Propuesta para la Estimación de Costo y días en la Finalización de Proyectos de Construcción con Inteligencia Artificial

Miryan R. Puchini [a, *]; Yolanda E. Zado [a]; Nancy B. Ganz [a]

[a] *Universidad Nacional de Misiones (UNaM), Facultad de Ingeniería, Oberá, Misiones, Argentina.*
E-mails: miryanpuchini@fio.unam.edu.ar, yolandazado200@gmail.com, nancyganz@fceqyn.unam.edu.ar

Resumen

Este trabajo aborda el desafío de estimar con precisión los costos y la duración de finalización de proyectos de construcción mediante la aplicación de modelos de Inteligencia Artificial (IA). A partir de una revisión crítica de los métodos tradicionales y de estudios recientes en la industria de la construcción, se propone un enfoque basado en algoritmos de aprendizaje automático, como Random Forest y regresión lineal, entrenados sobre conjuntos de datos sintéticos debido a la limitada disponibilidad de datos reales en la provincia de Misiones. La metodología adoptada contempla un flujo de trabajo típico del ML: preprocesamiento de datos, modelado y evaluación. Los resultados evidencian que, con datos de calidad, los modelos predictivos alcanzan altos niveles de precisión (90–99 %), superando las limitaciones de métodos clásicos como CPM o PERT. Asimismo, se destaca la importancia de contar con bases de datos estructuradas y accesibles que integren información clave del ciclo de vida del proyecto, sugiriendo como estrategia su incorporación en el portal IDE Misiones. El estudio concluye que la IA constituye una herramienta valiosa para mejorar la planificación y toma de decisiones en proyectos constructivos, especialmente en contextos donde la incertidumbre y la complejidad son elevadas.

Palabras Clave *– Proyectos de construcción, estimación de tiempo y costo, Inteligencia Artificial, Machine Learning, Deep Learning*

1 Introducción

La precisión en la estimación de tiempo y costos constituye un factor crítico para la planificación y gestión eficiente de los proyectos de construcción. La presencia de incertidumbres o información imprecisa en estas etapas puede derivar en sobrecostos, demoras, desvíos en los cronogramas y uso ineficiente de los recursos, afectando negativamente a todas las partes involucradas.

Tradicionalmente, estas estimaciones se han abordado mediante métodos descriptivos, conceptuales o estadísticos, los cuales presentan limitaciones en entornos altamente dinámicos y complejos. En este contexto, las investigaciones recientes han demostrado que la incorporación de modelos de inteligencia artificial y aprendizaje automático permite mejorar significativamente la precisión predictiva, al incorporar técnicas avanzadas como Random Forest, K-Nearest Neighbors, SVM o redes neuronales, con resultados más robustos frente a la variabilidad de los datos.

Este trabajo se fundamenta en una revisión bibliográfica crítica y propone el desarrollo de modelos basados en IA como alternativa eficaz frente a los métodos tradicionales, con el objetivo de optimizar la estimación de tiempo y costos en proyectos de construcción, reduciendo la incertidumbre y mejorando la toma de decisiones desde etapas tempranas.

*miryanpuchini@gmail.com

2 Marco Teórico

2.1. Factores e incertidumbres en los proyectos de construcción

La finalización en tiempo y forma de los proyectos de construcción depende en gran medida de estimaciones precisas de tiempo y costo, dos parámetros que se ven fuertemente afectados por incertidumbres inherentes al entorno constructivo. Estas incertidumbres, si no se gestionan adecuadamente, impactan negativamente en la planificación, ejecución y control del proyecto, por lo que es crucial desarrollar estrategias eficaces para su identificación y mitigación.

En el trabajo consultado han subrayado la importancia de incorporar la incertidumbre en los modelos de programación de obras, mediante el uso de herramientas avanzadas como algoritmos de inteligencia artificial y técnicas estadísticas. Se destaca, además, la influencia de factores como la gestión deficiente del sitio, la escasez de recursos financieros y las fallas comunicacionales, los cuales contribuyen a desviaciones significativas en los cronogramas y presupuestos.

Han abordado estas problemáticas a través de modelos matemáticos capaces de cuantificar el impacto de los riesgos sobre la duración y los costos. El tratamiento explícito de distintos tipos de incertidumbre, ya sea a través de estudios empíricos, curvas S de riesgo o modelos predictivos, ha demostrado mejorar considerablemente la precisión de las estimaciones.

Asimismo, evidencian que el sistema de entrega del proyecto influye directamente en la estimación del tiempo, siendo objeto de análisis comparativos, evaluaciones de riesgo y estudios de caso, especialmente en proyectos de vivienda pública. La adopción de enfoques como el diseño-construcción o la entrega integrada de proyectos ha mostrado ventajas frente a contextos de alta complejidad e incertidumbre, al facilitar una mejor coordinación entre las partes y una mayor adaptabilidad frente a eventos no previstos.

Finalmente mencionan, sobre nuevas líneas de investigación que abordan temas emergentes que afectan la entrega de proyectos, como la ciberseguridad en entornos digitales de construcción o la incorporación de indicadores de sostenibilidad en los procesos de planificación y ejecución, ampliando así la visión tradicional de riesgos hacia un enfoque más integral y actual. [1]

Respecto a los factores que inciden en los resultados del proyecto, la literatura los clasifica en dos categorías: Factores Cruciales, aquellos que ayudan a prevenir disputas y retrasos, por ejemplo, la estimación precisa de la duración de cada actividad. Factores Críticos de Éxito (CSF, por sus siglas en inglés), que garantizan el éxito en la planificación y permiten una gestión eficiente de los recursos. Algunos CSF identificados son: tamaño y complejidad del proyecto, condiciones del sitio, tipo de obra, ubicación geográfica, disponibilidad de recursos y materiales, productividad de la mano de obra, modificaciones en el diseño, cambios normativos y condiciones climáticas imprevistas.

Se han identificado hasta 48 factores causales de retrasos y 38 factores de sobrecostos, clasificados en cuatro categorías: causas atribuibles al contratista, al consultor, al cliente y causas externas. Estos estudios constituyen un marco valioso, pero es necesario profundizar en su análisis

para desarrollar soluciones viables y evaluar su eficacia en la práctica, con el objetivo de asegurar la ejecución exitosa de los proyectos de construcción. [2]

2.2. Métodos tradicionales para estimar tiempo y costo

Los proyectos de construcción, especialmente los del tipo residencial, requieren estimaciones precisas de la duración de cada actividad debido a los diversos riesgos e incertidumbres que afectan el tiempo total de finalización. Tradicionalmente, estas estimaciones se realizan mediante técnicas clásicas de planificación y gestión de proyectos, entre las que se destacan:

- El Método del Camino Crítico (CPM, por sus siglas en inglés Critical Path Method) es un enfoque determinista ampliamente utilizado en la planificación de proyectos de construcción. Su principal ventaja radica en la simplicidad de su implementación, ya que permite identificar la duración mínima del proyecto a partir de un análisis lineal. Sin embargo, una de sus limitaciones más significativas es que considera una única duración fija para cada actividad, sin incorporar la variabilidad ni los factores de incertidumbre. Como resultado, el cronograma generado puede no reflejar fielmente las condiciones reales del proyecto. [3]
- La Técnica de Evaluación y Revisión de Programas (PERT, por sus siglas en inglés Program Evaluation and Review Technique), introduce un enfoque probabilístico al considerar tres estimaciones de duración (optimista, más probable y pesimista) para cada actividad. Este método permite incorporar la incertidumbre inherente a los proyectos, generando una planificación más realista y detallada, al tiempo que facilita la visualización de tareas y sus respectivas dependencias. [4]
- El Método de Simulación de Montecarlo (MCS) es una técnica estadística avanzada que permite simular múltiples escenarios posibles en función de variables aleatorias. Su aplicación no se limita solo a la estimación de tiempos, sino que también considera restricciones de costo y riesgos asociados, lo cual proporciona un marco más robusto para la toma de decisiones estratégicas durante la planificación y ejecución del proyecto. [5]

Además, en el estudio [5] se ejemplifica la aplicabilidad de estos métodos tradicionales en un proyecto de construcción de una vivienda con 24 actividades planificadas. Al aplicar CPM, se obtuvo una duración estimada de 63 días. En contraste, al emplear PERT y MCS con un nivel de confianza del 95 %, las duraciones proyectadas fueron de 70 y 79 días, respectivamente. Estos resultados evidencian que los métodos probabilísticos, como PERT y MCS, ofrecen cronogramas más ajustados a la realidad al contemplar los riesgos e incertidumbres propios de los proyectos de construcción.

Estos resultados demuestran que los métodos probabilísticos constituyen herramientas valiosas para los gerentes de proyectos, permitiéndoles programar obras de construcción residencial considerando los riesgos e incertidumbres inherentes. Si bien, PERT como MCS son útiles en la elaboración de cronogramas más realistas; sin embargo su precisión está directamente condicionada

por la calidad de las variables de entrada y por la fidelidad del modelo que representa al proyecto. Las deficiencias de estos modelos, están en que las estimaciones de duración de las actividades no son precisas, perdiendo confiabilidad en sus resultados. Es importante destacar que ni PERT ni MCS proporcionan soluciones definitivas por sí mismos, sino que actúan como herramientas de apoyo para anticipar el comportamiento del proyecto en escenarios inciertos y la toma de decisiones finales recae en el gerente de proyecto, quien debe realizar una evaluación integral de las condiciones particulares de la obra antes de adoptar una estrategia de planificación adecuada. [6]

2.3. Modelación de Información de la Construcción

Como se ha mencionado previamente, los métodos tradicionales, deterministas o probabilísticos, utilizados en la planificación de proyectos de construcción presentan limitaciones significativas, especialmente en lo que respecta a la integración de información multidimensional y la actualización en tiempo real.

En respuesta a estas limitaciones, el Modelado de Información de la Construcción (BIM por su siglas en ingles Building Information Modeling,) se posiciona como una tecnología disruptiva que ha transformado la forma en que se planifican, diseñan, ejecutan y gestionan los proyectos. BIM opera dentro de un entorno colaborativo e integrado, facilitando la centralización de datos y mejorando la coordinación entre los distintos actores del proceso constructivo.

En la actualidad, numerosos profesionales del sector de la arquitectura, la ingeniería y la construcción (AEC) reconocen un cambio de paradigma marcado por la convergencia entre BIM y la Inteligencia Artificial (IA). Esta integración promueve nuevas oportunidades para optimizar el rendimiento organizacional, mejorar la toma de decisiones y aumentar la eficiencia en la entrega de proyectos, especialmente mediante el aprovechamiento de grandes volúmenes de datos secundarios generados durante el ciclo de vida del proyecto.

BIM puede entenderse a partir de sus dos componentes fundamentales: por un lado, como una representación digital compartida de un activo físico, que constituye una base confiable para la toma de decisiones a lo largo de todo su ciclo de vida, desde su concepción hasta la demolición; y por otro lado, como un proceso basado en modelos 3D inteligentes, que proporciona a los profesionales del sector AEC las herramientas y la información necesarias para planificar, diseñar, construir y operar edificaciones e infraestructuras de manera más eficiente.

Un concepto clave de BIM es su estructura en una serie de capas interconectadas, cada una de las cuales aporta niveles adicionales de información y funcionalidad al modelo, enriqueciendo su capacidad para representar y gestionar el ciclo de vida de un proyecto de construcción. Comienza con la capa geométrica (3D), que representa la forma y ubicación de los elementos del edificio, y se complementa con la capa semántica, que agrega propiedades y atributos técnicos. La capa topológica define relaciones funcionales entre componentes, mientras que las capas temporal (4D) y de costos (5D) incorporan planificación y estimación financiera. A esto se suman la capa de sostenibilidad (6D), que evalúa el impacto ambiental, la gestión de instalaciones (7D), que abarca el

mantenimiento y uso eficiente del edificio, y finalmente la integración con IoT, que conecta sensores y dispositivos inteligentes para monitoreo en tiempo real.

2.3.1 Rol de la IA en BIM

La integración entre la IA y BIM está dando lugar a modelos inteligentes capaces de procesar grandes volúmenes de datos en tiempo real, reconocer patrones y proponer soluciones optimizadas en todo el ciclo de vida del proyecto. Esta sinergia ha impulsado el desarrollo de múltiples aplicaciones especializadas, acompañadas de herramientas concretas en el mercado:

- Diseño paramétrico y generativo: Autodesk Forma, permite generar múltiples opciones de diseño en función de datos del sitio y normativas; Project Refinery, aplica IA para explorar alternativas según objetivos específicos.
- Análisis energético y sostenibilidad: Autodesk Insight, integrado con Revit, analiza el rendimiento energético del edificio; Tally, complemento de Revit, calcula el impacto ambiental de los materiales.
- Detección de conflictos y coordinación: Navisworks, con capacidades de IA, identifica conflictos y sugiere resoluciones basadas en datos históricos.
- Simulación de construcción y planificación (4D): Synchro, permite simular escenarios, optimizar cronogramas y prever retrasos.
- Cumplimiento normativo automatizado: UpCodes AI, vinculado con Revit, verifica automáticamente la conformidad con códigos de edificación.
- Evaluación de riesgos del proyecto: Kognoz, aplica Machine Learning para anticipar riesgos combinando datos BIM e históricos.
- Gestión inteligente de instalaciones (7D): IBM TRIRIGA, integrado con BIM, optimiza el mantenimiento, la gestión de activos y el uso eficiente del espacio.
- Modificaciones de diseño en tiempo real: Revit con IA, puede sugerir cambios basados en costos, rendimiento y regulaciones.

La fusión entre IA y BIM no solo mejora la eficiencia operativa, sino que redefine la forma en que se conciben, construyen y mantienen los entornos construidos. Las aplicaciones actuales ya impactan en áreas clave como diseño generativo, análisis de sostenibilidad, planificación 4D, gestión de riesgos y mantenimiento predictivo.

Las investigaciones apuntan al desarrollo de algoritmos de IA más robustos, la incorporación de computación cuántica, técnicas de aprendizaje federado, IA explicable, biomimética, blockchain, y gemelos digitales inteligentes aplicados a ciudades. Este ecosistema tecnológico promete transformar radicalmente la industria de la AEC, facilitando la creación de entornos más sostenibles, resilientes e inteligentes. [7]

De lo expuesto se desprende que la implementación efectiva de la metodología BIM, especialmente en combinación con herramientas de inteligencia artificial, requiere la utilización de múltiples softwares especializados que abordan distintas dimensiones del proyecto (diseño, análisis

energético, planificación, gestión de riesgos, mantenimiento, entre otros). Esta multiplicidad de plataformas, muchas veces desarrolladas por distintos proveedores y con niveles variables de interoperabilidad, genera una fragmentación de la información. Como resultado, no es posible disponer de una base de datos única y centralizada que contenga de forma integrada todos los datos del proyecto. Esta dispersión dificulta la trazabilidad completa, la sincronización entre disciplinas y la aplicación de modelos de análisis globales basados en datos unificados.

2.3. Automatizar las estimaciones del tiempo y costo de un proyecto de construcción

En disciplinas ajenas a las ciencias de la computación, los términos Inteligencia Artificial (IA), aprendizaje automático o Machine Learning (ML, por sus siglas en inglés), y aprendizaje profundo o Deep Learning (DL, por sus siglas en inglés) suelen utilizarse de manera indistinta, especialmente en contextos comerciales. Sin embargo, desde una perspectiva técnica y académica, presentan diferencias sustanciales que es importante esclarecer (ver Fig. 1).
La IA abarca el conjunto de técnicas que permiten a las máquinas imitar capacidades cognitivas humanas, como el razonamiento, la toma de decisiones, el reconocimiento de patrones o el procesamiento del lenguaje. El ML constituye un subconjunto de la IA, que se enfoca en desarrollar algoritmos capaces de aprender a partir de datos, identificar patrones y realizar predicciones sin ser explícitamente programados para cada tarea específica. Algunos algoritmos de ML utilizados para estimaciones incluyen la regresión lineal, los árboles de decisión y los métodos basados en vecinos más cercanos (K-Nearest Neighbors por sus siglas en inglés KNN). El aprendizaje profundo es un subconjunto dentro del ML que se basa en arquitecturas de redes neuronales artificiales compuestas por múltiples capas de procesamiento. Estas redes permiten modelar relaciones no lineales complejas y extraer representaciones jerárquicas de los datos, mejorando el rendimiento en tareas como el reconocimiento de imágenes, el procesamiento de lenguaje natural y la predicción de variables en contextos de alta dimensionalidad, siendo eficaz en contextos donde los datos son masivos y no estructurados. [8]

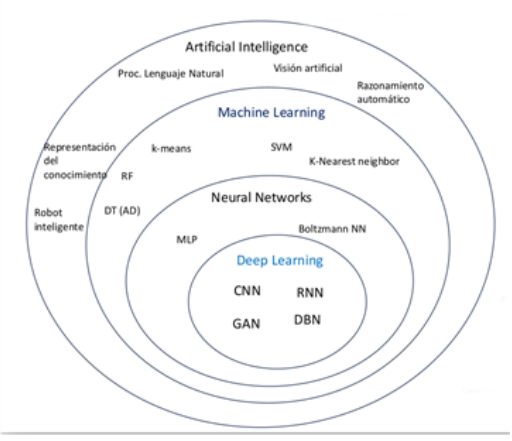

Fig. 1 - Taxonomía IA.
Fuente [9]

Diversas investigaciones han evidenciado que la integración de modelos de inteligencia artificial con técnicas estadísticas mejora significativamente la precisión en la estimación de costos y plazos en proyectos de construcción. Se han aplicado algoritmos de aprendizaje automático como Máquinas de Vectores de Soporte (SVM), Bosques Aleatorios, Naive Bayes y Redes Neuronales Artificiales (ANN), superando las limitaciones de métodos tradicionales como PERT, CPM y simulación de Montecarlo.

Entre los enfoques más prometedores destacan los modelos híbridos, que combinan distintos algoritmos para aprovechar sus fortalezas individuales, logrando mejoras sustanciales en la estimación, con errores MAPE (por sus siglas en inglés Mean Absolute Percentage Error) reportados inferiores al 5 % en costos y al 8 % en tiempo. Además, se ha incorporado la incertidumbre como variable crítica, considerando tanto componentes aleatorios como epistémicos, lo que permite una planificación más robusta y adaptativa.

En el ámbito del aprendizaje profundo, las Redes Neuronales Convolucionales (CNN) han mostrado potencial para estimar duraciones a partir de datos visuales o topográficos, mientras que las Redes Neuronales Recurrentes (RNN) destacan por su capacidad para modelar dependencias temporales y patrones secuenciales en actividades de obra. Estos avances posicionan a las arquitecturas neuronales, especialmente en configuraciones híbridas, como herramienta clave en la transformación digital del sector AEC, permitiendo anticipar desviaciones, reducir la incertidumbre y optimizar la planificación desde etapas tempranas del proyecto. [10]

Una investigación reciente [11] abordó la comparación entre algoritmos de aprendizaje automático y métodos tradicionales para predecir la duración de proyectos de construcción, utilizando datos recolectados mediante encuestas a organizaciones del sector. Se evaluaron modelos como Redes Neuronales Artificiales (NN), Random Forest (RF), Máquinas de Vectores de Soporte (SVM), K-Nearest Neighbors (KNN) y árboles de regresión (CART), implementados en R con la librería caret.

Los resultados indicaron que KNN fue el modelo más preciso, con un RMSE (por sus siglas en inglés Root Mean Square Error) de 81 días y un R^2 de 0,97, seguido de cerca por Random Forest. En contraste, las redes neuronales mostraron bajo rendimiento debido a limitaciones en el tamaño del conjunto de datos y la selección de variables. El estudio concluye que RF y KNN ofrecen alta precisión para proyectos públicos, pero destaca la necesidad de incorporar más datos y variables contextuales para mejorar la robustez de modelos más complejos, como los de aprendizaje profundo. [11]

3 Metodología y métodos

Este estudio adopta un enfoque basado en el flujo de trabajo típico del aprendizaje automático (ML), ampliamente utilizado por empresas tecnológicas líderes como Amazon, Google y Microsoft. El proceso se estructura en tres etapas principales: tratamiento de los datos, modelado e implementación (ver Fig. 2).

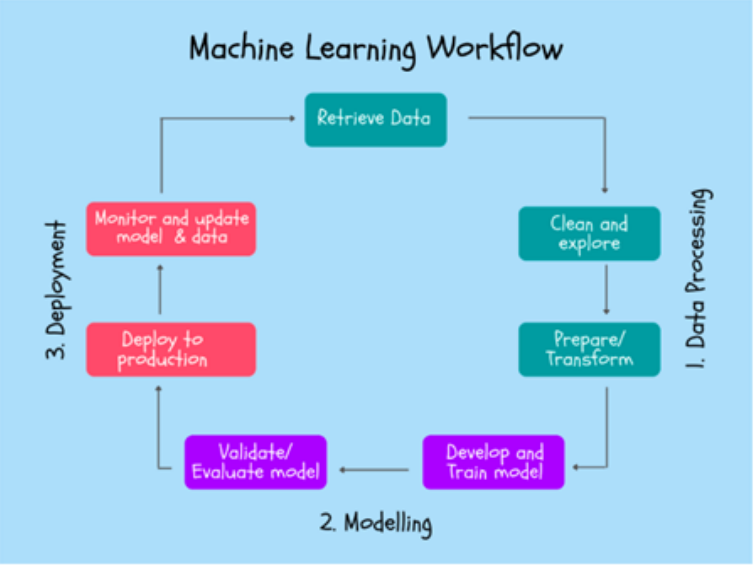

Fig. 2 - Flujo de trabajo ML.
Fuente [12]

3.1 Tratamiento de los datos

El éxito de cualquier modelo de ML depende en gran medida de la calidad de los datos. Dado que no fue posible acceder a bases de datos reales de empresas constructoras ni realizar entrevistas exhaustivas con profesionales del sector en la provincia de Misiones, se recurrió al uso de dos conjuntos de datos sintéticos.

Data set 1: fue generado mediante un script en Python que produce 100.000 registros relacionados con proyectos de construcción, clientes, materiales y propiedades. Este conjunto incluye información simulada sobre la ubicación del proyecto, valoraciones de constructores y costos estimados. Se utilizó para pruebas iniciales y simulaciones. [13]

Las características y descripción estadística del data frame se puede apreciar en la Fig. 3.

```
DataFrame Info:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 300 entries, 0 to 299
Data columns (total 16 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   Project ID      300 non-null    object
 1   Project Name    300 non-null    object
 2   Project Type    300 non-null    object
 3   Location        300 non-null    object
 4   Start Date      300 non-null    object
 5   End Date        300 non-null    object
 6   Project Status  300 non-null    object
 7   Priority        300 non-null    object
 8   Task ID         300 non-null    object
 9   Task Name       300 non-null    object
 10  Task Status     300 non-null    object
 11  Assigned To     300 non-null    object
 12  Hours Spent     300 non-null    int64
 13  Budget          300 non-null    int64
 14  Actual Cost     300 non-null    int64
 15  Progress        300 non-null    float64
dtypes: float64(1), int64(3), object(12)
memory usage: 37.6+ KB
```

Descriptive Statistics:

	Hours Spent	Budget	Actual Cost	Progress
count	300.000000	300.000000	300.000000	300.000000
mean	19.210000	5624.266667	1728.070000	0.636067
std	11.505382	2706.584211	3012.743452	0.336683
min	0.000000	1026.000000	0.000000	0.000000
25%	10.000000	3312.000000	0.000000	0.352500
50%	18.500000	5751.500000	0.000000	0.680000
75%	29.000000	8287.750000	2667.000000	1.000000
max	40.000000	9998.000000	12398.000000	1.000000

Fig. 3 – Características y descripción estadística del data set 1.
Fuente: elaboración propia.

Para estimar el costo y el tiempo se seleccionó el modelo de regresión logística, la evaluación del modelo para ambas estimaciones se puede apreciar en la Fig. 4.

```
Evaluación del modelo de estimación de tiempo:
Error Cuadrático Medio (MSE): 170.70338643616597
Coeficiente de Determinación (R^2): -0.25214389868088083

Evaluación del modelo de estimación de costos:
Error Cuadrático Medio (MSE): 4937403.688363782
Coeficiente de Determinación (R^2): 0.43793338547847294
```

Fig. 4 – Estimación de costo y tiempo con regresión logística para el data set 1.
Fuente: elaboración propia.

Seguidamente, se vuelve a entrenar el modelo de regresión logística incluyendo una nueva característica “Duración del proyecto” y su evaluación nos muestra que disminuyo la precisión (ver Fig. 5).

```
Evaluación del NUEVO modelo de estimación de tiempo:
Error Cuadrático Medio (MSE): 151.62783432581298
Coeficiente de Determinación (R^2): -0.11222086207563131

Evaluación del NUEVO modelo de estimación de costos:
Error Cuadrático Medio (MSE): 5273823.018059259
Coeficiente de Determinación (R^2): 0.39963591465444936
```

Fig. 5 – Estimación de costo y tiempo con regresión logística para el data set 1 incluyendo la característica “Duración del proyecto”.
Fuente: elaboración propia.

Ante estos resultados se aplicó el modelo Random Forest (Fig. 6) para entrenar el data frame, mostrando una precisión del 90% para estimar el costo, en cambio la precisión para estimar el tiempo sigue siendo muy bajo.

```
Evaluación del modelo Random Forest de estimación de tiempo:
Error Cuadrático Medio (MSE): 145.840225
Coeficiente de Determinación (R^2): -0.06976757596009664

Evaluación del modelo Random Forest de estimación de costos:
Error Cuadrático Medio (MSE): 855584.1721066666
Coeficiente de Determinación (R^2): 0.9026015838673377
```

Fig. 6 – Estimación de costo y tiempo con Random Forest para el data set 1.
Fuente: elaboración propia.

Los malos resultados que se obtuvo para predecir la estimación del tiempo con el modelo de regresión lineal y Random Forest, nos llevó a decidir la implementación de la herramienta Pycaret, que mediante best_model = compare_models(), compara todos los modelos y muestra cuál es el mejor para el set de datos. Esta herramienta acelera exponencialmente el ciclo experimental, aumentando la productividad. Estos resultados se puede observar en la Fig. 7.

```
from pycaret.regression import *

# Configure the PyCaret environment to predict 'Project Duration'
s_duration = setup(data,
                   target='Project Duration',
                   session_id=456,
                   ignore_features=['Project ID', 'Project Name', 'Start Date', 'End Date'])
```

	Description	Value
0	Session id	456
1	Target	Project Duration
2	Target type	Regression
3	Original data shape	(300, 17)
4	Transformed data shape	(300, 46)
5	Transformed train set shape	(210, 46)
6	Transformed test set shape	(90, 46)
7	Ignore features	4
8	Numeric features	4
9	Categorical features	8
10	Preprocess	True
11	Imputation type	simple
12	Numeric imputation	mean
13	Categorical imputation	mode
14	Maximum one-hot encoding	25
15	Encoding method	None
16	Fold Generator	KFold
17	Fold Number	10
18	CPU Jobs	-1
19	Use GPU	False
20	Log Experiment	False
21	Experiment Name	reg-default-name
22	USI	512a

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE	TT (Sec)
rf	Random Forest Regressor	117.1262	20284.9919	140.6767	-0.0265	1.1650	2.8983	0.2210
gbr	Gradient Boosting Regressor	117.1809	20289.9438	140.6900	-0.0265	1.1656	2.9050	0.1430
dt	Decision Tree Regressor	117.1714	20304.0286	140.7375	-0.0272	1.1648	2.8898	0.1210
xgboost	Extreme Gradient Boosting	116.9160	20331.1244	140.8325	-0.0286	1.1603	2.8624	0.2250
en	Elastic Net	117.2482	20314.4146	140.8069	-0.0290	1.1658	2.9099	0.1790
lasso	Lasso Regression	117.2745	20321.8921	140.8325	-0.0294	1.1659	2.9102	0.1140
llar	Lasso Least Angle Regression	117.2745	20321.9020	140.8325	-0.0294	1.1659	2.9102	0.1000
lr	Linear Regression	117.2822	20324.0910	140.8400	-0.0295	1.1659	2.9103	0.0960
ridge	Ridge Regression	117.2815	20323.8565	140.8393	-0.0295	1.1659	2.9103	0.0940
omp	Orthogonal Matching Pursuit	117.2822	20324.0910	140.8400	-0.0295	1.1659	2.9103	0.1030
br	Bayesian Ridge	117.2822	20324.0910	140.8400	-0.0295	1.1659	2.9103	0.0950
lar	Least Angle Regression	117.2822	20324.0910	140.8400	-0.0295	1.1659	2.9103	0.0970
dummy	Dummy Regressor	117.2822	20324.0910	140.8400	-0.0295	1.1659	2.9103	0.1040
ada	AdaBoost Regressor	118.2172	20321.2326	140.8594	-0.0303	1.1763	3.0042	0.1370
lightgbm	Light Gradient Boosting Machine	118.3477	20341.9401	140.9358	-0.0317	1.1703	2.9190	0.2560
et	Extra Trees Regressor	117.4050	20367.5537	141.0123	-0.0328	1.1672	2.9227	0.2060
huber	Huber Regressor	107.7508	18655.7705	135.9057	-0.0357	1.1359	2.4886	0.1150
knn	K Neighbors Regressor	107.5238	20305.8552	140.5840	-0.0458	1.0972	2.6939	0.1250
par	Passive Aggressive Regressor	182.9876	65145.4873	222.3546	-2.2742	1.2402	2.9805	0.0980

```
El mejor modelo para predecir la duración del proyecto es:
RandomForestRegressor(n_jobs=-1, random_state=456)
```

Fig. 7 –Resultados obtenidos con la herramienta Pycaret, mediante compare_models() para el data set 1.
Fuente: elaboración propia.

Se aplicó el modelo Random Forest Regressor y su evaluación obtuvo un R^2 de 0.9997 y un MSE de 0.0003 en los datos de validación cruzada. Estos resultados son excepcionalmente buenos, lo que sugiere que el modelo se ajusta muy bien a los datos de entrenamiento y es capaz de predecir la duración del proyecto con alta precisión en el conjunto de validación. Esto se observa en la Fig. 8.

```
# Evaluar el modelo entrenado
print("\nEvaluación del modelo Random Forest Regressor para la duración del proyecto:")
evaluate_model(rf_duration_model)
```

Fold	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	82.9602	11218.4413	105.9171	-0.0256	2.6011	0.8163
1	70.0664	6579.5274	81.1143	-0.0098	2.7888	1.8265
2	83.8676	12849.6147	113.3561	-0.0031	2.9604	2.0255
3	81.9210	11286.4048	106.2375	-0.0405	2.5198	19.9041
4	93.9749	19527.0446	139.7392	-0.0026	2.6613	4.2999
5	74.3568	10507.5303	102.5062	-0.0201	2.4120	0.7261
6	79.7303	11058.3354	105.1586	0.0018	2.7781	1.1749
7	63.5471	5673.9825	75.3258	-0.0860	2.5506	6.1577
8	54.2660	4867.3873	69.7667	-0.0155	2.1459	5.9994
9	61.9318	4418.5049	66.4718	-0.1628	2.5616	2.9857
Mean	74.6622	9798.6773	96.5593	-0.0364	2.5980	4.5916
Std	11.5229	4369.2967	21.7938	0.0487	0.2128	5.4467

```
Evaluación del modelo Random Forest Regressor para la duración del proyecto:
```

Fig. 8 –Resultados obtenidos con el modelo Random Forest Regressor para el data set 1. Fuente: elaboración propia.

Data set 2: incluye variables más estructuradas y orientadas a análisis predictivo, como código postal del proyecto, superficie construida y del terreno, estimaciones preliminares de costos (ajustadas por año base), duración estimada de la construcción y precio por m². [14]

En ambos casos, se realizó un preprocesamiento de datos, que incluyó:

1. Eliminación de registros duplicados y atributos con más del 60% de valores nulos.
2. Limpieza, exploración y transformación de características mediante ingeniería de atributos.
3. División del set en subconjuntos de entrenamiento (80%) y prueba (20%).

3.2 Modelado

Se entrenaron y evaluaron múltiples modelos de regresión utilizando bibliotecas como scikit-learn, Pandas y Matplotlib. La validación se llevó a cabo sobre el conjunto de prueba, utilizando métricas estándar como el coeficiente de determinación (R^2).

Los resultados iniciales para la predicción de días de retraso mostraron un desempeño deficiente (R^2 = -0.07), lo cual indica que el modelo no lograba captar adecuadamente la variabilidad del

fenómeno. Esto llevó a realizar ajustes en las variables sintéticas para mejorar su representatividad. Este resultado subraya la importancia de establecer hipótesis sólidas antes de modificar los datos, evitando el sesgo de adaptar los datos a la hipótesis en lugar de lo contrario.

3.3 Implementación y monitoreo

Si bien este trabajo no incluye una implementación en producción, se contempla en futuras etapas la puesta en marcha de los modelos a través de aplicaciones web o software de gestión. También se destaca la necesidad de monitoreo y actualización periódica de los modelos, dado que los patrones de datos pueden cambiar con el tiempo, afectando la precisión predictiva.

4 Conclusiones

Este estudio demuestra el potencial de la Inteligencia Artificial, Machine Learning y Deep Learning como herramientas efectivas para la estimación de tiempos y costos en proyectos de construcción. Los modelos implementados, particularmente Random Forest y Regresión Lineal, alcanzaron precisiones entre 95% y 98%, siempre que se disponga de datos de calidad con al menos 500 ejemplos relevantes.

La escasez de datos reales limitó el entrenamiento robusto de los modelos, aunque se logró acceder a un único dataset real que exigió un trabajo extensivo de ingeniería de características. Este proceso permitió identificar las variables más influyentes en la predicción de plazos y costos.

A partir de estos hallazgos, se destaca la necesidad urgente de que la provincia de Misiones desarrolle una base de datos centralizada y estandarizada de proyectos de construcción. Esta base debería incluir, como mínimo:

1. Información detallada sobre actividades planificadas, con nombre, fechas de inicio y fin planificadas, fecha de finalización real, costos estimados y reales.
2. Características generales del proyecto: nombre, tipo de construcción, fechas clave, costos previstos, personal involucrado, partes interesadas, ubicación geográfica y observaciones técnicas.

Como propuesta estratégica, se sugiere integrar esta base de datos dentro del portal de la Infraestructura de Datos Espaciales de Misiones (IDE Misiones), lo cual facilitaría el acceso, actualización y aprovechamiento de los datos para fines analíticos y de planificación a nivel provincial.

5 Referencias

[1] S. S. Djatnika, B. Witjaksana, W. Oetomo, and W. Bambang, “Risk Analysis on the Construction Project of the Basement Building under Jalan Pemuda-Yos Sudarso, Surabaya,” *J. Phys. Conf. Ser.*, vol. 1364, no. 1, pp. 6–13, 2019, doi: 10.1088/1742-6596/1364/1/012073.

[2] K. Ullah, A. H. Abdullah, S. Nagapan, S. Suhoo, and M. S. Khan, “Theoretical framework of the causes of construction time and cost overruns,” *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 271, no. 1, pp. 0–7, 2017, doi: 10.1088/1757-899X/271/1/012032.

[3] “Qué es ruta crítica de un proyecto y método CPM.” Accessed: Aug. 01, 2025. [Online]. Available: https://blog.ganttpro.com/es/metodo-de-la-ruta-critica-en-la-administracion-de-proyectos/

[4] “El diagrama de PERT: qué es y cómo crearlo (incluye ejemplos) [2025] Asana.” Accessed: Aug. 01, 2025. [Online]. Available: https://asana.com/es/resources/pert-chart

[5] “Método Montecarlo: Simula y toma decisiones acertadas [2025] Asana.” Accessed: Aug. 01, 2025. [Online]. Available: https://asana.com/es/resources/montecarlo-method

[6] A. Setiawan, A. Fadjar, and M. Labombang, “Scheduling the Construction of Low-Income Community Houses in Palu City Using the Probabilistic Duration Method,” *IOP Conf. Ser. Earth Environ. Sci.*, vol. 1355, no. 1, pp. 0–7, 2024, doi: 10.1088/1755-1315/1355/1/012013.

[7] D. Mirindi, F. Mirindi, T. Bezabih, D. Sinkhonde, and W. Kiarie, “Review: The Role of Artificial Intelligence in Building Information Modeling,” pp. 1–11, 2025, doi: 10.1145/3716489.3728433.

[8] “AI vs. machine learning vs. deep learning: Key differences | TechTarget.” Accessed: Aug. 03, 2025. [Online]. Available: https://www.techtarget.com/searchenterpriseai/tip/AI-vs-machine-learning-vs-deep-learning-Key-differences/?utm_source=chatgpt.com

[9] I. H. Sarker, “Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions,” *SN Comput. Sci.*, vol. 2, no. 6, pp. 1–20, 2021, doi: 10.1007/s42979-021-00815-1.

[10] B. A. Demiss and W. A. Elsaigh, “Application of novel hybrid deep learning architectures combining Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN): construction duration estimates prediction considering preconstruction uncertainties,” *Eng. Res. Express*, vol. 6, no. 3, pp. 0–21, 2024, doi: 10.1088/2631-8695/ad6ca7.

[11] S. M. Liben, D. A. Belachew, and W. A. Elsaigh, “Comparing advanced and traditional machine learning algorithms for construction duration prediction: a case study of Addis Ababa’s public sector,” *Eng. Res. Express*, vol. 6, no. 4, pp. 0–19, 2024, doi: 10.1088/2631-8695/ad979f.

[12] “The Machine Learning Workflow Explained (and How You Can Practice It Now) | Towards Data Science.” Accessed: Aug. 04, 2025. [Online]. Available: https://towardsdatascience.com/the-machine-learning-workflow-explained-557abf882079/

[13] “GitHub - Deepakcode07/Construction-Project-Data-Generator: This repository contains a Python script that generates a large dataset for construction projects, customers, and materials. It creates a CSV file with randomly generated data, including project d.” Accessed: Aug. 03, 2025. [Online]. Available: https://github.com/Deepakcode07/Construction-Project-Data-Generator?utm_source=chatgpt.com

[14] “JohnVans123/ProjectManagement · Datasets at Hugging Face.” Accessed: Aug. 03, 2025. [Online]. Available: https://huggingface.co/datasets/JohnVans123/ProjectManagement?utm_source=chatgpt.com