Page No.
Date

Devanshu Surana
PC-23, Panel C, Batch C1
1032210755

DEC Lab Assignment 2

Problem Statement : Use housing Dataset for Data Pre-processing. Apply Various data cleaning functions for following:

Handle missing values: Ignore, Defaults, Impute.
Handle duplicate: Identify, Remove smoothing noisy data, Resolve inconsistencies.

Objective

To clean data and make it noise free.
Prepare data for Analysis.

Theory:

What is Data Preprocessing?

→ Data preprocessing is a crucial step in data analysis and machine learning pipeline. It involves a series of operations and techniques applied to raw data to make it suitable for analysis or training machine learning models.

Here are some key steps and techniques involved in data preprocessing.

1. Data Cleaning
2. Data Transformation

Teacher's Sign.: _____

3. Data Reduction
4. Data Integration
5. Data formatting
6. Data Exploration
7. Handling Imbalanced data
8. Normalization
9. Data Splitting
10. Data Imputation.

Need of Data Pre-Processing

→ 1. Handling Missing data: Real world dataset often contain missing values due to various reasons such as sensor failures.

2. Removing duplicate data: Duplicate records can skew analysis result and mislead the model or user.

3. Data Transformation: Raw data often requires transf--ormation to make it suitable for analysis or modelling.

4. Data formatting: Ensure appropriate data for analysis / modelling.

5. Data Imputation: When missing values are present, data preprocessing techniques can impute or fill in those values.

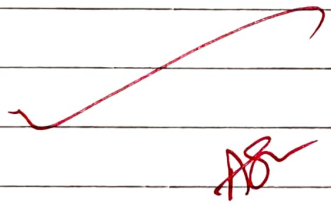
Page No.
Date

3. List of steps in Data Cleaning with example:

→ Data cleaning is process of correcting errors or inconsistency in data.

Steps:

- Removal of irrelevant or duplicate data
- fixing structural errors
- Dealing with missing data
- filtering out data outliers
- Validating data
- standardizing capitalization
- Converting data type
- Language translation



Teacher's Sign.: _____