Problem Set 3

John Esteban Londoño & William Aguirre

2024-05-26

Introducción

La inversión más importante para la mayoría de las personas es tener casa propia, y esta es una de las fuentes más comunes en la acumulación de capital. Por esto, conocer cuáles son los elementos que determinan el valor de este bien puede servir de guía para el proceso de planificación financiera.

Para determinar el valor de las viviendas, la literatura ha avanzado en dos aspectos principales: primero, en entender las variables más importantes, y segundo, en identificar el algoritmo que mejor se adapta a este tipo de problemas. En cuanto a las variables, los estudios utilizan las características propias del hogar, como el área construida y la cantidad de habitaciones, además de incluir variables que explican el entorno donde está ubicada la vivienda. En cuanto a los modelos utilizados, es común empezar con el modelo de regresión lineal y luego aumentar la complejidad, llegando hasta los algoritmos basados en árboles.

Para determinar las variables que más aportan a la predicción de precios de la vivienda, estas se clasifican primero en dos grupos: características propias del hogar y características del entorno donde está ubicado el hogar. Las más importantes del primer grupo son: el número de habitaciones, el área del hogar, el piso donde está ubicada, la cantidad de baños, si tiene garaje, la edad de la propiedad, el área construida, si tiene aire acondicionado y el estrato. De este grupo, las variables que más contribuyen a predecir el precio son el número de habitaciones, la cantidad de baños, el área construida y el estrato.

El segundo grupo de variables está relacionado con las características del entorno donde está ubicada la vivienda. En este grupo se utilizan variables como la distancia a estaciones de transporte y de educación, la proporción de personas de bajos recursos que viven cerca, el porcentaje de centros de negocio y la localización. La distancia al transporte es la característica que más peso tiene en este grupo.

El modelo que mejor rendimiento ha presentado para la predicción de los precios de las viviendas es el Random Forest. El trabajo desarrollado por Cueto (2022), utilizando solo características de la vivienda en cinco modelos, entre ellos redes neuronales y modelos basados en árboles, encuentra que el que tiene mejor rendimiento es el Random Forest. De la misma forma, el trabajo de Choy and Ho (2023), en el que además de las características de la vivienda se utiliza la variable de la distancia al transporte, concluye que el mejor modelo es el Random Forest. Alfaro-Navarro et al. (2020), Alzate (2019) y Giraldo (2023.) también encuentran que el Random Forest es el mejor modelo.

Sin embargo, hay otros algoritmos que también han tenido buen desempeño para predecir el precio de las viviendas. En el trabajo de Mohamed, Ibrahim, and Hagras (2023), se destaca el uso de redes neuronales. Bushan (2010) sugiere trabajar con XGBoost, mientras que el modelo Ripper mostró un rendimiento destacado en el trabajo de Park and Bae (2015). Además, tanto el Gradient Boosting como el Random Forest demostraron eficacia en el estudio de Alzate (2019).

El propósito de este artículo es realizar la predicción de precios de vivienda en Bogotá. En este proceso, se utilizan tanto variables propias de la vivienda como variables que describen el entorno donde está ubicada. Además, se hace uso extensivo de técnicas de procesamiento de lenguaje natural para aprovechar las descripciones de las viviendas.

Para este problema, siguiendo los principales resultados de la literatura, los modelos de Random Forest fueron los que mejor predijeron los precios de las viviendas. Sin embargo, los modelos más complejos tuvieron predicciones inferiores en comparación con los primeros. La precisión de las predicciones aumentaba conforme se agregaban más variables de contexto de la vivienda.

Datos

Agregación de las Nuevas Variables

Para modelar el precio de las viviendas en Bogotá, se cuenta con diversas características propias del hogar. En primer lugar, tenemos el precio de la vivienda, que es la variable que se va a modelar. Luego, disponemos de algunos atributos esenciales para iniciar el proceso de modelación, tales como el área de la vivienda, el número de habitaciones para dormir y el número de baños. También se puede caracterizar el tipo de vivienda, diferenciando si es una casa o un apartamento. Además, se tiene a disposición la descripción de la vivienda, de donde se pueden extraer datos que ayuden a describir mejor la propiedad. A partir de esta descripción, se ha agregado el número de habitaciones que tiene la vivienda.

Para conocer mejor el contexto en el que se ubica la vivienda, se creó la variable "estrato" a partir de la información del DANE. En primer lugar, se descargó la cartografía del Marco Geoestadístico Nacional del DANE y el Censo 2018 a nivel de manzana. Luego, se hizo un join espacial entre cada uno de los inmuebles y el Marco Geoestadístico Nacional para identificar la manzana de cada una de las viviendas. Posteriormente, con el código de manzana, se realizó un join con el Censo 2018 para identificar el estrato predominante.

La percepción de seguridad también es un aspecto clave para determinar la disposición a pagar por una vivienda. Se espera que, si hay mayor seguridad, el precio de la vivienda sea mayor. Para medir este fenómeno, se incluyeron una serie de variables relevantes. Con las cifras oficiales de la alcaldía, se obtuvo información geolocalizada de hurtos a personas y viviendas. Esta información solo está disponible a nivel de localidades y no georreferenciada de manera similar a la base de datos. Por ello, se utilizó la distancia para integrar la información de hurtos más cercanos. Además, a través de la información de Google, se calculó la distancia entre las viviendas y las estaciones de policía.

En resumen, para modelar el precio de las viviendas en Bogotá, se consideran variables propias de la vivienda, como el área, el número de habitaciones, el número de baños y el tipo de vivienda, además de variables contextuales como el estrato y la percepción de seguridad. Estas últimas se obtienen mediante un join espacial con datos del DANE y la alcaldía, así como a través de la distancia a estaciones de policía y zonas con mayor incidencia de hurtos. Este enfoque integral permite una mejor descripción y predicción del precio de las viviendas.

Analisis descriptivo de los datos

Precio de la vivienda

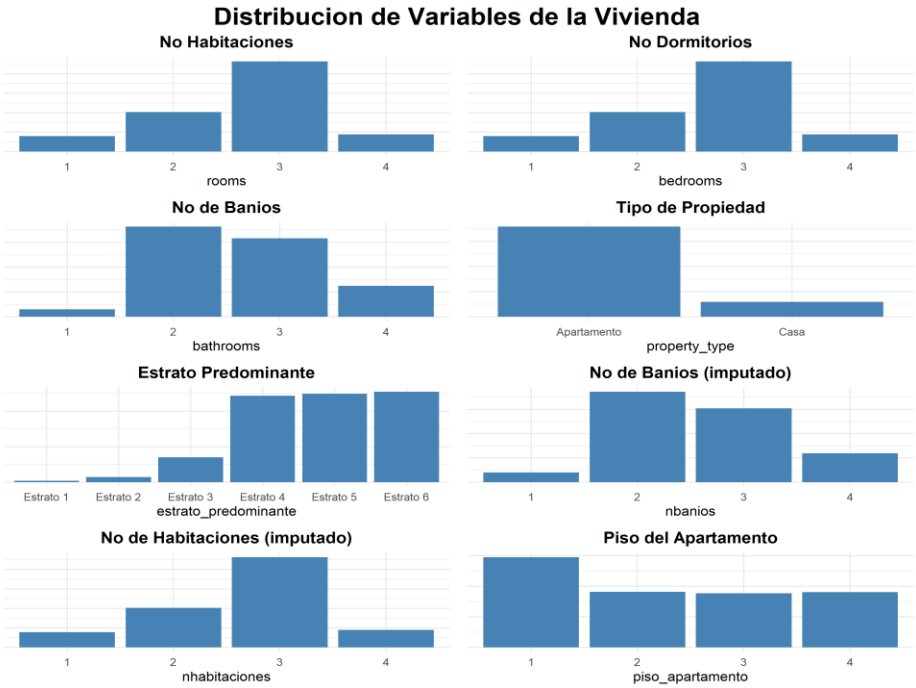
El objetivo de este artículo es predecir el precio de las viviendas en la ciudad de Bogotá. Esta variable está medida a lo largo de tres años distintos, lo que en términos reales no representa lo mismo debido a la inflación. Inicialmente, se intentó ajustar los precios a términos reales deflactándolos por la inflación, pero los resultados no fueron coherentes. Por ello, se optó por utilizar efectos fijos, lo que mejoró la predicción considerablemente.

Teniendo esto en cuenta, se observa que el precio de entrada para comprar una vivienda en Bogotá, tomando como referencia esta base de datos, es de 300 millones de pesos, con un máximo de 1650 millones y un precio promedio de aproximadamente 650 millones. En términos de precios de vivienda, Bogotá se encuentra por debajo del promedio nacional. Según el Índice de Precios de Vivienda del DANE, el índice de precios en Bogotá es de 126 puntos, mientras que el promedio nacional es de 130 puntos. Además, el aumento de precios durante estos tres años ha sido moderado, con un incremento promedio del 8%.

Resumen del Precio de las Viviendas	
Estadistico	Precio
Min.	$300,000,000.00
1st Qu.	$415,000,000.00
Median	$559,990,000.00
Mean	$654,534,675.29
3rd Qu.	$810,000,000.00
Max.	$1,650,000,000.00

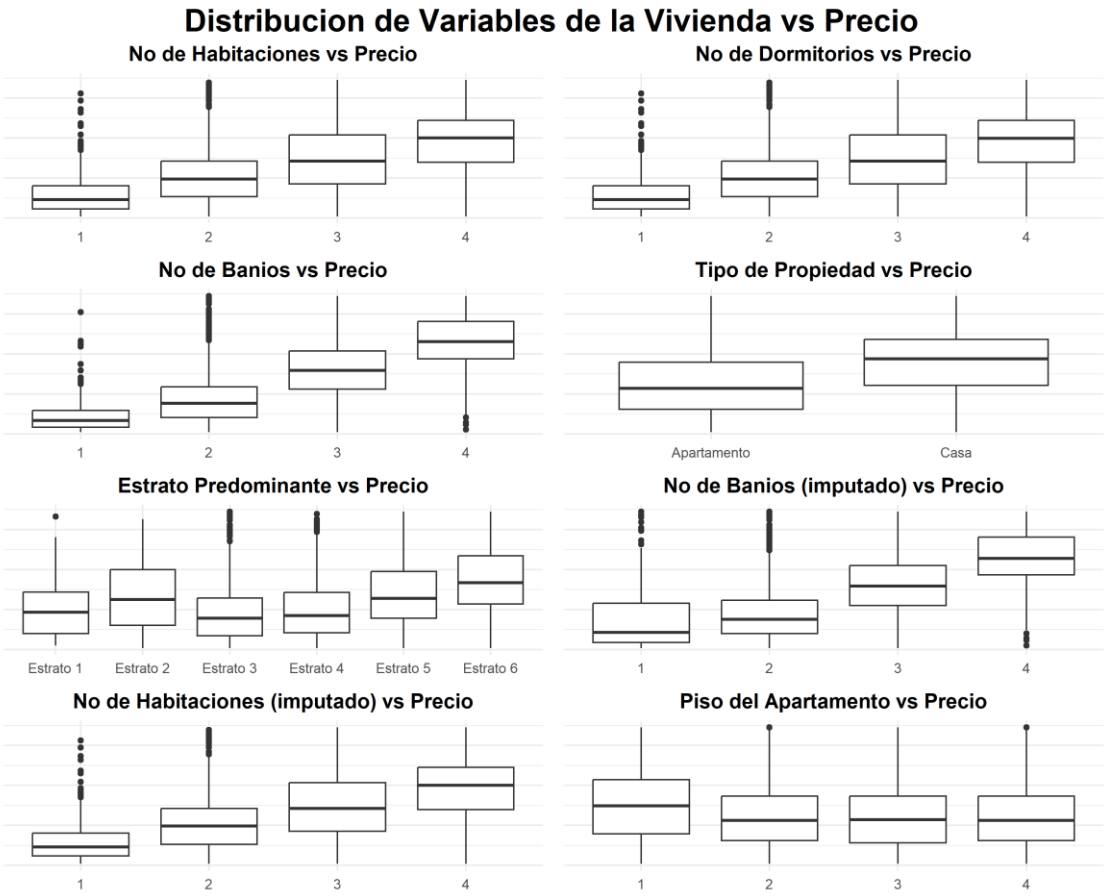
Caracteristicas del hogar

En primer lugar, vamos a analizar cómo se relacionan las principales variables del hogar con el precio de venta. Para este análisis, tomaremos en cuenta el número de habitaciones, el número de baños, el número de alcobas, el tipo de propiedad, el estrato y el piso donde se encuentra ubicada la vivienda. Inicialmente, imputaremos de forma sencilla las variables que tienen valores NA y luego procederemos con un análisis descriptivo bivariado.



Es interesante observar que la cantidad de habitaciones más común tanto para las habitaciones en general como para las habitaciones para dormir es 3. Además, la cantidad de baños tiende a ser similar, lo que podría explicarse por el hecho de que la mayoría de las viviendas en venta son apartamentos. Esta distribución de baños se concentra principalmente en 2 y 3, lo que sugiere una preferencia por una distribución equilibrada en los espacios de descanso y aseo.

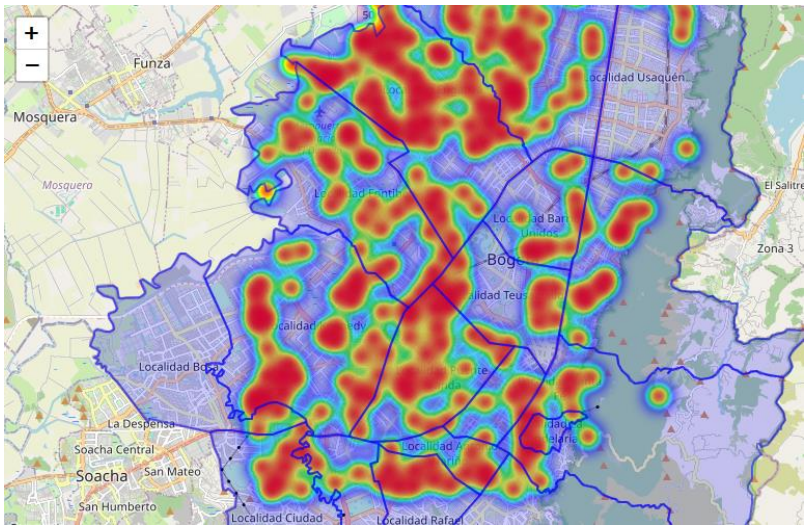
Es importante resaltar que la muestra de datos que tenemos disponible muestra una predominancia de apartamentos ubicados en zonas de Estrato 4, 5 y 6. Esto puede indicar una tendencia hacia viviendas de mayor estatus socioeconómico en la muestra analizada.



Es interesante observar que al analizar la relación entre las características de las viviendas y el logaritmo del precio de las viviendas, se encuentran relaciones estadísticas significativas con el número de baños, el número de habitaciones, el tipo de vivienda y el estrato socioeconómico.

Estos hallazgos son consistentes con lo que se ha observado en la literatura previa, donde se muestra que a medida que una vivienda tiene más habitaciones, su precio tiende a aumentar, lo mismo ocurre con la cantidad de baños. Además, para esta muestra en particular, se observa que el tipo de vivienda más costoso son las casas, aunque se debe tener en cuenta que hay pocos registros para este tipo de vivienda, lo que podría sesgar los resultados.

Por último, se nota que a medida que aumenta el estrato socioeconómico, el precio promedio de la vivienda tiende a aumentar, especialmente en los estratos 4 al 6, donde se tienen más datos disponibles. Esto sugiere una relación positiva entre el estrato y el precio de la vivienda en esta muestra analizada.



El mapa muestra la concentración de las viviendas que están siendo analizadas en este estudio. En primer lugar, la mayoría de las viviendas en venta se encuentran en el oeste, norte y sur de la ciudad. Como se puede observar en el mapa de calor, donde se coloca el logaritmo de los precios de las viviendas, entre más cerca estén las viviendas de la periferia de la ciudad, mayor es la intensidad del precio. Esto evidencia que la ubicación es un factor importante para la predicción de precios.

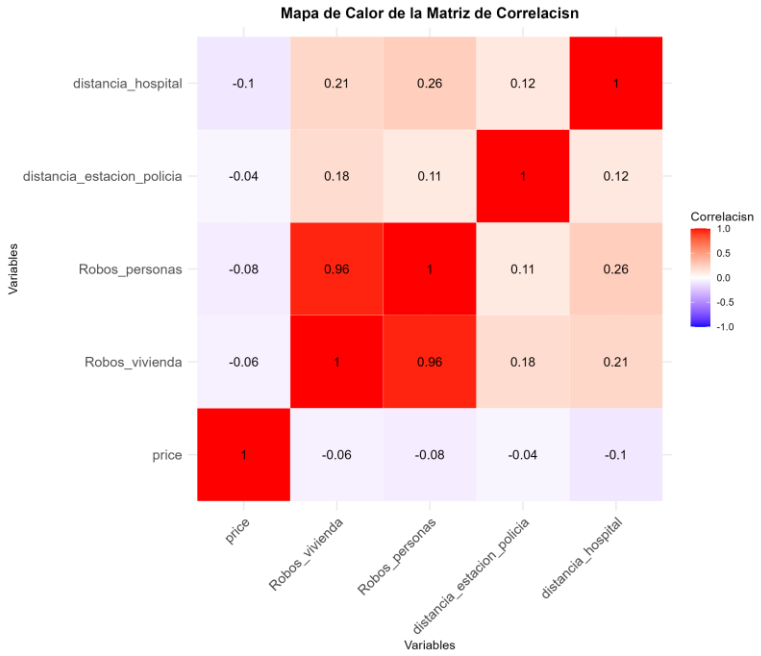
Seguridad y salud

Explorar la relación entre la seguridad del lugar donde está ubicada la vivienda y su costo es un enfoque importante en el análisis. Para esto, se trabajará desde dos perspectivas:

1. **Cantidad de Robos Georeferenciados**: Se utilizará la información proporcionada por la alcaldía de Bogotá sobre la cantidad de robos a personas y a casas georeferenciadas a nivel de localidad. Luego, mediante la distancia, se agregará la cantidad de robos que ocurren más cerca de la vivienda. Se espera que entre más cerca estén los robos, menor sea la percepción de seguridad y, por lo tanto, menor sea el precio de la vivienda.

2. **Distancia a la Estación de Policía más Cercana**: Se calculará la distancia mínima entre cada hogar y la estación de policía más cercana. Se espera que entre más cerca esté una vivienda de una estación de policía, mayor sea la percepción de seguridad y, por ende, mayor sea el precio de la vivienda.

3. **Distancia al Hospital más Cercano**: Además de la seguridad, la accesibilidad a servicios de salud también puede influir en el precio de una vivienda. Se calculará la distancia mínima entre cada hogar y el hospital más cercano. Se espera que una mayor proximidad a un hospital aumente la percepción de comodidad y seguridad para los residentes, lo que podría reflejarse en un aumento en el precio de la vivienda.



Con esta gráfica, podemos observar que existe una relación negativa entre el valor de una vivienda y la cantidad de robos cercanos que se han producido. Esto indica que la percepción de seguridad es un factor a tener en cuenta al momento de entender el valor de la vivienda. Del mismo modo, cuanto mayor sea la distancia entre la vivienda y la estación de policía, el precio de la vivienda tiende a bajar. Por lo tanto, se puede comprobar la hipótesis de que la percepción de seguridad es importante para entender el precio de la vivienda. Además, existe una relación negativa entre el precio de la vivienda y la distancia a los centros de salud. Como se explicaba en la revisión de literatura, este tipo de variables tienen un impacto significativo sobre el precio de la vivienda.

Modelos y resultados

Para entrenar los modelos se utilizó la técnica de validación cruzada, de manera que se pudiera reducir el efecto de la autocorrelación espacial sobre el aprendizaje de cada modelo. Teniendo en cuenta que el Marco Geoestadístico Nacional del DANE provee una agrupación geográfica a nivel de manzana, sección (conjuntos de manzanas) y sectores (conjuntos de secciones), se pudo utilizar cada sección como un fold que rompiera con la dependencia espacial del precio.

```
sectores<-unique(train$cod_sector)
folds<-createFolds(sectores,k=length(sectores),list = TRUE,returnTrain = TRUE)
```

Los modelos se entrenaron en primera instancia con la información extraida de la base de datos y del censo del DANE *nbanios*, *nhabitaciones*, *piso_apartamento* ; *estrato* . Dado que los precios de las viviendas también corresponden a diferentes meses de un horizonte de tiempo de tres años, se tuvo que incluir la variable *periodo* para tratar de establecer un efecto fijo del tiempo sobre el precio. Posteriormente se fueron incluyendo las variables de seguridad y salud, encontrando que hubo una mejora significativa en la predicción al incluir este tipo de variables.

Los algoritmos utilizados para las predicciones fueron *Regresión lineal*[*], *Elastic net* [**], *Random forest* [***]y *Xgboost* [****]. Cada uno de estos se entrenó realizando la validación cruzada con los folds anteriormente señalados que pueden ser observados en Mapa de folds. Adicionalmente se entrenaron *Redes neuronales*[*****], para estas redes se partió de una primera con una sola capa oculta y 25 nodos, la cual se fue complejizando hasta llegar una de tres capas ocultas, la primera de 25 nodos reduciendose diez nodos para la siguiente. En todos los casos el MAE obtenido no se redujo

significativamente.

Table 1: Resumen de resultados

Algoritmo	Variable objetivo	Predictores utilizados	MAE fuera de muestra
Regresión lineal	price	nbanios, nhabitaciones, piso_apartamento, estrato	860024814
Regresión lineal	log(price)	nbanios, nhabitaciones, piso_apartamento, estrato	295197752
Regresión lineal	log(price)	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo	294830576
Elastic net	price	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo	283134955
Elastic net	price	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia	268696039
Elastic net	log(price)	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia	276791195
Elastic net (caret)	price	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia	264342673
Random forest	price	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo	274216139
Random forest	price	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia	244889109
Random forest	log(price)	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia	252533214
Random forest (caret)	price	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia	249855878
Random forest	price	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia, distancia_hospital	235325152
Xgboost	price	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia	277450159
Red neuronal una capa oculta	log(price)	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia, distancia_hospital	274199430
Red neuronal tres capas ocultas	log(price)	nbanios, nhabitaciones, piso_apartamento, estrato, Periodo, Robos_vivienda, Robos_personas, distancia_estacion_policia, distancia_hospital	269680313

El modelo que mejor desempeño tuvo en las predicciones fu el Random Forest, para este se definió un bosque de 500 arbóles y 4 variables por partición. Posteriormente se ajustaron estos híperparámetros

para identificar que el MAE se reducia considerablemente, sin embargo, se concluyó que los cambios en el MAE dentro de muestra no eran lo suficientemente importantes en comparación con el costo computacional asumido. Por esta razón se decide mantener los hiperparametros iniciales en la predicción enviada a Kaggle.

Conclusiones y recomendaciones

Se logra evidenciar que las caracteristicas de una vivienda son un buen predictor de su precio, sin embargo, las variables del entorno también se convierten en un factor importante al momento de establecer su precio. Evidenciamos como hay aspectos del entorno que pueden afectar positivamente su precio, como la distancia a amenities como hospítales, mientras que existen otras variables que pueden impactar negativamente el precio, como el aumento de la criminalidad en la zona.

También observamos que en este tipo de problemas algoritmos que tienen un poca carga computacional como Elastic Net y la propia regresión lineal tienen un rendimiento aceptable frente a algunos modelos más complejos como Xgboost. Por lo que el tiempo dedicado a entrenar modelos podría bien utilizarse en mejorar la modelación en una regresión y obtener resultados mucho mejores y menos costosos.

La cantidad de variables del entorno incluidas impacta significativamente el rendimiento de los modelos, por lo que se recomienda en estos problemas dedicar un mayor tiempo a la construcción de un set de datos lo más completo posible. En este sentido, por restricciones de tiempo no fue posible incluir otras variables que podrían mejorar la predicción como la distancia a centros comerciales y zonas verdes.

Referencias

Alfaro-Navarro, Jose Luis, Emilio L. Cano, Esteban Alfaro-Cortes, Noelia Garcia, Matias Gamez, and Beatriz Larraz. 2020. “A Fully Automated Adjustment of Ensemble Methods in Machine Learning for Modeling Complex Real Estate Systems.” *Complexity* 2020. https://doi.org/10.1155/2020/5287263.

Alzate, Yuri. 2019. “Modelo Predicción Precios Viviendas.”

Choy, Lennon H. T., and Winky K. O. Ho. 2023. “The Use of Machine Learning in Real Estate Research.” *Land* 12 (March): 740. https://doi.org/10.3390/land12040740.

Cueto, Ana Bruno. 2022. “ANÁLISIS PREDICTIVO DEL PRECIO DE LA VIVIENDA EN LOS DISTRITOS DE CIUDAD LINEAL y LA LATINA CON MODELOS DE MACHINE LEARNING.”

Giraldo, Angie. n.d. “Prediccioin de Los Precios de Vivienda En La Ciudad de Medellin y El Area Metropolitana.”

Jha, Shashi Bhushan, Radu F Babiceanu, Vijay Pandey, and Rajesh Kumar Jha. n.d. “Housing Market Prediction Problem Using Different Machine Learning Algorithms: A Case Study.”

Mohamed, Hossam H., Ahmed H. Ibrahim, and Omar A. Hagras. 2023. “Forecasting the Real Estate Housing Prices Using a Novel Deep Learning Machine Model.” *Civil Engineering Journal (Iran)* 9: 46–64. https://doi.org/10.28991/CEJ-SP2023-09-04.

Park, Byeonghwa, and Jae Kwon Bae. 2015. “Using Machine Learning Algorithms for Housing Price Prediction: The Case of Fairfax County, Virginia Housing Data.” *Expert Systems with Applications* 42 (April): 2928–34. https://doi.org/10.1016/j.eswa.2014.11.040.