CS 6375

ASSIGNMENT ___1__________

Names of students in your group:

Wang, Xunde *xxw150130*
Zhou, Xiangru *xxz141830*

Number of free late days used: _____0_______________

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

```
CS6375_002_HW1\
-----Part I\
        -----Assignment 1-Part I.pdf
-----Part II\
        -----readme.pdf
        -----Report.pdf
        -----post\
                -----Data.java
        ----- pre\
                -----Data.java
                -----train-win.dat
                -----train2-win.dat
                -----test-win.dat
                -----test2-win.dat
```

Assumptions:

For creating the decision tree, new nodes will be added into tree recursively. The assumption for labeling leaf nodes is that only nodes with entropy above certain level α or entropy gain is above certain level β will be labeled as leaf node (no more splitting), which is pre-pruning. The values of α and β could be modified for obtaining optimized results for different dataset. To be noted that the values of α and β will be set to very low values so that pre-pruning will not take any effects for default tree creation.

Accomplished:

We first parsed the datasets into integer arrays. Then we built a decision tree using all the training dataset. Compared the training and testing datasets with the tree to get the accuracy of training and testing. Post-pruned the tree. Pre-pruned the tree.

Learned:

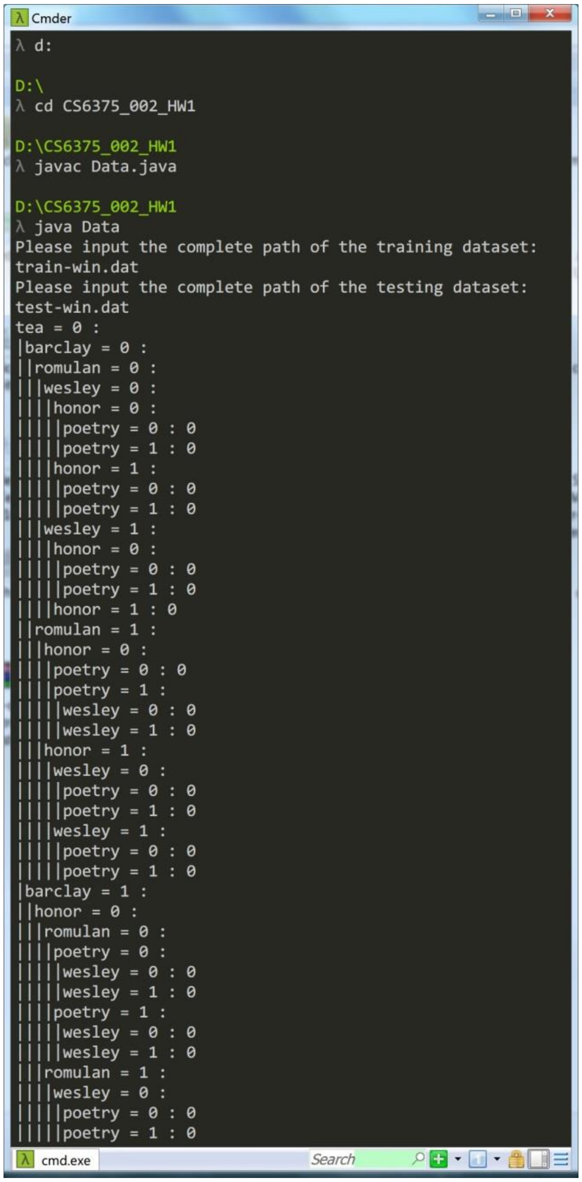
How to create a binary decision tree. How to predict the class label of dataset. Post-prune and pre-prune methods. For different datasets, we might use different prune method to get a better result.

Notes:

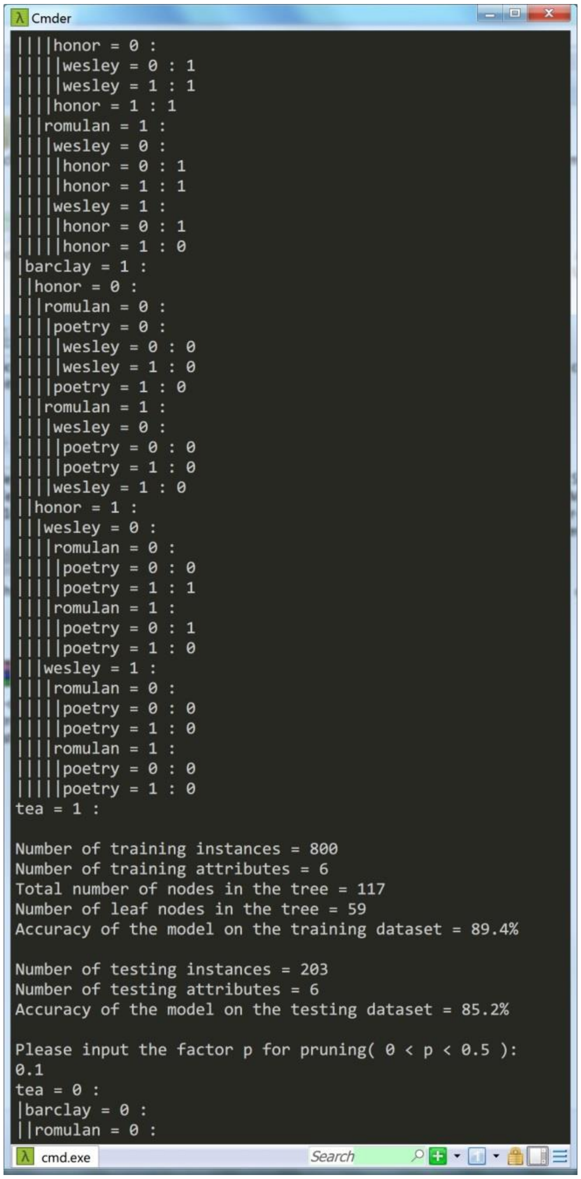
Only post-prune is required for this homework, however, we did not get very good result for post-prune. So we want to try other methods to get a better outcome.

Post-prune

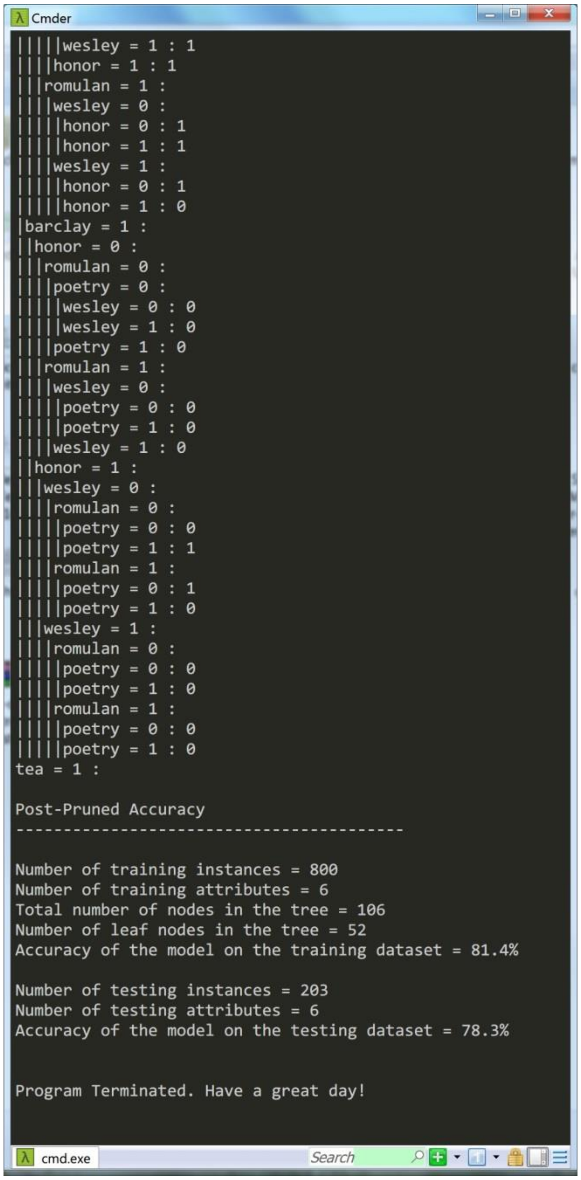
The assumption for post-pruning is that only less than 50% of total nodes are allowed for deletion and leaf node only. Post-pruning will not yield good results for both dataset. The accuracy decreased for both the training and testing datasets, also for both dataset 1 and dataset 2.



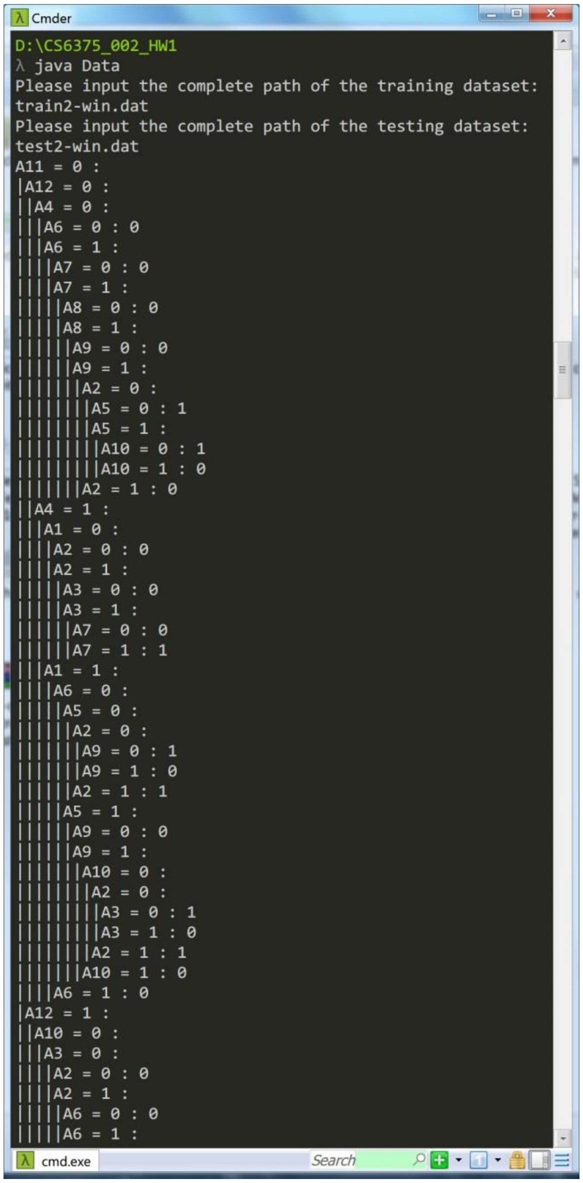
```
λ d:

D:\
λ cd CS6375_002_HW1

D:\CS6375_002_HW1
λ javac Data.java

D:\CS6375_002_HW1
λ java Data
Please input the complete path of the training dataset:
train-win.dat
Please input the complete path of the testing dataset:
test-win.dat
tea = 0 :
|barclay = 0 :
||romulan = 0 :
|||wesley = 0 :
||||honor = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
||||honor = 1 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
|||wesley = 1 :
||||honor = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
||||honor = 1 : 0
||romulan = 1 :
|||honor = 0 :
||||poetry = 0 : 0
||||poetry = 1 :
|||||wesley = 0 : 0
|||||wesley = 1 : 0
|||honor = 1 :
||||wesley = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
||||wesley = 1 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
|barclay = 1 :
||honor = 0 :
|||romulan = 0 :
||||poetry = 0 :
|||||wesley = 0 : 0
|||||wesley = 1 : 0
||||poetry = 1 :
|||||wesley = 0 : 0
|||||wesley = 1 : 0
|||romulan = 1 :
||||wesley = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
```



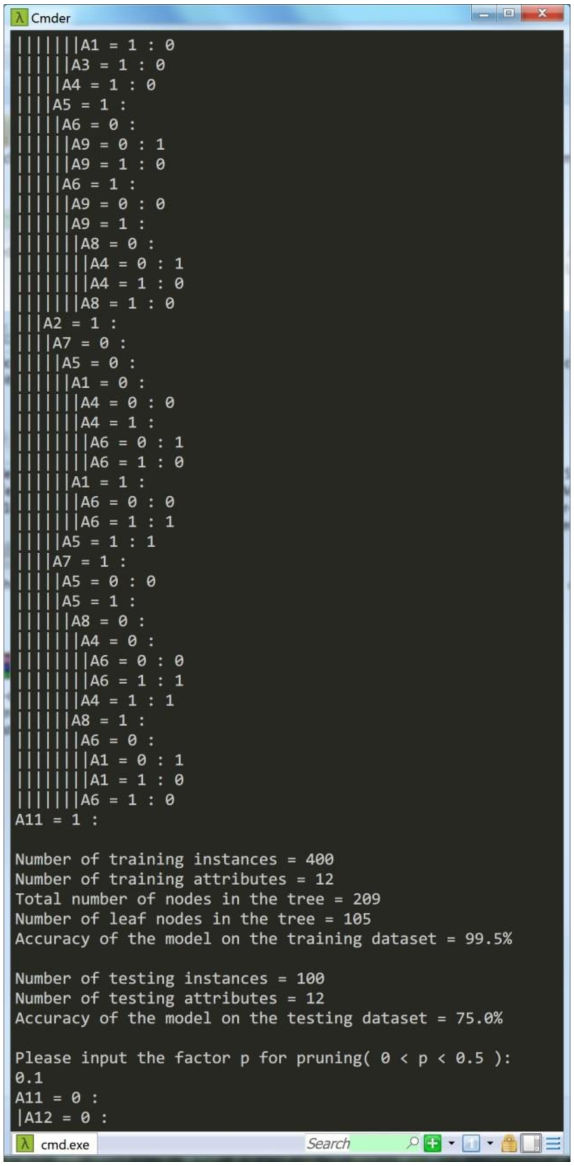
```
||||honor = 0 :
|||||wesley = 0 : 1
|||||wesley = 1 : 1
||||honor = 1 : 1
|||romulan = 1 :
||||wesley = 0 :
|||||honor = 0 : 1
|||||honor = 1 : 1
||||wesley = 1 :
|||||honor = 0 : 1
|||||honor = 1 : 0
|barclay = 1 :
||honor = 0 :
|||romulan = 0 :
||||poetry = 0 :
|||||wesley = 0 : 0
|||||wesley = 1 : 0
||||poetry = 1 : 0
|||romulan = 1 :
||||wesley = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
||||wesley = 1 : 0
||honor = 1 :
|||wesley = 0 :
||||romulan = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 1
||||romulan = 1 :
|||||poetry = 0 : 1
|||||poetry = 1 : 0
|||wesley = 1 :
||||romulan = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
||||romulan = 1 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
tea = 1 :

Number of training instances = 800
Number of training attributes = 6
Total number of nodes in the tree = 117
Number of leaf nodes in the tree = 59
Accuracy of the model on the training dataset = 89.4%

Number of testing instances = 203
Number of testing attributes = 6
Accuracy of the model on the testing dataset = 85.2%

Please input the factor p for pruning( 0 < p < 0.5 ):
0.1
tea = 0 :
|barclay = 0 :
||romulan = 0 :
```



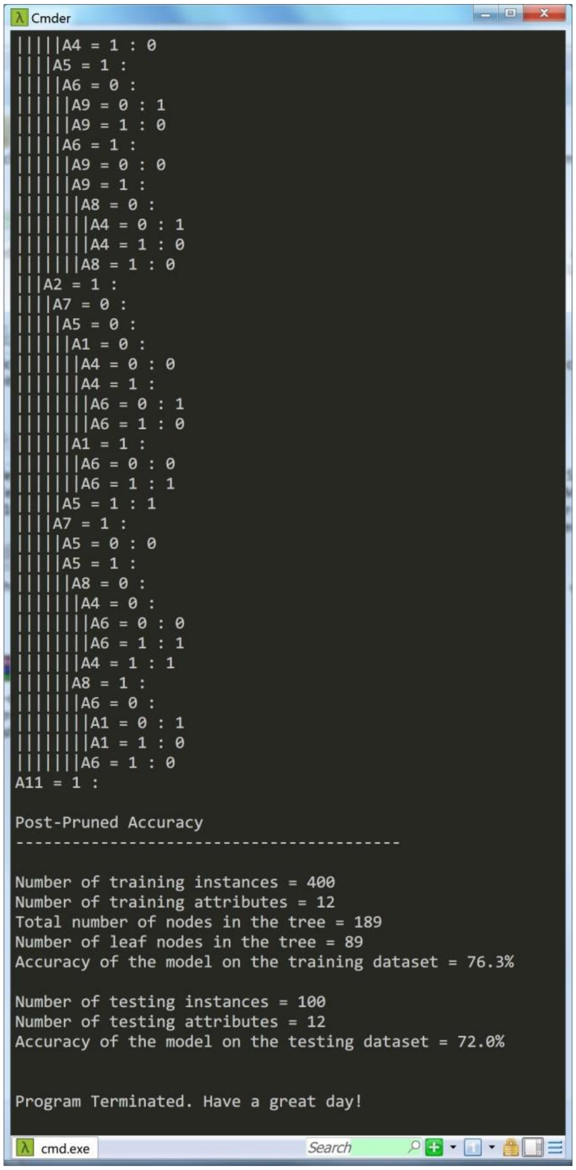
```
|||||wesley = 1 : 1
||||honor = 1 : 1
|||romulan = 1 :
||||wesley = 0 :
|||||honor = 0 : 1
|||||honor = 1 : 1
||||wesley = 1 :
|||||honor = 0 : 1
|||||honor = 1 : 0
|barclay = 1 :
||honor = 0 :
|||romulan = 0 :
||||poetry = 0 :
|||||wesley = 0 : 0
|||||wesley = 1 : 0
||||poetry = 1 : 0
|||romulan = 1 :
||||wesley = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
||||wesley = 1 : 0
||honor = 1 :
|||wesley = 0 :
||||romulan = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 1
||||romulan = 1 :
|||||poetry = 0 : 1
|||||poetry = 1 : 0
|||wesley = 1 :
||||romulan = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
||||romulan = 1 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
tea = 1 :

Post-Pruned Accuracy
-----------------------------------------

Number of training instances = 800
Number of training attributes = 6
Total number of nodes in the tree = 106
Number of leaf nodes in the tree = 52
Accuracy of the model on the training dataset = 81.4%

Number of testing instances = 203
Number of testing attributes = 6
Accuracy of the model on the testing dataset = 78.3%


Program Terminated. Have a great day!
```



```
D:\CS6375_002_HW1
λ java Data
Please input the complete path of the training dataset:
train2-win.dat
Please input the complete path of the testing dataset:
test2-win.dat
A11 = 0 :
|A12 = 0 :
||A4 = 0 :
|||A6 = 0 : 0
|||A6 = 1 :
||||A7 = 0 : 0
||||A7 = 1 :
|||||A8 = 0 : 0
|||||A8 = 1 :
||||||A9 = 0 : 0
||||||A9 = 1 :
|||||||A2 = 0 :
||||||||A5 = 0 : 1
||||||||A5 = 1 :
|||||||||A10 = 0 : 1
|||||||||A10 = 1 : 0
|||||||A2 = 1 : 0
||A4 = 1 :
|||A1 = 0 :
||||A2 = 0 : 0
||||A2 = 1 :
|||||A3 = 0 : 0
|||||A3 = 1 :
||||||A7 = 0 : 0
||||||A7 = 1 : 1
|||A1 = 1 :
||||A6 = 0 :
|||||A5 = 0 :
||||||A2 = 0 :
|||||||A9 = 0 : 1
|||||||A9 = 1 : 0
||||||A2 = 1 : 1
|||||A5 = 1 :
||||||A9 = 0 : 0
||||||A9 = 1 :
|||||||A10 = 0 :
||||||||A2 = 0 :
|||||||||A3 = 0 : 1
|||||||||A3 = 1 : 0
||||||||A2 = 1 : 1
|||||||A10 = 1 : 0
||||A6 = 1 : 0
|A12 = 1 :
||A10 = 0 :
|||A3 = 0 :
||||A2 = 0 : 0
||||A2 = 1 :
|||||A6 = 0 : 0
|||||A6 = 1 :
```



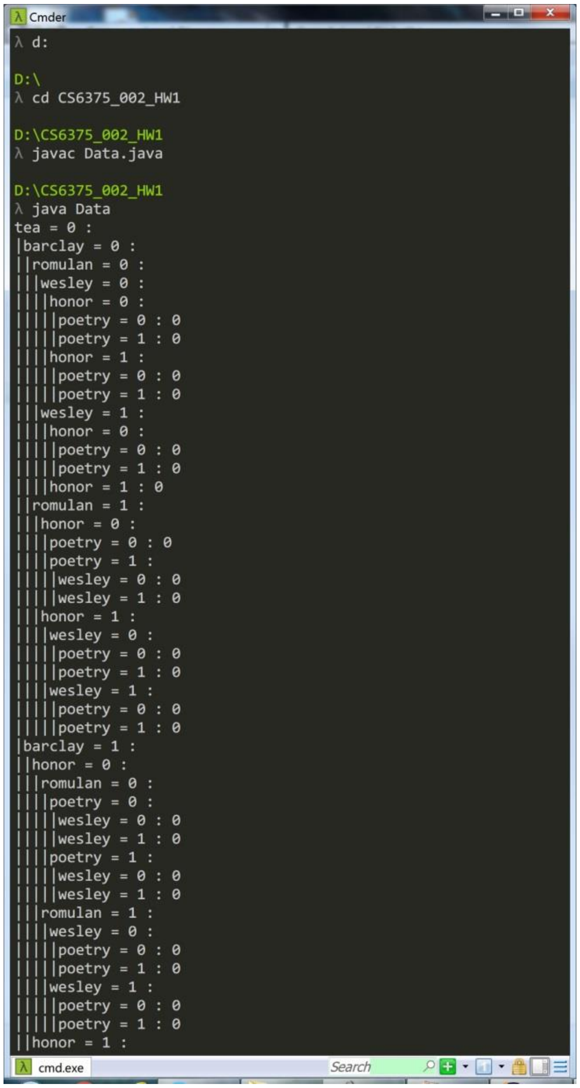
```
|||||||A1 = 1 : 0
||||||A3 = 1 : 0
|||||A4 = 1 : 0
||||A5 = 1 :
|||||A6 = 0 :
||||||A9 = 0 : 1
||||||A9 = 1 : 0
|||||A6 = 1 :
||||||A9 = 0 : 0
||||||A9 = 1 :
|||||||A8 = 0 :
||||||||A4 = 0 : 1
||||||||A4 = 1 : 0
|||||||A8 = 1 : 0
|||A2 = 1 :
||||A7 = 0 :
|||||A5 = 0 :
||||||A1 = 0 :
|||||||A4 = 0 : 0
|||||||A4 = 1 :
||||||||A6 = 0 : 1
||||||||A6 = 1 : 0
||||||A1 = 1 :
|||||||A6 = 0 : 0
|||||||A6 = 1 : 1
|||||A5 = 1 : 1
||||A7 = 1 :
|||||A5 = 0 : 0
|||||A5 = 1 :
||||||A8 = 0 :
|||||||A4 = 0 :
||||||||A6 = 0 : 0
||||||||A6 = 1 : 1
|||||||A4 = 1 : 1
||||||A8 = 1 :
|||||||A6 = 0 :
||||||||A1 = 0 : 1
||||||||A1 = 1 : 0
|||||||A6 = 1 : 0
A11 = 1 :

Number of training instances = 400
Number of training attributes = 12
Total number of nodes in the tree = 209
Number of leaf nodes in the tree = 105
Accuracy of the model on the training dataset = 99.5%

Number of testing instances = 100
Number of testing attributes = 12
Accuracy of the model on the testing dataset = 75.0%

Please input the factor p for pruning( 0 < p < 0.5 ):
0.1
A11 = 0 :
|A12 = 0 :
```



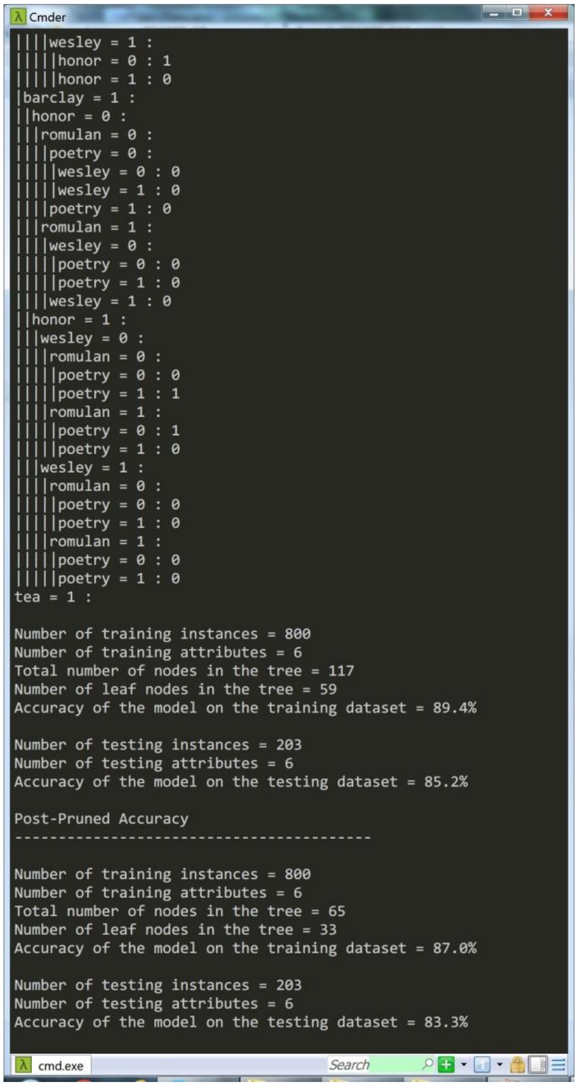
```
|||||A4 = 1 : 0
||||A5 = 1 :
|||||A6 = 0 :
||||||A9 = 0 : 1
||||||A9 = 1 : 0
|||||A6 = 1 :
||||||A9 = 0 : 0
||||||A9 = 1 :
|||||||A8 = 0 :
||||||||A4 = 0 : 1
||||||||A4 = 1 : 0
|||||||A8 = 1 : 0
|||A2 = 1 :
||||A7 = 0 :
|||||A5 = 0 :
||||||A1 = 0 :
|||||||A4 = 0 : 0
|||||||A4 = 1 :
||||||||A6 = 0 : 1
||||||||A6 = 1 : 0
||||||A1 = 1 :
|||||||A6 = 0 : 0
|||||||A6 = 1 : 1
|||||A5 = 1 : 1
||||A7 = 1 :
|||||A5 = 0 : 0
|||||A5 = 1 :
||||||A8 = 0 :
|||||||A4 = 0 :
||||||||A6 = 0 : 0
||||||||A6 = 1 : 1
|||||||A4 = 1 : 1
||||||A8 = 1 :
|||||||A6 = 0 :
||||||||A1 = 0 : 1
||||||||A1 = 1 : 0
|||||||A6 = 1 : 0
A11 = 1 :

Post-Pruned Accuracy
-----------------------------------------

Number of training instances = 400
Number of training attributes = 12
Total number of nodes in the tree = 189
Number of leaf nodes in the tree = 89
Accuracy of the model on the training dataset = 76.3%

Number of testing instances = 100
Number of testing attributes = 12
Accuracy of the model on the testing dataset = 72.0%


Program Terminated. Have a great day!
```

Pre-prune

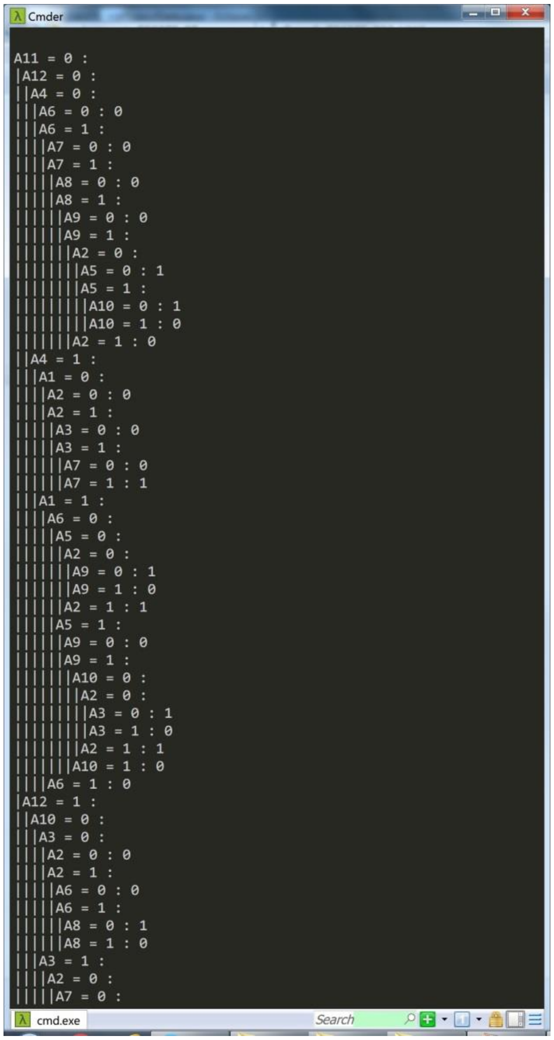
Pre-pruning will greatly increase the accuracy of the testing of dataset 2. This is because the training accuracy for dataset 1 is around 89% but 99.5% for dataset 2. This means that there are many contradictory samples in the dataset 1 so it is very hard to get a perfect tree for both training and testing. However, the accuracy of training for dataset 2 means that we can easily tradeoff accuracy of training for testing. As the result showed that the accuracy of training decreased from 99.5% to 94.5%, however, the accuracy of testing increased from 75% to 87% for dataset 2.



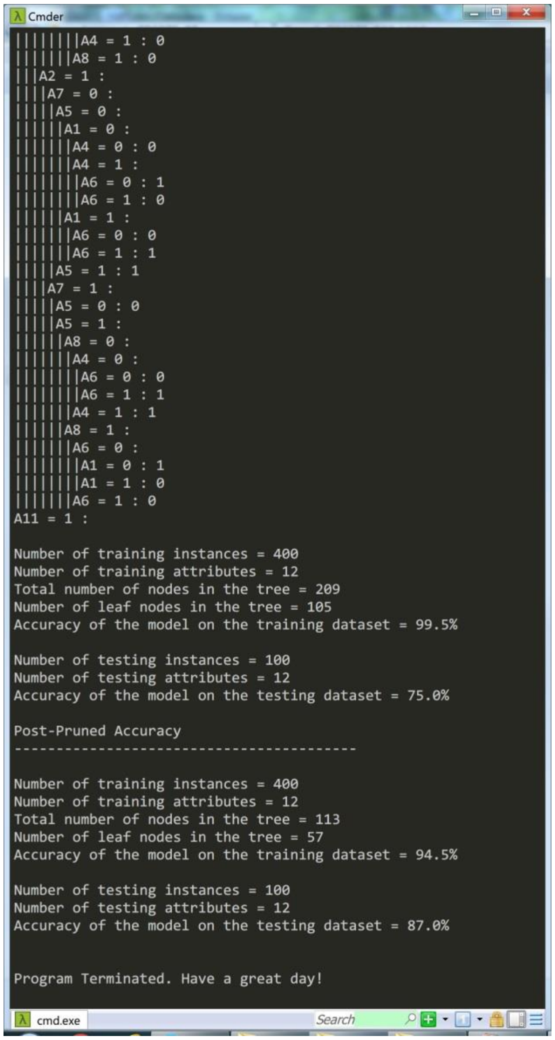
```
λ d:

D:\
λ cd CS6375_002_HW1

D:\CS6375_002_HW1
λ javac Data.java

D:\CS6375_002_HW1
λ java Data
tea = 0 :
|barclay = 0 :
||romulan = 0 :
|||wesley = 0 :
||||honor = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
||||honor = 1 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
|||wesley = 1 :
||||honor = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
||||honor = 1 : 0
||romulan = 1 :
|||honor = 0 :
||||poetry = 0 : 0
||||poetry = 1 :
|||||wesley = 0 : 0
|||||wesley = 1 : 0
|||honor = 1 :
||||wesley = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
||||wesley = 1 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
|barclay = 1 :
||honor = 0 :
|||romulan = 0 :
||||poetry = 0 :
|||||wesley = 0 : 0
|||||wesley = 1 : 0
||||poetry = 1 :
|||||wesley = 0 : 0
|||||wesley = 1 : 0
|||romulan = 1 :
||||wesley = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
||||wesley = 1 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
||honor = 1 :
```



```
||||wesley = 1 :
|||||honor = 0 : 1
|||||honor = 1 : 0
|barclay = 1 :
||honor = 0 :
|||romulan = 0 :
||||poetry = 0 :
|||||wesley = 0 : 0
|||||wesley = 1 : 0
||||poetry = 1 : 0
|||romulan = 1 :
||||wesley = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
||||wesley = 1 : 0
||honor = 1 :
|||wesley = 0 :
||||romulan = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 1
||||romulan = 1 :
|||||poetry = 0 : 1
|||||poetry = 1 : 0
|||wesley = 1 :
||||romulan = 0 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
||||romulan = 1 :
|||||poetry = 0 : 0
|||||poetry = 1 : 0
tea = 1 :

Number of training instances = 800
Number of training attributes = 6
Total number of nodes in the tree = 117
Number of leaf nodes in the tree = 59
Accuracy of the model on the training dataset = 89.4%

Number of testing instances = 203
Number of testing attributes = 6
Accuracy of the model on the testing dataset = 85.2%

Post-Pruned Accuracy
----------------------------------------

Number of training instances = 800
Number of training attributes = 6
Total number of nodes in the tree = 65
Number of leaf nodes in the tree = 33
Accuracy of the model on the training dataset = 87.0%

Number of testing instances = 203
Number of testing attributes = 6
Accuracy of the model on the testing dataset = 83.3%
```



```
A11 = 0 :
|A12 = 0 :
||A4 = 0 :
|||A6 = 0 : 0
|||A6 = 1 :
||||A7 = 0 : 0
||||A7 = 1 :
|||||A8 = 0 : 0
|||||A8 = 1 :
||||||A9 = 0 : 0
||||||A9 = 1 :
|||||||A2 = 0 :
||||||||A5 = 0 : 1
||||||||A5 = 1 :
|||||||||A10 = 0 : 1
|||||||||A10 = 1 : 0
|||||||A2 = 1 : 0
||A4 = 1 :
|||A1 = 0 :
||||A2 = 0 : 0
||||A2 = 1 :
|||||A3 = 0 : 0
|||||A3 = 1 :
||||||A7 = 0 : 0
||||||A7 = 1 : 1
|||A1 = 1 :
||||A6 = 0 :
|||||A5 = 0 :
||||||A2 = 0 :
|||||||A9 = 0 : 1
|||||||A9 = 1 : 0
||||||A2 = 1 : 1
|||||A5 = 1 :
||||||A9 = 0 : 0
||||||A9 = 1 :
|||||||A10 = 0 :
||||||||A2 = 0 :
|||||||||A3 = 0 : 1
|||||||||A3 = 1 : 0
||||||||A2 = 1 : 1
|||||||A10 = 1 : 0
||||A6 = 1 : 0
|A12 = 1 :
||A10 = 0 :
|||A3 = 0 :
||||A2 = 0 : 0
||||A2 = 1 :
|||||A6 = 0 : 0
|||||A6 = 1 :
||||||A8 = 0 : 1
||||||A8 = 1 : 0
|||A3 = 1 :
||||A2 = 0 :
|||||A7 = 0 :
```



```
||||||||A4 = 1 : 0
|||||||A8 = 1 : 0
|||A2 = 1 :
||||A7 = 0 :
|||||A5 = 0 :
||||||A1 = 0 :
|||||||A4 = 0 : 0
|||||||A4 = 1 :
||||||||A6 = 0 : 1
||||||||A6 = 1 : 0
||||||A1 = 1 :
|||||||A6 = 0 : 0
|||||||A6 = 1 : 1
|||||A5 = 1 : 1
||||A7 = 1 :
|||||A5 = 0 : 0
|||||A5 = 1 :
||||||A8 = 0 :
|||||||A4 = 0 :
||||||||A6 = 0 : 0
||||||||A6 = 1 : 1
|||||||A4 = 1 : 1
||||||A8 = 1 :
|||||||A6 = 0 :
||||||||A1 = 0 : 1
||||||||A1 = 1 : 0
|||||||A6 = 1 : 0
A11 = 1 :

Number of training instances = 400
Number of training attributes = 12
Total number of nodes in the tree = 209
Number of leaf nodes in the tree = 105
Accuracy of the model on the training dataset = 99.5%

Number of testing instances = 100
Number of testing attributes = 12
Accuracy of the model on the testing dataset = 75.0%

Post-Pruned Accuracy
-----------------------------------------

Number of training instances = 400
Number of training attributes = 12
Total number of nodes in the tree = 113
Number of leaf nodes in the tree = 57
Accuracy of the model on the training dataset = 94.5%

Number of testing instances = 100
Number of testing attributes = 12
Accuracy of the model on the testing dataset = 87.0%


Program Terminated. Have a great day!
```