Daniel Casas Bautista

Código: **202120803**

Lucia Fillippo Aguillón

Código: **202213187**

Irina Andrea Vélez López

Código:

Miguel Ángel Victoria Simbaqueva

Código:

**Problem Set 2**

Big Data & Machine Learning

**Fecha de entrega: 15 de julio de 2023**

**Resumen:** El presente informe presenta la solución al Problem Set 2 de la clase Big Data & Machine Learning, en donde se aplicaron diversos conceptos y herramientas para la limpieza de bases de datos y el desarrollo de un modelo de predicción del precio de las viviendas en la localidad de Chapinero en Bogotá.

En el repositorio GitHub que abajo se referencia contiene el presente informe junto a los scripts de RStudio donde se generaron todos los resultados. Este repositorio se encuentra en: <https://github.com/irivelez/PS2_Making_Money_with_ML.git>

1. **Introducción**

***The introduction briefly states the problem and if there are any antecedents. It briefly describes the data and its suitability to address the problem set question. It contains a preview of the results and main takeaways.***

El precio de una vivienda se determina por una serie de factores que incluyen el tamaño total del inmueble, el número de baños y la proximidad a lugares importantes dentro de la ciudad, entre otros aspectos relevantes. Estos elementos se constituyen como variables explicativas que buscan explicar el precio de la propiedad. Cada ciudad tiene sus propias características distintivas que influyen en la determinación de precios por parte de los oferentes del mercado; sin embargo, es posible desarrollar un modelo que permita analizar el impacto individual de cada una de estas características esenciales, con el objetivo de proporcionar a los compradores y vendedores del mercado inmobiliario de vivienda, una comprensión más precisa de cómo influyen estos factores en el precio de mercado del sector inmobiliario de vivienda.

De esta manera, será posible predecir el precio de uno de los mercados más relevantes en el desempeño económico de la ciudad y del país y además, de cara al análisis de las políticas públicas, determinar el precio de mercado de uno de los bienes utilizados como política contra cíclica ante los ciclos macroeconómicos.

Considerando lo anterior, el objetivo principal de este documento es construir un modelo predictivo de los precios de la vivienda, a partir del siguiente modelo, donde la función representa un vector C con las características que explican el precio de la vivienda:

Para lograr el anterior objetivo se utilizaron datos de Properati sobre características de las viviendas, del DANE sobre XXX y de la Alcaldía de Bogotá de Datos Abiertos[[1]](#footnote-2). Para esto fue necesario realizar una limpieza de las bases de datos para lograr tener una muestra y entrenamiento para evaluar los distintos modelos de predicción y elegir el más aproximado.

1. **Datos**

***In this problem set, you are required to add expand the variables in your data (remember to expand the training and testing data), at a minimum you have to add six extra variables:***

* ***At least 4 predictors coming from external sources; these can be from open street maps***
* ***At least 2 predictors coming from the title or description of the properties***

***Describe the data, it's suitability for the problem, and the sample construction process, including how the data was cleaned, combined, and how new variables were created.***

***Include a descriptive analysis of the data. At a minimum, you should include a descriptive statistics table and two maps with its interpretation. However, I expect a deep analysis that helps the reader understand the data, its variation, and the justification for your data choices. Use your professional knowledge to add value to this section. Do not present it as a \dry" list of ingredients.***

Se ha tomado una muestra de la página *Properati,* la cual contiene características de las viviendas como ciudad, precio, número de cuartos, número de baños, precio de la vivienda, así como algunas características más específicas, con el objetivo de limpiar y consolidar una muestra de prueba y una de entrenamiento que permita evaluar distintos modelos de predicción del precio de la vivienda; lo anterior, para la localidad de Bogotá en Chapinero.

Dentro de la información que tiene la base de datos se encuentran las siguientes variables relevantes.

* **Ubicación:** Esta variable puede influenciar el precio de la vivienda, pues, en caso de estar más cercano al centro de la ciudad o a una zona empresarial, puede variar el precio de la vivienda. En la base de datos se trata de una variable que da las coordenadas geométricas.
* **Número de cuartos:** Esta variable es relevante en el análisis precisamente porque influye en el tamaño de la propiedad y en la cantidad de personas que pueden habitarla.
* **Número de baños:** Esta variable es relevante en el análisis precisamente porque influye en el tamaño de la propiedad y en la cantidad de personas que pueden habitarla.
* **Tipo de propiedad**: La variable indica el tipo de propiedad si es apartamento o casa. El tipo de propiedad afecta el precio de la vivienda precisamente por distintos motivos como el tamaño, el número de amenidades y espacios de recreación privados, entre otros.

Tras limpiar las bases de datos y crear los polígonos de análisis, los cuales se obtienen de Open Street Map, se selecciona la información de la base de datos correspondiente a estas zonas. En la Figura 1 se muestran la información de muestra y entrenamiento para la zona.

Figura 1. Polígono de estudio

Mapa

Descripción generada automáticamente

Fuente: Elaboración propia mediante Rstudio con datos de Properati

Adicionalmente, en cuanto a los predictores provenientes de bases de datos externos, se ha tomado las siguientes variables de referencia:

* Cercanías a parques y zonas de recreación: Número de kilómetros desde las observaciones hasta los parques[[2]](#footnote-3).
* Cercanía a vías arteriales: Número de kilómetros desde las observaciones hasta las vías arteriales
* Cercanía a ciclorrutas: Número de kilómetros desde las observaciones hasta la ciclorruta más próxima[[3]](#footnote-4).

1. **Modelo y resultados**

***This section presents the model(s) submitted for evaluation. When writing this section up, include:***

* ***An explanation of the variables used to train this model, remember to use the variables you added in the previous section.***
* ***A detailed explanation on how it was trained, the selection of hyper-parameters, and any other relevant information.***
* ***A comparison to at least 4 other specifications submitted to Kaggle.***

1. **Conclusiones y recomendaciones**

***In this section, you briefly state the main takeaways of your work***

1. Alcaldía de Bogotá (2023). Mapa de referencia para Bogotá D.C. Datos abiertos. Disponible en: <https://datosabiertos.bogota.gov.co/dataset/mapa-de-referencia> (Consultado el 12 de julio de 2023). [↑](#footnote-ref-2)
2. Alcaldía de Bogotá (2023). Parques y zonas de recreación. Disponible en: <https://datosabiertos.bogota.gov.co/dataset/parque-bogota-d-c> (Consultado el 13 de julio de 2023) [↑](#footnote-ref-3)
3. Alcaldía de Bogotá (2023). Ciclorruta Bogotá D.C..Datos abiertos. Disponible en: <https://datosabiertos.bogota.gov.co/dataset/cicloruta-bogota-d-c> (Consultado el 13 de julio de 2023) [↑](#footnote-ref-4)