**Problem Set 3: Making Money with ML?**

1. **Introducción**

Este estudio se basa en un modelo predictivo que contribuye al mercado inmobiliario, particularmente a la compraventa de propiedades, cuyo objetivo es comprar la mayor cantidad de propiedades en el barrio de Chapinero en Bogotá, Colombia, invirtiendo lo menos posible. Los datos para este análisis se toman de una muestra de datos de propiedades individuales en Bogotá, los cuales se extraen de https://www.properati.com.co. Con este modelo de predicción se pretende generar evidencia para tomar las mejores decisiones de compra de inmuebles basándose en las variables que mayor influencia tienen en el precio. En este caso, se consideran variables del entorno de la vivienda, como la tasa de homicidio, tasa de hurto a residencias, número de colegios, hospitales y parques en la UPZ y el estrato, que depende de atributos como vías de acceso, puntos de transporte e infraestructura social. Por otro lado, se incluyen variables inherentes a la vivienda, tales como el área en metros cuadrados, si tiene parqueadero, terraza, depósito o patio.

En el presente documento se consideran cuatro (4) modelos predictivos y se profundiza en el que presenta mejor desempeño. Se evalúan una regresión lineal simple, regresiones lineales regularizadas de Lasso y Ridge y, finalmente, un modelo Random Forest. Como resultado del ejercicio, se obtiene que el modelo Random Forest presenta el mejor desempeño en la predicción de precios de vivienda en la localidad de Chapinero, con un RMSE (Root Mean Squared Error) de XXXX.

**Nota:** la base de datos usada, al igual que el script de R y el presente documento están disponibles en el repositorio de GitHub en el siguiente enlace:

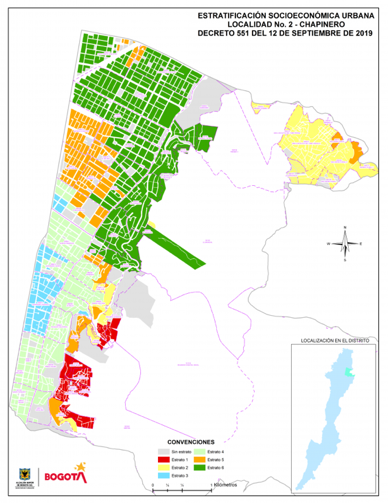
<https://github.com/Nelson1802/Repositorio_taller3.git>

**Contexto**

En Colombia, en el primer semestre de 2022 se vendieron en promedio 127.218 viviendas nuevas, representando un crecimiento de 2,5% en comparación con el mismo periodo del año inmediatamente anterior (Camacol, 2022). Adicionalmente, 7 de cada 10 viviendas vendidas fueron de interés social y las regiones que generaron más ventas fueron: Bogotá con 27 mil unidades, Valle con 18 mil, y Atlántico con 15 mil. Lo anterior representa 27,1 billones de inversión en vivienda en lo corrido del año (Camacol, 2022). Por otro lado, informes de la Cámara Colombiana de Construcción demuestran que el sector inmobiliario terminó el 2022 con un balance positivo, donde las ventas fueron 66.368 viviendas No VIS y 168.224 unidades de VIS (Semana, 2022). El promedio de las ventas ascendió en un 28% con relación a los últimos años y se estima la construcción de alrededor de 360.000 viviendas en construcción para el futuro.

Ahora bien, para entender el mercado inmobiliario en Bogotá es importante conocer algunos aspectos que pueden definir las preferencias de los consumidores a la hora de adquirir inmuebles. Para el caso objeto de análisis, cabe mencionar que Chapinero es una localidad que conforma el centro extendido de la capital y es el centro financiero, cultural y gastronómico de Bogotá, tiene cercanía con importantes lugares tanto para el trabajo como para el estudio y el esparcimiento (Properati, 2021). Esta localidad está ubicada en el centro norte de la ciudad: entre las Avenida Caracas – Autopista norte y los Cerros Orientales; y entre la Avenida 39 y la Calle 100 al norte.

**Grafica No. 1.** Estratificación socioeconómica urbana Localidad Chapinero



**Fuente:** Secretaría Distrital de Planeación de Bogotá

1. **Datos**
2. ***Descripción de las fuentes de datos***

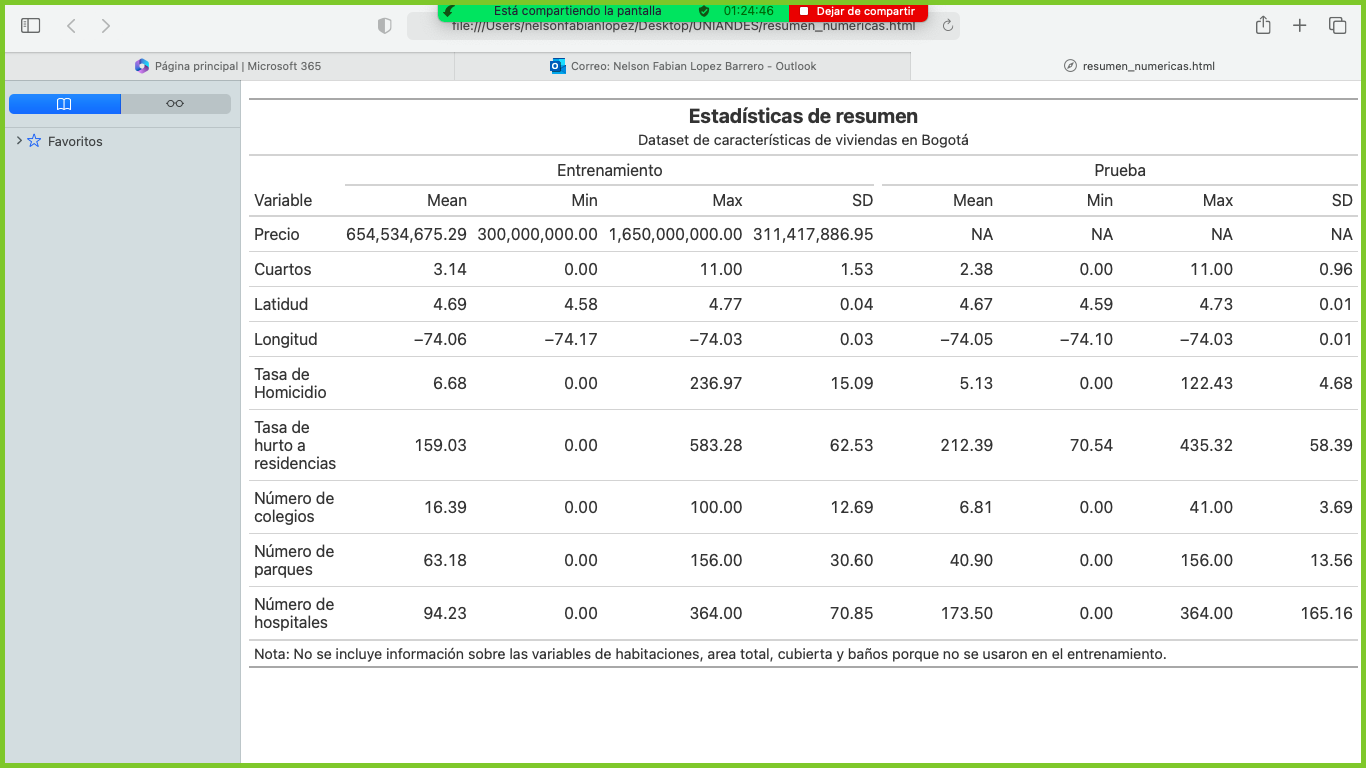
Para el desarrollo de este Problem Set se utilizarán los siguientes datos: por un lado, la tasa de homicidios y tasa de hurto a residencias por UPZ (Unidad de Planeación Zonal), entendida como una división administrativa de la ciudad que es más pequeña que las localidades, pero más grande que los barrios. Tanto la tasa de homicidios como la de hurto a residencias son tomadas de la Secretaría Distrital de Seguridad, Convivencia y Justicia y, las estimaciones de población de las UPZ de 2018 a 2021 se extrajeron de la Secretaría Distrital de Planeación. Por otro lado, para identificar la cercanía de parques, hospitales (IPS’s) y colegios, entre otros, se utilizaron datos provenientes del IDECA (Infraestructura de Datos Espaciales para el Distrito Capital) de la Unidad Administrativa Especial de Catastro Distrital. Para las fuentes que se derivan de los textos, se realiza una búsqueda de las descripciones de cada una de las viviendas, buscando por palabras clave como parqueadero, terraza, depósito y patio.

Adicionalmente, se utilizó el paquete sf (simple features) para trabajar con datos georreferenciados que permiten cargar la longitud y la latitud que aparece en el dataset de Kaggle, que a su vez, permite crear un objeto espacial que muestre en un mapa dónde están las viviendas, y luego, con la función step-join se identifican los polígonos donde están los puntos de las viviendas para asignar a cada casa el valor de tasa de homicidio, hurto a residencias, número de hospitales, número de parques y número de colegios que aparecen en cada UPZ.

1. ***Análisis descriptivo de los datos (estadísticas descriptivas)***

### Este trabajo tiene como propósito construir modelos predictivos de pobreza a nivel de hogar tomando como referencia los datos de la [Gran Encuesta Integrada de Hogares - GEIH](https://www.datos.gov.co/Estad-sticas-Nacionales/Gran-Encuesta-Integrada-de-Hogares-GEIH/mcpt-3dws) año 2018. Estos datos se encuentran segmentados en dos bases de datos, por un lado, se tiene la base *training* que proporciona información de la variable pobreza e ingreso, y por el otro lado, se tiene la base de datos *testing* que no proporciona estas variables y, por lo tanto, se realiza la predicción de los hogares pobres sobre esta última considerando el mejor modelo entrenado.

**Tabla No. 1.** Estadísticas descriptivas variables numéricas



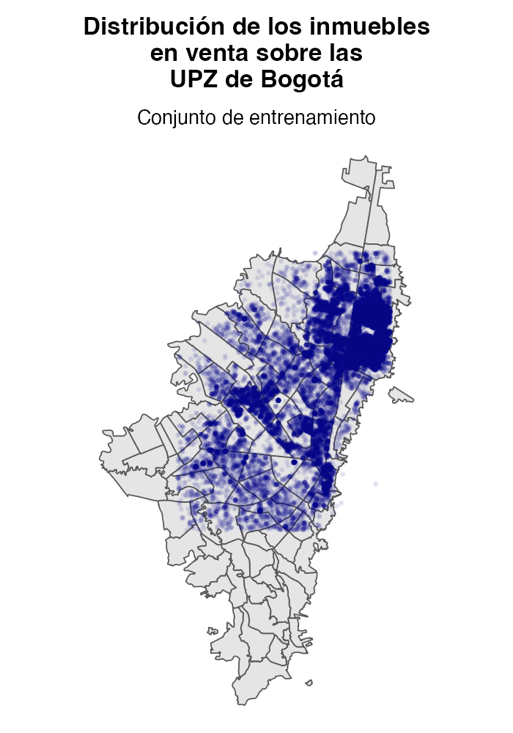
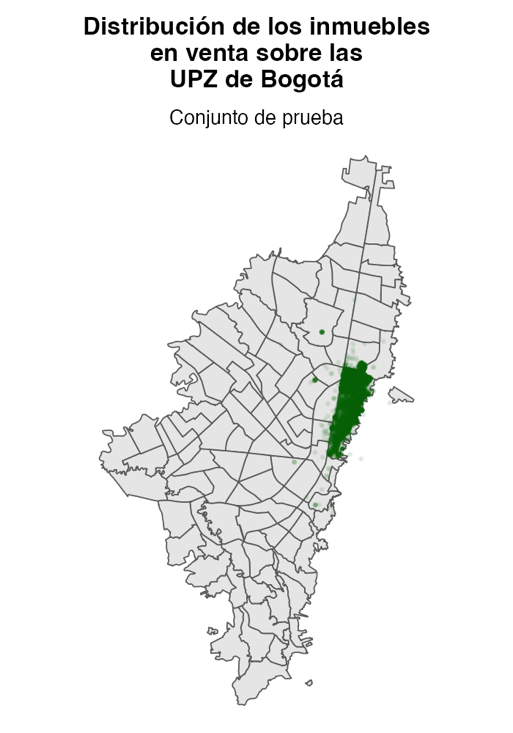
**Fuente:** R Studio

**Tabla No. 2.** Estadísticas descriptivas variables lógicas



**Fuente:** R Studio

**Grafica No. 3.** Conjunto de entrenamiento **Grafica No. 4.** Conjunto de prueba

**Fuente:** R Studio **Fuente:** R Studio

1. **Modelos y resultados**

medida de la diferencia entre los valores reales y los valores predichos por un modelo.

1. **Modelo final**
2. **Conclusión y recomendaciones**
3. **Bibliografía**

Camacol., (2022). Indicadores ventas de vivienda primer semestre 2022. Recuperado: <https://camacol.co/actualidad/noticias/indicadores-ventas-de-vivienda-primer-semestre-2022#:~:text=El%20primer%20semestre%20del%202022,lo%20revelan%20cifras%20de%20%23CoordenadaUrbana>.

Ciencuendras., (s.f.). Guía de barrio Chapinero en Bogotá. Recuperado de: <https://www.ciencuadras.com/blog/guia-de-barrio-chapinero-bogota>

DANE. (2022). Obtenido de https://www.dane.gov.co/index.php?option=com \_content&task=category&sectionid=101&id=604 &Itemid=1183

Kuhn, M. (2012). The caret package. R Foundation for Statistical Computing, Vienna, Austria.

Properti., (2021). ¿Por qué vivir en Chapinero? Recuperado de: <https://blog.properati.com.co/por-que-vivir-en-chapinero-recomendaciones/>

Rosen, S. (1974), ‘Hedonic prices and implicit markets: Product diferentiation in pure competition’, Journal of Political Economy.

Semana., (2022). Sector inmobiliario cierra el 2022 con balance positivo; ventas aumentaron en un 28%. Recuperado: <https://www.semana.com/economia/macroeconomia/articulo/sector-inmobiliario-cierra-el-2022-con-balance-positivo-ventas-aumentaron-en-un-28/202210/>