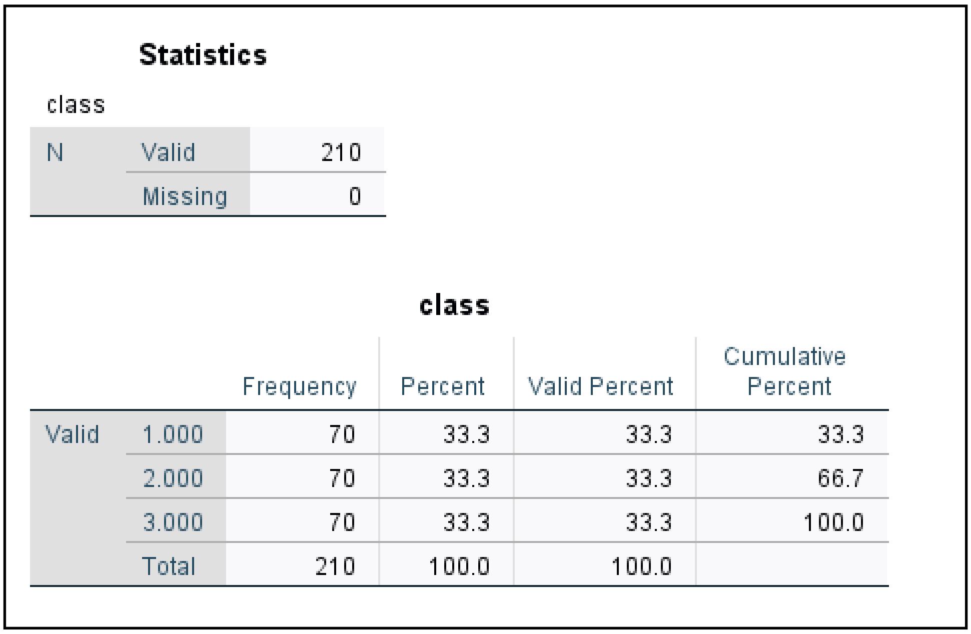
Decision tree

Section a)

Following is the class distribution for the diagnosis class.

There are total of 210 rows in this dataset.



Statistics

class

N	Valid	210
	Missing	0

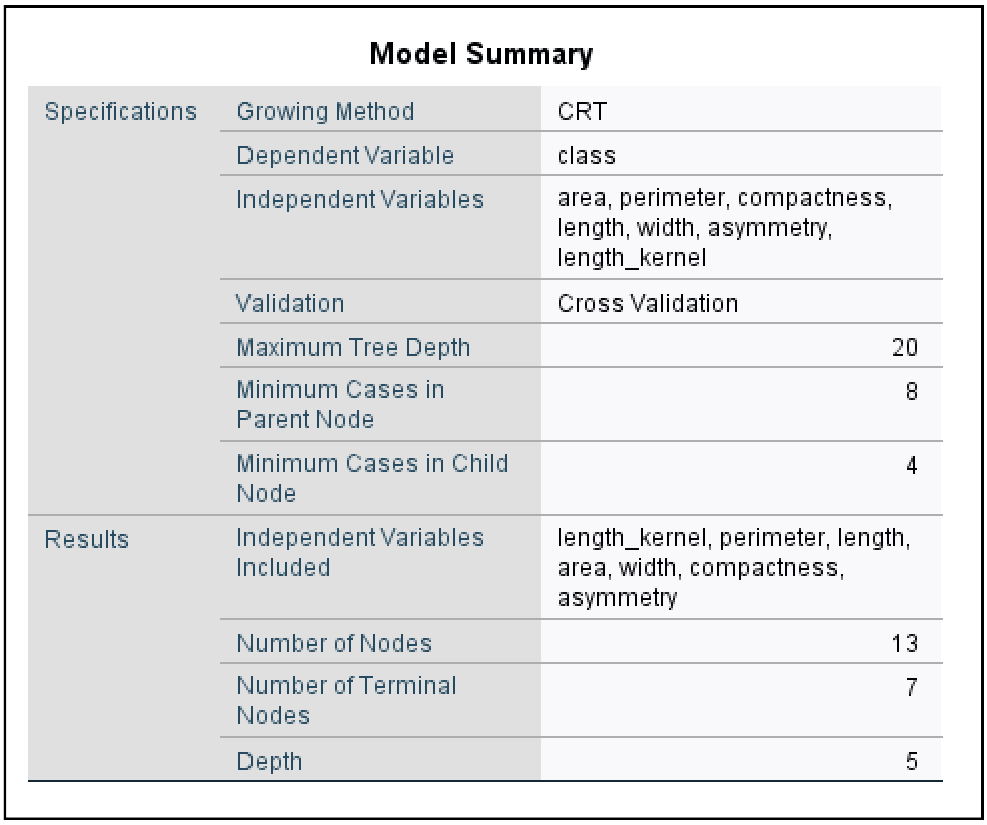
class

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1.000	70	33.3	33.3	33.3
	2.000	70	33.3	33.3	66.7
	3.000	70	33.3	33.3	100.0
	Total	210	100.0	100.0	

Each class has 70 records each.

When Parent Node = 8, Child Node = 4

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

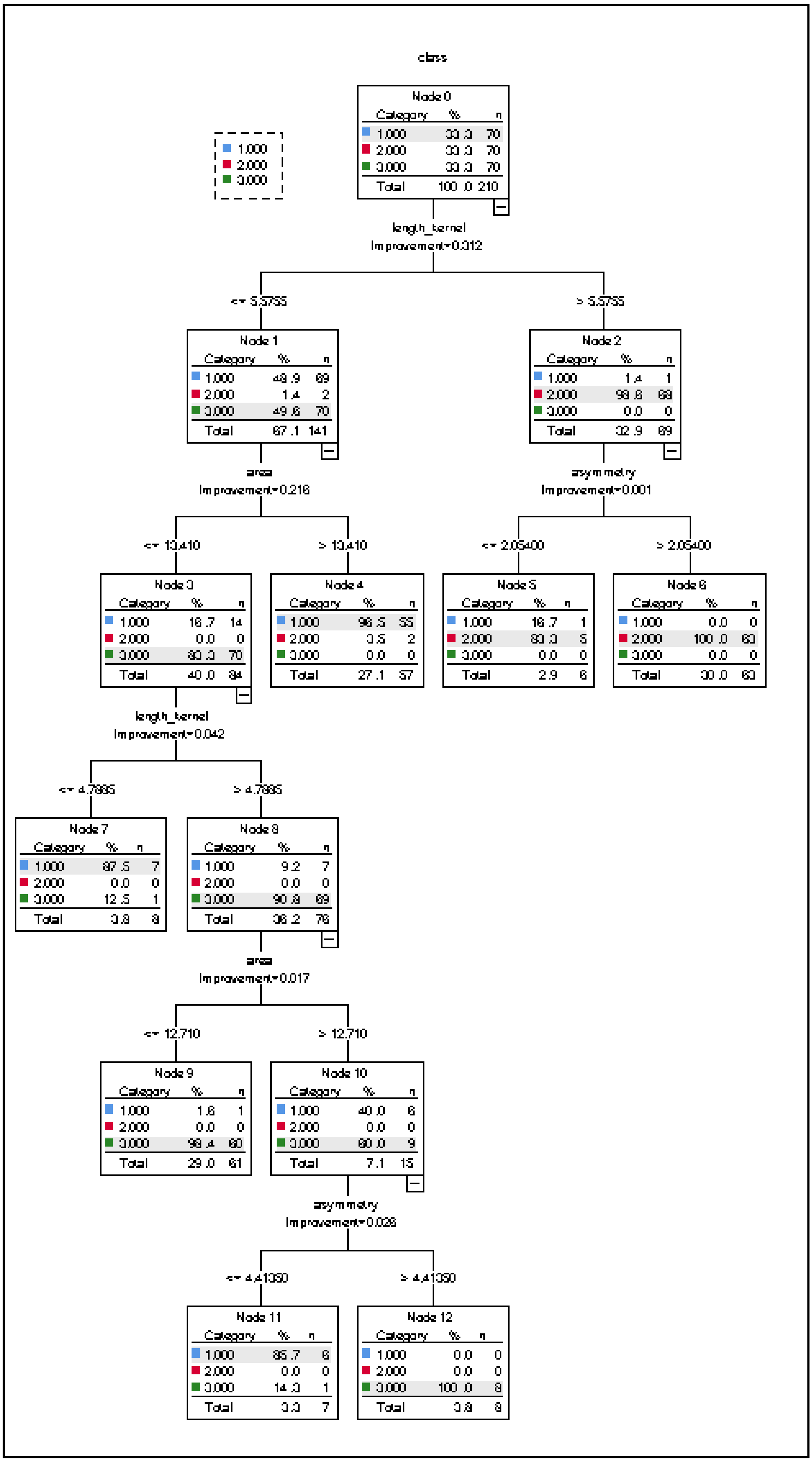


Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	class
	Independent Variables	area, perimeter, compactness, length, width, asymmetry, length_kernel
	Validation	Cross Validation
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	8
	Minimum Cases in Child Node	4
Results	Independent Variables Included	length_kernel, perimeter, length, area, width, compactness, asymmetry
	Number of Nodes	13
	Number of Terminal Nodes	7
	Depth	5

The decision tree has 13 nodes out of which 7 are terminal nodes. The depth of the tree is 5.

Decision Tree



Node 0

Category	%	n
🟦 1.000	33.3	70
🟥 2.000	33.3	70
🟩 3.000	33.3	70
Total	100.0	210

Node 1

Category	%	n
🟦 1.000	48.9	69
🟥 2.000	1.4	2
🟩 3.000	49.6	70
Total	67.1	141

Node 2

Category	%	n
🟦 1.000	1.4	1
🟥 2.000	98.6	68
🟩 3.000	0.0	0
Total	32.9	69

Node 3

Category	%	n
🟦 1.000	16.7	14
🟥 2.000	0.0	0
🟩 3.000	83.3	70
Total	40.0	84

Node 4

Category	%	n
🟦 1.000	96.5	55
🟥 2.000	3.5	2
🟩 3.000	0.0	0
Total	27.1	57

Node 5

Category	%	n
🟦 1.000	16.7	1
🟥 2.000	83.3	5
🟩 3.000	0.0	0
Total	2.9	6

Node 6

Category	%	n
🟦 1.000	0.0	0
🟥 2.000	100.0	63
🟩 3.000	0.0	0
Total	30.0	63

Node 7

Category	%	n
🟦 1.000	87.5	7
🟥 2.000	0.0	0
🟩 3.000	12.5	1
Total	3.8	8

Node 8

Category	%	n
🟦 1.000	9.2	7
🟥 2.000	0.0	0
🟩 3.000	90.8	69
Total	36.2	76

Node 9

Category	%	n
🟦 1.000	1.6	1
🟥 2.000	0.0	0
🟩 3.000	98.4	60
Total	29.0	61

Node 10

Category	%	n
🟦 1.000	40.0	6
🟥 2.000	0.0	0
🟩 3.000	60.0	9
Total	7.1	15

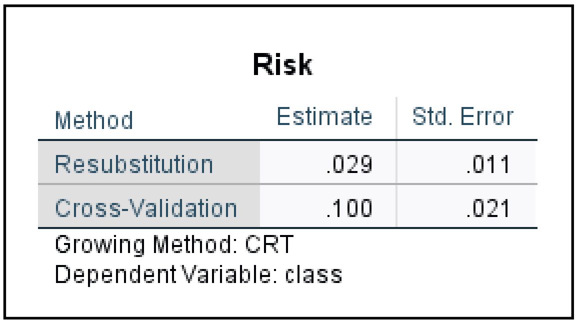
Node 11

Category	%	n
🟦 1.000	85.7	6
🟥 2.000	0.0	0
🟩 3.000	14.3	1
Total	3.3	7

Node 12

Category	%	n
🟦 1.000	0.0	0
🟥 2.000	0.0	0
🟩 3.000	100.0	8
Total	3.8	8

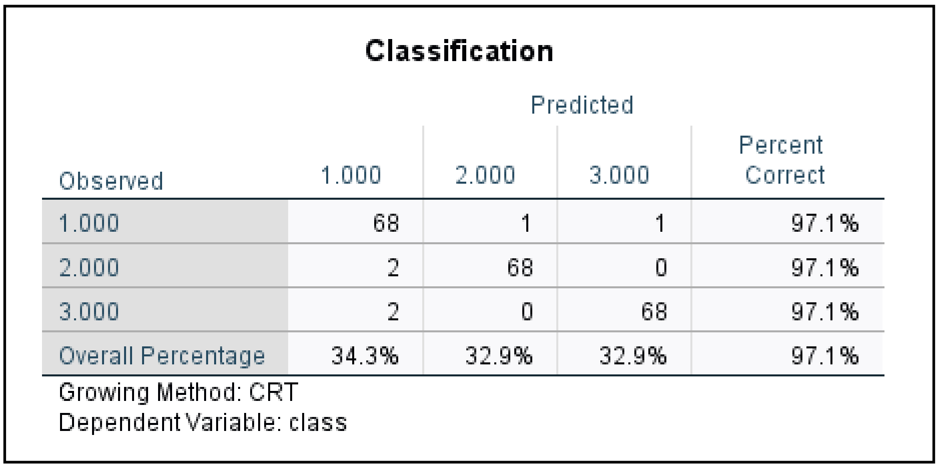
As can be seen in the tree, the first node is split based on our most important predictor, length_kernel. The value of impurity is calculated based on Gini with length_kernel node having a value of 0.312.
The variable with highest reduction of impurity is selected as splitting attribute.



Risk

Method	Estimate	Std. Error
Resubstitution	.029	.011
Cross-Validation	.100	.021

Growing Method: CRT
Dependent Variable: class



Classification

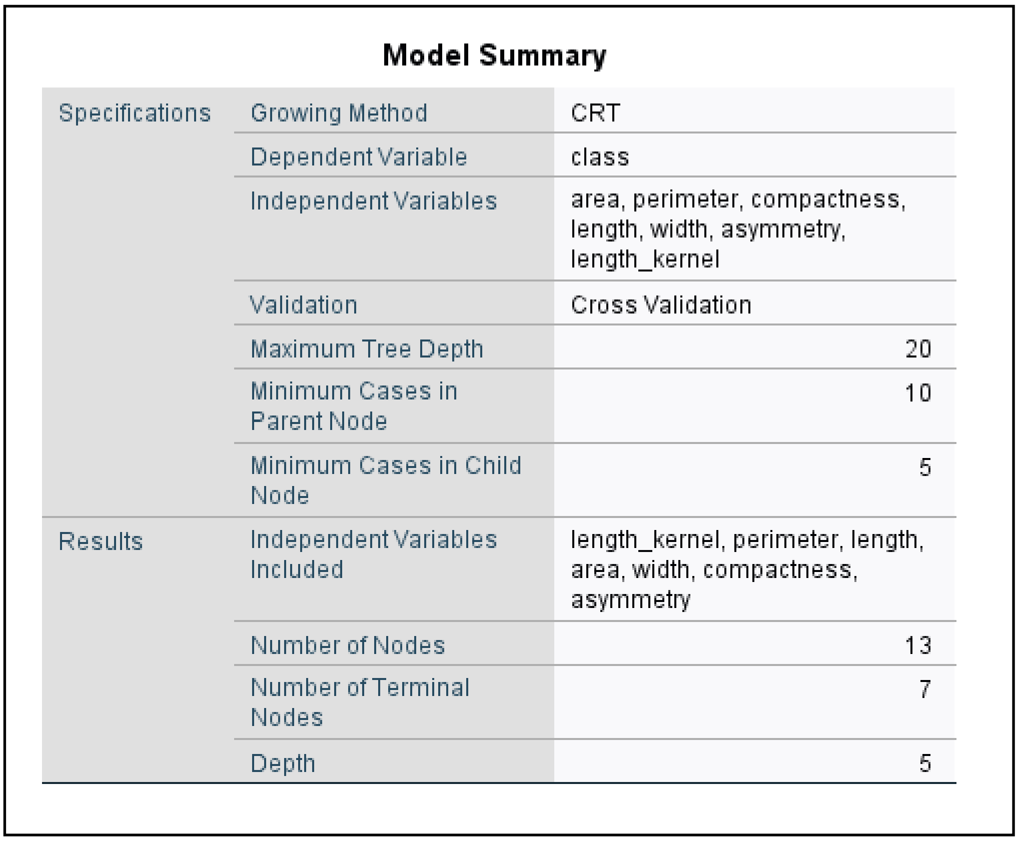
	Predicted			
Observed	1.000	2.000	3.000	Percent Correct
1.000	68	1	1	97.1%
2.000	2	68	0	97.1%
3.000	2	0	68	97.1%
Overall Percentage	34.3%	32.9%	32.9%	97.1%

Growing Method: CRT
Dependent Variable: class

According to the above classification matrix, the accuracy rate of this tree is 97.1%

When Parent Node = 10, Child Node = 5

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

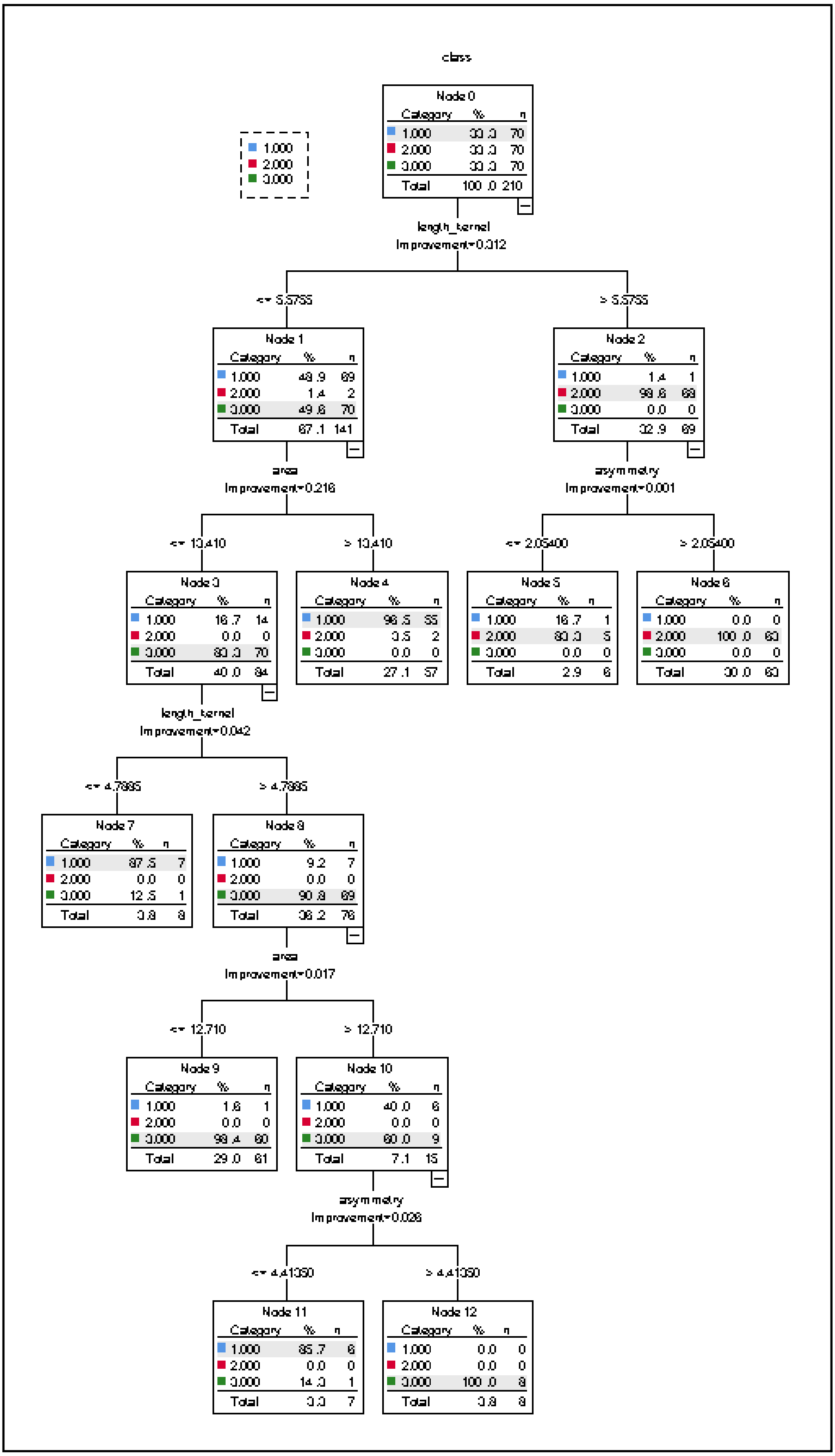


Model Summary

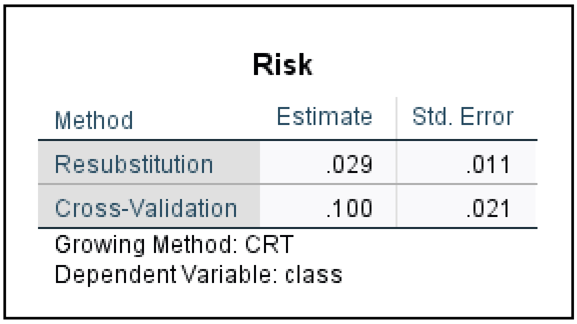
Specifications	Growing Method	CRT
	Dependent Variable	class
	Independent Variables	area, perimeter, compactness, length, width, asymmetry, length_kernel
	Validation	Cross Validation
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	length_kernel, perimeter, length, area, width, compactness, asymmetry
	Number of Nodes	13
	Number of Terminal Nodes	7
	Depth	5

The decision tree has 13 nodes out of which 7 are terminal nodes. The depth of the tree is 5.

Decision Tree


class
🟦 1.000
🟥 2.000
🟩 3.000
Node 0
Category % n
🟦 1.000 33.3 70
🟥 2.000 33.3 70
🟩 3.000 33.3 70
Total 100.0 210
length_kernel
Improvement=0.312
<= 5.5755
> 5.5755
Node 1
Category % n
🟦 1.000 48.9 69
🟥 2.000 1.4 2
🟩 3.000 49.6 70
Total 67.1 141
Node 2
Category % n
🟦 1.000 1.4 1
🟥 2.000 98.6 68
🟩 3.000 0.0 0
Total 32.9 69
area
Improvement=0.216
asymmetry
Improvement=0.001
<= 13.410
> 13.410
<= 2.05400
> 2.05400
Node 3
Category % n
🟦 1.000 16.7 14
🟥 2.000 0.0 0
🟩 3.000 83.3 70
Total 40.0 84
Node 4
Category % n
🟦 1.000 96.5 55
🟥 2.000 3.5 2
🟩 3.000 0.0 0
Total 27.1 57
Node 5
Category % n
🟦 1.000 16.7 1
🟥 2.000 83.3 5
🟩 3.000 0.0 0
Total 2.9 6
Node 6
Category % n
🟦 1.000 0.0 0
🟥 2.000 100.0 63
🟩 3.000 0.0 0
Total 30.0 63
length_kernel
Improvement=0.042
<= 4.7885
> 4.7885
Node 7
Category % n
🟦 1.000 87.5 7
🟥 2.000 0.0 0
🟩 3.000 12.5 1
Total 3.8 8
Node 8
Category % n
🟦 1.000 9.2 7
🟥 2.000 0.0 0
🟩 3.000 90.8 69
Total 36.2 76
area
Improvement=0.017
<= 12.710
> 12.710
Node 9
Category % n
🟦 1.000 1.6 1
🟥 2.000 0.0 0
🟩 3.000 98.4 60
Total 29.0 61
Node 10
Category % n
🟦 1.000 40.0 6
🟥 2.000 0.0 0
🟩 3.000 60.0 9
Total 7.1 15
asymmetry
Improvement=0.028
<= 4.41350
> 4.41350
Node 11
Category % n
🟦 1.000 85.7 6
🟥 2.000 0.0 0
🟩 3.000 14.3 1
Total 3.3 7
Node 12
Category % n
🟦 1.000 0.0 0
🟥 2.000 0.0 0
🟩 3.000 100.0 8
Total 3.8 8

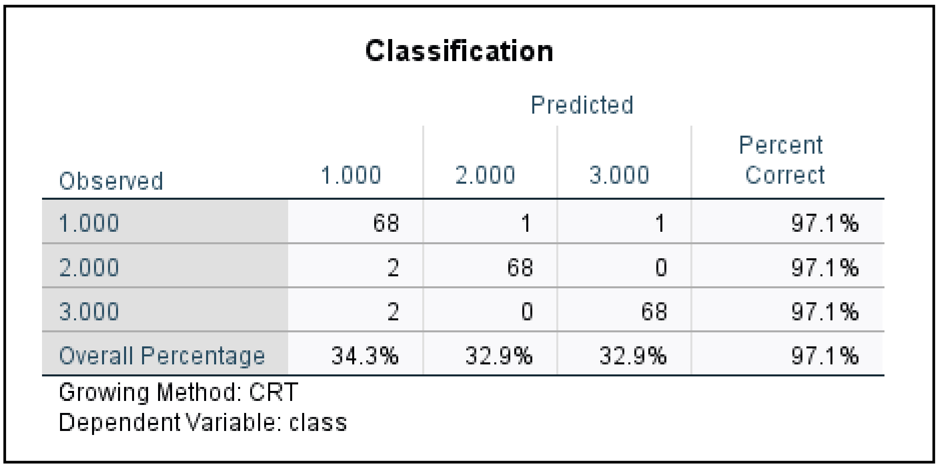
As can be seen in the tree, the first node is split based on our most important predictor, length_kernel. The value of impurity is calculated based on Gini with length_kernel node having a value of 0.312.
The variable with highest reduction of impurity is selected as splitting attribute.



Risk

Method	Estimate	Std. Error
Resubstitution	.029	.011
Cross-Validation	.100	.021

Growing Method: CRT
Dependent Variable: class



Classification

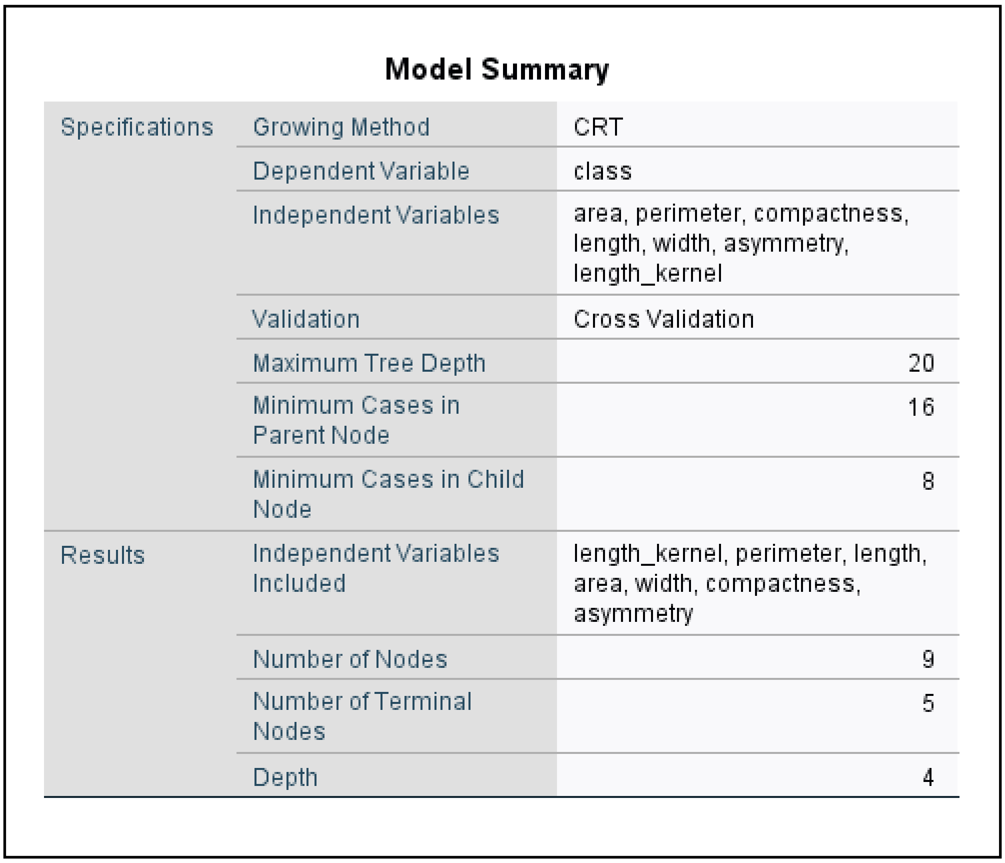
	Predicted			
Observed	1.000	2.000	3.000	Percent Correct
1.000	68	1	1	97.1%
2.000	2	68	0	97.1%
3.000	2	0	68	97.1%
Overall Percentage	34.3%	32.9%	32.9%	97.1%

Growing Method: CRT
Dependent Variable: class

According to the above classification matrix, the accuracy rate of this tree is 97.1%

When Parent Node = 16, Child Node = 8

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

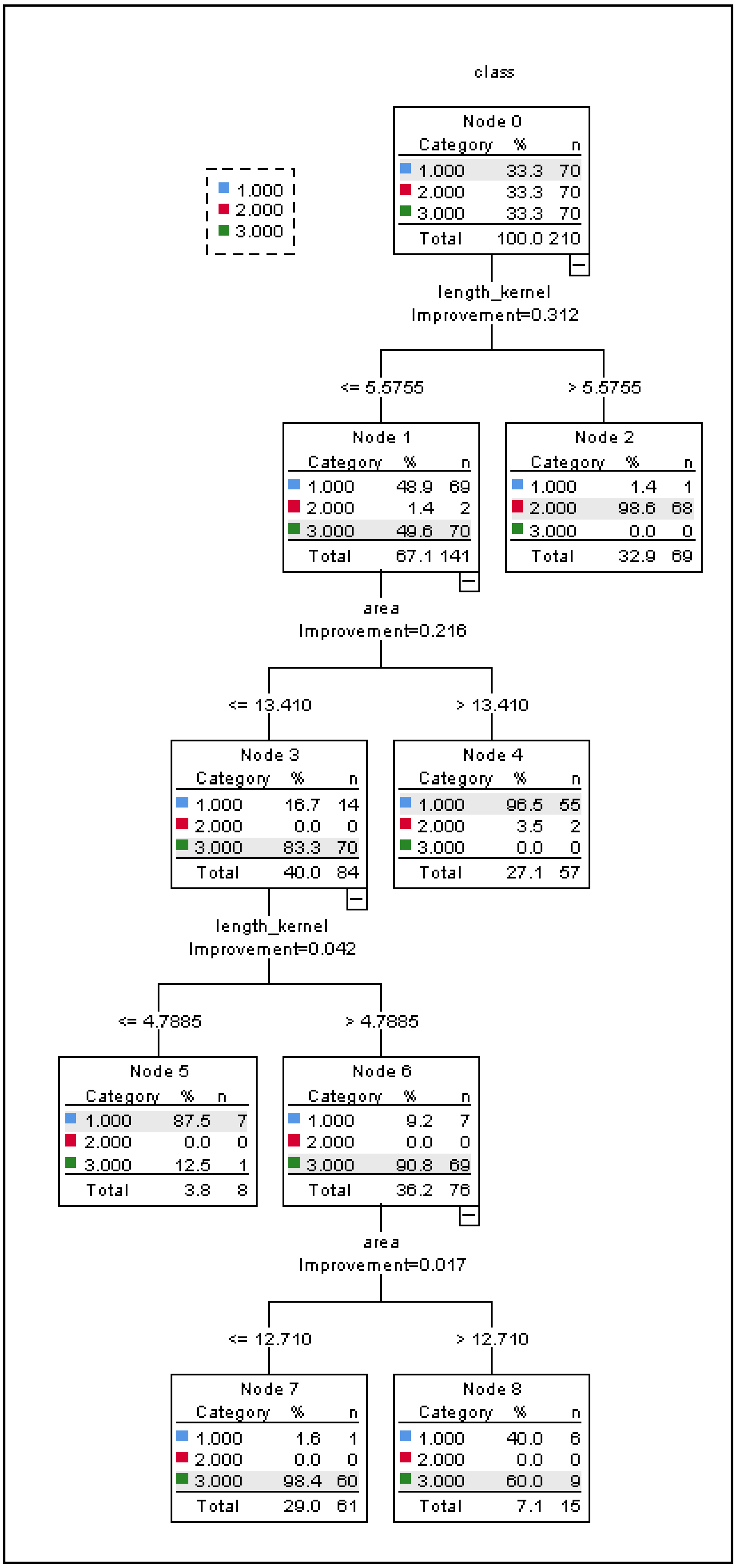


Model Summary

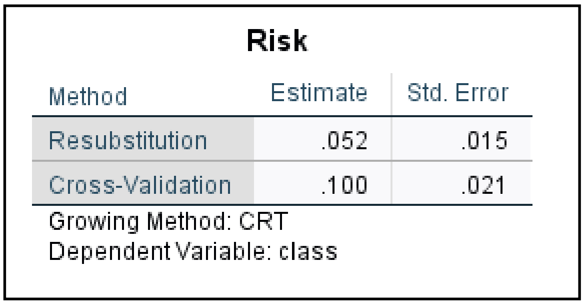
Specifications	Growing Method	CRT
	Dependent Variable	class
	Independent Variables	area, perimeter, compactness, length, width, asymmetry, length_kernel
	Validation	Cross Validation
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	16
	Minimum Cases in Child Node	8
Results	Independent Variables Included	length_kernel, perimeter, length, area, width, compactness, asymmetry
	Number of Nodes	9
	Number of Terminal Nodes	5
	Depth	4

The decision tree has 9 nodes out of which 5 are terminal nodes. The depth of the tree is 4.

Decision Tree


class
🟦 1.000
🟥 2.000
🟩 3.000
Node 0
Category % n
🟦 1.000 33.3 70
🟥 2.000 33.3 70
🟩 3.000 33.3 70
Total 100.0 210
length_kernel
Improvement=0.312
<= 5.5755
> 5.5755
Node 1
Category % n
🟦 1.000 48.9 69
🟥 2.000 1.4 2
🟩 3.000 49.6 70
Total 67.1 141
Node 2
Category % n
🟦 1.000 1.4 1
🟥 2.000 98.6 68
🟩 3.000 0.0 0
Total 32.9 69
area
Improvement=0.216
<= 13.410
> 13.410
Node 3
Category % n
🟦 1.000 16.7 14
🟥 2.000 0.0 0
🟩 3.000 83.3 70
Total 40.0 84
Node 4
Category % n
🟦 1.000 96.5 55
🟥 2.000 3.5 2
🟩 3.000 0.0 0
Total 27.1 57
length_kernel
Improvement=0.042
<= 4.7885
> 4.7885
Node 5
Category % n
🟦 1.000 87.5 7
🟥 2.000 0.0 0
🟩 3.000 12.5 1
Total 3.8 8
Node 6
Category % n
🟦 1.000 9.2 7
🟥 2.000 0.0 0
🟩 3.000 90.8 69
Total 36.2 76
area
Improvement=0.017
<= 12.710
> 12.710
Node 7
Category % n
🟦 1.000 1.6 1
🟥 2.000 0.0 0
🟩 3.000 98.4 60
Total 29.0 61
Node 8
Category % n
🟦 1.000 40.0 6
🟥 2.000 0.0 0
🟩 3.000 60.0 9
Total 7.1 15

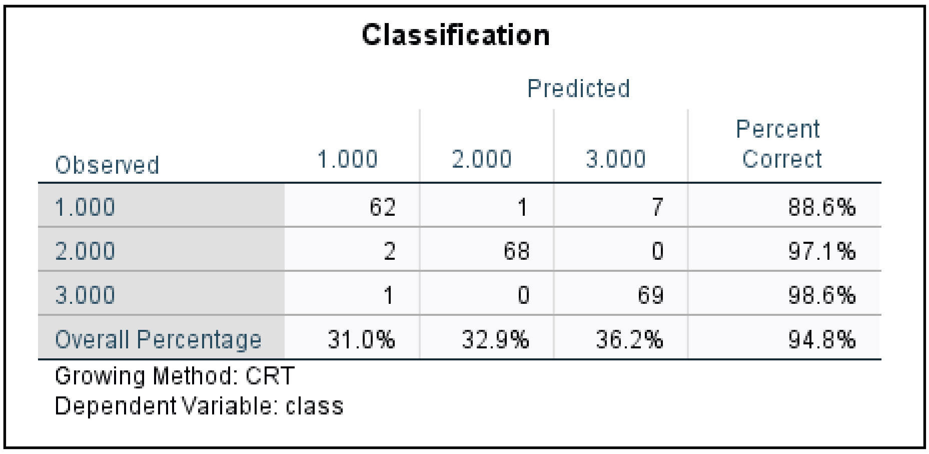
As can be seen in the tree, the first node is split based on our most important predictor, length_kernel. The value of impurity is calculated based on Gini with length_kernel node having a value of 0.312.
The variable with highest reduction of impurity is selected as splitting attribute.



Risk

Method	Estimate	Std. Error
Resubstitution	.052	.015
Cross-Validation	.100	.021

Growing Method: CRT
Dependent Variable: class



Classification

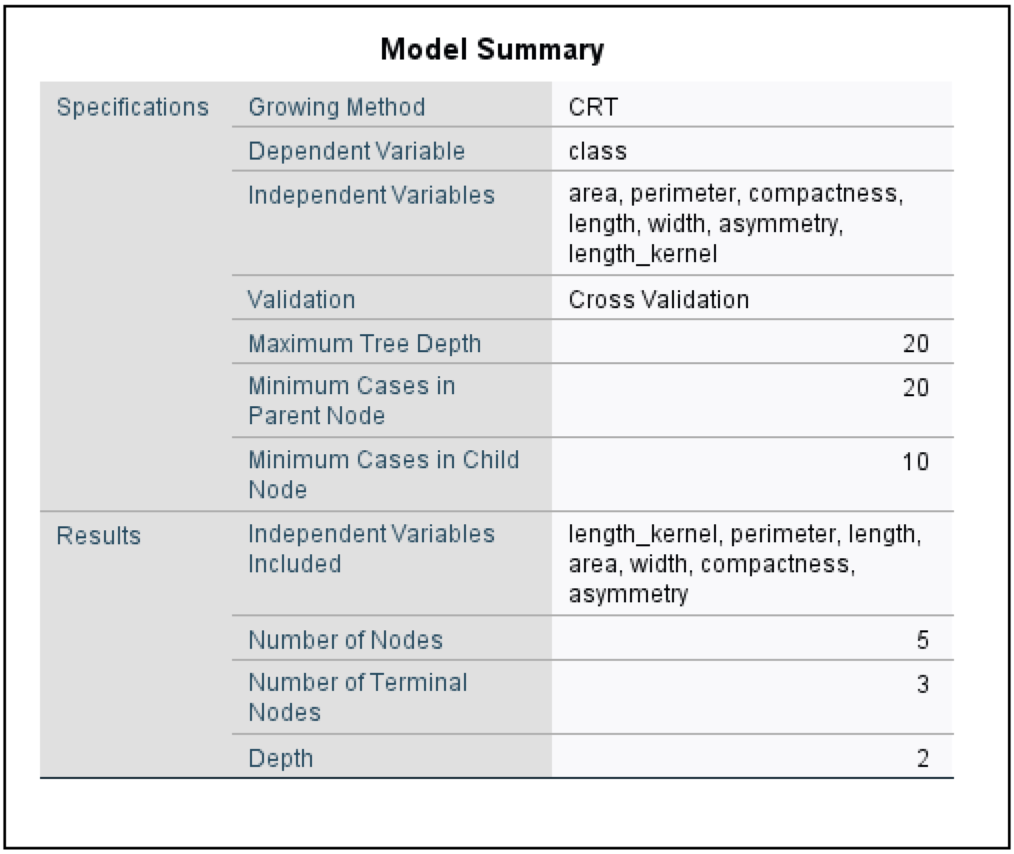
	Predicted			
Observed	1.000	2.000	3.000	Percent Correct
1.000	62	1	7	88.6%
2.000	2	68	0	97.1%
3.000	1	0	69	98.6%
Overall Percentage	31.0%	32.9%	36.2%	94.8%

Growing Method: CRT
Dependent Variable: class

According to the above classification matrix, the accuracy rate of this tree is 94.8%

When Parent Node = 20, Child Node = 10

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

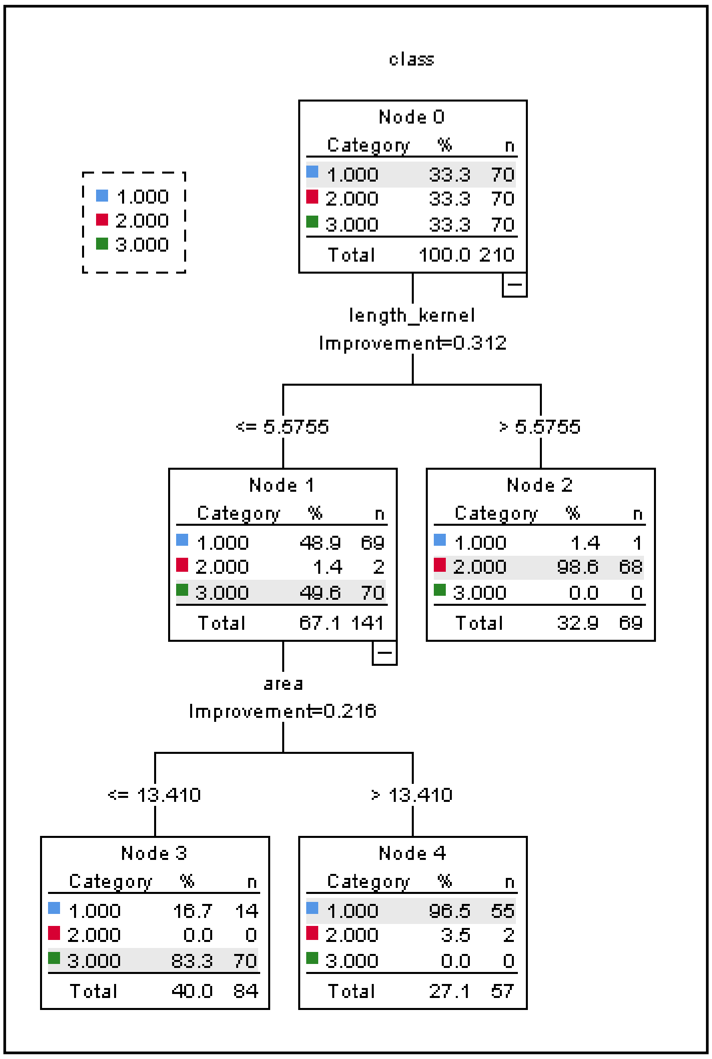


Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	class
	Independent Variables	area, perimeter, compactness, length, width, asymmetry, length_kernel
	Validation	Cross Validation
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	20
	Minimum Cases in Child Node	10
Results	Independent Variables Included	length_kernel, perimeter, length, area, width, compactness, asymmetry
	Number of Nodes	5
	Number of Terminal Nodes	3
	Depth	2

The decision tree has 5 nodes out of which 3 are terminal nodes. The depth of the tree is 2.

Decision Tree



Node 0

Category	%	n
🟦 1.000	33.3	70
🟥 2.000	33.3	70
🟩 3.000	33.3	70
Total	100.0	210

Node 1

Category	%	n
🟦 1.000	48.9	69
🟥 2.000	1.4	2
🟩 3.000	49.6	70
Total	67.1	141

Node 2

Category	%	n
🟦 1.000	1.4	1
🟥 2.000	98.6	68
🟩 3.000	0.0	0
Total	32.9	69

Node 3

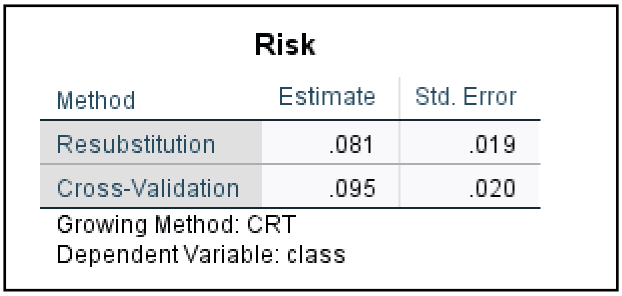
Category	%	n
🟦 1.000	16.7	14
🟥 2.000	0.0	0
🟩 3.000	83.3	70
Total	40.0	84

Node 4

Category	%	n
🟦 1.000	96.5	55
🟥 2.000	3.5	2
🟩 3.000	0.0	0
Total	27.1	57

As can be seen in the tree, the first node is split based on our most important predictor, length_kernel. The value of impurity is calculated based on Gini with length_kernel node having a value of 0.312.

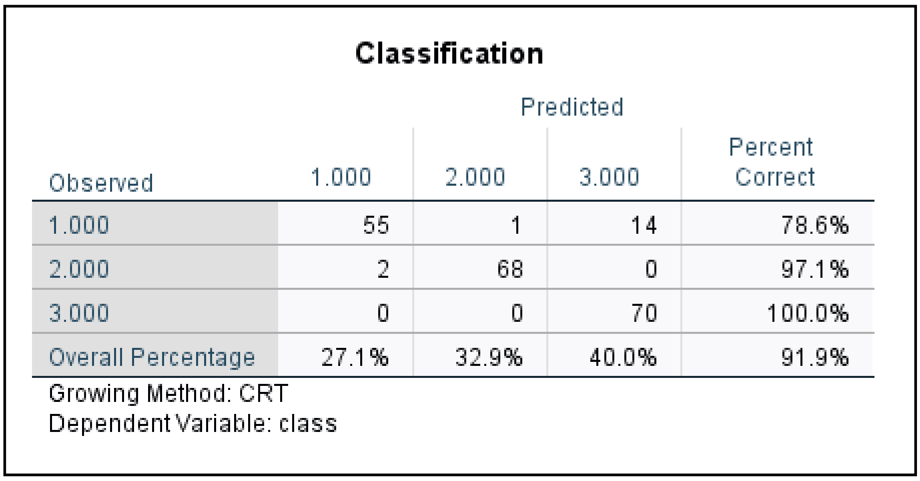
The variable with highest reduction of impurity is selected as splitting attribute.



Risk

Method	Estimate	Std. Error
Resubstitution	.081	.019
Cross-Validation	.095	.020

Growing Method: CRT
Dependent Variable: class



Classification

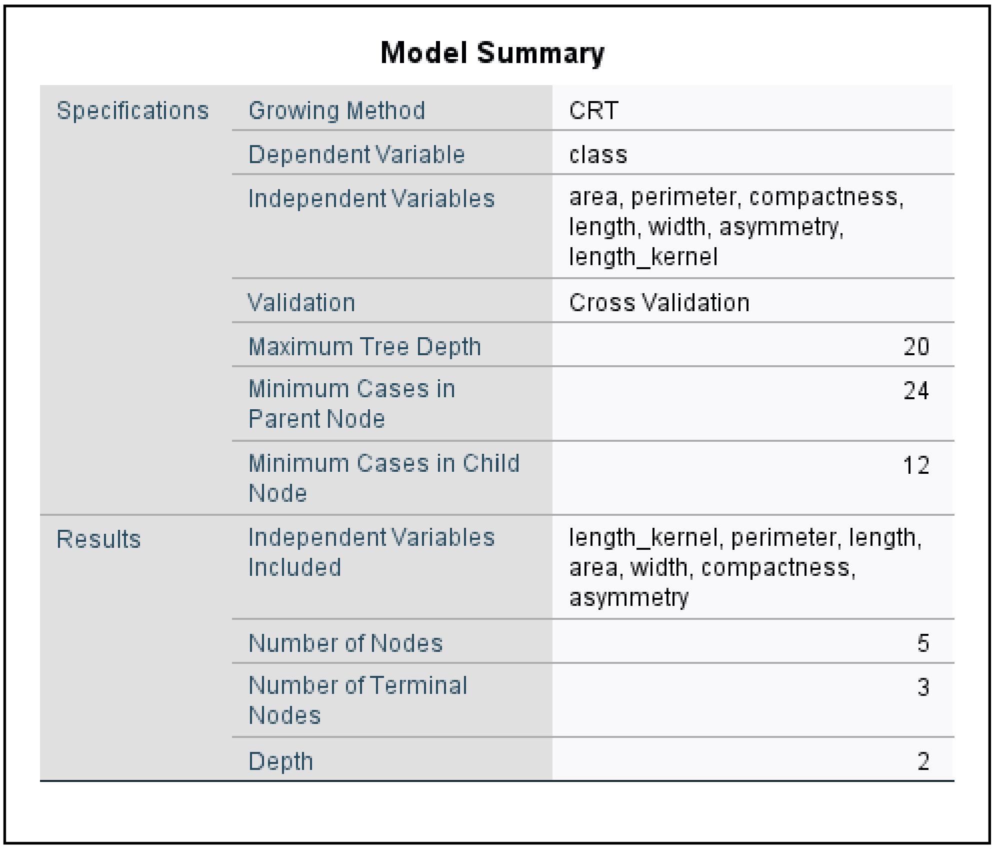
Observed	Predicted 1.000	 2.000	 3.000	 Percent Correct
1.000	55	1	14	78.6%
2.000	2	68	0	97.1%
3.000	0	0	70	100.0%
Overall Percentage	27.1%	32.9%	40.0%	91.9%

Growing Method: CRT
Dependent Variable: class

According to the above classification matrix, the accuracy rate of this tree is 91.9%

When Parent Node = 24, Child Node = 12

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

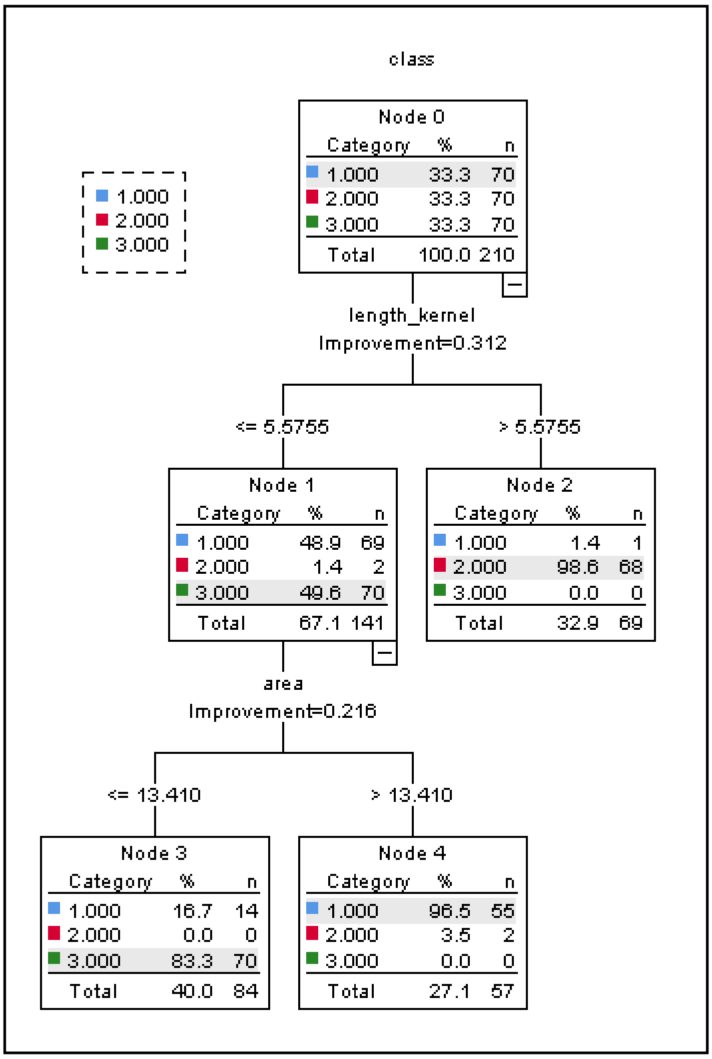


Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	class
	Independent Variables	area, perimeter, compactness, length, width, asymmetry, length_kernel
	Validation	Cross Validation
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	24
	Minimum Cases in Child Node	12
Results	Independent Variables Included	length_kernel, perimeter, length, area, width, compactness, asymmetry
	Number of Nodes	5
	Number of Terminal Nodes	3
	Depth	2

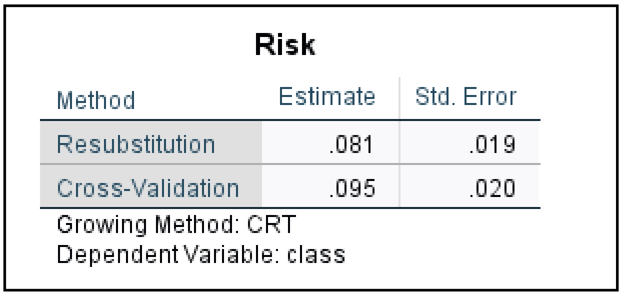
The decision tree has 5 nodes out of which 3 are terminal nodes. The depth of the tree is 2.

Decision Tree



As can be seen in the tree, the first node is split based on our most important predictor, length_kernel. The value of impurity is calculated based on Gini with length_kernel node having a value of 0.312.

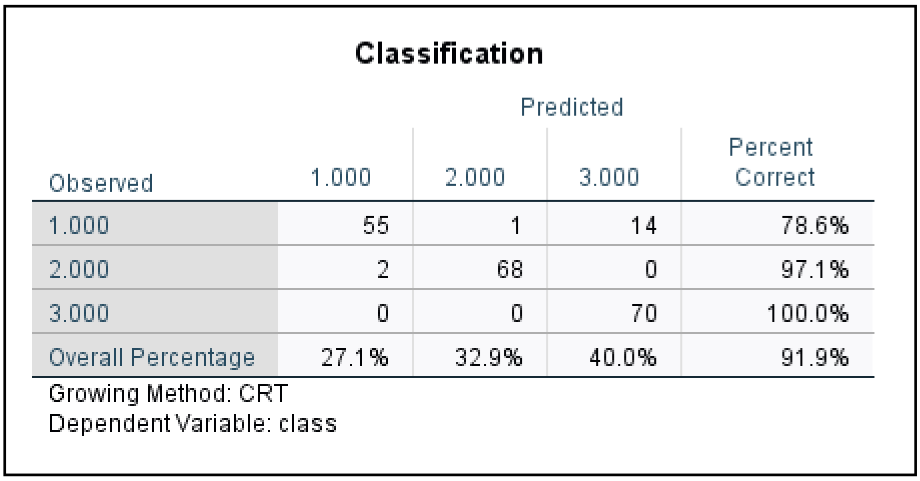
The variable with highest reduction of impurity is selected as splitting attribute.



Risk

Method	Estimate	Std. Error
Resubstitution	.081	.019
Cross-Validation	.095	.020

Growing Method: CRT
Dependent Variable: class



Classification

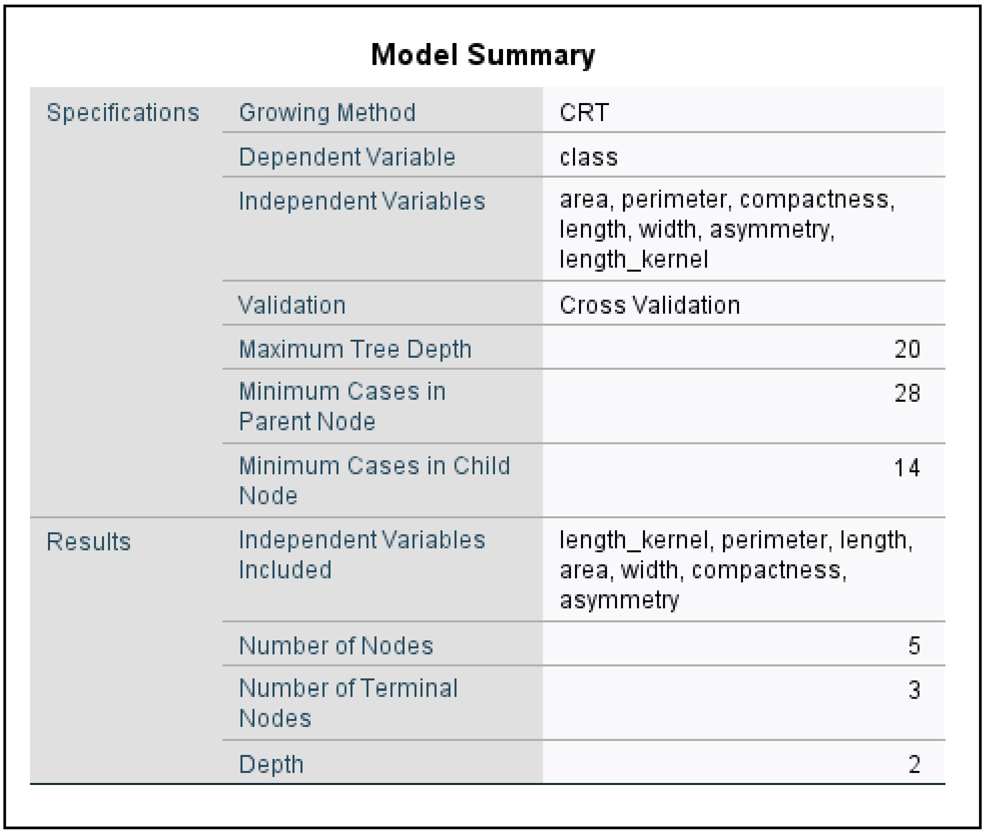
Observed	Predicted 1.000	2.000	3.000	Percent Correct
1.000	55	1	14	78.6%
2.000	2	68	0	97.1%
3.000	0	0	70	100.0%
Overall Percentage	27.1%	32.9%	40.0%	91.9%

Growing Method: CRT
Dependent Variable: class

According to the above classification matrix, the accuracy rate of this tree is 91.9%

When Parent Node = 28, Child Node = 14

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

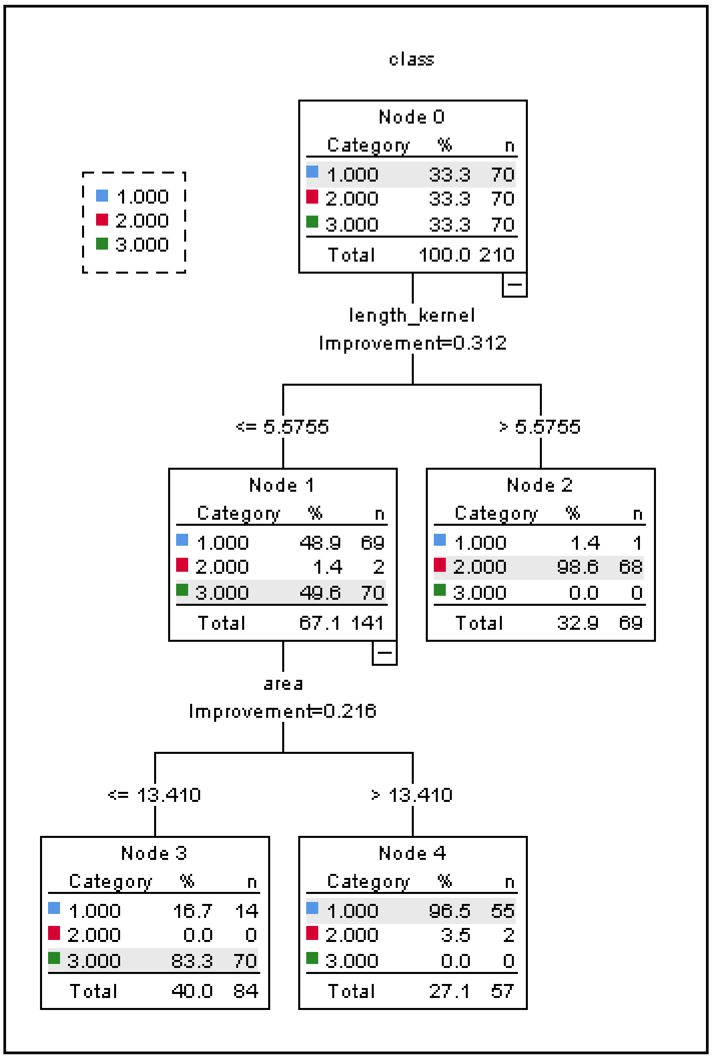


Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	class
	Independent Variables	area, perimeter, compactness, length, width, asymmetry, length_kernel
	Validation	Cross Validation
	Maximum Tree Depth	20
	Minimum Cases in Parent Node	28
	Minimum Cases in Child Node	14
Results	Independent Variables Included	length_kernel, perimeter, length, area, width, compactness, asymmetry
	Number of Nodes	5
	Number of Terminal Nodes	3
	Depth	2

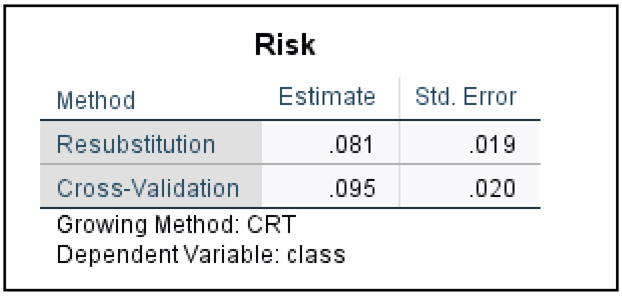
The decision tree has 5 nodes out of which 3 are terminal nodes. The depth of the tree is 2.

Decision Tree



As can be seen in the tree, the first node is split based on our most important predictor, length_kernel. The value of impurity is calculated based on Gini with length_kernel node having a value of 0.312.

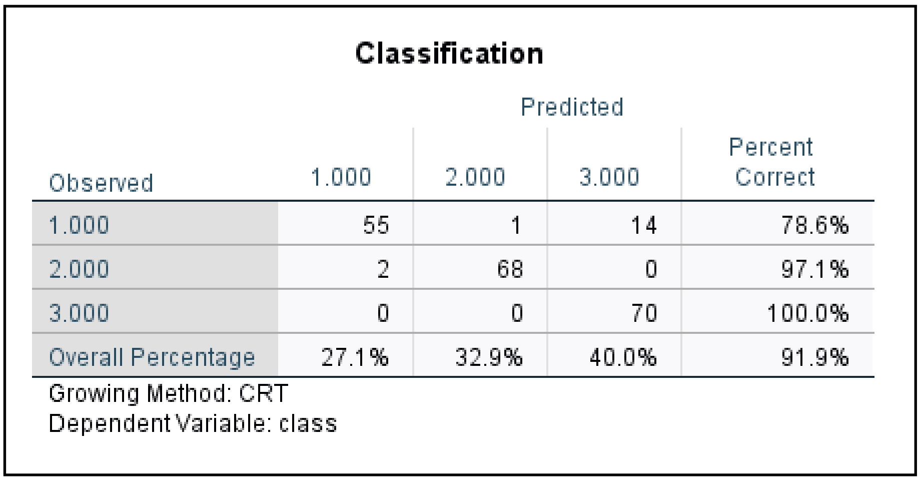
The variable with highest reduction of impurity is selected as splitting attribute.



Risk

Method	Estimate	Std. Error
Resubstitution	.081	.019
Cross-Validation	.095	.020

Growing Method: CRT
Dependent Variable: class



Classification

	Predicted			
Observed	1.000	2.000	3.000	Percent Correct
1.000	55	1	14	78.6%
2.000	2	68	0	97.1%
3.000	0	0	70	100.0%
Overall Percentage	27.1%	32.9%	40.0%	91.9%

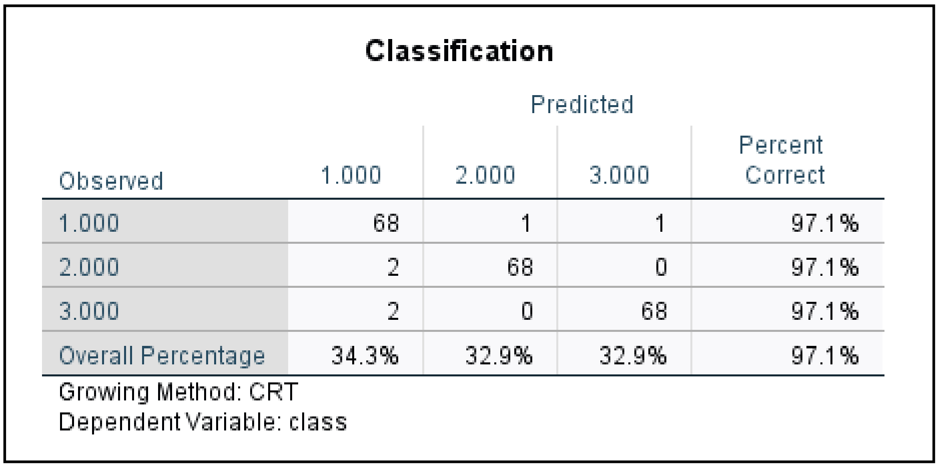
Growing Method: CRT
Dependent Variable: class

According to the above classification matrix, the accuracy rate of this tree is 91.9%

Section b)

There are various ways of measuring model performance (precision, recall, F1 Score, ROC Curve, etc).
Accuracy is one of the simple metric of measuring the performance of the model.

The best model is obtained when parent node = 8 and child node = 4. The decision tree has a best accuracy rate of 97.1%.



Classification

	Predicted			
Observed	1.000	2.000	3.000	Percent Correct
1.000	68	1	1	97.1%
2.000	2	68	0	97.1%
3.000	2	0	68	97.1%
Overall Percentage	34.3%	32.9%	32.9%	97.1%

Growing Method: CRT
Dependent Variable: class

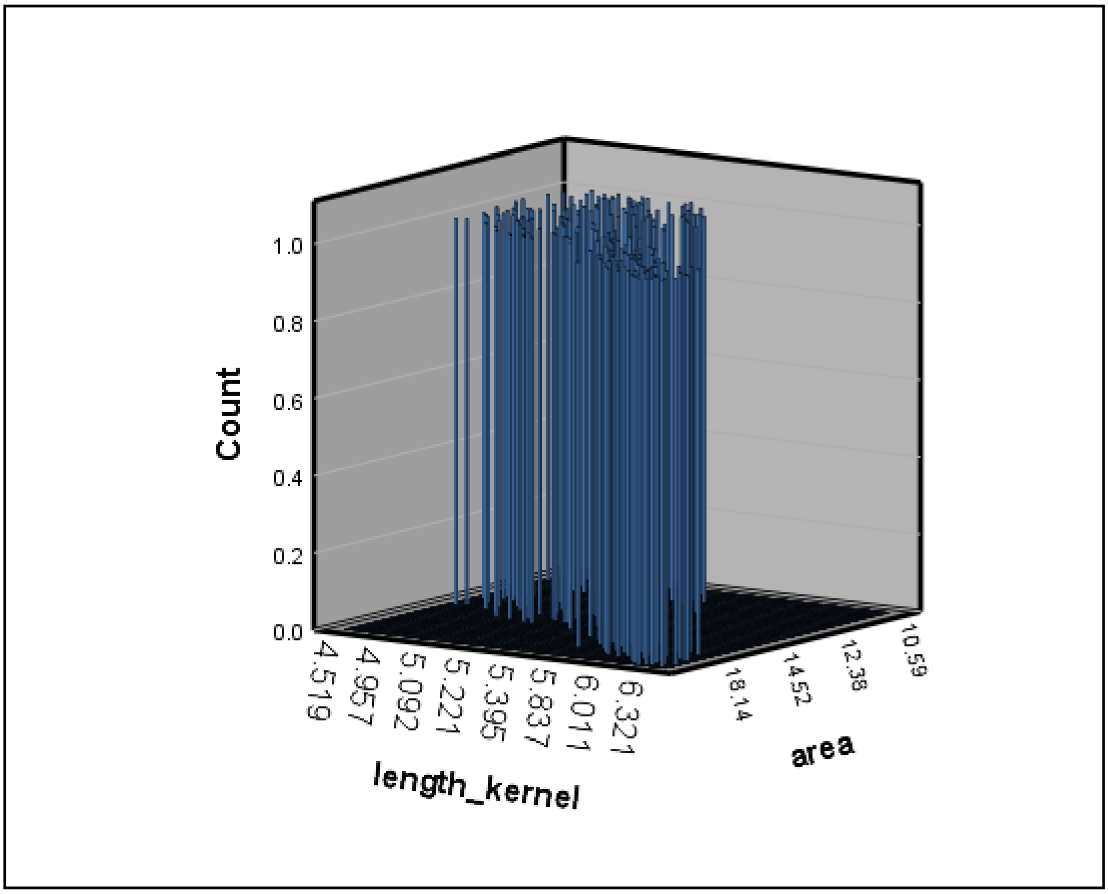
The decision tree has a few misclassification with class 1 having one each record misclassified as class 2 and class 3.
Class 2 has a couple of records misclassified as class 1 while class 3 has a couple of records misclassified as class 3.

Section c)

The three most important attributes for classifying wheat are as follows

- length_kernel: Index value – 0.312
- area: Index value – 0.216
- asymmetry: Index value – 0.001

Section d)



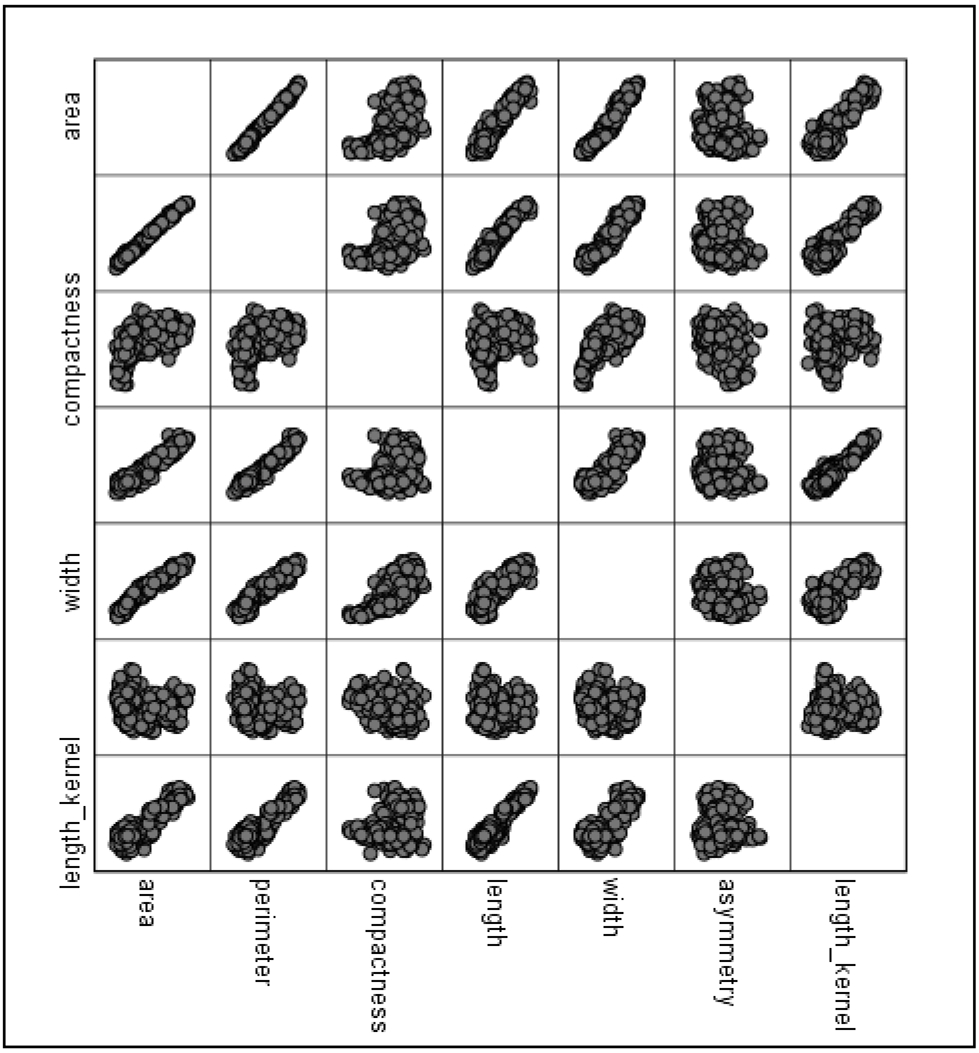
As can be seen in the above graph, we have used the most important variable area and length of the kernel as our x and z axis.

The plot is showing the number of cases based on the relationship between these two variables.

We can majority of the data point are right along the line with a few outliers.

Section e)

Correlation between variables



The graph shows us the relationship between various variables.
We can use this graph to detect multicollinearity between various variables.

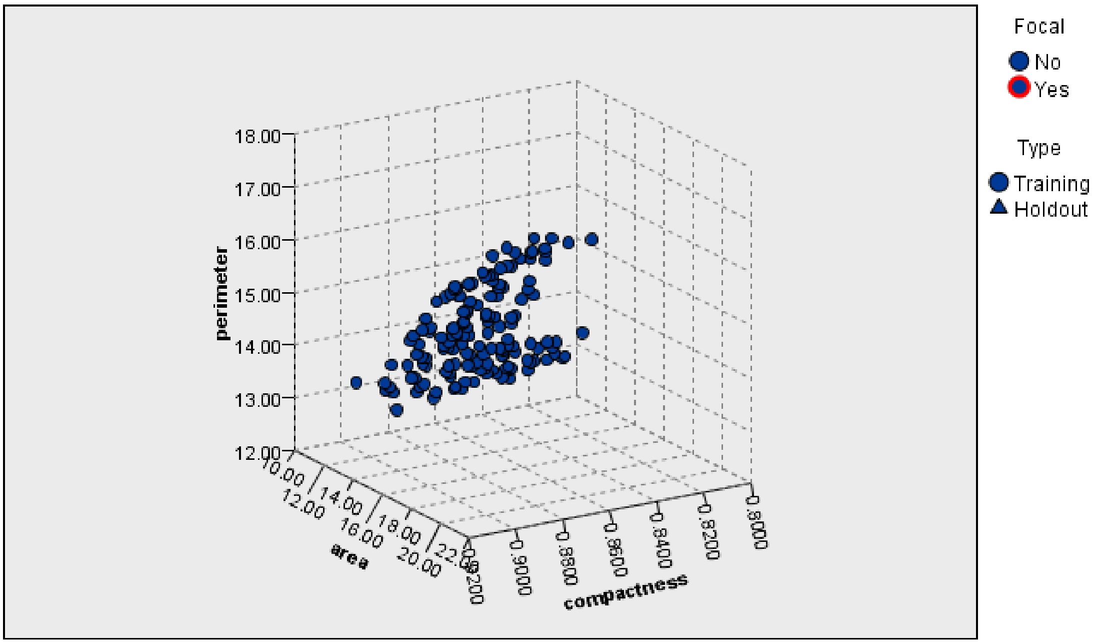
The graph shows a strong positive correlation between variables as below.
Area and perimeter
Area and length
Area and width
Length and perimeter
Length and length_kernel

The following variables show some kind of correlation between each other.
Length_kernel and area
Length_kernel and perimeter
Width and perimeter

K Nearest Neighbour

Predictor Space

Built Model: 3 selected predictors, K = 3



Select points to use as focal records

This chart is a lower-dimensional projection of the predictor space, which contains a total of 7 predictors.

As can be seen in the graph, the three most important predictors are used for classification.