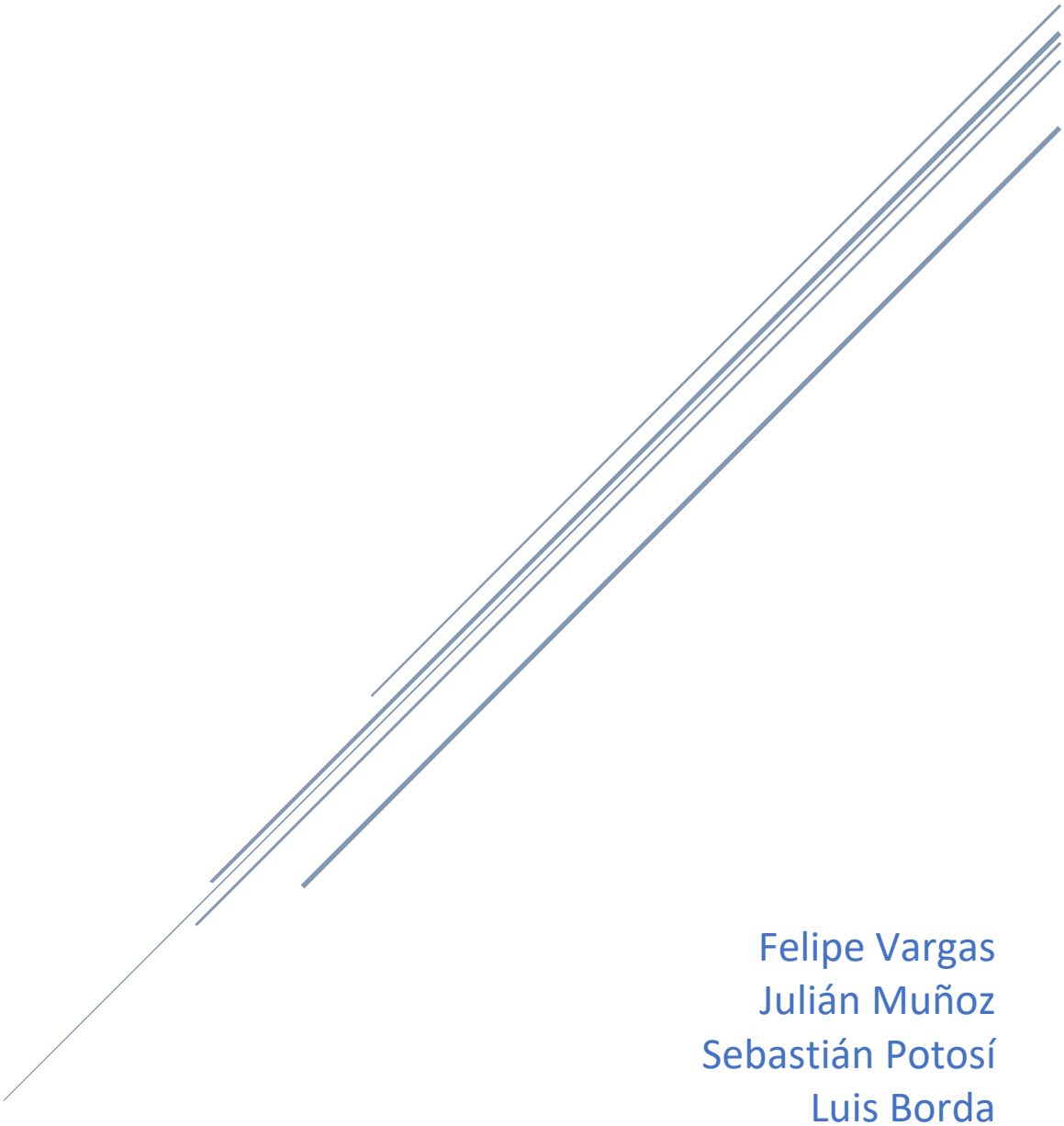
TALLER 2

BIG DATA Y MACHINE LEARNING PARA ECONOMÍA APLICADA



Felipe Vargas
Julián Muñoz
Sebastián Potosí
Luis Borda
MECA

Contenido

1 INTRODUCCIÓN

A diferencia de la mayoría de los bienes económicos, las viviendas se caracterizan por ser "bienes heterogéneos que poseen una diversidad de atributos físicos, funcionales, de localización y de durabilidad, a la vez que proveen una gama de servicios, como confort, seguridad, proximidad al empleo y medios de transporte, etc., que las hacen prácticamente únicas e irrepetibles" (Desormeaux, 2003). Sin embargo, dado que lo que se tranza es la vivienda como un todo, la mayoría de las veces no es posible observar los precios o valoraciones marginales objetivas de cada uno de ellos.

En consecuencia, es de interés conocer el precio implícito o hedónico de cada atributo que conforma la vivienda. Al respecto, Rosen define la teoría de precios hedónicos como un problema de economía del equilibrio espacial en el que todo el conjunto de precios implícitos guía las decisiones de ubicación tanto del consumidor como del productor (Rosen, 1974). La literatura reciente demuestra que factores como el número de dormitorios, baños, habitaciones de uso múltiple y cercanía a espacios como paraderos de buses, instituciones educativas, centros de salud, estaciones de policía o CAI, parques, bancos, cajeros, escenarios culturales, entre otros, influyen de forma significativa en el precio de las viviendas.

En este contexto, el presente documento desarrolla un ejercicio para predecir los precios de viviendas en la localidad de Chapinero en Bogotá haciendo uso de modelos de precios hedónicos y técnicas de aprendizaje de máquinas. El objetivo de los modelos planteados es comprar la mayor cantidad gastando lo menos posible. Para esto se utilizaron diferentes métodos: modelos lineales, técnicas de regularización, árboles de decisión, bosques y bagging, con el propósito de obtener una comprensión más profunda del espacio de predicción. La base de datos utilizada para este *Set,* cuenta con una muestra de 38.644 observaciones y 16 variables, además de algunas variables que se pueden extraer a partir de la descripción de cada inmueble en venta y otras adicionales, que se crearon con el propósito de optimizar la capacidad predictiva del modelo.

2 DATOS

2.1 Descripción de los datos, limpieza y proceso de construcción de la muestra

Con la información de las 38.644 propiedades de la ciudad de Bogotá obtenidos se evaluaron atributos tales como; precio, numero de baños, superficie, habitaciones, el tipo de propiedad (casa o apartamento), latitud, longitud y una descripción general del bien que proporcionan los vendedores del inmueble. Las anteriores son características comúnmente evaluadas a la hora de comprar el bien, no obstante, se precisa i) Completar de forma adecuada los datos faltantes en cada característica y ii) construir algunas variables espaciales para la construcción de los modelos. Lo anterior nos permitió tener un modelo lo suficientemente robusto en la labor de predicción.

Se realizó un análisis provisional de la base de datos en el cual se pudo identificar diferentes características de las casas y apartamentos en venta ubicados en la ciudad. La imputación de los datos se realizó de forma convencional según mediana y moda, donde: se evidencia en la base de train que para la variable rooms y bathrooms el valor que más frecuente es 3, es decir que la mayoría de los inmuebles tenía 3 cuartos y/o 3 baños y ese fue el parámetro para la imputación. De igual modo, se realizó para el resto de las variables teniendo en cuenta la mediana de la superficie total y cubierta.

Por último, los datos espaciales se tomaron de open streeet map. Como polígonos se usaron las variables parques, estaciones de Transmilenio y universidades. Para medir la distancia de esos polígonos se creó un centroide en cada uno de ello. Las demás variables se tomaron como un punto específico y se midió la distancia mínima hasta cada vivienda.

2.2 Análisis descriptivo de los datos

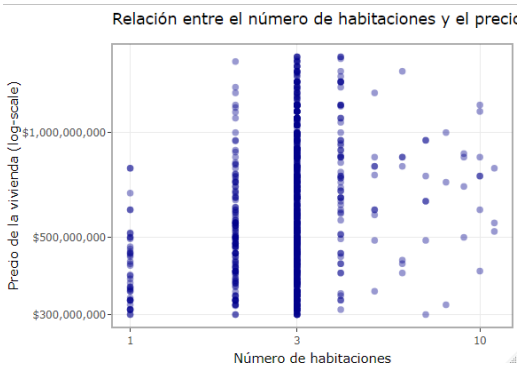
La información utilizada contiene datos de tres tipos para lograr medir cualidades de las viviendas de las que se tiene información: 1) atributos estructurales; 2) Atributos de vecindario; y 3) atributos de servicios locales. A continuación, se describe cada grupo de variables:

2.2.1 Atributos estructurales de la vivienda

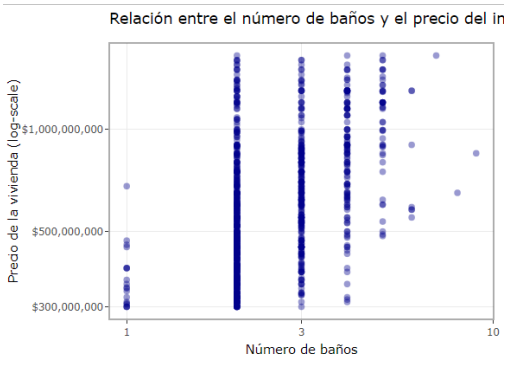
En la actualidad se pueden encontrar numerosos trabajos que abordan la determinación del precio de las viviendas desde distintos enfoques, siendo el modelo de precios hedónicos uno de los métodos más utilizados en la literatura. Estos modelos permiten conocer el precio implícito de cada uno de los atributos que componen los inmuebles donde las características físicas como el área, el número de habitaciones, número de baños y las zonas comunes, entre otras, juegan un papel fundamental en la conformación del precio de la vivienda (Toloza & Melo , 2021). En este sentido, para el ejercicio de predicción se tomaron las variables de número de habitaciones, número de baños y área de la vivienda. Para este caso particular, la media de habitaciones es 3 y la media de baños es 2. En la Figura 1 se evidencia que las viviendas o apartamentos con menor número de habitaciones y baños se tienen los menores precios.

Figura 1: Variables estructurales de la vivienda

Panel A: Relación entre número de habitaciones y precio vivienda



Panel B: Relación entre el número de baños y el precio de la vivienda



Fuente: Elaboración propia

Finalmente, el anexo 1, evidencia que la relación entre el área de la vivienda y su precio, es en principio positiva.

2.2.2 Atributos de vecindario

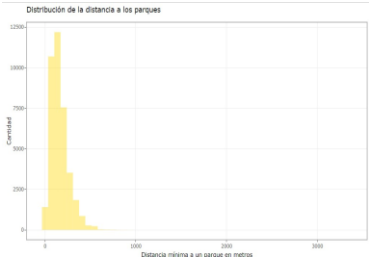
Parques

Los parques urbanos y sus espacios verdes cumplen una función más allá de lo ornamental, coadyuvando a mejorar la calidad del aire al tomar un papel de moderador de intercambio de aire, calor y humedad en el paisaje urbano. Al mismo tiempo toma un papel perceptual paisajístico que participa como deleite visual y por consiguiente mejorando la calidad de vida

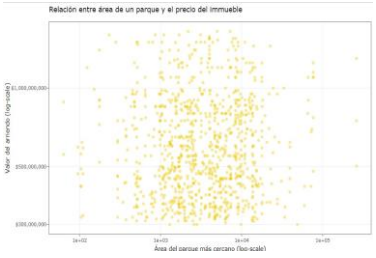
urbana. Se consideran zonas verdes a todas aquellas superficies de parques y jardines y otros espacios públicos (plazas, ramblas, interiores de manzana, etc.) dotadas de cobertura vegetal que estén localizados dentro de los límites del área urbana consolidada (Vidaurre & Olivera, 2018). En consecuencia, la variable parques se considera de gran importancia para predecir los precios de la vivienda, en la figura 1 se muestra su comportamiento. Cabe resaltar que su unidad de medida es el área de cada uno de los polígonos.

Figura 2: Parques

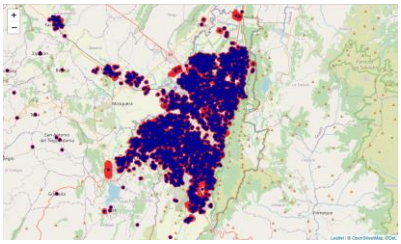
Panel A: Distribución distancia parques



Panel B: Relación entre área parque y precio vivienda



Panel C: Distribución espacial



Fuente: Elaboración propia

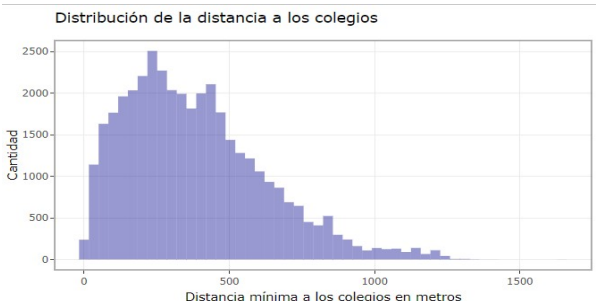
En el panel A, se evidencia que la distribución que la distancia de las viviendas a los parques se concentra alrededor de los 200 metros, su media es 160 metros y su mediana 139 metros. En el Panel B no se evidencia una clara relación entre el área del parque mas cercano y el valor del arriendo de las viviendas, sin embargo, si se vislumbra una distribución que tiende a la normal ya que los datos se concentran alrededor de su media. Finalmente, en el panel C se evidencia la distribución espacial de la variable.

Universidades, colegios y jardines

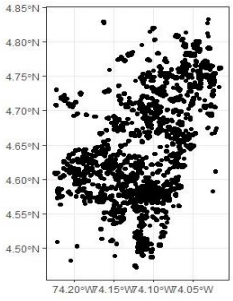
Algunos hogares o personas pueden valorar mejor el hecho de que una vivienda este cerca de universidades, colegios o algunas instituciones educativas ya que ellos mismos o sus hijos deben recorrer menores distancias para llegar a su lugar de estudio (Manfrino, 2021). En este sentido se considera la cercanía a este tipo de lugares crucial para predecir el precio de una vivienda. Su comportamiento se muestra en la Figura 3. En el panel A, se evidencia que la distancia mínima de las viviendas a colegios se concentra alrededor de los 250 metros y su distribución esta sesgada hacia la izquierda. Su mediana es 382 metros, su mediana 343 metros. Por su parte, en el panel B, no se vislumbra una relación clara entre el precio de la vivienda y esta variable,

Figura 3: Colegios

Panel A: Distribución de la variable colegios



Panel B: Relación precio y distancia a colegios



Fuente: Elaboración propia

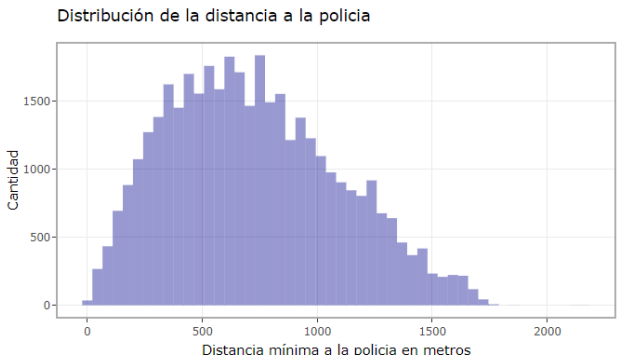
Para el caso de los jardines, su media es 572 metros y para el caso de las universidades es 1059 metros.

Estaciones de policía o Centros de Atención Inmediata (CAI)

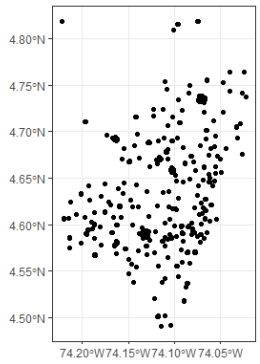
Algunas investigaciones han demostrado que la cercanía a estaciones de policía ayuda a incrementar la sensación de seguridad y esto tiene efectos en el precio de las viviendas (Arends, 2017); por lo cual se utilizó esta variable para el ejercicio. La Figura 4 muestra el comportamiento de esta variable. La distancia de las viviendas a la estación o CAI de policía mas cercana tiene una media de 722 metros y una mediana de 685 metros. El panel A de la figura evidencia que su distribución esta sesgada hacia la izquierda y los datos de concentran alrededor de los 600 metros aproximadamente. Por último, el panel B no se evidencia una relación clara entre la distancia a la policía y el precio de las viviendas.

Figura 4: Estaciones de policía

Panel A: Distribución



Panel B: Relación entre precio y distancia a estaciones de policía



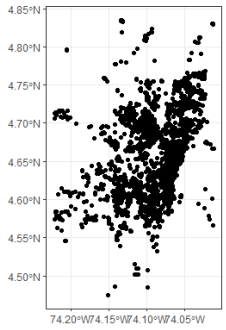
Fuente: Elaboración propia

2.2.3 Atributos de servicios locales

Restaurantes

Las actividades recreativas, de socialización o culturales también son un incentivo para atraer habitantes, por lo cual la cercanía a plazas públicas, restaurantes, entre otros son elementos que agregan valor a las viviendas (Ramírez, 2016). Por esta razón la cercanía a restaurantes es una variable a utilizar en el análisis. Su media es 572 metros y su mediana 481 metros. No se evidencia una relación clara entre esta variable con el precio de las viviendas.

Figura 5: Relación entre distancia a restaurantes y precio de las viviendas.



Fuente: Elaboración propia

Transmilenio

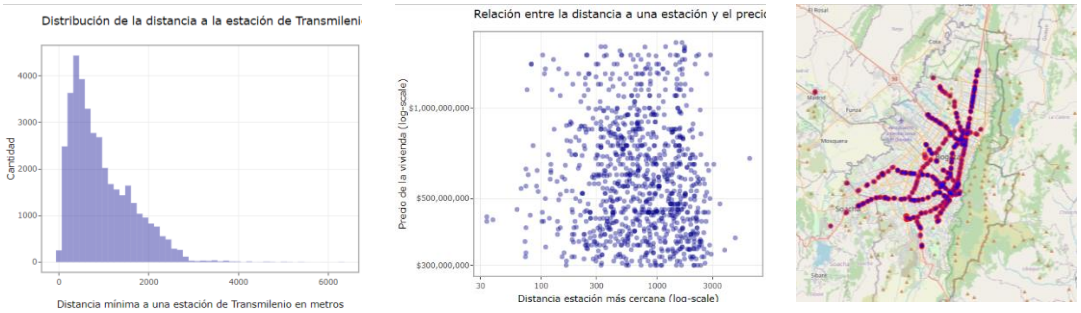
Se ha demostrado que el acceso a las estaciones de TransMilenio en Bogotá capitaliza el valor de las tierras estimando que, por cada cinco minutos extras de caminata hacia una estación disminuye el valor de las propiedades entre 6,8% y 9,3%. (Rodríguez & Targa, 2004). Por esta razón se considera crucial tener en cuenta esta variable para predecir los precios de la vivienda. La media de esta variable es 949 metros y su mediana 967 metros.

Figura 6: Transmilenio

Panel A: Distribución

Panel B: Relación entre la variable y el precio

Panel C: Distribución espacial de la variable



Fuente: Elaboración propia

Distancia a avenida principal

Diferentes estudios demuestran que la distancia de la vivienda a una avenida principal tiene gran influencia en su precio ya que en la medida que esta sea más corta, se puede acceder con mayor facilidad a diferentes medios de transporte y lugares estratégicos de la ciudad. Por esta razón es una de las variables utilizadas en el análisis. Para este caso particular, la media de esta variable es 233 metros y su mediana es 200 metros.

3 MODELOS Y RESULTADOS

En esta sección se exponen las variables utilizadas, el entrenamiento del modelo, selección de los hiperparámetros y finalmente se hace un análisis comparativo.

3.1 Variables utilizadas

Como se mencionó en la sección de datos, las variables utilizadas se dividieron en 3 grupos. Un primer grupo comprende variables relacionadas con aspectos estructurales de las viviendas tales como superficie total, superficie con cubierta, número de habitaciones, número de baños y tipo de propiedad, entendiendo que son los aspectos que en principio definen el precio de las viviendas. El segundo grupo está constituido por atributos del vecindario tales como distancia al parque más cercano, área del parque más cercano, distancia a avenida principal, distancia a universidad, distancia a una estación de policía, distancia a jardines infantiles. Por último, se utilizaron variables relacionadas con atributos de servicios locales como distancia a estación de Transmilenio y distancia a restaurantes.

3.2 Entrenamiento del modelo, selección de hiperparámetros y Análisis comparativo

El modelo con la mejor predicción dentro de los 7 modelos realizados fue un modelo de Super Lerner, el cual, es una técnica mediante la cual combinamos múltiples modelos de aprendizaje automático para mejorar la precisión de las predicciones. Lo que realizo el modelo fue identificar las fortalezas de diferentes modelos y mitigar sus debilidades, para de tal manera obtener un rendimiento de predicción más robusto y generalizado. Para la implementación del modelo realizamos el siguiente procedimiento:

1. **División de datos**: El conjunto de datos se divide en un conjunto de entrenamiento y un conjunto de validación o prueba. Los modelos base se entrenaron en el conjunto de entrenamiento. Para obtener las bases definitivas se realizó los métodos de limpieza respectivos ya mencionados, para obtener una base de entrenamiento 38.644 y la de testeo con 10.014.
2. **Generación de predicciones**: Se realizo el proceso de predicción con el método ya mencionado y se imputaron los datos a la base en cuestión, para así poder determinar
3. **Combinación de predicciones:** El Super Learner combina las predicciones de los modelos base de diversas maneras. Utilizando en nuestro caso la regresión de mínimos cuadrados no negativos y combinación convexa para fusionar las predicciones de los modelos base. La idea fue dar más peso a los modelos que funcionan bien en el conjunto de validación y menos peso a los modelos que no lo hacen.
4. **Evaluación de rendimiento**: Se evaluó el rendimiento del Super Learner combinado en el conjunto de validación o prueba utilizando métricas de evaluación apropiadas, como el error cuadrático medio (MSE) para regresión o la precisión y el área bajo la curva ROC (AUC) para clasificación.
5. **Ajuste de hiperparámetros:** Elegimos el modelo que por defecto iba a ser integrado en el Super Learner, para este caso seleccionamos el método NNLS (el valor predeterminado). Este es bajo la metodología de mínimos cuadrados no negativos basados en el algoritmo de Lawson-Hanson y el método dual de Goldfarb. Para este caso NNLS funcionara tanto para resultados gaussianos como binomiales
6. **Predicciones finales:** Ya el modelo de Super Learner ha sido entrenado y ajustado, se puede utilizar para realizar predicciones en nuevos datos.

Por otro lado, este modelo lo seleccionamos como el mejor debido a su valor del MAE; para este ejercicio diseñamos 7 modelos con sus respectivas métricas:

- Un primer modelo de OLS, el cual lo utilizamos para modelar la relación entre la variable de respuesta (Precio de la vivienda) y las demás variables predictoras (Explicadas anteriormente). **Obteniendo un RMSE: 268.102.637**
- Un segundo modelo de Ridge, en el cual buscábamos una penalización en los coeficientes del modelo con el objetivo de mejorar la calidad de las predicciones **Obteniendo un RMSE: 268.191.478**
- Un tercer modelo de Lasso es una técnica utilizada en modelos predictivos, especialmente en el contexto de regresión lineal, para abordar problemas como la selección de características, la reducción de la dimensionalidad y la prevención del sobreajuste. Buscamos introducir una penalización en los coeficientes del modelo con el objetivo de mejorar la calidad de las predicciones. **Obteniendo un RMSE: 268.239.019**
- Un cuarto modelo de Elastic Net, en el cual buscábamos combinar elementos de la regresión Ridge (L2) y la regresión Lasso (L1). Nuestro objetivo era abordar problemas en la selección de características, la reducción de la dimensionalidad y la prevención del sobreajuste, al mismo tiempo que aprovecha las ventajas de ambas técnicas. **Obteniendo un RMSE: 268.178.096**
- En el quinto y sexto modelo, se realizaron dos árboles de decisión, dado que, estos aprenden las formas funcionales no lineales. A partir de particiones recursivas binarias dividiendo el espacio variable por variable y escogiendo el error estandar menor. De los resultados, se evidencia que:

  Error promedio en la muestra, fue de 213.703.743

  Error promedio fuera de la muestra fue 213.489.943

- Finalmente, como último modelo alternativo corrimos un Random Forest, el cual al combinar múltiples árboles de decisión, puede reducir el sobreajuste (overfitting) y mejorar la capacidad de generalización, lo que nos llevo a predicciones más precisas

Y, por último, el modelo se Super Lerner, nuestro mejor modelo obtuvo **un MAE de 179.035.166,** convirtiéndolo así en nuestro mejor modelo.

4 CONCLUSIONES Y RECOMENDACIONES

Para los modelos que se usaron fue muy importante el uso de las variables de vecindario o de soportes urbanos, los cuales son relevantes a la hora de tomar la decisión de comprar una vivienda y , en consecuencia, con el precio final de las viviendas.

Entre tanto, las variables espaciales a simple vista no guardar una relación lineal, pero son de gran importancia por la correlación espacial que tienen. Estos análisis son usados regularmente para encontrar zonas homogéneas físicas, que ayudan a clasificar los territorios y demuestran en modelos como los utilizados, relaciones espaciales fundamentales a la hora de predecir precio de las viviendas u otras variables que dependen de la ubicación.

Por otro lado, los modelos de árboles tienen un buen comportamiento y para los datos utilizados no demostraron un alto costo computacional, por lo que en el trade off de costo computacional y eficiencia del modelo, puede resultar en un opción óptima.

El Super Learner es efectivo porque se beneficia de la diversidad de los modelos base, lo que lo hace más robusto frente a diferentes patrones de datos. Además, puede adaptarse a problemas de aprendizaje automático donde no está claro cuál es el mejor modelo para usar, ya que combina

varios en lugar de depender de uno solo. Sin embargo, su desventaja es que puede ser más costoso computacionalmente y requerir más tiempo de entrenamiento debido a la combinación de múltiples modelos

5 LINK REPOSITORIO

https://github.com/Luis-Borda/BIG_DATA_G_10_T2.git

6 BIBLIOGRAFÍA

Arends, L. (2017). *Una visión comparada de la vivienda.*

Desormeaux, D. (2003). *Precios Hedónicos e Índices de Precios de viviendas.*

Manfrino, M. (2021). *Aplicación del método de precios hedónicos para la estimación del valor de terrenos en barrios privados del conurbano bonaerense.*

Ramírez, E. (2016). *¿Cuál es la contribución del espacio público al precio de las viviendas?* CIDE.

Rodríguez, D., & Targa, F. (2004). Value of accesibility to Bogota's Bus Rapid Transit System.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, 34-55.

Toloza, J., & Melo , O. (2021). Determinantes del precio de vivienda nueva en Bogotá para el año 2019. *Ingeniería y ciencia.*

Vidaurre, R., & Olivera, S. (2018). Urban parks in La Paz city, Bolivia: Public policy applications. *Public policy applications.*

ANEXOS

Anexo 1