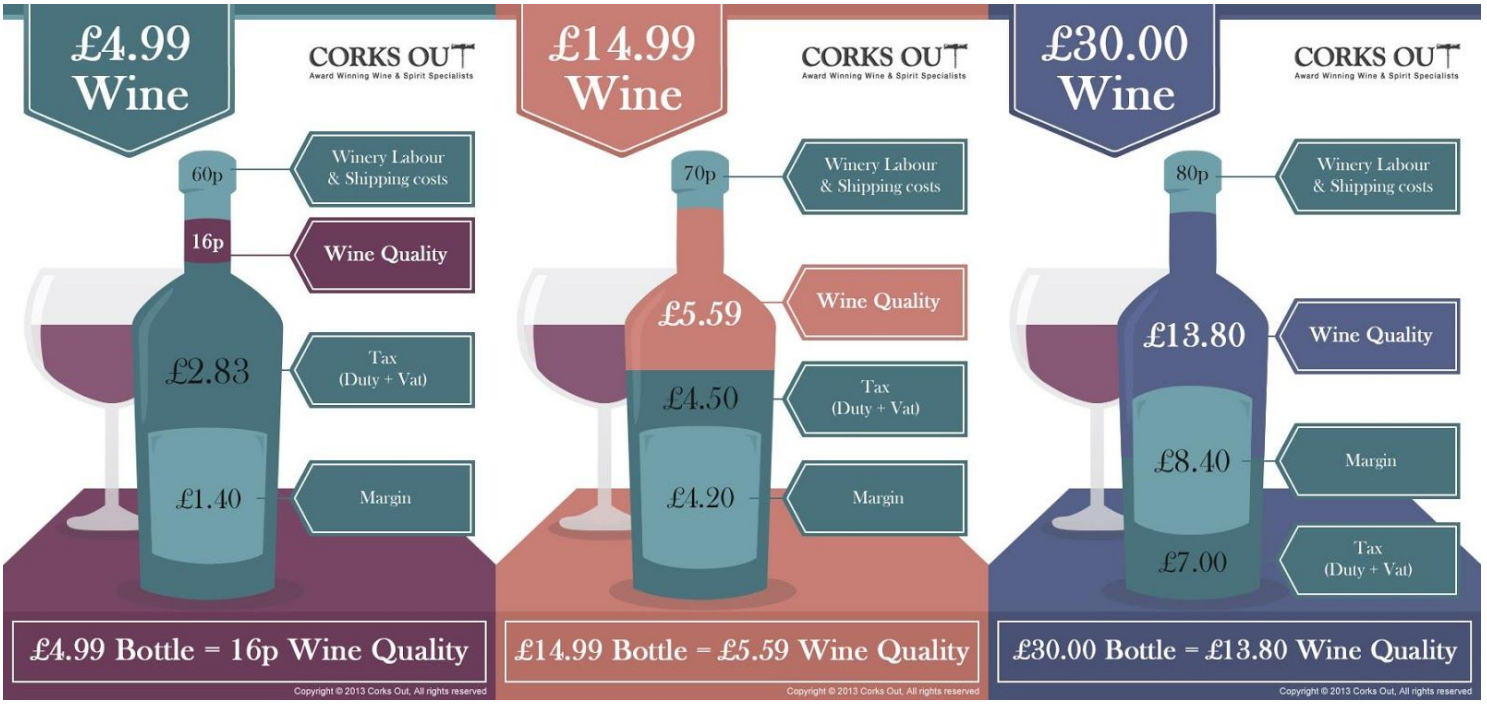
What Makes a Wine “Good”?

Classification Project

Xiaolu Yu
Aileen Fu



1. Introduction

Do you find it complicated when choosing wine? Have you wondered what are the chemical qualities that make a wine "good"? If your answer is yes, then you are in the right place! This paper aims to give a guidance on wine selection by looking at some wine physicochemical qualities and classifying the wine into three classes (high; medium; low). The accuracy for the wine classification on this dataset is 88.69%.

2. Summarize Data

2.1 Data Description

The dataset of *Red Wine Quality* comes from Kaggle[1]. There are 1599 observations and 12 variables. 11 variables are independent variables (numeric) describing the physicochemical levels of the red wine and 1 variable is dependent variable (factor) named as the quality of the red wine. In the original dataset, the wine quality is ranged from 3 to 8. For simplicity, we reclassified level 3 and 4 as low quality, 5 and 6 as medium quality and 7 and 8 as high quality and recreated a variable *quality.class* to denote it. The meaning of 11 independent variables is shown as below:

Table 1: Description of Independent Variables

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	found in small quantities, citric acid can add 'freshness' and flavor to wines
residual.sugar	the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet
chlorides	the amount of salt in the wine
free.sulfur.dioxide	the free form of SO2 exists in equilibrium between molecular SO2 and bisulfite ion; it prevents microbial growth and the oxidation of wine
total.sulfur.dioxide	amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine
density	the density of water is close to that of water depending on the percent alcohol and sugar content
pH	describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale
sulphates	a wine additive which can contribute to sulfur dioxide gas (S02) levels, with acts as an antimicrobial and antioxidant
alcohol	the percent alcohol content of the wine

[1] https://www.kaggle.com/uciml/red-wine-quality-cortez-et-al-2009
P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.

2.2 Create a training and validation dataset

We split out the original dataset by selecting 20% of the data for validation and the remaining 80% of data for training. The training dataset has 1281 observations and validation dataset has 318 observation.

2.3 Class distribution

The distribution of wine quality in training data is shown in Table 2. We can find the distribution is not even. The datapoint limitation in low-quality wine may cause inaccuracy classification.

Table 2: Distribution of the Red Wine Quality

Statistic	Levels	N	%
Quality.class	Low	51	3.98%
	Medium	1056	82.44%
	High	174	13.58%

2.4 Dataset Visualization

We made scatterplots to see the interaction of each pair of attributions and made boxplots, density plots to see how each attribution is distributed in each class. We also made radar chart to see the differences of each class, which is more intuitive. The charts are shown below. (Scatter plots is not shown here but Appendix because it's not precise due to the large number of independent variables)

From the radar chart, we can find that the high-quality red wine is likely to have high level of fixed acidity, alcohol, sulphates, citric acid and low level of volatile acidity, PH, density and chlorides. It seems to be logical. High fixed acidity is associated with low PH. Relatively more alcohol means the wine is more likely to carry you to a relaxed state of intoxication. High level of citric acid will add more 'freshness' and flavor to wines. More sulphates will increase wine's antioxidant and antibacterial properties. Low volatile acidity means low amount in acetic acid, which may lead to an unpleasant, vinegar taste if it's high. Less chlorides means low amount in salt, which may lead to an salty taste if it's high.

Chart 1: The Rader Chart of Three Classes

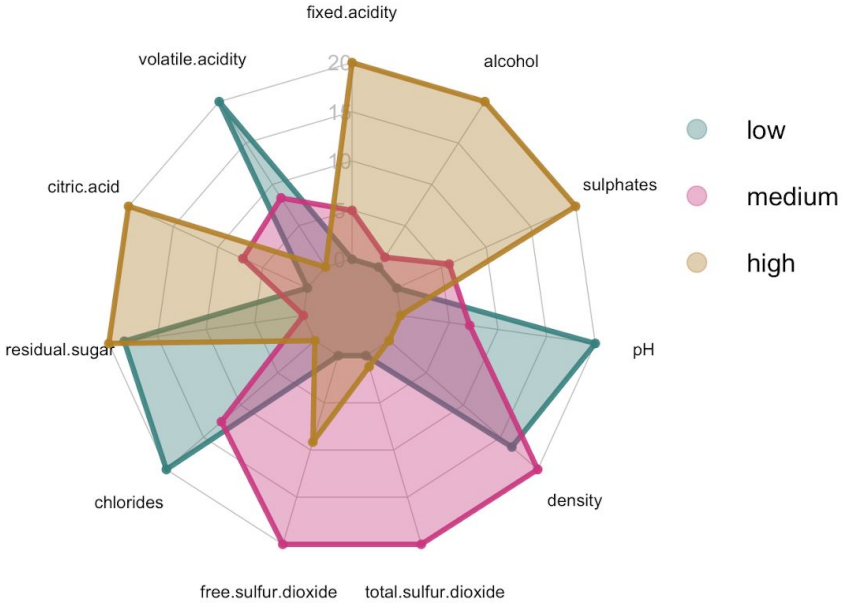


Chart 2: Box and Whisker Plots for Each Attribute

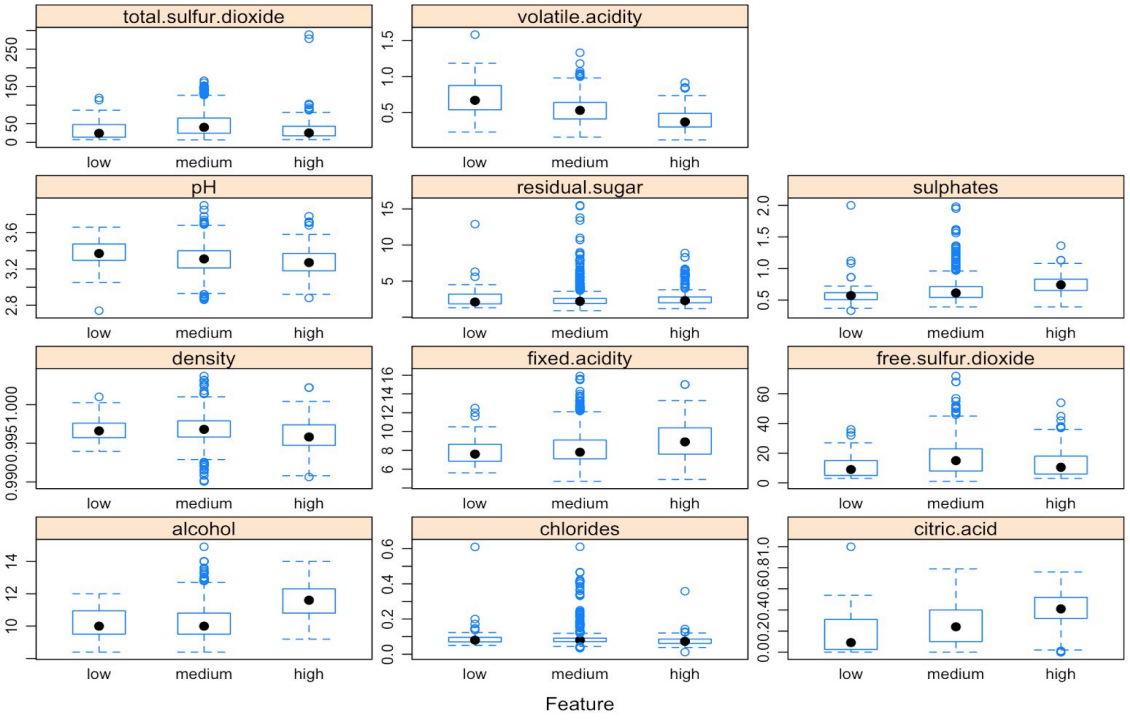
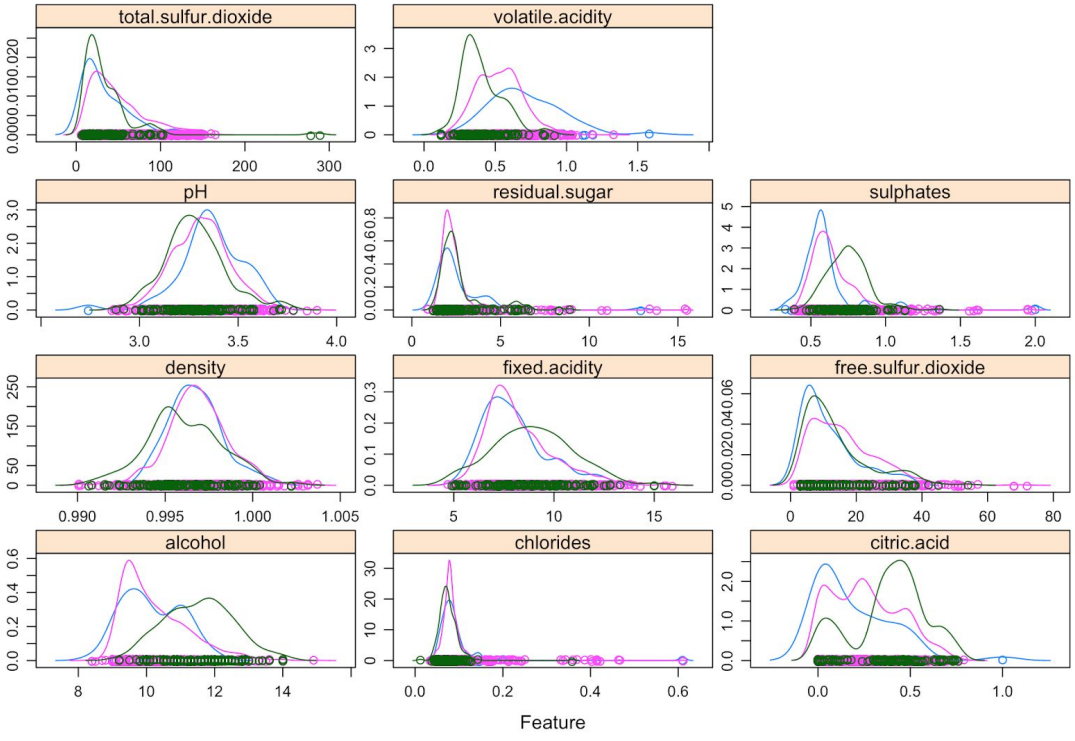


Chart 3: Density Plots for Each Attribute



3. Analysis

3.1 Transform and test harness

We used scale function to make each attribution standardized and used Box-Cox transformation to stabilize variance and make the data more normal-distribution-like. For test harness, we used 10-fold cross validation and repeated the process 3 times to make our models more accurate.

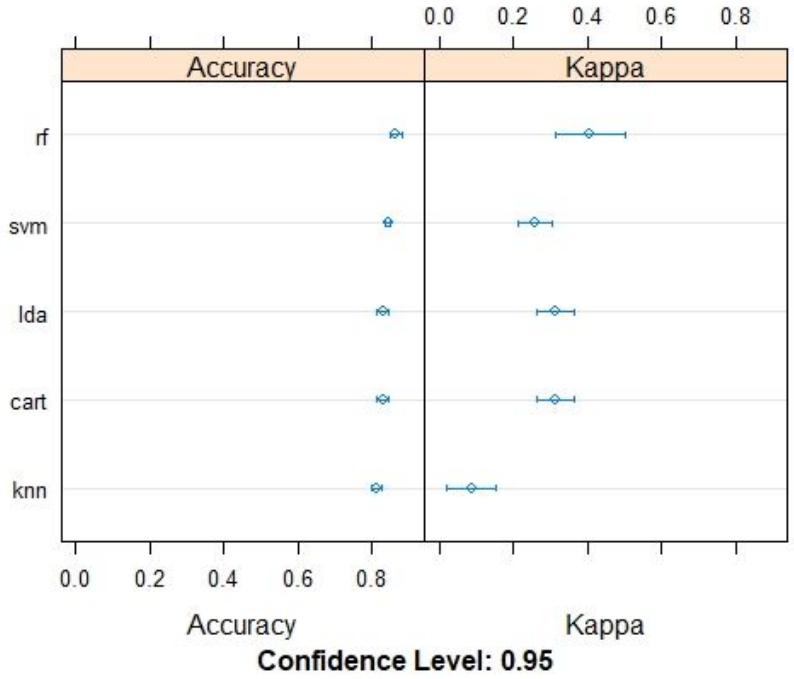
3.2 Build models

There are many machine learning algorithms which we can use to train and test our dataset. We chose five of them: linear (LDA), nonlinear (CART, kNN) and advanced (SVM, RF). We also made random number seed as 7 before each run to ensure that each algorithm would use the same data splits, which ensured the comparison of the results of these five algorithms.

3.3 Select the best model

Chart 3 shows the comparison of these five methods. We can find that the optimal model turns out to be the Random Forest Model[2] using mtry[3]=2 with 85.78% accuracy and 0.45 Kappa[4] value.

Chart 4: Comparison of five models



4. Results

Use the optimal classification model to make a prediction on the testing data, and there is 89.94% accuracy to the original data. In other words, this classification method brings around 89.94% accuracy in guiding you to classify the wine quality. Chart 5 shows the prediction result.

[2] RandomForest implements Breiman's random forest algorithm (based on Breiman and Cutler's original Fortran code) for classification and regression.
https://cran.r-project.org/web/packages/randomForest/randomForest.pdf

[3] Number of variables available for splitting at each tree node.
http://code.env.duke.edu/projects/mget/export/HEAD/MGET/Trunk/PythonPackage/dist/TracOnlineDocumentation/Documentation/ArcGISReference/RandomForestModel.FitToArcGISTable.html

[4]The Kappa statistic (or value) is a metric that compares an Observed Accuracy with an Expected Accuracy (random chance)
https://stats.stackexchange.com/questions/82162/cohens-kappa-in-plain-english

Chart 5: The Prediction Result of Random Forest Model

```
Overall Statistics

               Accuracy : 0.8994
                 95% CI : (0.8609, 0.9301)
    No Information Rate : 0.827
    P-Value [Acc > NIR] : 0.0001985

                  Kappa : 0.596
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: low Class: medium Class: high
Sensitivity             0.00000         0.9772      0.67442
Specificity             1.00000         0.5273      0.97818
Pos Pred Value              NaN         0.9081      0.82857
Neg Pred Value          0.96226         0.8286      0.95053
Prevalence              0.03774         0.8270      0.13522
Detection Rate          0.00000         0.8082      0.09119
Detection Prevalence    0.00000         0.8899      0.11006
Balanced Accuracy       0.50000         0.7522      0.82630
```

5. Limits and Improvements

5.1 Samples Limitation

The dataset includes 1599 samples. However, the quality percentage for each classes is not even in here. The medium quality class accounts for a major percentage with 82.44% and 3.98% for low class and 13.58% for high class. Data Insufficiency may lead to a less accurate result. The test result will be more representative and applicable with more samples. In addition, a larger samples may come with a higher accuracy in testing classifiers.

5.2 Category Classification

Here, we simply classify the wine variety into three categories, but in reality, a niche category may be more useful and doable. Nevertheless, since we have a limited numbers of samples, more categories will lower the accuracy when testing the classification. Hence, a niche category might be applicable when there is a larger dataset.

5.3 Value Number

Some chemicals' content in a wine are so low that it may not differ a lot among different types of wine.

5.4 Variables Limitation

This dataset focuses on chemical quantities and doesn't include measures like producing region and production year, which is a general way we use in selecting wine preliminary. Missing major variables can affect the accuracy and application of result.

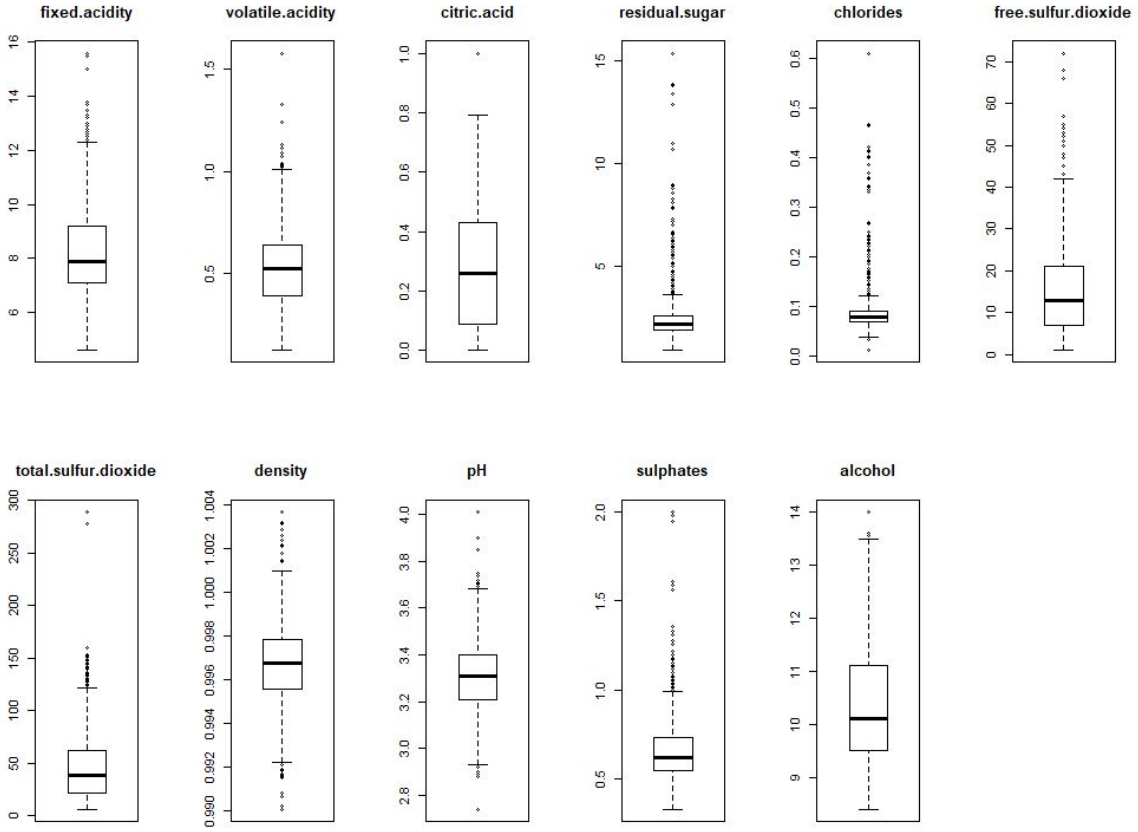
Appendix:

1. Table for Statistical Description of Original Dataset's Variables

Statistics	N	Mean	St. Dev.	Min	Max
fixed.acidity	1,599	8.320	1.741	4.600	15.900
volatile.acidity	1,599	0.528	0.179	0.120	1.580
citric.acid	1,599	0.271	0.195	0	1
residual.sugar	1,599	2.539	1.410	0.900	15.500
chlorides	1,599	0.087	0.047	0.012	0.611
free.sulfur.dioxide	1,599	15.875	10.460	1	72
total.sulfur.dioxide	1,599	46.468	32.895	6	289
density	1,599	0.997	0.002	0.990	1.004
pH	1,599	3.311	0.154	2.740	4.010
sulphates	1,599	0.658	0.170	0.330	2.000
alcohol	1,599	10.423	1.066	8.400	14.900
quality	1,599	5.636	0.808	3	8

(Source: Stargazer)

2. Boxplots for Numeric Variables



3. Scatterplot Matrix