Normalización de Datos

Nayelis Jaimes, Juan Camilo Orozco, Diego Fernandez

March 15, 2025

1 Resumen

Este documento presenta un procedimiento para la normalización de datos en un DataFrame de Pandas, asegurando que todas las columnas numéricas estén escaladas entre 0 y 1. La normalización es un paso crucial en el preprocesamiento de datos para modelos de aprendizaje automático y análisis estadísticos. Se describe la metodología utilizada y se analizan los resultados obtenidos.

2 Introducción

En el análisis de datos, es común encontrarse con valores numéricos en diferentes escalas, lo que puede afectar la interpretación de los resultados y el desempeño de los modelos de aprendizaje automático. Para solucionar esto, se utiliza la normalización, que ajusta los valores de las columnas numéricas dentro de un rango estandarizado, generalmente entre 0 y 1. Este documento describe un enfoque para normalizar datos en un DataFrame de Pandas utilizando Python.

3 Objetivos

Los objetivos principales de este trabajo son:

- Leer y limpiar datos de un archivo CSV.
- Identificar si las columnas numéricas ya están normalizadas.
- Aplicar normalización a las columnas que lo requieran.
- Mostrar los datos procesados para su análisis posterior.

4 Metodología

La normalización se realiza mediante las siguientes funciones en Python:

4.1 Verificación de Normalización

Para comprobar si una columna está normalizada, se usa la función:

```
def is_normalized(column):
    return column.min() >= 0 and column.max() <= 1
```

Esta función verifica si los valores de una columna están dentro del rango [0,1].

4.2 Proceso de Normalización

Si una columna no está normalizada, se aplica la siguiente transformación:

```
def normalize_dataframe(df):
    for col in df.columns:
        if not is_normalized(df[col]):
            col_min = df[col].min()
            col_max = df[col].max()
            if col_max != col_min:
                df[col] = (df[col] - col_min) / (col_max - col_min)
            else:
                df[col] = 0
    return df
```

5 Resultados y Discusión

El código fue probado en un conjunto de datos de viviendas en Medellín. Se observaron los siguientes resultados:

- Columnas previamente normalizadas se mantuvieron sin cambios.
- Columnas con valores fuera del rango fueron correctamente escaladas.
- Se evitó la división por cero en columnas con valores constantes.

6 Conclusiones

A través de la normalización se organizó la forma de interpretar la información evitando sesgos a las hora de interpretarla. De esta manera logramos obtener datos estandarizados para proporcionar al aprendizaje de maquina entre otras ciencias. Por último esta tabla se convierte en una base para futuros estudios inmobiliarios en Medellín.