Análisis de Datos de Casas en Medellín

ALDAIR SIERRA FONTALVO,LEISER ANGARITA MELENDREZ,ELIMELE

March 14, 2025

1 Introducción

Este análisis se enfoca en un conjunto de datos que contiene información sobre casas en Medellín, extraída de un archivo CSV disponible en GitHub. El objetivo es procesar el conjunto de datos para entender su estructura y limpiar cualquier dato erróneo o faltante.

2 Carga de Datos

Para comenzar, cargamos el conjunto de datos desde un archivo CSV en GitHub utilizando la biblioteca `pandas` de Python. El archivo se encuentra en el siguiente enlace:

`https://raw.githubusercontent.com/jrudascas/ml-infotep/main/Datasets/houses_medell`

El siguiente código se utiliza para cargar el archivo:

```
import pandas as pd
import numpy as np

# Cargar el archivo CSV desde la URL raw
Url = 'https://raw.githubusercontent.com/jrudascas/ml-infotep/main/Datasets/houses
data = pd.read_csv(Url)

# Ver los primeros registros
data.head()
```

3 Inspección Inicial

Una vez cargado el conjunto de datos, revisamos las primeras filas para obtener una visión general de la estructura de los datos. Esto se logra utilizando el método `head()` de pandas:

```
data.head()
```

4 Limpieza de Datos

El siguiente paso en el proceso de análisis fue limpiar las columnas que contienen una gran cantidad de valores nulos. Se decidió eliminar aquellas columnas donde más del 90% de los datos eran nulos. Para ello, se utiliza el método `dropna()` de pandas, con un umbral del 90% de valores no nulos. El código utilizado es el siguiente:

```
# Calcular el umbral de nulos (90%)
threshold = len(data) * 0.9

# Eliminar las columnas que tienen más del 90% de datos nulos
data_cleaned = data.dropna(axis=1, thresh=threshold)

# Ver los primeros registros después de limpiar
data_cleaned.head()
```

Este paso permite asegurar que solo las columnas con una cantidad significativa de datos se mantengan en el conjunto de datos.

5 Conclusión

El proceso de carga, inspección y limpieza de los datos fue exitoso. Las columnas con más del 90% de valores nulos fueron eliminadas, lo que garantiza que el análisis posterior sea más preciso y no esté sesgado por datos faltantes. Los datos ahora están listos para cualquier análisis exploratorio o modelado predictivo que se desee realizar.