CSE 5334
Data Mining

Name: Sagar Sharma
UTA ID: 1001626958

Note: The folder consists of the file **hw2.py** which consists of the code for the problem.

Language Used: Python 3

Instructions to run the different parts of the problem.

Part 2

To run the part for predictions on test data set but without the Scatter Plot and ROC curve, type:

```
python hw2.py 2
```

To run the part for predictions on test data set without the ROC curve but with Scatter Plot, type:

```
python hw2.py 2 Scatter
```

To run the part for predictions on test data set without the Scatter Plot but with ROC curve, type:

```
python hw2.py 2 ROC
```

Part3

To run the part for predictions on test data set and plot the accuracy vs sample size curve, type:

```
python hw2.py 3
```

Part4

To run the part for predictions on test data set but without the ROC curve, type:

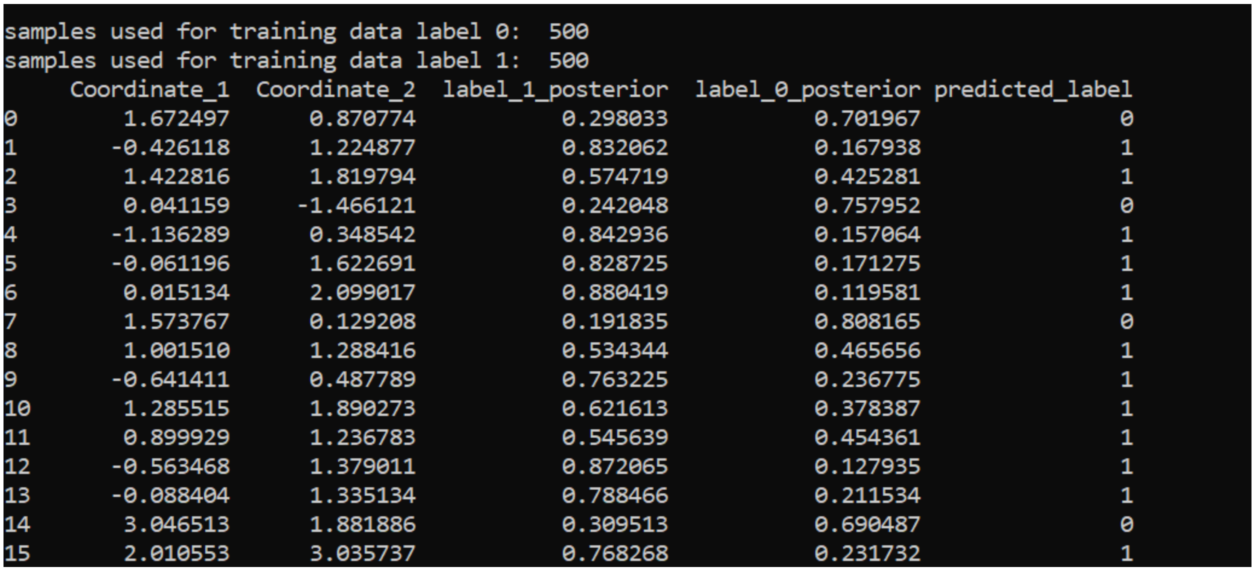
python hw2.py 4

To run the part for predictions on test data set and with the ROC curve, type:

python hw2.py 4 ROC

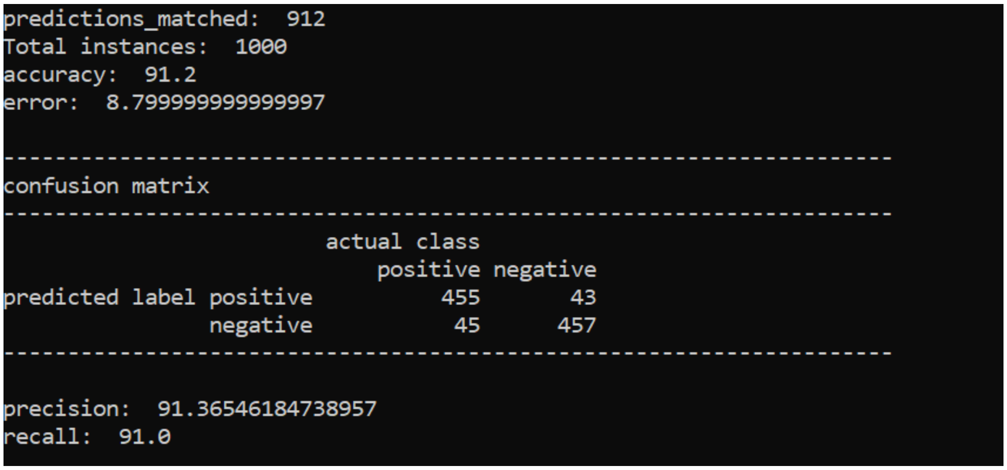
Task 2:

Snapshot of predictions made on the testing data set, snapshot showing for first 15 instances.



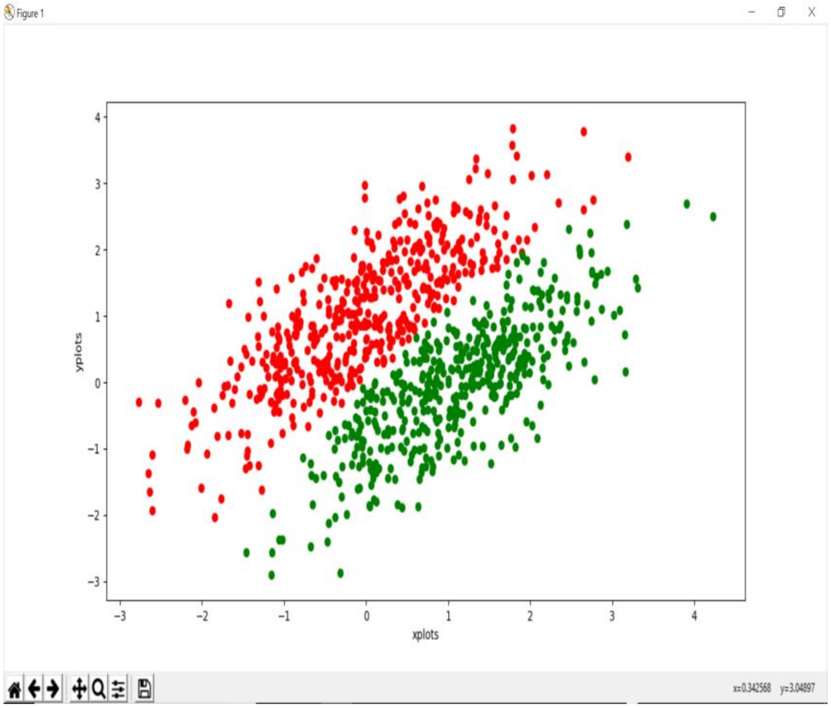
```
samples used for training data label 0:  500
samples used for training data label 1:  500
     Coordinate_1  Coordinate_2  label_1_posterior  label_0_posterior predicted_label
0        1.672497      0.870774           0.298033           0.701967               0
1       -0.426118      1.224877           0.832062           0.167938               1
2        1.422816      1.819794           0.574719           0.425281               1
3        0.041159     -1.466121           0.242048           0.757952               0
4       -1.136289      0.348542           0.842936           0.157064               1
5       -0.061196      1.622691           0.828725           0.171275               1
6        0.015134      2.099017           0.880419           0.119581               1
7        1.573767      0.129208           0.191835           0.808165               0
8        1.001510      1.288416           0.534344           0.465656               1
9       -0.641411      0.487789           0.763225           0.236775               1
10       1.285515      1.890273           0.621613           0.378387               1
11       0.899929      1.236783           0.545639           0.454361               1
12      -0.563468      1.379011           0.872065           0.127935               1
13      -0.088404      1.335134           0.788466           0.211534               1
14       3.046513      1.881886           0.309513           0.690487               0
15       2.010553      3.035737           0.768268           0.231732               1
```

Accuracy, Error, Precision, Recall, Confusion Matrix



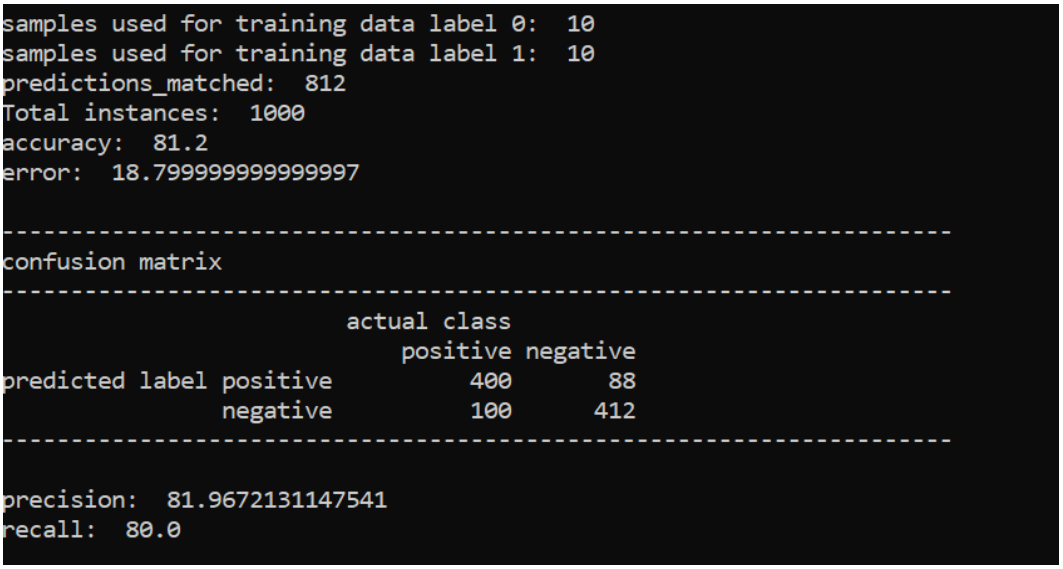
```
predictions_matched:  912
Total instances:  1000
accuracy:  91.2
error:  8.799999999999997

--------------------------------------------------------------------
confusion matrix
--------------------------------------------------------------------
                         actual class
                             positive negative
predicted label positive          455       43
                negative           45      457
--------------------------------------------------------------------

precision:  91.36546184738957
recall:  91.0
```

Scatter Plot



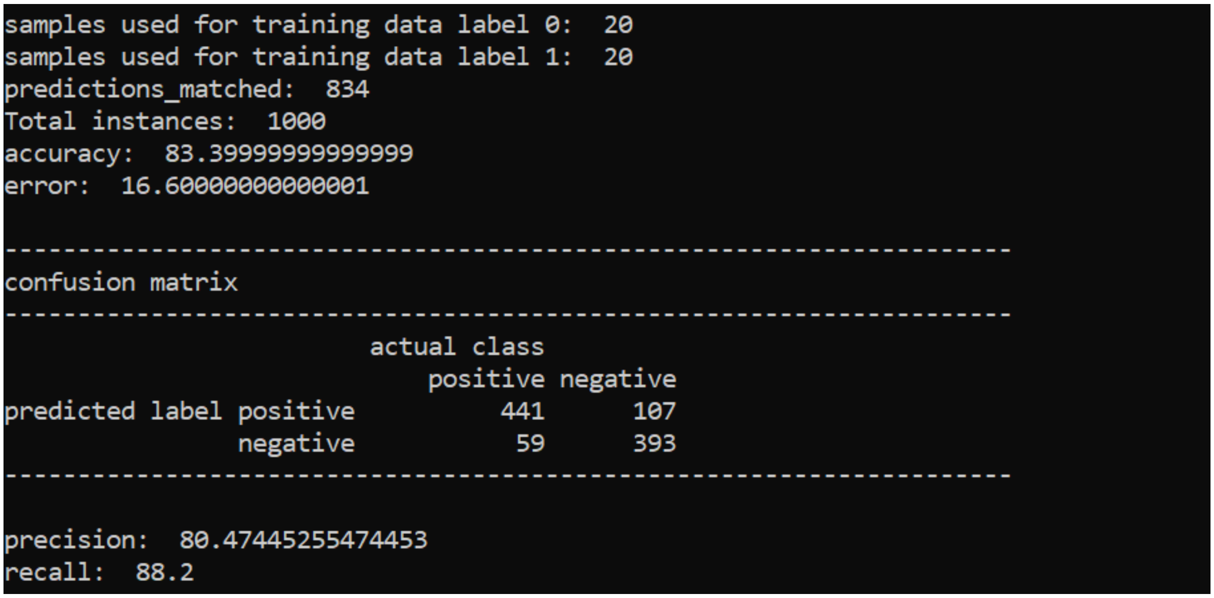
Task 3:

Performance Measures for Training Data set of sample size 10



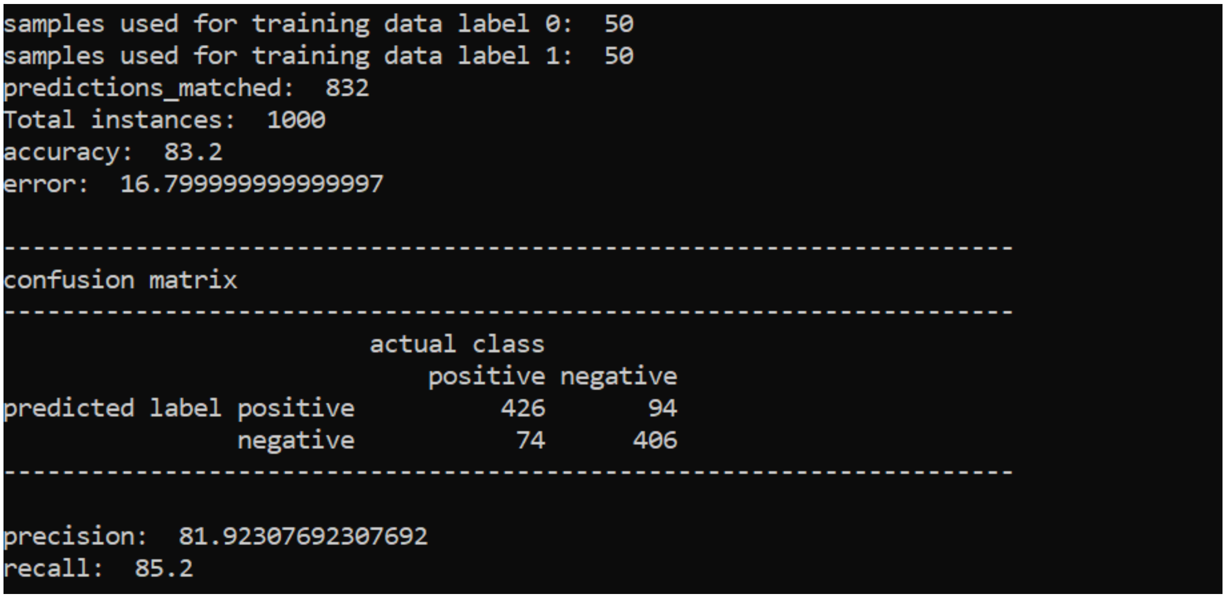
```
samples used for training data label 0:  10
samples used for training data label 1:  10
predictions_matched:  812
Total instances:  1000
accuracy:  81.2
error:  18.799999999999997

--------------------------------------------------------------------
confusion matrix
--------------------------------------------------------------------
                         actual class
                             positive negative
predicted label positive          400       88
                negative          100      412
--------------------------------------------------------------------

precision:  81.9672131147541
recall:  80.0
```

Performance Measures for Training Data set of sample size 20



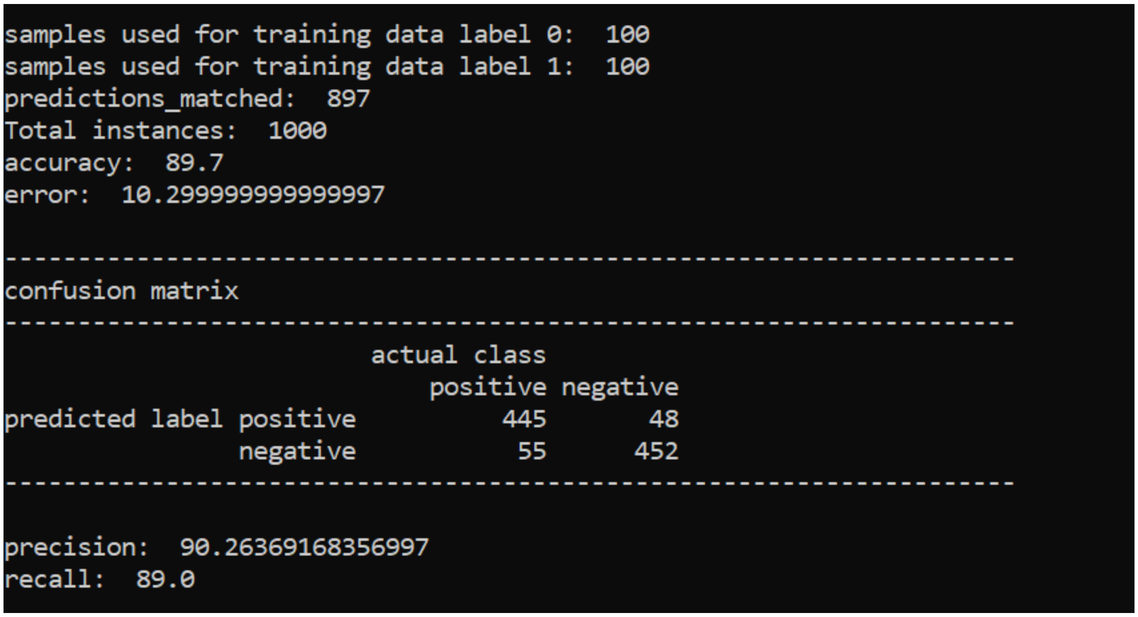
```
samples used for training data label 0:  20
samples used for training data label 1:  20
predictions_matched:  834
Total instances:  1000
accuracy:  83.39999999999999
error:  16.60000000000001

---------------------------------------------------------------------
confusion matrix
---------------------------------------------------------------------
                         actual class
                             positive negative
predicted label positive          441      107
                negative           59      393
---------------------------------------------------------------------

precision:  80.47445255474453
recall:  88.2
```

Performance Measures for Training Data set of sample size 50



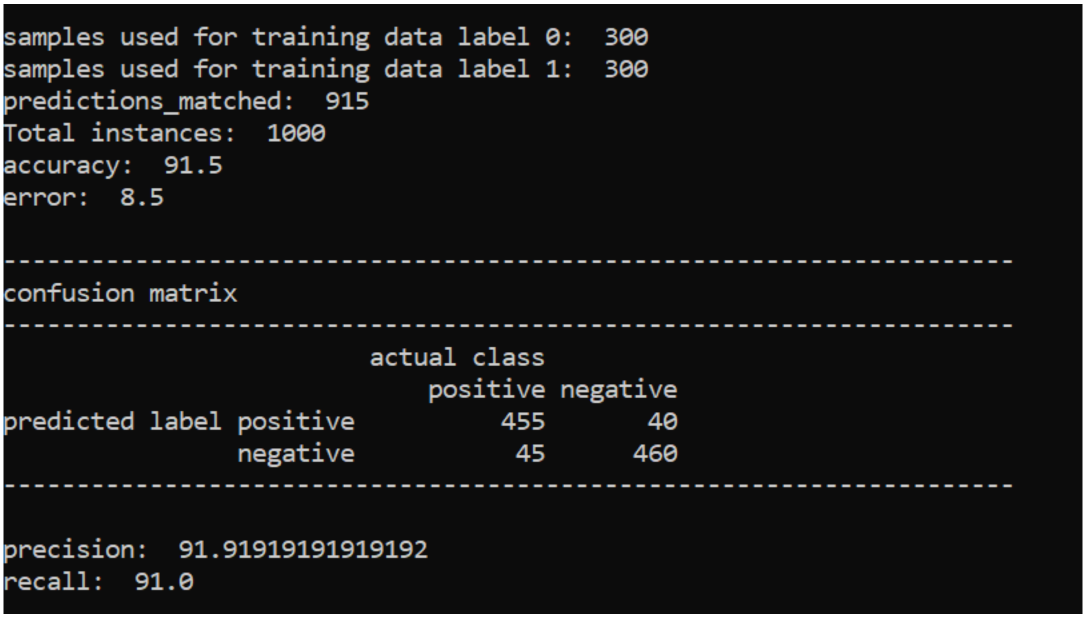
```
samples used for training data label 0:  50
samples used for training data label 1:  50
predictions_matched:  832
Total instances:  1000
accuracy:  83.2
error:  16.799999999999997

---------------------------------------------------------------------
confusion matrix
---------------------------------------------------------------------
                         actual class
                             positive negative
predicted label positive          426       94
                negative           74      406
---------------------------------------------------------------------

precision:  81.92307692307692
recall:  85.2
```

Performance Measures for Training Data set of sample size 100



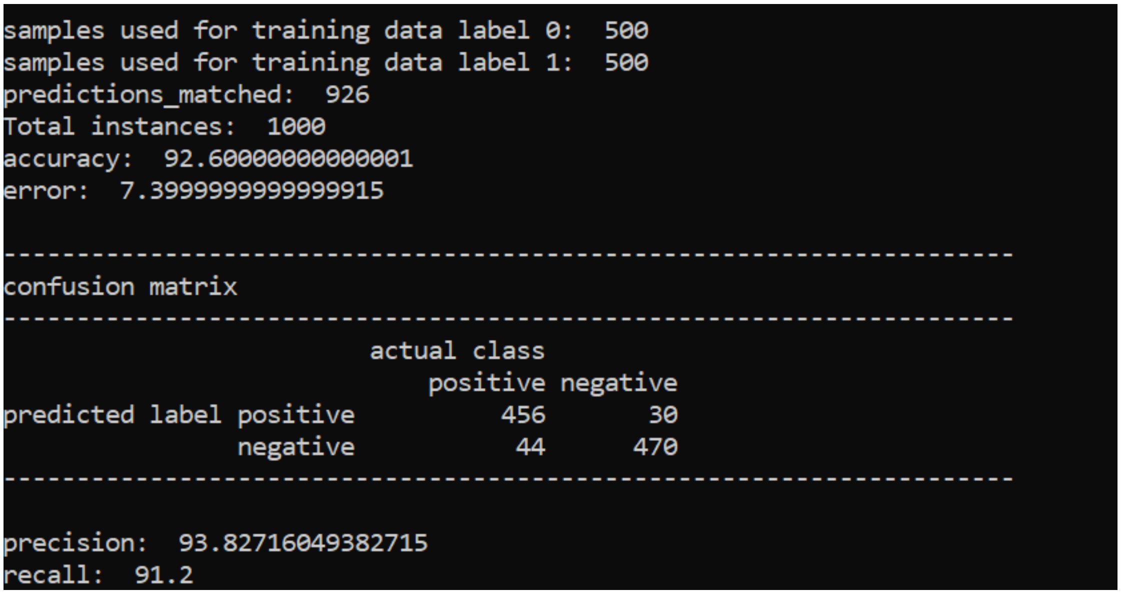
```
samples used for training data label 0:  100
samples used for training data label 1:  100
predictions_matched:  897
Total instances:  1000
accuracy:  89.7
error:  10.299999999999997

---------------------------------------------------------------------
confusion matrix
---------------------------------------------------------------------
                         actual class
                             positive negative
predicted label positive          445       48
                negative           55      452
---------------------------------------------------------------------

precision:  90.26369168356997
recall:  89.0
```

Performance Measures for Training Data set of sample size 300



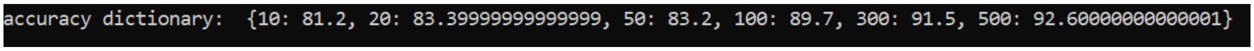
```
samples used for training data label 0:  300
samples used for training data label 1:  300
predictions_matched:  915
Total instances:  1000
accuracy:  91.5
error:  8.5

---------------------------------------------------------------------
confusion matrix
---------------------------------------------------------------------
                         actual class
                             positive negative
predicted label positive          455       40
                negative           45      460
---------------------------------------------------------------------

precision:  91.91919191919192
recall:  91.0
```

Performance Measures for Training Data set of sample size 500



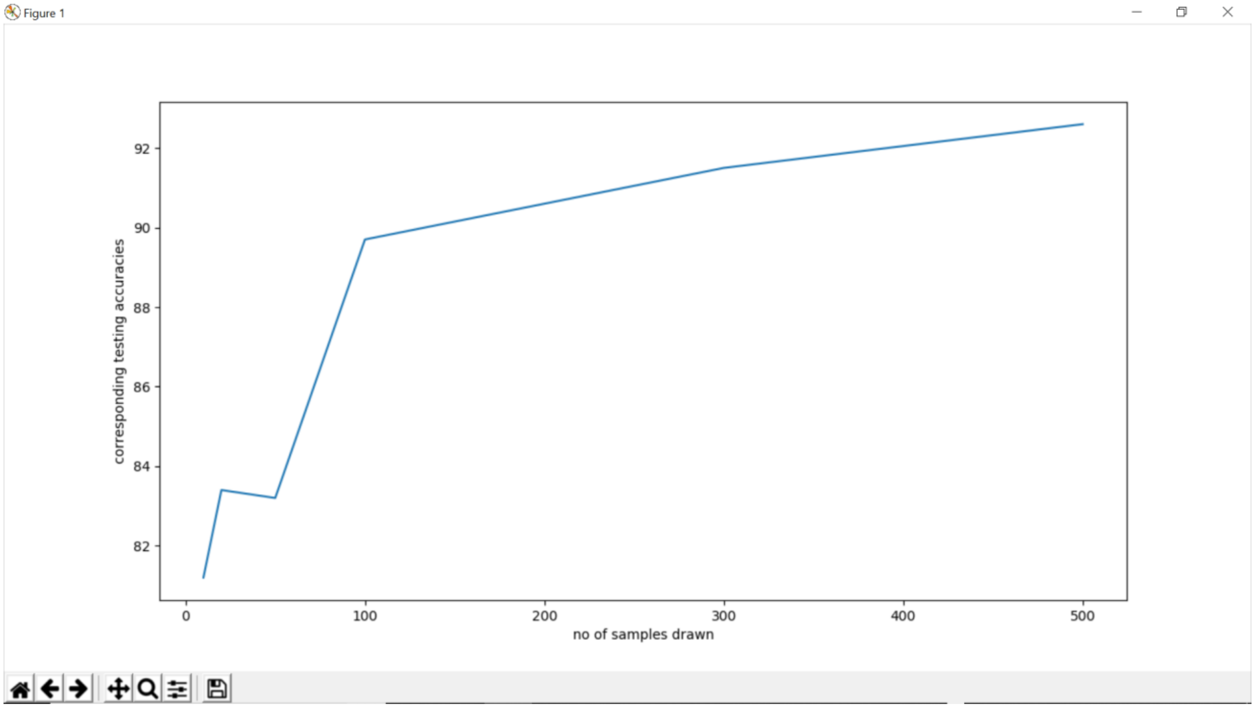
```
samples used for training data label 0:  500
samples used for training data label 1:  500
predictions_matched:  926
Total instances:  1000
accuracy:  92.60000000000001
error:  7.3999999999999915

---------------------------------------------------------------------
confusion matrix
---------------------------------------------------------------------
                         actual class
                             positive negative
predicted label positive          456        30
                negative           44       470
---------------------------------------------------------------------

precision:  93.82716049382715
recall:  91.2
```

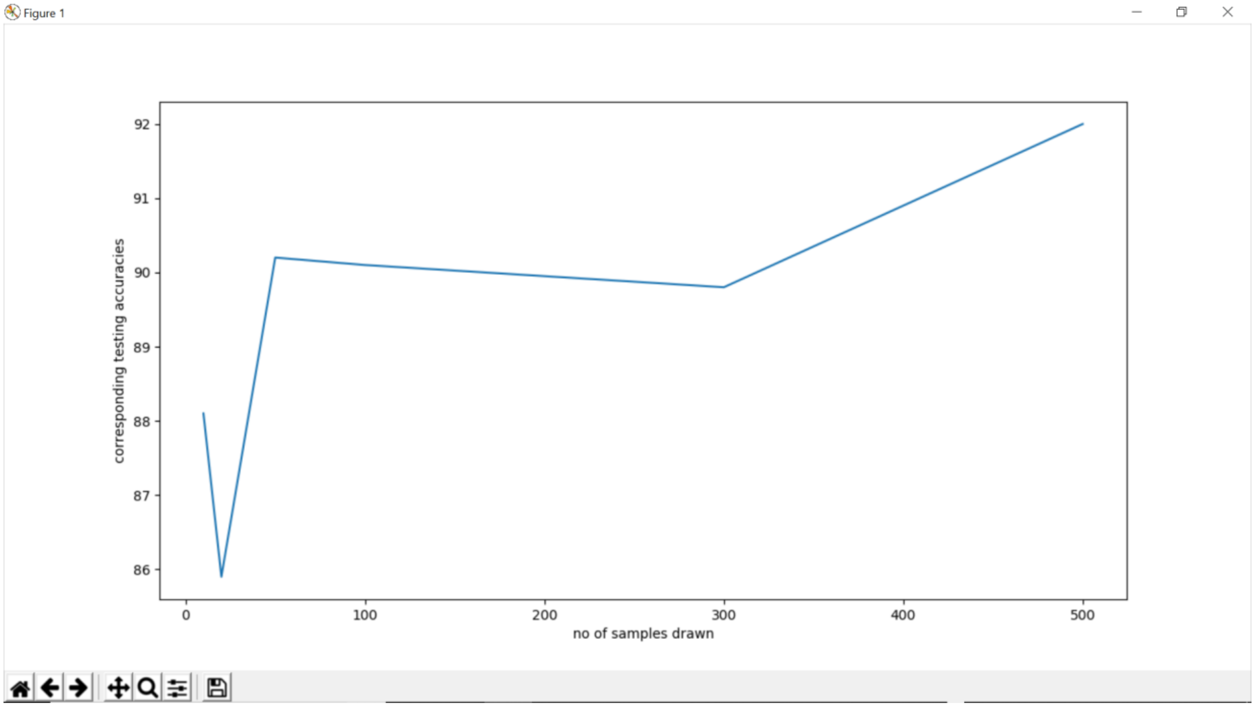
Dictionary consisting of accuracies for training data sets of different sample size.



```
accuracy dictionary:  {10: 81.2, 20: 83.39999999999999, 50: 83.2, 100: 89.7, 300: 91.5, 500: 92.60000000000001}
```

Accuracy vs Training data set sample size plots

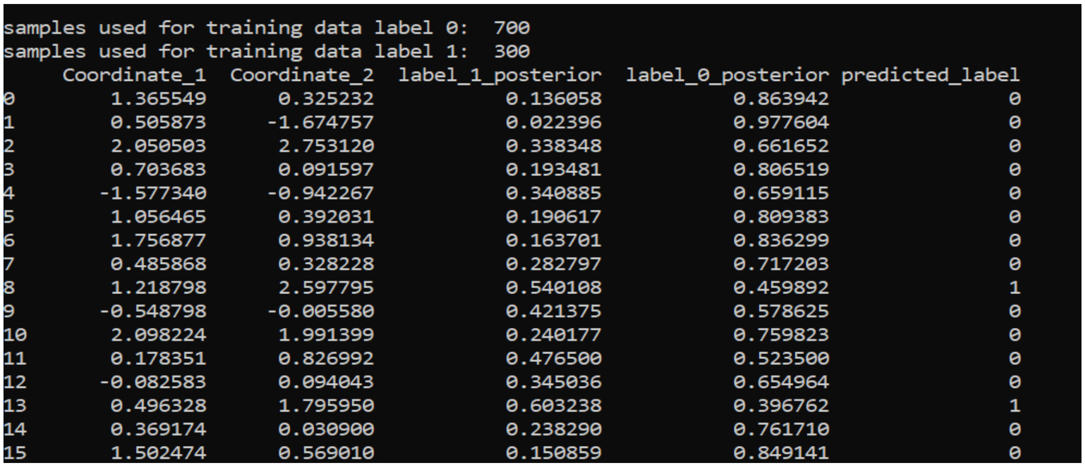




Observation: As the sample size for the training data set increases there is a general increasing trend in the accuracy of the model on the test data set. This proves that with more samples for the training data set we are able to compute the means and standard deviations for the attributes of the domain for the two sperate class labels more accurately, and thus the accuracy of the model increases while predicting the labels for the test data set. Also, since we are computing the accuracy for each sample sizes for the training data set only once, this does not give a good measure of how good a model will perform over the test data sets over a long run for that training data set of a certain sample size. Therefore, we observe that the accuracies for sample size between 10 to 100 is pretty random for every time we run the program. A good measure of the performance can be computed as we take the mean of the accuracies for the model over a test set for a training data set of certain sample size over a longer run.

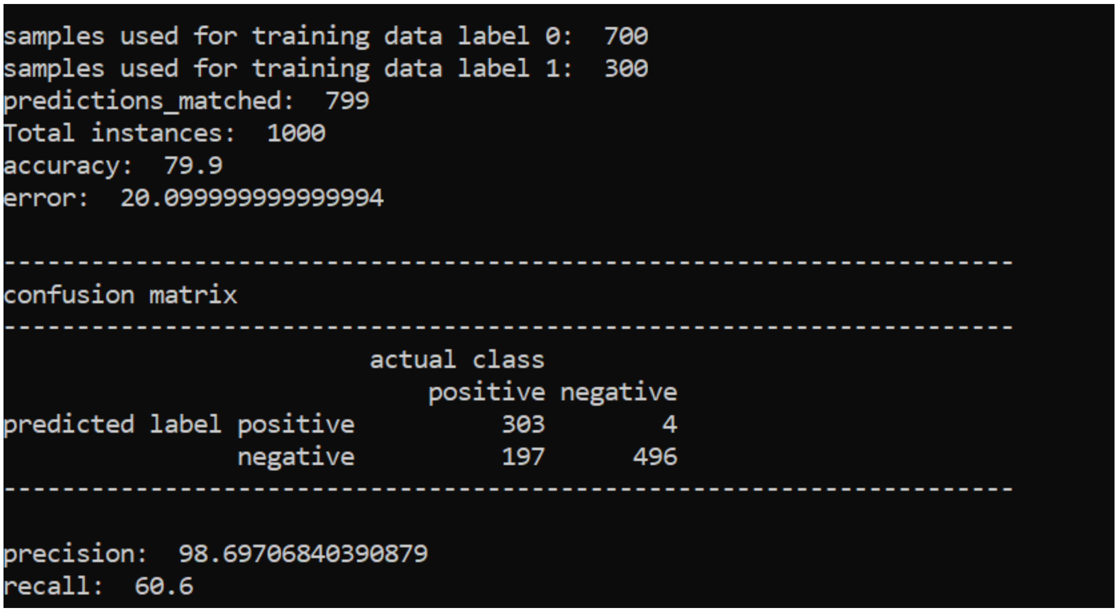
Part 4:

Snapshot of prediction made on the testing data set, snapshot showing for first 15 instances.



```
samples used for training data label 0:  700
samples used for training data label 1:  300
     Coordinate_1  Coordinate_2  label_1_posterior  label_0_posterior predicted_label
0        1.365549      0.325232           0.136058           0.863942               0
1        0.505873     -1.674757           0.022396           0.977604               0
2        2.050503      2.753120           0.338348           0.661652               0
3        0.703683      0.091597           0.193481           0.806519               0
4       -1.577340     -0.942267           0.340885           0.659115               0
5        1.056465      0.392031           0.190617           0.809383               0
6        1.756877      0.938134           0.163701           0.836299               0
7        0.485868      0.328228           0.282797           0.717203               0
8        1.218798      2.597795           0.540108           0.459892               1
9       -0.548798     -0.005580           0.421375           0.578625               0
10       2.098224      1.991399           0.240177           0.759823               0
11       0.178351      0.826992           0.476500           0.523500               0
12      -0.082583      0.094043           0.345036           0.654964               0
13       0.496328      1.795950           0.603238           0.396762               1
14       0.369174      0.030900           0.238290           0.761710               0
15       1.502474      0.569010           0.150859           0.849141               0
```

Performance Measures

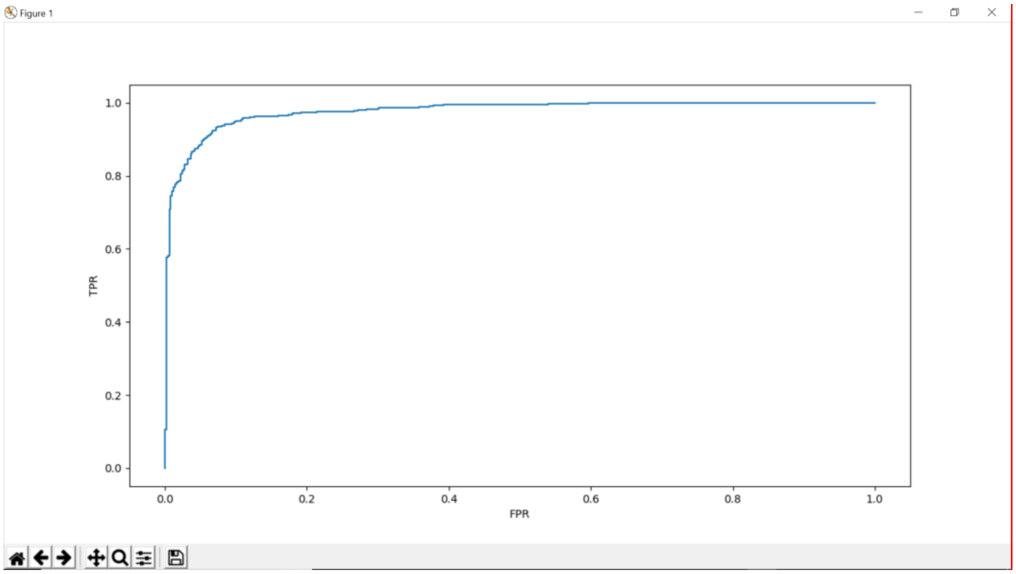


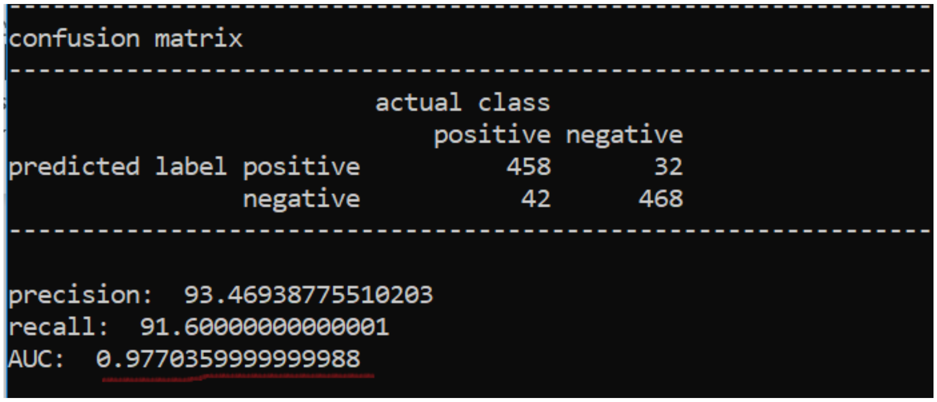
```
samples used for training data label 0:  700
samples used for training data label 1:  300
predictions_matched:  799
Total instances:  1000
accuracy:  79.9
error:  20.099999999999994

---------------------------------------------------------------------
confusion matrix
---------------------------------------------------------------------
                         actual class
                             positive negative
predicted label positive          303        4
                negative          197      496
---------------------------------------------------------------------

precision:  98.69706840390879
recall:  60.6
```

Observation: Since the training data set is skewed towards the samples from the class label 0, the means and standard deviations computed for the attributes for the class label 1 are not that accurate as compared to that for class label 0. Also the prior for label 1 is low as compared to prior for label 0 and thus the test instances which should have been predicted as positive are predicted as negative and the count for FN is pretty high which leads to a smaller count for TP and thus a relatively lower accuracy as compared to an uniformly distributed training data set.

Part 5:

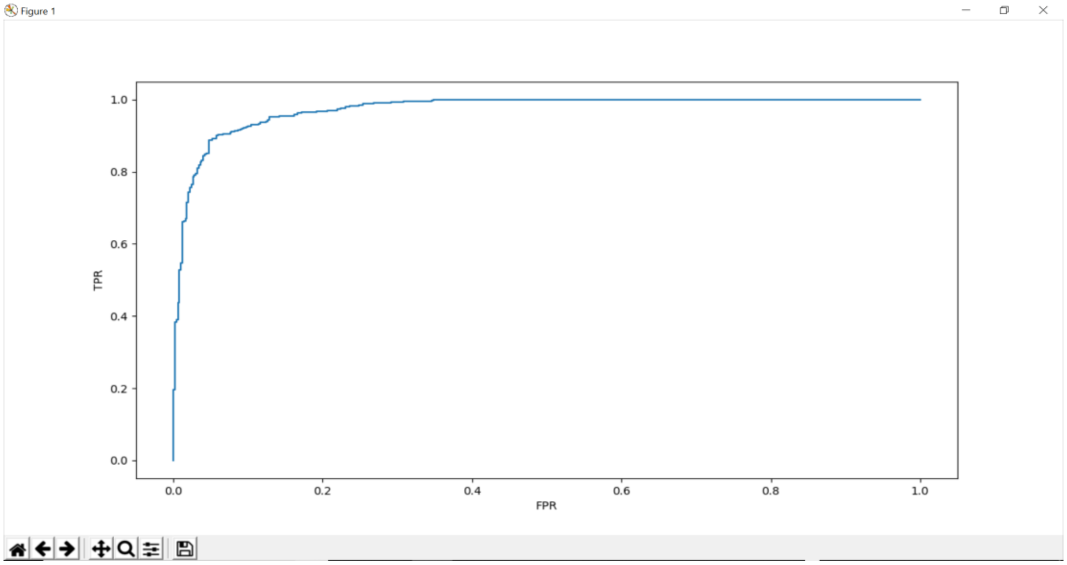
ROC Plot for Problem 2:

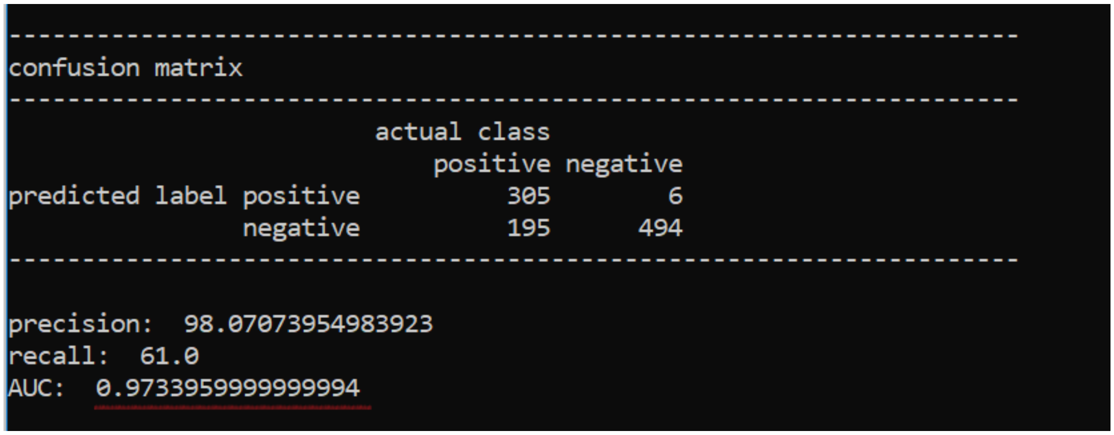




```
---------------------------------------------------------------
confusion matrix
---------------------------------------------------------------
                         actual class
                             positive negative
predicted label positive          458       32
                negative           42      468
---------------------------------------------------------------

precision:  93.46938775510203
recall:  91.60000000000001
AUC:  0.9770359999999988
```

ROC Plot for Problem 4:





```
---------------------------------------------------------------------
confusion matrix
---------------------------------------------------------------------
                         actual class
                             positive negative
predicted label positive          305        6
                negative          195      494
---------------------------------------------------------------------

precision:  98.07073954983923
recall:  61.0
AUC:  0.9733959999999994
```