AIN429 Data Mining Laboratory

Assignment 2: Comprehensive Data Preprocessing

Due Date: 05/11/2024 23:59

Aim of the Experiment

In this assignment, you will explore data preprocessing techniques, applying cleaning, scaling, and exploratory data analysis to real-world datasets. The goal is to transform raw data into a usable format and extract insights. You will document your process using comments and markdown cells. The assignment should be implemented as a single Jupyter Notebook.

Part 1: Dataset Import and Summary (10 Marks)

Tasks:

1. **Dataset Loading**: Load the provided datasets and print basic information (e.g., shape, data types, and missing values).

2. **Summary Statistics**: Generate summary statistics for all features (e.g., mean, median, standard deviation) and present an overview of the datasets' structure.

3. **Visualization:** Create visualizations (e.g., histograms, box plots) to show the distribution of key features.

4. **Initial Data Observations**: Briefly describe any initial insights or issues observed in the raw data (e.g., incomplete data, outliers, inconsistencies).

Part 2: Data Integration (30 Marks)

Tasks:

1. **Secondary Dataset Integration**:

    - Integrate the provided secondary dataset with the main dataset. Handle discrepancies such as different column orders, missing features, or mismatched data types.
    - Align the columns and ensure the datasets are properly merged. Resolve missing or incomplete columns by filling with default values, merging with placeholders, or dropping irrelevant columns.

2. **Post-Integration Check**:

    - Check for any further missing values or inconsistencies after the integration and document what adjustments were made.

Part 3: Data Preprocessing (30 Marks)

Tasks:

1. **Data Cleaning**: After integration, handle missing values at a feature level using techniques like filling with mean/median, predictive modeling, or dropping rows/columns with excessive missing data.
2. **Scaling**:
    - Apply at least one scaling method (e.g., Min-Max Scaling or Z-score Standardization) to the dataset.
    - Explain why this scaling method is suitable for the data and how it affects the feature distributions.
3. **Preprocessing Summary**: Summarize the preprocessing steps taken and compare the dataset before and after cleaning and scaling.

Part 4: Data Visualization and Analysis (30 Marks)

Tasks:

1. **Feature Distributions**:
    - Use visualizations to show the distribution of key features before and after preprocessing.
    - Comment on the general distribution and any significant changes caused by scaling.
2. **Correlation Analysis**:
    - Create a correlation matrix or heatmap to visualize relationships between features.
    - Interpret the correlation results, highlighting any strong relationships or patterns.
3. **Principal Component Analysis (PCA)**:
    - Apply PCA to the preprocessed dataset to reduce its dimensionality.
    - Visualize and discuss the impact of PCA on the dataset, including how much variance is captured by the principal components.
4. **Data Insights**: Discuss any meaningful insights gained from the analysis, such as trends, outliers, or correlations that could be useful for further machine learning or analysis tasks.

Report

Write concise explanations (in markdown cells) summarizing your findings, preprocessing steps, and analysis results. The report should be done alongside the previous tasks and should include:

1. A brief introduction to the datasets.
2. Key findings from the preprocessing steps.
3. Insights gained from the visualizations and analyses.
4. Any challenges faced and how you addressed them.

Grading Breakdown

- **Dataset Import and Summary**: 10%
- **Data Integration**: 30%
- **Preprocessing**: 30%
- **Visualization and Analysis**: 30%

Submission Requirements

- **Format**: Submit your assignment as `b<studentID.zip>` containing:
    - A single Jupyter Notebook (`assignment_2.ipynb`) with all code, visualizations, and markdown explanations.
- **Late Penalty**: First day %10, second day %20 penalty for late submissions.
- **Submission Method**: Use https://submit.cs.hacettepe.edu.tr.
    - **Not getting 1 from the submit system will lose you points!**
- **Important** The assignment must be original, INDIVIDUAL work. Duplicate or very similar assignments are both going to be punished. General discussion of the problem is allowed, but DO NOT SHARE answers, algorithms, or source codes.
- **Questions** You can ask your questions through the course's Piazza group and you are supposed to be aware of everything discussed in the group.

Helpful Links

Some basic tutorials for data preprocessing using Python:

- https://www.v7labs.com/blog/data-preprocessing-guide
- https://www.javatpoint.com/data-preprocessing-machine-learning
- https://www.tutorialspoint.com/machine_learning_with_python/machine_learning_with_python_data_preprocessing_analysis_visualization.htm
- https://www.analyticsvidhya.com/blog/2021/08/data-preprocessing-in-data-mining-a-hands-on-guide/

Notebooks for ML

- https://github.com/krasserm/machine-learning-notebooks
- https://www.kdnuggets.com/2016/04/top-10-ipython-nb-tutorials.html