Assignment 4

Use the GSE45827 dataset for primary invasive breast cancer. You are required to do the following using Scikit package:

1. Execute the prescribed data preprocessing steps previously implemented in Assignment 3 to prepare the dataset for analysis.

2. Implement Recursive Feature Elimination (RFE) with Support Vector Machine (SVM) as the estimator. Recursively select the most relevant features, and as part of this process, eliminate 50% of the least informative data.

3. Perform Grid Search with a linear SVM using the selected features, exploring different values of the regularization parameter C (e.g., C = 1, 10, 100), and discuss your results.

Bonus (over total assignments grades-1 Mark)

You are part of a medical research team focusing on identifying rare genetic disorders. The dataset represents genetic features extracted from patient samples. Due to the rarity of these disorders, the dataset is imbalanced, with the majority of patients being unaffected.

1. Read the biomedical dataset 'biomedical\_dataset.csv'.

2. Print and create a scatter plot to visualize the class distribution.

3. Implement the Synthetic Minority Over-sampling Technique (SMOTE) to oversample the minority class and balance the dataset.

4. Print and create a scatter plot to visualize the class distribution after applying SMOTE.

5. Compare 2 and 4 to observe the changes in class distribution.

6. Implement an SVM classifier from scikit-learn for balanced data and imbalanced data through

- RepeatedStratifiedKFold cross-validation with 10 splits and 3 repeats.

- SMOTE oversampling to handle imbalanced data.

Then compare the results.