Irina Andrea Vélez López – Código:

Miguel Angel Victoria Simbaqueva – Código:

Daniel Casas Bautista – Código: **202120803**

Lucia Fillippo – Código: **202213187**

**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 contribuyen a explicar el precio de la propiedad en cuestión. Cada ciudad tiene sus propias características distintivas que influyen en la fijación de precios por parte de los vendedores; 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 vendedores y compradores una comprensión más precisa de cómo influyen estos factores en las transacciones de bienes raíces, facilitando así las decisiones futuras de compra o venta de propiedades. Con objeto de 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 y datos del DANE sobre XXX. Para esto fue necesario realizar una limpieza de las bases de datos para lograr tener bases de datos de muestra y entrenamiento para evaluar los distintos modelos de predicción y elegir el adecuado.

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 sutability 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 encuentra:

**Ubicación:** Esta variable puede definir el precio de la vivienda, pues, en caso de estar más cercano al centro de la ciudad o a una zona empresarial, puede afectar el precio de la vivienda. En la base de datos se trata de una variable que da las coordenadas geométricas.

(…) aquí describir algunas otras

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

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