ISSN 2616-7433 Vol. 2, Issue 18: 72-81, DOI: 10.25236/FSST.2020.021813

Analysis of the Evaluation of Wine Quality Based on Different Statistical Methods

Zihao Liu, Yuemeng Wang, Shiyu Wu, Chujun Fang

University of California Davis 1 Shields Ave, Davis, CA 95618 US

ABSTRACT. *In the paper, the evaluation of wine quality is analyzed based on statistical methods, correlation, regression, classification and clustering. The accuracy of our four methods are approximately the same: Approximately 75 percent data match with the reality. In conclusion, wine quality depends on its chemistry components for most parts.*

KEYWORDS: *The evaluation of wine quality, Statistical methods, Correlation, Regression, Classification, Clustering*

1. Introduction

To master more knowledge on statistical learning and data analysis to deal with different types of data, the evaluation of wine quality is analyzed by statistical methods, correlation, regression, classification and clustering. In this project, we choose to analyze the red wine quality data from UCI machine learning center. The wine quality dataset is related to red variants of the Portuguese “Vinho Verde” wine. It has 1599 number of instances for red wine. The dataset contains 11 input variables which are based on physicochemical tests and one output variable.

Fixed Acidity: Non-volatile acid. It is a measure of the total concentration of titratable acids and free hydrogen ions in wine.

Volatile Acidity: Acetic acids and byproducts caused by bacteria in wine. This is a signal for mistakes made during the process because it can lead to an unpleasant taste.

Citric Acid: It acts as a preservative, and sometimes can bring “fresh” flavors to the finished wine, usually only in very small amount.

Residual Sugar: The natural glucose that left after fermentation. It is important to find the balance between sweetness and sourness in wine.

Chlorides: Chlorides cause the “salty” flavour in wine. Proper proportions can make the wine more palatable.

ISSN 2616-7433 Vol. 2, Issue 18: 72-81, DOI: 10.25236/FSST.2020.021813

Free Sulfur dioxide: The most commonly used preservative added by the winemakers to protect the negative influence of wine exposure under oxygen.

Total Sulfur dioxide: Combination of free sulfur dioxide and bound sulfur dioxide.

Density: The density of wine is close to the density of the water.

PH: a measure of the acidity or alkalinity of a solution.

Sulphates: A natural by-product of the fermentation process. It can prevent wine from oxidizing and help it maintain the fresh taste.

Alcohol: Alcoholic component of the red wine.

Quality: It is a qualitative score ranging from 0 to 10 that evaluate each wine.

2. Statistical Questions of Interest and Analysis Plan

The primary scientific question of interest in this project is to predict the wine quality using the physicochemical tests features, and to understand the relationships between numeric variables and their effectiveness on the quantity of red wine.

The wine quality data set is shown in Table 1.

Table 1 Summary of Data

fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide
Min. : 4.60	Min. :0.1200	Min. :0.000	Min. : 0.900	Min. :0.01200	Min. : 1.00
1st Qu.: 7.10	1st Qu.:0.3900	1st Qu.:0.090	1st Qu.: 1.900	1st Qu.:0.07000	1st Qu.: 7.00
Median : 7.90	Median :0.5200	Median :0.260	Median : 2.200	Median :0.07900	Median :14.00
Mean : 8.32	Mean :0.5278	Mean :0.271	Mean : 2.539	Mean :0.08747	Mean :15.87
3rd Qu.: 9.20	3rd Qu.:0.6400	3rd Qu.:0.420	3rd Qu.: 2.600	3rd Qu.:0.09000	3rd Qu.:21.00
Max. :15.90	Max. :1.5800	Max. :1.000	Max. :15.500	Max. :0.61100	Max. :72.00

total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
Min. : 6.00	Min. :0.9901	Min. :2.740	Min. :0.3300	Min. : 8.40	Min. :3.000
1st Qu.: 22.00	1st Qu.:0.9956	1st Qu.:3.210	1st Qu.:0.5500	1st Qu.: 9.50	1st Qu.:5.000
Median : 38.00	Median :0.9968	Median :3.310	Median :0.6200	Median :10.20	Median :6.000
Mean : 46.47	Mean :0.9967	Mean :3.311	Mean :0.6581	Mean :10.42	Mean :5.636
3rd Qu.: 62.00	3rd Qu.:0.9978	3rd Qu.:3.400	3rd Qu.:0.7300	3rd Qu.:11.10	3rd Qu.:6.000
Max. :289.00	Max. :1.0037	Max. :4.010	Max. :2.0000	Max. :14.90	Max. :8.000

For the complexity of data, it is hard to get real results only from a statistical point of view without a solid background knowledge in chemistry and winemaking. In an analogy, we decide to approach the real result as we cut the line to approximate the slope in Math. In order to solve this question, we use linear regression to find the correlation between numeric variables and then do the following:

First, treating scores as quantitative, we want to analyze the correlation between different variables and build a linear regression model to see how it works.

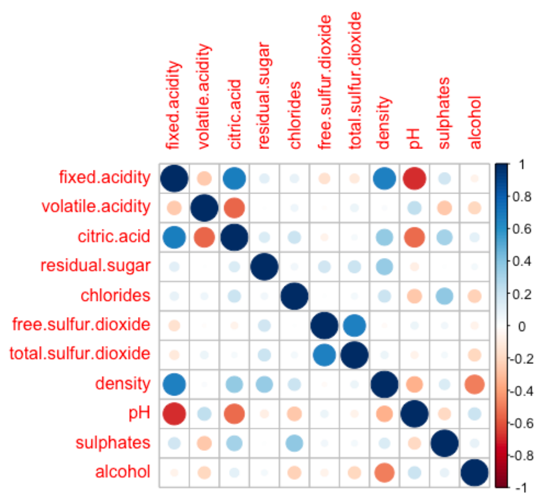
ISSN 2616-7433 Vol. 2, Issue 18: 72-81, DOI: 10.25236/FSST.2020.021813

Secondly, we want to use knowledge in machine learning to estimate the result when treating scores as qualitative.

3. Statistical Analysis

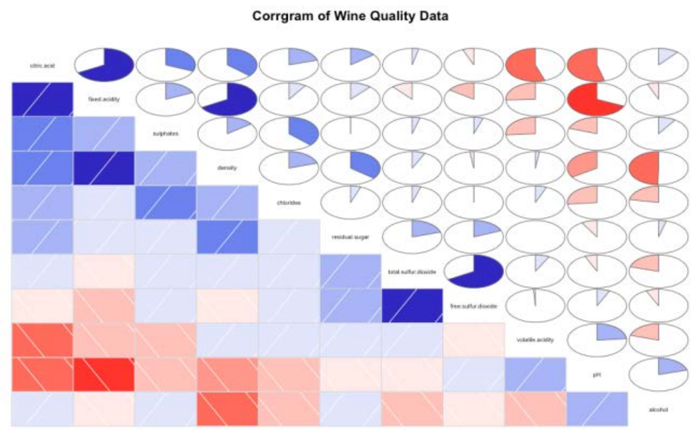
3.1 Correlation

The wine quality data set contains 11 numeric input variables and one numeric output variable; the 11 input variables which are fixed, acidity, volatile. acidity, citric. acid, residual. sugar, chlorides, free. sulfur. dioxide, total. sulfur. dioxide, density, pH, sulphates, and alcohol, as shown in Fig.1. And the output variable is quality. Firstly, we want to find out how the various numeric variables relate to each other by using correlation plot and correlation diagram on the 11 input variables. In correlation plot, color blue suggests positive correlation, and red color suggests negative correlation, and the size of the circle corresponds to the correlation coefficients. In observing the correlation plot and correlation gram, they suggest that there are some strong positive correlations between citric. acid and fixed. acidity, density and fixed. acidity, total. sulfur. dioxide and free. sulfur. dioxide. The plots also suggest that there are some strong negative correlations between pH and fixed. acidity, citric. acid and volatile. acidity, pH and citric. acid, alcohol and density. We used pairs function and scatter plot to plot the relationship between correlated variables to see their trends more closely.

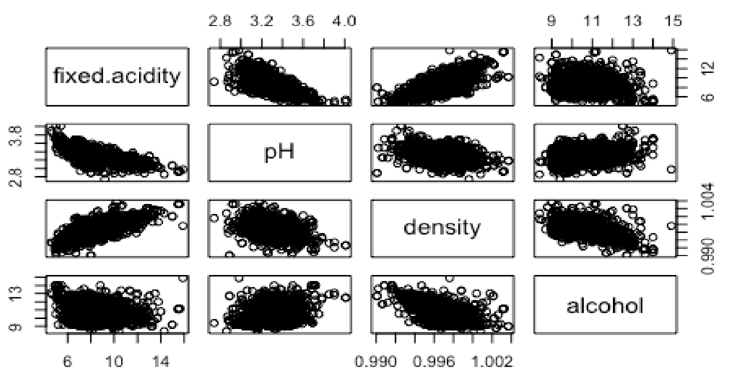


(a)

ISSN 2616-7433 Vol. 2, Issue 18: 72-81, DOI: 10.25236/FSST.2020.021813

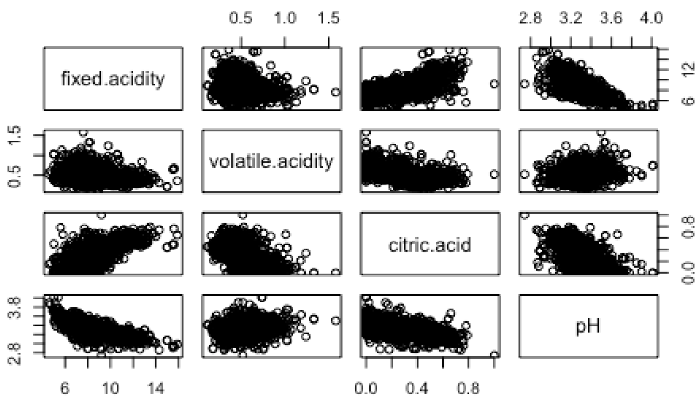


(b)

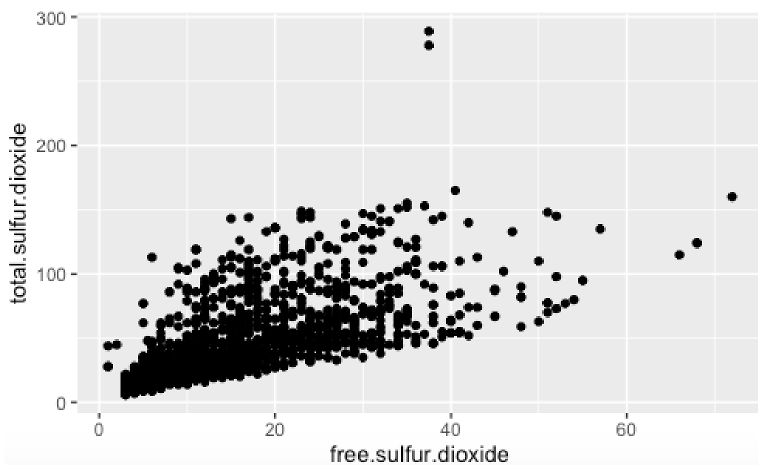


(c)

ISSN 2616-7433 Vol. 2, Issue 18: 72-81, DOI: 10.25236/FSST.2020.021813



(d)



(e)

Fig.1 The Wine Quality Data Set

To further know how these variables influence quality precisely, we go to the next stage of analysis of dataset pattern.

ISSN 2616-7433 Vol. 2, Issue 18: 72-81, DOI: 10.25236/FSST.2020.021813

3.2 Regression Analysis

In analyzing the wine data, we first partition the dataset into two subsets. We randomly select 80% of the data to be our training data and the remaining 20% to be our test data. Then we use family = binomial with the glm function for logistic regression on the training data. We then get the overall summary which is the image below. For the summary of logistic regression, we can clearly see that density has a very high p-value of 0.77406 and residual.sugar also has a very high p-value of 0.91388; this suggests that they are not statistically significant for this model. Other variables have low p-value which can help explain the model. Then we predict for test data using type = "response" to get class probabilities. We then convert prediction to class labels 0 or 1. At the end, we calculated the accuracy of logistic regression is 0.775, as shown in Table 2.

Table 2 the Result of Logistic Regression Analysis

```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)          19.150545  85.590617   0.224  0.82296
fixed.acidity         0.111649   0.105894   1.054  0.29173
volatile.acidity     -2.844256   0.534364  -5.323 1.02e-07 ***
citric.acid          -0.988874   0.630212  -1.569  0.11662
residual.sugar        0.006617   0.061182   0.108  0.91388
chlorides            -4.933251   1.738512  -2.838  0.00454 **
free.sulfur.dioxide   0.016142   0.009139   1.766  0.07734 .
total.sulfur.dioxide -0.014461   0.003119  -4.636 3.55e-06 ***
density             -25.095517  87.419908  -0.287  0.77406
pH                   -0.921976   0.793994  -1.161  0.24557
sulphates             2.567119   0.499801   5.136 2.80e-07 ***
alcohol               0.872373   0.113699   7.673 1.68e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

3.3 Classification

We also use the classification to check how to determine the quality of wine, and we find that the alcohol variable did the best job of dividing up the observations into the different cultivars.

The reason why we also chose to use the decision tree is that it can also show some non-linear relationship between the variables. Within this method, each tree corresponds to an attribute and the leaf of the decision tree means the certain classes. The result is shown as classification in Fig.2.

ISSN 2616-7433 Vol. 2, Issue 18: 72-81, DOI: 10.25236/FSST.2020.021813

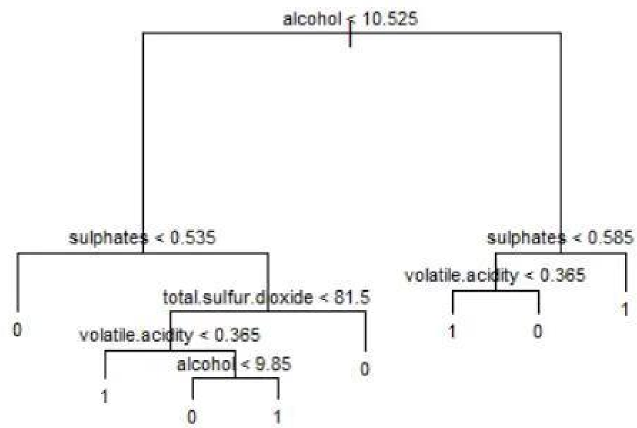
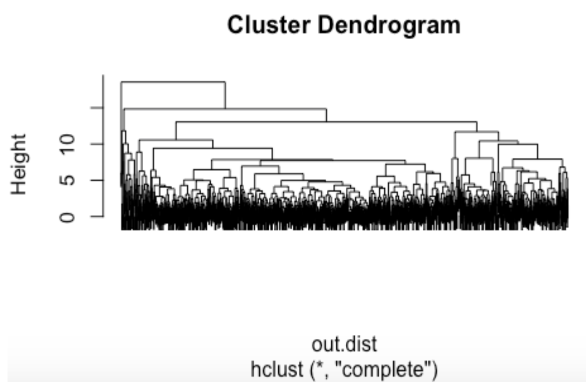


Fig.2 The Result of Classification Analysis

3.4 Clustering



ISSN 2616-7433 Vol. 2, Issue 18: 72-81, DOI: 10.25236/FSST.2020.021813

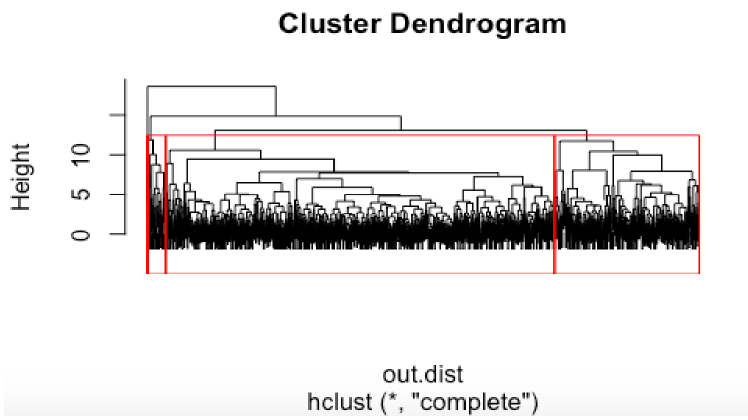


Fig.3 The Result of Clustering Analysis

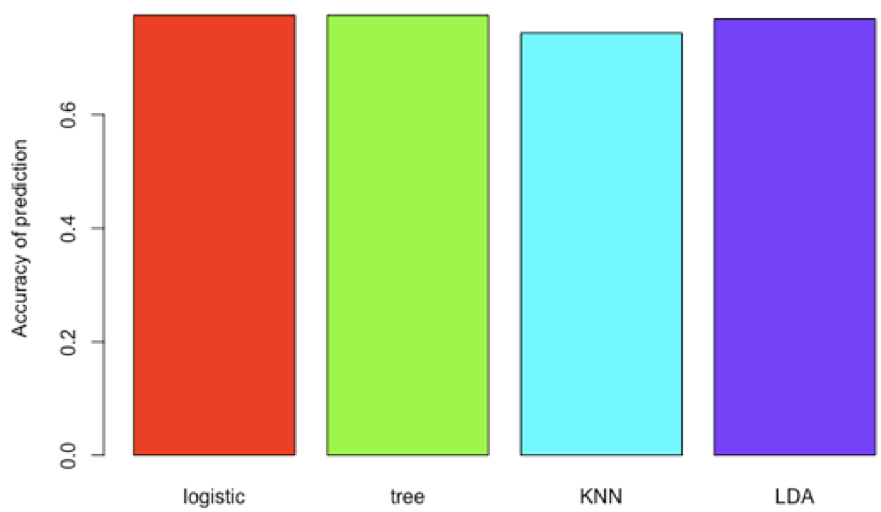
We checked whether the data can separate into four groups so we can conclude something. Looking at the dendrogram for the variable data, there are not clearly distinct groups; As shown in Fig.3, most of the observations in the right hand group are clustering together at about the same height. So for the date set, we can tell that there are too many variables and we can not conclude them into 4 groups. We conclude that clustering does not provide valuable information about our data.

4. Conclusion and Discussion

As shown Fig.4, the accuracy of our four methods are approximately the same: Approximately 75 percent data match with the reality. In conclusion, we can safely say that wine quality depends on its chemistry components for most parts.

How could we explain the fact that some parts cannot be explained by our models? The inference is that a human's subject feeling also decides how we evaluate the quality of wine. From the regression model, we find that density and sugar should be excluded from evaluation. Actually, in real life, we can often hear wine tasting experts talk about how “sweet and dense” the wine should be, but they often quarrel with each other and cannot force others to accept their standard. (Great deviation and hard to measure feels!) So, we also need to incorporate our scientific reasoning and knowledge on society into data analysis.

ISSN 2616-7433 Vol. 2, Issue 18: 72-81, DOI: 10.25236/FSST.2020.021813



logistic	tree	KNN	LDA
0.77500	0.77500	0.74375	0.76875

Fig.4 The Accuracy of Our Four Methods

In extension, this project encourages us to master more knowledge on statistical learning and data analysis to deal with different types of data. For example, when we dig information on clustering, we luckily find the method of tree, which makes cuts on the numeric value for the dataset." Unlike linear models, they map nonlinear relationships quite well. They are adaptable at solving any kind of problem at hand (classification or regression)."(Brid) This visualization of this method explains things very clearly.

References

[1] An Introduction to Corrplot Package: https://cran.r-project.org/web/packages/corrplot/vignettes/corrplot-intro.html
[2] Rajesh s. Brid. Decision Trees-a Simple Way to Visualize a Decision: https://medium.com/greyatom/decision-trees-a-simple-way-to-visualize-a-decision-dc506a403aeb
[3] A Graphical Explanation of How to Interpret a Dendrogram: https://casoilresource.lawr.ucdavis.edu/blog/graphical-explanation-how-interpret-dendrogram/

ISSN 2616-7433 Vol. 2, Issue 18: 72-81, DOI: 10.25236/FSST.2020.021813

[4] Huang Yi, Hu Erqin(2013).Study on statistical analysis method of Wine Quality Score.Journal of Yangtze University(Natural Science Edition), vol.10, no.2, pp. 24-26.
[5] LI Yun, LI Ji-ming, JIANG Zhong-jun(2009). Application of Statistical Analysis in the Evaluation of Grape Wine Quality. LIQUOR-MAKING SCIENCE & TECHNOLOGY, no. 4, pp.79-82.
[6] Qi Ya-e, Cao Zhi-qiang, Liu Ya-nan, Yang Ting, Ma Lei (2013). The Quality Evaluation Model of Grape Wine Based on Statistical Analysis of Physical and Chemical Indexes，Journal of Hexi University, vol.29, no.5, pp.58-65.