

**AI Saturdays** **| Red Wine Quality Group** **|** **June 19, 2020**

**Red Wine Quality Analysis Report**

**A MACHINE LEARNING PROJECT FOR COHORT FIVE**



# **Citation**

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.

Modeling wine preferences by data mining from physicochemical properties.

In Decision Support Systems>, Elsevier, 47(4):547-553. ISSN: 0167-9236.

Available at: [@Elsevier] http://dx.doi.org/10.1016/j.dss.2009.05.016

[Pre-press (pdf)] http://www3.dsi.uminho.pt/pcortez/winequality09.pdf

[bib] http://www3.dsi.uminho.pt/pcortez/dss09.bib

1. Title: Wine Quality

2. Sources

Created by: Paulo Cortez (Univ. Minho), António Cerdeira, Fernando Almeida, Telmo Matos and José Reis (CVRVV) @ 2009

3. Past Usage:

P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis.

Modeling wine preferences by data mining from physicochemical properties.

In Decision Support Systems>, Elsevier, 47(4):547-553. ISSN: 0167-9236.

4. Number of Instances: red wine - 1599; white wine - 4898.

5. Number of Attributes: 11 + output attribute

Note: several of the attributes may be correlated, thus it makes sense to apply some sort of

feature selection.

6. Attribute information:

For more information, read [Cortez et al., 2009].

Input variables (based on physicochemical tests):

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 – pH

10 - sulphates

11 - alcohol

Output variable (based on sensory data):

12 - quality (score between 0 and 10)

7. Missing Attribute Values: None

**Contents**

[**Citation** 2](#_Toc43501034)

[**Aim of the Project** 4](#_Toc43501035)

[**Group Members** 5](#_Toc43501036)

[**Methods Used** 6](#_Toc43501037)

[**Stage 1: Problem Identification** 6](#_Toc43501038)

[**Stage 2: Data Analysis** 7](#_Toc43501039)

[**Data Cleaning** 7](#_Toc43501040)

[**Feature Creation** 7](#_Toc43501041)

[**Data Exploration and Visualization** 7](#_Toc43501042)

[**Findings and Discoveries** 8](#_Toc43501043)

[**Stage 3: Machine Learning** 8](#_Toc43501044)

[**Data Pre-Processing** 8](#_Toc43501045)

[**Stage 3: Machine Learning** 10](#_Toc43501046)

[**Building Our Machine Learning Models** 10](#_Toc43501047)

[**Evaluating the Models** 10](#_Toc43501048)

[**Feature Selection** 10](#_Toc43501049)

[**Performance Review / Findings and Discoveries** 10](#_Toc43501050)

[**Improving Our Models** 11](#_Toc43501051)

[**Conclusion** 11](#_Toc43501052)

[**Difficulties** 12](#_Toc43501053)

[**Recommendations** 12](#_Toc43501054)

# **Aim of the Project**

High quality red wine has been part of social, religious, and cultural events for hundreds of years. Medieval monasteries believed that their monks lived longer partly because of their regular, moderate drinking of high-quality wine. There is ongoing research on improving existing methods used in distinguishing wine by quality. These days quality is mainly assigned score values by professional wine tasters who judge samples based on sensory data.

Obviously, this method of evaluation is highly subjective and it stresses the need for developing a highly effective machine learning model which predicts the quality score of various samples based on their properties. Such a model could be used alongside professional wine testers to increase the accuracy of testing through correlation of the scores gotten from both sources.

In this project, our aim is to determine which physicochemical properties make red wine 'good' by using some basic data exploratory and visualization analytical methods as well as some machine learning techniques. We were able to access red wine data of Portuguese "Vinho Verde" wine through Cortez et al., 2009 research to carry out this project.

# **Group Members**

* Olorondu Chizurum
* Babatunde Koiki
* Erinfolami Emmanuel
* Sandra Orji
* Obiora Maduakor
* Sodiq Agunbiade
* Modurotolu Olokode
* Orutorojo Okiogbero Michael
* Shukurat Bello

# **Methods Used**

We were able to successfully carry out the project using a workflow that comprised of three stages.

## **Stage 1: Problem Identification**

Problem Statement: “A restaurant in Nigeria is currently faced with the issue of quickly distinguishing

wine by quality. They reached out to your team to come up with a model that would help them quickly

identify high quality wine”

We began by importing all the python modules we needed to carry out the project. After carefully digesting the problem state given above, we retrieved the dataset and converted it into a well-structured dataframe using the Pandas framework. We proceeded to analyze the data frame and discovered that there were:

* 11 Input Variables or Features
* 1 Output Variable

We also found out that the output variable contained 6 unique integers within the range 3 to 8. The two discoveries made us realize that we were dealing with a **Multi-Class Classification Problem**. We also evaluated the entire data set by checking the distribution of each unique integers contained in the output variable and saw that it was an **Imbalanced Dataset.** This was due to the fact that the dataset (containing a total of 1599 wine test samples) was dominated by samples with a quality score of “**5**” (681 samples), and also “**6**” (638 samples) while a small number of wines samples scored “**3, 4, 7, 8**” (10, 53, 199 and 18 samples respectively).

The evaluation of the problem as a **Multi-Class Classification Problem with an Imbalanced Dataset** helped us to effectively chart out a solution.

## **Stage 2: Data Analysis**

### **Data Cleaning**

After carefully going through the dataset, we discovered the following:

* There were no missing values present in any row or column.
* The dataset only contained numerical data of type float and integer.
* The dataset was well ordered.

These discoveries made us realize that there wasn’t any cleaning to be done.

### **Feature Creation**

From our understanding of various machine learning algorithms, we knew the importance of simplifying our dataset to improve the performance of our machine learning model. We also got a tip from the dataset we acquired that samples with quality values higher than “**6.5**” could be classified as “**Good Wines**”. We decided to therefore to create a new output variable out of the “**quality**” column which we named “**verdict**”.

We went about this by looping through the quality column and ascribing a value of “**0**” on the **verdict** column for samples with quality values less than 6.5 and giving samples with a quality of 6.5 and above a verdict of “**1**”. The verdict column when used as a single output variable would change our classification problem to a **Binary Classification Problem** which are generally easier to work on.

### **Data Exploration and Visualization**

We then proceeded to explore our dataset by the following methods:

* Checking the correlation of all the columns in the dataset.
* Checking the correlation of the input variables to the quality scores as well as the verdict score.
* Evaluating the mean and variance of the input variables.
* Checking the count, mean, standard deviation, minimum values, 25th percentile, 50th percentile, 75th percentile and the maximum values of all input variables.

We then extracted a new dataframe from the original dataset and named it the “**gooddata**”. This data frame contained only samples with a verdict value of **1**. We also applied the data exploratory methods present above on this dataframe.

We visualized the data we were able to retrieve from the original dataset by building:

* Two count plots showing the quality and verdict score distribution respectively.
* A box plot of each input variable.
* A histogram for each input variable.
* A scatter plot matrix for each input variable.
* Heatmaps for the original and gooddata dataframes.
* Clustermaps for the original and gooddata dataframe.

### **Findings and Discoveries**

The results of our data exploration and visualization include the following:

1. Correlations pointing toward similarities between some input variables namely:

* Total Sulphur-dioxide and free Sulphur dioxide.
* Fixed acidity and citric acid.
* Chlorides and sulphates.
* PH and alcohol.

1. There was no obvious influencing input variable on the quality score though Alcohol and Volatile Acidity showed the highest direct and indirect correlation factor respectively to the

Quality score.

1. The quality score of samples was indeed imbalanced (there were a huge number of normal wine samples in the data with few bad and good samples).
2. There was a huge variance in the values of the eleven input variables (Total and free Sulphur-dioxide had a very high variance of **1082 and 108** while features like density had a variance value of **0.000004**
3. Also, the huge values of the total and free Sulphur-dioxide had the capacity to reduce the ability of our models to accurately calculate **feature importance**.

## **Stage 3: Machine Learning**

### **Data Pre-Processing**

#### **Splitting the data**

We started our pre-processing process by splitting our original dataset into 3 variables namely:

* **X -** which was a 2D array containing the values of all the **input variable**
* **y** **-** which contained an array of the values of the **quality scores**
* **z -** which contained an array of the values of the **verdict scores**

The next thing we did was to split the data set into **2** **groups** using the train test split function in scikit-learn. Both groups had separate training and test datasets.

* The first group had **z** as its only single output variable.
* The second group had **y** as its only single output variable.

**The essence of this was to carry out machine learning on the problem both as a multi-class classification problem and also as a binary classification problem while comparing the performance of our model on both instances.**

#### **Applying SMOTE Oversampling**

We moved on to make our best attempt at fixing the **Imbalanced** nature of our dataset by applying a **specialized sampling technique** on the 2 groups. We choose the **SMOTE Oversampling Technique** for this. The reason was due to the fact that our dataset was relatively small (1599 samples). We applied this technique on the training data of both groups and noticed an increase in the total number of training samples.

#### **Feature Scaling**

Referencing the result, we got from our data exploration, we had to carry out some feature scaling in the training data of both groups. This was done to put all the input variables on the same level and to avoid decreased accuracy when calculating feature importance. To achieve this, we made used of the **Standard Scaler function** of the scikit-learn package to **Regularize** the training data. This made all the input variables to have a value that fell within the range of **-1 and +1**. With this we were done with pre-processing our data.

## 

### **Building Our Machine Learning Models**

After pre-processing our data, we created five machine learning models and fitted them with the training data from the two groups. They include a:

* **Logistic Regression Model**
* **Support Vector Machine – Linear SVC Model**
* **Support Vector Machine – SVC Model**
* **Random Forest Classifier Model**
* **Decision Tree Model**

### **Evaluating the Models**

The metrics we used in evaluating each of these models include:

* The **Accuracy Score**
* The **F1 Score**
* The **Confusion Matrix**
* The **Mean Absolute Error Score**
* The **Classification Report**

### **Feature Selection**

We carried out feature selection on the models to determine the most influential physiochemical properties of red wine. This was done by:

* Finding the **.coef\_ property** of three of the models (Logistic Regression, SVM- linear SVM and SVM – SVC models)
* Finding the **.feature\_importancies\_ property** of two of its models (Decision Tree and Random Forest Classifier models)

The values obtained were matched to the respective input variables and converted to dataframe. The most important features found overall include Alcohol, volatile and fixed acidity, density, sulphates, total sulfur-dioxide, citric acid and chlorides (The two dominant features being Alcohol and Volatile Acidity confirming the results gotten from our data exploration/visualization.)

### **Performance Review / Findings and Discoveries**

All the models generally had higher metric scores when fitted with the dataset group whose **output** variable was the **verdict** column compared to when they were fitted with dataset group whose **output** variable was the **quality** column

For the first dataset group whose output variable was the **verdict** column, the **Support Vector Machine – SVC model** performed best with an **accuracy score** of **0.92** and an **F1 Score** of **0.923** followed by the **Random Forest Classifier Model (0.88, 0.896),** the **Decision Tree Model (0.875, 0.875),** the **Logistic Regression Model (0.74, 0.795)** and the **Support Vector Machine – Linear SVC model (0.73, 0.78).**

For the second dataset group whose output variable was the **quality** column, the **Random Forest Classifier Model** performed best with an **accuracy score** of **0.73** and an **F1 Score** of **0.72** followed bythe **Decision Tree Model (0.65, 0.65),** the **Logistic Regression Model (0.59, 0.59), Support Vector Machine – SVC model (0.58, 0.58)** and the **Support Vector Machine – Linear SVC model (0.34, 0.40).**

This shows that for our dataset the **Support Vector Machine – SVC model** was the best for our defined problem when it was in a **Binary Classification Format** while the **Random Forest Classifier Model** was the best performer for our problem in its original **Multi-Class Classification Format.**

Additionally, when evaluating our models for their **ability to predict good wine accurately,** from the respective confusion matrix plotted on the **dataset group** whose output was the **verdict column**, the **Logistic Regression Model** (25 correct predictions out of 30 good wine samples) and the **Support Vector Machine – Linear SVC model** (25 correct predictions out of 30 good wine samples)performed **best** followed by the **Random Forest Classifier Model** (23 correct predictions out of 30 good wine samples**), Decision Tree Model** (23 correct predictions out of 30 good wine samples) **and lastly** the **Support Vector Machine – SVC model** (18 correct predictions out of 30 good wine samples).

### **Improving Our Models**

After **improving the accuracy of our models with:**

* **GridSearch CV**
* **Cross Validation**

We ended up with a general improvement on the accuracy scores across board, the **Support Vector Machine – SVC model** stillperformed best with an **accuracy score** of **0.97** followed by the **Random Forest Classifier Model (0.93),** the **Decision Tree Model (0.895),** the **Logistic Regression Model (0.82)** and the **Support Vector Machine – Linear SVC model (0.82).**

# **Conclusion**

After careful evaluation of the dataset, visual plots and machine learning models, the six most influential physiochemical features that makes red wine good fall within:

* - Alcohol Percentage (9.2 - 14.9)
* - Volatile Acidity (0.12 - 0.915)
* - Sulphates (0.39 - 2.00)
* - Total Sulfur dioxide (6.0- 289.0)
* - Citric Acid (0.0 - 1.00)
* - Chlorides (0.012 - 0.358)

PH, Residual Sugars, Density and Fixed Acidity all have a large range across board so they won't really be helpful in pinpointing whether a wine is good or not. Free Sulphur Dioxide is a subset of total Sulphur dioxide.

Also, I would give 2 models to the Nigerian restaurant namely:

* The **Support Vector Machine Model - SVC Model –** For quick and accurate determination of which wines are good and which wines are not.
* The **Random Forest Classifier Model –** For accurate prediction of wine quality scores

## **Difficulties**

The major difficulties we encountered while carrying out this project included the following:

* The small sample size of the data.
* The imbalanced nature of the data which we tried to solve with SMOTE Oversampling.
* The data was taken from just one out of the thousands of different wine producers worldwide (Portuguese "Vinho Verde" wine).

## **Recommendations**

Our recommendation to the Nigerian restaurant would be to:

* Get a much larger, much diverse sample sized dataset which we could use to build a better preforming and generalized machine learning model as the model we developed is only applicable (to an extent) to the Portuguese "Vinho Verde" wine type.