Análisis y Proyección de Costos en Proyectos de Vivienda VIS y NO VIS Desarrollados por la Constructora Vértice Ingeniería y Construcción en el Área Metropolitana del Valle de Aburrá durante el Período 2016-2023

Ricardo Lastra Lopera

Eladio Yovera Yovera

Trabajo de Grado para optar al título de Especialista en Analítica y Ciencia de Datos

Asesor

Martha Lucia Rodriguez Lopez, PhD en Ingeniería Electrónica y Computación

Universidad de Antioquia

Facultad de Ingeniería

Especialización en Analítica y Ciencia de Datos

Medellín

2024

Cita	Lastra Lopera y Yovera Yovera, 2024 [1]
Referencia Estilo IEEE (2020)	[1] Lastra Lopera, R., y Yovera Yovera, E. "Análisis y Proyección de Costos en Proyectos de Vivienda VIS y NO VIS Desarrollados por la Constructora Vértice Ingeniería y Construcción en el Área Metropolitana del Valle de Aburrá durante el Período 2016-2023", Presencial, Especialización en Analítica y Ciencia de Datos, Universidad de Antioquia, Medellín, 2024.

Especilización en Analítica y Ciencia de Datos, Cohorte VII.

Universidad de Antioquia - www.udea.edu.co

Rector: John Jairo Arboleda Céspedes.
Decano/Director Julio César Saldarriaga Molina.
Jefe departamento: Diego José Luis Botia.

***Abstract*—The paper aims to estimate the cost of VIS and NON VIS housing projects using historical data from Constructor Vértice Ingeniería and Construcción, covering the period from 2016 to 2023. This study investigated the predictive power of a Random Forest Regressor for estimating total project costs in real estate development. Data preprocessing involved handling missing values, encoding categorical variables, and aggregating relevant features to create a comprehensive dataset. The Random Forest model, optimized through GridSearchCV, leveraged its ability to handle non-linear relationships and interactions between variables. The model evaluation metrics, including R-squared (R^2) and root mean square error (RMSE), revealed an R^2 value (0.938) and an RMSE of 129,195,291, demonstrating its strong predictive performance, which for a VIS project on average has a cost of 11,000,000,000 Colombian pesos and for a NON VIS project has an average cost of 40,000,000,000 Colombian pesos, it should be noted that a low RMSE indicates a better fit of the model. This represents a significant improvement over previous linear regression approaches. Future work will explore the inclusion of additional variables, such as macroeconomic indicators, and compare the Random Forest model to other ensemble methods like Gradient Boosting to further enhance predictive accuracy and applicability in the real estate sector.**

***Index Terms*—Regresión, Outliers, Machine, GridSearch, Random Forest Regression, Vivienda, VIS, NO VIS.**

I. Introducción

La industria de la construcción en Colombia enfrenta un desafío crítico en la estimación precisa de costos de proyectos, lo que genera repercusiones negativas en el sector. De acuerdo con un estudio de la Constructora Contex, en los últimos cinco años, la industria ha experimentado un panorama complejo debido a factores como la planeación ineficiente de proyectos, altas tasas de interés, disminución de subsidios y aumento en el precio de materiales [1]. Esto ha resultado en la liquidación de aproximadamente 150 constructoras y la reestructuración de 366 más por parte de la Superintendencia de Sociedades - SuperSociedades [1]. Por esta razón, se entiende que la estimación de costos en la construcción es un proceso complejo y sujeto a una gran cantidad de incertidumbres, lo que genera grandes retos, tales como:

- Sobrecostos: dónde el costo real de los proyectos supera significativamente lo presupuestado.

- Retrasos: dónde la imprecisión en la estimación de costos conduce a retrasos en la ejecución de los proyectos.

En este proyecto se realizó el análisis de un conjunto de datos históricos y la aplicación de técnicas de aprendizaje automático, se desarrolló modelos predictivos más precisos y confiables para estimar costos de manera eficiente. Asimismo, los datos fueron obtenidos de la Constructora desde el año 2016 hasta el año 2023. Las métricas de desempeño podrían incluir la precisión del modelo y el valor del costo de los proyectos futuros con características similares a los actuales.

II. Background

A continuación, se realiza la definición de los términos clave utilizados en el presente artículo, los cuales son fundamentales para comprender las técnicas y metodologías aplicadas en el análisis de datos y Machine learning.

- Coeficiente de determinación (R^2): Es una métrica estadística que indica la proporción de la variabilidad en la variable dependiente que es explicada por el modelo de regresión. Su valor oscila entre 0 y 1, donde un valor cercano a 1 indica que el modelo explica la mayor parte de la variabilidad observada en los datos, mientras que un valor cercano a 0 sugiere que el modelo explica poco de dicha variabilidad. [2].
- Root Mean Squared Error (RMSE): Es una métrica que mide la magnitud promedio del error en las predicciones de un modelo, expresado en las mismas unidades que la variable objetivo. Es ampliamente utilizado para evaluar modelos de regresión porque penaliza más los errores grandes que los pequeños. Un RMSE bajo indica un mejor ajuste del modelo. [3].
- La distribución de los residuos (Residuals Distribution) se refiere al análisis de los errores residuales de un modelo estadístico o de aprendizaje automático. Un residuo es la diferencia entre el valor observado (yi) y el valor predicho (ŷi) por el modelo. El análisis de la distribución de los residuos es fundamental para evaluar la calidad del modelo y validar sus supuestos. Idealmente, en un modelo bien ajustado, los residuos deben:
  - Tener una distribución aproximadamente normal.
  - Estar centrados en torno a cero.
  - No mostrar patrones (indicando homocedasticidad: varianza constante).
  - Ser independientes entre sí.

  Una distribución no normal de residuos, patrones en un gráfico de dispersión o varianza no constante puede sugerir problemas como mala especificación del modelo, presencia de relaciones no lineales o datos atípicos [4].
- Codificación de Variables Categóricas: Es el proceso de transformar datos categóricos (como nombres o categorías) en un formato numérico que los modelos de Machine Learning puedan procesar. Ejemplos de técnicas de codificación incluyen ”One-Hot Encoding” y ”Label Encoding” [5].
- Valores Atípicos: También conocidos como ”outliers”, son observaciones en los datos que se desvían significativamente de la tendencia general. Pueden ser resultado de errores en los datos o eventos excepcionales, y es importante tratarlos adecuadamente para evitar que distorsionen los resultados del modelo [5].
- Planificación Financiera en Proyectos de Construcción: Proceso mediante el cual se estiman, asignan y controlan los recursos financieros necesarios para la ejecución de proyectos de construcción. Involucra predicción de costos, presupuestos y análisis de rentabilidad [6].
- Insumos en Proyectos de Construcción: Son los materiales, mano de obra, equipos y otros recursos necesarios para la ejecución de un proyecto de construcción. El análisis detallado de insumos específicos ayuda a entender los costos y optimizar la planificación [6].
- Estimación de Costos: Proceso de calcular el costo aproximado de un proyecto basado en datos históricos,

especificaciones y modelos predictivos. Es clave para evitar sobrecostos y optimizar recursos en la construcción [6].

- Periodo de Referencia de Datos: Intervalo de tiempo específico durante el cual se recopilan datos históricos para su análisis. En este caso, comprende de 2016 a 2023, permitiendo identificar tendencias y patrones relevantes.
- VIS: Corresponde a las Viviendas de Interes Social. Esta viviendas generalmente son las más economicas dentro del segmento, y tambien lo usuarios pueden aplicar a subsidios etc [7].
- NO VIS: Corresponde a las Viviendas de NO Interes Social. Esta viviendas generalmente son las más costosas dentro del segmento [7].

III. Estado del Arte

La estimación precisa de costos en proyectos de construcción es fundamental para la planificación y ejecución eficiente. Diversos estudios han explorado la aplicación de técnicas de Machine Learning para mejorar la precisión en la predicción de costos.

Un estudio de 2020 realizó una revisión sistemática de técnicas de Machine Learning aplicadas a la estimación de costos en proyectos de construcción, destacando la necesidad de conjuntos de datos adecuados para modelar y prever costos con precisión [8].

En 2007, una investigación en la Universidad de Chile desarrolló un modelo basado en redes neuronales para predecir variaciones de costos en proyectos de construcción habitacional en altura. Este estudio también implementó modelos de regresión lineal múltiple para comparación, encontrando que las redes neuronales ofrecían una mayor precisión en las predicciones [9].

Un estudio de 2023, realizó un análisis y predicción del precio de la vivienda en Madrid implementando varios modelos de Machine Learning, incluyendo Random Forest, para predecir precios de viviendas en Madrid. Los resultados destacan que Random Forest ofrece una precisión notable en las predicciones, siendo uno de los modelos más efectivos en el estudio [10].

En 2024, una investigación en la Universidad Pontificia de Comillas desarrolló modelos de Predicción de precios Inmobiliarios utilizando técnicas de Machine Learning Este estudio desarrolló modelos predictivos para estimar precios de venta de viviendas en Madrid, utilizando la técnica de Random Forest. Los resultados indican que Random Forest es uno de los modelos con mejor desempeño en términos de precisión y error [11].

Además, se han desarrollado tutoriales que demuestran la aplicación de la regresión para la predicción de precios, utilizando herramientas como Model Builder para entrenar modelos de Machine Learning y evaluar su precisión [12].

Estos estudios resaltan la importancia de seleccionar variables relevantes y aplicar técnicas de preprocesamiento de datos para mejorar la precisión de los modelos predictivos en la estimación de costos de construcción. La reestructuración de variables y el enfoque en insumos específicos, como se propone en el presente proyecto, están alineados con las mejores prácticas identificadas en la literatura existente.

IV. Metodología

A. Dataset

Durante la identificación de fuentes de información, resulta que los datos de los proyectos se encuentran en una base de datos SQL Server On Premise, distribuidos en cuatro tablas, que al final eran parte de un modelo de datos tipo estrella (tres dimensiones y una tabla de hechos). La data comprende información desde el año 2016 hasta el año 2023. Para su uso se descargaron a través de archivos en formato CSV y en conjunto ocupan 500 MB de información. Los archivos contienen información de Nombre de Proyectos que incluyen 268 registros, Insumos que incluyen 110000 registros, Tipos de Costos que incluyen 3227 registros y finalmente la tabla de hechos Control Proyectos donde se encuentran los campos para relacionar las dimensiones y el valor del costo del proyecto, incluye un millón de registros. El modo de acceso a la data es mediante una máquina virtual, para ello se debe tener las credenciales que otorga la Constructora Vértice Ingeniería y Construcción para poder acceder. Las tablas son las siguientes: Proyectos, Insumos, Tipos de Costo, Control de Proyectos.

Dado que los datos originales están en diferentes tablas, se procede a realizar un proceso de ingeniería de datos para consolidar la información en un solo dataset, para ello se inicia con la lectura de los archivos en un notebook de la plataforma Databricks, luego se realiza joins entre las tablas, teniendo como base la tabla de hechos. Una vez realizado los joins, seleccionamos los campos que componen el dataset final. Los campos del dataset son los siguientes:

TABLE I
Estructura del Dataset a usar

Nombre Columna	Tipo Dato	Descripción
$\text{str}_{p}royecto$	varchar	Nombre del proyecto
$\text{str}_{v}is$	varchar	El proyecto es de tipo VIS (SI o NO)
$\text{str}_{i}nsumo_{d}etalle$	varchar	Descripción detallada asignada al insumo
$\text{str}_{i}nsumo_{a}grupado$	varchar	Descripción de la agrupación asignada al insumo
$\text{str}_{t}ipo_{c}osto$	varchar	Descripción del tipo de costo
$\text{val}_{t}otal_{i}nsumo$	int	Valor total del insumo

El dataset utilizado incluye datos de 64 registros relacionados con proyectos inmobiliarios. Las principales variables son:

- Variables categóricas: Nombre del proyecto, tipo de insumo, categoría del insumo, entre otras.
- Variable objetivo: val_total_insumo, que representa el costo total de los insumos en cada proyecto

B. Modelo Random Forest Regressor

El modelo Random Forest Regressor es un algoritmo de aprendizaje supervisado basado en el ensamblado de árboles de decisión. Su objetivo es realizar predicciones precisas al combinar los resultados de múltiples árboles de decisión,

lo que reduce problemas como el sobreajuste y mejora la generalización del modelo [13].

Hiperparámetros principales

- n_estimators (Número de árboles): Determina la cantidad de árboles en el bosque. Un mayor número mejora la estabilidad, pero incrementa el tiempo de entrenamiento.
- max_depth (Profundidad máxima): Controla la profundidad máxima de los árboles para evitar árboles extremadamente específicos.
- min_samples_split (Muestras mínimas para dividir): Número mínimo de muestras requeridas para dividir un nodo.
- min_samples_leaf (Muestras mínimas en una hoja): Ayuda a controlar el crecimiento excesivo de los árboles.
- max_features (Características por nodo): Especifica cuántas características considerar en cada división. [13].

El modelo principal usado en este proyecto fue Random Forest Regressor, optimizado con GridSearchCV. Este modelo no lineal es robusto frente a relaciones complejas entre variables, valores atípicos y multicolinealidad. Asimismo, fue elegido debido a:

- Complejidad de las relaciones entre variables: Los costos de proyectos inmobiliarios involucran múltiples factores no lineales.

- Robustez frente a multicolinealidad: A diferencia de la regresión lineal, Random Forest no se ve afectado por la alta correlación entre variables predictoras.

- Capacidad para manejar datos heterogéneos: Incluye tanto variables categóricas como numéricas sin necesidad de transformaciones exhaustivas.

En general, el Random Forest Regressor es una herramienta poderosa para abordar problemas de predicción en escenarios complejos y con ruido, como el análisis de costos inmobiliarios. Su capacidad para capturar relaciones no lineales lo hace ideal en contextos donde la simplicidad de los modelos lineales no es suficiente.

El proceso desarrollado para la implementación del modelo de forma general sigue los siguientes pasos:

- Preprocesamiento de datos:
  - Codificación de variables categóricas.
  - Imputación de valores faltantes y eliminación de duplicados.
  - Creación de una nueva variable objetivo: total_cost_per_project.
- División de datos:
  Separación en conjuntos de entrenamiento (80%) y prueba (20%).
- Optimización de hiperparámetros:
  Evaluación de configuraciones para n_estimators, max_depth, min_samples_split, y min_samples_leaf mediante validación cruzada.
- Entrenamiento y evaluación:
  - Ajuste del modelo en el conjunto de entrenamiento.
  - Evaluación en el conjunto de prueba con métricas como R^2 y RMSE.

V. Resultados y Discusión

A. Validacion del Modelo

El modelo Random Forest Regressor, optimizado mediante GridSearchCV, alcanzó un coeficiente de determinación R^2 de 0.938, este alto valor de R^2 indica que el modelo es capaz de explicar el 93.8% de la variabilidad en los costos de los proyectos inmobiliarios.

En cuanto al error cuadrático medio raíz (RMSE), el valor obtenido fue 129,195,291, lo que sugiere un buen nivel de precisión en las predicciones, considerando la magnitud de los costos asociados a proyectos inmobiliarios, el cual para un proyecto VIS en promedio tiene un costo de 11,000,000,000 pesos colombianos y para un proyecto NO VIS tiene un costo promedio de 40,000,000,000 pesos colombianos. Cabe destacar Un RMSE bajo indica un mejor ajuste del modelo.

Asimismo, en la Figura 1 se muestra el resultado de la Distribución Residual, donde se evidencia que esta centrado en torno a cero y tiene una distribución aproximadamente normal, lo que evidencia la buena calidad del modelo empleado.

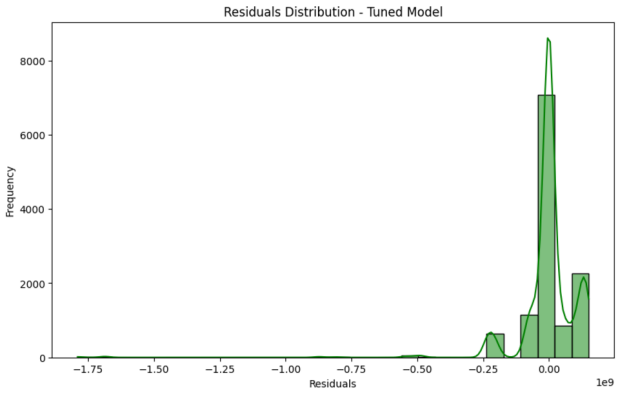

Fig. 1. Distribucion de Residuos.

La inclusión de un modelo no lineal como Random Forest permitió capturar interacciones más complejas entre las variables predictoras, lo que contribuyó al incremento en la precisión del modelo.

B. Desempeño del Modelo

La evaluación del modelo en el conjunto de prueba evidenció predicciones consistentes y ajustadas a los valores reales.

La Figura 2 compara los valores reales y predichos del conjunto de prueba. El gráfico muestra una alineación entre ambas series, destacando la precisión del modelo en la mayoría de los casos.

C. Discusión

El uso de Random Forest Regressor demostró ser una decisión acertada debido a su capacidad para capturar interacciones complejas entre las variables y manejar datos con ruido. Sin embargo, el RMSE aún puede mejorar con la integración de más variables explicativas, como indicadores

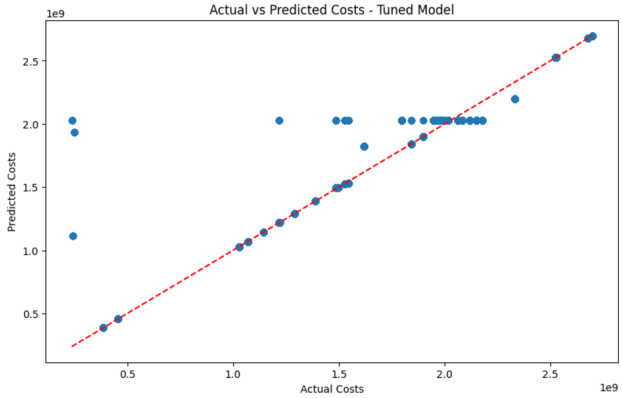

Fig. 2. Valores reales vs. predichos

macroeconómicos, y mediante la exploración de técnicas como Gradient Boosting (XGBoost o LightGBM).

El enfoque de este estudio destaca la importancia de utilizar modelos no lineales y estrategias avanzadas de preprocesamiento en problemas complejos como la estimación de costos. Además, el análisis de residuos y multicolinealidad contribuyó a garantizar la validez y estabilidad del modelo.

Conclusiones

Los resultados obtenidos con el modelo Random Forest Regressor evidencian un avance significativo en la capacidad de predecir costos en proyectos inmobiliarios. Con un R^2 de 0.938, el modelo establece una base sólida para su implementación en entornos productivos.

El análisis sugiere que la consideración de modelos no lineales es crucial para abordar la alta complejidad de los datos en el sector de la construcción. Además, la integración de técnicas avanzadas de preprocesamiento y optimización de hiperparámetros es fundamental para maximizar la capacidad predictiva.

En futuros trabajos, se recomienda explorar modelos de Gradient Boosting, como XGBoost o LightGBM, para comparar su desempeño con el modelo actual y continuar mejorando las predicciones. Asimismo, la incorporación de más variables relevantes, como indicadores económicos o geográficos, podría enriquecer aún más el modelo y su aplicabilidad.

References

[1] M.Garcia, “Las constructoras que no aguantaron la crisis y terminaron en la quiebra”, Las2Orillas, 2024. [En línea]. Disponible en: https://www.las2orillas.co/las-constructoras-que-no-aguantaron-la-crisis-y-terminaron-en-la-quiebra/ [Accedido: 25 Sep. 2024]

[2] N. R. Draper y H. Smith, Applied Regression Analysis, 3ª ed., Wiley-Interscience, 1998.

[3] J. J. Montero, J. M. Martínez, and F. Montes, ”Evaluating the accuracy of mathematical models: RMSE versus R^2,” Mathematics and Computers in Simulation, vol. 43, no. 4, pp. 495–502, 1997.

[4] C. Chatfield, The Analysis of Time Series: An Introduction, 6ª ed., Boca Raton, FL, USA: Chapman & Hall/CRC, 2003.

[5] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed., Morgan Kaufmann, 2016.

[6] N.Martin, “Estimación de costes en proyectos: técnicas y consejos”, banktrack.com, 2024. [En línea]. Disponible en: https://banktrack.com/blog/estimacion-costes [Accedido: 25 Sep. 2024]

[7] Contex, “Viviendas VIS y No VIS: ¿Cuál es la diferencia?”, Contex.com, 2024. [En línea]. Disponible en: https://contex.com.co/vivienda-vis-y-no-vis/ [Accedido: 01 Nov. 2024]

[8] S.Tayefeh, O.Mahdi “Cost estimation and prediction in construction projects: a systematic review on machine learning techniques”, banktrack.com, 2020. [En línea]. Disponible en: https://link.springer.com/article/10.1007/s42452-020-03497-1 [Accedido: 25 Sep. 2024]

[9] M.Jory “Predicción de las variaciones de costos para proyectos de construcción utilizando Redes Neuronales”, Universidad de Chile, 2007. [En línea]. Disponible en: https://repositorio.uchile.cl/tesis/uchile/2007/cf-jorymr/pdfAmont/cf-jorymr.pdf [Accedido: 26 Sep. 2024]

[10] A.Datsko “Análisis y Predicción del Precio de la Vivienda en Madrid Utilizando Técnicas de Exploración de Datos e Inteligencia Artificial Implementadas en Python”, Universidad Politecnica de Madrid, 2023. [En línea]. Disponible en: https://oa.upm.es/80281/?utm [Accedido: 27 Sep. 2024]

[11] B.Sicilia “Desarrollo de Modelos de Predicción de Precios Inmobiliarios utilizando técnicas de Machine Learning”, Universidad Pontificia de Comillas, 2020. [En línea]. Disponible en: https://repositorio.comillas.edu/xmlui/handle/11531/78964?utm [Accedido: 27 Sep. 2024]

[12] Microsoft, “Tutorial: Predicción de precios mediante regresión con Model Builder”, Learn Microsoft, 2024. [En línea]. Disponible en: https://learn.microsoft.com/es-es/dotnet/machine-learning/tutorials/predict-prices-with-model-builder [Accedido: 25 Sep. 2024]

[13] J.Amat, “Random Forest con Python”, CienciaDeDatos.Net, 2020. [En línea]. Disponible en: https://cienciadedatos.net/documentos/py08_random_forest_python [Accedido: 25 Sep. 2024]