Assignment 3

Kubra Iqbal

Due Date: Monday, October 22nd, by midnight

Total number of points: 55 points plus 5 for extra credit

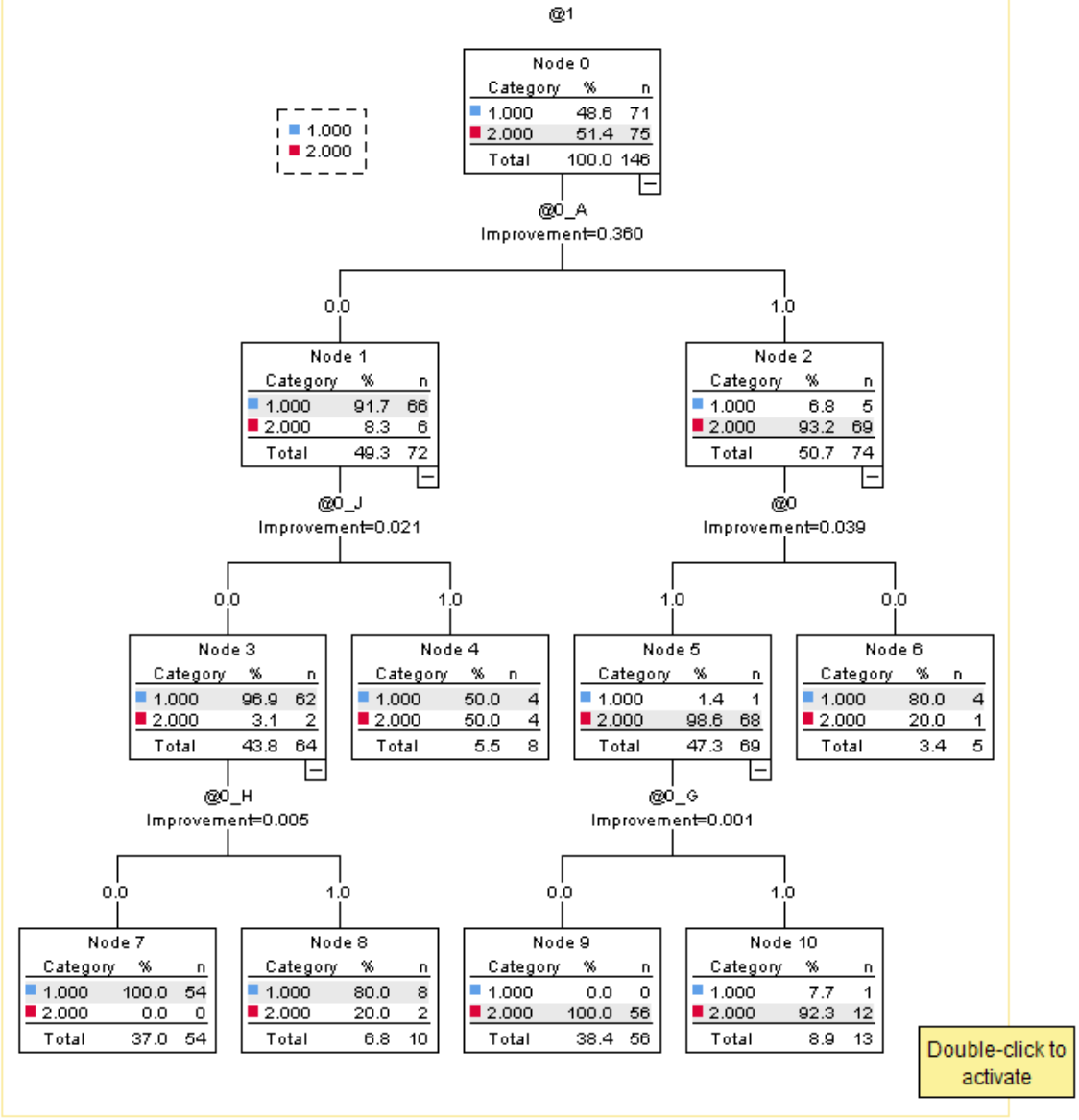
Problem 1 (20 points): This problem illustrates the classification approach by using decision trees and the Lupus data (you can download the data file "sledata" from D2L site, course documents for week 6). The data consists of 300 patient records. Each record contains 12 elements. The first 11 elements stand for different symptoms and the final element of each record indicates the diagnosis. Build a decision tree and report:

1) **The decision tree and the criteria used for building the tree for deciding the best split and the stopping condition (such as which impurity measure, how many cases for parents and children per node, etc)**

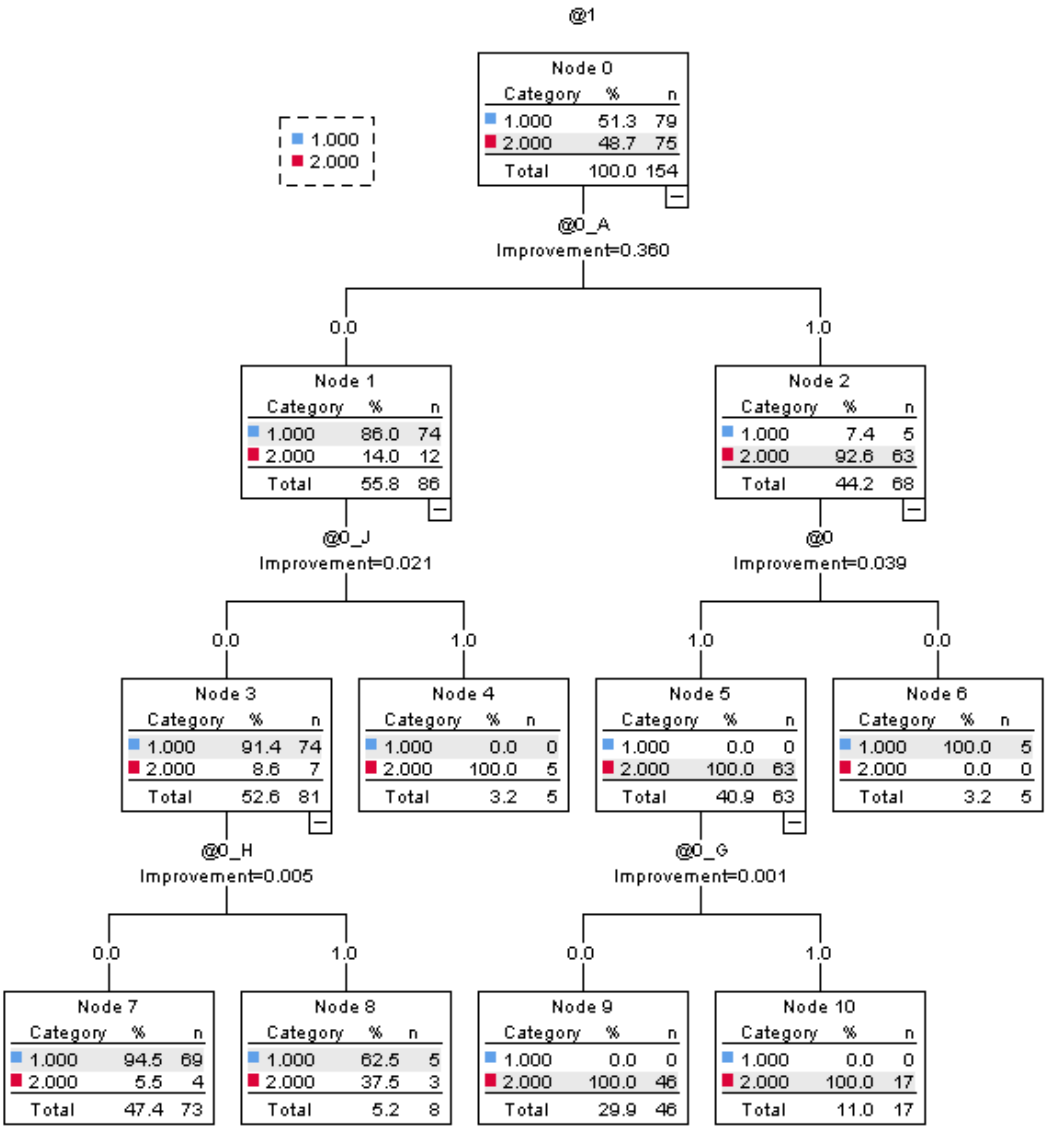
Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	@1
	Independent Variables	@0_J, @0_I, @0_H, @0_G, @0_F, @0_E, @0_D, @0_C, @0_B, @0_A, @0
	Validation	Split Sample
	Maximum Tree Depth	3
	Minimum Cases in Parent Node	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	@0_A, @0_D, @0, @0_B, @0_E, @0_J, @0_H, @0_F, @0_I, @0_C, @0_G
	Number of Nodes	11
	Number of Terminal Nodes	6
	Depth	3

Training Sample



Test Sample


@1
1.000
2.000
Node 0
Category % n
1.000 51.3 79
2.000 48.7 75
Total 100.0 154
@0_A
Improvement=0.360
0.0
1.0
Node 1
Category % n
1.000 86.0 74
2.000 14.0 12
Total 55.8 86
Node 2
Category % n
1.000 7.4 5
2.000 92.6 63
Total 44.2 68
@0_J
Improvement=0.021
@0
Improvement=0.039
0.0
1.0
1.0
0.0
Node 3
Category % n
1.000 91.4 74
2.000 8.6 7
Total 52.6 81
Node 4
Category % n
1.000 0.0 0
2.000 100.0 5
Total 3.2 5
Node 5
Category % n
1.000 0.0 0
2.000 100.0 63
Total 40.9 63
Node 6
Category % n
1.000 100.0 5
2.000 0.0 0
Total 3.2 5
@0_H
Improvement=0.005
@0_G
Improvement=0.001
0.0
1.0
0.0
1.0
Node 7
Category % n
1.000 94.5 69
2.000 5.5 4
Total 47.4 73
Node 8
Category % n
1.000 62.5 5
2.000 37.5 3
Total 5.2 8
Node 9
Category % n
1.000 0.0 0
2.000 100.0 46
Total 29.9 46
Node 10
Category % n
1.000 0.0 0
2.000 100.0 17
Total 11.0 17

Classification

Sample	Observed	Predicted 1	Predicted 2	Percent Correct
Training	1	70	1	98.6%
	2	7	68	90.7%
	Overall Percentage	52.7%	47.3%	94.5%
Test	1	79	0	100.0%
	2	12	63	84.0%
	Overall Percentage	59.1%	40.9%	92.2%

Growing Method: CRT
Dependent Variable: @1

Risk

Sample	Estimate	Std. Error
Training	.055	.019
Test	.078	.022

Growing Method: CRT
Dependent Variable: @1

2) How many nodes the final tree has and how many of them are terminal nodes;

The tree has 11 nodes with 6 of those being terminal nodes.

3) What are the most important three Lupus data features in building the tree? Explain your answer.

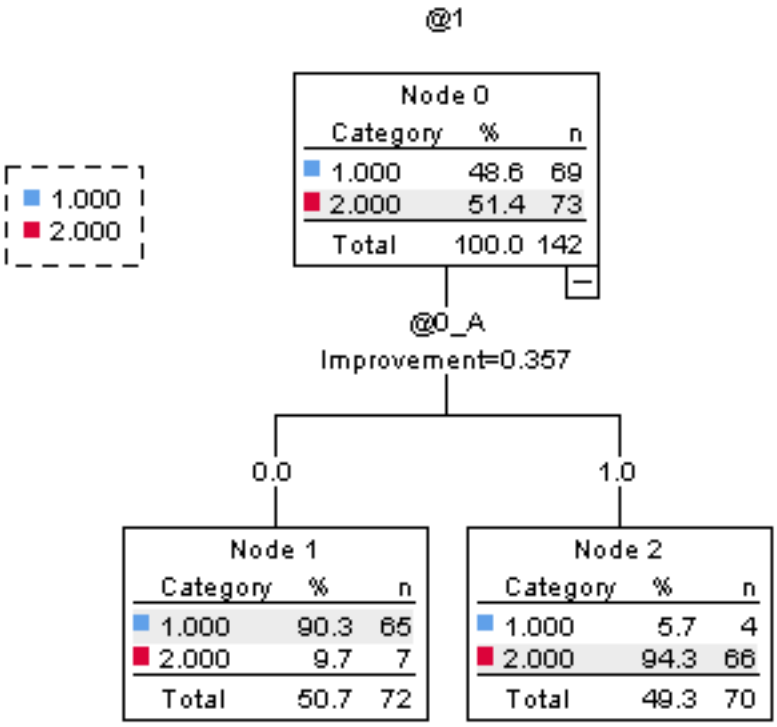
The most important features are : A, J, O as these are at the top nodes of the decision tree and a decision tree implicitly performs variable selection.

4) Increase the number of cases for each parent and child. What do you notice with the complexity of the tree? Does it increase? Explain your answer.

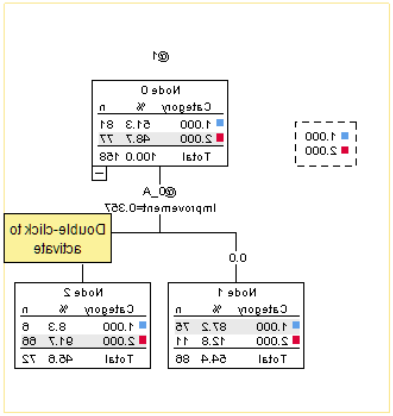
Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	@1
	Independent Variables	@0_J, @0_I, @0_H, @0_G, @0_F, @0_E, @0_D, @0_C, @0_B, @0_A, @0
	Validation	Split Sample
	Maximum Tree Depth	30
	Minimum Cases in Parent Node	20
	Minimum Cases in Child Node	10
Results	Independent Variables Included	@0_A, @0_J, @0_D, @0, @0_B, @0_H, @0_E, @0_C, @0_F, @0_I, @0_G
	Number of Nodes	3
	Number of Terminal Nodes	2
	Depth	1

Training Sample



Test Sample



Risk

Sample	Estimate	Std. Error
Training	.077	.022
Test	.108	.025

Growing Method: CRT
Dependent Variable: @1

Classification

Sample	Observed	Predicted 1	Predicted 2	Percent Correct
Training	1	65	4	94.2%
	2	7	66	90.4%
	Overall Percentage	50.7%	49.3%	92.3%
Test	1	75	6	92.6%
	2	11	66	85.7%
	Overall Percentage	54.4%	45.6%	89.2%

Growing Method: CRT
Dependent Variable: @1

The overall percentage decreased when the number of parent and child nodes were increased in this case.

Problem 2 (30 points): This problem illustrates the effect of the class imbalance of the accuracy of the decision trees. Download the red wine quality data from the UCI machine learning repository at: **http://archive.ics.uci.edu/ml/datasets/Wine+Quality**

1. **Report how many classes (treat each quality level as a different class) are and what is the distribution of these classes for the red wine data is.**

quality

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3	10	.6	.6	.6
	4	53	3.3	3.3	3.9
	5	681	42.6	42.6	46.5
	6	638	39.9	39.9	86.4
	7	199	12.4	12.4	98.9
	8	18	1.1	1.1	100.0
	Total	1599	100.0	100.0	

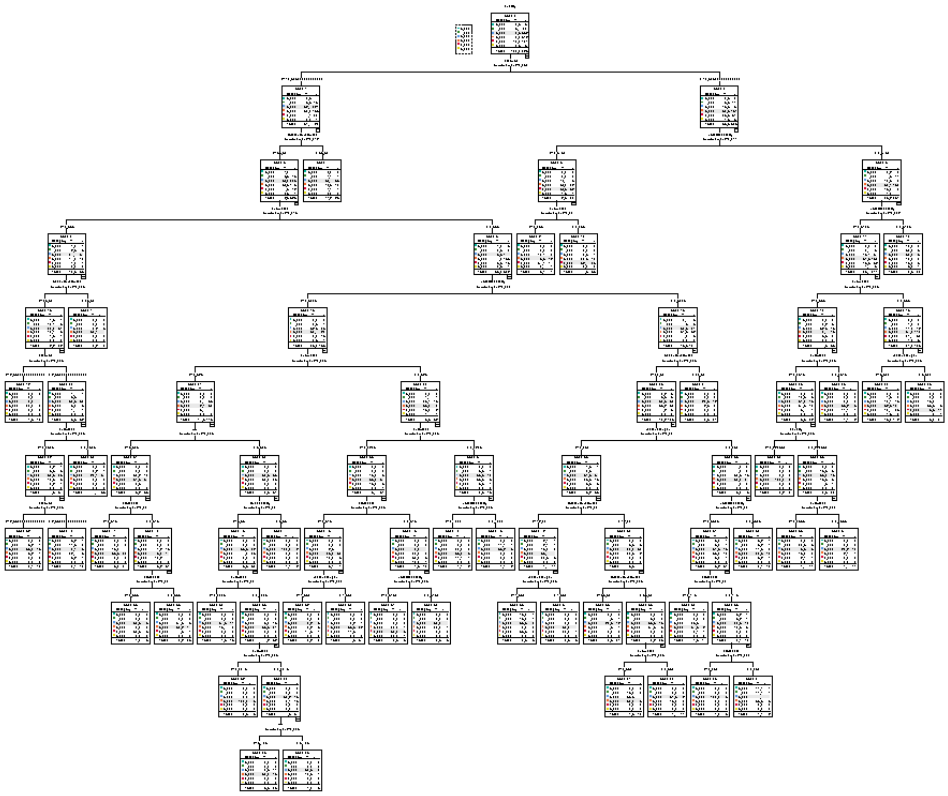
There are 6 classes that are included in the quality variable and the distribution is above.

2. **Repeat Problem 1 on the red wine data.**

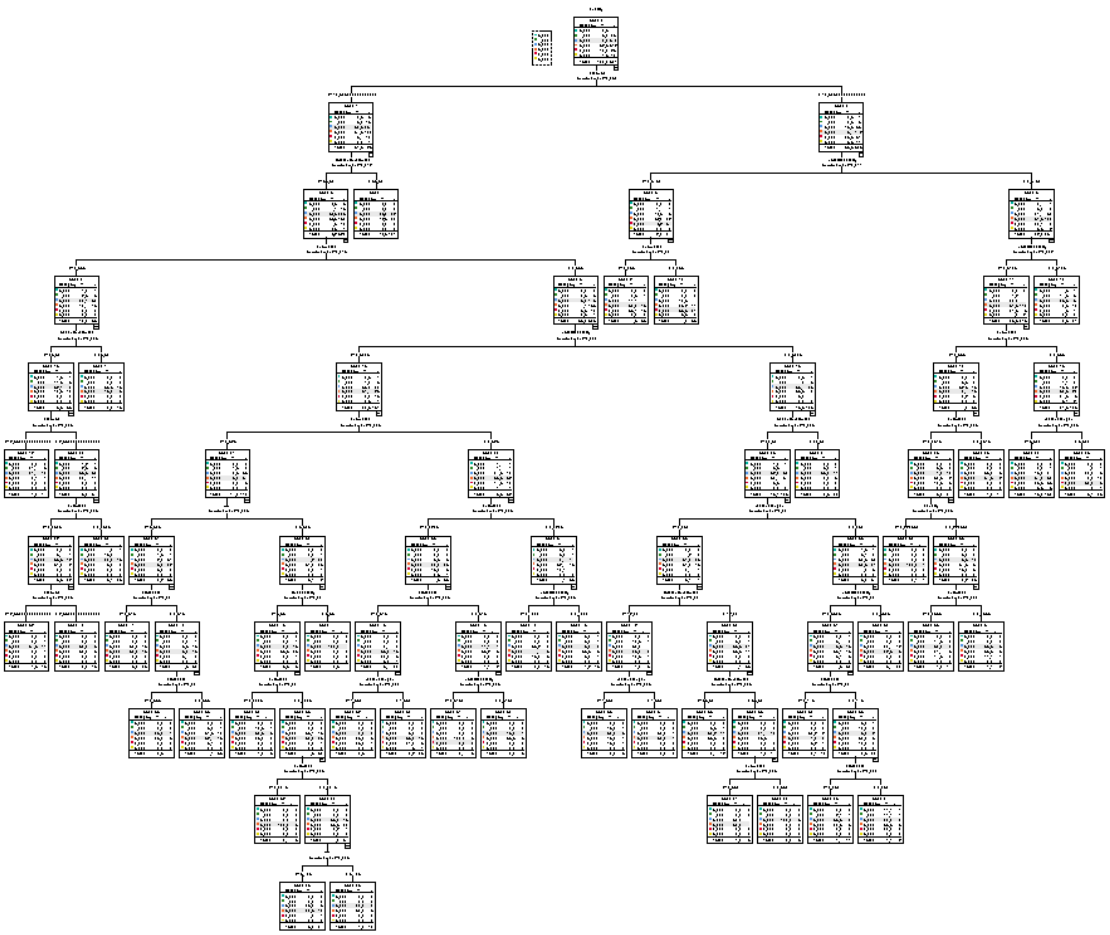
Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	quality
	Independent Variables	fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol
	Validation	Split Sample
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	alcohol, density, chlorides, volatileacidity, fixedacidity, sulphates, citricacid, totalsulfurdioxide, pH, residualsugar, freesulfurdioxide
	Number of Nodes	77
	Number of Terminal Nodes	39
	Depth	10

Training Sample



Test Sample



Risk

Sample	Estimate	Std. Error
Training	.281	.016
Test	.453	.018

Growing Method: CRT
Dependent Variable: quality

Classification

Sample	Observed	Predicted 3	4	5	6	7	8	Percent Correct
Training	3	0	0	5	1	0	0	0.0%
	4	0	0	22	5	0	0	0.0%
	5	0	0	273	61	5	0	80.5%
	6	0	0	38	262	19	0	82.1%
	7	0	0	12	50	39	0	38.6%
	8	0	0	1	3	2	0	0.0%
	Overall Percentage	0.0%	0.0%	44.0%	47.9%	8.1%	0.0%	71.9%
Test	3	0	0	4	0	0	0	0.0%
	4	0	0	22	3	1	0	0.0%
	5	0	0	207	126	9	0	60.5%
	6	0	0	93	202	24	0	63.3%
	7	0	0	9	60	29	0	29.6%
	8	0	0	0	9	3	0	0.0%
	Overall Percentage	0.0%	0.0%	41.8%	49.9%	8.2%	0.0%	54.7%

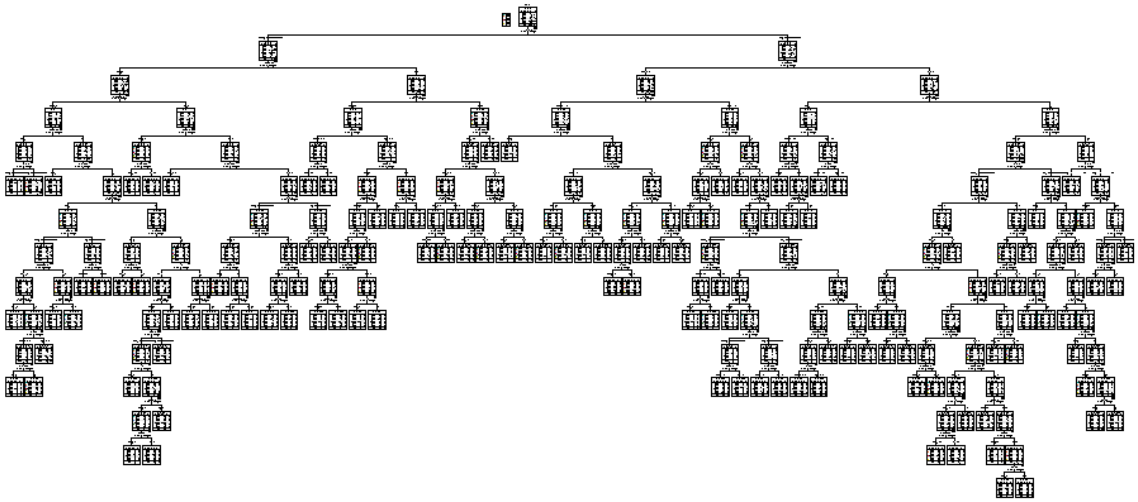
Growing Method: CRT
Dependent Variable: quality

3. Now bin the class variable in such a way that data is not so imbalanced with respect to the class variable. Repeat Problem 1 but on the wine data with less number of classes (the binned class variable).

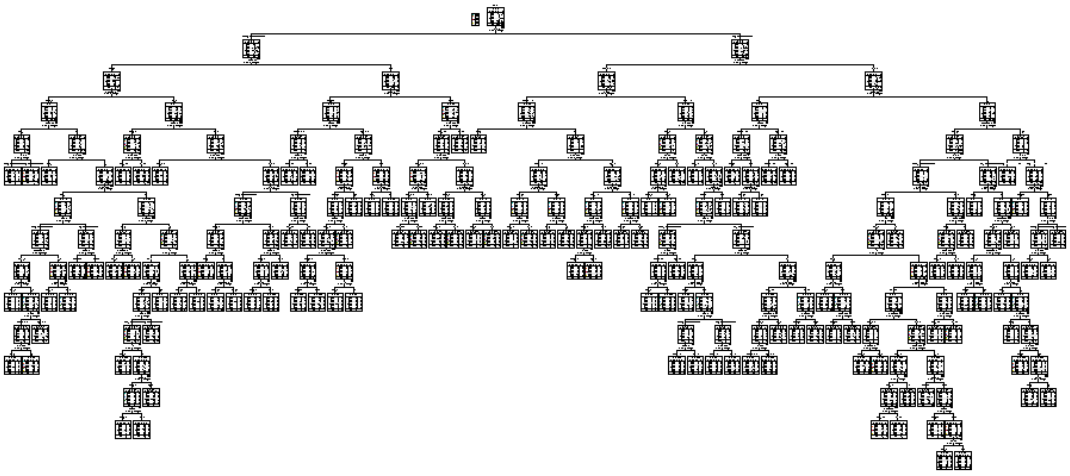
Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	quality
	Independent Variables	fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol
	Validation	Split Sample
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	5
	Minimum Cases in Child Node	2
Results	Independent Variables Included	alcohol, density, totalsulfurdioxide, chlorides, sulphates, volatileacidity, pH, residualsugar, citricacid, fixedacidity, freesulfurdioxide
	Number of Nodes	229
	Number of Terminal Nodes	115
	Depth	14

Training Sample



Test Sample



Risk

Sample	Estimate	Std. Error
Training	.118	.011
Test	.393	.017

Growing Method: CRT
Dependent Variable: quality

Sample	Observed	Predicted 3	4	5	6	7	8	Percent Correct
Training	3	0	0	3	1	0	0	0.0%
	4	0	12	7	6	3	0	42.9%
	5	0	2	293	28	1	0	90.4%
	6	0	0	21	318	3	0	93.0%
	7	0	1	3	11	87	0	85.3%
	8	0	0	1	1	3	2	28.6%
	Overall Percentage	0.0%	1.9%	40.6%	45.2%	12.0%	0.2%	88.2%
Test	3	0	0	4	2	0	0	0.0%
	4	0	2	13	8	2	0	8.0%
	5	0	9	245	92	11	0	68.6%
	6	0	5	74	190	27	0	64.2%
	7	0	1	9	43	44	0	45.4%
	8	0	0	0	6	5	0	0.0%
	Overall Percentage	0.0%	2.1%	43.6%	43.1%	11.2%	0.0%	60.7%

Growing Method: CRT
Dependent Variable: quality

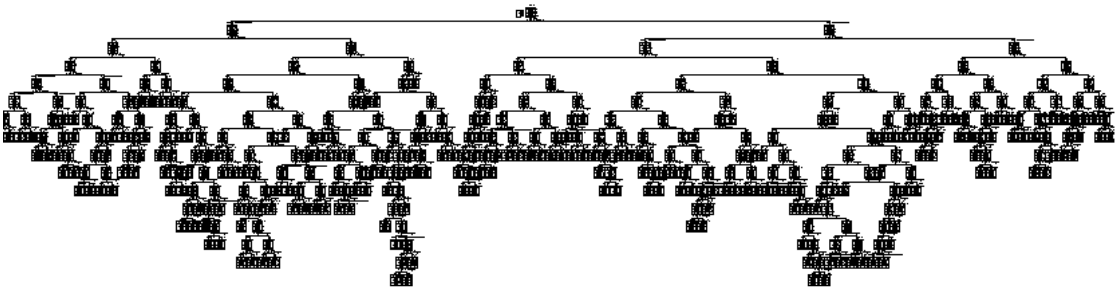
4. **How the performance of the best classification model on the original class variable compares with the accuracy of the best classification model on the binned classification variable?**

The reduced parent/child nodes end up increasing in accuracy for prediction and classification of the quality class and seen in the classification table above. (See question 3 – last table)

5. **Do you have any other ideas on how you can improve the results further?**
   Showing that your idea will actually work will be graded with five extra credit points.

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	quality
	Independent Variables	fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol
	Validation	None
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	5
	Minimum Cases in Child Node	2
Results	Independent Variables Included	alcohol, density, chlorides, volatileacidity, citricacid, sulphates, totalsulfurdioxide, fixedacidity, pH, residualsugar, freesulfurdioxide
	Number of Nodes	391
	Number of Terminal Nodes	196
	Depth	15



Risk

Estimate	Std. Error
.116	.008

Growing Method: CRT
Dependent Variable: quality

Classification

Observed	Predicted 3	4	5	6	7	8	Percent Correct
3	2	0	5	2	1	0	20.0%
4	0	22	17	14	0	0	41.5%
5	0	2	640	34	5	0	94.0%
6	0	1	51	570	14	2	89.3%
7	0	0	7	23	168	1	84.4%
8	0	0	0	3	3	12	66.7%
Overall Percentage	0.1%	1.6%	45.0%	40.4%	11.9%	0.9%	88.4%

Growing Method: CRT
Dependent Variable: quality

The results could be improved by creating a balanced dataset to pick the values train/test to remove any bias from not having a equal representation of each level. To make this happen, I re ran the classification tree on the entire set without any validation (it randomly selects) – this creates a better distribution of the wine quality rating. The overall percentage comes up to – 88.4% as seen from the table above. If compared to the training from the table from the part 2, there is a 16.5% increase and 33.7% increase from the test tree.
Another option could be to create special clusters that would further represent an equal distribution of the quality ratings before running the decision tree for the data set provided.

Problem 3 (5 points): Given the decision tree in Figure 1, show how the new examples in Table 1 would be classified by filling in the last column in the table. If an example cannot be classified, enter UNKNOWN in the last column. For each example, explain your answer by writing down the path from the root to the leaf that corresponds to that specific example.

1- Red does not show height or width in the decision tree : NO
2- Blue – Fat : Yes
3- Green – Short : No
4- Green – Tall : Yes
5- Blue does not show short or thin : No

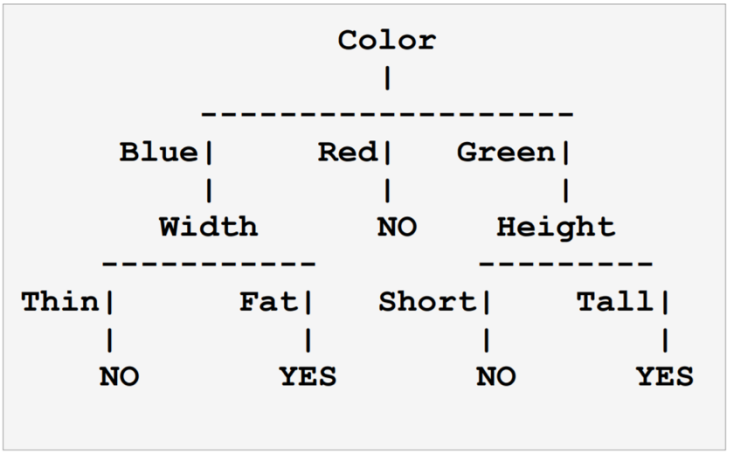


Figure 1: Decision tree

Table 1: Data for Problem #3

Example	Color	Height	Width	Class
A	Red	Short	Thin	**No**
B	Blue	Tall	Fat	**Yes**
C	Green	Short	Fat	**No**
D	Green	Tall	Thin	**Yes**
E	Blue	Short	Thin	**No**

Submission Instructions

1. Answer the problems and write your answers in a Word document.
2. Submit your file online at the website at **http://d2l.depaul.edu** and check your submission
3. Keep a copy of all your submissions!
4. If you have questions about the homework, email me BEFORE the deadline.
5. Late submissions are allowed with a 5%, 10%, and 15% penalty for a one day, two days, and three days, respectively.
6. No late work will be accepted after three days since the assignment was due.