
Data Sommeliers:
Uncorking the Mysteries of
Wine Quality

AGENDA

01. Overview

Objectives and applications

02. Exploratory Data Analysis

First sip

03. Solutions and Insights

Recommendations

OBJECTIVES

1. Do highly rated wines exhibit similar features?
2. Do the same features matter for red and white wines?
3. If we cluster the wines (without quality ratings), would all wines in a particular cluster have similar ratings?

APPLICATIONS

Vineyards

Winemakers

Sommeliers

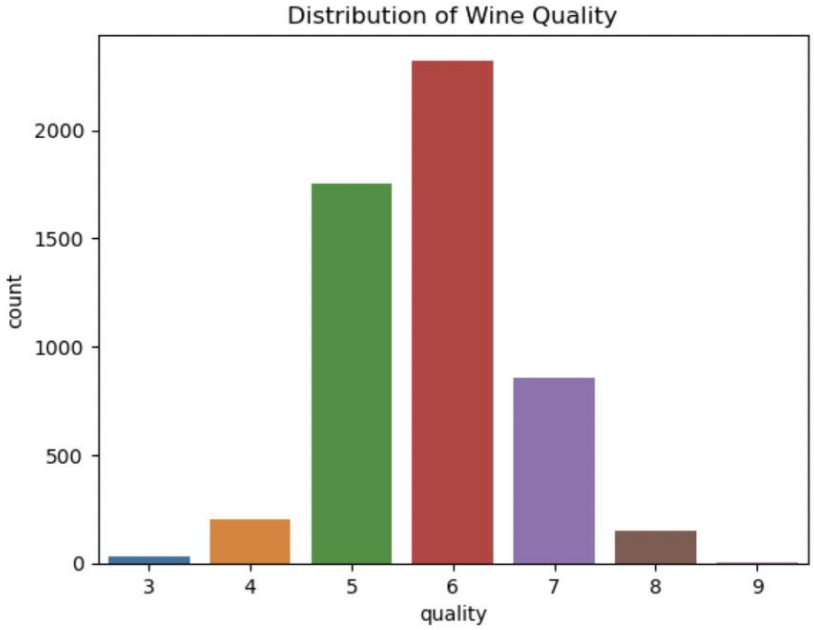
Restaurants

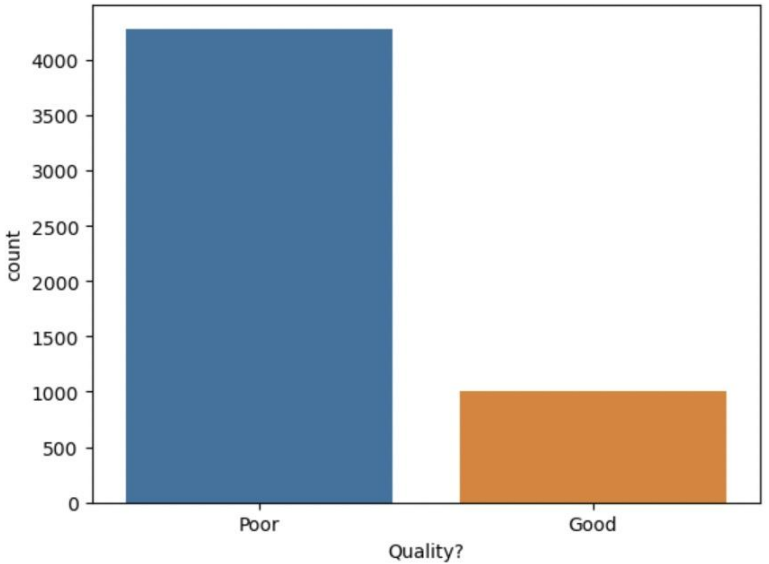
DATA SETS

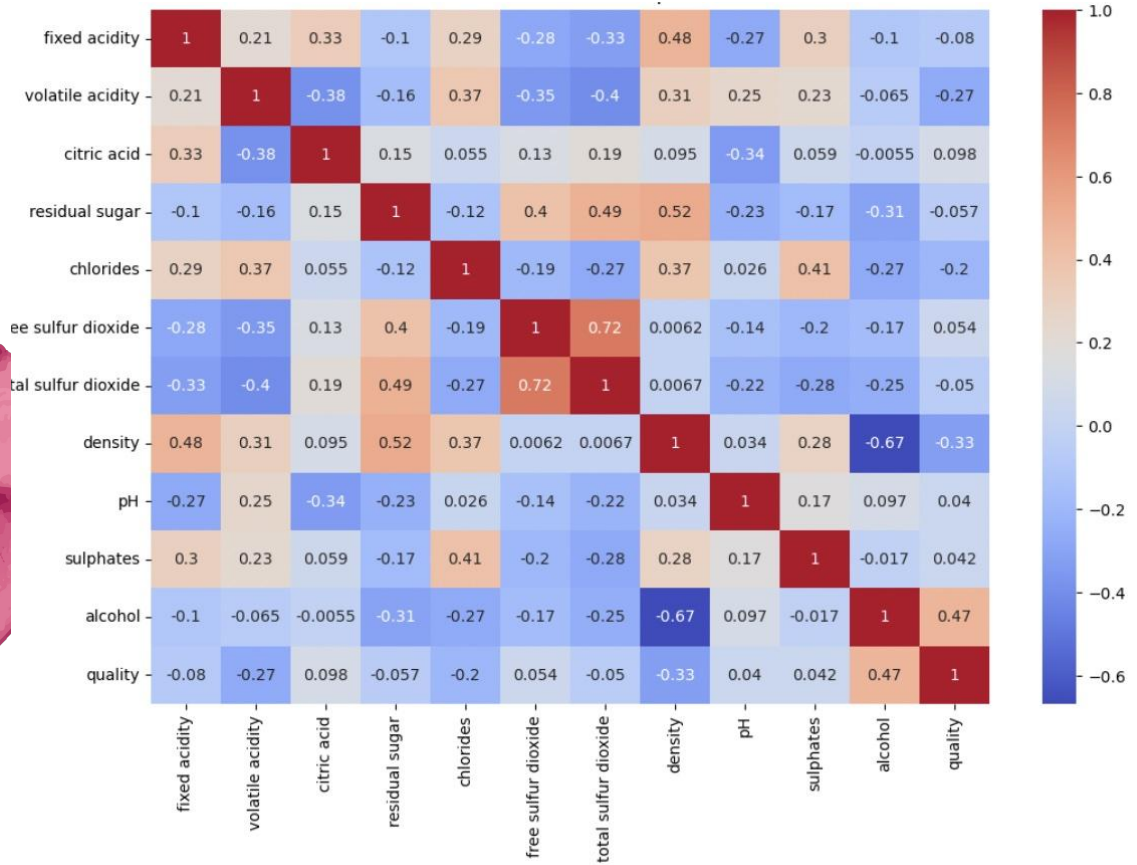
3961

1359

EXPLORATORY DATA ANALYSIS







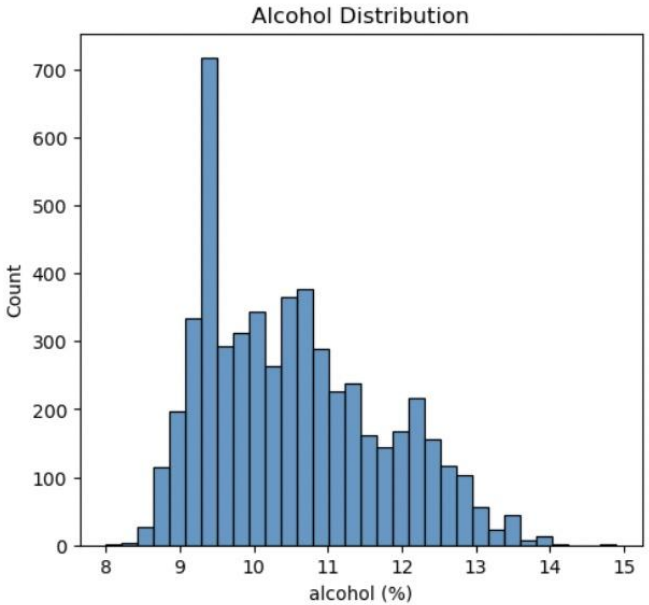
Alcohol

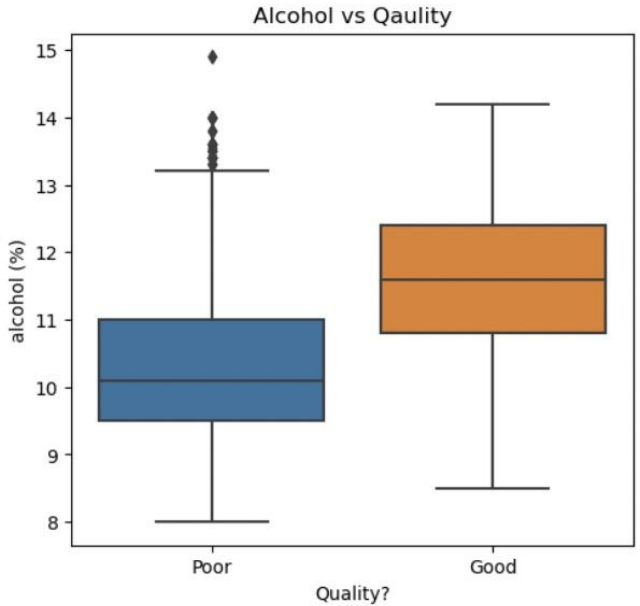
Density

Volatile Acidity

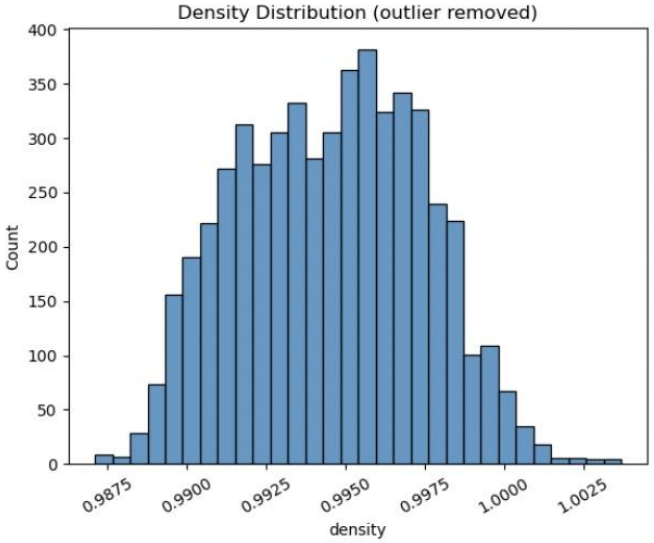
Chlorides

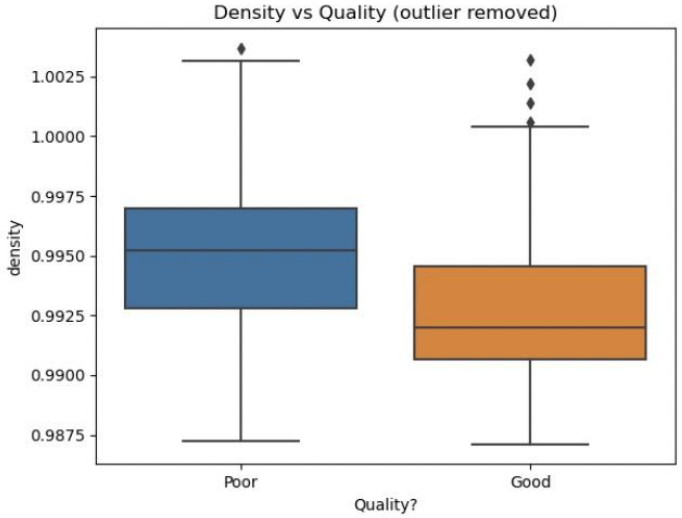
ALCOHOL vs QUALITY



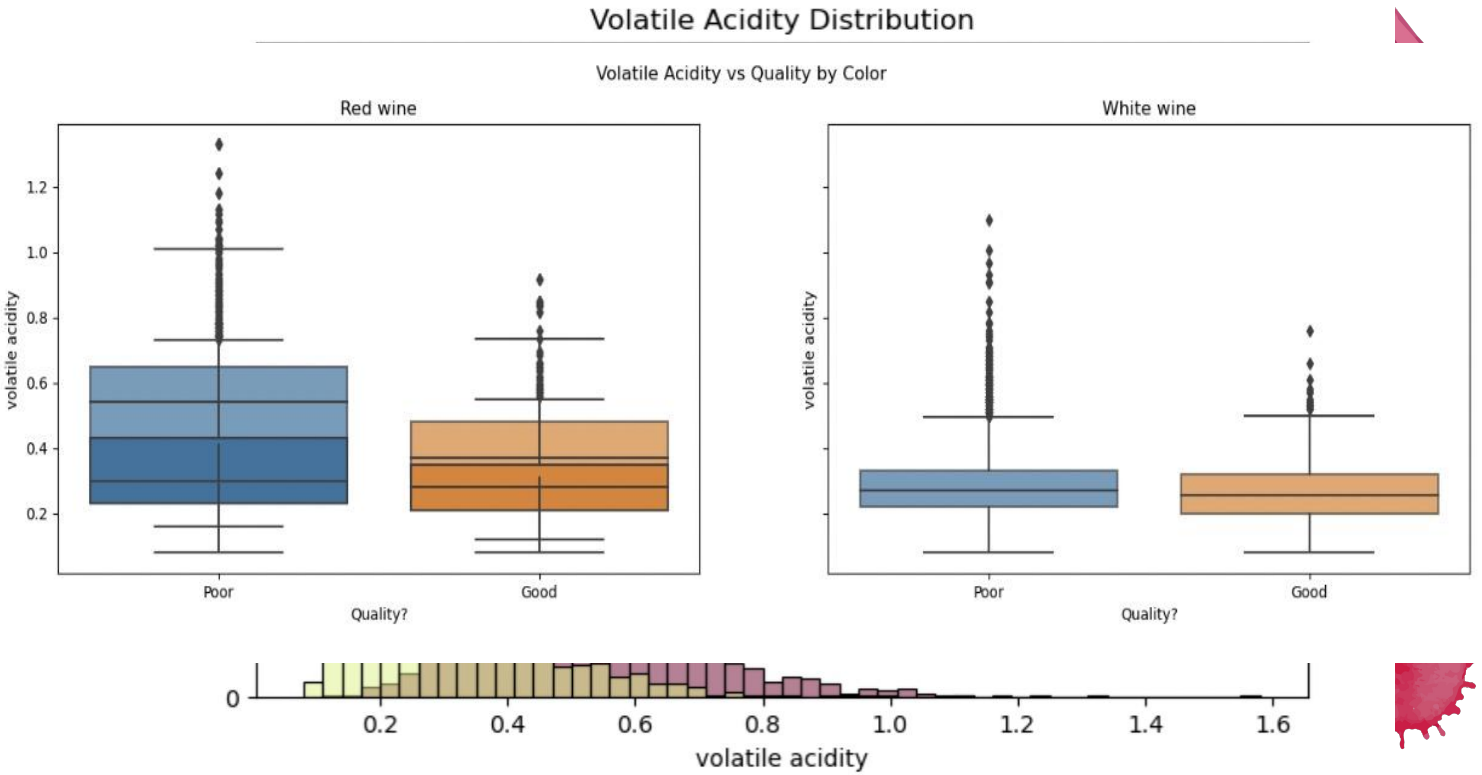


DENSITY vs QUALITY

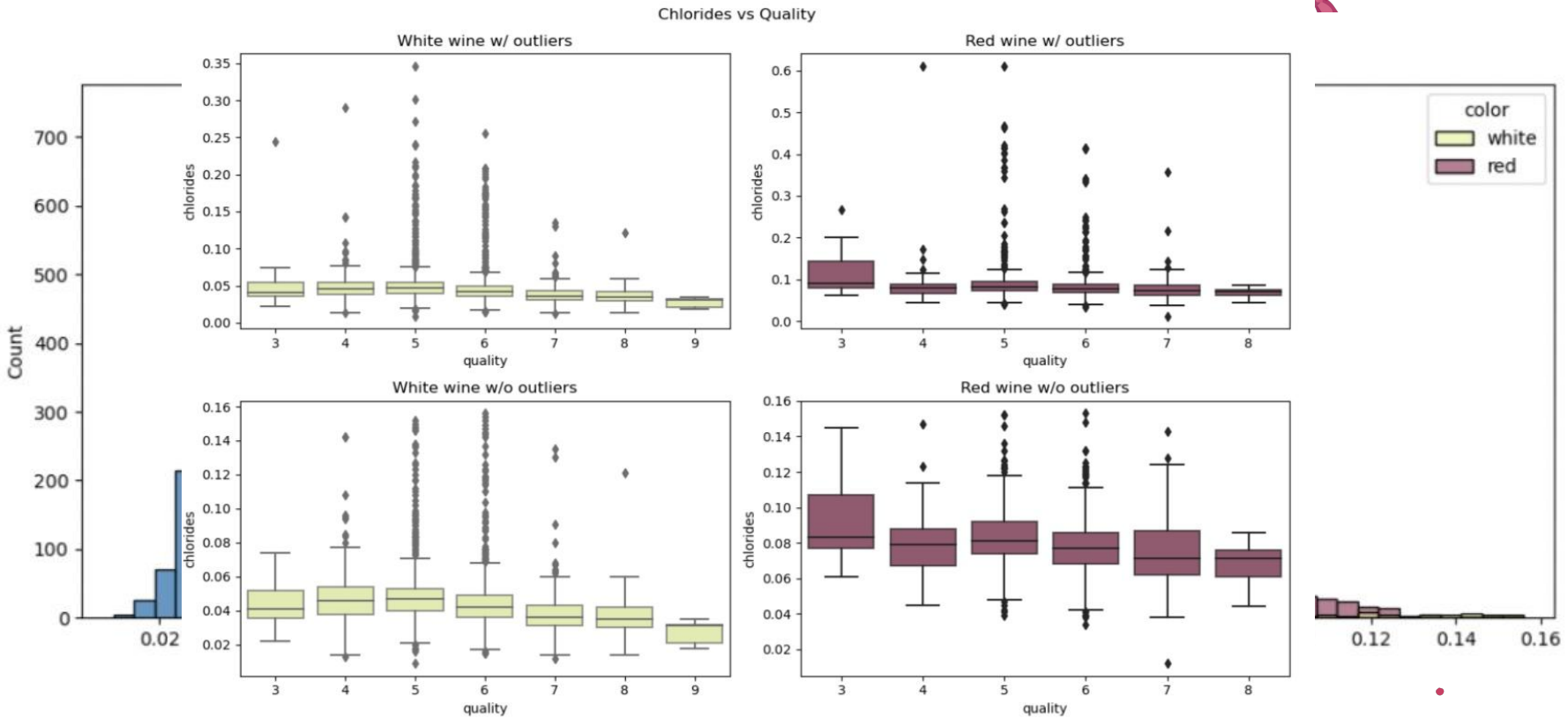




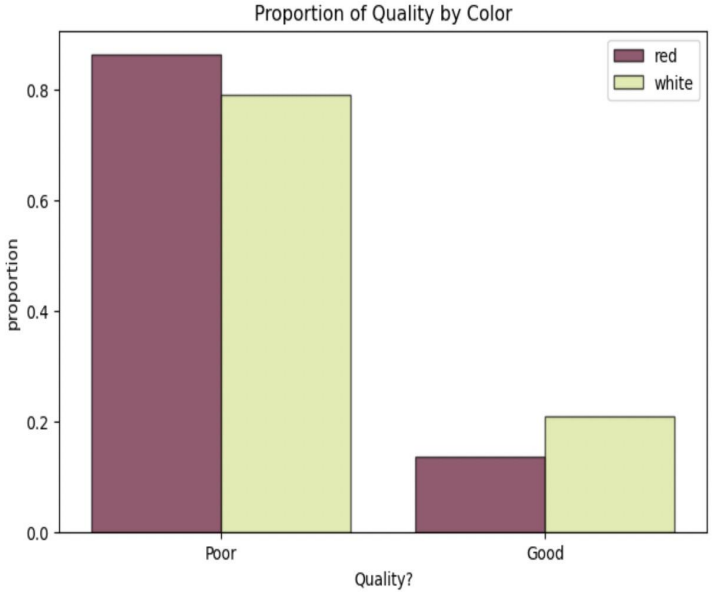
VOLATILE ACIDITY vs QUALITY

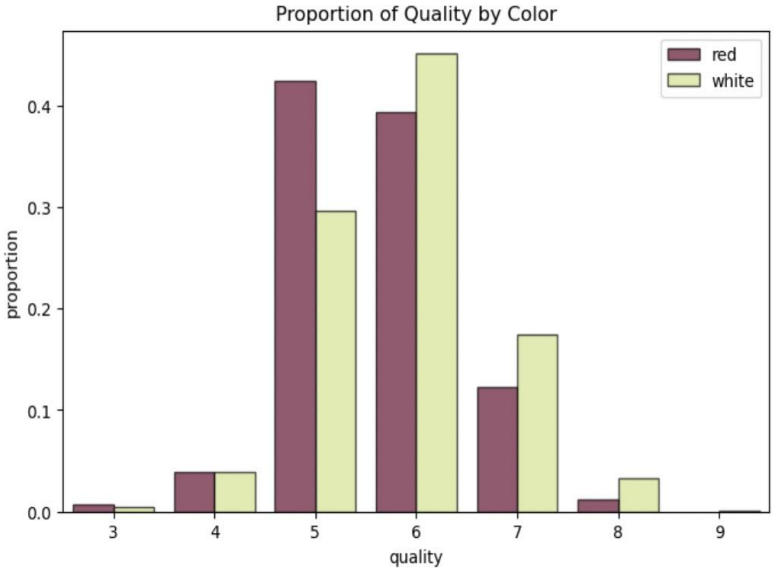


CHLORIDES vs QUALITY



RED vs WHITE


Proportion of Quality by Color
red
white
0.8
0.6
0.4
0.2
0.0
proportion
Poor
Good
Quality?


Proportion of Quality by Color
red
white
0.4
0.3
0.2
0.1
0.0
proportion
3
4
5
6
7
8
9
quality

WHAT IMPACTS WINE QUALITY?

Higher alcohol content => higher quality

Correlated with decreased density

Red wines: higher quality linked to a decrease in volatile acidity

A slightly elevated likelihood of being white wines

Logistic Regression

WHITE WINE

WHITE WINE - MODEL ACCURACY

81.57%

Training Set Accuracy

80.40%

Test Set Accuracy

P-VALUES AND COEFFICIENTS

```
Optimization terminated successfully.
         Current function value: 0.395041
         Iterations 7
                           Logit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:                 2772
Model:                          Logit   Df Residuals:                     2760
Method:                           MLE   Df Model:                           11
Date:                Fri, 04 Aug 2023   Pseudo R-squ.:                  0.2369
Time:                        21:54:33   Log-Likelihood:                -1095.1
converged:                       True   LL-Null:                       -1435.1
Covariance Type:            nonrobust   LLR p-value:                1.036e-138
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -1.8828      0.073    -25.845      0.000      -2.026      -1.740
x1             0.3813      0.101      3.769      0.000       0.183       0.580
x2            -0.3164      0.068     -4.655      0.000      -0.450      -0.183
x3            -0.0145      0.067     -0.216      0.829      -0.146       0.117
x4             1.1344      0.226      5.021      0.000       0.692       1.577
x5            -0.5872      0.132     -4.450      0.000      -0.846      -0.329
x6             0.2869      0.076      3.754      0.000       0.137       0.437
x7            -0.1590      0.091     -1.746      0.081      -0.338       0.020
x8            -1.5425      0.359     -4.301      0.000      -2.245      -0.840
x9             0.5486      0.084      6.507      0.000       0.383       0.714
x10            0.2549      0.053      4.798      0.000       0.151       0.359
x11            0.4345      0.182      2.383      0.017       0.077       0.792
==============================================================================
```

	Variable	P-Value
x1	fixed_acidity	1.640111e-04
x2	volatile_acidity	3.236556e-06
x3	citric_acid	8.289482e-01
x4	residual_sugar	5.141865e-07
x5	chlorides	8.574005e-06
x6	free_sulfur_dioxide	1.737174e-04
x7	total_sulfur_dioxide	8.084902e-02
x8	density	1.698186e-05
x9	pH	7.661145e-11
x10	sulphates	1.604116e-06
x11	alcohol	1.718325e-02

pH | Most Significant Variable

Residual Sugar, Sulfates, Volatile Acidity | Next Most Significant Variables

pH, residual sugar, and sulfates are positively associated with classification of “good wine”!
Volatile acidity are negatively associated with classification of “good wine”!

RED WINE

RED WINE - MODEL ACCURACY

88.34%

Training Set Accuracy

87.01%

Test Set Accuracy

P-VALUES AND COEFFICIENTS

```
Optimization terminated successfully.
         Current function value: 0.261783
         Iterations 8
                           Logit Regression Results
==============================================================================
Dep. Variable:                 target   No. Observations:                  951
Model:                          Logit   Df Residuals:                      939
Method:                          MLE   Df Model:                           11
Date:               Fri, 04 Aug 2023   Pseudo R-squ.:                  0.3650
Time:                       22:09:07   Log-Likelihood:                -248.96
converged:                      True   LL-Null:                       -392.06
Covariance Type:            nonrobust   LLR p-value:                6.997e-55
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
const         -3.1279      0.223    -14.049      0.000      -3.564      -2.692
x1             0.3274      0.284      1.154      0.249      -0.229       0.884
x2            -0.5836      0.199     -2.928      0.003      -0.974      -0.193
x3            -0.0243      0.222     -0.110      0.913      -0.459       0.411
x4             0.2238      0.140      1.598      0.110      -0.051       0.498
x5            -0.3297      0.181     -1.826      0.068      -0.684       0.024
x6             0.4059      0.188      2.160      0.031       0.038       0.774
x7            -1.1544      0.295     -3.913      0.000      -1.733      -0.576
x8            -0.1893      0.274     -0.690      0.490      -0.727       0.348
x9            -0.0955      0.211     -0.452      0.651      -0.510       0.319
x10            0.8373      0.134      6.239      0.000       0.574       1.100
x11            1.0544      0.195      5.418      0.000       0.673       1.436
==============================================================================
```

	Variable	P-Value
x1	fixed_acidity	2.486780e-01
x2	volatile_acidity	3.416577e-03
x3	citric_acid	9.127853e-01
x4	residual_sugar	1.100302e-01
x5	chlorides	6.786139e-02
x6	free_sulfur_dioxide	3.077258e-02
x7	total_sulfur_dioxide	9.111546e-05
x8	density	4.898839e-01
x9	pH	6.513099e-01
x10	sulphates	4.409700e-10
x11	alcohol	6.013860e-08

Sulfates | Most Significant Variable

Alcohol, Total Sulfur Dioxide, Volatile Acidity | Next Most Significant Variables

Sulfates and alcohol are positively associated with classification of “good wine”!
Total sulfur dioxide and volatile acidity are negatively associated with classification of “good wine”!

Random Forest

WHITE WINE

70:30

Train:Test Data Split
for 3961 Samples

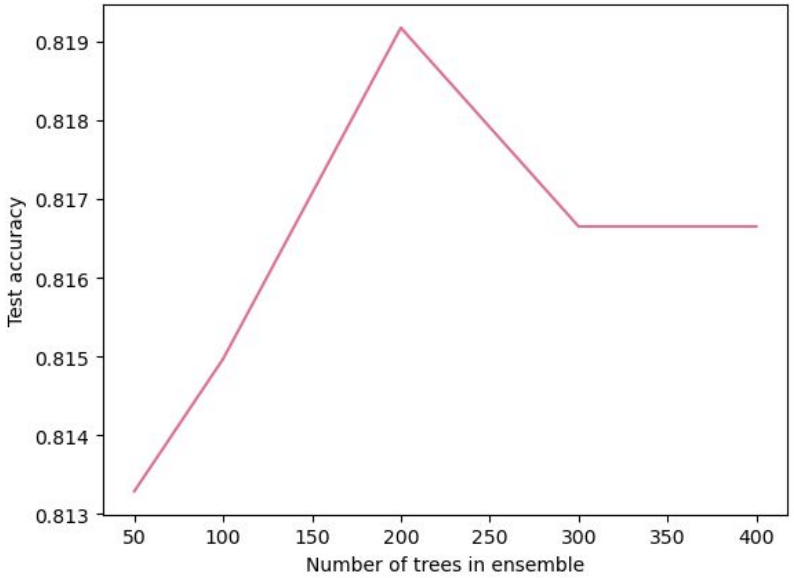
200

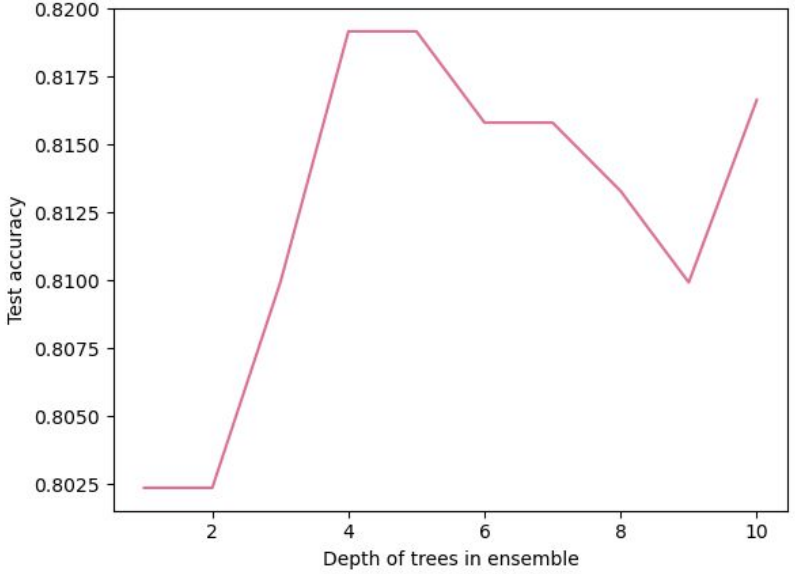
Number of Decision Trees

4

Maximum Depth of Trees

Deciding Model Parameters





MODEL ACCURACY

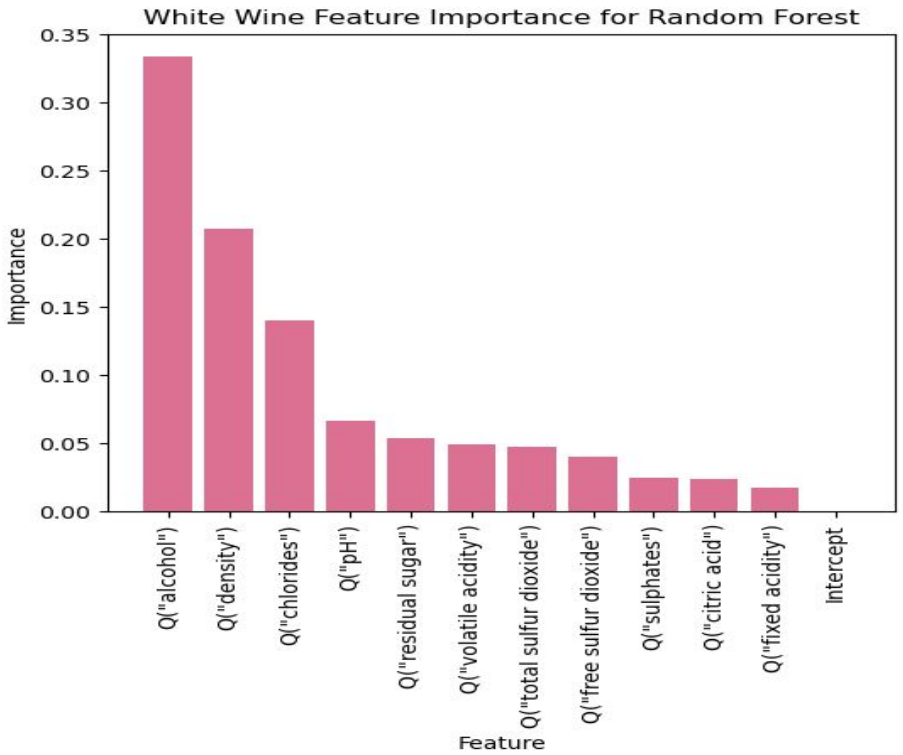
82.75%

Training Set Accuracy

81.74%

Test Set Accuracy

VARIABLE IMPORTANCE



Alcohol | Most Significant Variable

Density, Chlorides, Volatile Acidity | Next Most Significant Variables

There seems to be a drastic drop in significance between alcohol & other variables.

RED WINE

70:30

Train:Test Data Split
for 1359 Samples

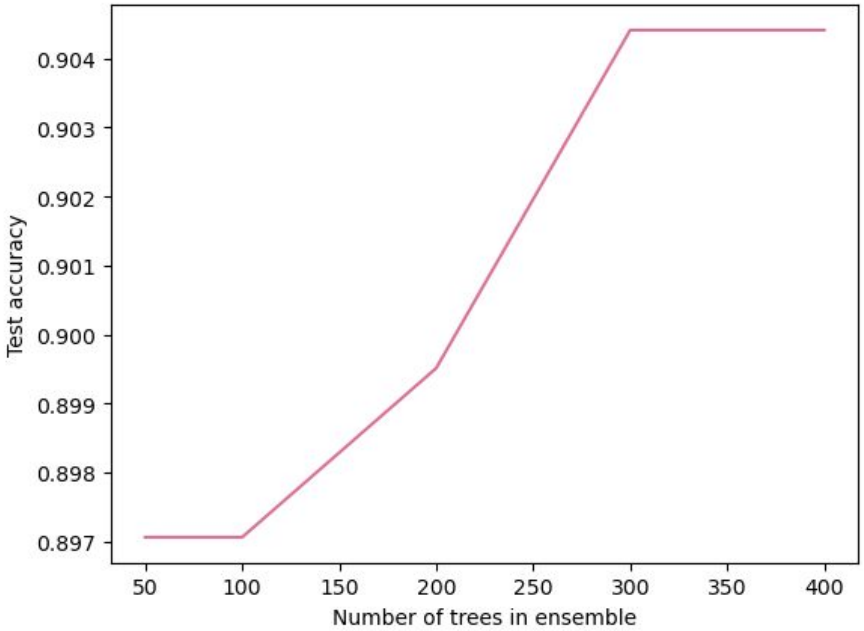
300

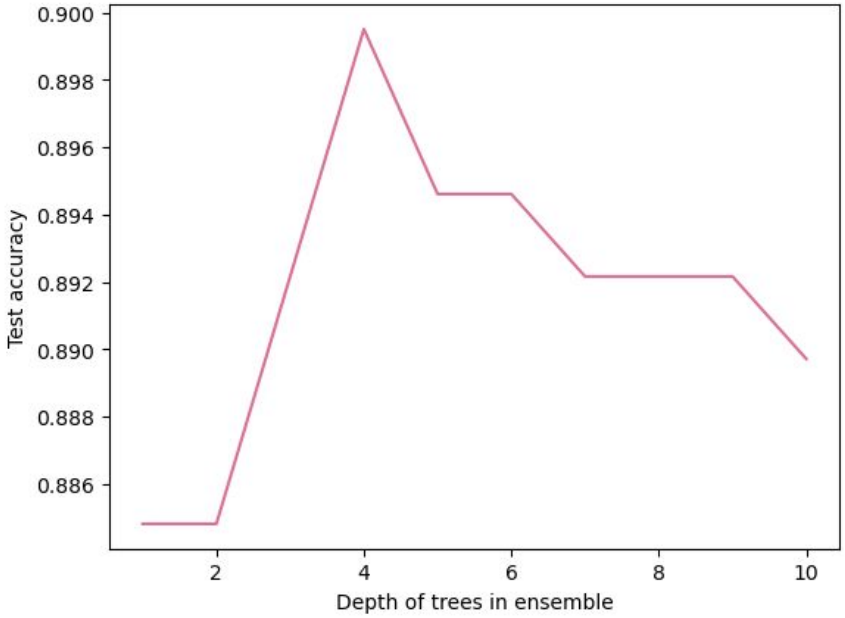
Number of Decision Trees

4

Maximum Depth of Trees

Deciding Model Parameters





MODEL ACCURACY

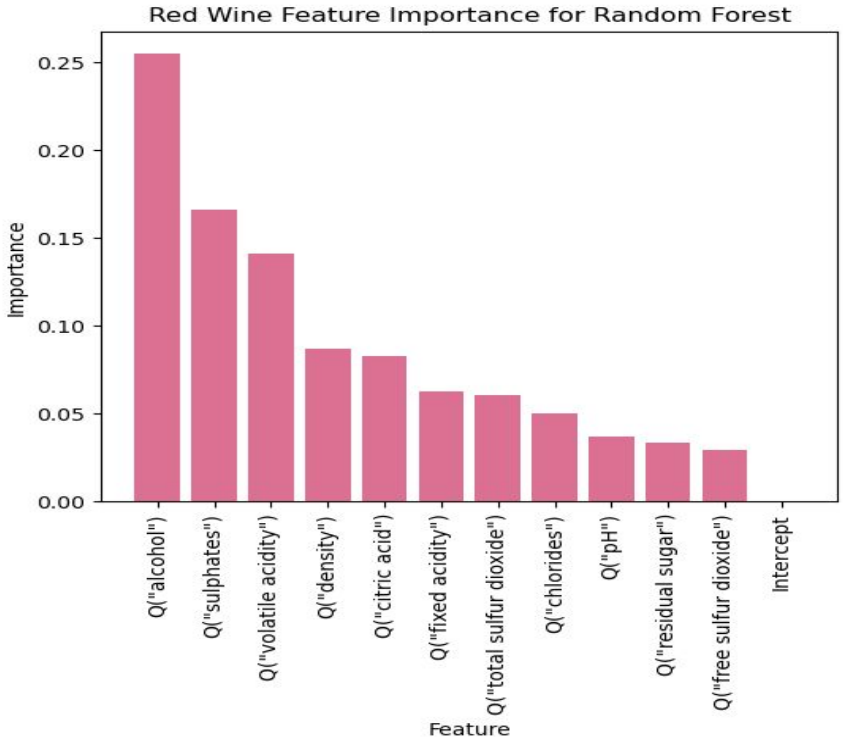
89.90%

Training Set Accuracy

88.72%

Test Set Accuracy

VARIABLE IMPORTANCE



Alcohol | Most Significant Variable

Sulphates, Volatile Acidity, Density | Next Most Significant Variables

There seems to be a drastic drop in significance between alcohol & other variables.

Boosting

AKA Gradient Boosting Model

WHITE WINE

70:30

Train:Test Data Split
for 3961 Samples

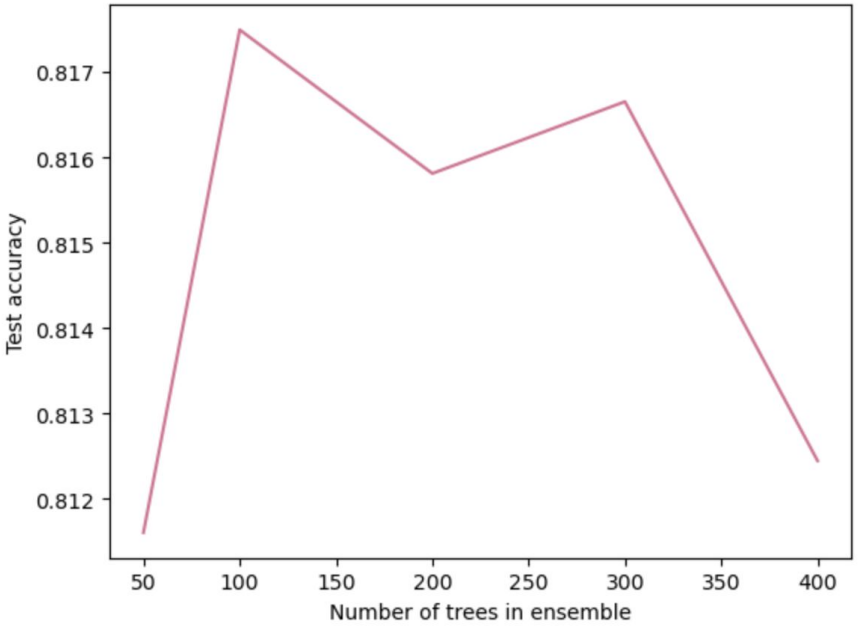
100

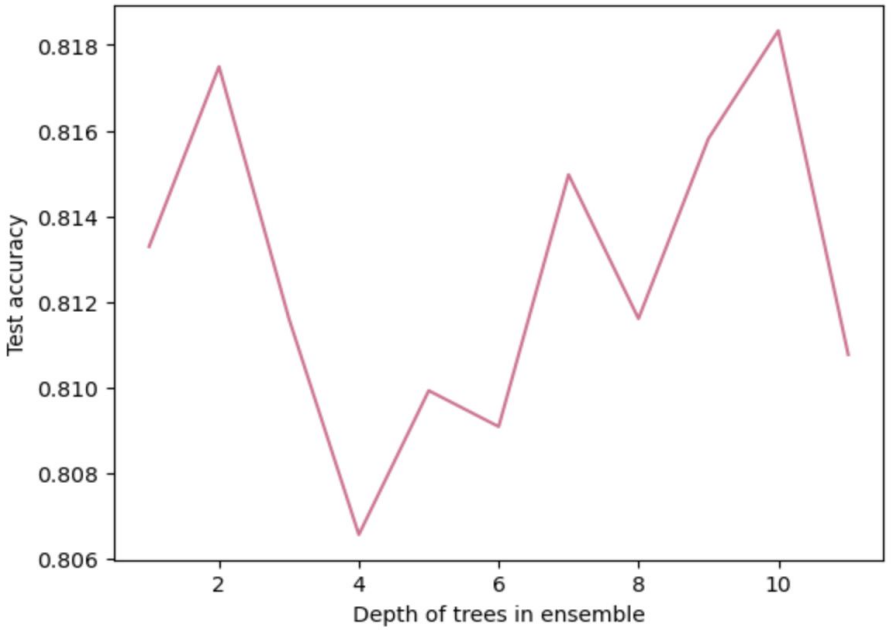
Number of Decision Trees

2

Maximum Depth of Trees

Deciding Model Parameters





MODEL ACCURACY

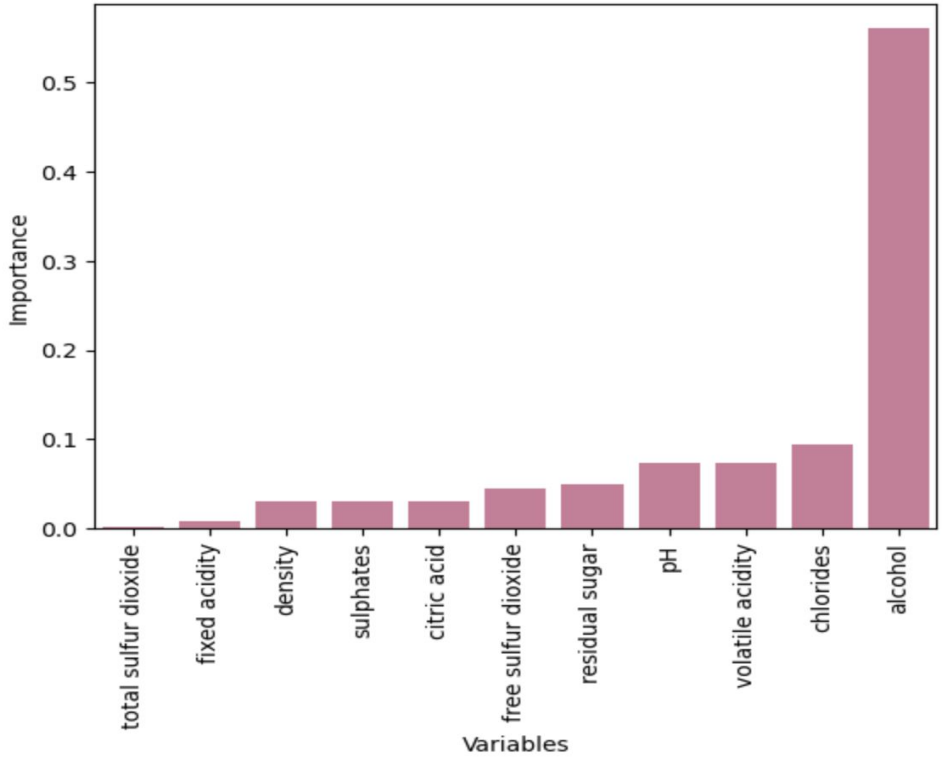
82.24%

Training Set Accuracy

81.75%

Test Set Accuracy

VARIABLE IMPORTANCE



Alcohol | Most Significant Variable

Chlorides, Volatile Acidity, pH | Next Most Significant Variables

There seems to be a drastic drop in significance between alcohol & other variables.

RED WINE

70:30

Train:Test Data Split
for 1359 Samples

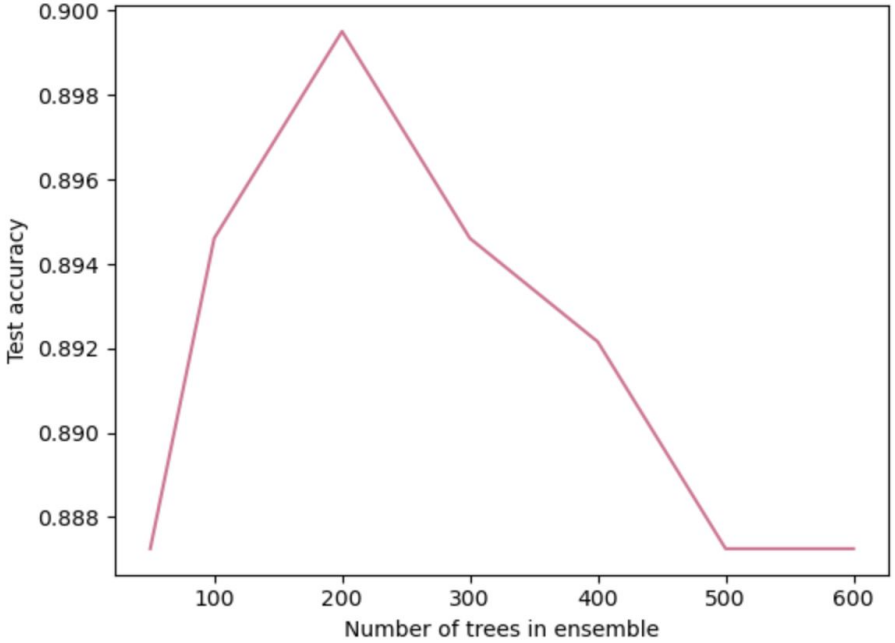
200

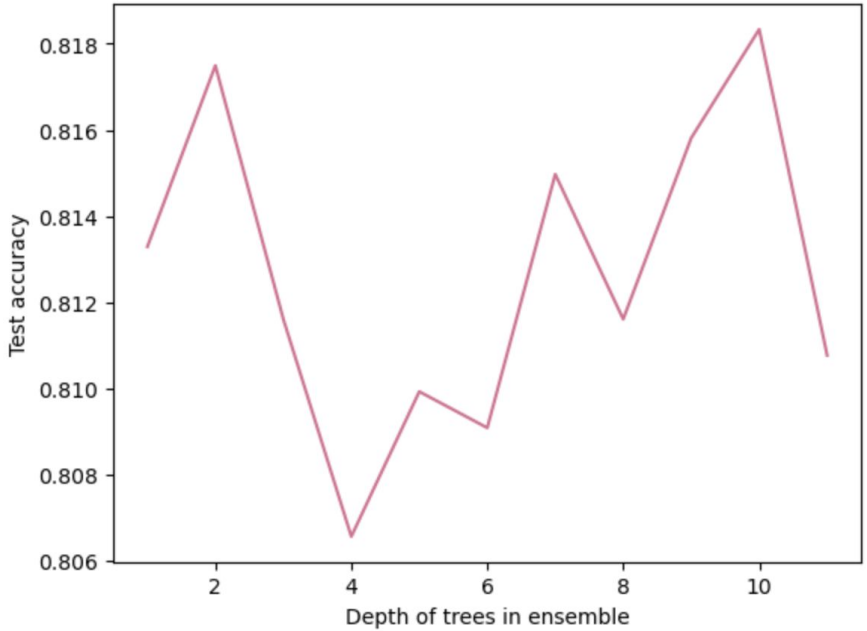
Number of Decision Trees

3

Maximum Depth of Trees

Deciding Model Parameters





MODEL ACCURACY

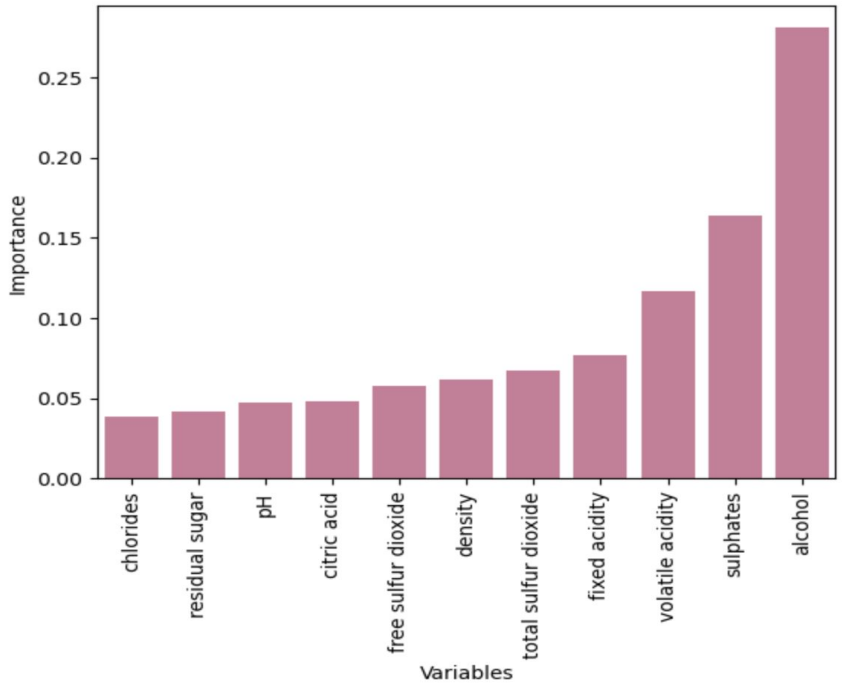
98.84%

Training Set Accuracy

89.95%

Test Set Accuracy

VARIABLE IMPORTANCE



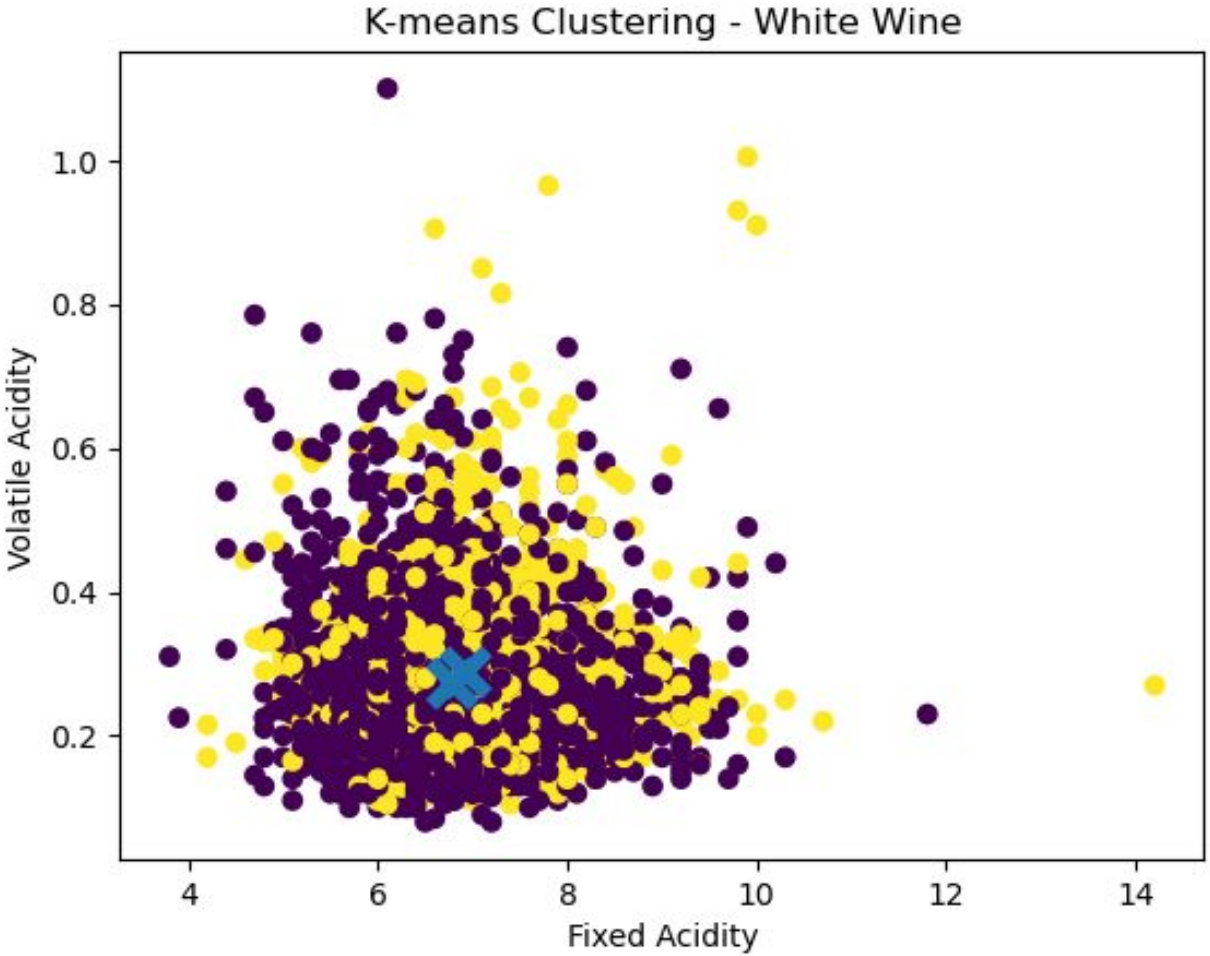
Alcohol | Most Significant Variable

Sulphates, Volatile Acidity, Fixed Acidity | Next Most Significant Variables

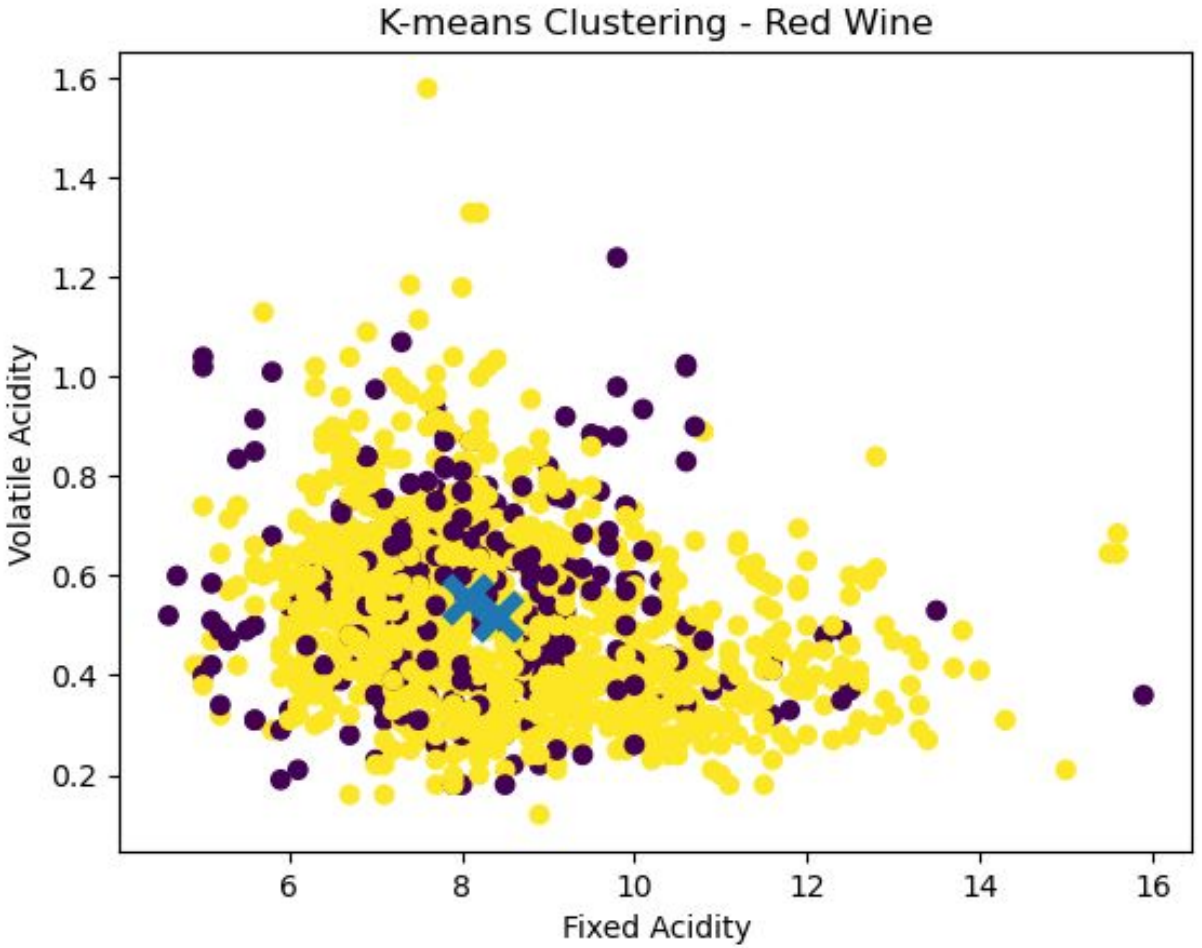
The drop in significance is not as drastic as White Wine Data

K-Means Clustering

47% Accuracy


K-means Clustering - White Wine
Volatile Acidity
1.0
0.8
0.6
0.4
0.2
4
6
8
10
12
14
Fixed Acidity

63% Accuracy


K-means Clustering - Red Wine
Volatile Acidity
Fixed Acidity
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
6
8
10
12
14
16

CONCLUSION

BEST MODELS

-White : Boosting (81.75)
-Red : Boosting (89.95)

BEST PREDICTORS

-White : Alcohol, Volatile Acidity, pH
-Red : Alcohol, Sulfates, Volatile Acidity

KEY TAKEAWAYS

- Quality improvement
- Product differentiation
- Marketing efforts

QUESTIONS?