UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Facultad de Estadística
Trabajo de Grado
Junio 2021

IMPLEMENTACIÓN DE METODOLOGÍAS ESTADÍSTICAS PARA PRONÓSTICOS DE PRECIOS DE VIVIENDAS EN BOGOTÁ

Daniel Isaac Martínez Arroyo.[a]
danielmartineza@usantotomas.edu.co

Wilmer Darío Pineda.[b]
Wilmerpineda@usantotomas.edu.co

Resumen

La adquisición de vivienda ha sido algo indispensable desde hace miles de años, desde que los humanos dejaron de ser nómadas y decidieron asentarse en un lugar en concreto, adicionalmente con el sistema capitalista las viviendas dejaron de ser solo importantes para vivir, ya que también se convirtieron en uno de los principales negocios. Esto es debido a que mueve grandes sumas de dinero y es de los negocios menos riesgosos, pero en contraparte suele no ser muy rentable pues proporcionalmente no presentan un gran incremento en el precio, adicionalmente no se tiene información acerca de cuáles viviendas son las que van a tener una mayor valorización, ni que características son las más relevantes a la hora de determinar el precio final del inmueble, si es el área, el estrato, el número de habitaciones, u otra que tal vez no sea tan evidente.

En el presente trabajo se implementaran diferentes metodologías estadísticas con el fin poder tener un mayor entendimiento sobre qué características determinan el valor del inmueble, pero para esto se necesita obtener la mayor información posible sobre las características de los apartamentos y su precio. Esto se pudo realizar gracias a que la galería inmobiliaria brindó sus bases de datos, las cuales lleva recolectando información sobre los inmuebles de varias ciudades del país desde el 2002. Con esta información se puede obtener una mejor perspectiva y una idea más clara y precisa sobre cuáles son los factores más importantes a la hora de determinar el precio de las viviendas de Bogotá, implementando distintas metodologías estadísticas como los modelos de regresión lineal múltiple, regresión ridge y lasso, vecinos más cercanos, árboles de regresión, random forest y redes neuronales multicapa, y al final comparar los resultados obtenidos entre ellos, para determinar cuál es la mejor metodología para determinar el precio de vivienda de Bogotá.

Después de determinar cuál es el mejor modelo y obtener sus resultados, es posible tomar mejores decisiones a la hora de invertir en una vivienda debido a que al tener mayor conocimiento sobre qué variables son más importantes a la hora de determinar el precio de vivienda, se puede realizar una mejor inversión al hacer mejor uso del dinero, usando la menor cantidad y en contraparte obteniendo la mayor valorización. Adicionalmente las personas que deseen usar sus ahorros para comprar un apartamento, podrán usar esta información para encontrar apartamentos con las características deseadas al mejor precio y tener una idea más clara de los valores reales de los precios de las viviendas, con el fin de no pagar más de lo debido.

***Palabras clave*:** Regresión Lineal Múltiple, Regresión Ridge, Regresión Lasso, Vecino Más Cercano, Árbol De Regresión, Random Forest, Red Neuronal Multicapa, vivienda, finca raíz.

[a]Estudiante Universidad Santo Tomás
[b]Profesor Universidad Santo Tomás

Abstract

The acquisition of housing has been something essential for miles of years, since humans stopped being nomads and decided to settle in a specific place, additionally with the capitalist system the houses stopped being only significant to live, they also became one big business. This is due to the fact that it moves large sums of money and is one of the least risky businesses, but on the other hand it is not usually very profitable, because proportionally it does not present a large increase in price, additionally there is no information about which housing transfers are those that are going to have a greater appreciation, or what characteristics are the most relevant on determining the final price of the property, whether it is the area, the social stratum, the number of rooms, or another that may not be so obvious. In this work, different statistical methodologies will be implemented in order to have a better understanding of what characteristics determine the value of the property, but for this it is necessary to obtain as much information as possible about the characteristics of the apartments and their price. This could be done thanks to the fact that the real estate gallery provided its databases, which have been collecting information on the properties of various cities in the country since 2002. With this information, a better perspective and a clearer and more precise idea about which are the most important factors when determining the price of houses in Bogotá, implementing different statistical methodologies, among those models of multiple linear regression, ridge and lasso regression, closest neighbors, 'regression trees ón, random forest and multilayer neural networks, and at the end compare the results obtained between them, to determine which is the best methodology to determine the house price in Bogotá. After determining which is the best model and obtaining its results, it is possible to make better decisions when investing in a home, because having greater knowledge about which variables are most important when determining the price of a home, a better investment can be made, by making better use of money, using the least amount and in return obtaining the highest valuation. Additionally, people who wish to use their savings to buy an apartment, will be able to use this information to find apartments with desired characteristics at the best price and have a clearer idea of the real values of the prices of the houses, in order not to pay more than it should.

***Keywords*:** Multiple Linear Regression, Regression Ridge,Lasso regression, k-Nearest Neighbors, Regression Tree, Random Forest, Multilayer Neural Network, living place, real estate.

1. Agradecimientos

De antemano, se desea hacer un agradecimiento a la galería inmobiliaria, debido a que ellos brindaron los datos y asesorías necesarias sin las que este trabajo no hubiera podido ser realizado. Por otro lado también se le agradece al tutor Wilmer Pineda, por el asesoramiento y apoyo brindado.

2. Índice

3. Introducción

Desde hace miles de años los humanos dejaron de ser nómadas con el fin de organizar sus civilizaciones en un espacio en concreto, con esto se dio el avance en la arquitectura, dejando de vivir en cuevas o chozas construidas con el único fin de cubrir de la lluvia y de los depredadores de esa noche, a construir hogares cada vez más resistentes y tecnológicos. Por el contrario, hoy en día no es novedad una casa, pues como dice (Garcia. A,2005) la vivienda es una necesidad social en cualquier parte del mundo actual; son pocas las comunidades estrictamente nómadas y aun ellos realizan ciertas formas de arquitectura efímera o se refugian en cuevas realizando adaptaciones al espacio creado naturalmente.

La vivienda no es solo una necesidad básica, también es un negocio, pues desde personas de bajos recursos que al crecer los hijos (que generalmente no llegan a estudiar hasta niveles superiores y por tanto comienzan a trabajar desde muy jóvenes) aportan dinero para la construcción de la vivienda constituyéndose en contribuyentes secundarios al ingreso familiar (Sudra, 1981:43), o en una perspectiva macroeconómica, como en el caso colombiano que el sector inmobiliario en Colombia es un motor estratégico para el desarrollo económico del país y un área prometedora para el futuro (FEDELONJAS, 2016). Según datos del Departamento Nacional de Estadística, la rama económica que más contribuye al PIB es Establecimientos financieros, empresas e inmobiliarias con un 20.9 %, de la cual, el sector inmobiliario genera el mayor aporte (8.5 %), seguido de empresas (6.3 %) y el sector financiero (4.7 %), valores promedio para el periodo 2000-2016 (DANE, 2017). (Casas. J , Torres. N. (2017)).

Por lo anterior, se puede evidenciar que la vivienda es el bien más importante en nuestrá epoca, por lo cual es de suma importancia tener una morada, pero a pesar de que no todos busquen una con las mismas condiciones, si hay algo que se tienen en común, que es la obtención de una con la mayor de las cualidades deseadas a un menor precio. Por esto mismo en este proyecto se desea hacer un estudio de los precios y características de las viviendas de Bogotá, con el fin de implementar diversos modelos estadísticos, para posteriormente comparar los resultados de estos y poder determinar cual de estos presenta las mejores estimaciones, dando así los resultados más precisos y un mejor pronóstico del valor del inmueble .

Para poder determinar cual es la mejor metodología para poder pronosticar los precios, se usa de referencia trabajos realizados anteriormente, un ejemplo de esto es el realizado por Zornoza(2018) donde realizan una predicción del precio de una vivienda mediante un proceso de ciencia de datos, en este proyecto se trabaja con una base de datos disponible (como una competición en la página Kaggle.com), denominada "House Prices", y que describe el proceso de venta de propiedades inmobiliarias individuales en la localidad de Ames (Iowa), entre los años 2006 y 2010. Para dicho proyecto se usan metodologías de minería de datos, como árboles de regresión, vecinos más cercanos (KNN) y Random Forest, entre otros. Los anteriormente mencionados son los que se aplicaran en este trabajo, adicionalmente se implementará otra metodología de minería de datos, que es una de las técnicas mÃ¡s usadas hoy en día, se trata de las redes neuronales, la cual según (Tahir, ul-Hassan, Asghar Saqib, 2016, pág. 50) y (Molino, Cardoso, Ruíz, Sánchez, 2014, pág. 1) es capaz de resolver funciones altamente no lineales en un tiempo corto porque aprenden de los datos que son difíciles de expresar matemáticamente, son âherramientas poderosas para el análisis de señales y la modelación de sistemasâ.

Adicionalmente se tiene en cuenta el trabajo de (Grajales Alzate, 2019), la cual realizó el trabajo de maestría "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". Para este proyecto se implementó una regresión lineal, además también se implementaron árboles de regresión y Random Forest. Tomando en cuenta la regresión lineal, se plantea resolver los problemas de sobreajuste con Modelos Ridge y lasso, pues como menciona (Carrasco Carrasco, 2016), los métodos de regularización son utilizados para la selección del modelo y para evitar el sobreajuste en las técnicas predictivas. A la hora de estimar los coeficientes de regresión por mínimos cuadrados, se pueden presentar problemas de colinealidad, este problema impide obtener estimaciones y predicciones fiables a través de mínimos cuadrados, por lo que se ha de recurrir a los métodos de regresión regularizada como son Ridge y Lasso.

Por otro lado debido a que Bogotá es una ciudad muy desigual, donde hay localidades como Suba, que tienen todos los estratos socioeconómicos, donde un estrato bajo puede estar a una cuadra de uno alto, se ha decidido realizar un cluster espacial, el cual consiste en agrupar las viviendas por localización y ciertas características, como dice (Vaca Ramírez, 2015) La ciudad es un concepto que involucra una gran variedad de aspectos: sociales, económicos, arquitectónicos, geográficos, ambientales, políticos, etc. De ahí la complejidad de estudiarla, pues no sólo basta caracterizar cada uno de dichos aspectos, sino también entender las relaciones que existen entre ellos y su evolución. Para esto mismo se aplica la metodología de DBSCAN jerárquico, pues (Joshi et al,2009) proponen una extensión al algoritmo DBSCAN para agrupar polígonos.

El documento está dividido en seis secciones, empezando con la introducción, posteriormente se muestra los objetivos que tiene el trabajo, la justificación del porqué se hizo, el marco teórico donde se hace una descripción de las distintas metodologías utilizadas, y ya finalmente se muestra el trabajo realizado, empezando con la metodología y finalizando con los resultados obtenidos, para concluir el trabajo con las conclusiones obtenidas con la realización del proyecto.

4. Objetivos

4.1. General

- Proponer distintas metodologías estadísticas, con el fin de obtener la mejor estimación de los precios de vivienda de Bogotá D.C. A la vez que se determina cuales son las variables más influyentes a la hora de establecer dicho precio.

4.2. Específicos

- Contrastar los resultados obtenidos por los distintos modelos, con el fin de identificar cual es la metodología que mejor se acopla al caso de la capital colombiana.
- Identificar las diversas características que presentan los inmuebles, con el fin de conocer cuales presentan una mayor influencia en el precio final de este.

5. Justificación

El negocio de inmuebles es uno de los negocios más seguros que existen ya que las únicas veces que los bienes raíces se devaluaron fue por culpa de la burbuja inmobiliaria, pero por ejemplo la crisis financiera del 2008 fue provocada principalmente por el mal sistema de préstamos que había en esos momentos en Estados Unidos. Sin embargo, en casos normales es un negocio muy seguro, en el cual no solo se gana la valorización a través de los años, sino que también se puede arrendar y así ganar más dinero, o se puede vivir en esta y así disfrutar de esta mientras se valoriza.

Pero ningún negocio es perfecto, pues invertir en finca raíz requiere de una gran inversión, normalmente se debe hacer un gran esfuerzo para comprar el inmueble, por lo cual la decisión de cual comprar es de suma importancia dado que se entra en un costo de oportunidad ya que a pesar de que todas las viviendas pueden ser rentables, no todas tienen la misma valoración, no cuestan lo que debería, son fáciles de vender o arrendar, ni son seguras, entre otros diversos que cada una puede tener. Por esto mismo se plantea tener un mecanismo estadístico que facilite la toma de decisiones, que brinde una mejor perspectiva, con la cual la inversión pueda ser lo más acertada posible.

El problema radica en que estos cálculos estadísticos no son totalmente precisos pues el precio de la vivienda es muy impreciso, para empezar depende de la oferta y demanda la cual no puede ser medida

matemáticamente, adicionalmente hay variables que pueden tener más importancia en una época que en otra. Debido a esto se elabora una gran cantidad de modelos, con el fin de observar cual es el más preciso, esto con el obejtivo de tener los mejores pronósticos y entendimiento de los precios.

El resultado de este estudio es de gran utilidad para una gran variedad de individuos dado que es útil desde la persona que quiere comprar su hogar, pero por esto mismo quiere buscar cual es el mejor precio que podría obtener dependiendo de las particularidades que desee, así decidiendo si prefiere pagar más por cierta característica o si prefiere un costo menor, además de no pagar más de lo debido. Adicionalmente también sirve para todos los inversionistas que deseen comprar un inmueble para luego venderlo, observando cuando una vivienda esté a un precio mucho menor al que en verdad vale, observando proyectos de construcción para comprar o valorando la peculiaridad que tendrán en un futuro.

6. Marco Teórico

Para determinar qué variables son las más importantes a la hora de establecer el precio, se requiere implementar diferentes metodologías, asÃ se aplican modelos lineales, adicionalmente se aplican metodologías de machine learning, aprovechando la gran información obtenida, con el fin de crear modelos de aprendizaje que con ayuda de la información adquirida anteriormente puedan predecir valores futuros. Con estos resultados se aplican criterios de evaluación y así ver cual de estas metodologías es la mejor. Por otra parte, también se aplican metodologías de cluster para poder agrupar las viviendas en lugares cercanos y con características en común. No obstante, antes de esto se requiere de la implementación de metodologías que permitan seleccionar las variables a partir de un criterio estadístico ya que así se reduce la dimensión de los datos sin perder información.

6.1. Vivienda

El sector de la vivienda es reconocido por su enorme influencia en la estabilidad y desempeño de las economías. Tanto a nivel internacional como local, se han observado episodios donde auges desproporcionados en el mercado de la vivienda han contribuido al desarrollo de crisis financieras y económicas.(Castaño, J., Laverde, M., Morales, M,A ., Yaruro, A,M,2013)). Y es que la vivienda es uno de los activos más importantes para las personas, sino el más importante, pues este no tiene riesgo de pérdida, por lo cual le brinda seguridad al propietario, adicionalmente le brinda un techo, calor y seguridad en caso de vivir en ella, por otro lado, puede significar un ingreso extra, en caso de arrendar el inmueble o dedicarse a compra y venta, con el fin de ganar la valorización. La vivienda es un bien de consumo durable que en la gran mayoría de países constituye el activo más importante de los portafolios de los hogares (Holmes, Otero Panagiotidis, 2011). Igualmente, la localización de las viviendas puede tener efecto sobre el acceso a oportunidades laborales, sociales y a bienes públicos locales (OSullivan, 2012).

A pesar de la importancia del sector inmobiliario, es frecuente encontrar deficiencias en la disponibilidad y análisis de los precios de alquiler y venta de las viviendas, lo que dificulta la predicción de los retornos de las inversiones en el sector, así como sus posibles efectos en otros sectores de la economía (Caplin, Chopra, Leahy, LeCun Thampy, 2008). Pues el comportamiento de las viviendas es demasiado variable, depende de la ciudad, estrato, ubicaciún, si es casa o apartamento. Según DiPasquale y Wheaton (1996), la vivienda se caracteriza por ser un bien de carácter heterogéneo, en el sentido que se diferencia en términos de tamaño y otras características, así como por su localización. Mientras en otros mercados el precio se define según una cantidad fija del bien (e.g. precio por una libra de naranjas o por un galón de gasolina), en los mercados de vivienda se observa el gasto por cada vivienda individual.(Cárdenas Rubio, J. A., Chaux Guzmán, F. J., Otero, J, 2019).

El mercado de vivienda y la evolución de su precio es un tema recurrente en la literatura internacional dado su impacto en la economía. De acuerdo con Syz (2008) la vivienda representa cerca de un tercio del total de la riqueza en el mundo y según (Eurostat,2013) es el principal componente dentro del gasto

de los hogares. De otro lado, según Hill (2011) el comportamiento del mercado inmobiliario tiene una fuerte repercusión en el resto de la economía; los ciclos de dicho mercado afectan el nivel de consumo, la distribución de la riqueza, y la estabilidad financiera de los países. Así mismo, este autor sugiere que los índices de precios de vivienda constituyen una herramienta útil para la política fiscal y monetaria, así como para los mercados financieros, en la medida que actúan como un barómetro del estado de la economía (Castaño, J., Laverde, M., Morales, M,A ., Yaruro, A,M, 2013).

6.2. DBSCAN Jerarquico

El algoritmo de agrupación espacial basada en densidad de aplicaciones con ruido, o en inglés Density-based spatial clustering of applications with noise(DBSCAN). Es un algoritmo de agrupamiento por densidad, el cual busca regiones donde se presente una gran concentración de puntos, que se encuentren alejadas ya sea en distancias o características de otras. Según (Cedeño Mata, sf) estos métodos permiten construir agrupaciones densas y de cierto tamaño, y a diferencia de la mayoría de algoritmos, no producen agrupaciones totales de los datos, sino que pueden dejar algunos elementos sin clasificar en el caso de que estén muy lejos de los centroides de los grupos, de acuerdo con unos parámetros especificados. Pero para poder ejecutar este algoritmo se debe especificar el epsilon, que es el radio máximo de la vecindad, y se debe precisar el mínimo de elementos requeridos dentro del epsilon, para poder formar el cluster.

Este algoritmo fue propuesto por Martin Ester, Hans-Peter Kriegel, Jörg Sander y Xiaowei Xu, con el propósito de que fuera fácil de implementar, pues este algoritmo se basa en la búsqueda de puntos centrales o "core" en los grupos. Estos puntos centrales poseen un área o región de vecindad para un radio determinado que contiene al menos un número mínimo de puntos, es decir, que su área de vecindad excede un umbral determinado (M. Ester, H.-P. Kriegel, S. Jorg, and X. Xu, 1996). Otra cosa a tener en cuenta es que la densidad de los puntos depende del radio de la región de vecindad especificada. De este modo, si el radio es suficientemente grande todos los puntos tendrán una densidad igual al número de puntos total del conjunto de datos. Por el contrario, si es muy pequeño todos los puntos tendrán una densidad igual a 1, es decir, el punto se encontrará aislado. (M. Gallardo Campos, 2009).

El DBSCAN ya ha sido aplicado con éxito en varios casos, unos de estos son la detección de usos en las tierras a partir de imágenes satélite, la creación de perfiles de usuarios en Internet mediante la agrupación de sesiones Web, o el agrupamiento de bases de datos de imágenes en histogramas en color facilitando la búsqueda de imágenes similares (H. Kriegel, 2000).

El principal objetivo al implementar el DBSCAN es encontrar todos los puntos centrales, estos puntos centrales son aquellos que tienen una región de vecindad que contiene un número mínimo de puntos para el radio determinado anteriormente. Por otro lado, la forma de la región de vecindad está determinada por la elección de la medida de la distancia entre dos puntos, $dist(p, q)$. Un ejemplo de esto es cuando se utiliza la distancia de Manhattan, DBSCAN tiende a formar grupos rectangulares.

La región de vecindad de un punto p perteneciente a una base de datos D dado un radio Eps se define como:

$$N_{Eps}(P) = \{q \in D | dis(p, q)\} \leq Eps$$

Con lo definido anteriormente se da una primera versión del algoritmo, donde se busca que para cada punto se cumpla el criterio del mínimo de puntos en la vecindad, aunque toca tener en cuenta que este algoritmo no funciona debido a que en el centro de los grupos se genera una mayor densidad que en los bordes, por esto mismo en el conjunto de datos se pueden distinguir tres tipos de puntos:

- *Puntos centrales o core.*: Son los puntos que están ubicados en el interior del grupo. Para poder identificar a un punto central, se observa a aquel que número de puntos en la región de vecindad definida por el radio *Eps* supera el umbral definido por el $MinPts$. Mejor dicho, en el caso que se cumple que $N_{Eps}(p) \geq MinPts$.

- *Puntos frontera o de borde*: Estos puntos se caracterizan por no ser puntos centrales, sino que pertenecen a la región de vecindad de uno o más puntos centrales.
- *Puntos de ruido*: Estos puntos se distinguen por no ser ni centrales ni de frontera, o sea que no no pertenecen a la vecindad de ningún punto.

Para que el algoritmo tenga un correcto funcionamiento, se necesita que para cada punto p, de un grupo C, exista otro punto q en C, de tal manera de p de encuentre dentro de la región de vecindad q, ademas que esta región de vecindad contenga el mínimos de puntos estipulados. Con esto se define:

$Definición1$: Un punto p es densamente alcanzable de manera directa (directly density-reachable) desde un punto q dados el radio Eps y el número mńimo de puntos $MinPts$ si:

- $p \in N_{Eps}(q)$, esto significa que p pertenece a la región de vecindad de q
- $|N_{Eps}(q)| \geq MinPts$, es decir q es un punto central (M.Ester,1996)

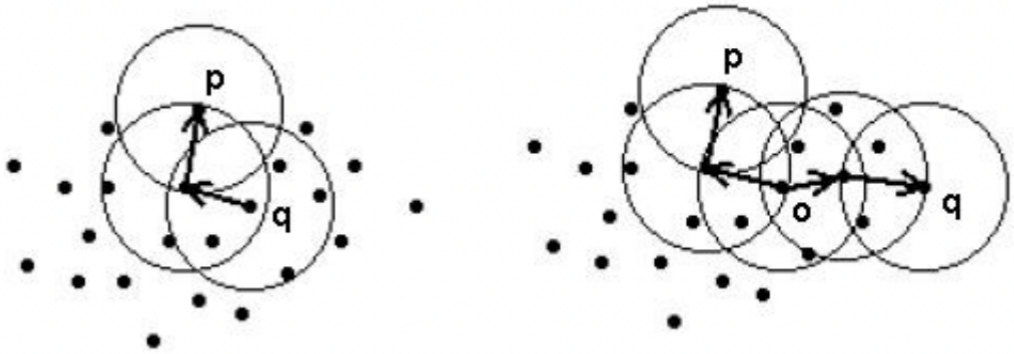
$Definición2$: Un punto p es densamente alcanzable (density-reachable) desde un punto q dado el radio Eps y el mínimo de puntos $MinPts$, en el caso que exista una secuencia de puntos $p_1, \cdots, p_n, p_1 = q, p_n = p$ tal que p_{i+1} es densamente alcanzable directamente desde p_i

Dos puntos de borde de un mismo grupo pueden no ser densamente alcanzables entre ellos, esto es debido a que existe la posibilidad de que la condición de que sean centrales no se cumpla para ambos. Por otro lado, debe haber un punto central en el grupo C desde el cual los puntos de borde sean densamente alcanzables. Con lo anterior se da la siguiente definición:

$Definicion3$: Un punto p está densamente conectado (density-connected) a un punto q dados el radio Eps y el número mínimo de puntos $MinPts$, si existe un punto tal que ambos p y q, son densamente alcanzables desde o dados Eps y $MinPts$. La relación densamente conectado es una relación sistemática y reflexiva(M.Ester,1996).

Para poder darse una idea más clara de esto, se cuenta con las siguientes figuras. En la figura [a] se observa como el punto p es densamente alcanzable para el punto q, pero por otro lado q no es densamente alcanzable para p, debido a que la región de vecindad de p, no cuenta con el mínimo de puntos $MinPts$. La figura [b], evidencia como p y q se encuentran densamente conectadas desde el punto o

Imagen de (Gallardo Campos, M., 2009)



(a) Punto densamente alcanzable. (b) Puntos densamente conectados.

Después de ser definidos los conceptos de alcanzabilidad y conectividad, se puede proseguir a dar una definición formal de grupo y ruido, dados en el conjunto de datos D, el radio de la región de vecindad Eps y el mínimo de puntos $MinPts$.

*Definición*4: Dado un conjunto de datos D, se declara que un grupo C con radio *Eps* y el mínimo de puntos $MinPts$, es un subconjunto no vacío de D, que cumple las siguientes condiciones:

- $\forall_{p,q}$ si $p \in C$ es densamente alcanzable desde p, dados *Eps* y $MinPts$, entonces $q \in C$ (Maximitud)
- $\forall_{p,q} \in C : q$ está densamente conectado a p dados *Eps* y $MinPts$. (Conectividad)(M.Ester,1996).

$Definicion5$ Sean $C_1, \cdots, C_K$ los grupos del conjunto de datos D dados los parámetros *Eps* y $MinPts$, $i = 1, \cdots, K$. Entonces, se define el ruido como aquellos datos pertenecientes a D que no forman parte de ningún grupo C_i , es decir: $ruido = p \in D | \forall_i : p \notin C_i$ (M.Ester,1996).

Cabe resaltar que dado un grupo C con un radio *Eps* y un mínimo de puntos $MinPts$, contiene este último debido a que C cuenta con al menos un punto p, el cual debe estar conectado a otro punto o (El cual puede ser el mismo punto p). Por lo cual, al menos un punto o debe cumplir la condición de que punto central, lo cual resulta en que la región de vecindad o contenga al menos el mínimo de puntos $MinPts$.

Para validar la corrección del algoritmo de agrupamiento que se quiere definir, son necesarios los siguientes lemas. Estos lemas contemplan que dados los parámetros *Eps* y $MinPts$, se puede determinar un grupo, al seguirse dos pasos esenciales. Lo primero es elegir un punto aleatorio del conjunto de datos, que cumpla con la condición de ser un punto central. La siguiente instancia es determinar todos los puntos que son densamente alcanzables desde el punto inicial, con esto se obtiene el grupo.

*Lema*1 Sea p un punto en D y $|N_{Eps}(p)| \geq MinPts$. Entonces el conjunto $O = \{o | o \in D$ y o es densamente alcanzable desde p dados *Eps* y $MinPts\}$ es un grupo dados *Eps* y $MinPts$(M.Ester,1996).

No hay que ignorar que al haber un grupo C, con *Eps* y $MinPts$ conocidos, sea determinado únicamente por alguno de sus puntos centrales. No obstante, todos los puntos punto C, son densamente alcanzables desde cualquiera de los otros puntos centrales C, por lo cual un grupo C, contiene textualmente los puntos que son densamente alcanzables desde cualquiera de los puntos centrales de C.

*Lema*2 Sea C un grupo tal que son conocidos *Eps* y $MinPts$, y sea p cualquier punto en C tal que $|N_{Eps}(p)| \geq MinPts$. Entonces C es igual al conjunto definido por: $O = \{o | o \in D$ y o es densamente alcanzable desde p, dados *Eps* y $MinPts\}$(M.Ester,1996).

6.3. Selección de variables

Para seleccionar variables de un conjunto de datos, se cuenta con múltiples metodologías que buscan reducir la dimensionalidad a partir del conjunto óptimo que contenga las variables con mejor calidad de información, las cuales minimicen la pérdida de información y a su vez maximicen la variabilidad de los datos (James, Witten, Hastie Tibshinari, 2013). De tal forma, se requiere de un criterio de selección que evalúe todo el subconjunto de variables hasta lograr la combinación óptima. Para esto, se han desarrollado diferentes métodos, a continuación se nombrarán los principales.

6.3.1. Método de Forward Selection

Es un algoritmo de búsqueda secuencial, inicialmente considera cada una de las variables individualmente y selecciona la que le de mejor criterio de información, de igual forma va agregando las variables que le permiten tener un mejor criterio. El algoritmo finaliza cuando obtiene la mejor combinación de variables, las cuales en este caso minimicen el criterio de información de Akaike (AIC) (Gualdron, 2006).

El AIC está dado por la siguiente fórmula:

$$AIC = 2k - 2ln(L)$$

Donde:

- k Es el número de parámetros del modelo
- L Es el máximo valor de la función de verosimilitud para el modelo estimado.

6.3.2. Método de Backward Selection

Al igual que le método de Forward Selection, es un algoritmo de búsqueda secuencial, sin embargo es todo lo contrario que este método. El Backward comienza con el modelo completo y va eliminando variables a partir del criterio de selección de AIC hasta que llega a la combinación de variables que lo minimicen (Gualdron, 2006).

6.3.3. Método Stepwise Selection

El método de selección Stepwise es una combinación de los dos métodos mencionados anteriormente. Es decir, consiste en agregar y eliminar variables, así una variable que se incluyó y que luego fue eliminada puede volver a ser incluida en las combinaciones posteriores del modelo. De igual forma, el críterio de selección es minimizar el AIC.

6.4. Modelos

6.4.1. Regresíon lineal multiple

La regresión es uno de los pilares estadísticos más modernos el cual hace referencia al análisis simultáneo de dos o más variables relacionadas entre sí. (Pat Fernandez, Martínez Menchaca, Pat Fernández, Martínez Luis, 2013). Adicionalmente la regresión lineal múltiple es la expansión del modelo de regresión simple a k variables explicativas, la cual trata de ajustar modelos lineales o linealizables entre una variable dependiente y más de una variables independientes. En este tipo de modelos es importante testar la heterocedasticidad, la multicolinealidad y la especificación (Montero Granados. R, 2016).

En el modelo de regresión múltiple existe una variable dependiente y y otras variables independientes x, para entender mejor esto se muestra su estructura, la cual es:

$$y = f(x_1, \cdots, x_k) + \epsilon$$

Donde:

- y: Es la variable explicada, dependiente o respuesta.
- $x_1, \cdots, x_k$: Son las variables explicativas, regresores o variables independientes.
- $y = f(x_1, \cdots, x_k)$: Es la parte determinista del modelo.
- ϵ: Representa el error aleatorio. Contiene el efecto sobre y de todas las variables distintas de $x_1, \cdots, x_k$ (Carrasquilla-Batista, A,2016).

Pero como tal el modelo es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_0 k x_k$$

Donde:

- y: : es la variable dependiente, la cual también es denominada variable respuesta.
- x_i: es la variable independiente i, la cual también se llama exploratoria.
- β_i: es el coeficiente del modelo para la variable x_i

Tanto la variable independiente como las independientes deben ser métricas, aunque las independientes también pueden tener valores cualitativos(Forero Gómez, G.,2020).

Por último, según (Garcia, Morales Serrano, González Cavazos, 2013) las etapas para poder realizar una regresión lineal múltiple, son:

1. Identificar problema o área de oportunidad.
2. Seleccionar las variables dependientes e independientes.
3. Recolectar variables.
4. Realizar análisis descriptivo del tipo de relación entre variables
5. Seleccionar método.
6. Calcular coeficientes del modelo de regresión lineal múltiple para construir la función.
7. Identificar problemas de colinealidad o multicolinealidad.
8. Realizar prueba global de la ecuación.
9. Efectuar pruebas individuales de los coeficientes.
10. Probar cumplimiento de los supuestos del análisis.
11. Interpretar coeficientes de determinación, correlación, determinación ajustado y error estándar.
12. Analizar los coeficientes de la ecuación de regresión.
13. Elaborar pronósticos puntuales y por intervalo.

6.4.2. Regresíon Ridge

La regresón Ridge fue propuesta originalmente por Hoerl y Kennard, con el propósito de eludir los efectos adversos que provoca la colinealidad en un modelo lineal que fuera estimado por mínimos cuadrados, es decir p<n. Esta regresión es extremadamente similar a mínimos cuadrados, pero la diferencia radica en que los coeficientes estimados por Ridge $\hat{\beta}^{ridge}$, son valores que minimizan.

$$\hat{\beta}^{ridge} = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 - \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^2$$

Donde $\lambda \geq 0$ es el parámetro de contracción que se determinará por separado.

El método Ridge tiende a contraer los coeficientes de regresión al incluir el término de penalización en la función objetivo: cuanto mayor sea λ, mayor penalización y, por tanto, mayor contracción de los coeficientes. (Carrasco Carrasco, M.,2016)

Una forma similar de plasmar el problema de Ridge es:

$$\hat{\beta} = arg_{\beta}min\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \quad s.a\sum_{j=1}^{p}\beta_j^2 \leq t,$$

donde t es el parámetro de penalización por complejidad.

Sea $\hat{\beta}^{MCO} = (X'X)^{-1}X'Y$ la estimación por mínimos cuadrados de β. Se propone que la baja estabilidad de $\hat{\beta}^{MCO}$, puede ser resuelta añadiendo una constante $\lambda \geq 0$, a cada uno de los términos de la diagonal $(X'X)$, aunque antes de invertir la matriz resulta que:

$$\hat{\beta}^{Ridge}(\lambda) = (X'X + \lambda I_p)^{-1}X'Y$$

donde I es la matriz $p * p$. Hay que tener en cuenta que esta penalización se le aplica a los coeficientes $\beta_1, \cdots, \beta_p$ pero no a β_0. Ajustamos el modelo sin término independiente, estimando este mediante $\bar{y} = sum_{i=1}^{n} \frac{y_i}{n}$. Es excluido de la penalización para evitar que el resultado dependa del origen en la variable Y. (Carrasco Carrasco, M.,2016).

Este difiere al método de mínimos cuadrados, en qué Ridge crea un conjunto diferente para cada λ, en vez de producir un único vector de coeficientes estimados.

si = 0 se deduce que es la metodología de mínimos cuadrados. Si por otro lado $\lambda \to \infty, \hat{\beta} \to 0$ y hay un estimador sesgado de β. Es decir, se introduce un sesgo con la consecuencia de que se reduce la varianza. Con el fin de impedir que la penalización varíe por los cambios de escalas de las variables, ya que normalmente estas se estandarizan previamente (media 0 varianza 1). También cabe resaltar que Y se supone centrada, por lo cual la matriz X tendrá p columnas y no $p + 1$. La ventaja de Ridge es que comúnmente genera predicciones más precisas que MCO y una mejor selección de variables. Al tener estimado el coeficiente, se debe buscar el valor de λ, tal que $0 < \lambda < \infty$, esto con el fin de minimizar el error esperado de la predicción. Esto es aplicable para este método y para los otros de validación cruzada.

Aunque cabe resaltar que esta metodología presenta un inconveniente, el cual es que contrae todos los coeficientes a 0, pero por más pequeños e irrelevantes que sean no llega a conseguir la nulidad de ninguno de ellos. Esto provoca que no se produzca una selección de variables, dejando en el modelo final todas las variables. Esto ocasiona inconvenientes en estudios en los que se tiene un gran número de p variables explicativas o predictoras. Para solucionar este problema se propone la regresión Lasso.

6.4.3. Regresíon Lasso

Con el propósito de encontrar una regresión lineal la cual con ayuda de la contracción de los coeficientes, logre estabilizar las estimaciones y predicciones, a la vez que realiza la selección de variables. (Tibshirani,R. ,1996) propuso la técnica Lasso (least absolute shrinkage and selection operator). Es una técnica de regresión lineal regularizada, como Ridge, con una leve diferencia en la penalización que trae consecuencias importantes.

Esto se da principalmente desde tal valor del parámetro de complejidad, el estimador de Lasso realiza estimaciones nulas para algunos coeficientes y no nulas para otros, con esto Lasso realiza una elección de variables de forma continua. Lasso reduce la variabilidad de las estimaciones por la reducción de los coeficientes y al mismo tiempo produce modelos interpretables por la reducción de algunos coeficientes a cero (Carrasco Carrasco, M.,2016).

El auge en los últimos años en la investigación y aplicación de tecnicas Lasso se debe principalmente, a la existencia de problemas donde $p\, n$ y al desarrollo paralelo de algoritmos eficientes (Tibshirani,R. ,1996).

Lasso soluciona el problema de mínimos cuadrados del vector de coeficientes:

$$\hat{\beta}^{lasso} min_{\beta}\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2\} \quad s.a \sum_{j=1}|\beta_j| \leq s$$

O de una manera semejante, minimizando

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij})^2 + \lambda \sum_{j=1}|\beta_j| = RSS + \lambda \sum_{j=1}|\beta_j|$$

Siendo s y $\lambda \geq 0$ los respectivos parámetros de penalización por complejidad. De nuevo reparametrizamos la constante β_0 mediante la estandarización de los predictores, y por tanto, la solución $\hat{\beta}_0 = \bar{y}$ (Carrasco Carrasco, M.,2016).

Comúnmente, los modelos generalizados Lasso presentan una mayor facilidad a la hora de interpretarlos en comparación con los modelos Ridge. Al igual que en Ridge, se debe seleccionar el mejor valor de λ, por el método de validación cruzada.

La ventaja de Lasso es que genera un modelo fácil de interpretar y con una buena precisión, pero al igual que todas las metodologías presenta varias desventajas, como:

- En el caso $p > n$, Lasso selecciona a lo sumo n variables antes de saturarse, debido a la naturaleza del problema de optimización convexa. Esto parece ser una limitación para un método de selección de variables. Además, Lasso no está bien definido a menos que el límite de la norma L1 de los coeficientes sea menor que un cierto valor.
- Si hay un grupo de variables entre las cuales las correlaciones por parejas son muy altas, entonces Lasso tiende a seleccionar sólo una variable del grupo, sin importar cuál de ellas selecciona.
- Para el caso $n > p$, si hay una alta correlación entre predictores, se ha observado que, en general, la predicción a través de regresión Ridge resulta más óptima que la obtenida a través de Lasso. (Carrasco Carrasco, M.,2016).

Se han realizado algunos estudios con el fin de comparar Ridge y Lasso y se ha llegado a las siguientes conclusiones:

- Al hacer Lasso una selección de variables, posee una gran ventaja sobre Ridge, debido a que genera modelos más simples y de mayor interpretabilidad. Por otro lado, no hay un modelo predominante sobre el otro.
- Es de esperarse que en caso de tener una base de datos que cuente con pocos predictores, tiene coeficientes sustanciales y los demás predictores tuvieran coeficientes muy bajos o incluso cero, sería de mayor utilidad la regresión Lasso.
- En caso de tener muchos factores predictivos, que tengan coeficientes de un tamaño aproximado, se recomienda la regresión Ridge. Aunque cabe resaltar que el número de predictores que tienen relación con la variable respuesta, no se conocen con anterioridad.
- La validación cruzada es útil a la hora de determinar que metodología es mejor para determinado caso.
- Al tener una estimación de mínimos cuadrados con una alta varianza , es de gran utilidad utilizar la regresión Ridge o Lasso, con la desventaja de tener un pequeño aumento en el sesgo. Aunque igualmente esto puede provocar una mejor predicción.

6.4.4. Vecino más cercano

El algoritmo de vecinos más cercanos, o por sus siglas en inglés k-NN (k-Nearest Neighbors), es uno de los algoritmos de clasificación más utilizados hoy en día, debido a que este clasifica los datos deseados a través de ejemplos conocidos con anterioridad, llamado conjunto de entrenamiento o train. Este método calcula la distancia de cada dato a clasificar con todos los ejemplos obtenidos de la base de entrenamiento, con eso clasifica el dato a pronosticar con base en los datos de entrenamiento más cercanos al conjunto train.

Para empezar, el algoritmo K-NN calcula la distancia entre cada dato nuevo a clasificar con los datos de entrenamiento. Los nuevos datos tienen una serie de variables, las cuales tienen cada una un distinto

rango de valores, por lo cual un rango de valores muy alto o muy pequeño puede contrastar con las demÃ¡s escalas de valores. Por esto mismo es necesario normalizar todas las variables, para poder dejarlas en un rango entre 0 y 1, y con esto todas las variables afectarían de igual manera a la hora de buscar el vecino más cercano. Para esto a cada variable se le resta el valor mínimo que pueda tener y se divide por la diferencia entre el valor máximo y mínimo.

$$e_k^j = \frac{e_k^j - min^j}{max^j - min^j}, con\ k = \{1, \cdots, P\}$$

Una vez que todas las variables tienen el mismo rango, se puede medir la distancia del nuevo dato a clasificar con respecto a los datos de entrenamiento. Para poder calcular esa distancia se pueden implementar los siguientes métodos:

- **Distancia Euclídea:** Es la que se utiliza de manera predeterminada, es la unión en línea recta entre los dos puntos.

$$d_e(e_i, e') = \sqrt{\sum_{j=1}^{N}(e_i^j - e'^j)^2}$$

- **Distancia Manhattan:**: La distancia entre dos puntos es la suma de las diferencias absolutas entre sus coordenadas.

$$d_m(e_i, e') = \sum_{j=1}^{N} |e_i^j - e'^j|$$

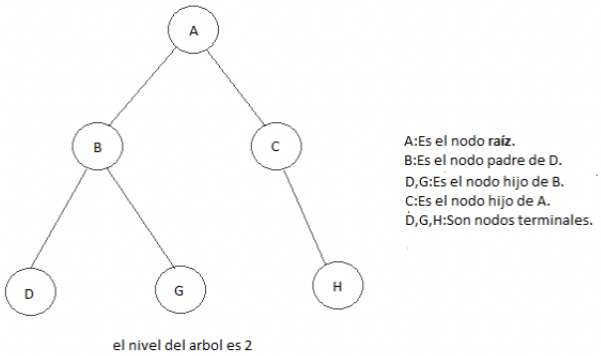
- **Distancia Canberra:** Similar a la distancia Manhattan pero más receptiva a los puntos de datos más cercanos al origen.

$$d(p, q) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

- **Distancia Máximo:** La distancia entre dos puntos se calcula como la diferencia mayor en cualquiera de las dimensiones de coordenadas (Berástegui Arbeloa, G ,2018)

6.4.5. Árboles de regresión

Un árbol es un gráfico conexo acíclico dirigido formado por un conjunto finito de nodos conectados, pero para poder entender mejor se muestra la siguiente imagen tomada de (Ejea Carbonell, D, 2017)



Definiendo sus partes, para empezar un nodo es la unidad sobre la que se construye el árbol, adicionalmente este nodo puede tener varios nodos, los cuales se les denomina nodo hijo, en contraparte el nodo padre es aquel que posee al menos un nodo hijo y el nodo sin padre se llama nodo raíz y solo existe uno, pues es del cual parte el árbol. Por otro lado existen también nodos hermanos, que son aquellos que tienen el mismo padre. Para finalizar existen los nodos terminales u hoja , estos son los últimos de la ramificación, es decir aquellos que no tienen ramificaciones o nodos hijos.

Adicionalmente existen otras partes a tener en cuenta en los árboles, por un lado están las ramas, las cuales son el camino(unión de arcos simples) desde la raíz a un nodo hoja. También hay que tener en cuenta el grado de un nodo, el cual es el número de descendientes directos que posee cada nodo, y el nivel el cual es número de arcos que han de recorrer desde la raíz, a la cual se le presupone nivel 1. Por último el grado de un árbol es el máximo grado de sus nodos, y la altura es el máximo de los niveles.

Con lo anterior, se puede concluir que cada nodo no terminal, va a representar una proposición lógica sobre una variable predictora, adem´as cada camino a una región u hoja diferente del espacio que es la intersección de la secuencia de proposiciones lógicas. Por otro lado los árboles de regresión derivan su nombre por el aspecto que tienen. Los árboles se pueden ver como una regla de análisis no paramétrico la cual se basan en dividir el espacio predictor en una serie de regiones simples las cuales incluyen observaciones con valores similares para la variable respuesta.

Por último hay que tener en cuenta que para poder hacer una predicción dentro de cada hoja se utiliza la media de los datos de entrenamiento que están ubicados en la misma región del espacio multidimensional de predictores en caso de los árboles de regresión y se le asignará una observación a la clase más votada en los árboles de clasificaci´on.

6.4.6. Random Forest

El método de Random Forest realiza principalmente clasificación y regresión, es decir explica el comportamiento de una variable a partir del realizamiento de otras. También, es utilizado para reducir dimensionalidad (cuando se tiene gran volumen de datos), identificar outliers, estimar datos faltantes, entre otras funciones de exploración de los datos (Sullivan, 2017).

Para la clasificación, el algoritmo emplea una amplia cantidad de árboles de decisión múltiples, es decir realiza un ensamble de modelos de predicción débiles, en este caso árboles de decisión, en donde a partir del agrupamiento se genera un modelo de predicción fuerte, siendo así una mejora de los árboles de decisión.

Cuando se emplea el Random Forest para clasificación, cada árbol de decisión elige una clase, de tal forma que finalmente el algoritmo escoge la clase con mayor cantidad de votos. Es decir, el número de casos de un conjunto de datos es N y se toma una muestra aleatoria con reemplazo (Boostrap) de tamaño n, y de las M variables de entrada se selecciona aleatoriamente m, tal que $m < M$ sean específicas para cada nodo. Esta muestra funciona como un conjunto de datos que construye el árbol a la vez de que aumenta el tamaño de este lo más que puede, adicionalmente en el algoritmo no se realiza podamiento del árbol (Orellana, 2018).

Para realizar el algoritmo, por lo general se dividen los datos en dos grupos homogéneos, en donde se obtiene un conjunto de entrenamiento y otro de prueba, esto se realiza con el fin de validar que el proceso no presente problemas de sobre ajuste y sesgo. Además se debe considerar que el conjunto de prueba debe corresponder a un tercio de los datos y son conocidos como OOB (Out of Bag).

Cabe mencionar que se obtienen mejores resultados cuando se realiza clasificación en lugar de regresión, debido a que las predicciones están limitadas al rango del conjunto de datos de entrenamiento.

6.4.7. Redes Neuronales

Las Redes Neuronales Artificiales (RNA), también conocidas como ANN (Artificial Neural Networks), están inspiradas como dice su nombre en las redes neuronales biológicas del cerebro humano, debido a que están construidas de tal manera que se comportan de forma similar en sus funciones más comunes. Estas se encuentran formadas por un conjunto de elementos llamados nodos o neuronas, las cuales están conectadas entre sí por conexiones que tienen un valor numérico modificable llamado peso.

La actividad que una unidad de procesamiento o neurona artificial realiza en un sistema de este tipo es simple, normalmente consiste en sumar los valores de las entradas (inputs) que recibe de otras unidades conectadas a ella, comparar esta cantidad con el valor umbral y, si lo iguala o supera, enviar activación o salida (output) a las unidades a las que esté conectada. Tanto las entradas que la unidad recibe como las salidas que envía dependen a su vez del peso o fuerza de las conexiones por las cuales se realizan dichas operaciones. (MONTAÑO MORENO. J, 2002)

Las redes neuronales realizan tres etapas, con el fin de generar el aprendizaje y poder realizar la predicción, estas etapas son:

- **Aprender:** Adquirir el conocimiento de una cosa por medio del estudio, ejercicio o experiencia. Las ANN pueden cambiar su comportamiento en función del entorno. Se les muestra un conjunto de entradas y ellas mismas se ajustan para producir unas salidas consistentes.
- **Generalizar:** Extender o ampliar una cosa. Las ANN generalizan automáticamente debido a su propia estructura y naturaleza. Estas redes pueden ofrecer, dentro de un margen, respuestas correctas a entradas que presentan pequeñas variaciones debido a los efectos de ruido o distorsión.
- **Abstraer:** Aislar mentalmente o considerar por separado las cualidades de un objeto. Algunas ANN son capaces de abstraer la esencia de un conjunto de entradas que aparentemente no presentan aspectos comunes o relativos (Basogain Olabe. X, s.f).

7. Metodología

7.0.1. Base de datos

Para el trabajo se necesito la mayor cantidad de información posible sobre el precio de los apartamentos de Bogotá y sus características, y se obtuvo una gran informaciń con ayuda de la empresa privada **Galería Inmobiliaria** la cual se dedica a censar viviendas nuevas en 18 ciudades en Colombia y Ciudad de Panamá, además de hacer seguimiento a vivienda usadas en 5 ciudades del país y mantener actualizada la cobertura de locales comerciales, bodegas y oficinas (La Galería inmobiliaria, s. f.).

La base de datos brindada contaba con la información de los precios de apartamentos, casas y lotes de Bogotá y varios municipios de Cundinamarca desde febrero del 2002 hasta enero del 2021. Adicionalmente contaba con una diversidad de características de los inmuebles que brindaban características acerca la ubicación, las empresas que gestionaron los proyectos y las financiaciones, particularidades físicas de la vivienda entre otra serie de variables lo cual brindaba una clara idea sobre cada inmueble.

Al empezar con el proyecto se realizaron diversos cambios a la base de datos, como realizar un filtrado de ésta, el cual consistió en solo dejar los datos propios para el estudio, los cuales serían los apartamentos que se encontraran en Bogotá y sus municipios propiamente aledaños, donde se buscaba que estos lugares contarán con precios y características similares a las de la capital, además de que sus habitantes tengan la posibilidad de ir a Bogotá a trabajar diariamente. Por esto mismo se trabajó con los municipios de Soacha, Funza, Mosquera, Cota, Chía y La Calera.

Posteriormente se prosiguió a eliminar las viviendas que tuvieran datos faltantes en alguna de las variables, esto se pudo realizar debido a que la base de datos era muy grande por lo cual no había una

pérdida de información considerable, además que era más recomendable que simular los datos, ya que esta simulación podría afectar los pronósticos. Por último, se hicieron arreglos generales a la base, como dejar el piso de los apartamentos y la cantidad de garajes en la misma escala dado que la mayoría de apartamentos tenia área 0 en la terraza y balcones y en ocasiones tenían terrazas o balcones muy grandes, que generaban datos atípicos, no se tomó en cuenta el área sino que se tiene en cuenta si tenía o no esta característica.

Después de tener la base arreglada, se prosiguió a tomar las mejores variables, por lo cual se realizó un modelo lineal múltiple para observar cuáles variables son significativas para determinar el precio, posteriormente se realizaron pruebas de criterio de seleccin como el forward, backward y el stepwise, con el fin de hacer un filtrado mayor de variables. Luego de esto, se seleccionaron las variables estrato, área, baños sin terminar, baños completos, garajes, piso, fecha, la cantidad de ventas del inmueble por mes, la ubicación y si contaba con patio, jardín, terraza o balcón, estudio, o con depósito.

7.0.2. Descriptivos

Después de obtener los datos con los que se pretende trabajar, se prosigue a hacer una visualización de los datos para tener una idea más clara de estos y tener un primer acercamiento, con lo cual se tiene una mayor claridad de las características y comportamiento de las viviendas de Bogotá. Para empezar se observaron las variables numéricas, y para tener una visión general de estas se visualizó el mínimo, máximo, la media y mediana. Adicionalmente se calculó la matriz de correlaciones con el fin de ver la relación que había entre variables, principalmente de las otras con el precio.

Posterior a eso se realizaron unos gráficos, donde se observa la relación del precio con el estrato, el año, el área y el número de alcobas, mostrando en promedio cuanto es el valor de los inmuebles, dependiendo de las diferentes características y ver la tendencia que se va marcando.

7.0.3. Modelos

Al tener claridad sobre los datos, se decide dividir la base en 2 con la aplicación de un muestreo aleatorio simple (MAS), una va a contar con el 80 % de los datos, la cual será la base de entrenamiento (train) que sirve como su nombre lo indica para entrenar los modelos. Mientras que el 20 % restante, se utilizará para probar que tan eficaces son estos modelos, la cual será la base de prueba (test). Después de entrenar todos los modelos con la base de prueba y aplicarlos a la base test, se compara qué tan precisos son las predicciones con la prueba de “error absoluto medio porcentual” (MAPE) y el “error cuadrático medio” (RSME), para posteriormente observar cual de los modelos implementados presenta los menores errores y así se concluye cual es el mejor.

Se comienza aplicando el modelo más simple, una regresión lineal múltiple, con el fin de intentar predecir el precio con las demás variables, posteriormente se aplican modelos Ridge y Lasso, con el caso de comprobar si hay variables que no necesiten el mismo peso y por esto buscar una mejor predicción en el modelo. Luego, se aplican metodologías de machine learning aprovechando el tamaño de la base, se empieza aplicando vecinos más cercanos, para proseguir con árboles de decisión y random forest y por último redes neuronales.

7.0.4. clustering

Para poder mejorar los pronósticos se tomó en cuenta el componente espacial debido a que Bogotá es una ciudad muy desigual donde un estrato bajo puede estar al lado de uno alto, donde un apartamento lujoso puede estar al lado de uno sencillo, en general donde hay una gran disimilitud entre las viviendas. Por lo anteriormente dicho, se plantea un DBSCAN jerárquico con el objetivo de crear grupos de viviendas, los cuales no estén conformados solo por la cercanía geográfica entre ellas, sino que se busca que estos

grupos tengan características en común, buscando crear conjuntos muy homogéneos que entre ellos sean muy heterogéneos.

En este trabajo se realizó el DBSCAN agrupando las viviendas por localización, estrato, el área, número de alcobas, cantidad de garajes, el año y el piso. Este clustering se realizó con un mínimo de puntos de 11 y dio como resultado 111 clúster. Con la segmentación anterior se vuelve a realizar el procedimiento de entrenar los modelos y posteriormente comparar los resultados obtenidos, así se puede evidenciar sí la componente espacial mejora o no las predicciones.

8. Resultados

8.0.1. Descriptivos

Para empezar se realiza una observación de los datos, primero con una tabla que muestra su máximo, mínimo, media y mediana, posteriormente con la matriz de correlación y por último con unas gráficas.

Tabla 1. Descriptivos

Variable	Mínimo	Mediana	Media	Máximo
Área	14.2 m^2	67.8 m^2	76.0 m^2	599.0 m^2
Alcobas	1	3	2.4	4
Baños Completos	0	2	2.01	8
Baños Sin Terminar	0	0	0.106	2
Garajes	0	1	1.13	5
Piso	1	4	5.45	50
Precio en Millones	24.8	251.0	328.0	8454.3

Para entender mejor esta tabla cabe aclarar primero la diferencia entre la mediana y la media. La mediana es el dato que se encuentra en la mitad al haber organizado estos de menor a mayor. Por otro lado, la media es el promedio de estos datos, que resulta de sumar todos los valores y dividirlos por n, que es la cantidad de datos sumados.

Al observar la tabla podemos evidenciar que hay varios datos atípicos (Datos fuera de lo normal), pues se evidencia que en promedio los apartamentos tienen un área de 76 m^2 y que el de la mitad es de 67,8 m^2, pero es evidente que hay apartamentos demasiados grandes ya que el de mayor tamaño es de 599 m^2, el cual es casi 8 veces el promedio. Por otro lado, también es de extrañar la vivienda que tiene 8 baños, la que se encuentra en el piso 50 y la que tiene un precio de 8454.3 millones.

Para tener claridad sobre esto, se identificaron estos casos. Se observó que el apartamento de mayor precio (8454.3 millones), esta ubicado en el barrio de los rosales, el " Edificio 5-93", es estrato 6 y con un area de 469 m^2 , piso 15. Además se observó que el conjunto con 8 baños es el "Camino del Bosque Aptos", también estrato 6 y con un área de 300 m^2. Por último, se evidencia que los pisos 50 son del Bacata, el edificio más alto Bogotá.

A continuación se realiza la matriz de correlación, para poder observar qué variables presentan mayor relación a la hora de determinar el precio:

Tabla 2. Matriz de correlaciones

A primera vista se observa que de las variables numéricas la que menos ejerce una influencia sobre el precio es el piso y el número de alcobas, lo cual puede sorprender a muchos. En contraparte, las variables que más ejercen influencia sobre el valor del inmueble son el área, seguido por la cantidad de garajes y

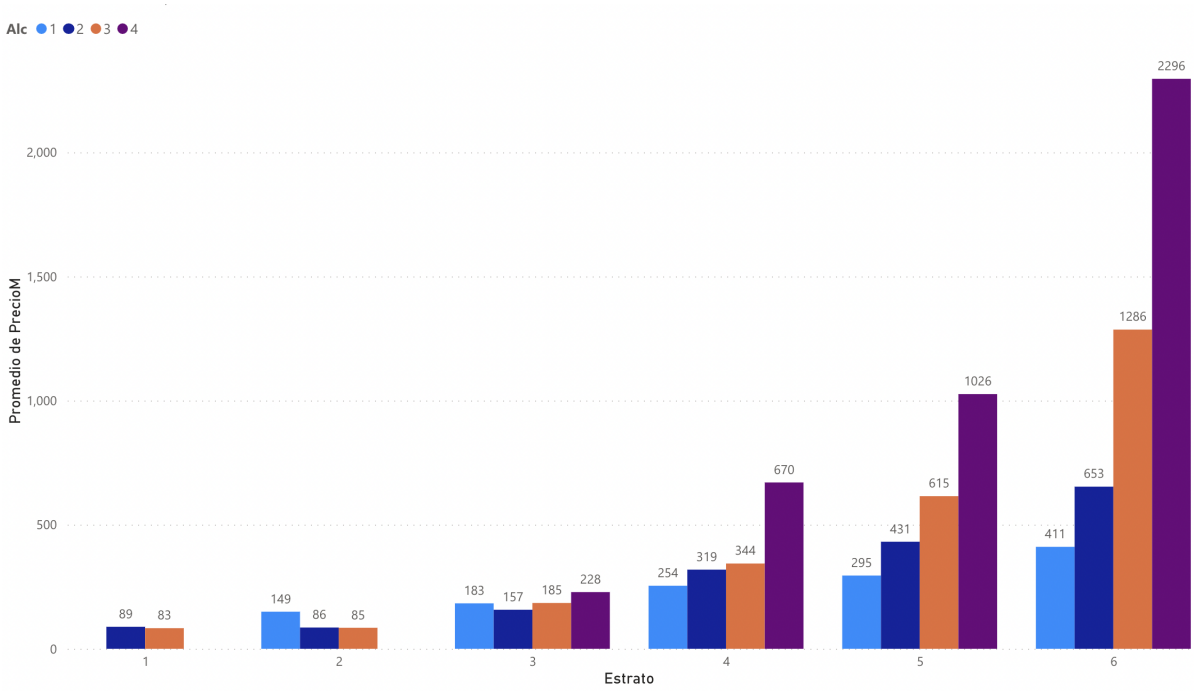
Correlación	Área	Alcobas	B C	B S T	Garajes	Piso	Precio
Área	1.000000	0.445809	0.812972	-0.200600	0.781874	0.040367	0.813052
Alcobas	0.445809	1.000000	0.328155	0.149628	0.189022	0.045402	0.138844
Baños Completos	0.812972	0.328155	1.000000	-0.426243	0.787340	0.039905	0.681970
Baños Sin Terminar	-0.200600	0.149628	-0.426243	1.000000	-0.375138	-0.081423	-0.239493
Garajes	0.781874	0.189022	0.787340	-0.375138	1.000000	0.027577	0.717884
Piso	0.040367	0.045402	0.039905	-0.081423	0.027577	1.000000	0.072755
Precio en Millones	0.813052	0.138844	0.681970	-0.239493	0.717884	0.072755	1.000000

la cantidad de baños completados. Por último, se evidencia que aunque entre más baños completados mayor es el precio, en contraparte entre más baños haya sin terminar, menor será el valor.

Adicionalmente la matriz también nos permite observar otras relaciones que hay entre las variables, como que al ser un apartamento grande es de esperarse que tenga más baños y tiene una mayor posibilidad de tener más parqueaderos. Por otro lado se observa que una alta cantidad de baños sin terminar puede significar con una vivienda de bajas características, pues entre más baños en este estado posea el apartamento, este tiende a tener menos parqueaderos, ser un apartamento pequeño y ser un inmueble de menor valor.

A continuación se muestran unas gráficas para observar de mejor manera algunas características y comportamientos.

Gráfico 1. Comportamiento del precio por estrato



Para empezar se realiza un gráfico entre el precio promedio de los inmuebles y el número de alcobas, dividido por los diferentes estratos, como se puede evidenciar no es el mismo comportamiento en todos los sectores. Algo que es sorprendente es que en los estratos 1 y 2 a medida que hay más habitaciones, el precio es menor y que en el estrato 3 el valor de 1 habitación sea mayor que con 2 y casi el mismo que con 3. Por otro lado el resto de estratos si tienen el comportamiento esperado, aunque de estos cabe resaltar el gran salto que se presenta cuando hay 4 habitaciones, pues se ve un gran incremento en el precio cuando se pasa de 3 a 4 habitaciones.

Adicionalmente, se evidencia con esto que entre mayor sea el estrato, el cambio en el precio por cantidad de habitaciones es mayor, pues se observa que no hay una gran diferencia en los precios del estrato 3, pues cabe resaltar que el precio del inmueble no es tan alto, en comparación a la diferencia entre los precios los estratos 5 y 6, que es donde se evidencia una mayor discrepancia entre los precios. Por último, se puede evidenciar una gran diferencia entre los precios y el estrato, pues se observa que el incremento en el precio es casi del doble desde el estrato 3, mientras que el 1 y 2 si son más homogéneos.

Prosiguiendo con la segunda gráfica, en esta se puede observar el comportamiento del precio en la última década categorizado por estratos. Aquí se puede evidenciar lo anteriormente dicho, que a medida que va incrementando el estrato, también aumenta la diferencia de precios entre ellos. Adicionalmente, se observa que los precios tienen el comportamiento esperado, que a medida que pasan los años las viviendas van incrementando su valor, aunque es notorio el efecto que tuvo el Covid-19 sobre el precio de estos dado que se aprecia que al estrato 2 y 6 no lo afecta, el estrato 3 se estancó y los 4 y 5 vieron una caída en sus precios.

Adicionalmente se contempla que no todos los estratos se afectan igual, pues un ejemplo es que en el 2017 el estrato 6 tuvo una caída, mientras los demás a excepción del 2 seguían su crecimiento normal, además en el 2019 solo el estrato 4 vio una baja en sus precios. Por lo anterior se puede evidenciar que normalmente todos suelen tener un incremento en el precio a través de los años y a excepción de un acontecimiento tan drástico como el Covid 19, los golpes en los precios son independientes entre sí.

Gráfico 2. Comportamiento del precio a través de los años categorizado por estrato

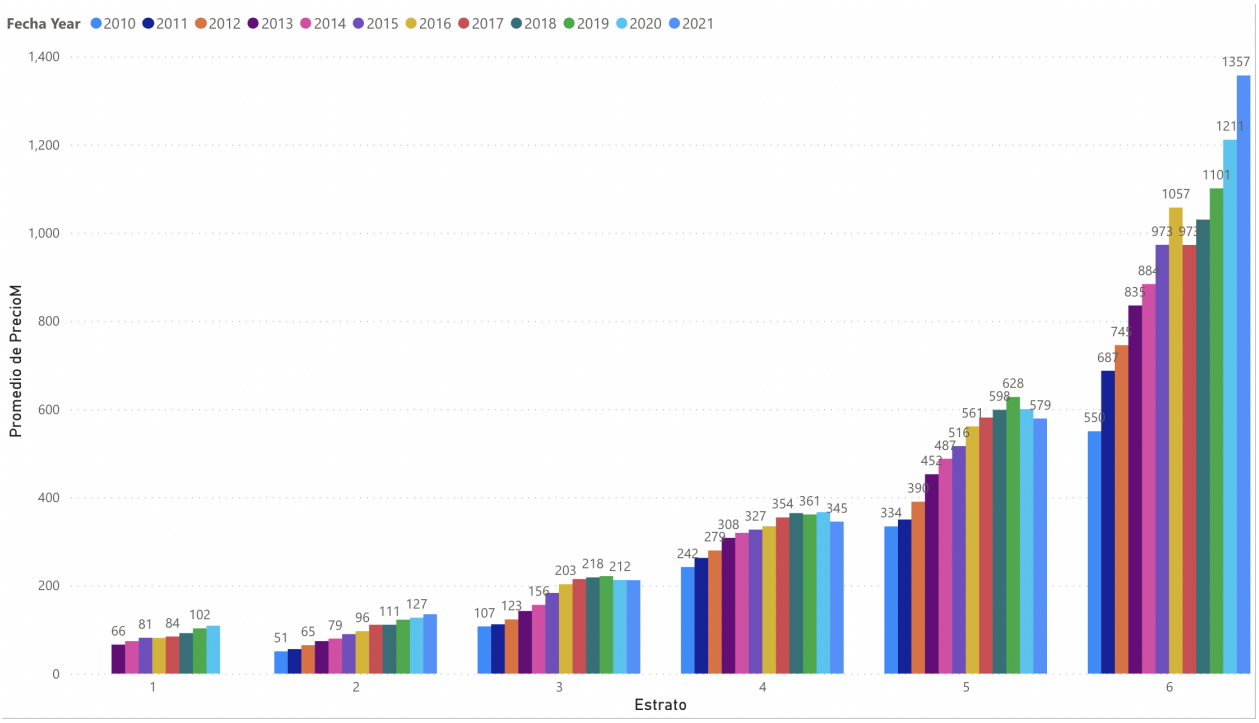


Gráfico 3. Comportamiento del precio por área

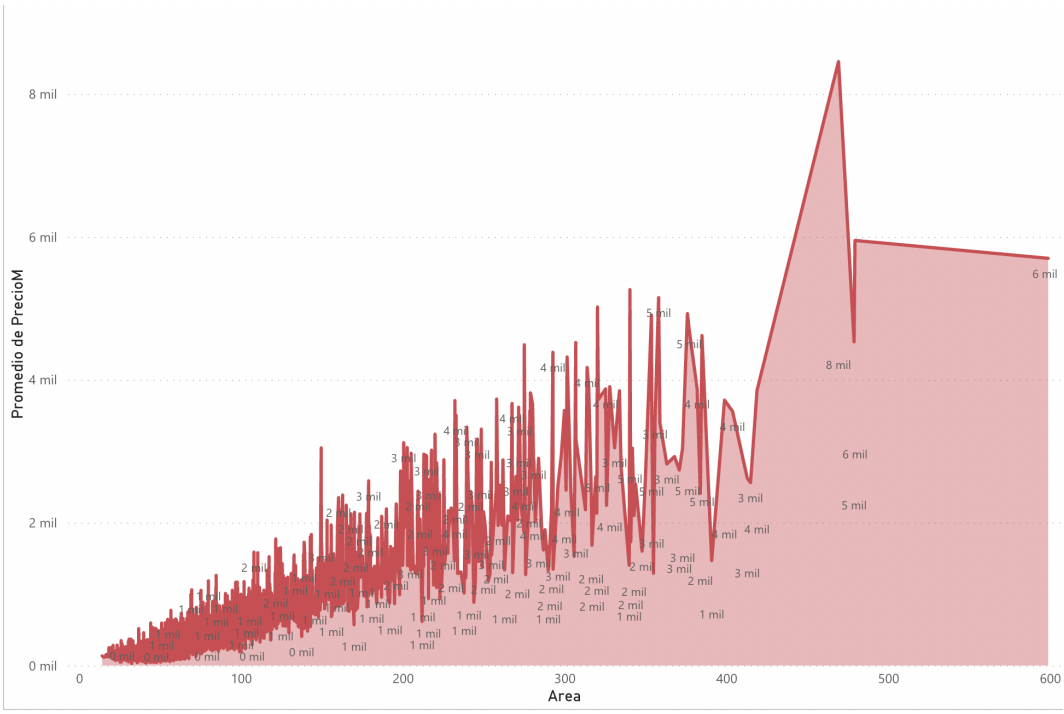
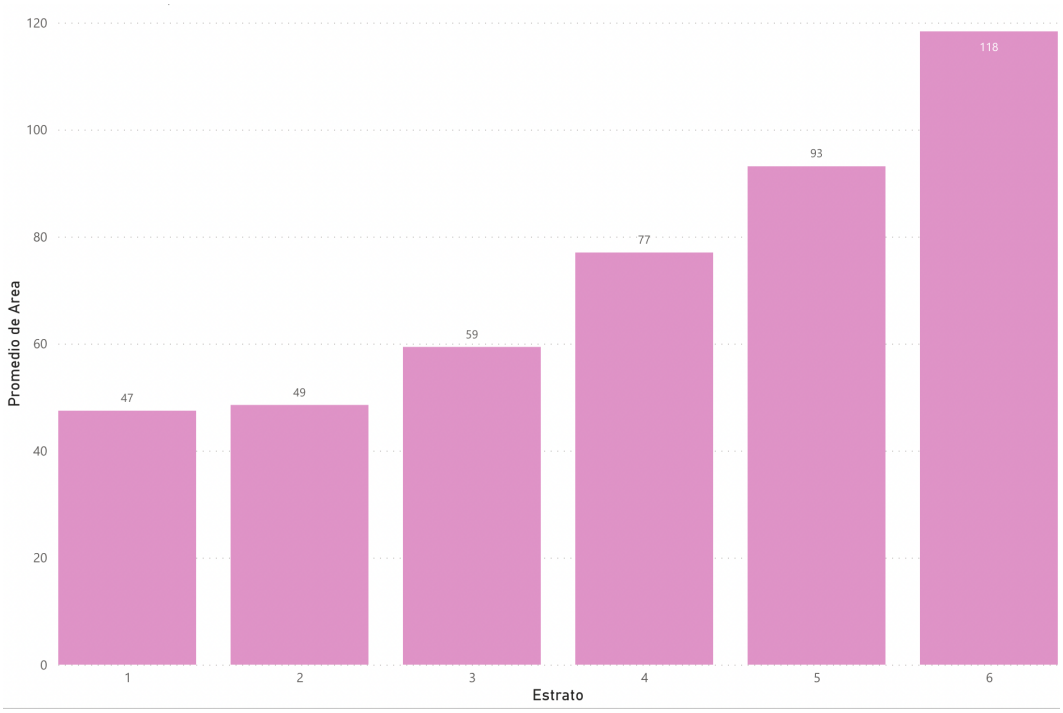


Gráfico 4. Comportamiento del área por estrato



La relación que existe entre el precio, área y estrato se muestra en dos gráficos (Gráfico 3 y 4), para que tenga una mejor visualización. En este caso ocurre lo esperado, a medida que sube el estrato los apartamentos tienden a ser más grandes, y como en la mayoría de los casos, a medida que sube el estrato también incrementa la diferencia entre ellos. Por otro lado, también se observa que a pesar de que haya una gran volatilidad en los precios y el área (probablemente debido al estrato y otras características), esta tiene una tendencia creciente, por lo cual se comprueba que en la mayoría de los casos que es mayor el área también se tiene un precio más alto.

8.0.2. Modelos

Como se mencionó anteriormente se realizó un modelo lineal múltiple, una regresión Ridge, una lasso, se entrenó el modelo con vecinos más cercanos, árbol de regresión, random forest y redes neuronales. Estos fueron entrenados con dos bases de entrenamiento, una sin aplicar cluster y otra con el DBSCAN . Después de aplicar los modelos a las bases de prueba se calculó el RSME (Error cuadrático medio) y el MAPE (Error absoluto medio porcentual) y se hace la comparación entre todos con el fin de determinar cuál es el que presenta el menor error, por ende el que da los resultados más precisos.

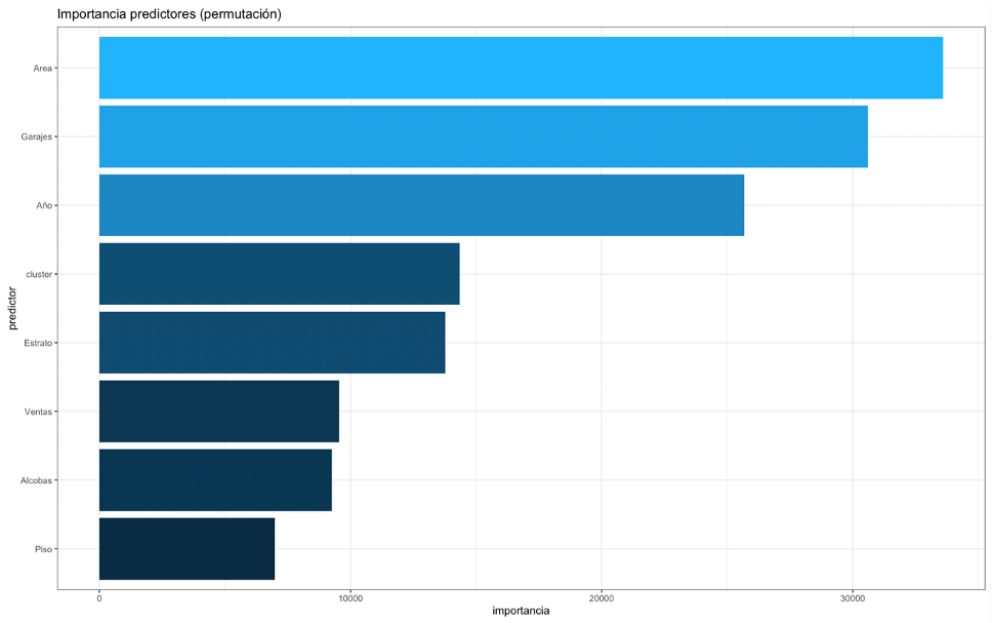
En la Tabla 3 se puede evidenciar la gran diferencia en estimaciones que se da al aplicar el DBSCAN, por lo cual se recomienza su aplicación para tener mejores resultados, por otro lado también se observa que las metodologías de machine learning son mucho más precisas que las regresiones empleadas debido a que la base de datos al ser tan grande genera un muy buen entrenamiento de los modelos, adicionalmente la combinación entre variables afecta de sobremanera el precio, por lo cual se evidencia que no existe una relación lineal entre las características y el precio del inmueble.

Tabla 3. Resultados y comparación de los modelos

.	Sin Clustering		DBSCAN	
Modelo	MAPE	RSME	MAPE	RSME
Lineal	0.4385135	159.0221	0.1135286	73.49984
Ridge	0.4012881	166.9025	0.1307968	78.76077
Lasso	0.3912518	161.9088	0.1288001	74.2221
KNN	0.0813622	68.36507	0.0002084037	28.06952
Árbol	0.37632	166.3094	0.1811463	72.11232
RF	0.08930116	63.36703	0.0027638	17.4
Red Neuronal	0.2013011	111.9784	0.06143072	35.9038

Por otro lado, se observa que los modelos que presentan los mejores resultados son la metodología de vecinos más cercanos (KNN) y el random forest. En este proyecto se seleccionó como mejor modelo el random forest, que da un error en la precisión de 17 millones, y adicionalmente presenta una carga computacional menor que los vecinos más cercanos. El modelo de random forest fue entrenado con 50 árboles, una profundidad de 20 y un número de predictores en cada división de 3.

Gráfico 5. Importancia de las variables



Por último se muestra en la gráfica 5 cuales son las variables más influyentes a la hora de determinar el precio, como se puede evidenciar la más significativa es el área, pues como se vio en los descriptivos mostraba una gran correlación entre estas variables, además de verse la tendencia creciente. Posteriormente, se aprecia que los garajes y el año también ejercen una gran influencia en el precio por lo cual se hace la recomendación de que si venden por separado el apartamento y parqueadero, comprar así sea para arrendar parqueaderos, pues en los años posteriores, cuando ya se haya completado la venta de los parqueaderos, los que tengan este beneficio van a tener un precio considerablemente mayor de quienes no lo tengan.

Por otro lado, se ve una influencia menor de la ubicación y del estrato, que a pesar de ser variables muy importantes, no son tan determinantes a la hora de determinar el precio. Ya posteriormente de las variables tomadas en cuenta, las menos influyentes son la cantidad de apartamentos vendidos, el número de alcobas, aunque cabe aclarar que cuando pasa de 3 a 4 si hay un mayor incremento, y el piso del

inmueble, aunque este último podria ser mas significativo al tener una mejor diferenciación en la base de datos. .

9. Conclusiones

Al realizar el anterior trabajo se observó que es extremadamente complicado hacer una precisión totalmente acertada de un valor del inmueble, pues hay una variedad de factores externos como econímicos, construcciones nuevas en el sector, seguridad de la zona, entre otras que no se pueden predecir. Pero por otra parte, se pudo hacer un acercamiento al valor del inmueble teniendo algunas características básicas, considerando que entre más información se tenga, mejor podría ser el modelo, puesto que el problema de la recolección en los pisos afecta la precisión de esa variable.

Por otro lado, hay variables que podrían ser influyentes y no son tenidas en cuenta, como saber si la ventana recibe el sol en las mañanas, como muchos prefieren, si la vista es hacia dentro del conjunto o hacia afuera, si tiene colegios, centros comerciales, estaciones de transmilenio u otras estructuras de importancia cerca. Así, como se pudo observar la ubicación de la vivienda es de suma importancia ya que la zona donde está ubicada es indispensable a la hora de determinar el precio, pues así existan dos edificios con las mismas características, al estar separados por unas cuadras puede afectar drásticamente el precio.

Adicionalmente, se reafirma la importancia del trabajo de la Galería Inmobiliaria debido a que sin esta información hubiera sido imposible realizar este trabajo, puesto que no solo se limitan a recolectar los datos, sino que la base de datos tiene una gran cantidad de registros, lo cual provoca que las metodologías de machine learning sean tan efectivas, sumandole la ventaja de que con el tiempo van a dar mejores resultados, al tener más datos que entrenen estos modelos.

Recitando lo dicho en el trabajo, se evidencio que hay una gran desigualdad en la ciudad, puesto que hay una gran diferencia de los precios entre los estratos. Pero esto no es lo que más determina el valor del inmueble dado que lo que ejerce una mayor influencia es el tamaño del apartamento. Posteriormente se recuerda la importancia de comprar parqueaderos en los proyectos que vendan esto por separado, pues el garaje es la segunda variable más influyente a la hora de determinar el valor del apartamento.

También se corroboró la antítesis de que a través de los años sube el precio de los inmuebles, o al menos en la mayoría de los casos, pues a excepción de caso extremos como el Covid-19 o una burbuja inmobiliaria, o casos puntuales en donde un sector socioeconómico se ve afectado, estos valores tienden a la alza.

Por último se valora que este trabajo puede ser usado por diferentes tipos de individuos, pues es útil desde la persona que se quiera comprar su primera vivienda y desee conseguir la que tenga las características más importantes para él o ella, a el menor precio. También es útil para personas o empresas que deseen invertir en viviendas sobre planos o que se encuentren en remate y esperen recibir en los años próximos la mayor valorización posible. Por último se recomienda para la empresa o inversionistas que deseen construir un edificio, para que lo equipen con la suma de características que incremente el precio y con esto sus ganancias sean mayores.

10. Referencias

- Abbott. J., Klaibber. H (2011). An Embarrassment Of Riches: Confronting Omitted Variable Bias And Multiscale Capitalization In Hedonic Price Models
- Basogain Olabe, X. REDES NEURONALES ARTIFICIALES Y SUS APLICACIONES. Escuela Superior de Ingeniería de Bilbao, EHU.
- Berástegui Arbeloa, G. (2018). Implementación del algoritmo de los k vecinos más cercanos (k-NN) y estimación del mejor valor local de k para su cálculo. Universidad publica de Navarra.

- Calhoun C. A., 2001, "Property Valuation Methods and Data in The United States", Housing Finance International, 16(2): 12 - 23
- Caplin, A., Chopra, S., Leahy, J. V., LeCun, Y., Thampy, T. (2008). Machine learning and the spatial structure of house prices and housing returns. Disponible en http://dx.doi.org/10.2139/ssrn.1316046
- Cárdenas Rubio, J. A., Chaux Guzmán, F. J., Otero, J. (2019). Construcción de una base de datos de precios y características de vivienda para Colombia. Revista de Economía del Rosario, 22(1), 75-100. doi: . http://dx.doi.org/10.12804/revistas. urosario.edu.co/economia/a.7768
- Casas. J., Torres. N. (Casas Bazurto, J., Torres Amezquita, N. (2017). Plan de negocio Inmobiliaria Casas Torres S.A.S. Retrieved from https://ciencia.lasalle.edu.co/finanzas_comercio/94)Plan de negocio Inmobiliaria Casas Torres S.A.S
- -Castaño, J., Laverde, M., Morales, M,A ., Yaruro, A,M. (2013). Índice de Precios de la Vivienda Nueva para Bogotá: Metodología de Precios Hedónicos*. Reporte de estabilidad financiera
- Carrasquilla-Batista, A; Chacón-Rodríguez, A; NúñezMontero, K; Gómez-Espinoza, O; Valverde, J; GuerreroBarrantes M. Regresión lineal simple y múltiple: aplicación en la predicción de variables naturales relacionadas con el crecimiento microalgal . Tecnología en Marcha. Encuentro de Investigación y Extensión 2016. Pág 33-45
- arrasco Carrasco, M. (2016). TÃCNICAS DE REGULARIZACIÃN EN REGRESIÃN: IMPLEMENTACIÓN Y APLICACIONES. Universidad de Sevilla.
- Ceular, N. Caridad Ocerín, JM (2001). Un análisis del mercado de la vivienda a través de redes neuronales artificiales. Estudios de Economía Aplicada, 18 (2), 41-66. [Fecha de Consulta 3 de Febrero de 2021]. ISSN: 1133-3197. Disponible en: https://www.redalyc.org/articulo.oa?id=301/3011821
- Cedeño Mata, M. Técnicas de clustering aplicadas a la discriminación de pigmentos en espectroscopía Raman. Universitat PolitÈcnica de Catalunya (UPC)
- Chaphalkar, N. B., Sandbhor, S. (2013). Use of artificial intelligence in real property valuation. International Journal of Engineering and Technology, 5(3), 2334â2337.
- DANE. (2017). Departamento Administrativo Nacional de Estadistica. Recuperado el 03 de Marzo de 2017, de DANE Cuentas Trimestrales Nacionales Producto Intero Bruto: http://www.dane.gov.co/index.php/estadisticasportema/cuentasnacionales/cuentasnacionalestrimestrales#pibporramadeactividad
- DiPasquale, D., Wheaton, w. c. (1996). Urban economics and real estate markets. Englewood Cliffs, nj: Prentice Hall.
- Ejea Carbonell, D. (2017). Árboles de Regresión. Algunos algoritmos y extensiones a métodos de consenso.. Universidad Zaragoza.
- Eurostat (2013), âHandbook on Residential Property Price Indices (RPPI)â, Methodologies and Working Papers
- FEDELONJAS. (16 de Septiembre de 2016). Federación Colombiana de Lonjas de Propiedad Raiz. Recuperado el 05 de Marzo de 2017, de Noticias: Sector inmobiliario demostró ser clave para impulsar desarrollo económico del país: http://www.fedelonjas.org.co/sector-inmobiliariodemostro-ser-clave-para-impulsar-desarrollo-economico-del-país
- Forero Gómez, G. (2020). MODELO DE REGRESIÓN LINEAL MÚLTIPLE PARA EL PRONÃSTICO DE VENTAS DE BOLSAS ECOLÃGICAS PARA LA EMPRESA BOLECO SA, EN LA CIUDAD DE BOGOTÁ DC. UNIVERSIDAD COOPERATIVA DE COLOMBIA
- Gallardo Campos, M., 2009. Aplicaci´on De Técnicas De Clustering Para La Mejora De Aprendizaje. Universidad Carlos III De Madrid.

- GARCÍA, A (2005). Vivienda, familia, identidad. La casa como prolongación de las relaciones humanas. Trayectorias, VII (17), 43-56. [Fecha de Consulta 2 de Febrero de 2021]. ISSN: 2007-1205. Disponible en: https://www.redalyc.org/articulo.oa?id=607/60722197006
- García, N., Gamez, M., Alfaro, E. (2006). Redes neuronales artificiales: un largo e irregular camino hasta la actualidad. En H. d. estadística
- Garzón , J. M., Cardozo, D. (2019). Cerros orientales de Bogotá: una aplicación hedónica al precio de la vivienda para la localidad de Chapinero para el año 2018. Retrieved from https://ciencia.lasalle .edu.co/economia/907
- Gigi Andres . (2020, 4 julio). Web Scraping: encontrando la propiedad más barata. Muestreo, encuestas y estadśticas oficiales. https://predictive.home.blog/2020/07/03/web-scraping-encontrando-la-propiedad-mas-barata/amp/?__twitter_impression=true
- Grajales Alzate, Y. (2019). 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 (MAESTRÍA). UNIVERSIDAD PONTIFICIA BOLIVARIANA
- Gualdron, O. (2006). Desarrollo de diferentes métodos de selección de variables para sistemas multisensoriales. Tarragona, España: Universitat Rovira I Virgili. Tomado en: https://www.tdx.cat/bitstream/handle/10803/8473/Tesis Oscar Gualdron.pdf?...1
- H. Kriegel, âDensity-based cluster- and outlier analysis.â Website, 2000. http://www.dbs.informatik.uni-muenchen.de/Forschung/KDD/ Clustering/index.html.
- Hill, R. (2011), âHedonic Price Indexes for housingâ, Statistics Working Papers
- Holmes, M. J., Otero, J., Panagiotidis, T. (2011). Investigating regional house price convergence in the United States: Evidence from a pair-wise approach. Economic Modelling, 28(6), 2369-2376
- James, G., Witten, D., Hastie, T. and Tibshinari, R. (2013) An Introduction to Statistical Learning (with applications in R). California, United States: Springer New York Heidelberg Dordrecht London.
- Joshi, D., Samal, A. K., Soh, L. K. (2009, March). Density-based clustering of polygons. In Computational Intelligence and Data Mining, 2009.CIDM'09. IEEE Symposium on (pp.171-178). IEEE.
- Knight, J. R., Dombrow, J. Sirmans, C. F. (1995), âA varying parameters approach to constructing house price indexesâ, Real Estate Economics 23(2), 187â205.
- La Galería inmobiliaria. (s. f.). La Galería Inmobiliaria. Recuperado 3 de julio de 2021, de http://www.fincaraiz.com/
- Lever, G. (2000). Determinantes del precio de la vivienda en Santiago: Una estimación Hedónica. Paper. Santiago, Chile: Editorial.
- M. Ester, H.-P. Kriegel, S. Jorg, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," 1996.
- M. Gallardo Campos, "Aplicación de técnicas de clustering para la mejora del aprendizaje", Universidad Carlos III de Madrid, Leganés, 2009
- Molino, E., Cardoso, J., Ruíz, A., Sánchez, J. (2014). Comparación de redes neuronales artificiales y análisis armónico para el pronóstico del nivel del mar (estero de Urías, Mazatl'an, México). Ciencias Marinas, 254.
- MONTAÑO MORENO, J. (2002). Redes Neuronales Artificiales aplicadas al Análisis de Datos. UNIVERSITAT DE LES ILLES BALEARS.

- Montero Granados. R (2016): Modelos de regresión lineal múltiple. Documentos de Trabajo en Economía Aplicada. Universidad de Granada. España
- OâSullivan, A. (2012). Urban economics. Boston: McGraw-Hill, Irwin. Sevillano Pérez, F. (2015). Big Data. EconomÃa industrial, 395, 71-86
- Pat Fernandez, L. A., Martínez Menchaca, A. H., Pat Fernández, J. M., Martínez Luis, D. (2013). Introducción a los Modelos de Regresión. Ciudad del Carmen: Plaza y Valdes. Obtenido de https://ebookcentral.proquest.com
- Peterson, S., Flanagan, A. B. (2009). Neural Network Hedonic Pricing Models in Mass Real Estate Appraisal. Journal of Real Estate Research, 31(2), 147â165. Retrieved from http://ares.metapress.com /index/M3H27210W6411373.pdf
- Rosen S., 1974, "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition", Journal of Political Economics, 82: 34 â 55
- RUMELHART, D.E.; HINTON, G.E. y MCCLELLAND, J.L. (1986): Learning representations by backpropagation. Nature, 323, pp. 533-536
- Stanley M., Alastair A., Dylan M. and D. Patterson, 1998, "Neural Networks: The Prediction of Residential Values", Journal of Property Valuation Investment, 16(1): 57 â 70
- Sullivan, W. (2017) Machine Learning For Beginners Algorithms, Decision Tree Random Forest Introduction. Carolina del Sur, United States: CreateSpace.
- Syz, J. (2008), Property Derivatives: Pricing, Hedging and applications, John Wiley and Sons Inc
- Viridiana, P, L. (2017) EL USO DE WEB SCRAPING PARA LA EXTRACCIÓN DE DATOS, (11)
- Tahir, F., ul-Hassan, T., Asghar Saqib, M. (2016). Optimal scheduling of electrical power in energy-deficient scenarios using artificial neural network and Bootstrap aggregating
- Tibshirani,R. (1996). Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society. Series B (Methodological), 58(1): pÃ¡gs. 267-288
- Vaca RamÃrez, F. (2015). Segmentación geométrica de la ciudad de Quito por medio de un método de aprendizaje no supervisado. UNIVERSIDAD SAN FRANCISCO DE QUITO USFQ.
- Zornoza Martínez, A. (2020). Predicción del precio de una vivienda mediante un proceso de ciencia de datos. ESCUELA SUPERIOR DE INGENIERÍA INFORMÃTIC

Recibido:
Aceptado: