DATA MINING ASSIGNMENT 2

Naive Bayes Classification

TASK 1: Train the training dataset using naive bayes and observe classifier output.

PROCEDURE:

1) Open Weka GUI Chooser.

2) Select EXPLORER present in Applications.

3) Select Preprocess Tab.

4) Go to OPEN file and browse the file that is already stored in the system “credit-g.arff”.

5) Go to Classify tab.

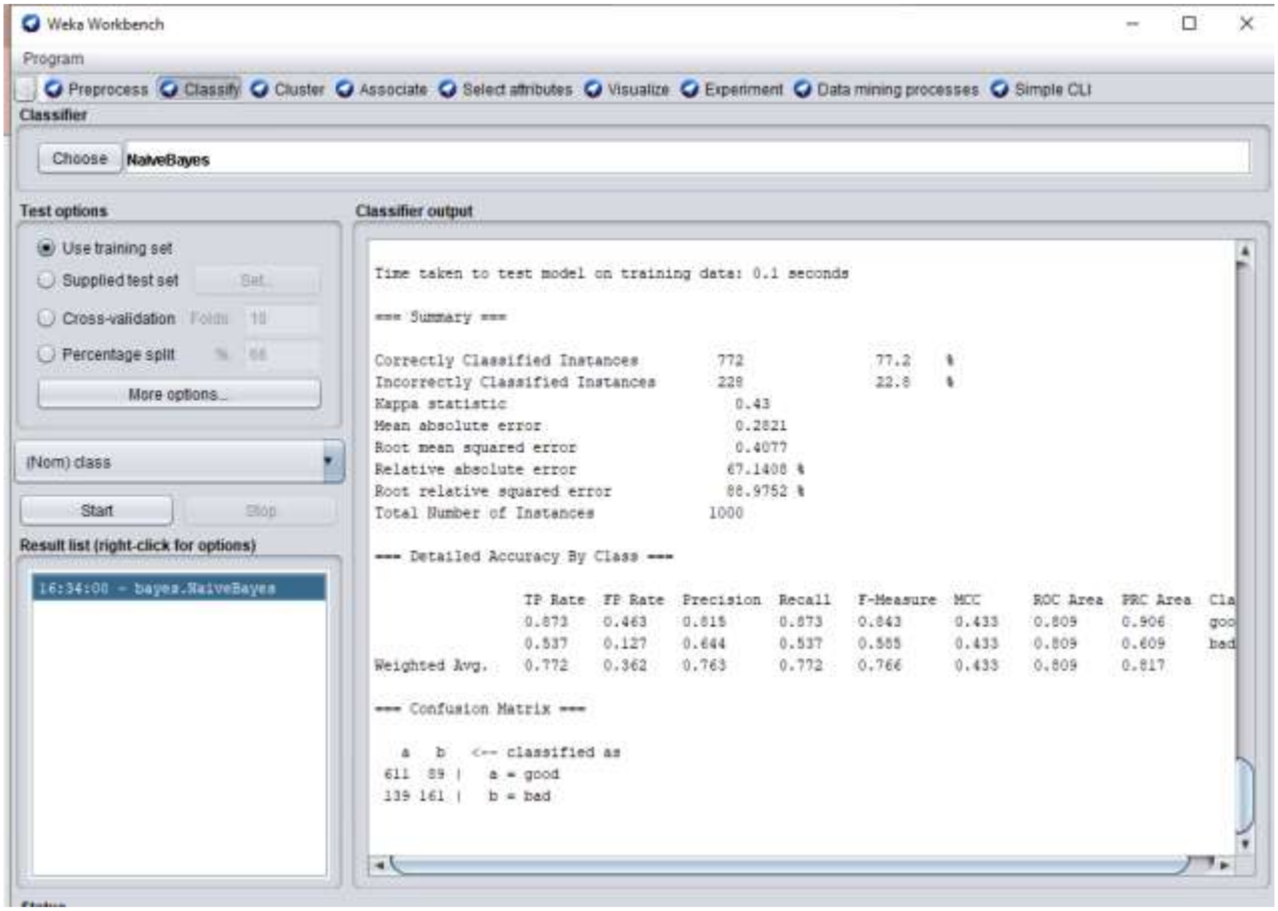
6) Select by clicking the button choose and select bayes NaiveBayes.

7) Select Test options “Use training set”.

8) Select class attribute.

9) Click Start.

10) Now we can see the output details in the Classifier output.



Classifier: NaiveBayes

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

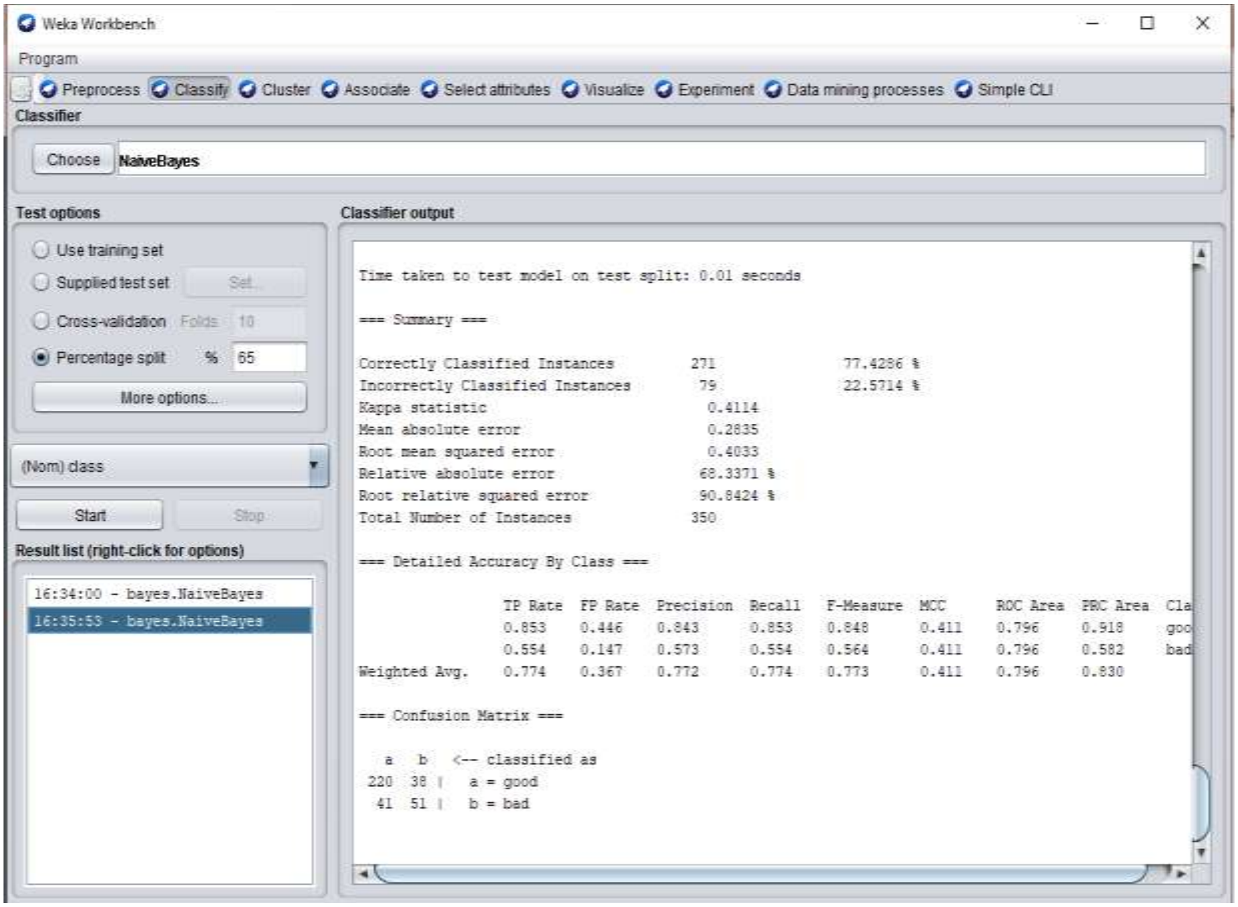
(Nom) class

16:34:00 - bayes.NaiveBayes

```
Time taken to test model on training data: 0.1 seconds

=== Summary ===

Correctly Classified Instances         772               77.2    %
Incorrectly Classified Instances       228               22.8    %
Kappa statistic                          0.43
Mean absolute error                      0.2821
Root mean squared error                  0.4077
Relative absolute error                 67.1408 %
Root relative squared error             88.9752 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.873    0.463    0.815      0.873    0.843      0.433    0.809     0.906     goo
                 0.537    0.127    0.644      0.537    0.585      0.433    0.809     0.609     bad
Weighted Avg.    0.772    0.362    0.763      0.772    0.766      0.433    0.809     0.817

=== Confusion Matrix ===

   a   b   <-- classified as
 611  89 |   a = good
 139 161 |   b = bad
```

TASK 2: Train a Decision Tree using percentage split and report your results. Increase percentage split by 5% upto 80% starting from 65% and check at which percentage split we are getting the best accuracy.

1. When percentage split is 65%, the accuracy is 77.4286%



NaiveBayes

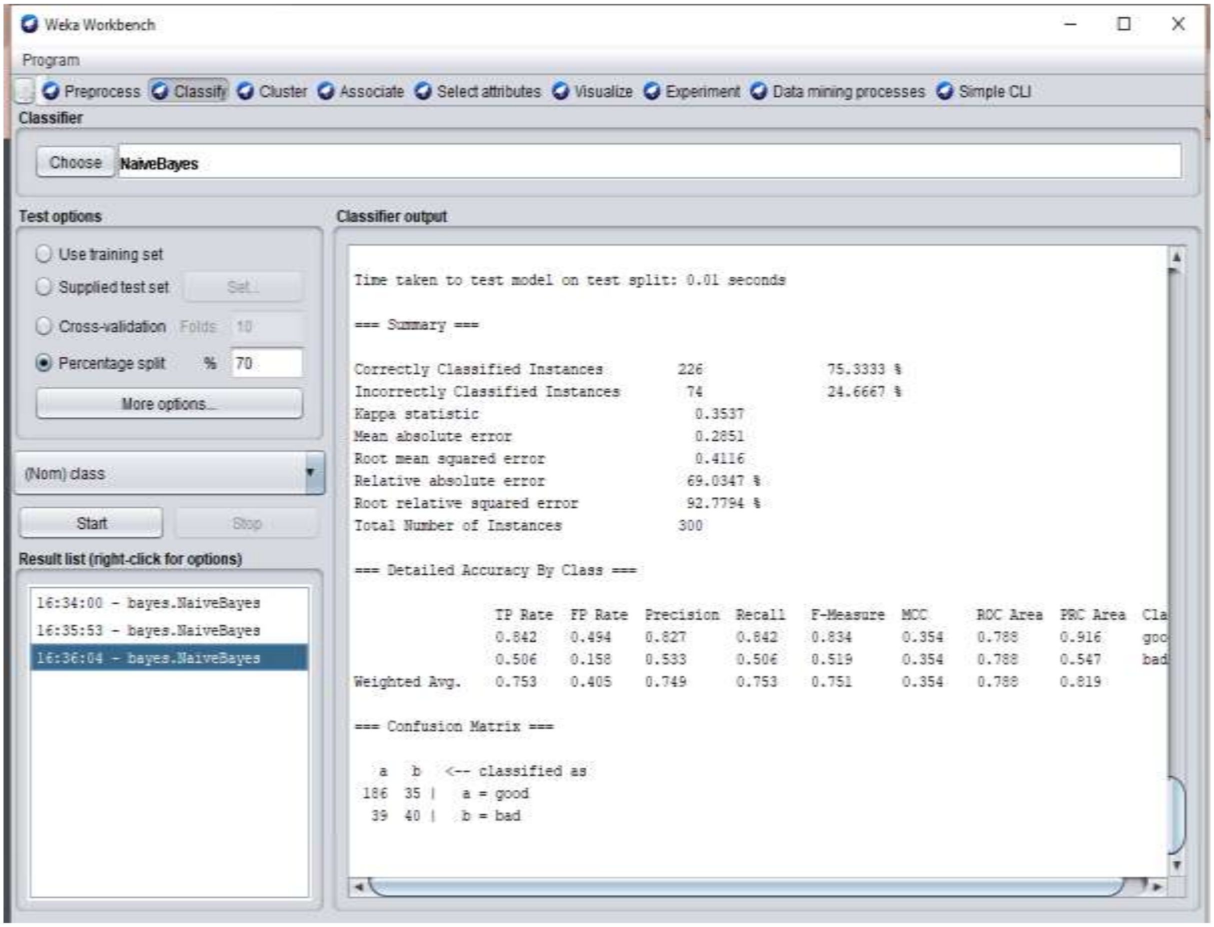
○ Use training set
○ Supplied test set
○ Cross-validation Folds 10
● Percentage split % 65

(Nom) class

16:34:00 - bayes.NaiveBayes
16:35:53 - bayes.NaiveBayes

```
Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         271               77.4286 %
Incorrectly Classified Instances        79               22.5714 %
Kappa statistic                          0.4114
Mean absolute error                      0.2835
Root mean squared error                  0.4033
Relative absolute error                 68.3371 %
Root relative squared error             90.8424 %
Total Number of Instances              350

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.853    0.446    0.843      0.853    0.848      0.411    0.796     0.918     goo
                 0.554    0.147    0.573      0.554    0.564      0.411    0.796     0.582     bad
Weighted Avg.    0.774    0.367    0.772      0.774    0.773      0.411    0.796     0.830

=== Confusion Matrix ===

   a   b   <-- classified as
 220  38 |   a = good
  41  51 |   b = bad
```

2. When percentage split is 70%, the accuracy is 75.3333%



Classifier: NaiveBayes

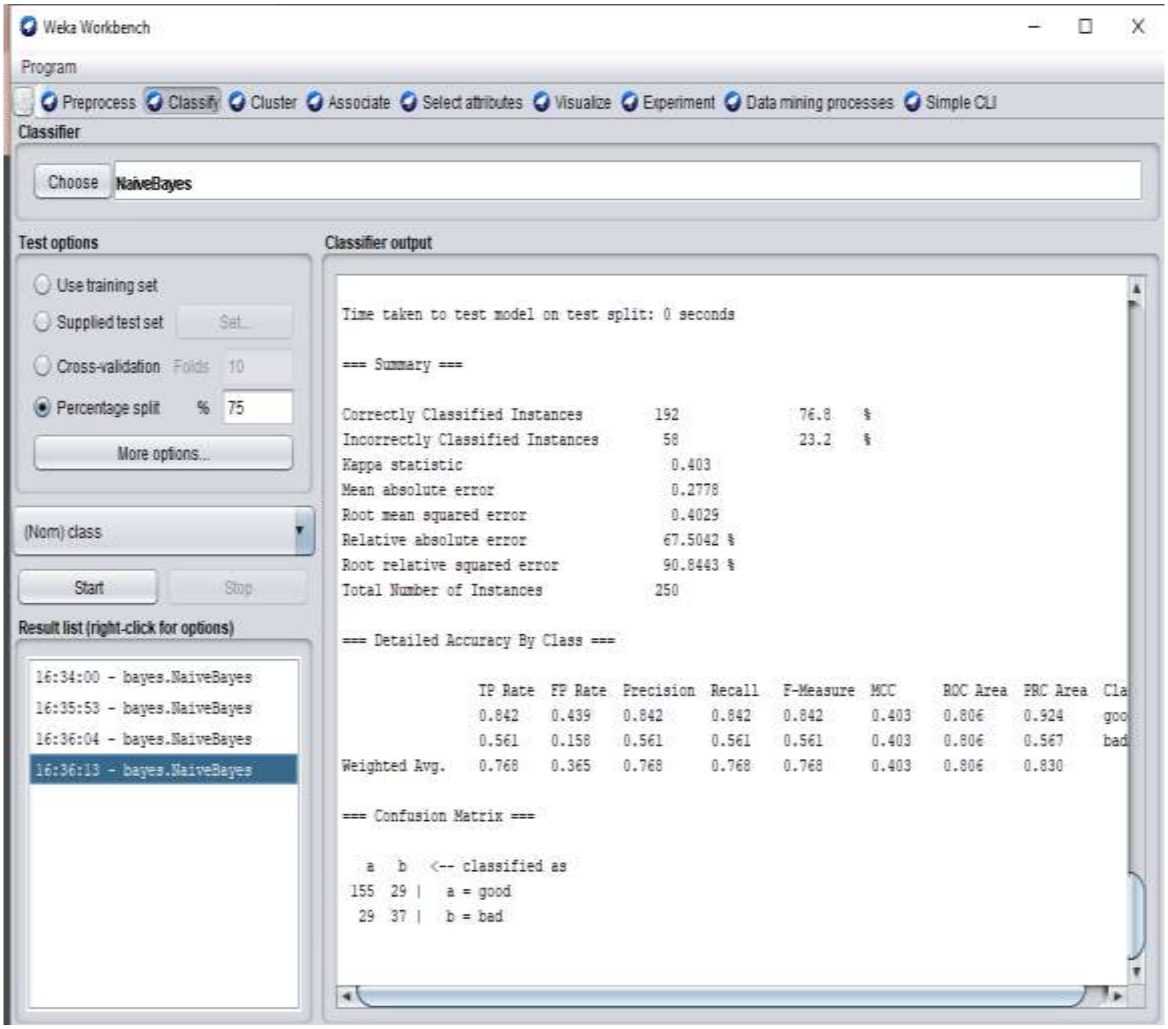
● Percentage split % 70

(Nom) class

```
16:34:00 - bayes.NaiveBayes
16:35:53 - bayes.NaiveBayes
16:36:04 - bayes.NaiveBayes
```

```
Time taken to test model on test split: 0.01 seconds

=== Summary ===

Correctly Classified Instances         226               75.3333 %
Incorrectly Classified Instances        74               24.6667 %
Kappa statistic                          0.3537
Mean absolute error                      0.2851
Root mean squared error                  0.4116
Relative absolute error                 69.0347 %
Root relative squared error             92.7794 %
Total Number of Instances              300

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.842    0.494    0.827      0.842    0.834      0.354    0.788     0.916     goo
                 0.506    0.158    0.533      0.506    0.519      0.354    0.788     0.547     bad
Weighted Avg.    0.753    0.405    0.749      0.753    0.751      0.354    0.788     0.819

=== Confusion Matrix ===

   a   b   <-- classified as
 186  35 |   a = good
  39  40 |   b = bad
```

3. When percentage split is 75%, the accuracy is 76.8%



Classifier: NaiveBayes

● Percentage split % 75

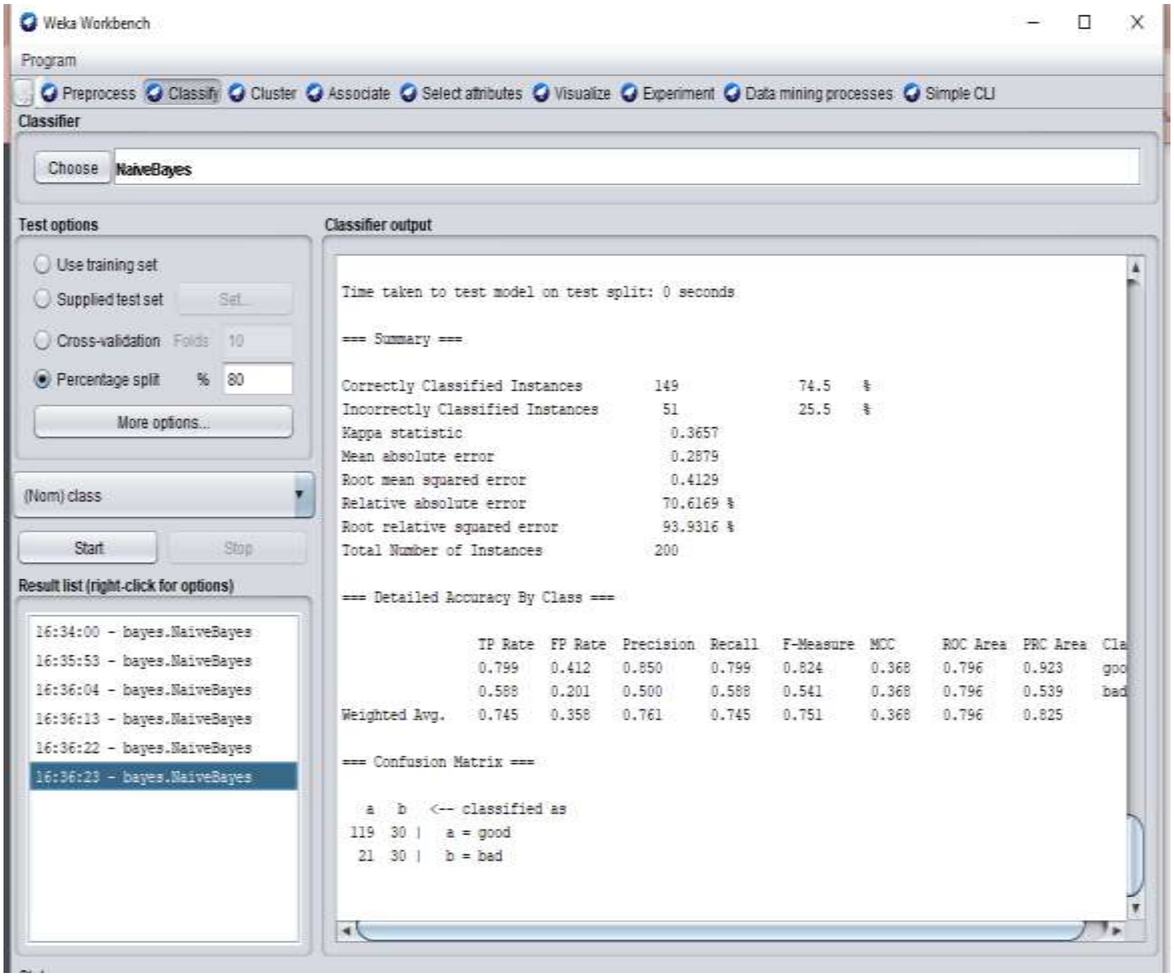
Folds 10

(Nom) class

```
16:34:00 - bayes.NaiveBayes
16:35:53 - bayes.NaiveBayes
16:36:04 - bayes.NaiveBayes
16:36:13 - bayes.NaiveBayes
```

```
Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances        192               76.8    %
Incorrectly Classified Instances       58               23.2    %
Kappa statistic                         0.403
Mean absolute error                     0.2778
Root mean squared error                 0.4029
Relative absolute error                67.5042 %
Root relative squared error            90.8443 %
Total Number of Instances             250

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.842    0.439    0.842      0.842    0.842      0.403    0.806     0.924     goo
                 0.561    0.158    0.561      0.561    0.561      0.403    0.806     0.567     bad
Weighted Avg.    0.768    0.365    0.768      0.768    0.768      0.403    0.806     0.830

=== Confusion Matrix ===

   a   b   <-- classified as
 155  29 |   a = good
  29  37 |   b = bad
```

4. When percentage split is 80%, the accuracy is 74.5%



NaiveBayes

○ Use training set
○ Supplied test set Set...
○ Cross-validation Folds 10
● Percentage split % 80

(Nom) class

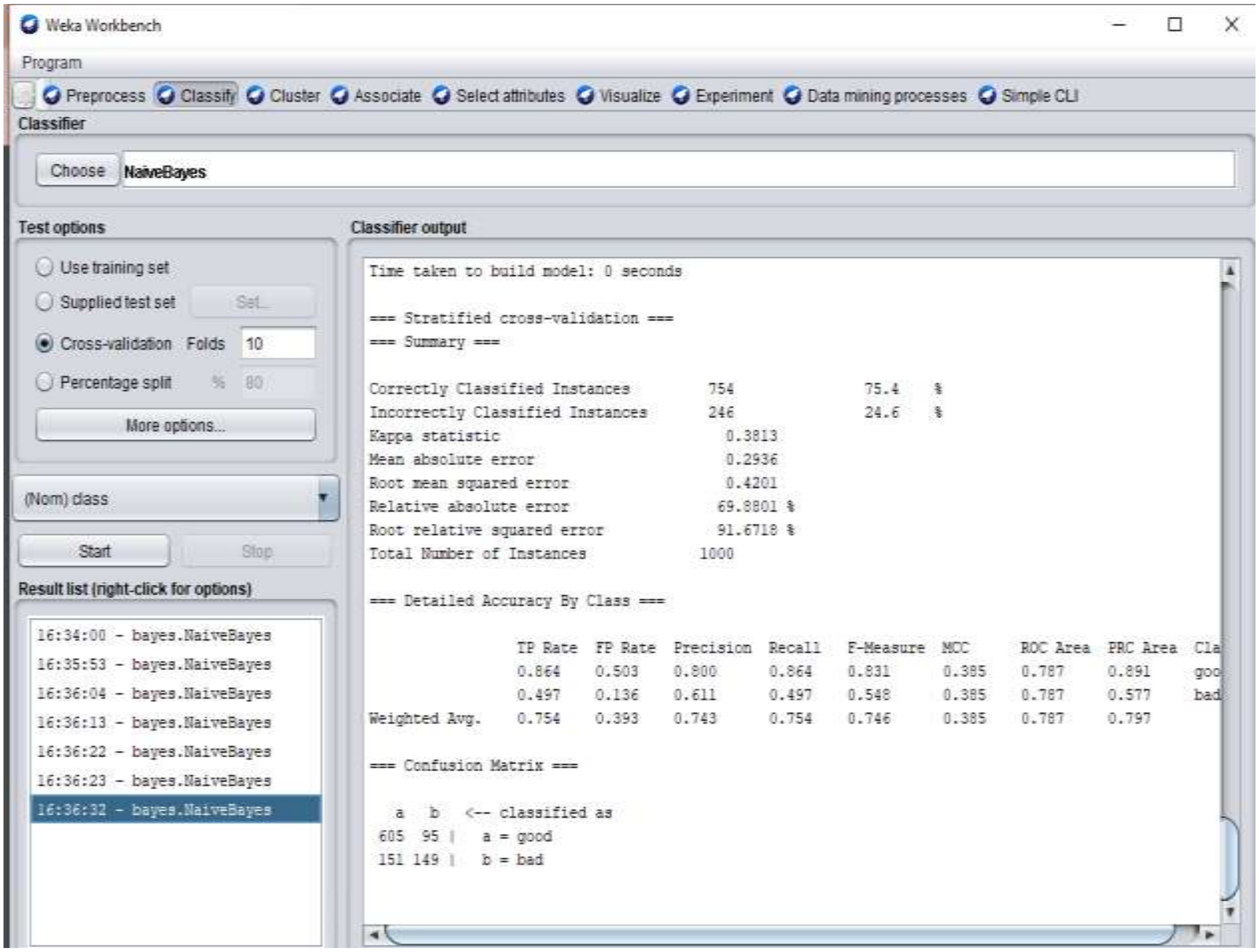
```
16:34:00 - bayes.NaiveBayes
16:35:53 - bayes.NaiveBayes
16:36:04 - bayes.NaiveBayes
16:36:13 - bayes.NaiveBayes
16:36:22 - bayes.NaiveBayes
16:36:23 - bayes.NaiveBayes
```

```
Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         149               74.5    %
Incorrectly Classified Instances        51               25.5    %
Kappa statistic                          0.3657
Mean absolute error                      0.2879
Root mean squared error                  0.4129
Relative absolute error                 70.6169 %
Root relative squared error             93.9316 %
Total Number of Instances              200

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.799    0.412    0.850      0.799    0.824      0.368    0.796     0.923     goo
                 0.588    0.201    0.500      0.588    0.541      0.368    0.796     0.539     bad
Weighted Avg.    0.745    0.358    0.761      0.745    0.751      0.368    0.796     0.825

=== Confusion Matrix ===

   a   b   <-- classified as
 119  30 |   a = good
  21  30 |   b = bad
```

CONCLUSION: When the percentage split is 65%, the accuracy is high(77.4286%).

TASK 3: Train a Decision Tree using cross validation and report your results.

1. When cross validation folds : 10, accuracy is 75.4%



Choose NaiveBayes

● Cross-validation Folds 10

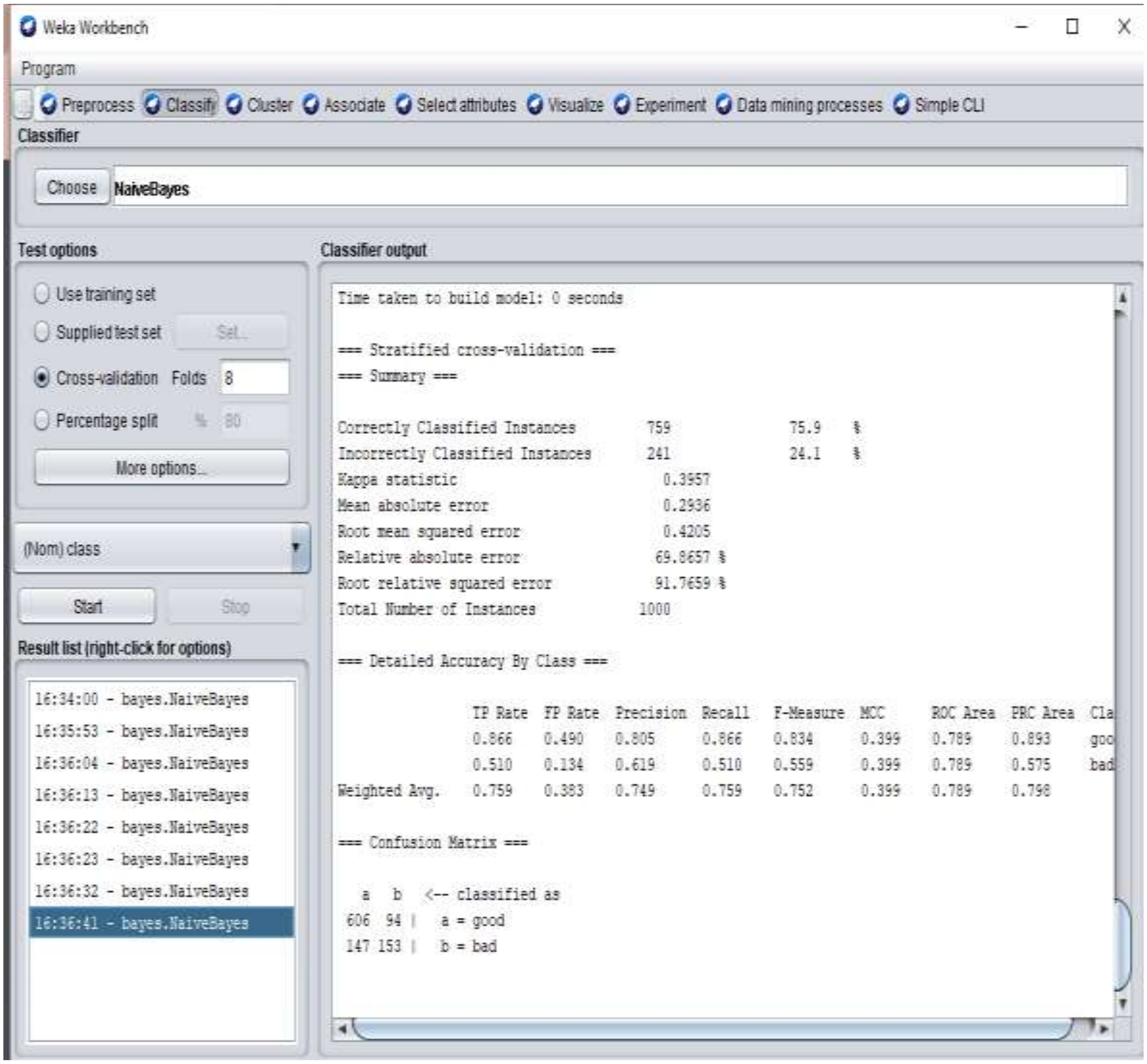
% 80

(Nom) class

```
16:34:00 - bayes.NaiveBayes
16:35:53 - bayes.NaiveBayes
16:36:04 - bayes.NaiveBayes
16:36:13 - bayes.NaiveBayes
16:36:22 - bayes.NaiveBayes
16:36:23 - bayes.NaiveBayes
16:36:32 - bayes.NaiveBayes
```

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         754               75.4    %
Incorrectly Classified Instances       246               24.6    %
Kappa statistic                          0.3813
Mean absolute error                      0.2936
Root mean squared error                  0.4201
Relative absolute error                 69.8801 %
Root relative squared error             91.6718 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.864    0.503    0.800      0.864    0.831      0.385    0.787     0.891     goo
                 0.497    0.136    0.611      0.497    0.548      0.385    0.787     0.577     bad
Weighted Avg.    0.754    0.393    0.743      0.754    0.746      0.385    0.787     0.797

=== Confusion Matrix ===

   a   b   <-- classified as
 605  95 |   a = good
 151 149 |   b = bad
```

2. When cross validation folds : 8, accuracy is 75.9%



Classifier: NaiveBayes

○ Use training set
○ Supplied test set   Set...
● Cross-validation   Folds 8
○ Percentage split   % 80

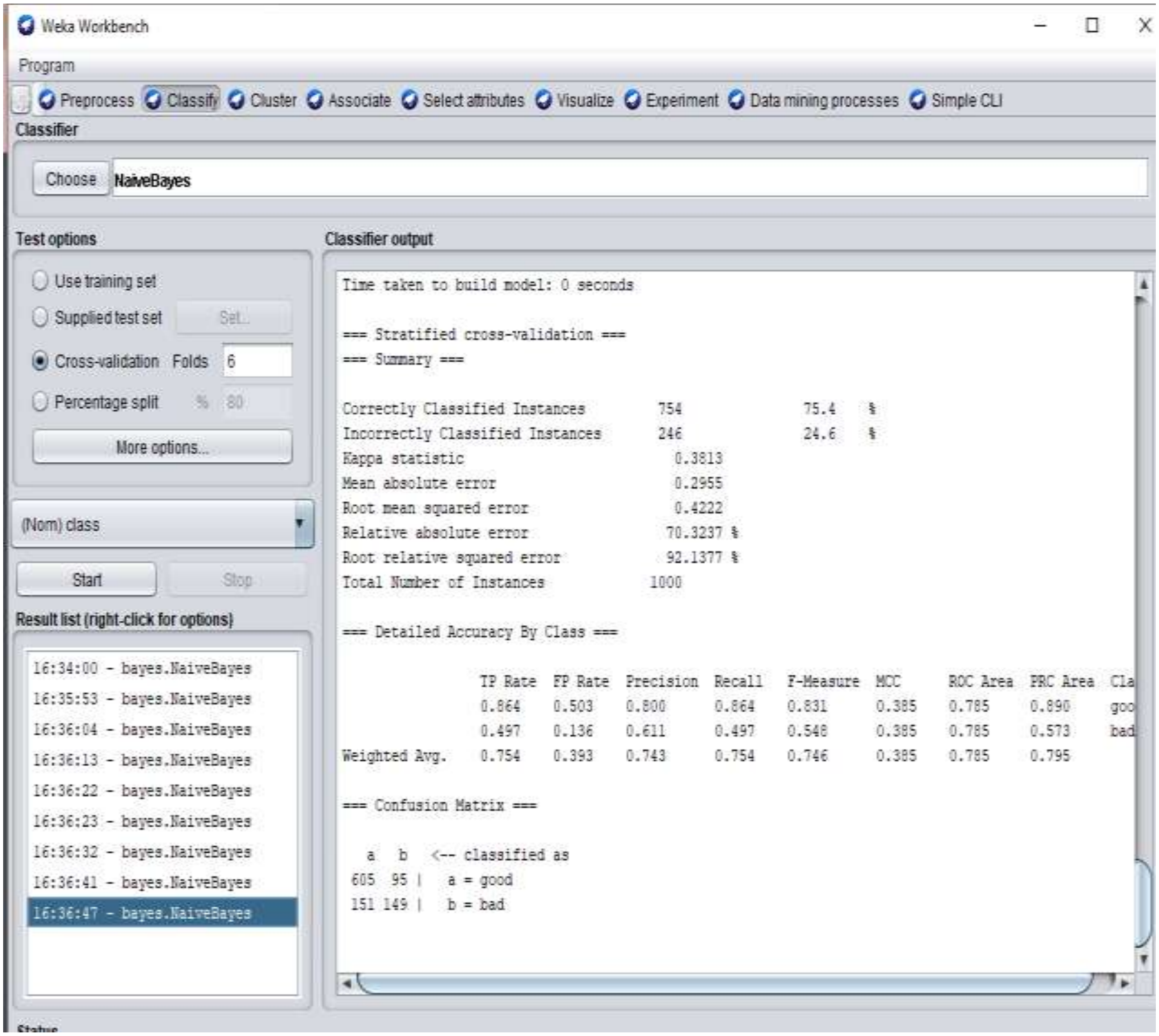
(Nom) class

Result list (right-click for options)

```
16:34:00 - bayes.NaiveBayes
16:35:53 - bayes.NaiveBayes
16:36:04 - bayes.NaiveBayes
16:36:13 - bayes.NaiveBayes
16:36:22 - bayes.NaiveBayes
16:36:23 - bayes.NaiveBayes
16:36:32 - bayes.NaiveBayes
16:36:41 - bayes.NaiveBayes
```

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         759               75.9    %
Incorrectly Classified Instances       241               24.1    %
Kappa statistic                          0.3957
Mean absolute error                      0.2936
Root mean squared error                  0.4205
Relative absolute error                 69.8657 %
Root relative squared error             91.7659 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.866    0.490    0.805      0.866    0.834      0.399    0.789     0.893     goo
                 0.510    0.134    0.619      0.510    0.559      0.399    0.789     0.575     bad
Weighted Avg.    0.759    0.383    0.749      0.759    0.752      0.399    0.789     0.798

=== Confusion Matrix ===

   a   b   <-- classified as
 606  94 |   a = good
 147 153 |   b = bad
```

3. When cross validation folds : 6, accuracy is 75.4%



Classifier: NaiveBayes

○ Use training set
○ Supplied test set
● Cross-validation Folds 6
○ Percentage split % 80

(Nom) class

```
16:34:00 - bayes.NaiveBayes
16:35:53 - bayes.NaiveBayes
16:36:04 - bayes.NaiveBayes
16:36:13 - bayes.NaiveBayes
16:36:22 - bayes.NaiveBayes
16:36:23 - bayes.NaiveBayes
16:36:32 - bayes.NaiveBayes
16:36:41 - bayes.NaiveBayes
16:36:47 - bayes.NaiveBayes
```

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         754               75.4    %
Incorrectly Classified Instances       246               24.6    %
Kappa statistic                          0.3813
Mean absolute error                      0.2955
Root mean squared error                  0.4222
Relative absolute error                 70.3237 %
Root relative squared error             92.1377 %
Total Number of Instances             1000

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0.864    0.503    0.800      0.864    0.831      0.385    0.785     0.890     goo
                 0.497    0.136    0.611      0.497    0.548      0.385    0.785     0.573     bad
Weighted Avg.    0.754    0.393    0.743      0.754    0.746      0.385    0.785     0.795

=== Confusion Matrix ===

   a   b   <-- classified as
 605  95 |   a = good
 151 149 |   b = bad
```

CONCLUSION: The accuracy is high(75.9%) when cross validation folds: 8