Assignment No. 2

Problem Statement

For a given dataset perform Data Pre-Processing - Data Cleaning.
Apply various data cleaning functions to:
i) Handle missing values or null values (ignore, defaults, impute)
ii) Handle duplicates (identify, remove)

Objectives

1. To explore various Data Cleaning methods.
2. To explore the operations for handling missing data using Python.

Theory

1. Missing values in a dataset

Missing values occur when no data value is stored for a variable in an observation. This can happen for various reasons, such as data entry errors, non-responses in surveys, or system errors during data collection.

Types of Missing Data:

- MCAR (Missing Completely At Random): The missingness is unrelated to any data, observed or unobserved. Each instance of missing data is random; no pattern.

- MAR (Missing At Random): The missingness is related to the observed data but not to the missing data itself.

- MNAR (Missing Not At Random): The missingness is related to the value of the missing data itself.

2. Data Cleaning

Is the process of identifying and correcting (or removing) inaccuracies, inconsistencies, and errors in the dataset. This ensures the data is accurate, complete, and ready for analysis.

Page No.
Date :

Steps

- Identify missing data
- Handle missing data
  - Imputation
  - Removal
- Correcting inaccuracies
- Remove duplicates
- Standardise formats

3. Simple Imputer

Is a tool provided by the ~~Sci-kit~~ Scikit-learn library in Python that replaces missing values with a specific value or a constant.

The simple imputer fills in missing values by calculating a statistic (mean, median, or mode) based on the non-missing values in the column.

CONCLUSION

Data cleaning operations were performed on the given dataset .csv file using Python.

FAQs

1. Explain the advantages of data preprocessing.

- Improved data quality.

  Preprocessing enhances the quality of data by removing noise, correcting inconsistencies, and handling missing values, leading to more accurate and reliable results.

- Better model performance.

  ML models can learn more effectively, leading to better performance, higher accuracy, and more robust predictions.

Page No.
Date :

- Reduced complexity
  Data is simplified, making it easier to analyse by reducing dimensionality, normalizing values, and encoding categorical variables.

- Efficiency in analysis
  Preprocessed data reduces the time and computational resources needed for data analysis, as it removes irrelevant or redundant information.

- Enhanced interpretability
  By scaling, normalisation, and encoding, the data is made more interpretable, allowing analysts to better understand the underlying patterns.

2. Explain various data cleaning techniques.

i. Handling Missing Values
- Imputation: Replace missing values with statistical measures or with values predicted by ML models.
- Deletion: Remove rows or columns with missing values if the missing data is minimal or insignificant.
- Forward / Backward Fill: Missing values can be filled by the previous or next available value.

ii. Removing Duplicates
Identify and remove duplicate rows to prevent bias.

iii. Standardising Data
Convert data into a consistent format.

iv. Handling Outliers
Detect and handle by capping them, or transforming them.

Page No.
Date :

3. What is an outlier?
Is a data point that significantly differs from the other data points in a dataset. They can be caused by variability in the data or by measurement errors.

Types
- Univariate Outliers - Outliers detected within a single variable.
- Multivariate - Outliers detected within a combination of variables.

4. Give the importance of handling missing data.
- Maintaining data integrity.
- Avoiding bias.
- Improving model performance.
- Ensuring consistency.
- Better interpretation.
- Minimizing data loss.