Assignment 7: Logic Based Approaches

Team: Venkatesh Duvvuri ([VED14@pitt.edu](mailto:VED14@pitt.edu)); Haifa Alnasser ([HIA11@pitt.edu](mailto:HIA11@pitt.edu)); Gopi Tata ([GKT3@pitt.edu](mailto:GKT3@pitt.edu))

1. Iris Data Summary:

After loading the given *iris* data file (iris.txt) into the Weka Explorer Machine Learning tool it has been found that there are 150 instances with five attributes in the data namely SW, SW, PL,PW and class.

The tool provides further details about each selected attribute as below.

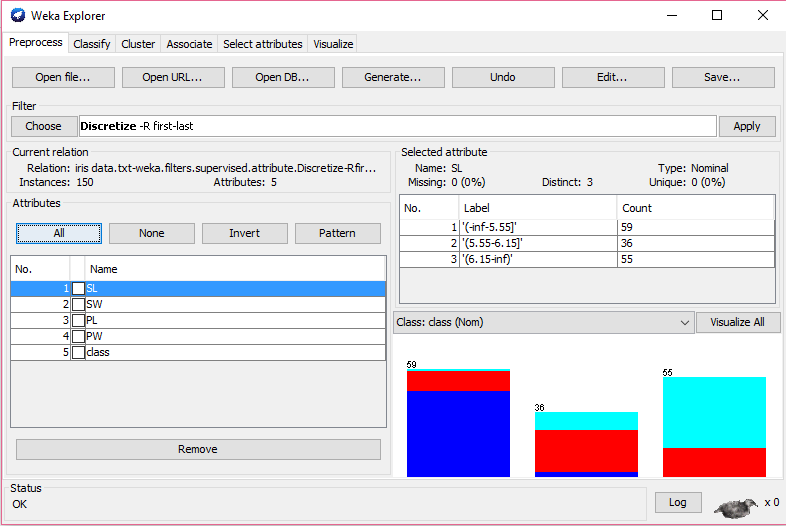
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Type | Missing | Distinct | Unique | Other Details |
| Class | Nominal | Zero (0% of ) instances in the data for which this attribute is unspecified | There are three different values that the data contains for this attribute. | There are 0 (0% of ) instances in the data having a value for this attribute that no other  instances have. | Count:  Iris\_setosa = 50 Iris\_versicolor = 50  Iris\_virginica = 50 |
| SW | Numeric | Zero (0% of) instances in the data for which this attribute is unspecified. | There are 35 number of different values that the data contains for this attribute. | There are 9 (6% of ) instances in the data having a value for this attribute that no other  instances have. | Minimum = 4.3  Maximum = 7.9  Mean = 5.843  Std. Deviation = 0.828 |
| SL | Numeric | Zero (0% of ) instances in the data for which this attribute is unspecified | There are 23 number of different values that the data contains for this attribute | There are 5 (3% of ) instances in the data having a value for this attribute that no other  instances have. | Minimum = 2.0  Maximum = 4.4  Mean = 3.054  Std. Deviation = 0.434 |
| PL | Numeric | Zero (0% of ) instances in the data for which this attribute is  unspecified | There are 43 number of different values that the data contains for this attribute | There are 10 (7% of) instances in the data having a value for this attribute that no other  instances have. | Minimum = 1.0  Maximum = 6.9  Mean = 3.759  Std. Deviation = 1.764 |
| PW | Numeric | Zero (0% of ) instances in the data for which this attribute is  unspecified | There are 22 number of different values that the data contains for this attribute | There are 2 (1% of) instances in the data having a value for this attribute that no other  instances have. | Minimum = 0.1  Maximum = 2.1  Mean = 1.199  Std. Deviation = 0.763 |

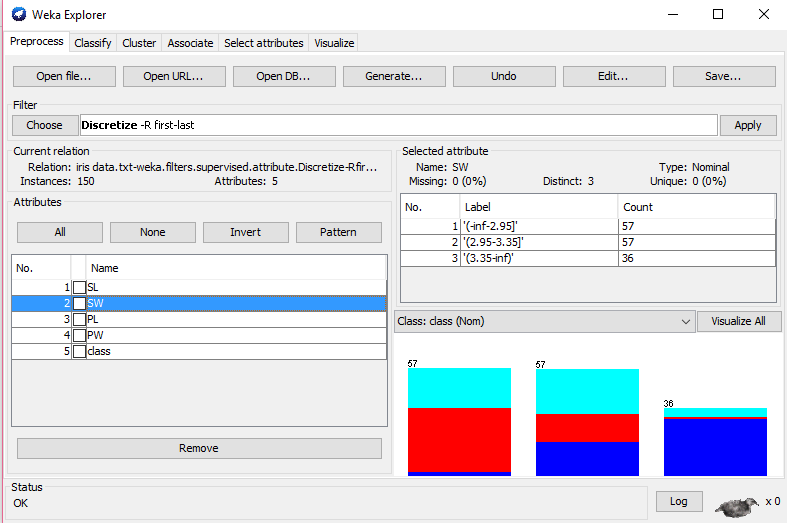
Applying Filter to the Attributes:

1. Discretizing Numeric data to Nominal data for attribute SL, SW, PL and PW:

The screen print below shows the Discretizing the attribute SL into nominal and the results as below.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Name | Type | Missing | Distinct | Unique | Other Details |
| SL | Nominal | Zero (0% of ) instances in the data for which this attribute is unspecified | There are 3 number of different values that the data contains for this attribute | There are 0 (0% of ) instances in the data having a value for this attribute that no other  instances have. | Count:  =========  (-inf-5.5) = 59  (5.55-6.15) = 36  (6.15-inf) = 55 |
| SW | Nominal | Zero (0% of ) instances in the data for which this attribute is unspecified | There are 3 number of different values that the data contains for this attribute | There are 0 (0% of ) instances in the data having a value for this attribute that no other  instances have. | Count:  =========  (-inf-2.95) = 57  (2.95-3.35) = 57  (3.35-inf) = 36 |
| PL | Nominal | Zero (0% of ) instances in the data for which this attribute is unspecified | There are 3 number of different values that the data contains for this attribute | There are 0 (0% of ) instances in the data having a value for this attribute that no other  instances have. | Count:  =========  (-inf-2.45) = 50  (2.45-4.75) = 45  (4.75-inf) = 55 |
| PW | Nominal | Zero (0% of ) instances in the data for which this attribute is unspecified | There are 3 number of different values that the data contains for this attribute | There are 0 (0% of ) instances in the data having a value for this attribute that no other  instances have. | Count:  =========  (-inf-0.8) = 50  (0.8-1.75) = 54  (1.75-inf) = 46 |





1. Classification of Iris data:

Choosing the Clssifiers 🡪 Tree 🡪J48 and using the cross-validation method as Leave-Out-One mthod by setting the Folds equal to the number of instances (150). The algorithm used is the J48 and the results are as below.

Classifier model is a pruned decision tree in textual form that was produced on the full training data. The first split is on the ‘PW’ attribute, at the second level, the splits is on ‘PL’. In the tree structure, there are 5 number of leaves and the size of the Tree is 9.The time taken to build this model is 0.02 seconds.

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: iris data.txt

Instances: 150

Attributes: 5

SL

SW

PL

PW

class

Test mode:150-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree

------------------

PW <= 0.6: Iris\_setosa (50.0)

PW > 0.6

| PW <= 1.7

| | PL <= 4.9: Iris\_versicolor (48.0/1.0)

| | PL > 4.9

| | | PW <= 1.5: Iris\_virginica (3.0)

| | | PW > 1.5: Iris\_versicolor (3.0/1.0)

| PW > 1.7: Iris\_virginica (46.0/1.0)

Number of Leaves : 5

Size of the tree : 9

Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 143 95.3333 %

Incorrectly Classified Instances 7 4.6667 %

Kappa statistic 0.93

Mean absolute error 0.0389

Root mean squared error 0.171

Relative absolute error 8.7011 %

Root relative squared error 36.0413 %

Total Number of Instances 150

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.98 0 1 0.98 0.99 0.99 Iris\_setosa

0.94 0.04 0.922 0.94 0.931 0.937 Iris\_versicolor

0.94 0.03 0.94 0.94 0.94 0.941 Iris\_virginica

Weighted Avg. 0.953 0.023 0.954 0.953 0.954 0.956

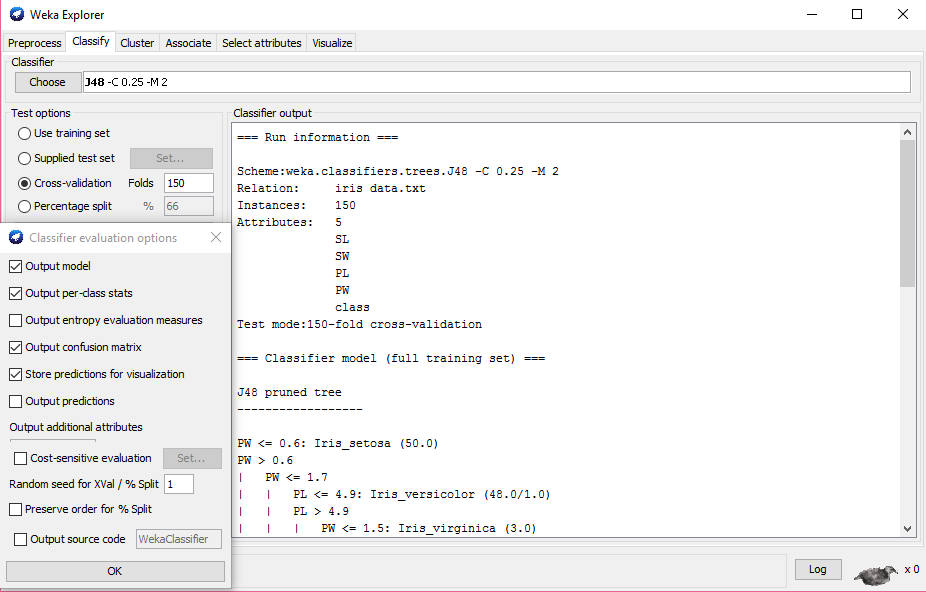
=== Confusion Matrix ===

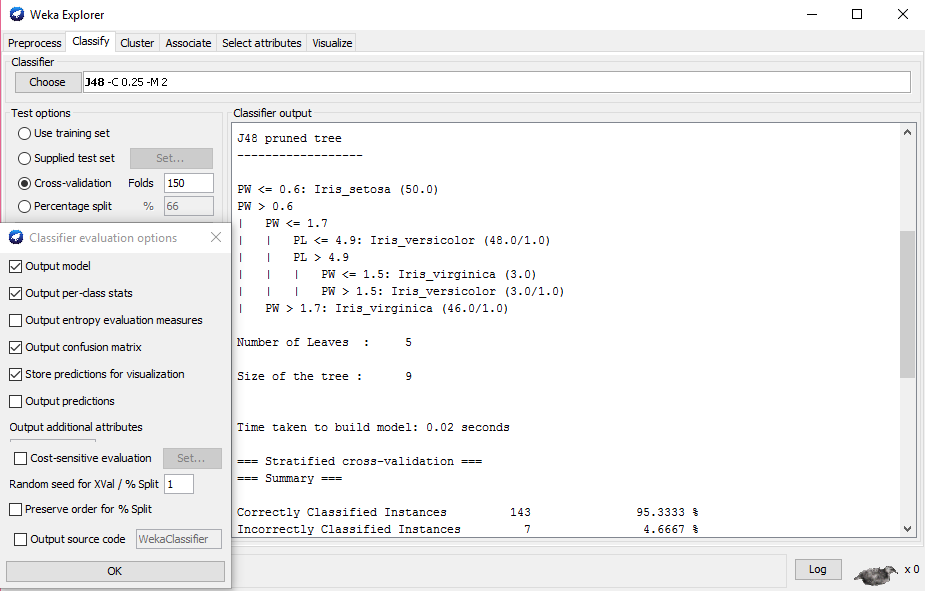
a b c <-- classified as

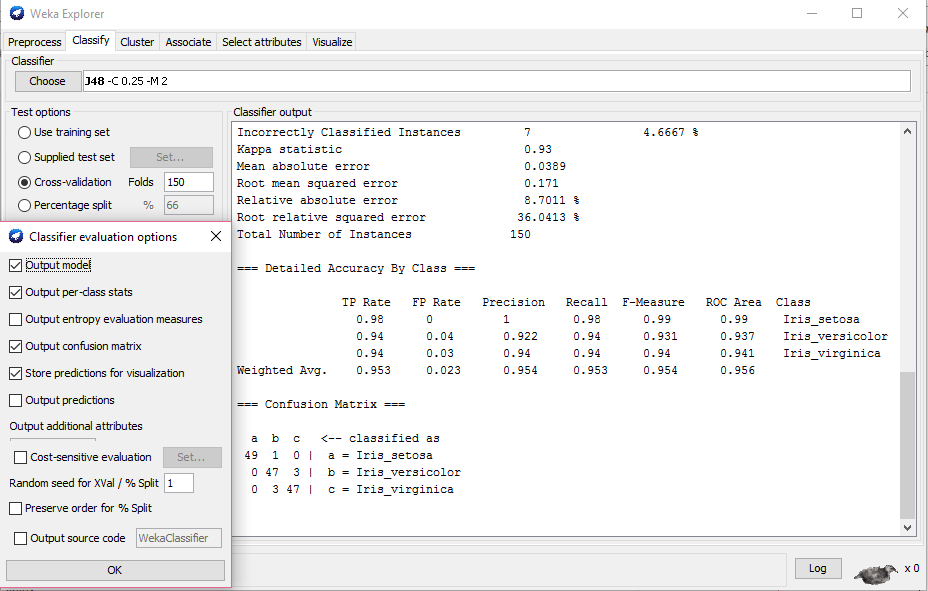
49 1 0 | a = Iris\_setosa

0 47 3 | b = Iris\_versicolor

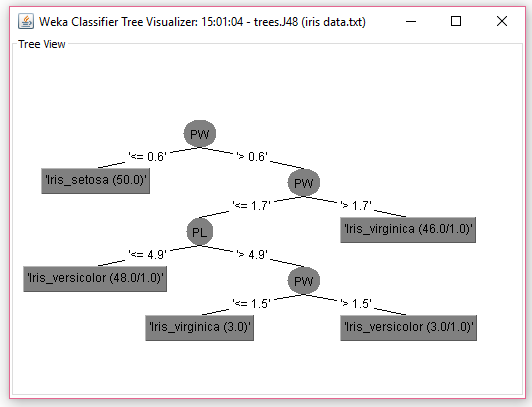
0 3 47 | c = Iris\_virginica





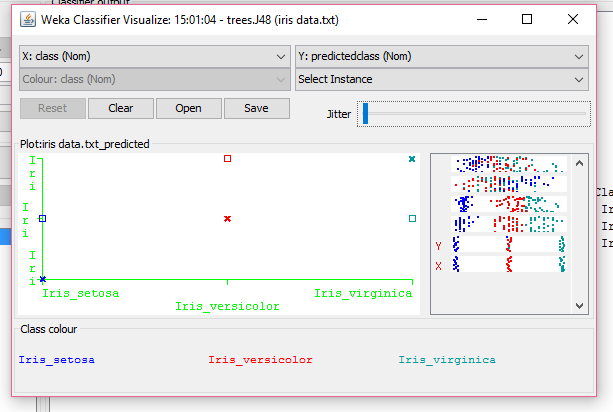


Graphical representation of the Classification Tree produced is as below.

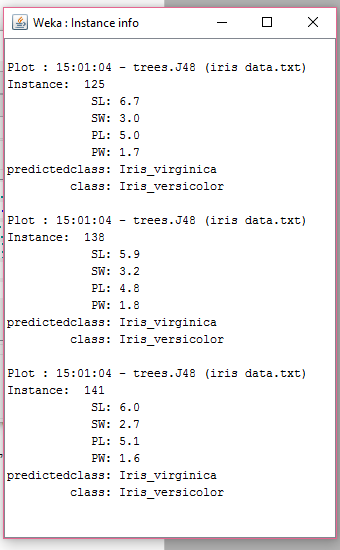


Visualization of Error Classification is as below:

The correctly classified instances are *crosses* and incorrectly classified instances are *squares*.



The data for the incorrectly classified instance (the Red square) is as below for instances 125, 138 and 141 all these three data instances are of class iris\_versicolor; however predicted incorrectly as of class iris\_virginica.



E) Classification of Iris data using RandomForest:

Choose the Classification algorithm RandomForest with the cross-validation folds set to 150. The output results are as below.

=== Run information ===

Scheme:weka.classifiers.trees.RandomForest -I 100 -K 0 -S 1

Relation: iris data.txt

Instances: 150

Attributes: 5

SL

SW

PL

PW

class

Test mode:150-fold cross-validation

=== Classifier model (full training set) ===

Random forest of 100 trees, each constructed while considering 3 random features.

Out of bag error: 0.0467

Time taken to build model: 0.11 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 143 95.3333 %

Incorrectly Classified Instances 7 4.6667 %

Kappa statistic 0.93

Mean absolute error 0.0388

Root mean squared error 0.159

Relative absolute error 8.6829 %

Root relative squared error 33.5132 %

Total Number of Instances 150

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

1 0 1 1 1 1 Iris\_setosa

0.94 0.04 0.922 0.94 0.931 0.991 Iris\_versicolor

0.92 0.03 0.939 0.92 0.929 0.991 Iris\_virginica

Weighted Avg. 0.953 0.023 0.953 0.953 0.953 0.994

=== Confusion Matrix ===

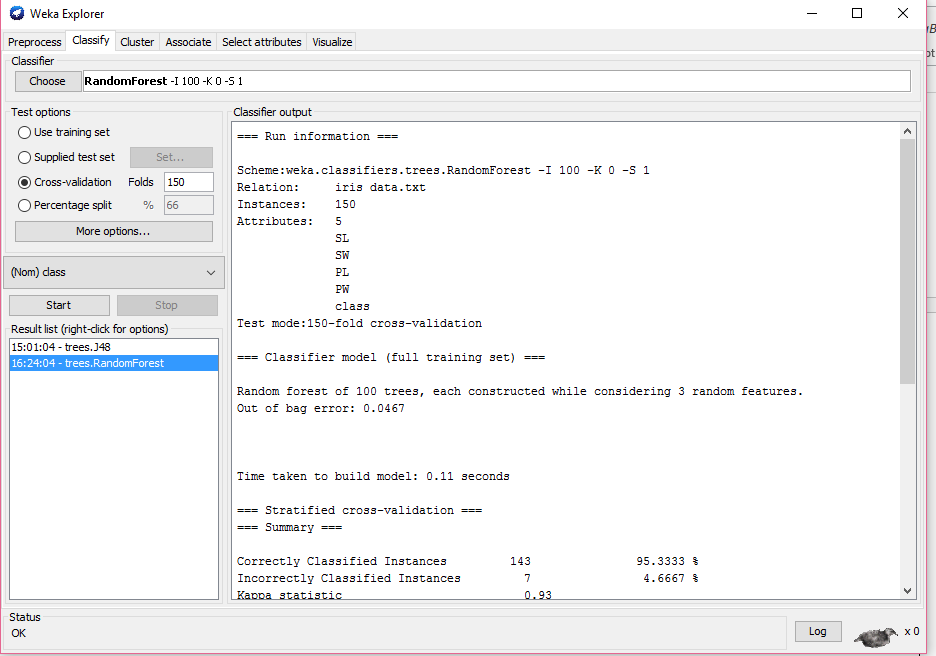
a b c <-- classified as

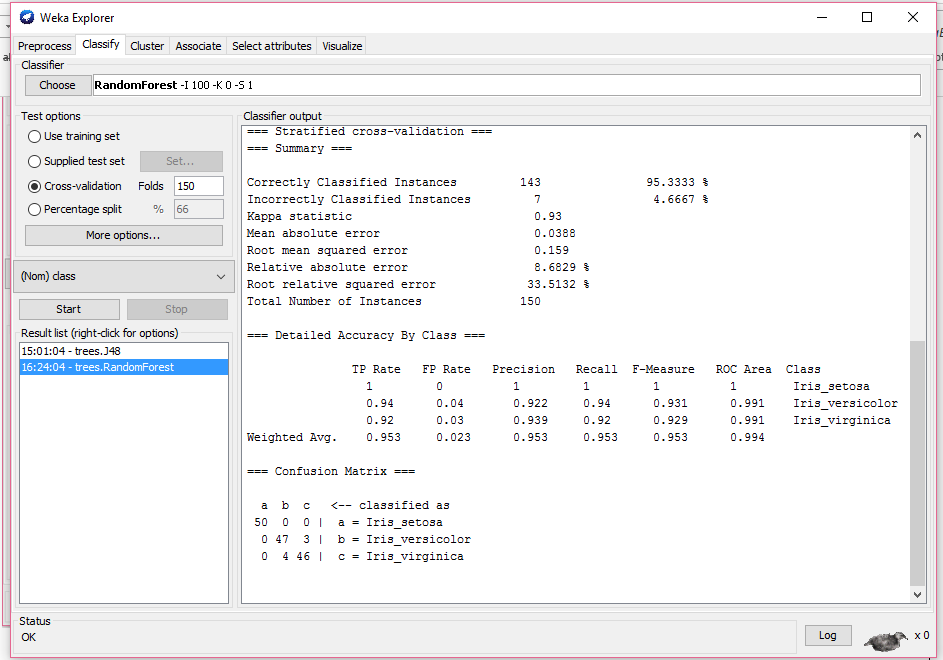
50 0 0 | a = Iris\_setosa

0 47 3 | b = Iris\_versicolor

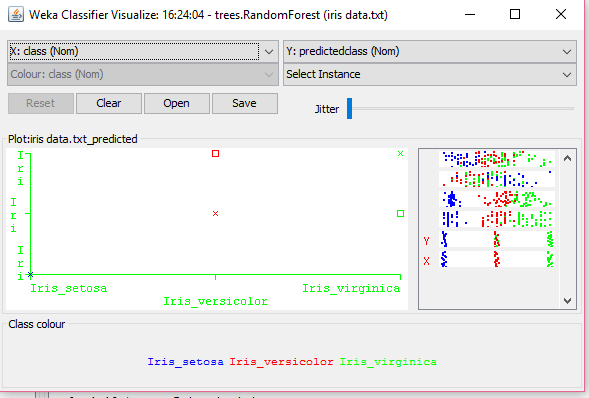
0 4 46 | c = Iris\_virginica

The screen prints of the output of RandomForest Classifiction is





The visual classification of errors are as below:



F) Classification of Iris data using JRip rule:

The output summary is as below

=== Run information ===

Scheme:weka.classifiers.rules.JRip -F 3 -N 2.0 -O 2 -S 1

Relation: iris data.txt

Instances: 150

Attributes: 5

SL

SW

PL

PW

class

Test mode:150-fold cross-validation

=== Classifier model (full training set) ===

JRIP rules:

===========

(PL <= 1.9) => class=Iris\_setosa (50.0/0.0)

(PW >= 1.8) => class=Iris\_virginica (46.0/1.0)

(PL >= 5) => class=Iris\_virginica (6.0/2.0)

=> class=Iris\_versicolor (48.0/1.0)

Number of Rules : 4

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 141 94 %

Incorrectly Classified Instances 9 6 %

Kappa statistic 0.91

Mean absolute error 0.0529

Root mean squared error 0.1975

Relative absolute error 11.8175 %

Root relative squared error 41.6207 %

Total Number of Instances 150

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

1 0.02 0.962 1 0.98 0.995 Iris\_setosa

0.88 0.03 0.936 0.88 0.907 0.93 Iris\_versicolor

0.94 0.04 0.922 0.94 0.931 0.94 Iris\_virginica

Weighted Avg. 0.94 0.03 0.94 0.94 0.939 0.955

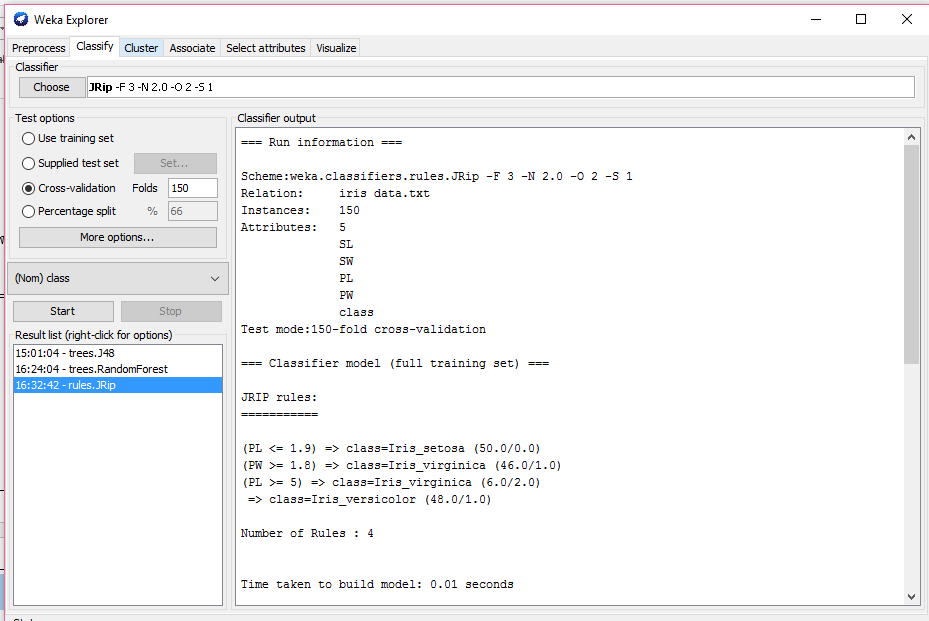
=== Confusion Matrix ===

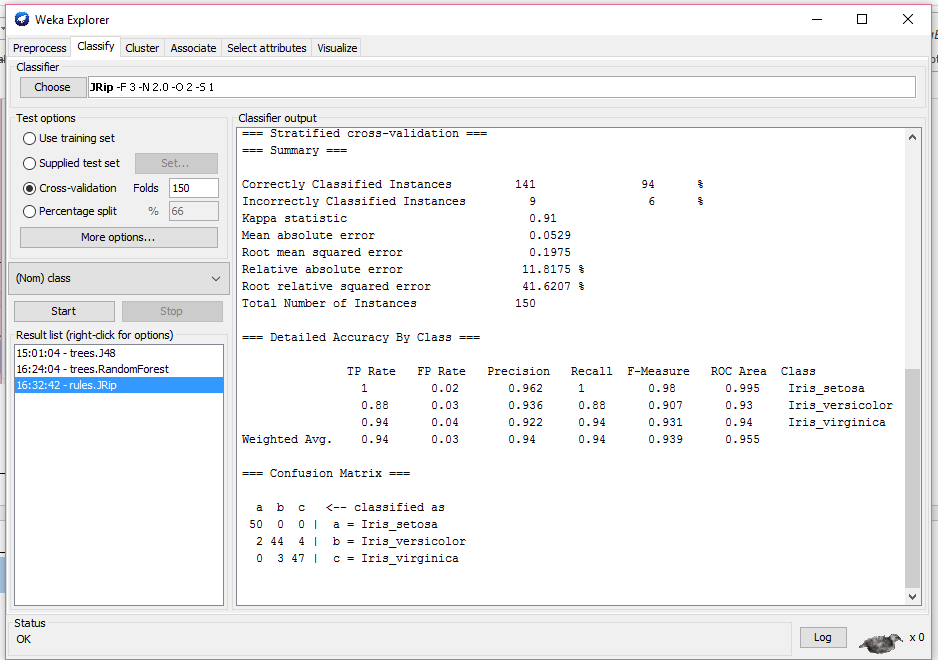
a b c <-- classified as

50 0 0 | a = Iris\_setosa

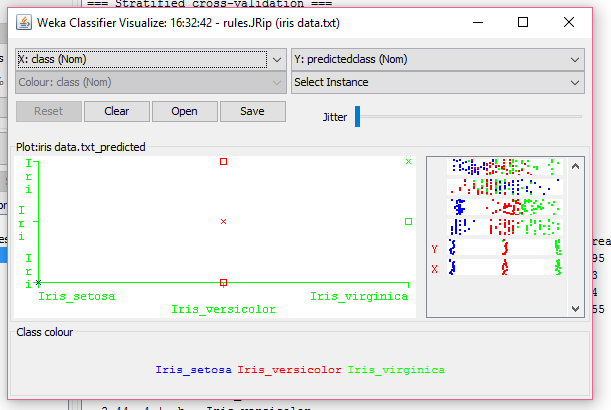
2 44 4 | b = Iris\_versicolor

0 3 47 | c = Iris\_virginica





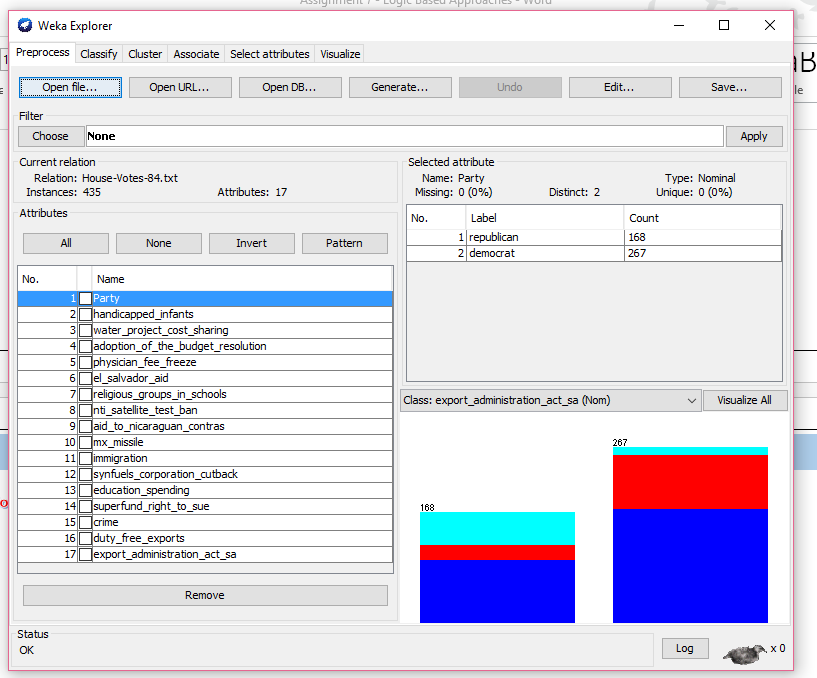
Visualization of errors from JRip rule.



**House Votes 84 Data Summary:**

From the screen print below there are 17 different attributes in the data. All the 17 attributes of the type Nominal. There are two parties Democratic and Republican with votes 267 and 168 respectively. All the other attributes are different issues and classified as whether voted *yes*, *no* or *withdrawn*.

In the following screen you would see the results and corresponding screen prints of Classification rules that are different from those used for analyzing the Iris data explained. We will be using Classification algorithm REPTree, Classification rules ConjunctiveRule and Zero R with the cross-validation folds set to 435.



1. House Votes 84 Classification using REPTree:

=== Run information ===

Scheme:weka.classifiers.trees.REPTree -M 2 -V 0.001 -N 3 -S 1 -L -1

Relation: House-Votes-84.txt

Instances: 435

Attributes: 17

Party

handicapped\_infants

water\_project\_cost\_sharing

adoption\_of\_the\_budget\_resolution

physician\_fee\_freeze

el\_salvador\_aid

religious\_groups\_in\_schools

nti\_satellite\_test\_ban

aid\_to\_nicaraguan\_contras

mx\_missile

immigration

synfuels\_corporation\_cutback

education\_spending

superfund\_right\_to\_sue

crime

duty\_free\_exports

export\_administration\_act\_sa

Test mode:435-fold cross-validation

=== Classifier model (full training set) ===

REPTree

============

aid\_to\_nicaraguan\_contras = n

| nti\_satellite\_test\_ban = n

| | crime = y

| | | duty\_free\_exports = n

| | | | superfund\_right\_to\_sue = y

| | | | | education\_spending = y

| | | | | | Party = republican

| | | | | | | water\_project\_cost\_sharing = y : y (28/13) [12/6]

| | | | | | | water\_project\_cost\_sharing = n : n (24/13) [10/4]

| | | | | | | water\_project\_cost\_sharing = w : y (7/4) [4/1]

| | | | | | Party = democrat : y (6/2) [3/2]

| | | | | education\_spending = n : y (9/5) [3/0]

| | | | | education\_spending = w : n (3/0) [3/1]

| | | | superfund\_right\_to\_sue = w : n (2/0) [1/1]

| | | | superfund\_right\_to\_sue = n : y (7/1) [2/1]

| | | duty\_free\_exports = y : y (9/4) [2/2]

| | | duty\_free\_exports = w : y (4/2) [1/0]

| | crime = n : y (3/0) [3/2]

| | crime = w : w (2/1) [1/0]

| nti\_satellite\_test\_ban = y : y (14/2) [12/1]

| nti\_satellite\_test\_ban = w : n (1/0) [2/1]

aid\_to\_nicaraguan\_contras = y : y (159/50) [83/24]

aid\_to\_nicaraguan\_contras = w : w (12/6) [3/0]

Size of the tree : 24

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 269 61.8391 %

Incorrectly Classified Instances 166 38.1609 %

Kappa statistic 0.0929

Mean absolute error 0.3285

Root mean squared error 0.415

Relative absolute error 90.8824 %

Root relative squared error 97.5932 %

Total Number of Instances 435

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

0.933 0.855 0.639 0.933 0.758 0.596 y

0.154 0.045 0.516 0.154 0.237 0.599 w

0.032 0.024 0.182 0.032 0.055 0.796 n

Weighted Avg. 0.618 0.543 0.544 0.618 0.533 0.625

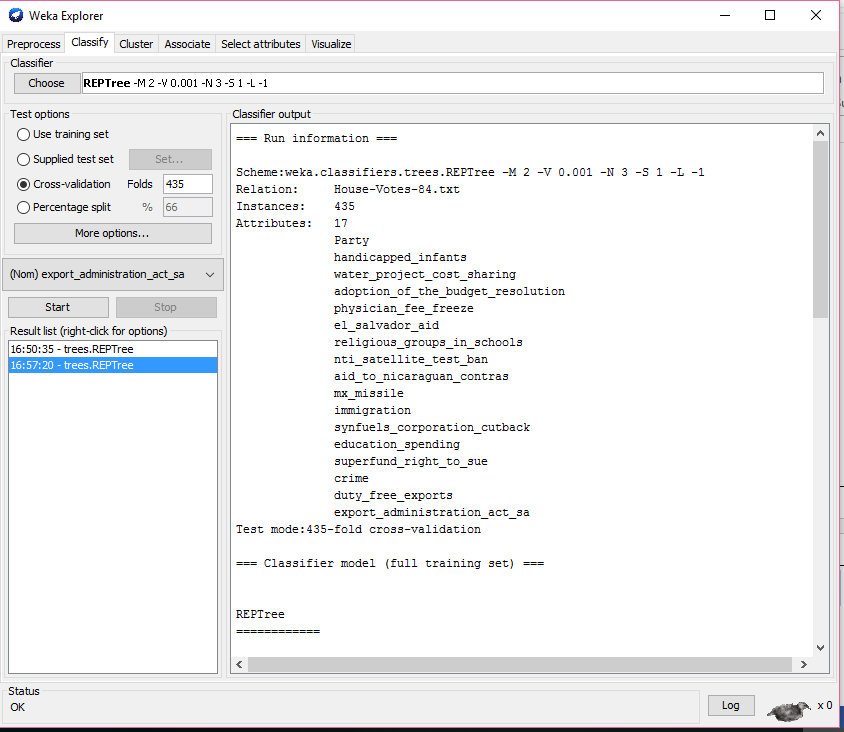
=== Confusion Matrix ===

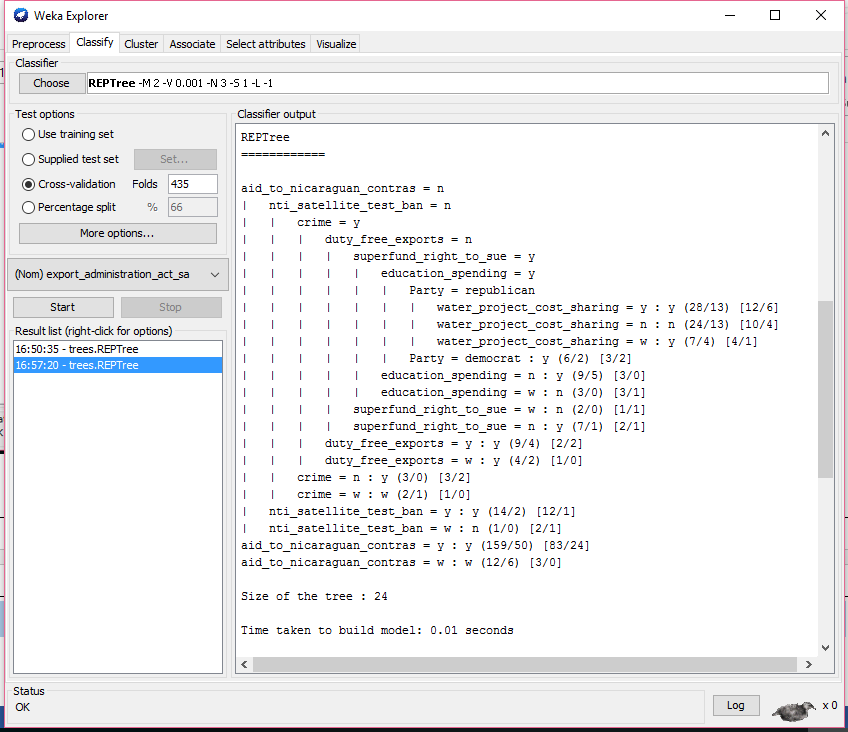
a b c <-- classified as

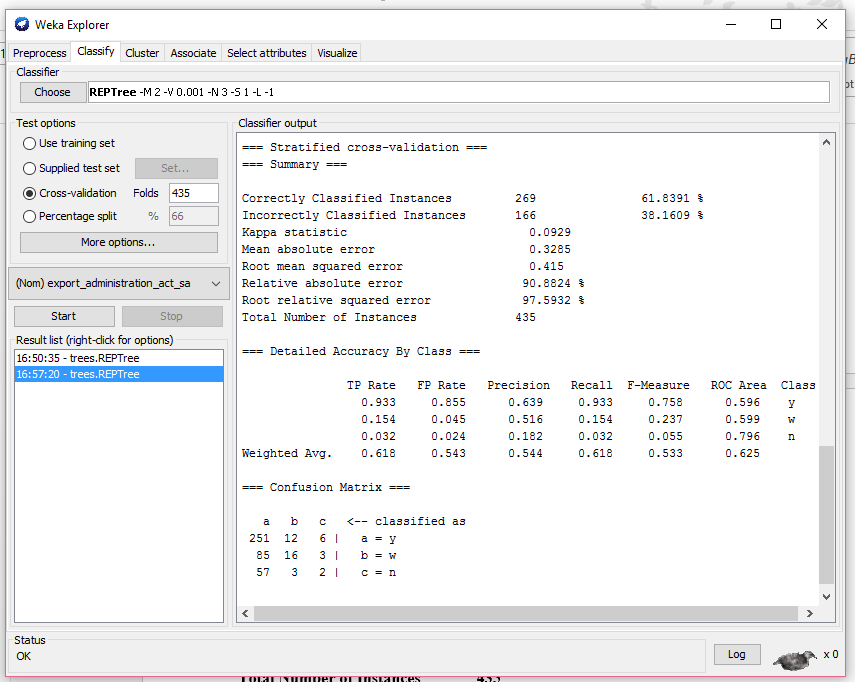
251 12 6 | a = y

85 16 3 | b = w

57 3 2 | c = n









1. Classification House Votes 84 data using ConjunctiveRule:

The output of Classification is as below.

=== Run information ===

Scheme:weka.classifiers.rules.ConjunctiveRule -N 3 -M 2.0 -P -1 -S 1

Relation: House-Votes-84.txt

Instances: 435

Attributes: 17

Party

handicapped\_infants

water\_project\_cost\_sharing

adoption\_of\_the\_budget\_resolution

physician\_fee\_freeze

el\_salvador\_aid

religious\_groups\_in\_schools

nti\_satellite\_test\_ban

aid\_to\_nicaraguan\_contras

mx\_missile

immigration

synfuels\_corporation\_cutback

education\_spending

superfund\_right\_to\_sue

crime

duty\_free\_exports

export\_administration\_act\_sa

Test mode:435-fold cross-validation

=== Classifier model (full training set) ===

Single conjunctive rule learner:

--------------------------------

=> export\_administration\_act\_sa = y

Class distributions:

Covered by the rule:

y w n

0.617241 0.241379 0.141379

Not covered by the rule:

y w n

0 0 0

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 269 61.8391 %

Incorrectly Classified Instances 166 38.1609 %

Kappa statistic 0

Mean absolute error 0.3616

Root mean squared error 0.4257

Relative absolute error 100.027 %

Root relative squared error 100.1091 %

Total Number of Instances 435

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

1 1 0.618 1 0.764 0.105 y

0 0 0 0 0 0.126 w

0 0 0 0 0 0.106 n

Weighted Avg. 0.618 0.618 0.382 0.618 0.473 0.11

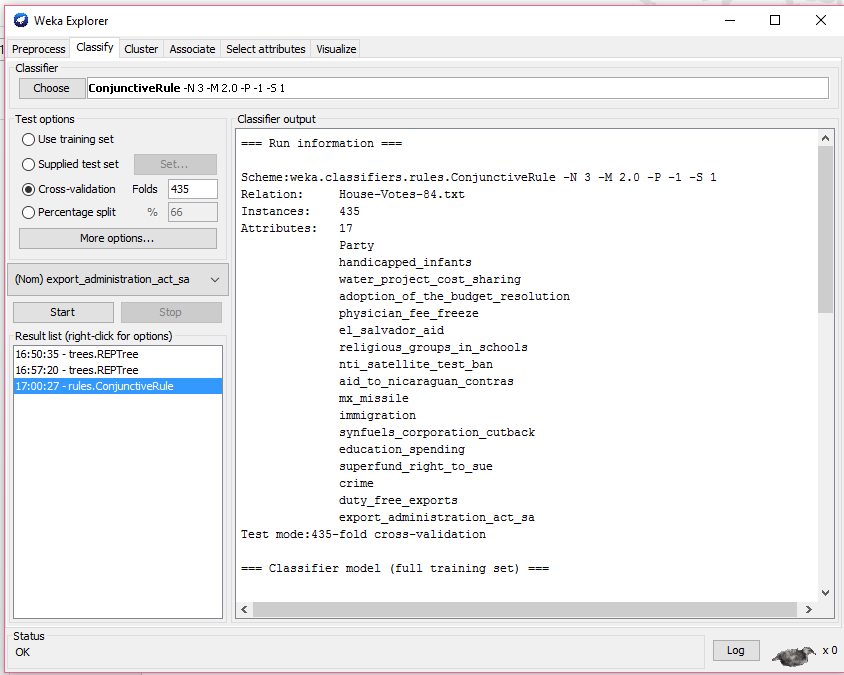
=== Confusion Matrix ===

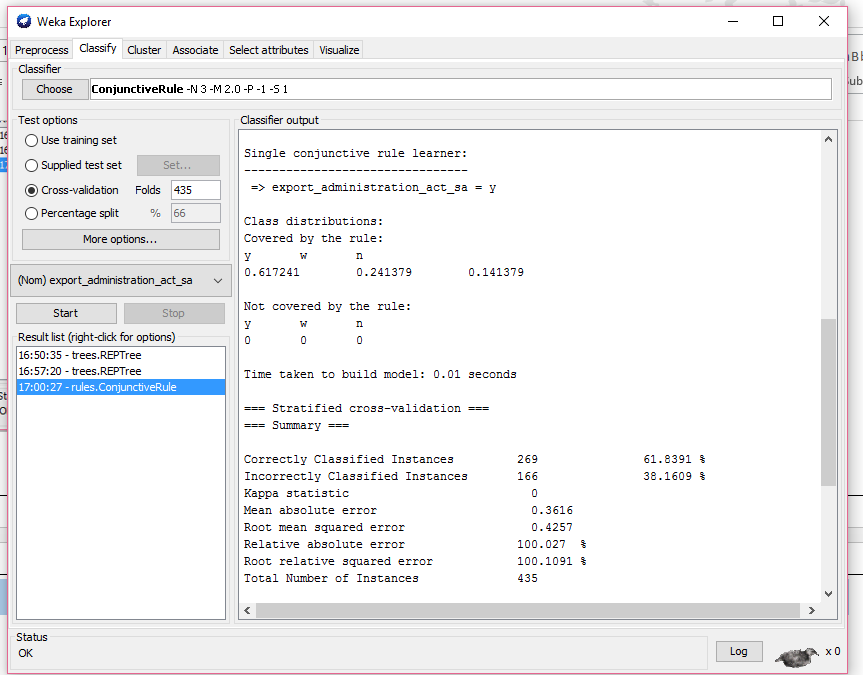
a b c <-- classified as

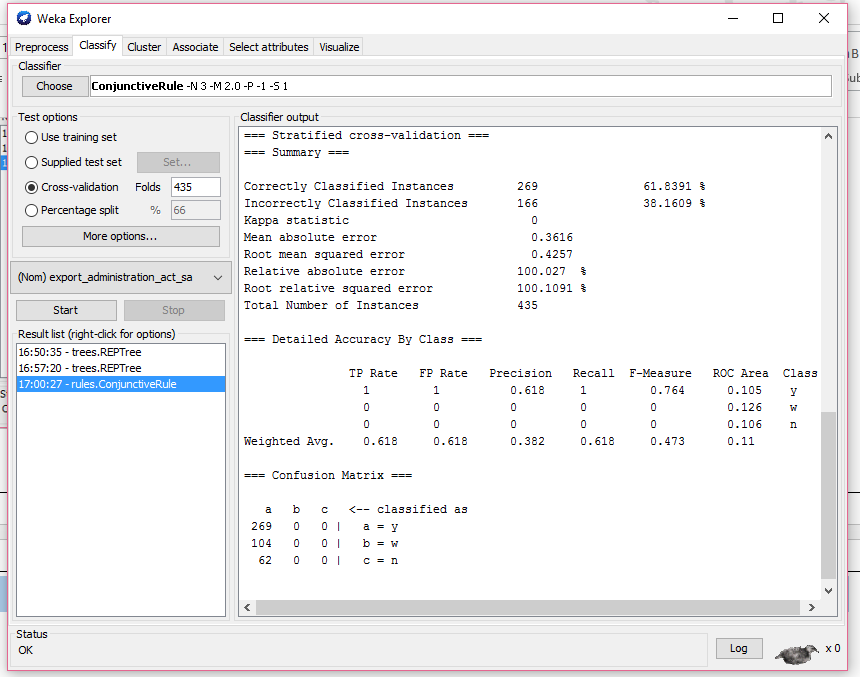
269 0 0 | a = y

104 0 0 | b = w

62 0 0 | c = n









1. House Votes 84 Classification using ZeroR:

The output summary of the Classification is as below

=== Run information ===

Scheme:weka.classifiers.rules.ZeroR

Relation: House-Votes-84.txt

Instances: 435

Attributes: 17

Party

handicapped\_infants

water\_project\_cost\_sharing

adoption\_of\_the\_budget\_resolution

physician\_fee\_freeze

el\_salvador\_aid

religious\_groups\_in\_schools

nti\_satellite\_test\_ban

aid\_to\_nicaraguan\_contras

mx\_missile

immigration

synfuels\_corporation\_cutback

education\_spending

superfund\_right\_to\_sue

crime

duty\_free\_exports

export\_administration\_act\_sa

Test mode:435-fold cross-validation

=== Classifier model (full training set) ===

ZeroR predicts class value: y

Time taken to build model: 0 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances 269 61.8391 %

Incorrectly Classified Instances 166 38.1609 %

Kappa statistic 0

Mean absolute error 0.3615

Root mean squared error 0.4253

Relative absolute error 100 %

Root relative squared error 100 %

Total Number of Instances 435

=== Detailed Accuracy By Class ===

TP Rate FP Rate Precision Recall F-Measure ROC Area Class

1 1 0.618 1 0.764 0 y

0 0 0 0 0 0 w

0 0 0 0 0 0 n

Weighted Avg. 0.618 0.618 0.382 0.618 0.473 0

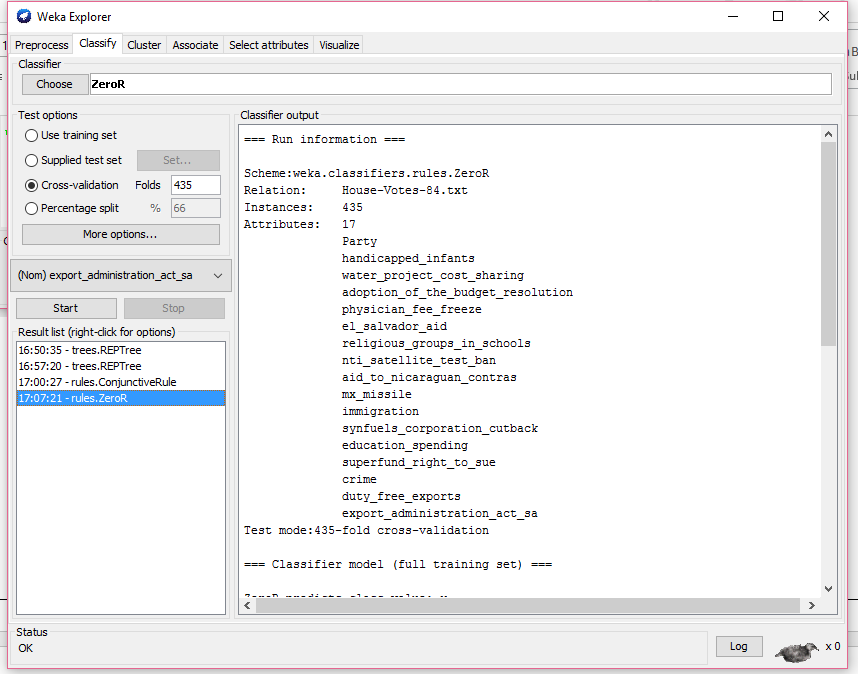
=== Confusion Matrix ===

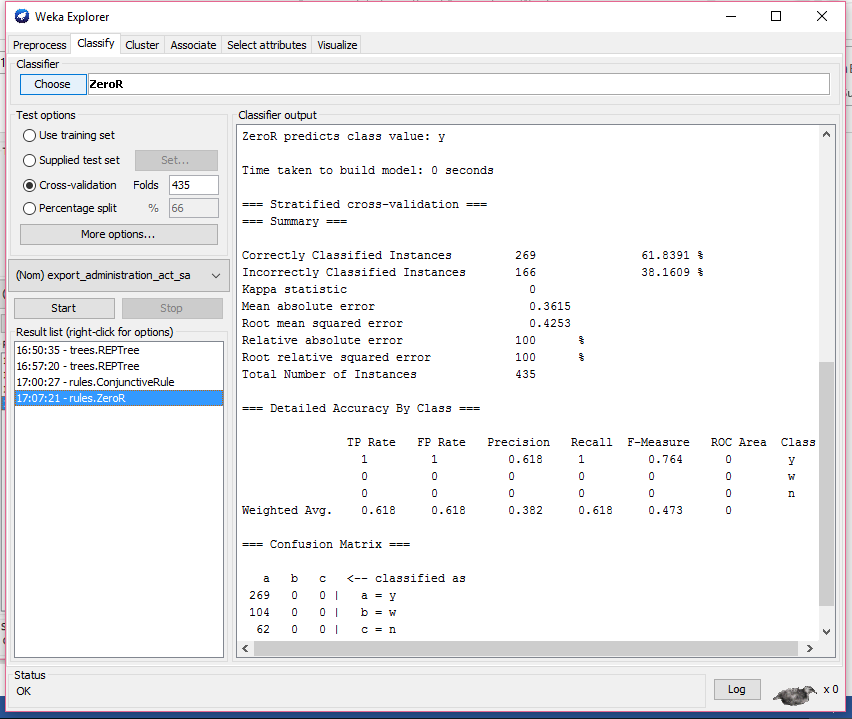
a b c <-- classified as

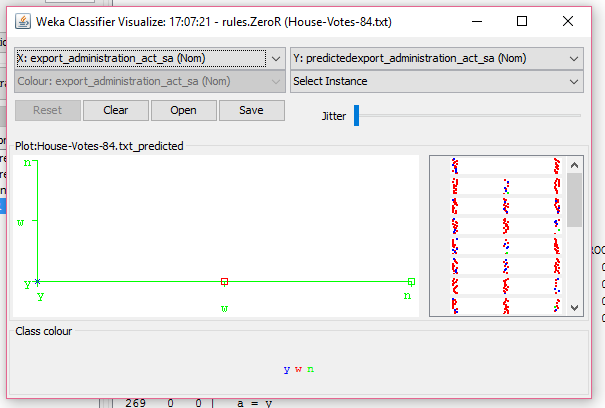
269 0 0 | a = y

104 0 0 | b = w

62 0 0 | c = n







Conclusions of Classification of both the data using different Classification algorithms.

The following is the summary of the prediction accuracy for both the Iris and House Votes 84 data.

For Iris data both the Tree Classifications (J48 and RandomForest) have predicted with same accuracy with for Correct and Incorrect Instances with 95.3333 % and 4.6667%. However the Root Mean Squared Error is slightly less for RandomForest with 0.151 compared to J48 and also less than that of Jrip Rule. Hence we conclude RandomForest is the most accurate Classification of all these three trials.

For House Votes 84 data – all the three Classification algorithms predicted with same accuracy of 61.8391% and 38.1609% for Correct and Incorrect instances. However the Root Mean Squared Error is the lowest for REPTree, though very slight margin, we believe this has best accuracy of these three Classification algorithms for this data.

Iris data - Tree-J48:

=== Summary ===

Correctly Classified Instances 143 95.3333 %

Incorrectly Classified Instances 7 4.6667 %

Kappa statistic 0.93

Mean absolute error 0.0389

Root mean squared error 0.171

Relative absolute error 8.7011 %

Root relative squared error 36.0413 %

-------------------------------------------------------------------------

Iris data - Tree-Random Forest:

=== Summary ===

Correctly Classified Instances 143 95.3333 %

Incorrectly Classified Instances 7 4.6667 %

Kappa statistic 0.93

Mean absolute error 0.0388

Root mean squared error 0.159

Relative absolute error 8.6829 %

Root relative squared error 33.5132 %

Total Number of Instances 150

----------------------------------------------------------------------------

Iris data - JRip Rule

=== Summary ===

Correctly Classified Instances 141 94 %

Incorrectly Classified Instances 9 6 %

Kappa statistic 0.91

Mean absolute error 0.0529

Root mean squared error 0.1975

Relative absolute error 11.8175 %

Root relative squared error 41.6207 %

Total Number of Instances 150

--------------------------------------------------------------------------------

House Votes 84 – REPTree:

=== Summary ===

Correctly Classified Instances 269 61.8391 %

Incorrectly Classified Instances 166 38.1609 %

Kappa statistic 0.0929

Mean absolute error 0.3285

Root mean squared error 0.415

Relative absolute error 90.8824 %

Root relative squared error 97.5932 %

Total Number of Instances 435

--------------------------------------------------------------------------------

House Votes 84 – Conjunctive Rule

=== Summary ===

Correctly Classified Instances 269 61.8391 %

Incorrectly Classified Instances 166 38.1609 %

Kappa statistic 0

Mean absolute error 0.3616

Root mean squared error 0.4257

Relative absolute error 100.027 %

Root relative squared error 100.1091 %

Total Number of Instances 435

----------------------------------------------------------------------------------

House Votes 84 – ZeroR

=== Summary ===

Correctly Classified Instances 269 61.8391 %

Incorrectly Classified Instances 166 38.1609 %

Kappa statistic 0

Mean absolute error 0.3615

Root mean squared error 0.4253

Relative absolute error 100 %

Root relative squared error 100 %

Total Number of Instances 435

----------------------------------------------------------------------------------

Use logic-based approaches (such as those listed under the group of rules and trees in Weka) to build classification models for the following two data sets, originating from the Irvine Machine Leaning Repository (http://archive.ics.uci.edu/ml/):

1. Iris (http://archive.ics.uci.edu/ml/datasets/Iris)

2. Congressional Voting Records (http://archive.ics.uci.edu/ml/datasets/Congressional+Voting+Records)

(For your convenience, I have placed both data sets in text format on this website.)

1. The models should help in classification of iris flowers (1) and congressmen (2) based on flower properties and voting record, respectively.
2. Please clean up and discretize the data, if needed (if you need domain knowledge, you can read more about the data sets in the descriptions placed at UCI Repository).
3. Make sure to try different learning algorithms and their parameters.
4. Once you have learned models, check and report their classification accuracy (by means of a cross-validation method, such as "leave-one-out").
5. Try to improve the accuracy by using different learning approaches, different values of parameters, and injecting outside knowledge.
6. Report the best classification accuracy that you have been able to achieve for each of the two data sets along with the methods that gave you the best accuracy.
7. Make sure to summarize your observations and conclusions.