

# **Final Group Project: Full Data Mining Project Analysis on Vinho Verde Wine from Portugal**

ISM 6136 Data Mining  
Dr. Kiran Garimella  
May 2<sup>nd</sup>, 2021  
University of South Florida

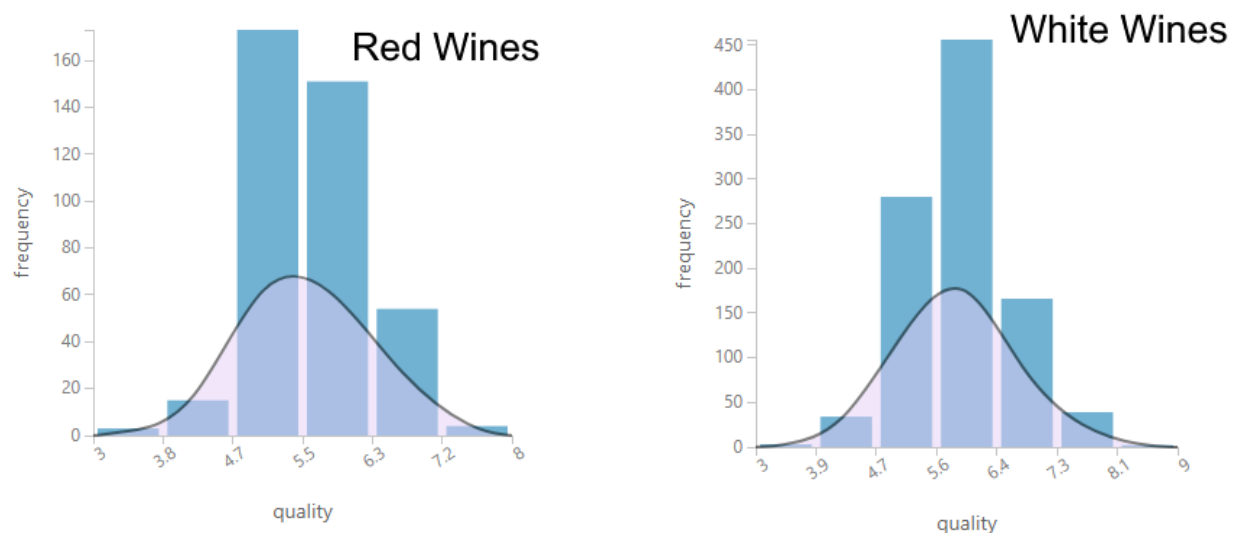
## **Final Project Group 1:**

Ethan Anderson  
William Collins  
Sujhan Das  
Raunak Ghosh  
Arpita Perugu  
Zachary Smith  
Jeffrey West

## **Background:**

Portugal is famous for several things including seafood, beaches, and soccer legends, but is known globally in the culinary world for Port wine from the Douro Valley region. However, Portugal has vineyards across most of the countryside, and produces a vast variety of wines from the sweet, fortified Port to light-bodied Bucelas, and the fruit-driven Tinto Beira. With 31 distinct “Denominação de Origem Controlada,” also called a DOC or Controlled Denomination of Origin, Portugal has the infrastructure to be a wine-producing powerhouse. Why then is the country not as renowned for its wine as countries such as France, Spain, or Italy? According to winefolly.com, each DOC means “the wine comes from a strictly defined geographical area with recommended and permitted grapes and maximum wine yields to control quality,” much like the Champagne region of France. Additionally, Portuguese wine expert Rui Falcao mentions that Portugal has more grape varieties than any other country; with over three hundred, the grapes are still pressed by foot instead of using presses, and the climate and soil mean Portuguese wines are generally more acidic and taste fresher than more basic wines. This analysis will take an in-depth study of how Portugal can increase market shares of the global wine market by using predictors to determine what attributes generally make a wine rated higher, as well as a decision tree and forest to predict the rating of a wine based on the different attributes of the wine. The wine that will be used is specifically the red and white variants of the Vinho Verde wine, a wine from the Minho region of Portugal, in the northwest of the country.

## **Motivation:**



Unfortunately, when rated on a scale from one to ten by critics, the average Portuguese white Vinho Verde wine sits rated at about a 5.9 out of 980 samples, with two of the samples rated at a 9.0. The average Portuguese red Vinho Verde wine sits rated at 5.6 out of 320 samples, with the top four rated wines only sitting at an 8.0. Currently, Portugal holds about a 3% market share of the global wine market, while France and Italy combined hold nearly fifty percent of the entire market value. The purpose of the experimentation is to attempt to not only predict the quality rating of any particular bottle of Vinho Verde based on the specific physiochemical properties of the wine, but also give the ability to adjust wine production in a manner to better fit the highest rated wines. If more acidic Vinho Verde wines are generally rated higher, then it's safe to make the attempt to reduce the pH in the wines to bring them to















a more desirable acidity. If a higher alcohol content is preferred, make the attempt to try and increase the proof of the wine to better align with the higher rated bottles of both red and white Vinho Verde.















### **Solution Methodology and Metrics:**

In order to predict the quality of wine, and therefore evaluate the metrics that make for a better Vinho Verde, both red and white variants, a pair of datasets were pulled from the UCI Machine Learning Repository, which is meant to model wine qualities based on physicochemical tests. To study the quality of the samples compared with the chemical compositions of all samples in both datasets, a series of six experiments were made: three for the red Vinho Verde, and three for the white Vinho Verde. Each dataset utilized a two-class boosted decision tree and a two-class decision forest passed through a one-vs-all multiclass, and a multiclass decision forest for classification and prediction, as well as a logistic regression for prediction. To create these classifications and the regression, we trained our data in an 80%/20% split, and utilized random number seed 1369. The main comparisons that will be shown are quality versus total sulfur dioxide, alcohol percentage, sulphate, and pH. In the decision forests, there are also predictive models to estimate the quality rating of any wine. The overall and average accuracy for each experiment will be added into the comparison of experiments, and the confusion matrices will be added into the experiment summary sheet and conclusions.

### **Dataset Description:**

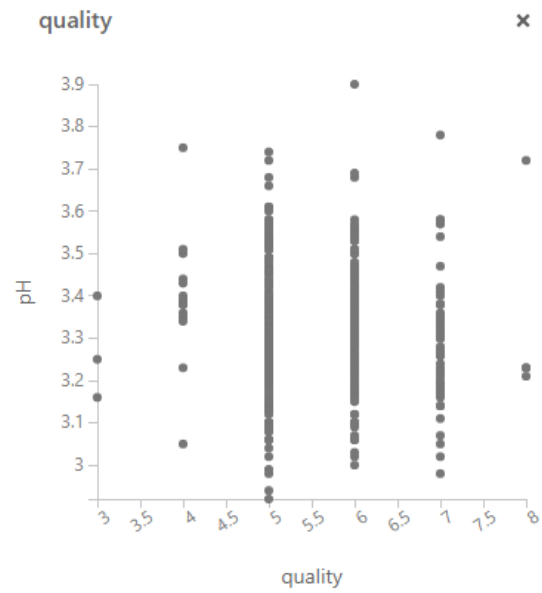
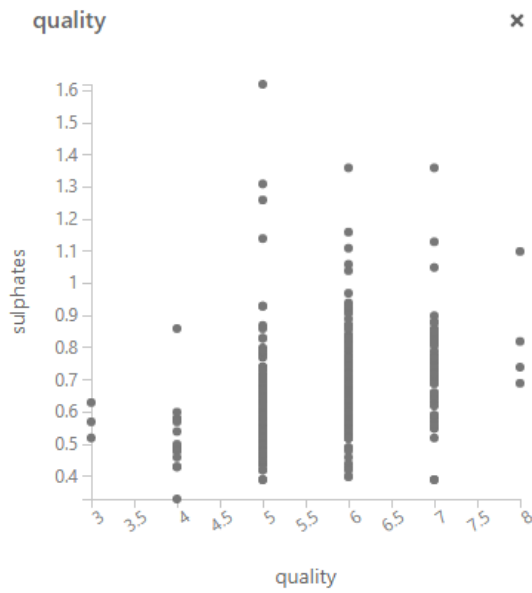
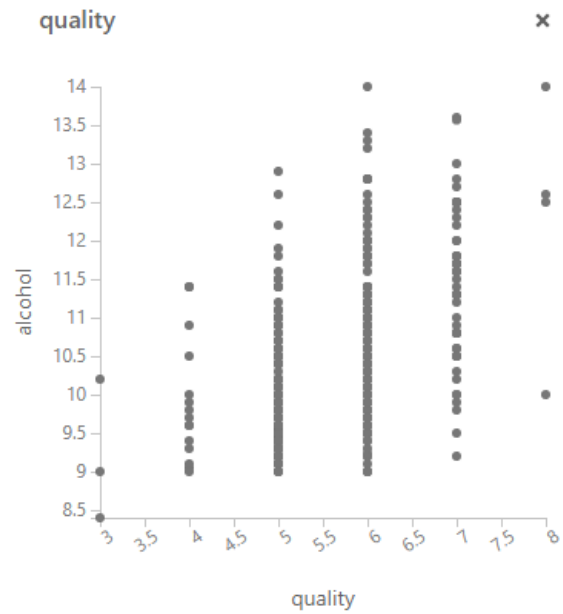
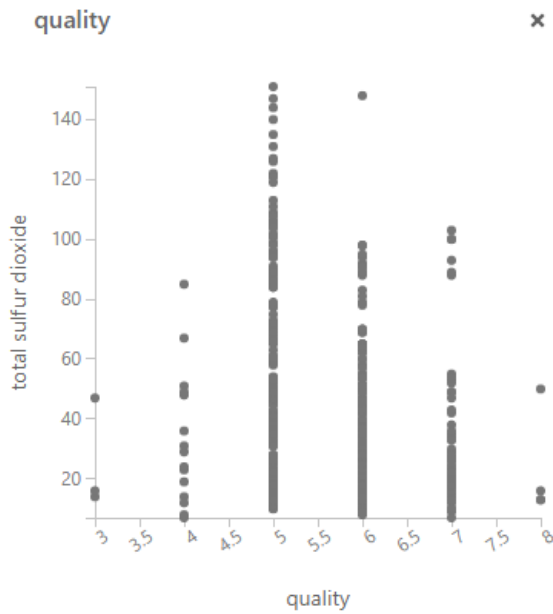
The two data sets used in this project were sourced from the UCI Machine Learning Repository but originally recorded from a vineyard in Portugal where a series of physicochemical tests were performed on the Vinho Verde brand of red and white wines. These physicochemical tests recorded continuous data through objective testing (such as pH level) on 11 different variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol content. In addition, a 12<sup>th</sup> variable of an ordinal data type was recorded that ranked the quality of each wine sample on a scale of 1-10 based off wine critic's ratings. Although the quality variable was kept in its original format for most of the experiments that were performed in this project, quality was recoded as a categorical binary variable for the multiclass logistic regression test using the "Edit Metadata" and "Group Categorical Values" module in Microsoft's Azure Machine Learning Studio. In the logistic regression test, values of 1-5 were recoded as 'Poor Quality' wine while values of 6-10 were recoded as 'Good Quality' wine. In total the data set for white wine contains approximately 4,800 hundred rows of data and the red wine data set contains approximately 1,600 rows of data. These data sets are ordered but not balanced by class, where average quality wines outnumber those of excellent or bad quality wines.

| Red Wine Dataset  |   |   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature   | Count   | Unique Value Count  | Missing Value Count   | Min   | Max   | Mean  | Mean Deviation  | 1st Quartile  | Median  | 3rd Quartile  | Mode  | Range   | Sample Variance   |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| fixed acidity   | 1599  | 96  | 0   | 4.6   | 15.9  | 8.319637  | 1.360136  | 7.1   | 7.9   | 9.2   | 7.2   | 11.3  | 3.031416  |
| volatile acidity  | 1599  | 143   | 0   | 0.12  | 1.58  | 0.527821  | 0.142391  | 0.39  | 0.52  | 0.64  | 0.6   | 1.46  | 0.032062  |
| citric acid   | 1599  | 80  | 0   | 0   | 1   | 0.270976  | 0.164654  | 0.09  | 0.26  | 0.42  | 0   | 1   | 0.037947  |
| residual sugar  | 1599  | 91  | 0   | 0.9   | 15.5  | 2.538806  | 0.764065  | 1.9   | 2.2   | 2.6   | 2   | 14.6  | 1.987897  |
| chlorides   | 1599  | 153   | 0   | 0.012   | 0.611   | 0.087467  | 0.021773  | 0.07  | 0.079   | 0.09  | 0.08  | 0.599   | 0.002215  |
| free sulfur dioxide   | 1599  | 60  | 0   | 1   | 72  | 15.874922   | 8.187527  | 7   | 14  | 21  | 6   | 71  | 109.41488   |
| total sulfur dioxide  | 1599  | 144   | 0   | 6   | 289   | 46.467792   | 25.354053   | 22  | 38  | 62  | 28  | 283   | 1082.1023   |
| density   | 1599  | 436   | 0   | 0.99007   | 1.00369   | 0.996747  | 0.001433  | 0.9956  | 0.99675   | 0.997835  | 0.9972  | 0.01362   | 0.000004  |
| pH  | 1599  | 89  | 0   | 2.74  | 4.01  | 3.311113  | 0.119769  | 3.21  | 3.31  | 3.4   | 3.3   | 1.27  | 0.023835  |
| sulphates   | 1599  | 96  | 0   | 0.33  | 2   | 0.658149  | 0.119094  | 0.55  | 0.62  | 0.73  | 0.6   | 1.67  | 0.028733  |
| alcohol   | 1599  | 65  | 0   | 8.4   | 14.9  | 10.422983   | 0.877969  | 9.5   | 10.2  | 11.1  | 9.5   | 6.5   | 1.135647  |
| quality   | 1599  | 6   | 0   | 3   | 8   | 5.636023  | 0.683178  | 5   | 6   | 6   | 5   | 5   | 0.652168  |

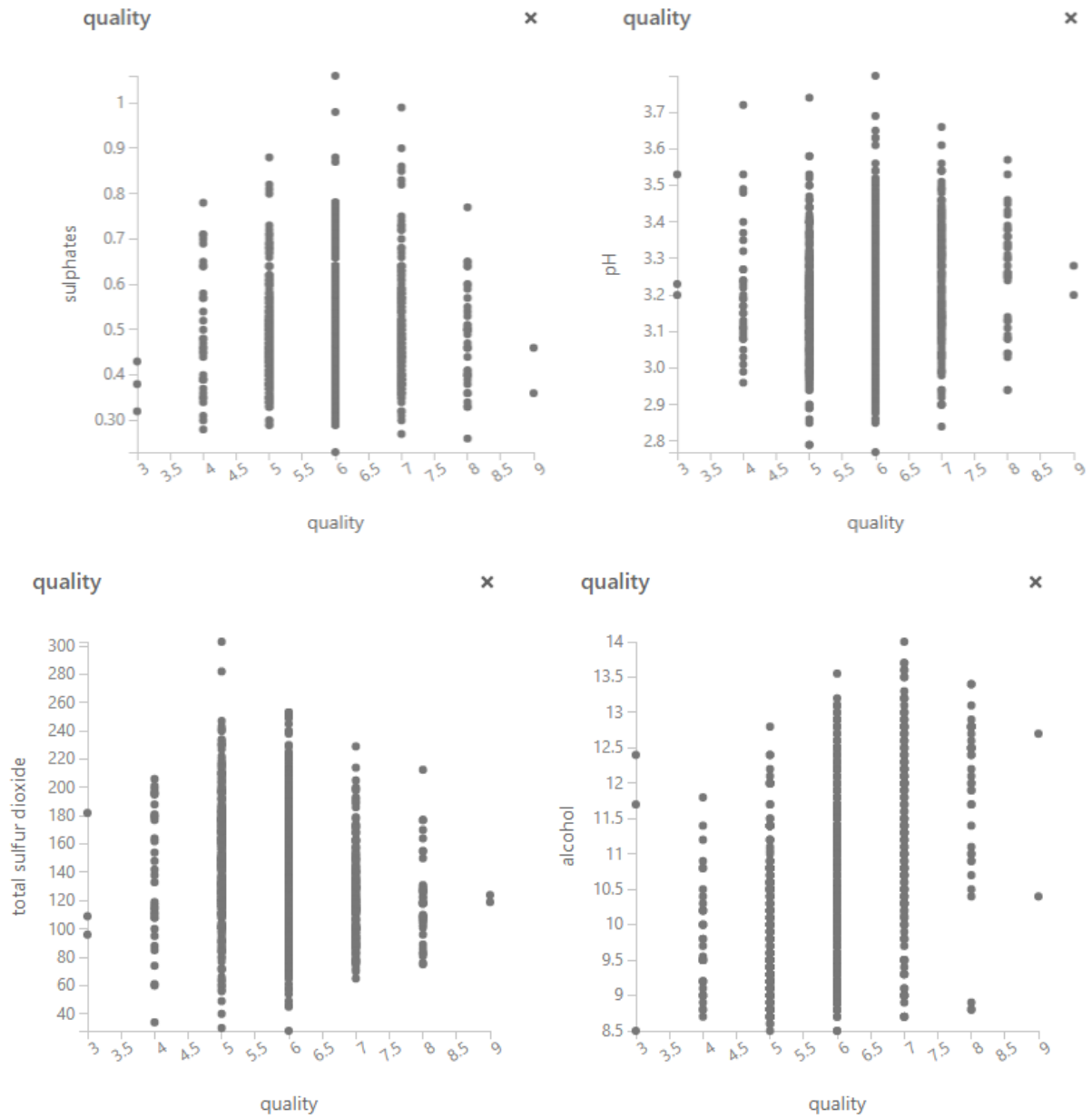
| White Wine Dataset  |   |   |   |   |   |   |   |   |   |   |   |   |   |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Feature   | Count   | Unique Value Count  | Missing Value Count   | Min   | Max   | Mean  | Mean Deviation  | 1st Quartile  | Median  | 3rd Quartile  | Mode  | Range   | Sample Variance   |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| fixed acidity   | 4898  | 68  | 0   | 3.8   | 14.2  | 6.854788  | 0.644923  | 6.3   | 6.8   | 7.3   | 6.8   | 10.4  | 0.712114  |
| volatile acidity  | 4898  | 125   | 0   | 0.08  | 1.1   | 0.278241  | 0.073803  | 0.21  | 0.26  | 0.32  | 0.28  | 1.02  | 0.01016   |
| citric acid   | 4898  | 87  | 0   | 0   | 1.66  | 0.334192  | 0.086411  | 0.27  | 0.32  | 0.39  | 0.3   | 1.66  | 0.014646  |
| residual sugar  | 4898  | 310   | 0   | 0.6   | 65.8  | 6.391415  | 4.227788  | 1.7   | 5.2   | 9.9   | 1.2   | 65.2  | 25.72577  |
| chlorides   | 4898  | 160   | 0   | 0.009   | 0.346   | 0.045772  | 0.011462  | 0.036   | 0.043   | 0.05  | 0.044   | 0.337   | 0.000477  |
| free sulfur dioxide   | 4898  | 132   | 0   | 2   | 289   | 35.308085   | 13.149156   | 23  | 34  | 46  | 29  | 287   | 289.24272   |
| total sulfur dioxide  | 4898  | 251   | 0   | 9   | 440   | 138.360657  | 34.282442   | 108   | 134   | 167   | 111   | 431   | 1806.0854   |
| density   | 4898  | 890   | 0   | 0.98711   | 1.03898   | 0.994027  | 0.002447  | 0.991723  | 0.99374   | 0.9961  | 0.992   | 0.05187   | 0.000009  |
| pH  | 4898  | 103   | 0   | 2.72  | 3.82  | 3.188267  | 0.118204  | 3.09  | 3.18  | 3.28  | 3.14  | 1.1   | 0.022801  |
| sulphates   | 4898  | 79  | 0   | 0.22  | 1.08  | 0.489847  | 0.087504  | 0.41  | 0.47  | 0.55  | 0.5   | 0.86  | 0.013025  |
| alcohol   | 4898  | 103   | 0   | 8   | 14.2  | 10.514267   | 1.034213  | 9.5   | 10.4  | 11.4  | 9.4   | 6.2   | 1.514427  |
| quality   | 4898  | 7   | 0   | 3   | 9   | 5.877909  | 0.670793  | 5   | 6   | 6   | 6   | 6   | 0.784356  |

## Comparison of Experiments:

### Red Wines



# White Wines



## Evaluation Results (Red Wine):

### 1) Two Class Decision Forest

Red Wine > Evaluate Model > Evaluation results

#### Metrics

|                          |          |
|--------------------------|----------|
| Overall accuracy         | 0.634375 |
| Average accuracy         | 0.870125 |
| Micro-averaged precision | 0.634375 |
| Macro-averaged precision | 0.38102  |
| Micro-averaged recall    | 0.634375 |
| Macro-averaged recall    | 0.318642 |

Two Class Decision Forest

#### Confusion Matrix

|              |   | Predicted Class |      |       |       |       |
|--------------|---|-----------------|------|-------|-------|-------|
|              |   | 0               | 1    | 2     | 3     | 4     |
| Actual Class | 0 |                 |      | 50.0% | 50.0% |       |
|              | 1 | 7.1%            | 7.1% | 57.1% | 28.6% |       |
|              | 2 |                 | 1.5% | 72.1% | 24.3% | 2.2%  |
|              | 3 |                 |      | 24.0% | 71.1% | 4.1%  |
|              | 4 |                 |      | 6.8%  | 50.0% | 40.9% |

Red Wine > Evaluate Model > Evaluation results

#### Metrics

|                          |          |
|--------------------------|----------|
| Overall accuracy         | 0.64375  |
| Average accuracy         | 0.88125  |
| Micro-averaged precision | 0.64375  |
| Macro-averaged precision | 0.337675 |
| Micro-averaged recall    | 0.64375  |
| Macro-averaged recall    | 0.316764 |

Two Class Decision Forest (2)

#### Confusion Matrix

|              |   | Predicted Class |      |       |       |       |
|--------------|---|-----------------|------|-------|-------|-------|
|              |   | 0               | 1    | 2     | 3     | 4     |
| Actual Class | 0 |                 |      | 50.0% | 50.0% |       |
|              | 1 | 7.1%            |      | 57.1% | 35.7% |       |
|              | 2 |                 | 0.7% | 73.5% | 25.0% | 0.7%  |
|              | 3 |                 |      | 24.8% | 71.1% | 4.1%  |
|              | 4 |                 |      | 6.8%  | 45.5% | 45.5% |

After changing the number of trees from 8 to 12, there is a slight change of metrics in the results of the model of the Two Class Decision Forest. The overall accuracy changes from 0.634 to 0.643. The average accuracy also changes from 0.878 to 0.881. From the confusion matrix, we can see that there is an increase in the true positives.

### 2) Two Class Boosted Decision Tree

Red Wine > Evaluate Model > Evaluation results

#### Metrics

|                          |          |
|--------------------------|----------|
| Overall accuracy         | 0.653125 |
| Average accuracy         | 0.884375 |
| Micro-averaged precision | 0.653125 |
| Macro-averaged precision | 0.336817 |
| Micro-averaged recall    | 0.653125 |
| Macro-averaged recall    | 0.325869 |

Two Class Boosted Decision Tree

#### Confusion Matrix

|              |   | Predicted Class |      |       |       |       |
|--------------|---|-----------------|------|-------|-------|-------|
|              |   | 0               | 1    | 2     | 3     | 4     |
| Actual Class | 0 |                 |      | 50.0% | 50.0% |       |
|              | 1 | 7.1%            |      | 64.3% | 28.6% |       |
|              | 2 |                 | 0.7% | 72.8% | 25.0% | 1.5%  |
|              | 3 |                 | 1.7% | 19.0% | 72.7% | 5.8%  |
|              | 4 |                 |      | 4.5%  | 43.2% | 50.0% |

Red Wine > Evaluate Model > Evaluation results

#### Metrics

|                          |          |
|--------------------------|----------|
| Overall accuracy         | 0.646875 |
| Average accuracy         | 0.882292 |
| Micro-averaged precision | 0.646875 |
| Macro-averaged precision | 0.327858 |
| Micro-averaged recall    | 0.646875 |
| Macro-averaged recall    | 0.326132 |

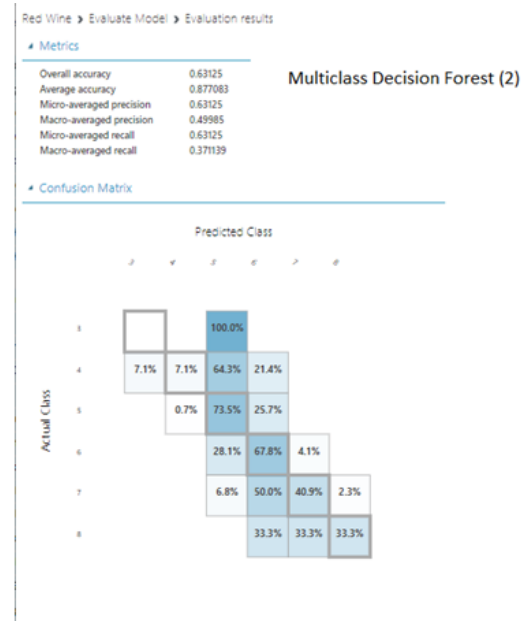
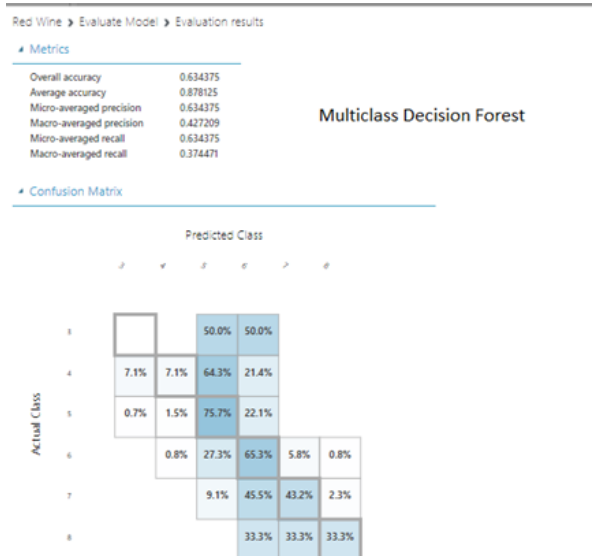
Two Class Boosted Decision Tree (2)

#### Confusion Matrix

|              |   | Predicted Class |      |       |       |       |
|--------------|---|-----------------|------|-------|-------|-------|
|              |   | 0               | 1    | 2     | 3     | 4     |
| Actual Class | 0 |                 |      | 50.0% | 50.0% |       |
|              | 1 | 14.3%           |      | 50.0% | 35.7% |       |
|              | 2 |                 | 1.7% | 69.9% | 27.9% | 1.5%  |
|              | 3 |                 |      | 1.7%  | 73.6% | 8.3%  |
|              | 4 |                 |      | 4.5%  | 40.9% | 52.3% |

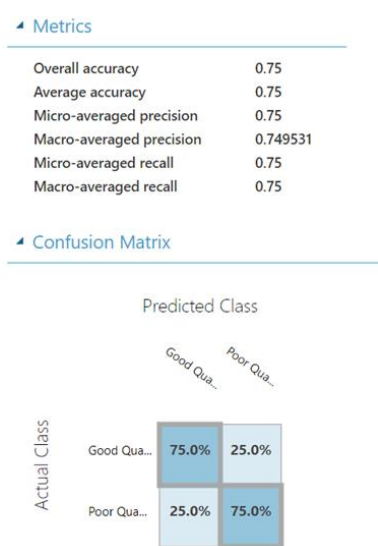
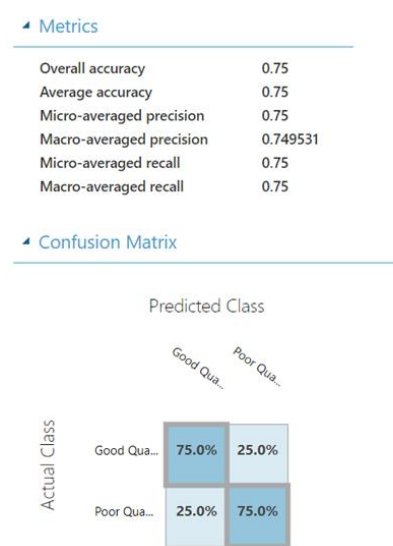
After changing the number of trees from 100 to 200 and the number of leaves from 20 to 10, there is a slight change of metrics in the results of the model of the Two Class Boosted Decision Tree. The overall accuracy changes from 0.653 to 0.646. The average accuracy also changes from 0.653 to 0.646.

### 3) Multiclass Decision Forest



After changing the number of trees from 8 to 12, there is a slight change of metrics in the results of the model of the Multiclass Decision Forest. The overall accuracy changes from 0.634 to 0.631. The average accuracy also changes from 0.878 to 0.877.

### 4) Multiclass Logistic Regression

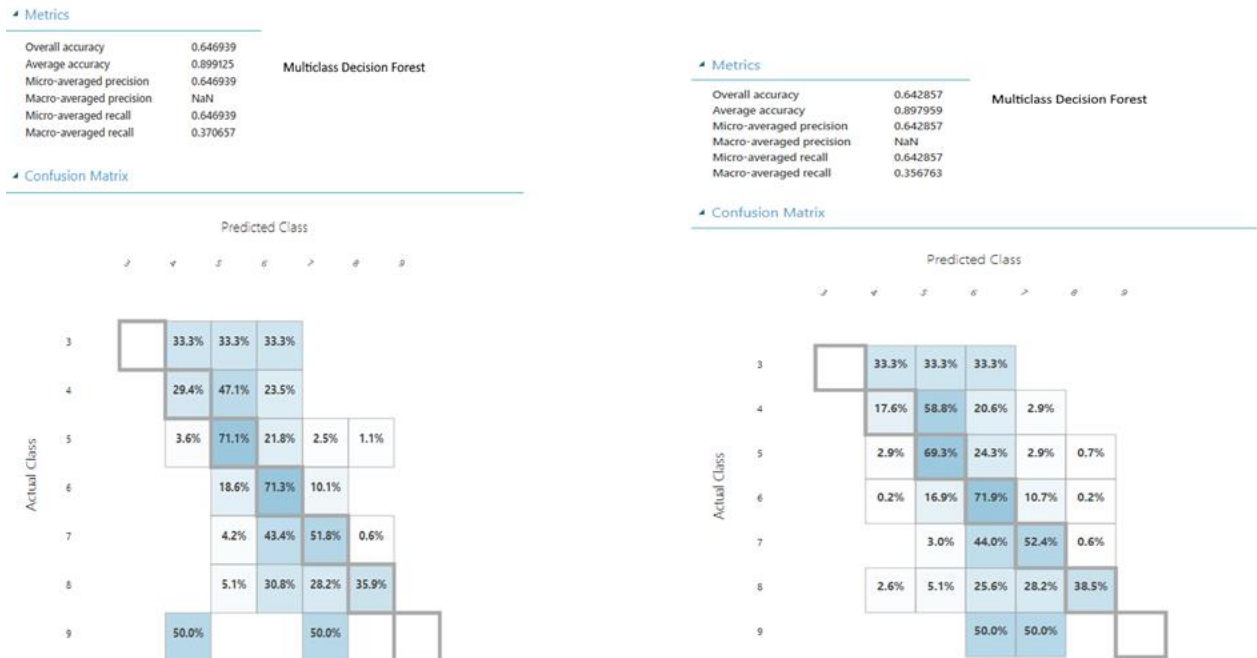




The results on the left include the 'pH' parameter and the right results do not include it. Since there is no change in accuracy, we can drop the 'pH' column to reduce the number of independent variables and still get an accurate result.

## Evaluation Results (White Wine):

### 1) Multiclass Decision Forest



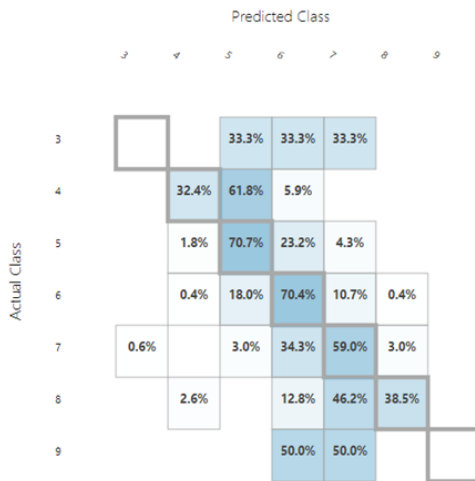
After changing the number of trees from 8 to 12, there is a slight change of metrics in the results of the model of the Multiclass Decision Forest. The overall accuracy changes from 0.646 to 0.642. The average accuracy also changes from 0.899 to 0.897.

## 2) Two Class Boosted Decision Tree

### Metrics

|                          |          |                                 |
|--------------------------|----------|---------------------------------|
| Overall accuracy         | 0.656122 | Two Class Boosted Decision Tree |
| Average accuracy         | 0.901749 |                                 |
| Micro-averaged precision | 0.656122 |                                 |
| Macro-averaged precision | NaN      |                                 |
| Micro-averaged recall    | 0.656122 |                                 |
| Macro-averaged recall    | 0.387085 |                                 |

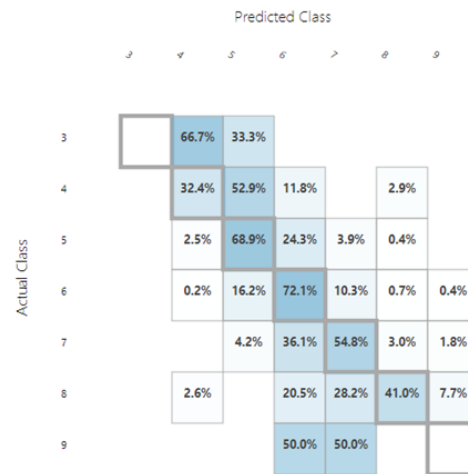
### Confusion Matrix



### Metrics

|                          |          |                                 |
|--------------------------|----------|---------------------------------|
| Overall accuracy         | 0.653061 | Two-Class Boosted Decision Tree |
| Average accuracy         | 0.900875 |                                 |
| Micro-averaged precision | 0.653061 |                                 |
| Macro-averaged precision | NaN      |                                 |
| Micro-averaged recall    | 0.653061 |                                 |
| Macro-averaged recall    | 0.384679 |                                 |

### Confusion Matrix



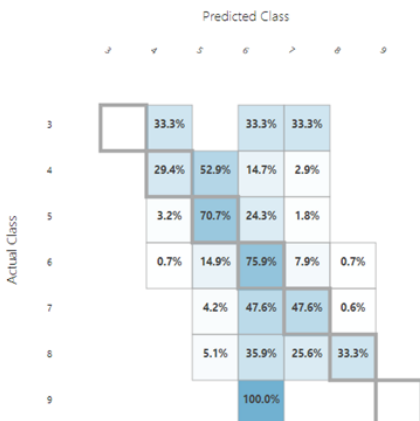
After changing the number of trees from 100 to 200 and the number of leaves from 20 to 10, there is a slight change of metrics in the results of the model of the Two Class Boosted Decision Tree. The overall accuracy changes from 0.653 to 0.646. The average accuracy also changes from 0.884 to 0.882. From the confusion matrix, we can see that there is an increase in the true positives for 2 out of 3 instances.

## 3) Two Class Decision Forest

### Metrics

|                          |          |                           |
|--------------------------|----------|---------------------------|
| Overall accuracy         | 0.659184 | Two Class Decision Forest |
| Average accuracy         | 0.902624 |                           |
| Micro-averaged precision | 0.659184 |                           |
| Macro-averaged precision | NaN      |                           |
| Micro-averaged recall    | 0.659184 |                           |
| Macro-averaged recall    | 0.367038 |                           |

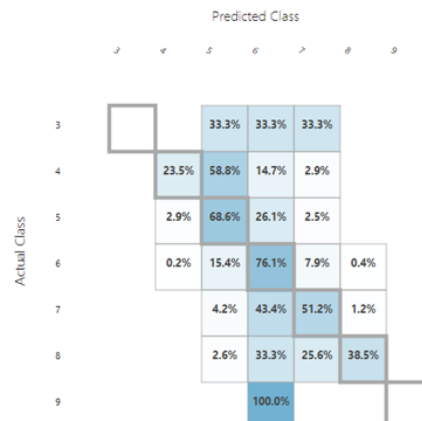
### Confusion Matrix



### Metrics

|                          |          |                           |
|--------------------------|----------|---------------------------|
| Overall accuracy         | 0.660204 | Two-Class Decision Forest |
| Average accuracy         | 0.902915 |                           |
| Micro-averaged precision | 0.660204 |                           |
| Macro-averaged precision | NaN      |                           |
| Micro-averaged recall    | 0.660204 |                           |
| Macro-averaged recall    | 0.368377 |                           |

### Confusion Matrix



After changing the number of trees from 8 to 12, there is a slight change of metrics in the results of the model of the Two Class Decision Forest. The overall accuracy changes from 0.659 to 0.66. The average accuracy remains constant. From the confusion matrix, we can see that there is an increase in the true positives in 3 out of 5 instances.

#### 4) Multiclass Logistic Regression

White Wine > Evaluate Model > Evaluation results

##### Metrics

|                          |          |
|--------------------------|----------|
| Overall accuracy         | 0.741837 |
| Average accuracy         | 0.741837 |
| Micro-averaged precision | 0.741837 |
| Macro-averaged precision | 0.703586 |
| Micro-averaged recall    | 0.741837 |
| Macro-averaged recall    | 0.690669 |

##### Confusion Matrix

|              |             | Predicted Class |             |
|--------------|-------------|-----------------|-------------|
|              |             | Good Qua...     | Poor Qua... |
| Actual Class | Good Qua... | 83.6%           | 16.4%       |
|              | Poor Qua... | 45.4%           | 54.6%       |

White Wine > Evaluate Model > Evaluation results

##### Metrics

|                          |          |
|--------------------------|----------|
| Overall accuracy         | 0.744898 |
| Average accuracy         | 0.744898 |
| Micro-averaged precision | 0.744898 |
| Macro-averaged precision | 0.70732  |
| Micro-averaged recall    | 0.744898 |
| Macro-averaged recall    | 0.693754 |

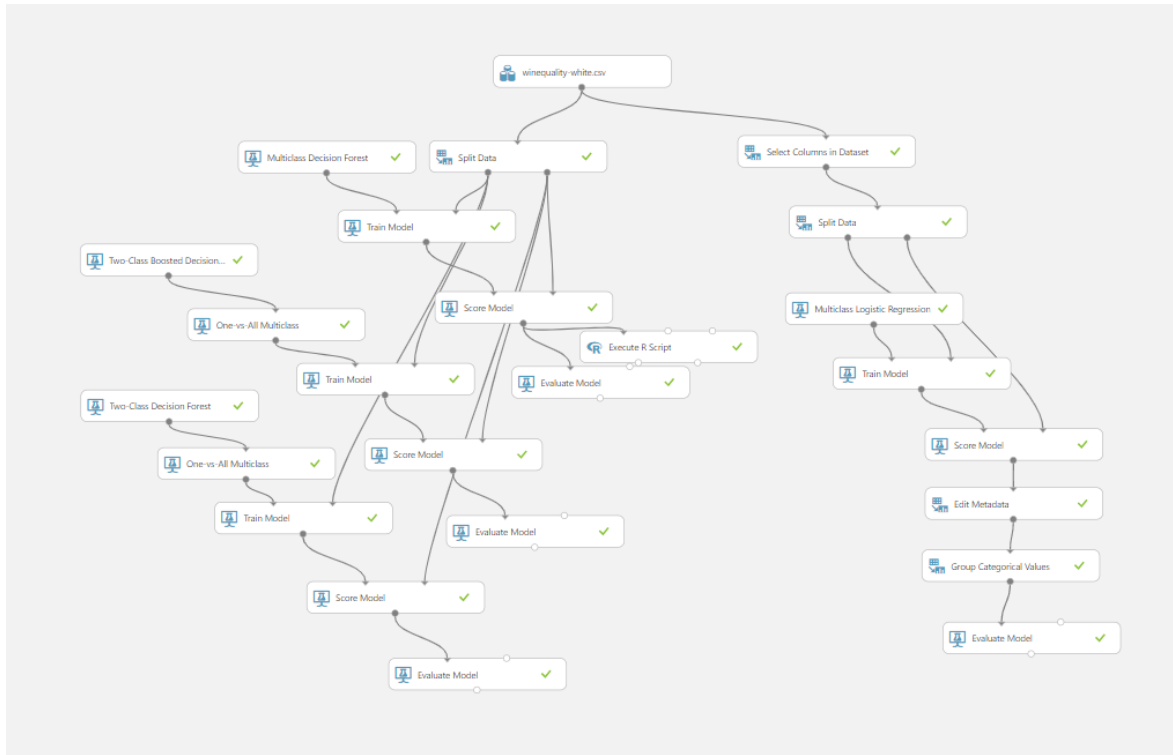
##### Confusion Matrix

|              |             | Predicted Class |             |
|--------------|-------------|-----------------|-------------|
|              |             | Good Qua...     | Poor Qua... |
| Actual Class | Good Qua... | 83.9%           | 16.1%       |
|              | Poor Qua... | 45.1%           | 54.9%       |

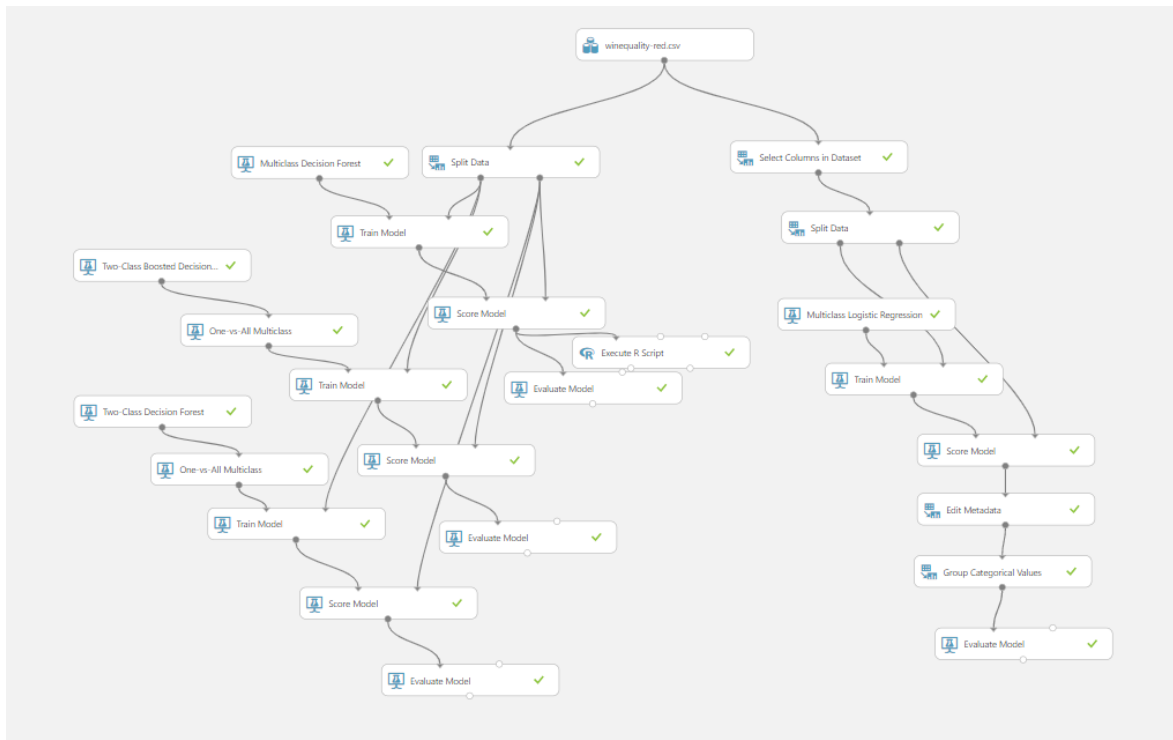
The results on the right include the 'pH' parameter and the left results do not include it. Since there is no change in accuracy, we can drop the 'pH' column to reduce the number of independent variables and still get an accurate result.

### Azure Experiments:

### 1) White Wine dataset



## 2) Red Wine dataset



## Experiment Summary Sheet:

| <b>White Wine</b>               |                                       |          |                            |           |        |          |
|---------------------------------|---------------------------------------|----------|----------------------------|-----------|--------|----------|
| Algorithm                       | Modifying Factor                      | Accuracy | Misclassification rate (%) | Precision | Recall | F1 Score |
| Multiclass Decision forest      | Number of decision trees : 8          | 0.8991   | 10.09                      | 0.6469    | 0.6469 | 0.6469   |
| Two-class Boosted Decision Tree | Number of trees constructed: 100      | 0.9017   | 9.83                       | 0.6561    | 0.6561 | 0.6561   |
|                                 | Maximum number of leaves per tree: 20 |          |                            |           |        |          |
| Two-class Decision Forest       | Number of decision trees : 8          | 0.9026   | 9.74                       | 0.6591    | 0.6591 | 0.6591   |
| Logistic Regression             | With Ph                               | 0.7448   | 25.52                      | 0.7448    | 0.7448 | 0.7448   |
| Logistic Regression             | Without Ph                            | 0.7418   | 25.82                      | 0.7418    | 0.7418 | 0.7418   |
|                                 |                                       |          |                            |           |        |          |
| Algorithm                       | Modifying Factor                      | Accuracy | Misclassification rate (%) | Precision | Recall | F1 Score |
| Multiclass Decision forest      | Number of decision trees : 12         | 0.8979   | 10.21                      | 0.6428    | 0.6428 | 0.6428   |
| Two-class Boosted Decision Tree | Number of trees constructed: 200      | 0.9008   | 9.92                       | 0.653     | 0.653  | 0.653    |
|                                 | Maximum number of leaves per tree: 10 |          |                            |           |        |          |
| Two-class Decision Forest       | Number of decision trees : 12         | 0.9029   | 9.71                       | 0.6602    | 0.6602 | 0.6602   |
| Logistic Regression             | With Ph                               | 0.7448   | 25.52                      | 0.7448    | 0.7448 | 0.7448   |
| Logistic Regression             | Without Ph                            | 0.7418   | 25.82                      | 0.7418    | 0.7418 | 0.7418   |
|                                 |                                       |          |                            |           |        |          |
| <b>Red Wine</b>                 |                                       |          |                            |           |        |          |
| Algorithm                       | Modifying Factor                      | Accuracy | Misclassification rate (%) | Precision | Recall | F1 Score |
| Multiclass Decision forest      | Number of decision trees : 8          | 0.8781   | 12.19                      | 0.6343    | 0.6343 | 0.6343   |
| Two-class Boosted Decision Tree | Number of trees constructed: 100      | 0.8843   | 11.57                      | 0.6531    | 0.6531 | 0.6531   |
|                                 | Maximum number of leaves per tree: 20 |          |                            |           |        |          |
| Two-class Decision Forest       | Number of decision trees : 8          | 0.8781   | 12.19                      | 0.6343    | 0.6343 | 0.6343   |
| Logistic Regression             | With Ph                               | 0.75     | 25                         | 0.75      | 0.75   | 0.75     |
| Logistic Regression             | Without Ph                            | 0.75     | 25                         | 0.75      | 0.75   | 0.75     |
|                                 |                                       |          |                            |           |        |          |
| Algorithm                       | Modifying Factor                      | Accuracy | Misclassification rate (%) | Precision | Recall | F1 Score |
| Multiclass Decision forest      | Number of decision trees : 12         | 0.877    | 12.3                       | 0.6312    | 0.6312 | 0.6312   |
| Two-class Boosted Decision Tree | Number of trees constructed: 200      | 0.8822   | 11.78                      | 0.6468    | 0.6468 | 0.6468   |
|                                 | Maximum number of leaves per tree: 10 |          |                            |           |        |          |
| Two-class Decision Forest       | Number of decision trees : 12         | 0.8812   | 11.88                      | 0.6437    | 0.6437 | 0.6437   |
| Logistic Regression             | With Ph                               | 0.75     | 25                         | 0.75      | 0.75   | 0.75     |
| Logistic Regression             | Without Ph                            | 0.75     | 25                         | 0.75      | 0.75   | 0.75     |

## **Conclusions:**

### **Factors of Wine Quality**

From our analysis we can conclude that the most statistically significant factors for high quality wines are total sulphur dioxide content and alcohol content for both red wine and white wine. Sulphates and pH are also somewhat related to high quality, but the results are not significant.

The total sulphur dioxide measure is more statistically significant for red wine than white wine. For red wine, there is resemblance to a bell-shaped distribution; average quality red wine, around 5/10, has high total sulphur dioxide content, while high quality red wine has medium to low total sulphur dioxide content. For white wine, the trend is similar but reduced. Average quality white wine has a wide range of sulphur dioxide content from approximately 20 to 300, while high quality white wine has a tighter spread of approximately 80 to 180. We can conclude from these results that to match the qualities of premium wine, Portuguese manufacturers need to control the range of total sulphur dioxide and maintain mid-range total content. For red wine, this should be emphasized further as it is statistically more significant.

The alcohol content is more statistically significant for white wine compared to red wine; as the alcohol content increases, the quality of the wine also increases. For red wine, this trend is visible but the relative ranges of alcohol content across various qualities are wider and less distinct. For white wine however, specifically at the 8/10 quality category, there is a clear separation between low alcohol content and high alcohol content, and the alcohol content difference trend from low quality to high quality is more noticeable. This signifies that a defining characteristic of high-quality wine is higher alcohol content; thus, Portuguese wineries should modify their content accordingly.

In order to compete with France and Italy's dominance of the wine market, wineries in Portugal need to make these discussed changes to their Vinho Verde red and white wines. Costs associated with implementing these chemical changes will be moderate; some wines require slight modifications of their chemical content, while others simply need stricter quality control of ingredients. Both types of adjustments will require additional monitoring throughout the wine production process. Although these changes have a nominal monetary cost, the benefit of being able to compete with the bigger players in the market is far more valuable. With improved wine products, Portugal will be able to increase their market share and begin to compete with larger wine-producing states such as France, Spain, and Italy.

### **Model Conclusions**

From our experiments, we can state that the two-class decision forest with number of decision trees as 12 provides the best accuracy rate with the lowest misclassification rate for the white wine dataset compared to the rest of the models. Regarding the red wine dataset, we find that the most accurate model with the lowest misclassification rate is the two-class boosted decision tree with number of trees constructed as 100 and maximum number of leaves per tree as 20.

From the logistic regression experiments, we have also established that the pH factor does not change significantly for the results of the model and is therefore best left out. This showcases the principle of parsimony in our model.

The imbalanced classification in our data set provides a pertinent issue which needs to be highlighted regarding the evaluation of our model. Although we do not believe the effects of the imbalanced class to be detrimental to our predictive validity of our model, we do recognize some improvements can be made. The most applicable solution for this issue would be to collect more data belonging to highly rated wine quality samples and more data on poorly rated wine quality samples. By adding more data from these two classes into our model we can bring balance to our data set so that our model's predictive validity can be improved regarding wine quality estimation in future experiments to serve our business needs.

#### Citations

Cortez, Paulo. *UCI Machine Learning Repository: Wine Quality Data Set*, 2009, [archive.ics.uci.edu/ml/datasets/wine+quality](https://archive.ics.uci.edu/ml/datasets/wine+quality).

Lessrof. "What Makes Portugal Different from Other Wine Countries?: BKWine Magazine |." *BKWine Magazine*, 3 Dec. 2020, [www.bkwine.com/features/wine-regions/portugal-different-wine-countries/](https://www.bkwine.com/features/wine-regions/portugal-different-wine-countries/).

Puckette, Madeline. "The Wines of Portugal (Organized by Region)." *Wine Folly*, 3 Dec. 2020, [winefolly.com/deep-dive/what-wines-to-drink-from-portugal-by-region/](https://winefolly.com/deep-dive/what-wines-to-drink-from-portugal-by-region/).

*Wine Producing Countries 2021*, 2021, [worldpopulationreview.com/country-rankings/wine-producing-countries](https://worldpopulationreview.com/country-rankings/wine-producing-countries).