**Homework 4**

**Name: Megha Godwal**

**UID: U01695574**

### Classification

The file [vote.arff](http://csis.pace.edu/~benjamin/teaching/cs619/webfiles/vote.arff) contains the voting records of members of the U.S. House of Representatives. Each example consists of the attributes indicating the yes/no/absent vote for each bill considered by congress and the political party of this representative. You will use this data to learn a decision tree that predicts the political party of the representative based on his/her vote.

1. Use the voting data to train a decision tree to predict the political party (Democrat or Republican) based on the voting record. Set the option -U of J48 so that no pruning is performed. (Otherwise, the trees are pruned by default.) Randomly select 25% of the members of congress for training, and the rest for testing. In order to do this, you have to split the ARFF file vote.arff into several ARFF files. Rerun this experiment several times and you will notice the impact of different random splits of the data into training and testing sets. Report the sizes and accuracies of the trees from 5 runs.

**Ans:**

By splitting and randomly shuffling the same file along with 5 different runs on the same data set with 25% of training data below are the stats:

**Run 1:**

Size of Tree: 37

Number of Leaves: 19

Correctly Classified Instances 303 Accuracy 92.9448 %

Incorrectly Classified Instances 23 Inaccuracy 7.0552 %

**Run 2:**

Size of Tree: 37

Number of Leaves: 19

Correctly Classified Instances 306 Accuracy 93.865 %

Incorrectly Classified Instances 20 Inaccuracy 6.1325 %

**Run 3:**

Size of Tree: 37

Number of Leaves: 19

Correctly Classified Instances 305 Accuracy 93.5583 %

Incorrectly Classified Instances 22 Inaccuracy 6.4417 %

**Run 4:**

Size of Tree: 37

Number of Leaves: 19

Correctly Classified Instances 310 Accuracy 95.092 %

Incorrectly Classified Instances 16 Inaccuracy 4.908 %

**Run 5:**

Size of Tree: 37

Number of Leaves: 19

Correctly Classified Instances 313 Accuracy 96.0123 %

Incorrectly Classified Instances 13 Inaccuracy 3.9877 %

1. Measure the impact of training set size on the accuracy and the size of the learned tree (no pruning, so use option -U and use 60% of the data for testing.) Consider training set sizes in the range 0-40% (include at least the values .02, .1, .2, .3 and .4 for the training fractions.) Because of the high variance due to random splits, repeat the experiment with ten different random seeds for each training set size, then report the mean, maximum and minimum accuracies at each training set size. Turn in two plots, showing how accuracy varies with training set size.

**Ans:**

|  |  |
| --- | --- |
| **Training Set Size** | **Accuracy %** |
| 0.01 | 61.2529 |
| 0.02 | 80.7512 |
| 0.03 | 95.7346 |
| 0.1 | 95.3964 |
| 0.15 | 94.3243 |
| 0.2 | 94.2529 |
| 0.25 | 93.865 |
| 0.3 | 93.75 |
| 0.35 | 94.3463 |
| 0.4 | 94.636 |

Minimum Accuracy = 95.7346

Maximum Accuracy = 61.2529

Mean Accuracy = 898.3096/10 = 89.86