DCET, SoCSE, MITWPU
(T.Y.B.Tech.) Semester-5, AY 2024-25
DE [CET3005B]: Lab Assignment No. 03

Problem Statement: : Performing various tasks for Data Preprocessing with data transformations as using python.

Objectives:

1) To explore various data transformation techniques.
2) To explore the operations on a dataset file using data integration, normalization and data reduction, while handling data redundancy using Python libraries.

Theory: *(students should elaborate on following points)*

- Data Correlation
- Pearson Correlation (pandas Implementation, NumPy and SciPy Implementation)
- Visualization of Correlation (Heatmaps of Correlation Matrices)
- Feature Normalization and their techniques
- Feature Selection

Implementation tasks to be performed:

1. Load/read and display the dataset (with missing values)
2. Identify duplicates
3. Handle data redundancy (drop the records with duplicates)
4. To perform correlation analysis (Pearson's ratio)
5. Display the heat map
6. Data visualization – techniques (e.g. – histogram plot, etc.)
7. Perform min-max scaling
8. Perform ZScore scaling
9. Perform data smoothing using binning method
10. Perform feature reduction with selection to select most significant features.

Execution: *(students should elaborate on following points)*

- Program implementation in python
- Output study

Conclusion:

FAQs:

Q1. Compare the two normalization techniques
Q2. State the significance of Data Scaling.
Q3. Explain the process of Feature Extraction
Q4. Benefits and Techniques of Binning in Python
Q5. What is Data leakage, how to avoid any data leakage during the model testing process.
Q6. Which technique we should use: Normalization or Standardization?
Q7. What are the benefits of Correlation Analysis?
Q8. What are the different kinds of correlation analysis? Discuss their strength and weakness.
Q9. What are the factors that affect a Correlation Analysis?
Q10. Write a short note on

a. The correlation coefficient
b. The p-value