CS5483 Data Warehousing & Data Mining – Assignment 1

Name: Poon Darren Yuk Ho

Student ID: 52833413

==========

## Glass Identification

Data set description: number of attributes: 9 (RI, Na, Mg, Al, Si, K, Ca, Ba, Fe)

number of instances:214

classes: “1”, “2”, “3”, “4”, “5”, “6”, “7”

1) Construct different decision trees based on different partitions of the data set into training and test sets.

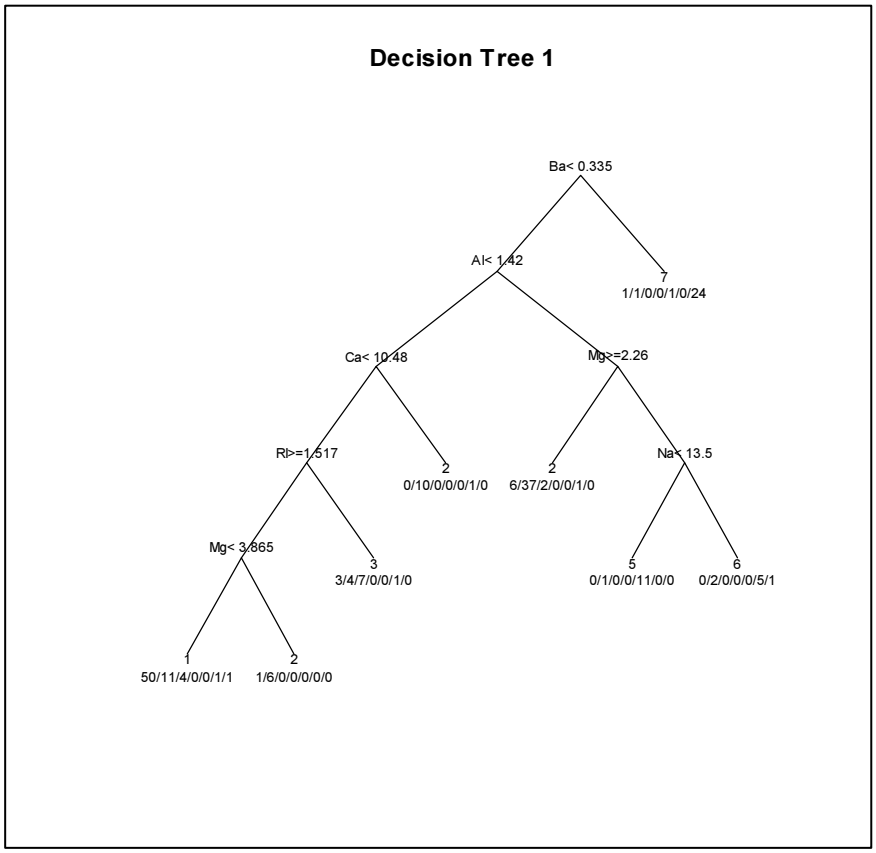
- I will be using 10-fold cross validation to this data set.
- Each partition contains 21-22 rows of data.
- For each run, one of the partitions is used for testing. The rest of them are used for training. This will be repeated for 10 times so that each partition is used for testing exactly once only.

Results of decision trees

Decision Tree 1:

- uses data[22-214] for training, and data[1-21] for testing.
- No. of leave nodes: 8
- No. of training records: 193
- No. of misclassified training records: 43
- Training error rate: 43/193=22.28%

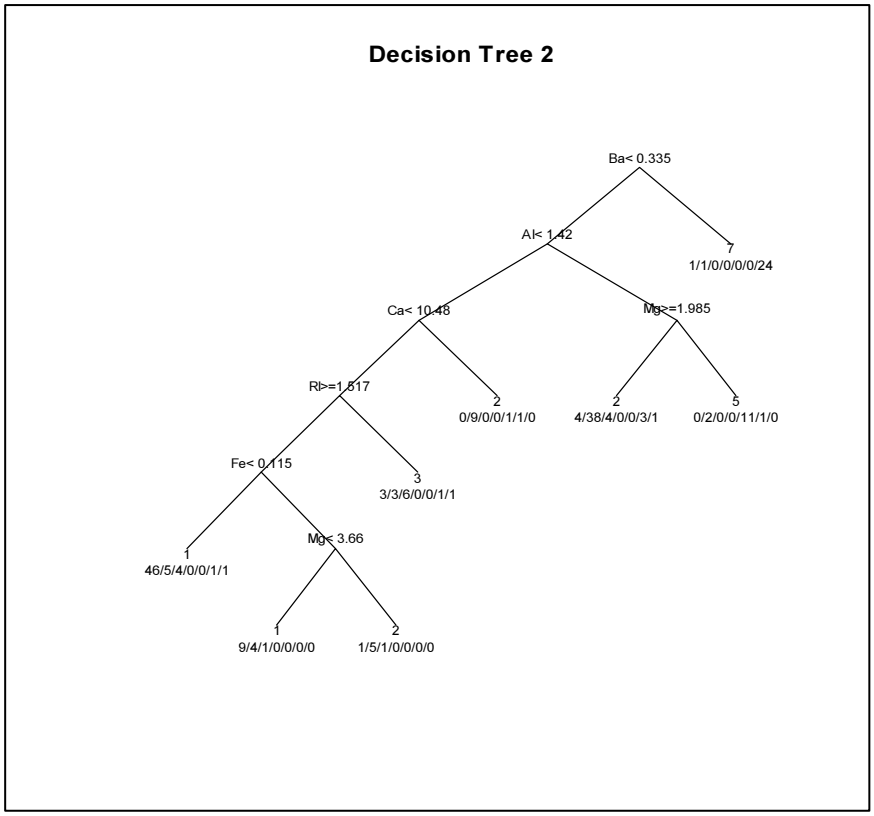
- No. of testing records: 21
- No. of correctly classified testing records: 15
- No. of misclassified testing records: 6
- Classifier error: 6/21 = 28.57%



Decision Tree 2:

- uses data[1-21, 43-214] for training, and data[22-42] for testing.
- No. of leave nodes: 8
- No. of training records: 193
- No. of misclassified training records: 45
- Training error rate: 23.32%

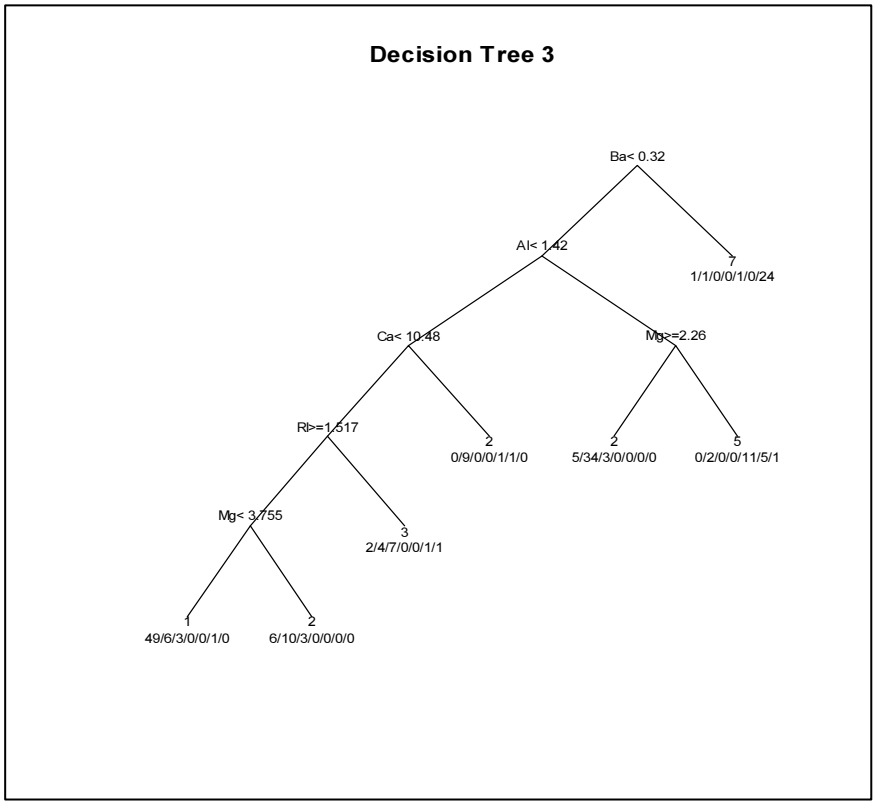
- No. of testing records: 21
- No. of correctly classified testing records: 13
- No. of misclassified testing records: 8
- Classifier error: 8/21 = 38.09%




Decision Tree 3:

- uses data[1-42, 65-214] for training, and data[43-64] for testing.
- No. of leave nodes: 7
- No. of training records: 192
- No. of misclassified training records: 48
- Training error rate: 25%

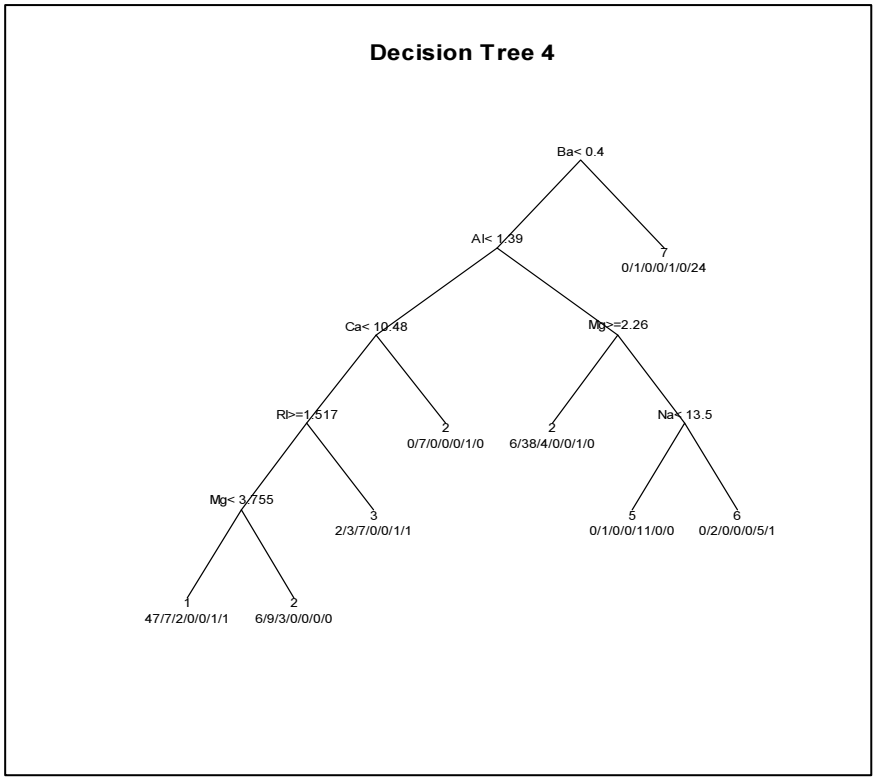
- No. of testing records: 22
- No. of correctly classified testing records: 13
- No. of misclassified testing records: 9
- Classifier error: 9/22 = 40.91%



Decision Tree 4:

- uses data[1-64, 86-214] for training, and data[65-85] for testing.
- No. of leave nodes: 8
- No. of training records: 193
- No. of misclassified training records: 52
- Training error rate: 26.94%

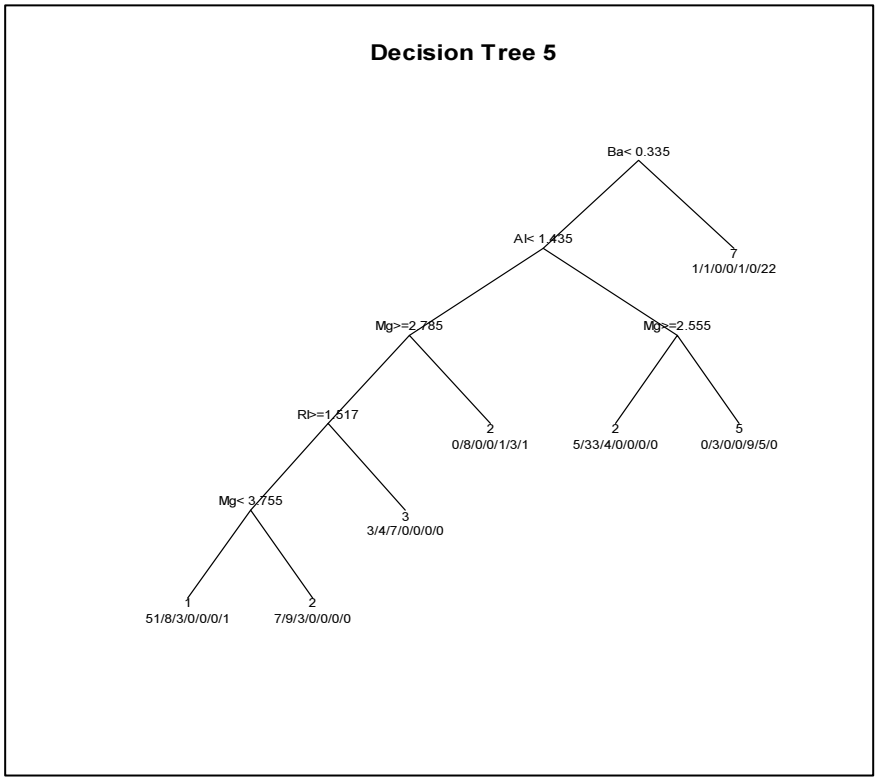
- No. of testing records: 21
- No. of correctly classified testing records: 14
- No. of misclassified testing records: 7
- Classifier error: 7/21 = 33.33%




Decision Tree 5:

- uses data[1-85, 107-214] for training, and data[86-106] for testing.
- No. of leave nodes: 7
- No. of training records: 193
- No. of misclassified training records: 54
- Training error rate: 27.98%

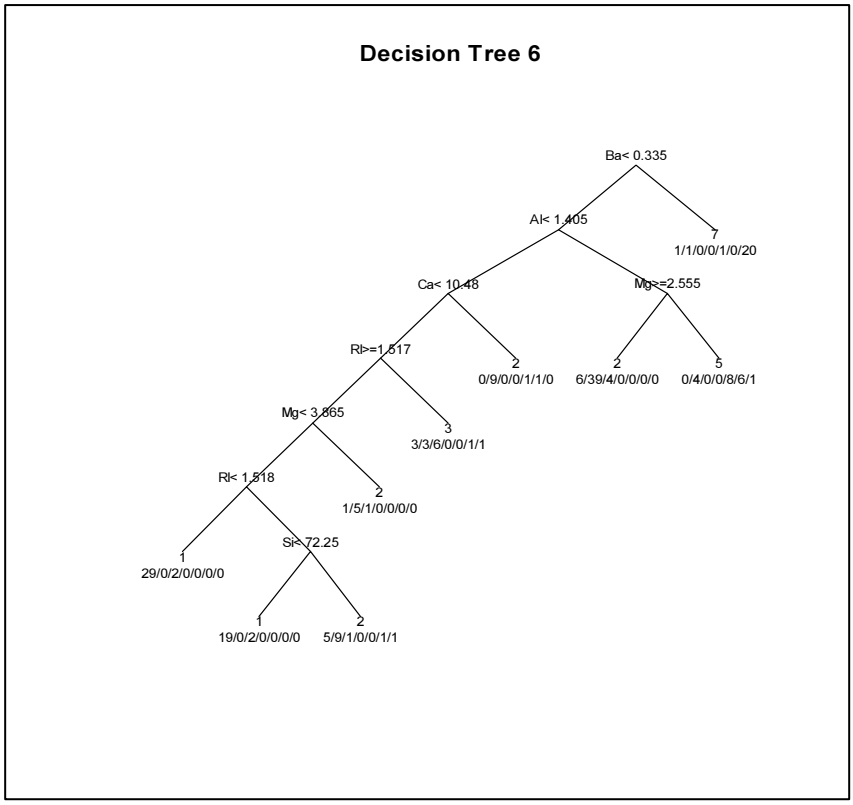
- No. of testing records: 21
- No. of correctly classified testing records: 14
- No. of misclassified testing records: 7
- Classifier error: 7/21 = 33.33%



Decision Tree 6:

- uses data[1-106, 129-214] for training, and data[107-128] for testing.
- No. of leave nodes: 9
- No. of training records: 192
- No. of misclassified training records: 48
- Training error rate: 25%

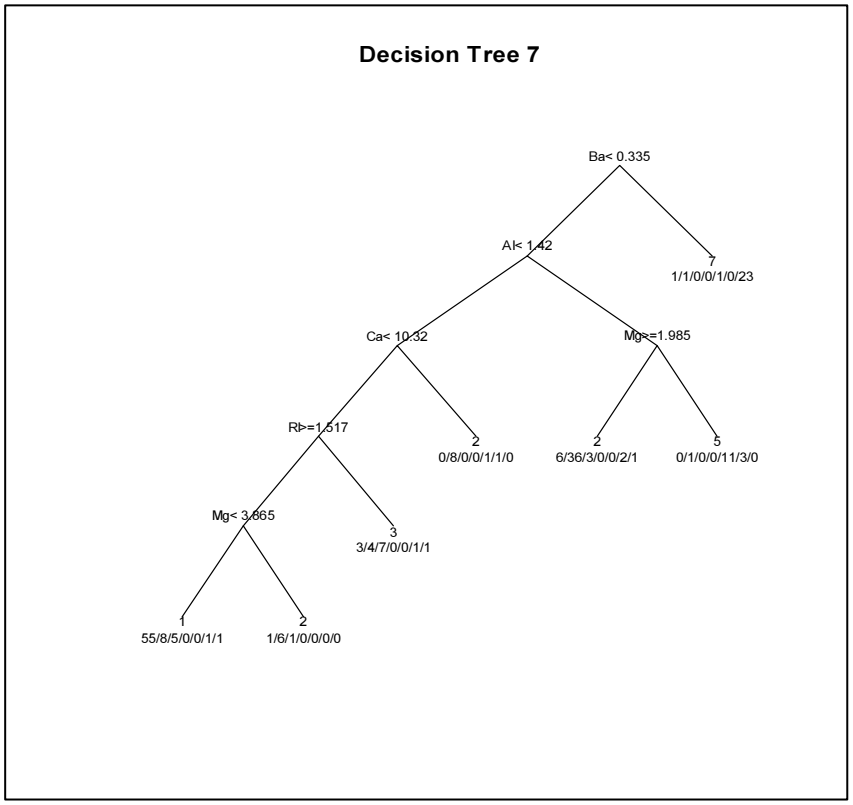
- No. of testing records: 22
- No. of correctly classified testing records: 17
- No. of misclassified testing records: 5
- Classifier error: 5/22 = 22.73%




Decision Tree 7:

- uses data[1-128, 150-214] for training, and data[129-149] for testing.
- No. of leave nodes: 7
- No. of training records: 193
- No. of misclassified training records: 47

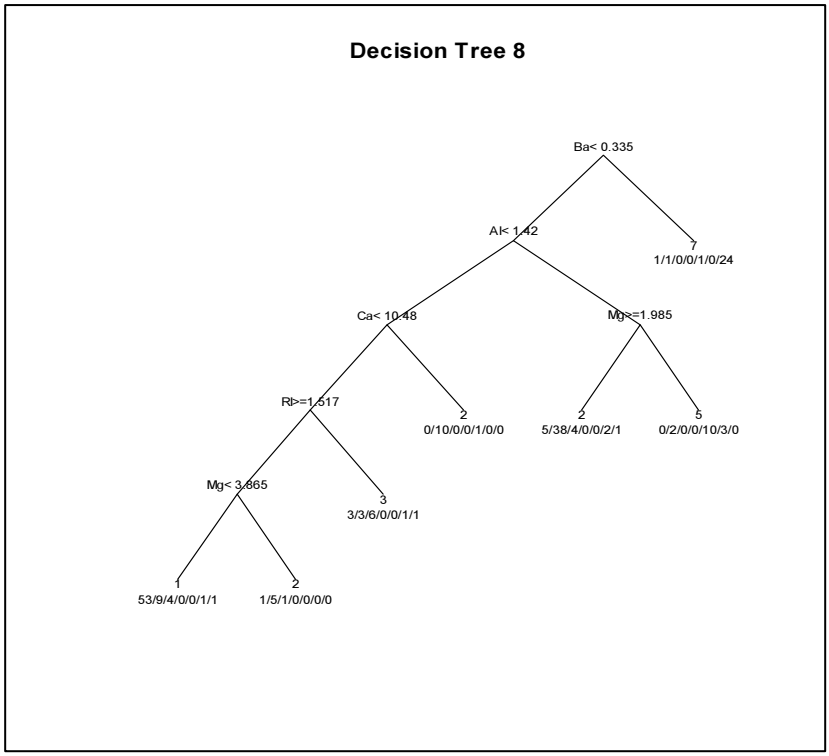
- No. of testing records: 21
- No. of correctly classified testing records: 15
- No. of misclassified testing records: 6
- Classifier error: 6/21 = 28.57%



Decision Tree 8:

- uses data[1-149, 172-214] for training, and data[150-171] for testing.
- No. of leave nodes: 7
- No. of training records: 192
- No. of misclassified training records: 46
- Training error rate: 23.96%

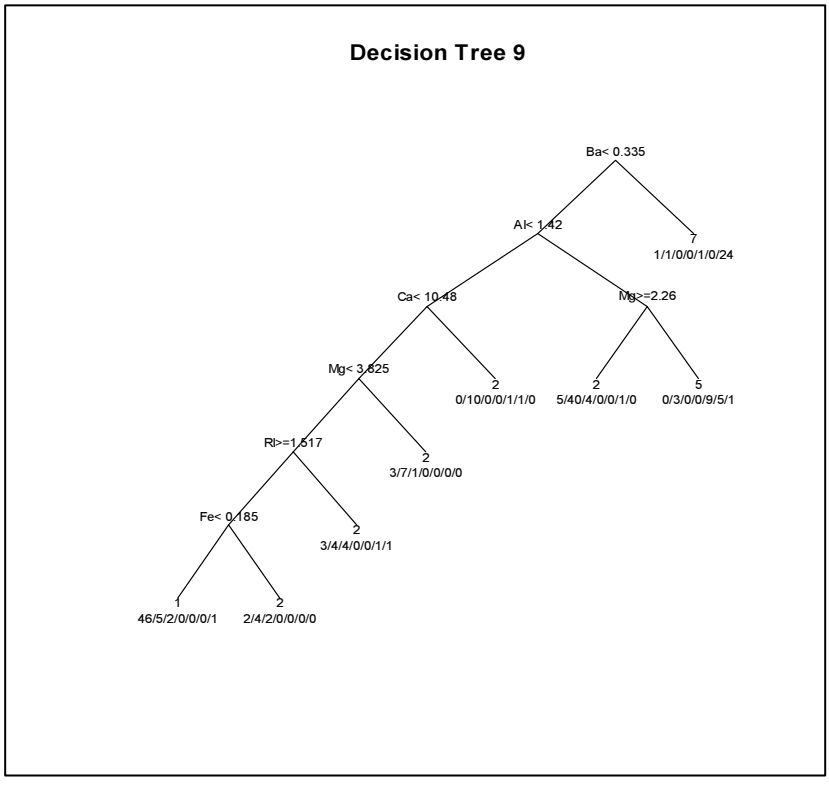
- No. of testing records: 22
- No. of correctly classified testing records: 15
- No. of misclassified testing records: 7
- Classifier error: 7/22 = 31.82%




Decision Tree 9:

- uses data[1-171, 193-214] for training, and data[172-192] for testing.
- No. of leave nodes: 8
- No. of training records: 193
- No. of misclassified training records: 49
- Training error rate: 25.39%

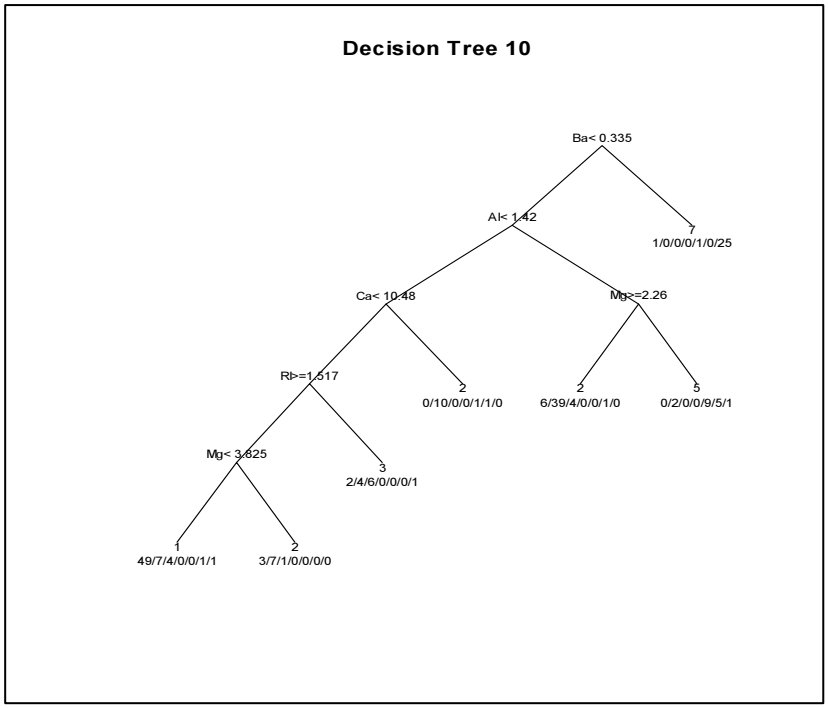
- No. of testing records: 21
- No. of correctly classified testing records: 13
- No. of misclassified testing records: 8
- Classifier error: 8/21 = 38.10%



Decision Tree 10:

- uses data[1-192] for training, and data[193-214] for testing.
- No. of leave nodes: 7
- No. of training records: 192
- No. of misclassified training records: 47
- Training error rate: 24.48%

- No. of testing records: 22
- No. of correctly classified testing records: 14
- No. of misclassified testing records: 8
- Classifier error: 8/22 = 36.36%




2) Compare the structures and classification performance of these different trees.

The structures of these 10 trees are identical, in which all the trees have 7-9 leave nodes. The root node of each tree has selected "Ba<0.335" as their first split. Also, all root nodes' left child are using "AI < 1.42" as the second split, and "7" (a leave node) to be its right child.

Moreover, all of these trees have similar training error rate in the range of 23% to 30%. The classifier error is also identical with an average of 33% (sum of error/ k) error rate.

3) For some of these trees, analyze the class distribution of the records and the associated classification error at selected nodes.

Tree 1 class distribution: 193 training records

Class 1: 67 (50 correctly classified, 17 misclassified), classification error = 17/67 = 25.37%

Class 2: 65 (53 correctly classified, 12 misclassified), classification error = 12/65 = 18.46%

Class 3: 14 (7 correctly classified, 7 misclassified), classification error = 7/14 = 50%

Class 4: 0

Class 5: 12 (11 correctly classified, 1 misclassified), classification error = 1/12 = 8.33%

Class 6: 8 (5 correctly classified, 3 misclassified), classification error = 3/8 = 37.5%

Class 7: 27 (24 correctly classified, 3 misclassified), classification error = 3/27 = 11.11%

Tree 3 class distribution: 192 training records

Class 1: 59 (49 correctly classified, 10 misclassified), classification error = 10/59 = 16.95%

Class 2: 73 (53 correctly classified, 20 misclassified), classification error = 20/73 = 27.4%

Class 3: 14 (7 correctly classified, 7 misclassified), classification error = 7/14 = 50%

Class 4: 0

Class 5: 19 (11 correctly classified, 8 misclassified), classification error = 8/19 = 42.11%

Class 6: 0

Class 7: 27 (24 correctly classified, 3 misclassified), classification error = 3/27 = 11.11%

Tree 5 class distribution: 193 training records

Class 1: 63 (51 correctly classified, 12 misclassified), classification error = 12/63 = 19.05%

Class 2: 74 (50 correctly classified, 24 misclassified), classification error = 24/74 = 32.43%

Class 3: 14 (7 correctly classified, 7 misclassified), classification error = 7/14 = 50%

Class 4: 0

Class 5: 17 (9 correctly classified, 8 misclassified), classification error = 8/17 = 47.06%

Class 6: 0

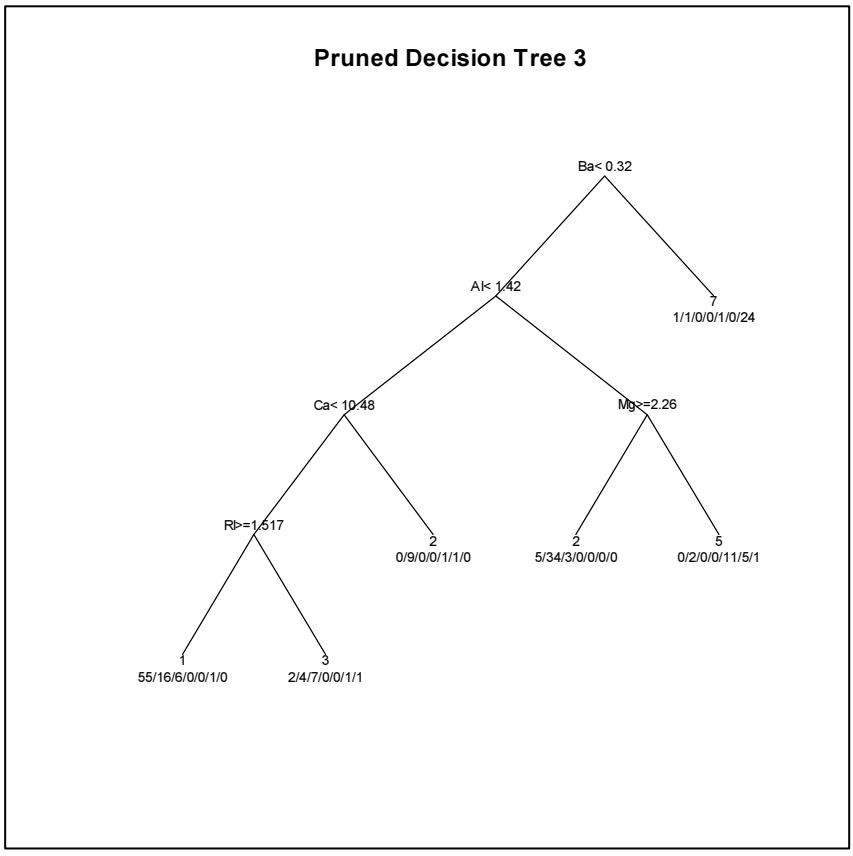
Class 7: 25 (22 correctly classified, 3 misclassified), classification error = 3/25 = 12%

4) Apply pruning to some of the trees you have constructed. For each of these trees, compare its classification performance before and after pruning.

Pruned Decision Tree 3

*Error before pruning:*

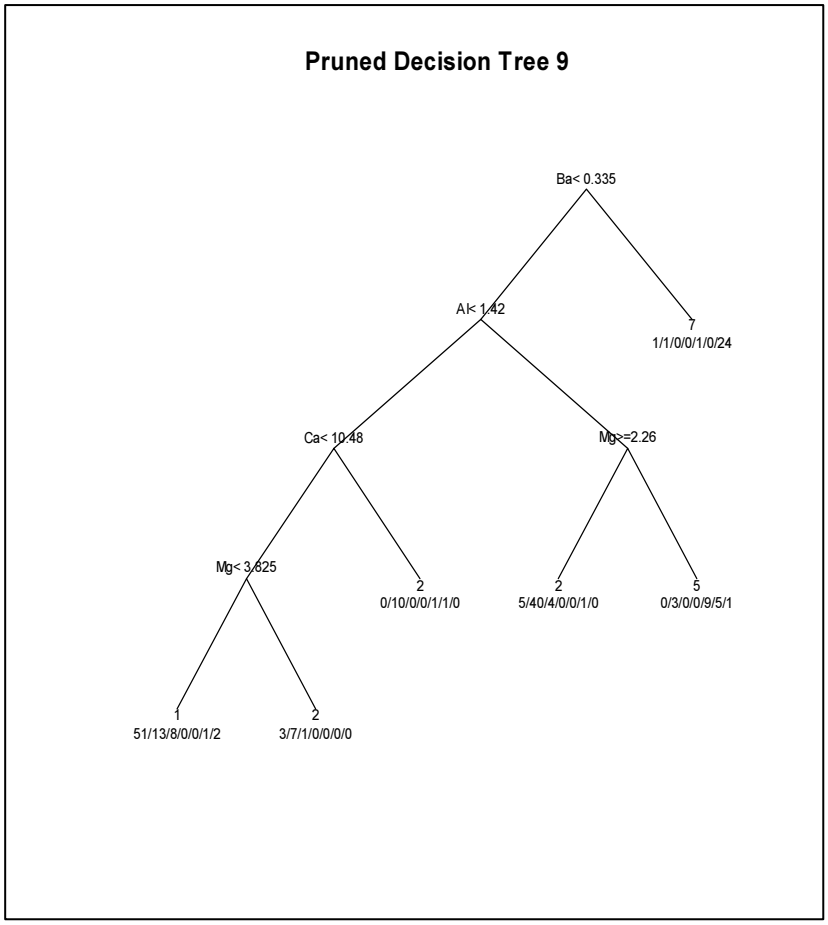
- No. of training records: 192
- No. of leave nodes: 7
- No. of misclassified training records: 48
- Training error rate: 25%
- No. of testing records: 22
- No. of correctly classified testing records: 13
- No. of misclassified testing records: 9
- Classifier error: 9/22 = 40.91%




*Error after pruning:*

- No. of leave nodes: 6
- No. of misclassified training records: 52
- Training error rate: 27.08%
- No. of correctly classified testing records: 15
- No. of misclassified testing records: 7
- Classifier error: 7/22 = 31.82%
- Therefore, the overfitting problem is relieved. As we can see, even though the training error increases, the testing error decreases.

Pruned Decision Tree 9



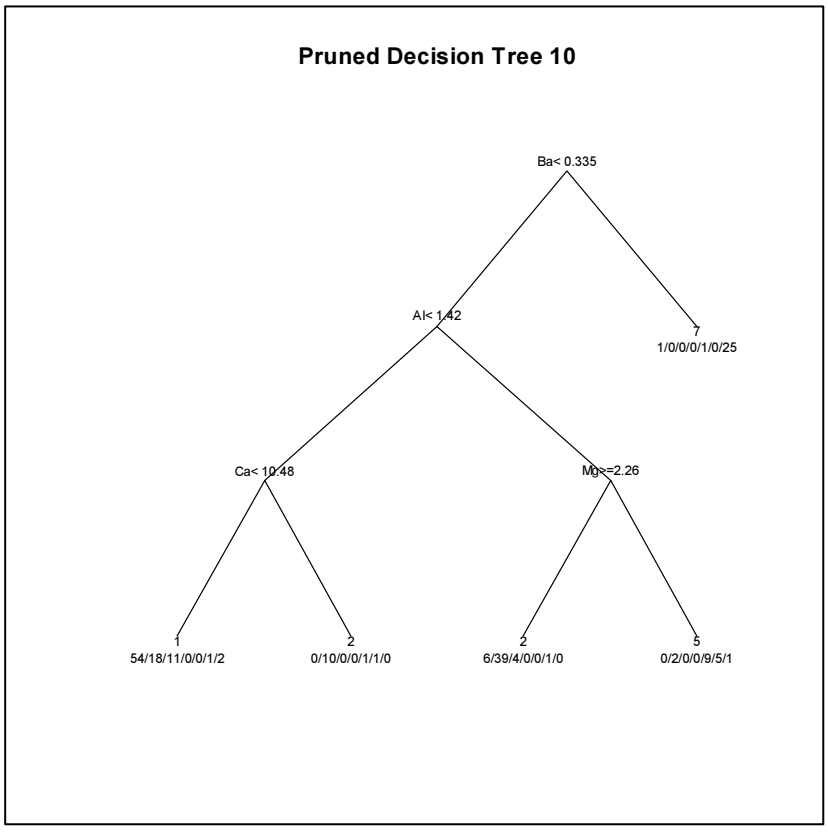

*Error before pruning:*

- No. of training records: 193
- No. of leave nodes: 8
- No. of misclassified training records: 49
- Training error rate: 25.39%
- No. of testing records: 21
- No. of correctly classified testing records: 13
- No. of misclassified testing records: 8
- Classifier error: 8/21 = 38.10%

*Error after pruning:*

- No. of leave nodes: 6
- No. of misclassified training records: 52
- Training error rate: 26.94%
- No. of correctly classified testing records: 14
- No. of misclassified testing records: 7
- Classifier error: 7/21 = 33.33%
- Therefore, the overfitting problem is relieved. As we can see, even though the training error increases, the testing error decreases.

Pruned Decision Tree 10




*Error before pruning:*

- No. of training records: 192
- No. of leave nodes: 7
- No. of misclassified training records: 47
- Training error rate: 24.48%
- No. of testing records: 22
- No. of correctly classified testing records: 14
- No. of misclassified testing records: 8
- Classifier error: 8/22 = 36.36%

*Error after pruning:*

- No. of leave nodes: 5
- No. of misclassified training records: 55
- Training error rate: 28.65%
- No. of correctly classified testing records: 14
- No. of misclassified testing records: 8
- Classifier error: 8/22 = 36.36%
- Even though the training error increases after the tree is pruned, the classifier error does not decrease at all. This tree may suffer in other problems rather than overfitting, ex. not enough data to train/not enough meaningful attributes to train.

---

## Wine

Data set description: number of attributes: 13 (Al, Ma, Ash, Aoa, Mag, TP, F, NP, P, CI, H, OD, Pro)
number of instances: 178
classes: "1", "2", "3"

1) Construct different decision trees based on different partitions of the data set into training and test sets.

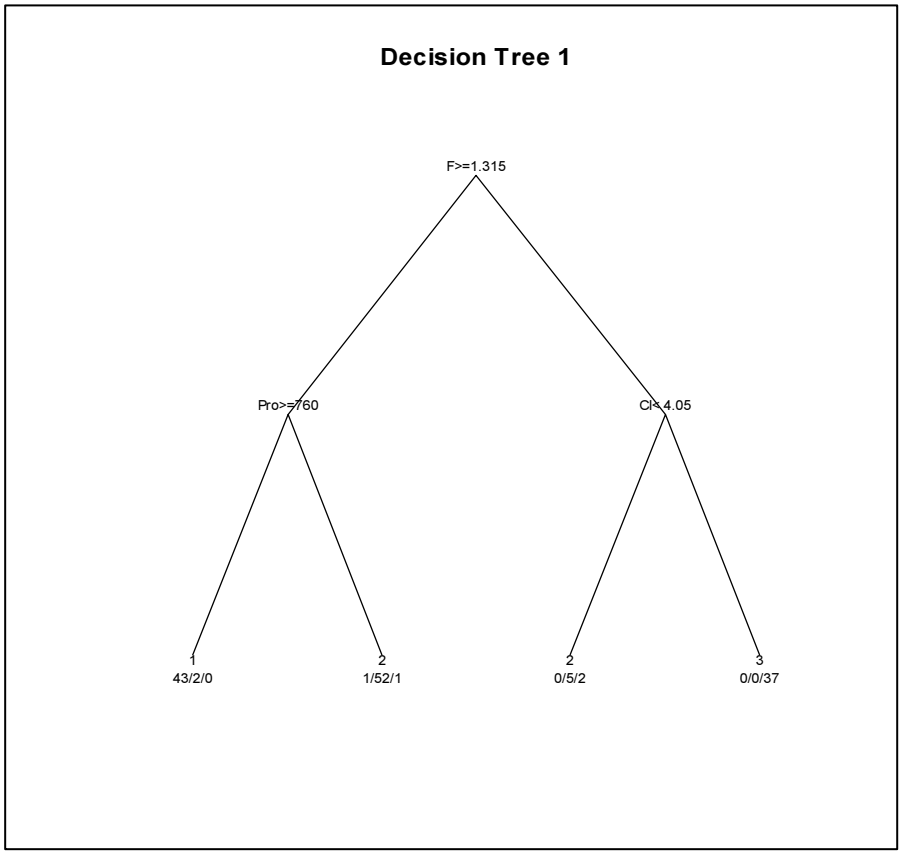
- I will be using 5-fold cross validation to this data set.
- Each partition contains 34-35 rows of data.
- For each run, one of the partitions is used for testing. The rest of them are used for training. This will be repeated for 5 times so that each partition is used for testing exactly once only.

Results of decision trees

Decision Tree 1:

- uses data[36-178] for training, and data[1-35] for testing.
- No. of leave nodes: 4
- No. of training records: 143
- No. of misclassified training records: 6
- Training error rate: 6/143 = 4.2%

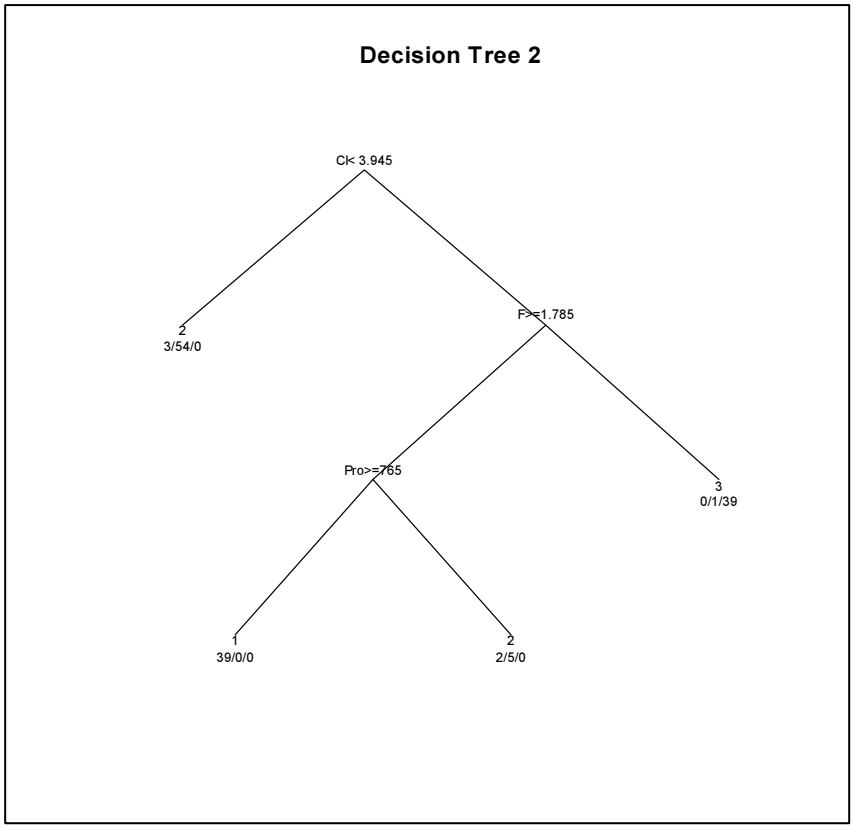
- No. of testing records: 35
- No. of correctly classified testing records: 31
- No. of misclassified testing records: 4
- Classifier error: 4/35 = 11.43%



Decision Tree 2:

- uses data[1-35, 71-178] for training, and data[36-70] for testing.
- No. of leave nodes: 4
- No. of training records: 143
- No. of misclassified training records: 6
- Training error rate: 6/143 = 4.2%

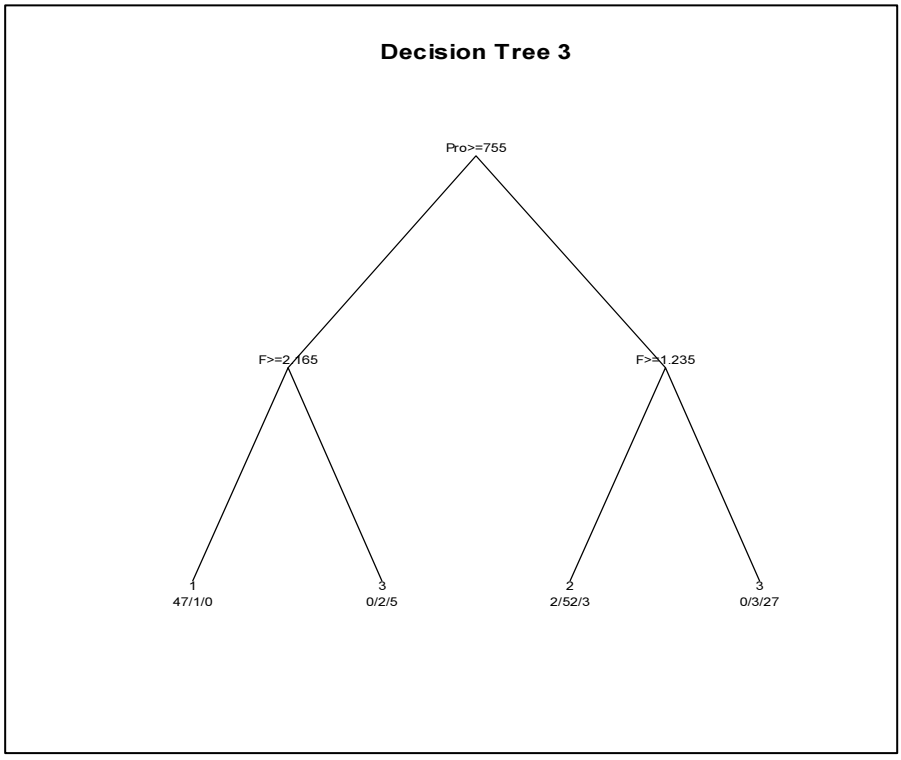
- No. of testing records: 35
- No. of correctly classified testing records: 29
- No. of misclassified testing records: 6
- Classifier error: 6/35 = 17.14%




Decision Tree 3:

- uses data[1-70, 107-178] for training, and data[71-106] for testing.
- No. of leave nodes: 4
- No. of training records: 142
- No. of misclassified training records: 11
- Training error rate: 11/142 = 7.75%

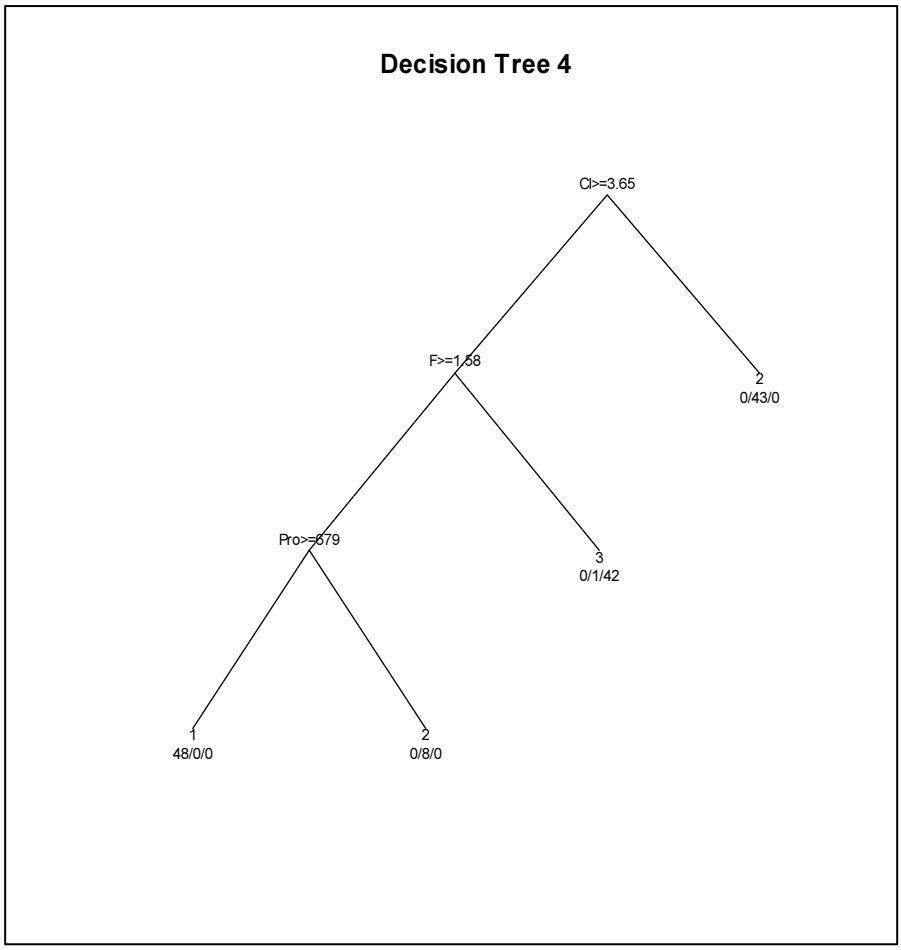
- No. of testing records: 36
- No. of correctly classified testing records: 32
- No. of misclassified testing records: 4
- Classifier error: 4/36 = 11.11%



Decision Tree 4:

- uses data[1-106, 143-178] for training, and data[107-142] for testing.
- No. of leave nodes: 4
- No. of training records: 142
- No. of misclassified training records: 1
- Training error rate: 1/142 = 0.7%

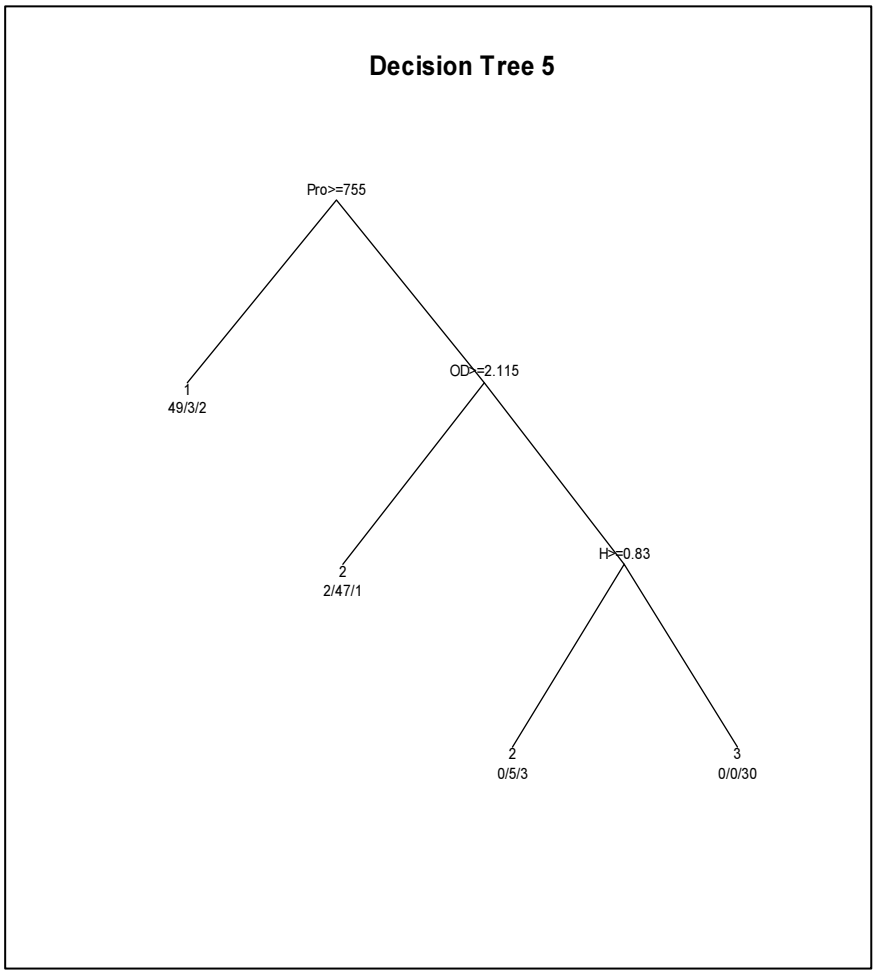
- No. of testing records: 36
- No. of correctly classified testing records: 32
- No. of misclassified testing records: 4
- Classifier error: 4/36 = 11.11%




Decision Tree 5:

- uses data[1-142] for training, and data[143-178] for testing.
- No. of leave nodes: 4
- No. of training records: 142
- No. of misclassified training records: 11
- Training error rate: 11/142 = 7.75%

- No. of testing records: 36
- No. of correctly classified testing records: 28
- No. of misclassified testing records: 8
- Classifier error: 8/36 = 22.22%



2) Compare the structures and classification performance of these different trees.

Even though the structures of these 5 trees are not quite identical (ex. the way of how nodes are splitted, the first attribute chose for splitting), they have some similar properties, and for example, all trees have the same number of leave nodes with size of 4. Also, all these trees have relatively low training error rate and classifier rate.

3) For some of these trees, analyze the class distribution of the records and the associated classification error at selected nodes.

Tree 1 class distribution: 143 training records

Class 1: 45 (43 correctly classified, 2 misclassified), classification error = 2/45 = 4.44%

Class 2: 61 (57 correctly classified, 4 misclassified), classification error = 4/61 = 6.56%

Class 3: 37 (37 correctly classified, 0 misclassified), classification error = 0/37 = 0%

Tree 3 class distribution: 142 training records

Class 1: 48 (47 correctly classified, 1 misclassified), classification error = 1/48 = 2.08%

Class 2: 57 (52 correctly classified, 5 misclassified), classification error = 5/57 = 8.77%

Class 3: 37 (32 correctly classified, 5 misclassified), classification error = 5/37 = 13.51%

Tree 5 class distribution: 142 training records

Class 1: 54 (49 correctly classified, 5 misclassified), classification error = 5/54 = 9.26%

Class 2: 58 (52 correctly classified, 6 misclassified), classification error = 6/58 = 10.34%

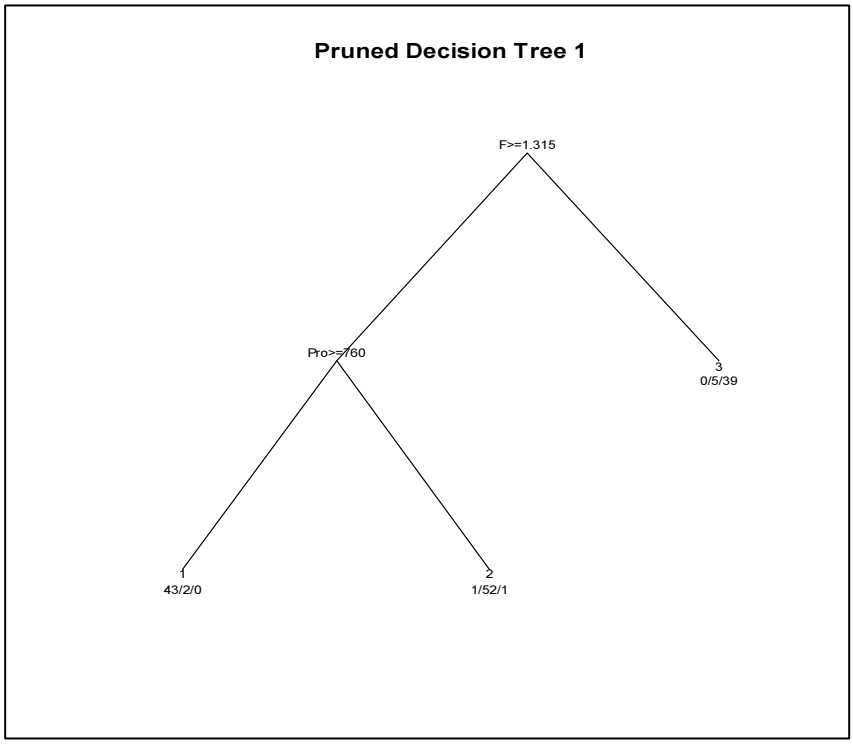
Class 3: 30 (30 correctly classified, 0 misclassified), classification error = 0/30 = 0%

4) Apply pruning to some of the trees you have constructed. For each of these trees, compare its classification performance before and after pruning.

Pruned Decision Tree 1

*Error before pruning:*

- No. of training records: 143
- No. of leave nodes: 4
- No. of misclassified training records: 6
- Training error rate: 4.2%
- No. of testing records: 35
- No. of correctly classified testing records: 31
- No. of misclassified testing records: 4
- Classifier error: 4/35 = 11.43%



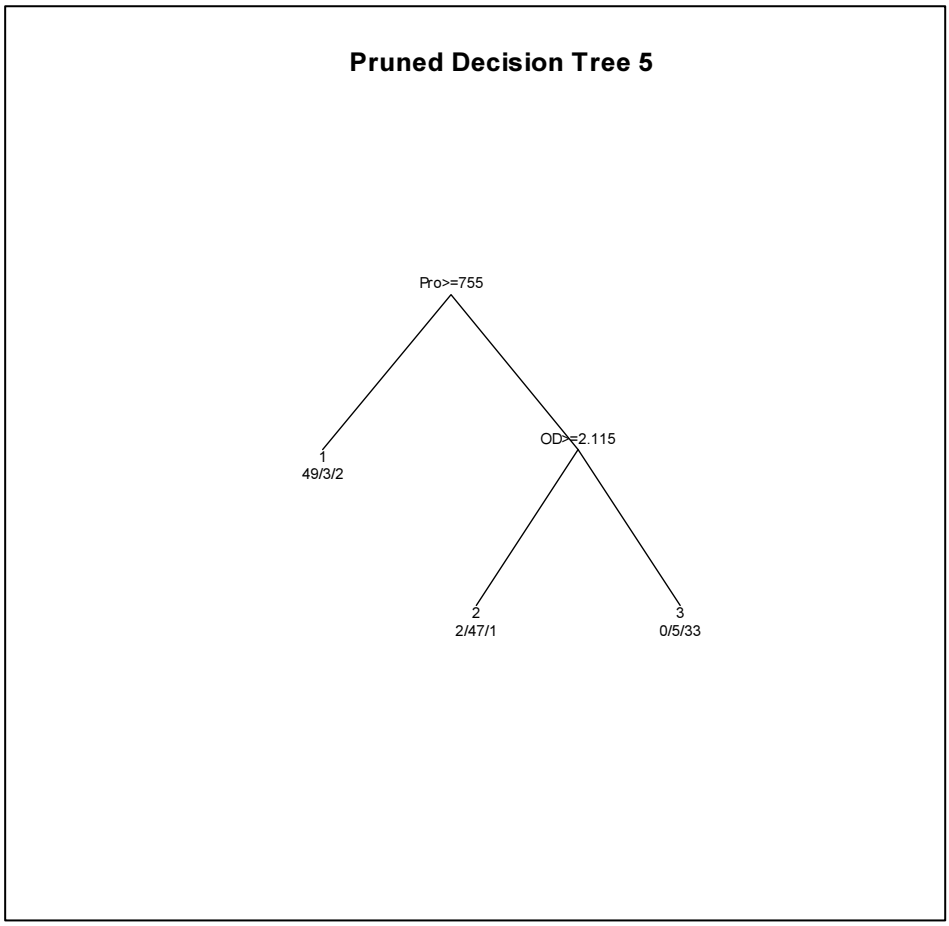
*Error after pruning:*

- No. of leave nodes: 3
- No. of misclassified training records: 9
- Training error rate: 6.29%
- No. of correctly classified testing records: 28
- No. of misclassified testing records: 7
- Classifier error: 7/35 = 20%
- Therefore, the training error and classifier both increase, which means pruning the tree would just cause the underfitting problem more serious.

Pruned Decision Tree 5

*Error before pruning:*

- No. of training records: 142
- No. of leave nodes: 4
- No. of misclassified training records: 11
- Training error rate: 7.75%
- No. of testing records: 36
- No. of correctly classified testing records: 28
- No. of misclassified testing records: 8
- Classifier error: 8/36 = 22.22%




*Error after pruning:*

- No. of leave nodes: 3
- No. of misclassified training records: 13
- Training error rate: 9.15%
- No. of correctly classified testing records: 29
- No. of misclassified testing records: 7
- Classifier error: 7/36 = 19.44%
- Therefore, the overfitting problem is relieved. As we can see, even the training error increases after pruning, the classifier error decreases.