Data Mining CSE-5334-001

Assignment 2-Report

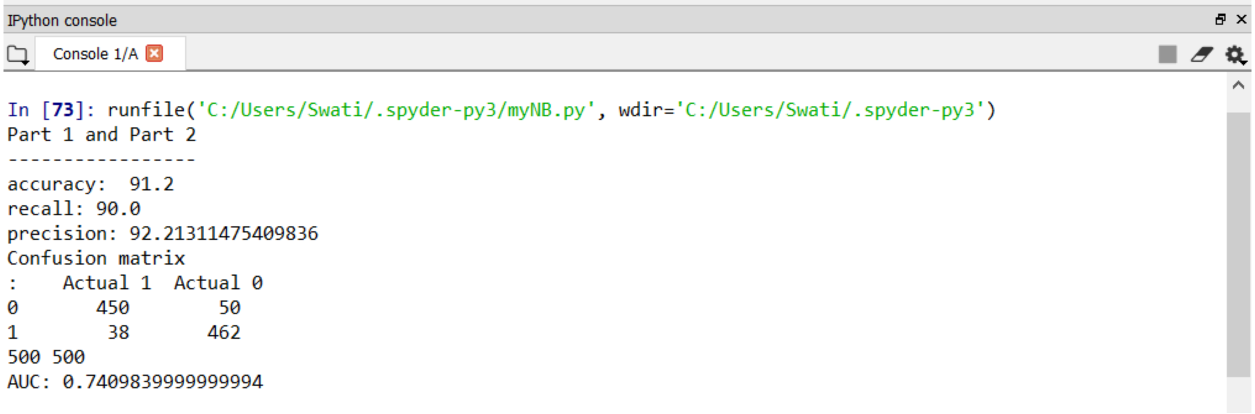
Name: Swati Sriram Mani

UTA Id: 1001648136

Algorithm:

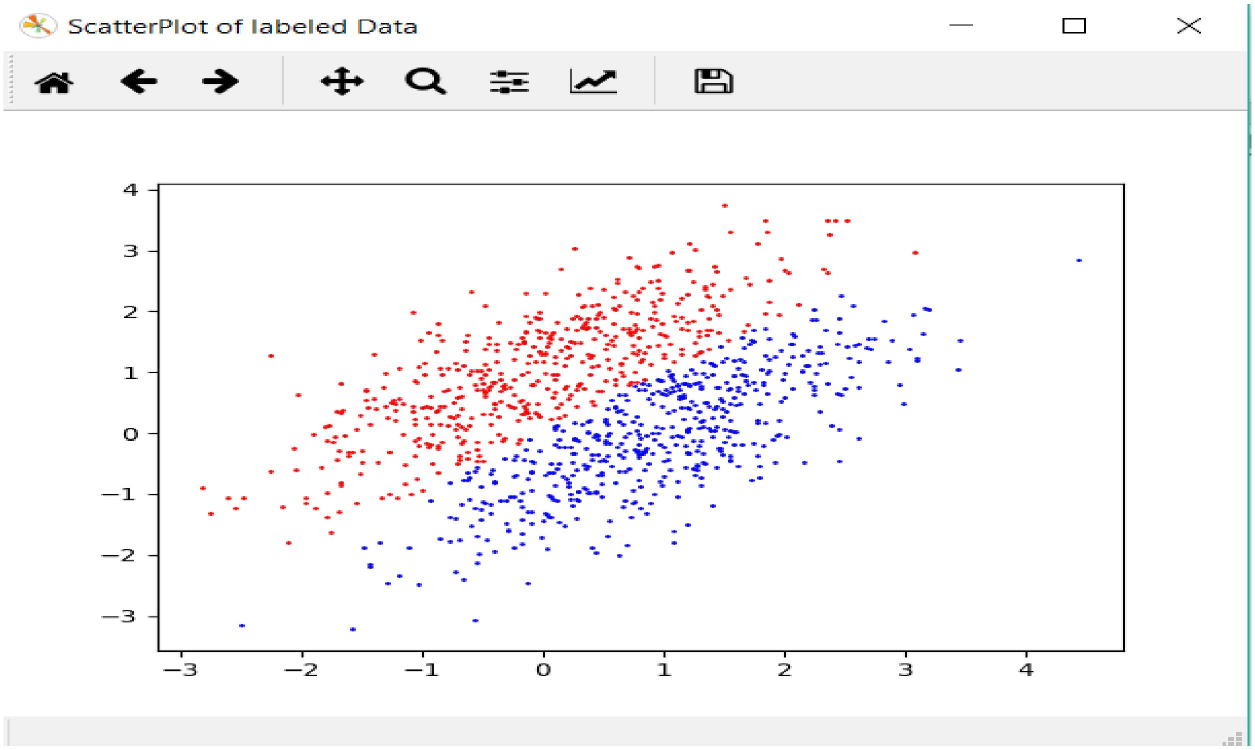
1. Initially,we have to generate the training data(X) and the test data(X_test)
2. We need to create zeros and ones for assigning the labels for both the training (Y) and test(Y_test)
3. We create a dataframe wherein the data as well as the labels as present so that we can split the data according to the labels and then do further calculations
4. Training: we have to calculate the mean and the standard deviation of each of the data in the labels 0 and 1.
5. Testing: we have to find the gaussian pdf using the data point, the corresponding mean, standard deviation and multiply with the corresponding priori value inorder to get the posterior value.
6. Next we need to predict, so we compare the class 0 probability and class 1 probability and assign the prediction accordingly to class 0 or class 1.
7. We need to calculate the values of accuracy, recall,precision and the confusion matrix which is done by finding the true positive values,true negative values,false negative and true negative values. Then the confusion matrix is displayed on the console.
8. To calculate the ROC, we need to find the actual positive and actual negative values from the actual column in your test frame and then find the true positive rate and false positive rates which is then plotted.
9. To calculate the Area Under Roc, we use the formula: ((tpr)*(fpr-fpr_previous))

Question 2:



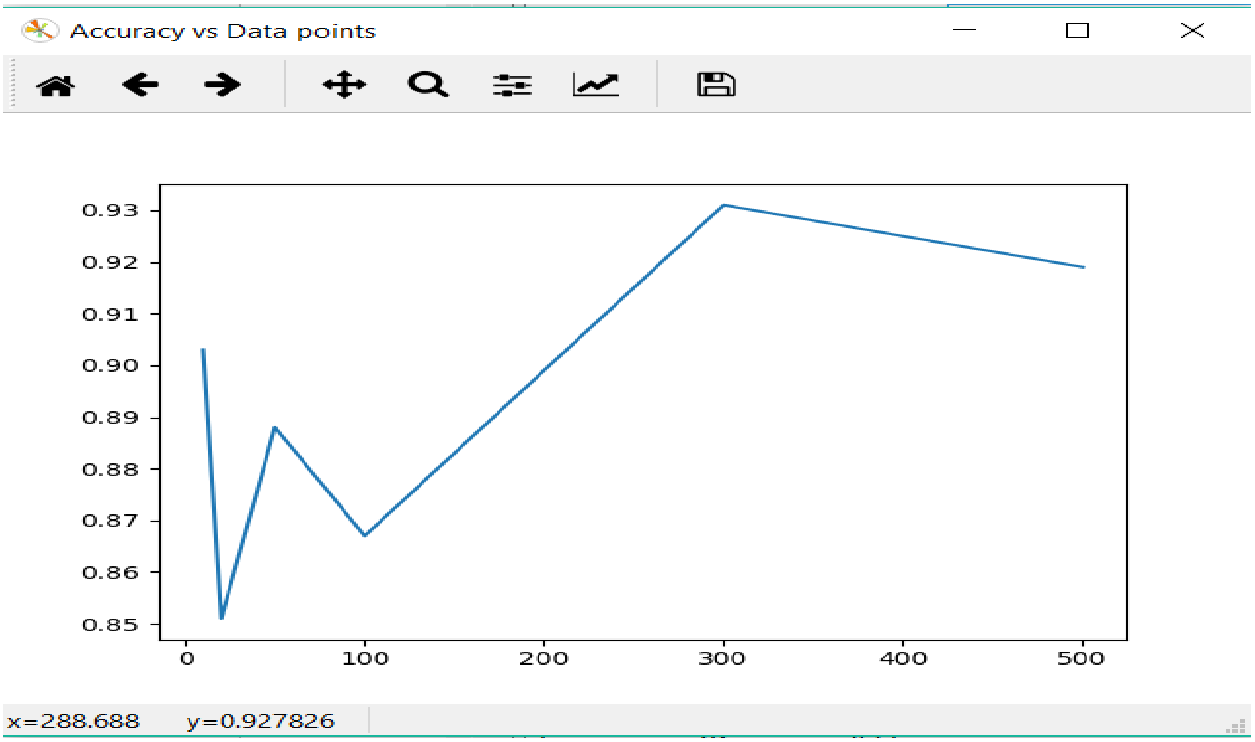
```
In [73]: runfile('C:/Users/Swati/.spyder-py3/myNB.py', wdir='C:/Users/Swati/.spyder-py3')
Part 1 and Part 2
-----------------
accuracy:  91.2
recall: 90.0
precision: 92.21311475409836
Confusion matrix
:    Actual 1  Actual 0
0        450        50
1         38       462
500 500
AUC: 0.7409839999999994
```

Scatter plot:



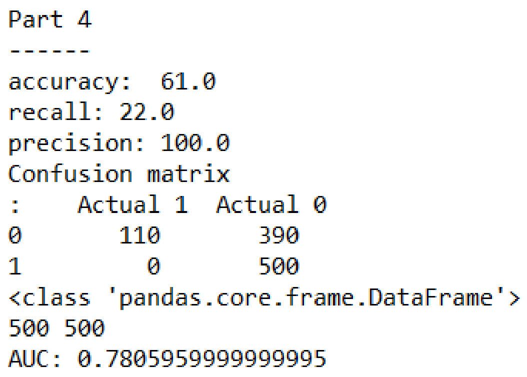
Question 3:

Accuracy Vs Data Points:



Observation: The graph above says that the accuracy initially increases and the decreases and then continuously decreases upto a certain point and then decreases again. As we keep training the accuracy keeps increases.

Question 4:

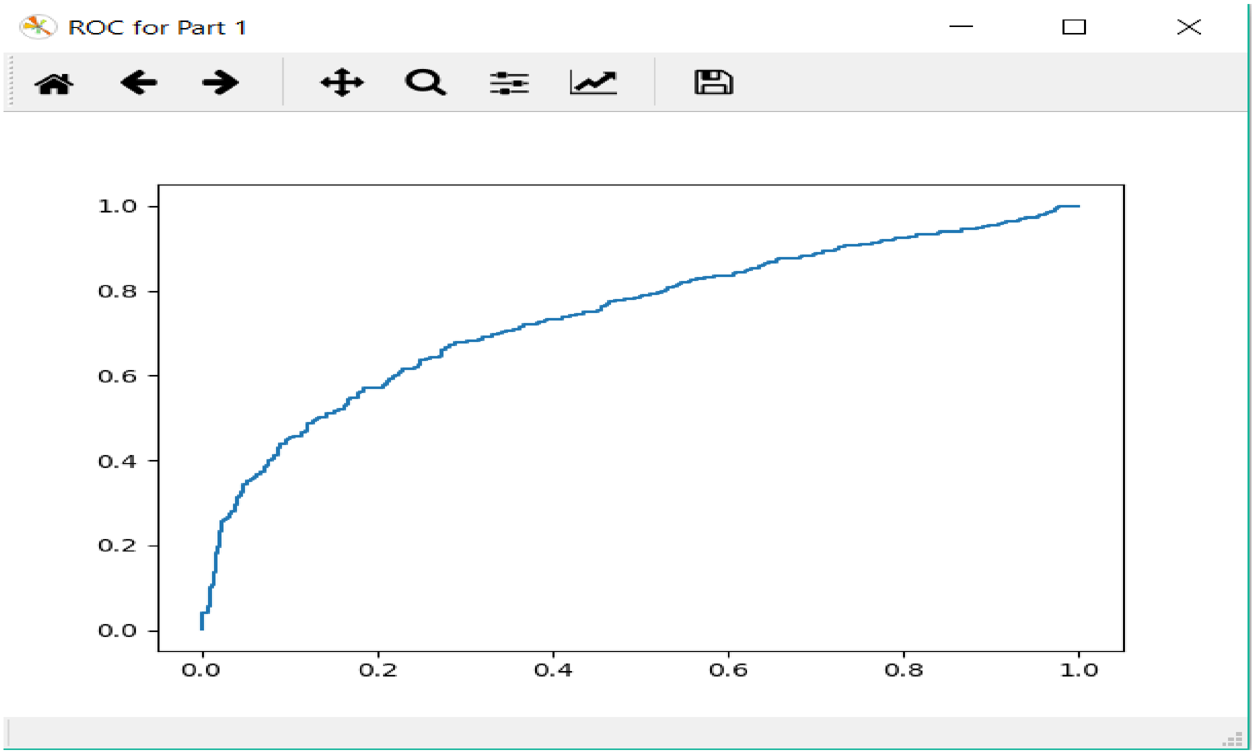


```
Part 4
------
accuracy:  61.0
recall: 22.0
precision: 100.0
Confusion matrix
:     Actual 1  Actual 0
0        110       390
1          0       500
<class 'pandas.core.frame.DataFrame'>
500 500
AUC: 0.7805959999999995
```

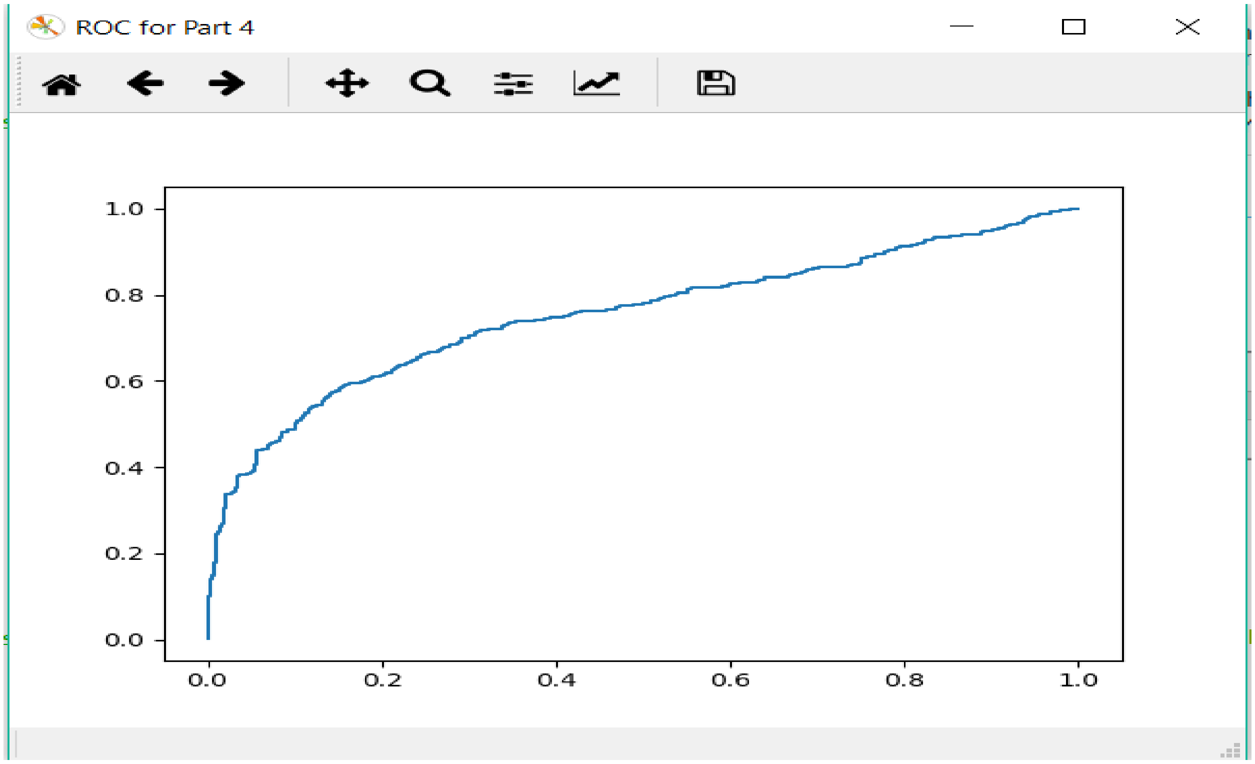
Observation: There is a difference in division in the labelling of the data thus the accuracy changes accordingly. The accuracy decreases as compared to the accuracy we have found in the part 1 of the problem.

Question 5:

Roc part 1:



ROC Part 4:


ROC for Part 4
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0