AVANCE PROYECTO ANÁLISIS DE DATOS

JAVIER CAMILO TIQUE QUIMBAYA - 20221678024

JOSÉ LUIS HERNÁNDEZ CLAROS - 20221678036

En la entrega de este avance se puede evidenciar el desarrollo del análisis exploratorio y preprocesamiento a los datos para el dataset relacionado a la problemática de la compra y venta de vivienda, además de una idea de modelo machine learning de predicción como forma de uso de estos datos.

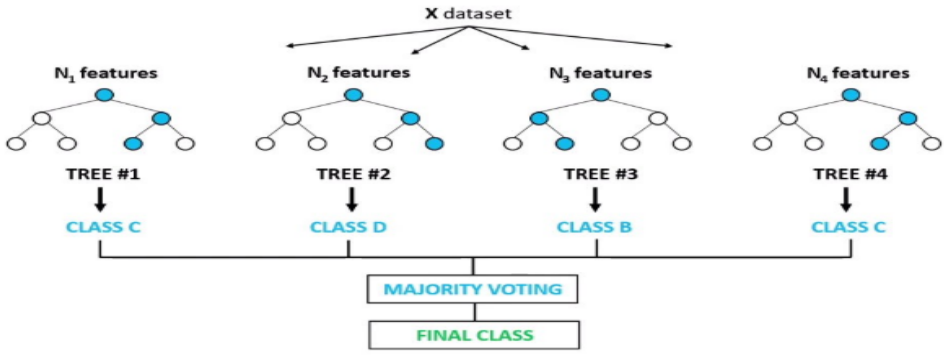
Se comenzó con una depuración inicial, eliminando características con valores redundantes y de poco aporte a la problemática, seguidamente se realizó una validación de datos faltantes y nulos y se modificaron, rellenaron, reemplazaron o eliminaron dependiendo de su valor de utilidad al análisis de datos. Posteriormente como forma de análisis exploratorio se comparó la cantidad de registros en base a características básicas de una vivienda (habitaciones, baños, estrato, etc) y se graficaron, para de esta forma poder deducir las condiciones más relevantes para tener en cuenta al momento de la predicción.

Posteriormente se realizaron transformaciones a las características que se utilizarán en la predicción, se modificaron los tipos de datos, se redondearon, se organizaron, especialmente en los precios y valores numéricos de forma que quedaron listos para realizar operaciones sobre ellos.

Para iniciar en la construcción del modelo machine learning se realizó nuevamente una selección de características, de forma que nos fuera posible reducir el número de entradas al modelo. Igualmente se espera poder implementar técnicas de selección de características intrínsecas como Random Forest Importance, para argumentar la selección de las más relevantes. Se va a utilizar el aprendizaje supervisado para desarrollar un modelo Machine Learning de predicción haciendo uso de las características de definición para una propiedad (habitaciones, baños, terrazas, etc) al igual que características categóricas (estrato, metros cuadrados, condiciones exteriores, etc) para aproximar una variable de salida precio, en este caso estaríamos hablando de una regresión multivariante.

Debido a la cantidad de características y la diferencia entre ellas se optó por elegir como técnica de machine learning un modelo Random Forest. Este modelo está formado por múltiples árboles de decisión individuales y cada uno de estos árboles es entrenado con una muestra ligeramente diferente de los datos de entrenamiento, es decir se predecirá el precio en base a características de definición (número de habitaciones, número de baños, existencia de terraza, altillo, etc) y características categóricas (estrato, cantidad de metros cuadrados, ubicación, etc) y de esta forma realizar predicciones sobre nuevas

observaciones, se combinan las predicciones de todos los árboles que conforman el modelo.



Se pretende hacer despliegue de la solución mediante https://mybinder.org/ el cual permite visualizar el notebook y además interactuar con él para ver su funcionamiento.