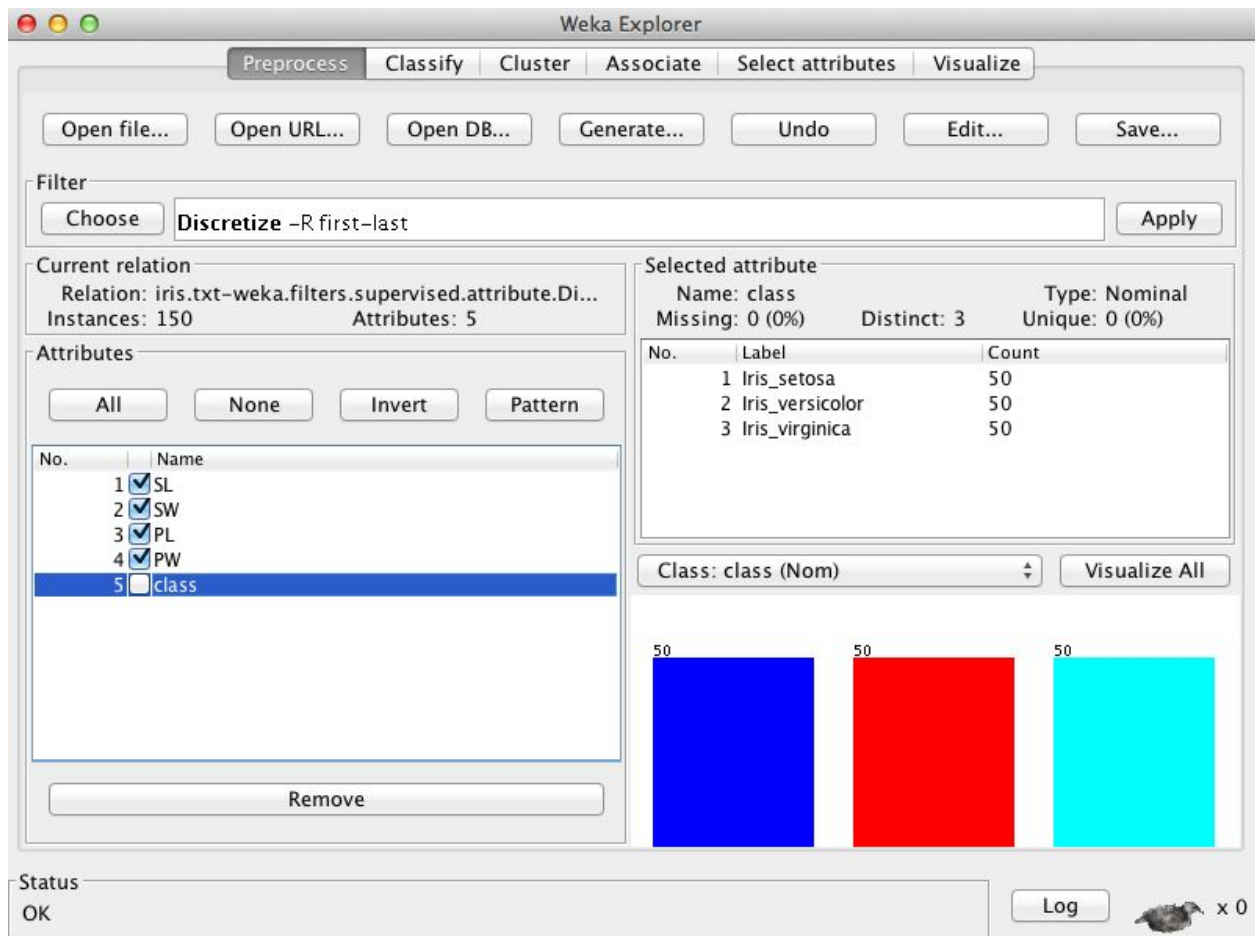


Assignment 7

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 ====

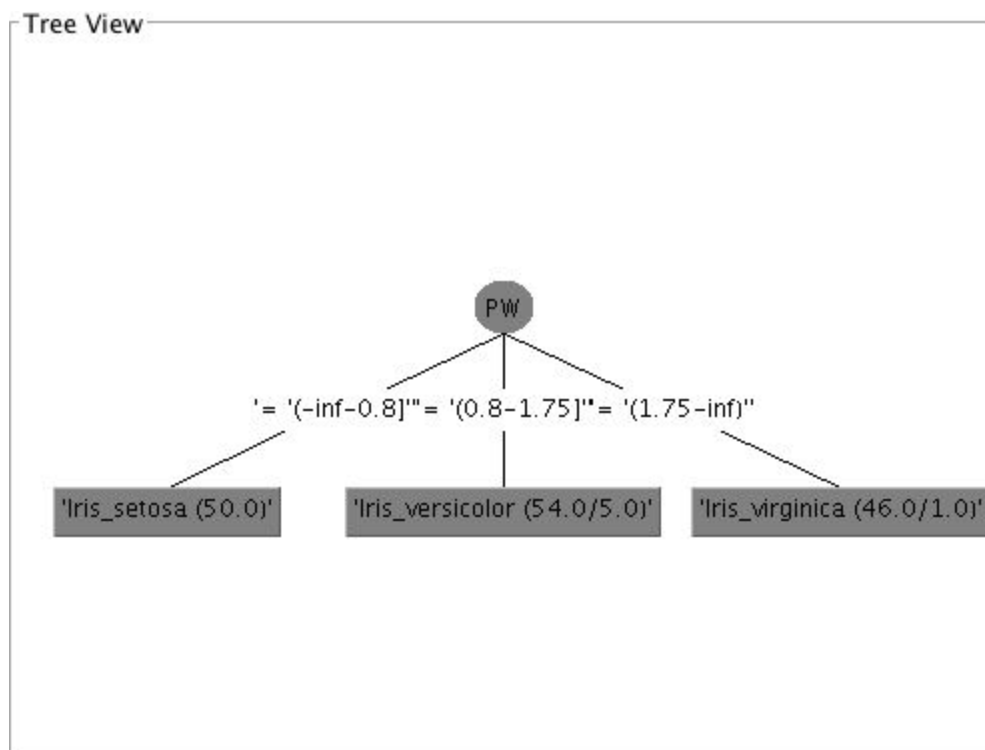
	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	0	1	1	1	1	Iris_setosa
	0.92	0.05	0.902	0.92	0.911	0.938	Iris_versicolor
	0.9	0.04	0.918	0.9	0.909	0.943	Iris_virginica
Weighted Avg.	0.94	0.03	0.94	0.94	0.94	0.96	

==== 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



2. Cross-validation: folds = 10, Use training set: Classifier = BFTree
Accuracy rate : 95.33%

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier is 'RandomTree -K 0 -M 1.0 -S 1'. The 'Test options' section has 'Cross-validation' selected with 'Folds' set to 10. The 'Classifier output' section displays the following metrics:

Metric	Value	Percentage
Correctly Classified Instances	143	95.3333 %
Incorrectly Classified Instances	7	4.6667 %
Kappa statistic	0.93	
Mean absolute error	0.0442	
Root mean squared error	0.1725	
Relative absolute error	9.954 %	
Root relative squared error	36.603 %	
Total Number of Instances	150	

Below the metrics is a 'Detailed Accuracy By Class' table:

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	1	0	1	1	1	1	Iris_setosa
0.96	0.05	0.906	0.96	0.932	0.957	Iris_versicolor	
0.9	0.02	0.957	0.9	0.928	0.957	Iris_virginica	
Weighted Avg.	0.953	0.023	0.954	0.953	0.953	0.971	

A 'Confusion Matrix' is also shown:

```

a b c <-- classified as
50 0 0 | a = Iris_setosa
0 48 2 | b = Iris_versicolor
0 5 45 | c = Iris_virginica
  
```

3. Cross-validation: folds = 10, Classifier = LADTree
Accuracy rate = 94%

The screenshot shows the Weka Explorer interface with the 'Classify' tab selected. The classifier is 'RandomTree -K 0 -M 1.0 -S 1'. The 'Test options' section has 'Cross-validation' selected with 'Folds' set to 10. The 'Classifier output' section displays the following metrics:

Metric	Value	Percentage
Correctly Classified Instances	141	94 %
Incorrectly Classified Instances	9	6 %
Kappa statistic	0.91	
Mean absolute error	0.0456	
Root mean squared error	0.1723	
Relative absolute error	10.2584 %	
Root relative squared error	36.5539 %	
Total Number of Instances	150	

Below the metrics is a 'Detailed Accuracy By Class' table:

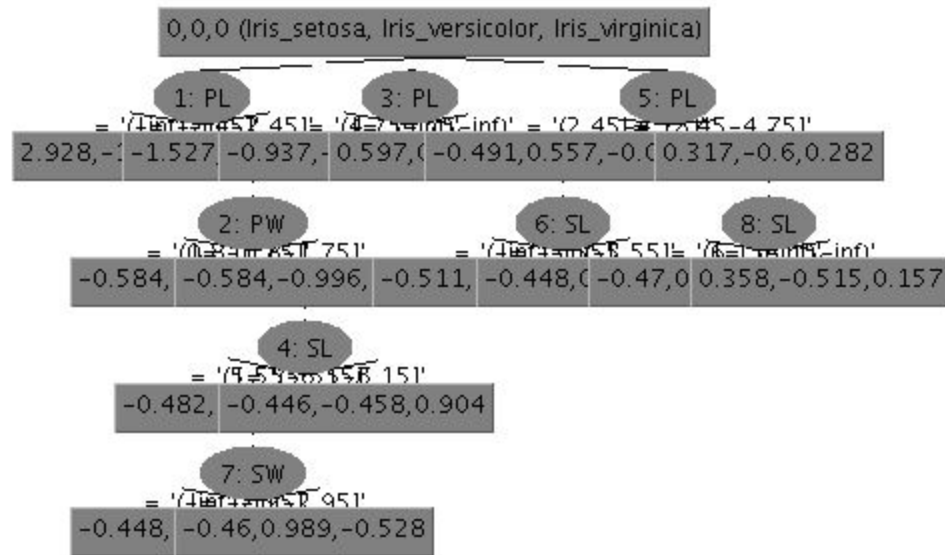
	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	1	0	1	1	1	1	Iris_setosa
0.92	0.05	0.902	0.92	0.911	0.963	Iris_versicolor	
0.9	0.04	0.918	0.9	0.909	0.963	Iris_virginica	
Weighted Avg.	0.94	0.03	0.94	0.94	0.94	0.975	

A 'Confusion Matrix' is also shown:

```

a b c <-- classified as
50 0 0 | a = Iris_setosa
0 46 4 | b = Iris_versicolor
0 5 45 | c = Iris_virginica
  
```

Tree View



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

Classifier
Choose RandomTree -K 0 -M 1.0 -S 1

Test options
☐ Use training set
☐ Supplied test set Set...
☒ Cross-validation Folds 10
☐ Percentage split % 66
 More options...

Classifier output

Correctly Classified Instances	142	94.6667 %
Incorrectly Classified Instances	8	5.3333 %
Kappa statistic	0.92	
Mean absolute error	0.0393	
Root mean squared error	0.1588	
Relative absolute error	8.8391 %	
Root relative squared error	33.6957 %	
Total Number of Instances	150	

=== Detailed Accuracy By Class ===

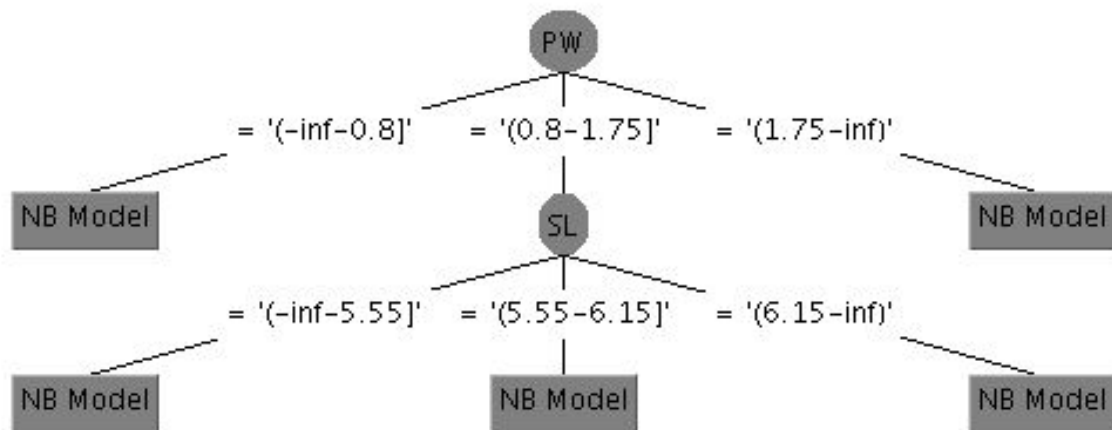
	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	1	0	1	1	1	1	Iris_setosa
0.94	0.05	0.904	0.94	0.922	0.981	Iris_versicolor	
0.9	0.03	0.938	0.9	0.918	0.983	Iris_virginica	
Weighted Avg.	0.947	0.027	0.947	0.947	0.947	0.988	

=== Confusion Matrix ===

	a	b	c	<-- classified as
50	0	0		a = Iris_setosa
0	47	3		b = Iris_versicolor
0	5	45		c = Iris_virginica

Result list (right-click for options)
 21:31:21 - trees.J48
 21:40:11 - trees.J48graft
 21:40:40 - trees.BFTree
 21:40:48 - trees.DecisionStump
 21:40:57 - trees.LADTree
 21:41:02 - trees.NBTree
 21:41:16 - trees.RandomForest
 21:41:26 - trees.RandomTree
 21:41:30 - trees.RandomTree

Status: OK Log x 0



Classification of congressmen

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

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) 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

Classifier output

```

physician_fee_freeze = w: democrat (11.0/3.0)
physician_fee_freeze = n: democrat (247.0/2.0)

Number of Leaves :    11
Size of the tree :    16

Time taken to build model: 0.01 seconds

=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      413      94.9425 %
Incorrectly Classified Instances    22       5.0575 %
Kappa statistic                    0.894
Mean absolute error                 0.068
Root mean squared error             0.2051
Relative absolute error             14.3367 %
Root relative squared error        42.1278 %
Total Number of Instances          435

=== Detailed Accuracy By Class ===
               TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
Weighted Avg.   0.952   0.052    0.92      0.952   0.936     0.963   republican
                0.948   0.048    0.969    0.948   0.958     0.963   democrat

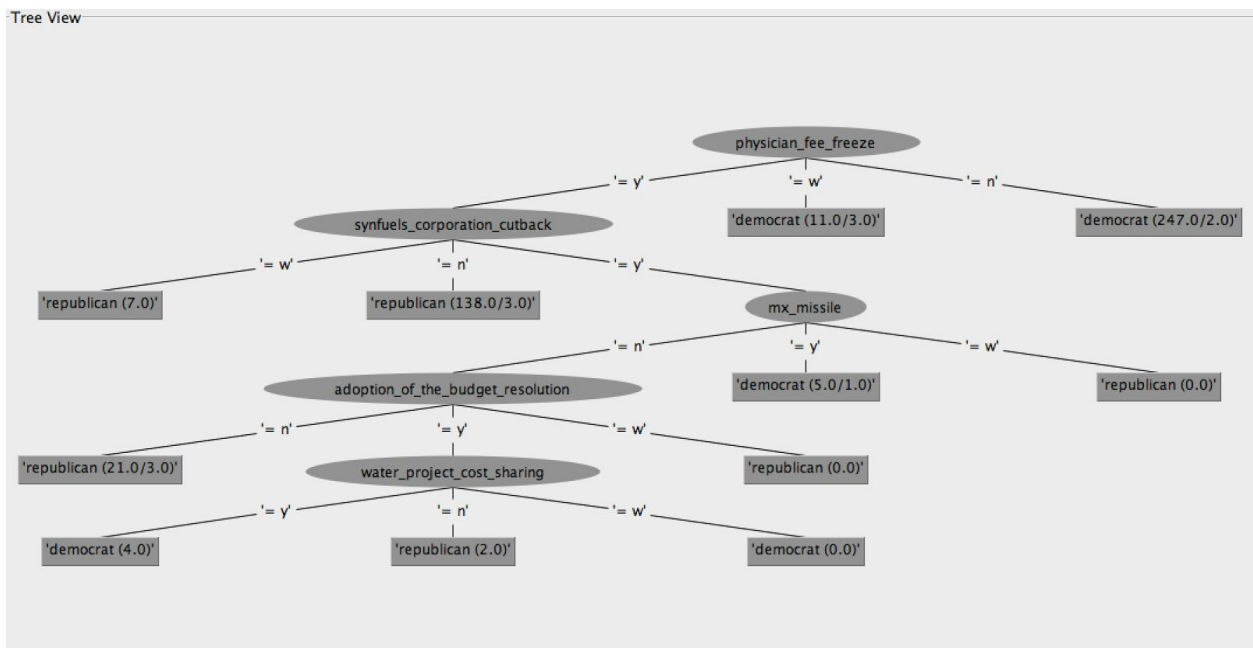
```

=== Confusion Matrix ===

a \ b	160	8	14	253
a = republican	160	8		
a = democrat			14	253

Status
OK

Log x 0



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

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 **20**
☐ 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**

Classifier output

```

physician_fee_freeze = w: democrat (11.0/3.0)
physician_fee_freeze = n: democrat (247.0/2.0)

Number of Leaves :    11
Size of the tree :    16

Time taken to build model: 0 seconds


=== Stratified cross-validation ===
=== Summary ===
Correctly Classified Instances      414      95.1724 %
Incorrectly Classified Instances    21      4.8276 %
Kappa statistic                    0.8987
Mean absolute error                 0.0638
Root mean squared error             0.2009
Relative absolute error             13.4559 %
Root relative squared error         41.2529 %
Total Number of Instances          435

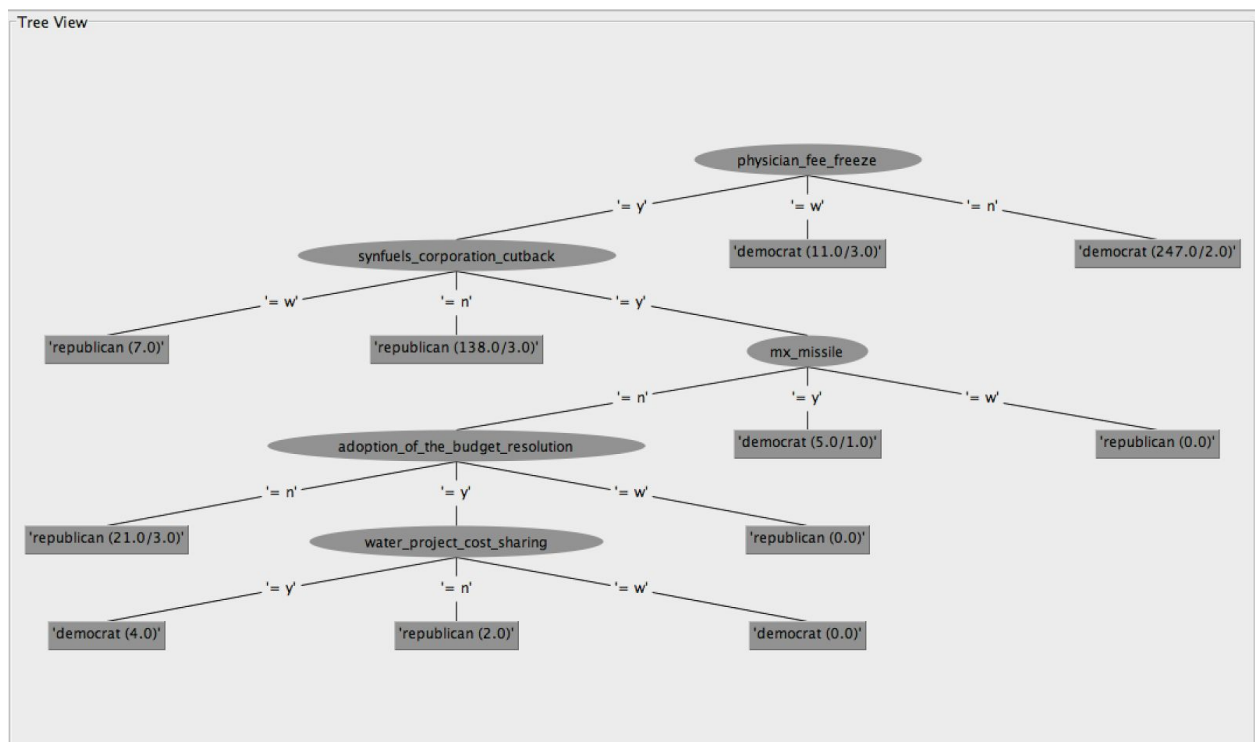
=== Detailed Accuracy By Class ===
                TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
Weighted Avg.   0.952    0.049    0.925     0.952    0.938     0.962    republican
                0.951    0.048    0.969     0.951    0.96      0.962    democrat

=== Confusion Matrix ===
  a  b  |-- classified as
160 8  | a = republican
 13 254 | b = democrat

```

Status
OK

Log  x 0



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

```

| adoption_of_the_budget_resolution = w : democrat (1/0) [0/0]
| 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

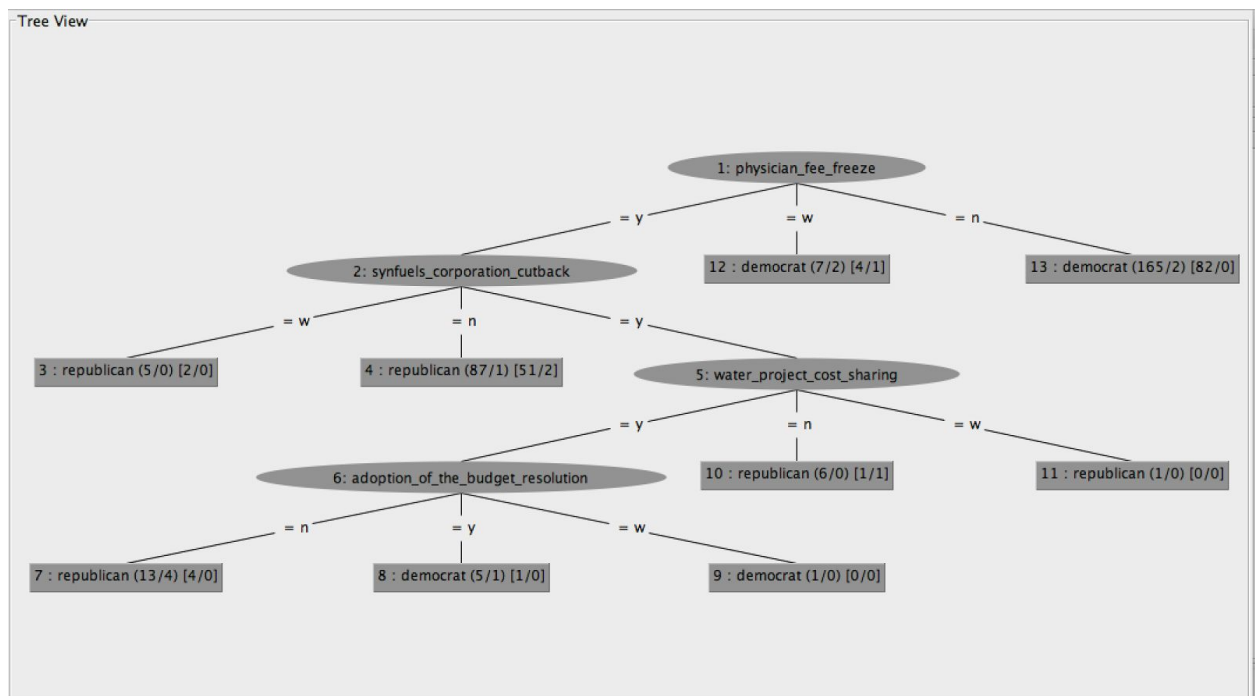
=== Detailed Accuracy By Class ===
               TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
Weighted Avg.   0.946   0.052   0.919    0.946   0.933     0.957   republican
                0.948   0.054   0.966   0.948   0.957     0.957   democrat

=== Confusion Matrix ===
  a  b  <-- classified as
159  9  a = republican
 14 253 b = democrat

```

Status
OK

Log x 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_missile 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.