Lab 4 Summary – Data Mining and Machine Learning

In Lab 4, we had to work on Classifier testing and evaluation using python functions in our project.

The first step was to split the dataset into two subsets - training data and test data. Each subset was further divided into X and y, where X contains the chosen features and y is the prediction target.

We then inspected five random instances in the training data for any inconsistencies.

The prediction algorithm we used was Logistic Regression, a linear machine-learning model.

The next step was to formulate the confusion matrix on the results obtained.

The results obtained were as follows:

**TP** = 75 **FP** = 30

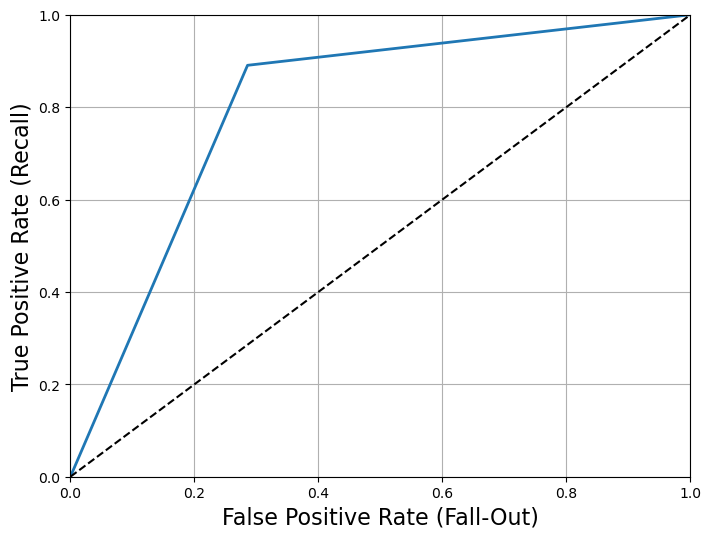
**FN** = 24 **TN** = 122

Precision score = 0.8

Recall Score = 0.89

F1 Score = 0.84

The following is the ROC curve obtained:



The algorithm used here has outputted a high number of True Positives and True Negatives. This means that it is noticeably efficient in detecting whether a case is positive or negative.

Since this algorithm aims to predict the risk of a heart attack, high priority must be given to minimize the False Negative, since missing the detection of a heart attack can have fatal consequences. On the other hand, while False positives are also to be avoided, it has a lower priority because it is better to be safe than sorry. In the real world, a false or true positive can help lead the patient to a doctor for further checkup.

We have a good F1 score of 0.84, which indicates that the precision and recall are balanced out. But since this is a dataset on heart attack analysis, a higher F1 score would be preferred.

We have a high precision score of 0.8, which means most cases detected as positive are actually positive. This is good, since it reduces false alarms.

Our recall value (0.89), is a high number which indicates most cases that are positive will be detected. This value has higher precedence over the precision score since we want to detect as many true cases as possible.