Section a)

Following is the class distribution for the diagnosis column (Last column).

Value 1 has 150 rows
Value 2 has 150 rows

Frequencies

Statistics

D

N	Valid	300
	Missing	0

D

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	150	50.0	50.0	50.0
	2	150	50.0	50.0	100.0
	Total	300	100.0	100.0	

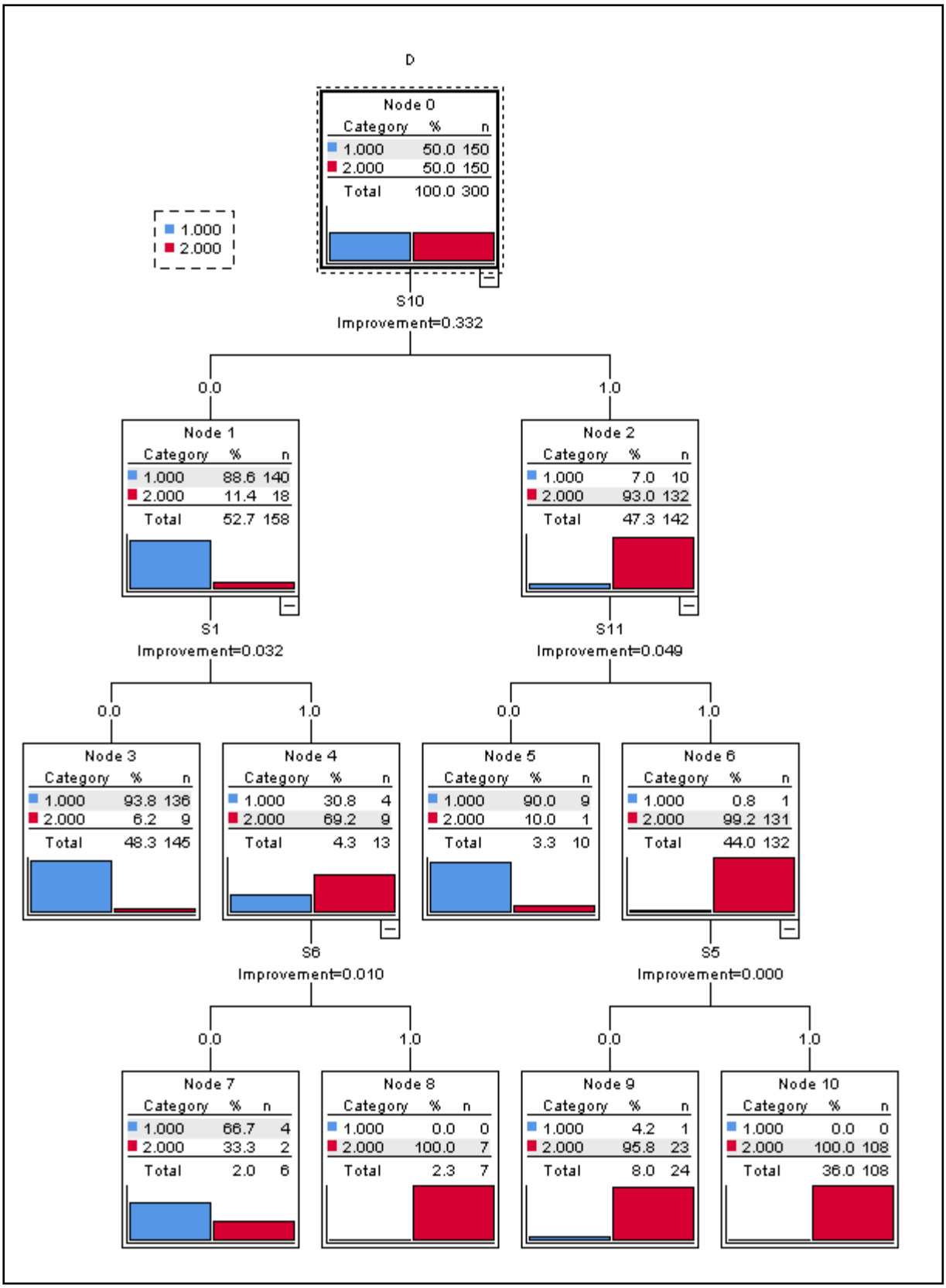
When Parent Node = 10, Child Node = 5

The model summary shows all the variables were included in the tree.

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	D
	Independent Variables	S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11
	Validation	Cross Validation
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	S10, S7, S1, S9, S11, S6, S3, S8, S5, S4, S2
	Number of Nodes	11
	Number of Terminal Nodes	6
	Depth	3

Decision Tree



There are 11 nodes for this particular tree out of which 6 are terminal nodes. The root node shows there are 150 tuples in category 1 and 150 tuples in category 2.

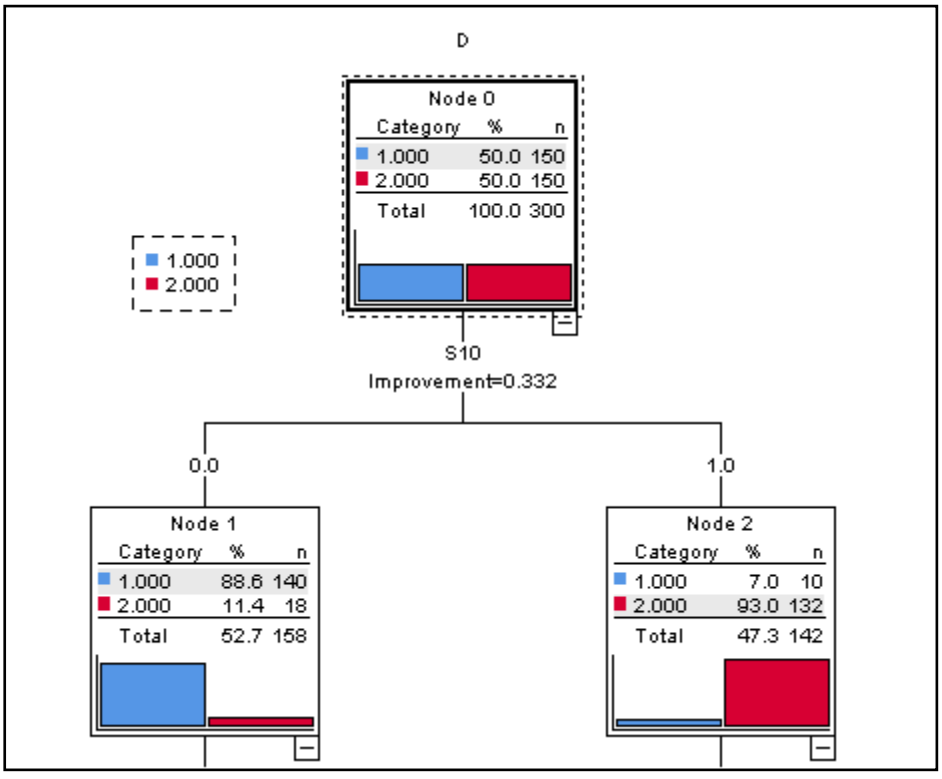
As can be seen, the first node is split based on our most important predictor, S10. The value of impurity is calculated based on Gini with S10 node having a value of 0.332.
The variable with highest reduction of impurity is selected as splitting attribute.

S10: Number of cases
S10 node has 158 tuples with value 0 and 142 records with value 1.

S10

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	158	52.7	52.7	52.7
	1	142	47.3	47.3	100.0
	Total	300	100.0	100.0	

Since S10 consists of binary value and is split into two 1 and 0.



<u>S10:</u> 0 value

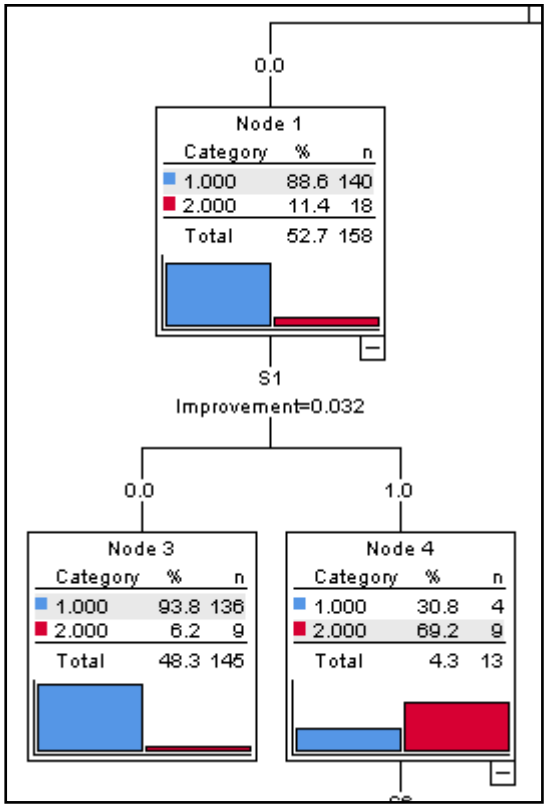
Based on the Gini index value, S1 is the most impure variable with a value of 0.032.

S1: Number of cases
S1 node has 215 tuples with value 0 and 85 records with value 1.

S1

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	215	71.7	71.7	71.7
	1	85	28.3	28.3	100.0
	Total	300	100.0	100.0	

S1 is further split into 0 and 1.



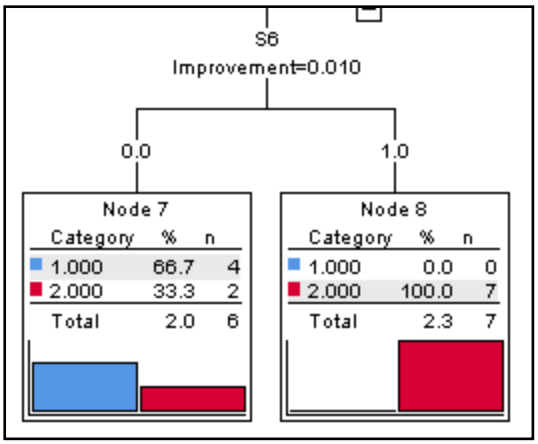
S1: 0 value
S1 with 0 value is a homogenous node with no further split. Therefore it is a leaf node. Most of the records belong to class 1.

S1: 1 value
S1 node with 1 value if further split into node S6. S6 column having a Gini index of 0.010

S6: Number of cases
S6 node has 195 tuples with value 0 and 105 records with value 1.

S6

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	195	65.0	65.0	65.0
	1	105	35.0	35.0	100.0
	Total	300	100.0	100.0	



S6: 0 value
S6 with 0 value is a homogenous node with no further split. Therefore it is a leaf node.

S6: 1 value
S6 with 1 value is a homogenous node with no further split. Therefore it is a leaf node. All the records belong to class 2.

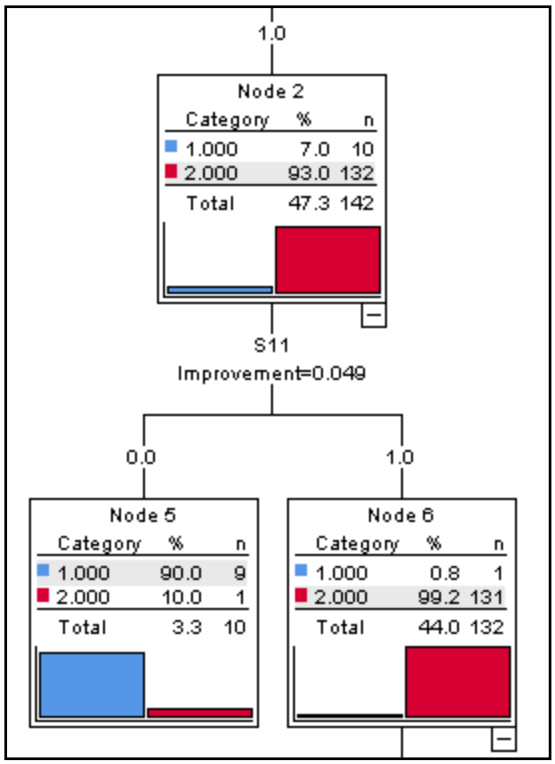
S10: 1 value

Based on the Gini index value, S11 is the most impure variable with a value of 0.049.

S11: Number of cases
S11 node has 81 tuples with value 0 and 219 records with value 1.

S11

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	81	27.0	27.0	27.0
	1	219	73.0	73.0	100.0
	Total	300	100.0	100.0	



S11 is further split into 0 and 1.

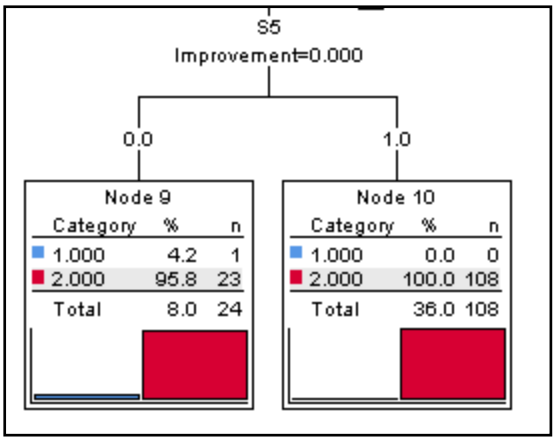
S11: 0 value
S11 with 0 value is a homogenous node with no further split. Therefore it is a leaf node. Most of the records belong to class 1.

S11: 1 value
S11 node with 1 value if further split into node S5.

S5: Number of cases
S5 node has 92 tuples with value 0 and 208 records with value 1.

S5

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0	92	30.7	30.7	30.7
	1	208	69.3	69.3	100.0
	Total	300	100.0	100.0	



S5: 0 value
S5 with 0 value is a homogenous node with no further split. Therefore it is a leaf node.

S5: 1 value
S5 with 1 value is a homogenous node with no further split. Therefore it is a leaf node. All the records belong to class 2.

The stopping criteria is when a terminal node in which all cases have the same value for the dependent variable is a homogenous node that requires no further splitting because it is "pure".

A confusion matrix is used to compare model predictions to determine the effectiveness of a classification model. Thus, it can be used to evaluate the performance of supervised classification processes.
The confusion matrix shows an accuracy rate of 95.7%.

Classification

Observed	Predicted 1	Predicted 2	Percent Correct
1	149	1	99.3%
2	12	138	92.0%
Overall Percentage	53.7%	46.3%	95.7%

Growing Method: CRT
Dependent Variable: D

Section b)

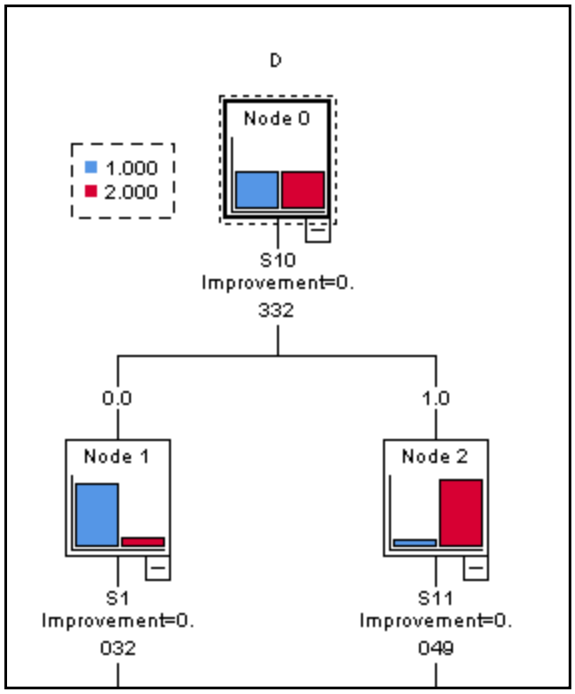
There are 11 nodes for this particular tree out of which 6 are terminal nodes. The root node shows there are 150 tuples in category 1 and 150 tuples in category 2.

As can be seen, the first node is split based on our most important predictor, S10. The value of impurity is calculated based on Gini with S10 node having a value of 0.332.

Section c)

Lupus is autoimmune heterogeneous disease and also a multi system disorder which predominantly affects women.

The three important features for this dataset are S10, S1 and S11.



The attribute with the highest reduction of impurity is selected as the splitting attribute. For this particular dataset, S10 is the most impure node at 0.332.
Variable S1 and S11 are the next most import attribute with highest reduction of impurity value at 0.032 and 0.049 respectively.

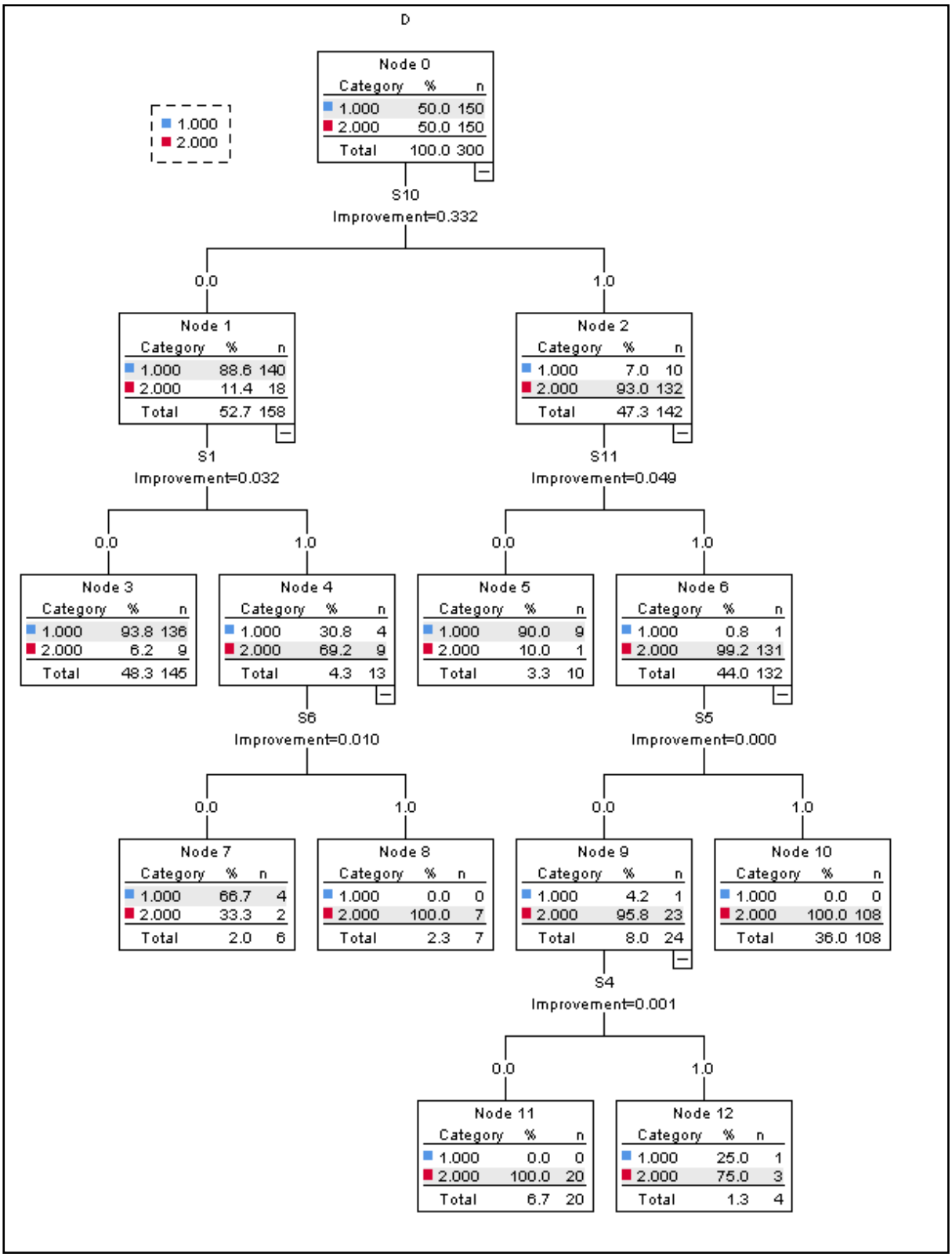
Section d)

As shown in Section a), **when Parent Node = 10, Child Node = 5**
The complexity of decision tree was 11 nodes with 6 terminal nodes. The depth of the tree is 3.

When Parent Node = 8, Child Node = 4
The complexity of decision tree is reduced. The decision tree has 13 nodes, out of which 7 are terminal nodes. The depth of the tree is 4.
The accuracy rate is 95.7%.

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	D
	Independent Variables	S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11
	Validation	Cross Validation
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	8
	Minimum Cases in Child Node	4
Results	Independent Variables Included	S10, S7, S1, S9, S11, S6, S3, S8, S5, S4, S2
	Number of Nodes	13
	Number of Terminal Nodes	7
	Depth	4



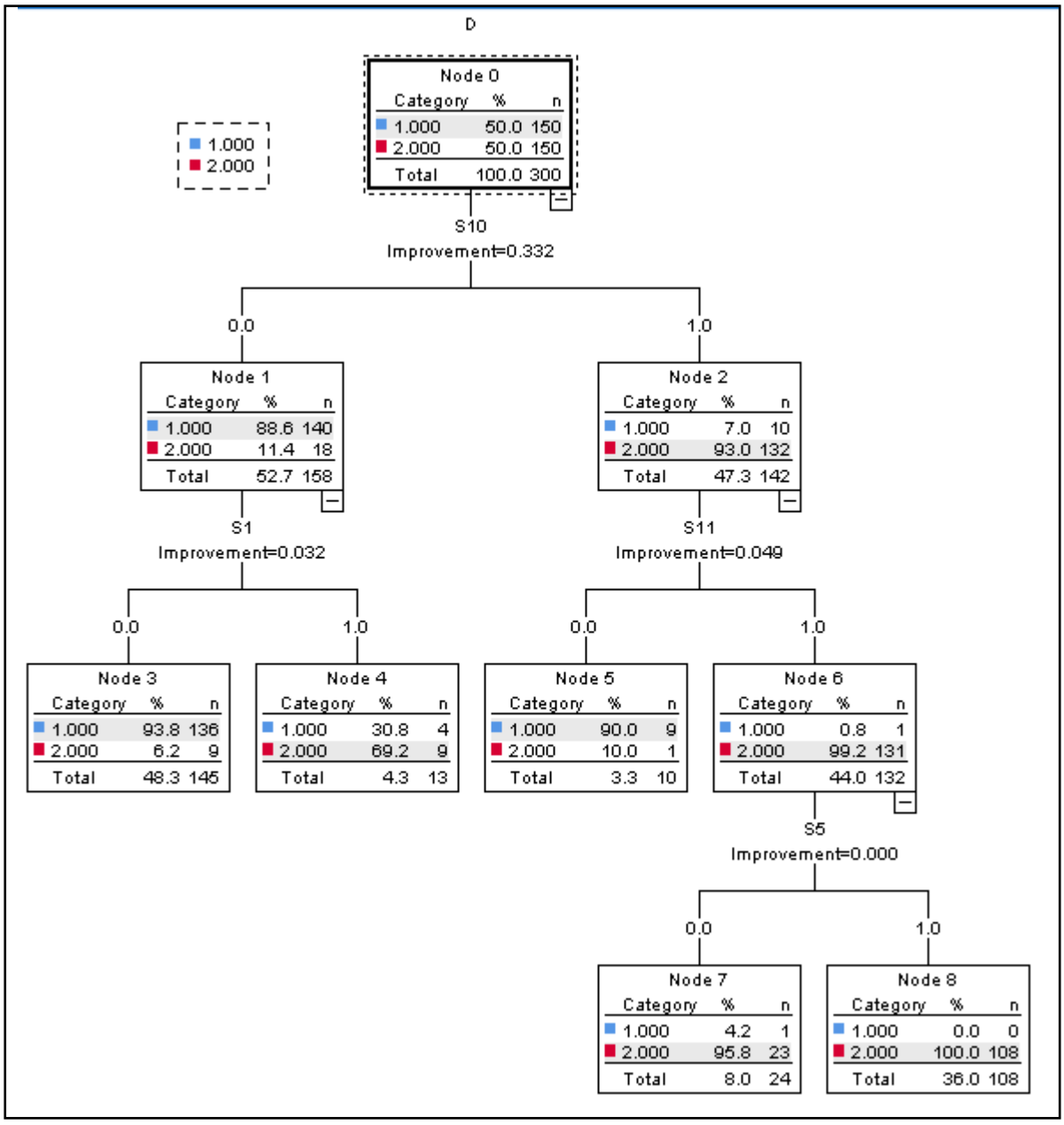
When Parent Node = 14, Child Node = 7

The complexity of decision tree is reduced. The decision tree has 9 nodes, out of which 5 are terminal nodes. The depth of the tree is 3.

The accuracy rate is 95%.

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	D
	Independent Variables	S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11
	Validation	Cross Validation
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	14
	Minimum Cases in Child Node	7
Results	Independent Variables Included	S10, S7, S1, S9, S11, S6, S3, S8, S5, S4, S2
	Number of Nodes	9
	Number of Terminal Nodes	5
	Depth	3



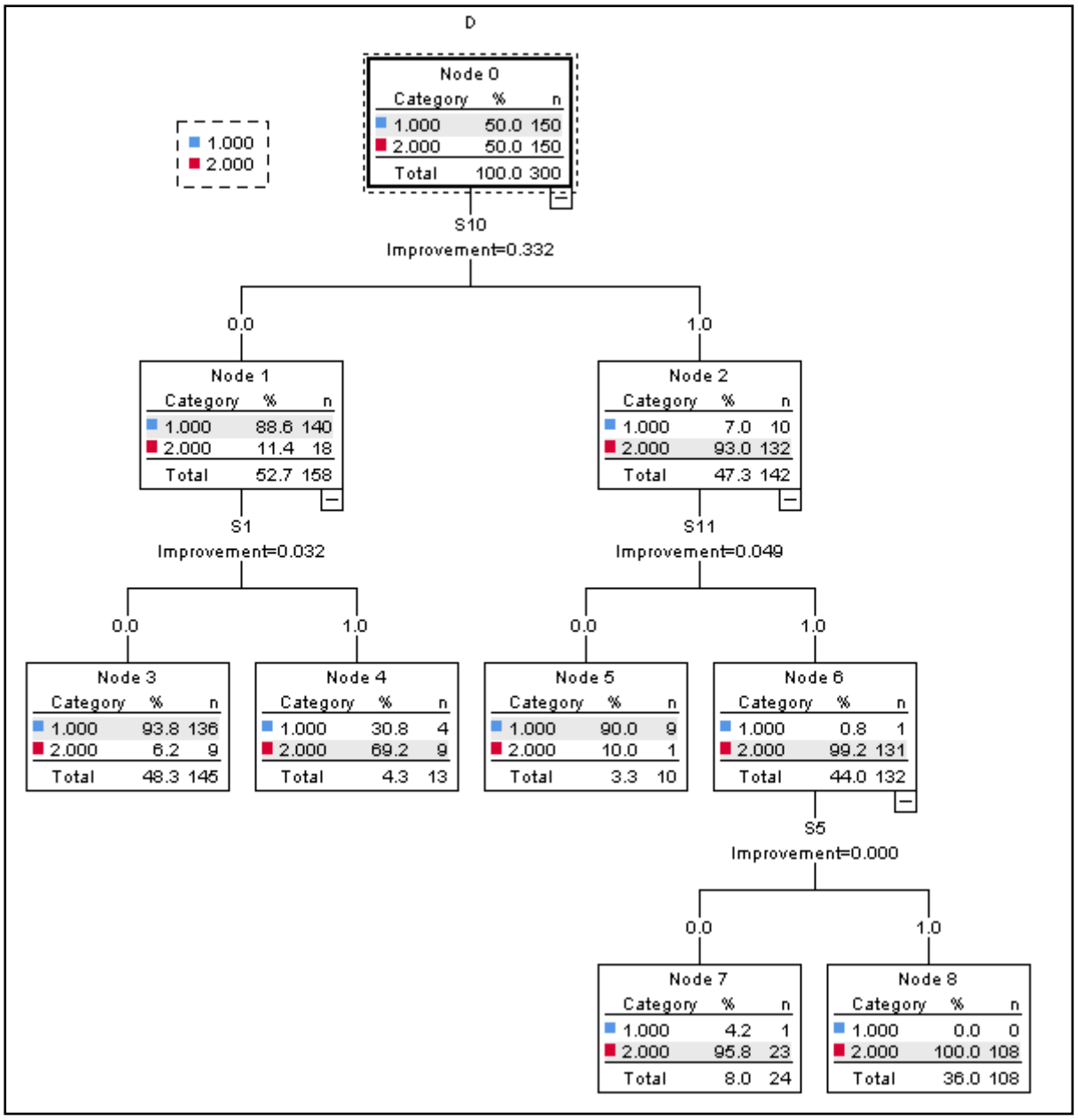
When Parent Node = 18, Child Node = 9

The complexity of decision tree is reduced. The decision tree has 9 nodes, out of which 5 are terminal nodes. The depth of the tree is 3.

The accuracy rate is 95%.

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	D
	Independent Variables	S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11
	Validation	Cross Validation
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	18
	Minimum Cases in Child Node	9
Results	Independent Variables Included	S10, S7, S1, S9, S11, S6, S3, S8, S5, S4, S2
	Number of Nodes	9
	Number of Terminal Nodes	5
	Depth	3



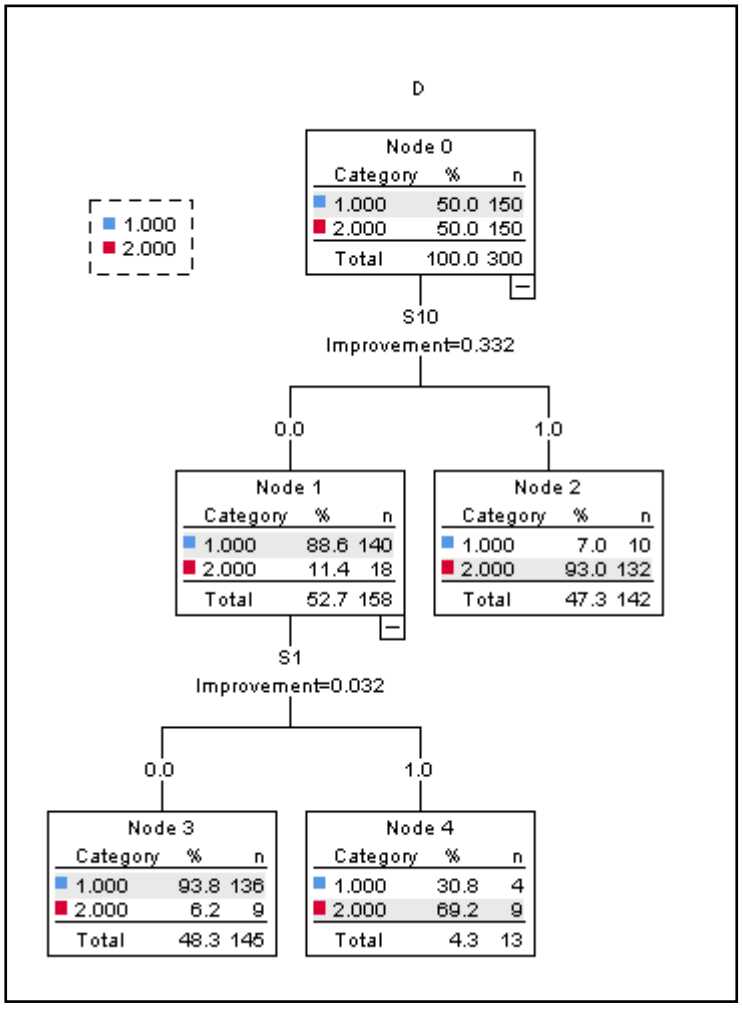
When Parent Node = 24, Child Node = 12

The complexity of decision tree is reduced. The decision tree has 9 nodes, out of which 5 are terminal nodes. The depth of the tree is 3.

The accuracy rate is reduced to 92.3%

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	D
	Independent Variables	S1, S2, S3, S4, S5, S6, S7, S8, S9, S10, S11
	Validation	Cross Validation
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	24
	Minimum Cases in Child Node	12
Results	Independent Variables Included	S10, S7, S1, S9, S11, S6, S3, S8, S5, S4, S2
	Number of Nodes	5
	Number of Terminal Nodes	3
	Depth	2



Conclusion: As we increase the count of parent nodes and child nodes, the complexity of the decision tree decreases but the accuracy rate also decreases.
The complexity is least when parent node = 24, child node = 12. At the same time accuracy rate is least at 92.3%
The complexity is high when parent node = 8, child node = 4. The accuracy rate is best at 95.7%

Problem 2

Section a)

Statistics

quality

N	Valid	1599
	Missing	0

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	

Total number of records: 1599
Number of cases: 6
Value 3 has least number of records at 10, while value 4 has 53 records.
Maximum numbers of records are concentrated in between 5 and 6 at 681 and 638 respectively.
Value 7 has 199 records and value 8 has 18 records.

Section b)

When Parent Node = 10, Child Node = 5
The model summary shows all the variables were included in the tree.

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	quality
	Independent Variables	fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol
	Validation	Cross Validation
	Maximum Tree Depth	25
	Minimum Cases in Parent Node	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	alcohol, density, chlorides, volatileacidity, citricacid, sulphates, totalsulfurdioxide, fixedacidity, pH, residualsugar, freesulfurdioxide
	Number of Nodes	123
	Number of Terminal Nodes	62
	Depth	11

The decision tree for Red Wine Quality data is quite complex.
There are total of 123 nodes in this decision tree and 62 terminal nodes. The depth of this tree is 11.
As can be seen, the first node is split based on our most important predictor, Alcohol. The value of impurity is calculated based on Gini with Alcohol node having a value of 0.059.
The variable with highest reduction of impurity is selected as splitting attribute.

problem_2.xml

The splitting value for alcohol variable is <= 10.25 and > 10.25.

Since the class is poorly balanced most of the prediction is concentrated on value 5 and value 6.
The accuracy rate of this tree is poor at 73%.

Classification

Observed	Predicted 3	4	5	6	7	8	Percent Correct
3	0	1	7	1	1	0	0.0%
4	0	15	23	14	1	0	28.3%
5	0	6	572	91	12	0	84.0%
6	0	1	151	443	42	1	69.4%
7	0	3	15	48	133	0	66.8%
8	0	0	1	5	8	4	22.2%
Overall Percentage	0.0%	1.6%	48.1%	37.6%	12.3%	0.3%	73.0%

Growing Method: CRT
Dependent Variable: quality

Alcohol, Sulphates and Total Sulphur Dioxide are the most important variables in this dataset. Modifying the number of parent and child nodes.

When Parent Node = 8 and Child Node = 4

Attached is the decision tree diagram

problem_2_8.xml

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	quality
	Independent Variables	fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol
	Validation	Cross Validation
	Maximum Tree Depth	25
	Minimum Cases in Parent Node	8
	Minimum Cases in Child Node	4
Results	Independent Variables Included	alcohol, density, chlorides, volatileacidity, citricacid, sulphates, totalsulfurdioxide, fixedacidity, pH, residualsugar, freesulfurdioxide
	Number of Nodes	157
	Number of Terminal Nodes	79
	Depth	11

There are total of 157 nodes in this decision tree and 79 terminal nodes. The depth of this tree is 11.

Classification

	Predicted						
Observed	3	4	5	6	7	8	Percent Correct
3	0	1	7	1	1	0	0.0%
4	0	15	23	11	4	0	28.3%
5	0	6	592	74	9	0	86.9%
6	0	1	146	465	25	1	72.9%
7	0	3	15	45	135	1	67.8%
8	0	0	1	6	4	7	38.9%
Overall Percentage	0.0%	1.6%	49.0%	37.6%	11.1%	0.6%	75.9%

Growing Method: CRT
Dependent Variable: quality

The accuracy rate improves to 75.9%
The complexity of this decision tree worsens.

When Parent Node = 20 and Child Node = 10

Attached is the decision tree diagram

problem_2_20.xml

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	quality
	Independent Variables	fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol
	Validation	Cross Validation
	Maximum Tree Depth	25
	Minimum Cases in Parent Node	20
	Minimum Cases in Child Node	10
Results	Independent Variables Included	alcohol, density, chlorides, volatileacidity, citricacid, sulphates, totalsulfurdioxide, fixedacidity, pH, residualsugar, freesulfurdioxide
	Number of Nodes	77
	Number of Terminal Nodes	39
	Depth	9

There are total of 77 nodes in this decision tree and 39 terminal nodes. The depth of this tree is 9.

Classification

	Predicted						
Observed	3	4	5	6	7	8	Percent Correct
3	0	0	9	1	0	0	0.0%
4	0	0	35	18	0	0	0.0%
5	0	0	581	90	10	0	85.3%
6	0	0	195	409	34	0	64.1%
7	0	0	34	55	110	0	55.3%
8	0	0	1	9	8	0	0.0%
Overall Percentage	0.0%	0.0%	53.5%	36.4%	10.1%	0.0%	68.8%

Growing Method: CRT
Dependent Variable: quality

The accuracy rate reduces to 68.8%
The complexity of this decision tree is slightly improved with respect to the depth and terminal nodes.

When Parent Node = 24 and Child Node = 12

Attached is the decision tree diagram

problem_2_24.xml

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	quality
	Independent Variables	fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol
	Validation	Cross Validation
	Maximum Tree Depth	25
	Minimum Cases in Parent Node	24
	Minimum Cases in Child Node	12
Results	Independent Variables Included	alcohol, density, chlorides, volatileacidity, citricacid, sulphates, totalsulfurdioxide, fixedacidity, pH, residualsugar, freesulfurdioxide
	Number of Nodes	69
	Number of Terminal Nodes	35
	Depth	9

There are total of 69 nodes in this decision tree and 35 terminal nodes. The depth of this tree is 9.

Classification

Observed	Predicted 3	4	5	6	7	8	Percent Correct
3	0	0	9	1	0	0	0.0%
4	0	0	33	20	0	0	0.0%
5	0	0	574	97	10	0	84.3%
6	0	0	191	416	31	0	65.2%
7	0	0	34	61	104	0	52.3%
8	0	0	1	10	7	0	0.0%
Overall Percentage	0.0%	0.0%	52.7%	37.8%	9.5%	0.0%	68.4%

Growing Method: CRT
Dependent Variable: quality

The accuracy rate is slightly reduced to 68.4%
The complexity of this decision tree is improved as compared to the initial tree with parent node = 10 with 35 terminal nodes and depth of the tree 9.

When Parent Node = 30 and Child Node = 15

Attached is the decision tree diagram

problem_2_30.xml

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	quality
	Independent Variables	fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol
	Validation	Cross Validation
	Maximum Tree Depth	25
	Minimum Cases in Parent Node	30
	Minimum Cases in Child Node	15
Results	Independent Variables Included	alcohol, density, chlorides, volatileacidity, citricacid, sulphates, totalsulfurdioxide, fixedacidity, pH, residualsugar, freesulfurdioxide
	Number of Nodes	61
	Number of Terminal Nodes	31
	Depth	8

There are total of 61 nodes in this decision tree and 31 terminal nodes. The depth of this tree is 8.

Classification

	Predicted						
Observed	3	4	5	6	7	8	Percent Correct
3	0	0	9	1	0	0	0.0%
4	0	0	32	21	0	0	0.0%
5	0	0	572	99	10	0	84.0%
6	0	0	194	413	31	0	64.7%
7	0	0	31	64	104	0	52.3%
8	0	0	1	10	7	0	0.0%
Overall Percentage	0.0%	0.0%	52.5%	38.0%	9.5%	0.0%	68.1%

Growing Method: CRT
Dependent Variable: quality

The accuracy rate slightly reduces at 68.1%
The complexity of this decision tree is improved as compared to the initial tree with parent node = 10.

As the complexity of the tree improves, the accuracy rate reduces.

The complexity is worst with 157 nodes out of which 79 are terminal nodes when parent node = 8 and child node = 4. The accuracy rate is best at 75.9%
The complexity is least with 61 nodes out of which 31 re terminal nodes when parent node = 30 and child node = 15. The accuracy rate is least at 68.1%.
When parent node = 10 and child node = 5, the accuracy rate is 73% with 123 nodes out of which 62 are terminal nodes.

Section c)

Values for Quality variable (class): 3, 4, 5, 6, 7, 8

Equal Depth Partitioning: Dividing into 3 bins

Bin 1: 3, 4

Bin 2: 5, 6

Bin 3: 7, 8

Smoothing by means

Bin 1: 3.5, 3.5

Bin 2: 5.5, 5.5

Bin 3: 7.5, 7.5

We have replaced the values of the quality variables with the values found by Smoothing of means.

When Parent Node = 10, Child Node = 5
The model summary shows all the variables were included in the tree.

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	quality_binned
	Independent Variables	fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol
	Validation	Cross Validation
	Maximum Tree Depth	25
	Minimum Cases in Parent Node	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	alcohol, density, fixedacidity, chlorides, pH, totalsulfurdioxide, volatileacidity, residualsugar, freesulfurdioxide, citricacid, sulphates
	Number of Nodes	71
	Number of Terminal Nodes	36
	Depth	9

Using the binned class variable, we have a total of 71 nodes out of which 36 are terminal nodes. The depth of the tree is 9. As can be seen, the first node is split based on our most important predictor, Alcohol. The value of impurity is calculated based on Gini with Alcohol node having a value of 0.032. The variable with highest reduction of impurity is selected as splitting attribute.

problem_2_bin.xml

The splitting value for alcohol variable is <= 11.55 and > 11.55.

Classification

	Predicted			
Observed	3.50	5.50	7.50	Percent Correct
3.50	11	52	0	17.5%
5.50	1	1277	41	96.8%
7.50	0	75	142	65.4%
Overall Percentage	0.8%	87.8%	11.4%	89.4%

Growing Method: CRT
Dependent Variable: quality_binned

When Parent Node = 8 and Child Node = 4

Attached is the decision tree diagram

problem_2_8_bin.xml

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	quality_binned
	Independent Variables	fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol
	Validation	Cross Validation
	Maximum Tree Depth	25
	Minimum Cases in Parent Node	8
	Minimum Cases in Child Node	4
Results	Independent Variables Included	alcohol, density, fixedacidity, chlorides, pH, totalsulfurdioxide, volatileacidity, residualsugar, freesulfurdioxide, citricacid, sulphates
	Number of Nodes	79
	Number of Terminal Nodes	40
	Depth	9

There are total 79 nodes in this tree out of which 40 are terminal nodes. The depth of tree is 9.

Classification

	Predicted			
Observed	3.50	5.50	7.50	Percent Correct
3.50	20	43	0	31.7%
5.50	3	1278	38	96.9%
7.50	0	72	145	66.8%
Overall Percentage	1.4%	87.1%	11.4%	90.2%

Growing Method: CRT
Dependent Variable: quality_binned

The accuracy rate is slightly reduced to 90.2%.
The accuracy rate is improved from 89.4% to 90.2%.

When Parent Node = 20 and Child Node = 10

Attached is the decision tree diagram

problem_2_20_bin.xml

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	quality_binned
	Independent Variables	fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol
	Validation	Cross Validation
	Maximum Tree Depth	25
	Minimum Cases in Parent Node	20
	Minimum Cases in Child Node	10
Results	Independent Variables Included	alcohol, density, fixedacidity, chlorides, pH, totalsulfurdioxide, volatileacidity, residualsugar, freesulfurdioxide, citricacid, sulphates
	Number of Nodes	45
	Number of Terminal Nodes	23
	Depth	8

The number of total nodes have almost reduced to half 45 out of which 23 are terminal nodes.
The depth of the tree is 8.

Classification

Observed	Predicted 3.50	5.50	7.50	Percent Correct
3.50	15	48	0	23.8%
5.50	13	1261	45	95.6%
7.50	0	94	123	56.7%
Overall Percentage	1.8%	87.7%	10.5%	87.5%

Growing Method: CRT
Dependent Variable: quality_binned

The accuracy rate is 87.5%.

When Parent Node = 24 and Child Node = 12

Attached is the decision tree diagram

problem_2_24_bin.xml

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	quality_binned
	Independent Variables	fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol
	Validation	Cross Validation
	Maximum Tree Depth	25
	Minimum Cases in Parent Node	24
	Minimum Cases in Child Node	12
Results	Independent Variables Included	alcohol, density, fixedacidity, chlorides, pH, totalsulfurdioxide, volatileacidity, residualsugar, freesulfurdioxide, citricacid, sulphates
	Number of Nodes	41
	Number of Terminal Nodes	21
	Depth	8

The total numbers of nodes are 41 while the terminal nodes are 21.

The depth of the tree is 8.

Classification

Observed	Predicted 3.50	5.50	7.50	Percent Correct
3.50	15	48	0	23.8%
5.50	13	1264	42	95.8%
7.50	0	101	116	53.5%
Overall Percentage	1.8%	88.4%	9.9%	87.2%

Growing Method: CRT
Dependent Variable: quality_binned

The accuracy rate of this decision tree is 87.2%.

When Parent Node = 30 and Child Node = 15

Attached is the decision tree diagram

problem_2_30_bin.xml

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	quality_binned
	Independent Variables	fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol
	Validation	Cross Validation
	Maximum Tree Depth	25
	Minimum Cases in Parent Node	30
	Minimum Cases in Child Node	15
Results	Independent Variables Included	alcohol, density, fixedacidity, chlorides, pH, totalsulfurdioxide, volatileacidity, residualsugar, freesulfurdioxide, citricacid, sulphates
	Number of Nodes	37
	Number of Terminal Nodes	19
	Depth	7

The total numbers of nodes are 37 while the terminal nodes are 19.
The depth of the tree is 7.

Classification

Observed	Predicted 3.50	5.50	7.50	Percent Correct
3.50	15	48	0	23.8%
5.50	13	1264	42	95.8%
7.50	0	101	116	53.5%
Overall Percentage	1.8%	88.4%	9.9%	87.2%

Growing Method: CRT
Dependent Variable: quality_binned

The accuracy rate remain unchanged compared to parent node = 24 at 87.2%.

As the complexity of the tree improves, the accuracy rate reduces.
The complexity is worst with 79 nodes out of which 40 are terminal nodes when parent node = 8 and child node = 4. The accuracy rate is best at 90.2%
The complexity is least with 37 nodes out of which 19 are terminal nodes when parent node = 30 and child node = 15. The accuracy rate is least at 87.2%.
When parent node = 10 and child node = 5, the accuracy rate is 89.4% with 71 nodes and 36 terminal nodes.

Section d)

Original Variable
The complexity is worst with 157 nodes out of which 79 are terminal nodes when parent node = 8 and child node = 4. The accuracy rate is best at 75.9%
The complexity is least with 61 nodes out of which 31 re terminal nodes when parent node = 30 and child node = 15. The accuracy rate is least at 68.1%.
When parent node = 10 and child node = 5, the accuracy rate is 73% with 123 nodes out of which 62 are terminal nodes

Binned Class Variable
The complexity is worst with 79 nodes out of which 40 are terminal nodes when parent node = 8 and child node = 4. The accuracy rate is best at 90.2%
The complexity is least with 37 nodes out of which 19 are terminal nodes when parent node = 30 and child node = 15. The accuracy rate is least at 87.2%.
When parent node = 10 and child node = 5, the accuracy rate is 89.4%.

The binned variable is production a decision tree with better accuracy rate. The complexity of the tree with binned variable is better compared to the one with original variable.

Section e)

Imbalanced data typically refers to a classification problem where the number of observations per class is not equally distributed; often you'll have a large amount of data/observations for one class

(referred to as the *majority class*), and much fewer observations for one or more other classes (referred to as the *minority classes*).
In this particular dataset, the number of records for value 5, 6 and 7 are 681, 638 and 199 respectively. The amount of data compared to value 3, 4 and 8 are quite large.
The amount of data for value 3, 4 and 8 are 10, 53 and 18 respectively.

We would be using the undersampling method to resolve the class imbalance problem.
As part of undersampling, we will randomly duplicate records so that the number of records match the majority class.

As part of oversampling, we will randomly select records (delete remaining) so that the number of records match the minority class.

Once there is a balance of records between different values, we would regenerate the decision tree.

quality

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	3.00	19	17.3	17.3	17.3
	4.00	18	16.4	16.4	33.6
	5.00	18	16.4	16.4	50.0
	6.00	19	17.3	17.3	67.3
	7.00	18	16.4	16.4	83.6
	8.00	18	16.4	16.4	100.0
	Total	110	100.0	100.0	

Total number of records: 110
Number of cases: 6
Value 3 has 19 records, while value 4 has 18 records.
Value 6 has 19 records while value 5, 7 and 8 has 18 records each.

When Parent Node = 10, Child Node = 5

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	quality
	Independent Variables	fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol
	Validation	Cross Validation
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	alcohol, density, fixedacidity, chlorides, pH, volatileacidity, citricacid, freesulfurdioxide, totalsulfurdioxide, sulphates, residualsugar
	Number of Nodes	21
	Number of Terminal Nodes	11
	Depth	6

The number of nodes in this tree is 21 out of which 11 are terminal nodes.
The depth of the tree is 6.

Attached is the decision tree diagram

problem_2e.xml

Classification

	Predicted						
Observed	3.00	4.00	5.00	6.00	7.00	8.00	Percent Correct
3.00	17	2	0	0	0	0	89.5%
4.00	4	13	1	0	0	0	72.2%
5.00	0	4	10	4	0	0	55.6%
6.00	1	0	3	15	0	0	78.9%
7.00	0	1	0	6	10	1	55.6%
8.00	0	0	0	0	2	16	88.9%
Overall Percentage	20.0%	18.2%	12.7%	22.7%	10.9%	15.5%	73.6%

Growing Method: CRT
Dependent Variable: quality

The accuracy rate has improved by 0.6% to 73.6% compared to the original tree.

When Parent Node = 8 and Child Node = 4

Attached is the decision tree diagram

problem_2e_8_bin.xml

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	quality
	Independent Variables	fixedacidity, volatileacidity, citricacid, residualsugar, chlorides, freesulfurdioxide, totalsulfurdioxide, density, pH, sulphates, alcohol
	Validation	Cross Validation
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	8
	Minimum Cases in Child Node	4
Results	Independent Variables Included	alcohol, density, fixedacidity, chlorides, pH, volatileacidity, citricacid, freesulfurdioxide, totalsulfurdioxide, sulphates, residualsugar
	Number of Nodes	21
	Number of Terminal Nodes	11
	Depth	6

The number of nodes in this tree is 21 out of which 11 are terminal nodes.
The depth of the tree is 6.

Classification

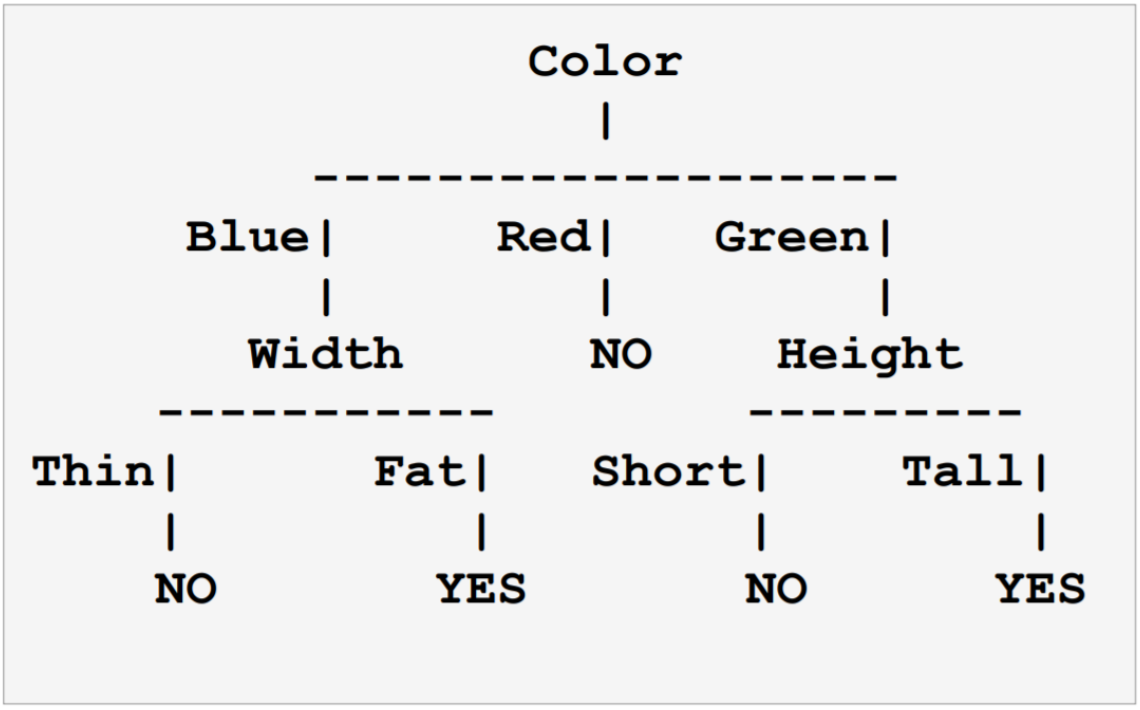
Observed	Predicted 3.00	4.00	5.00	6.00	7.00	8.00	Percent Correct
3.00	17	2	0	0	0	0	89.5%
4.00	4	13	1	0	0	0	72.2%
5.00	0	4	10	4	0	0	55.6%
6.00	1	0	3	15	0	0	78.9%
7.00	0	1	0	6	10	1	55.6%
8.00	0	0	0	0	2	16	88.9%
Overall Percentage	20.0%	18.2%	12.7%	22.7%	10.9%	15.5%	73.6%

Growing Method: CRT
Dependent Variable: quality

There is no change in the accuracy rate compared to the previous tree with parent node = 10.

Conclusion: We are resolving the issue with class imbalance by either oversampling or under sampling the dataset. Following the class balancing, the complexity of the tree with respect to number of nodes and terminal nodes reduces.
The accuracy rate when compared to original model slightly improves.

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

Example A
If Color = “Red” then Class = “NO”

Example B
If Color = “Blue” and Width = “Fat” then Class = “YES”

Example C
If Color = “Green” and Height = “Short” then Class = “NO”

Example D
If Color = “Green” then Class = “YES”

Example E
If Color = “Blue” and Width = “Thin” then Class = “NO”