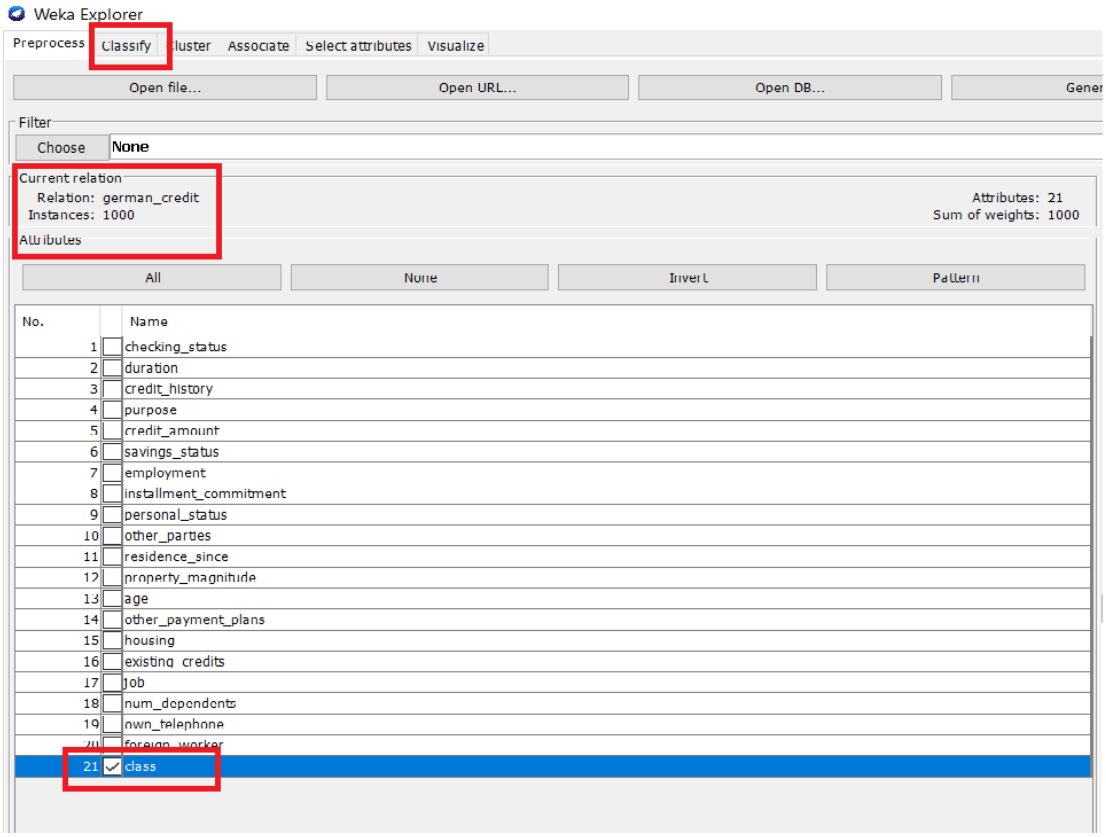
Data Mining Assignment – II

German Credit Dataset:

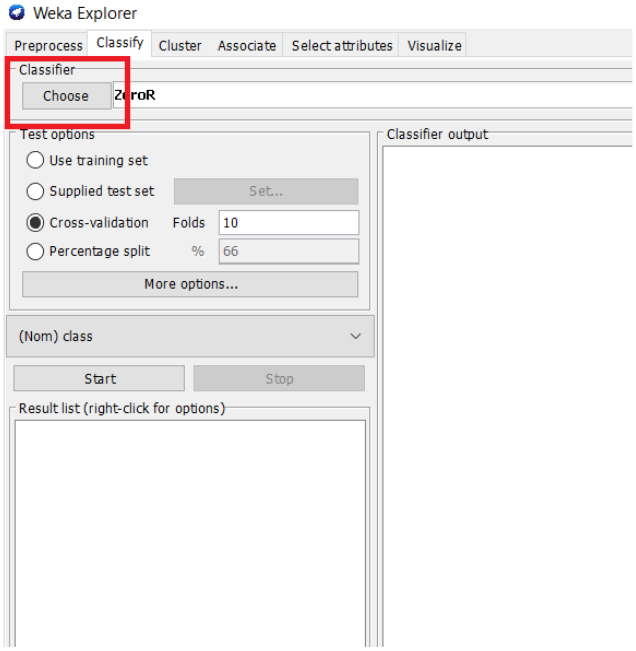
Steps:

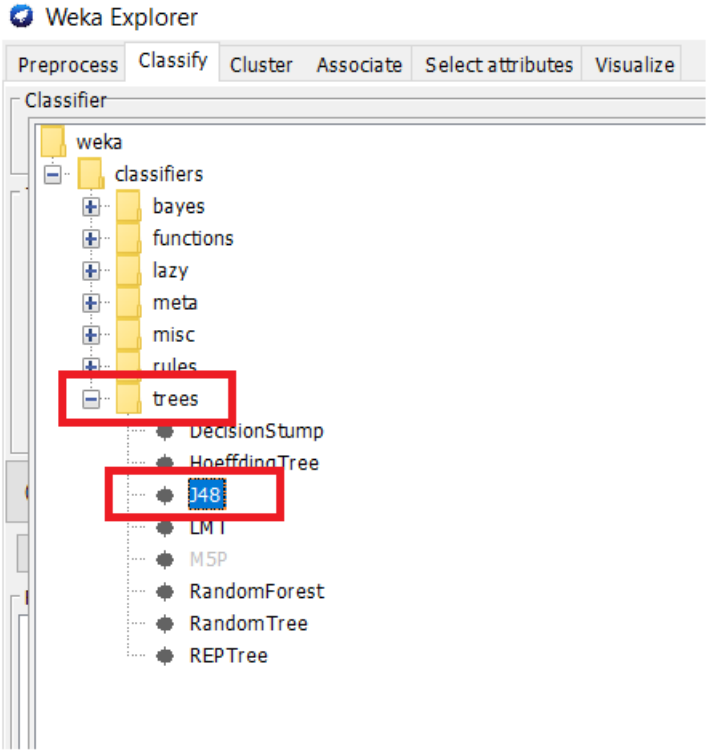
Open german credit dataset in weka. Select class attribute and then click classify.



Decision Tree:

Click choose option in the classify, select trees and then select J48, as shown below.

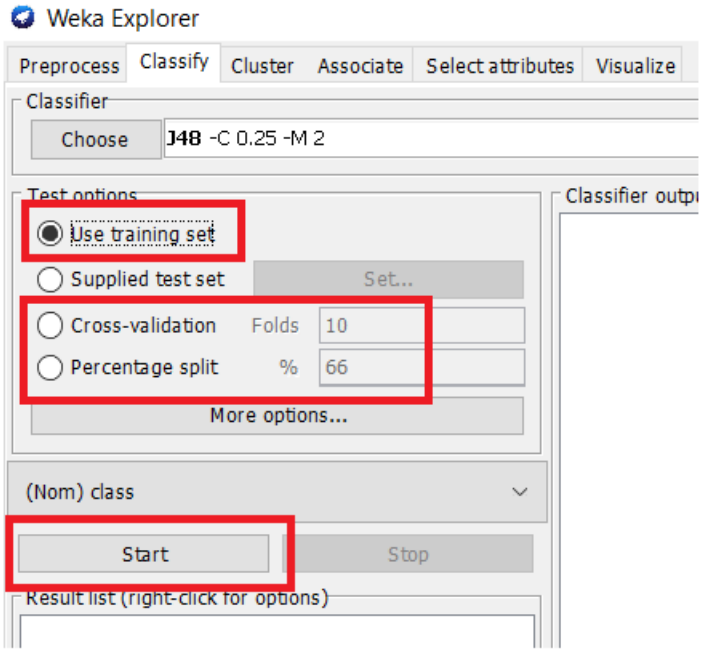


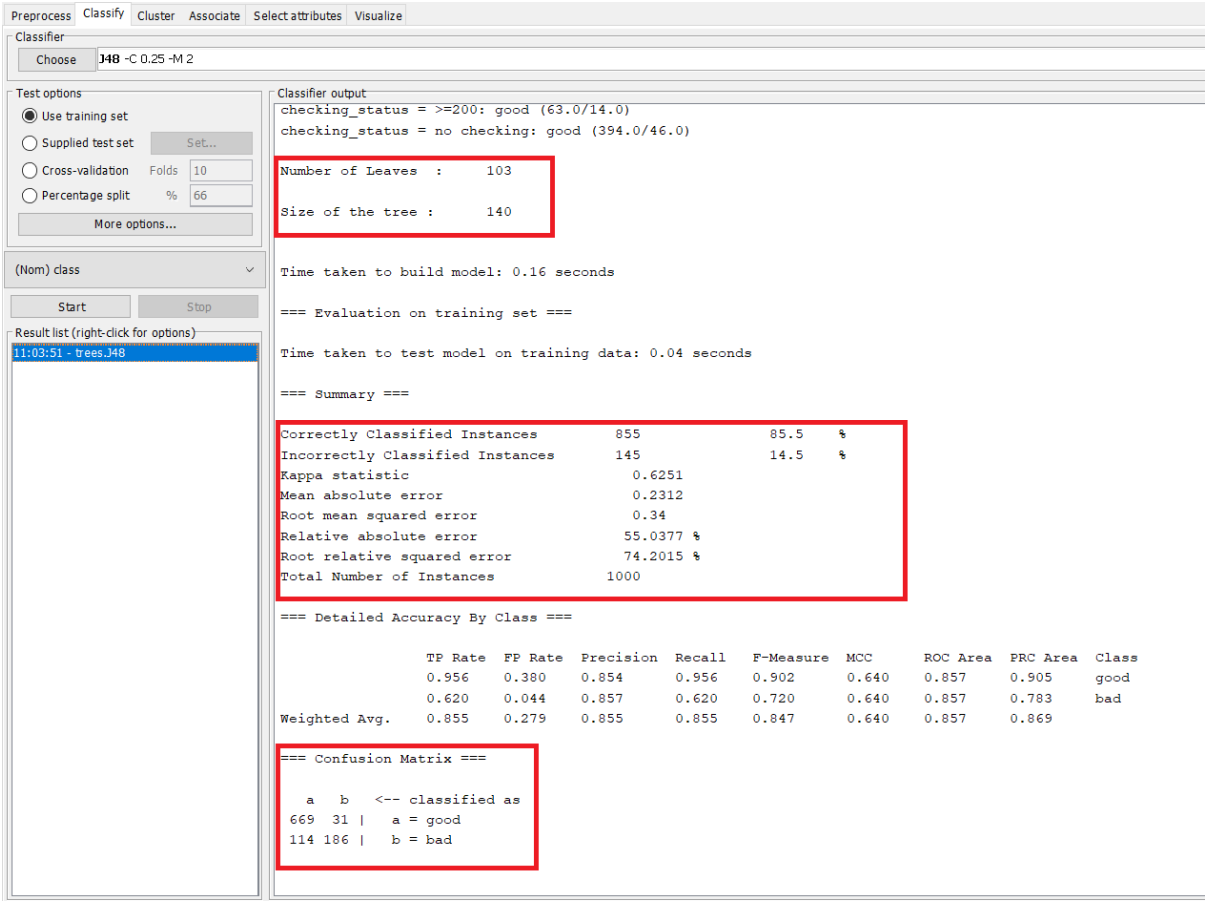


In test options select use training set and the start.

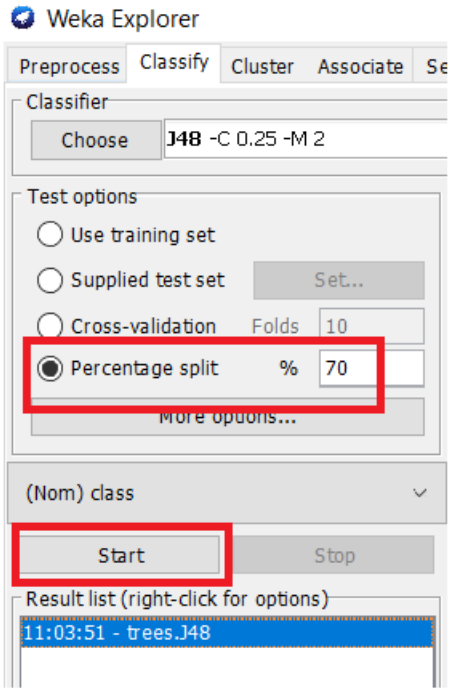
The number of leaves in tree are 103 and the size of the tree is 140.

Accuracy is 85.5% when percentage split is 66%.


Weka Explorer
Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
Classifier outpu
(Nom) class
Start
Stop
Result list (right-click for options)


Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
J48 -C 0.25 -M 2
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) class
Start
Stop
Result list (right-click for options)
11:03:51 - trees.J48
Classifier output
checking_status = >=200: good (63.0/14.0)
checking_status = no checking: good (394.0/46.0)
Number of Leaves : 103
Size of the tree : 140
Time taken to build model: 0.16 seconds
=== Evaluation on training set ===
Time taken to test model on training data: 0.04 seconds
=== Summary ===
Correctly Classified Instances 855 85.5 %
Incorrectly Classified Instances 145 14.5 %
Kappa statistic 0.6251
Mean absolute error 0.2312
Root mean squared error 0.34
Relative absolute error 55.0377 %
Root relative squared error 74.2015 %
Total Number of Instances 1000
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC Area Class
0.956 0.380 0.854 0.956 0.902 0.640 0.857 0.905 good
0.620 0.044 0.857 0.620 0.720 0.640 0.857 0.783 bad
Weighted Avg. 0.855 0.279 0.855 0.855 0.847 0.640 0.857 0.869
=== Confusion Matrix ===
a b <-- classified as
669 31 | a = good
114 186 | b = bad

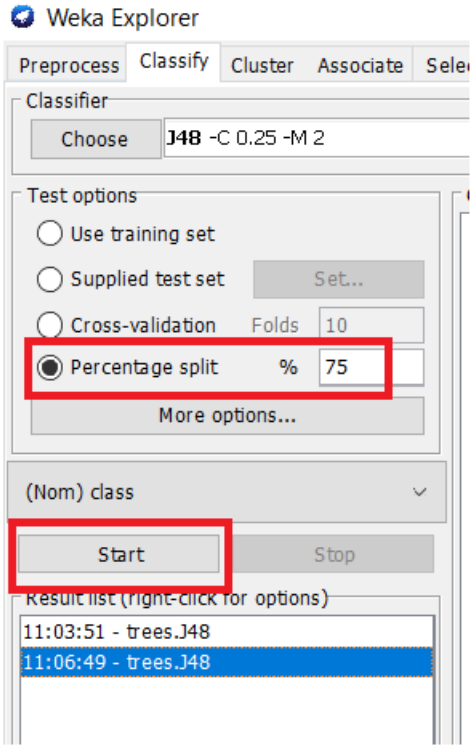
Accuracy is 73.66% when percentage split is 70%.



Classifier output

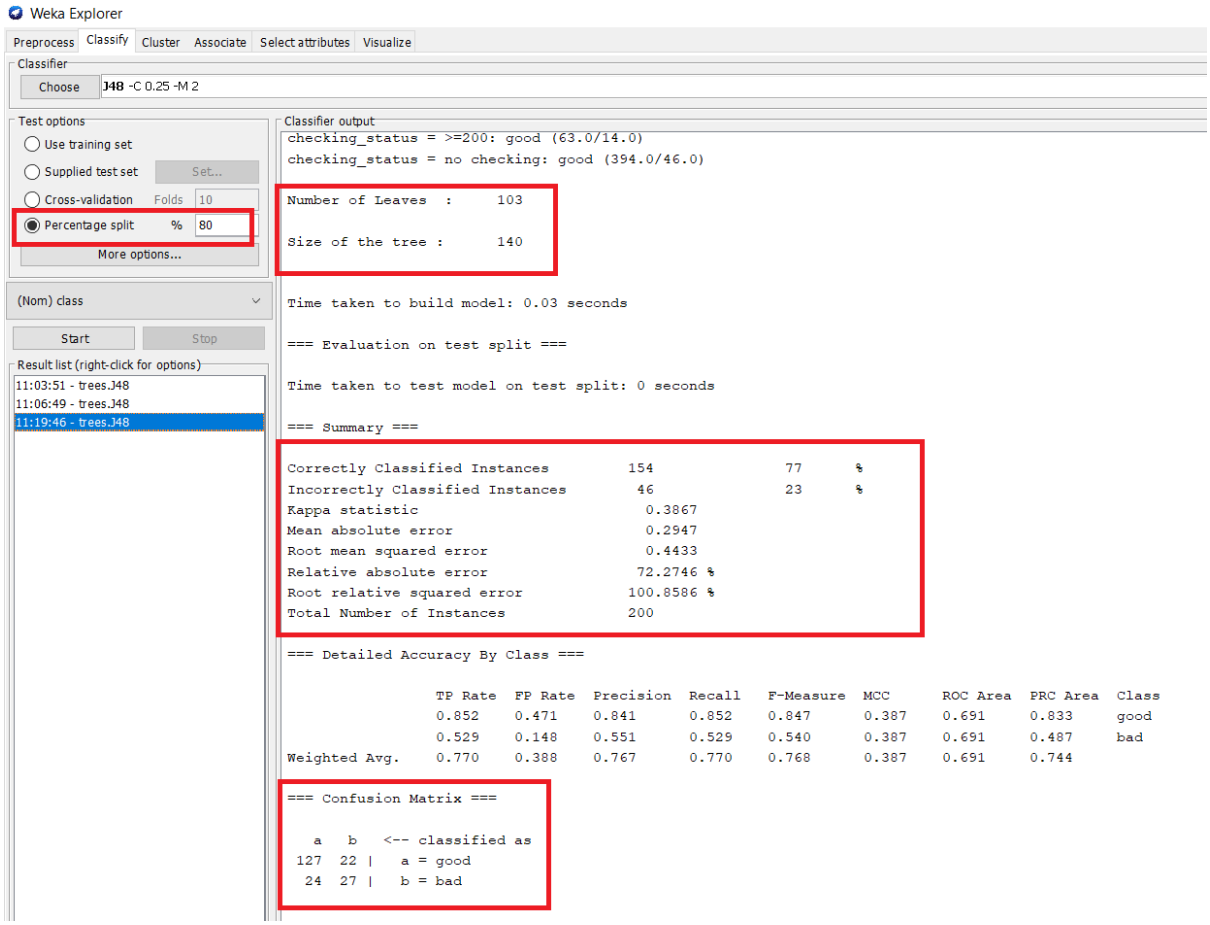
```
checking_status = >=200: good (63.0/14.0)
checking_status = no checking: good (394.0/46.0)

Number of Leaves  :     103

Size of the tree :      140


Time taken to build model: 0.03 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         221               73.6667 %
Incorrectly Classified Instances        79               26.3333 %
Kappa statistic                          0.2579
Mean absolute error                      0.323
Root mean squared error                  0.47
Relative absolute error                 78.2126 %
Root relative squared error            105.9524 %
Total Number of Instances              300

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.869    0.633    0.793      0.869    0.829      0.263    0.636     0.794     good
                 0.367    0.131    0.500      0.367    0.423      0.263    0.636     0.424     bad
Weighted Avg.    0.737    0.501    0.716      0.737    0.722      0.263    0.636     0.696

=== Confusion Matrix ===

   a   b   <-- classified as
 192  29 |   a = good
  50  29 |   b = bad
```

Accuracy is 76% when percentage split is 75%.

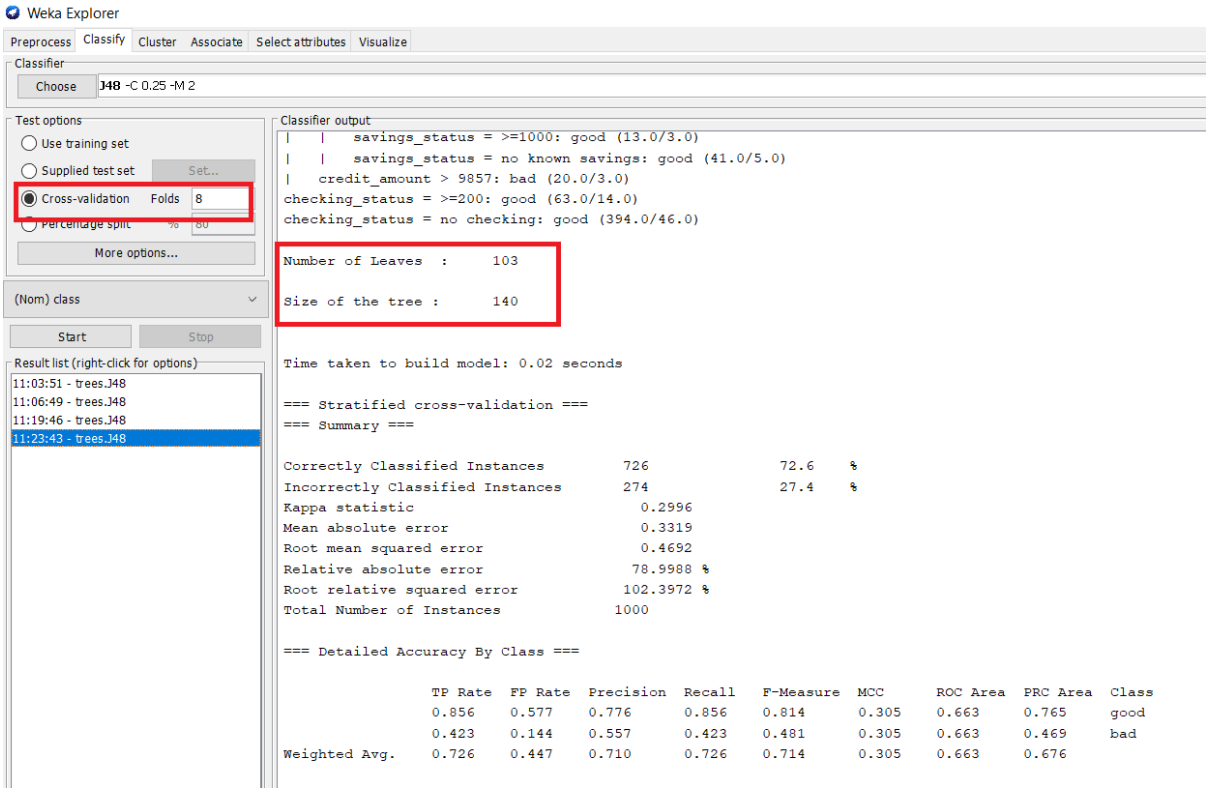



```
Number of Leaves  :     103

Size of the tree :      140


Time taken to build model: 0.01 seconds

=== Evaluation on test split ===

Time taken to test model on test split: 0 seconds

=== Summary ===

Correctly Classified Instances         190               76       %
Incorrectly Classified Instances        60               24       %
Kappa statistic                          0.3232
Mean absolute error                      0.3073
Root mean squared error                  0.4365
Relative absolute error                 74.6884 %
Root relative squared error             98.4212 %
Total Number of Instances              250

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.886    0.591    0.807      0.886    0.845      0.330    0.673     0.820     good
                 0.409    0.114    0.563      0.409    0.474      0.330    0.673     0.478     bad
Weighted Avg.    0.760    0.465    0.742      0.760    0.747      0.330    0.673     0.730

=== Confusion Matrix ===

   a   b   <-- classified as
 163  21 |   a = good
  39  27 |   b = bad
```

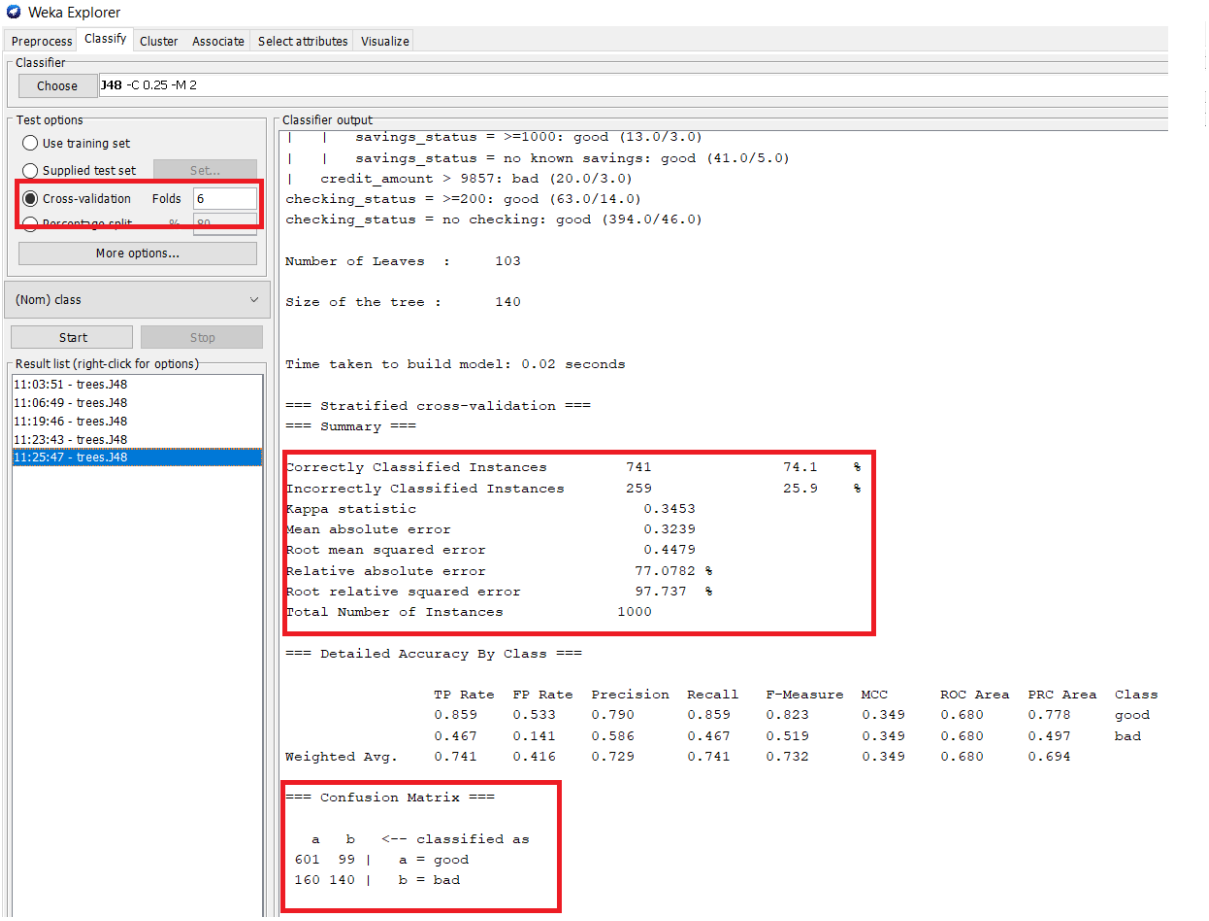
Accuracy is 77% when percentage split is 80%.



Accuracy is 72.6% when cross validation is 8.

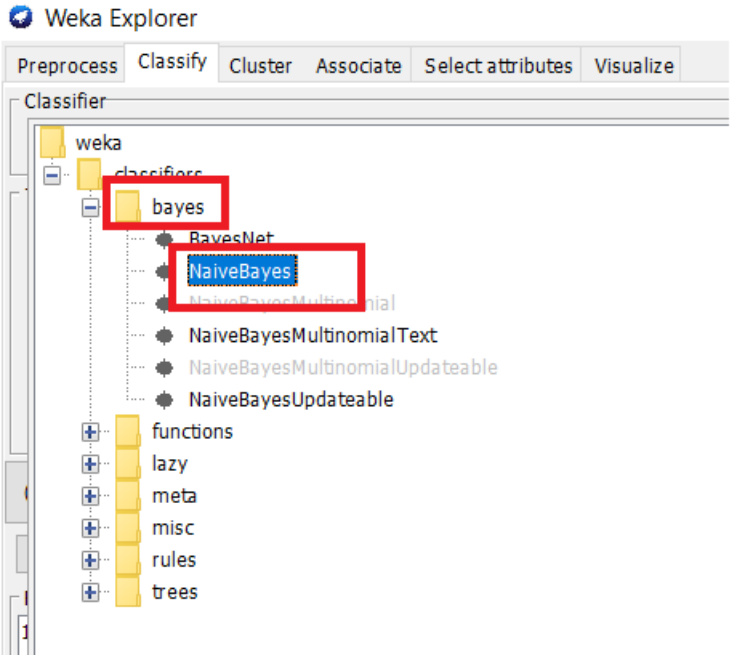


Accuracy is 74.1% when cross validation is 6.

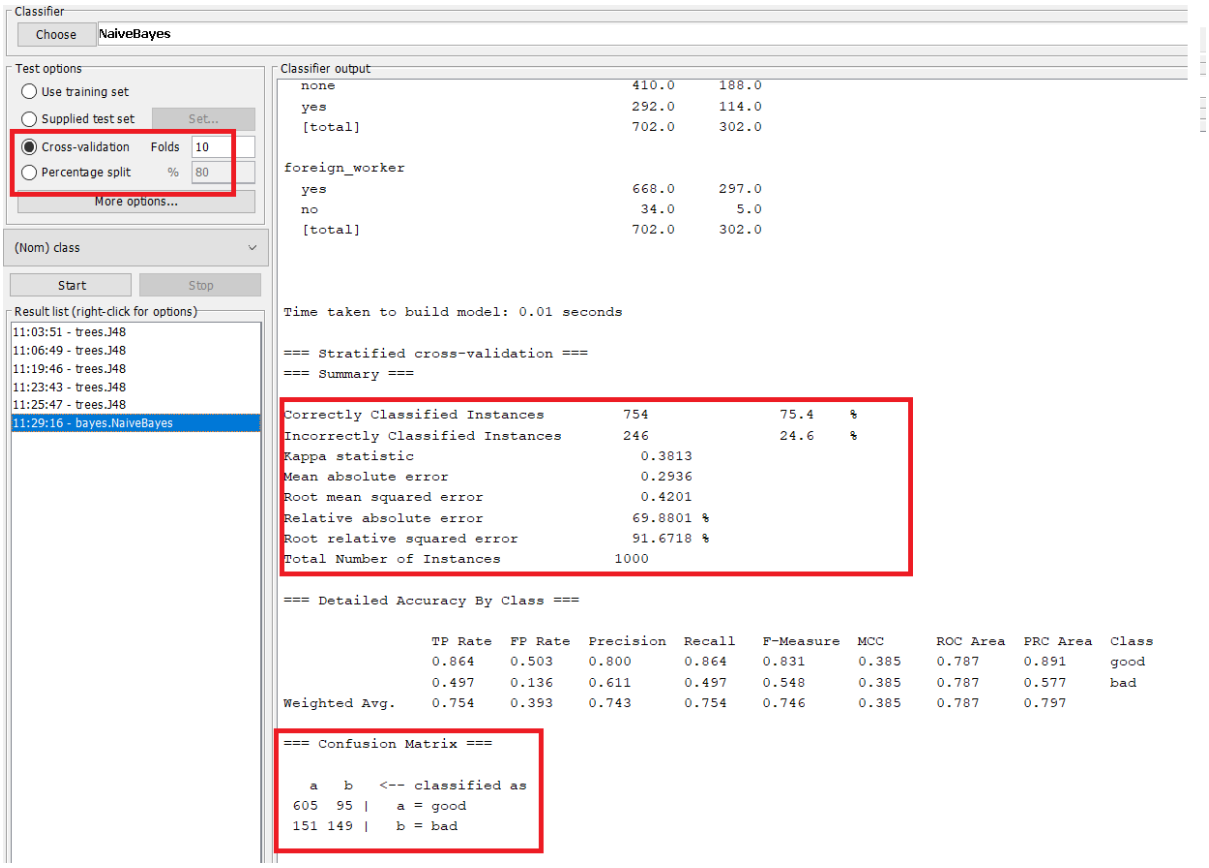


Naïve Bayes:

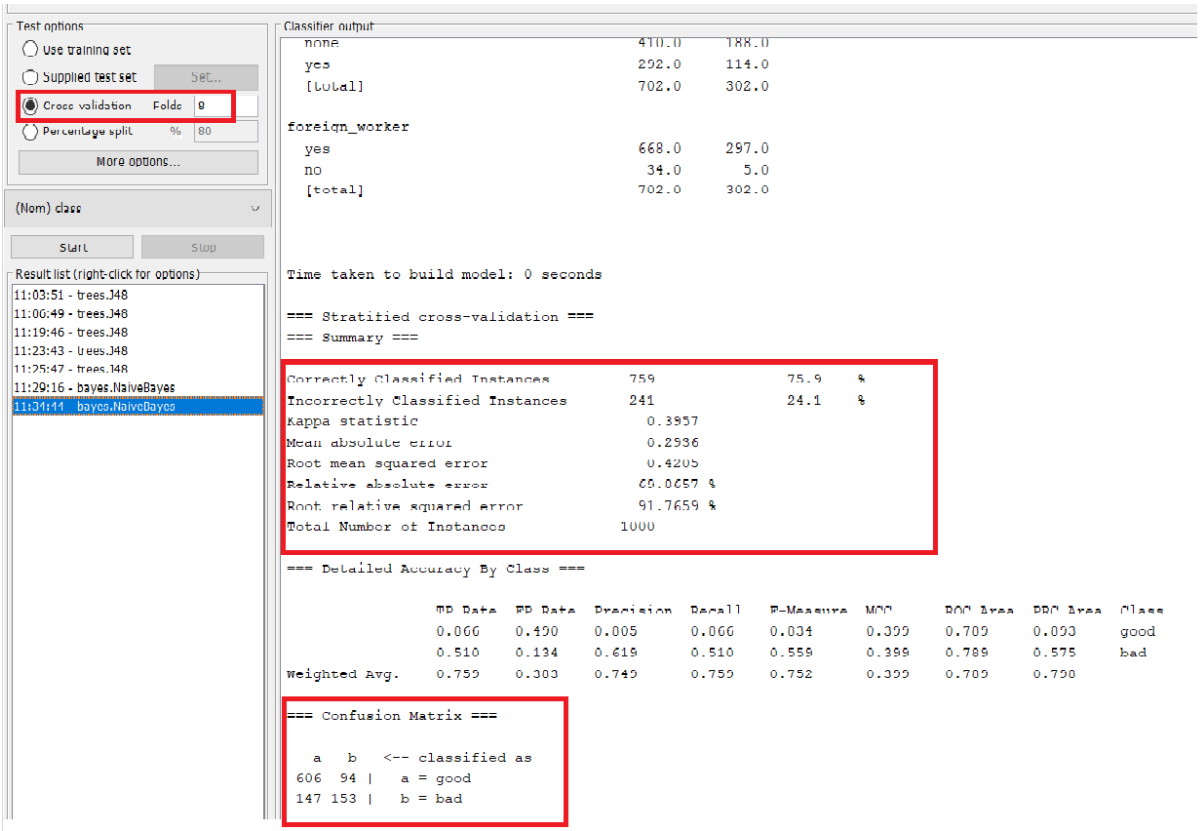
Click choose option in the classify, select bayes and then select naivebayes, as shown below.



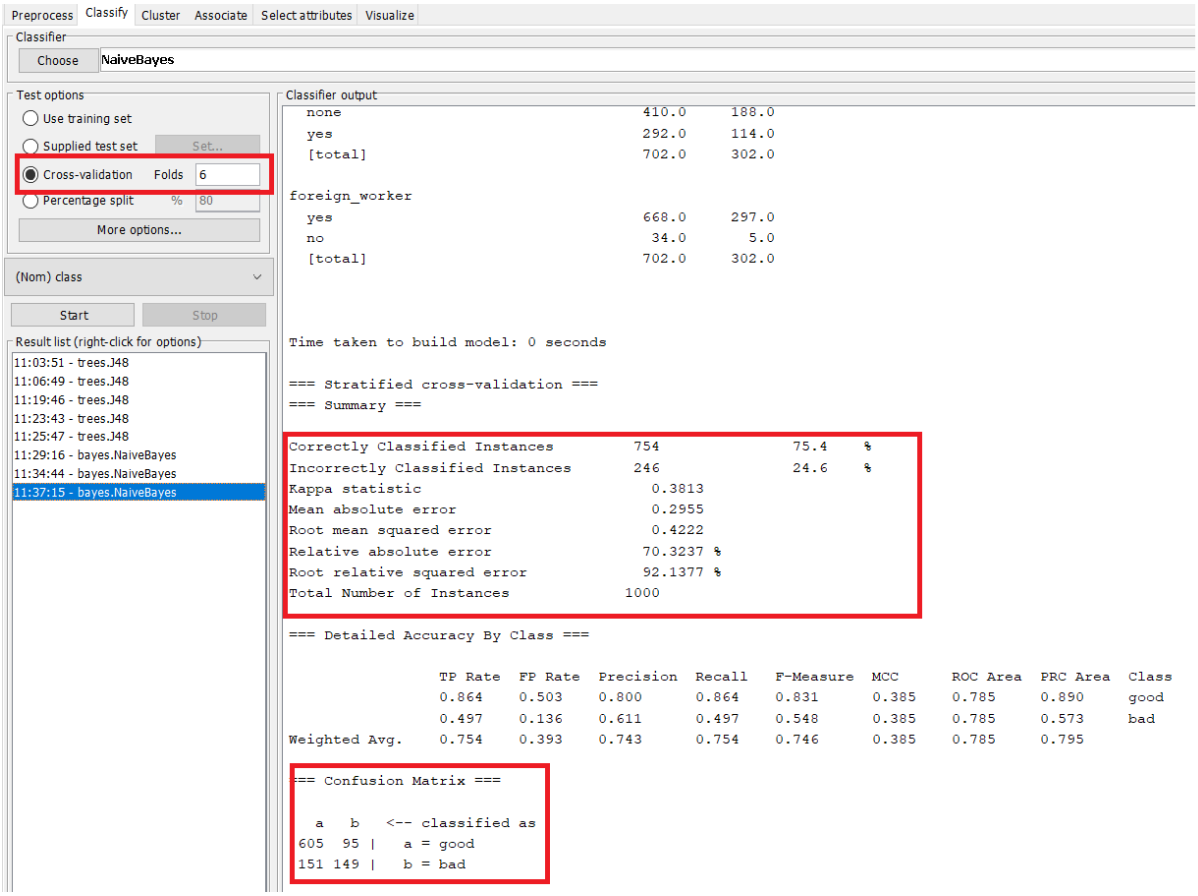
Accuracy is 75.4% when cross validation is 10.



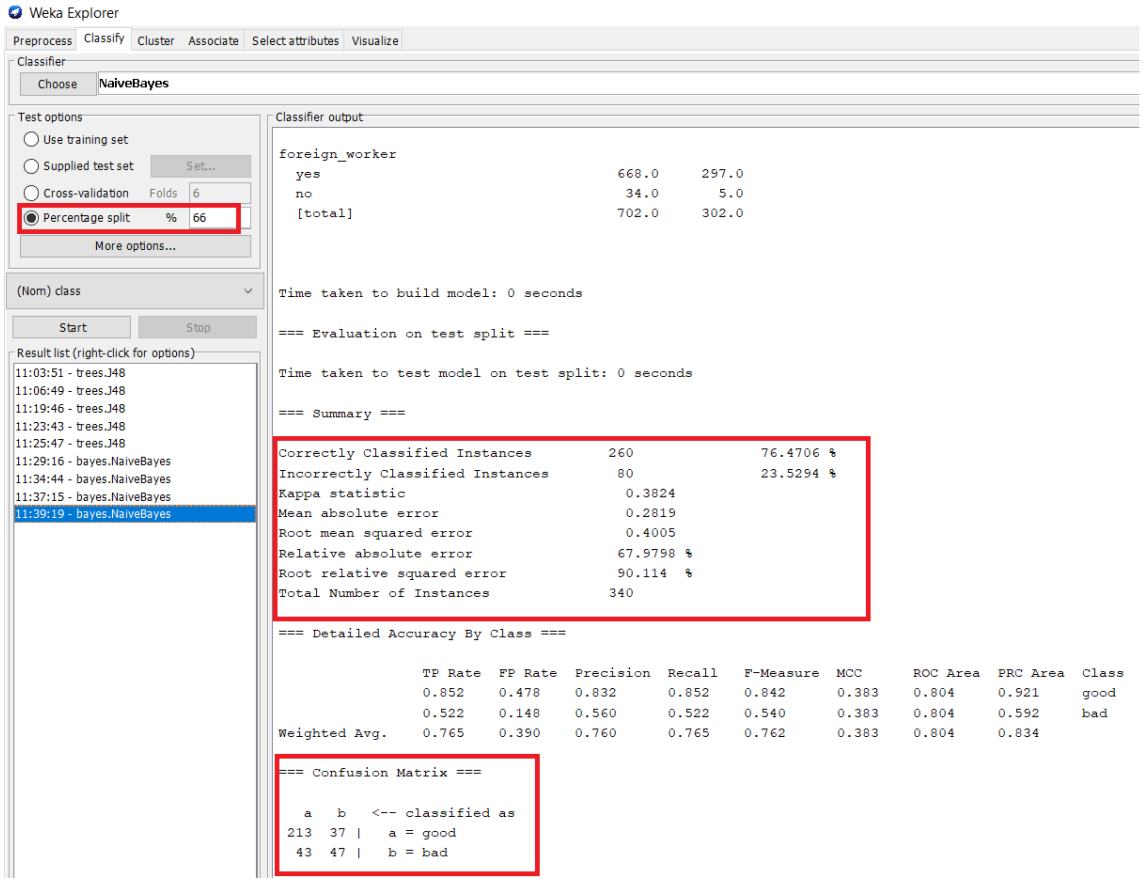
Accuracy is 75.9% when cross validation is 8.



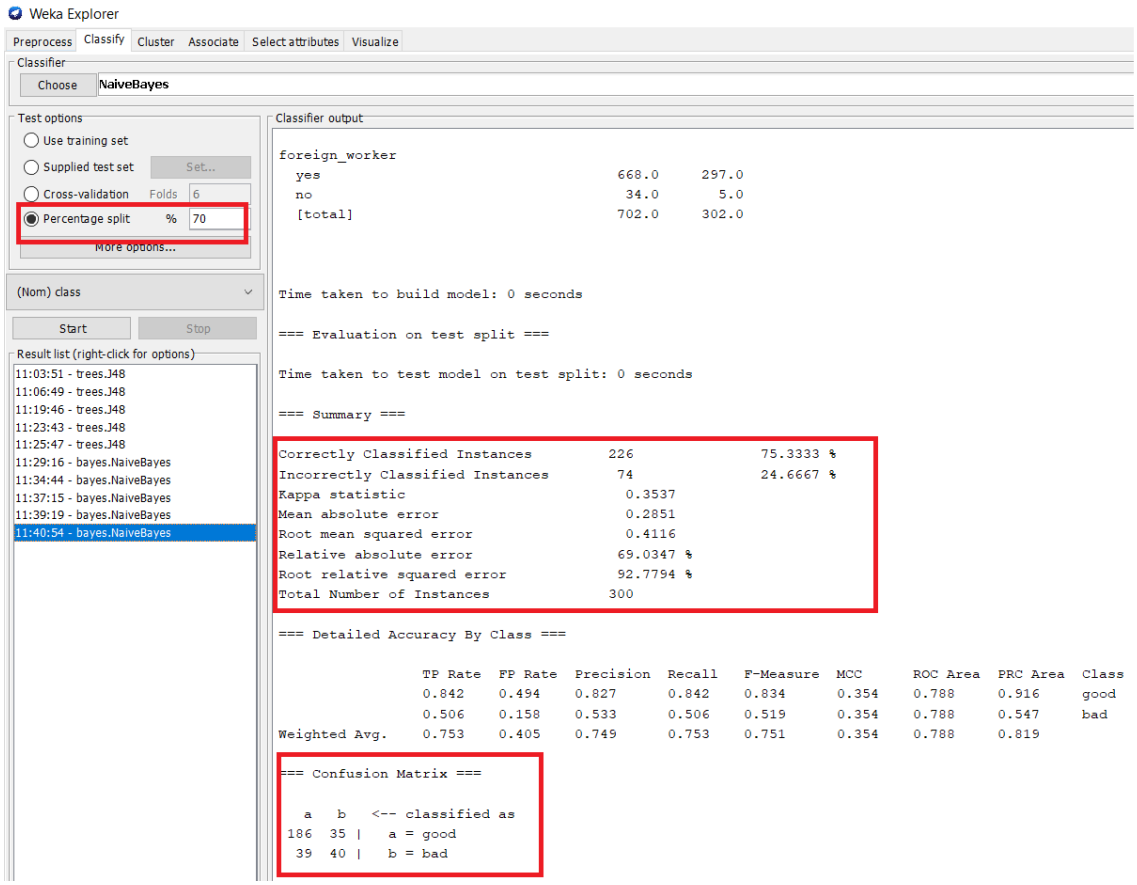
Accuracy is 75.4% when cross validation is 6.



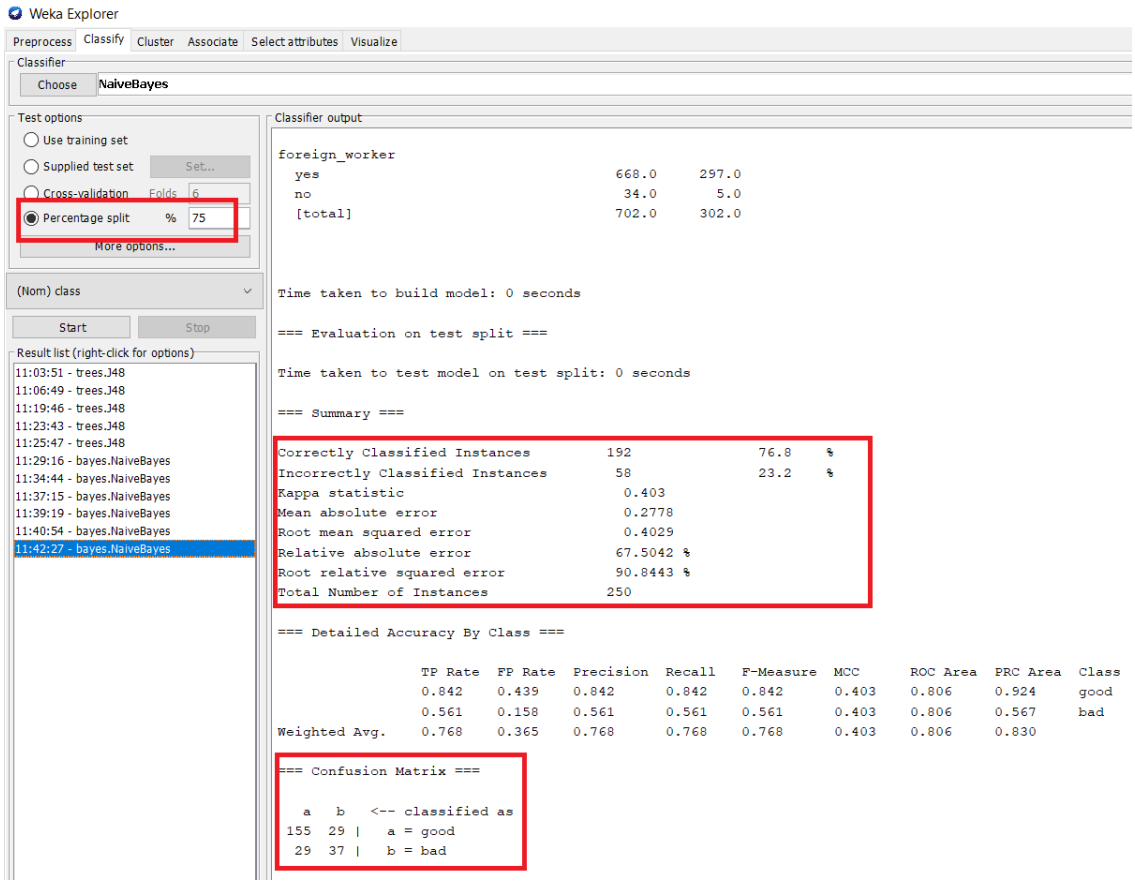
Accuracy is 76.47% when percentage split is 66.



Accuracy is 75.33% when percentage split is 70.



Accuracy is 76.8% when percentage split is 75.



Accuracy is 74.5% when percentage split is 80.