IST707 – Data Analytics

Classification Analysis of Red Wine Quality

Yafei Lou

QingQing Hu

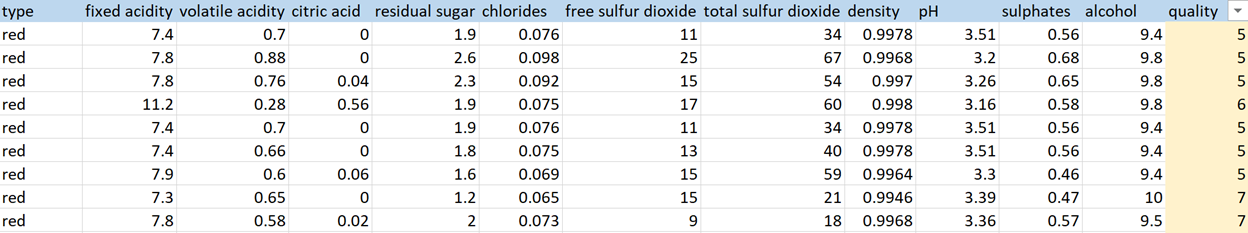
## **Introduction**

Since traditionally the standards of performance that wine experts use to judge wine quality include the following: sweetness, acidity, tannin, and alcohol, people usually taste the wine to see if it is balanced when nothing sticks out, like harsh tannin or too much sweetness. Therefore, wine quality is easier to detect than define, because the quality is defined subjectively and influenced by extrinsic factors. Nonetheless, novices, unlike experts, cannot easily determine what great characteristics a wine should possess but still pay for a high price blindly. What is not certain is if this is a complex example of the blind leading the blind. However, if we know what influences and signifies wine quality, we will be in a better position to make good purchases. Then, our appreciation for wines will deepen once we are familiar with wine quality levels and how wines vary in taste from region to region.

The data mining task will focus on classification mainly. By doing the classification task,we can map wines that contain certain characteristics into higher or lower score ranges. In addition, we will also use a clustering algorithm to determine what factors influence the score the most and possibly find the pattern among the data points.

In this project, we will use a data set downloaded from the UCI machine learning library. This data set is about the red variety of Portuguese "Vinho Verde" wine. Due to privacy and logic issues, only physicochemical (input) and sensory (output) variables are available (for example, no data on grape type, wine brand, wine selling price, etc.). The dataset consists of 1599 rows and 12 Input variables (based on physicochemical tests) and one Output variable (based on sensory data). Input variables include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol. quality (score between 0 and 10) is the Output variable.

Here is an example of the data, as shown below:



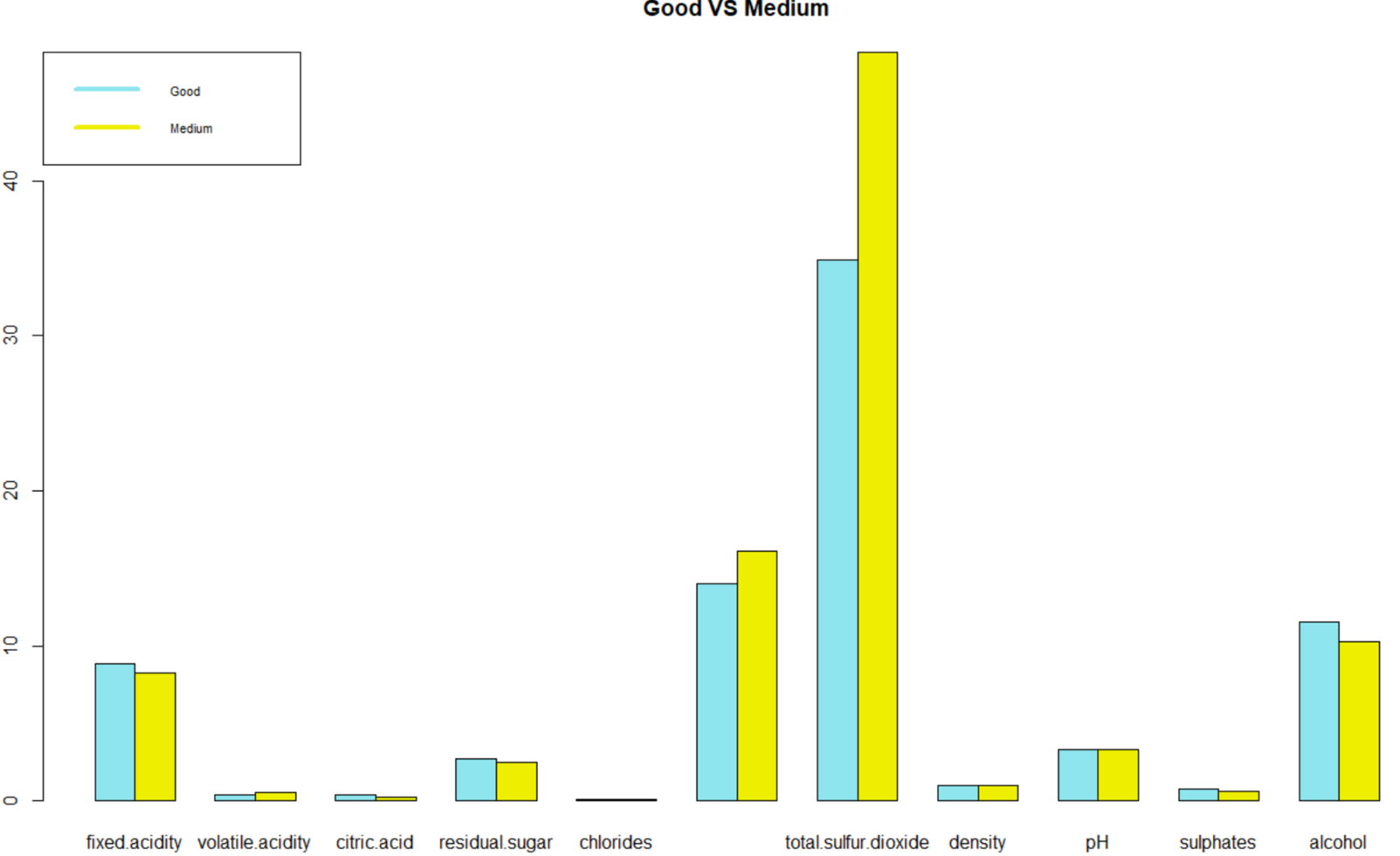
Based on this data set, we will use two models, classification and clustering, to analyze the factors that mostly affect the quality of wine. First, we divided the wine into two classes, Good (7~8) with 217 rows and Bad (3~6) with 1382 rows, we set the quality column (score range 3~8) as the target variable. In the classification task, we will analyze and compare the mean distribution of the different attributes of the two grades to find the factors that have the most influence on the quality of wine. Taking pH as an example, we plot the pH distribution for both good and bad levels. Based on this, see if the two grades have a significantly different distribution in pH, that is, whether the pH value significantly affects the quality of the wine grade classification. For other variables, we used the same method for comparison and analysis. In the association rule model, we will take the two levels of wine quality as RHS, and take all the remaining 10 attributes as the analysis factors, LHS. Through the comparison of support, confidence and lift, we can find the key factors to classify the wine grade.

Finally, we will construct a Confusion Matrix to calculate Accuracy, Precision, Sensitivity, Specificity and Score. According to the calculated value, the quality of the model is tested and evaluated.

## **Objective**

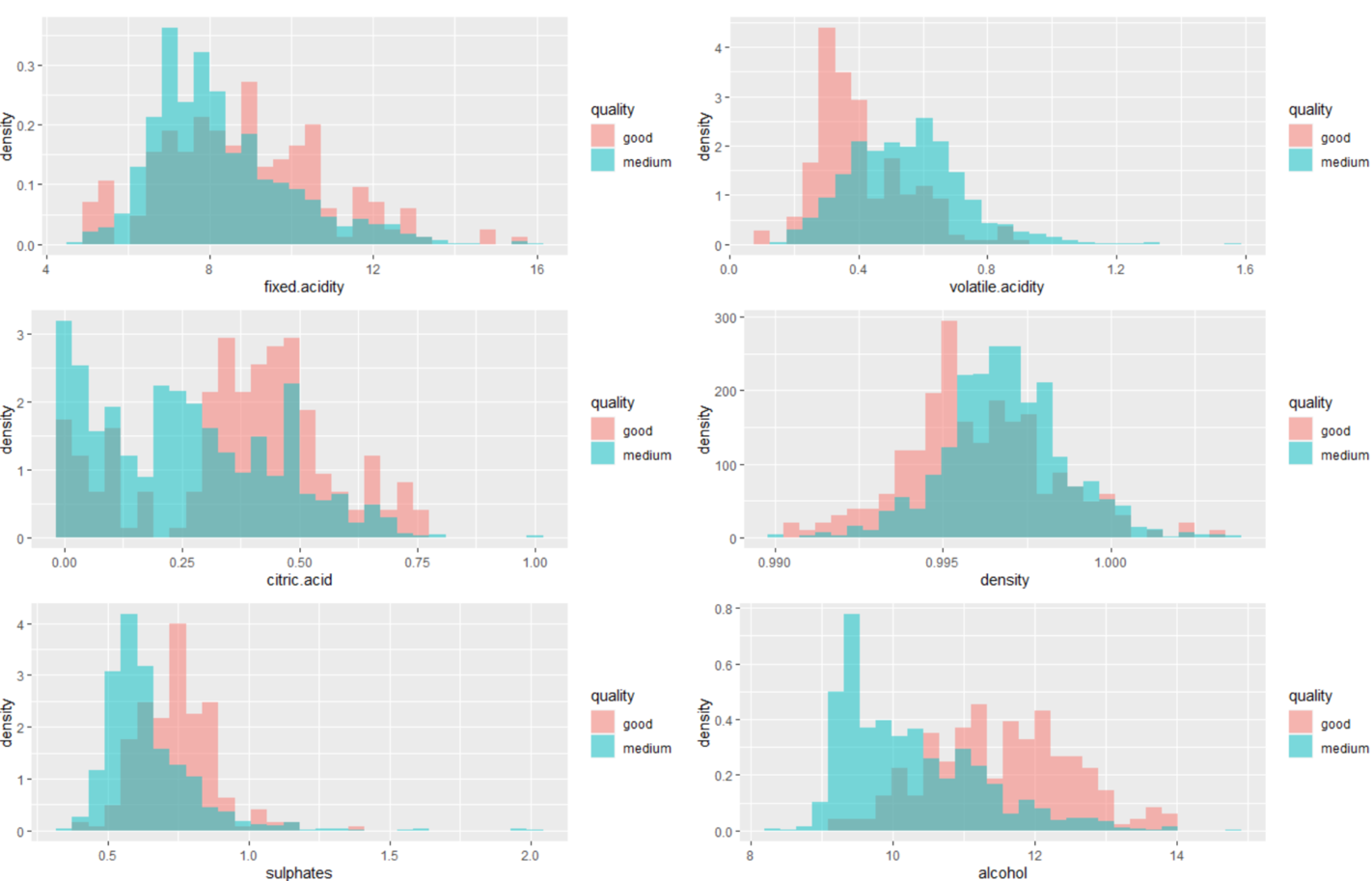
In this study, the models will predict wine’s quality score between 3 and 8, given the attributes such as acidity, residual sugar, chlorides, density, pH and alcohol. As a result, we could apply the model to impartially evaluate the wine quality using the scientific method.Software R is used to predict the red wine quality via classification algorithms and logistic regression. Classification algorithms of Decision Tree, Naïve Bayes, and Association Data Mining are chosen to classify.

## **Preliminary Analysis for dataset**

****

Here we set an averaged level for 11 red wine components, and then got the histogram above. We found that in the two grades of red wine, there are some obvious differences in the ingredients like total sulfur dioxide,free sulfur dioxide,alcohol,fixed acidity.

Next, In order to better highlight the differences in the amounts of various components of the two different grades of red wine, we made the following histogram based on the original data. This can not only see the components with significant differences, but also roughly see how much the corresponding ingredients are.

****

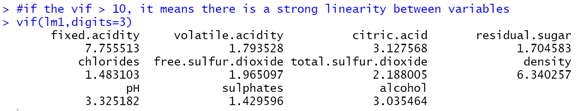
Here we just choose the six ingredients that vary from good to bad. From the bar chart above, if the shadow is smaller, it means that there are bigger differences between the two classes of red wine. Besides, we can see the percentage and range of each ingredient for two grades of red wine. Take alcohol for example,most percent of bad have less alcohol than the good quality does.

## **Methodology**

### Data Preparation

The data contains 12 numeric independent variables and 1 integer dependent variable. For the continuous variables, we separate all possible values into a number of bins, each having the same width and storing it into meaningful groups. Doing so can handle outliers. While for the dependent variable, we convert it into two level factors, in this case, wine quality greater than or equal to 7 would be defined as good, otherwise it will be bad.

Additionally, there are 6 rows that contain null values, because we have more than 1000 rows of data, so those rows are removed to prevent model error. After that, we want to check if a multicollinearity problem occurs in our dataset, because if the degree of correlation between variables is high enough, it could cause problems when we fit the regression model and interpret the results. To identify which variables are affected by multicollinearity and the strength of the correlation, the variance inflation factor method is used. As the result shown below, all features’ VIF is less than 10, which means each feature cannot be explained by all other features jointly. Therefore, there is no cause for collinearity in the case.

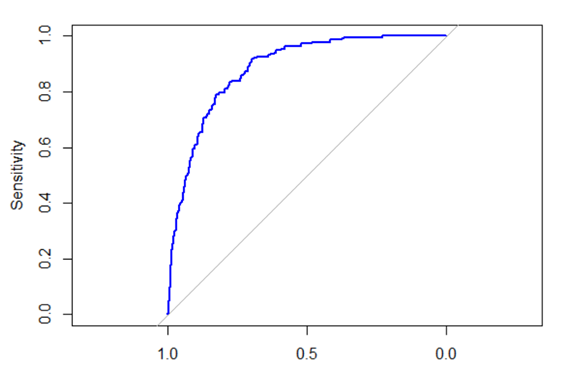


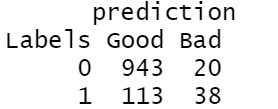
Lastly, data is randomly divided into train and test datasets with a ratio of 7 to 3. Training data includes 1114 rows, while 479 records are distributed to the test data.

### Model Results

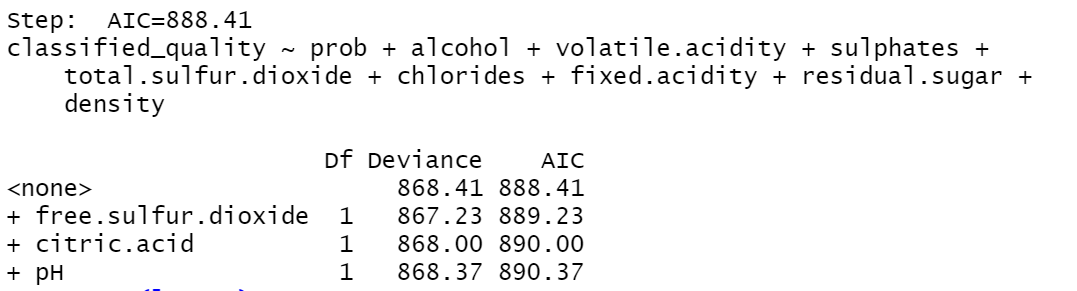
1. Logistic regression

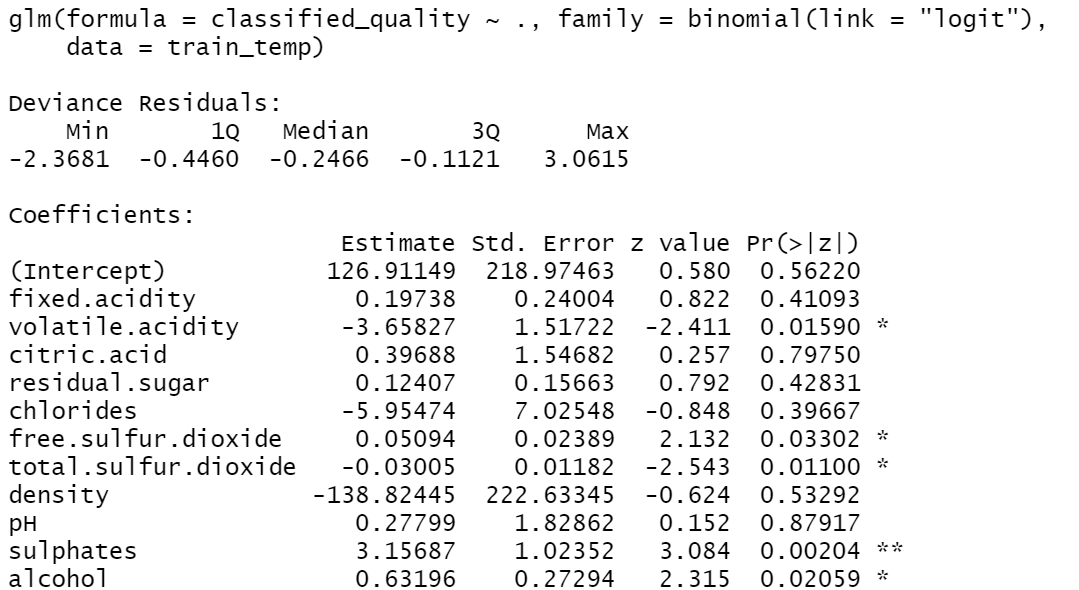
The area under the curve is 0.8824, which indicates the model is capable of distinguishing between classes. Also from the prediction results, we can conclude that this model has a great performance predicting the good and bad wine quality. To improve the model accuracy, we perform the stepwise regression model to include all the potential independent variables in the model and eliminate those that are not statistically significant.





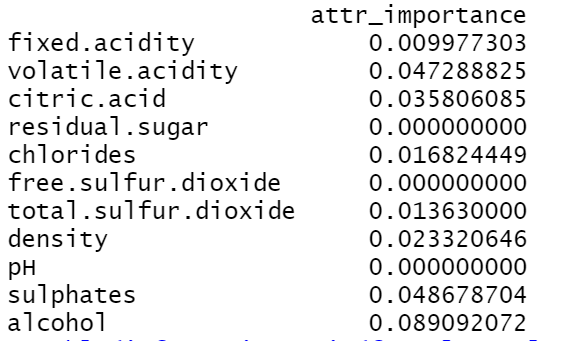
While the stepwise regression model is performed, it tells that the best set of predictors are shown below. Therefore, after we limit the number of independent variables in the logistic regression function, the model accuracy indeed has been improved.





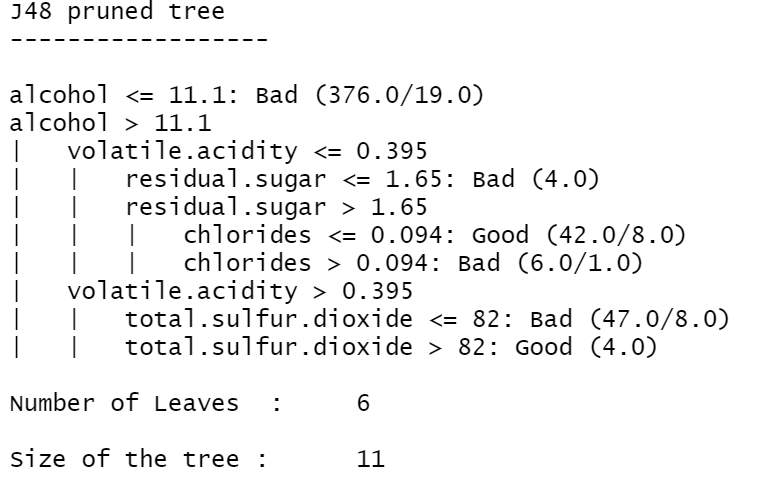
1. **Decision Tree**

The information gain contribution of each variable indicates that alcohol, sulphates and volatile acidity are the top three most important factors of the wine quality assessment. The interesting fact that the logistic regression result also shows the same conclusion if we look at the statistical significant P value for each variable.



The pruned decision tree is listed below. From this diagram, the level of alcohol is the root node. The tree infers

that alcohol level less than or equal to 11.1 is classified as bad. If the alcohol level is greater than 11.1 and volatile acidity is less than 0.395, the residual sugar level less than 1.65 will also be classified as bad. In addition, If the alcohol level is greater than 11.1, volatile acidity is less than 0.395, and the residual sugar level greater than 1.65, the chlorides is less than 0.094, it will be classified as good.

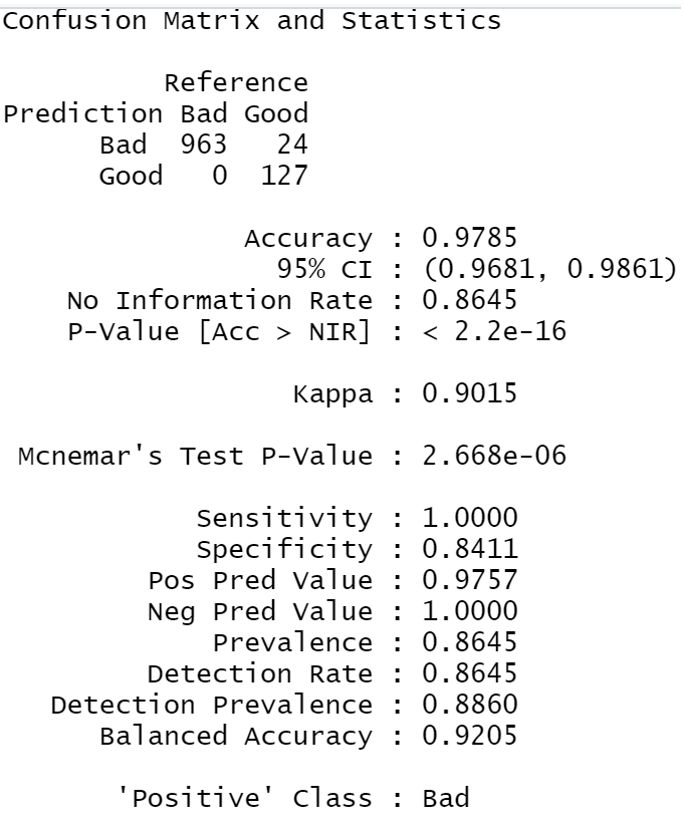


To find the best split point and have the max information gain, we iterate through the C array range from 0.01 to 0.5 and M from 2 to 5. After the iteration, we find the best performance model that has the highest percentage of correctness. Then the model is evaluated on testing data, as the result. The confusion matrix is shown below.

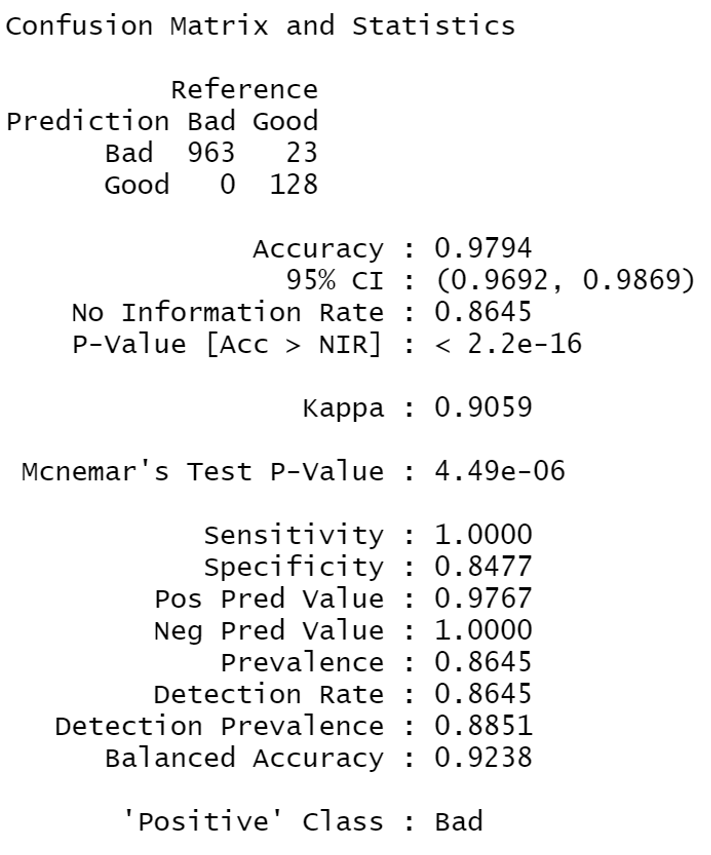


1. **Naïve Bayes**

Since the columns we want to focus on are all numeric types, we choose to use a naive Bayes model to analyze and predict the data. First, we build a model based on 30%(479 rows) of the train set and then make predictions on the remaining 70%(1174rows) of the test set. We use the confusion matrix to test the accuracy of the model. The result is shown in the figure below. The accuracy of the model is 97.85%.



Considering that the recording of the train set and the test set is asymmetric, we standardized the two data sets respectively. Similarly, we get the following results. The accuracy of the model is 97.94%.

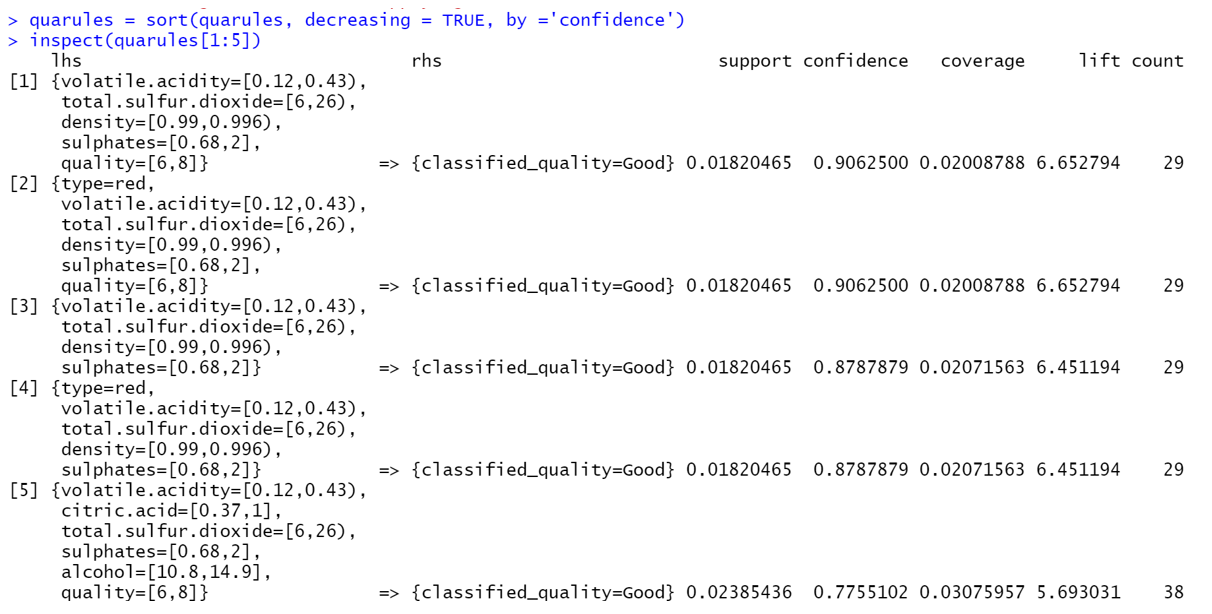


Compared with the former result, there seems to be no improvement for this data set. Because the accuracy is over 95%, we can safely draw a conclusion that the Naive Bayes model has a good performance to predict the quality of red wine.

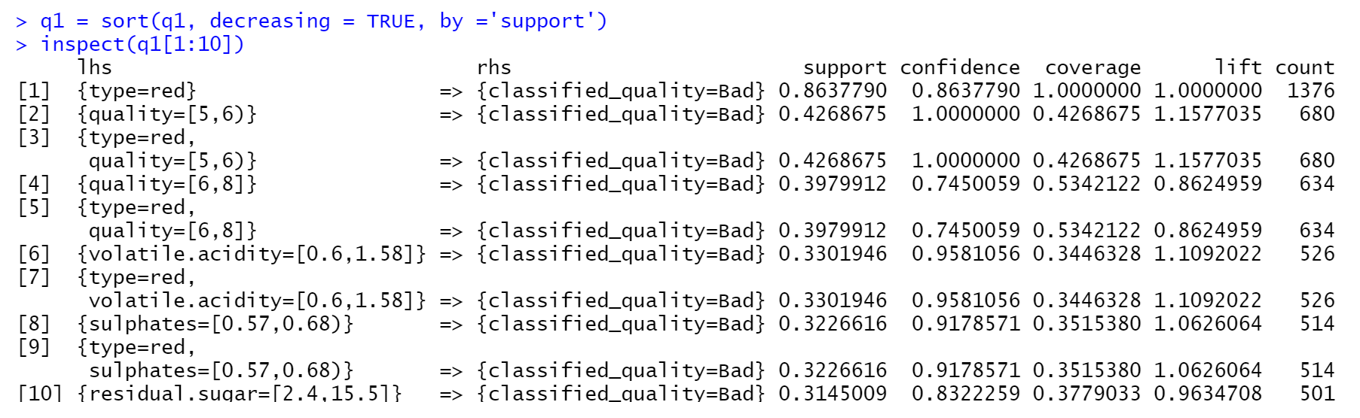
1. **Association Rule Data Mining**

In order to find out which ingredients are the key to the quality of red wine, we also selected data association mining to analyze the two levels separately.

On the one hand, for good quality of red wine, here are the top 5 rules ranking by confidence. Confidence reflects how likely the research object (RHS) will appear when LHS appears. According to the results, we found that the following components are within a certain range , volatile.acidity[0.12,0.43)，total.sulfur.dioxide=[6,26)，density=[0.99,0.996)，sulphates=[0.68,2]，alcohol=[10.8,14.9], the quality of red wine is better.



On the other hand, we also found the top10 rules about bad quality. We can see that when the content of the following ingredients is in the corresponding range，such as volatile.acidity=[0.6,1.58],sulphates=[0.57,0.68),residual.sugar=[2.4,15.5], the quality of red wine may be poor.



## **Model Evaluation**

|  |  |
| --- | --- |
| **Model** | **Overall Accuracy** |
| Logistic Regression | 88.24% |
| Decision Tree | 85.18% |
| Naïve Bayes | 97.94% |

## **Conclusion**

To sum up, based on the results of four models( logistic regression, Decision tree, Naive Bayes, and association rule data mining) from the dataset, we found that some ingredients like volatile acidity, sulfates, and acidity are significant factors to identify whether the red wine is good or not. Furthermore, we can also conclude that when the Sulphates and Alcohol indexes tend to be large and the Volatile acidity index is small, the grade of red wine is higher. The analysis result of the model is almost consistent with our preliminary analysis. As for association rule mining, we can conclude that we can state which conditions have to be met in order to be considered as good wine, as seen in the above example. The algorithm depends on the support and confidence we set as well as the question we ask. A final reflection as a learning outcome, is that the model has better performance when predicting red wine with bad quality.Therefore, more factors need to be included to better distinguish the wine quality.