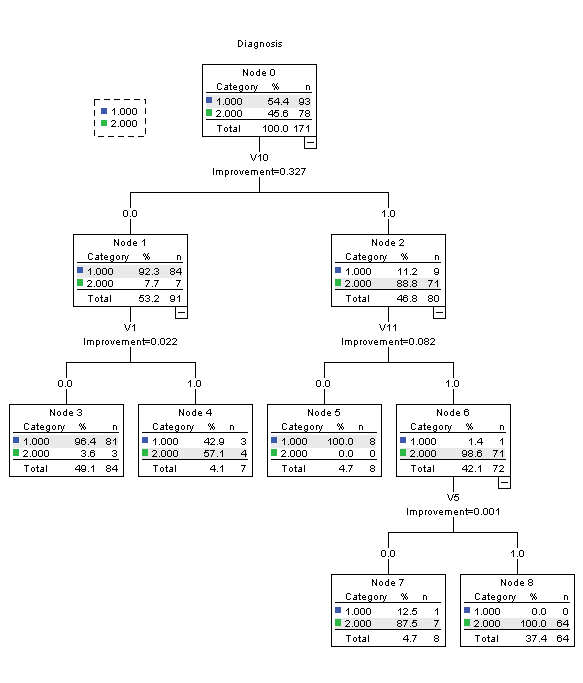
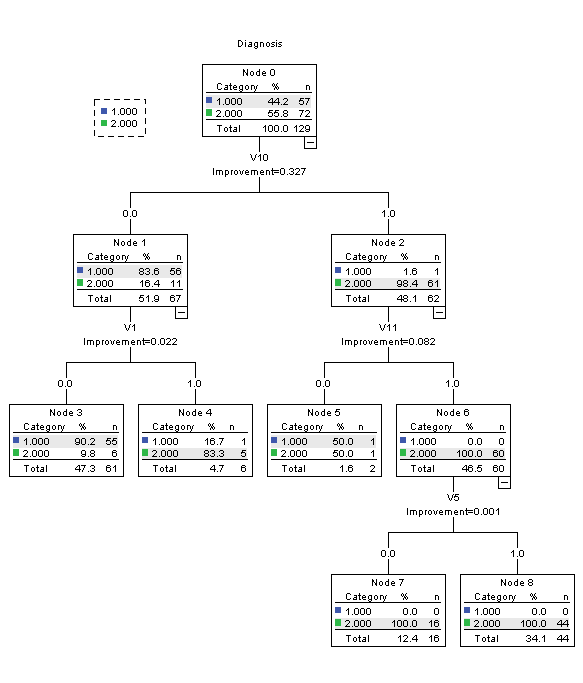
1.

1.1. Train set:



The final output for the test set:



For the criteria, I used the following:

|  |  |  |
| --- | --- | --- |
| **Model Summary** | | |
| Specifications | Growing Method | CRT |
| Dependent Variable | Diagnosis |
| 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 | 30 |
| Minimum Cases in Child Node | 6 |
| Results | Independent Variables Included | V10, V1, V9, V7, V3, V6, V11, V5, V4, V8, V2 |
| Number of Nodes | 9 |
| Number of Terminal Nodes | 5 |
| Depth | 3 |

I used the automatic stopping condition. I also used 60% training data and 40% test data split.

This is the best decision tree because it has the lower risk and std error than other decision trees that I created. It’s also 93.8% correct on the test data set.

|  |  |  |
| --- | --- | --- |
| **Risk** | | |
| Sample | Estimate | Std. Error |
| Training | .041 | .015 |
| Test | .062 | .021 |
| Growing Method: CRT  Dependent Variable: Diagnosis | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification** | | | | |
| Sample | Observed | Predicted | | |
| 1 | 2 | Percent Correct |
| Training | 1 | 89 | 4 | 95.7% |
| 2 | 3 | 75 | 96.2% |
| Overall Percentage | 53.8% | 46.2% | 95.9% |
| Test | 1 | 56 | 1 | 98.2% |
| 2 | 7 | 65 | 90.3% |
| Overall Percentage | 48.8% | 51.2% | 93.8% |
| Growing Method: CRT  Dependent Variable: Diagnosis | | | | |

1.2. The final tree has 9 nodes and 5 terminal nodes.

1.3.

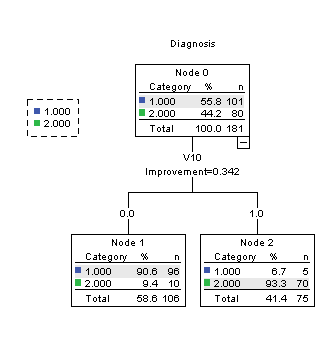
|  |  |  |
| --- | --- | --- |
| **Independent Variable Importance** | | |
| Independent Variable | Importance | Normalized Importance |
| V10 | .327 | 100.0% |
| V11 | .259 | 79.1% |
| V1 | .181 | 55.4% |
| V9 | .147 | 44.8% |
| V7 | .141 | 43.0% |
| V3 | .101 | 30.8% |
| V6 | .100 | 30.5% |
| V5 | .081 | 24.8% |
| V4 | .069 | 21.1% |
| V8 | .048 | 14.6% |
| V2 | .040 | 12.2% |
| Growing Method: CRT  Dependent Variable: Diagnosis | | |

The most important variables are V10, V11 and V1 because they have the highest importance.

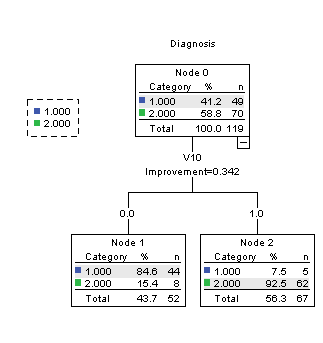
4.

I increased the parameters to parent = 50 and child =10. This decreases the number of nodes because it allows for more tuples to be included in each parent and child node.

Train set:



Test set:



|  |  |  |
| --- | --- | --- |
| **Risk** | | |
| Sample | Estimate | Std. Error |
| Training | .083 | .020 |
| Test | .109 | .029 |
| Growing Method: CRT  Dependent Variable: Diagnosis | | |

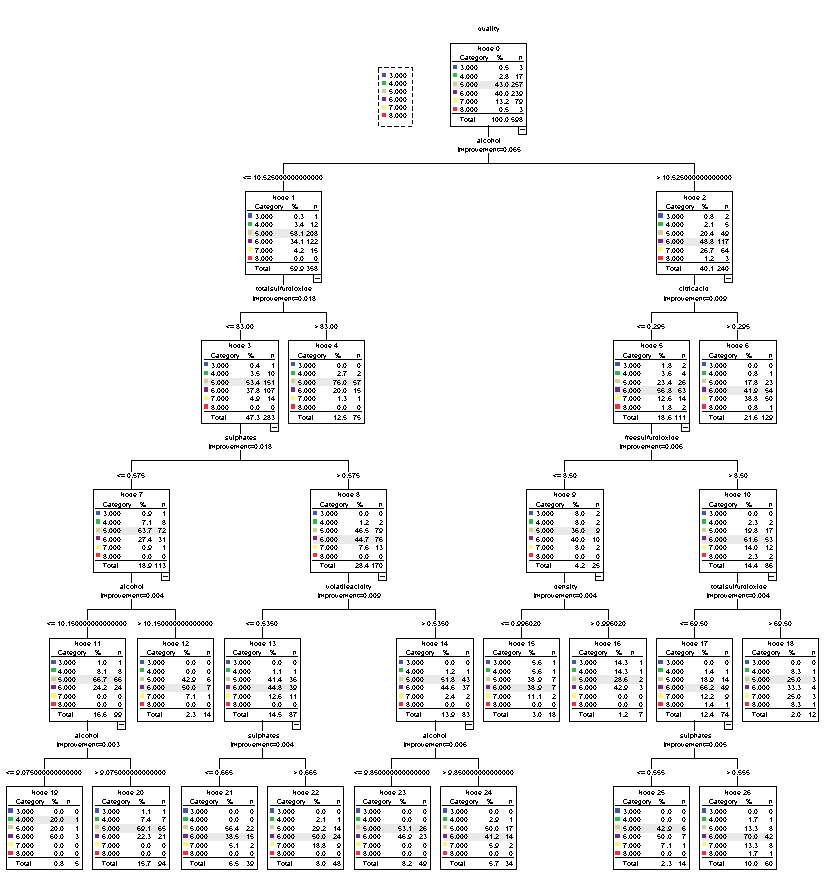
2.

2.1. This distribution for red wine is:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **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 |  |

There are 6 classes. With most of the data points being between 5-7.

* 1. Repeat the Decision tree you used on the Lupus data



I used the same specifications as with the Lupus data to build this decision tree. The final number of nodes is 27.

|  |  |  |
| --- | --- | --- |
| **Risk** | | |
| Sample | Estimate | Std. Error |
| Training | .370 | .015 |
| Test | .460 | .020 |
| Growing Method: CRT  Dependent Variable: quality | | |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Classification** | | | | | | | | |
| Sample | Observed | Predicted | | | | | | |
| 3 | 4 | 5 | 6 | 7 | 8 | Percent Correct |
| Training | 3 | 0 | 1 | 3 | 3 | 0 | 0 | 0.0% |
| 4 | 0 | 2 | 20 | 14 | 0 | 0 | 5.6% |
| 5 | 0 | 1 | 308 | 115 | 0 | 0 | 72.6% |
| 6 | 0 | 2 | 76 | 321 | 0 | 0 | 80.5% |
| 7 | 0 | 0 | 8 | 112 | 0 | 0 | 0.0% |
| 8 | 0 | 0 | 1 | 14 | 0 | 0 | 0.0% |
| Overall Percentage | 0.0% | 0.6% | 41.6% | 57.8% | 0.0% | 0.0% | 63.0% |
| Test | 3 | 0 | 0 | 2 | 1 | 0 | 0 | 0.0% |
| 4 | 0 | 1 | 11 | 5 | 0 | 0 | 5.9% |
| 5 | 0 | 1 | 159 | 97 | 0 | 0 | 61.9% |
| 6 | 0 | 3 | 73 | 163 | 0 | 0 | 68.2% |
| 7 | 0 | 0 | 5 | 74 | 0 | 0 | 0.0% |
| 8 | 0 | 0 | 1 | 2 | 0 | 0 | 0.0% |
| Overall Percentage | 0.0% | 0.8% | 42.0% | 57.2% | 0.0% | 0.0% | 54.0% |
| Growing Method: CRT  Dependent Variable: quality | | | | | | | | |

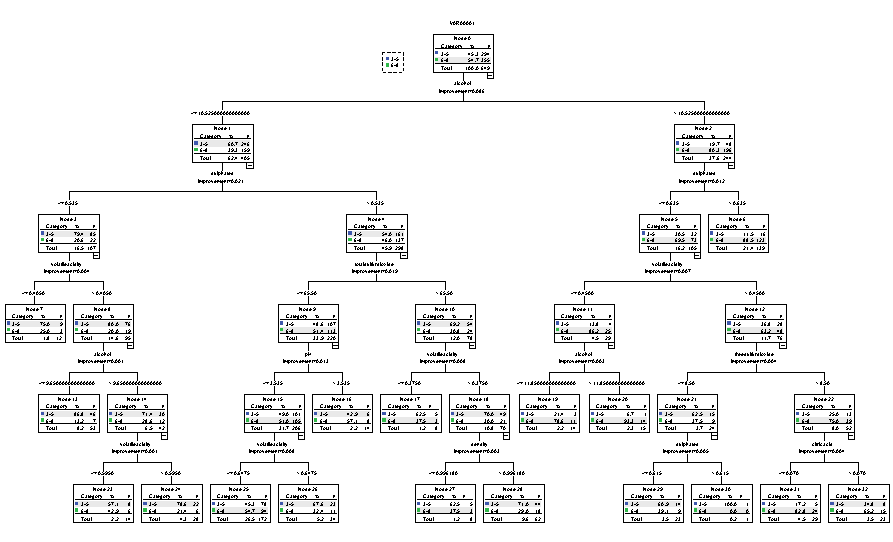
The most important variables were alcohol, sulphates and volatile\_acidity.

|  |  |  |
| --- | --- | --- |
| **Independent Variable Importance** | | |
| Independent Variable | Importance | Normalized Importance |
| alcohol | .086 | 100.0% |
| sulphates | .068 | 79.2% |
| volatileacidity | .065 | 75.7% |
| totalsulfurdioxide | .038 | 44.3% |
| density | .035 | 41.1% |
| chlorides | .021 | 24.7% |
| citricacid | .020 | 23.4% |
| pH | .016 | 18.4% |
| freesulfurdioxide | .014 | 16.8% |
| fixedacidity | .014 | 16.4% |
| residualsugar | .009 | 10.6% |
| Growing Method: CRT  Dependent Variable: quality | | |

* 1. I am using the following bins for the next decision tree

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Bin** | | | | | |
|  | | Frequency | Percent | Valid Percent | Cumulative Percent |
| Valid | 3-5 | 744 | 46.5 | 46.5 | 46.5 |
| 6-8 | 855 | 53.5 | 53.5 | 100.0 |
| Total | 1599 | 100.0 | 100.0 |  |

Test result:



|  |  |  |
| --- | --- | --- |
| **Risk** | | |
| Sample | Estimate | Std. Error |
| Training | .220 | .013 |
| Test | .304 | .018 |
| Growing Method: CRT  Dependent Variable: VAR00001 | | |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Classification** | | | | |
| Sample | Observed | Predicted | | |
| 3-5 | 6-8 | Percent Correct |
| Training | 3-5 | 313 | 137 | 69.6% |
| 6-8 | 72 | 428 | 85.6% |
| Overall Percentage | 40.5% | 59.5% | 78.0% |
| Test | 3-5 | 180 | 114 | 61.2% |
| 6-8 | 83 | 272 | 76.6% |
| Overall Percentage | 40.5% | 59.5% | 69.6% |
| Growing Method: CRT  Dependent Variable: VAR00001 | | | | |

There are 33 nodes and the model has the same 3 most important variables as the first model.

|  |  |  |
| --- | --- | --- |
| **Independent Variable Importance** | | |
| Independent Variable | Importance | Normalized Importance |
| alcohol | .095 | 100.0% |
| volatileacidity | .080 | 84.8% |
| sulphates | .080 | 84.3% |
| citricacid | .040 | 42.1% |
| density | .038 | 40.1% |
| totalsulfurdioxide | .027 | 28.5% |
| chlorides | .026 | 27.6% |
| pH | .020 | 21.0% |
| freesulfurdioxide | .014 | 15.2% |
| residualsugar | .008 | 8.0% |
| fixedacidity | .005 | 5.0% |
| Growing Method: CRT  Dependent Variable: VAR00001 | | |

2.4 The accuracy of the first model was 54% while the accuracy of the second model was 69%. The second model also had a better percent correct than the first model without the binning.

3.

a. Feature selection involves choosing which features or variables you are going to use for your anaylsis. There are a few different ways to choose the variables including forward and backward selection. Feature extraction is using some of the features you have chosen to work with and manipulating them to create variables. One example is the principal component analysis which creates new variables that are linear combinations of the originals.

b. Training data is the data that we use to train our algorithm or decision tree. We then take the algorithm, decision tree, etc and apply it to the test data. Essentially training data trains the data and then the training is applied to the test data.

c. Parametric data reduction uses a model instead of the actual data so only the data parameters need to be stored. Non-parametric data reduction uses graphs or pictures such as histograms and clustering to represent the data.

d. Uniform binning is when you separate the tuples into bins of the same frequency.

Non-uniform binning is when you separate the tuples into bins with the same width, which could result in having bins of different sizes.

e. Covariance matrix and correlation matrix essentially measure the same thing. The correlation matrix shows the linear relationship between variables. The covariance matrix