ASSIGNMENT

Title: Feature Scaling, Selection, and Constant Column Removal in Machine Learning with the Breast Cancer Dataset

Objective: In this assignment, you will learn how to preprocess and select features in machine learning using Python. You will practice techniques such as feature scaling, feature selection using correlation analysis, and removal of constant columns with the breast cancer dataset from scikit-learn.

Tasks:

1. Import the necessary libraries: Start by importing the required libraries, including pandas, numpy, matplotlib, scikit-learn, and seaborn.

2. Load the dataset: Load the breast cancer dataset into a pandas DataFrame using the load_breast_cancer() function from scikit-learn (sklearn.datasets()).

3. Remove constant columns: Remove any columns in the dataset that have constant values.

4. Encode categorical features: If the dataset contains any categorical features, encode them using one-hot encoding or label encoding.

5. Feature scaling: Scale the numerical features using scikit-learn's StandardScaler or MinMaxScaler function.

6. Correlation analysis: Use seaborn's heatmap() function to plot a heatmap of the correlation matrix between the features. Identify highly correlated features and remove one of them.

7. Feature selection: Use scikit-learn's SelectKBest or SelectPercentile function to select the k best or top percentile features based on their correlation with the target variable.

8. Bonus: Choose a different dataset and perform feature scaling, selection, and constant column removal using different techniques. Be creative!

Submission:

Submit a Jupyter Notebook file (.ipynb) containing your code, a brief explanation of your thought process for each task, and any necessary comments for clarity. Make sure to test your code and provide examples of the output for each task.