Supervised Learning:

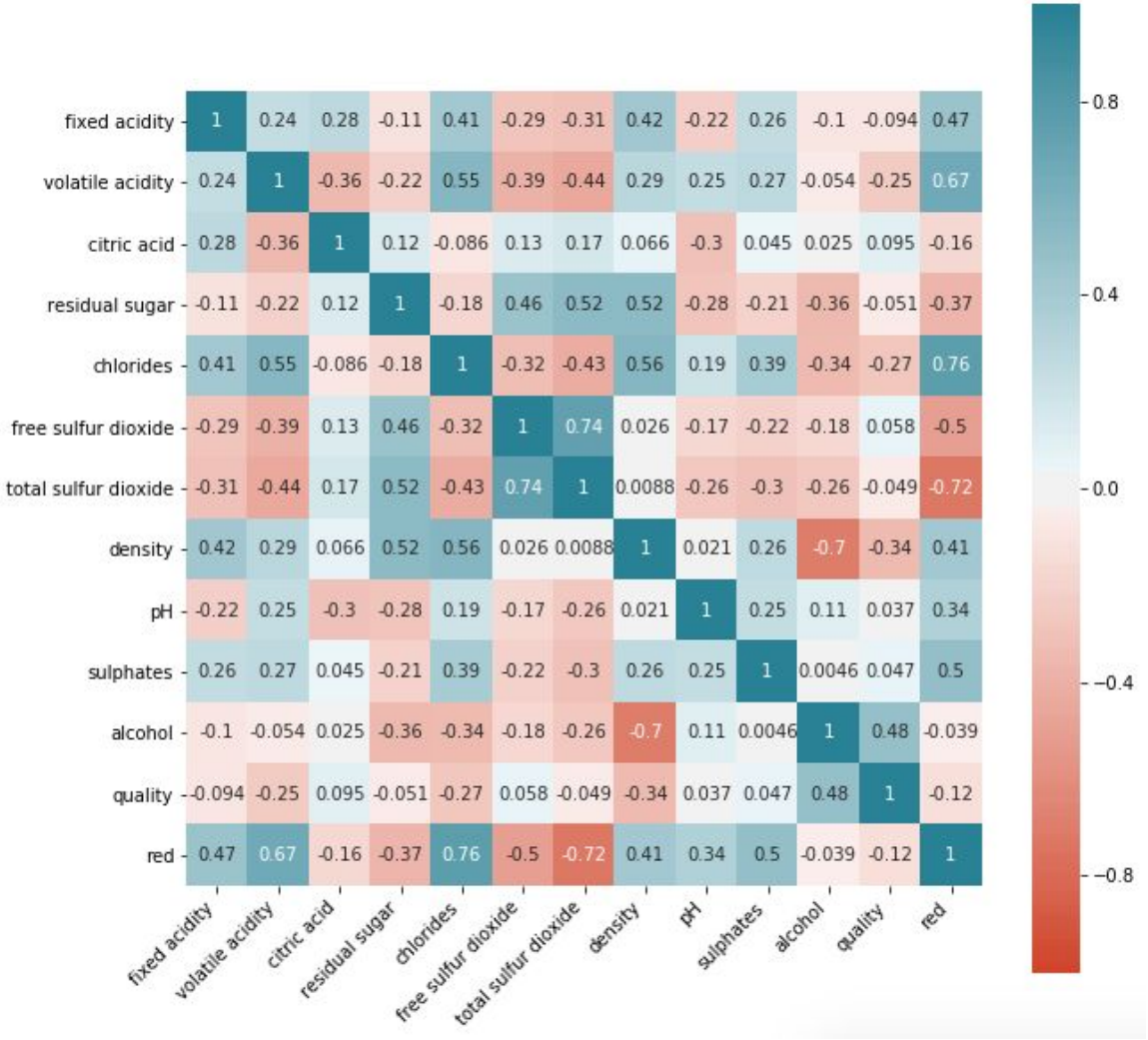
Predicting Wine Color

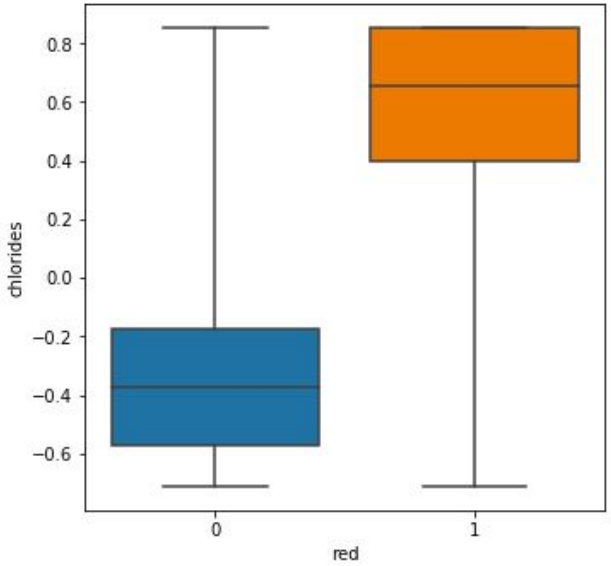
Motivation

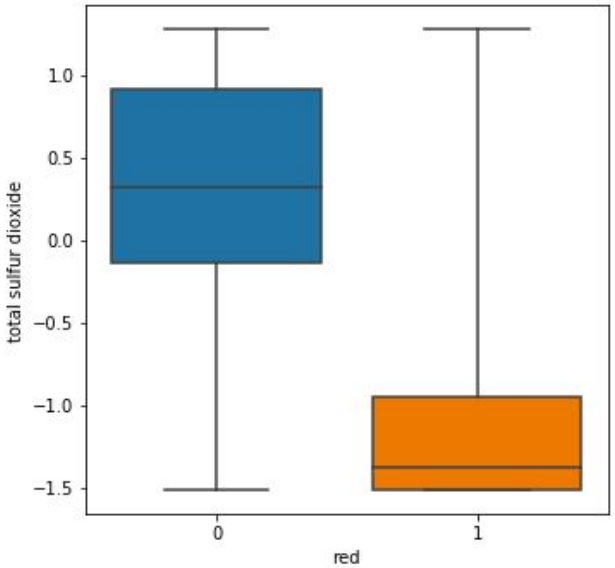
- Humans can differentiate wines by their color and flavor, but can we teach a computer to distinguish them by what chemicals they contain?
  - What model would be best for this classification problem?
  - What variables are most important to this classification?

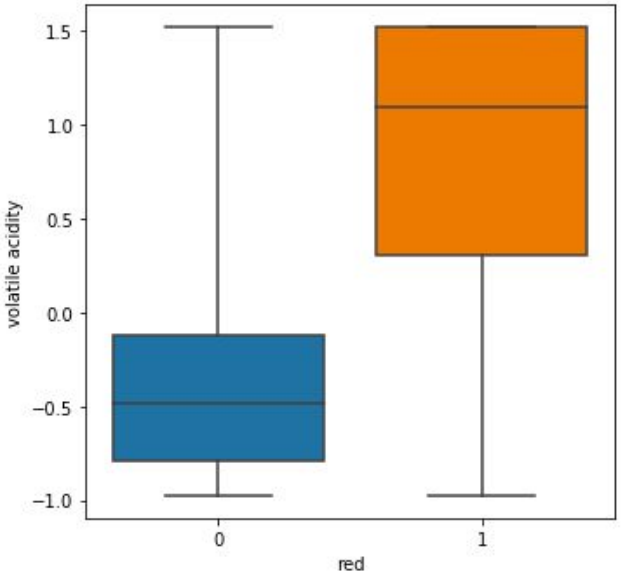
Data Source

- UC Irvine Machine Learning Repository - Wine Quality
  - Red Wine & White Wine - Imbalanced Set
  - 11 Features Regarding the Chemical Composition of the Wine
  - 1 Feature Regarding the Overall “Rating” of the Wine’s Flavor
- Modifications
  - Indicator Variable
  - Standardization
  - Winsorization


fixed acidity 1 0.24 0.28 -0.11 0.41 -0.29 -0.31 0.42 -0.22 0.26 -0.1 -0.094 0.47
volatile acidity 0.24 1 -0.36 -0.22 0.55 -0.39 -0.44 0.29 0.25 0.27 -0.054 -0.25 0.67
citric acid 0.28 -0.36 1 0.12 -0.086 0.13 0.17 0.066 -0.3 0.045 0.025 0.095 -0.16
residual sugar -0.11 -0.22 0.12 1 -0.18 0.46 0.52 0.52 -0.28 -0.21 -0.36 -0.051 -0.37
chlorides 0.41 0.55 -0.086 -0.18 1 -0.32 -0.43 0.56 0.19 0.39 -0.34 -0.27 0.76
free sulfur dioxide -0.29 -0.39 0.13 0.46 -0.32 1 0.74 0.026 -0.17 -0.22 -0.18 0.058 -0.5
total sulfur dioxide -0.31 -0.44 0.17 0.52 -0.43 0.74 1 0.0088 -0.26 -0.3 -0.26 -0.049 -0.72
density 0.42 0.29 0.066 0.52 0.56 0.026 0.0088 1 0.021 0.26 -0.7 -0.34 0.41
pH -0.22 0.25 -0.3 -0.28 0.19 -0.17 -0.26 0.021 1 0.25 0.11 0.037 0.34
sulphates 0.26 0.27 0.045 -0.21 0.39 -0.22 -0.3 0.26 0.25 1 0.0046 0.047 0.5
alcohol -0.1 -0.054 0.025 -0.36 -0.34 -0.18 -0.26 -0.7 0.11 0.0046 1 0.48 -0.039
quality -0.094 -0.25 0.095 -0.051 -0.27 0.058 -0.049 -0.34 0.037 0.047 0.48 1 -0.12
red 0.47 0.67 -0.16 -0.37 0.76 -0.5 -0.72 0.41 0.34 0.5 -0.039 -0.12 1
fixed acidity
volatile acidity
citric acid
residual sugar
chlorides
free sulfur dioxide
total sulfur dioxide
density
pH
sulphates
alcohol
quality
red
0.8
0.4
0.0
−0.4
−0.8


chlorides
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0
1
red


total sulfur dioxide
1.0
0.5
0.0
−0.5
−1.0
−1.5
0
1
red


volatile acidity
1.5
1.0
0.5
0.0
−0.5
−1.0
0
1
red

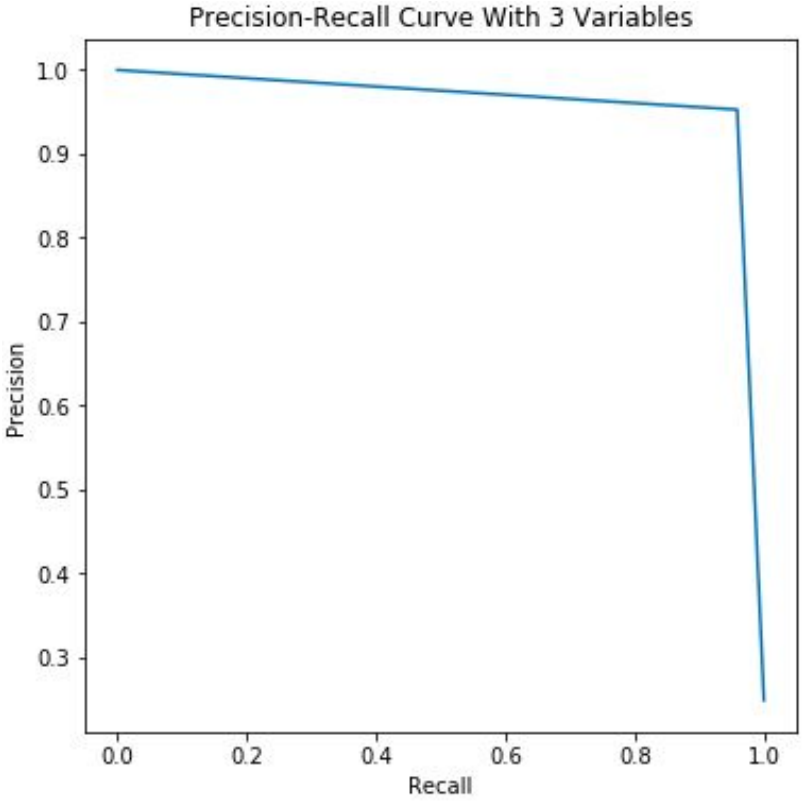
Linear Classification

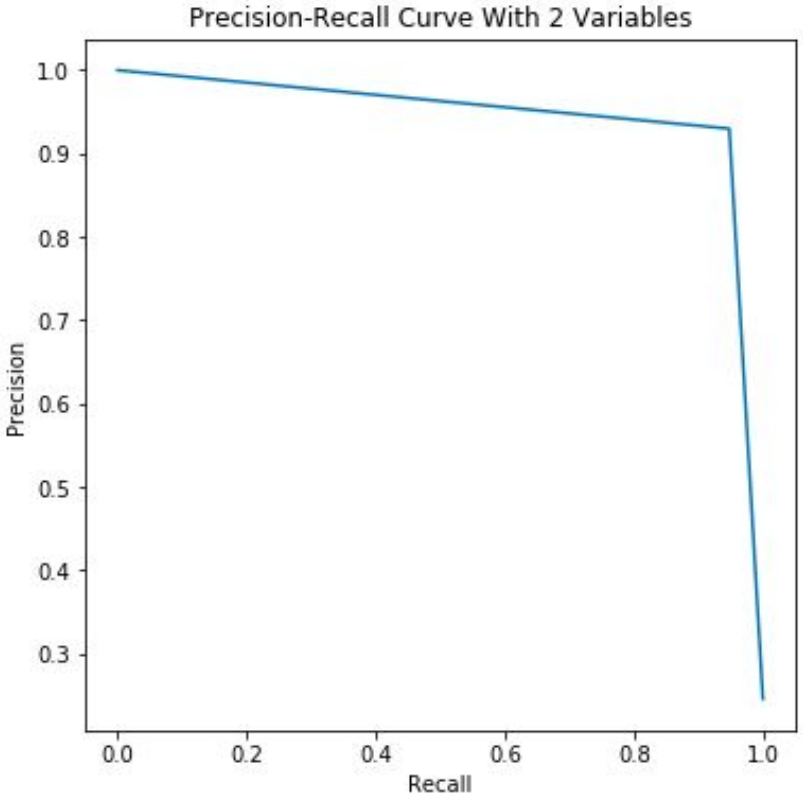
Using 3 Variables

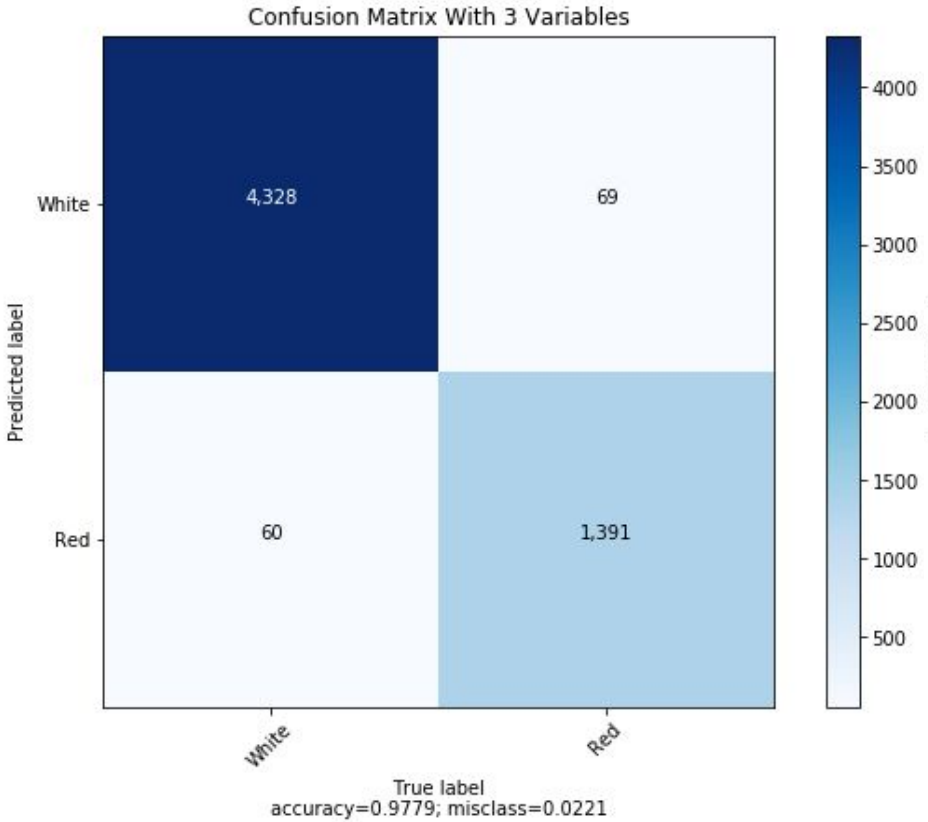
```
Score on training data:  0.987673343605547
Score on test data:  0.9779411764705882
Mean cross validatiion score: 0.986
```

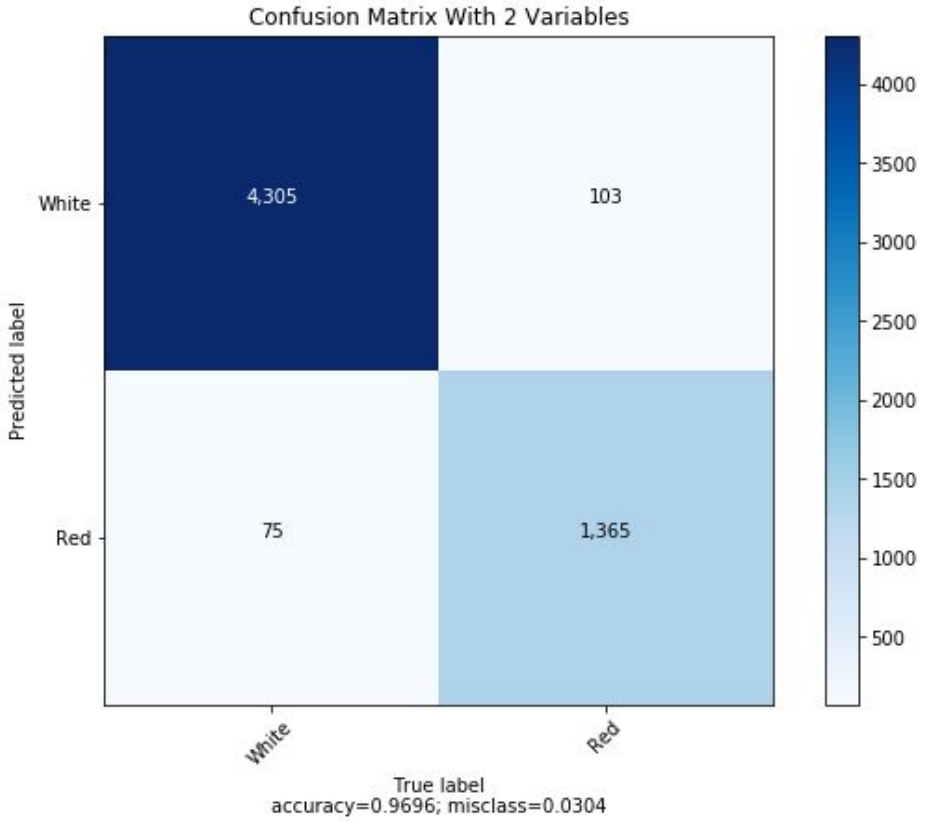
Using 2 Variables

```
Score on training data:  0.9599383667180277
Score on test data:  0.969562243502052
Mean cross validatiion score: 0.958
```


Precision-Recall Curve With 3 Variables
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0

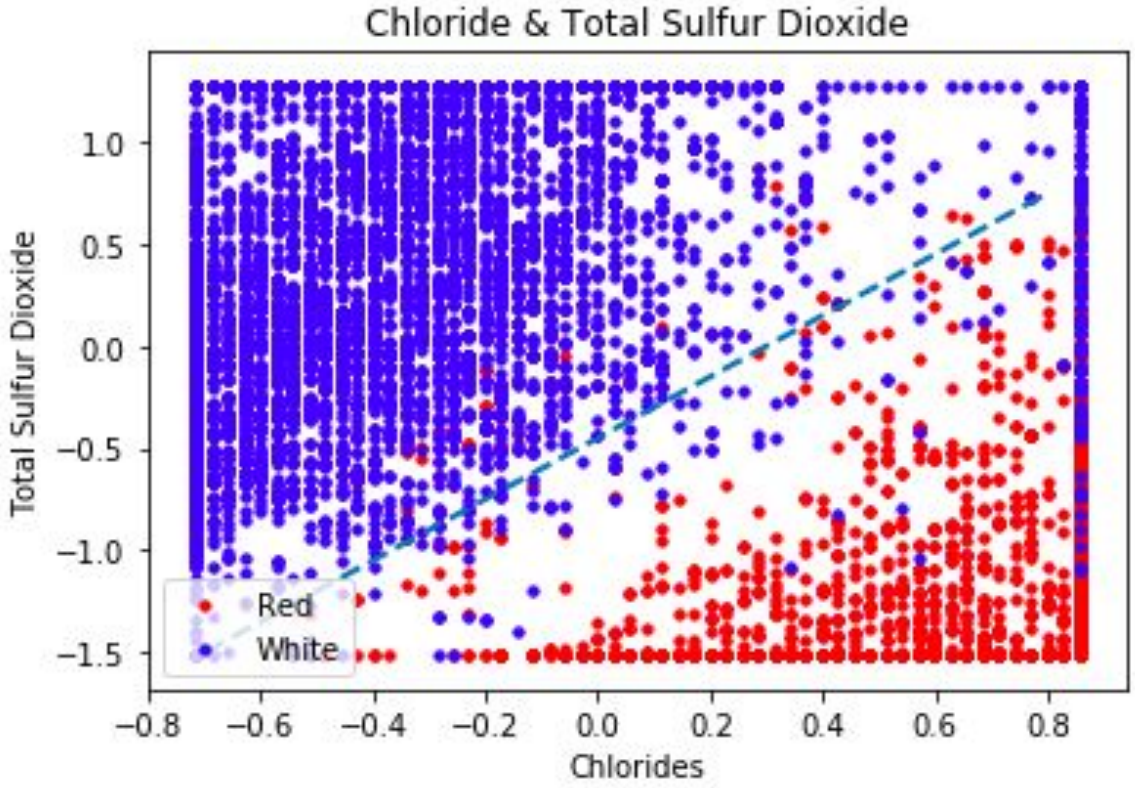

Precision-Recall Curve With 2 Variables
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0


Confusion Matrix With 3 Variables
White
Red
4,328
69
60
1,391
Predicted label
White
Red
True label
accuracy=0.9779; misclass=0.0221
4000
3500
3000
2500
2000
1500
1000
500


Confusion Matrix With 2 Variables
White
Red
4,305
103
75
1,365
Predicted label
White
Red
True label
accuracy=0.9696; misclass=0.0304
4000
3500
3000
2500
2000
1500
1000
500

It appears that the model is overfit when I use 3 variables, so I will continue the analysis by using 2 variables - chlorides and total sulfur dioxide.

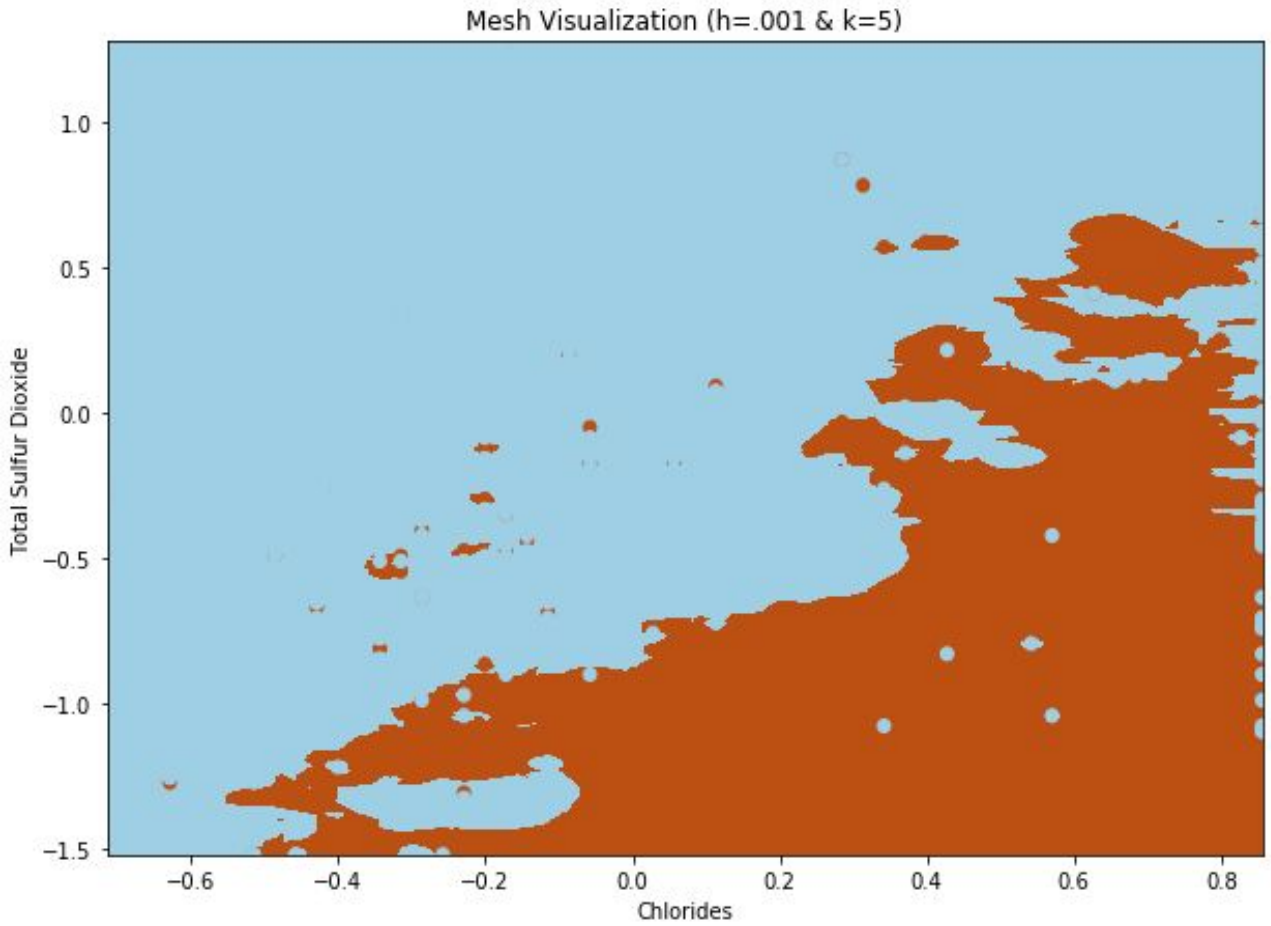
SVC

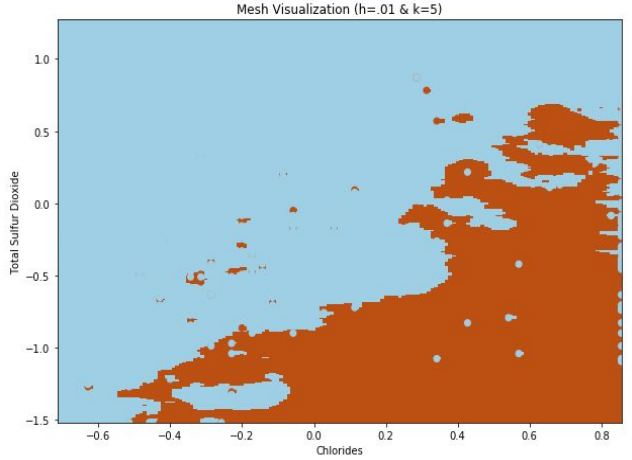

Chloride & Total Sulfur Dioxide
Total Sulfur Dioxide
1.0
0.5
0.0
−0.5
−1.0
−1.5
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
Chlorides
Red
White

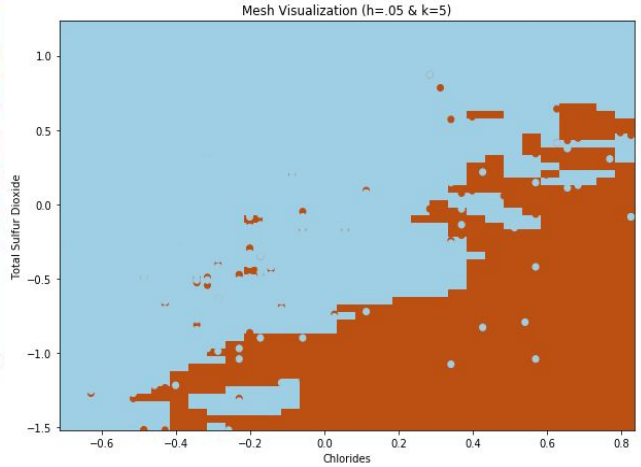
There doesn’t appear to be a margin that perfectly splits the data, so SVC might not be the best approach. What we do notice, however, is the similarity between points of the same class.

KNN Neighbors

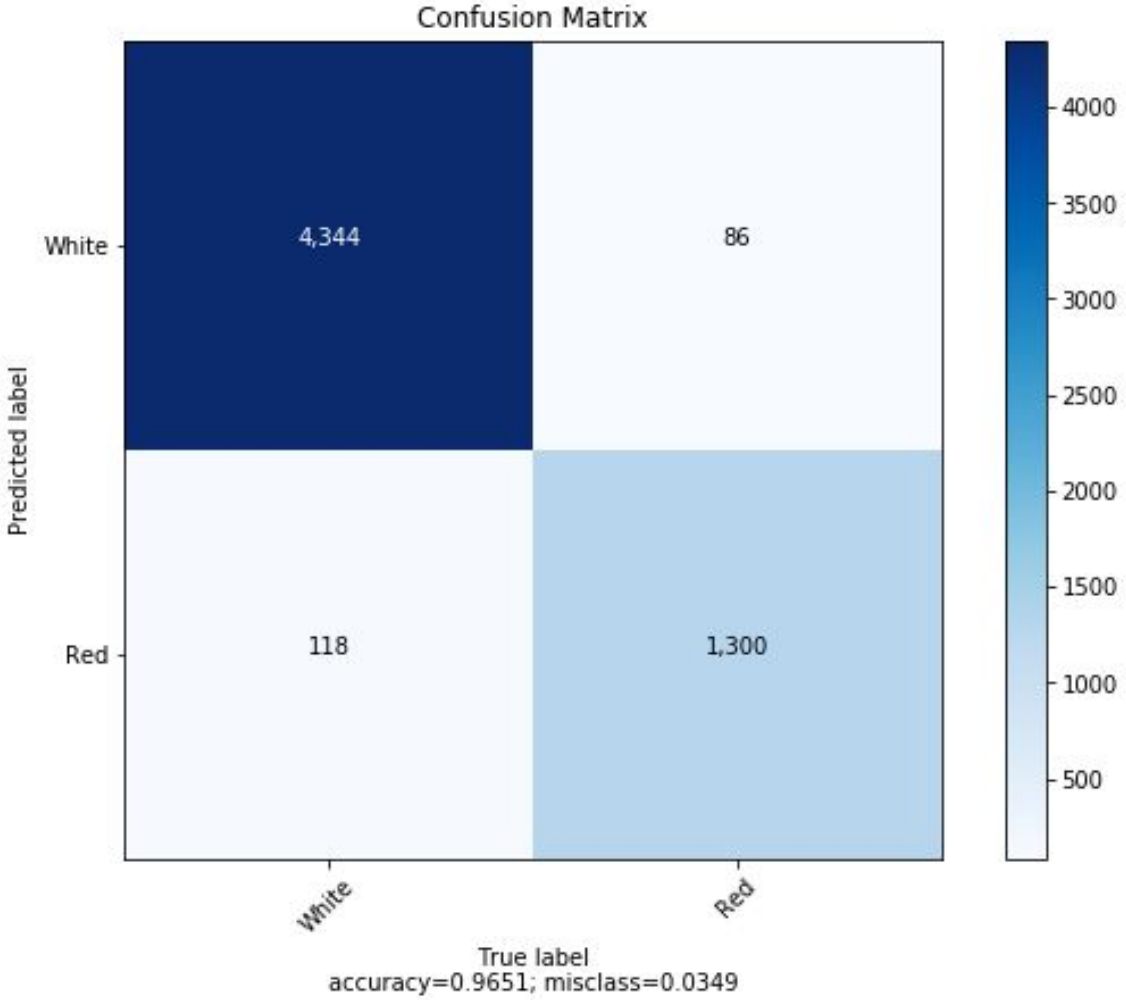
	No Weights	Distance
K =5	Score on training data: 0.9768875192604006 Score on test data: 0.9805061559507524 Mean cross validation score: 0.9687553739562977	Score on training data: 0.9892141756548536 Score on test data: 0.9904240766073872 Mean cross validation score: 0.9682940723633564
K=10	Score on training data: 0.9830508474576272 Score on test data: 0.9752051983584131 Mean cross validation score: 0.9687553739562977	Score on training data: 0.9953775038520801 Score on test data: 0.9897400820793434 Mean cross validation score: 0.9689095754130397

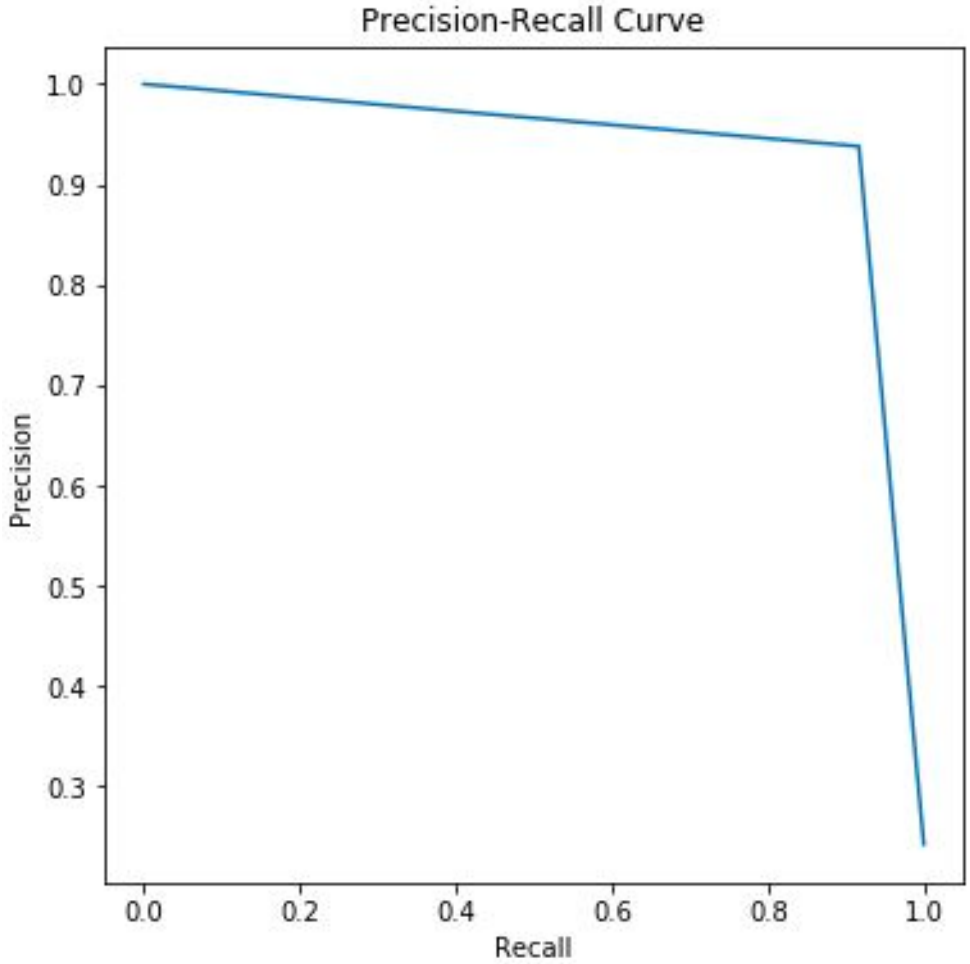

Mesh Visualization (h=.001 & k=5)
Total Sulfur Dioxide
Chlorides
1.0
0.5
0.0
−0.5
−1.0
−1.5
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8


Mesh Visualization (h=.01 & k=5)
Total Sulfur Dioxide
Chlorides

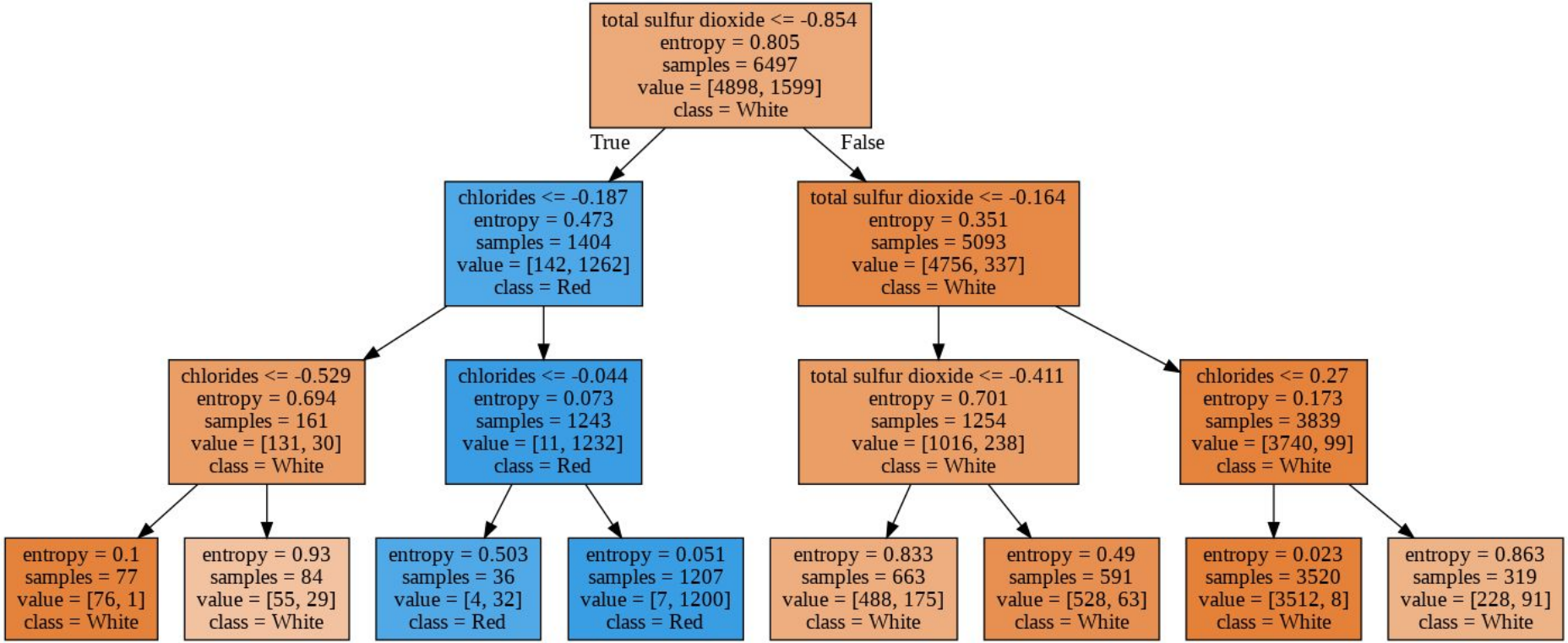

Mesh Visualization (h=.05 & k=5)
Total Sulfur Dioxide
Chlorides

Decreasing the step size makes the mesh smoother; however, the boundaries of the graph continue to be problematic due to the standardization of the data set. In general, however, it appears red wines have less sulfur dioxide and more chloride than white wine.


Confusion Matrix
White
Red
4,344
86
118
1,300
Predicted label
White
Red
True label
accuracy=0.9651; misclass=0.0349
4000
3500
3000
2500
2000
1500
1000
500


Precision-Recall Curve
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0

Tree Models


total sulfur dioxide <= -0.854
entropy = 0.805
samples = 6497
value = [4898, 1599]
class = White
True
False
chlorides <= -0.187
entropy = 0.473
samples = 1404
value = [142, 1262]
class = Red
total sulfur dioxide <= -0.164
entropy = 0.351
samples = 5093
value = [4756, 337]
class = White
chlorides <= -0.529
entropy = 0.694
samples = 161
value = [131, 30]
class = White
chlorides <= -0.044
entropy = 0.073
samples = 1243
value = [11, 1232]
class = Red
total sulfur dioxide <= -0.411
entropy = 0.701
samples = 1254
value = [1016, 238]
class = White
chlorides <= 0.27
entropy = 0.173
samples = 3839
value = [3740, 99]
class = White
entropy = 0.1
samples = 77
value = [76, 1]
class = White
entropy = 0.93
samples = 84
value = [55, 29]
class = White
entropy = 0.503
samples = 36
value = [4, 32]
class = Red
entropy = 0.051
samples = 1207
value = [7, 1200]
class = Red
entropy = 0.833
samples = 663
value = [488, 175]
class = White
entropy = 0.49
samples = 591
value = [528, 63]
class = White
entropy = 0.023
samples = 3520
value = [3512, 8]
class = White
entropy = 0.863
samples = 319
value = [228, 91]
class = White

Notice the criteria at each node is either related to chlorides or total sulfur dioxide. This reaffirms that these variables are preferred to the Decision Tree Model, just as they were in our previous models.

Decision Tree

Depth = 2

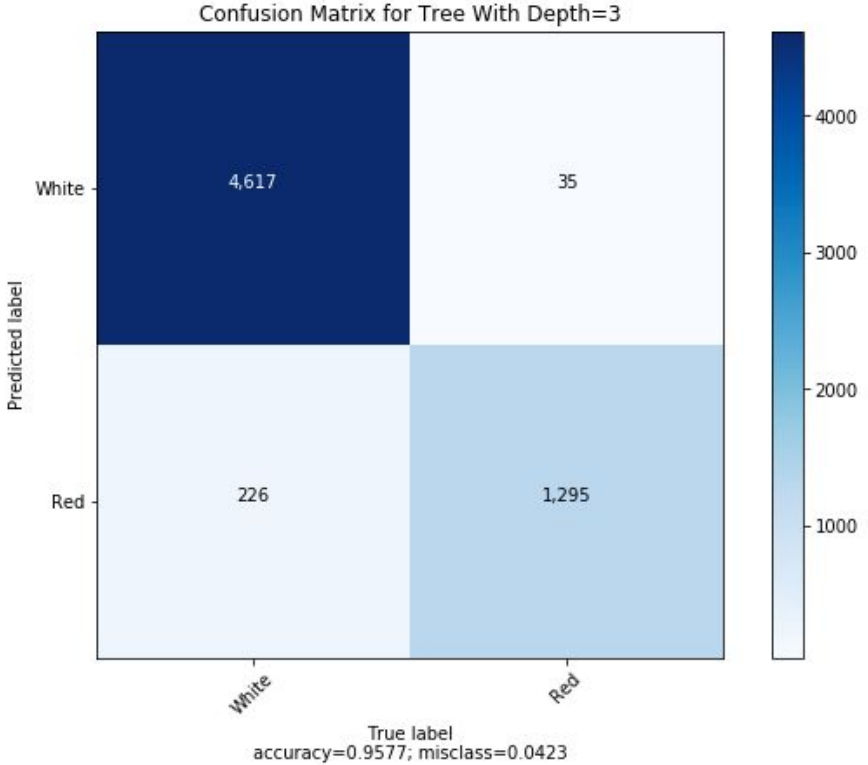
```
Score on training data:  0.9567901234567902
Score on test data:  0.9562611372104325
Mean cross validation score: 0.940278912773139
```

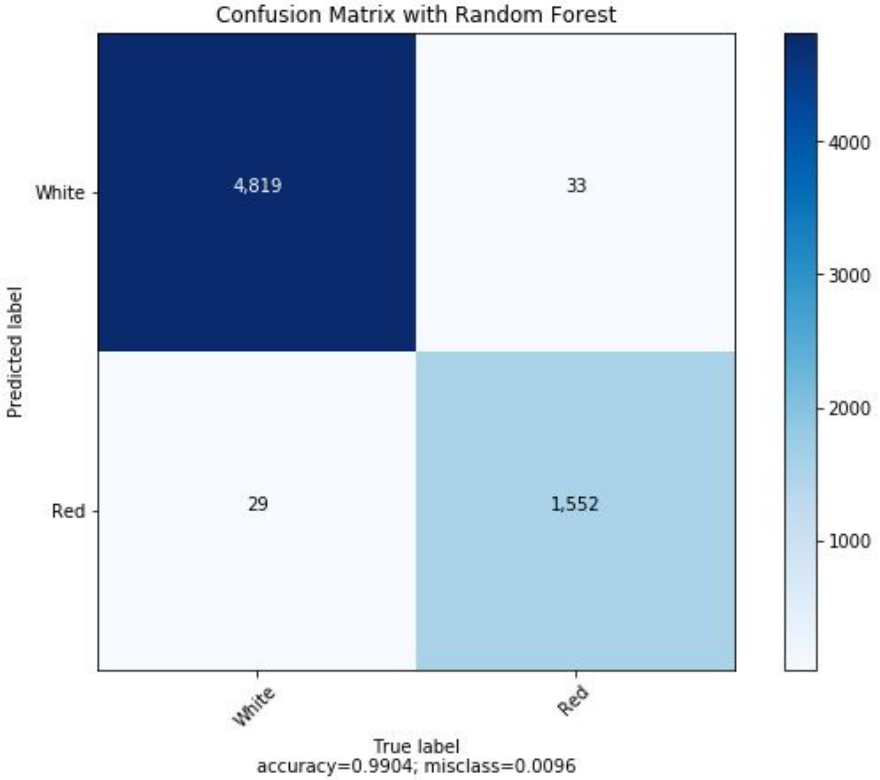
Depth = 3

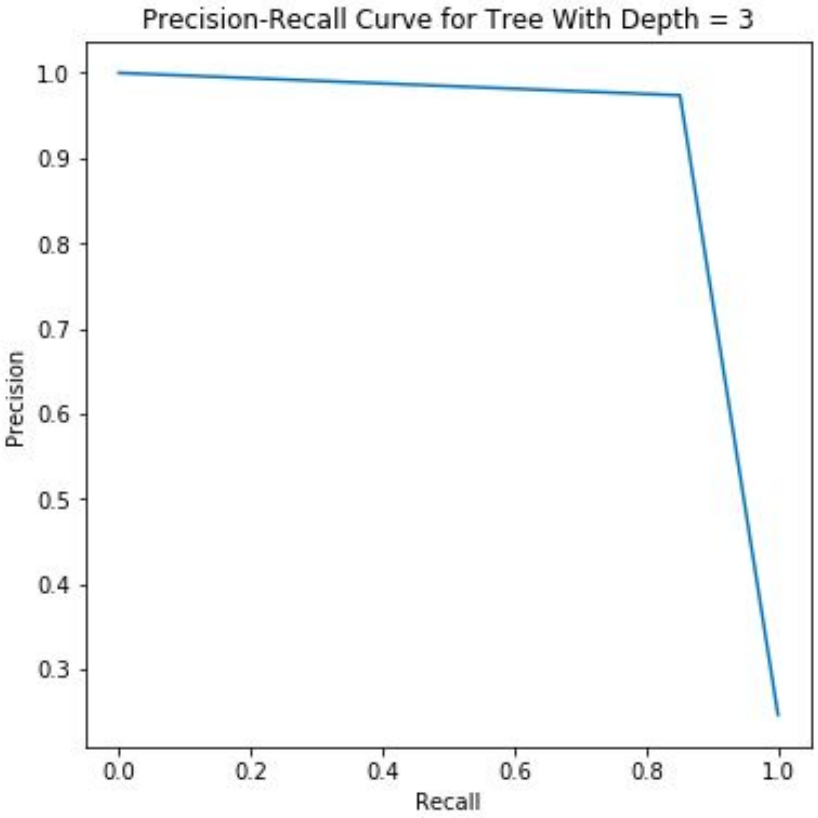
```
Score on training data:  0.9444444444444444
Score on test data:  0.9577190993034181
Mean cross validation score: 0.9595202226564814
```

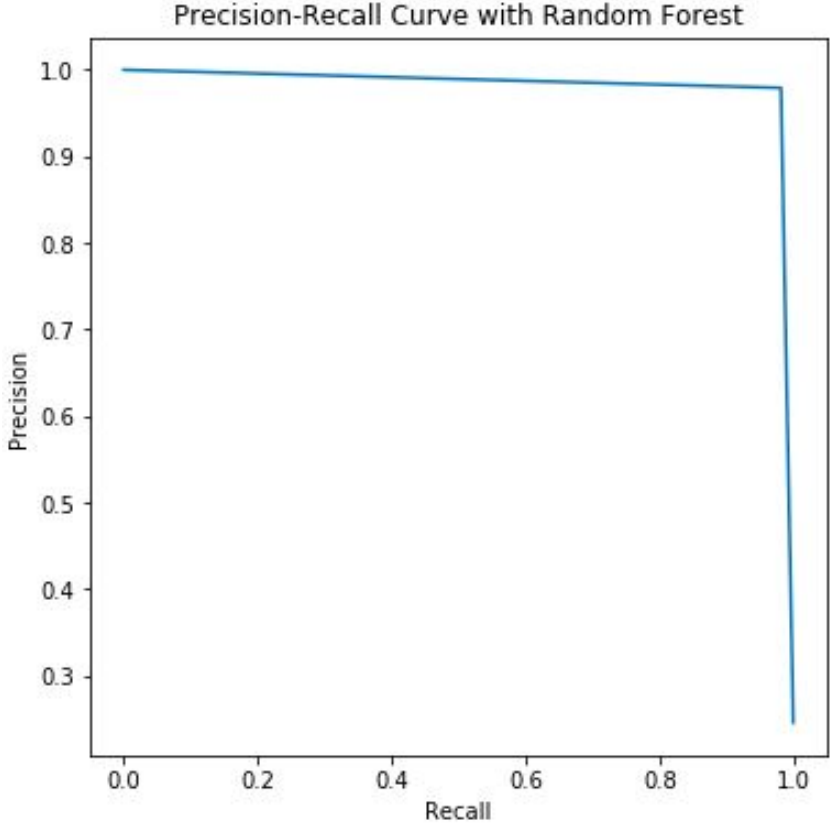
Random Forest

```
Score on training data:  0.984375
Score on test data:  0.9902067464635473
Mean cross validation score: 0.9682934801918636
```


Confusion Matrix for Tree With Depth=3
White
Red
4,617
35
226
1,295
Predicted label
White
Red
True label
accuracy=0.9577; misclass=0.0423
4000
3000
2000
1000


Confusion Matrix with Random Forest
White
Red
4,819
33
29
1,552
Predicted label
White
Red
True label
accuracy=0.9904; misclass=0.0096
4000
3000
2000
1000


Precision-Recall Curve for Tree With Depth = 3
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0


Precision-Recall Curve with Random Forest
Precision
Recall
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0

The Random Forest Model seems to perfect the Decision Tree Model through repetition without costing too much time.

Conclusion

- At a chemical level, the differences between red wine & white wine are very clear to a machine, thus classifications are extremely effective in nearly every model
    - Chlorides and Sulfur Dioxide appear to be the most influential variables in this classification problem
- The Random Forest Model performed at the highest level of accuracy without being overfit to the data