Red & White Wine: Predicting Quality

Max Schemitsch, DATA 440L 200

The Dataset

- Obtained from Kaggle
- Originally available from
  UCI Machine Learning Repository
- Paulo Cortez, University of Minho (Portugal)

About the Data

- Contains information about both red & white wines
- Portuguese “Vinho Verde” wine
- Does not include sensory characteristics
- Grape type, wine brand, price, etc
- Each wine is given a quality score, from 0 to 10
- However, the no wine was given scores 0, 1, 2, or 10

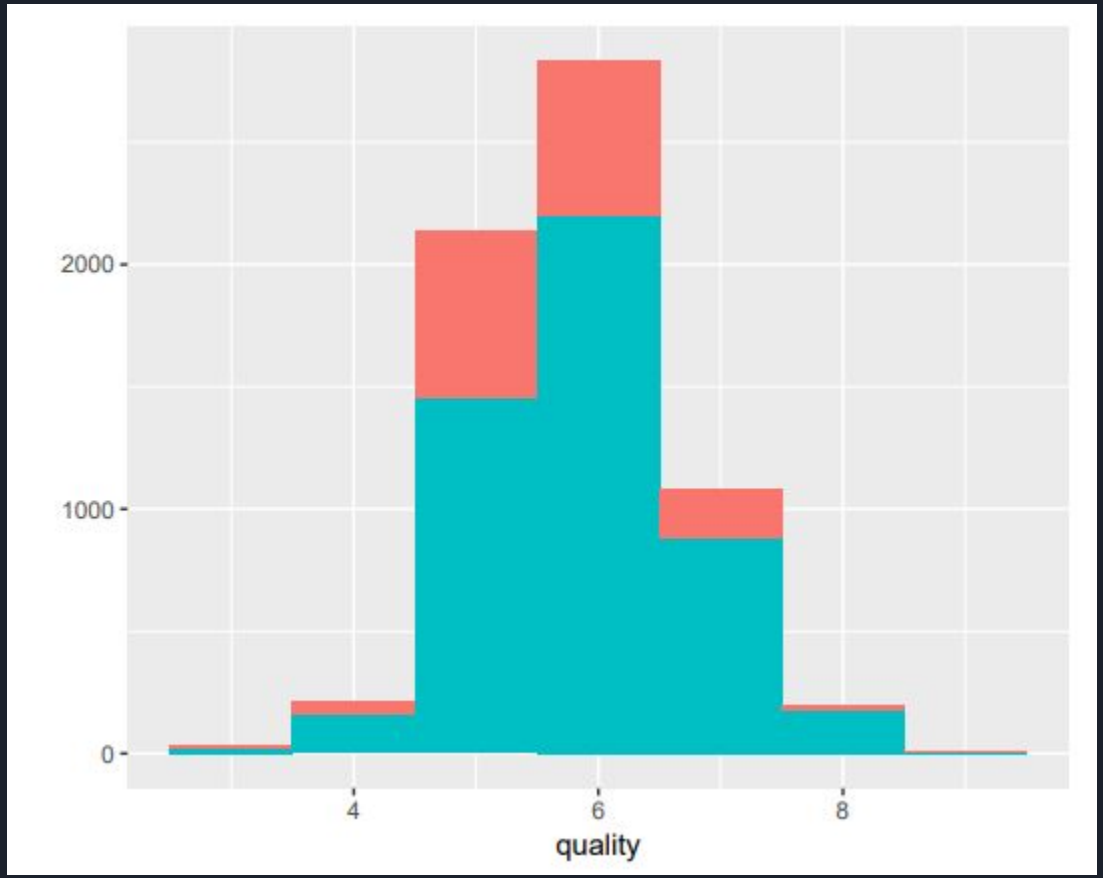
Data Attributes

- Fixed Acidity: Most acids involved with wine (do not evaporate readily)
- Volatile Acidity: Acetic Acid, which can lead to unpleasant vinegar taste
- Citric Acid: Usually in small quantities, can add freshness to wine
- Residual Sugar: Amount of sugar remaining after fermentation
- Chlorides: Amount of salt in the wine

Data Attributes (cont)

- Free Sulfur Dioxide: Free form SO_2 , prevents microbial growth and oxidation
- Total Sulfur Dioxide: Both free form SO_2 and bound SO_2
- Density: Can be close to water depending on alcohol and sugar content
- pH: How acidic or basic the wine is from 0 to 14 (usually 3-4)
- Sulphates: Wine additive that contributes to SO_2 levels (antioxidant)
- Alcohol: Percent alcohol content of the wine

Normality

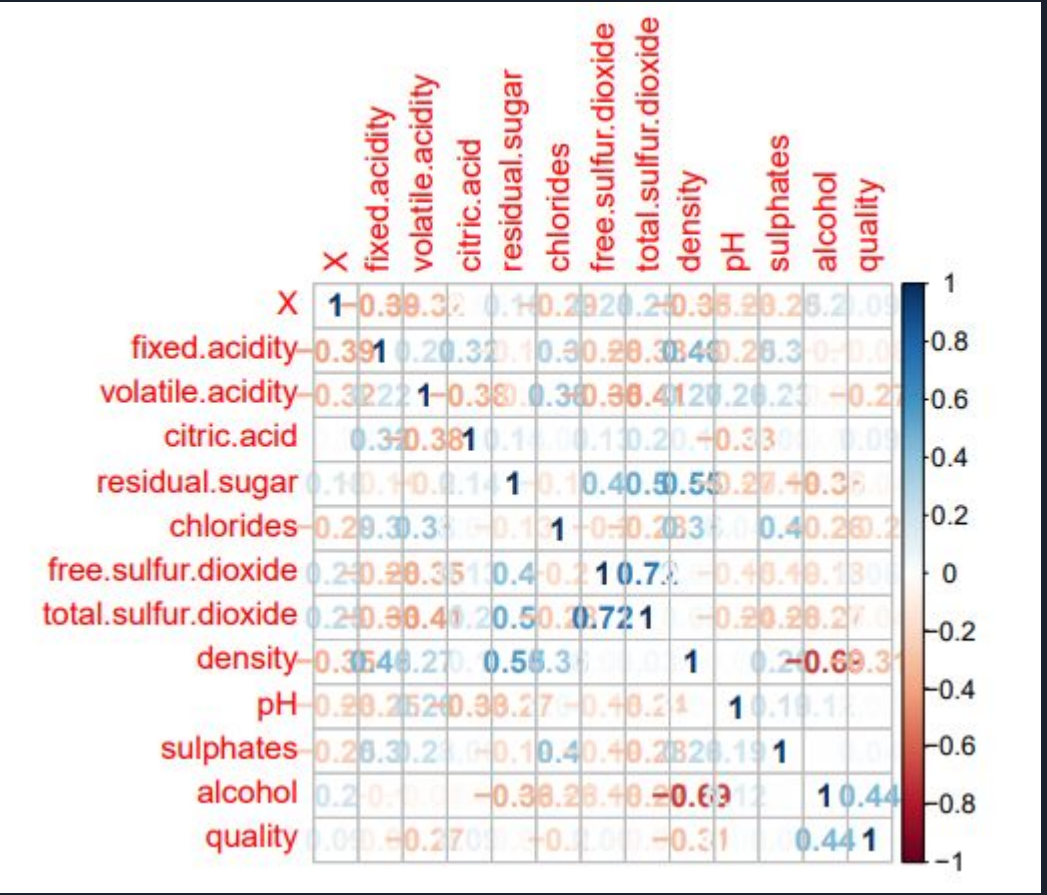


Data Tool of Choice

Choice of Methods

- Multiple Linear Regression
  - showcase weaknesses (correlation problems)
- k-Nearest Neighbors
  - strengths in categorizing
- Random Forest
  - decision trees

Correlation Matrix



Prominent Correlations for Quality

- Alcohol (0.44)
- Density (-0.31)
- Volatile Acidity (-0.27)

Linear Regression

- Model 1: All Variables
  - Adjusted R-square Value = 0.2909
- Model 2: Alcohol
  - Adjusted R-square Value = 0.1973
- Model 3: Alcohol, Volatile Acidity, Density
  - Adjusted R-square Value = 0.267

- Best model with 4 variables?

- Model 4: Alcohol, Volatile Acidity, Density, Sulphates
  - Adjusted R-square Value = 0.27

- Cross Validation also confirms Model 4 as the best

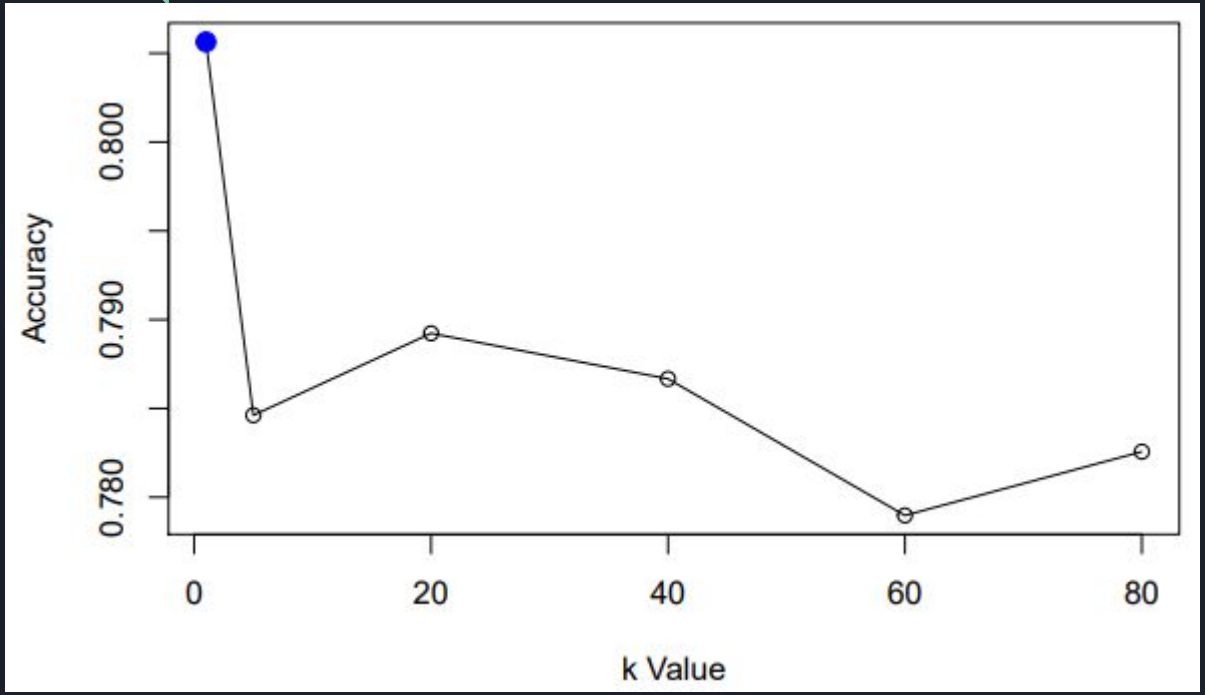
k-Nearest Neighbors

- Split data into groups:
  - 3,4 = “Poor”
  - 5, 6 = “Average”
  - 7, 8, 9 = “Good”

- Use all variables

- K-values: 1, 5, 20, 40, 60, 80

Graph of Accuracies




```
mean(knn1==y_test)
```

```
## [1] 0.805641
```

```
table(knn1, y_test)
```

```
##          y_test
## knn1      Average Good Poor
##   Average    1319  150   63
##   Good        117  235    1
##   Poor         46    2   17
```

Conclusions? (What about Random Forest?)

- It’s my last thing to do
- Might show that decision trees could be useful for something like this
- Otherwise, KNN looks much better in terms of categorizing
- Wine is great
- Pablo please don’t be mad at me for using R

Image References

https://upload.wikimedia.org/wikipedia/commons/1/11/Alvarinho_Vinho_Verde_Quinta_de_Carape%C3%A7os.jpg

https://cdn-images-1.medium.com/max/1200/0*ftOal7fKVCNtJr4N.png