Dataset Cleaning and Missing Value Handling

Objective

The objective of this task is to clean the datasets, handle missing values, and prepare the data for machine learning.

Datasets Used

- House Prices Dataset
- Medical Appointment No Shows Dataset

Tools Used

- Python
- Pandas
- NumPy
- Matplotlib
- VS Code
- Jupyter Notebook

Dataset Analysis

House Prices Dataset

- The dataset contains information about houses such as price, area, bedrooms, bathrooms, and amenities.
- Numerical features include price, area, bedrooms, bathrooms, stories, and parking.
- Categorical features include mainroad, guestroom, basement, hotwaterheating, airconditioning, prefarea, and furnishingstatus.
- Binary features include mainroad, guestroom, basement, hotwaterheating, airconditioning, prefarea.

- No missing values were found in any column.
- The dataset is suitable for regression problems.

Medical Appointment No Shows Dataset

- The dataset contains information about patients and their appointment attendance.
- Numerical features include Age and SMS_received.
- Categorical features include Gender and Neighbourhood.
- Binary features include Scholarship, Hipertension, Diabetes, Alcoholism, Handcap, and No-show.
- Missing values were minimal; some invalid rows (e.g., negative ages) were removed.
- The target variable is No-show, which was converted to binary (0 → attended, 1 → missed).
- The dataset is suitable for classification problems.

Conclusion

This task helped in understanding how data cleaning and missing value handling are performed before applying machine learning models. Proper preprocessing improves data quality and model performance.