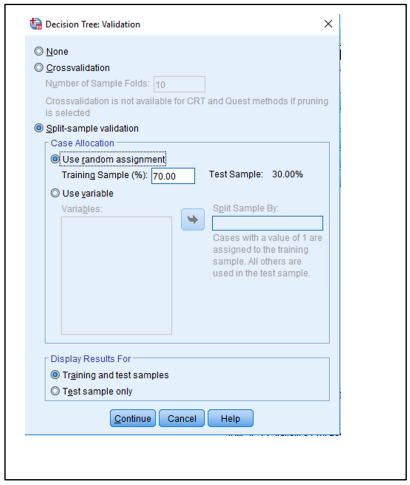
1.Problem

1)

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	V12
	Independent Variables	V1, V2, V3, V4, V5, V6, V7, V8, V9, V10, V11
	Validation	Split Sample
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	V10, V1, V9, V7, V11, V6, V3, V8, V5, V2, V4
	Number of Nodes	13
	Number of Terminal Nodes	7
	Depth	4



Risk

Sample	Estimate	Std. Error
Training	.043	.014
Test	.056	.024

Growing Method: CRT
Dependent Variable: V12

Classification

		Predicted		
Sample	Observed	1	2	Percent Correct
Training	1	104	4	96.3%
	2	5	98	95.1%
	Overall Percentage	51.7%	48.3%	95.7%
Test	1	37	5	88.1%
	2	0	47	100.0%
	Overall Percentage	41.6%	58.4%	94.4%

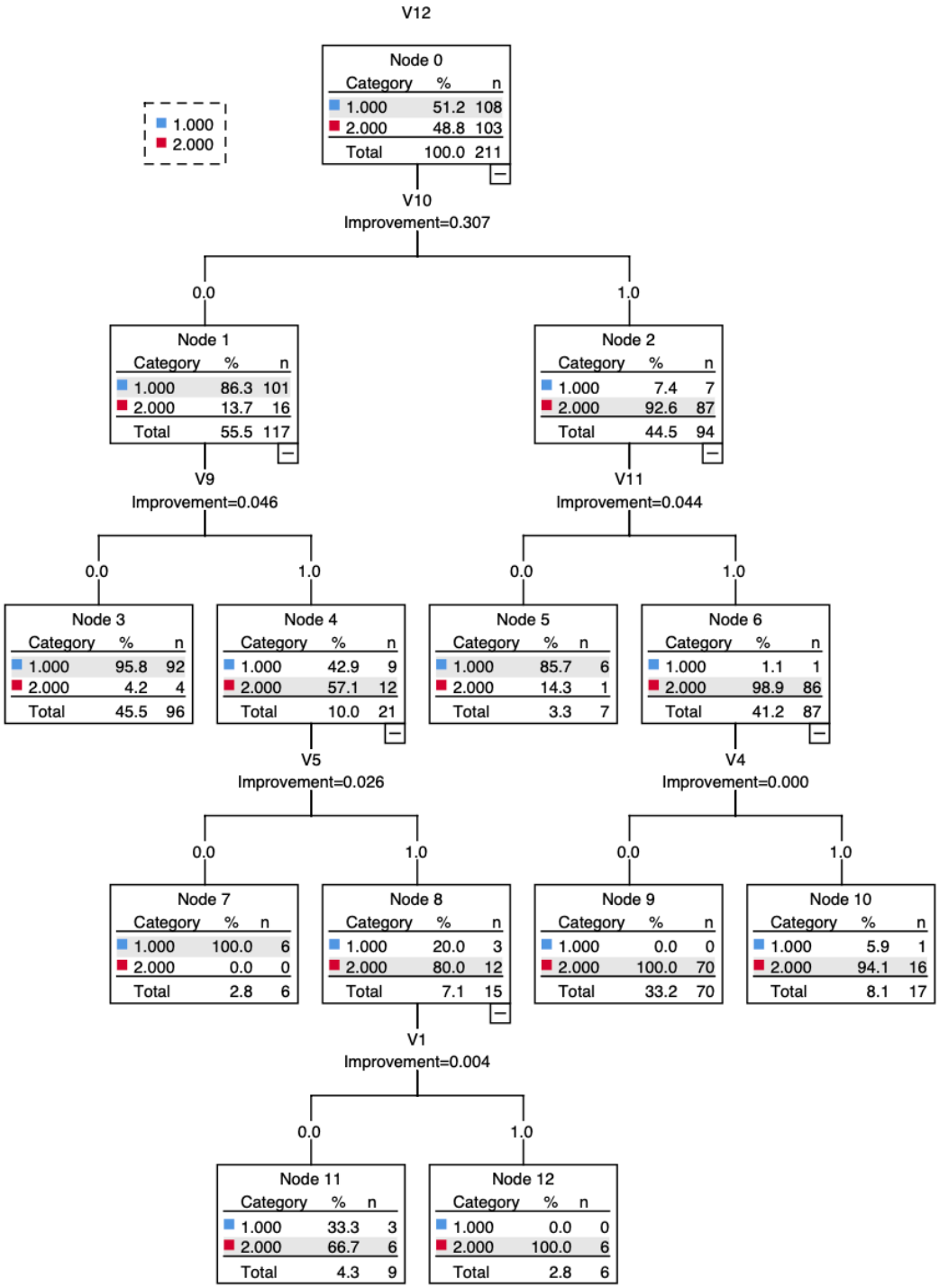
Growing Method: CRT
Dependent Variable: V12

Independent Variable Importance

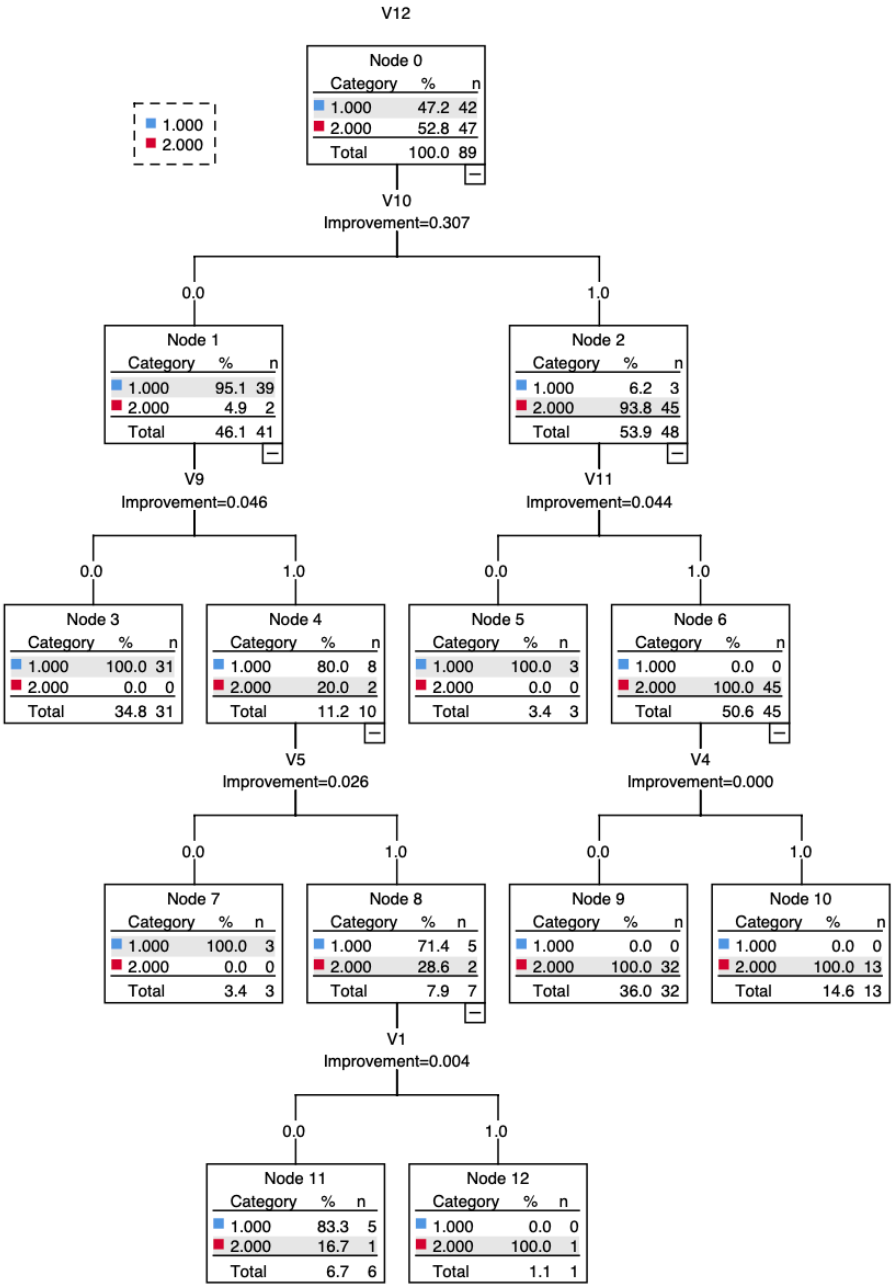
Independent Variable	Importance	Normalized Importance
V10	.307	100.0%
V11	.220	71.4%
V9	.204	66.3%
V1	.197	64.0%
V7	.126	41.0%
V6	.100	32.5%
V5	.079	25.7%
V8	.072	23.5%
V3	.062	20.2%
V2	.027	8.9%
V4	.023	7.4%

Growing Method: CRT
Dependent Variable: V12

Training Set:



Test set:


V12
1.000
2.000
Node 0
Category % n
1.000 47.2 42
2.000 52.8 47
Total 100.0 89
V10
Improvement=0.307
0.0
1.0
Node 1
Category % n
1.000 95.1 39
2.000 4.9 2
Total 46.1 41
Node 2
Category % n
1.000 6.2 3
2.000 93.8 45
Total 53.9 48
V9
Improvement=0.046
V11
Improvement=0.044
Node 3
Category % n
1.000 100.0 31
2.000 0.0 0
Total 34.8 31
Node 4
Category % n
1.000 80.0 8
2.000 20.0 2
Total 11.2 10
Node 5
Category % n
1.000 100.0 3
2.000 0.0 0
Total 3.4 3
Node 6
Category % n
1.000 0.0 0
2.000 100.0 45
Total 50.6 45
V5
Improvement=0.026
V4
Improvement=0.000
Node 7
Category % n
1.000 100.0 3
2.000 0.0 0
Total 3.4 3
Node 8
Category % n
1.000 71.4 5
2.000 28.6 2
Total 7.9 7
Node 9
Category % n
1.000 0.0 0
2.000 100.0 32
Total 36.0 32
Node 10
Category % n
1.000 0.0 0
2.000 100.0 13
Total 14.6 13
V1
Improvement=0.004
Node 11
Category % n
1.000 83.3 5
2.000 16.7 1
Total 6.7 6
Node 12
Category % n
1.000 0.0 0
2.000 100.0 1
Total 1.1 1

The data is divided into test and train data 70% and 30 % respectively.
The overall correct percentage of the model reach 94.4%.
The model error rate is 0.043 for training set and 0.056 for test data set.

2)

The final tree has 13 nodes and 7 terminal nodes

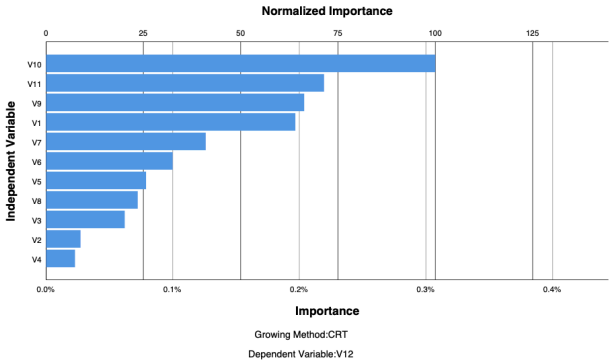
3)

Top 3 importance

Independent Variable Importance

Independent Variable	Importance	Normalized Importance
V10	.307	100.0%
V11	.220	71.4%
V9	.204	66.3%
V1	.197	64.0%
V7	.126	41.0%
V6	.100	32.5%
V5	.079	25.7%
V8	.072	23.5%
V3	.062	20.2%
V2	.027	8.9%
V4	.023	7.4%

Growing Method: CRT
Dependent Variable: V12



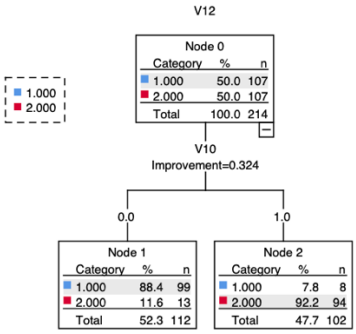
The three most important data features in building the tree are respectively V10, V11 and V9.

4)

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	V12
	Independent Variables	V1, V2, V3, V4, V5, V6, V7, V8, V9, V10, V11
	Validation	Split Sample
	Maximum Tree Depth	5
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	V10, V7, V9, V1, V11, V6, V3, V8, V4, V2, V5
	Number of Nodes	3
	Number of Terminal Nodes	2
	Depth	1

Train dataset

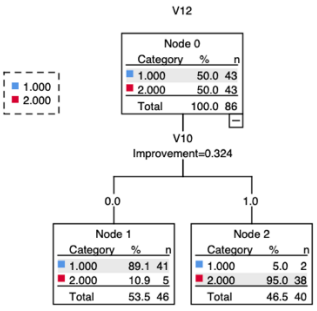


Risk

Sample	Estimate	Std. Error
Training	.098	.020
Test	.081	.029

Growing Method: CRT
Dependent Variable: V12

Test dataset



Classification

Sample	Observed	Predicted 1	Predicted 2	Percent Correct
Training	1	99	8	92.5%
	2	13	94	87.9%
	Overall Percentage	52.3%	47.7%	90.2%
Test	1	41	2	95.3%
	2	5	38	88.4%
	Overall Percentage	53.5%	46.5%	91.9%

Growing Method: CRT
Dependent Variable: V12

When we increase the parameters of cases allowed in parent and child nodes, the complexity given by the number of nodes reduce. By increasing the minimum of cases, in the nodes, we give more tolerance in the decision tree and therefore remove complexity.

Independent Variable Importance

Independent Variable	Importance	Normalized Importance
V10	.324	100.0%
V11	.191	59.0%
V9	.148	45.8%
V1	.148	45.6%
V7	.147	45.4%
V6	.093	28.9%
V3	.075	23.2%
V8	.054	16.8%
V4	.039	12.0%
V5	.037	11.3%
V2	.030	9.4%

Growing Method: CRT
Dependent Variable: V12

2. Problem

1. The red wine data contains of 6 classes [quality level 3 to 8] .

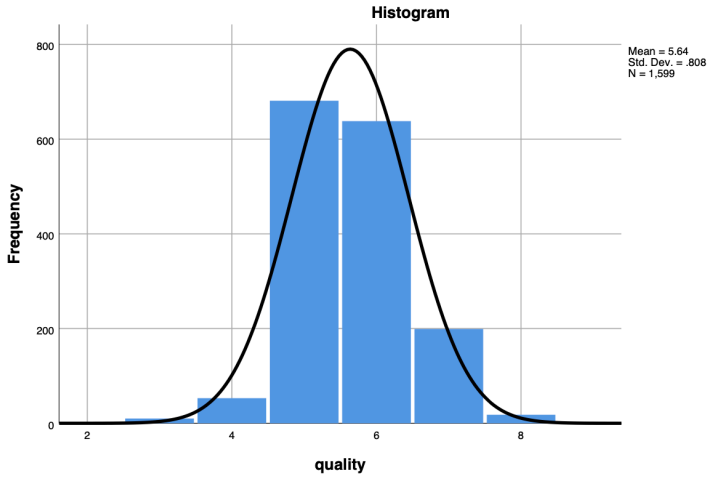
   The classes 5 and 6 have a very high frequency so the data is imbalanced.

   Mean of distribution – 5.64

   Std deviation – 0.808

quality

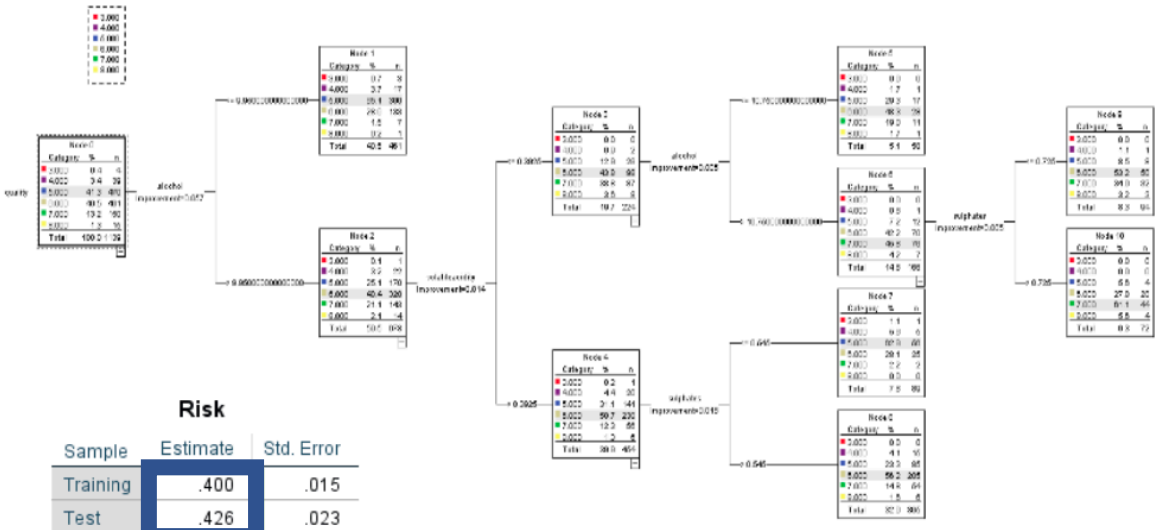
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3	10	.6	.6	.6
	4	53	3.3	3.3	3.9
	5	681	42.6	42.6	46.5
	6	638	39.9	39.9	86.4
	7	199	12.4	12.4	98.9
	8	18	1.1	1.1	100.0
	Total	1599	100.0	100.0	



2)

Model Summary

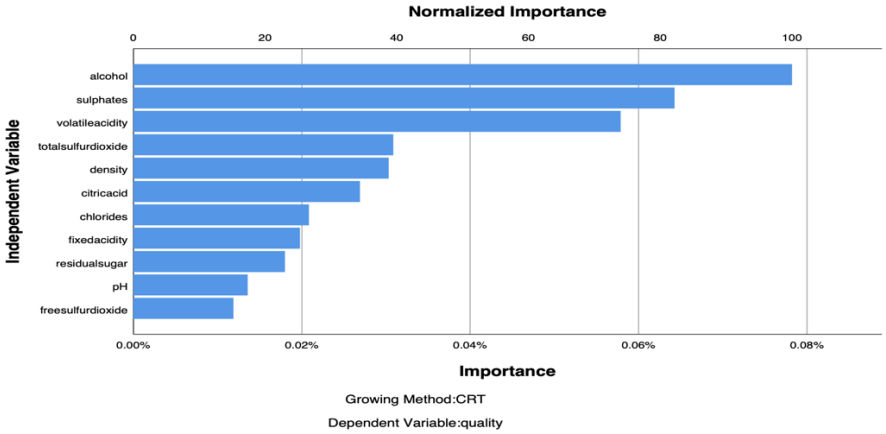
Specifications	Growing Method	CRT
	Dependent Variable	quality
	Independent Variables	fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol
	Validation	Split Sample
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	alcohol, totalsulfurdioxide, density, sulphates, chlorides, pH, residualsugar, freesulfurdioxide, citricacid, volatileacidity, fixedacidity
	Number of Nodes	11
	Number of Terminal Nodes	6
	Depth	4



Risk

Sample	Estimate	Std. Error
Training	.400	.015
Test	.426	.023

Growing Method: CRT
Dependent Variable: quality



Growing Method:CRT
Dependent Variable:quality

2.1) Training data - 70% of data using CART
Gini index - 0.0001 for impurity threshold
For the stopping condition, the minimum number of cases for parent and child nodes are 100 and 50.
The maximum depth is 20.
Risk rate for training set is 0.400 and for test set is 0.426.

2.2) The final tree has 11 nodes, in which there are 6 terminal nodes.

2.3) The most important three features are ranked as following:
1- alcohol
2-sulphates
3- volatileacidity

Independent Variable Importance

Independent Variable	Importance	Normalized Importance
alcohol	.078	100.0%
sulphates	.064	82.2%
volatileacidity	.058	74.0%
totalsulfurdioxide	.031	39.5%
density	.030	38.8%
citricacid	.027	34.4%
chlorides	.021	26.6%
fixedacidity	.020	25.3%
residualsugar	.018	23.0%
pH	.014	17.3%
freesulfurdioxide	.012	15.2%

Growing Method: CRT
Dependent Variable: quality

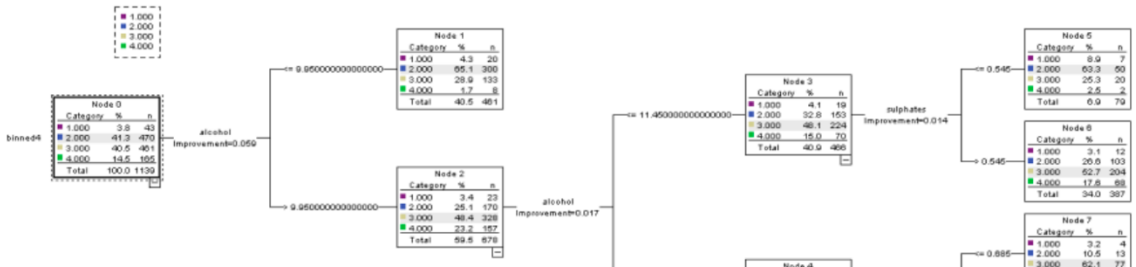
2.4) when we increase the parameters for the number of parent cases from 100 to 200 and child cases from 50 to 100, the new tree is smaller than the final model. When the parameters parent and child cases are increased the nodes of tree will be decreased this is the algorithm will stop after the cases becomes fewer than the threshold. Thus, the larger threshold makes a smaller tree.

3.1)

The original class has the level 3 to 8.
The class variable are binned into 4 bins
classes 3,4 for bin 1
class 5 for bin 2,
class 6 for bin 3
classes 7,8 for bin 4.

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	binned4
	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	100
	Minimum Cases in Child Node	50
Results	Independent Variables Included	alcohol, totalsulfurdioxide, density, sulphates, chlorides, pH, residualsugar, freesulfurdioxide, citricacid, fixedacidity, volatileacidity
	Number of Nodes	9
	Number of Terminal Nodes	5
	Depth	3



Risk

Sample	Estimate	Std. Error
Training	.396	.014
Test	.409	.023

Growing Method: CRT
Dependent Variable: binned4

Gini index - 0.0001 for impurity threshold
Test data = 70%
Minimum number of cases for parent and child nodes are 100 and 50The maximum depth is 20.
The final tree has an accuracy rate of 73.7%.

3.2) The final tree has 9 nodes, and 5 terminal nodes.

3.3) The most important three features are ranked as following:
1- alcohol
2-sulphates
3- volatileacidity

Independent Variable Importance

Independent Variable	Importance	Normalized Importance
alcohol	.078	100.0%
sulphates	.064	82.2%
volatileacidity	.058	74.0%
totalsulfurdioxide	.031	39.5%
density	.030	38.8%
citricacid	.027	34.4%
chlorides	.021	26.6%
fixedacidity	.020	25.3%
residualsugar	.018	23.0%
pH	.014	17.3%
freesulfurdioxide	.012	15.2%

Growing Method: CRT
Dependent Variable: quality

3.4) When the parameters parent and child cases are increased the nodes of tree will be decreased this is the algorithm will stop after the cases becomes fewer than the threshold. Thus, the larger threshold makes a smaller tree.

4. The original class variable has a poor performance because it has a class imbalance.
The performance of binned variable is better than the original variable because the binned class is more balanced than the original variable.
The overall accuracy rate of the binned class from test set is 59.1% which is higher than the rate of 57.4% from the original class.

Original Data

		Predicted						
Sample	Observed	3	4	5	6	7	8	Percent Correct
Training	3	0	0	4	0	0	0	0.0%
	4	0	0	22	17	0	0	0.0%
	5	0	0	356	110	4	0	75.7%
	6	0	0	158	283	20	0	61.4%
	7	0	0	9	97	44	0	29.3%
	8	0	0	1	10	4	0	0.0%
	Overall Percentage	0.0%	0.0%	48.3%	45.4%	6.3%	0.0%	60.0%
Test	3	0	0	6	0	0	0	0.0%
	4	0	0	9	5	0	0	0.0%
	5	0	0	158	52	1	0	74.9%
	6	0	0	69	92	16	0	52.0%
	7	0	0	5	30	14	0	28.6%
	8	0	0	0	3	0	0	0.0%
	Overall Percentage	0.0%	0.0%	53.7%	39.6%	6.7%	0.0%	57.4%

Binned Data

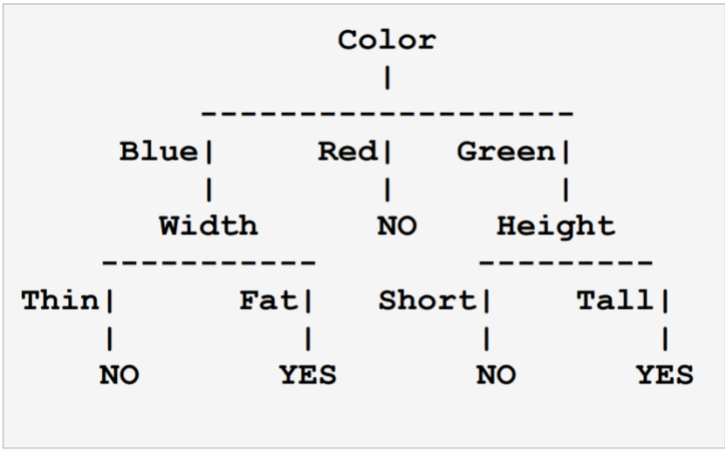
		Predicted				
Sample	Observed	1.00	2.00	3.00	4.00	Percent Correct
Training	1.00	0	27	16	0	0.0%
	2.00	0	350	116	4	74.5%
	3.00	0	153	281	27	61.0%
	4.00	0	10	98	57	34.5%
	Overall Percentage	0.0%	47.4%	44.9%	7.7%	60.4%
Test	1.00	0	15	5	0	0.0%
	2.00	0	159	52	0	75.4%
	3.00	0	71	91	15	51.4%
	4.00	0	5	25	22	42.3%
	Overall Percentage	0.0%	54.3%	37.6%	8.0%	59.1%

5.

Ways to improve the results further

1. We can further bin the class variable into 2 bins so that we can improve our model.
2. Features selection could be used and therefore, make the dataset lighter and easier to compute, to understand and analyze.
3. To improve the accuracy, the user could use sensitivity and specificity since the distribution of the data set is strongly centered.
4. To improve the quality of the results, would be to use other validation techniques such as Bootstrap or cross-validation techniques.

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.



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