# UCI Wine Dataset Analysis

Jingyan Liu(jl3292), Haoyang Li(hl2237), Jiachen Zhao(jz799), Xiaomeng Qin(xq72)

**Abstract**

This project analyses the UCI wine quality data set. For convenience, we will merge the dataset of white wine and red wine together.

In this project, we will first explore the data by looking into the correlation between variables, then we will preprocess data with Principal Data Analysis. After that, we will apply the results of factor analysis to build linear regression models using Lasso Regression and Ridge Regression. Since the quality is recorded in a discrete way, we also applied classification methods including Support Vector Machine and K-Nearest Neighbour to predict the quality.

At last, we will evaluate the performance of each method.

## 1. Exploring the Data Set

### *1.1. Variables*

By the R output, we can get the variables which are fixed.acidity, volatile.acidity, citric.acid, residual.sugar, chlorides, free. sulfur.dioxides, total.sulfur.dioxide, density, pH, sulphates, alcohol, quality.

fixed.acidity - most acids involved with wine or fixed or nonvolatile (do not evaporate readily)

volatile.acidity - the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste.

citric.acid - a colorless weak organic acid.

residual.sugar - from natural grape sugars leftover in a wine after the alcoholic fermentation finishes.

chlorides - Chemicals produced during wine fermentation.

free.sulfur.dioxides - Chemicals produced during wine fermentation.

total sulfur dioxide - Chemicals produced during wine fermentation.

density - the substance's mass per unit of volume.

pH - a scale used to designate the acidity or alkalinity of wine.

sulphates - Chemicals produced during wine fermentation.

alcohol - amount of alcohol by volume.

quality - the grade of the wine

### *1.2. Correlation*

From the plot of correlation, wine quality has largest correlation with alcohol, and also has relatively high correlation with chlorides, volatile acidity and density.

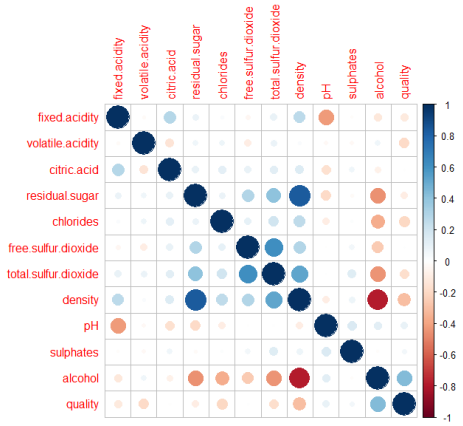


Figure 1: Correlation between Variables

The following figures show the correlation between quality and the relevant variables fitted by linear models.

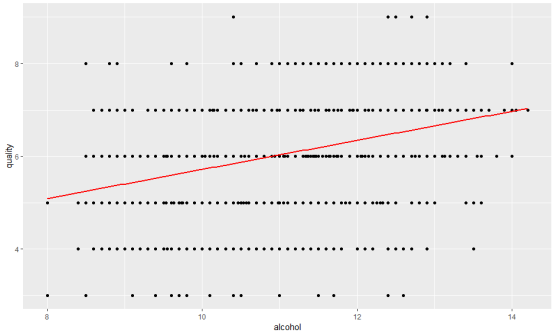


Figure 2: Correlation between Wine Quality and Alcohol

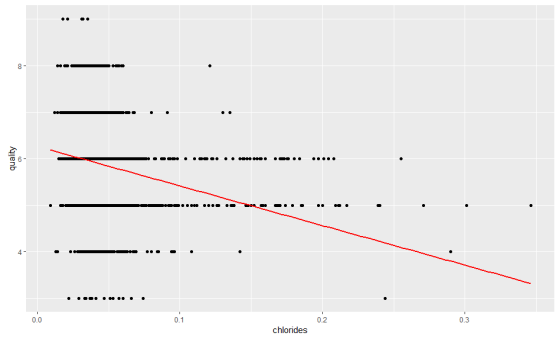


Figure 3: Correlation between Wine Quality and Chlorides

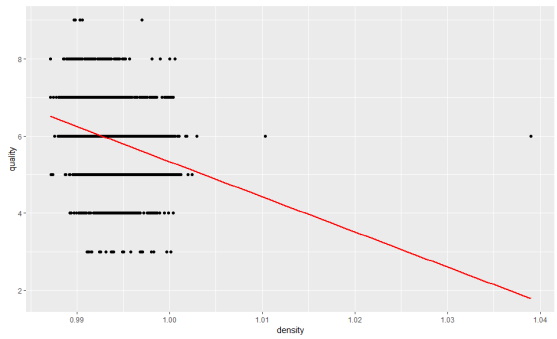


Figure 4: Correlation between Wine Quality and Density

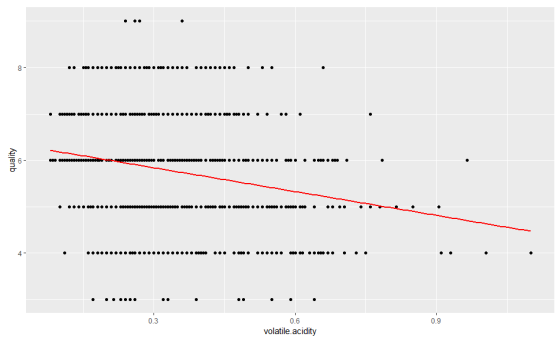


Figure 5: Correlation between Wine Quality and Volatile acidity

Since alcohol is the variable with highest correlation, variables correlated to it also worth looking into. They are chlorides, density, total sulfur dioxide, residual sugar and free sulfur dioxide. The following are relationships between alcohol and the relevant variables. The lines in the figures are fitted by linear models.

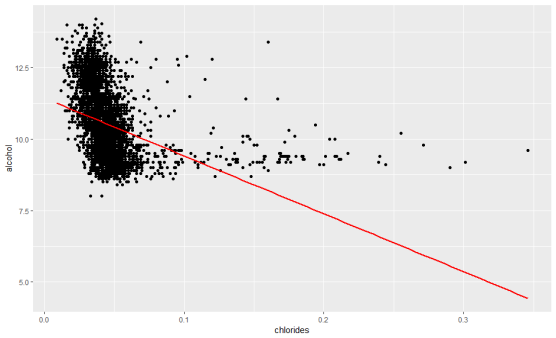


Figure 6: Correlation between Alcohol and Chlorides

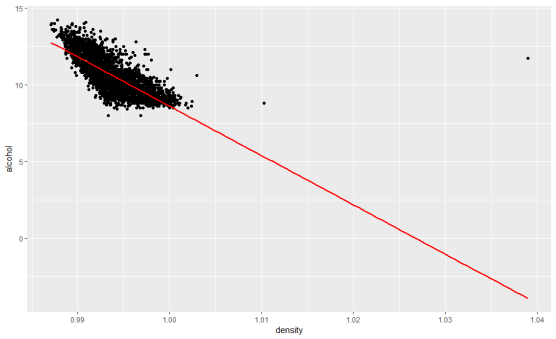


Figure 7: Correlation between Alcohol and Density

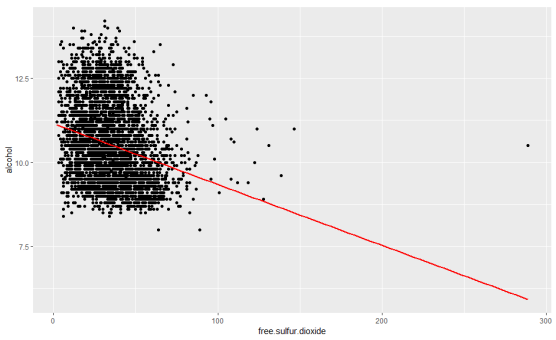


Figure 8: Correlation between Wine Alcohol and Free Sulfur Dioxide

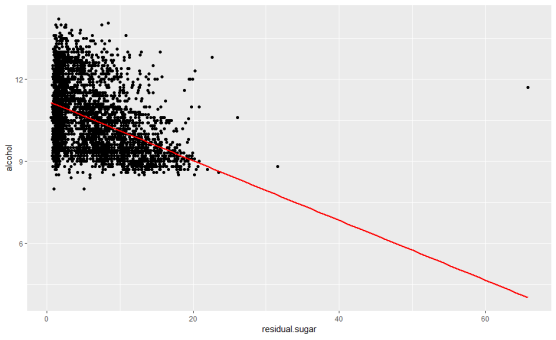


Figure 9: Correlation between Alcohol and Residual Sugar

### 1.3. Variable Names Used in the Project

| Input Variables | | |
|---|---|---|
| Names | Names in code | Description |
| Fixed acidity | $fix\_ac$ | Most acids involved with wine or fixed or nonvolatile (do not evaporate readily) |
| Volatile acidity | $vol\_ac$ | The amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste |
| Citric acid | $ci\_ac$ | A colorless weak organic acid |
| Residual sugar | $res\_sugar$ | From natural grape sugars leftover in a wine after the alcoholic fermentation finishes |
| Chlorides | $ch$ | Chemicals produced during wine fermentation |
| Free sulfur dioxide | $free\_sd$ | Chemicals produced during wine fermentatio |
| Total sulfur dioxide | $total\_sd$ | Chemicals produced during wine fermentation |
| Density | $density$ | The substance's mass per unit of volume |
| pH | $pH$ | A scale used to designate the acidity or alkalinity of wine |
| Sulphates | $sp$ | Chemicals produced during wine fermentation |
| Alcohol | $alcohol$ | Amount of alcohol by volume. |
| Output Variable | | |
| Quality | $y$ | The grade of the wine |

## 2. Dimension Reduction

### 2.1. Principal Component Analysis

We try to reduce dimension and get a subset of variables which will influence the quality of wines more compared with the rest variables. We applied Principle Component Analysis in order to address the influential variables.

Before that, we find the variance inflation factors (VIF) of each variable. According to Figure 10, It is clear than the VIF of density is greater than 5, which are two potentially concern.

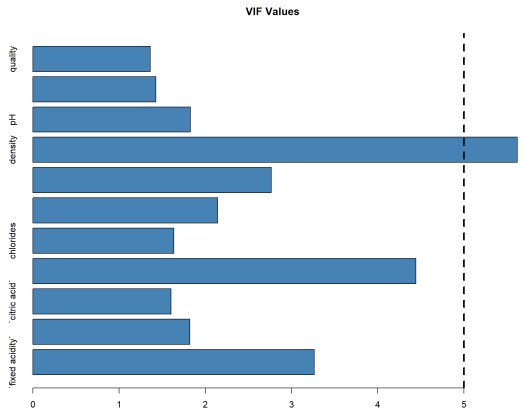


Figure 10: Barplot for VIF Values

To figure out a subset of variables that mainly impact on our result, we applied PCA at this stage. First, we would like to clarify the assumptions for PCA: (1) variables are continuous; (2) There is a linear relationship between all variables; (3) sampling adequacy; (4) no significant outliers. Based on the assumptions, we prefer PCA with 6 factors, the 6 principle components can explain almost 100% variance. PCA gives us more information about the training data, and is more suitable for further analysis.

Now, we use a scree plot to decide on a dimension reduction, and justify the choice. According to Figure 11, 12, and 13, it shows the scree plots for individual proportion of variance, accumulative proportion of variance, and eigenvalues. We believe a good model should be able to explain at least 50% variance, and the desired dimension cannot be too close to the original dimension. Based on these criteria, the appropriate number of components should be 2.

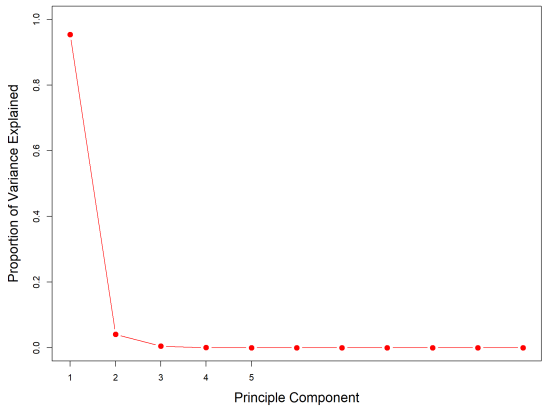


Figure 11: individual proportion of variance

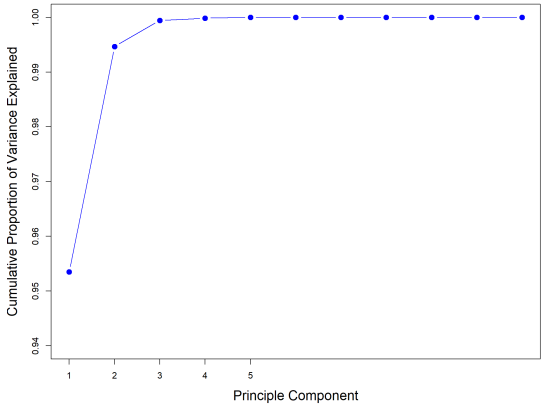


Figure 12: accumulative proportion of variance

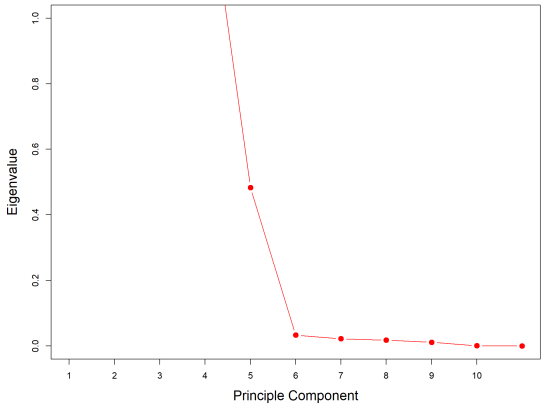


Figure 13: eigenvalues

### *2.2. Factor Analysis*

In order to find the critical factors, we applied factor analysis. Based on the previous result form PCA, we prefer to choose the factor model with 2 factors. The first factors assigns high loading to fixed acidity, total sulfur dioxide, and free sulfur dioxide. The second factors puts more wights on density, residual sugar, and alcohol.

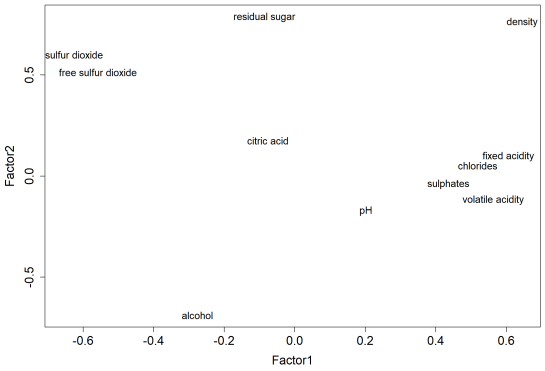


Figure 14: factors plot

Based on Figure X, the scatter plot of factor scores, most points are grouped in the left-bottom region and we assume this is a normal region for these two factors. We detect extreme values with either high content for one of the factors, or unbalanced content. We suggest a closer look at these individuals for specific reasons.

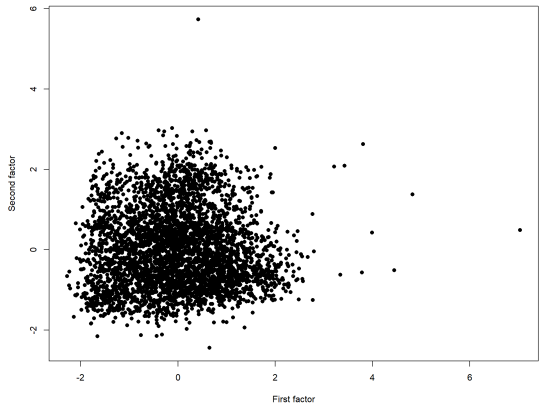


Figure 15: scatter plot of factor scores

Also, we should emphasize that the two-factor model can explain about 99% variance of the original data. This dimension reduction approach via factor model is quite reliable.

## 3. Models and Performance

By observing the relation between variables, it exists a multivariate relation between the factors and the response. According to the dimension reduction, we reduced variables into 6. And we figure our the six crucial factors, fixed acidity, total sulfur dioxide, free sulfur dioxide, density, residual sugar, and alcohol. At this stage, we will apply the six substances into our linear models.

### *3.1. Linear Model - Lasso Regression*

Lasso regression is a regression analysis method that performs variable selection and regularization. Next, using the LASSO method, I devised a model that performs variable selection and regularization. The equation of lasso regression is shown below.

$$RSS + \lambda \sum_{j=1}^{p} |\beta_j|$$

This second term in the equation is known as a shrinkage penalty. We choose 6 factors which are fixed acidity, total sulfur dioxide, free sulfur dioxide, density, residual sugar, and alcohol finding by dimension reduction to fit the model. In lasso regression, we select a value for $\lambda$ that produces the lowest possible test MSE (mean squared error). Figure X1 visualizes how the value of $\lambda$ affects the test MSE. We can find that $\lambda = 0.00047$ will minimize the test MSE. Then the coefficient values for each variable are shown in Figure. In the end, we compute the test MSE by using the testing data.

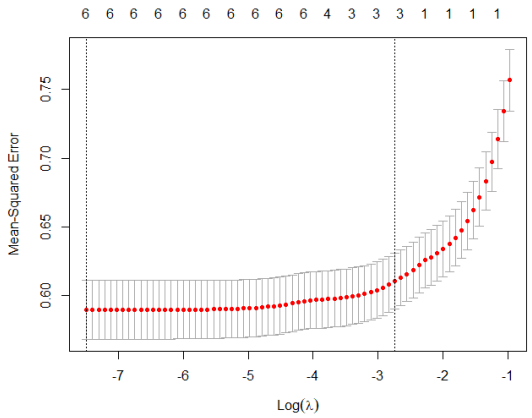


Figure 16: Best lambda of lasso regression

### *3.2. Linear Model - Ridge Regression*

Ridge regression is a method we can use to fit a regression model when multicollinearity is present in the data. The equation of ridge regression is shown below.

$$RSS + \lambda \sum_{j=1}^{p} \beta_j^2$$

This second term in the equation is known as a shrinkage penalty. In ridge regression, we select a value for $\lambda$ that produces the lowest possible test MSE (mean squared error). According to the Figure X2, it visualized how the value of $\lambda$ affect the test MSE. We can find the $\lambda = 0.01$ will minimize the test MSE. Then the coefficient values for each variable are shown in Figure. At the end, we compute the test MSE by using the testing data.

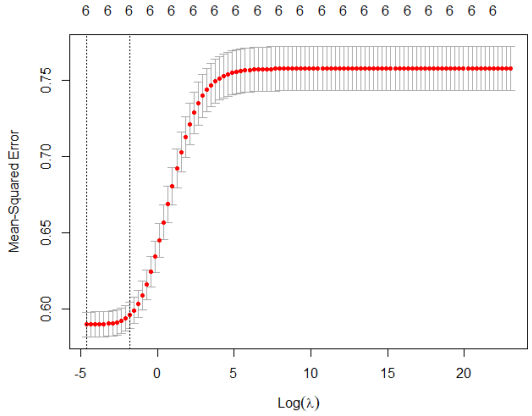


Figure 17: Best lambda of ridge regression

### *3.3. Classification - KNN*

Our first classification approach used K-nearest neighbors. Since the attributes have different value ranges, we first standardize all the attributes, except the outcome variable quality, to interpret the data. Then we split the data sets into training set and test set with $p = 0.8$.

Then, we try to find an appropriate value of K. The root mean squared error(RMSE) corresponds to the square root of the average difference between the observed known outcome values and the predicted values. A lower RMSE will give us more accurate model regarding our training data sets.

We randomly select 6 different values of $k = 3, 5, 7, 9, 12, 20$ and plot how the RMSE of the models change with the value of k. From the plot, we notice that the curve bends at 7. It means that we can choose more than 7 clusters but our RMSE will not dramatically decrease.

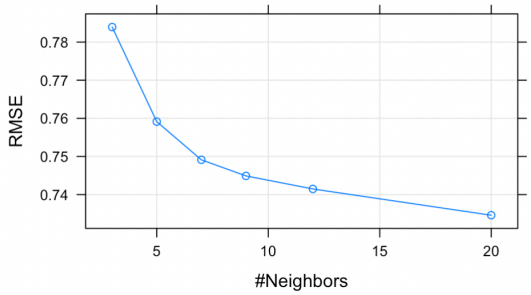


Figure 18: Best number of neighbors regarding to the RMSE

With k=7, we built an KNN prediction model starting from 2 predictors to all 11 predictors regarding to our training data set. From the plots, we can find some overlaps, which represent a not very high accuracy of the KNN model. By comparing the prediction results of the KNN models to the results in the test data set, the overall accu-

racy of this model is 0.4974. Among all predictions, 50% of the models got right.

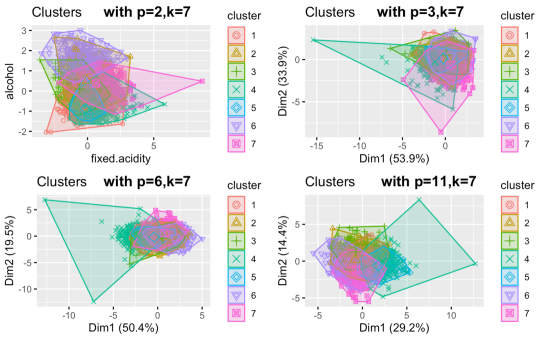


Figure 19: KNN models for different values of p

### *3.4. Classification - Decision Tree*

The next classification approach is Decision Tree. We also randomly split the original data sets into the training data and the test data. In both data set, the amount of each level is about the same. The level of wine quality ranged from 3 to 9 and mainly distributed in range 5 to 7.

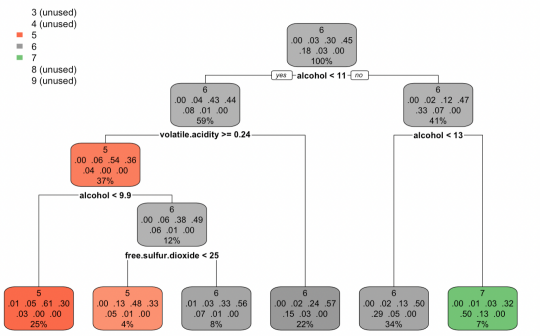


Figure 20: Decision tree

We build a decision tree model from our training data. The accuracy is calculated by comparing its resulting quality level from the input variables to the actual quality level from the test data sets. The accuracy for this decision tree model is 0.5189, which is about 52%.

### *3.5. Classification - SVM*

Another classification approach we use is Support Vector Machine. To do a classification, we first converted the numeric values into categorical values.

To ensure better performance of testing set and avoid overfitting, we used 5-fold CV while training the model.For parameter tuning, we attempted with $Cost = 4, 6, 8, 10, 12$ and $sigma = 0.05, 0.1, 0.5, 1, 2$. Out of the 25 models, the one with $cost = 6$ and $sigma = 1$ has best training set performance with accuracy of 0.6117681 and the one with $cost = 10$ and $sigma = 2$ has best testing set performance with accuracy 0.6307075.

## 4. Discussion and Conclusion

### *4.1. Regression Model Selection*

Compared the coefficient values in lasso and ridge regression models, the results are close to each other. As a result, we need to compare the test MSE values for ridge and lasso. Lasso regression model have a smaller test MSE value. So we choose ridge regression model.

### *4.2. Classification Model Selection*

Our motivation for doing classification: although the result, quality, are numeric values, it is discrete and only has 7 different levels. Thus we consider classification is possible.

A downside for classification is that the model would treat different classes regardless of how close their quality are. The difference between level 3 and level 9 would be considered the same as the difference between level 3 and level 4, which is not applicable to real life scenario.

In practice, we found that the SVM model we build performs slightly better than the decision tree model, one at 63%, the other at 52%.