ASDS 6302

Assignment-1

Objective: To perform data analysis on the Breast Cancer dataset using Python, including data exploration, sampling, normalization, PCA, and visualization.

Submission Format: .py or .ipynb containing code snippets, explanations, and visualizations.

Submission Deadline: Late submissions will incur a penalty of 10%. (Due Wednesday Sep, 12th , 9 pm)

QUESTIONS

1. Import the Breast Cancer dataset from the sklearn module for this assignment.

2. Display the first and last 10 rows of the dataset. Additionally, check the data types of all features and identify any missing values. Explain your findings in detail.

3. Apply three sampling methods—Random Sampling, Stratified Sampling, and Systematic Sampling—to generate 150 random samples from the Breast Cancer dataset. Explain the steps and rationale behind each sampling method.

4. Remove the target variable from the dataset and create a correlation matrix. Identify and explain the top 3 pairs of features with the highest correlation. Discuss any potential implications of these correlations for further analysis.

5. Normalize the dataset using three different normalization methods, such as Standardization, Min-Max Scaling, and Robust Scaling. Provide the normalized datasets and compare their distributions using appropriate visualizations. Discuss the advantages and disadvantages of each normalization method.

6. Apply Principal Component Analysis (PCA) to the standardized dataset. Share your conclusions based on the analysis, including the cumulative explained variance. Discuss how many principal components are needed to explain at least 85% of the variance.

7. Create a visual representation illustrating the explained variance by each principal component. Include a title, x-axis label, and y-axis label in your graph for clarity. Additionally, provide a scree plot and discuss its significance.

8. Apply another dimensionality reduction technique, such as t-SNE or LDA, and repeat questions 6 and 7. Compare the results with PCA, discussing the strengths and weaknesses of each technique.