Effective Utilization of J48 Decision Tree Classifier in Data Mining using WEKA

Mahmut Emir ARSLAN

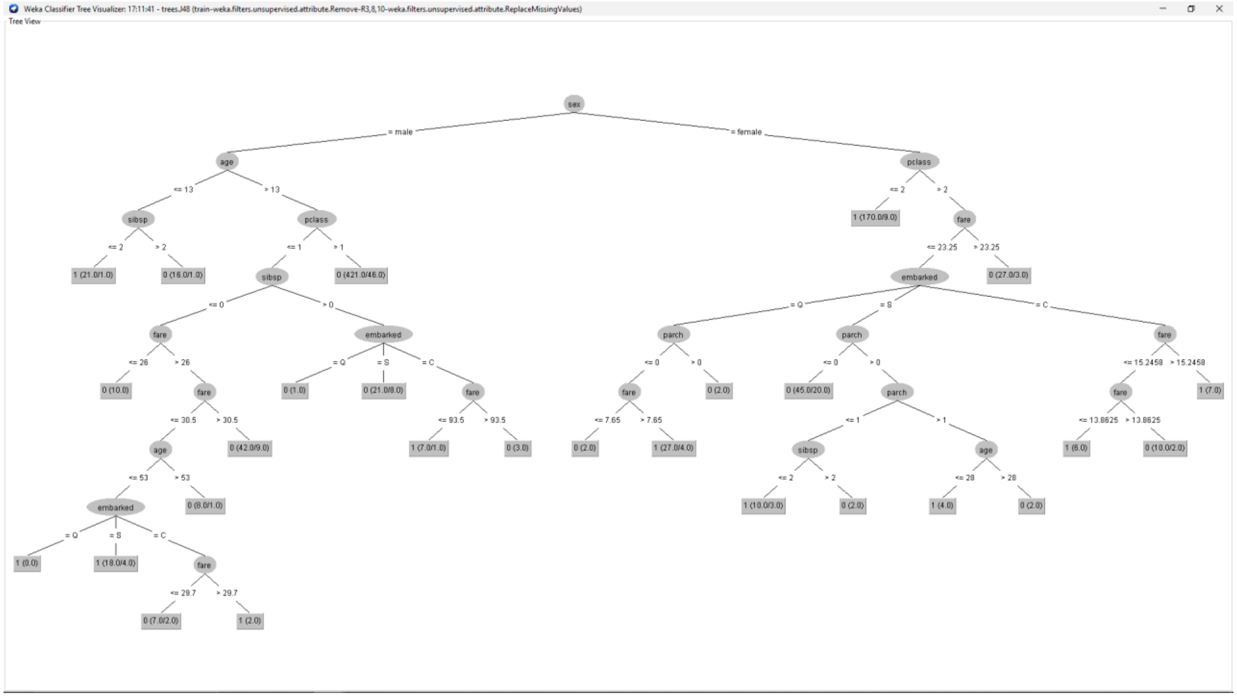
In this study, the performance and effectiveness of models created using different data preprocessing steps and various parameter settings of the J48 classifier have been systematically evaluated. This study, which holds fundamental importance in the field of data mining, contributes to our understanding of the impact of different preprocessing methods and J48 algorithm parameters on classification performance. The obtained results provide valuable insights for optimizing decision-making processes and achieving more accurate classification outcomes in data mining applications.

The aim of this study is to accurately classify the survivors of the Titanic disaster based on features such as age, gender, or ticket types.

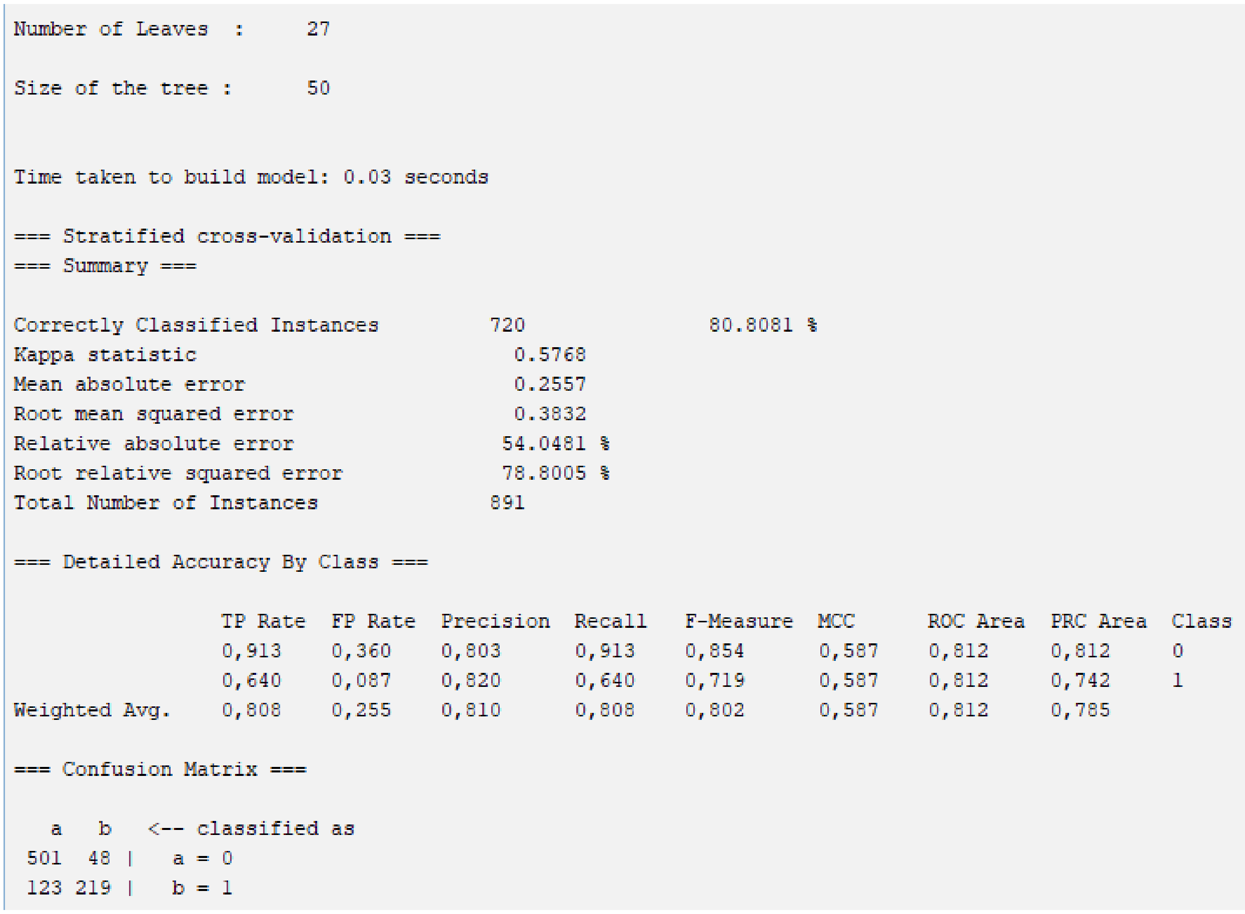
Model	Data pre-processing steps	J48 parameter	Evaluation on training data
1	Name, Ticket, and Cabin features were removed, and missing values in the remaining features were deleted	Default setting	10-fold CV accuracy 80.808%
2	By removing the Name, Ticket, and Cabin features, missing values in the remaining attributes were deleted. Furthermore, the 'Discretize' filter was applied to divide numeric values into five equal-frequency bins and converted into binary format.	Default setting	10-fold CV accuracy 82.155%
3	The features Name, Ticket, and Cabin were removed, and missing values in the remaining attributes were deleted. The 'Discretize' filter was applied to divide numeric values into five equal-frequency bins and convert them into binary format.	“The Confidence Factor” parameter was set to a value of 0.05.	10-fold CV accuracy 83.165%

Model-1)

During the creation of Model-1, to make the J48 algorithm functional, features with 'string' values were removed, and the 'ReplaceMissingValues' filter was applied to handle missing values. Normalization processes were found to have no impact on the decision tree, so no normalization was performed. The decision tree created in this manner and the obtained Weka results are as follows:



Model-1) Decision Tree

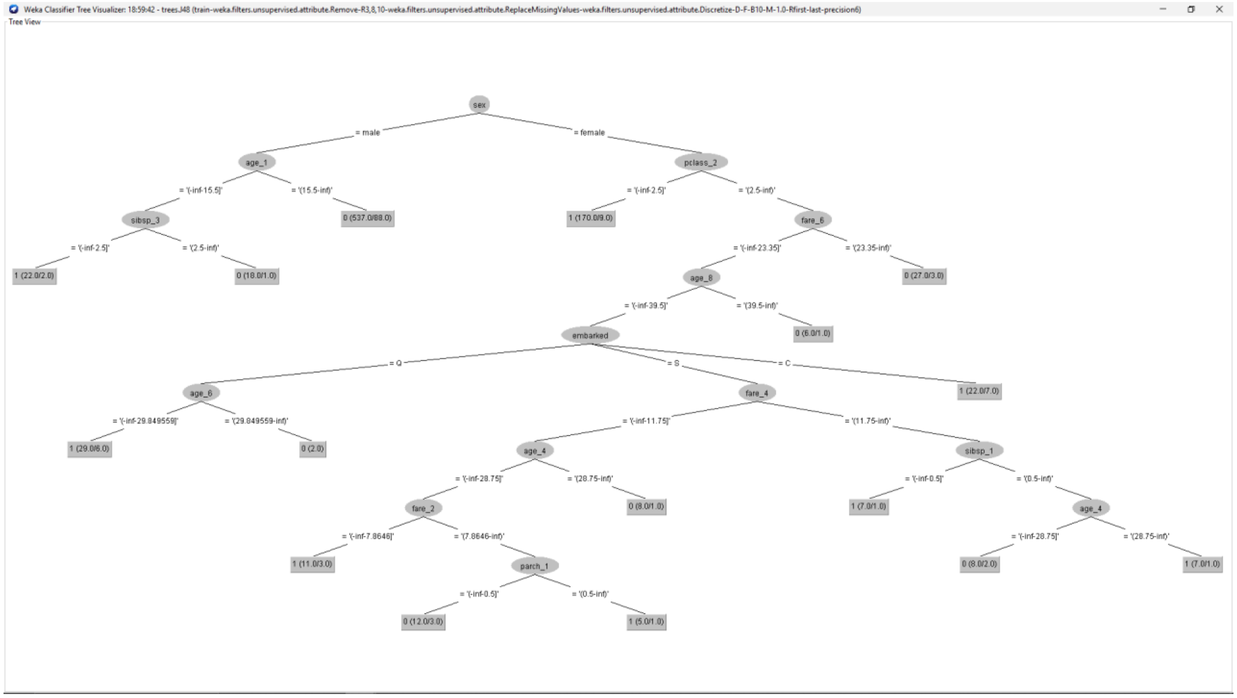


```
Number of Leaves  :     27

Size of the tree :      50


Time taken to build model: 0.03 seconds

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

Correctly Classified Instances         720               80.8081 %
Kappa statistic                          0.5768
Mean absolute error                      0.2557
Root mean squared error                  0.3832
Relative absolute error                 54.0481 %
Root relative squared error             78.8005 %
Total Number of Instances              891

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,913    0,360    0,803      0,913    0,854      0,587    0,812     0,812     0
                 0,640    0,087    0,820      0,640    0,719      0,587    0,812     0,742     1
Weighted Avg.    0,808    0,255    0,810      0,808    0,802      0,587    0,812     0,785

=== Confusion Matrix ===

   a   b   <-- classified as
 501  48 |   a = 0
 123 219 |   b = 1
```

Model-1) Weka Results

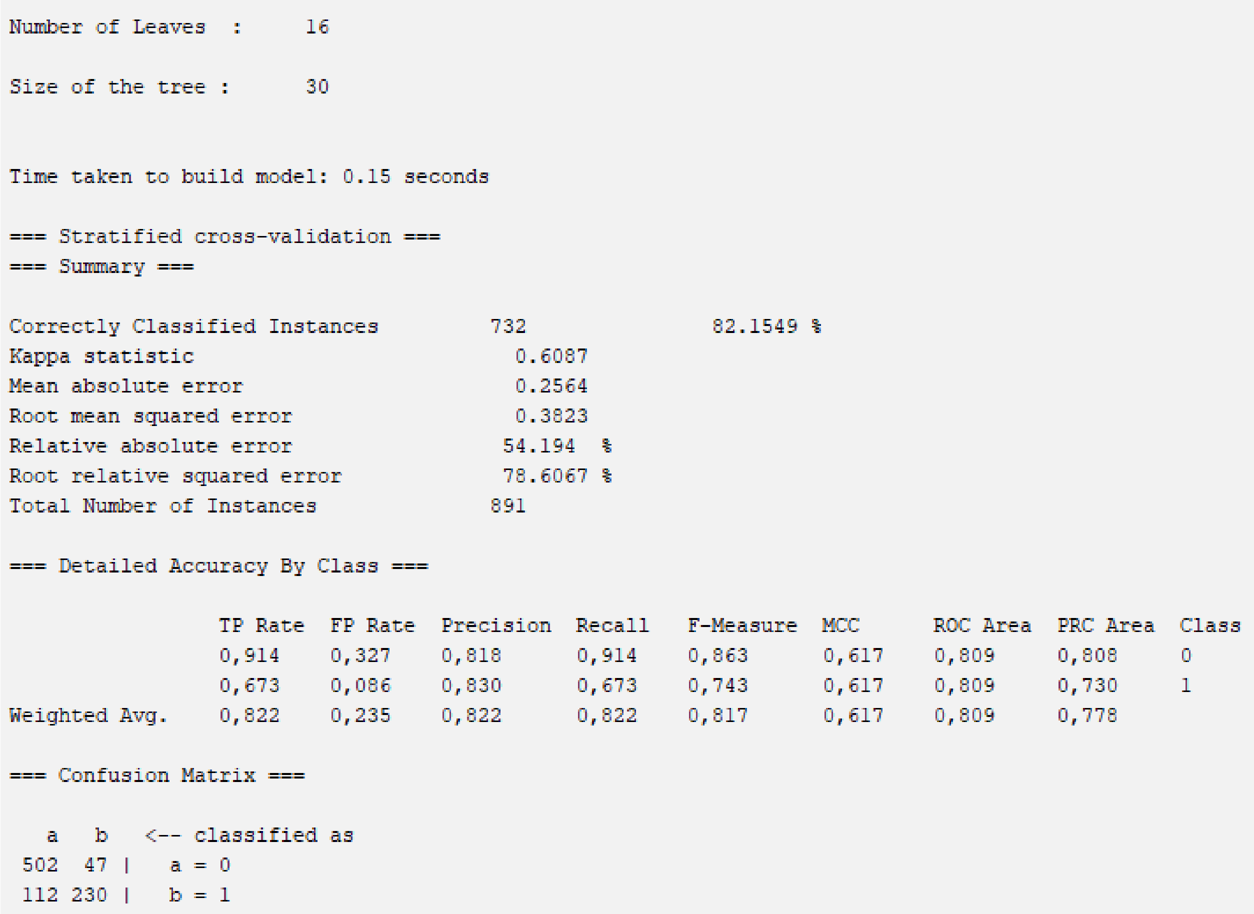
Model-2)

In this model, "Discretize" methods were applied. First, the J48 algorithm was run with the default values of the Discretize filter, and it was observed that the accuracy rate dropped to 78.788%. Subsequently, the feature with the default value of "Equal-Width Binning" was transformed into "Equal-Frequency Binning," and the "MakeBinary" feature was set to "True" to activate it. The "Bins" parameter was kept at its default value of 10 as it yielded the best result. With these changes, the accuracy rate increased to 82.155%. Additionally, the size of the decision tree reduced from 50 to 30, and the number of leaves decreased from 27 to 16, indicating a reduction in the complexity of the decision tree. The process of equal-frequency binning and converting to binary format improved the decision tree by dividing the data into more balanced partitions. As a result, the decision tree created and the Weka results are as follows:



Weka Classifier Tree Visualizer: 18:59:42 - trees.J48 (train-weka.filters.unsupervised.attribute.Remove-R3,8,10-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Discretize-D-F-B10-M-1.0-Rfirst-last-precision6)

Model-2) Decision Tree

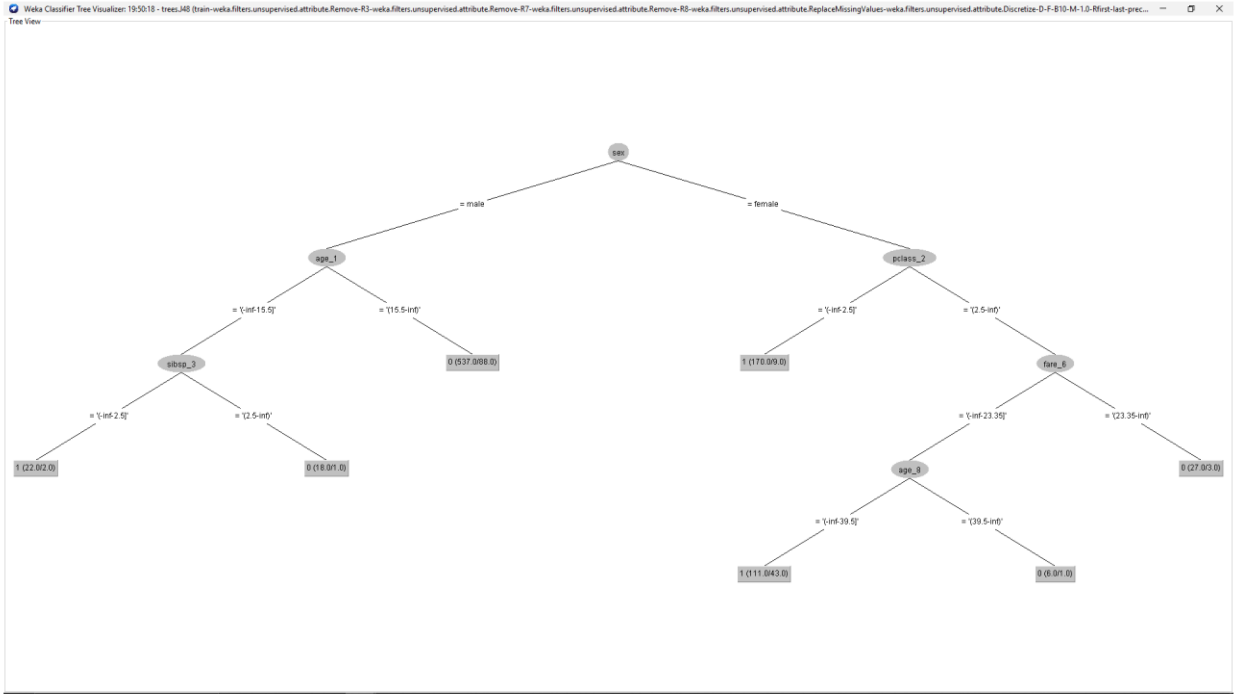


```
Number of Leaves  :     16

Size of the tree :      30


Time taken to build model: 0.15 seconds

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

Correctly Classified Instances         732               82.1549 %
Kappa statistic                          0.6087
Mean absolute error                      0.2564
Root mean squared error                  0.3823
Relative absolute error                 54.194  %
Root relative squared error             78.6067 %
Total Number of Instances              891

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,914    0,327    0,818      0,914    0,863      0,617    0,809     0,808     0
                 0,673    0,086    0,830      0,673    0,743      0,617    0,809     0,730     1
Weighted Avg.    0,822    0,235    0,822      0,822    0,817      0,617    0,809     0,778

=== Confusion Matrix ===

   a   b   <-- classified as
 502  47 |   a = 0
 112 230 |   b = 1
```

Model-2) Weka Results

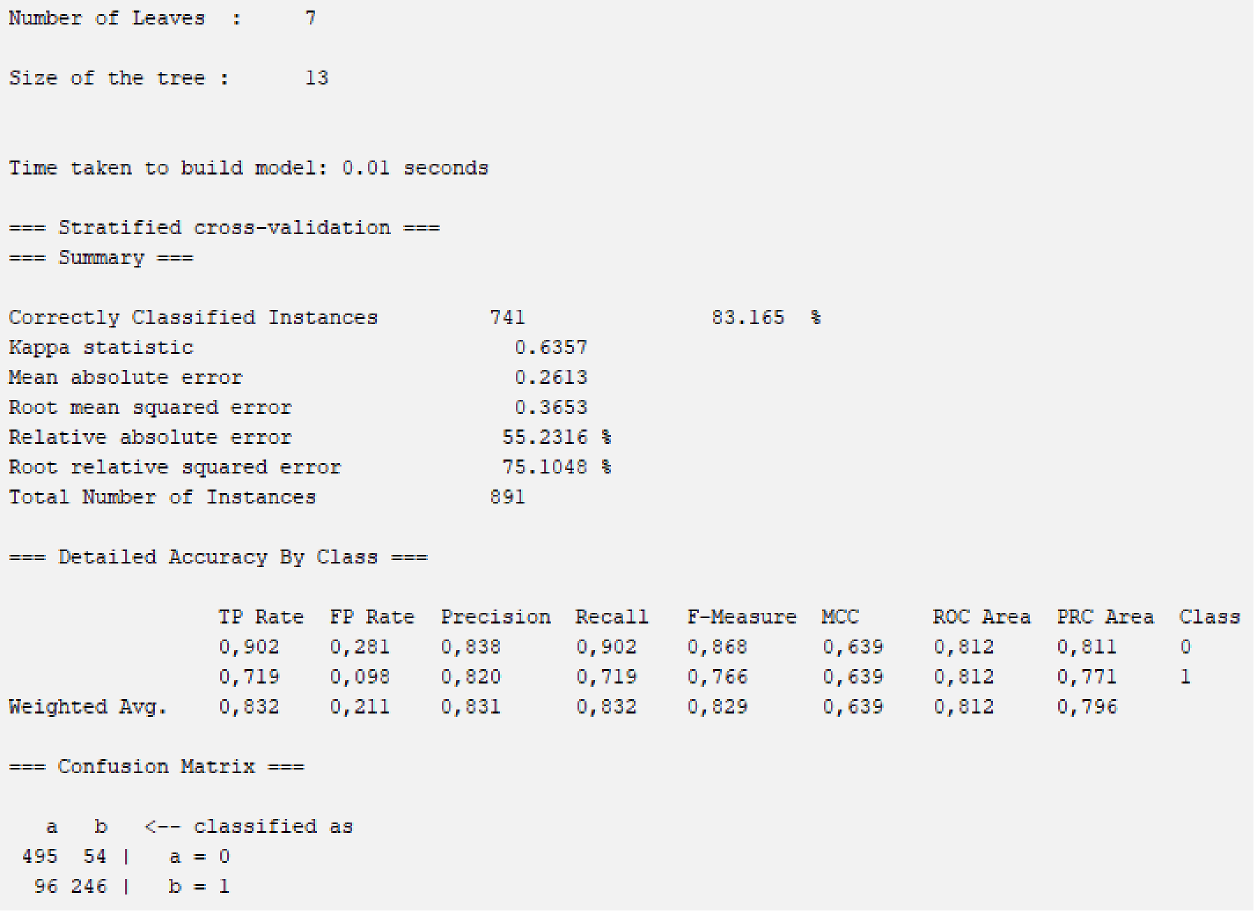
Model-3)

In the previous models, it was observed that the best results were obtained from filters. Therefore, in this model, the focus was shifted to the J48 parameters. Different parameter configurations were explored to find the highest accuracy rate. Through these experiments, the "Confidence Factor" parameter was identified as the sole J48 parameter that positively influenced the dataset. By changing the default value of the Confidence Factor parameter from "0.25" to "0.05", the highest accuracy rate was achieved, and the size and number of leaves of the decision tree were further reduced, leading to a decrease in complexity. Lowering the Confidence Factor parameter resulted in better performance because it indicated that the training data did not represent all possible events well and allowed the decision tree to be pruned more aggressively, reducing its complexity. Below are the obtained decision tree and Weka results:



Weka Classifier Tree Visualizer: 19:50:18 - trees.J48 (train-weka.filters.unsupervised.attribute.Remove-R3-weka.filters.unsupervised.attribute.Remove-R7-weka.filters.unsupervised.attribute.Remove-R8-weka.filters.unsupervised.attribute.ReplaceMissingValues-weka.filters.unsupervised.attribute.Discretize-D-F-B10-M-1.0-Rfirst-last-prec...

Model-3) Decision Tree



```
Number of Leaves  :     7

Size of the tree :      13


Time taken to build model: 0.01 seconds

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

Correctly Classified Instances         741               83.165  %
Kappa statistic                          0.6357
Mean absolute error                      0.2613
Root mean squared error                  0.3653
Relative absolute error                 55.2316 %
Root relative squared error             75.1048 %
Total Number of Instances              891

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,902    0,281    0,838      0,902    0,868      0,639    0,812     0,811     0
                 0,719    0,098    0,820      0,719    0,766      0,639    0,812     0,771     1
Weighted Avg.    0,832    0,211    0,831      0,832    0,829      0,639    0,812     0,796

=== Confusion Matrix ===

   a   b   <-- classified as
 495  54 |   a = 0
  96 246 |   b = 1
```

Model-3) Weka Results