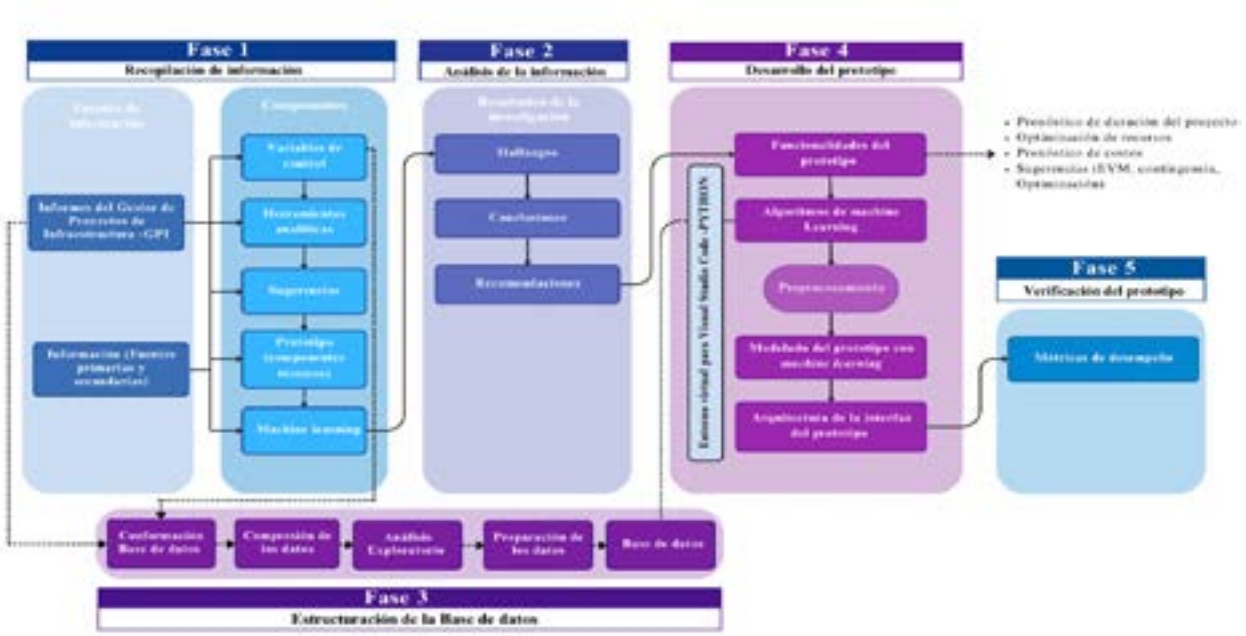
Este contexto evidencia que se trata de un sector rezagado con una ventana de oportunidad estratégica en el subsector de carreteras, en donde un factor asociado a la baja productividad es la baja adopción de herramientas analíticas

[3]. Teniendo en cuenta este escenario se hace pertinente la implementación de soluciones analíticas basadas en *machine learning* (ML) que contribuyan a cambiar dicha tendencia. Aunque la literatura incluye estudios que exploran el uso del ML en la estimación de costos y gestión de proyectos, aún son escasos los trabajos que van más allá de la teoría. Por ello, el propósito de la investigación que da origen a este artículo es desarrollar un prototipo de herramienta analítica, sustentado en técnicas de *machine learning*, para el seguimiento de proyectos de infraestructura vial en Colombia. El prototipo está diseñado para estimar el tiempo final del proyecto, optimizar recursos, pronosticar desviaciones de costo y ofrecer recomendaciones automatizadas, con base en datos históricos del Gestor de Proyectos de Infraestructura (GPI) del Ministerio de Transporte (CAMACOL, 2018).

Metodología

En la Figura 1, se detallan las fases del proyecto que guiaron la construcción del prototipo.

Figura 1. Fases del proyecto



Fuente: Autores

Fase 1: Recopilación de información

Se recolecta información proveniente de fuentes primarias y secundarias, incluyendo informes del Gestor de Proyectos de Infraestructura (GPI) del Ministerio de Transporte, literatura académica y normativa técnica, así como entrevistas con expertos del sector. Esta información es insumo para construir una base conceptual sólida que orienta el desarrollo del prototipo.

Fase 2: Análisis de la información

La información recopilada es analizada y se identifican variables clave para el control de proyectos, considerando indicadores contractuales, financieros, técnicos y climáticos.

Posteriormente, se establecen criterios de clasificación y selección de variables con potencial de ser incorporadas en modelos predictivos. Esto asegura que la información procesada tenga valor práctico y pertinencia para la herramienta analítica.

Fase 3: Estructuración de la base de datos

Se seleccionan 29 proyectos del GPI que cumplen criterios de completitud, coherencia y calidad, eliminando datos duplicados, corrigiendo inconsistencias y estandarizando formatos; esta depuración permite consolidar una base de datos homogénea y confiable. La base de datos final se estructura siguiendo la metodología CRISP-DM, para ello se realizan procesos propios del análisis exploratorio identificando datos faltantes, valores atípicos, relaciones entre variables y análisis de distribución de los datos. Posteriormente se realiza la imputación de los datos faltantes a través de métodos estadísticos, estandarización del formato de los campos y normalización de variables categóricas asegurando que los datos queden listos para las fases posteriores de desarrollo, entrenamiento y validación de los algoritmos de *machine learning*. La estructura de la base de datos comprende 621 registros y 37 campos, cuya clasificación se detalla en la tabla 1.

Tabla 1. Clasificación de las variables de la base de datos

N°	Clasificación	N° de Variables
1	**Datos fijos del proyecto**	10
2	**Datos de control temporal**	2
3	**Indicadores de avance**	7
4	**Recursos y productividad**	11
5	**Análisis y recomendaciones**	7
	Total	37

Fuente: Autores

Fase 4: Desarrollo del prototipo:

Modelo de pronóstico de duración de proyectos

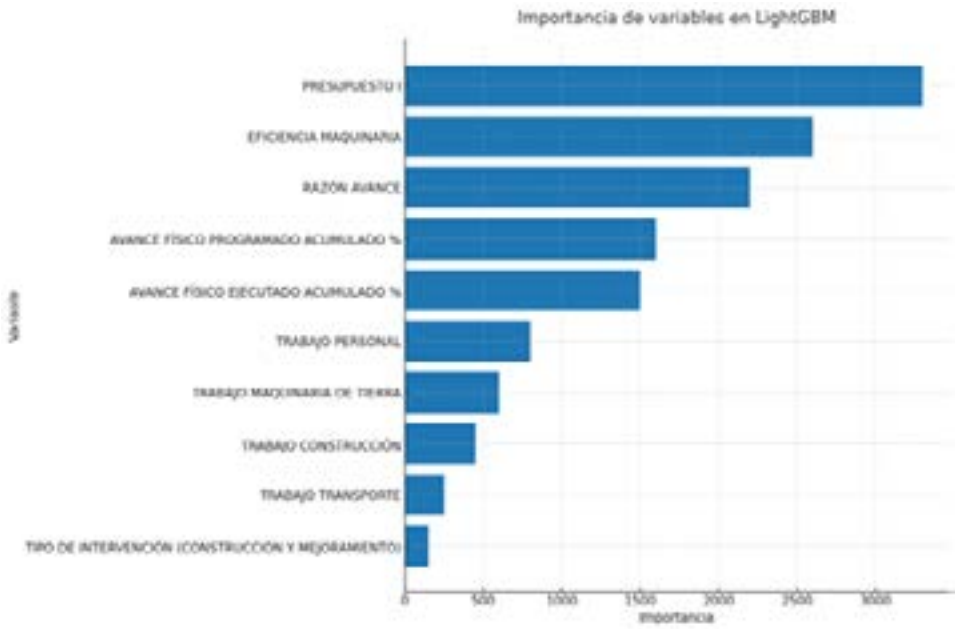
El modelo tiene como finalidad pronosticar la duración total estimada de proyectos de infraestructura vial en Colombia a partir de variables operativas, financieras se desarrolla un modelo buscando brindar una estimación más precisa de los tiempos de ejecución para apoyar la planificación, el control de avance y la gestión de riesgos. Se incluyen aspectos como la cantidad de maquinaria y personal utilizados, junto con los días de trabajo respectivos, además, se contemplan indicadores financieros como el presupuesto inicial y el avance físico programado y ejecutado.

En el proceso de preprocesamiento del modelo, la utilización de recursos se muestra con mucha mayor exactitud gracias a las variables derivadas que son generadas, estas variables incluyen a todo el trabajo realizado por la maquinaria de transporte, construcción, movimiento de tierra y personal. Todas son de tipo numérico, además se calculan como el producto entre cantidad y días operativos. Además, se consideran la razón de avance, que relaciona el avance ejecutado y programado, y la eficiencia de maquinaria, que es el cociente entre el trabajo total de maquinaria y el presupuesto inicial del proyecto; ambas de tipo decimal. Además, las variables categóricas fueron codificadas mediante *One-Hot Encoding* para que el modelo interprete correctamente sus diferencias. Con el fin de asegurar que

las mismas transformaciones se apliquen consistentemente al recibir nuevos datos antes de realizar predicciones, este preprocesamiento se encapsula en un script independiente.

Mediante el uso del algoritmo LightGBM, para pronosticar la duración de proyectos se evalúa la importancia relativa de cada variable. Las variables más destacadas son el presupuesto inicial, la eficiencia del uso de maquinaria y la razón de avance, como se aprecia en la figura 2. Este análisis permite identificar los factores más influyentes y ofrece una guía valiosa para ajustar las características del modelo.

Figura 2. Importancia de variables en *LightGBM*



Fuente: Autores

El algoritmo *LightGBM* [10] es ideal para escenarios complejos de proyectos de infraestructura por su alta precisión, capacidad de manejar restricciones lógicas (al aumentar maquinaria o personal, la duración tiende a disminuir) y su eficiencia computacional. El modelo basado en *LightGBM* funciona a partir de árboles de decisión que utiliza el método de *Gradient Boosting.* Este enfoque construye modelos de forma secuencial, donde cada nuevo árbol se entrena para corregir los errores del anterior. Para entrenar este modelo de predicción de duración de proyectos, se utiliza una división de la base de datos en un 70% para entrenamiento y un 30% para prueba, garantizando un conjunto representativo para evaluar el rendimiento; además se aplican restricciones monótonas que controlan la dirección de la relación entre variables independientes y la variable objetivo. Estas restricciones permiten asegurar que ciertas características, como el trabajo total de maquinaria y personal, mantengan una relación inversa con la duración, mayor trabajo implica menor tiempo (Rodrigo, 2023).

El modelo se configura con parámetros optimizados para balancear sesgo y varianza, se definen *n_estimators*=700 para construir un número suficiente de árboles que capture relaciones complejas, y *max_depth*=10 para controlar la profundidad de cada árbol y evitar sobreajuste, también la tasa de aprendizaje (*learning_rate*=0.05) se mantiene moderada para favorecer un aprendizaje progresivo y finalmente,

se incluye *monotone_constraints=monotonic_constraints* para forzar la coherencia del modelo respecto al comportamiento esperado de las variables (Amat Rodrigo, 2023).

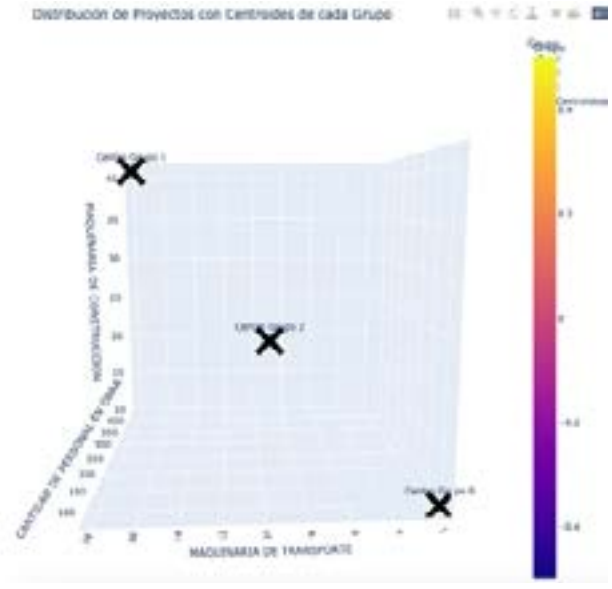
Modelo optimización de recursos

El modelo tiene como finalidad clasificar proyectos de infraestructura vial según su eficiencia en el uso de recursos, se desarrolla un modelo de *clustering* utilizando el algoritmo *K-Means (MacQueen,1967)*. Este modelo agrupa proyectos en función del uso de los recursos operativos (cantidad de personal y cantidad de maquinaria de transporte, construcción y movimiento de tierra) con el objetivo de identificar patrones de eficiencia que permitan sugerir mejoras hacia un uso óptimo y balanceado de estos recursos (Jain, 2010).

El algoritmo *K-Means*, se configura con tres grupos (*clusters*), se asigna cada proyecto a un grupo especifico (alto, balanceado y bajo). Esta clasificación permite identificar patrones de asignación de recursos que pueden aprovecharse para formular recomendaciones orientadas a optimizar la distribución de personal y maquinaria. Para el desarrollo del modelo los datos se escalan mediante *StandardScaler*, con el fin de garantizar que todas las variables tengan el mismo peso en el cálculo de distancias siendo especialmente relevante debido a que *K-Means* utiliza la distancia euclidiana como criterio de agrupamiento y es sensible a la escala de los datos. La estandarización aplicada transforma cada variable para que presente media (μ) igual a 0 y desviación estándar (σ) igual a 1, asegurando que variables con rangos diferentes, por ejemplo, personal (0–1000) y maquinaria de transporte (0–20) contribuyan de manera equilibrada al proceso de agrupamiento (Han ed al., 2012).

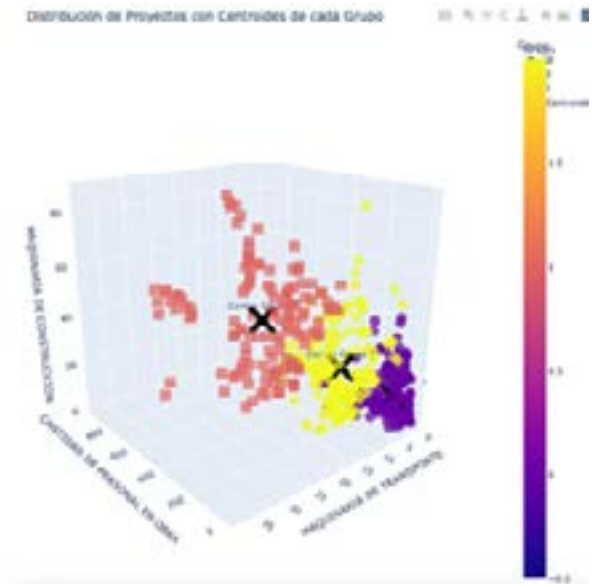
La figura 3 presenta la ubicación de los centroides de los grupos una vez el modelo llega a la convergencia después de las iteraciones realizadas. La figura 4 presenta la nube de puntos correspondientes a la distribución de los proyectos de la base de datos ya clasificados en los tres grupos.

Figura 3. Centroides de los grupos



Fuente: Autores

Figura 4. Nube de puntos de la base de datos



Fuente: Autores

Modelo de pronóstico de costos del proyecto

El modelo tiene como finalidad estimar el presupuesto final ejecutado en proyectos de infraestructura vial en Colombia, a partir de un factor de ajuste calculado con base en el presupuesto inicial y variables de desempeño físico y financiero. Este factor estimado permite anticipar desviaciones presupuestales y mejorar la planificación y el control financiero en las etapas de ejecución del proyecto.

Se selecciona el algoritmo de regresión lineal, por su alta interpretabilidad y comportamiento proporcional, siendo adecuado para explicar relaciones directas entre variables de entrada y el factor de incremento presupuestal. El modelo de regresión lineal utiliza el factor de costo, el cual se define como la relación entre el presupuesto final ejecutado y el presupuesto inicial de un proyecto (Mirjalili & Raschka, 2020).

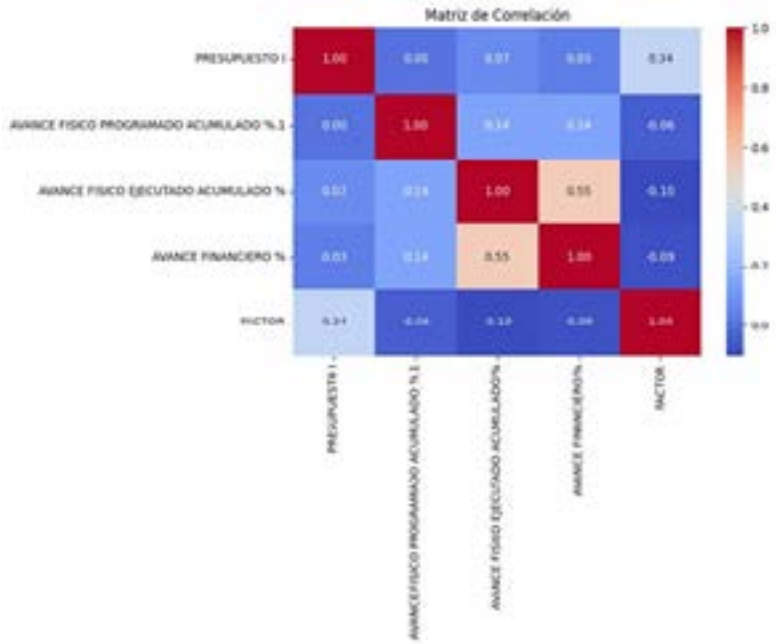
La forma general del modelo es la siguiente:

$$FACTOR = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \ ... \ + \beta_n * X_n + \in \quad (1)$$

Donde:

X_i representan las variables independientes, tales como los avances físicos y financieros del proyecto, β_i son los coeficientes del modelo, los cuales se determinan a partir de los datos durante el proceso de entrenamiento y ϵ corresponde al término de error, que captura las desviaciones no explicadas por el modelo (Google, s.f.) . Para respaldar el modelo de regresión lineal propuesto y garantizar que las variables seleccionadas contribuyen efectivamente a la estimación del factor de ajuste presupuestal, se incorpora una gráfica (Figura 5) que permiten analizar las relaciones entre los datos. La matriz de correlación muestra tres aspectos clave: hay una correlación moderada entre el presupuesto inicial y el factor de ajuste (0.34), una correlación notable entre el progreso físico realizado y el progreso financiero (0.55), y una correlación mínima del progreso físico planificado con las otras variables, lo que evidencia que su impacto en la proyección del costo final es restringido.

Figura 5. Ejemplo gráfico correlación de variables.



Fuente: Autores

El proceso para entrenar el modelo de regresión lineal comienza dividiendo los datos en dos partes: un 80% para entrenar y un 20% para probar. Se usa la regresión lineal simple a través de LinearRegression(), y los datos de prueba ayudan a ajustar el modelo. Finalmente, al aplicar el método fit(), el modelo aprende los coeficientes que muestran cómo se relacionan las variables independientes con el factor de costo. Esta relación genera una función que se usa para predecir dicho factor, lo cual ayuda a estimar el presupuesto final del proyecto.. Esta estimación se calcula mediante la siguiente fórmula:

*Presupuesto final predicho = Presupuesto inicial * Factor (*2)

Modelo de generación de sugerencias

El modelo tiene como finalidad apoyar la toma de decisiones en la gestión de proyectos de infraestructura vial para ello se utiliza un enfoque de desarrollo bajo un modelo de clasificación multiclase Random Forest a partir del análisis de variables contractuales, climáticas, operativas y de desempeño con el fin de realizar una generación automática de acciones técnicas recomendadas en tres dimensiones de control: metodología del valor ganado (EVM), contingencia y optimización (Cutler et al., 2007).

Las variables seleccionadas de la base de datos, obedecen a los factores determinantes en el seguimiento y control de los proyectos de infraestructura, las cuales son: tipo de intervención, zona geográfica, duración del proyecto en meses, prórroga en meses, suspensión en meses, adición presupuestal, presupuesto inicial, presupuesto final ejecutado, cantidad de maquinaria de transporte, maquinaria de construcción, maquinaria de movimiento de tierra, cantidad de personal en obra, horas de lluvia al mes, índice de desempeño de costos (CPI), índice de desempeño de cronograma (SPI), variación de costos (CV) y variación de cronograma (SV). Las variables anteriores permiten lograr una cobertura total que garantiza que el modelo capture la complejidad de las condiciones de ejecución y genere recomendaciones ajustadas al perfil específico de cada proyecto. Las variables categóricas “zona geográfica” y “tipo de intervención” no pueden ser procesadas directamente por el modelo, ya que el algoritmo Random Forest requiere entradas numéricas, por esta razón se aplica Label Encoding, que asigna un valor entero único a cada categoría, preservando la distinción entre clases (Liaw & Wiener, 2002).

Posteriormente, dado que las clases objetivo presentan un desbalance significativo debido a que algunas categorías de recomendación están subrepresentadas, se emplea SMOTENC (Synthetic Minority Over-sampling Technique for Nominal and Continuous), bajo este método se generan ejemplos sintéticos para las clases minoritarias, respetando la naturaleza mixta de la base de datos (variables numéricas y categóricas), evitando que el modelo sesgue sus predicciones hacia las clases mayoritarias (Rodriguez ed al., 2012).

Una vez preprocesados y balanceados, los datos se dividen en 80 % para entrenamiento y 20 % para validación. Esta partición permite que el modelo entrene con la mayoría de los datos, capturando patrones complejos de interacción entre variables, mientras que el 20 % reservado se utiliza para evaluar su capacidad de generalización, es decir, su rendimiento sobre información que no ha visto previamente, asegurando robustez y reduciendo el riesgo de overfitting. Se desarrollaron tres modelos Random Forest entrenados individualmente para predecir la recomendación más adecuada en cada dimensión, cada modelo consta de 300 arboles de decisión con una profundidad máxima definida en 12 y mínimo 5 muestras

para dividir un nodo con el objetivo de controlar la complejidad (Rodriguez ed al., 2012).

Resultados

Una vez presentadas todas las fases de la figura 1, se procede a exponer los resultados a continuación.

Finalizado el desarrollo y entrenamiento de cada uno de los modelos de machine learning, estos se integran en 4 pantallas diseñadas para que el usuario ingrese los datos de las variables requeridas en cada uno y pueda consumir los modelos en tiempo real y conocer el resultado de los pronósticos de cada modelo.

La interfaz del sistema expone un gráfico 3D, que se presenta en la figura 6, donde se ve la respuesta al estimativo de duración de tiempo según los datos ingresados por el usuario y compara su pronóstico frente a los datos históricos de proyectos similares. Esto contribuye a anticipar la duración del proyecto con base en datos reales, incluso durante su ejecución, facilitando ajustes tempranos al cronograma para evitar desviaciones por medio de la planificación operativa al evaluar si los recursos asignados son adecuados, proporcionando respaldo técnico para prórrogas o cambios contractuales.

Figura 6. Respuesta modelo del pronóstico de duración del proyecto.

Fuente: Autores

La efectividad de esta funcionalidad se respalda en el desempeño del modelo evaluado mediante tres métricas clave: el coeficiente de determinación (R^2), el error absoluto medio (MAE) y la precisión estimada *(accuracy)*, reforzando la utilidad del modelo como herramienta de apoyo para la toma de decisiones en la gerencia de proyectos (Mirjalili & Raschka, 2020).

Los resultados obtenidos de R^2 Score de 0.8923 indica que el modelo explica el 89.23% de la variabilidad en la duración de los proyectos. El MAE de 4.09 meses señala que, en promedio, las predicciones difieren de la duración real por ese valor y por último, el *accuracy* del 87.76% refleja que la mayoría de las predicciones se encuentran dentro de un margen de error aceptable (Mirjalili & Raschka, 2020)..

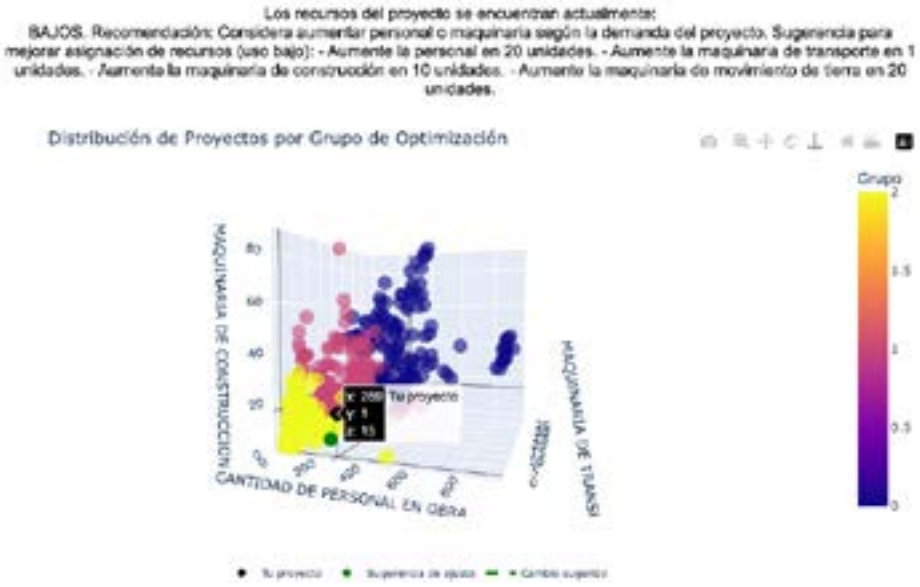
En la figura 7 se presenta la respuesta del modelo de optimización de recursos basada en una combinación de datos ingresados. El proyecto se presenta al usuario gráficamente clasificado en alguno de los 3 tipos de agrupación junto con

una recomendación aritmética de reorganización de los recursos operativos dependiendo de la clasificación realizada por el modelo.

Cada proyecto clasificado en uno de los tres grupos (Grupo_Optimización) puede analizarse para:

- Detectar infra asignación de recursos, evitando limitaciones operativas.
- Mantener buenas prácticas observadas en el grupo balanceado (óptimo).
- Evaluar excesos de recursos, proponiendo estrategias de ajuste sin comprometer la calidad.

Figura 7. Respuesta del modelo de optimización de recursos frente a datos ingresados



Fuente: Autores

Este tipo de análisis aporta un valor significativo en casos reales de proyectos de infraestructura vial, ya que permite anticipar desviaciones operativas, proponer estrategias de reasignación basadas en datos históricos y patrones de eficiencia observados. En la práctica, esta información puede apoyar la toma de decisiones proactiva, optimizando la ejecución, reduciendo riesgos de retrasos y mejorando el rendimiento global del proyecto sin incurrir en sobrecostos innecesarios

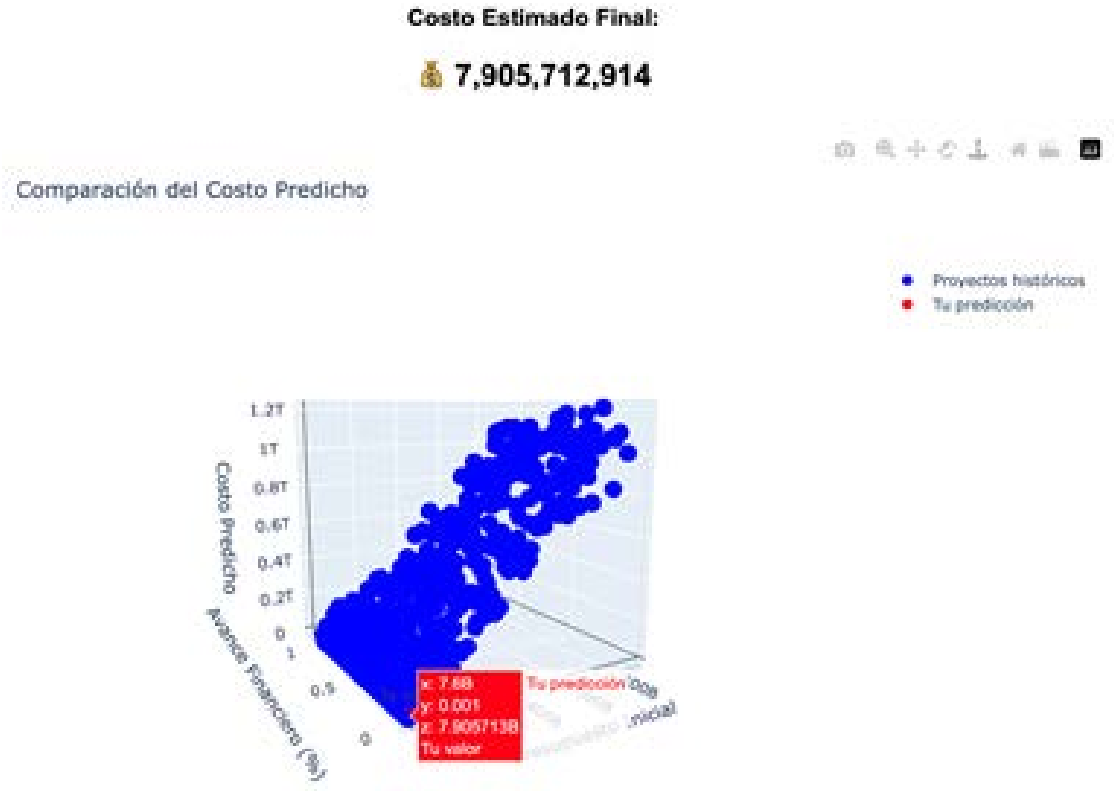
Dado que el modelo de optimización de recursos implementado se fundamenta en un enfoque no supervisado mediante técnicas de *clustering*, su desempeño fes evaluado a través de tres métricas internas ampliamente reconocidas en la literatura: *Silhouette Score* (Rousseeuw, 1987), *Calinski-Harabasz Inde*x (Calinski & Harabasz,1974) y *Davies-Bouldin Index (*Davies & Bouldin,1979). Estas métricas permiten analizar la calidad de las agrupaciones en términos de cohesión y separación, sin requerir etiquetas verdaderas.

El *Silhouette Score* arrojó un valor de 0.325 que refleja una estructura de agrupamiento aceptable y coherente con la complejidad de los datos, aunque sin segmentación altamente definida. El *Calinski-Harabasz Index* de 483.75 confirma

clústeres bien diferenciados y compactos, acorde con los análisis tridimensionales. El Davies-Bouldin Index de 1.17, dentro del rango aceptable, valida una separación suficiente para sustentar recomendaciones operativas en proyectos de infraestructura.

La respuesta del modelo de pronóstico de costos se presenta en la figura 8. Esta visualización en 3D permite comparar el valor estimado del proyecto actual frente a los costos de proyectos de la base de datos, el usuario puede comprender fácilmente cómo se comporta su proyecto en relación con datos históricos de otros proyectos, lo que facilita una lectura más intuitiva del resultado. Esta herramienta representa un apoyo fundamental para el seguimiento y control financiero del proyecto. Permite anticipar posibles sobrecostos según el desempeño actual, tomar decisiones presupuestales más informadas como reasignaciones o ajustes de contingencia, mejorar la planificación del flujo de caja y brinda sustento para justificar desviaciones presupuestales mediante análisis comparativos y proyecciones basadas en datos reales.

Figura 8. Respuesta modelo del pronóstico de costo del proyecto.

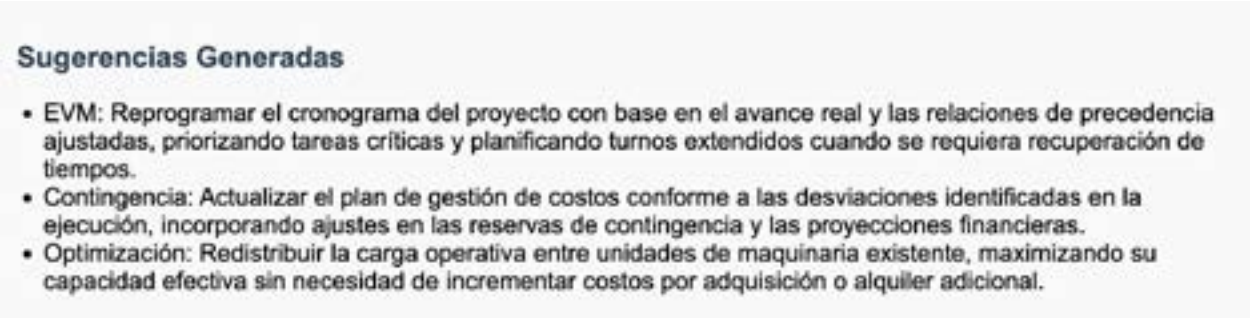


Fuente: Autores

La validación del desempeño del modelo de regresión lineal se realiza a partir de tres métricas. Primero, el Error Absoluto Medio (MAE) fue de aproximadamente 250 millones de dólares. Luego, se obtiene un coeficiente de determinación (R^2) de 0.84. Por último, la precisión estimada es de alrededor del 92.5%, lo que significa que, en promedio, las predicciones del modelo están cerca del 92.5% del valor real del presupuesto final. Esto demuestra que el modelo es útil en tareas de control y estimación financiera. En general, estas métricas muestran que el modelo funciona de manera sólida y confiable.

La evaluación de los modelos de recomendaciones desarrollados evidencia un desempeño sólido y consistente en las tres áreas analizadas: metodología del valor ganado (EVM), contingencia y optimización de recursos. Cada modelo fue entrenado para clasificar acciones técnicas específicas, con métricas que reflejan una buena capacidad de generalización frente a datos no vistos. Si bien los resultados muestran variabilidad en el rendimiento de ciertas clases debido a similitudes semánticas, el conjunto de modelos mantiene un equilibrio adecuado entre precisión, sensibilidad y capacidad predictiva. La figura 9 muestra la respuesta de los modelos *Random Forest* frente a los datos ingresados de un escenario de un proyecto.

Figura 9. Respuesta modelos *Random Forest*



Fuente: Autores

En términos generales, estos modelos constituyen una herramienta funcional para generar recomendaciones automatizadas y coherentes con el contexto operativo de proyectos de infraestructura vial. Los tres modelos de sugerencias presentan desempeños consistentes y aceptables. El modelo EVM alcanza un *accuracy* del 80 %, con métricas equilibradas de *precision*, *recall* y *f1-score* (0.80 en promedio), evidenciando buena capacidad para identificar y clasificar recomendaciones técnicas. El modelo de Contingencia obtiene una *accuracy* del 79 %, manteniendo estabilidad en sus métricas globales (*precision y recall* ≈0.78–0.79), lo que indica solidez al predecir acciones frente a riesgos y eventos no planificados. Por su parte, el modelo de Optimización de recursos logra una *accuracy* del 76 %, con *f1-score* promedio de 0.75, mostrando un desempeño ligeramente inferior pero aceptable considerando la complejidad y variabilidad en las estrategias de asignación de recursos. En conjunto los tres modelos demuestran consistencia en la predicción y aplicabilidad práctica en contextos de proyectos viales.

Conclusiones

El prototipo, fundamentado en cuatro modelos de aprendizaje automático, evidencia que la utilización de datos históricos optimiza la gestión de proyectos de infraestructura vial. Además, resalta que la calidad y coherencia de los datos son cruciales para crear análisis predictivos, mejorar la administración y prever desviaciones. Las evaluaciones efectuadas evidencian que los modelos creados proporcionan una robusta estimación, lo que permite prever con exactitud tanto la duración como los costos de los proyectos, favoreciendo una óptima planificación y administración de riesgos. La optimización de recursos mediante agrupamiento no supervisado permite clasificar proyectos y sugerir ajustes operativos basados en datos reales, promoviendo una mejora continua que se fundamenta en evidencia. Además, los modelos de recomendación, especialmente para

acciones críticas, brindan soporte técnico fiable, ayudando a estandarizar buenas prácticas y a identificar áreas de mejora de forma oportuna a lo largo del proyecto.

El análisis conjunto de las métricas de desempeño incluyendo R^2, MAE, *accuracy*, f1-score y los índices internos de agrupamiento evidencia que los modelos desarrollados ofrecen pronósticos fiables y generalizables. Los altos valores de R^2 y *accuracy*, junto a bajos MAE, aseguran un buen ajuste a los datos como errores mínimos y relevantes para la práctica del sector, asimismo, los índices de *clustering* reflejan agrupaciones suficientemente compactas y separadas, lo que valida la pertinencia de las recomendaciones operativas.

Dado que el prototipo desarrollado se basa en un conjunto de datos estático, una línea de trabajo futuro recomendable es la incorporación de mecanismos de realimentación que permitan la actualización continua de los modelos. Esto facilitaría el aprendizaje a partir de nuevos proyectos registrados, mejorando progresivamente la precisión de los pronósticos.

Referencias bibliográficas

Abed, Y. G., Hasan, T. M., & Zehawi, R. N. (2022). Machine learning algorithms for constructions cost prediction: A systematic review. International Journal of Nonlinear Analysis and Applications, 13(2), 2205–2218. https://doi.org/10.22075/ijnaa.2022.27673.3684

Amat Rodrigo, J. (2023). Gradient boosting con Python. https://dev.cienciadedatos.net/documentos/py09_gradient_boosting_python

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). https://doi.org/10.56556/jtie.v2i2.511

Calinski, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. Communications in Statistics, 3(1), 1–27. https://doi.org/10.1080/03610927408827101

Cámara Colombiana de la Construcción [CAMACOL]. (2018). Informe de productividad en el sector de la construcción. https://camacol.co/informe-productividad

Camacol. (2024). Informe económico No. 119: Coyuntura y retos para el sector de la construcción en 2024. https://camacol.co/informe-economico-119

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007). Random forests for classification in ecology. Ecology, 88(11), 2783–2792. https://doi.org/10.1890/07-0539.1

Departamento Administrativo Nacional de Estadística [DANE]. (2022). Encuesta Pulso Empresarial: Uso de tecnologías de inteligencia artificial por sectores económicos. https://www.dane.gov.co/files/investigaciones/boletines/pulso-empresarial/presentacion-pulso-empresarial-oct22-nov22.pdf

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

EY Americas. (2021, noviembre 3). AI: Construction's new frontier of digital enablement. https://www.ey.com/en_us/construction-real-estate/ai-constructions-new-frontier-of-digital-enablement

Google for Developers. (s. f.). Machine learning: Regresión lineal. https://developers.google.com/machine-learning/crash-course/linear-regression?hl=es-419

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann.

Instituto Iberoamericano de Mercados de Valores [IIMV]. (2017). La financiación de las micro, pequeñas y medianas empresas a través de los mercados de capitales en Iberoamérica. Fundación IIMV. https://scioteca.caf.com/handle/123456789/1454

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30. https://dl.acm.org/doi/10.5555/3294996.3295074

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R News, 2(3), 18–22.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1(Statistics), 281–297.

Mirjalili, V., & Raschka, S. (2020). Python machine learning (1st ed.). Marcombo. https://www.perlego.com/book/2152522/python-machine-learning-pdf

Munawar, H. S., Ullah, F., Qayyum, S., & Shahzad, D. (2022). Big data in construction: Current applications and future opportunities. Big Data and Cognitive Computing, 6(1), 18. https://doi.org/10.3390/bdcc6010018

Rodriguez-Galiano, V. F., Ghimire, B., Rogan, J., Chica-Olmo, M., & Rigol-Sanchez, J. P. (2012). An assessment of the effectiveness of a random forest classifier for land-cover classification. ISPRS Journal of Photogrammetry and Remote Sensing, 67, 93–104. https://doi.org/10.1016/j.isprsjprs.2011.11.002

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Vsimple. (2024, agosto 29). McKinsey study shows the construction industry is in a productivity rut—Here's how Vsimple can help. Vsimple Insights. https://vsimple.com/insights/mckinsey-study-shows-the-construction-industry-is-in-a-productivity-rut

Zhou, Z.-H. (2021). Ensemble methods: Foundations and algorithms. CRC Press. https://doi.org/10.1201/9781003033350