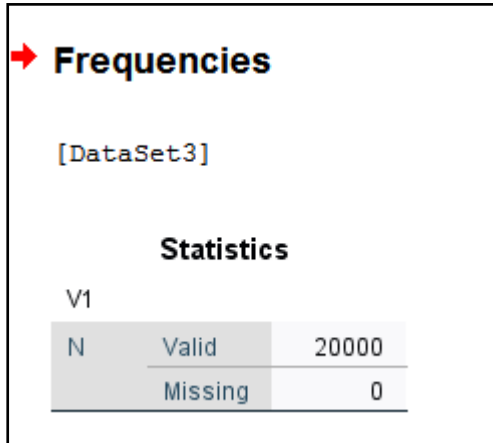


Classification Model

Section a)

Following is the class distribution for the diagnosis letter column (lettr capital letter (26 values from A to Z)).

There are total of 20000 rows in this dataset.



The screenshot shows the SPSS 'Frequencies' window for variable V1. It displays the dataset name '[DataSet3]' and the variable 'V1'. Below this, a table titled 'Statistics' shows the distribution of V1. The table has two rows: 'Valid' with a count of 20000, and 'Missing' with a count of 0.

Statistics		
V1		
N	Valid	20000
	Missing	0

Following is the distribution of the records between the different variables.

Letter					
		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	789	3.9	3.9	3.9
	B	766	3.8	3.8	7.8
	C	736	3.7	3.7	11.5
	D	805	4.0	4.0	15.5
	E	768	3.8	3.8	19.3
	F	775	3.9	3.9	23.2
	G	773	3.9	3.9	27.1
	H	734	3.7	3.7	30.7
	I	755	3.8	3.8	34.5
	J	747	3.7	3.7	38.2
	K	739	3.7	3.7	41.9
	L	761	3.8	3.8	45.7
	M	792	4.0	4.0	49.7
	N	783	3.9	3.9	53.6
	O	753	3.8	3.8	57.4
	P	803	4.0	4.0	61.4
	Q	783	3.9	3.9	65.3
	R	758	3.8	3.8	69.1
	S	748	3.7	3.7	72.8
	T	796	4.0	4.0	76.8
	U	813	4.1	4.1	80.9
	V	764	3.8	3.8	84.7
	W	752	3.8	3.8	88.5
	X	787	3.9	3.9	92.4
	Y	786	3.9	3.9	96.3
	Z	734	3.7	3.7	100.0
	Total	20000	100.0	100.0	

The data is divided into training and testing sample with an 80 – 20 % ratio.

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	lettr
	Independent Variables	x_box, y_box, width, high, onpix, x_bar, y_bar, x2bar, y2bar, xybar, x2ybr, xy2br, x_ege, xegvy, y_ege, yegvx
	Validation	Split Sample
	Maximum Tree Depth	35
	Minimum Cases in Parent Node	10
	Minimum Cases in Child Node	5
Results	Independent Variables Included	x2ybr, y_bar, xybar, x_bar, y2bar, x_ege, xy2br, x2bar, xegvy, y_ege, yegvx, onpix, high, width, y_box, x_box
	Number of Nodes	1143
	Number of Terminal Nodes	572
	Depth	24

The decision tree for letter prediction data is quite complex.

There are total of 1143 nodes in this decision tree and 572 terminal nodes. The depth of this tree is 24.

Risk Estimate

Risk		
Sample	Estimate	Std. Error
Training	.118	.003
Test	.177	.006
Growing Method: CRT		
Dependent Variable: lettr		

The risk estimate of 0.118 in training sample indicates that the letter category predicted by the model (right or wrong) is wrong for 11.8% of the cases. So the “risk” of misclassifying a letter is approximately 12%.

The risk estimate of 0.177 in testing sample indicates that the letter category predicted by the model (right or wrong) is wrong for 17.7% of the cases. So the “risk” of misclassifying a letter is approximately 18%.

Classification Matrix



classification_10.xlsx

The results in the classification table are consistent with the risk estimate.

The table shows that the training sample classifies approximately 88.2% of the letters correctly (highlighted in Yellow).

The table shows that the testing sample classifies approximately 82.3% of the letters correctly (highlighted in Blue).

Decision Tree

Training Sample



training_10.xml

As can be seen in the train sample, the first node is split based on our most important predictor, x2ybr. The value of impurity is calculated based on Gini with x2ybr node having a value of 0.022. The variable with highest reduction of impurity is selected as splitting attribute.

Testing Sample



testing_10.xml

As can be seen in the test sample, the first node is split based on our most important predictor, x2ybr. The value of impurity is calculated based on Gini with x2ybr node having a value of 0.022.

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	lettr
	Independent Variables	x_box, y_box, width, high, onpix, x_bar, y_bar, x2bar, y2bar, xybar, x2ybr, xy2br, x_ege, xegvy, y_ege, yegvx
	Validation	Split Sample
	Maximum Tree Depth	30
	Minimum Cases in Parent Node	20
	Minimum Cases in Child Node	10
Results	Independent Variables Included	x2ybr, y_bar, xybar, x_bar, x_box, width, y2bar, x_ege, xy2br, x2bar, xegvy, y_ege, yegvx, onpix, y_box, high
	Number of Nodes	701
	Number of Terminal Nodes	351
	Depth	20

The decision tree for letter prediction data is quite complex. There are total of 701 nodes in this decision tree and 351 terminal nodes. The depth of this tree is 20.

Risk Estimate

Risk		
Sample	Estimate	Std. Error
Training	.165	.003
Test	.213	.006
Growing Method: CRT		
Dependent Variable: lettr		

The risk estimate of 0.165 in training sample indicates that the letter category predicted by the model (right or wrong) is wrong for 16.5% of the cases. So the “risk” of misclassifying a letter is approximately 16%.

The risk estimate of 0.213 in testing sample indicates that the letter category predicted by the model (right or wrong) is wrong for 21.3% of the cases. So the “risk” of misclassifying a letter is approximately 21%.

Classification Matrix



classification_20.xlsx

The results in the classification table are consistent with the risk estimate.

The table shows that the training sample classifies approximately 83.5% of the letters correctly (highlighted in Yellow).

The table shows that the testing sample classifies approximately 78.7% of the letters correctly (highlighted in Blue).

Decision Tree

Training Sample



training_20.xml

As can be seen in the train sample, the first node is split based on our most important predictor, x2ybr. The value of impurity is calculated based on Gini with x2ybr node having a value of 0.022. The variable with highest reduction of impurity is selected as splitting attribute.

Testing Sample



testing_20.xml

As can be seen in the test sample, the first node is split based on our most important predictor, x2ybr. The value of impurity is calculated based on Gini with x2ybr node having a value of 0.022.

When Parent Node = 30, Child Node = 15

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

Model Summary		
Specifications	Growing Method	CRT
	Dependent Variable	lettr
	Independent Variables	x_box, y_box, width, high, onpix, x_bar, y_bar, x2bar, y2bar, xybar, x2ybr, xy2br, x_ege, xegvy, y_ege, yegvx
	Validation	Split Sample
	Maximum Tree Depth	30
	Minimum Cases in Parent Node	30
	Minimum Cases in Child Node	15
Results	Independent Variables Included	x2ybr, y_bar, xybar, x_bar, y2bar, x_ege, xy2br, x2bar, xegvy, y_ege, yegvx, high, y_box, x_box, onpix, width
	Number of Nodes	509
	Number of Terminal Nodes	255
	Depth	19

There are total of 509 nodes in this decision tree and 255 terminal nodes. The depth of this tree is 19.

Risk Estimate

Risk		
Sample	Estimate	Std. Error
Training	.203	.003
Test	.239	.007
Growing Method: CRT		
Dependent Variable: lettr		

The risk estimate of 0.203 in training sample indicates that the letter category predicted by the model (right or wrong) is wrong for 20.3% of the cases. So the “risk” of misclassifying a letter is approximately 20%.

The risk estimate of 0.239 in testing sample indicates that the letter category predicted by the model (right or wrong) is wrong for 23.9% of the cases. So the “risk” of misclassifying a letter is approximately 24%.

Classification Matrix



classification_30.xlsx

The results in the classification table are consistent with the risk estimate.

The table shows that the training sample classifies approximately 79.7% of the letters correctly (highlighted in Yellow).

The table shows that the testing sample classifies approximately 76.17% of the letters correctly (highlighted in Blue).

Decision Tree

Training Sample



training_30.xml

As can be seen in the train sample, the first node is split based on our most important predictor, x2ybr. The value of impurity is calculated based on Gini with x2ybr node having a value of 0.022. The variable with highest reduction of impurity is selected as splitting attribute.

Testing Sample



testing_30.xml

As can be seen in the test sample, the first node is split based on our most important predictor, x2ybr. The value of impurity is calculated based on Gini with x2ybr node having a value of 0.022.

When Parent Node = 40, Child Node = 20

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

Model Summary		
Specifications	Growing Method	CRT
	Dependent Variable	lettr
	Independent Variables	x_box, y_box, width, high, onpix, x_bar, y_bar, x2bar, y2bar, xybar, x2ybr, xy2br, x_ege, xegvy, y_ege, yegvx
	Validation	Split Sample
	Maximum Tree Depth	30
	Minimum Cases in Parent Node	40
	Minimum Cases in Child Node	20
Results	Independent Variables Included	x2ybr, y_bar, xybar, x_bar, y2bar, x_ege, xy2br, x2bar, y_ege, xegvy, yegvx, y_box, high, onpix, width, x_box
	Number of Nodes	389
	Number of Terminal Nodes	195
	Depth	21

There are total of 389 nodes in this decision tree and 195 terminal nodes. The depth of this tree is 21.

Risk Estimate

Risk		
Sample	Estimate	Std. Error
Training	.235	.003
Test	.259	.007
Growing Method: CRT		
Dependent Variable: lettr		

The risk estimate of 0.235 in training sample indicates that the letter category predicted by the model (right or wrong) is wrong for 23.5% of the cases. So the “risk” of misclassifying a letter is approximately 23%.

The risk estimate of 0.259 in testing sample indicates that the letter category predicted by the model (right or wrong) is wrong for 25.9% of the cases. So the “risk” of misclassifying a letter is approximately 26%.

Classification Matrix



classification_40.xlsx

The results in the classification table are consistent with the risk estimate.

The table shows that the training sample classifies approximately 76.5% of the letters correctly (highlighted in Yellow).

The table shows that the testing sample classifies approximately 74.1% of the letters correctly (highlighted in Blue).

Decision Tree

Training Sample



training_40.xml

As can be seen in the train sample, the first node is split based on our most important predictor, x2ybr. The value of impurity is calculated based on Gini with x2ybr node having a value of 0.022. The variable with highest reduction of impurity is selected as splitting attribute.

Testing Sample



testing_40.xml

As can be seen in the test sample, the first node is split based on our most important predictor, x2ybr. The value of impurity is calculated based on Gini with x2ybr node having a value of 0.022.

When Parent Node = 50, Child Node = 25

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

Model Summary

Specifications	Growing Method	CRT
	Dependent Variable	lettr
	Independent Variables	x_box, y_box, width, high, onpix, x_bar, y_bar, x2bar, y2bar, xybar, x2ybr, xy2br, x_ege, xegvy, y_ege, yegvx
	Validation	Split Sample
	Maximum Tree Depth	30
	Minimum Cases in Parent Node	50
	Minimum Cases in Child Node	25
Results	Independent Variables Included	x2ybr, y_bar, xybar, x_bar, y2bar, x_ege, xy2br, x2bar, y_ege, xegvy, yegvx, y_box, high, onpix, width, x_box
	Number of Nodes	303
	Number of Terminal Nodes	152
	Depth	18

There are total of 303 nodes in this decision tree and 152 terminal nodes. The depth of this tree is 18.

Risk Estimate

Risk		
Sample	Estimate	Std. Error
Training	.275	.004
Test	.293	.007
Growing Method: CRT		
Dependent Variable: lettr		

The risk estimate of 0.275 in training sample indicates that the letter category predicted by the model (right or wrong) is wrong for 27.5% of the cases. So the “risk” of misclassifying a letter is approximately 27%.

The risk estimate of 0.293 in testing sample indicates that the letter category predicted by the model (right or wrong) is wrong for 29.3% of the cases. So the “risk” of misclassifying a letter is approximately 29%.

Classification Matrix



classification_50.xlsx

The results in the classification table are consistent with the risk estimate.

The table shows that the training sample classifies approximately 72.5% of the letters correctly (highlighted in Yellow).

The table shows that the testing sample classifies approximately 70.7% of the letters correctly (highlighted in Blue).

Decision Tree

Training Sample



training_50.xml

As can be seen in the train sample, the first node is split based on our most important predictor, x2ybr. The value of impurity is calculated based on Gini with x2ybr node having a value of 0.022. The variable with highest reduction of impurity is selected as splitting attribute.

Testing Sample



testing_50.xml

As can be seen in the test sample, the first node is split based on our most important predictor, x2ybr. The value of impurity is calculated based on Gini with x2ybr node having a value of 0.022.

When Parent Node = 60, Child Node = 30

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

Model Summary		
Specifications	Growing Method	CRT
	Dependent Variable	lettr
	Independent Variables	x_box, y_box, width, high, onpix, x_bar, y_bar, x2bar, y2bar, xybar, x2ybr, xy2br, x_ege, xegvy, y_ege, yegvx
	Validation	Split Sample
	Maximum Tree Depth	30
	Minimum Cases in Parent Node	60
	Minimum Cases in Child Node	30
Results	Independent Variables Included	x2ybr, y_bar, xybar, x_bar, y2bar, x_ege, xy2br, x2bar, xegvy, y_ege, yegvx, width, high, y_box, x_box, onpix
	Number of Nodes	243
	Number of Terminal Nodes	122
	Depth	16

There are total of 243 nodes in this decision tree and 122 terminal nodes. The depth of this tree is 16.

Risk Estimate

Risk		
Sample	Estimate	Std. Error
Training	.286	.004
Test	.306	.007
Growing Method: CRT		
Dependent Variable: lettr		

The risk estimate of 0.286 in training sample indicates that the letter category predicted by the model (right or wrong) is wrong for 28.6% of the cases. So the “risk” of misclassifying a letter is approximately 29%.

The risk estimate of 0.306 in testing sample indicates that the letter category predicted by the model (right or wrong) is wrong for 30.6% of the cases. So the “risk” of misclassifying a letter is approximately 31%.

Classification Matrix



classification_60.xlsx

The results in the classification table are consistent with the risk estimate.

The table shows that the training sample classifies approximately 71.4% of the letters correctly (highlighted in Yellow).

The table shows that the testing sample classifies approximately 69.4% of the letters correctly (highlighted in Blue).

Decision Tree

Training Sample



training_60.xml

As can be seen in the train sample, the first node is split based on our most important predictor, x2ybr. The value of impurity is calculated based on Gini with x2ybr node having a value of 0.022. The variable with highest reduction of impurity is selected as splitting attribute.

Testing Sample



testing_60.xml

As can be seen in the test sample, the first node is split based on our most important predictor, x2ybr. The value of impurity is calculated based on Gini with x2ybr node having a value of 0.022.

Conclusion: According to me, the model should at least have an accuracy of 80% and minimum misclassification error.

Also the accuracy rate predicted by the training and testing set should have a minimum difference. The best configuration model is obtained when the number parent node is fixed at 20 and child node at 10. The accuracy rate for this model is around 80% and the difference of misclassification matrix between the training and testing set is minimum.

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.

We have an accuracy of around 79% on the training set and ~ 76% on the testing set when parent node = 30 and child node = 15. Also there is minimum difference of misclassification error between the training and testing set.

The difference in accuracy rate between the training and testing set drastically reduces from 4.8% when parent node = 20 and child node = 10 to 3.6% when parent node = 30 and child node = 15. Therefore I am choosing this as my best model.

Following is the misclassification matrix when parent node = 30 and child node = 15

Misclassification Matrix

Risk		
Sample	Estimate	Std. Error
Training	.203	.003
Test	.239	.007
Growing Method: CRT		
Dependent Variable: lettr		

The risk estimate of 0.203 in training sample indicates that the letter category predicted by the model (right or wrong) is wrong for 20.3% of the cases. So the “risk” of misclassifying a letter is approximately 20%.

The risk estimate of 0.239 in testing sample indicates that the letter category predicted by the model (right or wrong) is wrong for 23.9% of the cases. So the “risk” of misclassifying a letter is approximately 24%.

Section c)

The three most important attributes for recognizing letters are as follows

- x2ybr: Index value – 0.022
- y-bar: Index value – 0.019
- y2bar: Index value – 0.023