Universidad de los Andes
Facultad de Economía
Big Data y Machine Learning para Economia Aplicada
PhD. Ignacio Sarmiento-Barbieri

Merit Tejeda: 202210104
Celin Hernández: 202210067

Repositorio Url: https://github.com/chernan77/Taller_2

Problem Set 2: Making Money with ML?

"It's all about location location location!!!"

I. Introducción

El objetivo principal de este trabajo es realizar predicciones de los precios de venta de propiedades (casas y apartamentos) de la localidad de Chapinero; de acuerdo a la Cámara de Comercio de Bogotá, Chapinero representa el 5% del área total de Bogotá y es una zona donde predominan estratos altos, el 45% de los predios corresponden a estrato 6, un 30.8% a estrato 4 y un 11.7% se ubica en estrato 5; por tanto, es la localidad de Bogotá con el mayor índice de Condiciones de Vida. [1] En este contexto, los precios de las propiedades de Chapinero y según el Observatorio Técnico Catastral de Bogotá (2020) se ubican en los más alto de Bogotá, solo superados por los precios en la localidad de Usaquen.

Para predecir los precios de casas y apartamentos en Chapinero se utilizará el planteamiento de Rosen (1974), en el cual existe una matriz de características, que incluye variables relacionadas con atributos estructurales (número de habitaciones, baños, área construida, etc.), servicios públicos del vecindario (por ejemplo, calidad de la escuela local) y servicios locales (por ejemplo, delincuencia, calidad del aire, etc.). Adicionalmente, es importante vincular dentro del modelo hedónico el análisis de datos espaciales (Delago, et al, 2021) para tener una visión más amplia del comportamiento de los precios en el mercado. La idea de incluir el componente espacial se refiere a la relación que tiene la fijación de precios de propiedades en función de las características de la región donde se ellas se ubican, tal el caso de los accesos o cercanías a centros educativos, transporte público y/o centros comerciales, factores que guardan una relación positiva, al menos intuitivamente, con el precio del inmueble. Esta perspectiva se ha empleado por diversos autores con el fin de cuantificar la calidad del vecindario, tal el caso de la relación de los efectos del vecindario con los precios de la vivienda (Dubin, 1992).

Este trabajo busca generar predicciones de los precios de casas y apartamentos de Chapinero, a partir de la relación del precio con datos estructurales y características espaciales de las otras localidades de Bogotá. Esto plantea un desafío importante en términos del modelamiento econométrico, dada la heterogeneidad de las localidades de Bogotá y como al combinar los datos estructurales y características espaciales de estas zonas se pueden predecir los precios de Chapinero; para tal propósito se construirán una serie de modelos econométricos que incluyen Linear Regression, Ridge, Lasso, Elastic Net, CART, Random Forest y Boosting models. El conjunto de datos a utilizar proviene de https://www.properati.com.co que contienen una muestra de datos de propiedades para Bogotá, Colombia.

[1] Cámara de Comercio de Bogotá, Perfil Económico y Empresarial, localidad de Chapinero.

II. Datos

i. Proceso de Adquisición de los datos

Los datos utilizados para este análisis comprenden información de propiedades (casas y apartamentos) de la ciudad de Bogotá, Colombia, en el período de abril 2019 a agosto 2021. A los cuales se pudo acceder descargando las bases de datos en formato .csv de https://www.kaggle.com/competitions/uniandes-bdml-202320-ps2/data. Los datos disponibles incluyen la localización o identificación espacial de las propiedades capturada mediante las referencias de la latitud y longitud, estas variables permitieron identificar la localidad en la cual se ubican las propiedades. sin embargo, la misma presentaba una serie de falencias, como irregularidades en los caracteres de los datos numéricos y de texto, valores atípicos y valores perdidos en diferentes variables y otros problemas que se abordaron según los casos en particular.

ii. Proceso de depuración y limpieza de los datos

La base de datos de entrenamiento contiene 38,644 observaciones y 16 variables; en tanto, los datos de prueba incluyen 10,286 observaciones para igual número de variables, se realizó el siguiente proceso de depuración, limpieza e imputación para los valores atípicos, el cual se describe a continuación:

- Identificación de la cantidad de valores perdidos en cada una de las variables, el precio de las propiedades, la variable de interés a explicar y pronosticar se constató que no presenta valores perdidos.
- En la determinación del precio de venta es el número de habitaciones es relevante, esta variable no posee valores perdidos, pero si observaciones iguales a cero (0), lo cual carece sentido, ya que no puede existir una casa o apartamento sin habitaciones; estos registros fueron sustituidos con la mediana de esta variable.
- Otra variable relevante, es el número de baños, la cantidad de valores perdidos equivalen al 26.1% del total de observaciones, al revisar las características de esta variable se encontró que el valor modal es 2 y que representa el 30% de los datos, este valor se imputó para los valores perdidos.
- El Área en Metros Cuadrados (M^2) en la base de datos original aparece representada por "suface total" y "surface_covered", ambas variables muestran una cantidad representativa de valores perdidos. La depuración e imputación de esta variable consistió en:
  - Combinar los valores registrados de "suface total" y "surface_covered, de lo cual se obtuvo una variable con cobertura del 30% del total de datos.
  - Extraer expresiones regulares y/o valores numéricos seguidos de patrones para identificar el área, a partir de la descripción.
  - Se construyó para casas y apartamentos la variable Metros Cuadrados por Habitación ($M^2\_H_i$):

$$M^2\_H_i = \frac{A_i}{H_i} \quad (1)$$

  - Se obtiene para las propiedades el promedio de $M^2\_H_i$, excluyendo valore atípicos, manteniendo solo las filas donde la variable $M^2\_H_i$ es menor o igual a su percentil 95, este procedimiento permite obtener un promedio de $M^2\_H_i$, que excluye los valores extremos encontrados en los datos.
  - Del punto anterior, se calcula un promedio global de $M^2\_H_i$, fijando un rango de tolerancia de $\pm$ 2 desviaciones estándar sobre el promedio de $M^2\_H_i$, esta es una regla empírica, considera que los valores encontrados dentro de una banda con dos desviaciones típicas de la media, el valor obtenido se imputó para todos los valores por encima y/o debajo de los umbrales.
  - Finalmente, la variable $M^2\_H_i$ normalizada se multiplica por el número de habitaciones para obtener la variable Área en M^2.

- Los datos de latitud y longitud presentaban una serie de caracteres erróneos, por lo cual se removieron los puntos y se corrigieron las ubicaciones de los decimales.

iii. Descripción de los datos

Las variables que explicaran el precio de casas y apartamentos se presentan a continuación:

$$P_{j,i} = f(H_i, B_i, A_i, M^2\_H_i, T_i, BQ_i, G_i, Gy_i, S_i, CH_i, PS_i, Lat_i, Lon_i, DP_i, DT_i, DS_i, DC_i, DU_i, DR_i, DB_i, E_i, Año_i) \quad (2)$$

Donde:

$\mathbf{P_{j,i}}$: Es el precio de la i propiedad, con j = Casas y Apartamentos

1) Datos Estructurales:

1. $\mathbf{H_i}$: Es el numero de habitaciones de la i propiedad
2. $\mathbf{B_i}$: Es el numero de baños de la i propiedad
3. $\mathbf{A_i}$: Es el Area en Metros Cuadrados (M^2) de la i propiedad
4. $\mathbf{M^2\_H_i}$: Metros Cuadrados (M^2) por habitación de la i propiedad

Dentro de las variables que describen datos estructurales de las propiedades y, a partir de la descripción de las propiedades se construyeron las siguientes variables binarias:

5. $\mathbf{T_i}$: Variable binaria, mide si la i propiedad tiene Terraza ($1 =$ si y $0 =$ no)
6. $\mathbf{BQ_i}$: Variable binaria, mide si la i propiedad tiene sala BBQ ($1 =$ si y $0 =$ no)
7. $\mathbf{G_i}$: Variable binaria, mide si la i propiedad tiene Garaje ($1 =$ si y $0 =$ no)
8. $\mathbf{Gy_i}$: Variable binaria, mide si la i propiedad tiene Gimnasio ($1 =$ si y $0 =$ no)
9. $\mathbf{S_i}$: Variable binaria, mide si la i propiedad tiene Seguridad Privada ($1 =$ si y $0 =$ no)
10. $\mathbf{CH_i}$: Variable binaria, mide si la i propiedad tiene Chiminea ($1 =$ si y $0 =$ no)
11. $\mathbf{PS_i}$: Variable binaria, mide si la i propiedad tiene Piscina ($1 =$ si y $0 =$ no)

2) Características Espaciales

12. $\mathbf{Lat_i}$: Es la latitud de la i propiedad
13. $\mathbf{Lon_i}$: Es la longitud de la i propiedad
14. $\mathbf{DP_i}$: Distancia a Parques en Metros de la i propiedad
15. $\mathbf{DT_i}$: Distancia al Transporte Público en Metros de la i propiedad
16. $\mathbf{DS_i}$: Distancia a Supermercados y otros establecimientos en Metros de la i propiedad
17. $\mathbf{DC_i}$: Distancia a Centros Comerciales en Metros de la i propiedad
18. $\mathbf{DU_i}$: Distancia a Centros Educativos en Metros de la i propiedad
19. $\mathbf{DR_i}$: Distancia a Restaurantes en Metros de la i propiedad
20. $\mathbf{DB_i}$: Distancia a Bancos en Metros de la i propiedad

Las variables de distancias se obtuvieron considerando los límites geográficos de las ubicaciones de propiedades en Bogotá, de acuerdo a sus registros de latitud y longitud. Primero, se utiliza la función getbb() para obtener los límites geográficos de Bogotá, luego, con la función opq() se crea una consulta de Overpass API que busca los tipos lugares, como restaurantes, centros comerciales y educativos, etc. dentro de los límites geográficos establecidos. Adicionalmente, a partir de los datos

de latitud y longitud se obtuvo la localidad correspondiente a cada propiedad.

21. $\mathbf{E_{i,t}}$: Es el Estrato Socioeconómico de la i propiedad (1: Bajo-bajo; 2: Bajo; 3: Medio-bajo; 4: Medio; 5: Medio-alto y 6: Alto)

Finalmente, y de acuerdo a datos de la caracterización socioeconómica de la Alcaldía Mayor de Bogotá, se identificó el estrato promedio de cada localidad.

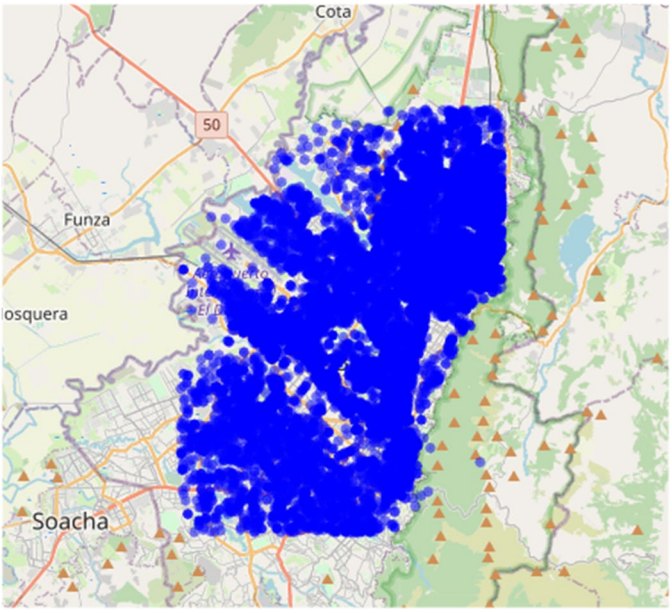
22. $\mathbf{Año_i}$: año de registro de la i propiedad.

Análisis descriptivo de los datos

Distribución espacial de las propiedades

La localización de la propiedad es una característica espacial importante ya que permite conocer la ubicación exacta de las casas y apartamentos ofrecidos en Bogotá, Colombia; esto facilita el análisis de los precios por localidades. Adicionalmente, la información espacial permite relacionar la ubicación de la casa y/o apartamentos con otros sitios donde se desarrollan actividades económicas, culturales o de esparcimiento que tiene influencia en el precio de mercado de los inmuebles.

Gráfica 1: Localización de Casas y Apartamentos en Bogotá



En la gráfica 3 se observa el mapa de Bogotá, en el cual se puede apreciar la distribución de apartamentos y casa en venta, en el cual se pueden

Tabla 1: Apartamentos: Estadísticas de variables seleccionadas

Estadísticas	N	Mean	St. Dev.	Min	Max
Precio	28,093	620,780,098.00	294,379,968.00	300,000,000	1,650,000,000
Precio_M2	28,093	4,945,436.00	2,263,592.00	542,209	44,852,378
Habitaciones	28,093	3	0.78	1	11
Baños	28,093	2	0.79	1	11
Área	28,093	137.224	58.757	15.161	571.735
M2_por_Habitación	28,093	53.16	18.535	14	95

Dist_Parques	28,093	162.436	102.752	0.991	3,344.62
Dist_Transp_Público	28,093	8,467.65	3,143.53	26.193	15,649.56
Dist_Establecimientos	28,093	174.332	117.549	0.99	1,539.72
Dist_C_Comerc	28,093	541.615	329.281	1.216	3,803.65
Dist_Centros_Educ	28,093	268.243	165.331	0.579	1,357.69
Dist_Restaurantes	28,093	219.491	150.244	1.171	1,506.62
Dist_Bancos	28,093	618.33	443.044	3.498	3,999.40
Estrato	28,093	4	0.44	2	4

La tabla 1 presenta un conjunto de estadísticos de las variables cuantitativas seleccionadas y construidas de la base de datos de propiedades, de acuerdo con esta información, el precio promedio de los apartamentos es de COP $620,780,098.00, al obtener el coeficiente de variación:

$$C.V = \frac{\text{St. Dev.}}{\text{Mean}} \rightarrow C.V = \frac{294{,}379{,}968.00}{620{,}780{,}098.00} \rightarrow C.V = 0.4742 \quad (4)$$

Un C.V alto indica una mayor heterogeneidad de los precios; en tanto que menor C.V, implicaría homogeneidad en esta la variable. Dado que este coeficiente es 0.4742, lo cual se considera alto, ya que denota una fuerte variabilidad en los precios, en este sentido, el promedio no es representativo del conjunto de datos y existe una importante heterogeneidad en esta variable.

El promedio de habitaciones en estos apartamentos es de 3 y el de baños 2, el máximo de ambas variables es 11, en el caso de los baños se observa una mayor dispersión, respecto a las habitaciones. En cuanto a la variable área, la cual esta medida en M^2, el promedio de esta variable es 137.22, al comparar con datos publicados por el Departamento Administrativo Nacional de Estadísticas (DANE) de las licencias de construcción para el destino vivienda, para el período 2005-2023 el promedio de área construida fue 131.55; por tanto, los datos utilizados en este análisis no están alejados de la información oficial. Por otra parte, se puede observar que el estrato predominante es 4, siendo el estrato mínimo 2.

Tabla 2: Casas: Estadísticas de variables seleccionadas

Estadísticas	N	Mean	St. Dev.	Min	Max
Precio	9,071	764,700,216.00	336,092,666.00	300,000,000	1,650,000,000
Precio_M2	9,071	4,192,018.00	3,582,735.00	541,015	63,093,779
Habitaciones	9,071	5	2.04	1	11
Baños	9,071	3	1.35	1	13
Área	9,071	214.589	84.678	12.68	874.283
M2_por_Habitación	9,071	49.5	16.69	6.34	79.48
Dist_Parques	9,071	156.181	96.435	1.345	941.08
Dist_Transp_Público	9,071	6,828.36	3,524.77	80.127	15,763.01
Dist_Establecimientos	9,071	168.91	120.193	0.669	914.96
Dist_C_Comerc	9,071	677.756	416.734	3.167	2,312.61
Dist_Centros_Educ	9,071	229.866	162.413	1.761	1,243.90
Dist_Restaurantes	9,071	209.036	162.7	1.625	1,220.53
Dist_Bancos	9,071	796.975	586.254	5.002	4,216.41
Estrato	9,071	4	0.48	2	4

En la tabla 2 se presentan un conjunto de estadísticas descriptivas de las casas, el precio promedio de estas propiedades es superior al de los apartamentos y nuevamente, se observa una importante dispersión en esta variable, la significativa brecha entre el precio mínimo y el máximo indica una

amplia diversidad de tipos de casas disponibles para la venta. Como es de esperarse, en promedio las casas muestran una mayor extensión en M^2 y poseen más habitaciones y baños que los apartamentos.

El precio por M^2 de las casas es inferior al de los apartamentos, de acuerdo con Analítica de Habi (2022), el valor promedio del metro cuadrado en las principales ciudades de Colombia es COP $4,409,361.00; es decir que si obtenemos el precio promedio por M^2 para casas y apartamentos (COP $4,568,727.00), encontramos resultados muy similares. Por otra parte, similar que, en el caso de los apartamentos, continúa prevaleciendo el estrato 4 como punto de ubicación prevaleciente de las propiedades. Al comparar las distancias a lugares seleccionados, entre propiedades estos varían según el lugar indicado, en términos promedio, las casas tienen una mayor proximidad al transporte público, pero guardan una mayor distancia de centros comerciales y bancos; por tanto, se podrían esperar efectos heterogéneos de cada una de las distancias en función del tipo de propiedad.

El precio por M^2 de casas y apartamentos revela la existencia de inmuebles con diferentes extensiones, pero con un valor de mercado influenciado por otras características, que dan heterogeneidad al valor del M^2. Al analizar los precios por M^2 se observa que las casas con precios más elevados se encuentran en la localidad de Usaquen; en cambio, la localidad Bosa muestra los precios más bajos, en el caso de Usaquen de las zonas más referentes del estrato 4, incluye el 17% de todas las casas de la muestra de datos. En el caso de los apartamentos, el precio más alto nuevamente se ubica en San Cristóbal y corresponde a un apartamento altamente lujoso, con características especiales que hace que su precio en el mercado este por encima del promedio de las demás localidades. La mayor cantidad de apartamentos se ubican en la localidad de Usaquen; sin embargo, el precio por M2 se ubica en el promedio. Por su parte, los precios más bajos de apartamento se localizan en Suba, la cual es una zona de estrato 3, que concentra un 37% de los apartamentos de este estrato.

III. MODELOS Y RESULTADOS.

Para estimar los precios de apartamentos y viviendas en la localidad de Chapinero, se emplearon diversos enfoques de modelado, incluyendo regresiones lineales, Lasso, Elastic Net, Boosting, Random Forest y Gradient Descent. Estos modelos se basaron en un conjunto de variables que abarcan tanto las características intrínsecas de las propiedades como aspectos geoespaciales del resto de localidades de Bogotá. Dado que se asume que existen diferencias significativas en los precios de apartamentos y casas la mayoría de las estimaciones se dividieron entre casas y apartamentos y luego se agregaron en función del codificador de propiedad ($property\_id$). De todos los modelos utilizados el que tuvo un mejor rendimiento fuera de muestra fue el de Ramdom Forest. Para cuya estimación se utilizó la biblioteca Random Forest de Rstudio que utiliza técnica de aprendizaje automático robusta y versátil que crea una colección, o "bosque," de árboles de decisión. Cada árbol se entrena de manera independiente en un subconjunto aleatorio de los datos de entrenamiento, utilizando también un subconjunto aleatorio de las características, es decir, las variables predictoras. Este proceso introduce una dosis de aleatoriedad en la construcción de los árboles. Uno de los beneficios fundamentales de Random Forest es su capacidad para mitigar el sobreajuste, un problema común en modelos basados en árboles de decisión. Al entrenar los árboles en subconjuntos aleatorios de datos y características, el modelo evita el aprendizaje de ruido presente en los datos, lo que le permite generalizar mejor a nuevos datos no vistos.

Para emplear esta metodología y realizar estimaciones fuera de muestra, se llevaron a cabo varios procedimientos de prueba:

1. División en Conjuntos de Entrenamiento y Prueba: Los datos se dividieron en dos conjuntos, con el 70% de los datos destinados al entrenamiento y el 30% restante reservado para pruebas, tanto en el caso de apartamentos como en el de casas.

2. Utilización de la Localidad de Usaquén para Entrenamiento: Se seleccionó la localidad de Usaquén como conjunto de entrenamiento, ya que compartía similitudes en características con Chapinero. Posteriormente, se utilizó este modelo entrenado para estimar los valores en las el resto de localidades de estrato 4, considerándolas como datos de prueba.
3. Filtrado por Estrato 4 y División en Conjuntos: Se aplicó un filtro a la base de datos, segmentando los datos en estrato 4. Nuevamente, se dividió en conjuntos de entrenamiento y prueba, con el 70% de los datos utilizados para entrenamiento y el 30% restante para las pruebas, tanto para apartamentos como para casas.
4. Estimación en Datos Completos: En este enfoque, no se realizó una división entre apartamentos y casas, sino que se consideraron todos los datos en conjunto. El procedimiento fue análogo, con el 70% de los datos destinados al entrenamiento y el 30% restante reservado para las pruebas.

Estos métodos de evaluación y segmentación se llevaron a cabo para evaluar y comparar la capacidad predictiva de los modelos en diferentes escenarios y asegurar su robustez en la estimación de los precios de viviendas en las distintas localidades y estratos.

Para desarrollar este modelo, se seleccionó un conjunto de variables predictoras que creemos que son relevantes para predecir los precios de viviendas y apartamentos. Estas variables incluyen datos como el estrato, la latitud, la longitud, el número de habitaciones y habitaciones al cuadrado, los metros cuadrados por habitación, y diversas características relacionadas con las propiedades, como la presencia de terraza, garaje, sala BBQ, gimnasio, chimenea y seguridad. También consideramos distancias a estaciones de transporte público, establecimientos, centros comerciales, centros educativos, restaurantes y bancos.

La elección de los hiperparámetros del modelo Random Forest, se basó en consideraciones de rendimiento y capacidad de generalización. Se optó por un valor específico de 500 árboles para el hiperparámetro ntree, como un compromiso que equilibra la precisión del modelo con el costo computacional. Además, el valor de mtry se calculó utilizando la raíz cuadrada del número de variables menos uno, una regla comúnmente empleada para determinar el número de variables a considerar en cada división de un árbol.

Los resultados de las predicciones fuera de muestra, medidos a través del Error Absoluto Medio (MAE), utilizando los conjuntos de entrenamiento y prueba previamente mencionados (con una división del 70% y 30%, respectivamente), indican que el mejor modelo es Random Forest implementado con la biblioteca correspondiente, en lugar de Ranger, específicamente para el estrato 4. Esto sugiere que el modelo estimado con esta metodologia es el que mejor refleja las características físicas y geoespaciales de la localidad de Chapinero. Aunque en nuestro ejercicio de estimación encontramos modelos con MAEs más bajos, en la competición, Random Forest demostró ser el que presentó el menor error de pronóstico.

Tabla 3: Errores fuera de muestra con diversas metodologías

Muestra Completa (Apartamentos y Casas)		**Estrato 4**		**Modelo Total Sin Divisón**	
Modelo	MAE	Modelo	MAE	Modelo	MAE
Ramdom Forest	89,890,078.00	Ramdom Forest	95,746,447.00	Ramdom Forest	97,634,532.00
Ramdom Forest 2	83,499,760.00	Ramdom Forest 2	95,169,783.00	Ramdom Forest 2	99,054,081.00
Gradient Descent	92,860,928.00	Gradient Descent	92,810,413.00	Gradient Descent	92,860,928.00
OLS	168,406,401.00	OLS	165,841,831.00	Muestra Usaquen	
Ridge	171,333,841.00	Ridge	167,582,472.00	Modelo	MAE
Lasso	168,462,170.00	Lasso	169,891,322.00	Ramdom Forest	328,727,992.00
Elastic Net	168,348,307.00	Elastic Net	168,242,239.00	Ramdom Forest 2	133,069,629.00
Boosting	147,172,665.00	Boosting	143,321,238.00	Gradient Descent	195,875,250.00

Asimismo, se llevó a cabo un análisis de la importancia de las variables utilizando el parámetro "importance". Al evaluar la relevancia de las variables en la predicción en diversos conjuntos de datos, se destaca que el área y la ubicación geoespacial, representadas por la longitud y latitud, emergen como factores críticos en la valoración de las propiedades. Este hallazgo sugiere que tanto la ubicación como el espacio físico desempeñan un papel esencial en la determinación de precios en el mercado inmobiliario de Chapinero (Vea Apéndices 9-11). Además, se observa que la cantidad de baños y la proximidad a servicios esenciales, como el transporte público, bancos y restaurantes, ejercen una influencia significativa en la fijación de precios. Estos resultados proporcionan una valiosa percepción de los factores más influyentes que dan forma al mercado inmobiliario en la localidad de Chapinero.

IV. Conclusiones y recomendaciones

- Se utilizaron diversos enfoques de modelado, incluyendo regresiones lineales, Lasso, Elastic Net, Boosting, Random Forest y Gradient Descent, para estimar los precios de apartamentos y viviendas en la localidad de Chapinero. Estos modelos se basaron en un conjunto de variables que abarcan las características estructurales de las propiedades y datos espaciales de otras localidades de Bogotá.
- En general, los modelos que mejor predijeron los precios de las propiedades fueron aquellos basados en el estrato 4 y que utilizaron la muestra completa sin filtrar por estrato específico. Se identificó que variables como el área, la ubicación geográfica, el número de baños, la distancia al transporte público y la cercanía a establecimientos comerciales tenían un impacto significativo en la predicción de precios. Además, se observó que estos modelos obtuvieron un mejor rendimiento en términos de predicción de precios.
- El modelo Random Forest basado en la clasificación por estrato 4 fue el que mostró el mejor resultado en la predicción de precios. Sin embargo, no es el modelo más eficiente en términos computacionales, ya que la metodología de Gradient Descent ofreció un resultado y más eficiente en uso de recursos computacionales.
- En general, estos enfoques de modelado permitieron obtener estimaciones precisas de los precios de las propiedades en la localidad de Chapinero, lo que es fundamental en el mercado inmobiliario.
- Se recomienda dedicar un mayor espacio de tiempo a la depuración y revisión de los datos básicos, así como el acceso a otras fuentes de información más amplias y precisas para mejorar los resultados.

V. Bibliografía

1. Delagado, J,. Martinez, O., Romer, J. (2021). Determinantes del precio de la vivienda nueva en Bogotá para el año 2019: una aproximación a través de un modelo semiparamétrico de regresión espacial. ISSN:1794-9165., ing. cienc., vol. 17, no. 34, pp.23–52, julio-diciembre. 2021. https://publicaciones.eafit.edu.co/index.php/ingciencia/article/view/6772/5323

2. Departamento Administrativo Nacional de Estadísticas (DANE). Agosto (2023). Estadísticas de Licencias de Construcción (ELIC). https://www.dane.gov.co/index.php/estadisticas-por-tema/construccion/licencias-de-construccion

3. Dubin, R. (1992). Spatial autocorrelation and neighborhood quality. Regional Science and Urban Economics, Volume 22, Issue 3, Pages 433-452, ISSN 0166-0462, https://doi.org/10.1016/0166-0462(92)90038-3.

4. Quevedo, A. (2022). Analítica de Habi. ¿Cuánto es el valor metro cuadrado en Colombia? https://habi.co/blog/donde-es-mas-barata-la-vivienda-en-colombia

5. Observatorio Técnico Catastral., Alcaldía Mayor de Bogotá (2020). Dinámica inmobiliaria Bogotá región - 2017 – 2020. Un Estudio de la oferta de vivienda usada en Bogotá y la región de Cundinamarca en el periodo 2017 – 2020. https://www.catastrobogota.gov.co/sites/default/files/Dinamica%20Inmobiliaria%20Bogota%20Region%202017%20-%202020_20210331_VersionDiseno_v1_20210520.pdf

6. Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. Journal of Political Economy., Volume 82, Number 1. https://www.journals.uchicago.edu/doi/pdf/10.1086/260169

VI. Apendice: Otras Metodologías

Modelo de Regresión Lineal

Se utilizaron 19 recetas las cuales incluyen variables como el estrato, ubicación geográfica, el número de habitaciones, el número de baños, el área de las casas, una variable temporal (año), la distancia a parques, transporte público, establecimientos comerciales, centros educativos, restaurantes y bancos. Además, se utilizaron transformaciones de algunas variables, como el cuadrado de la cantidad de habitaciones (Habitaciones2) y la relación entre el área y el número de habitaciones (M2_por_Habitacion).

De los cuatro conjuntos de datos resultantes, los modelos que mejor predijeron fueron aquellos que se basaron en el estrato 4 y utilizaron la muestra total (dividiéndose entre apartamentos y casas), sin filtrar por estrato específico. Además, se dividió el conjunto de datos en un conjunto de entrenamiento (70% de los datos) y un conjunto de prueba (30% de los datos) para evaluar el rendimiento de los modelos.

El modelo obtenido en ambos casos tiene en cuenta diversas variables predictoras, incluyendo:

$$\begin{aligned} lP_{j,i} = {} & E_i + H_i + H_i^2 + M^2\_H_i + T_i + G_i + BQ_i + BQ_{iT_i} + Gy_i + CH_i + S_i + DP_i + DT_i + DS_i \\ & + DC_i + DU_i + DR_i + DB_i + Lat_i + Lon_i + Año_i \end{aligned}$$

Apéndice 1: Modelo de Regresión Lineal Base Completa y Estrato 4

Apartamentos

Variable Dependiente : Log(Precio)		
	(1) Base Completa	(2) Estrato 4
Estrato	-0.104***	
	-0.005	
Habitaciones	0.326***	0.320***
	-0.008	-0.008
Habitaciones^2	-0.028***	-0.028***
	-0.001	-0.001
Baños	0.183***	0.175***
	-0.003	-0.003
Latitud	0.003***	0.003***
	-0.0001	-0.0001
Longitud	-0.867***	-0.208
	-0.184	-0.196
M2_por_Habitación	7.536***	7.273***
	-0.131	-0.146
Terraza	0.003	0.103***
	-0.003	-0.005
Garaje	0.116***	-0.043***
	-0.004	-0.005
Sala_BBQ	-0.040***	-0.021***
	-0.004	-0.006
Gimnasio	-0.021***	0.094***
	-0.005	-0.005
Chimenea	0.082***	0.069***
	-0.004	-0.005
Seguridad	0.068***	-0.055***
	-0.004	-0.006
Dist_Parques	-0.052***	-0.001
	-0.005	-0.003
Dist_Transp_Público	0.0001***	0.00005**
	-0.00002	-0.00002
Año	-0.00005***	-0.00005***
	0.0000	0.0000
Dist_Establecimientos	0.0003***	0.0004***
	-0.00002	-0.00002
Dist_C_Comerc	-0.00001**	-0.00004***
	-0.00001	-0.00001
Dist_Centros_Educ	0.0001***	0.0001***
	-0.00001	-0.00001
Dist_Restaurantes	0.0001***	0.0001***
	-0.00002	-0.00002
Dist_Bancos	0.00000	0.0001***
	-0.00001	-0.00001
Constante	574.918***	560.304***
	-10.932	-12.201
Observaciones	28,093	20,780
R^2	0.463	0.448

Note: *p<0.1; **p<0.05; ***p<0.01

Casas

Variable Dependiente : Log(Precio)		
	(1) Base Completa	(2) Estrato 4
Estrato	0.015*	
	-0.009	
Habitaciones	0.126***	0.133***
	-0.009	-0.01
Habitaciones^2	-0.008***	-0.008***
	-0.001	-0.001
Baños	0.079***	0.070***
	-0.003	-0.003
Latitud	2.511***	0.008***
	-0.165	-0.0003
Longitud	4.673***	2.638***
	-0.163	-0.185
M2_por_Habitación	0.008***	4.221***
	-0.0003	-0.186
Terraza	-0.021**	-0.037***
	-0.009	-0.011
Garaje	-0.016**	-0.009
	-0.008	-0.009
Sala_BBQ	0.084***	0.100***
	-0.016	-0.021
Gimnasio	-0.079***	-0.057**
	-0.018	-0.023
Chimenea	0.121***	0.104***
	-0.01	-0.012
Seguridad	-0.041***	-0.058***
	-0.014	-0.018
Año	0.044***	0.051***
	-0.005	-0.005
Dist_Parques	0.00004	-0.0002***
	-0.00004	-0.00005
Dist_Transp_Público	-0.0001***	-0.0001***
	0.0000	0.0000
Dist_Establecimientos	0.0004***	0.0004***
	-0.00004	-0.00005
Dist_C_Comerc	0.00000	-0.00002
	-0.00001	-0.00001
Dist_Centros_Educ	0.0001***	0.0001***
	-0.00002	-0.00003
Dist_Restaurantes	-0.0001***	-0.0002***
	-0.00003	-0.00004
Dist_Bancos	-0.0001***	-0.0001***
	-0.00001	-0.00001
Constant	265.142***	217.793***
	-15.882	-18.129
Observaciones	9071	6423
R^2	0.405	0.363

Note: *p<0.1; **p<0.05; ***p<0.01

En la Tabla 4, se presentan los resultados para la base de datos completa y el estrato 4, en ambos casos segmentando entre casas y apartamentos. En el contexto de apartamentos, se destaca que el modelo que incorpora toda la muestra tiene una mejor bondad de ajuste, con un R^2 de 0.463, lo que indica que el conjunto de variables utilizadas explica el 46.3% de la variabilidad en los precios de los apartamentos.

En ambas bases de datos, la variable "M2_por_Habitación" sobresale como un predictor altamente significativo del precio de los apartamentos, con coeficientes de 7.53 y 7.23, respectivamente. Esto indica que un aumento de un metro cuadrado adicional por habitación está asociado con un aumento en el precio de los apartamentos en un rango que oscila entre el 7.23% y el 7.53%. Además, la importancia económica de esta variable se refleja en su significancia del 0.36%, lo que sugiere que un incremento de un metro cuadrado en esta relación genera una desviación de la media de los precios de los apartamentos en ese mismo valor.

Asimismo, tanto el número de baños y habitaciones como la presencia de chimenea, la distancia a transporte público y establecimientos tienen significancia estadística y ejercen un impacto positivo en los precios de los apartamentos. Sin embargo, en relación al resto de las variables, la relación con los precios de los apartamentos difiere entre los conjuntos de datos (base completa y estrato 4), lo que requiere de una evaluación más detallada para comprender completamente su influencia.

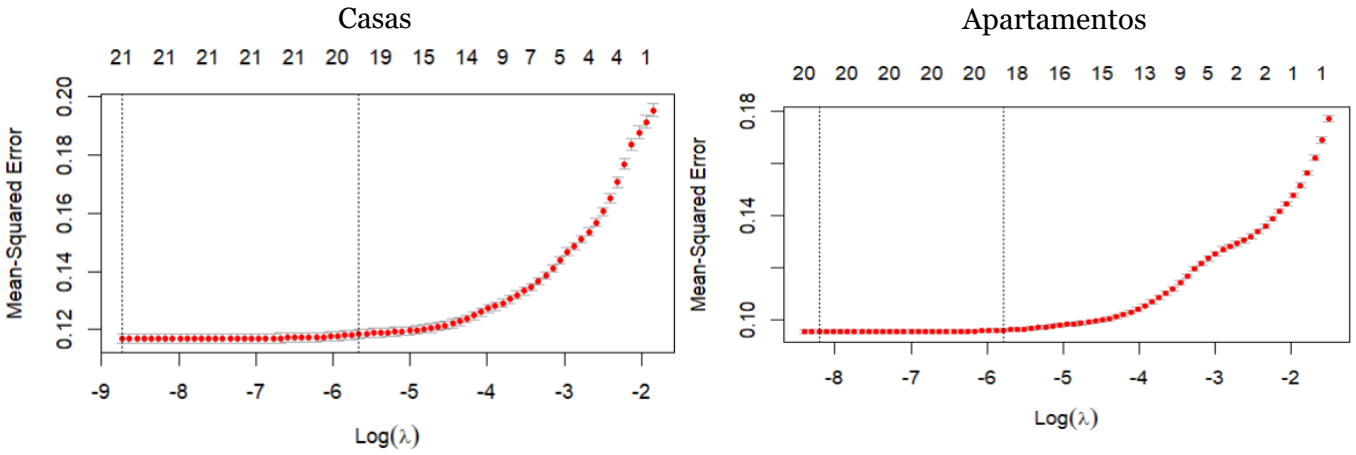
En el caso de las casas, observamos que el modelo con el mejor ajuste de datos es aquel que incluye tanto apartamentos como casas, sin realizar un filtrado por estrato 4. Este modelo presenta un coeficiente de determinación (R^2) de 0.40, lo que significa que las variables empleadas para predecir los precios explican el 40% de la variabilidad en dicha variable. Se observa que variables como el número de habitaciones, la cantidad de baños, el tamaño por habitación, la ubicación geográfica, la presencia de chimenea y el año de construcción tienen un impacto positivo en los precios de las viviendas. Esto implica que un incremento en estas variables se traduce en un aumento en el precio de las propiedades. Además, la distancia a parques, transporte público, establecimientos, centros educativos, restaurantes y bancos también ejerce una influencia positiva en los precios, aunque en menor medida.

Específicamente para el estrato 4, las variables relacionadas con el número de habitaciones, baños, el tamaño por habitación, la ubicación geográfica, la presencia de una sala de BBQ, la chimenea, el año de construcción y la distancia a establecimientos muestran una influencia positiva en los precios de las viviendas. Nuevamente, un aumento en estas variables se traduce en un incremento en los precios de las casas. Asimismo, las variables que representan la distancia a parques, transporte público, centros educativos, restaurantes y bancos también desempeñan un papel positivo en la determinación de los precios de las viviendas.

Modelos Lasso

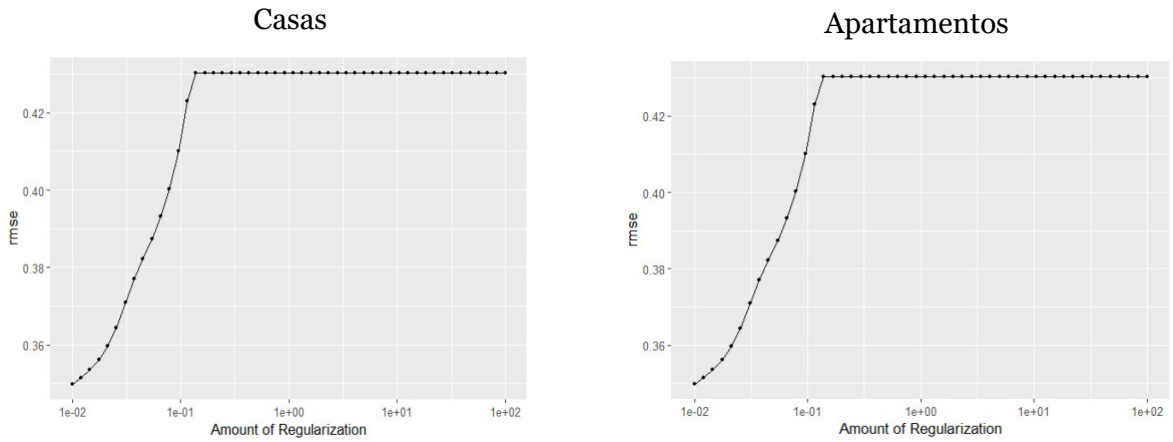
En el marco de este análisis, se han ajustado dos modelos de regresión Lasso para la predicción de precios de viviendas. El primer modelo (Modelo 1) se ha desarrollado utilizando la muestra total de apartamentos y casas donde el conjunto de entrenamiento se divide 70% y de prueba en un 30%. En este caso, se ha aplicado una validación cruzada para seleccionar el valor óptimo de lambda (0.0001615243 para casas y 0.0002726315 para apartamentos), hiperparámetro de regularización, lo que ha permitido optimizar el rendimiento del modelo.

Apendice 2: Lambda Óptimo: Base Completa



Por otro lado, el Modelo 2 se ha centrado de manera específica en el estrato 4, realizando una distinción entre apartamentos y casas, y ha dividido el conjunto de datos entrenamiento (70%) y prueba (30%). A diferencia del Modelo 1, este segundo enfoque ha involucrado un proceso de ajuste de hiperparámetros mediante la utilización de una cuadrícula de valores de penalización. Ambos modelos han sido evaluados con respecto a su capacidad para predecir los precios de las viviendas, y las predicciones resultantes se han aplicado a un conjunto de datos de prueba.

Apéndice 3: Sintonización de Hiperparámetros Modelo Lasso



Al observar el Gráfico 10, se identifica claramente el punto donde el error raíz cuadrado medio (RMSE) alcanza su valor mínimo, lo que indica el valor óptimo de lambda que maximiza la precisión de nuestras predicciones. Además, este gráfico permite evaluar la estabilidad del modelo en un rango de valores de penalización, lo que es fundamental para comprender cómo estos modelos responden a diferentes niveles de regularización.

Modelos Ridge

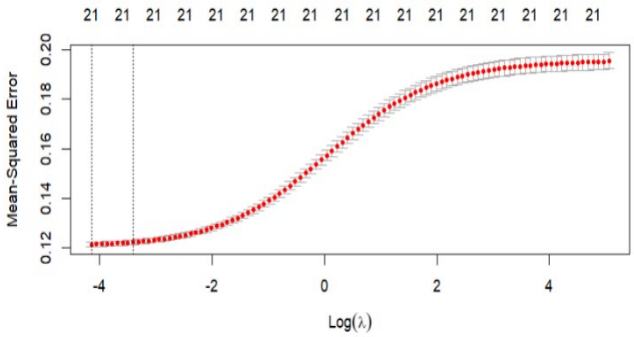
Se llevaron a cabo estimaciones de precios de viviendas utilizando el modelo de regresión Ridge en dos conjuntos de datos distintos: el conjunto de datos completo. Estos conjuntos de datos incluyeron tanto apartamentos como casas.

1. Conjunto de Datos Completo:

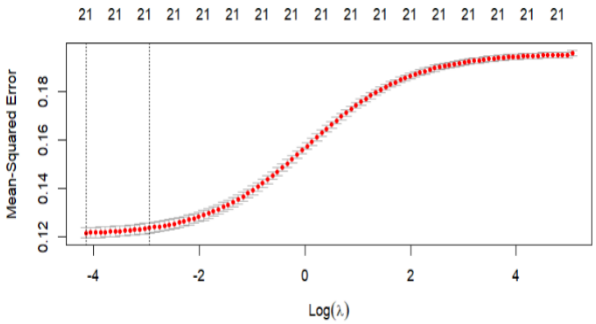
El proceso de ajuste de los modelos Ridge incluyó la búsqueda del valor óptimo de lambda, que es el hiperparámetro de penalización que controla la magnitud de los coeficientes en el modelo. Se utilizó la función cv.glmnet para realizar una validación cruzada y seleccionar el valor de lambda que minimiza el error. Una vez obtenido el valor óptimo de lambda, se ajustaron los modelos Ridge con ese valor y se realizaron predicciones para los conjuntos de entrenamiento correspondientes. El proceso de búsqueda de hiperparámetros para seleccionar el mejor valor de lambda (0.02211398 para apartamentos y 0.01578109 para casas) se representa visualmente en el siguiente Gráfico:

Apéndice 4: Lambda Óptimo Ridge: Base Completa

Apartamento



Casa

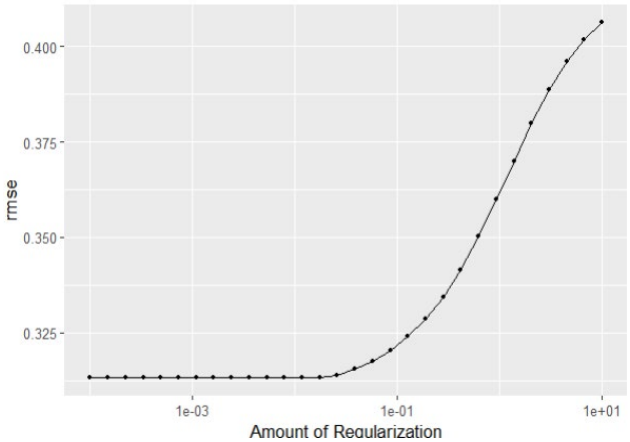


2. Estrato 4

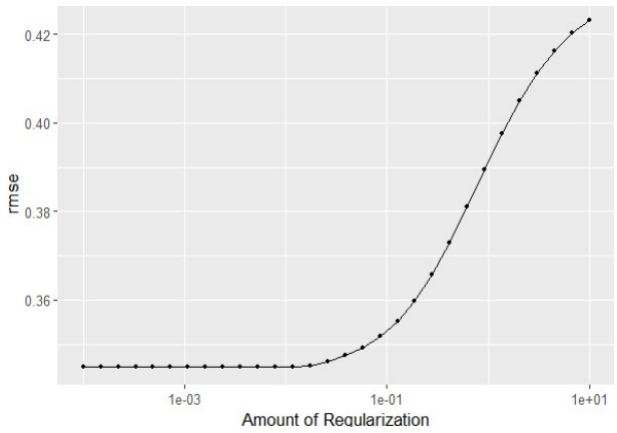
Se enfocó específicamente en el estrato 4 y se diferenció entre apartamentos y casas, dividiendo el conjunto de entrenamiento en un 70% y el de prueba en un 30%. Para este enfoque particular en el estrato 4, se llevó a cabo un proceso de ajuste de hiperparámetros mediante una cuadrícula de valores de penalización. A continuación, se presenta un gráfico que ilustra el proceso de afinación de hiperparámetros para el modelo de Regresión Ridge aplicado a los apartamentos y casas del Estrato 4. Este gráfico muestra cómo varía el error cuadrático medio (RMSE) en función de diferentes valores de penalización, lo que permite seleccionar el hiperparámetro óptimo del modelo.

Apéndice 5: Afinación de Hiperparámetros - Regresión Ridge

Apartamentos



Casas



Modelos Elastic Net

En el proceso de análisis, se llevaron a cabo modelos de regresión Elastic Net para estimar los precios de apartamentos y casas, utilizando dos conjuntos de datos diferentes para apartamentos y casas:

- Base de datos completa
- Estrato 4

En ambos casos, se construyeron modelos de regresión Elastic Net empleando un conjunto de variables predictoras. En la configuración de estos modelos, se estableció el parámetro alpha en 0.5, lo que resulta en una ponderación equilibrada entre las técnicas de regularización L1 y L2. Asimismo, se efectuó un proceso de validación cruzada con el propósito de identificar el valor óptimo de lambda, el factor de regularización. Una vez que se determinó el valor óptimo de lambda, se ajustaron los modelos Elastic Net y se realizaron las predicciones correspondientes para estimar los precios de las propiedades en cuestión.

Al analizar las predicciones dentro de la muestra, se evidencia un ajuste más preciso en el caso de la base de datos completa. Sin embargo, al evaluar los errores en las predicciones fuera de la muestra, se observa que la estimación basada únicamente en el estrato 4 presenta un ligeramente menor nivel de error.

Este hallazgo sugiere que el modelo ajustado a la base de datos completa se adapta de manera óptima a los datos utilizados durante el entrenamiento, lo que se traduce en un buen ajuste dentro de la muestra. Por otro lado, el modelo que se enfoca exclusivamente en el estrato 4 logra una mayor precisión al predecir valores fuera de la muestra, lo que indica su capacidad para generalizar de manera efectiva a datos no utilizados durante el entrenamiento.

Metodología Boosting

Se aplica la metodología de Boosting Tree para construir modelos de predicción de precios para apartamentos y casas. Estos modelos basados tienden a ofrecer un alto rendimiento en términos de precisión predictiva. Esto es especialmente útil en aplicaciones donde la precisión es fundamental, como es la predicción de precios inmobiliarios, donde pequeñas diferencias pueden tener un impacto significativo.

Los modelos se entrenan utilizando una base de datos completa y el estrato 4. Para optimizar el rendimiento del modelo, se exploran diferentes valores de hiperparámetros como el número de iteraciones de boosting, la profundidad máxima del árbol y el factor de aprendizaje. El proceso de ajuste de hiperparámetros implica experimentar con diferentes combinaciones de valores para mstop, maxdepth, y nu para determinar cuáles producen el mejor rendimiento del modelo. El valor óptimo se selecciona a través de la validación cruzada (CV), que evalúa cómo se comporta el modelo en diferentes subconjuntos de datos.

Técnicas de Ensamblaje.

Al emplear estas técnicas para realizar estimaciones fuera de muestra, se llevaron a cabo varios procedimientos de prueba:

5. División en Conjuntos de Entrenamiento y Prueba: Los datos se dividieron en dos conjuntos, con el 70% de los datos destinados al entrenamiento y el 30% restante reservado para pruebas, tanto en el caso de apartamentos como en el de casas.
6. Utilización de la Localidad de Usaquén para Entrenamiento: Se seleccionó la localidad de Usaquén como conjunto de entrenamiento, ya que compartía similitudes en características con Chapinero. Posteriormente, se utilizó este modelo entrenado para estimar los valores en las el resto de localidades de estrato 4, considerándolas como datos de prueba.

7. Filtrado por Estrato 4 y División en Conjuntos: Se aplicó un filtro a la base de datos, segmentando los datos en estrato 4. Nuevamente, se dividió en conjuntos de entrenamiento y prueba, con el 70% de los datos utilizados para entrenamiento y el 30% restante para las pruebas, tanto para apartamentos como para casas.
8. Estimación en Datos Completos: En este enfoque, no se realizó una división entre apartamentos y casas, sino que se consideraron todos los datos en conjunto. El procedimiento fue análogo, con el 70% de los datos destinados al entrenamiento y el 30% restante reservado para las pruebas.

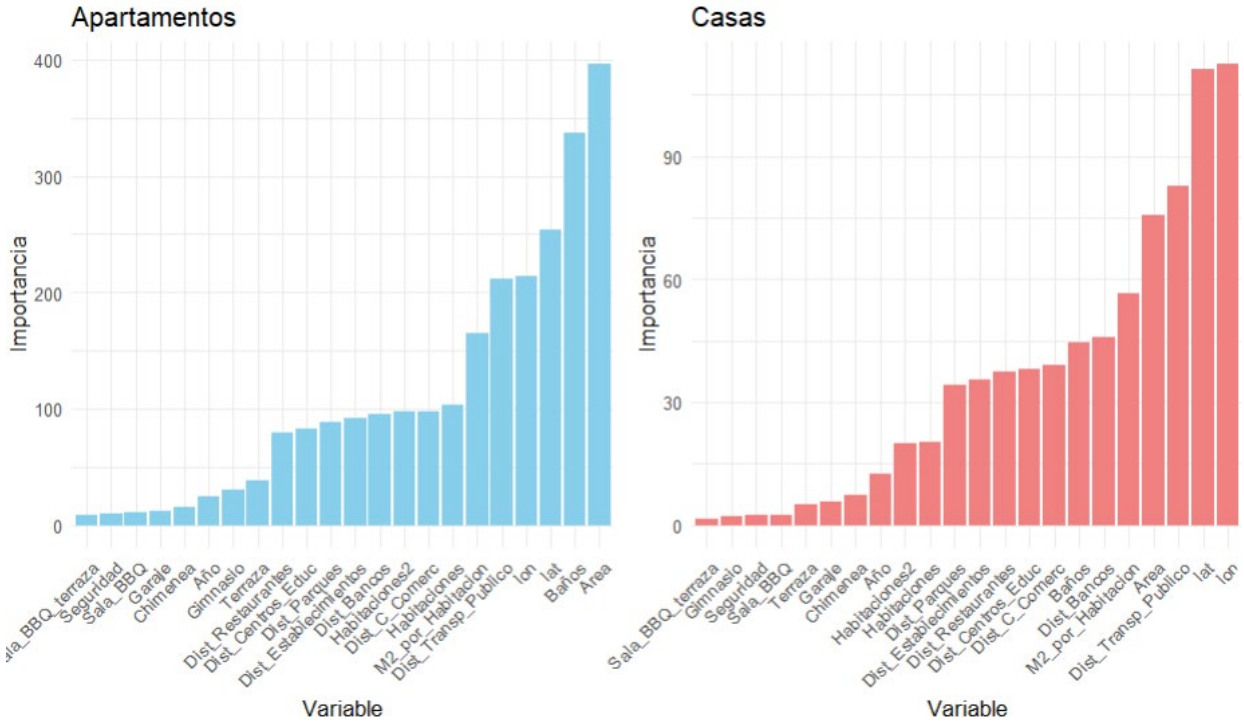
Estos métodos de evaluación y segmentación se llevaron a cabo para evaluar y comparar la capacidad predictiva de los modelos en diferentes escenarios y asegurar su robustez en la estimación de los precios de viviendas en las distintas localidades y estratos.

Random Forest (Bosque Aleatorio)
Algoritmo Ranger

Se aplicó el método "ranger" para ajustar los modelos de regresión. Estos incluyeron el número de variables aleatorias seleccionadas en cada división (mtry), la regla de división utilizada para separar nodos (por ejemplo, "variance"), y el tamaño mínimo de nodos terminales (min.node.size). La exploración de estos hiperparámetros tenía como objetivo encontrar la configuración óptima que maximizara el rendimiento del modelo en términos de precisión de las predicciones. Esto se logró mediante la validación cruzada y la búsqueda en la cuadrícula para evaluar cómo diferentes valores de estos hiperparámetros influyen en el rendimiento en conjuntos de entrenamiento y prueba.

En el caso de los apartamentos, las variables más importantes son el área de los apartamentos y su ubicación geográfica (latitud); también tienen un impacto significativo en los precios, el número de baños, la distancia al transporte público y la cercanía a establecimientos comerciales. Esto indica que la comodidad, el acceso al transporte y la ubicación geográfica son aspectos fundamentales en la fijación de precios de los apartamentos. De igual manera, para las casas, la importancia de las variables se mantiene centrada en su ubicación geográfica, distancia a los centros de transporte público, M2 por Habitación, distancia a bancos y baños.

Apéndice 6: Importancia de Variable Modelo Ranger Estrato 4

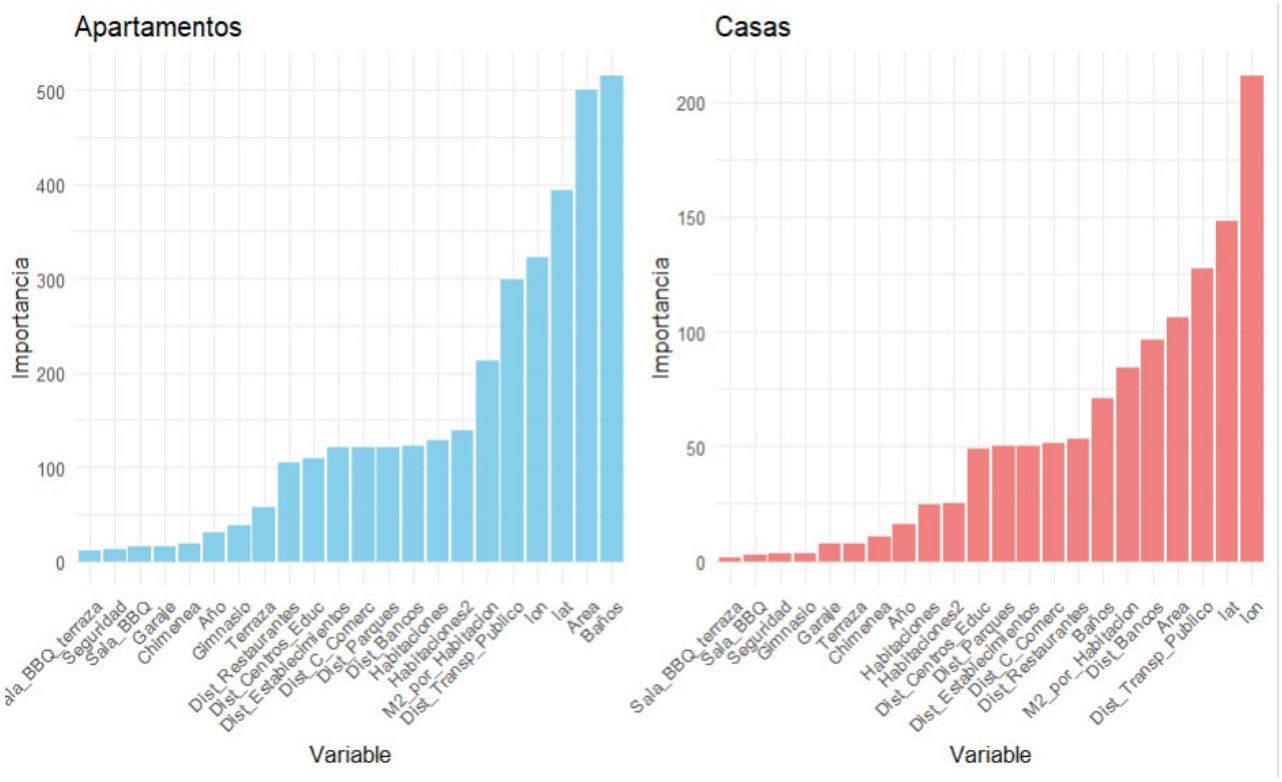


Para el Modelo Ranger 2 se divide el conjunto de datos de entrenamiento de apartamentos y casas en ambos casos utilizando en entrenamiento 0.7 de la muestra y 0.3 para la prueba. Las variables más importantes para la estimación de apartamento son el área, la latitud, los baños, y la distancia al transporte Público. En tanto que para casas las variables más importantes son la ubicación geográfica, la distancia al transporte público y la distancia a bancos. Esto sugiere que la superficie del inmueble y su ubicación geográfica tienen un fuerte impacto en el precio.

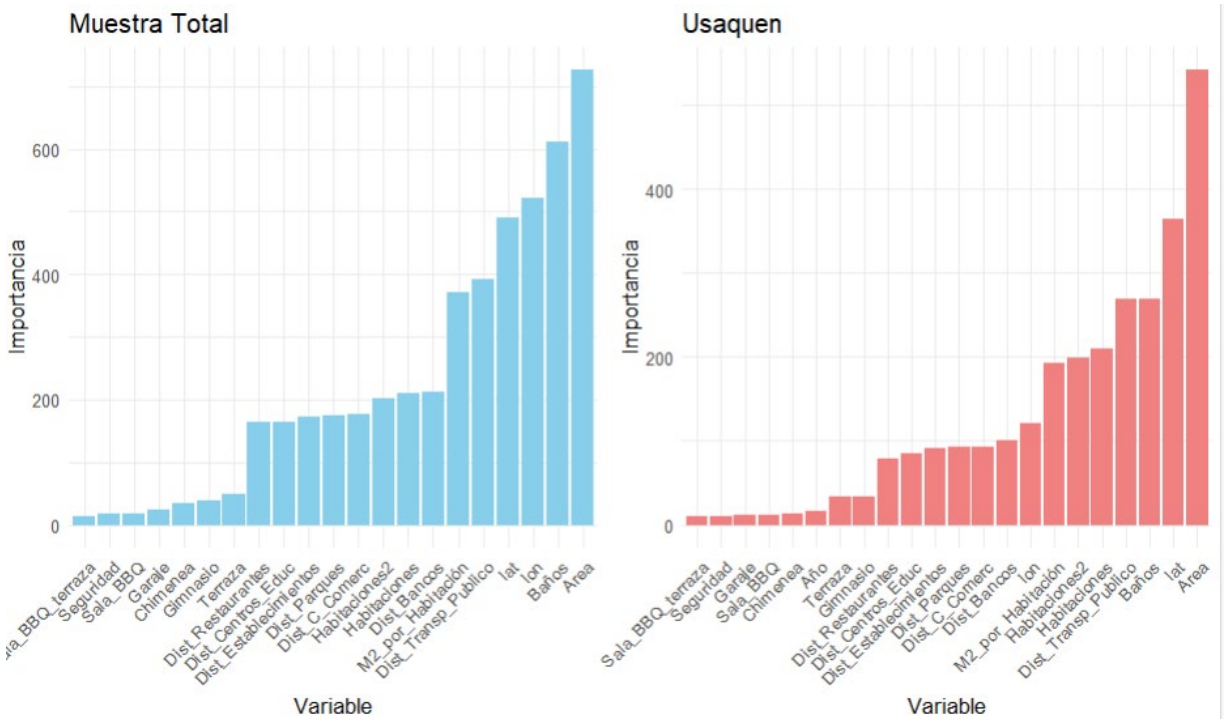
Para estimar el tercer modelo, se realizó una división de la muestra completa, sin distinción entre apartamentos y casas. Esta decisión se tomó con el propósito de evaluar si un conjunto de entrenamiento más amplio sería capaz de capturar la variabilidad presente en los datos. La muestra se dividió en un conjunto de entrenamiento, representando el 70% de los datos, y un conjunto de prueba, que comprendió el 30% restante.

En el caso del último modelo, se optó por utilizar la localidad de Usaquén como conjunto de entrenamiento, mientras que el resto de la muestra perteneciente al estrato 4 se asignó como conjunto de prueba. Esta estrategia permitió evaluar el desempeño del modelo en un contexto específico y contrastarlo con datos de una subpoblación particular, en aras de una evaluación más precisa. Al ver la importancia de las variables en estos dos últimos modelos, se observa que las variables que más explican los precios de los inmuebles son el área, la ubicación geográfica y el número de baños. Cabe señalar que se realizó el ejercicio de eliminar algunas características cuyo peso no era importante; sin embargo, el resultado de las predicciones fuera de muestra no mejoro.

Apéndice 7: Importancia de Variable Modelo Ranger 2 Estrato Muestra

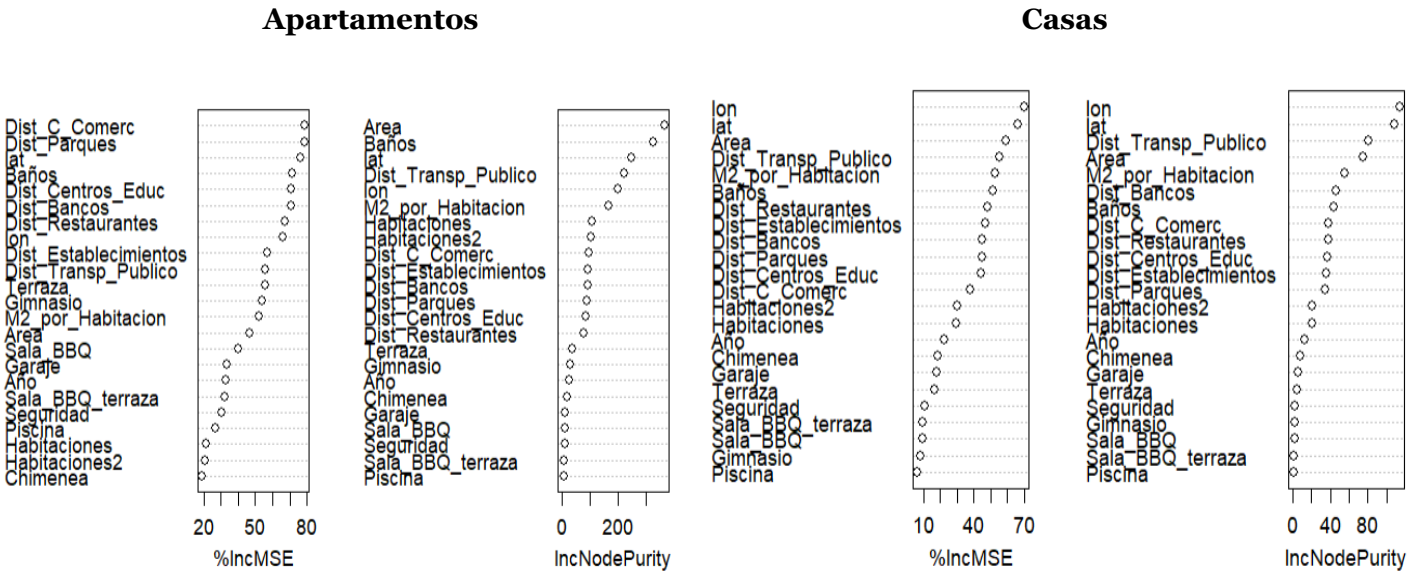


Apéndice 8: Importancia de Variable Modelo Ranger 3 (Sin División) y Ranger 4

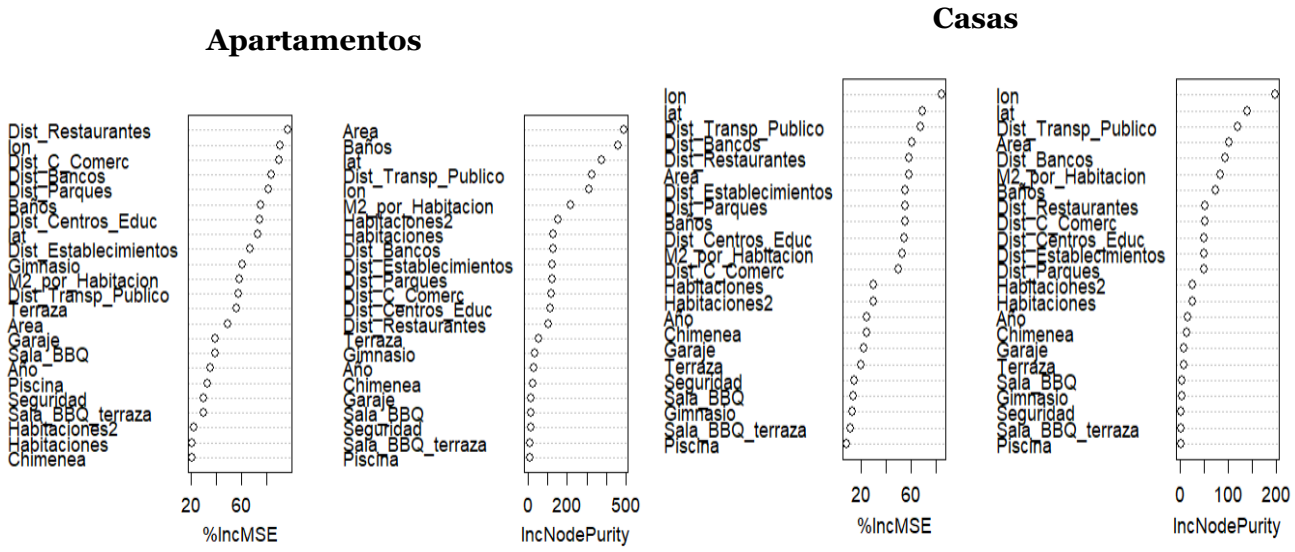


Por otra parte, al analizar la importancia de las variables en la predicción para distintos conjuntos de datos, se observa que las variables que explican en mayor medida el precio de las vivienda y apartamentos en términos generales: el área, la ubicación (longitud, latitud), el número de baños, y las distancias al transporte público, bancos y restaurantes.

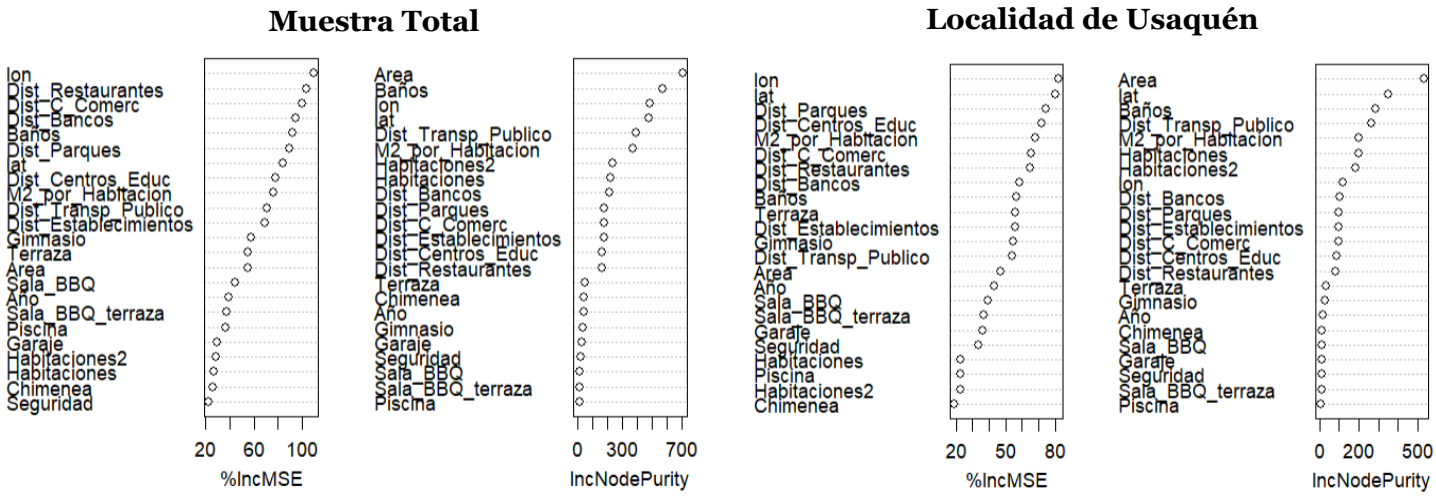
Apéndice 9: Importancia de las Variables Modelo Random Forest: Estrato 4



Apéndice 10: Importancia de las Variables Modelo Random Forest: Total Apartamentos y Casas



Apéndice 11: Importancia de las Variables Modelo Random Forest: Muestra Total y Usaquén



Gradient Boosting

En el contexto del análisis de precios de propiedades en la localidad de Chapinero, se implementó un enfoque de modelado basado en el algoritmo XGBoost que es un método de aprendizaje automático que se utiliza para tareas de regresión y clasificación. Para predecir los precios de los inmuebles de esta localidad se utilizaron los mismos criterios de división de datos de entrenamiento y prueba que se emplearon en los modelos de Random Forest. En ambos casos, se dividió el conjunto de datos en una proporción del 70% para datos de entrenamiento y del 30% para prueba. Esto permitió mantener la coherencia y la comparabilidad entre los modelos de Random Forest y los modelos de XGBoost, asegurando que las evaluaciones se realizaran en las mismas condiciones y se pudieran comparar de manera efectiva.

Para los modelos de XGBoost empleados, se llevaron a cabo múltiples pruebas con el fin de seleccionar

los hiperparámetros más adecuados, considerando la calidad de ajuste a los datos. Utilizando un número fijo de árboles (nrounds) establecido en 200, una tasa de aprendizaje (eta) de 0.1 y una profundidad máxima del árbol (max_depth) de 14, se observó un excelente rendimiento en todos los modelos, excepto en el caso de la clasificación de Usaquén. En este último, se logró un mejor rendimiento al ajustar los hiperparámetros con una profundidad de 14, nrounds de 100 y una tasa de aprendizaje eta de 0.1.

Apéndice 12: Resultados MAE Modelos Gradient Boosting Primera Prueba

Modelos	MAE	Hiperparámetros
Modelo XGBoost Estrato 4	88,202,413.00	max_depth=14
Modelo XGBoost Muestra Total Apart y Casas	92,810,443.00	nrounds=200
Modelo XGBoost Muestra Total Sin División	92,860,928.00	eta=0.1
Modelo XGBoost Usaquén	195,875,250.00	

Apéndice 13: Resultados MAE Modelos Gradient Boosting Segunda Prueba

Modelos	MAE	Hiperparámetros
Modelo XGBoost Estrato 4	93,087,622.00	max_depth=10
Modelo XGBoost Muestra Total Apart y Casas	97,713,874.00	nrounds=200
Modelo XGBoost Muestra Total Sin División	100,328,289.00	eta=0.1
Modelo XGBoost Usaquén	200,008,879.00	

Apéndice 14: Resultados MAE Modelos Gradient Boosting Tercer Prueba

Modelos	MAE	Hiperparámetros
Modelo XGBoost Estrato 4	92,539,578.00	max_depth=14
Modelo XGBoost Muestra Total Apart y Casas	97,345,970.00	nrounds=100
Modelo XGBoost Muestra Total Sin División	97,771,358.00	eta=0.1
Modelo XGBoost Usaquén	193,388,912.00	

Apéndice 15: Resultados MAE Modelos Gradient Boosting Cuarta Prueba

Modelos	MAE	Hiperparámetros
Modelo XGBoost Estrato 4	90,956,891.00	max_depth=14
Modelo XGBoost Muestra Total Apart y Casas	95,476,831.00	nrounds=200
Modelo XGBoost Muestra Total Sin División	93,657,428.00	eta=0.3
Modelo XGBoost Usaquén	223,216,250.00	