# Programming Assignment 1

**Usernames**

pruputta

mkaranam

Decision tree at depth 1:

tree = {12: {0.0: 4.0, 8.0: 7.0, 2.0: 2.0, 4.0: 1.0, 6.0: 6.0}}

Decision tree at depth 2:

tree = {12: {0.0: {1: {0.0: 4.0}}, 8.0: 7.0, 2.0: {1: {0.0: 1.0, 1.0: 2.0}}, 4.0: {1: {0.0: 1.0}}, 6.0: {1: {0.0: 6.0}}}}

**Compile and Run code:**

$python decisionTreeLatest.py zoo-train.csv 1 zoo-test.csv

$python <python filename> <training csv file name> <depth> <test file name>

**Question 1**

The output of the classifier for depth = 1 is as follows:

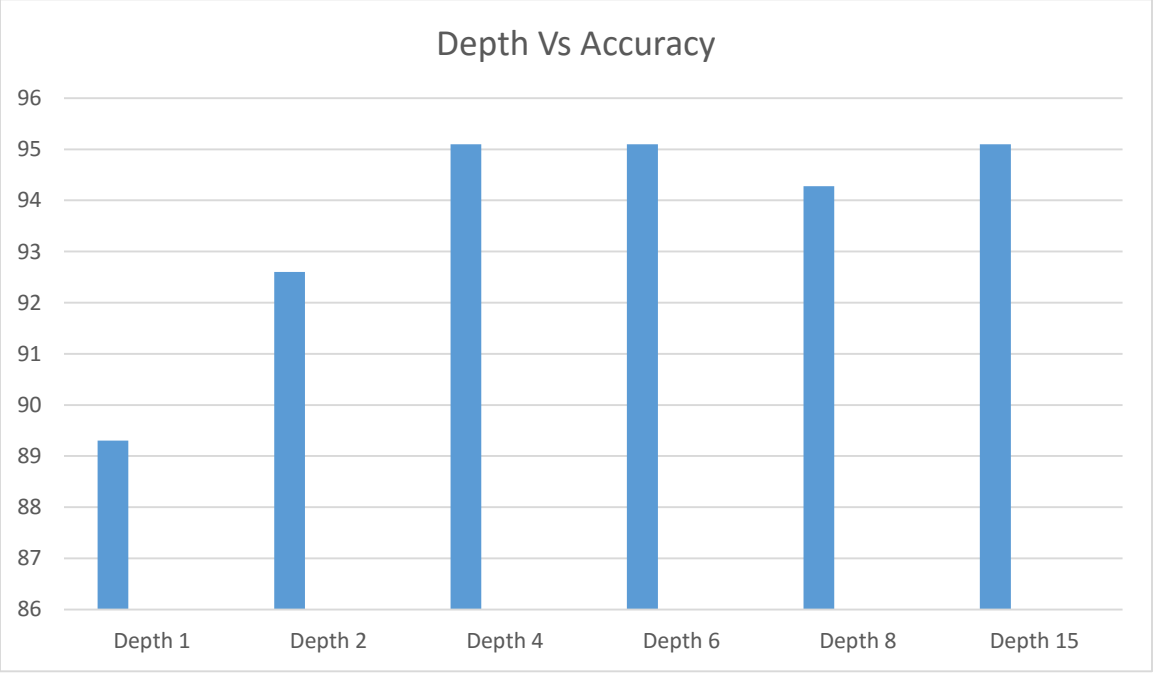
```
[pruputta@silo AppliedML]$ python decisionTreeLatest.py zoo-train.csv 1 zoo-test.csv
Decision Tree =  {12: {0.0: 4.0, 8.0: 7.0, 2.0: 2.0, 4.0: 1.0, 6.0: 6.0}}
Confusion matrix =
[[ 8.  4.  0.  2.  0.  0.  0.]
 [ 0.  7.  0.  0.  0.  0.  0.]
 [ 2.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  4.  0.  0.  0.]
 [ 1.  0.  0.  0.  0.  0.  0.]
 [ 0.  0.  0.  0.  0.  3.  0.]
 [ 1.  0.  0.  3.  0.  0.  0.]]
true positives =  [8.0, 7.0, 0.0, 4.0, 0.0, 3.0, 0.0]
true negatives =  [17.0, 24.0, 33.0, 26.0, 34.0, 32.0, 31.0]
false positives =  [4.0, 4.0, 0.0, 5.0, 0.0, 0.0, 0.0]
false negatives  =  [6.0, 0.0, 2.0, 0.0, 1.0, 0.0, 4.0]
total true pos =  22.0
total true neg =  197.0
Number of misclassifications =  13.0
Accuracy =  0.89387755102
[pruputta@silo AppliedML]$
```

The output of the classifier for depth = 2 is as follows:

```
[pruputta@silo AppliedML]$ python decisionTreeLatest.py zoo-train.csv 2 zoo-test.csv
Decision Tree =  {12: {0.0: {1: {0.0: 4.0}}, 8.0: 7.0, 2.0: {1: {0.0: 1.0, 1.0: 2.0}}, 4.0: {1: {0.0: 1.0}}, 6.0: {1: {0.0: 6.0}}}}
Confusion matrix =
[[ 12.   0.   0.   2.   0.   0.   0.]
 [  0.   7.   0.   0.   0.   0.   0.]
 [  2.   0.   0.   0.   0.   0.   0.]
 [  0.   0.   0.   4.   0.   0.   0.]
 [  1.   0.   0.   0.   0.   0.   0.]
 [  0.   0.   0.   0.   0.   3.   0.]
 [  1.   0.   0.   3.   0.   0.   0.]]
true positives =  [12.0, 7.0, 0.0, 4.0, 0.0, 3.0, 0.0]
true negatives =  [17.0, 28.0, 33.0, 26.0, 34.0, 32.0, 31.0]
false positives =  [4.0, 0.0, 0.0, 5.0, 0.0, 0.0, 0.0]
false negatives  =  [2.0, 0.0, 2.0, 0.0, 1.0, 0.0, 4.0]
total true pos =  26.0
total true neg =  201.0
Number of misclassifications =  9.0
Accuracy =  0.926530612245
[pruputta@silo AppliedML]$
```

The output of the classifier for depth = 4 is as follows:

```
[pruputta@silo AppliedML]$ python decisionTreeLatest.py zoo-train.csv 4 zoo-test.csv
Decision Tree =  {12: {0.0: {1: {0.0: {2: {0.0: {3: {0.0: 3.0, 1.0: 1.0}}, 1.0: {3: {0.0: 4.0}}}}}}, 8.0: 7.0, 2.0: {1: {0.0: 1.0,
1.0, 1.0: {3: {0.0: 5.0, 1.0: 1.0}}}}}}, 6.0: {1: {0.0: {2: {1.0: {3: {0.0: 6.0}}}}}}}
Confusion matrix =
[[ 14.   0.   0.   0.   0.   0.   0.]
 [  0.   7.   0.   0.   0.   0.   0.]
 [  0.   0.   0.   0.   2.   0.   0.]
 [  0.   0.   0.   4.   0.   0.   0.]
 [  0.   0.   0.   0.   1.   0.   0.]
 [  0.   0.   0.   0.   0.   3.   0.]
 [  1.   0.   0.   3.   0.   0.   0.]]
true positives =  [14.0, 7.0, 0.0, 4.0, 1.0, 3.0, 0.0]
true negatives =  [20.0, 28.0, 33.0, 28.0, 32.0, 32.0, 31.0]
false positives =  [1.0, 0.0, 0.0, 3.0, 2.0, 0.0, 0.0]
false negatives  =  [0.0, 0.0, 2.0, 0.0, 0.0, 0.0, 4.0]
total true pos =  29.0
total true neg =  204.0
Number of misclassifications =  6.0
Accuracy =  0.951020408163
[pruputta@silo AppliedML]$
```

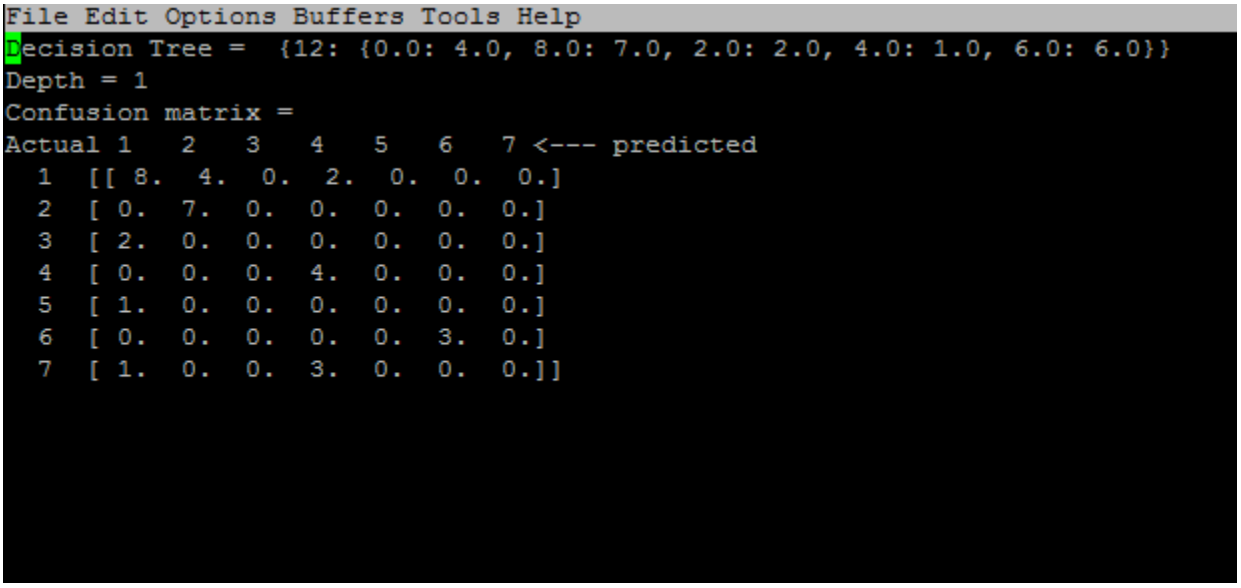


The error rates for different depths are as follows:

| Depth | Number of misclassifications |
|---|---|
| **1** | **13** |
| **2** | **9** |
| **4** | **6** |
| **6** | **6** |
| **8** | **7** |
| **15** | **7** |

**Question 2**

For Depth = 1, Confusion matrix and Decision Tree ouputs are as follows:




For Depth = 2, Confusion matrix and Decision Tree ouputs are as follows:

```
Decision Tree =  {12: {0.0: {1: {0.0: 4.0}}, 8.0: 7.0, 2.0: {1: {0.0: 1.0, 1.0: 2.0}}, 4.0: {1: {0.0: 1.0}}, 6.0: {1: {0.0: 6.0}}}}
Depth = 2
Confusion matrix =
Actual     1    2    3    4    5    6    7   <---- Predicted
   1    [[ 12.   0.   0.   2.   0.   0.   0.]
   2     [  0.   7.   0.   0.   0.   0.   0.]
   3     [  2.   0.   0.   0.   0.   0.   0.]
   4     [  0.   0.   0.   4.   0.   0.   0.]
   5     [  1.   0.   0.   0.   0.   0.   0.]
   6     [  0.   0.   0.   0.   0.   3.   0.]
   7     [  1.   0.   0.   3.   0.   0.   0.]]
```

**Question 3**

Output from Weka

```
J48 pruned tree
------------------

attribute_0 <= 0
|   attribute_2 <= 0
|   |   attribute_6 <= 0
|   |   |   attribute_3 <= 0: 7 (7.0/1.0)
|   |   |   attribute_3 > 0: 6 (4.0)
|   |   attribute_6 > 0
|   |   |   attribute_10 <= 0
|   |   |   |   attribute_11 <= 2: 3 (3.0)
|   |   |   |   attribute_11 > 2: 5 (3.0)
|   |   |   attribute_10 > 0: 4 (9.0)
|   attribute_2 > 0: 1 (27.0)
attribute_0 > 0: 2 (13.0)

Number of Leaves  :     7

Size of the tree :      13


Time taken to build model: 0.01 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances          32               91.4286 %
Incorrectly Classified Instances         3                8.5714 %
Kappa statistic                          0.8873
Mean absolute error                      0.028
Root mean squared error                  0.1483
Relative absolute error                 12.7242 %
Root relative squared error             44.8771 %
Total Number of Instances               35

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1         1         1          1        1
                 1         0          1         1         1          1        2
                 0         0          0         0         0          0.5      3
                 1         0          1         1         1          1        4
                 1         0.029      0.5       1         0.667      0.985    5
                 0.667     0          1         0.667     0.8        0.974    6
                 1         0.065      0.667     1         0.8        0.968    7
Weighted Avg.    0.914     0.008      0.89      0.914     0.893      0.965

=== Confusion Matrix ===

  a  b  c  d  e  f  g   <-- classified as
 14  0  0  0  0  0  0 |  a = 1
  0  7  0  0  0  0  0 |  b = 2
  0  0  0  0  1  0  1 |  c = 3
  0  0  0  4  0  0  0 |  d = 4
  0  0  0  0  1  0  0 |  e = 5
  0  0  0  0  0  2  1 |  f = 6
  0  0  0  0  0  0  4 |  g = 7
```

**Question 4**

Confusion matrix and decision tree with our own dataset(Cancer Dataset)

```
[pruputta@silo AppliedML]$ python decisionTreeLatest.py cancer-train1.csv 1 cancer-test1.csv
Decision Tree =  {2: {1.0: 2.0, 2.0: 2.0, 3.0: 1.0, 4.0: 1.0, 5.0: 1.0, 6.0: 1.0, 7.0: 1.0, 8.0: 1.0, 9.0: 1.0, 10.0: 1.0}}
Confusion matrix =
[[  44.    0.]
 [   8.  147.]]
true positives =  [44.0, 147.0]
true negatives =  [147.0, 44.0]
false positives =  [8.0, 0.0]
false negatives  =  [0.0, 8.0]
total true pos =  191.0
total true neg =  191.0
Number of misclassifications =  8.0
Accuracy =  0.959798994975
[pruputta@silo AppliedML]$
```

Question 3 with our own dataset(Cancer Dataset)

```
J48 pruned tree
------------------

attribute_1 <= 2
|   attribute_5 <= 3: 2 (263.0/2.0)
|   attribute_5 > 3
|   |   attribute_0 <= 3: 2 (7.0)
|   |   attribute_0 > 3
|   |   |   attribute_6 <= 2
|   |   |   |   attribute_3 <= 3: 1 (2.0)
|   |   |   |   attribute_3 > 3: 2 (2.0)
|   |   |   attribute_6 > 2: 1 (8.0)
attribute_1 > 2
|   attribute_5 <= 3
|   |   attribute_8 <= 4
|   |   |   attribute_2 <= 2: 2 (12.0)
|   |   |   attribute_2 > 2
|   |   |   |   attribute_8 <= 1
|   |   |   |   |   attribute_0 <= 5
|   |   |   |   |   |   attribute_3 <= 4: 2 (10.0/1.0)
|   |   |   |   |   |   attribute_3 > 4: 1 (5.0/1.0)
|   |   |   |   |   attribute_0 > 5: 1 (14.0/3.0)
|   |   |   |   attribute_8 > 1: 1 (6.0)
|   |   attribute_8 > 4: 1 (6.0)
|   attribute_5 > 3: 1 (165.0/8.0)

Number of Leaves  :     12

Size of the tree :      23


Time taken to build model: 0.04 seconds

=== Evaluation on test set ===
=== Summary ===

Correctly Classified Instances         197               98.995  %
Incorrectly Classified Instances         2                1.005  %
Kappa statistic                          0.9713
Mean absolute error                      0.0292
Root mean squared error                  0.0972
Relative absolute error                  6.6256 %
Root relative squared error             21.6047 %
Total Number of Instances              199

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0.013      0.957     1         0.978      0.996    1
                 0.987     0          1         0.987     0.994      0.996    2
Weighted Avg.    0.99      0.003      0.99      0.99      0.99       0.996

=== Confusion Matrix ===

   a   b   <-- classified as
  44   0 |   a = 1
   2 153 |   b = 2
```