3325 - Sharon Philip

INDEX PAGE

RED-WINE QUALITY

Sharon Philip
3325
MSc-IT(PART-1)

ABSTRACT

The quality of a wine is important for the consumers as well as the wine industry. The traditional (expert) way of measuring wine quality is time-consuming. Nowadays, machine learning models are important tools to replace human tasks. In this case, there are several features to predict the wine quality but the entire features will not be relevant for better prediction.So, our thesis work is focusing on what Wine features are important to get promising results. For the purpose of classification model and evaluation of the relevant features,In this study, we used two wine quality datasets red wine and white wine. Finally, we achieved the artificial neural network (ANN) algorithm has better prediction results than the Support Vector Machine (SVM) algorithm and the Naïve Bayes (NB) algorithm for both red wine and white wine datasets.

INTRODUCTION

The quality of the wine is a very important part for the consumers as well as the manufacturing industries. Industries are increasing their sales using product quality certification. Nowadays, all over the world wine is a regularly used beverage and the industries are using the certification of product quality to increase their value in the market. Previously, testing of product quality will be done at the end of the production, this is time taking process and it requires a lot of resources such as the need for various human experts for the assessment of product quality which makes this process very expensive. Every human has their own opinion about the test, so identifying the quality of the wine based on humans experts it is a challenging task.

There are several features to predict the wine quality but the entire features will not be relevant for better prediction.

The research aims to what wine features are important to get the promising result by implementing the machine learning classification algorithms such as Support Vector Machine (SVM), Naïve Bayes (NB), and Artificial Neural Network (ANN), using the wine quality Dataset.

The wine quality dataset is publically available on the UCI machine learning repository (Cortez et al., 2009). The dataset has two files red wine and white wine variants of the Portuguese “Vinho Verde” wine. It contains a large collection of datasets that have been used for the machine learning community. The red wine dataset contains 1599 instances and the white wine dataset contains 4898 instances. Both files contain 11 input features and 1 output feature. Input features are based on the physicochemical tests and output variable based on sensory data is scaled in 11 quality classes from 0 to 10 (0-very bad to 10-very good).

Feature selection is the popular data preprocessing step for generally (Wolf and Shashua, 2005). To build the model it selects the subset of relevant features. According to the weighted of the relevance of the features, and with relatively low weighting features will be removed. This process will simplify the model and reduce the training time, and increase the performance of the model (Panday et al., 2018).

About Dataset

Context

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones).

his dataset is also available from the UCI machine learning repository, https://archive.ics.uci.edu/ml/datasets/wine+quality , I just shared it to kaggle for convenience. (If I am mistaken and the public license type disallowed me from doing so, I will take this down if requested.)

Content

For more information, read [Cortez et al., 2009].

Input variables (based on physicochemical tests):

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 - alcohol

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

OBJECTIVES

KNIME is a great tool (GUI) that can be used for this.

1 - File Reader (for csv) to linear correlation node and to interactive histogram for basic EDA.

2- File Reader to 'Rule Engine Node' to turn the 10 point scale to dichotomous variable (good wine and rest), the code to put in the rule engine is something like this:

- **$quality$ > 6.5 => "good**
- **TRUE => "bad"**

3- Rule Engine Node output to input of Column Filter node to filter out your original 10 point feature (this prevent leaking)

4- Column Filter Node output to input of Partitioning Node (your standard train/tes split, e.g. 75%/25%, choose 'random' or 'stratified')

5- Partitioning Node train data split output to input of Train data split to input Decision Tree Learner node and

6- Partitioning Node test data split output to input Decision Tree predictor Node

7- Decision Tree learner Node output to input Decision Tree Node input

8- Decision Tree output to input ROC Node.. (here you can evaluate your model base on AUC value)

RESEARCH METHODOLOGY

The following research question and hypothesis are formulated.
1. What wine features are important to get a promising result?
The researchers have used a neural network for the regression task but for the classification task neural network was never used.
Hypothetically, the current prediction model that has been obtained by Researchers will be improved by using the neural network.

To address the research question the following objectives are formulated :

- To balance the dataset.
- To analyze the impact of the features.
- To optimize the classification models through hyperparameter tuning.
- To model and evaluate the approaches.

DATA ANALYSIS AND INTERPRETATION

DATA

	A	B	C	D	E	F	G	H	I	J	K	L
1	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
2	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
3	7.8	0.88	0	2.6	0.098	25	67	0.9968	3.2	0.68	9.8	5
4	7.8	0.76	0.04	2.3	0.092	15	54	0.997	3.26	0.65	9.8	5
5	11.2	0.28	0.56	1.9	0.075	17	60	0.998	3.16	0.58	9.8	6
6	7.4	0.7	0	1.9	0.076	11	34	0.9978	3.51	0.56	9.4	5
7	7.4	0.66	0	1.8	0.075	13	40	0.9978	3.51	0.56	9.4	5
8	7.9	0.6	0.06	1.6	0.069	15	59	0.9964	3.3	0.46	9.4	5
9	7.3	0.65	0	1.2	0.065	15	21	0.9946	3.39	0.47	10	7
10	7.8	0.58	0.02	2	0.073	9	18	0.9968	3.36	0.57	9.5	7
11	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
12	6.7	0.58	0.08	1.8	0.097	15	65	0.9959	3.28	0.54	9.2	5
13	7.5	0.5	0.36	6.1	0.071	17	102	0.9978	3.35	0.8	10.5	5
14	5.6	0.615	0	1.6	0.089	16	59	0.9943	3.58	0.52	9.9	5
15	7.8	0.61	0.29	1.6	0.114	9	29	0.9974	3.26	1.56	9.1	5
16	8.9	0.62	0.18	3.8	0.176	52	145	0.9986	3.16	0.88	9.2	5
17	8.9	0.62	0.19	3.9	0.17	51	148	0.9986	3.17	0.93	9.2	5
18	8.5	0.28	0.56	1.8	0.092	35	103	0.9969	3.3	0.75	10.5	7
19	8.1	0.56	0.28	1.7	0.368	16	56	0.9968	3.11	1.28	9.3	5
20	7.4	0.59	0.08	4.4	0.086	6	29	0.9974	3.38	0.5	9	4
21	7.9	0.32	0.51	1.8	0.341	17	56	0.9969	3.04	1.08	9.2	6
22	8.9	0.22	0.48	1.8	0.077	29	60	0.9968	3.39	0.53	9.4	6
23	7.6	0.39	0.31	2.3	0.082	23	71	0.9982	3.52	0.65	9.7	5
24	7.9	0.43	0.21	1.6	0.106	10	37	0.9966	3.17	0.91	9.5	5
25	8.5	0.49	0.11	2.3	0.084	9	67	0.9968	3.17	0.53	9.4	5
26	6.9	0.4	0.14	2.4	0.085	21	40	0.9968	3.43	0.63	9.7	6
27	6.3	0.39	0.16	1.4	0.08	11	23	0.9955	3.34	0.56	9.3	5
28	7.6	0.41	0.24	1.8	0.08	4	11	0.9962	3.28	0.59	9.5	5

	A	B	C	D	E	F	G	H	I	J	K	L
1573	6.4	0.38	0.14	2.2	0.038	15	25	0.99514	3.44	0.65	11.1	6
1574	7.3	0.69	0.32	2.2	0.069	35	104	0.99632	3.33	0.51	9.5	5
1575	6	0.58	0.2	2.4	0.075	15	50	0.99467	3.58	0.67	12.5	6
1576	5.6	0.31	0.78	13.9	0.074	23	92	0.99677	3.39	0.48	10.5	6
1577	7.5	0.52	0.4	2.2	0.06	12	20	0.99474	3.26	0.64	11.8	6
1578	8	0.3	0.63	1.6	0.081	16	29	0.99588	3.3	0.78	10.8	6
1579	6.2	0.7	0.15	5.1	0.076	13	27	0.99622	3.54	0.6	11.9	6
1580	6.8	0.67	0.15	1.8	0.118	13	20	0.9954	3.42	0.67	11.3	6
1581	6.2	0.56	0.09	1.7	0.053	24	32	0.99402	3.54	0.6	11.3	5
1582	7.4	0.35	0.33	2.4	0.068	9	26	0.9947	3.36	0.6	11.9	6
1583	6.2	0.56	0.09	1.7	0.053	24	32	0.99402	3.54	0.6	11.3	5
1584	6.1	0.715	0.1	2.6	0.053	13	27	0.99362	3.57	0.5	11.9	5
1585	6.2	0.46	0.29	2.1	0.074	32	98	0.99578	3.33	0.62	9.8	5
1586	6.7	0.32	0.44	2.4	0.061	24	34	0.99484	3.29	0.8	11.6	7
1587	7.2	0.39	0.44	2.6	0.066	22	48	0.99494	3.3	0.84	11.5	6
1588	7.5	0.31	0.41	2.4	0.065	34	60	0.99492	3.34	0.85	11.4	6
1589	5.8	0.61	0.11	1.8	0.066	18	28	0.99483	3.55	0.66	10.9	6
1590	7.2	0.66	0.33	2.5	0.068	34	102	0.99414	3.27	0.78	12.8	6
1591	6.6	0.725	0.2	7.8	0.073	29	79	0.9977	3.29	0.54	9.2	5
1592	6.3	0.55	0.15	1.8	0.077	26	35	0.99314	3.32	0.82	11.6	6
1593	5.4	0.74	0.09	1.7	0.089	16	26	0.99402	3.67	0.56	11.6	6
1594	6.3	0.51	0.13	2.3	0.076	29	40	0.99574	3.42	0.75	11	6
1595	6.8	0.62	0.08	1.9	0.068	28	38	0.99651	3.42	0.82	9.5	6
1596	6.2	0.6	0.08	2	0.09	32	44	0.9949	3.45	0.58	10.5	5
1597	5.9	0.55	0.1	2.2	0.062	39	51	0.99512	3.52	0.76	11.2	6
1598	6.3	0.51	0.13	2.3	0.076	29	40	0.99574	3.42	0.75	11	6
1599	5.9	0.645	0.12	2	0.075	32	44	0.99547	3.57	0.71	10.2	5
1600	6	0.31	0.47	3.6	0.067	18	42	0.99549	3.39	0.66	11	6

Importing data in r tool

```
> data=read.csv(file.choose())
> data
```

RGui (64-bit) - [R Console]

File Edit View Misc Packages Windows Help

491	9.3	0.775	0.27	2.80	0.078
492	9.2	0.410	0.50	2.50	0.055
493	8.9	0.400	0.51	2.60	0.052
494	8.7	0.690	0.31	3.00	0.086
495	6.5	0.390	0.23	8.30	0.051
496	10.7	0.350	0.53	2.60	0.070
497	7.8	0.520	0.25	1.90	0.081
498	7.2	0.340	0.32	2.50	0.090
499	10.7	0.350	0.53	2.60	0.070
500	8.7	0.690	0.31	3.00	0.086
501	7.8	0.520	0.25	1.90	0.081
502	10.4	0.440	0.73	6.55	0.074
503	10.4	0.440	0.73	6.55	0.074
504	10.5	0.260	0.47	1.90	0.078
505	10.5	0.240	0.42	1.80	0.077
506	10.2	0.490	0.63	2.90	0.072
507	10.4	0.240	0.46	1.80	0.075
508	11.2	0.670	0.55	2.30	0.084
509	10.0	0.590	0.31	2.20	0.090
510	13.3	0.290	0.75	2.80	0.084
511	12.4	0.420	0.49	4.60	0.073
512	10.0	0.590	0.31	2.20	0.090
513	10.7	0.400	0.48	2.10	0.125
514	10.5	0.510	0.64	2.40	0.107
515	10.5	0.510	0.64	2.40	0.107
516	8.5	0.655	0.49	6.10	0.122
517	12.5	0.600	0.49	4.30	0.100
518	10.4	0.610	0.49	2.10	0.200
519	10.9	0.210	0.49	2.80	0.088
520	7.3	0.365	0.49	2.50	0.088
521	9.8	0.250	0.49	2.70	0.088
522	7.6	0.410	0.49	2.00	0.088
523	8.2	0.390	0.49	2.30	0.099
524	9.3	0.400	0.49	2.50	0.085
525	9.2	0.430	0.49	2.40	0.086
526	10.4	0.640	0.24	2.80	0.105
527	7.3	0.365	0.49	2.50	0.088
528	7.0	0.380	0.49	2.50	0.097
529	8.2	0.420	0.49	2.60	0.084
530	9.9	0.630	0.24	2.40	0.077
531	9.1	0.220	0.24	2.10	0.078
532	11.9	0.380	0.49	2.70	0.098
533	11.9	0.380	0.49	2.70	0.098
534	10.3	0.270	0.24	2.10	0.072
535	10.0	0.480	0.24	2.70	0.102
536	9.1	0.220	0.24	2.10	0.078

RGui (64-bit) - [R Console]

File Edit View Misc Packages Windows Help

1554	18.0	28.0	0.99765	3.41	0.60	9.400000
1555	15.0	23.0	0.99627	3.54	0.60	11.000000
1556	15.0	24.0	0.99514	3.44	0.68	10.550000
1557	12.0	20.0	0.99636	3.53	0.56	9.900000
1558	15.0	23.0	0.99627	3.54	0.60	11.000000
1559	66.0	115.0	0.99787	3.22	0.56	9.500000
1560	31.0	131.0	0.99622	3.21	0.52	9.900000
1561	31.0	131.0	0.99622	3.21	0.52	9.900000
1562	31.0	131.0	0.99622	3.21	0.52	9.900000
1563	12.0	20.0	0.99546	3.29	0.54	10.100000
1564	12.0	20.0	0.99546	3.29	0.54	10.100000
1565	12.0	20.0	0.99546	3.29	0.54	10.100000
1566	26.0	42.0	0.99489	3.39	0.82	10.900000
1567	24.0	52.0	0.99494	3.34	0.71	11.200000
1568	12.0	20.0	0.99546	3.29	0.54	10.100000
1569	25.0	42.0	0.99629	3.34	0.59	9.200000
1570	15.0	34.0	0.99396	3.48	0.57	11.500000
1571	19.0	35.0	0.99340	3.37	0.93	12.400000
1572	15.0	25.0	0.99514	3.44	0.65	11.100000
1573	35.0	104.0	0.99632	3.33	0.51	9.500000
1574	15.0	50.0	0.99467	3.58	0.67	12.500000
1575	23.0	92.0	0.99677	3.39	0.48	10.500000
1576	12.0	20.0	0.99474	3.26	0.64	11.800000
1577	16.0	29.0	0.99588	3.30	0.78	10.800000
1578	13.0	27.0	0.99622	3.54	0.60	11.900000
1579	13.0	20.0	0.99540	3.42	0.67	11.300000
1580	24.0	32.0	0.99402	3.54	0.60	11.300000
1581	9.0	26.0	0.99470	3.36	0.60	11.900000
1582	24.0	32.0	0.99402	3.54	0.60	11.300000
1583	13.0	27.0	0.99362	3.57	0.50	11.900000
1584	32.0	98.0	0.99578	3.33	0.62	9.800000
1585	24.0	34.0	0.99484	3.29	0.80	11.600000
1586	22.0	48.0	0.99494	3.30	0.84	11.500000
1587	34.0	60.0	0.99492	3.34	0.85	11.400000
1588	18.0	28.0	0.99483	3.55	0.66	10.900000
1589	34.0	102.0	0.99414	3.27	0.78	12.800000
1590	29.0	79.0	0.99770	3.29	0.54	9.200000
1591	26.0	35.0	0.99314	3.32	0.82	11.600000
1592	16.0	26.0	0.99402	3.67	0.56	11.600000
1593	29.0	40.0	0.99574	3.42	0.75	11.000000
1594	28.0	38.0	0.99651	3.42	0.82	9.500000
1595	32.0	44.0	0.99490	3.45	0.58	10.500000
1596	39.0	51.0	0.99512	3.52	0.76	11.200000
1597	29.0	40.0	0.99574	3.42	0.75	11.000000
1598	32.0	44.0	0.99547	3.57	0.71	10.200000
1599	18.0	42.0	0.99549	3.39	0.66	11.000000

Type here to search

Summary Of data

```
> summary(data)
 fixed.acidity   volatile.acidity  citric.acid    residual.sugar     chlorides       free.sulfur.dioxide total.sulfur.dioxide    density            pH          sulphates     
 Min.   : 4.60   Min.   :0.1200   Min.   :0.000   Min.   : 0.900   Min.   :0.01200   Min.   : 1.00       Min.   :  6.00       Min.   :0.9901   Min.   :2.740   Min.   :0.3300  
 1st Qu.: 7.10   1st Qu.:0.3900   1st Qu.:0.090   1st Qu.: 1.900   1st Qu.:0.07000   1st Qu.: 7.00       1st Qu.: 22.00       1st Qu.:0.9956   1st Qu.:3.210   1st Qu.:0.5500  
 Median : 7.90   Median :0.5200   Median :0.260   Median : 2.200   Median :0.07900   Median :14.00       Median : 38.00       Median :0.9968   Median :3.310   Median :0.6200  
 Mean   : 8.32   Mean   :0.5278   Mean   :0.271   Mean   : 2.539   Mean   :0.08747   Mean   :15.87       Mean   : 46.47       Mean   :0.9967   Mean   :3.311   Mean   :0.6581  
 3rd Qu.: 9.20   3rd Qu.:0.6400   3rd Qu.:0.420   3rd Qu.: 2.600   3rd Qu.:0.09000   3rd Qu.:21.00       3rd Qu.: 62.00       3rd Qu.:0.9978   3rd Qu.:3.400   3rd Qu.:0.7300  
 Max.   :15.90   Max.   :1.5800   Max.   :1.000   Max.   :15.500   Max.   :0.61100   Max.   :72.00       Max.   :289.00       Max.   :1.0037   Max.   :4.010   Max.   :2.0000  
    alcohol         quality     
 Min.   : 8.40   Min.   :3.000  
 1st Qu.: 9.50   1st Qu.:5.000  
 Median :10.20   Median :6.000  
 Mean   :10.42   Mean   :5.636  
 3rd Qu.:11.10   3rd Qu.:6.000  
 Max.   :14.90   Max.   :8.000  
```

1) Mean (Alcohol)

```
> mean(data$alcohol)
[1] 10.42298
> 
```

2) Median (Alcohol)

```
> median(data$alcohol)
[1] 10.2
```

3) Sd (Alcohol)

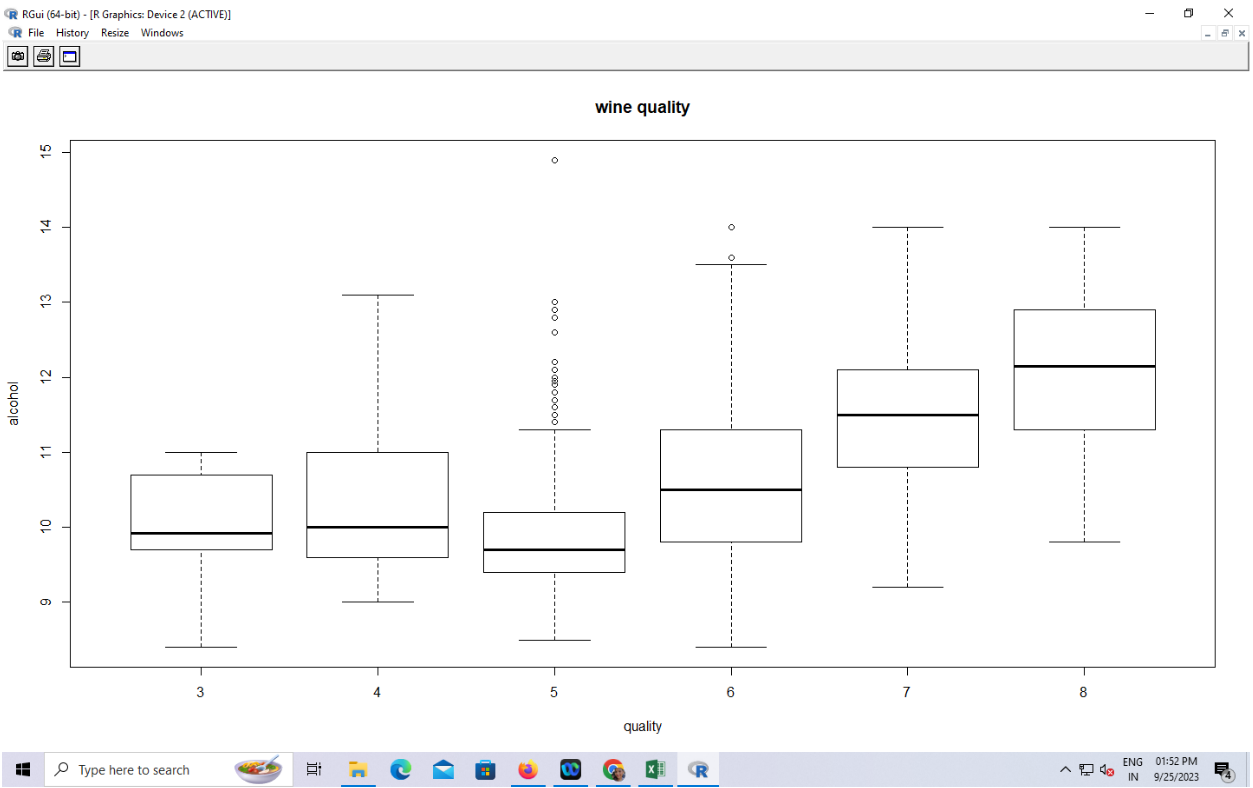
```
> sd(data$alcohol)
[1] 1.065668
```

4) Cv (Alcohol)

```
> cv<-sd(data$alcohol)/mean(data$alcohol)*100
> cv
[1] 10.22421
> 
```

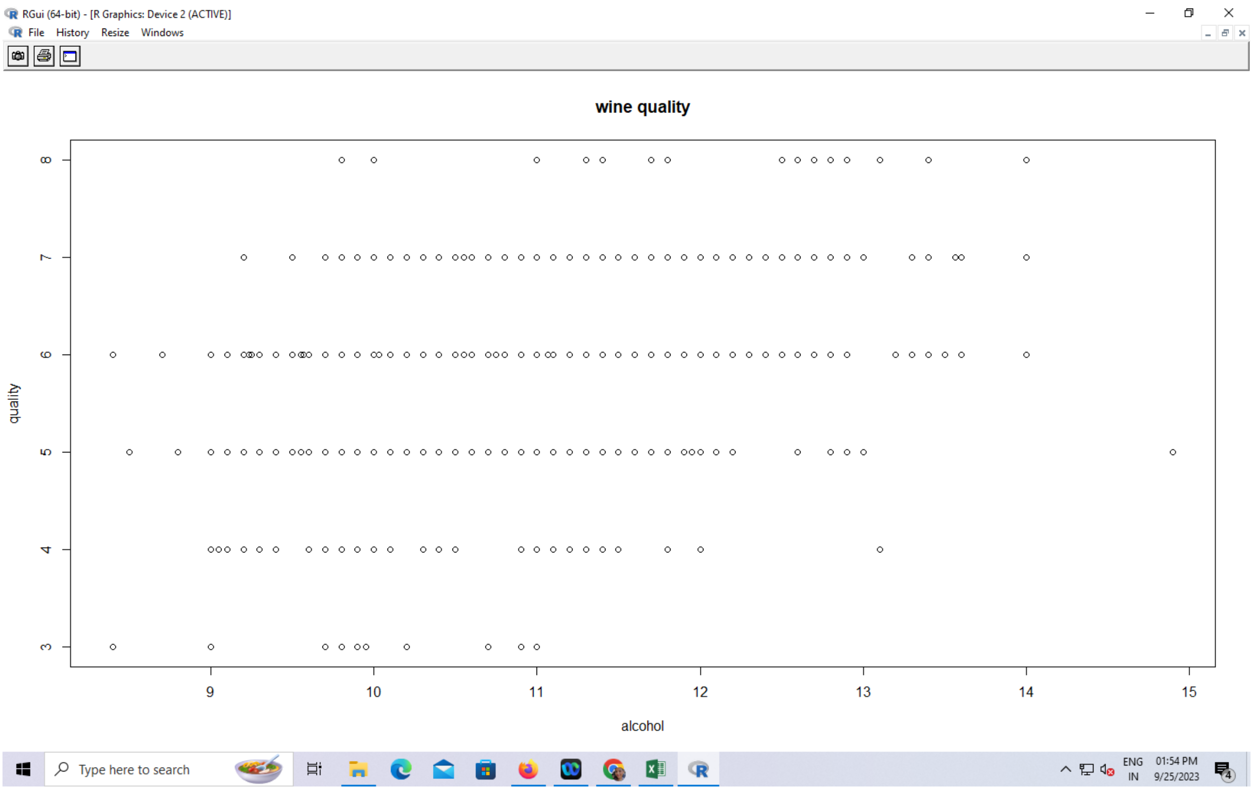
5) Boxplot (Alcohol as per Quality)

```
>
>
> boxplot(data$alcohol ~ data$quality, data=data, main="wine quality", xlab="quality", ylab="alcohol")
>
```



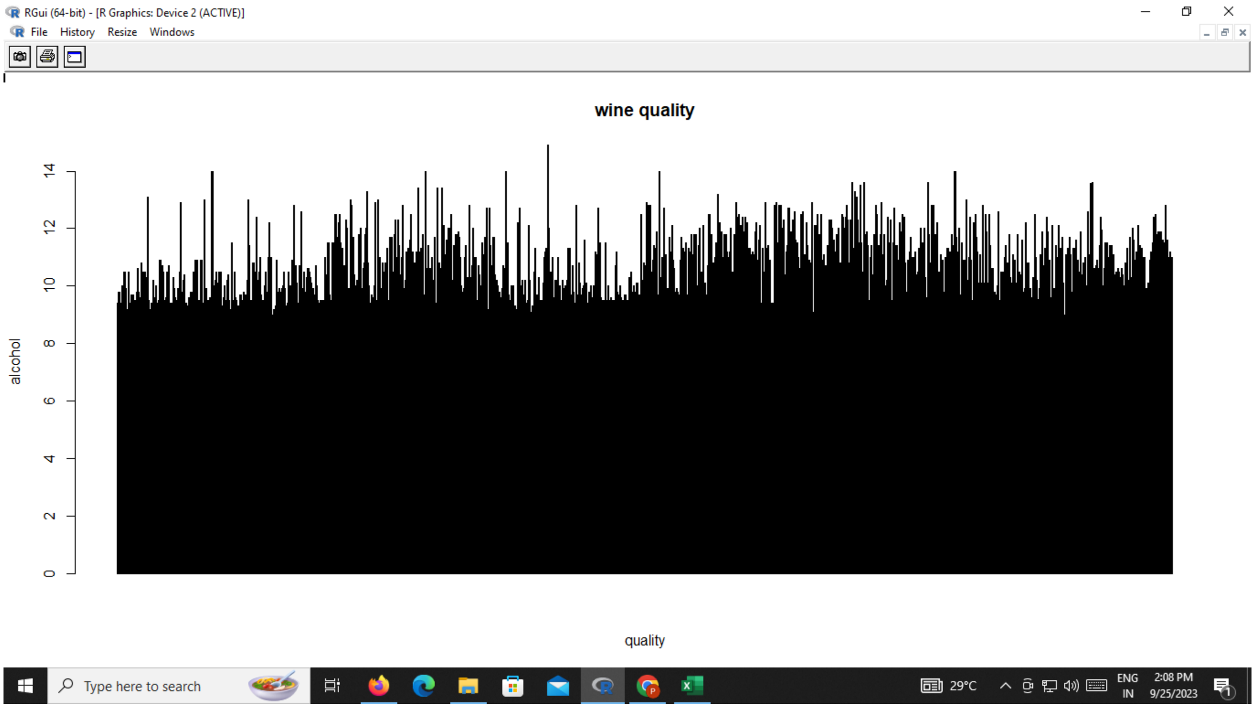
6) Scatterplot(Quality as per Alcohol)

```
>
> plot(x=data$alcohol, y=data$quality, xlab="alcohol",ylab="quality", main="wine quality")
>
```



7) Bar Graph (Alcohol as Quality)

```
> barplot(data$alcohol,name.arg=data$quality,col="green",border="black",main="wine quality",xlab="quality",ylab="alcohol")
```



8) Hypothesis Testing - Chi-square test

H0 :- Alcohol and quality are dependent
H1 :- Alcohol and quality are Not dependent

```
> chisq.test(data$alcohol,data$quality)

        Pearson's Chi-squared test

data:  data$alcohol and data$quality
X-squared = 1124.5, df = 320, p-value < 2.2e-16

Warning message:
In chisq.test(data$alcohol, data$quality) :
  Chi-squared approximation may be incorrect
> 
```

P-<0.05,reject H0
Hence Alcohol and quality are not dependent

9) T - test

H0: SIGNIFICANT difference between two group
H1: NO SIGNIFICANT difference between two group

```
> t.test(data$alcohol,data$quality)

        Welch Two Sample t-test

data:  data$alcohol and data$quality
t = 143.16, df = 2978.2, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.721397 4.852524
sample estimates:
mean of x mean of y 
10.422983  5.636023 

> |
```

P-<0.05,reject H0
Hence, NO SIGNIFICANT difference between two group

10) ANOVA

```
> anova(lm(data$alcohol~data$quality))
Analysis of Variance Table

Response: data$alcohol
               Df  Sum Sq Mean Sq F value    Pr(>F)    
data$quality    1  411.47  411.47  468.27 < 2.2e-16 ***
Residuals    1597 1403.30    0.88                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

FINDINGS & CONCLUSIONS

At the end of the story I can say that Wine quality is a very complex study.Good wine is more than a perfect combination of different chemical components. Future improvement can be made if more data can be collected on both low-quality and high-quality wine.If the data set has more records on both the low end and high end, the quality of analysis can be improved. We can be more certain about whether there is a significant correlation between a chemical component and the wine quality.

End of Red Wine Quality Analysis Final

REFERENCE

- Guidance from **Prof. Sabitha Praveen**

- Kaggle.com
  Dataset Link
  :https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009?select=winequality-red.csv
- Geekforgeeks.Org
- Wikipedia

THANK YOU