Do You Know The Quality of Your Wine?

Wine Not?

Ashley Ngo, Inara Ramji, Jose Torrealba, Kyra Pollak, Sharnpreet Gill

Data Set Overview

*The supervised data set was collected from a repo form UCI Irvin, who compiled it from **science direct***

*This data set is related to **red and white variants of Portuguese "Vinho Verde" wine**.*

*The goal is to **model wine quality** based on **physiochemical tests***

*The data set collected was a **supervised data set** as it includes labelled data inputs.*

*There are **12 attributes**, and the classes are not balanced (more ~Average~ wines than ~Excellent~ or ~Poor~ wines)*

*The **same testing and training models** were used when assessing each type of model using set.seed.*

Our Methodology

2 DATA SETS WERE COLLECTED, A WHITE WINE AND A RED WINE DATA SET

The Data set consists of many different features: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, alcohol and then the corresponding quality that an expert who tasted the wine has assigned to the specific white or red wine.

1. Data set was collected
2. Data set was stored in a CSV file – 1599 red wine entries, 4898 white wine entries
3. Digital Visualization was completed
4. Different models were tested
5. The testing and training model were chosen, and the model was evaluated using MSE
6. The results were interpreted, and a recommendation was constructed

VISUALIZING THE DATA

WHITE WINE SCATTER PLOTS
SCATTER PLOTS OF THE QUALITY VERSUS EVERY FEATURE WAS RUN

WHITE WINE CORRELATION MATRIX
A CORRELATION MATRIX WAS CONSTRUCTED TO SEE IF WE SHOULD REMOVE CORRELATED VARIABLES

RED WINE SCATTER PLOTS
SCATTER PLOTS OF THE QUALITY VERSUS EVERY FEATURE WAS RUN

RED WINE CORRELATION MATRIX
A CORRELATION MATRIX WAS CONSTRUCTED TO SEE IF WE SHOULD REMOVE CORRELATED VARIABLES

WHITE WINE SCATTER PLOTS

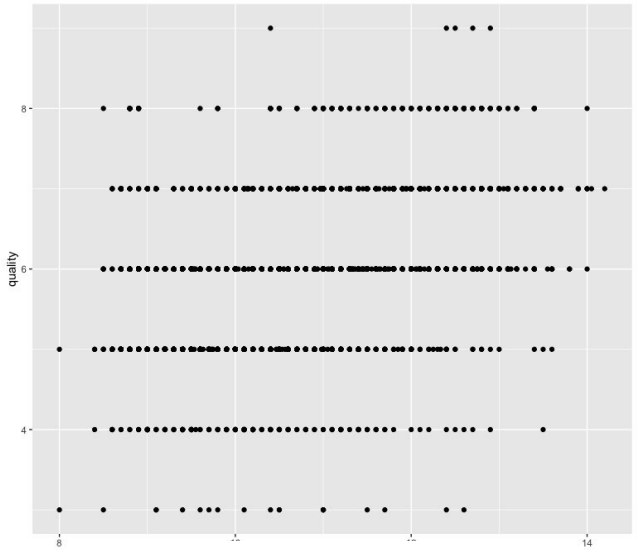
WHITE WINE SCATTER PLOTS
SCATTER PLOTS OF THE QUALITY VERSUS EVERY FEATURE WAS RUN

WHITE WINE CORRELATION MATRIX
A CORRELATION MATRIX WAS CONSTRUCTED TO SEE IF WE SHOULD REMOVE CORRELATED VARIABLES

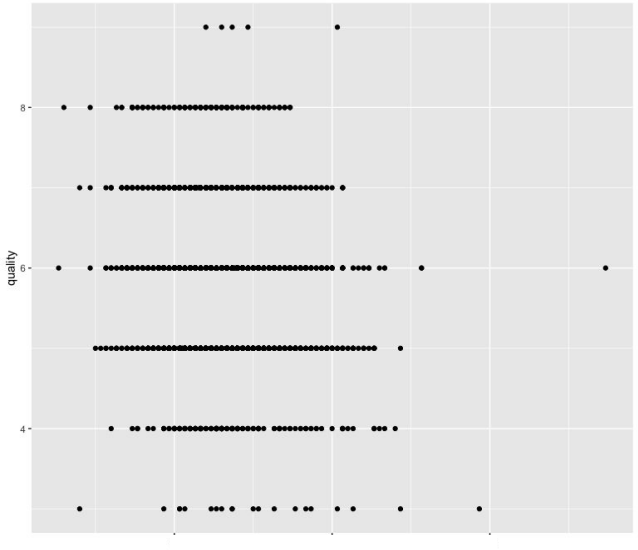
RED WINE SCATTER PLOTS
SCATTER PLOTS OF THE QUALITY VERSUS EVERY FEATURE WAS RUN

RED WINE CORRELATION MATRIX
A CORRELATION MATRIX WAS CONSTRUCTED TO SEE IF WE SHOULD REMOVE CORRELATED VARIABLES

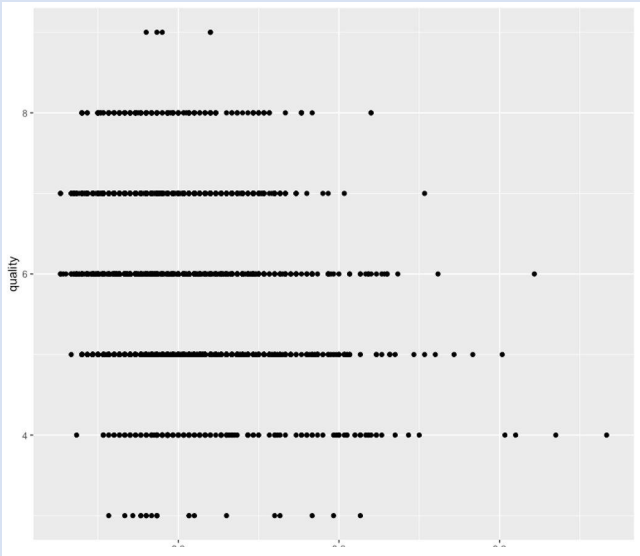
WHITE WINE QUALITY VS. INPUT FEATURES



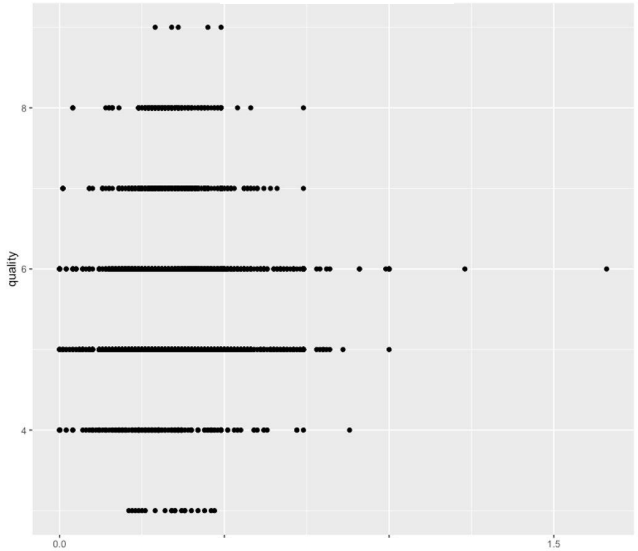
ALCOHOL



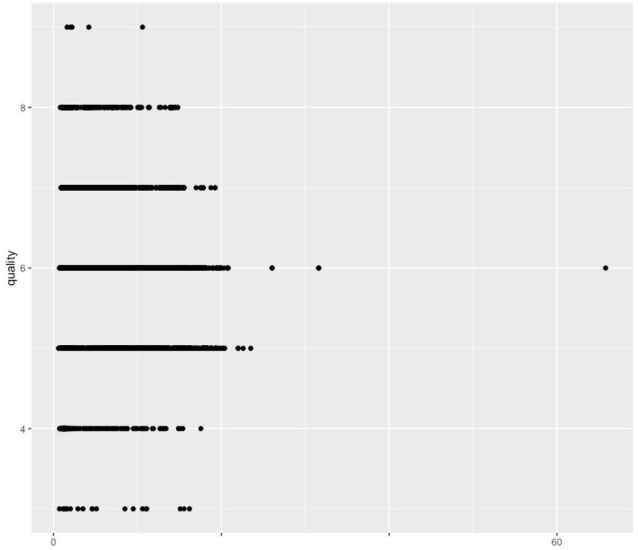
FIXED ACIDITY



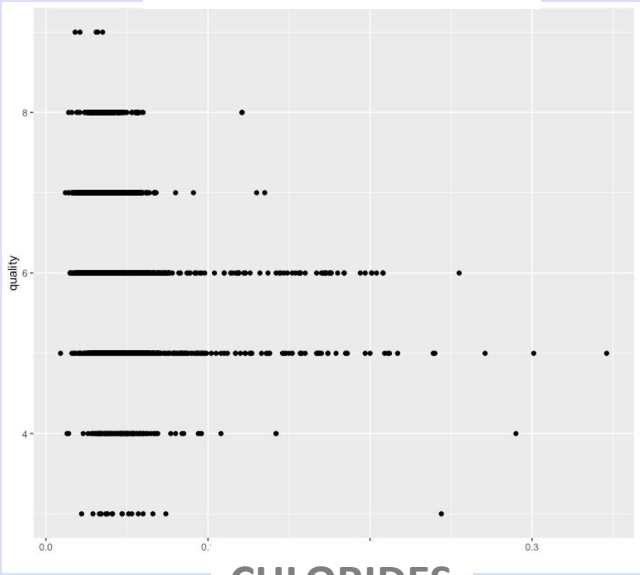
VOLATILE ACIDITY



CITRIC ACID

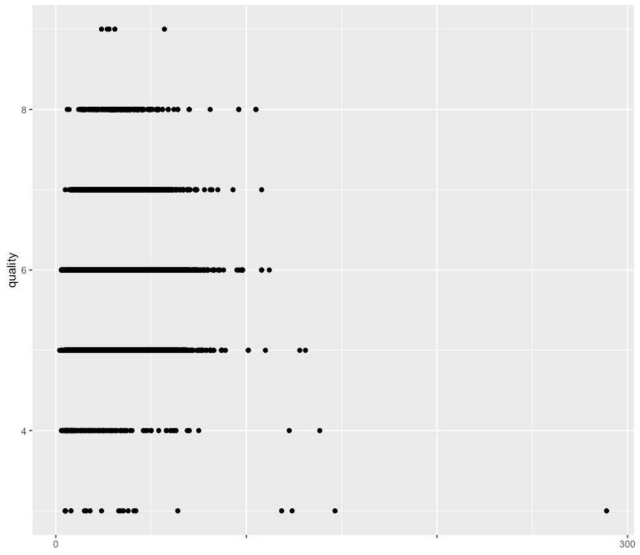


RESIDUAL SUGAR

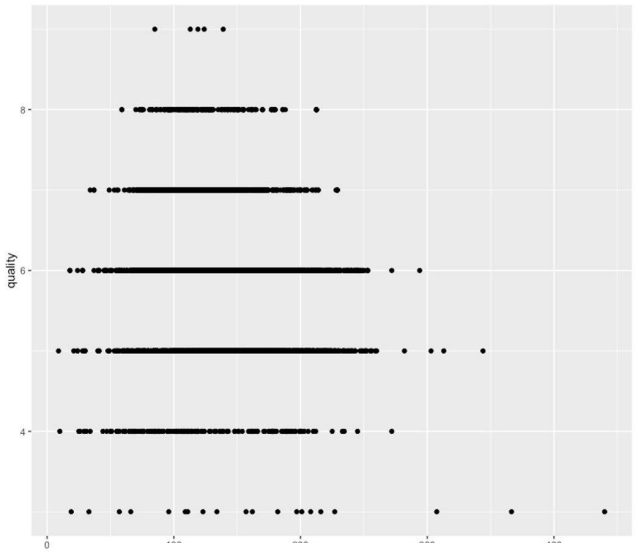


CHLORIDES

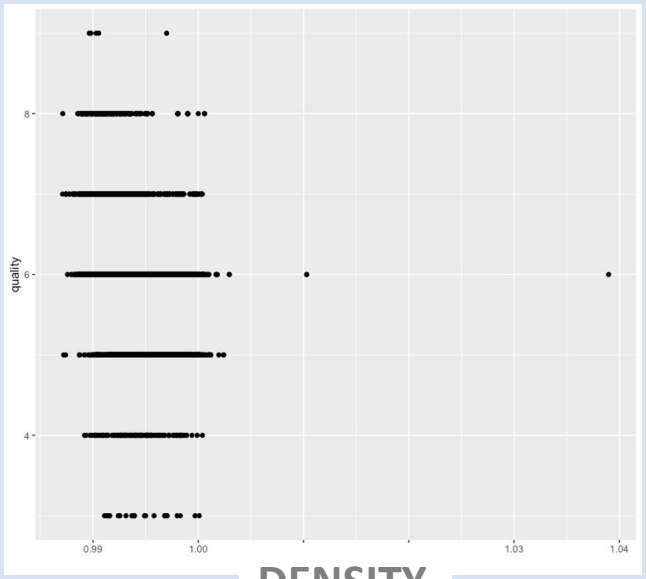
WHITE WINE QUALITY VS. INPUT FEATURES (CONT.)



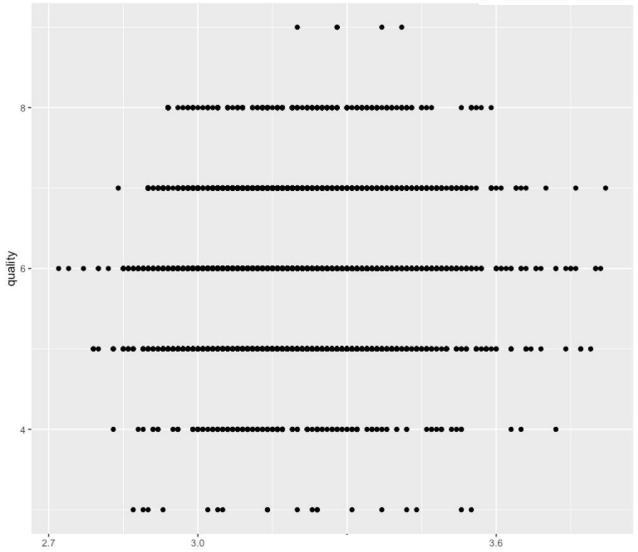
FREE SULFUR DIOXIDE



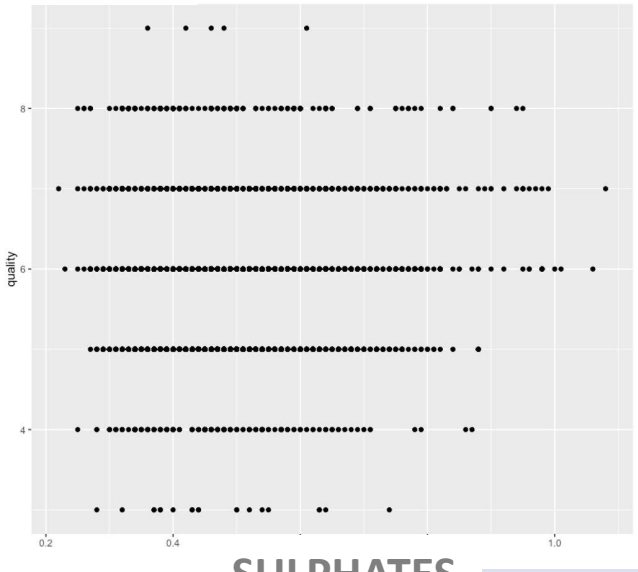
TOTAL SULFUR DIOXIDE



DENSITY



pH



SULPHATES

WHITE WINE CORRELATION MATRIX

WHITE WINE SCATTER PLOTS
SCATTER PLOTS OF THE QUALITY VERSUS EVERY FEATURE WAS RUN

WHITE WINE CORRELATION MATRIX
A CORRELATION MATRIX WAS CONSTRUCTED TO SEE IF WE SHOULD REMOVE CORRELATED VARIABLES

RED WINE SCATTER PLOTS
SCATTER PLOTS OF THE QUALITY VERSUS EVERY FEATURE WAS RUN

RED WINE CORRELATION MATRIX
A CORRELATION MATRIX WAS CONSTRUCTED TO SEE IF WE SHOULD REMOVE CORRELATED VARIABLES

THE CORRELATION: <0.4 OR >0.4

	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
fixed.acidity	0.09	0.27	-0.43	-0.02	-0.12	-0.11
volatile.acidity	0.09	0.03	-0.03	-0.04	0.07	-0.19
citric.acid	0.12	0.15	-0.16	0.06	-0.08	-0.01
residual.sugar	0.40	0.84	-0.19	-0.03	-0.45	-0.10
chlorides	0.20	0.26	-0.09	0.02	-0.36	-0.21
free.sulfur.dioxide	0.62	0.29	0.00	0.06	-0.25	0.01
total.sulfur.dioxide	1.00	0.53	0.00	0.13	-0.45	-0.17
density	0.53	1.00	-0.09	0.07	-0.78	-0.31
pH	0.00	-0.09	1.00	0.16	0.12	0.10
sulphates	0.13	0.07	0.16	1.00	-0.02	0.05
alcohol	-0.45	-0.78	0.12	-0.02	1.00	0.44
quality	-0.17	-0.31	0.10	0.05	0.44	1.00

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide
fixed.acidity	1.00	-0.02	0.29	0.09	0.02	-0.05
volatile.acidity	-0.02	1.00	-0.15	0.06	0.07	-0.10
citric.acid	0.29	-0.15	1.00	0.09	0.11	0.09
residual.sugar	0.09	0.06	0.09	1.00	0.09	0.30
chlorides	0.02	0.07	0.11	0.09	1.00	0.10
free.sulfur.dioxide	-0.05	-0.10	0.09	0.30	0.10	1.00
total.sulfur.dioxide	0.09	0.09	0.12	0.40	0.20	0.62
density	0.27	0.03	0.15	0.84	0.26	0.29
pH	-0.43	-0.03	-0.16	-0.19	-0.09	0.00
sulphates	-0.02	-0.04	0.06	-0.03	0.02	0.06
alcohol	-0.12	0.07	-0.08	-0.45	-0.36	-0.25
quality	-0.11	-0.19	-0.01	-0.10	-0.21	0.01

RED WINE SCATTER PLOTS

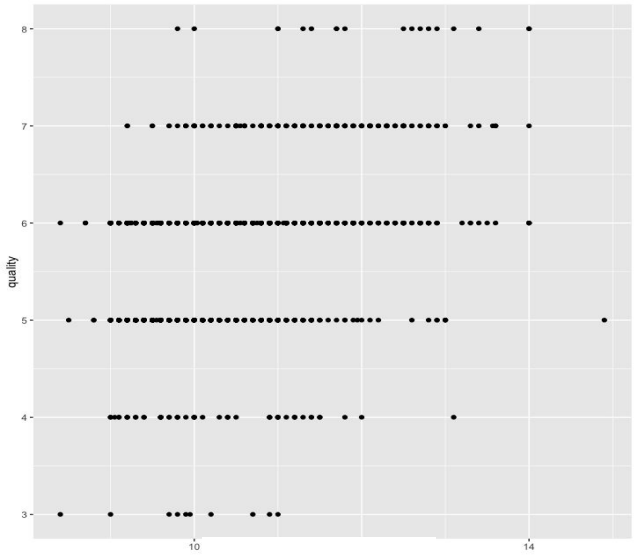
WHITE WINE SCATTER PLOTS
SCATTER PLOTS OF THE QUALITY VERSUS EVERY FEATURE WAS RUN

WHITE WINE CORRELATION MATRIX
A CORRELATION MATRIX WAS CONSTRUCTED TO SEE IF WE SHOULD REMOVE CORRELATED VARIABLES

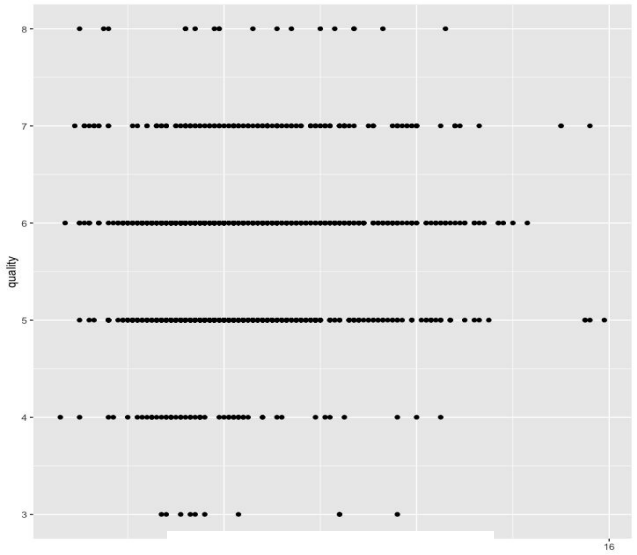
RED WINE SCATTER PLOTS
SCATTER PLOTS OF THE QUALITY VERSUS EVERY FEATURE WAS RUN

RED WINE CORRELATION MATRIX
A CORRELATION MATRIX WAS CONSTRUCTED TO SEE IF WE SHOULD REMOVE CORRELATED VARIABLES

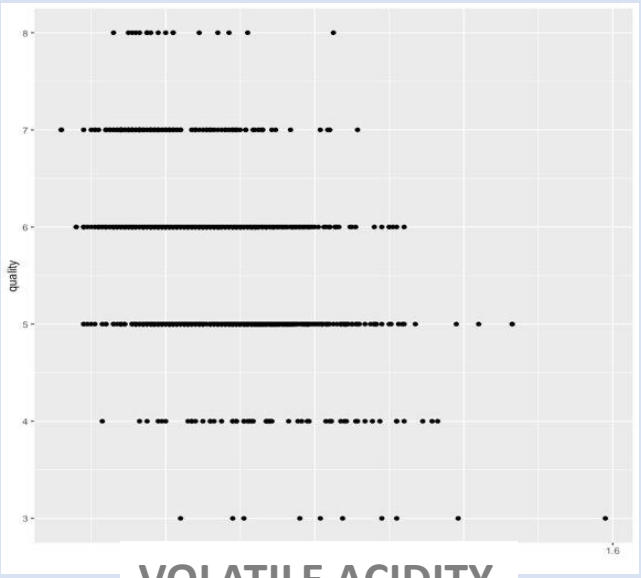
RED WINE QUALITY VS. INPUT FEATURES



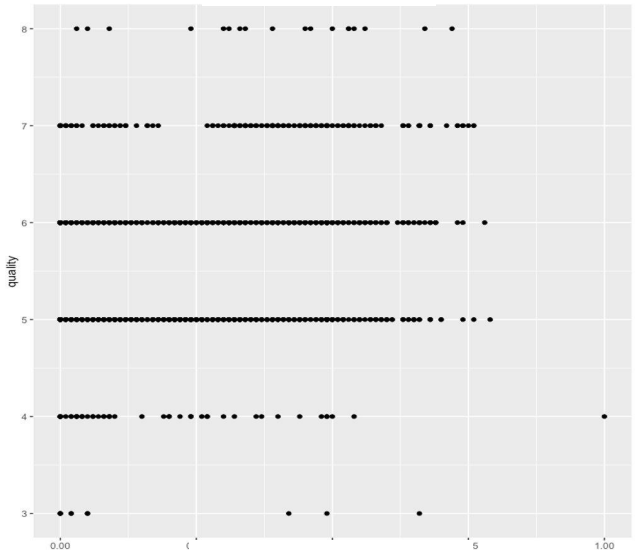
ALCOHOL



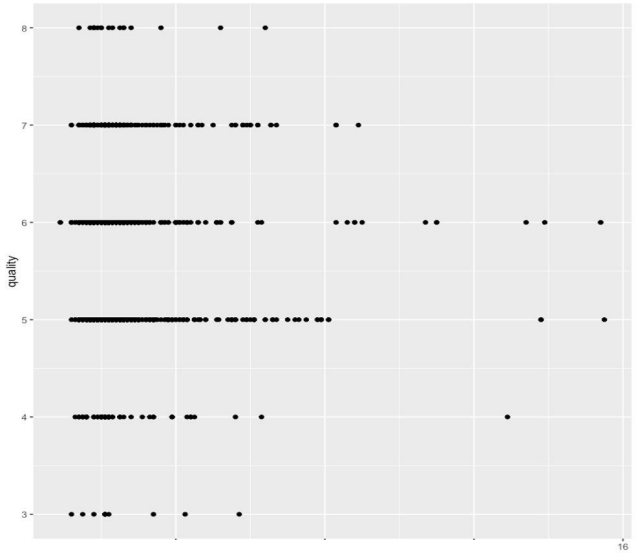
FIXED ACIDITY



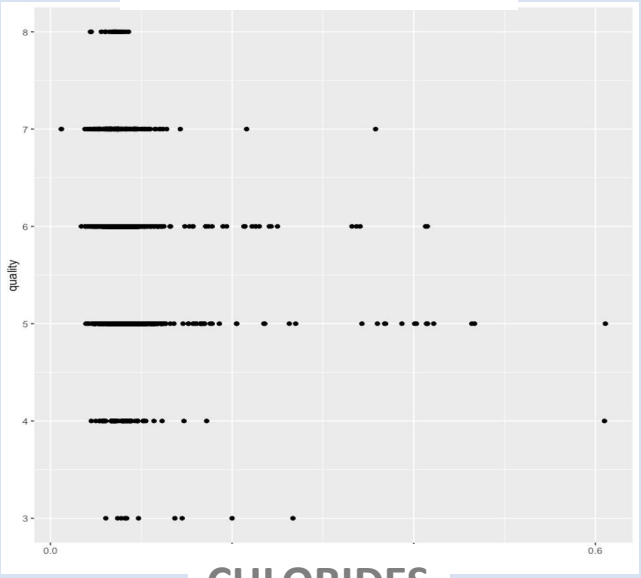
VOLATILE ACIDITY



CITRIC ACID

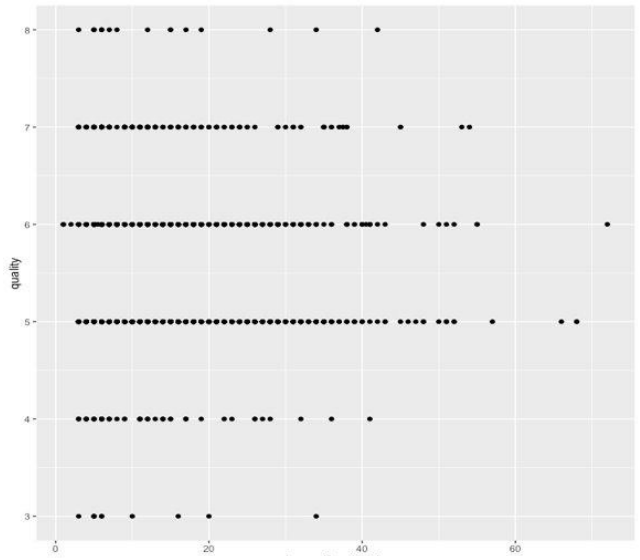


RESIDUAL SUGAR

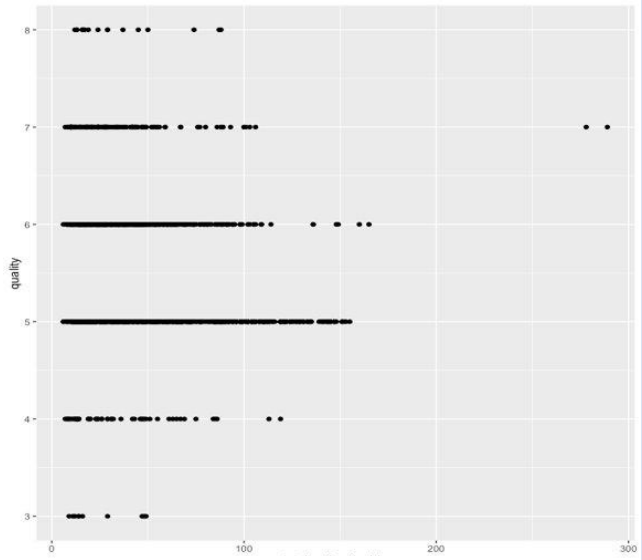


CHLORIDES

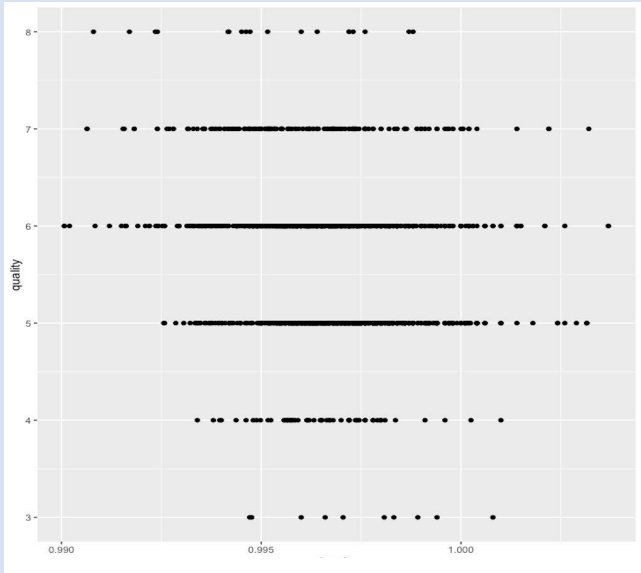
RED WINE QUALITY VS. INPUT FEATURES (CONT.)



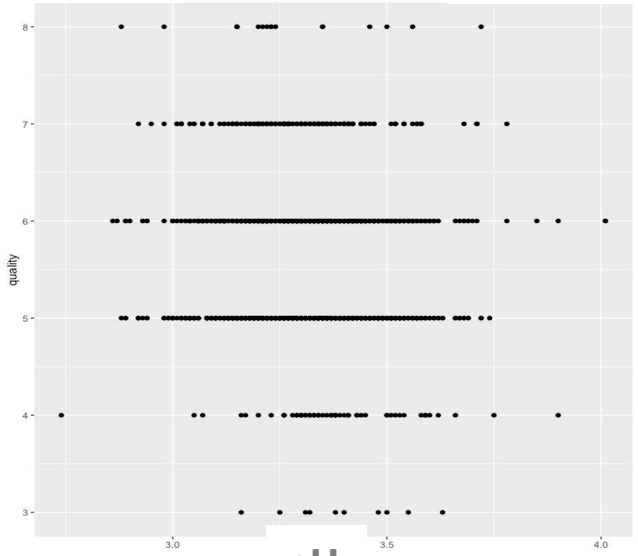
FREE SULFUR DIOXIDE



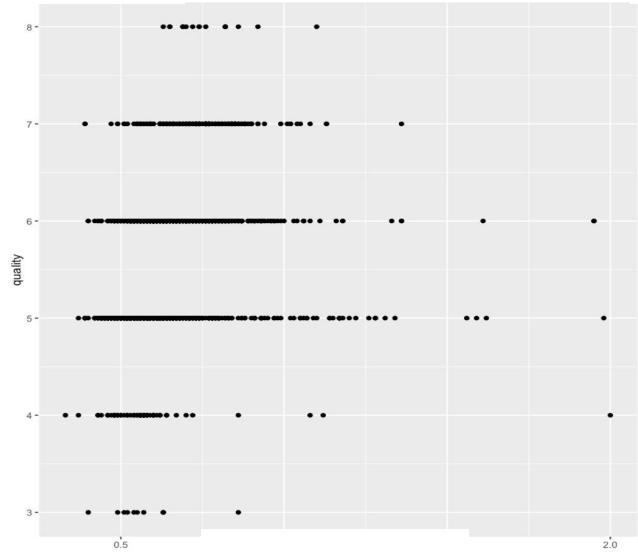
TOTAL SULFUR DIOXIDE



DENSITY



pH



SULPHATES

RED WINE CORRELATION MATRIX

WHITE WINE SCATTER PLOTS
SCATTER PLOTS OF THE QUALITY VERSUS EVERY FEATURE WAS RUN

WHITE WINE CORRELATION MATRIX
A CORRELATION MATRIX WAS CONSTRUCTED TO SEE IF WE SHOULD REMOVE CORRELATED VARIABLES

RED WINE SCATTER PLOTS
SCATTER PLOTS OF THE QUALITY VERSUS EVERY FEATURE WAS RUN

RED WINE CORRELATION MATRIX
A CORRELATION MATRIX WAS CONSTRUCTED TO SEE IF WE SHOULD REMOVE CORRELATED VARIABLES

THE CORRELATION: <0.4 OR >0.4

	total.sulfur.dioxide	density	pH	sulphates	alcohol	quality
fixed.acidity	-0.11	0.67	-0.68	0.18	-0.06	0.12
volatile.acidity	0.08	0.02	0.23	-0.26	-0.20	-0.39
citric.acid	0.04	0.36	-0.54	0.31	0.11	0.23
residual.sugar	0.20	0.36	-0.09	0.01	0.04	0.01
chlorides	0.05	0.20	-0.27	0.37	-0.22	-0.13
free.sulfur.dioxide	0.67	-0.02	0.07	0.05	-0.07	-0.05
total.sulfur.dioxide	1.00	0.07	-0.07	0.04	-0.21	-0.19
density	0.07	1.00	-0.34	0.15	-0.50	-0.17
pH	-0.07	-0.34	1.00	-0.20	0.21	-0.06
sulphates	0.04	0.15	-0.20	1.00	0.09	0.25
alcohol	-0.21	-0.50	0.21	0.09	1.00	0.48
quality	-0.19	-0.17	-0.06	0.25	0.48	1.00

	fixed.acidity	volatile.acidity	citric.acid	residual.sugar	chlorides	free.sulfur.dioxide
fixed.acidity	1.00	-0.26	0.67	0.11	0.09	-0.15
volatile.acidity	-0.26	1.00	-0.55	0.00	0.06	-0.01
citric.acid	0.67	-0.55	1.00	0.14	0.20	-0.06
residual.sugar	0.11	0.00	0.14	1.00	0.06	0.19
chlorides	0.09	0.06	0.20	0.06	1.00	0.01
free.sulfur.dioxide	-0.15	-0.01	-0.06	0.19	0.01	1.00
total.sulfur.dioxide	-0.11	0.08	0.04	0.20	0.05	0.67
density	0.67	0.02	0.36	0.36	0.20	-0.02
pH	-0.68	0.23	-0.54	-0.09	-0.27	0.07
sulphates	0.18	-0.26	0.31	0.01	0.37	0.05
alcohol	-0.06	-0.20	0.11	0.04	-0.22	-0.07
quality	0.12	-0.39	0.23	0.01	-0.13	-0.05

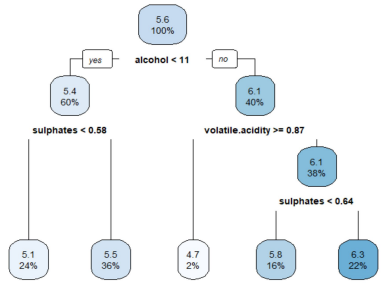
THE MODELS UNDER INVESTIGATION

ALL MSE VALUES FROM EACH MODEL WERE COMPARED

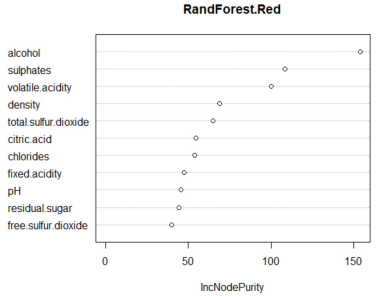
Linear Regression

```
Residuals:
     Min       1Q   Median       3Q      Max
-2.59831 -0.36813 -0.04217  0.45912  1.95424

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.6101506  0.4441768  10.379  < 2e-16 ***
volatile.acidity     -0.9993316  0.1110479  -8.999  < 2e-16 ***
chlorides            -2.1810949  0.4292109  -5.082 4.30e-07 ***
total.sulfur.dioxide -0.0025547  0.0005558  -4.596 4.73e-06 ***
pH                   -0.5269361  0.1304501  -4.039 5.68e-05 ***
sulphates             0.8958229  0.1218501   7.352 3.49e-13 ***
alcohol               0.2906447  0.0184610  15.744  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.644 on 1272 degrees of freedom
Multiple R-squared:  0.3622,    Adjusted R-squared:  0.3592
F-statistic: 120.4 on 6 and 1272 DF,  p-value: < 2.2e-16
```

Decision Tree



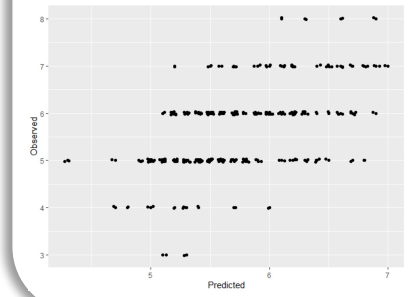
Random Forest



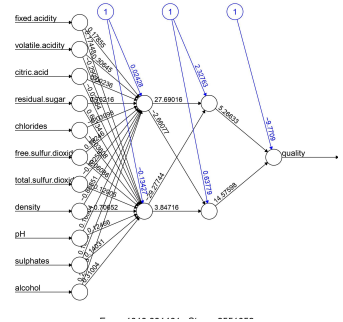
SVM

```
Call:
svm(formula = quality ~ ., data = trainingSetRed)


Parameters:
   SVM-Type:  eps-regression
 SVM-Kernel:  radial
       cost:  1
      gamma:  0.09090909
    epsilon:  0.1


Number of Support Vectors:  1074
```

K Nearest Neighbor



Neural Networks



WHITE DATA SET

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.8348 -0.4934 -0.0379  0.4637  3.1143

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.502e+02  1.880e+01   7.987 1.71e-15 ***
fixed.acidity         6.552e-02  2.087e-02   3.139  0.00171 **
volatile.acidity     -1.863e+00  1.138e-01 -16.373  < 2e-16 ***
citric.acid           2.209e-02  9.577e-02   0.231  0.81759
residual.sugar        8.148e-02  7.527e-03  10.825  < 2e-16 ***
chlorides            -2.473e-01  5.465e-01  -0.452  0.65097
free.sulfur.dioxide   3.733e-03  8.441e-04   4.422 9.99e-06 ***
total.sulfur.dioxide -2.857e-04  3.781e-04  -0.756  0.44979
density              -1.503e+02  1.907e+01  -7.879 4.04e-15 ***
pH                    6.863e-01  1.054e-01   6.513 8.10e-11 ***
sulphates             6.315e-01  1.004e-01   6.291 3.44e-10 ***
alcohol               1.935e-01  2.422e-02   7.988 1.70e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7514 on 4886 degrees of freedom
Multiple R-squared:  0.2819,    Adjusted R-squared:  0.2803
F-statistic: 174.3 on 11 and 4886 DF,  p-value: < 2.2e-16
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-3.6570 -0.5004 -0.0375  0.4756  3.1095

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           1.537e+02  1.960e+01   7.843 5.64e-15 ***
fixed.acidity         6.919e-02  2.237e-02   3.093 0.001995 **
volatile.acidity     -1.904e+00  1.246e-01 -15.289  < 2e-16 ***
residual.sugar        8.245e-02  7.980e-03  10.331  < 2e-16 ***
free.sulfur.dioxide   2.639e-03  7.469e-04   3.533 0.000416 ***
density              -1.540e+02  1.987e+01  -7.753 1.14e-14 ***
pH                    7.474e-01  1.150e-01   6.500 9.05e-11 ***
sulphates             6.000e-01  1.116e-01   5.374 8.14e-08 ***
alcohol               1.924e-01  2.612e-02   7.364 2.16e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7506 on 3909 degrees of freedom
Multiple R-squared:  0.2797,    Adjusted R-squared:  0.2782
F-statistic: 189.7 on 8 and 3909 DF,  p-value: < 2.2e-16
```

LINEAR REGRESSION

- Initially run models on full datasets to identify and remove statistically insignificant features (P value > 0.05)

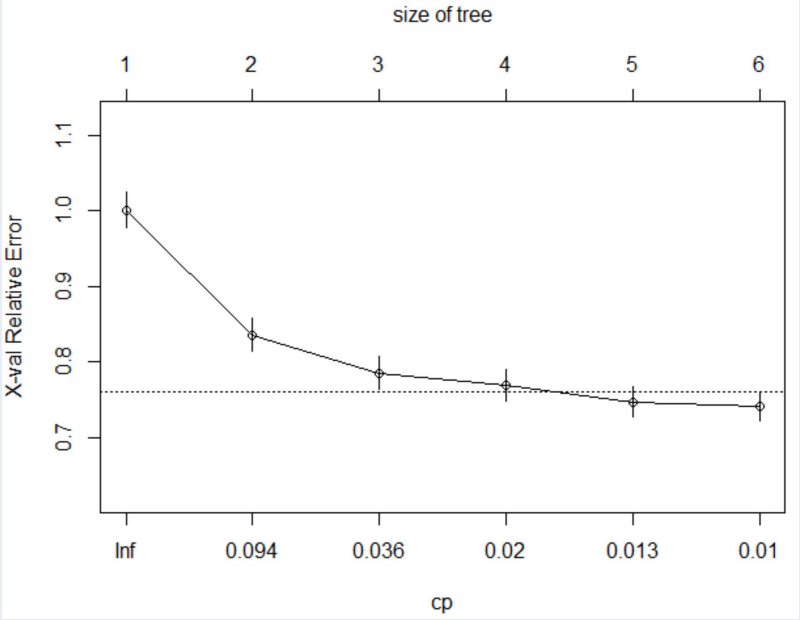
- Note that this model assumes a linear relationship between the output and input variables (which is not necessarily the case)

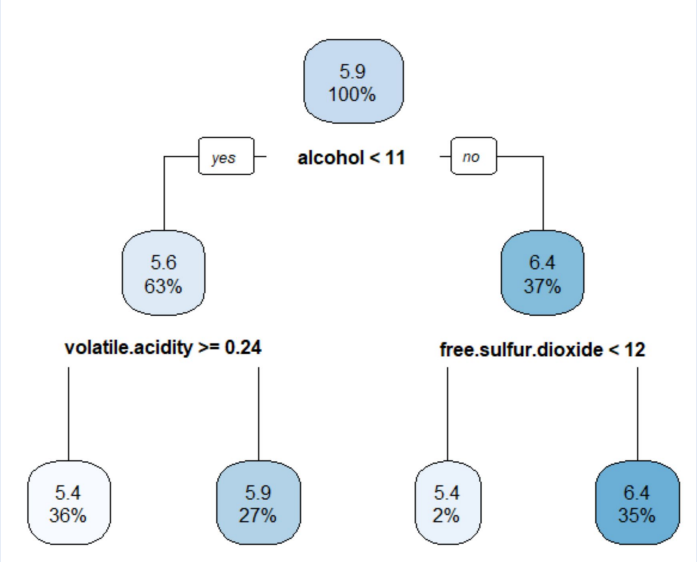
Generated MSEs*

White Wine	Red Wine
0.5731	0.3901

*Used 5-fold validation to generate average MSEs

WHITE DATA SET





DECISION TREE

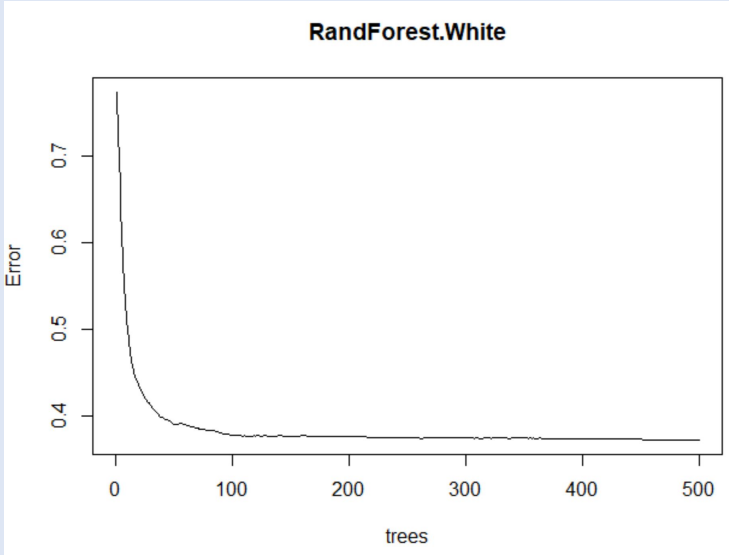
- Initially run models on full datasets to plot cp and prune tree to optimized cp level to avoid overfitting (cp = 0.02 for white, 0.025 for red)
- Useful in providing a visual representation of the specific features and exact thresholds used to predict final wine quality

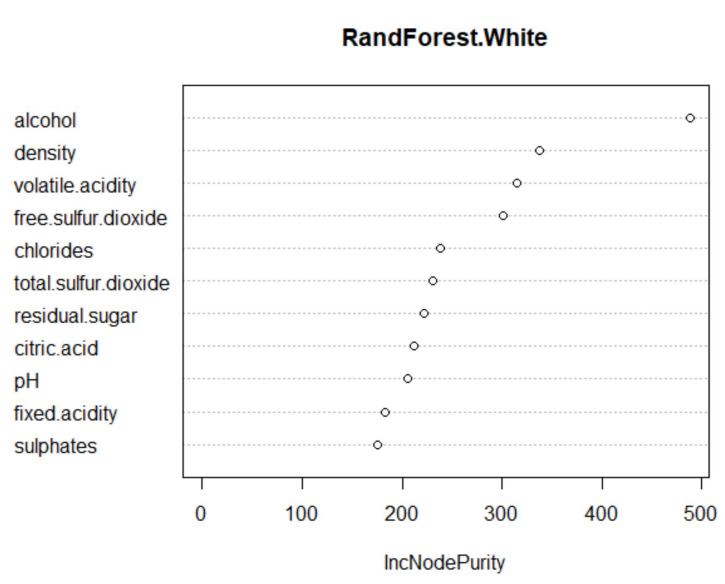
Generated MSEs

White Wine	Red Wine
0.6241	0.4590

*Used 5-fold validation to generate average MSEs

WHITE DATA SET





RANDOM FOREST

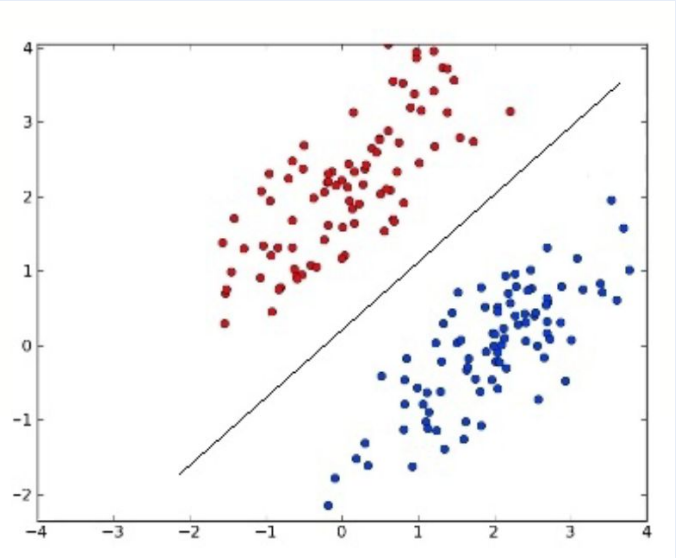
- Like decision trees, these models initially ran on full datasets and were plotted to identify & prune for the minimal number of trees to save on performance (200 for both)
- Using the variable importance function, we can also determine the order of impact each feature has on the predicted outcome

Generated MSEs

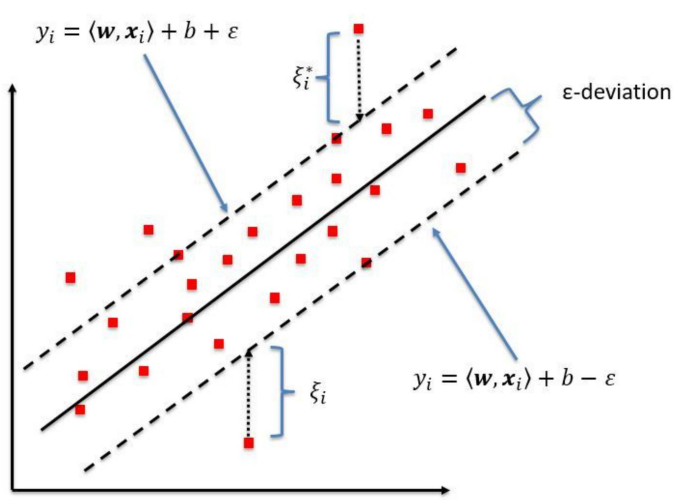
White Wine	Red Wine
0.3903	0.3037

*Used 5-fold validation to generate average MSEs

SUPPORT VECTOR MACHINE



SUPPORT VECTOR REGRESSION



SVR (Support Vector Regression)

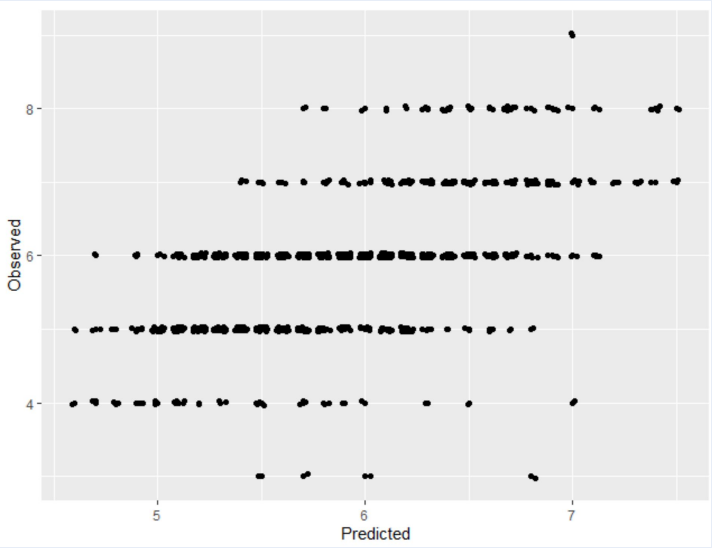
- Unlike SVM which only works with binary classification, SVR is a model that uses the same underlying concepts but now maximizes the distance between the support vectors to the regressed curve to predict a continuous output
- Through tuning, the radial function was chosen for best results

Generated MSEs

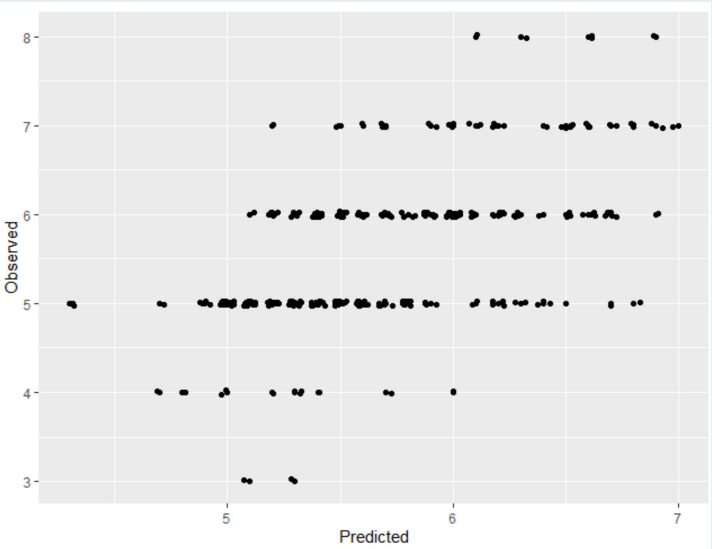
White Wine	Red Wine
0.4853	0.3785

*Used 5-fold validation to generate average MSEs

WHITE DATA SET



RED DATA SET



K-NEAREST NEIGHBOUR

- Run models on standardized dataset and tune parameters to identify optimized k value (k = 10)

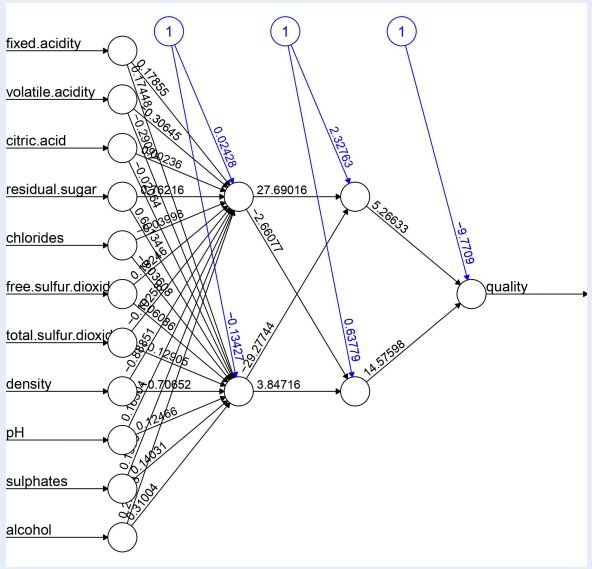
- Predicted vs. Observed graphs can be generated to visually identify accuracy of model (look for large clusters around intersections where x ≈ y)

Generated MSEs

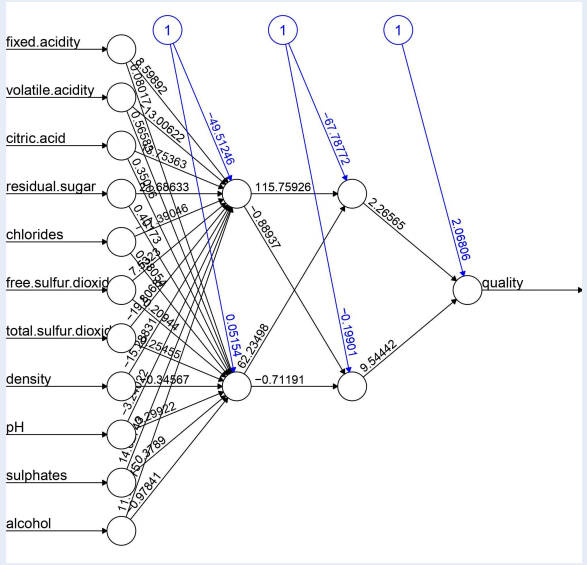
White Wine	Red Wine
0.5218	0.4049

*Used 5-fold validation to generate average MSEs

WHITE DATA SET



RED DATA SET



NEURAL NETWORKS

- Run models on standardized dataset and utilize hidden layer structure of 2,2 (parameter tuning was not feasible due to long compile times)
- Due to the computational complexity of running a neural network on large datasets (30min+ runtime), the MSE values below are generated from a single run

Generated MSEs

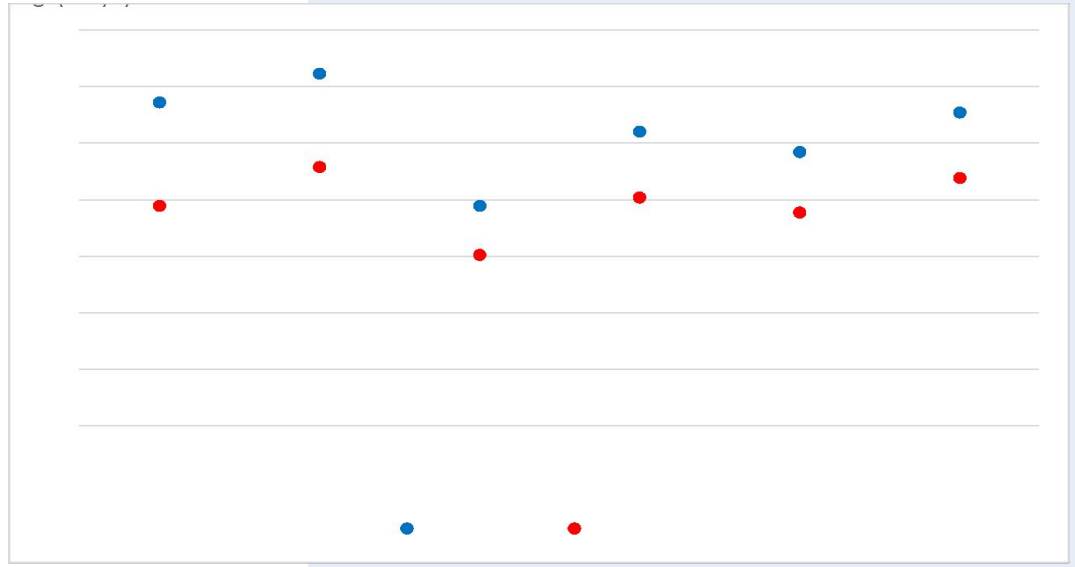
White Wine	Red Wine
0.5552	0.4395

ANALYSIS OF RESULTS

MSE VALUES ARE COMPARED ACROSS MODELS

- Models produced a collection of MSEs with values ranging between **0.3037 and 0.6241**
- All models consistently produced **lower MSE values for red wine compared to white**
- The **random forest model produced the lowest MSE for both red and white wine**, while the decision tree model produced the highest MSEs

Model	White Wine MSE	Red Wine MSE
Linear Regression	0.5731	0.3901
Decision Tree	0.6241	0.4590
Random Forest	0.3903	0.3037
KNN	0.5218	0.4049
SVM	0.4853	0.3785
Neural Network	0.5552	0.4395



IMPORTANT FEATURES FOR WINE QUALITY

ALCOHOL – BOTH

Alcohol content stands out as the **strongest determinant** of quality

SULPHATES – HIGH RED, LOW WHITE

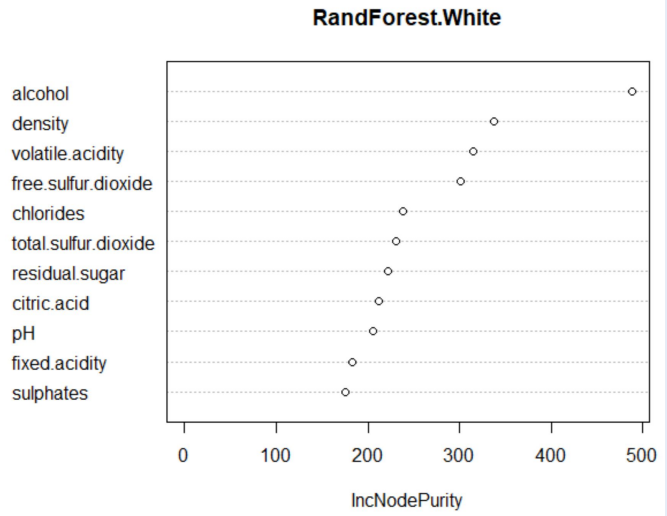
Sulphates have been added to wine since the 1800s to **protect against oxidization**; sulphate content alters the colour and taste of the wine, with **higher content distorting the taste of wine**

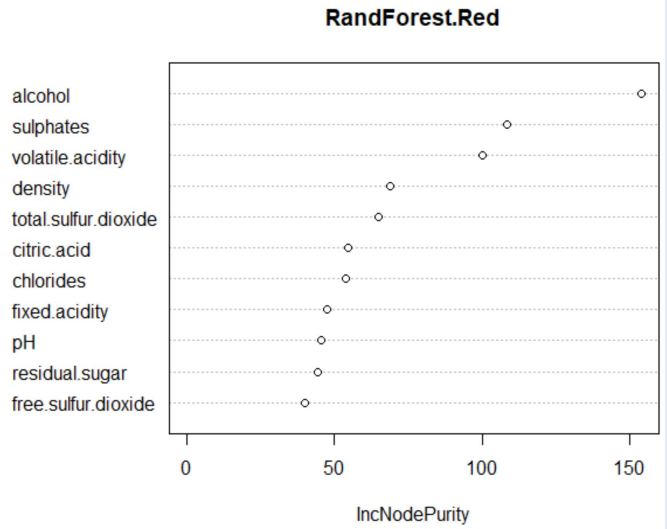
VOLATILE ACIDITY – BOTH

Measures the low weight molecules that are **released from the wine to produce its odor**. This can be manipulated by **adjusting acetic acid levels**

DENSITY – BOTH

Used to **measure sugar content**; higher density wines tend to be higher quality





BUSINESS APPLICATIONS

THE MODEL CAN BE USED FOR RESEARCHING, DEVELOPING, AND PRICING WINE

Research and Development Benchmarking

- Managers can set standard levels for less important features to lower cost of production
- Vary levels of important features based on desired quality

Optimal Pricing Strategy

- Using additional price data, the relationship between important features and price can be mapped to set the optimal price