Assignment 3: Data Preprocessing, PCA, and Modeling

DSCI 6601: Data Science

Submission Instructions

In this assignment, you will explore various data pre-processing techniques followed by applying regression or classification models to make predictions. The assignment will evaluate how data pre-processing and dimensionality reduction using PCA affects the performance of the models. The goal is to build a Decision Tree model for either classification or regression, depending on the nature of your dataset. You will evaluate the model's performance and analyze how the features contribute to the predictions. Submit your Jupyter Notebook file and datasets, including all code explanations and reflections. Include a final reflection discussing challenges and lessons learned., and ensure to include both your Python code and the dataset in your submission.

Assignment Tasks

Based on your chosen dataset:

1. **Data Selection:**
   - Select a real-world dataset with at least 15 features and at least 100 samples. Optionally, you may use two different datasets for data integration (bonus points).
2. **Data Preprocessing:**
   - Handle missing values using imputation methods and explain the impact on the dataset. Encode categorical features. Optionally, integrate two datasets and explain the steps taken. Discuss steps such as handling missing values, encoding categorical features, and scaling features if necessary.
   - If there are no missing values in your dataset, then replace S percentage of your data with NaN values and then use KNN to impute the missing values. Justify the effect of imputation as S changes.
   - Provide the code for each preprocessing step and explain why these steps are important for decision trees.

3. **Exploratory Analysis and PCA:**

    - Perform visualizations like histograms, box plots, and correlation matrices. Apply PCA to reduce dimensionality. Discuss the effect of PCA and visualize the explained variance. Split the dataset into training and testing sets. Train two different models (regression or classification depending on your dataset) on the original and PCA-reduced datasets. Evaluate the models using metrics like MSE, R-squared, accuracy, precision, and F1-score. Compare performance on original vs PCA-reduced datasets.
    - Determine the optimal number of principal components and justify the optimal number via experiments in your notebook.
    - Include the code and interpretation of the results.

4. **Feature Importance:**

    - Decision Trees provide feature importance scores that help understand which features contribute the most to the predictions.
    - Extract and visualize the feature importance for the dataset.
    - Discuss how you would use this information to either refine the model or explain the results to a non-technical audience.
    - Provide code and an interpretation of the feature importance scores.

Submission Instructions

Your submission should include the following:

- Comprehensive responses to all the tasks outlined above, accompanied by the Python code and outputs from your Jupyter Notebook.
- Clear and concise explanations of each step taken, with well-documented text and comments included in your Jupyter Notebook for every step.
- A thorough analysis of the results within the Jupyter Notebook.
- The dataset(s) used for the assignment.