

**Big Data and Machine Learning for Applied Economics – MECA 4107 – 2022-13**  
**Maestría en Economía Aplicada – Universidad de los Andes**  
**Taller 3 – Julio 16, 2022 – Equipo 12**

**Ingrid L. Molano, 200511102 / Jorge E. García, 201310645 / Camilo Villa, 201818624**

**Repositorio: [https://github.com/camilovillam/MECA\\_BD\\_PS3](https://github.com/camilovillam/MECA_BD_PS3)**

## **1. Introducción:**

El presente taller busca identificar el mejor modelo de predicción del precio de la vivienda para la UPZ Chapinero en Bogotá y la comuna El Poblado en Medellín. Los métodos utilizados para la implementación los diferentes modelos de predicción están basados en algoritmos de aprendizaje de máquinas y estimación bajo validación cruzada. La información inicial para el entrenamiento de los modelos proviene de <https://www.properati.com.co>. Estas bases de datos presentaron desafíos importantes en términos de datos faltantes. Para esto fue necesario recuperar información a través del método de expresiones regulares e imputar datos cuando no fue posible continuar con la recuperación. Dado que la base solo contaba con variables que describían características propias de la vivienda, se consultaron fuentes externas para complementar con variables que explican el precio de la vivienda según su ubicación y calificación normativa, y según su distancia con respecto a soportes de calidad urbana de la ciudad e inmuebles dotacionales. El mejor modelo de predicción identificado es un *Random Forest* y de los modelos evaluados es el que logra el mayor número de viviendas compradas y presenta la menor diferencia la predicción del dinero gastado y el dinero gastado si la compra se realizara con el precio de lista de la base de datos.

## **2. Datos:**

Para el tratamiento de las bases de datos, inicialmente se separaron entre Bogotá y Medellín dado que son dos mercados inmobiliarios diferentes. En particular, el metro cuadrado construido en Medellín es menor al de Bogotá.

Separar la base de entrenamiento permitió identificar variables externas de cada ciudad. Para el caso de Bogotá, la información se extrajo del portal de datos abiertos (<https://datosabiertos.bogota.gov.co/>), y las variables seleccionadas fueron localidad, unidad de planeamiento zonal, estrato, el uso del suelo, el índice y puntaje de inseguridad nocturna y el valor promedio del avalúo comercial del suelo por manzana.

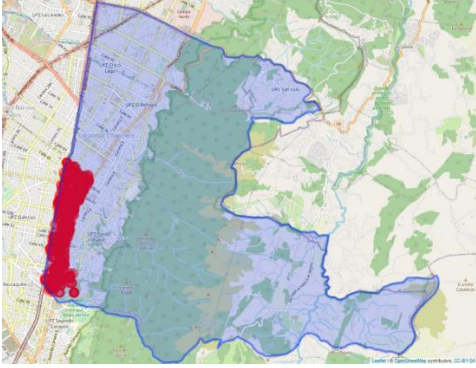
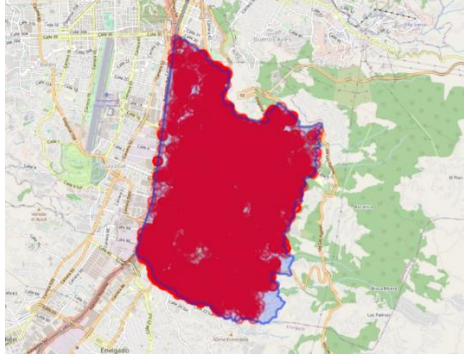
Para el caso de Medellín, la información se extrajo de Observatorio Inmobiliario de Medellín -OIME (<http://catastrooime.blogspot.com/>) y de la Alcaldía de Medellín *OpenData* (<https://geomedellin-m-medellin.opendata.arcgis.com/>). Del primero, se obtuvieron ofertas inmobiliarias entre 2018 y 2022 para construir una estimación de avalúo por manzana. Del segundo portal, las variables seleccionadas fueron las comunas, el barrio, el estrato y el uso del suelo.

De igual forma, desde *Open Street Map* para Bogotá y Medellín se identificaron las estaciones de transporte público (SITP/Metro de Medellín), clínicas, hospitales y centros comerciales. Para cada vivienda, se calculó la distancia a cada uno de los puntos de referencia mencionados.

Por otro lado, en un análisis inicial de las bases de datos dadas se identificaron cuatro variables de interés con un alto porcentaje de datos faltantes. Estas son: área cubierta (81,2%); área total (74,2%); número de cuarto (49,8%), y; baños(28%). Estas variables se clasifican dentro de la categoría *características inherentes a la vivienda o atributos físicos* y se tomaron teniendo en cuenta la literatura económica.

De estas variables se recuperaron datos considerables con la descripción, utilizando el método de expresiones regulares. Las observaciones restantes se imputaron por manzana. Si la manzana no estaba disponible en la información, se realizó una imputación por el siguiente nivel de área identificado.

Entre los grandes retos para conformar una buena base para trabajar los modelos de predicción fue recuperar las observaciones de la información espacial obtenida de datos del *OpenData* de Medellín y de Datos Abiertos de Bogotá. En particular, después de aplicar una corrección de geometrías y hacer un *spacial join*, quedaban un 30% aproximado de observaciones faltantes para las variables de manzanas, estrato, avalúo y usos del suelo. Para corregir esto, a las observaciones con NAs se les aplicó un *spacial join* con el método del vecino más cercano (*st\_nn*) con un umbral de distancia de 50 metros cuadrados. Esto permitió reducir en más de un 95% los NAs de estas variables. Lograr esto permitió una imputación con un menor nivel de variabilidad.

Localidad de Chapinero y observaciones base Test	Comuna El Poblado y observaciones base Test
	

Nota: Las observaciones de la base de entrenamiento se distribuyen en un 80% de Bogotá y el 20% de Medellín. La base de datos de prueba y sobre la cual se entregan la predicción está distribuida en un 93% Medellín (El poblado) y un 7% Bogotá (Chapinero).

De este modo, las variables seleccionadas para conformar la base de datos comprenden aquellas que describen las características propias de la vivienda y su ubicación. Estas variables se consideran importantes dado que el caso del área cubierta y área total, Figueroa (1992) sostiene que esta variable permite captar las variaciones en el metro cuadrado de la vivienda, por lo que por cada metro cuadrado de superficie y área construida aumentará el tamaño de la vivienda y este impactará en el precio de la misma. De igual manera, la variable *número de cuartos* es importante para el mismo autor, puesto que, además de ser un indicador del tamaño de la vivienda, también puede ser un indicador de la calidad de la vivienda, lo que impacta el precio de la misma. Por último, dentro de esta categoría de atributos de la vivienda, Favela, Galindo, Herrera y Rizo (2010) sostienen que el número de baños es uno de los atributos que el consumidor tiene en cuenta al momento de adquirir una vivienda y este guarda una relación directa con el precio de la misma. De igual forma, las *variables de ubicación de la vivienda* se seleccionaron para responder a los modelos de economía urbana (Modelo monocéntrico, dinámicas del mercado inmobiliario, economía espacial y de aglomeraciones).

### 3. Modelos y resultados:

Se construyeron modelos predictivos para Bogotá, Medellín y para Bogotá y Medellín en conjunto. Esta aproximación se hizo en el objetivo de identificar si al hacer un modelo por ciudad se obtenían mejores resultados que realizando un único modelo para las dos ciudades. Dentro de los modelos y algoritmos empleados están: OLS con validación cruzada *K-fold*; lasso; ridge; elastic net; árboles de predicción; random forests y XGBoost.

El mejor modelo se identificó bajo un Random Forest que se ajusta bajo la validación cruzada de datos. Entre los modelos evaluados, éste logra el mayor número de viviendas compradas (76,6%) y presenta la menor diferencia en la predicción del dinero gastado y el dinero gastado si la compra se realizara con el precio de lista de la base de datos. Este modelo empleó la siguiente forma funcional:

$$\begin{aligned} \text{precio}_{prop} = & \beta_0 + \beta_1 \text{baños} + \beta_2 \text{habitaciones} + \beta_3 \text{parqueaderos} + \beta_4 \text{área}_{m2} + \beta_5 \text{estrato} \\ & + \beta_6 \text{dist}_{trans.público} + \beta_7 \text{dist}_{parques} + \beta_8 \text{distancia}_{hospitales} \\ & + \beta_9 \text{distancia}_{centros.comerciales} \end{aligned}$$

El sentido económico de dichas variables se discutió en la sección anterior. Los demás modelos probados fueron diferentes combinaciones de variables y diferentes algoritmos de entrenamiento y predicción. La siguiente tabla presenta las comparaciones entre algunos de los modelos con mejor desempeño:

**Tabla 1. Comparación del desempeño de diferentes modelos de predicción (precios en millones de pesos)**

Modelo	Total precio de lista propiedades compradas	Precio total pagado	Diferencia Precio lista – pagado	Propiedades compradas	% propiedades compradas	Precio promedio
Elastic Net 1	6.395.755	10.061.116	- 3.665.361	15.431	72	652,00
Lasso 1	6.515.060	10.082.376	- 3.567.317	14.618	68	689,72
OLS-CV 4	6.880.268	10.120.048	- 3.239.780	14.181	66	713,63
<b>Random Forest 1</b>	<b>8.615.237</b>	<b>10.233.493</b>	<b>- 1.618.256</b>	<b>16.361</b>	<b>77</b>	<b>625,48</b>
Random Forest 2	8.533.069	10.210.844	- 1.677.775	16.272	76	627,51
Ridge 1	6.492.296	10.057.130	- 3.564.834	14.922	70	673,98
Tree 1	7.203.027	9.534.689	- 2.331.662	14.689	69	649,10
XGBoost 1	8.127.214	9.866.287	- 1.739.073	14.537	68	678,70

\*Se resalta en negrilla el modelo con el mejor desempeño.

La tabla anterior se puede resumir de la siguiente manera para el mejor modelo: compró el 76% de propiedades de la base de test, con un precio promedio de \$625,5 millones de pesos. Por estas viviendas, pagó un exceso de \$1.6 billones de pesos (teniendo en cuenta que pagó en total 10,2 billones por propiedades cuyo precio de lista era de 8,6 billones). Si bien el algoritmo pagó en exceso, fue el modelo que menos pagó en esta línea.

Finalmente, sobre este modelo Random Forest con mejor desempeño, cabe mencionar que empleó los siguientes hiperparámetros: validación cruzada con K=5; número de variables aleatorias para formar los distintos árboles  $mtry = \sqrt{10}$ , en donde 10 era el número de predictores en este modelo particular; *tunegrid* de expansión con base en el anterior valor de *mtry*.

#### 4. Conclusiones y recomendaciones:

Con este tercer taller, pudimos explorar otra dimensión del Big Data, no tanto caracterizada por el volumen y cantidad de observaciones, sino por su complejidad y diferentes dimensiones: los datos de carácter geográfico. Este tipo de datos y las diferentes funcionalidades desarrolladas en R para su tratamiento permiten abordar otro tipo de problemas económicos, como el desarrollado en este taller. Además, el hecho de poder agregar a las observaciones variables de corte geográfico, como la proximidad a diferentes puntos de interés, la localización específica, la pertenencia de un punto a determinado lugar o subdivisión administrativa, entre otros, permite contar con información valiosa a la hora de hacer predicciones o clasificaciones para la solución de problemas económicos. Esto lo pudimos observar de primera mano en este taller: gracias a determinar la manzana a la que pertenecía cada una de las viviendas de las bases de datos, pudimos luego imputar información promediada por manzana, como el área, número de baños, cuartos, parqueaderos, entre otros. Es decir, se le puede agregar la dimensión geográfica a la información y así poder desarrollar análisis mucho más completos. Un comentario adicional sobre los datos geográficos es que las operaciones relacionadas con la medición de distancias y la identificación de vecinos son de

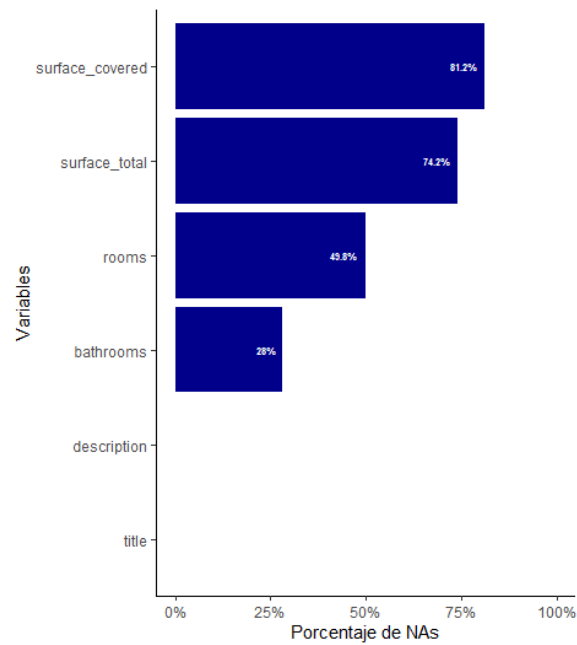
carácter exponencial, y por ende, la demanda de tiempo y capacidad de cómputo crece considerablemente conforme crecen el número de observaciones. Aquí una recomendación fundamental es, siempre que sea posible, dividir el problema, es decir, elaborar subconjuntos de datos que tengan un sentido económico o administrativo para disminuir de manera dramática el tiempo de procesamiento. Así, después de dividir los datos por comuna o localidad, pudimos en nuestro caso hallar las manzanas para todas las observaciones de Medellín y Bogotá en menos de dos horas, cuando la operación para la base completa, sin fraccionar, superó las 30 horas.

Desde el punto de vista de los algoritmos de machine learning, pudimos en esta ocasión constatar cómo un *random forest* obtuvo el mejor desempeño de acuerdo con los criterios específicos del problema en cuestión. Esto es fundamental resaltarlo, ya que no hay un único modelo o algoritmo que sea mejor en todos los casos; es decir, es importante siempre situar la selección del mejor modelo en el contexto específico del problema.

Finalmente, este taller nos permite concluir que la aproximación de Big Data y Machine Learning para abordar problemas económicos tiene un gran potencial y una inmensa gama de aplicaciones. Intentar definir un modelo econométrico preciso, con base en la teoría económica, para la predicción de inmuebles puede ser bastante complejo, mientras que la aproximación aquí desarrollada nos permitió, por lo menos en las bases de prueba, alcanzar resultados bastante interesantes. Por supuesto, una combinación inteligente de ambos enfoques puede ser exitosa para la solución de muchos problemas.

Anexos:

Ilustración 1. Tabla de porcentaje de datos faltantes de Base Train



Fuente: Elaboración propia

Tabla 1. Proporción de estratos Base Test

Characteristic	N	Overall, N = 11,150 <sup>i</sup>	Bogotá D.C, N = 793 <sup>i</sup>	Medellín, N = 10,357 <sup>i</sup>
ESTRATO	11,148			
0		35 (0.3%)	35 (4.4%)	0 (0%)
2		33 (0.3%)	0 (0%)	33 (0.3%)
3		707 (6.3%)	434 (55%)	273 (2.6%)
4		667 (6.0%)	324 (41%)	343 (3.3%)
5		1,383 (12%)	0 (0%)	1,383 (13%)
6		8,323 (75%)	0 (0%)	8,323 (80%)
Unknown		2	0	2
n (%)				

Fuente: Elaboración propia

Tabla 2. Proporción de estratos por Base Train

Characteristic	N	Overall, N = 107,567 <sup>i</sup>	Bogotá D.C, N = 86,211 <sup>i</sup>	Medellín, N = 21,356 <sup>i</sup>
ESTRATO	106,999			
0		6,821 (6.4%)	6,821 (8.0%)	0 (0%)
1		548 (0.5%)	407 (0.5%)	141 (0.7%)
2		8,097 (7.6%)	6,752 (7.9%)	1,345 (6.3%)
3		20,853 (19%)	15,662 (18%)	5,191 (24%)
4		24,165 (23%)	17,537 (20%)	6,628 (31%)
5		22,864 (21%)	15,617 (18%)	7,247 (34%)
6		23,651 (22%)	22,850 (27%)	801 (3.8%)

Unknown	568	565	3
n (%)			

Fuente: Elaboración propia

*Tabla 3. Proporción de tipo de propiedad por ciudad de Base Train*

Characteristic	N	Overall, N = 107,567 <sup>i</sup>	Bogotá D.C, N = 86,211 <sup>i</sup>	Medellín, N = 21,356 <sup>i</sup>
property_type	107,567			
Apartamento		81,577 (76%)	65,156 (76%)	16,421 (77%)
Casa		25,990 (24%)	21,055 (24%)	4,935 (23%)
n (%)				

Fuente: Elaboración propia

*Tabla 4. Proporción de tipo de propiedad por ciudad de Base Test*

Characteristic	N	Overall, N = 11,150 <sup>i</sup>	Bogotá D.C, N = 793 <sup>i</sup>	Medellín, N = 10,357 <sup>i</sup>
property_type	11,150			
Apartamento		9,658 (87%)	735 (93%)	8,923 (86%)
Casa		1,492 (13%)	58 (7.3%)	1,434 (14%)
n (%)				

Fuente: Elaboración propia

*Tabla 5. Media del área vivienda por ciudad base Train*

Characteristic	N = 107,567 <sup>i</sup>	Media área vivienda
13		
Bogotá D.C	86,211 (80%)	152
Medellín	21,356 (20%)	122
n (%)		

Fuente: Elaboración propia

*Tabla 6. Media del área vivienda por ciudad base Test*

Characteristic	N = 11,150 <sup>i</sup>	Media área vivienda
13		
Bogotá D.C	793 (7.1%)	78
Medellín	10,357 (93%)	228
n (%)		

Fuente: Elaboración propia

*Tabla 7. Media del área vivienda por ciudad base Train*

Characteristic	N = 107,567 <sup>i</sup>	Media pecio vivienda
13		
Bogotá D.C	86,211 (80%)	764,283,984
Medellín	21,356 (20%)	405,774,604
n (%)		

Fuente: Elaboración propia

## BIBLIOGRAFÍA:

Favela, A., Galindo, C., Herrera, D., Rizo, J., (2010). Determinantes del Precio de la vivienda en la zona metropolitana de Monterrey. Recuperado el 25 de julio del 2022, de <http://ree.economiatic.com/A2N2/207291.pdf>

Figuerola, E., Lever, G., (1992). Determinantes del Precio de la vivienda en Santiago: una estimación hedónica. Recuperado el 25 de julio del 2022, de [https://repositorio.uchile.cl/bitstream/handle/2250/128244/Eugenio\\_Figuerola\\_B.pdf?sequence=1&isAllowed=y](https://repositorio.uchile.cl/bitstream/handle/2250/128244/Eugenio_Figuerola_B.pdf?sequence=1&isAllowed=y)