CUNY Baruch College

Term Project

Red Wine Quality

JiaRui (Jesse) Shao

STA 3000 EMWA

Víctor Peña

18 December 2019

1. **Executive summary:**

In this project, I plan to analyze the data and evaluate which chemical properties classify a wine as "good" quality by utilizing three different methods.

2. **Data description:**

I found this dataset on Kaggle and it's already the usable csv format so I just need to download it to my laptop and load it in RStudio. This dataset consists of 12 variables and 1,599 observations. Among the 12 variables, quality is the response variable.

3. **Data Analysis**

A. Exploratory Data Analysis & Data Visualization

First, I created a summary table with the following statistics for each variable: minimum, maximum, mean, standard deviation, and median. This revealed some interesting observations. For example, I saw outliers in free sulfur dioxide and total sulfur dioxide. I also saw that pH levels varied, indicating there are tart and sweet wines present (Exhibit 1). After that, I made a ggpairs and a ggcorrplot to explore the relationships among the variables and see if there was a possibility of any correlation between the continuous variables (Graph B & C in EXHIBIT 2). Nothing appeared to be highly correlated (>0.70 correlation value). However, there are some pairs of variables that decently correlated. For example, fixed acidity is decently correlated (>0.5 or <-0.5) with citric acidity, density and pH (fixed.acidity and PH are negatively correlated), volatile acidity is decently correlated with citric acid. Also, total sulfur dioxide is decently correlated with free sulfur dioxide. I also created a histogram to visualize how the wine in this dataset falls into one of the following six quality categories: 3, 4, 5, 6, 7, and 8 (Graph A in EXHIBIT 2). Although the quality is a numeric variable, here I'm more interested in what makes a wine "good", so converting quality a categorical variable is more meaningful. Instead of having six responses for the quality variable, I converted it to a binary response falling into the two

descriptive categories, good quality and bad quality. A wine is categorized as a “good” wine as the one with quality larger than or equal to 6 and “bad” wine as the one quality smaller than 6. Now I have 855 “good” quality wines and 744 “bad” quality wines.

After converting the quality variable into a qualitative variable, I created another ggpairs plot with quality as color to see what’re the potential important variables to decide wine quality (Graph D in EXHIBIT 2). It's clear to see that alcohol looks like an important factor when deciding quality, the alcohol level, sulphates level and citric acidity of "good quality" red wine are higher than "bad quality". Total sulfur dioxide and volatile acidity of "good quality" red wine are lower than "bad quality".

B. Building models

I applied logistic regression, classification, and random forest to build the model to predict red wine quality and estimate what are some variables that are important to decide wine quality.

I. Logistic Regression

a. Model building

First, I fit the first model which includes all 11 independent variables. I used significant level of 0.05 as the cut off for significance when deciding the variables that I’m going to use for the second model. Based on the results, I can see that not every independent variable is statistically significant. These variables have high p-values: p-value for fixed acidity is 0.17, for residual sugar is 0.3, for density is 0.53 and for Ph is 0.6 (Exhibit 3). Then I fit the second model which includes 7 independent variables: alcohol, volatile acidity, chlorides, citric acid, free sulfur dioxide, total sulfur dioxide and sulphates. However, I found that p-value for citric acid is 0.45 which is much higher than alpha of 0.05 and it is not statistically significant in this second model (Exhibit 3). Next, I fit the third model which includes 6 variables: alcohol, volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide and sulphates, excluding critic acid. Based on

the results I see that all variables are statistically significant. Their associated p-values are very low (Exhibit 3).

b. Model fit

I also checked for multicollinearity in each model by computing variance inflation factor. The VIF score for the first model is 7.767512, which is problematic. VIF scores for predictors in the second model and the third model are below are 5 which indicates there's no multicollinearity (Exhibit 6).

c. Model Interpretation

I estimated these logistic regression models by computing their AIC and BIC. The Exhibit 5 shows that the first model gets 1679.625 on AIC, 1744.151 on BIC, the second model gets 1680.155 on AIC, 1723.172 on BIC, and the third gets 1678.716 on AIC and 1716.346 on BIC. The third model has smaller AIC and BIC. This is an evidence that the third model is a better fit to the data than the first model and the second model. Next, I evaluated the prediction accuracy of these models. I used holdout method here by randomly selecting 80% of the observations for training and the remaining as testing set. I evaluated the performance of the models using test set and I produced a confusion matrix to determine how many observations Ire correctly classified. The accuracy for the first model and the second model is 73.125 % and 72.5%, the accuracy for the third model is 73.4375% (Exhibit 5). The accuracy for the third model is slightly higher compare to the other two models.

II. Classification

a. Model building

I also used classification (decision tree) method to build the model because the accuracy of the logistic regression model is not very high. First, I set best as 4 and the accuracy is only 70%

(Exhibit 8). Then I set best to 8, and the final overall accuracy is 73.4375%, which is better than the first classification model but still not good enough to me (Exhibit 8).

b. Model fit

Although the accuracy of the second classification model is not better than the third logistic regression model, it provides us a clearer view and more information to see which variables are more important through the decision tree.

c. Model Interpretation

The Exhibit 7 tells us the prediction of in what cases wine quality will be "good", in what cases wine quality will be "bad". It indicates alcohol more than 10.525 always leads to good wine, and alcohol less than 10.525 follows various paths (Exhibit 7). This plot shows us that the level of alcohol in wine is what matters for making good wine. The overall accuracy of the model is 73.4375% and recall rate is 52.31% (Exhibit 8).

III. Random Forest

a. Model building

There' s only one random forest model.

b. Model fit

Random forest reduces the variance of random forest by averaging many trees. So next, I implemented random forest method which combines a lot of decision trees.

c. Model Interpretation

The accuracy of the model is 82.1875% (Exhibit 9), which means I successfully improved the model from the one in decision tree. In the plot of the feature importance of the variable (Exhibit

10), I see that alcohol is the most important variable affecting the quality, and following sulphates with the highest mean decrease in Accuracy and Gini.

4. Conclusions

I used Logistic Regression, Classification, and Random Forest exploring this dataset to determine the factors that are important for wine quality. According to the results I got, random forest gave us the highest accuracy that is 82.1875%, followed by the 73.4375% rate by logistic regression and decision tree. As for the important factors, alcohol, sulphates, total sulfur dioxide, volatile acidity seem to be more important than other factors when deciding wine quality.

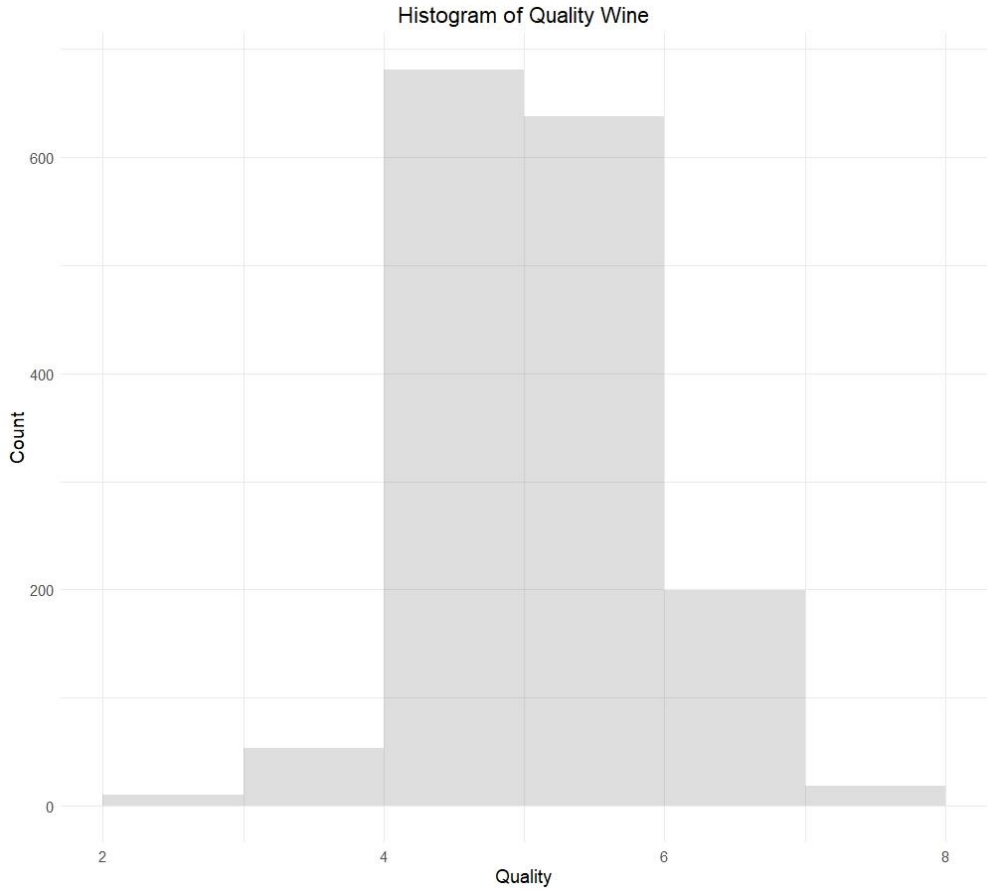
APPENDIX

EXHIBIT 1: Summary Table

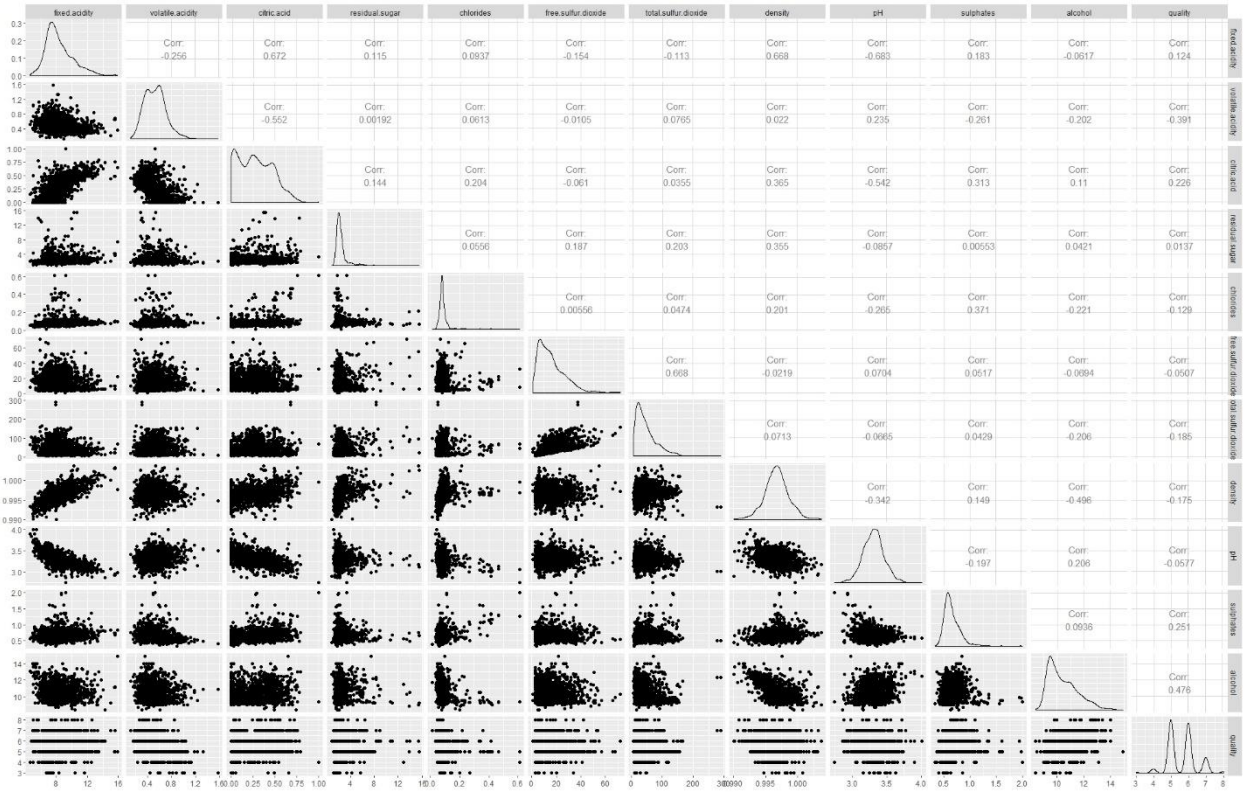
	Fixed Acidity	Volatile Acidity	Citric Acid	Residual Sugar	Chlorides	Free Sulfur Dioxide
Mean	8.32	0.53	0.27	2.54	0.09	15.87
Min	4.60	0.12	0.00	0.90	0.01	1.00
Max	15.90	1.58	1.00	15.50	0.61	72.00
Median	7.90	0.52	0.26	2.20	0.08	14.00
SD	1.74	0.18	0.19	1.40	0.05	10.46

	Total Sulfur Dioxide	Density	pH	Sulphates	Alcohol
Mean	46.47	0.99	3.31	0.66	10.42
Min	6.00	0.99	2.74	0.33	8.40
Max	289.00	1.00	4.01	2.00	14.90
Median	38.00	0.99	3.31	0.62	10.20
SD	32.89	0.00	0.15	0.17	1.07

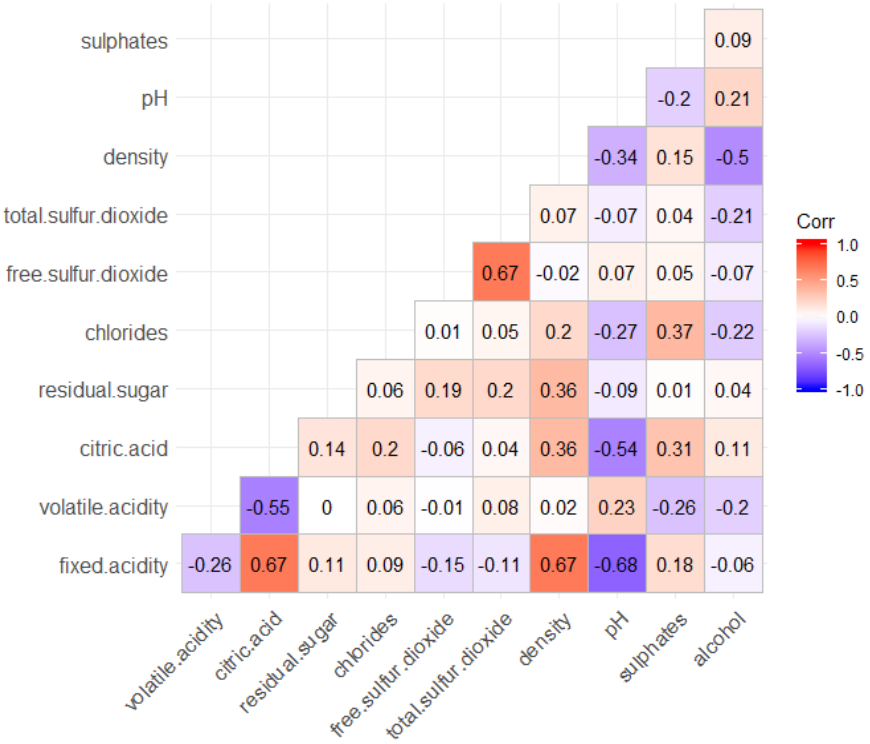
EXHIBIT 2: Exploratory Analysis Visualizations



Graph A



Graph B



Graph C

Graph D

EXHIBIT 3: Logistic Regression

Model 1

Coefficients	P-value
(Intercept)	0.58890
Alcohol	< 2e-16 ***
Fixed Acidity	0.16736
Volatile Acidity	1.79e-11 ***
Citric Acid	0.02354 *
Residual Sugar	0.30351
Chlorides	0.01259 *
Free Sulfur Dioxide	0.00698 **
Total Sulfur Dioxide	1.29e-08 ***
Density	0.53024
pH	0.59717
Sulphates	6.36e-10 ***

Model 2

Coefficients	P-value
(Intercept)	< 2e-16 ***
Alcohol	< 2e-16 ***
Volatile Acidity	1.24e-11 ***
Citric Acid	0.45448
Chlorides	0.00431 **
Free Sulfur Dioxide	0.00566 **
Total Sulfur Dioxide	6.48e-10 ***
Sulphates	2.16e-10 ***

Model 3

Coefficients	P-value
(Intercept)	< 2e-16 ***
Alcohol	< 2e-16 ***
Volatile Acidity	5.60e-15 ***
Chlorides	0.00202 **
Free Sulfur Dioxide	0.00281 **
Total Sulfur Dioxide	9.95e-11 ***
Sulphates	2.47e-10 ***

EXHIBIT 4: Model Estimation

	Model 1	Model 2	Model 3
AIC	1679.625	1680.155	1678.716
BIC	1744.151	1723.172	1716.356

EXHIBIT 5: Holdout Method

Model 1

	True Bad	True Good
Predicted Bad	122	39
Predicted Good	47	112

Model 2

	True Bad	True Good
Predicted Bad	120	39
Predicted Good	49	112

Model 3

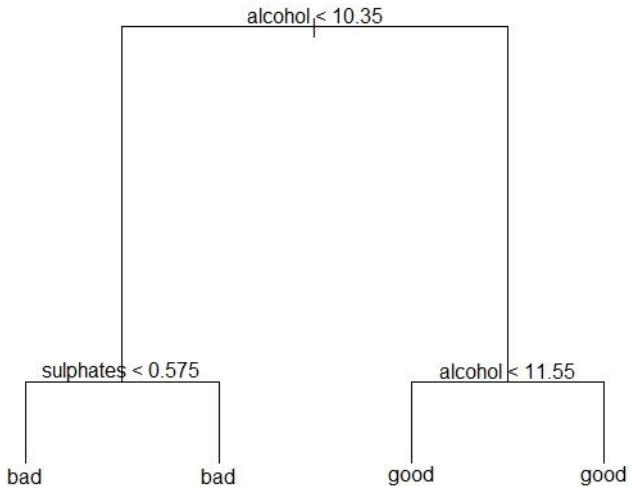
	True Bad	True Good
Predicted Bad	123	39

Predicted Good	46	112

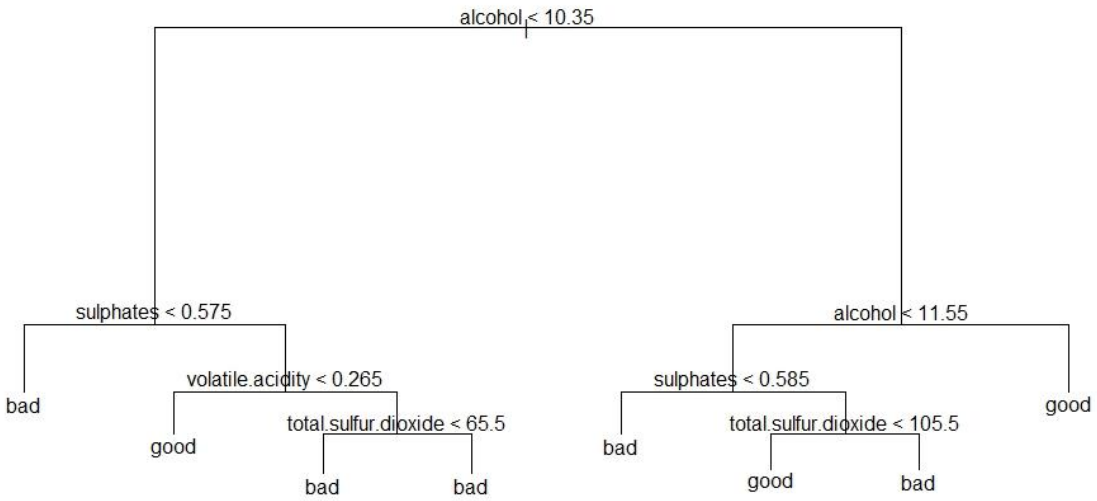
EXHIBIT 6: Variance Inflation Factor

	Model 1	**Model 2**	**Model 3**
VIF	7.767512	1.641316	1.139014

EXHIBIT 7: Decision Tree



Graph A



Graph B

EXHIBIT 8: Confusion Matrix for Decision Tree

Best = 4	**True Bad**	**True Good**
Predicted Bad	101	80
Predicted Good	68	71
Best = 8	**True Bad**	**True Good**
Predicted Bad	101	72
Predicted Good	68	79

EXHIBIT 9: Confusion Matrix for Random Forest

	True Bad	**True Good**
Predicted Bad	128	29
Predicted Good	26	137

EXHIBIT 10: Variance Importance