**Introduction**

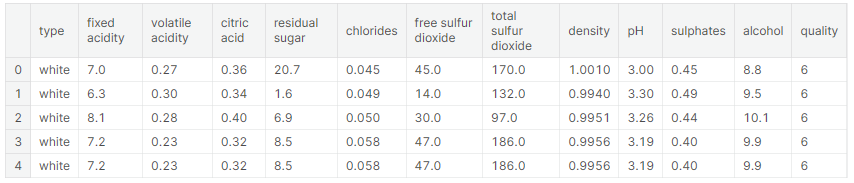
The dataset is related to red and white variants of the Portuguese "Vinho Verde" wine. The reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.)

This dataset is viewed as a binary classification task: define the quality of wine.

**1. Dataset and features**

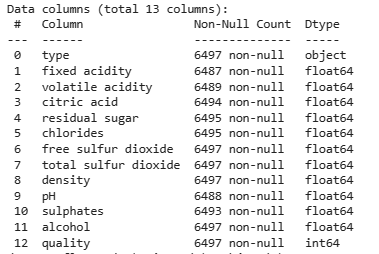
Input variables (based on physicochemical tests):  
1 - fixed acidity  
2 - volatile acidity  
3 - citric acid  
4 - residual sugar  
5 - chlorides  
6 - free sulfur dioxide  
7 - total sulfur dioxide  
8 - density  
9 - pH  
10 - sulphates  
11 - alcohol  
Output variable (based on sensory data):  
12 - quality (score between 0 and 10)

**Table 1. Main features**



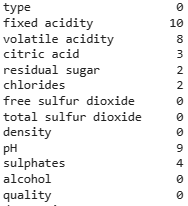
After downloading the dataset, we conduct a primary analysis of the features (Picture 1). It is necessary to understand what types of features exist and whether there are null values among features.

**Picture 1. Main features**



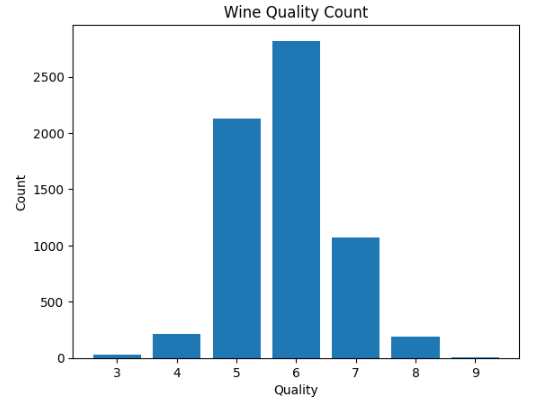
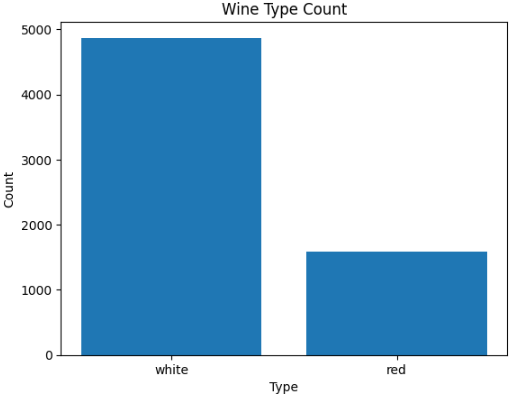
We determine the total number of missing values (Picture 2). Among the 6497 values, the total number of omissions is very small, so they can be deleted.

**Picture 2. Number of omissions**



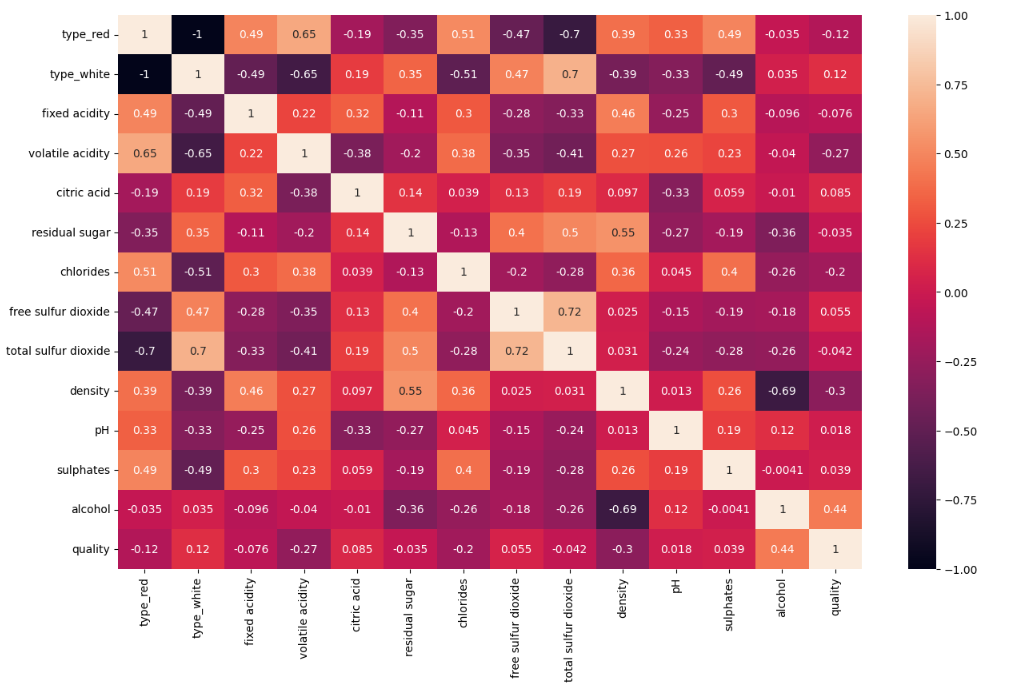
We can notice a very uneven distribution of values among the "Type" attribute. White wine in the sample is more than 2 times more than red wine. As for the quality of the wine, the wine of average quality prevails (from 5 to 7). In turn, tails – low quality (from 3 to 4) and very high-quality wine (from 8 to 9) make up a smaller part of the sample.

**Picture 3. Wine types and quality**



The following features (Picture 4) of "Alcohol" & "Density", "free sulfur dioxide" & "total sulfur dioxide", "density" & "residual sugar" demonstrate the greatest relationship. Given the high relationship, individual features can be removed to accelerate the model processing without loss of accuracy of predictions.

**Picture 4. Correlation Matrix**



**2. Defining a task and building a model**

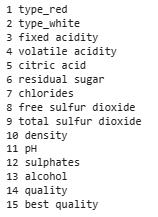
Task: It is necessary to identify a good/ bad wine.[[1]](#footnote-1)

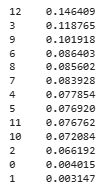
Model: RandomForestClassifier()

Features:

To train the model, it is proposed to use the most important features defined by the RandomForestClassifier algorithm, precisely

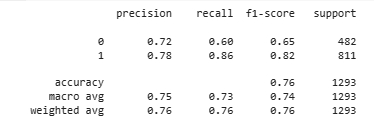
['sulphates', 'fixed acidity', 'total sulfur dioxide', 'residual sugar']

**Picture 5. Importance of features**



**3. Conclusions and further directions of research**

The accuracy of the model is 0.76. This means that the model is able to classify wine into good and bad with up to 76% accuracy. Not a bad result, but there is something to strive for. To improve the quality of the results, you can try other models, as well as other signs.



1. Good wine (from 6 to 9) and bad wine (strictly less than 6) [↑](#footnote-ref-1)