CO324 Pattern Recognition
Assignment-1

Siddhant Verma
2K18/EC/167

March 6, 2021

Exercise 1

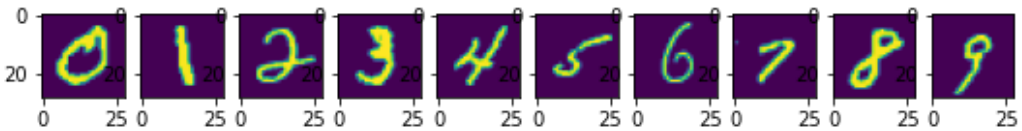
Design a Bayes classifier for classifying the numbers in the MNIST handwritten digit recognition databases.

Dataset link: http://yann.lecun.com/exdb/mnist/

1. Design a classifier to distinguish between 0 and 1. Use the training database available for both to design the classifier.
2. Use the testing database to compute the classification accuracy ((TP+TN)/ (TP+TN+FP+FN)) of the Bayesian model.
3. Plot the ROC curves between FAR vs GAR.
4. Repeat the above three parts for classifying the digits 3 and 8.
5. Compare and contrast the results of the two classifiers.

Solution:

Data set Overview:



The MNIST Data set has over 60,000 samples in the the form of 28x28 images in the training set and about 10,000 samples in the testing set. The respective labels are the values of the digits that have been captured as images, which means that there are 10 classes 0-9 in the entire dataset.

The introductory part of the question suggests that we need to build as classifier for the entirety of the data set for handwritten digit recognition.

The model achieved 55.8% accuracy.

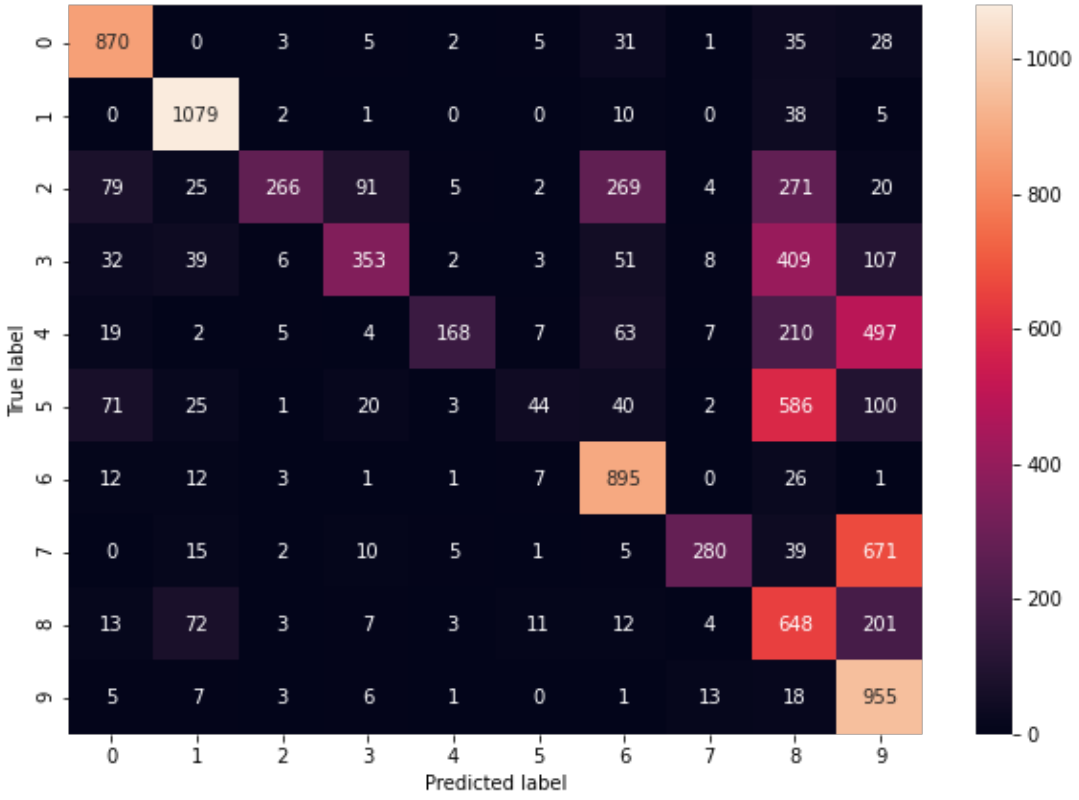


Figure 1: Confusion Matrix for the classifier

For the following part, since the intent of the question it to design a classifier which is inherently trained on samples belonging to classes for digits 0 and 1 only, both training and testing data sets were filtered, thus the training set only had 12665 samples left and the testing set had 2115 samples left.A new Bayes' classifier model was trained and it achieved 98.77% accuracy score. Same steps were repeated for digits 3 and 8 as per (d) of the question.

Comparison

Bayesian-01 Model
TP = 976 TN=1113 FP=4 FN=22 Accuracy = 99.87%
The ROC curve for the 0-1 classifier had an area under curve value of 99.22%.

Bayesian-38 Model
TP = 435 TN=952 FP=575 FN=22 Accuracy = 69.90%
The ROC curve for the 3-8 classifier had an area under curve value of 72.32%.

Visualisations for 0-1 classifier

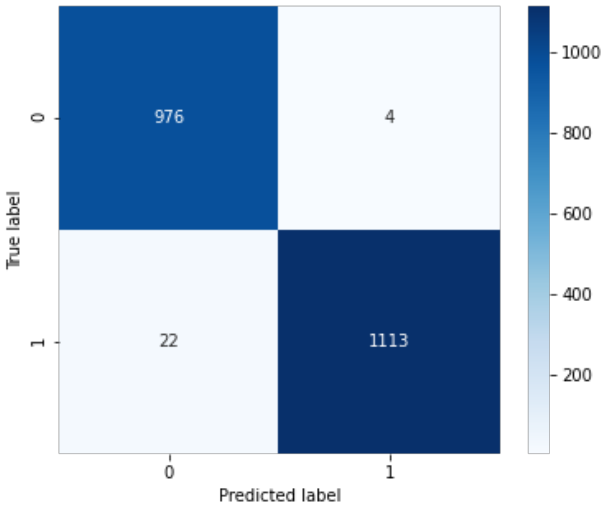


Figure 2: Confusion Matrix for the 0-1 classifier

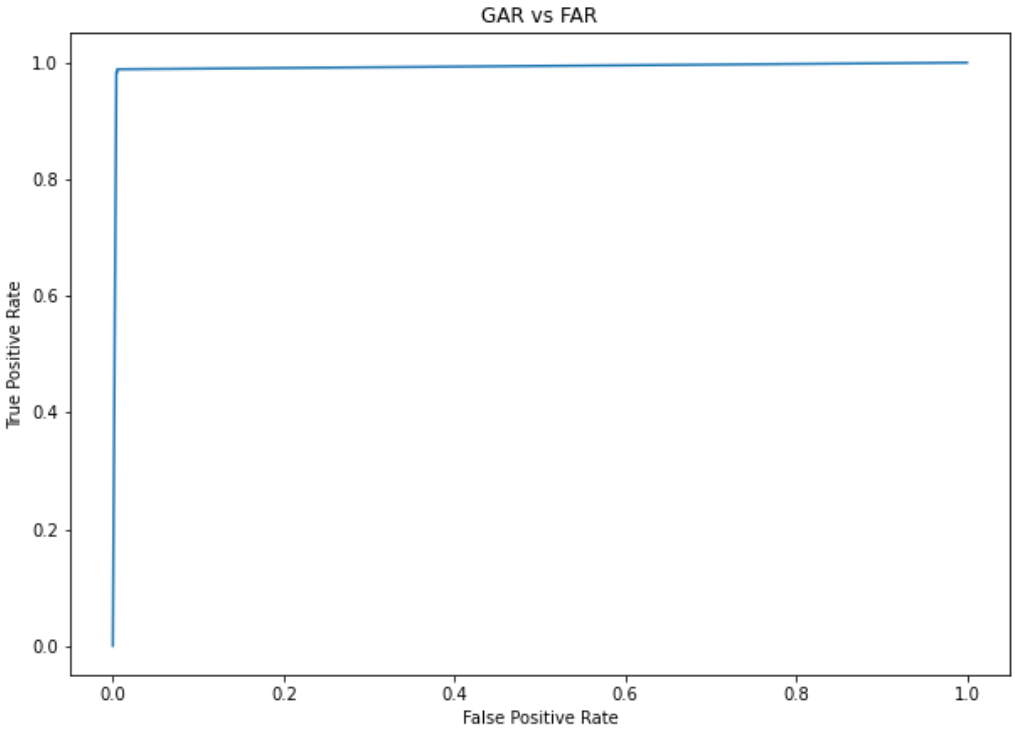


Figure 3: ROC for the 0-1 classifier

Visualisations for 3-8 classifier

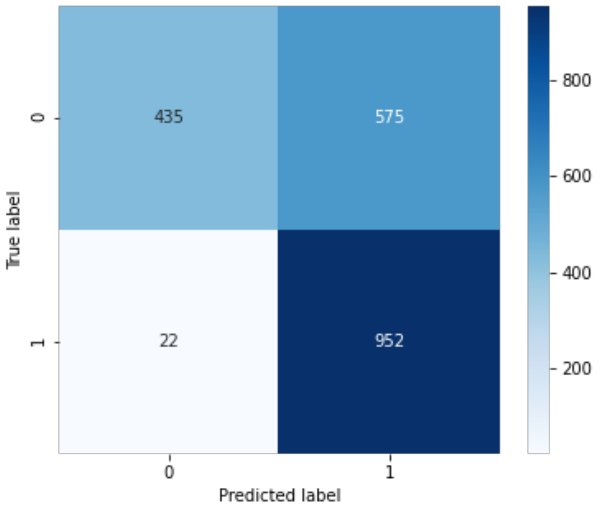


Figure 4: Confusion Matrix for the 3-8 classifier

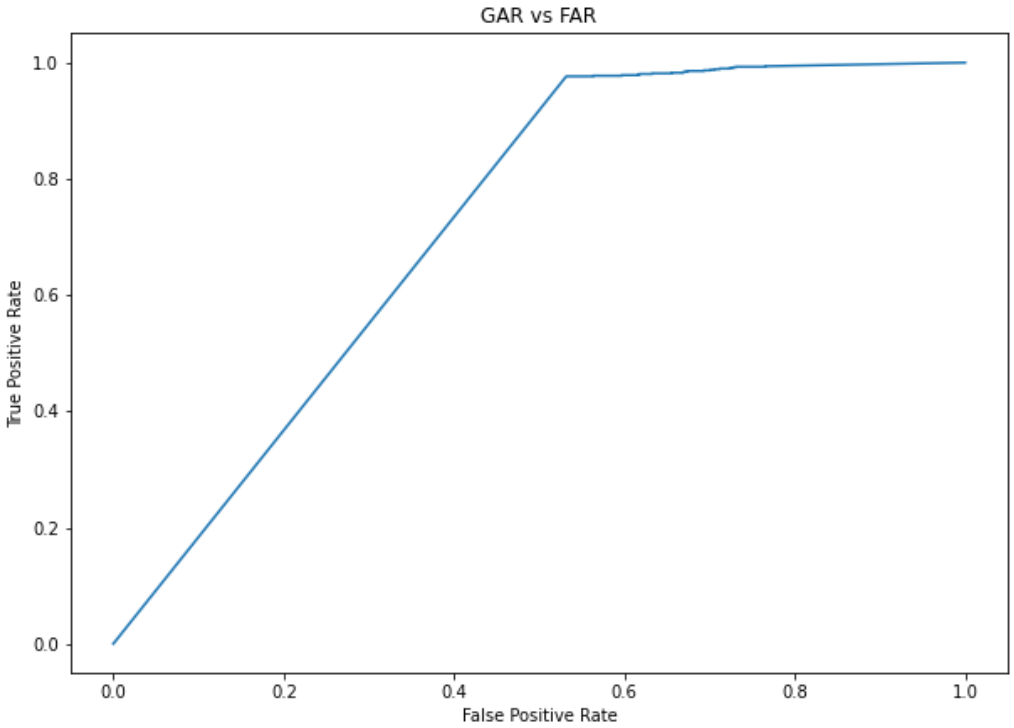


Figure 5: ROC for the 3-8 classifier

Exercise 2

Design a Bayes classifier for classifying bank notes as genuine or forged (https://archive.ics.uci.edu/ml/datasets/banknote+authentication).

1. Use 50% of the database from both the classes for training and the remaining 50% for testing. Classes are equiprobable.
2. Use the testing database to compute the classification accuracy ((TP+TN)/(TP+TN+FP+FN)) of the Bayesian model.
3. Plot the ROC curves between FAR vs GAR.
4. Repeat the above three parts when the prior of genuine class is 0.9 and forged is 0.1.
5. Compare and contrast the results of the two classifiers.

Solution:
Although the question says that the classes are equiprobable, they were not. Upon calculating the class distributions, the genuine to forged class ratio was about 55-45.

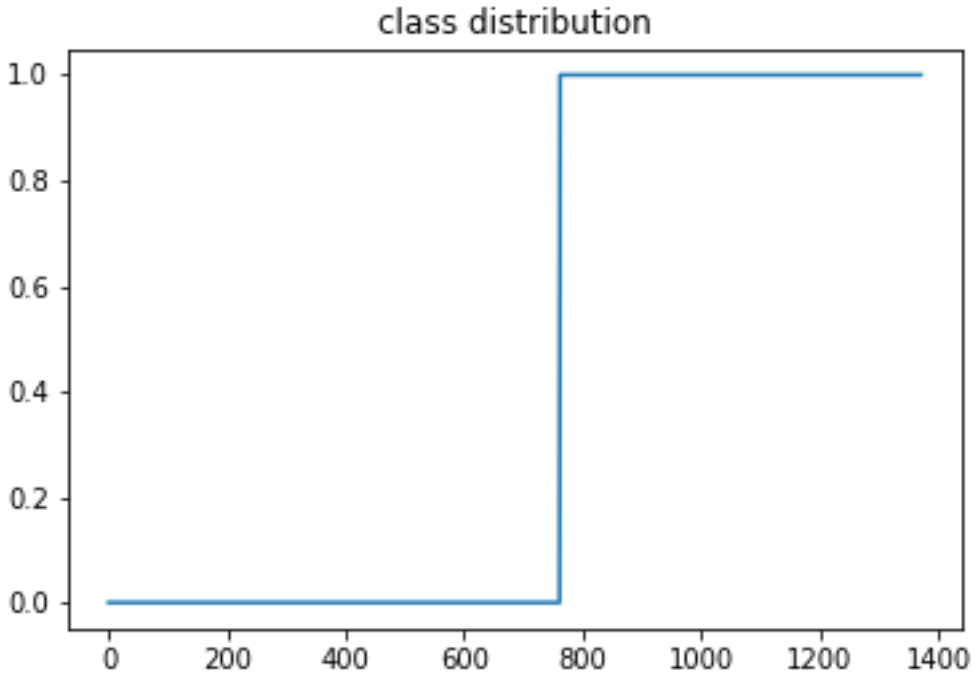


Figure 6: Class Distribution

The data set contrains the following features:

- variance
- skewness
- curtosis
- entropy
- class

Data Visualisation

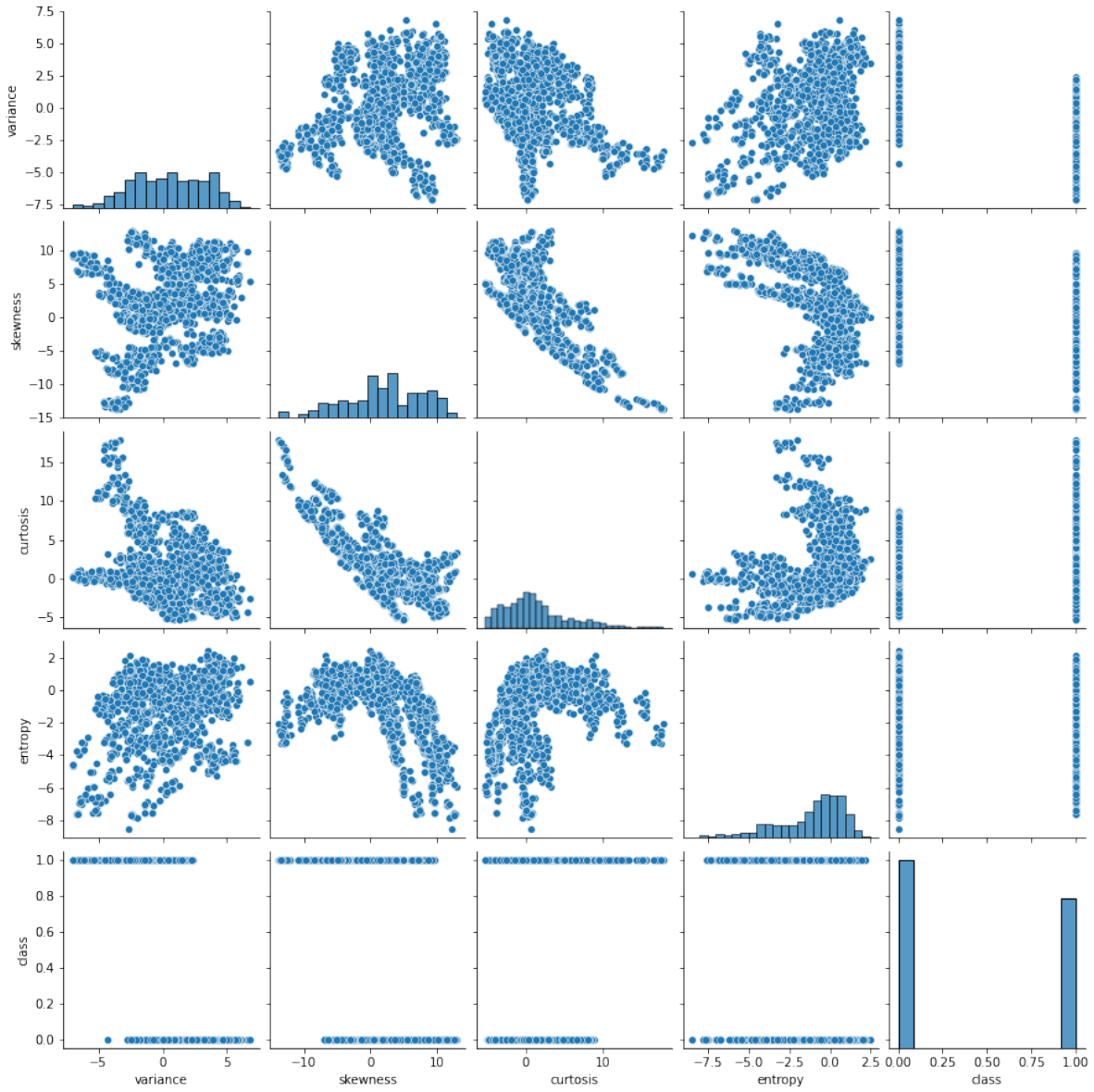


Figure 7: Pair plot of data set

The data was then shuffled and split in to 2 halves which are further used as testing and training sets. The bayesian classifier is then trained using the sklearn library.

50-50 classifier

The classifier achieved an accuracy of 84.98%.
TP = 340
TN = 243
FP = 42
FN = 61
The ROC area under curve = 93.26%

According to part (d), the data set was reshuffled and redistributed such that the prior of genuine class is 0.9 and that of forged class is 0.1. This way 762 samples of genuine class and 84 samples of forged class were taken into the data set. This was further shuffled and split into training and testing data sets.

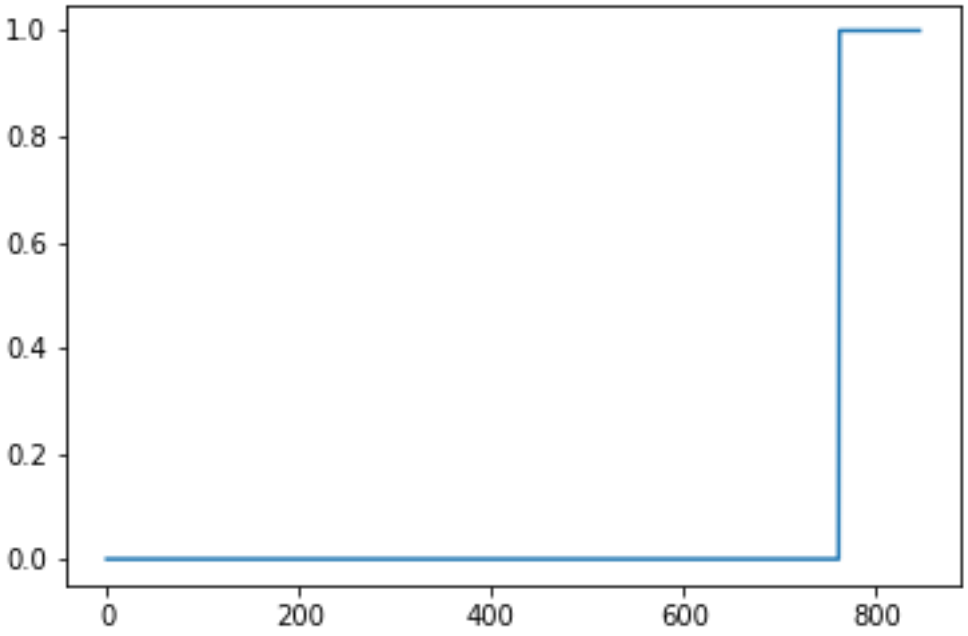


Figure 8: Class Distribution

90-10 classifier

The classifier achieved an accuracy of 93.14%.
TP = 372
TN = 22
FP = 7
FN = 22
The ROC area under curve = 95.53%

Visualisations for 50-50 classifier

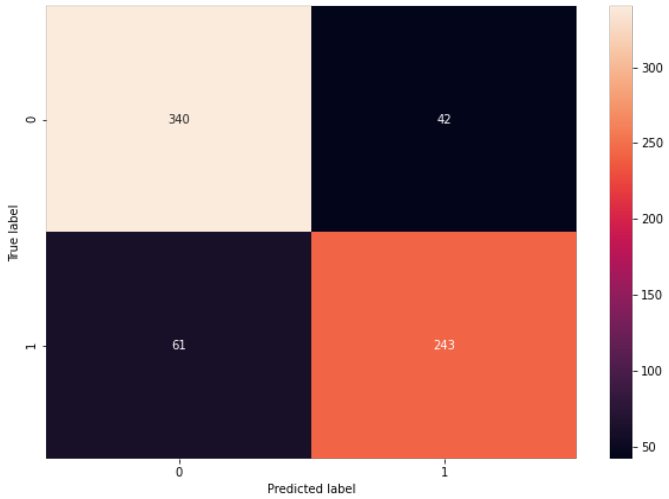


Figure 9: Confusion Matrix for the 50-50 classifier

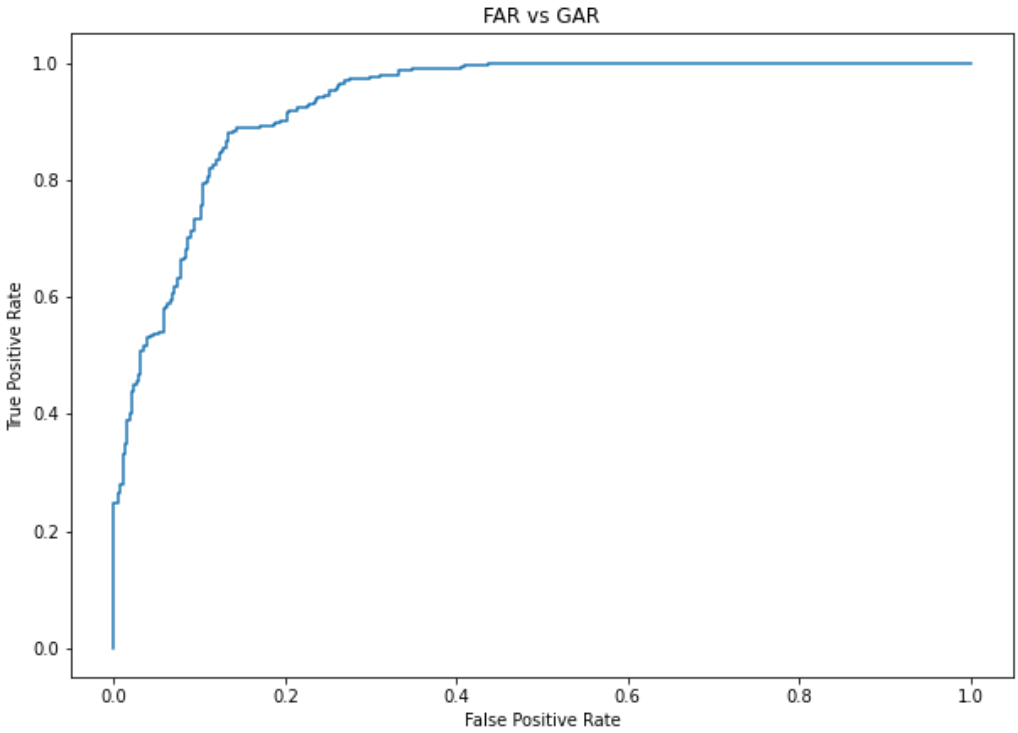


Figure 10: ROC for the 50-50 classifier

Visualisations for 90-1 classifier

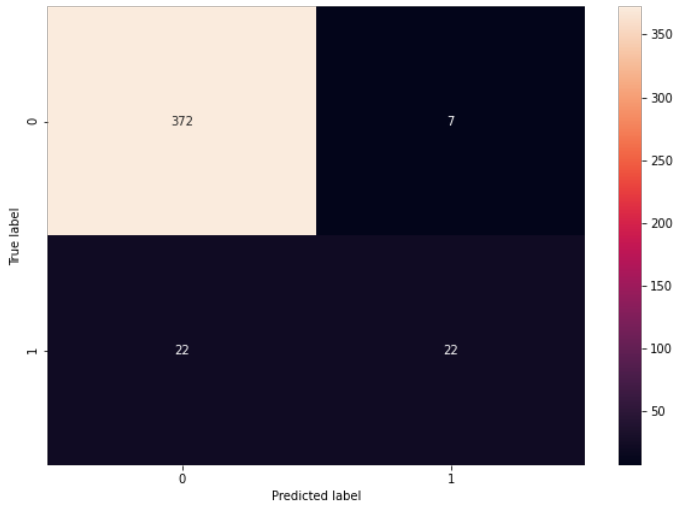


Figure 11: Confusion Matrix for the 90-10 classifier

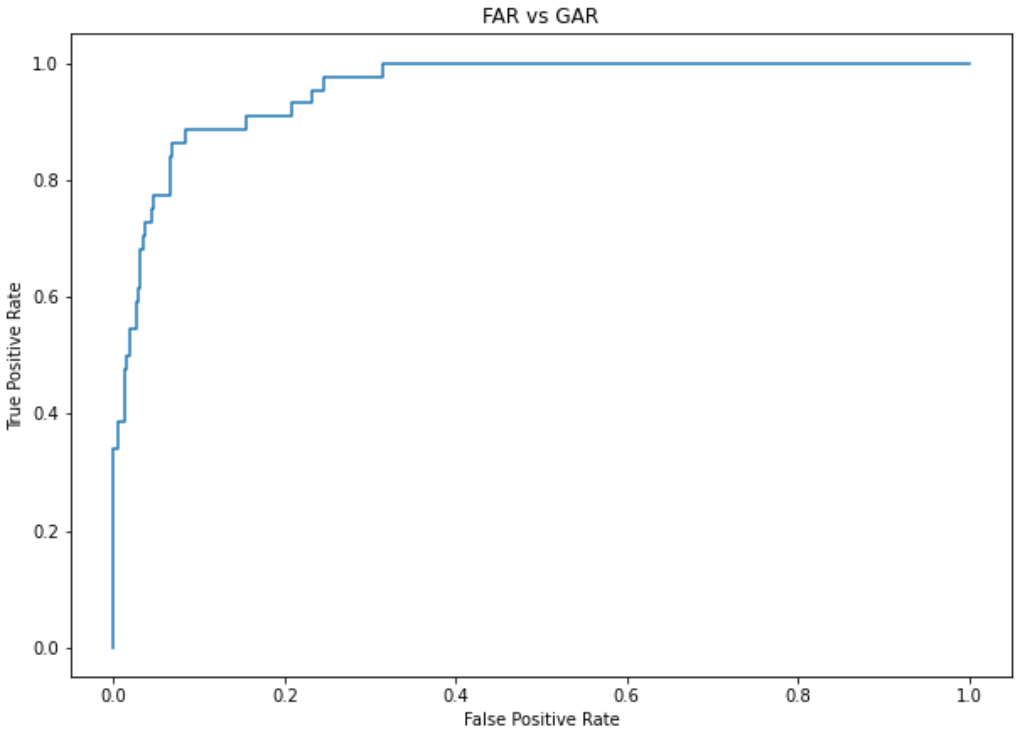


Figure 12: ROC for the 90-10 classifier