Artículo

Modelo de Aprendizaje Automático para la Predicción de Precios de Vivienda en la Ciudad de Bogotá

Mónica Andrea Nieto[1,*]

[1]Facultad de Ingeniería y Ciencias Básicas, Fundación Universitaria Los Libertadores
[*]Autor de correspondencia: manietot01@libertadores.edu.co

LOS LIBERTADORES
FUNDACIÓN UNIVERSITARIA
Acreditación Institucional en Alta Calidad

Facultad de Ingeniería y Ciencias Básicas

APUNTES DE CIENCIA E INGENIERÍA

Resumen

Como cualquier otro mercado, el sector inmobiliario atiende a las dinámicas de la oferta y la demanda, las cuales varían según factores propios de la vivienda. Con base a lo anterior, se ha identificado una necesidad de información en cuanto a precios, relativa a las dinámicas del mercado inmobiliario. Para esto, se pretende desarrollar herramientas que le permitan tanto a agentes inmobiliarios, compradores, vendedores, constructores y demás participantes del sector, optimizar sus procesos de decisión de cara a la alternativa que más se ajusta a los intereses de cada uno de ellos. En este documento se desarrolla un modelo de Machine Learning que permite predecir los precios de vivienda en la ciudad de Bogotá, lo cual facilita la toma de decisiones en cuanto a la compra de vivienda usada y, a su vez, pretende ser una herramienta que las empresas o personas involucradas en el sector inmobiliario, implementen para auspiciar el valor comercial de un bien en determinada zona. De esta manera, el modelo permitirá que el valor de un bien tenga argumentos válidos y no simples especulaciones a la hora de tomar decisiones de cara a compra y venta de inmuebles. Mediante la técnica del raspado de web se obtienen los datos (directamente de la página web fuente), con los cuales se realiza un análisis de cada una de las variables. A partir de lo anterior, se construye el modelo Machine Learning que más se ajusta al estudio, este caso fue un modelo Light Gradiente Boosting, el cual fue sometido a entrenamiento y testeo, dando como resultado un error (MAPE) del 15.58 %.

Palabras clave: Machine Learning, raspado de web

Recibido: 7 de febrero de 2022
Aceptado: 10 de abril de 2022
Publicado: 30 de junio de 2022

Como citar este articulo

M. Nieto, "Modelo de Aprendizaje Automático para la Predicción de Precios de Vivienda en la Ciudad de Bogotá", *Revista Apuntes de Ciencia e Ingeniería*, *vol. 1*, *no. 1*, pp. 63-71, Jun. 2022. doi: 10.37511/apuntesci.v1n1a6

1. Introducción

Actualmente la determinación de un precio de vivienda en la ciudad de Bogotá se ve afectado por distintas variables tales como el área, la ubicación, el estado de las vías, el diseño de la propiedad, las facilidades de financiación y otras más relacionadas a términos macroeconómicos [1].

Es por ello que el proceso que establece y fija el valor de un inmueble suele generar confusión en cuanto a la toma de decisiones de si comprar y vender una propiedad. De ahí que no en pocas ocasiones se suelen cometer errores que perjudican y benefician a los participantes de dicho intercambio, al ser este un mercado que no está exento de las especulaciones. Con base en esto, es imprescindible contar con un mínimo conocimiento del mercado y sus dinámicas que clarifique el proceso y conduzca a la mejor decisión posible [2].

En referencia a lo anterior, es imprescindible que todos los actores que intervienen en los procesos del sector inmobiliario conozcan no solamente las dinámicas microeconómicas, sino también comprendan el trasfondo por el cual se rigen y son fijadas dichas condiciones. Por lo tanto, es preciso aclarar que son los gobiernos y demás autoridades quienes permiten y regulan cada uno de los aspectos tanto económicos, como sociales, medioambientales y de diversa índole. Con base en lo anterior, el gobierno colombiano establece mediante el DECRETO 046 del 2020 [3], que "el estado será el encargado de incentivar la compra de vivienda nueva a través de planes de gobierno", los cuales serán determinados a corto, mediano y largo plazo [4].

El DANE ha venido analizando las variaciones de precios en el mercado inmobiliario mediante la implementación del índice de precios de vivienda nueva (IPVN), el cual ha publicado recientemente un informe en el que se evidencia un incremento del 3,3 % en comparación con el trimestre anterior. A su vez, el Banco de la República de Colombia implemento en el año 1986, un indicador de precios de la vivienda usada (IPVU) debido a que no se contaba con indicadores para este tipo de inmueble, lo cual impedía crear estadísticas fundamentadas en datos históricos en cuanto al análisis y evaluación del comportamiento de los precios de la vivienda [2],[5].

En efecto las variaciones de precio en Colombia, principalmente en las grandes ciudades como lo son Bogotá, Medellín, Cali, etc, han ido aumentando considerablemente. Lo anterior se debe principalmente a que cada día llegan más y más personas a dichas ciudades, de lo cual se puede evidenciar un considerable aumento en los precios en relación con la oferta disponible, la cual generalmente es menor ya que no crece en proporción al aumento poblacional [6].

Un estudio que se realizó en el departamento de Antioquia para predecir los precios de vivienda en el municipio de Rionegro utilizó como base metodológica la técnica del scraping para realizar una comparación de inmuebles en la zona. Mediante la utilización de variables numéricas tales como el área privada, el área construida, el tipo de vivienda y el estrato socioeconómico, dicho estudio encontró que algunos de modelos que más se ajustaban al resultado esperado, eran de regresión lineal, y árboles de decisiones [5]. En otro estudio realizado en la Universidad Nacional de Colombia sobre modelos de aprendizaje estadísticos, se analizaron los precios de vivienda nueva en Bogotá según distintas variables tanto numéricas como categóricas, concluyendo que los mejores modelos para este caso fueron el árbol de regresión y el de máquina vectorial de soporte, ya que estos lograron explicar las variaciones del precio de manera precisa [7].

Es posible resumir de manera bastante precisa las características más importantes dentro de la decisión de compra de inmuebles en la ciudad de Bogotá y se ha establecido que dentro de las más relevantes y según la tendencia actual, se destaca la ubicación por sobre otras variables como el área, el número de habitaciones de baños, garaje, etc., que, aunque son relevantes, no determinan directamente el precio de la vivienda como sí lo hace su localización [8].

2. Metodología

2.1. Datos

Como primera instancia, la información fue obtenida directamente de la página Metro Cuadrado, cuya función permite a las personas interesadas en la compra y venta de inmuebles acceder a las distintas propiedades disponibles en el mercado. Adicionalmente, esta página brinda a los usuarios vendedores la opción de publicar su inmueble con el objetivo de acceder a posibles compradores. Sin embargo, el vendedor deberá pagar por la suscripción según sea su necesidad, de acuerdo con el tiempo que el anuncio estará visible para el comprador.

Por otro lado y debido a la necesidad de optimizar el tiempo y los recursos, la recolección de datos no fue realizada manualmente sino que por el contrario, se ha empleado una técnica automática llamada raspado de Web , la cual permite optimizar dicho proceso. Los pasos que se llevaron a cabo fueron: i) limpieza de los datos, ii) realización de análisis descriptivo, iii) modelación.

Para el análisis, desarrollo e implementación del modelo machine learning, se utilizaron las librerías de Python Matplotlib, Pandas, Numpy, Seaborn, SweetViz y PyCaret.

Variable original	Nueva variable (Descripción)	Tipo
Objetivo: mvalorventa	Log_mvalorventa (Log10)	Numérica
Predictoras: mtipoinmueble	No Aplica	Categórica
marea	Log_marea (Log10)	Numérica
mnrocuartos	No Aplica	Numérica
mnrobanos	No Aplica	Numérica
mnrogaraje	No Aplica	Numérica
mzona	No Aplica	Categórica
mbarrio	No Aplica	Categórica
mnombrecomunbarrio	No Aplica	Categórica

Tabla 1: Variables asociadas a las viviendas.

2.2. Análisis Exploratorio de los Datos (EDA)

Al realizar la descripción del dataset inicial, sin ningún tipo de manipulación, se obtiene la Tabla 2 que resume de manera breve los datos a nivel cuantitativo:

index	mvalorventa	marea	mnrocuartos	mnrobanos	mnrogarajes
count	9546	9546	9543	9546	9546
mean	1787898225,87	213,59	3,10	3,28	2,09
std	32305265199,02	1160,87	0,90	1,15	1,11
min	1100000	0	1	0	0
25 %	520000000	88,30	3	2	1
50 %	890000000	149	3	3	2
75 %	1580000000	244	4	4	3
max	2700000000000	101800	5	5	4

Tabla 2: Descripción del dataset.

Al analizar esta información se prevé la necesidad de realizar un EDA para adquirir una mayor comprensión de los datos, pues aunque la información establece los parámetros generales de cada una de

las variables, no es posible establecer relación entre cada una de ellas, a su vez que es posible realizar una mejor visualización de dichas relaciones. Con base en lo anteriormente dicho, se procede a realizar el EDA, con la finalidad de determinar el tamaño de la base de datos, la cantidad de variables según su tipo (categóricas y numéricas) y sus correlaciones respectivas.

Con el propósito de establecer las asociaciones, se realiza un mapa de calor (*heatmap* en inglés) de las correlaciones entre las variables categóricas, representadas por cuadrados y numéricas simétricas, representadas por círculos. Las asociaciones categóricas miden el coeficiente de incertidumbre y la tasa de correlación en una escala de 0 a 1, mientras que las numéricas simétricas se determinan en un rango entre -1 y 1.

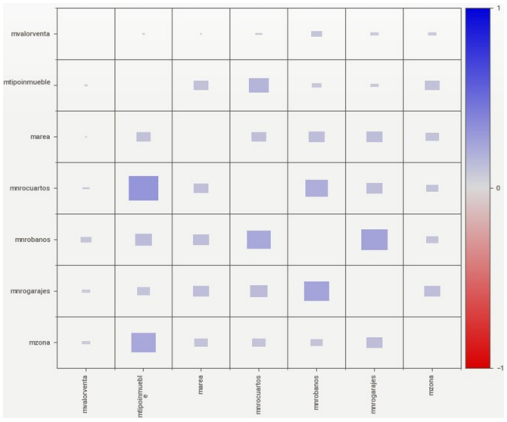

Figura 1: Definición y transformación de variables.

Al observar la Figura 1, se determina que el mayor valor de correlación se presenta entre las variables mnrocuartos y mtipoinmueble, donde el 55 % de los datos representan propiedades que tienen tres cuartos o habitaciones. También se observa una asociación moderada entre las variables numéricas mnrobanos y nmrogarajes, para las que el número de propiedades con tres baños representan el 29 % de la muestra, mientras que el menor número de baños es 1 con un 4 %. Por último, es de utilidad resaltar la relación entre las variables mzona y mtipoinmueble, donde el 77 % (7.306) de los datos son apartamentos, mientras que el 23 % (2.240) son del tipo casa.

2.3. Preparación inicial de los datos y las variables

Como se pudo observar en la sección anterior (ver Tabla 2), las variables numéricas presentan un sesgo positivo (a la derecha), por lo cual se sugiere realizar una transformación de las variables mvalorventa y marea. Dicha transformación se realiza con el logaritmo en base diez, el cual permite que los datos con asimetría positiva sean más normales, facilita la explicación de una curva en un modelo lineal y estabiliza la variación dentro de los grupos. La tabla 3, evidencia la estabilización de los datos en las nuevas variables *log_marea* y *log_valorventa*. Adicionalmente, como se observa en la Figura 2 la distribución de los datos en dichas variables queda balanceada, por lo que podemos dar inicio a la imputación de las demás variables.

index	mvalorventa	marea	cuartos	baños	garajes	log_marea	log_mvalorventa
count	9200	9200	9197	9200	9200	9200	9200
mean	1515696558	221,62	3,10	3,28	2,10	2,19	8,96
std	17068050637	1181,75	0,90	1,15	1,11	0,30	0,36
min	1100000	15	1	0	0	1,18	6,04
0,25	520000000	93	3	2	1	1,97	8,72
0,50	890000000	154	3	3	2	2,19	8,95
0,75	1582875000	249	4	4	3	2,40	9,20
max	1300000000000	101800	5	5	4	5,01	12,11

Tabla 3: Descripción del dataset

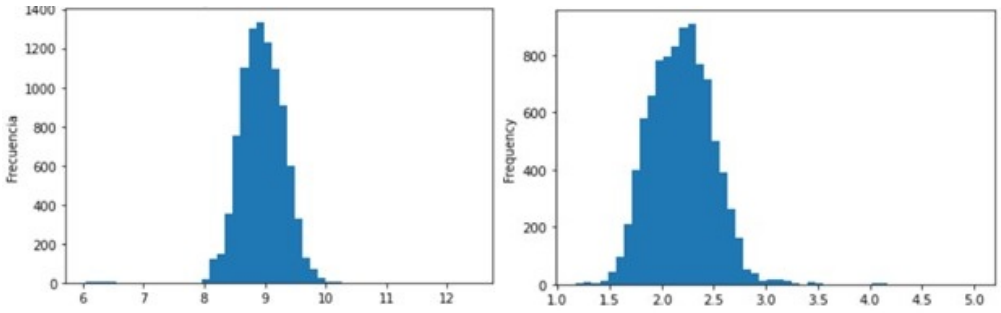

Figura 2: Histograma de *log_mvalorventa* y *log_marea*.

Como hemos mencionado, se realizan algunas imputaciones debido a que se encontraron valores nulos para las variables mnrcuartos (3) y mrobanos (27), para lo cual se usó la media (mean) con el fin de asignar este resultado a dichos valores nulos. En seguida, se realizó una imputación en la variable mzona debido a la gran cantidad de datos nulos (487), para la cual se usó el valor de la moda (Norte). Apartir de esto, se realiza el primer modelo.

2.4. Preprocesamiento y modelación

Como primera instancia, se procede a la creación del modelo base, el cual es una regresión lineal con descenso de gradiente estocástico (SGD). Dicha regresión buscará de manera aleatoria el valor mínimo de los parámetros asociados a las variables involucradas en la función de coste. Lo anterior se realiza mediante la selección de las variables de entrada y salida, *log_marea* y *log_mvalorventa*, respectivamente. Con base a dichos parámetros, se pronostica el valor de venta de una casa o apartamento. Adicionalmente, es preciso establecer una medida de error en dicho pronóstico, por lo cual para este ejercicio se implementa el MAPE. Con esta información, es posible interpretar dichos parámetros (intercepto y pendiente) para finalmente poder realizar la evaluación sobre los datos de testeo.

Por otro lado, es de utilidad contrastar el desempeño de dicho modelo a partir del desarrollo de un modelo automático. Para este fin, se implementa la librería PyCaret con la base inicial (sin imputaciones), solamente se deja *log_marea* y *log_valorventa*, además de las variables iniciales. Para la ejecución del análisis respectivo fue preciso definir las variables objetivo (*log_valorventa*) y las predictoras (otras). Esta librería permite llevar a cabo desde la preparación de los datos, hasta el despliegue del modelo final en poco tiempo.

Según PyCaret, para el presente proyecto el mejor modelo es el Light Gradient Boosting Machine. Dicho modelo se conforma por un conjunto de árboles de decisión individuales, los cuales han sido entrenados de manera secuencial, tal que cada árbol trata de reducir el error del anterior. Por lo tanto, la predicción es obtenida agregando las predicciones individuales de cada uno de los árboles que conforman dicho modelo.

3. Resultados

De primera mano, se analizan los resultados suministrados por parte del modelo base el cual, como ya se ha mencionado, implementa un Gradiente Estocástico Descendente que determina los parámetros de la función de coste para dos de las variables consideradas, *log_valorventa* y *log_marea*. Las métricas obtenidas para los parámetros intercepto y pendiente son 6.77 y 1.01, respectivamente. Con base a lo anterior, el pronóstico determina que el valor de una casa o apartamento de 32m es de 200 millones, mientras que para una propiedad de 1000m nos arrojará un valor de 6300 millones, aproximadamente.

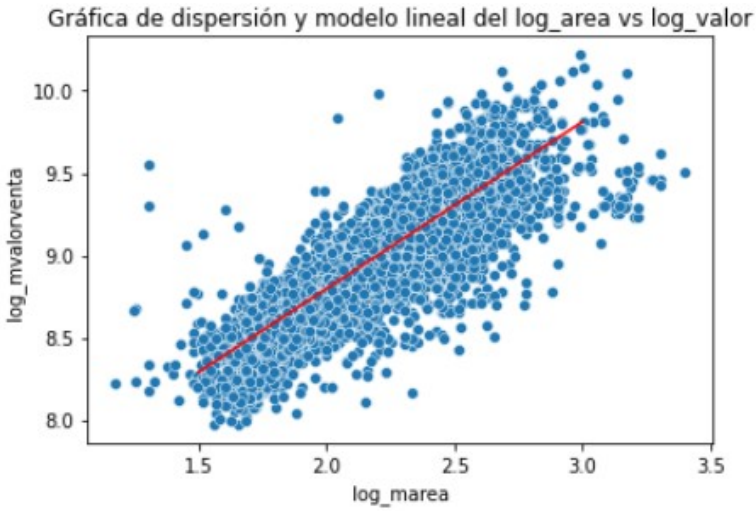

Figura 3: Modelo base.

Ahora, debemos determinar el error porcentual absoluto medio (MAPE), el cual proporciona la precisión de nuestro del pronóstico anterior. Este se calcula mediante la siguiente relación matemática:

$$MAPE = \frac{1}{n}\sum\left(\frac{|log\_mvalorventa - pronostico|}{|log\_mvalorventa|}\right)$$

Teniendo en cuenta la relación anterior, el valor del MAPE para este modelo se establece en un valor de 15.58 %. La figura 4 muestra el comportamiento de los residuales con relación al valor real y al valor pronosticado.

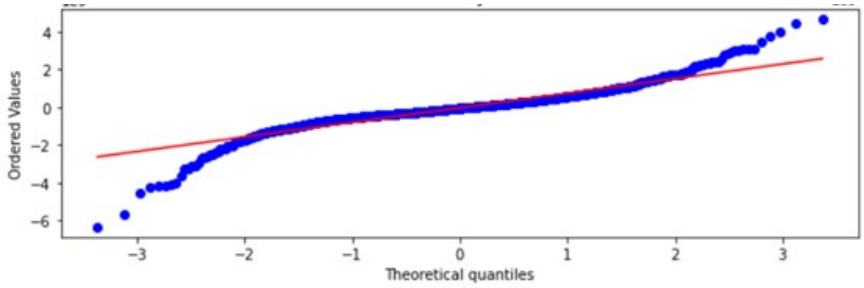

Figura 4: Gráfica de residuales.

Por otro lado, se realiza un modelo automático supervisado, cuya aplicación compara distintos tipos modelos (ver Tabla 4), de los cuales se han seleccionado los cinco primeros para los que se presentan los resultados del MAE, R2 y MAPE. Sin embargo, se determina que el mejor modelo es el Light Gradient Boosting Machine. La Tabla 4 muestra las métricas de regresión que mejor resumen los resultados para los cinco modelos con mejor desempeño.

Model	MAE	R2	MAPE
Light Gradient Boosting Machine	0.0806	0.8996	0.009
Extra Trees Regressor	0.0798	0.8952	0.0089
Random Forest Regressor	0.0809	0.8923	0.009
Bayesian Ridge	0.0849	0.886	0.0094
Ridge Regression	0.0855	0.8848	0.0095
Modelo base (SGD)	-	-	0.1558

Tabla 4: Comparación de modelos con PyCaret

Con base al análisis anterior, el modelo con mejor puntaje promedio para las métricas MAE, R2 y MAPE, es el LGBM. Por lo tanto, se continúa trabajando con dicho modelo para la realización de la predicción inicial del 70 % (6.414) de los datos para entrenamiento y del 30 % (2.749) para testeo. Es preciso recordar, que se eliminaron algunos valores nulos.

Una vez evaluado el mejor modelo, Pycaret arroja distintos tipos de gráficos que permiten realizar un mejor análisis de los resultados. En este caso, se selecciona un gráfico de residuales (Ver Figura 5), el cual nos permite observar la distribución de los residuales a lo largo del eje x, cuya disposición uniforme en el mismo, permite confirmar que un modelo lineal es apropiado para este ejercicio.

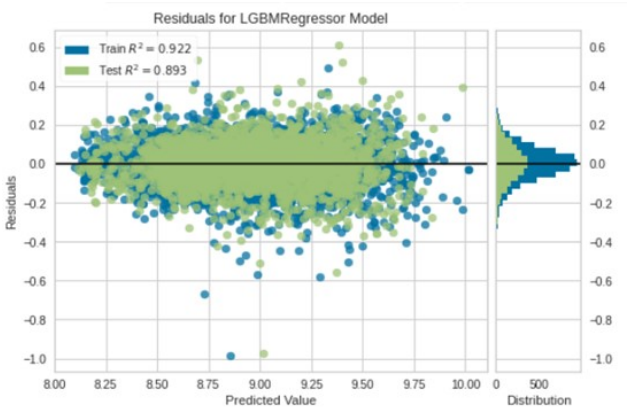

Figura 5: Análisis de residuales.

Finalmente, es preciso realizar una interpretación del modelo seleccionado de manera local, lo cual quiere decir que por cada dato se obtiene la contribución que realiza cada variable en la predicción entregada por el modelo. En este caso, al observar la Figura 6 se establece que la variable que más contribuye al modelo desarrollado en este ejercicio es *log_marea*.

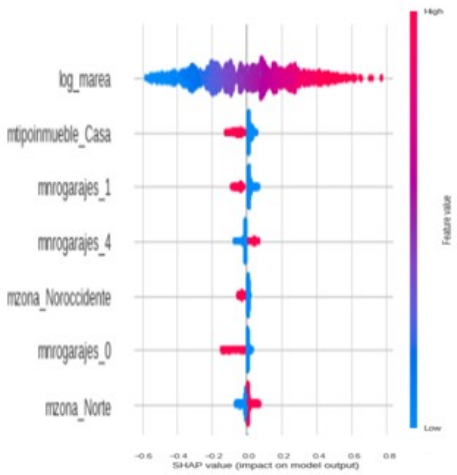

Figura 6: SHAP Value.

4. Análisis de resultados

En el presente estudio se busca establecer un modelo automático supervisado que permita pronosticar el precio o valor de un inmueble en relación al área, teniendo en cuenta otras variables como número de cuartos, baños, garaje, tipo de inmueble, ubicación, etc. Los resultados presentados en este documento evidencian que los modelos que más se ajustan al caso de estudio son Light Gradient Boosting Machine, Extra Tree Regressor y Random Forest. Al realizar una comparación entre dichos modelos, se determinó

que el MAPE en cada uno de los casos fue de 0.9 %, 0.91 % y 0.92 % respectivamente. Por lo tanto, podemos concluir que al comparar con el modelo base, cuyo MAPE fue de 15.58 %, se establece que el modelo a partir de árboles de decisión es bastante mejor que el modelo de regresión lineal inicialmente planteado. Al aplicar la librería Pycaret, el modelo de regresión lineal fue uno de los modelos peor evaluados, con un MAPE de 89 %.

Una vez se ha entrenado el modelo, es preciso realizar una última comprobación, pero en este caso con los datos de testeo, los cuales no han sido vistos por el modelo y así poder evidenciar que tan acertado es el pronóstico de este. Es por eso, que en la Tabla 5 se presentan algunos de los pronósticos para constatar el error de estos.

Área (m2)	Valor real (COP)	Valor pronóstico (COP)	MAPE
416,87	2.041.737.944,67	2.041.737.944,67	0
75,86	371.535.229,10	398.107.170,55	7,15
79,43	309.029.543,25	338.844.156,14	9,65
173,78	1.174.897.554,94	1.288.249.551,69	9,65
100	758.577.575,03	660.693.448,01	12,90
70,79	416.869.383,47	478.630.092,32	14,82
281,84	2.137.962.089,50	1.659.586.907,44	22,38
389,05	1.148.153.621,50	1.445.439.770,75	25,89
371,54	5.011.872.336,27	3.467.368.504,53	30,82
436,52	1.548.816.618,91	2.041.737.944,67	31,83

Tabla 5: Pronóstico del modelo ML automático supervisado.

Como se evidencia en la Tabla 5, las viviendas clasificadas con mayor área tienden a tener un pronóstico con un error más grande que las de áreas más reducidas. Lo anterior se puede deber a que, como hemos visto en la Figura 5, algunos registros están mucho más dispersos para áreas grandes. Sin embargo, la mayor cantidad de registros se encuentran concentrados a lo largo de la línea que representa el valor pronosticado.

Con base en lo anterior, podemos establecer que aunque el modelo ha emitido un pronóstico con un bajo porcentaje de error, es posible incurrir en sesgos debido a la dispersión de ciertos valores atípicos determinados por el área, la zona y el barrio. Esto último requiere que los individuos que intervengan en este mercado deban incluir dentro de sus variables a analizar, atributos relacionados al estrato socioeconómico, cercanía a vías principales, centros comerciales, hospitales, etc. Esto último puede implicar un mayor tiempo de compra-venta del inmueble, así como pérdidas económicas sustanciales debido a una clasificación pobre y errónea.

5. Conclusiones

Se determina que las variables que más explican el precio de una vivienda en la ciudad de Bogotá son el área y la ubicación. A su vez, variables como el número de habitaciones, baños, garajes, etc, no alteran de manera directa el precio de un inmueble. Por otro lado, es importante tener en cuenta la cercanía a vías principales, centros comerciales, bibliotecas, hospitales, etc (ya que esto incrementa la plusvalía de la propiedad de manera significativa), aunque esto no sea evaluado en el presente estudio.

Gracias a la optimización de modelos automáticos supervisados se puede establecer que los más apropiados para este tipo de estudios, son todos aquellos que se relacionen a los modelos de árboles de decisión, más específicamente árboles de regresión, ya que como hemos evidenciado, el análisis de compra-venta de inmuebles se enfoca en el valor de este, es decir a una variable objetivo.

Finalmente, se puede concluir que este tipo de modelos pueden ser aplicados a otro tipo de inmuebles tales como los de vacancia en centros comerciales, debido a que se comporta de manera análoga al sector

de la vivienda, pues se tienen en cuenta variables de área, valor de venta y ubicación.

Referencias

[1] Banco de la República. (2019) "Índice de precios de la vivienda usada (ipvu)". [Online]. Available: https://www.banrep.gov.co/es/estadisticas/indice-precios-vivienda-usada-ipvu

[2] N. S. B. Vargas, ""métodos de aprendizaje estadístico para analizar los precios de la vivienda nueva en bogotá entre 2008 y 2019 según las características de los inmuebles, de su ubicación y entorno"," Master's thesis, Universidad Nacional de Colombia, 2021. [Online]. Available: https://repositorio.unal.edu.co/handle/unal/79630

[3] C. Bilbao. (2000) "relación entre el precio de venta de una vivienda y sus características". [Online]. Available: https://dialnet.unirioja.es/servlet/articulo?codigo=3344762l

[4] G. Duque. (2014) "la vivienda social y sus determinantes". [Online]. Available: https://repositorio.unal.edu.co/bitstream/handle/unal/50845/gonzaloduqueescobar.201446.pdf

[5] Y. V. G. Alzate, ""modelo de predicción de precios de viviendas en el municipio de rionegro para apoyar la toma de decisiones de compra y venta de propiedad raíz"," Master's thesis, Institución Universitaria Salazar y Herrera, 2019. [Online]. Available: http://hdl.handle.net/20.500.11912/5285

[6] J. Rodrigo. (2020, Oct.) "gradient boosting con python". [Online]. Available: https://cienciadedatos.net/documentos/py09_gradient_boosting_python.html

[7] P. García, "Implementación de un modelo machine learning para la estimación del valor del metro cuadrado de un inmueble ubicado en Cundinamarca," Master's thesis, Universidad de Los Andes, Bogotá, Colombia, 2021. [Online]. Available: https://repositorio.uniandes.edu.co/bitstream/handle/1992/55114/25539.pdf?sequence=1

[8] S. D. R. Soto, "Modelo de Predicción del Precio de la Vivienda en el Valle de San Nicolás," Master's thesis, Universidad de Antioquia, Medellín, Colombia, 2021. [Online]. Available: https://bibliotecadigital.udea.edu.co/bitstream/10495/24674/7/DavidEstefania_2021_PrecioViviendaOriente.pdf