Final Term Project Report

Supervised Learning

Data Warehousing and Data Mining

Name: Samin, MD. Towhidur Rahman

ID: 17-34181-1

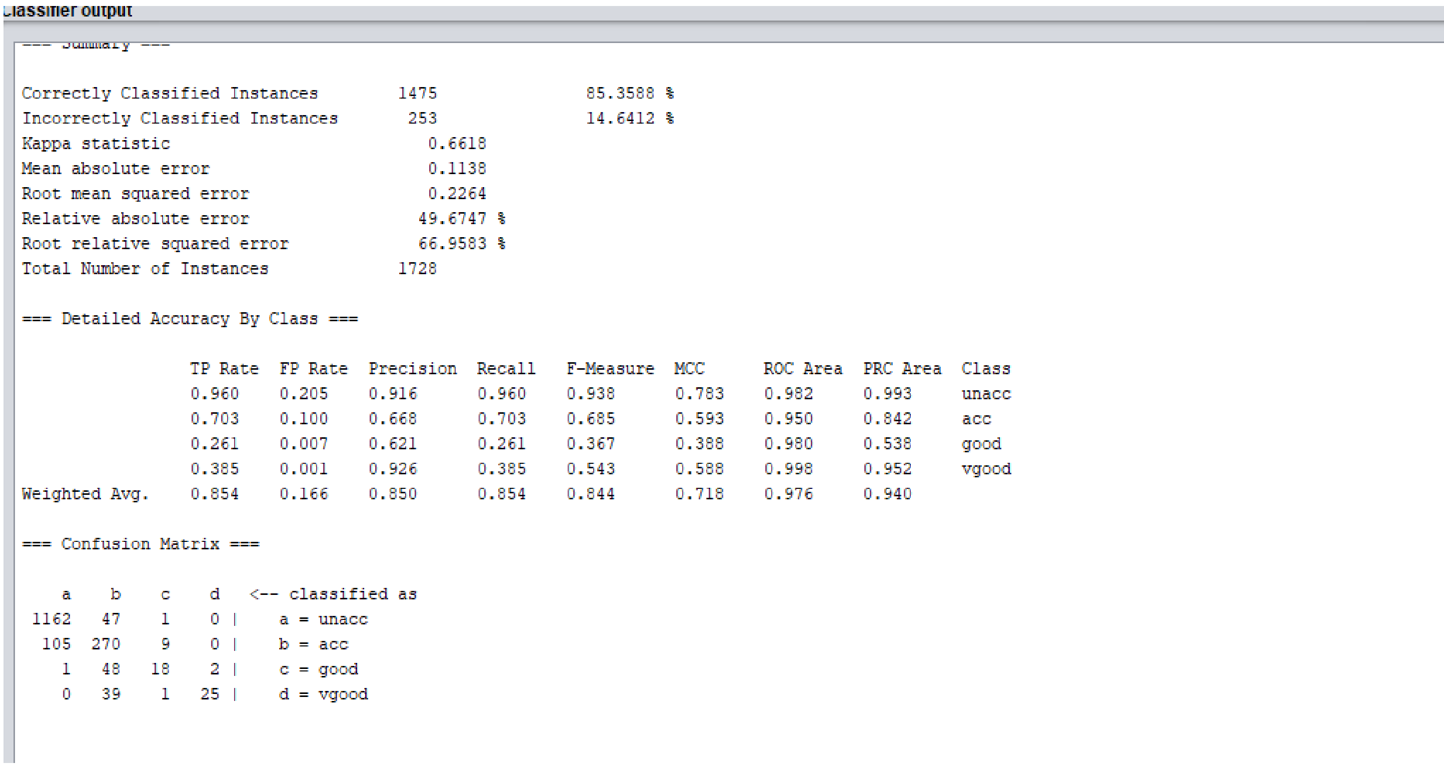
Section: C

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.

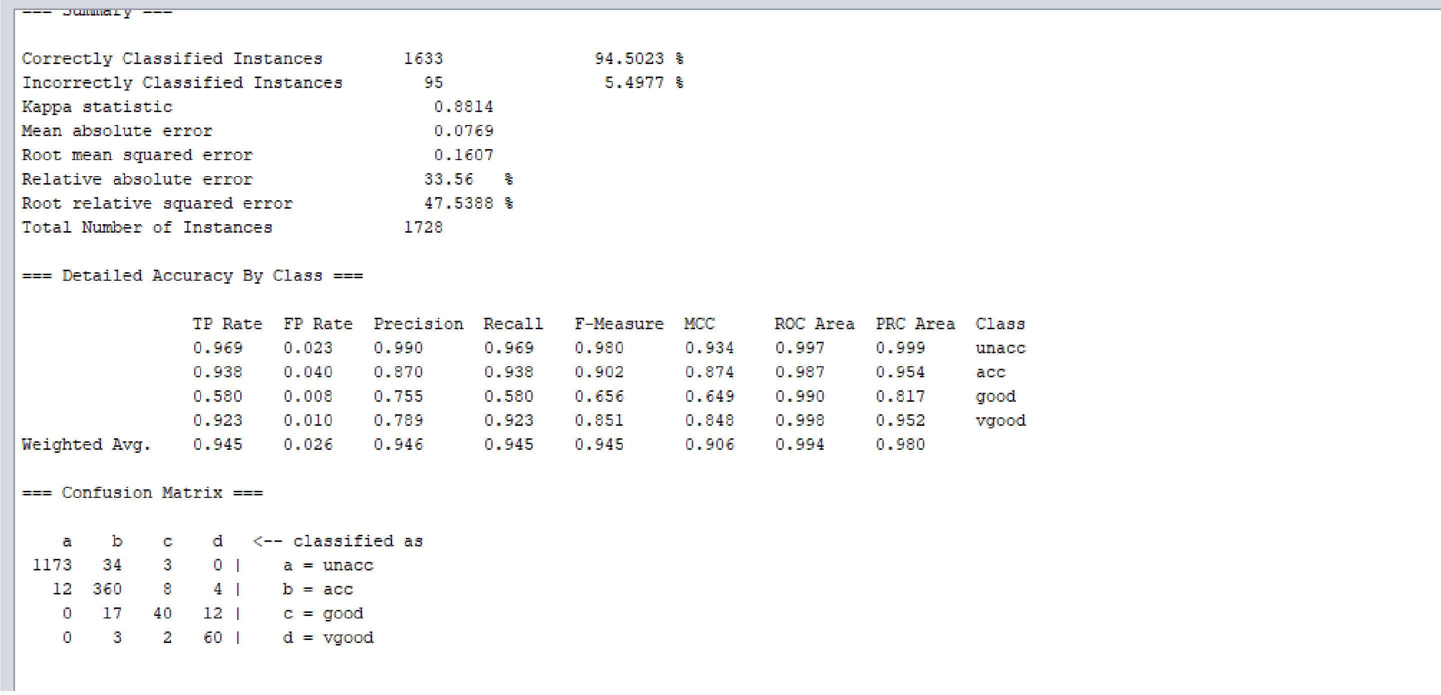
Methods:

Naive Bayes Classifier:



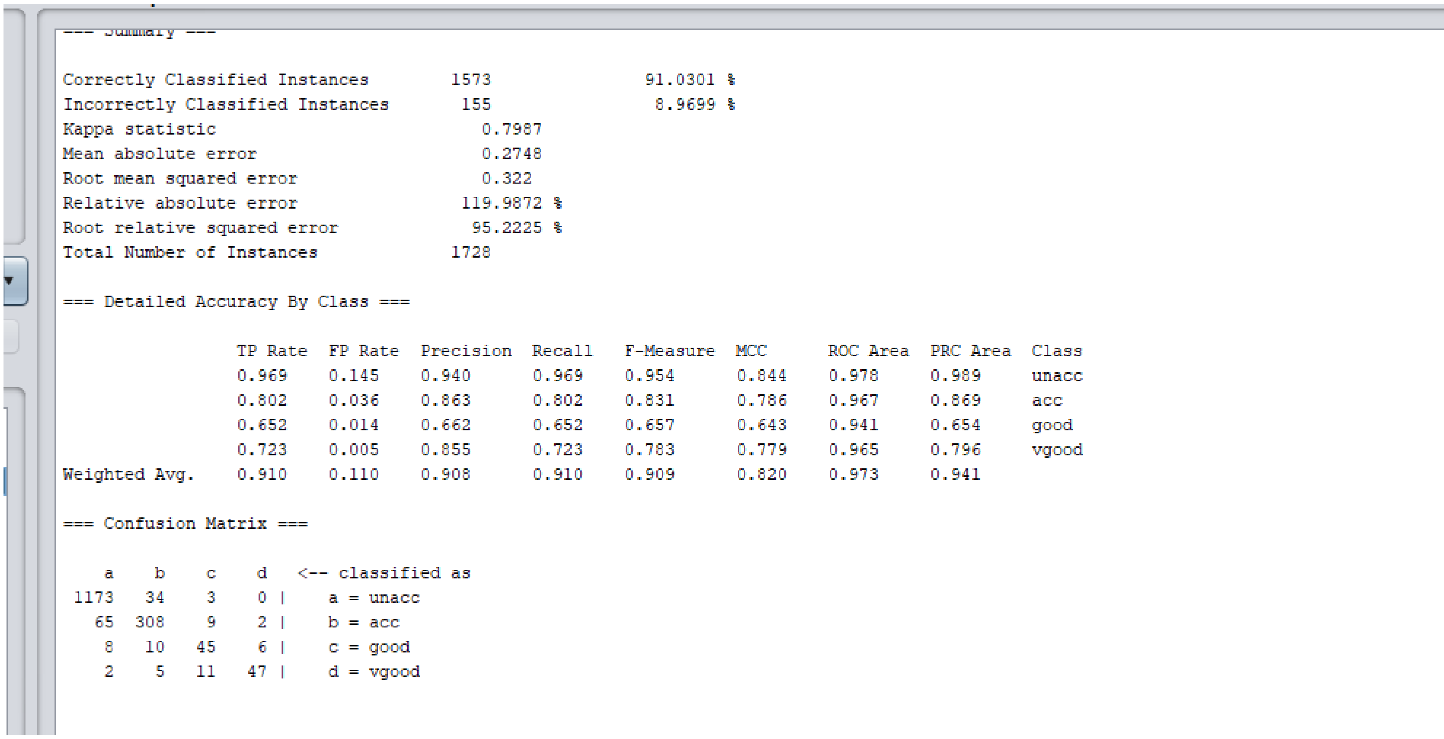
```
=== Summary ===

Correctly Classified Instances        1475               85.3588 %
Incorrectly Classified Instances       253               14.6412 %
Kappa statistic                          0.6618
Mean absolute error                      0.1138
Root mean squared error                  0.2264
Relative absolute error                 49.6747 %
Root relative squared error             66.9583 %
Total Number of Instances             1728

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.960    0.205    0.916      0.960    0.938      0.783    0.982     0.993     unacc
                 0.703    0.100    0.668      0.703    0.685      0.593    0.950     0.842     acc
                 0.261    0.007    0.621      0.261    0.367      0.388    0.980     0.538     good
                 0.385    0.001    0.926      0.385    0.543      0.588    0.998     0.952     vgood
Weighted Avg.    0.854    0.166    0.850      0.854    0.844      0.718    0.976     0.940

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1162   47    1    0 |    a = unacc
  105  270    9    0 |    b = acc
    1   48   18    2 |    c = good
    0   39    1   25 |    d = vgood
```

Random Forest Classifier:



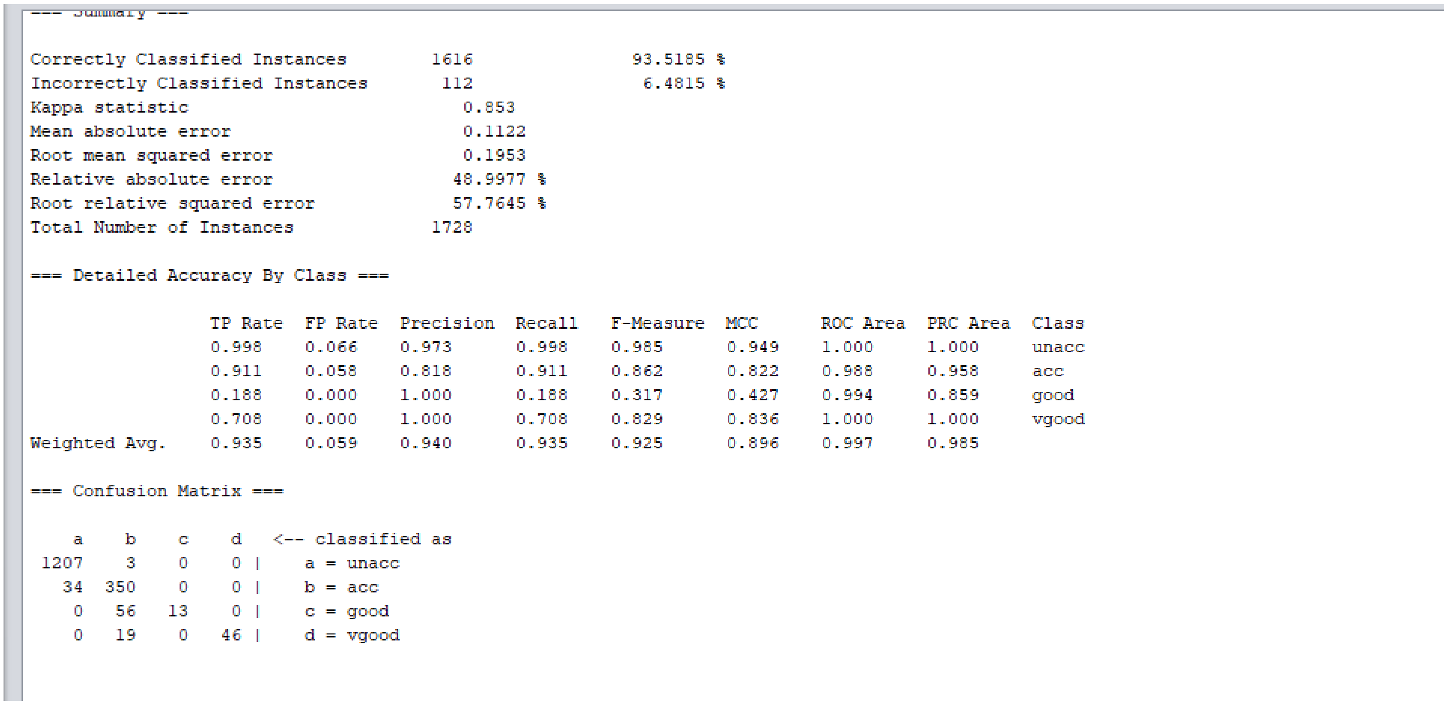
```
=== Summary ===

Correctly Classified Instances        1633               94.5023 %
Incorrectly Classified Instances        95                5.4977 %
Kappa statistic                          0.8814
Mean absolute error                      0.0769
Root mean squared error                  0.1607
Relative absolute error                 33.56   %
Root relative squared error             47.5388 %
Total Number of Instances             1728

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.969    0.023    0.990      0.969    0.980      0.934    0.997     0.999     unacc
                 0.938    0.040    0.870      0.938    0.902      0.874    0.987     0.954     acc
                 0.580    0.008    0.755      0.580    0.656      0.649    0.990     0.817     good
                 0.923    0.010    0.789      0.923    0.851      0.848    0.998     0.952     vgood
Weighted Avg.    0.945    0.026    0.946      0.945    0.945      0.906    0.994     0.980

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1173   34    3    0 |    a = unacc
   12  360    8    4 |    b = acc
    0   17   40   12 |    c = good
    0    3    2   60 |    d = vgood
```

Decision Table Classifier:



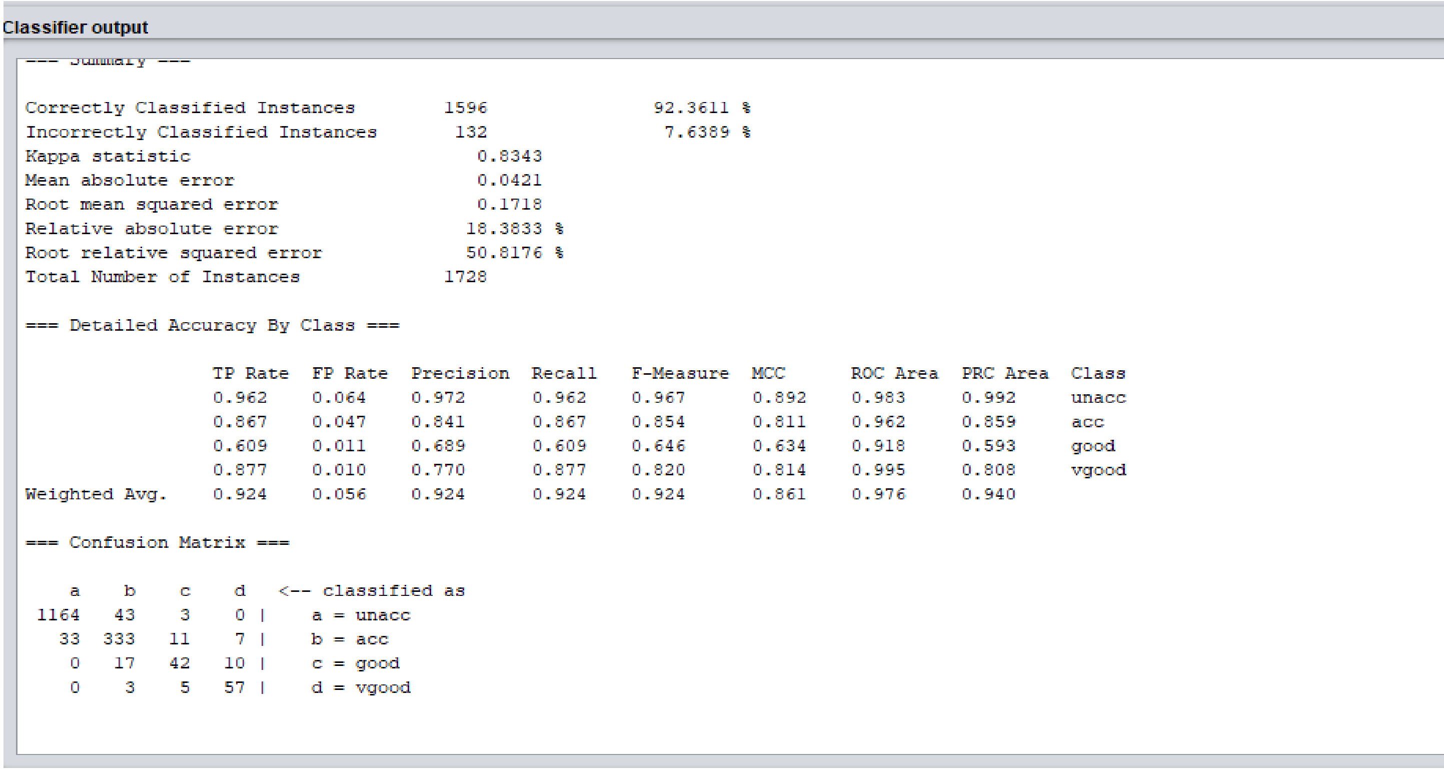
```
=== Summary ===

Correctly Classified Instances        1573               91.0301 %
Incorrectly Classified Instances       155                8.9699 %
Kappa statistic                          0.7987
Mean absolute error                      0.2748
Root mean squared error                  0.322
Relative absolute error                119.9872 %
Root relative squared error             95.2225 %
Total Number of Instances             1728

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.969    0.145    0.940      0.969    0.954      0.844    0.978     0.989     unacc
                 0.802    0.036    0.863      0.802    0.831      0.786    0.967     0.869     acc
                 0.652    0.014    0.662      0.652    0.657      0.643    0.941     0.654     good
                 0.723    0.005    0.855      0.723    0.783      0.779    0.965     0.796     vgood
Weighted Avg.    0.910    0.110    0.908      0.910    0.909      0.820    0.973     0.941

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1173   34    3    0 |    a = unacc
   65  308    9    2 |    b = acc
    8   10   45    6 |    c = good
    2    5   11   47 |    d = vgood
```

KNN Classifier:



```
=== Summary ===

Correctly Classified Instances        1616               93.5185 %
Incorrectly Classified Instances       112                6.4815 %
Kappa statistic                          0.853
Mean absolute error                      0.1122
Root mean squared error                  0.1953
Relative absolute error                 48.9977 %
Root relative squared error             57.7645 %
Total Number of Instances             1728

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.998    0.066    0.973      0.998    0.985      0.949    1.000     1.000     unacc
                 0.911    0.058    0.818      0.911    0.862      0.822    0.988     0.958     acc
                 0.188    0.000    1.000      0.188    0.317      0.427    0.994     0.859     good
                 0.708    0.000    1.000      0.708    0.829      0.836    1.000     1.000     vgood
Weighted Avg.    0.935    0.059    0.940      0.935    0.925      0.896    0.997     0.985

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1207    3    0    0 |    a = unacc
   34  350    0    0 |    b = acc
    0   56   13    0 |    c = good
    0   19    0   46 |    d = vgood
```

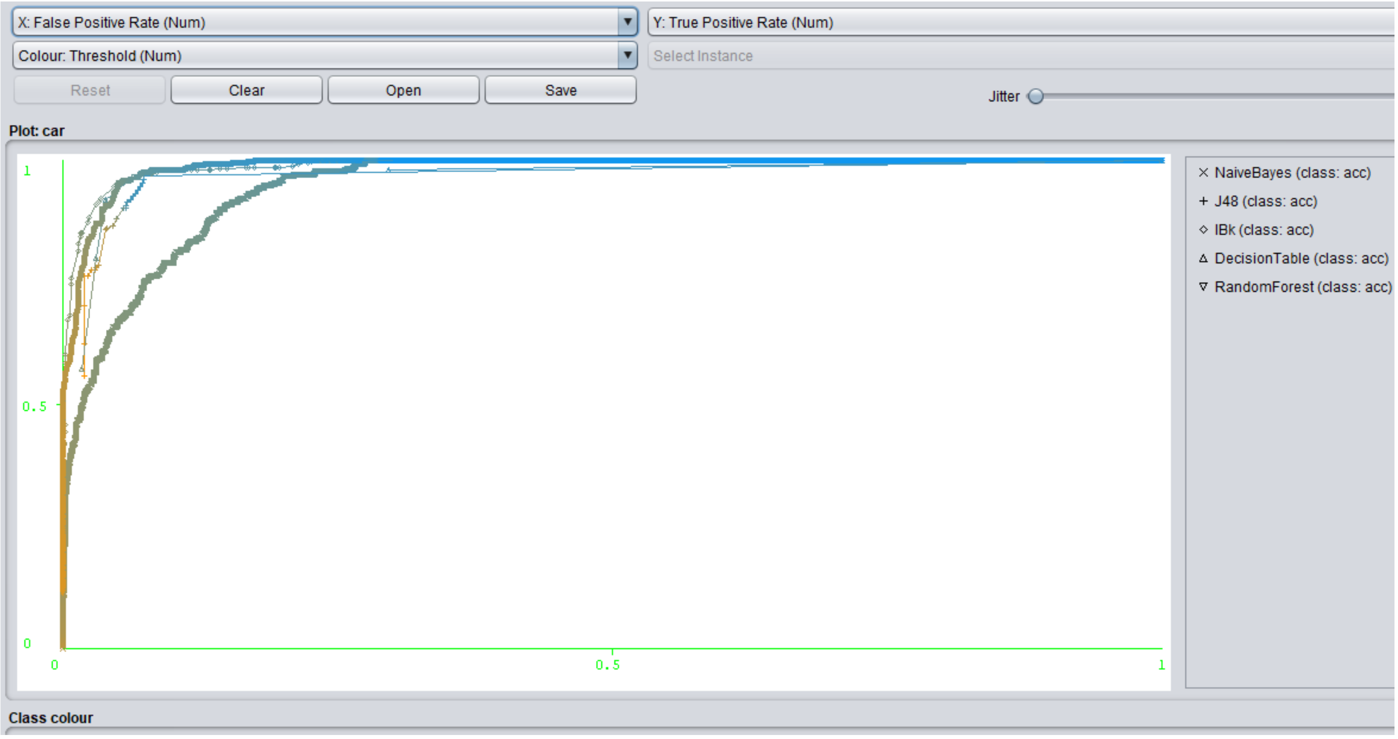
Decision Tree (J48) Classifier:



```
=== Summary ===

Correctly Classified Instances        1596               92.3611 %
Incorrectly Classified Instances       132                7.6389 %
Kappa statistic                          0.8343
Mean absolute error                      0.0421
Root mean squared error                  0.1718
Relative absolute error                 18.3833 %
Root relative squared error             50.8176 %
Total Number of Instances             1728

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.962    0.064    0.972      0.962    0.967      0.892    0.983     0.992     unacc
                 0.867    0.047    0.841      0.867    0.854      0.811    0.962     0.859     acc
                 0.609    0.011    0.689      0.609    0.646      0.634    0.918     0.593     good
                 0.877    0.010    0.770      0.877    0.820      0.814    0.995     0.808     vgood
Weighted Avg.    0.924    0.056    0.924      0.924    0.924      0.861    0.976     0.940

=== Confusion Matrix ===

    a    b    c    d   <-- classified as
 1164   43    3    0 |    a = unacc
   33  333   11    7 |    b = acc
    0   17   42   10 |    c = good
    0    3    5   57 |    d = vgood
```

ROC Curves:

For Class acc:



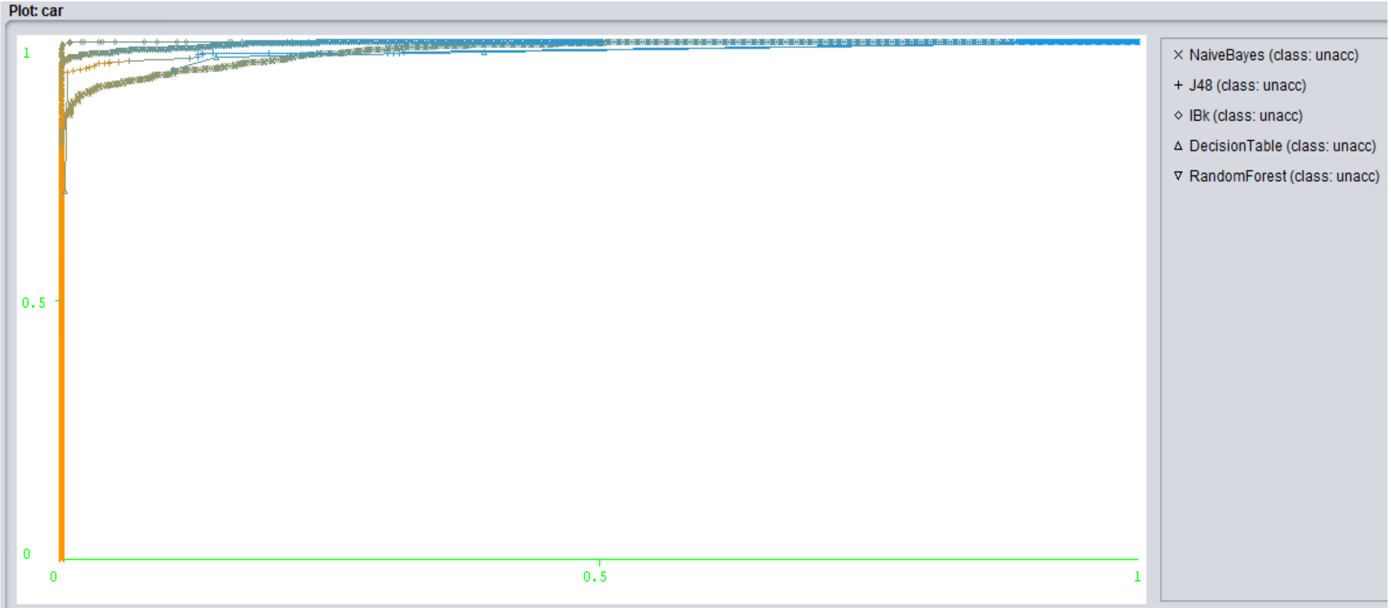
X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

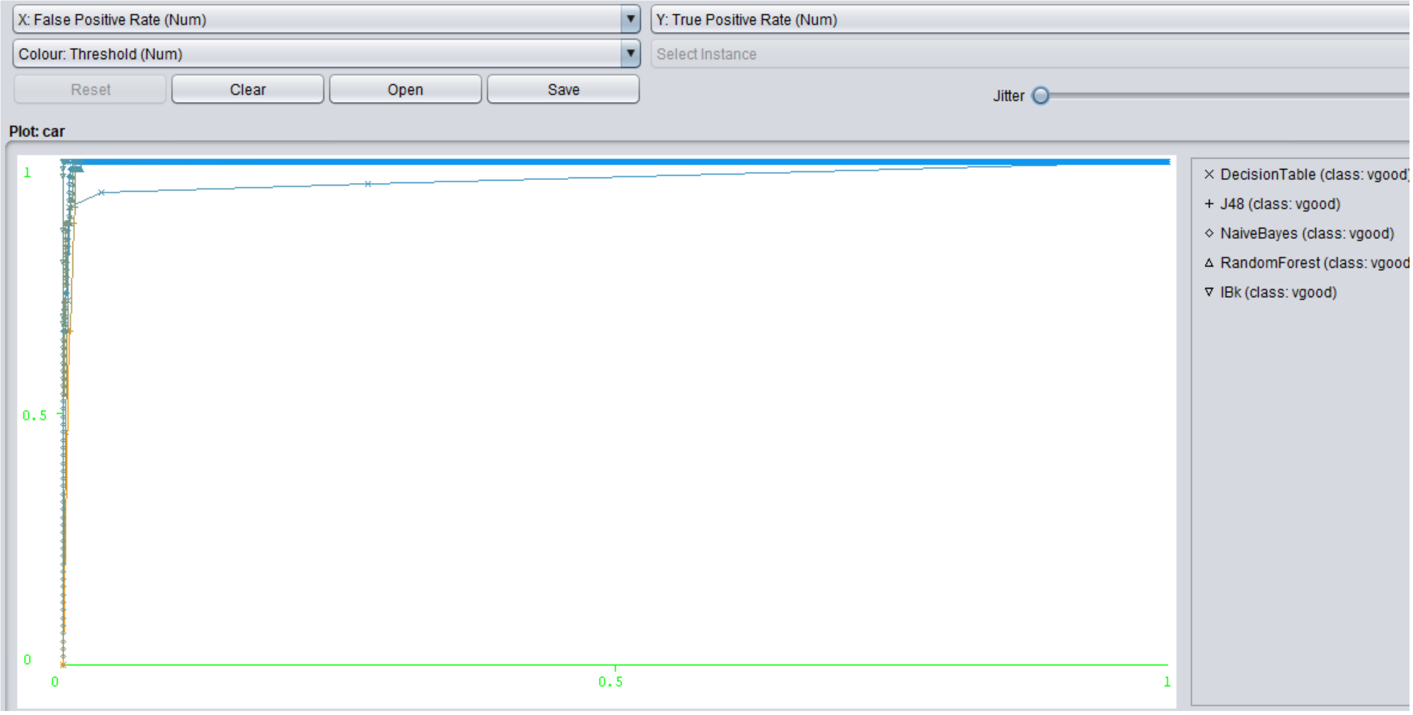
Select Instance

For Class unacc:



Plot: car
1
0.5
0
0
0.5
1
× NaiveBayes (class: unacc)
+ J48 (class: unacc)
◇ IBk (class: unacc)
△ DecisionTable (class: unacc)
▽ RandomForest (class: unacc)

For Class vgood:

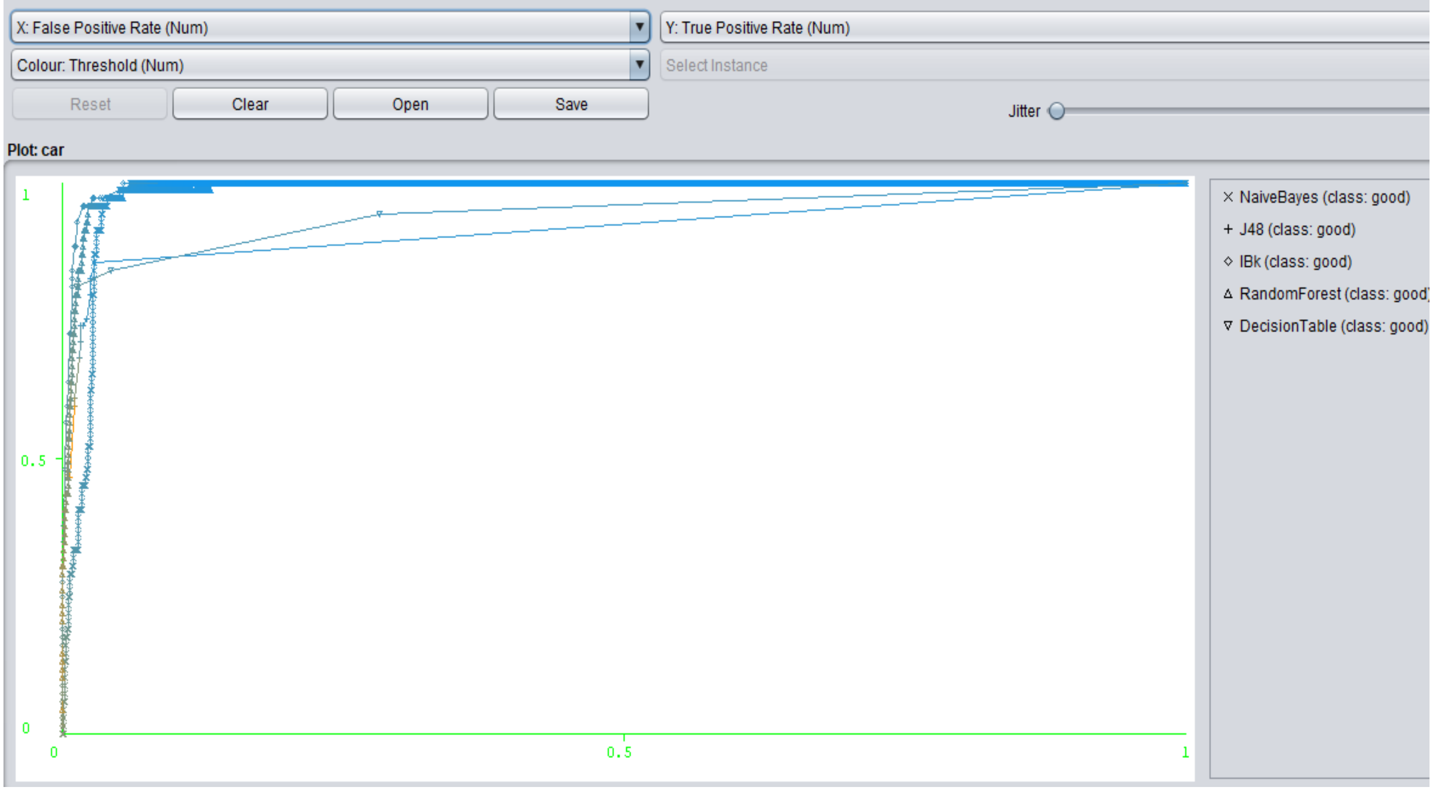


X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

For Class good:



X: False Positive Rate (Num)

Y: True Positive Rate (Num)

Colour: Threshold (Num)

Comments:

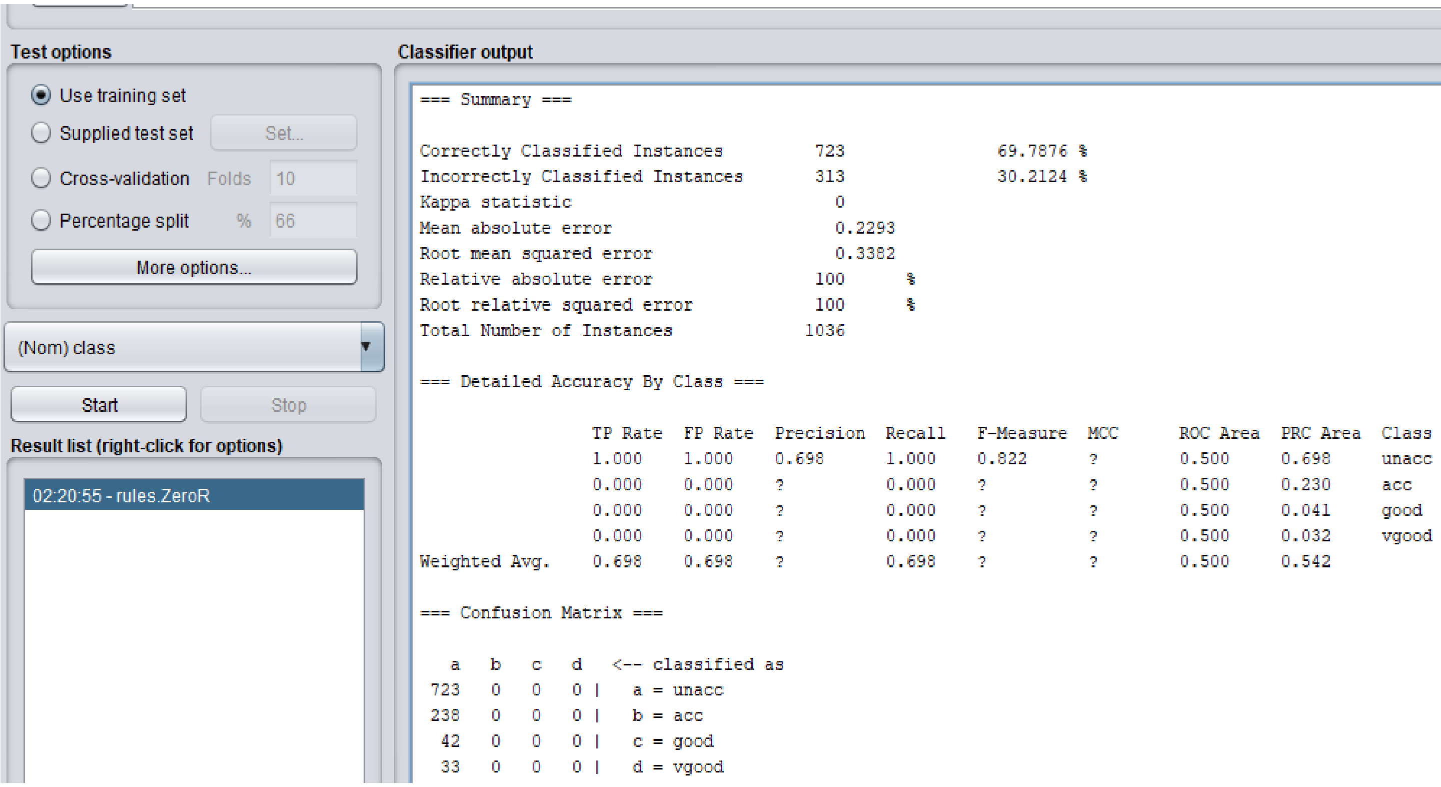
In this report here the weighted average of true positive rate of classifier Random Forest is 0.945 which is the highest and the false positive rate is 0.026 which is the lowest value between these classifiers. It has also the most correctly classified instances which is 94.5023%. The ROC curve of the 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:



Test options

● Use training set
○ Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

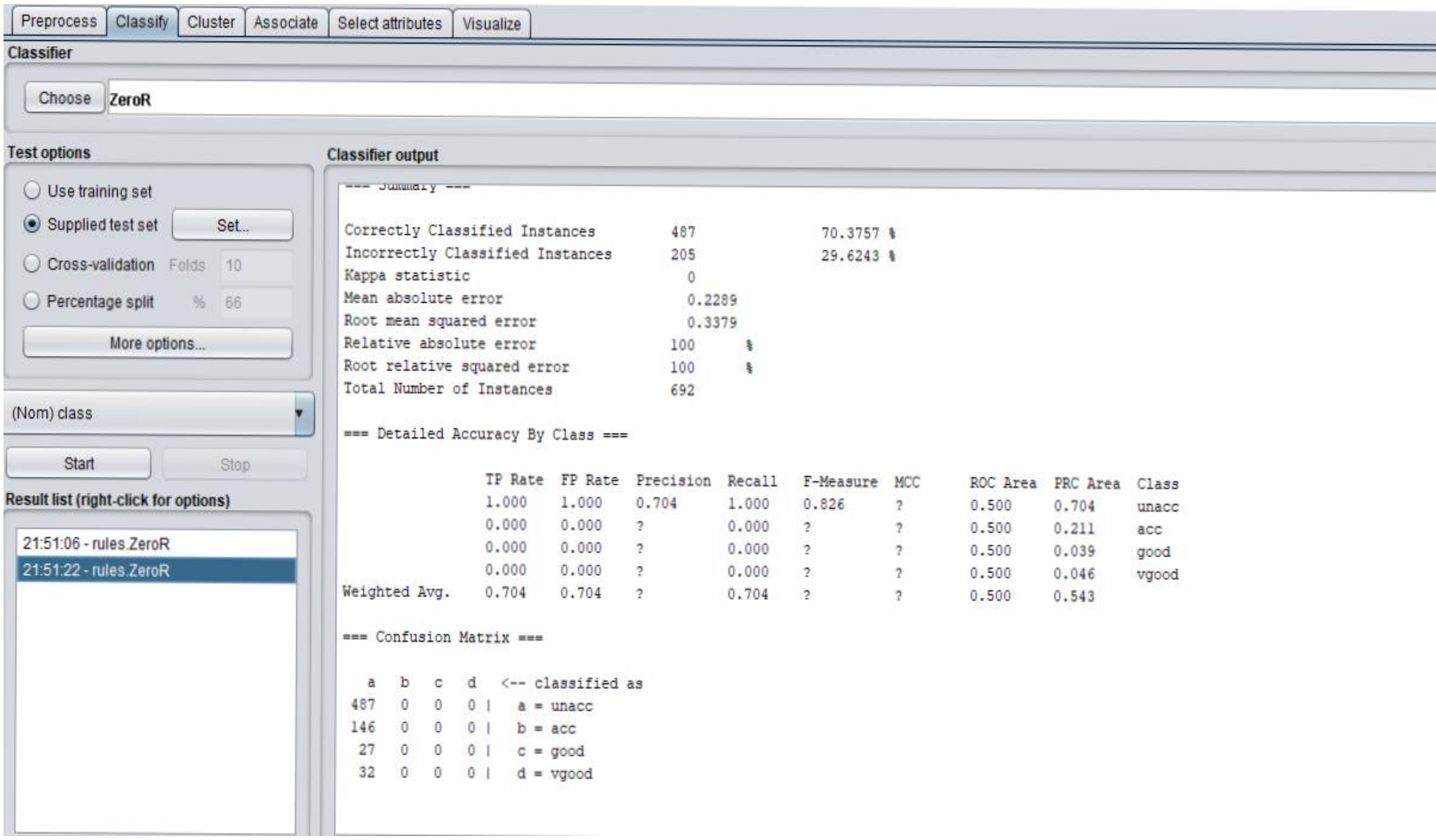
Result list (right-click for options)

02:20:55 - rules.ZeroR

Classifier output

```
=== Summary ===

Correctly Classified Instances         723               69.7876 %
Incorrectly Classified Instances       313               30.2124 %
Kappa statistic                          0
Mean absolute error                      0.2293
Root mean squared error                  0.3382
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances             1036

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.698      1.000    0.822      ?        0.500     0.698     unacc
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.230     acc
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.041     good
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.032     vgood
Weighted Avg.    0.698    0.698    ?          0.698    ?          ?        0.500     0.542

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 723   0   0   0 |   a = unacc
 238   0   0   0 |   b = acc
  42   0   0   0 |   c = good
  33   0   0   0 |   d = vgood
```

For Test Data set:



ZeroR

○ Use training set
● Supplied test set
○ Cross-validation Folds 10
○ Percentage split % 66

(Nom) class

21:51:06 - rules.ZeroR
21:51:22 - rules.ZeroR

```
=== Summary ===

Correctly Classified Instances         487               70.3757 %
Incorrectly Classified Instances       205               29.6243 %
Kappa statistic                          0
Mean absolute error                      0.2289
Root mean squared error                  0.3379
Relative absolute error                100      %
Root relative squared error            100      %
Total Number of Instances              692

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    1.000    0.704      1.000    0.826      ?        0.500     0.704     unacc
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.211     acc
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.039     good
                 0.000    0.000    ?          0.000    ?          ?        0.500     0.046     vgood
Weighted Avg.    0.704    0.704    ?          0.704    ?          ?        0.500     0.543

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 487   0   0   0 |   a = unacc
 146   0   0   0 |   b = acc
  27   0   0   0 |   c = good
  32   0   0   0 |   d = vgood
```

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.