Question - 4

a) Since the data is only a binary classification, it maybe is almost linearly seperable maybe with a few outliers.

Method used = Soft Margin SVM
Library = svm.SVC
Kernel used = Linear

Accuracy = 0.9787735849056604
Number of Support Vectors = 28
Support Vectors from class 1 = 14
Support Vectors from class 2 = 14

Since the SVM tries to maximize the margin between two almost linearly seperable data, the output line we get from SVM is the most optimum one that can classify two classes better. I think this maybe is the reason why the accuracy on test data is so high

b) We took different sets of data of length [50,100 , 200 ,800]

Kernel used = Linear

But the Accuracy remained constant by taking first 50, 100, 200, 800 datasets.

This maybe because the support vectors for defining the boundary are the same points in all these datasets. So, everytime the accuracy on the test dataset is same, it maybe that the decision boundary did not change when we train with those datapoints.

c) Proofs and graphs are given in the Jupyter Notebook Submitted

d) For training the SVM we took RBF kernel with different biases (C) in the dual of SVM.
The least training error (highest training accuracy) occurred at C = 10**6
The least Test error (highest test accuracy) occurred at C = 100

The train accuracy is going on increasing as we increase the value of our bias. This maybe because the curve has overfitted the data and hence the test accuracy went on decreasing as we go from C = 100 to C = 10**6