Annie Steenson
COM S 474
Lab 2

How to replicate part 2 experiments

I included 10 data sets named split1-training, split1-test, split2-training, split2-test, …. split5-training, split5-test.

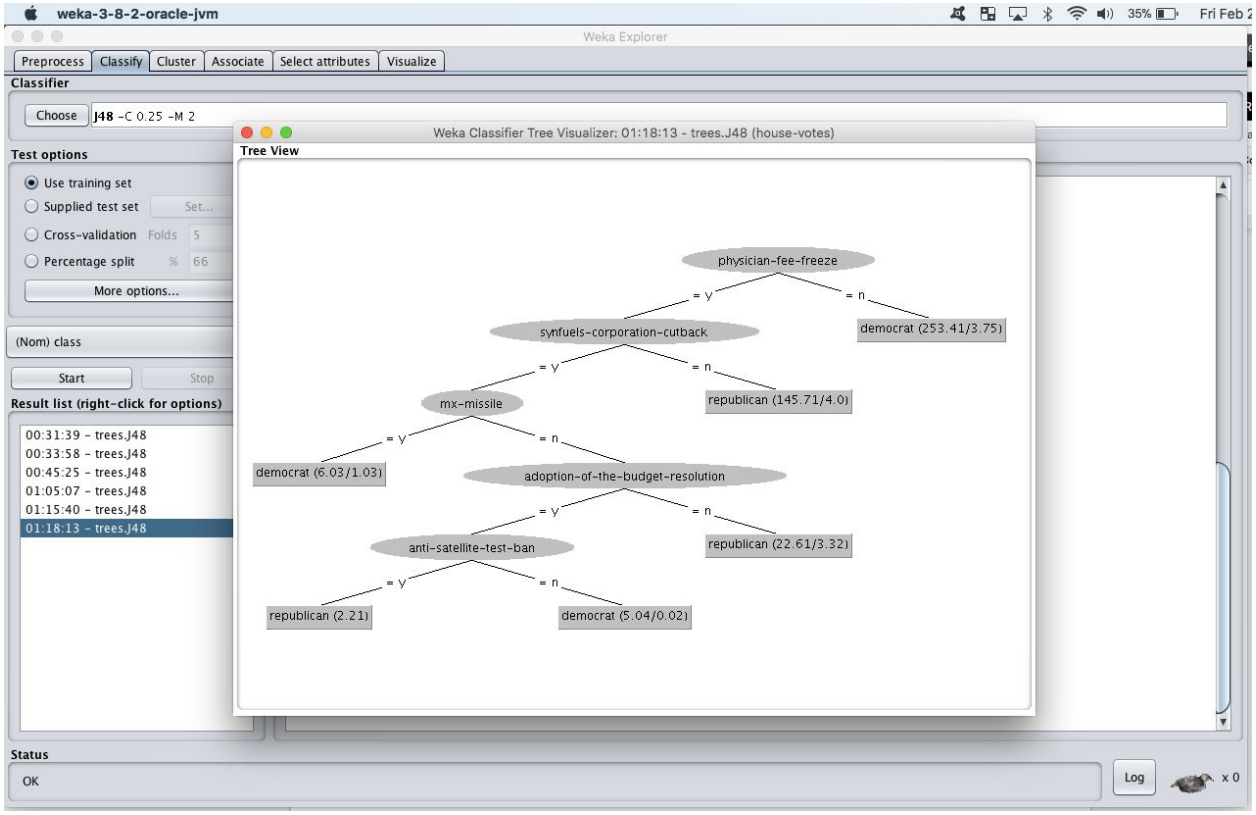
As they are named, the training set is used to train the model when using split i and the test set is used to test the model when using split i.

I used J48 and 5-fold-validation in Weka for all experiments.

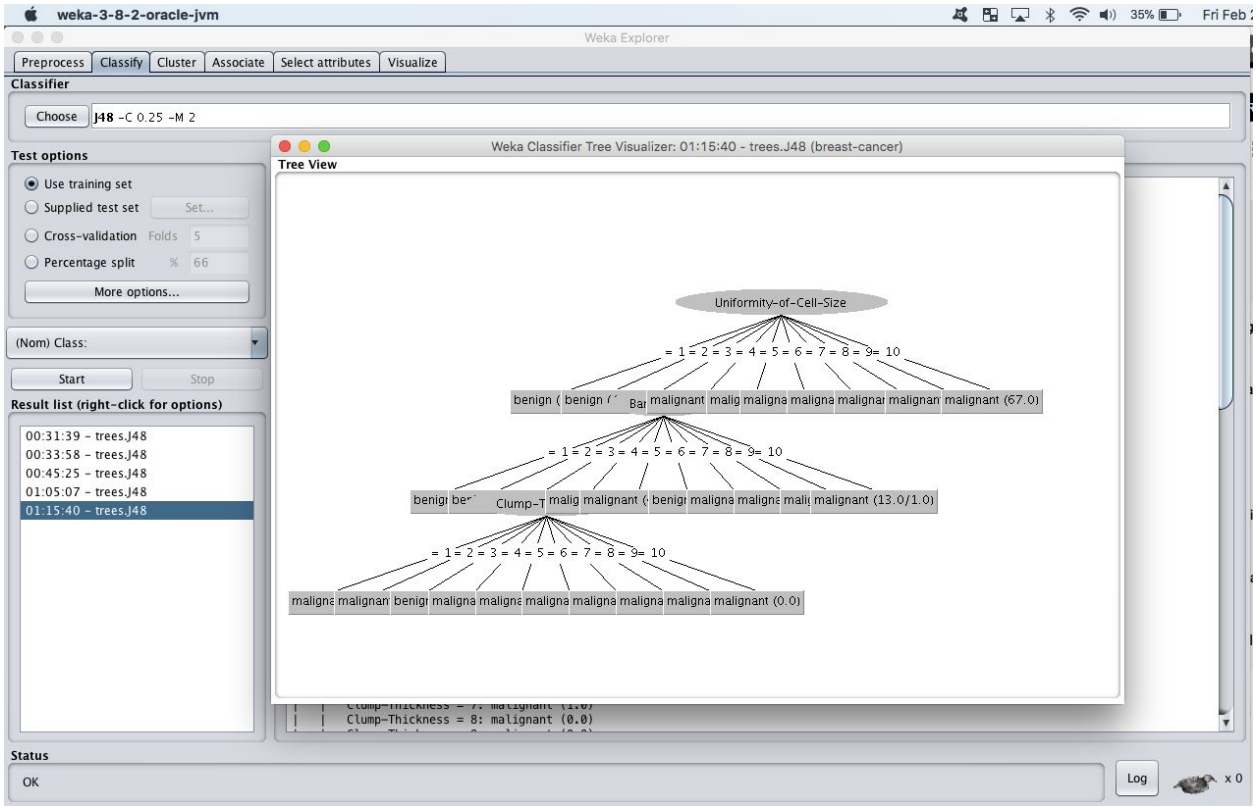
Write Up

1. Learn Decision Tree classifiers on the two data sets (for example, using J48 in Weka). Visualize the tree constructed by the decision tree algorithm. Are there some interesting rules that make sense based on what you understand about the data?

Votes:

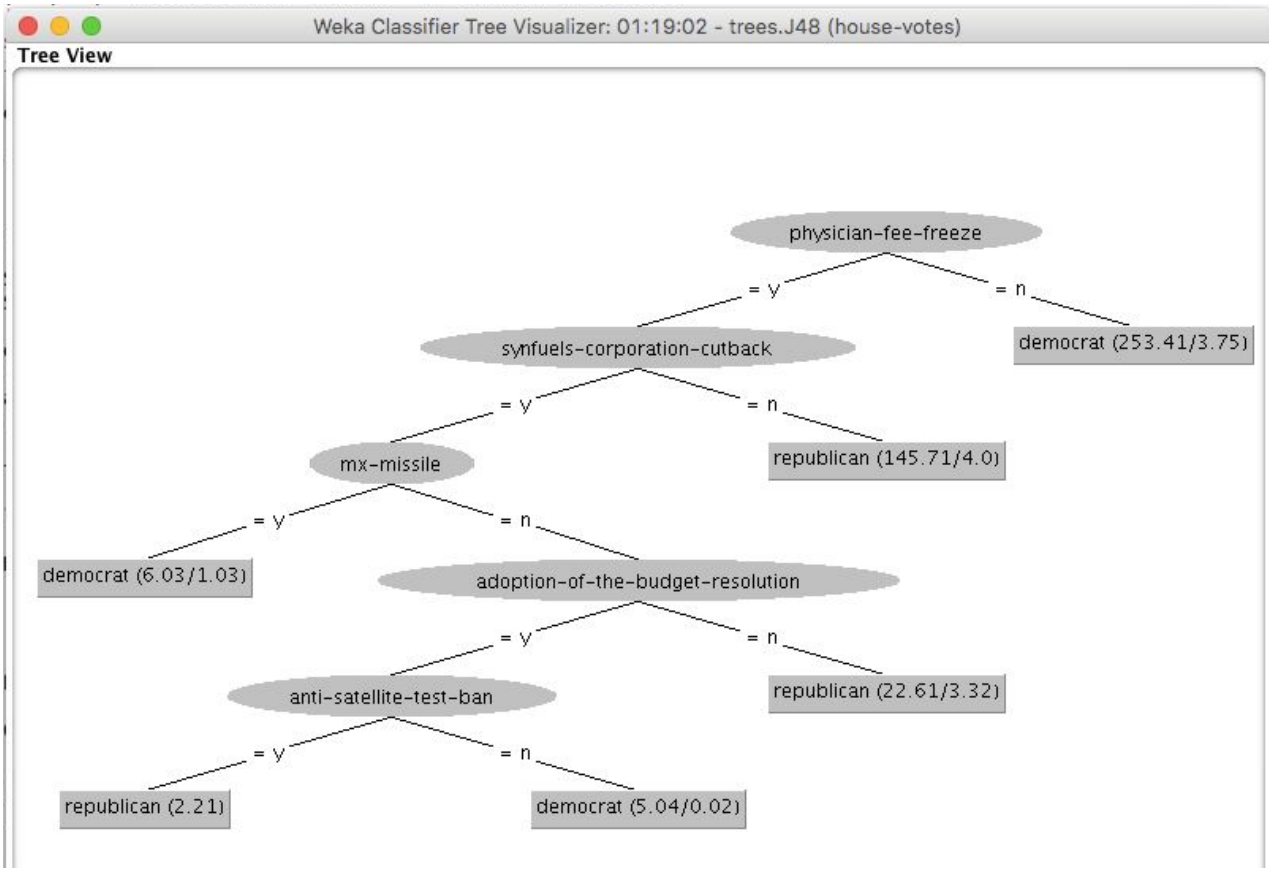


Breast Cancer:



2. Perform the following 5-fold cross-validation experiments with the Congressional Voting Records Data Set to study the stability of decision tree learning algorithm over the variability of data samples.

The tree constructed using all the data points:



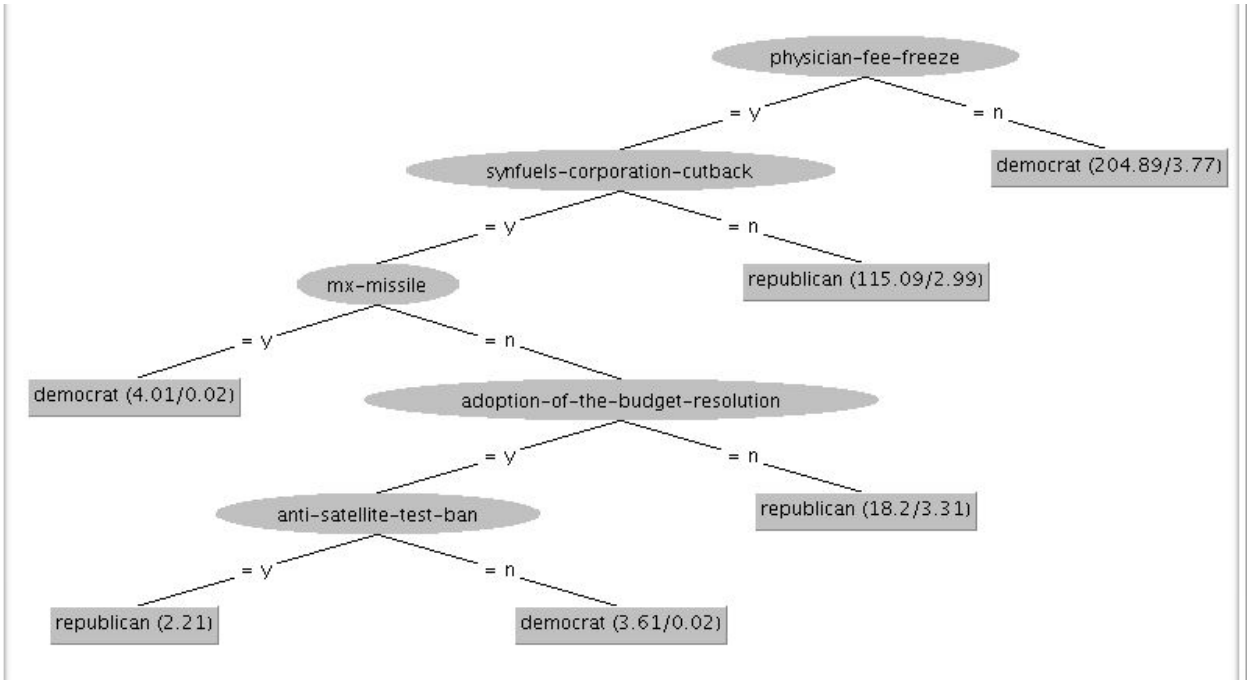
(a) Randomly split the dataset into 5 data sets of (roughly) equal size D1, D2, . . . , D5.

I divided the data by:
D1: the first 108 points
D2 : next 107 points
D3: next 107 points
D4: next 107 points
D5: last 111 points

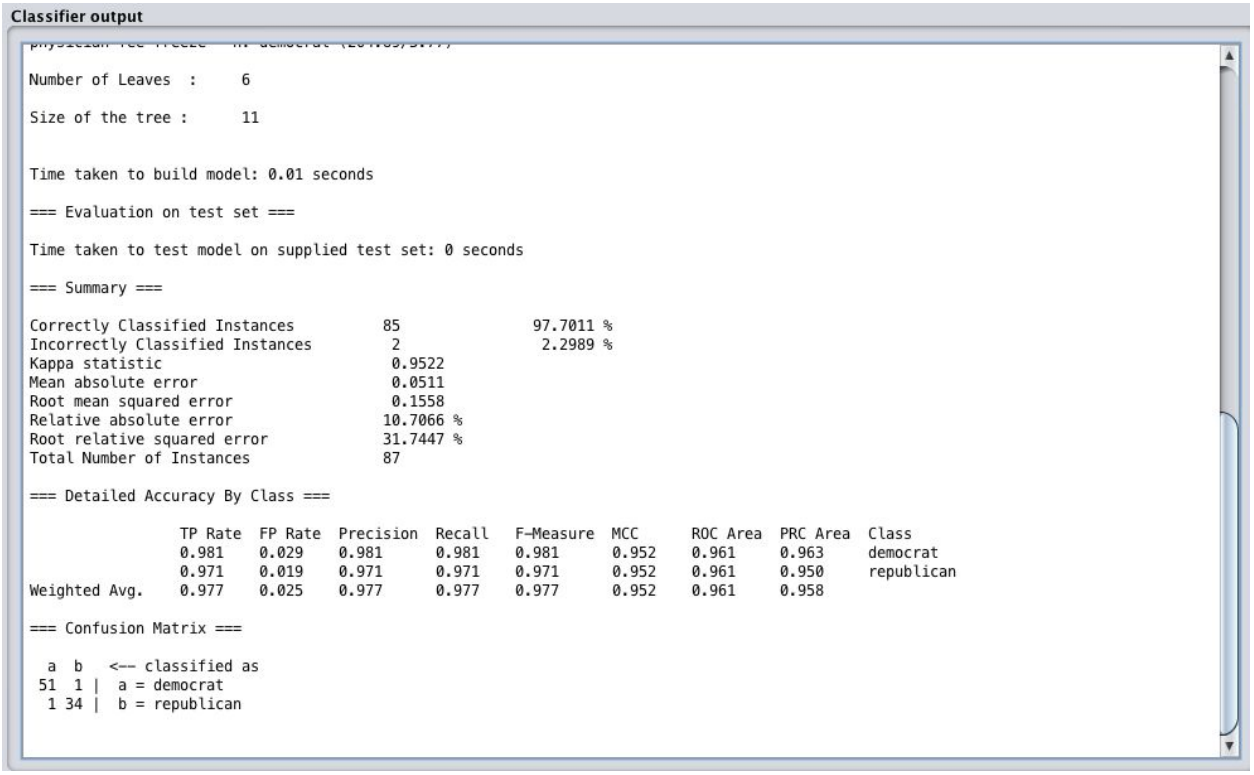
(b) For i = 1, 2, . . . , 5, each time use Di as test data and the rest as training data to learn a decision tree and measure its accuracy pi.

D1 as test data

Training results:



Test results:

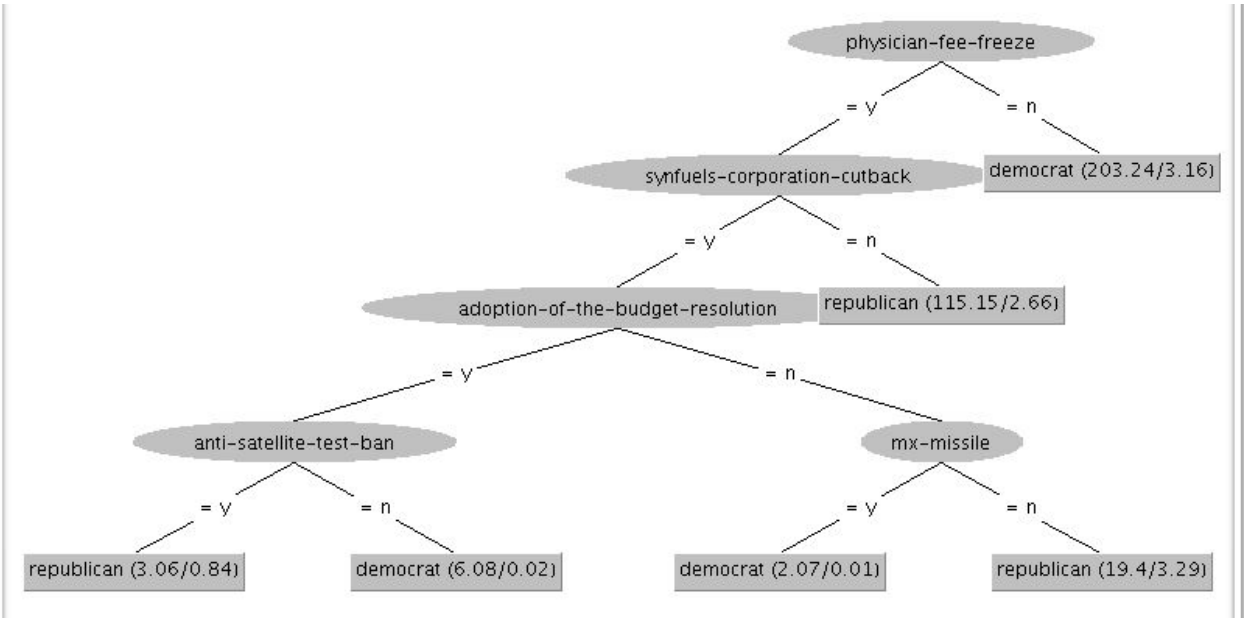


Classifier output

```
Number of Leaves  :     6

Size of the tree :      11


Time taken to build model: 0.01 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances          85               97.7011 %
Incorrectly Classified Instances         2                2.2989 %
Kappa statistic                          0.9522
Mean absolute error                      0.0511
Root mean squared error                  0.1558
Relative absolute error                 10.7066 %
Root relative squared error             31.7447 %
Total Number of Instances               87

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.981    0.029    0.981      0.981    0.981      0.952    0.961     0.963     democrat
                 0.971    0.019    0.971      0.971    0.971      0.952    0.961     0.950     republican
Weighted Avg.    0.977    0.025    0.977      0.977    0.977      0.952    0.961     0.958

=== Confusion Matrix ===

  a  b   <-- classified as
 51  1 |  a = democrat
  1 34 |  b = republican
```

D2 as test data

Training results:



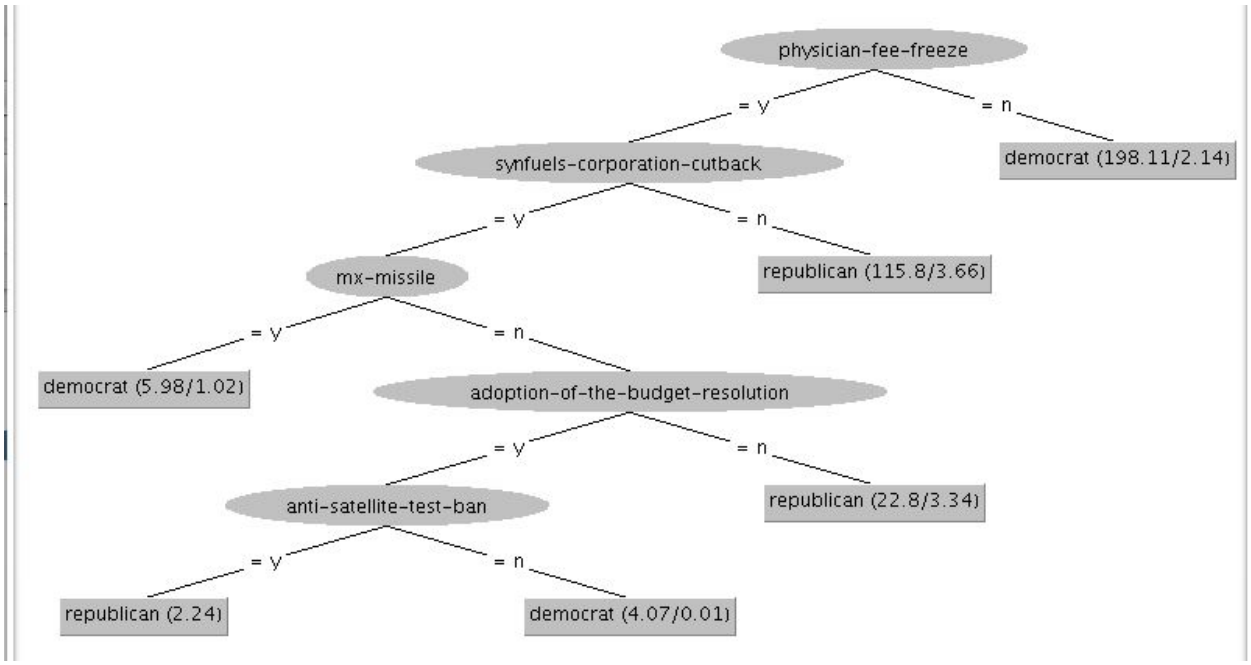
Test results:

Classifier output

```
Number of Leaves  :     6

Size of the tree :      11


Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances          84               97.6744 %
Incorrectly Classified Instances         2                2.3256 %
Kappa statistic                          0.9514
Mean absolute error                      0.0488
Root mean squared error                  0.1385
Relative absolute error                 10.2506 %
Root relative squared error             28.3135 %
Total Number of Instances               86

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.981    0.029    0.981      0.981    0.981      0.951    0.989     0.992     democrat
                 0.971    0.019    0.971      0.971    0.971      0.951    0.989     0.966     republican
Weighted Avg.    0.977    0.025    0.977      0.977    0.977      0.951    0.989     0.982

=== Confusion Matrix ===

  a  b   <-- classified as
 51  1 |  a = democrat
  1 33 |  b = republican
```

D3 as test data

Training results:



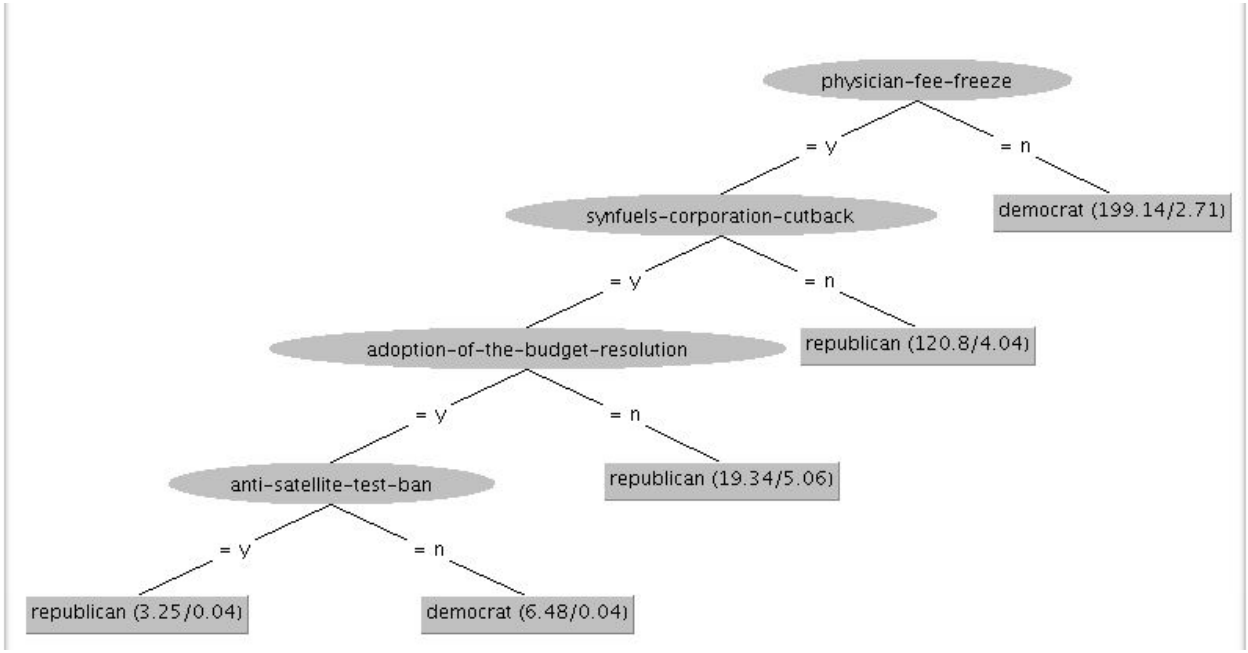
Test results:

Classifier output

```
physician-fee-freeze = n: democrat (130.11/2.11)

Number of Leaves  :     6

Size of the tree :      11


Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances          84               97.6744 %
Incorrectly Classified Instances         2                2.3256 %
Kappa statistic                          0.9488
Mean absolute error                      0.0399
Root mean squared error                  0.1321
Relative absolute error                  8.4895 %
Root relative squared error             27.456  %
Total Number of Instances               86

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 1.000    0.065    0.965      1.000    0.982      0.950    0.984     0.982     democrat
                 0.935    0.000    1.000      0.935    0.967      0.950    0.984     0.980     republican
Weighted Avg.    0.977    0.041    0.978      0.977    0.977      0.950    0.984     0.981

=== Confusion Matrix ===

  a  b   <-- classified as
 55  0 |  a = democrat
  2 29 |  b = republican
```

D4 as test data

Training results:



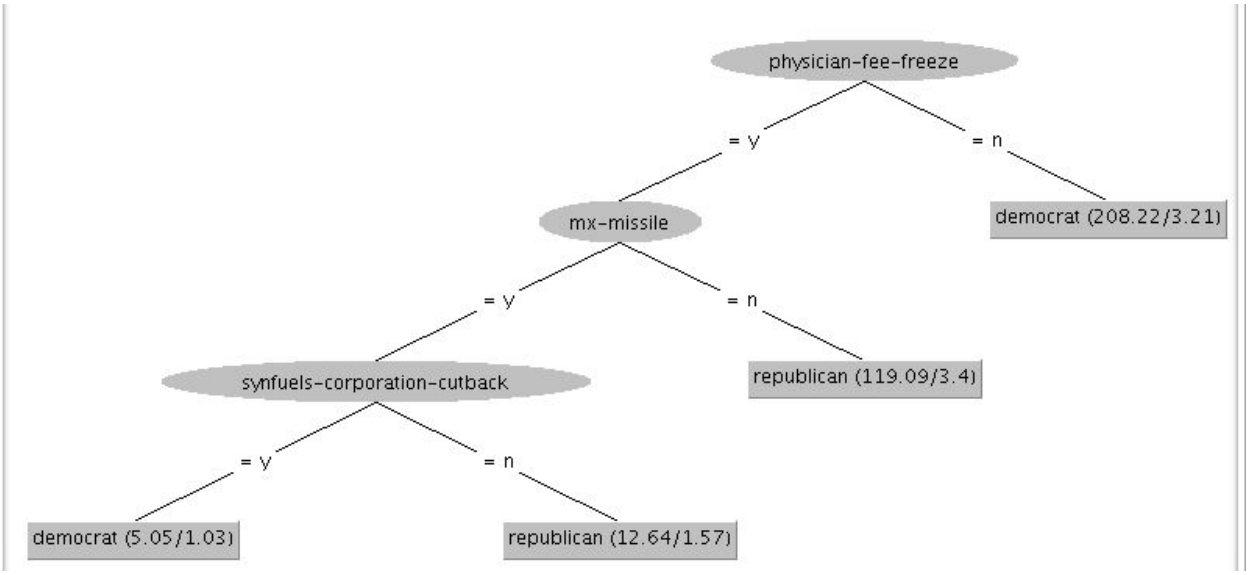
Test results:

Classifier output

```
Number of Leaves  :     5

Size of the tree :      9


Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances          84               97.6744 %
Incorrectly Classified Instances         2                2.3256 %
Kappa statistic                          0.9496
Mean absolute error                      0.0633
Root mean squared error                  0.1625
Relative absolute error                 13.4581 %
Root relative squared error             33.7733 %
Total Number of Instances               86

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.982    0.032    0.982      0.982    0.982      0.950    0.982     0.980     democrat
                 0.968    0.018    0.968      0.968    0.968      0.950    0.982     0.974     republican
Weighted Avg.    0.977    0.027    0.977      0.977    0.977      0.950    0.982     0.978

=== Confusion Matrix ===

  a  b   <-- classified as
 54  1 |  a = democrat
  1 30 |  b = republican
```

D5 as test data

Training results:



Test results:

Classifier output

```
Number of Leaves  :     4

Size of the tree :      7


Time taken to build model: 0 seconds

=== Evaluation on test set ===

Time taken to test model on supplied test set: 0 seconds

=== Summary ===

Correctly Classified Instances          83               92.2222 %
Incorrectly Classified Instances         7                7.7778 %
Kappa statistic                          0.8426
Mean absolute error                      0.1115
Root mean squared error                  0.2728
Relative absolute error                 23.2972 %
Root relative squared error             55.3363 %
Total Number of Instances               90

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.887    0.027    0.979      0.887    0.931      0.848    0.931     0.952     democrat
                 0.973    0.113    0.857      0.973    0.911      0.848    0.931     0.836     republican
Weighted Avg.    0.922    0.062    0.929      0.922    0.923      0.848    0.931     0.904

=== Confusion Matrix ===

  a  b   <-- classified as
 47  6 |  a = democrat
  1 36 |  b = republican
```

(c) Visualize the five trees constructed. Do the five trees differ with each other and with the tree constructed using all the data?

Trees visualized above. Yes, the five trees differ slightly with each other and with the tree constructed using all the data. Overall, “physician-fee-freeze” is still the best indicator of what party the voter is in. “synfuels-corporation-cutback” seems to be consistently the second best indicator of what party the voter is in (with one exception with my last 5-fold split). The “mx-missile” is also consistently the third best indicator with one exception. My first and three data split is almost identical to the tree that was constructed using all the data.

(d) Compute the average of p1, . . . , p5 as the 5-fold cross-validation estimation of the accuracy of the Decision Tree classifier. Report 95% confidence interval.

Accuracy for D1 = 97.7011 %
Accuracy for D2 = 97.6744 %
Accuracy for D3 = 97.6744 %
Accuracy for D4 = 97.6744 %
Accuracy for D5 = 92.2222 %

(97.7011 + 97.6744 + 97.6744 + 97.6744 + 92.2222) / 5 = 96.58%

Total number of samples = 434

D1 Confidence Interval:
n = 434 - 108 = 326 trials
error_s(h) = r / n = 97.7011 %
standard_deviation = sqrt((error_s(h) * 1-error_s(h)) / n) = 0.83 %
[97.7011 % - 0.83 %, 97.7011 % + 0.83 %] ⇒ [96.87 %, 98.53 %]

D2 Confidence Interval:
n = 434 - 107 = 327 trials
error_s(h) = r / n = 97.6744 %
standard_deviation = sqrt((error_s(h) * 1-error_s(h)) / n) = 0.83 %
[97.6744 % - 0.83 %, 97.6744 % + 0.83 %] ⇒ [96.68 %, 98.50 %]

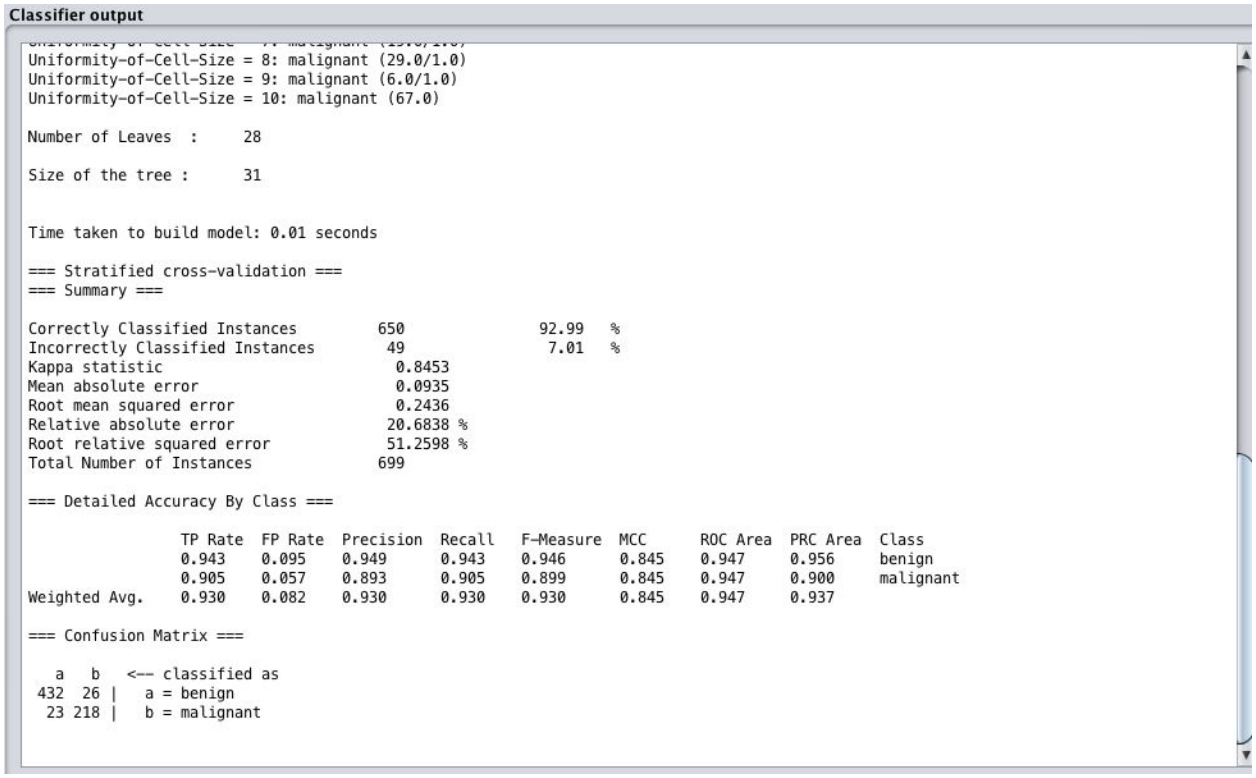
D3 Confidence Interval:
n = 434 - 107 = 327 trials
error_s(h) = r / n = 97.6744 %
standard_deviation = sqrt((error_s(h) * 1-error_s(h)) / n) = 0.83 %
[97.6744 % - 0.83 %, 97.6744 % + 0.83 %] ⇒ [96.68 %, 98.50 %]

D4 Confidence Interval:
n = 434 - 107 = 327 trials
error_s(h) = r / n = 97.6744 %
standard_deviation = sqrt((error_s(h) * 1-error_s(h)) / n) = 0.83 %
[97.6744 % - 0.83 %, 97.6744 % + 0.83 %] ⇒ [96.68 %, 98.50 %]

D5 Confidence Interval:
n = 434 - 111 = 323 trials
error_s(h) = r / n = 92.2222 %
standard_deviation = sqrt((error_s(h) * 1-error_s(h)) / n) = 1.49 %
[92.2222 % - 1.49 %, 92.2222 % + 1.49 %] ⇒ [90.73 %, 93.71 %]

3. Report the accuracy of the Decision Tree classifier on the Breast Cancer Wisconsin (Original) Data Set using 5-fold cross validation. Report 95% confidence interval.

Accuracy: 92.99 %


Classifier output

```
Uniformity-of-Cell-Size = 8: malignant (29.0/1.0)
Uniformity-of-Cell-Size = 9: malignant (6.0/1.0)
Uniformity-of-Cell-Size = 10: malignant (67.0)

Number of Leaves  :     28

Size of the tree :      31


Time taken to build model: 0.01 seconds

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

Correctly Classified Instances         650               92.99   %
Incorrectly Classified Instances        49                7.01   %
Kappa statistic                          0.8453
Mean absolute error                      0.0935
Root mean squared error                  0.2436
Relative absolute error                 20.6838 %
Root relative squared error             51.2598 %
Total Number of Instances              699

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.943    0.095    0.949      0.943    0.946      0.845    0.947     0.956     benign
                 0.905    0.057    0.893      0.905    0.899      0.845    0.947     0.900     malignant
Weighted Avg.    0.930    0.082    0.930      0.930    0.930      0.845    0.947     0.937

=== Confusion Matrix ===

   a   b   <-- classified as
 432  26 |   a = benign
  23 218 |   b = malignant
```

Confidence Interval:
n = 688 instances
error_s(h) = r / n = 92.99 %
standard_deviation = sqrt((error_s(h) * 1-error_s(h)) / n) = 0.97 %
[92.99 % - 0.97 %, 92.99 % + 0.97 %] ⇒ [92.02 % , 93.96 %]

4. Both datasets have missing values. C4.5 applies method of fractional instances during training and testing. A different method for dealing with missing values is to preprocess the data and fill them in with the most common values or average values (Weka has a missing values filter to do this). Now preprocess data by filling in missing values, and then compute the accuracy of the Decision Tree classifier on the two data sets using 5-fold cross validation. How does filling missing values affect the performance of the classifiers?

Filling in missing values through Weka, decreased the accuracy for the breast cancer data and slightly improved the voting data tree accuracy by about 0.7%.

Voting 5-fold results after preprocessing missing values:

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

Correctly Classified Instances         419               96.3218 %
Incorrectly Classified Instances        16                3.6782 %
Kappa statistic                          0.9224
Mean absolute error                      0.061
Root mean squared error                  0.1885
Relative absolute error                 12.8693 %
Root relative squared error             38.7223 %
Total Number of Instances              435

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.970    0.048    0.970      0.970    0.970      0.922    0.953     0.948     democrat
                 0.952    0.030    0.952      0.952    0.952      0.922    0.953     0.922     republican
Weighted Avg.    0.963    0.041    0.963      0.963    0.963      0.922    0.953     0.938

=== Confusion Matrix ===

   a   b   <-- classified as
 259   8 |   a = democrat
   8 160 |   b = republican
```

Breast Cancer 5-fold results after preprocessing missing values:

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

Correctly Classified Instances         642               91.8455 %
Incorrectly Classified Instances        57                8.1545 %
Kappa statistic                          0.819
Mean absolute error                      0.0967
Root mean squared error                  0.2559
Relative absolute error                 21.3866 %
Root relative squared error             53.8372 %
Total Number of Instances              699

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0.941    0.124    0.935      0.941    0.938      0.819    0.947     0.957     benign
                 0.876    0.059    0.887      0.876    0.881      0.819    0.947     0.896     malignant
Weighted Avg.    0.918    0.102    0.918      0.918    0.918      0.819    0.947     0.936

=== Confusion Matrix ===

   a   b   <-- classified as
 431  27 |   a = benign
  30 211 |   b = malignant
```