Name: Amy Aumpansub
Due: May 22, 2018

IS 467: Assignment 3

Assignment 3

Problem 1

1) Build the best decision tree you can and explain what makes it the best. Show what criteria you used including the number of cases in parents and children and depth and stopping condition.

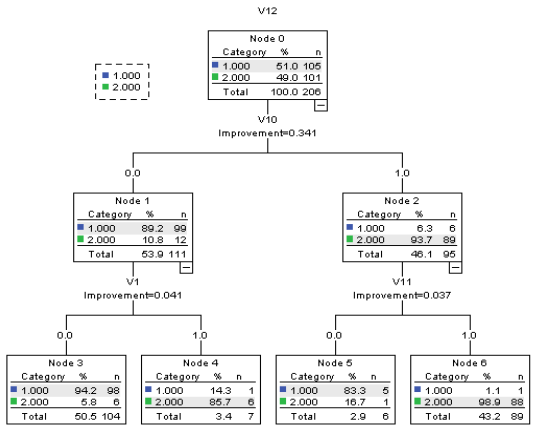
Model Summary

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, V7, V9, V11, V6, V3, V8, V4, V5, V2
	Number of Nodes	7
	Number of Terminal Nodes	4
	Depth	2

Risk

Sample	Estimate	Std. Error
Training	.044	.014
Test	.064	.025

Growing Method: CRT
Dependent Variable: V12



The final decision tree is built from the training set (70% of data) using CART with Gini index with the threshold of 0.0001, so it will stop if the current node does not improve impurity beyond 0.0001. For the stopping condition, the pre-pruning is used to avoid overfitting. I use the number of parent node of 10 and the number of child node of 5, and the maximum depth of 5. So, it will stop splitting if the parent node has less than 10 cases and the child node contains less than 5 cases. Post-pruning is also used in SPSS.

This is the best decision tree because it has the lowest error rate and the highest accuracy rate. The model error rate is 0.044 for training set which is less than 0.05. The error rate for test set of 0.064 is slightly higher. Overall, this tree provides a high accuracy as the accuracy rates for training set and test set are 95.6% and 93.6%, respectively.

2) How many nodes does the final tree have and how many of them are terminal nodes?

The final tree has 7 nodes, in which there are 4 terminal nodes.

3) What are the most important three Lupus data features in building the tree?

Independent Variable Importance

Independent Variable	Importance	Normalized Importance
V10	.341	100.0%
V11	.217	63.5%
V1	.204	59.7%

The most important features are ranked as following:
1st V10(variable #10), 2nd V11(variable #11), 3rd V1 (variable #1).
As the table shows the percent of importance, the V10 has the highest position in the final tree and is the most important (34.1%) variable for classification. The higher position, the more importance of variables.

4) Increase the parameters that let you set the number of cases allowed in parent and child nodes.

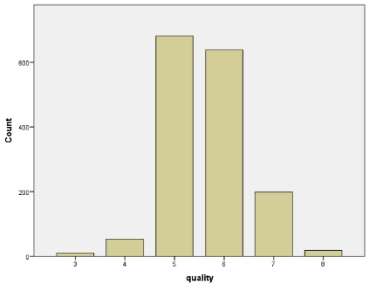
When we increase the parameters, we increase the minimum number of cases in parent and child nodes. So, the complexity of tree and the number of node decrease. The algorithm will stop splitting further if those nodes contains the cases fewer than the threshold. So, the larger threshold makes a smaller tree, but may provide a lower accuracy. For example, when we increase the parameters for the number of parent node from 10 to 20 and child node from 5 to 10, the new tree is smaller than the final model in 1) and contains only 5 nodes, but the new tree has a higher error rate.

Name: Amy Aumpansub
Due: May 22, 2018

IS 467: Assignment 3

Problem 2

1) **Consider each quality level of wine to be a different class. Report how many classes there are and what is the distribution of these classes for the red wine data**



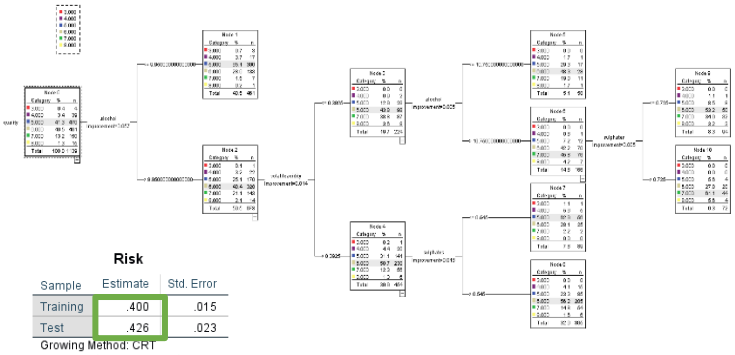
		Frequency	Percent
Valid	3	10	.6
	4	53	3.3
	5	681	42.6
	6	638	39.9
	7	199	12.4
	8	18	1.1
	Total	1599	100.0

The red wine data contains the quality level 3 to 8. Thus, there are 6 classes. This class variable is imbalanced because the class "level 5" and "level 6" have a very high frequency, which contains the majority cases of data. The frequency table above shows the frequency of each class. Class "3" contain 10 cases. Class "4" contains 53 cases. Class "5" contains 681 cases. Class "6" contains 638 cases. Class "7" contains 199 cases. Class "8" contains 18 cases.

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	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

2.1) The final tree is built from the training data (70% of data) using CART with 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, respectively. The maximum depth is 20. I select this tree because compared to other models it has the lower error rate of 40% for the train set and 42.6% for the test set. However, this final tree does not have a high accuracy rate (60% rate for training set).

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:
1st alcohol, 2nd sulphates, 3rd totalsulfurdioxide. As the table shows the percent of importance, the alcohol has the highest position in the final tree and is the most important (6.3%) variable for classification. The higher position, the more importance of variables.

Independent Variable Importance

Independent Variable	Importance	Normalized Importance
alcohol	.063	100.0%
sulphates	.059	93.8%
totalsulfurdioxide	.028	43.9%
volatileacidity	.020	31.5%
citricacid	.017	27.7%
density	.009	14.2%
fixedacidity	.005	7.4%
freesulfurdioxide	.004	6.4%
chlorides	.003	5.0%
residualsugar	.002	2.5%
pH	.001	1.8%

Growing Method: CRT
Dependent Variable: quality

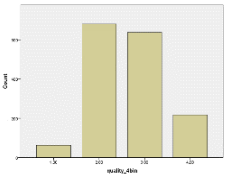
2.4) When we increase the parameters, we increase the minimum number of cases in parent and child nodes. So, the complexity of tree and the number of nodes will decrease. This is because the algorithm will stop splitting further if those nodes contains the cases (tuples) fewer than the threshold. So, the larger threshold makes a smaller tree. For example, when we increase the parameters for the number of parent node from 100 to 200 and child node from 50 to 100, the new tree is smaller than the final model in 2.1) and contains only 7 nodes, but it also has a lower accuracy.

Name: Amy Aumpansub
Due: May 22, 2018

IS 467: Assignment 3

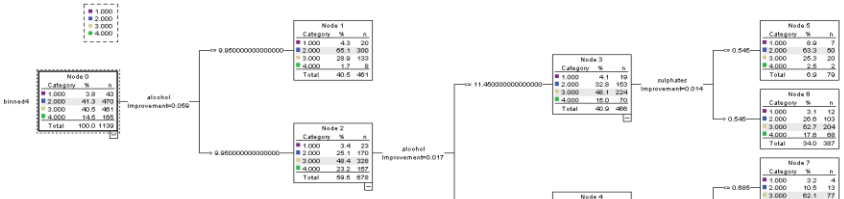
3. Now bin the class variable in such a way that data is not so imbalanced with respect to the class variable. Repeat Problem 1 but on the data you have processed with this smoothing.

The original class variable has the level 3 to 8, which is imbalanced because the class level 5 and 6 contains majority of data. First, I bin the class variable into 4 bins including level 3-4 for bin 1, level 5 for bin 2, level 6 for bin 3, and level 7-8 for bin 4. The frequency of binned class is as following Bin 1(3.9%), Bin 2(42.6%), Bin 3(39.9%), Bin 4(13.6%). The decision tree slightly improves after binning.



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

3.1) The final tree is built using CART with Gini index (0.0001 for impurity threshold) with 70% of data (training set). For the stopping condition, the minimum number of cases for parent and child nodes are 100 and 50, respectively. The maximum depth is 20. I select this tree because compared to other models it has the lower error rate of 26.23%. 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:
1st alcohol, 2nd sulphates, 3rd totalsulfurdioxide. As the table shows the percent of importance, the alcohol has the highest position in the final tree and is the most important (7.6% of importance rate) for classification. The higher position, the more importance of variables.

Independent Variable Importance

Independent Variable	Importance	Normalized Importance
alcohol	.076	100.0%
sulphates	.056	74.1%
totalsulfurdioxide	.023	30.2%
density	.010	12.6%
citricacid	.007	9.6%
volatileacidity	.006	8.2%
fixedacidity	.004	5.8%
chlorides	.004	5.0%
residualsugar	.003	3.3%
freesulfurdioxide	.002	2.3%
pH	.001	1.5%

Growing Method: CRT
Dependent Variable: binned4

3.4) When we increase the parameters, we increase the minimum number of cases in parent and child nodes. So, the complexity of tree and the number of nodes will decrease. This is because the algorithm will stop splitting further if those nodes contains the cases (tuples) fewer than the threshold. So, the larger threshold makes a smaller tree, but it may have a lower accuracy.

4. How does the performance of the original class variable compare with the binned class variable?
The original class variable has a poorer performance because it has a class imbalance, which the model does not detect any samples from class level "3","4", "8" because these classes contains only few cases. The performance of binned variable is better than the original variable because the binned class is more balanced than the original variable. The error rate of the new model from binned class variable is lower. The accuracy rate of the binned class variable is higher. 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, indicating that the classifier has a higher accuracy to recognize classes of binned variable.

Original Variable

Sample	Observed	Predicted 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 Variable

Sample	Observed	Predicted 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. (10% bonus credit) Do you have any other ideas on how you can improve the results further?

First, we can improve the model by assigning more weights for a class that is more important or has a high frequency using the option function in classify trees (SPSS).

Second, we can further bin the class variable into 2 bins. First bin contains the class level 3-5. Second bin contains the class level 6-8. If we bin the original class into 2 bins, we can further improve the model by applying the threshold-moving method to select the new decision threshold. ROC curve can help to select the new optimal threshold, then select the points on ROC curve as the new threshold regarding the sensitivity and specificity rate, which will improve the accuracy of the tree.

Problem 3

a) Feature selection is an approach of selecting a subset from original features, so that the probability distribution of different classes of those features is as close as to the original distribution of all features. Whereas, feature extraction is an approach of transforming the existing features into a lower dimension space or extracting a set of new uncorrelated attributes which can still capture most variance of the data.

b) Training data is the set of tuples that is used to build up the model. Whereas the testing data is held back from the training of the model, thus it is independent. It is used to give an unbiased evaluation of the model built from training data. The test set is also called validation set if it is used to select the model.

c) Parametric reduction techniques such as regression assume a fitted model for the data, in which this technique determines the parameters, store only the parameters, and discard the data.
Non-parametric reduction techniques do not assume the model for the data when selecting a smaller form that represents the entire dataset such as clustering.

d) Uniform binning divides ranges into interval of equal size, so that every bin has the same width. Whereas, non-uniform binning combines bins until they are reached a certain threshold of counts.

e) Covariance matrix is a matrix in which its element in i, j represents the covariance between i-th and j-th elements of a random vector (assessing how much variables change together). While, correlation matrix shows the correlation coefficient between variables, used to examine the strength relationship between variables. Each element of correlation matrix represents the correlation coefficient between variables and can be thought as the covariance of standardized variables.