**ISE465 HW#3 Report**

***Bolun Xu*** [***box215@lehigh.edu***](mailto:box215@lehigh.edu)

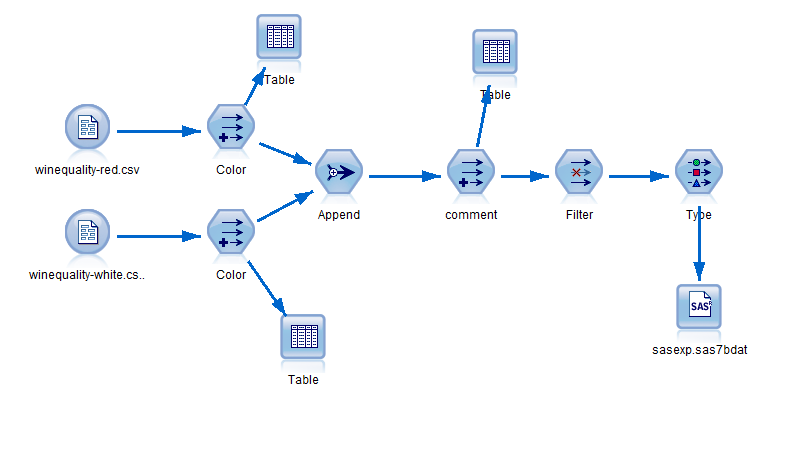
**2.**A:

(1). Use IBM SPSS to make the data fit our requirement.

i) Derive node is used to create a new variable ‘Color’. First data set is red, and second is white.

ii) Append node combines the two given data sets. Then used another Derive node to create the new variable ‘Comment’, set quality>=7 to be T and else to be F.

iii) Import the new data set as SAS file.



(2). Use SAS Enterprise Miner to model.

i) Use SAS Code to see the correlation of variables. We get this :

Correlation Coefficients |r| > 0.5: Density, alcohol; residual\_sugar, Density; free\_sulfur\_dioxide, total\_sulfur\_dioxide;

So we drop Density and free\_sulfur\_dioxide.

ii) Partition data into 70/30 training/validation data set and used decision tree to establish model.

iii) Model types: 3 Common different types of nominal criterion; 3 decision trees After dropping the high correlative variables; 3 decision trees after Variable Selection node to eliminated variables that not highly correlated to the target. Total 9 models.

iv) Comparison node is used to compare the decision trees described above.

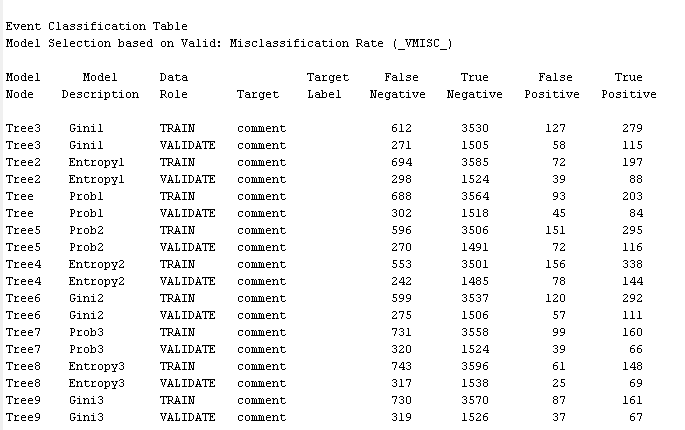
**3.**A:

From StatExplore Node, through the high correlation, we can find alcohol and sugar may partly decide the density of wine.

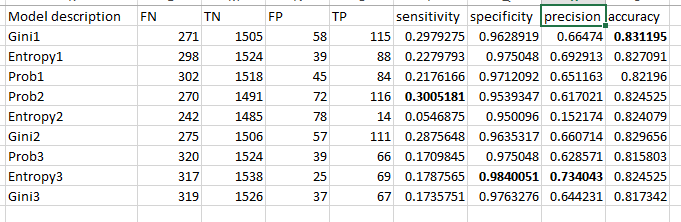
From the results of Graphboard node in SPSS and MultiPlot Node in Enterprise Miner, we get proper relationship between variables and the target:

High alcohol ~ high quality; Low chlorides ~ high quality; High or low pH ~ low quality; low volatile acidity wine ~ high quality; Low residual sugar ~ high quality; Other variables are not quite related.

Our goal is to predict which wine will be of high quality. Thus, high-quality wine is what we care about. We we should use precision to evaluate the model, meanwhile we need keep accuracy in an acceptable range. From the chart/table below we can see the precision, sensitivity, specificity and accuracy are shown below.



Comparison Node result with 9 models ↑



The table above shows that the accuracy of all decision trees are similar to each other, while the Gini1 decision tree built based on manual dropped data has the highest accuracy. Yet I think Entropy with variable selection is the best. It has the highest precision value and its accuracy is also acceptable (>80%). The result shows that by using this model (Entropy decision tree with variable selection), about 68% of the wine that I predicted to be high quality are true high quality wine. With highest sensitivity model above, its precision is much lower. That shows the correctness of the high quality wine prediction is worse than the chosen one.