CEE 298
Final Report: Wine Quality
Professor Bauchy
Benson Zu

1)Introduction:

The quality of the wine is personal. The criteria of wine seem mostly to be decided by the consumers. Few people can make a sterling decision without tasting the wine themselves. Personally, quality depends on both sensory and context-depending, such as the aroma of the wine and your circumstances. Some people like this experience in exploring unknown tastes and flavors of the wine, but the unpredictable quality of the wine confused people while making selections on wines. Indeed, people justify the quality of the wine based on their outlooks, price, and other reviews, which usually seem to be inconsistent. This project is working on analyzing the quality of the wine based on its attributes determined by the physicochemical tests. Based on the objective tests, 11 different characteristics of wines (eg. pH) will be analyzed to determine the general quality of them

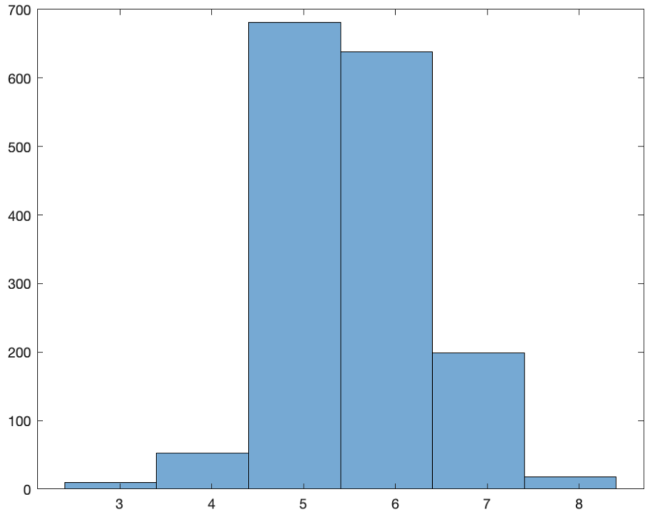
2)Methods

2.1 Data Acquisition

This dataset is downloaded from the research "Modeling wine preferences by data mining from physicochemical properties" in Decision Support Systems, Elsevier【47(4):547-553. ISSN: 0167-9236】. It is publicly available for research and can be downloaded from the UC Irvine Machine Learning Repository datasets, in which 1599 samples of red wines are created. 11 different characteristic, including its 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) sulfates, 11) alcohol content.
The inputs include objective tests (e.g. PH values) and the output is based on sensory data (median of at least 3 evaluations made by wine experts). Each expert graded the wine quality between 0 (very bad) and 10 (very excellent).
[Picture of the data]

2.2 Inspection of the data

Even though the dataset intends to classify and differentiate the quality of the wines quantitatively in the range from 0 (very bad) to 10 (very excellent), the distribution of the data (A the graph shows below) is extremely unevenly distributed. More than 1250 samples are rated as 5 or 6 points out of 10, which entails 80 percent of the data. Additionally, there are no wines that are rated as very bad (1 or 2 points) or vercy excellent (9 or 10 points). The



other samples for each grade are too few to make a justified analysis. Therefore, instead of using a linear regression model for analysis, this project will apply logistic regression to classify the wines into “Above Average Quality” and “the Others” to learn the quality of the wines in terms of its objective attributes.

2.3 Pre-Classification

The original dataset was processed to distinguish the high quality of wine and low quality of the wine in terms of the median of 5 points by using R language. `1` represents the” Above Average Quality” of the wine, and `0` represents “the Others”.

```
red_wine <- read_csv("~/Desktop/2020_Spring/CEE298/FinalProject/winequality-red_new.csv")
red_wine <- mutate(red_wine, Quality=ifelse(quality>5, 1, 0))
write_csv(red_wine, "~/Desktop/2020_Spring/CEE298/FinalProject/winequality-red_new2.csv")
```

[Picture of the data]

2.4 Classification

MATLA_R2020a was used to execute the machine learning program. The `winequality-red_new2.csv` file was imported by using `readmatrix` function. The input objective attributes are assigned as `x` and the class of quality is assigned as `y`. Since the logistic model needed all positive inputs and y values are assigned as 0 and 1, new `y` is created as `y + 1`.

```
clc
clear all
close all
data=readmatrix('winequality-red_new2.csv');
x = data(:,1:11);
y=data(:,13);
y=y+1;
m=length(y);
```

Then, 60, 70, 80, 90 percent of the data would be used for training purposes. Random selection of the training data are assigned as `xtrain and ytrain`. The rest of them would be stored in `xtest` and `test.` Then, multinomial logistic regression was used to give the coefficient of the model, which is stored in `B.` Coefficient in B was used to calculate the predicted features of the wine quality, and the results were stored in ztrain. and Applying the education of hypothesis that `1.0./(1.0+exp(-ztrain))` ,the hypothesis is assigned as `htrain.`

```
idx=randperm(m);P=0.7;
xtrain=x(idx(1:round(P*m)),:);
ytrain=y(idx(1:round(P*m)),:);
xtest=x(idx(round(P*m)+1:end),:);
ytest=y(idx(round(P*m)+1:end),:);
B = mnrfit(xtrain,ytrain);
```

```
mtrain=length(ytrain);

xtrain2=[ones(mtrain,1) xtrain];

ztrain = xtrain2*B;

htrain=1.0./(1.0+exp(-ztrain));
```

Then the same methodology was used to predict the values for the test sets

```
%test sets

mtest=length(ytest);

xtest2=[ones(mtest,1) xtest];%new training feature vetor

ztest = xtest2*B;%beta0+beta1+beta2

%hypothesis

htest=1.0./(1.0+exp(-ztest)); %sigmoid (z)
```

To evaluate the effectiveness of the prediction, its precision and recall values would be calculated. Accuracy indicates the percentage of correct prediction of “Above Average Quality” wine, and the amount of revelation of “Above Average Quality” wine among all samples. If `htest or htrain` is larger than 0.5, the wine would be considered as an “Above Average Quality” wine; otherwise, it belongs to the other class.

```
%Train set

 ytrainpred=htrain < 0.5;%output 0 and 1, 1 means TRUE, 0 means FALSE

ytrainpred=ytrainpred+1;

accuracy_training=mean(double(ytrainpred == ytrain));%equals 1 when
true (equal) otherwise 0 (false)

%test sets

ytestpred=htest < 0.5;%output 0 and 1, 1 means TRUE, 0 means FALSE

ytestpred=ytestpred+1;

accuracy_test=mean(double(ytestpred == ytest));

%precision and recall

true_rainy=sum(double(ytest==2));

predicted_rainy=sum(double(ytestpred==2));

true_positive = sum(double(ytest==2).*double(ytestpred==2));

%presion

precision = true_positive/predicted_rainy

%recall

recall=true_positive/true_rainy
```

Then, the results of recall and precision values are used to calculate the F.

```
F1=2*precision*recall/(precision+recall)
```

3)Results

The test set for choosing 60, 70, 80, 90 percent of the data are shown below:

Training Data Portion	Precision	Recall	F
60% of the dataset	0.7545	0.7522	0.7534
70% of the dataset	0.7823	0.7760	0.7791
80% of the dataset	0.7866	0.7588	0.7725
90% of the dataset	0.7654	0.7470	0.7561

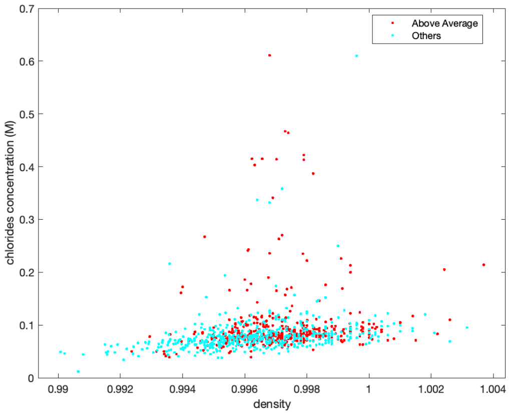
Based on the results, choosing 70% percent of data as the training sets provides us the best results with the highest accuracy. The coefficient values of the 70% test set are shown below:

B0	-57.4622798415630
B1	-0.146440046510185
B2	3.61207000378626
B3	1.34288973413268
B4	-0.0538438703499179
B5	2.61634267123348
B6	-0.0187513561685463
B7	0.0172142161183537
B8	66.0882338535486
B9	0.0282249011119674
B10	-2.52189425375297
B11	-0.832297663404792

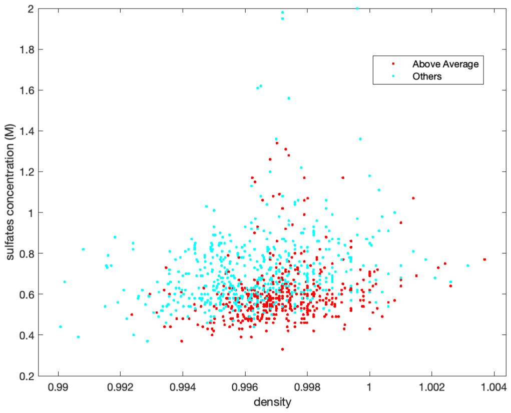
4)Discussion

As we can see from the B values table, input 2(volatile acidity), 5(chlorides), 8(density), and 10(sulfates) have a relatively significant effect on the results. The 2 dimension graphs defined by pairs of these four attributes are plotted to visualize the results find the most important factor in determining the wine quality.

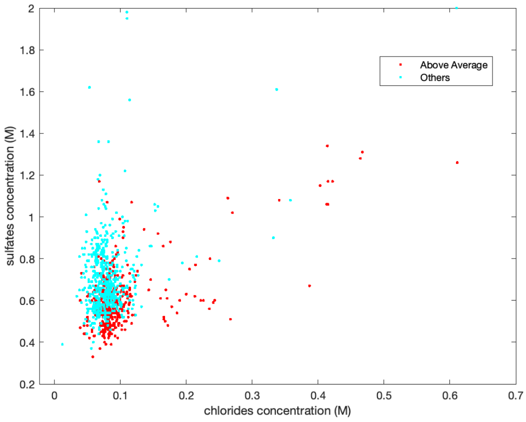
8(density) v.s. 5(chlorides)



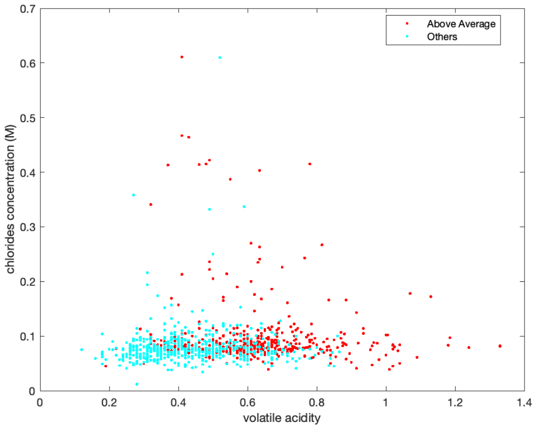
8(density) v.s 10(sulfates)



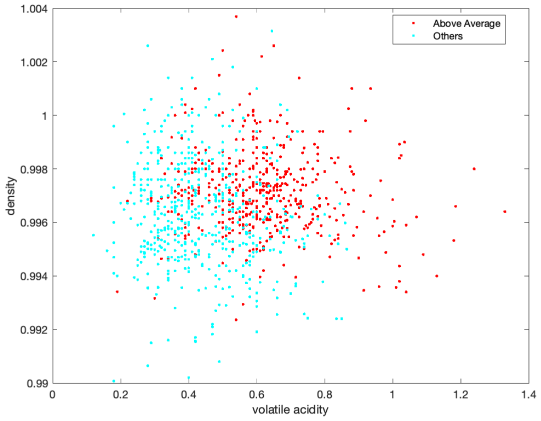
5(chlorides) v.s 10(sulfates)



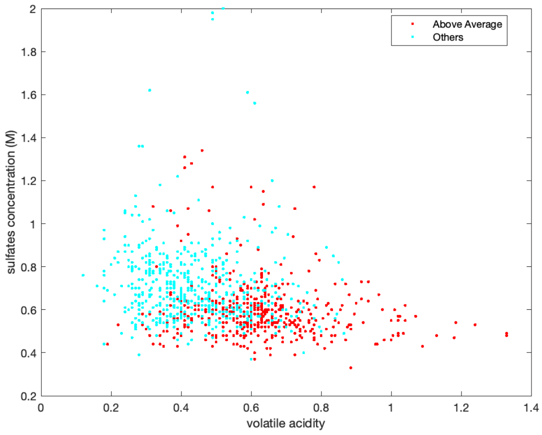
2(volatile acidity) vs. 5(chlorides)



2(volatile acidity) vs. 8(density)



2(volatile acidity) vs. 10(sulfates)



As the graphs above show, the density, chloride concentrations, and sulfates concentrations show a very vague relationship with each other. However, it is surprised to see that higher volatile acidity(VA) always shows a trend to have a better quality of the wine. It seems that it plays a relatively more significant role across all attributes of the wine. It seems that from the concentration from 0.2 g/L to 0.3 g/L, volatile acidity has little effect on wine's quality. At higher levels, however, the VA can give the wine a better taste in the concentration from 0.4g/L to 0.8 g/L. Nevertheless, these results of the study do not show the direct causation relationship between VA and the quality of the wine.

With additional literature reviews about the effect of Volatile acidity on the quality of the wine, I found out that “volatile acidity is derived from acids of the acetic series present in wine” (OIA 2009). Based on OIA, the volatile acidity of the wine should be kept low because an excessive amount of the volatile acidity will cause an unpleasant vinegar taste of the wine (OIA 2009). Typically, the acetic acid in the wine in concentrations is ranging from 0.2g/L to 0.8g/L, and the maximum boundary should be at 1.2 g/L (Boulton et al. 1996). If the acetic acid concentration is above 0.9 g/L, it will produce a discernible bitter and sour aftertaste in wine, which correspond with our study results that the wine samples with volatile acidity from 0.4g/L to 0.8 g/L provides the most “Above Average Quality” of the wine.

5)Conclusion

This is a project to study the sensory output based on the objective inputs by using logistic regression model in machine learning. According to our report, the wine with above-average quality shares the features of volatile acidity ranging from 0.4 to 0.8 g/L. The study results could be explored and applied to oversee conditions of the wine fermentation. By supervising the chemical composition in the wine, the manufacturers can monitor the quality of their wine based on quantitative analysis instead of inconsistent personal sensory data. However, the prediction results based on the objective tests need a further chemical study to prove their correlation and causation effect with the wine quality.

6)References

Bely M, Rinaldi A, Dubourdieu D (2003) Influence of assimilable nitrogen on volatile acidity production by Saccharomyces cerevisiae during high sugar fermentation. J Biosci Bioeng 96:507–512

Office Internationale de la Vigne et du Vin (2009) Compendium of international methods of wine and must analysis. Vol1 OIV, Paris, p 419

Vilela-Moura, A., Schuller, D., Mendes-Faia, A. et al. The impact of acetate metabolism on yeast fermentative performance and wine quality: reduction of volatile acidity of grape musts and wines. Appl Microbiol Biotechnol 89, 271–280 (2011). https://doi.org/10.1007/s00253-010-2898-3