REPORT - LAB4

B20BB047

1. Import the relevant libraries of python
    a. Matplotlib for plotting graphs
    b. Pandas for data set manipulation
    c. Numpy for vector algebra

2. Look for any Null values in data

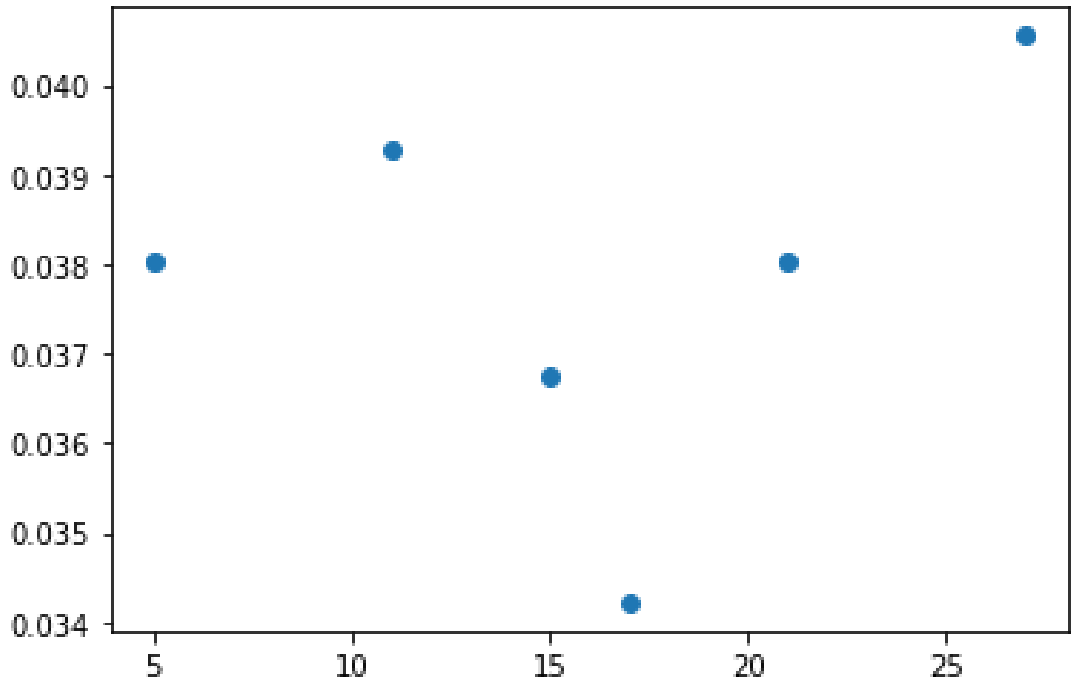
3. None found, but many values = 0 which logically do not make sense (Eg. BP = 0 or SkinThickness = 0)

4. Impute these incorrect values with mean

5. Apply standard scaling to prevent any bias towards features with larger values or larger ranges.

6. Write a function ***euclidean_dist(v1,v2)*** to find the euclidean distance between any two points/instances of data.

7. Write a function ***knn(X,u,k)***
    a. This function runs through k = [5,7,11,15,17,21]
    b. For each k, it enters another loop which iterates over the testing or validation set. For each row ***u*** of this set:
        i. Calculate euclidean distances with all instances of training set
        ii. Find the smallest k distances
        iii. From the frequency of each class, assign the corresponding class label.
    c. Calculate the confusion matrix and accuracy for all values of k.
    d. Plot error rate vs k to obtain the optimum k.



Optimum k obtained = 17