**ST563 Final Project Report**

**Group 3**

**Introduction**

The purpose of this report is to explore various statistical learning methods predicting red wine quality rating and evaluate which method is most accurate (low variance and low bias). A data set derived from the UCI Machine Learning Repository’ Red Wine Data was used. The methods examined in this report include Multiple Linear and Logistic Regression, Linear and Quadratic Discriminant Analysis (LDA and QDA), K-nearest Neighbors (KNN), Decision Trees, and Support Vector Machines (SVM).

**Wine Data**

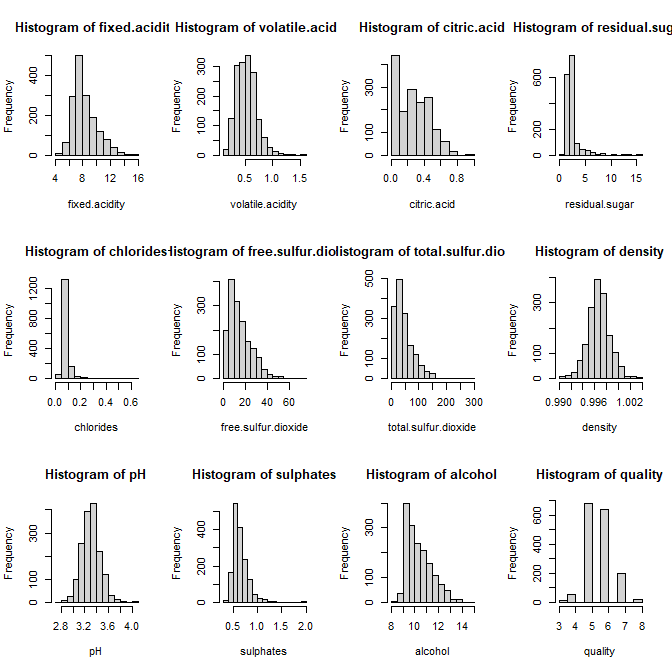
The data set has 1,599 different red wine varieties, and 11 chemical features of wines including Fixed Acidity, Volatile Acidity, Citric Acid, Residual Sugar, Chlorides, Free Sulfur Dioxide, Total Sulfur Dioxide, Density, pH, Sulphates, and Alcohol. These chemical properties were used to predict our response variable, Quality, which is in ordinal scale ranging from 1 (worst) to 10 (best) in its raw form that represents quality ratings of each wine. Table 1 presents all variables (11 predictors and 1 response) and the descriptive statistics including minimum, maximum, mean and standard deviation of each. Figure 1 presents histogram of variables in the data, majority variables are right skewed except for pH and quality, which are normally distributed.

**Table 1. Descriptive Statistics of Red Wine Data**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Property | Mean | SD | Min | Max |
| Fixed acidity | 8.32 | 1.74 | 4.6 | 15.9 |
| Volatile acidity | 0.53 | 0.18 | 0.12 | 1.58 |
| Citric acid | 0.27 | 0.2 | 0 | 1 |
| Residual sugar | 2.54 | 1.41 | 0.9 | 15.5 |
| Chlorides | 0.09 | 0.05 | 0.01 | 0.61 |
| Free sulfur dioxide | 15.87 | 10.46 | 1 | 72 |
| Total sulfur dioxide | 46.47 | 32.89 | 6 | 289 |
| Density | 1 | 0 | 0.99 | 1 |
| pH | 3.31 | 0.15 | 2.74 | 4.01 |
| Sulphates | 0.66 | 0.17 | 0.33 | 2 |
| Alcohol | 10.42 | 1.07 | 8.4 | 14.9 |
| Quality | 5.64 | 0.81 | 3 | 8 |

Note: All numbers are rounding up to two decimals.

**Figure 1. Histogram of Red Wine Data**



Reponses variable (quality) categorization…

**Methods and Results**

The raw data were split randomly into a training and test set. The training data used 80% of the original set, whereas the test data used the remaining 20%. A seed of 20 was used on all analyses in this report for replication and consistency. For classification methods, the response variable quality was split into “low” and “high” categories. Wines ranked with a quality of five or lower were categorized as “low” whereas wines with a quality of six or higher categorized as “high.” A three-category scheme was considered, but was discarded due to the lack of spread in the distribution of quality. A vast majority of observations were classified as “medium,” leaving the “high” and “low” categories with very low counts.

*Multiple Linear Regression*

When regressing all of the predictors in the dataset on quality with multiple linear regression, it was found that volatile acidity, chlorides, free sulfur dioxide, total sulfur dioxide, sulphates, and alcohol were all significant predictors.

*Multiple Logistic Regression*

*Linear/Quadratic Discriminant Analysis*

*K-nearest Neighbors*

*Decision Trees*

*Support Vector Machines*

**Table 2. Models and Test Misclassification Error Rate**

|  |  |  |  |
| --- | --- | --- | --- |
| Test Misclassification Error Rate | | | |
| Models |  |  |  |
| Linear Regression | 0.356 |  |  |
| Logistic Regression |  |  |  |
| LDA |  |  |  |
| QDA |  |  |  |
| KNN |  |  |  |
| Decision Tree |  |  |  |
| Support Vector Machines |  |  |  |

**Discussion**