**Final Term Project Report**

**Supervised Learning**

**Data Warehousing and Data Mining**

**Name: Samin, MD. Towhidur Rahman**

**ID: 17-34181-1**

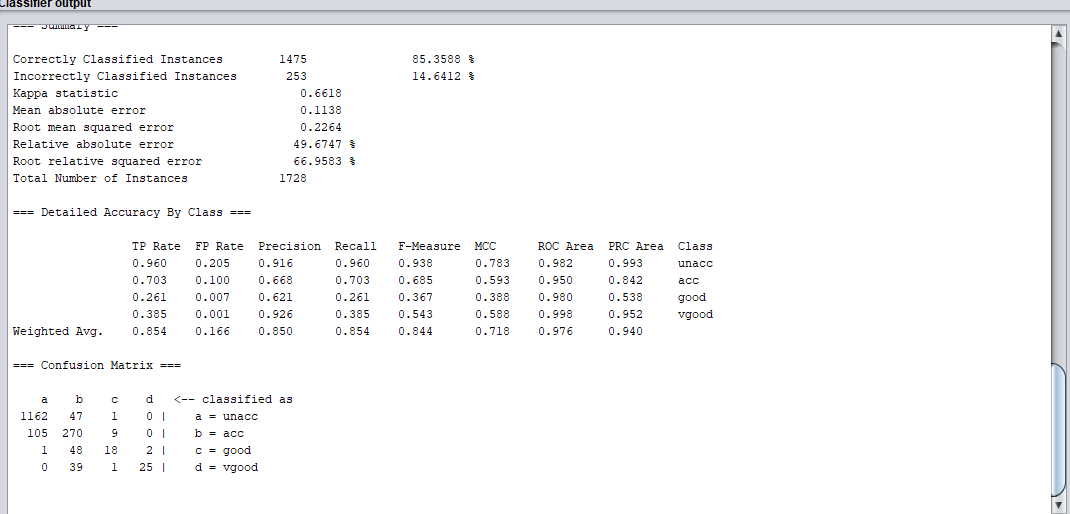
**Section: C**

**Project Definition:**

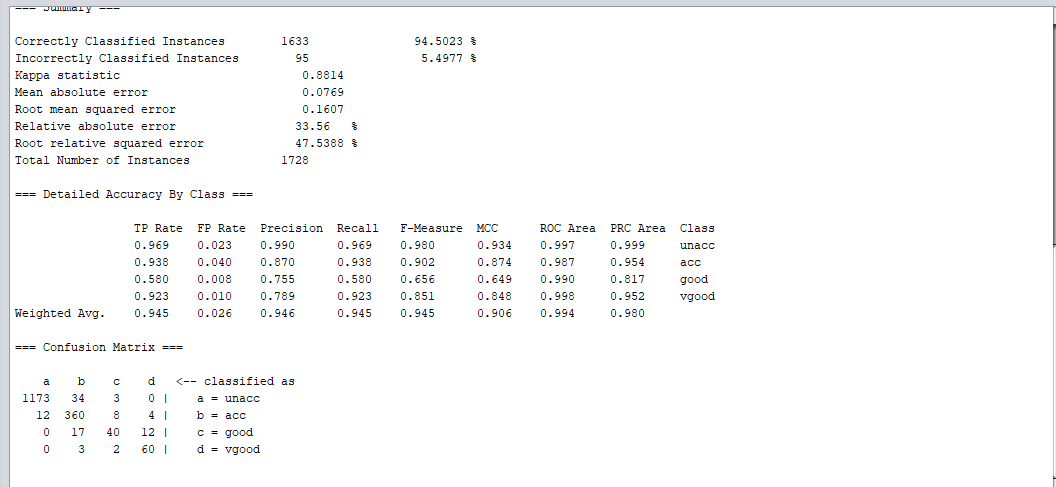
Data mining is a computer assisted process of digging through and analyzing a set of data and extracting the meaning of data. In this modern era data mining plays an important role to analyze the data with different types of algorithms and predict it’s result. In this report the data set of Car Evaluation from UCI repository was used to analyze the data by using five different supervised classifier algorithms. The goal of this project is to find out the best predictive result of this dataset by using these classifiers and also find out which classifier has the best performance among them.

**Methods:**

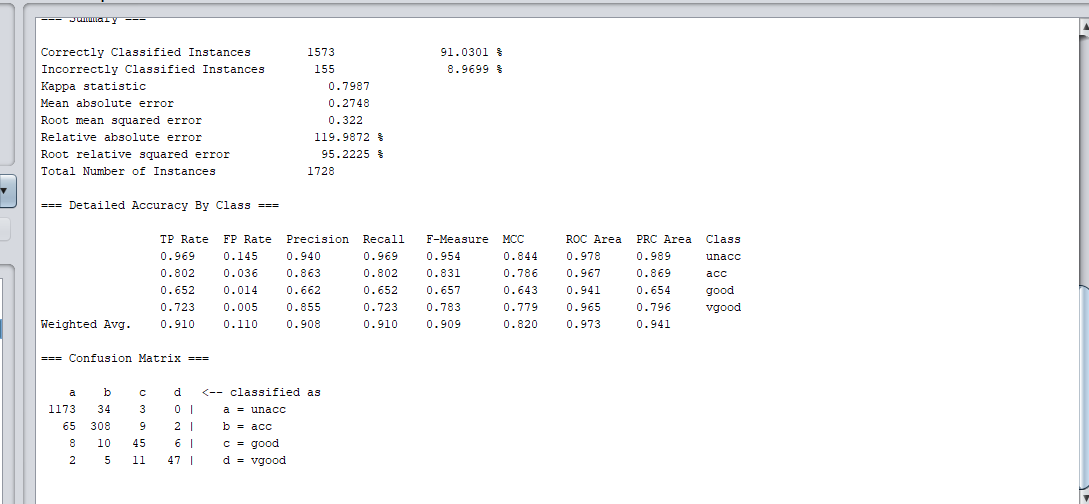
Naive Bayes Classifier:



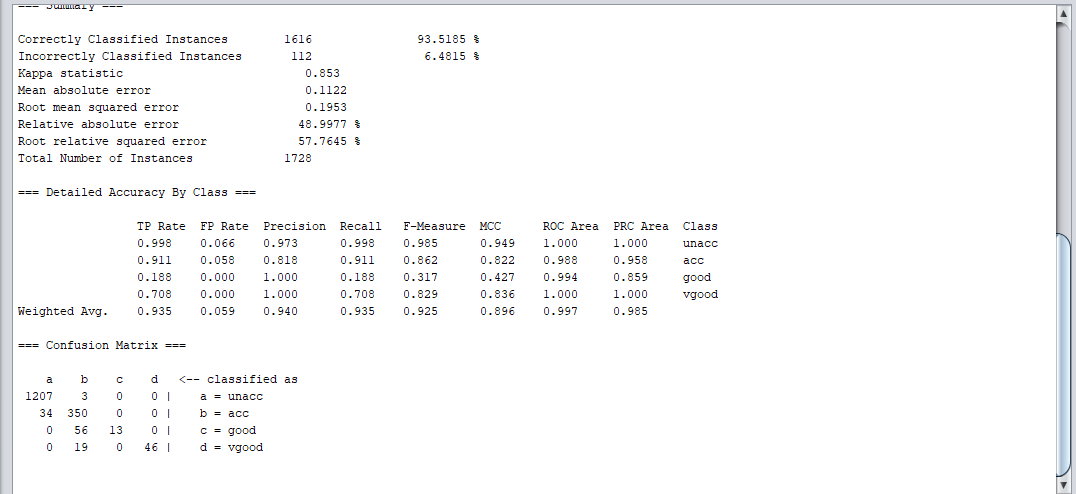
Random Forest Classifier:



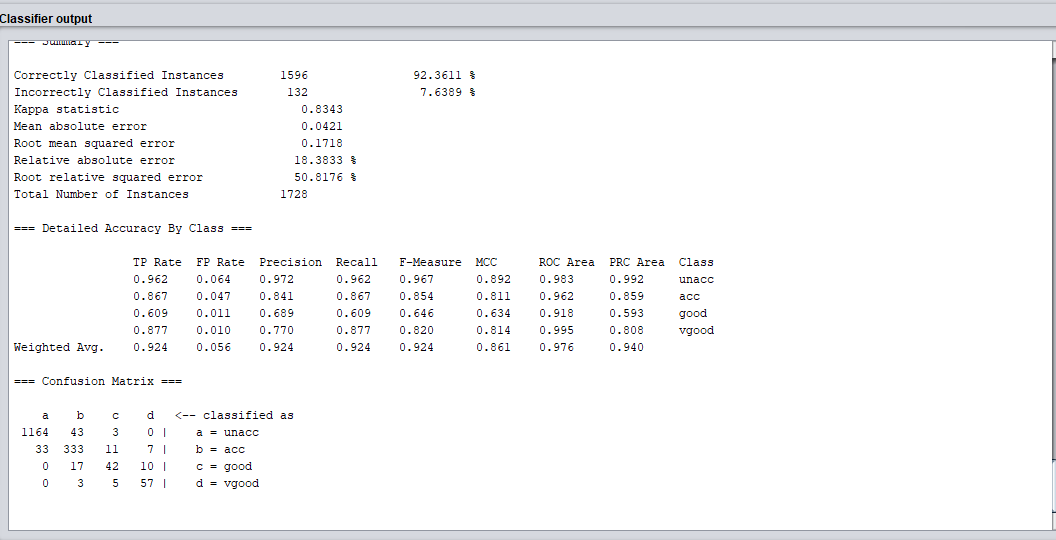
Decision Table Classifier:



KNN Classifier:

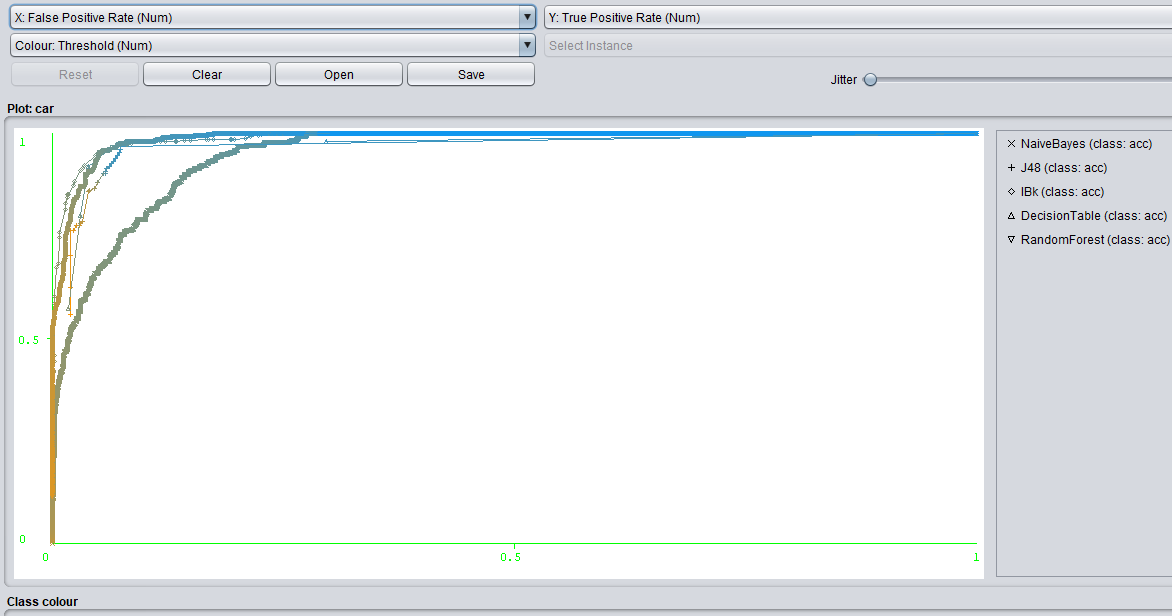


Decision Tree (J48) Classifier:

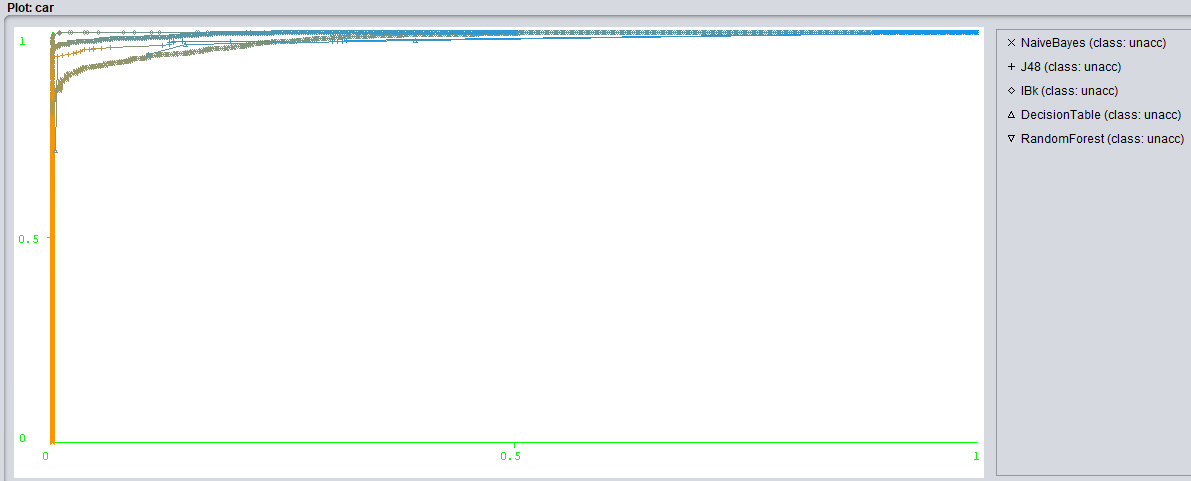


**ROC Curves:**

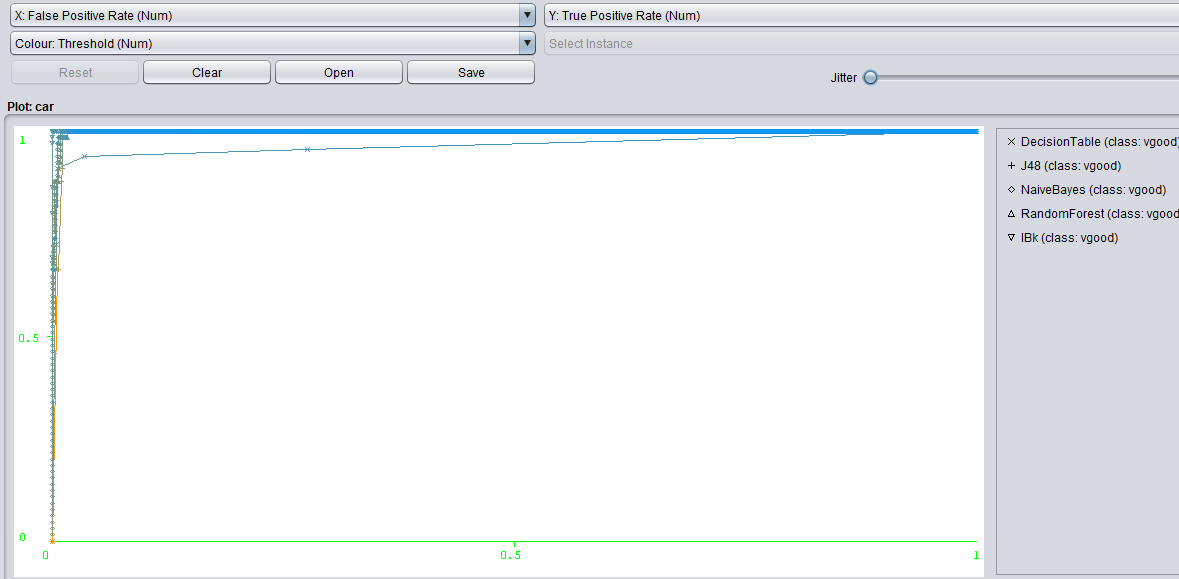
For Class acc:



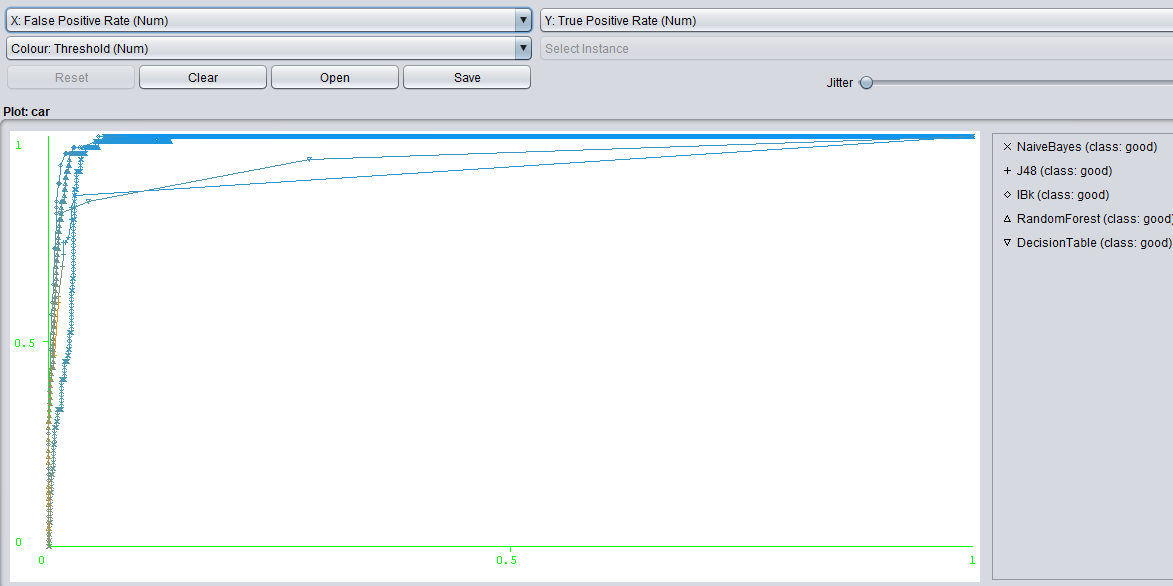
For Class unacc:



For Class vgood:



For Class good:



**Comments:**

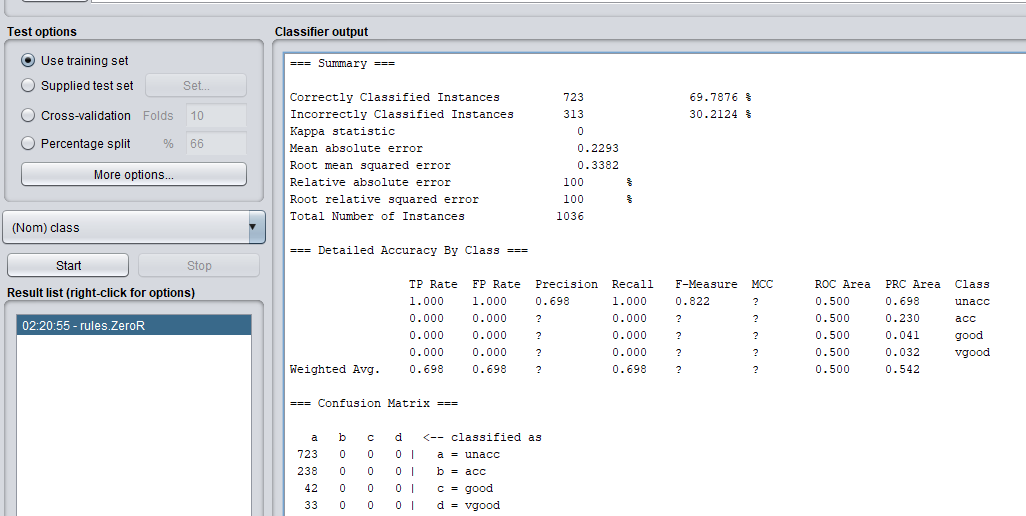
In this report here the weighted average of true positive rate of classifier Random Forest is 0.945 which is the highest and the false positive rate is 0.026 which is the lowest value between these classifiers. It has also the most correctly classified instances which is 94.5023%. The ROC curve of the Random Forest is also the closest to the ideal point (0,1). The ideal point (0,1) represents 100% sensitivity (no false negatives) and 100% specificity (no false positives).

Among the other classifiers the TP rate and the FP rate of Naïve Bayes are 0.854 and 0.166 which are the worst weighted average value among the classifiers.

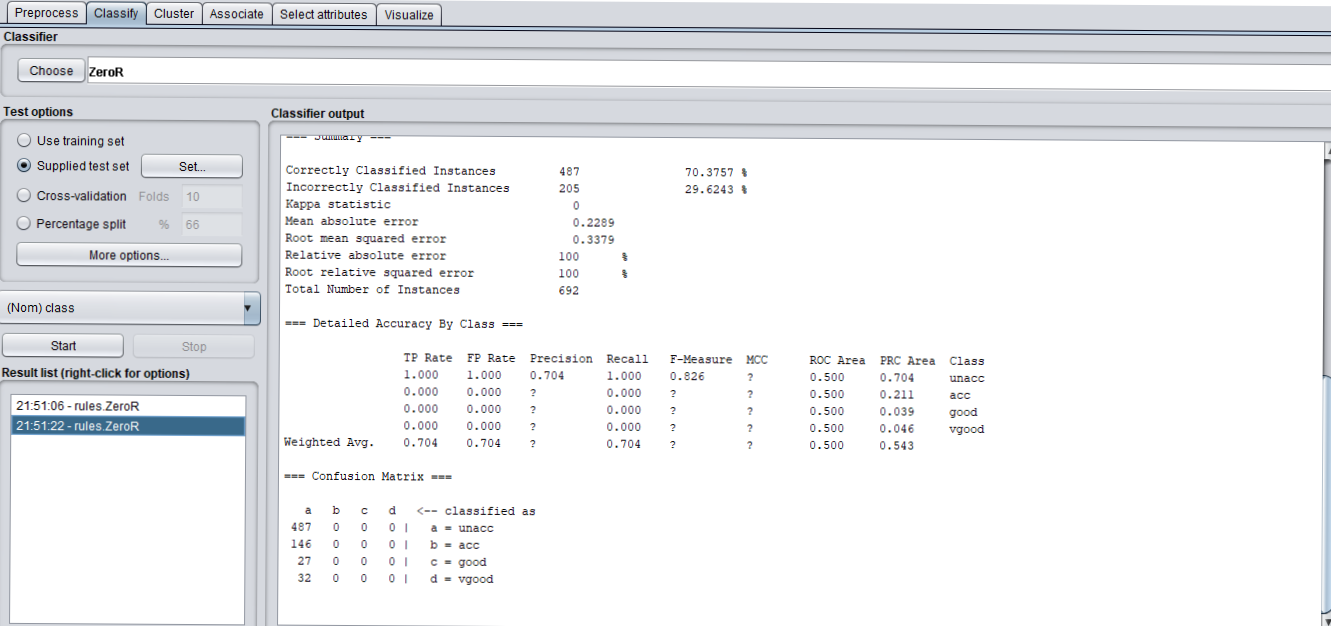
Finally, it can be said that by analyzing the data set the predicting the result, Random Forest classifier would be the best classifier.

**Additional Task:**

For Training Data set:



For Test Data set:



**Comments:**

For creating a Test data set 40% data of the data set was used and remaining 60% data was used for creating Training data set. 69.7876% instances of Training data set were correctly classified where 70.3757% instances of Test data set were correctly classified. The comparison isn’t too big but if we check the weighted average of False positive rate of the Test data set is 0.704 where the FP rate of Training data set is 0.698, which is lower and also much better than the Test data set.