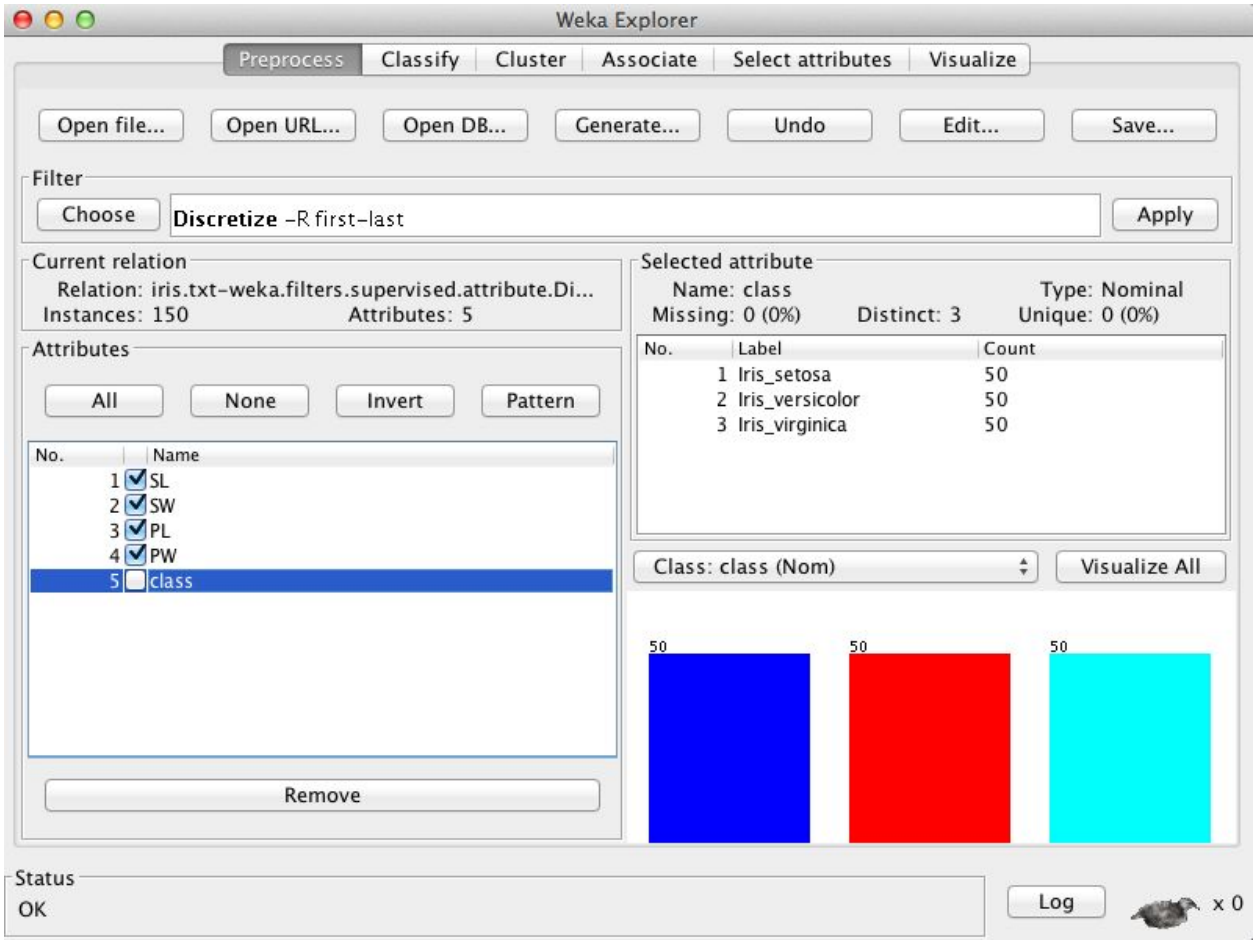
Assignment 7

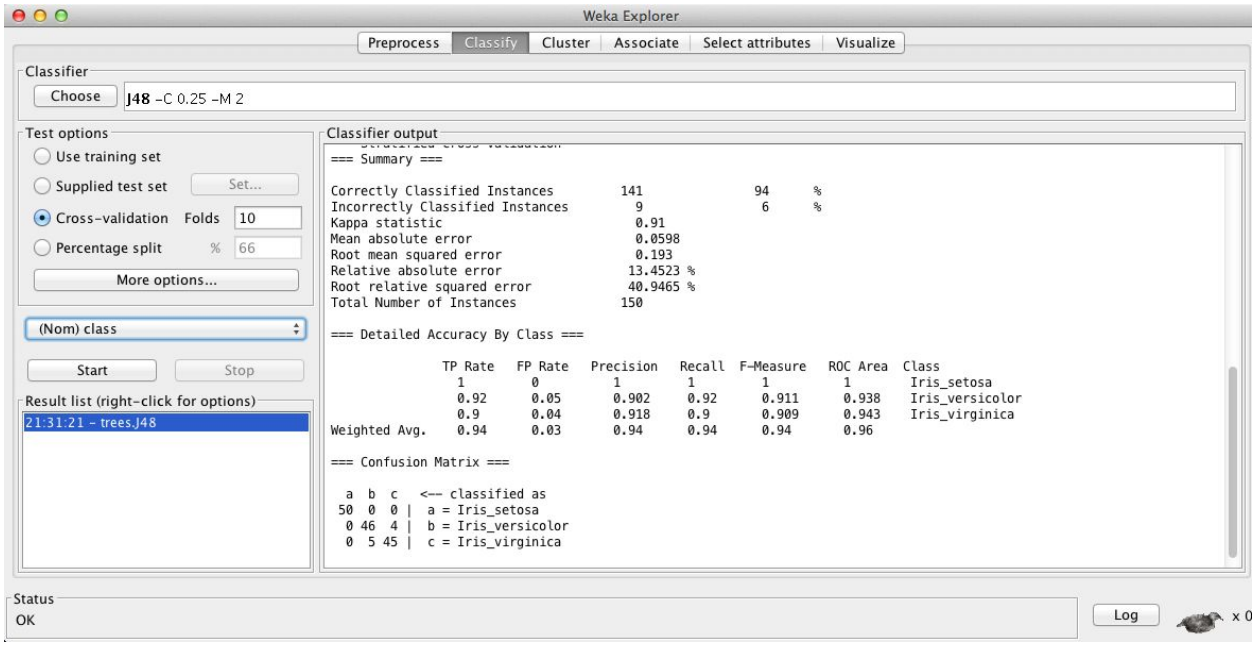
Zhaoyan Ai (zha4), JaJan Hsu (jah247), Leilei Liu (lel74)

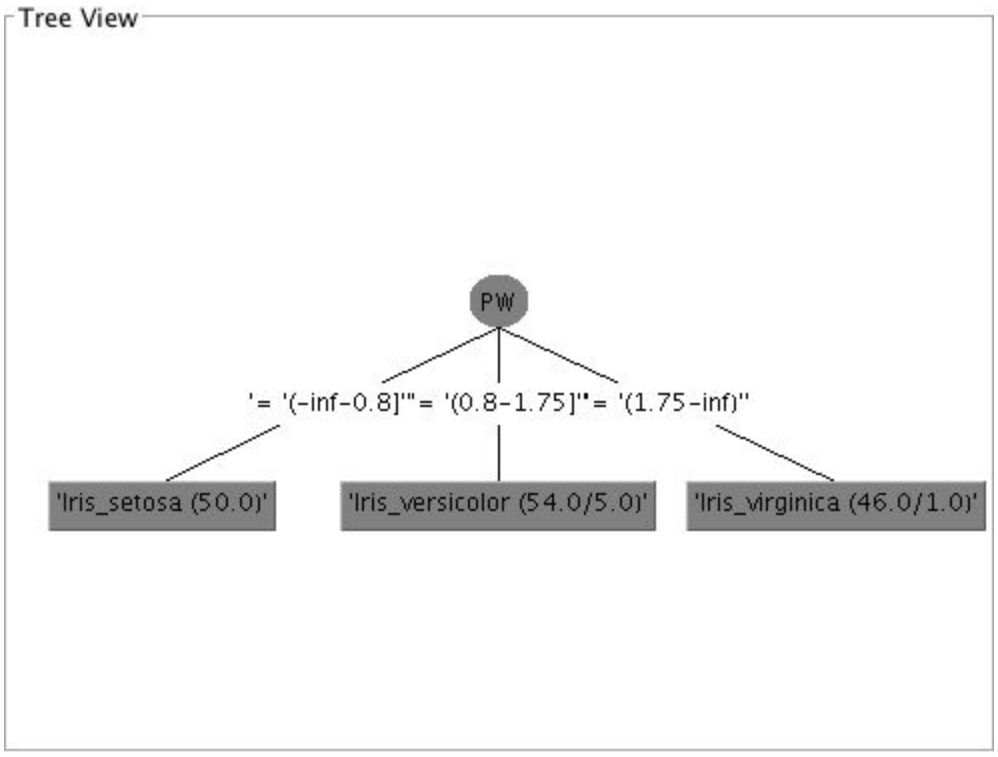
Classification of iris flowers

Firstly we cleaned up the data and discretized it.



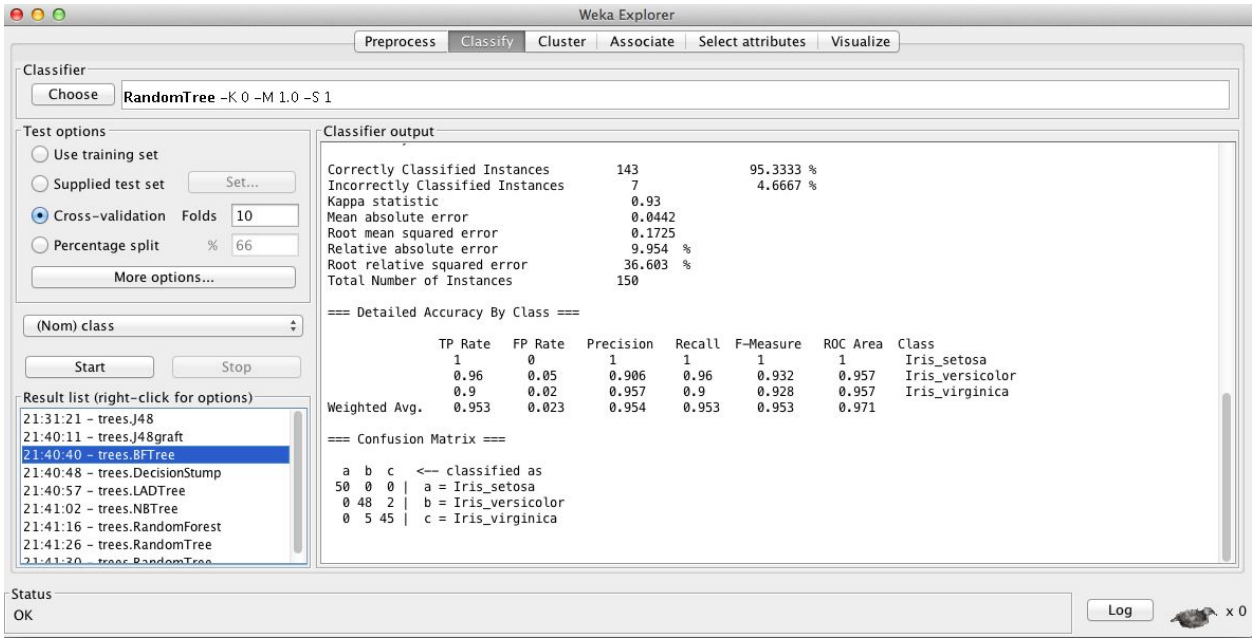
1. Cross-validation: folds = 10, Classifier = J48
   Accuracy rate：94%


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...
(Nom) class
Start
Stop
Result list (right-click for options)
21:31:21 - trees.J48
Classifier output
=== Summary ===
Correctly Classified Instances 141 94 %
Incorrectly Classified Instances 9 6 %
Kappa statistic 0.91
Mean absolute error 0.0598
Root mean squared error 0.193
Relative absolute error 13.4523 %
Root relative squared error 40.9465 %
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
0.92 0.05 0.902 0.92 0.911 0.938 Iris_versicolor
0.9 0.04 0.918 0.9 0.909 0.943 Iris_virginica
Weighted Avg. 0.94 0.03 0.94 0.94 0.94 0.96
=== Confusion Matrix ===
a b c <-- classified as
50 0 0 | a = Iris_setosa
0 46 4 | b = Iris_versicolor
0 5 45 | c = Iris_virginica
Status
OK
Log
x 0


Tree View
PW
'= '(-inf-0.8]'"= '(0.8-1.75]'"= '(1.75-inf)'
'Iris_setosa (50.0)'
'Iris_versicolor (54.0/5.0)'
'Iris_virginica (46.0/1.0)'

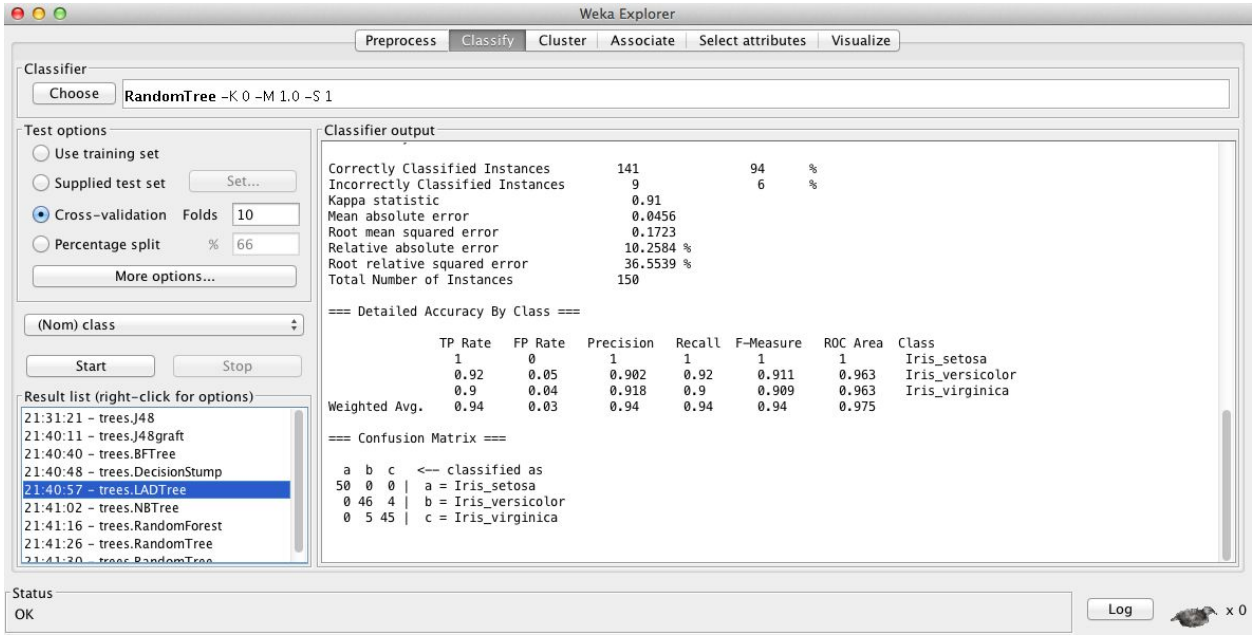
2. Cross-validation: folds = 10, Use training set: Classifier = BFTree

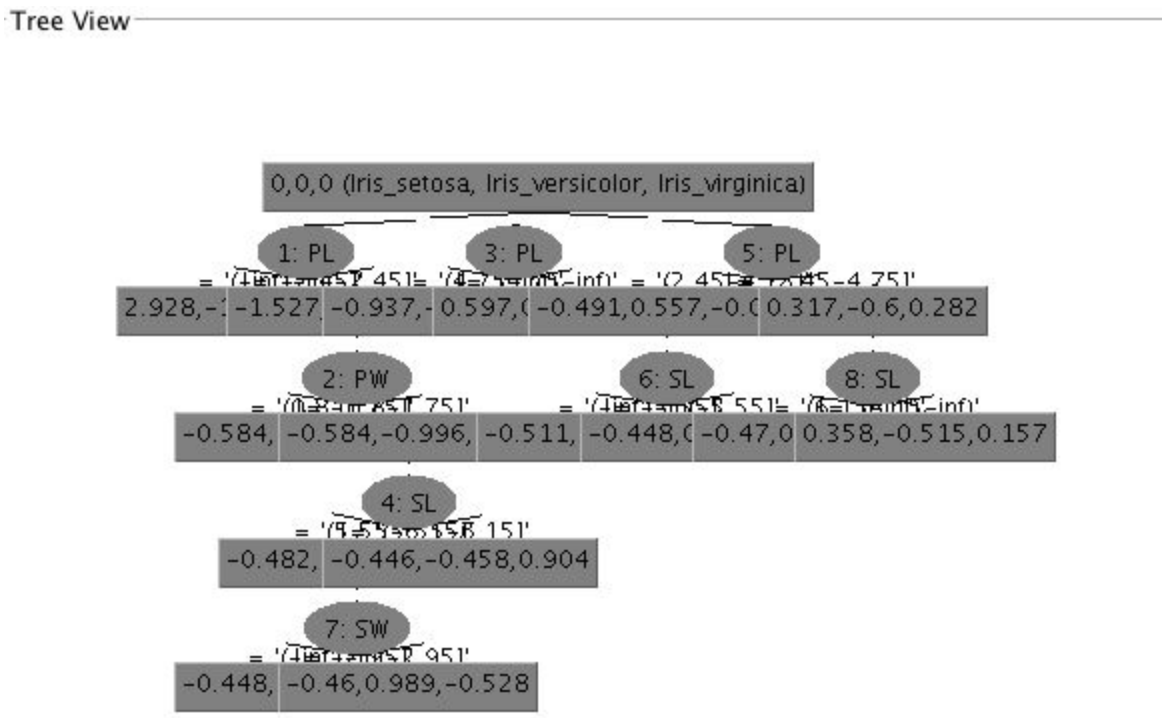
Accuracy rate：95.33%



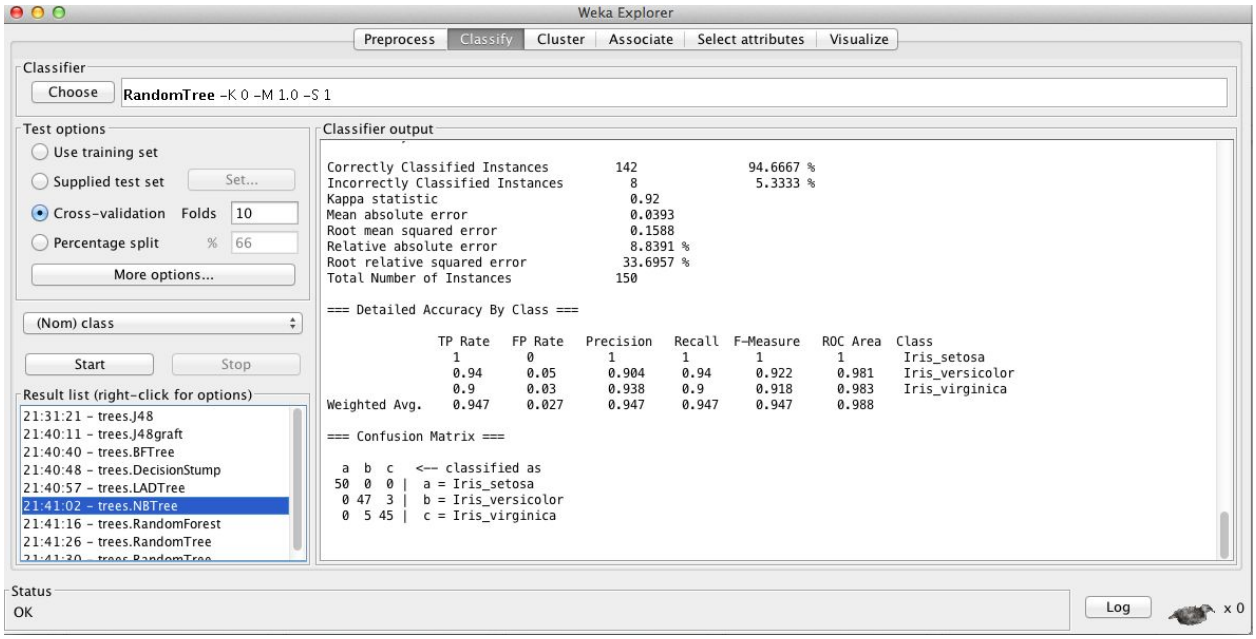
3. Cross-validation: folds = 10, Classifier = LADTree

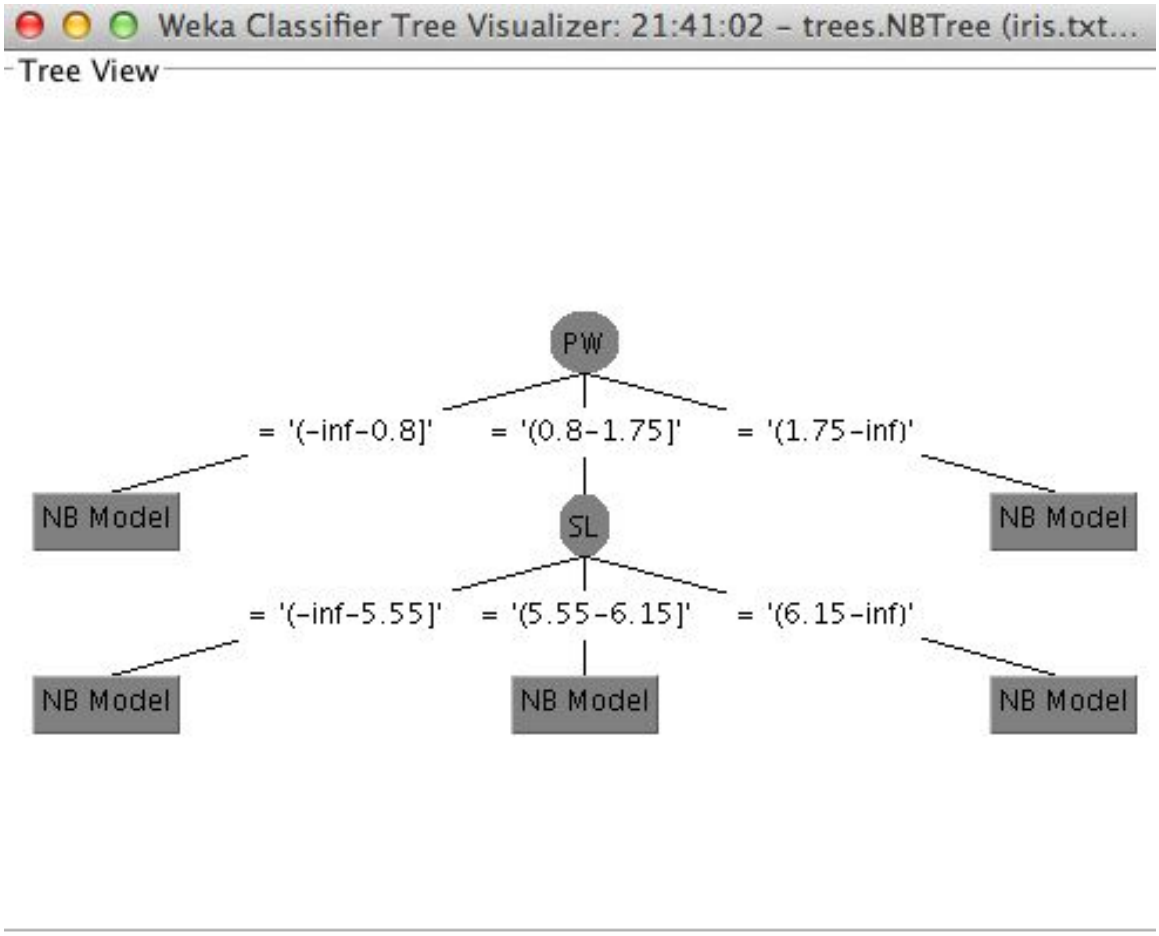
Accuracy rate = 94%





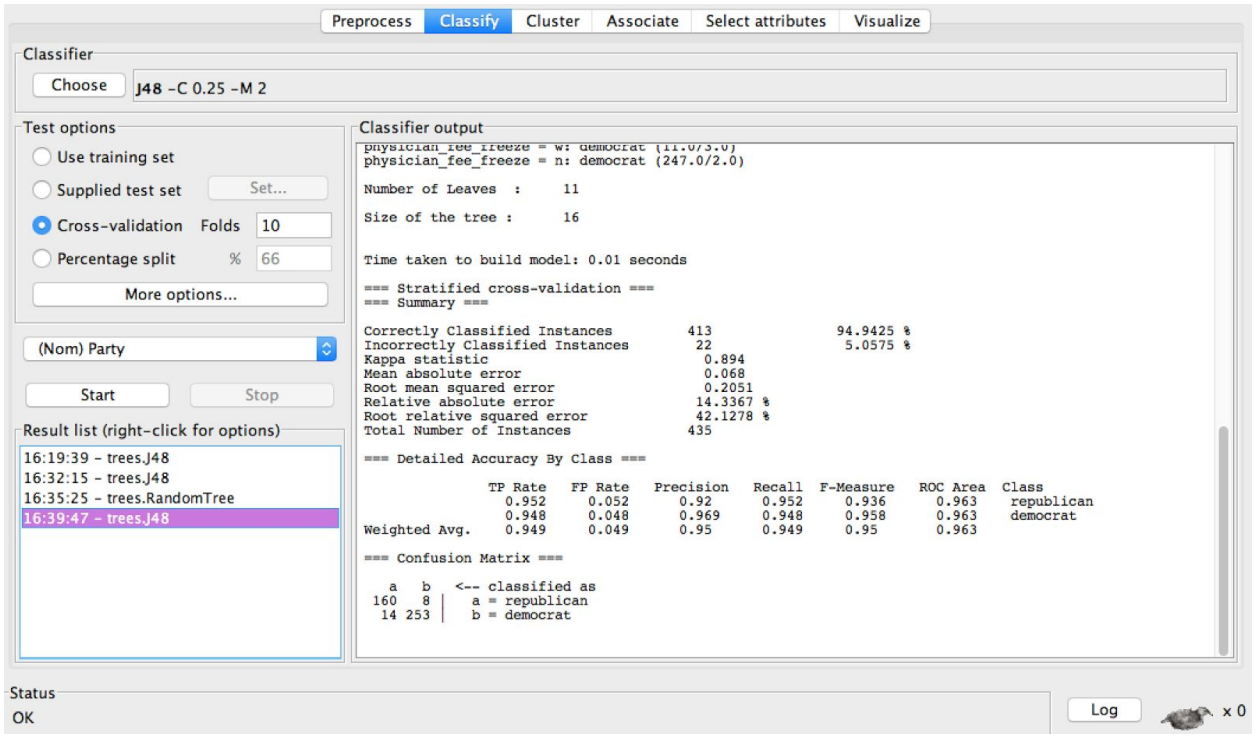
4. Cross-validation: folds = 10, Classifier = NBTree
   Accuracy rate = 94.6667%

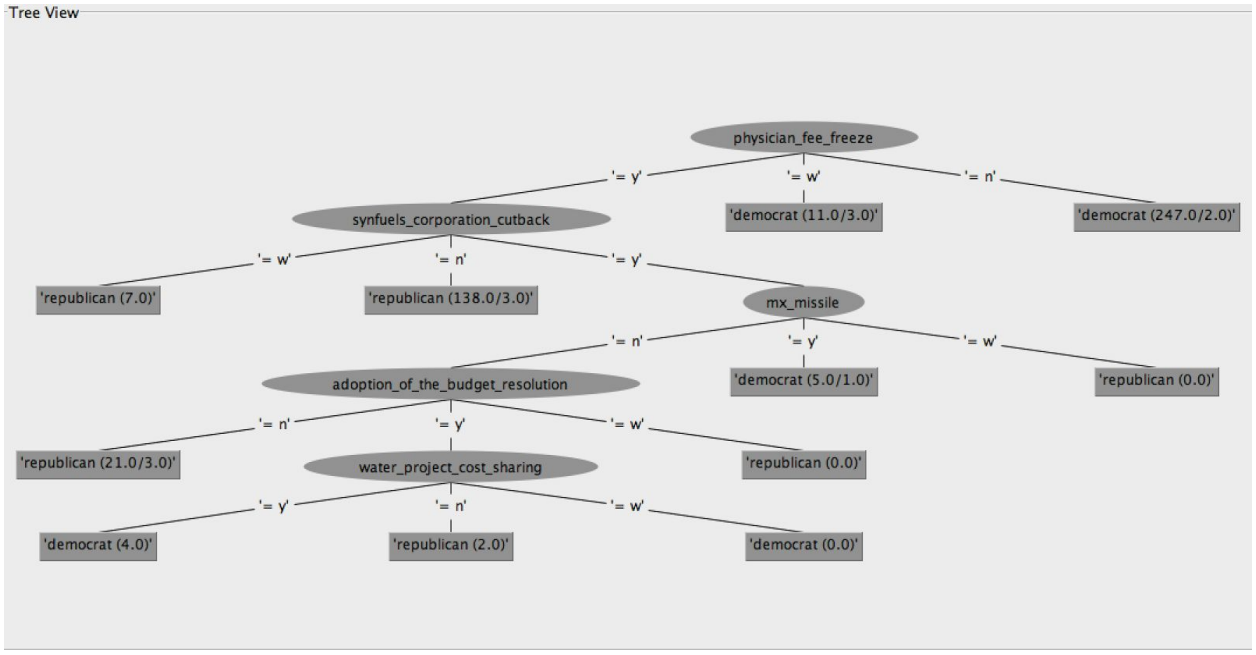




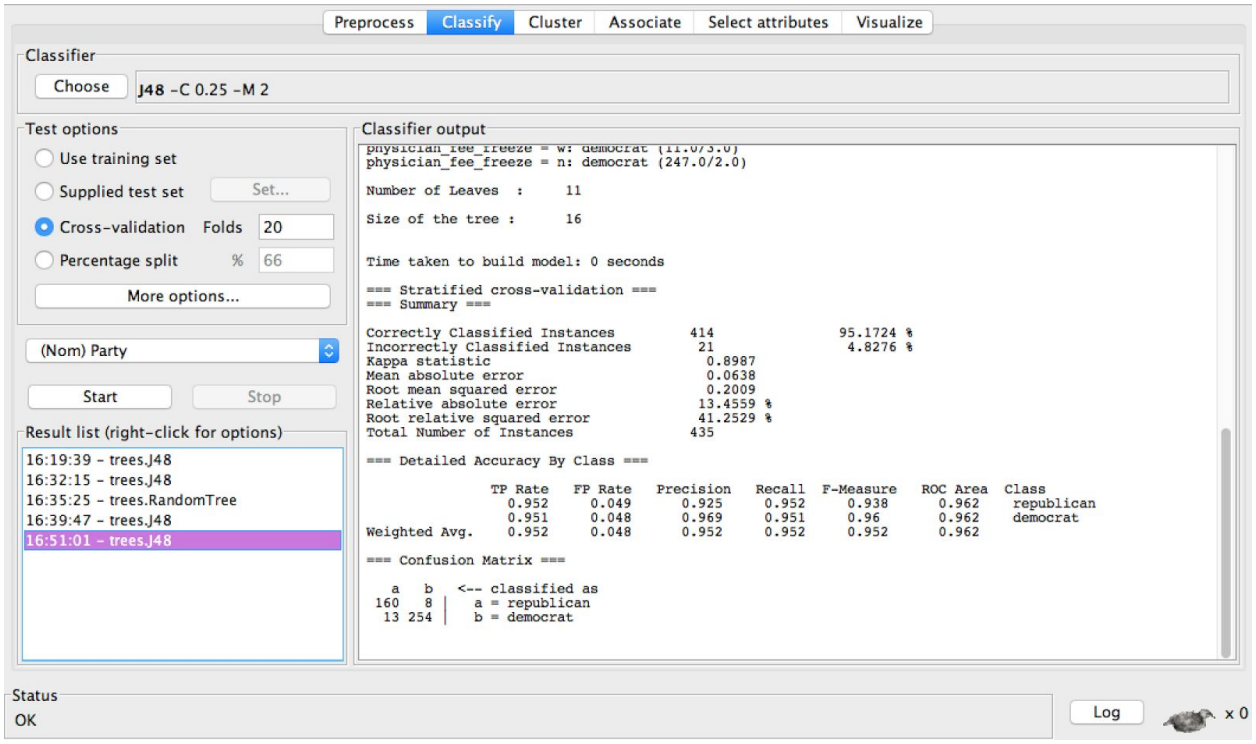
Classification of congressmen

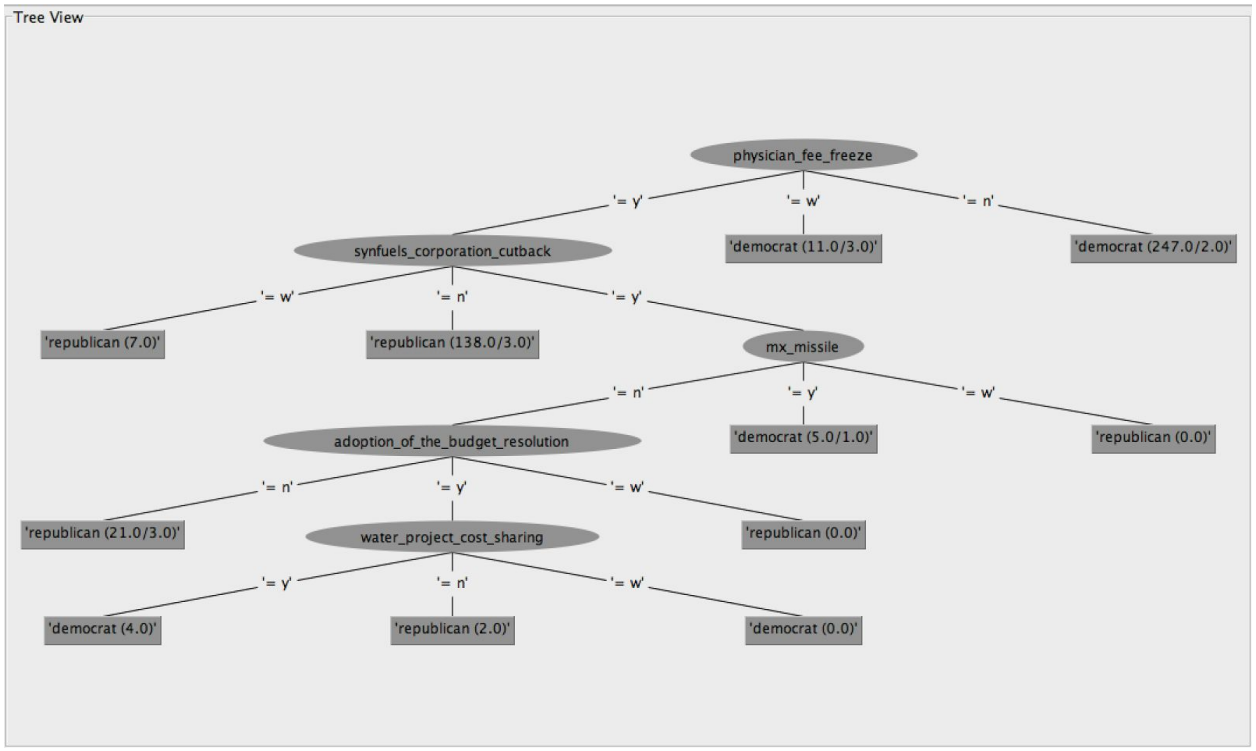
1. Cross-validation: folds = 10, Classifier = J48
   Accuracy rate：94.9425%



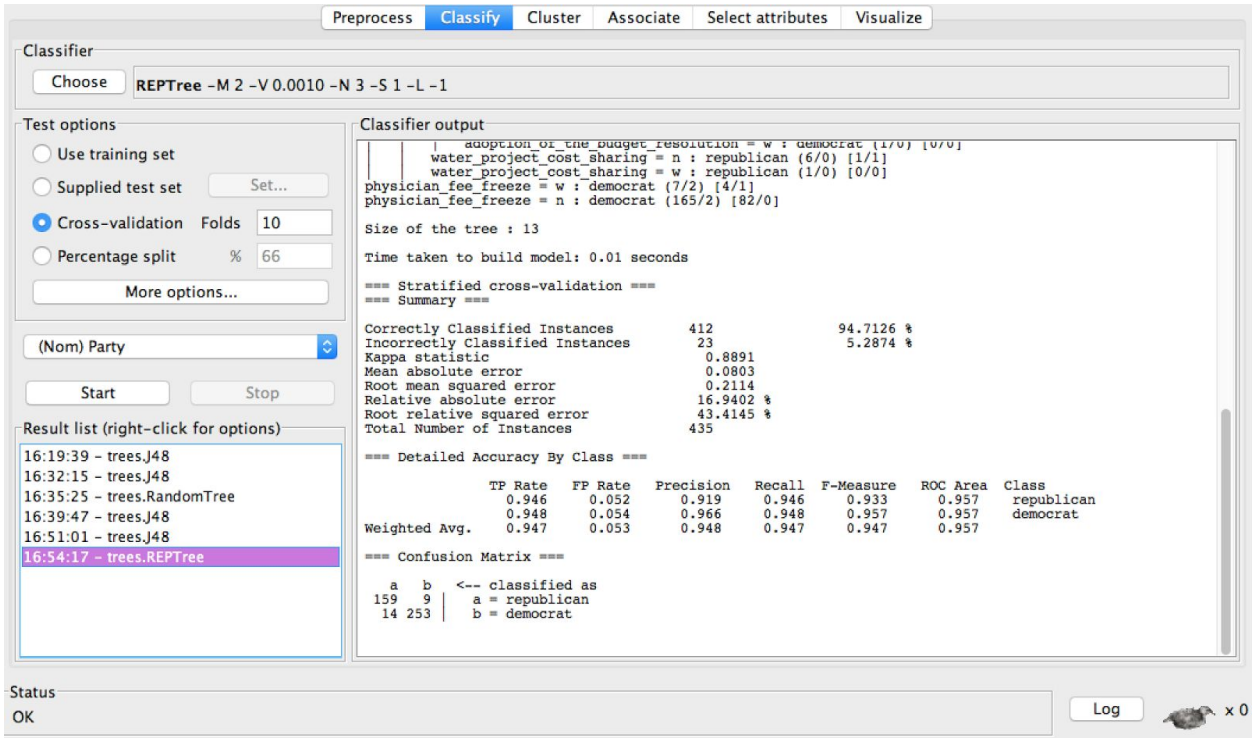


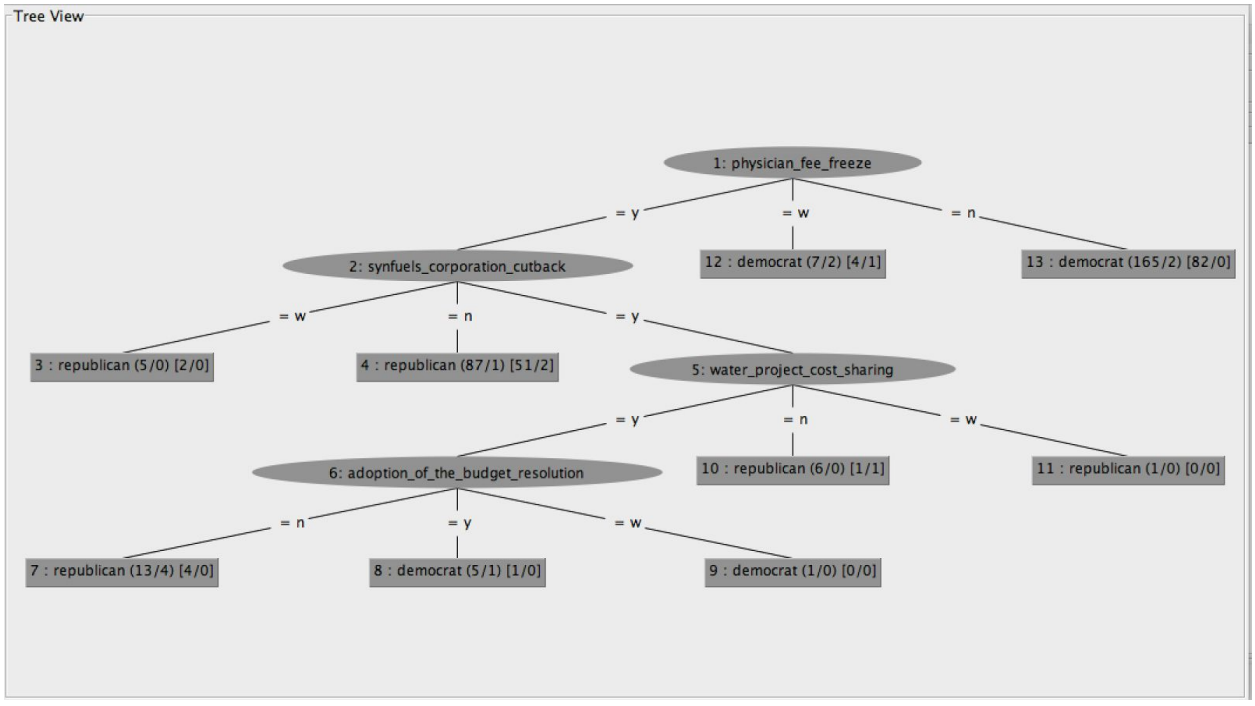
2. Cross-validation: folds = 10, Classifier = J48
   Accuracy rate：95.1724%





3. Cross-validation: folds = 10, Classifier = REPTree
   Accuracy rate：94.7126%


Preprocess
Classify
Cluster
Associate
Select attributes
Visualize
Classifier
Choose
REPTree -M 2 -V 0.0010 -N 3 -S 1 -L -1
Test options
Use training set
Supplied test set
Set...
Cross-validation
Folds
10
Percentage split
%
66
More options...
(Nom) Party
Start
Stop
Result list (right-click for options)
16:19:39 - trees.J48
16:32:15 - trees.J48
16:35:25 - trees.RandomTree
16:39:47 - trees.J48
16:51:01 - trees.J48
16:54:17 - trees.REPTree
Classifier output
water_project_cost_sharing = n : republican (6/0) [1/1]
water_project_cost_sharing = w : republican (1/0) [0/0]
physician_fee_freeze = w : democrat (7/2) [4/1]
physician_fee_freeze = n : democrat (165/2) [82/0]
Size of the tree : 13
Time taken to build model: 0.01 seconds
=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances 412 94.7126 %
Incorrectly Classified Instances 23 5.2874 %
Kappa statistic 0.8891
Mean absolute error 0.0803
Root mean squared error 0.2114
Relative absolute error 16.9402 %
Root relative squared error 43.4145 %
Total Number of Instances 435
=== Detailed Accuracy By Class ===
TP Rate FP Rate Precision Recall F-Measure ROC Area Class
0.946 0.052 0.919 0.946 0.933 0.957 republican
0.948 0.054 0.966 0.948 0.957 0.957 democrat
Weighted Avg. 0.947 0.053 0.948 0.947 0.947 0.957
=== Confusion Matrix ===
a b <-- classified as
159 9 | a = republican
14 253 | b = democrat
Status
OK
Log
x 0


Tree View
1: physician_fee_freeze
= y
= w
= n
2: synfuels_corporation_cutback
12 : democrat (7/2) [4/1]
13 : democrat (165/2) [82/0]
= w
= n
= y
3 : republican (5/0) [2/0]
4 : republican (87/1) [51/2]
5: water_project_cost_sharing
= y
= n
= w
6: adoption_of_the_budget_resolution
10 : republican (6/0) [1/1]
11 : republican (1/0) [0/0]
= n
= y
= w
7 : republican (13/4) [4/0]
8 : democrat (5/1) [1/0]
9 : democrat (1/0) [0/0]

Summary

Iris flower

We used four classification methods - J48, BFTree, LADTree, NBTree, to analyze the iris flower data. The BFTree has best accuracy but it does not have Tree view. The second best tree is NBTree (94.6667% accuracy). From those Tree views, we can see the variable PW may be an important factor. It might have decisive influence on the overall result in the Iris flower data.

Congressmen

It is easy to see that no matter what algorithm we used (J48 or REPTree), physician_fee_freeze is always the most important factor to classify democrat and republican, followed by synfueis_corporation_cutback. The result from classification done with J48 has an additional mx_missle factor, compared with the result from REPTree algorithm. Both algorithms consider adoption_of_the_budget_resolution and water_project_cost_sharing important factors to classify democrat and republican, nevertheless result from J48 puts adoption_of_the_budget_resolution ahead of water_projection_cost_sharing while result from REPTree interprets the other way. To sum up, J48 algorithm achieved a more reliable result judging from the accuracy rate.