Wine Quality Analysis – Red wine - dataset

The following analysis refers to the Red Wine Quality dataset, a public *dataset* that contains physical-chemical data of red wines and their quality. Data were collected from northern Portugal and had almost 1600 observations. The dataset can be downloaded on the https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/ (winequality-red.csv). The machine learning technique - *multiple linear regression made in the* python programming language for the *analysis* of the dataset was *used.*

The goal of the study is to assemble a *multiple linear regression model* and evaluate which of *the features* are the most relevant for the production of red wine.

Table 1 - Dataset Data

	fixed acidity	volatile acidity	citric acid	residual sugar	chlorides	free sulfur dioxide	total sulfur dioxide	density	pH	sulphates	alcohol	quality
0	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
1	7.8	0.880	0.00	2.6	0.098	25.0	67.0	0.99680	3.20	0.68	9.8	5
2	7.8	0.760	0.04	2.3	0.092	15.0	54.0	0.99700	3.26	0.65	9.8	5
3	11.2	0.280	0.56	1.9	0.075	17.0	60.0	0.99800	3.16	0.58	9.8	6
4	7.4	0.700	0.00	1.9	0.076	11.0	34.0	0.99780	3.51	0.56	9.4	5
...	...	...	...	...	...	...	...	...	...	...	...	...
1594	6.2	0.600	0.08	2.0	0.090	32.0	44.0	0.99490	3.45	0.58	10.5	5
1595	5.9	0.550	0.10	2.2	0.062	39.0	51.0	0.99512	3.52	0.76	11.2	6
1596	6.3	0.510	0.13	2.3	0.076	29.0	40.0	0.99574	3.42	0.75	11.0	6
1597	5.9	0.645	0.12	2.0	0.075	32.0	44.0	0.99547	3.57	0.71	10.2	5
1598	6.0	0.310	0.47	3.6	0.067	18.0	42.0	0.99549	3.39	0.66	11.0	6

In table 1, we see the features of the dataset and the target - quality. The features are:

1 - fixed acidity: amount of tartaric acid in wine (g/dm^3)

2 - volatile acidity: amount of acetic acid in wine (g/dm^3)

3 - citric acid: amount of citric acid in wine (g/dm^3)

4 - residual sugar: amount of sugar remaining after fermentation steps (g/dm^3)

5 - chlorides: amount of salt inwine (g/dm^3)

6 - free sulfur dioxide: amount of sulphur dioxide in wine (mg/dm^3)

7 - total sulphhur dioxide: total amount of sulphur dioxide in wine (mg/dm^3)

8 - density: wine density (g/cm^3)

9 - pH: pH of wine

10 - sulphates: additive for wines (g/dm^3)

11 - alcohol: alcoholic percentage of wine (% in volumes)

Quality is classified from 0 to 10, being 0 the best quality wine and 10 the best quality possible.

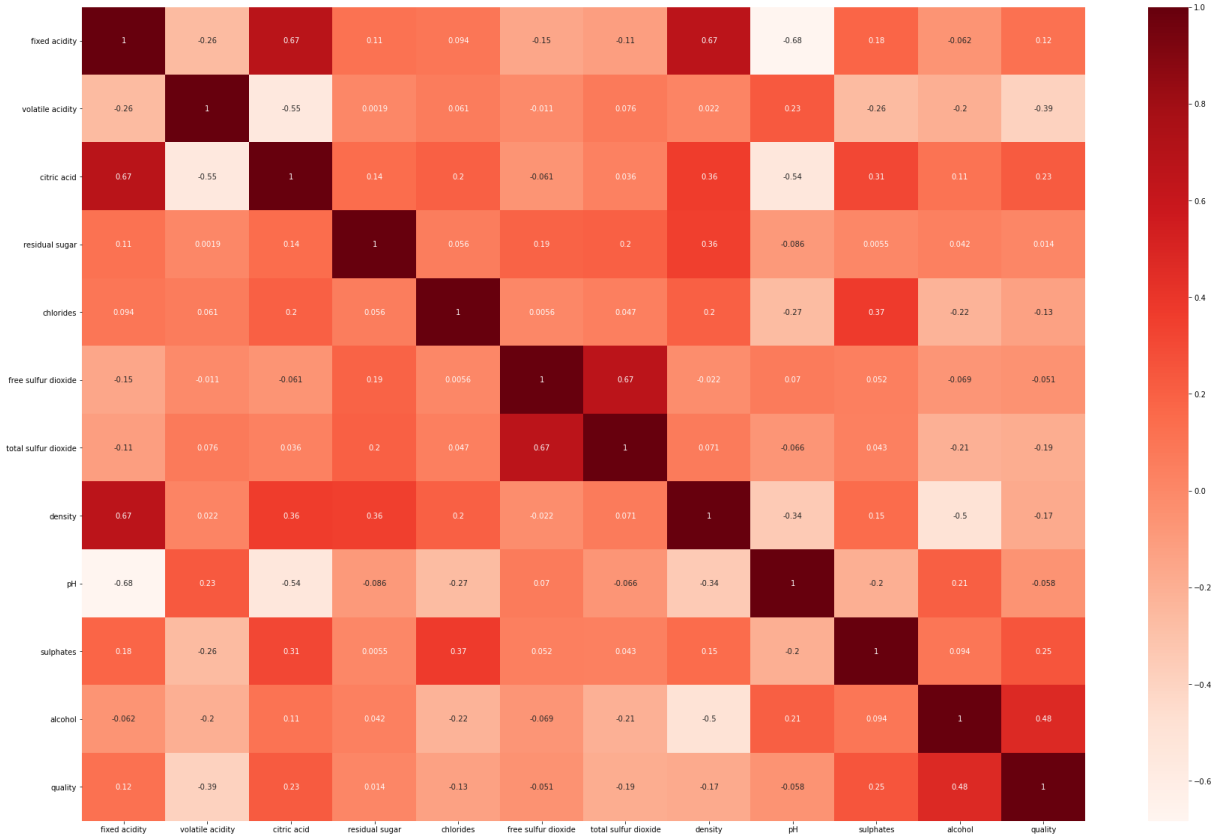

Figure 1 - Heatmap of the dataset

To check for correlations between the features and the target, a heatmap was made. In it, 1 means that there is a very positive correlation and, -1, very negative. We can observe that alcohol, sulphates, citric acid have positive correlation and, mainly, volatile acidity negative correlation. These relationships will be investigated.

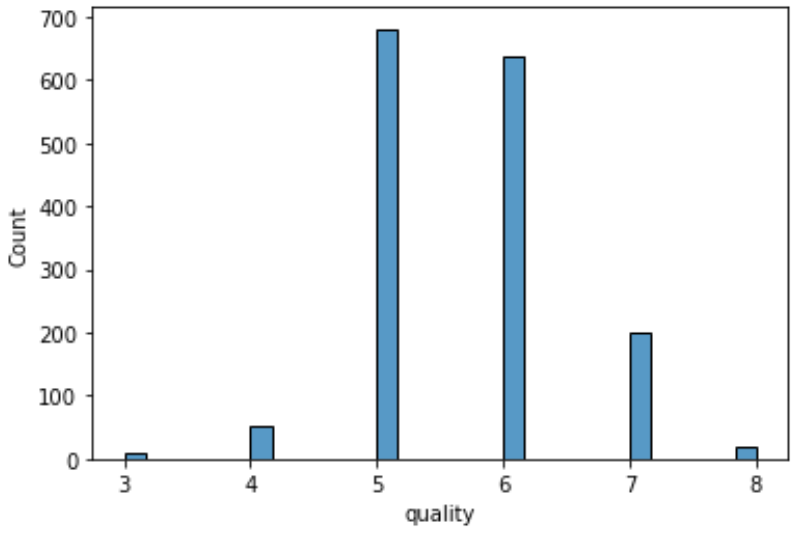

Figure 2 - Histogram of quality

In Figure 2, we have the histogram of quality and, right away, we see that the number of observations of quality 5 or 6 is much higher than the rest, so we have an unbalanced dataset that may impair the accuracy of the model to be implemented.

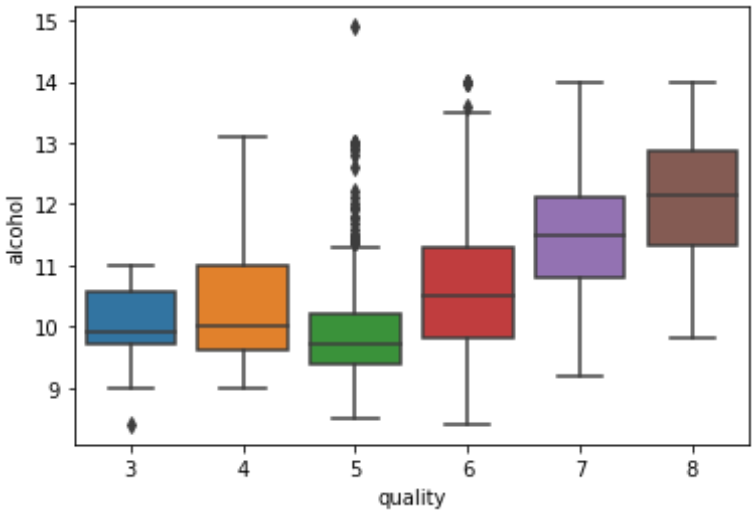

Figure 3 - Boxplot of alcohol and quality

Investigating what was discovered in Figure 1, we have figure 3, that the higher the alcoholic percentage of the wine, the better its quality.

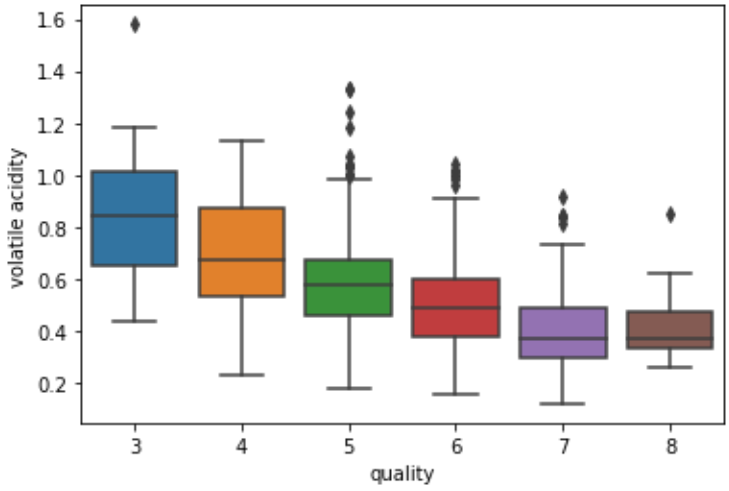

Figure 4 - Boxplot of volatile acidity and quality

Volatile acidity is the amount of acetic acid in the wine and, the higher its amount, the greater the taste of vinegar, spoiling the taste of the wine. This is what we see in Figure 4, the higher amount of acetic acid, the lower the quality of the wine.

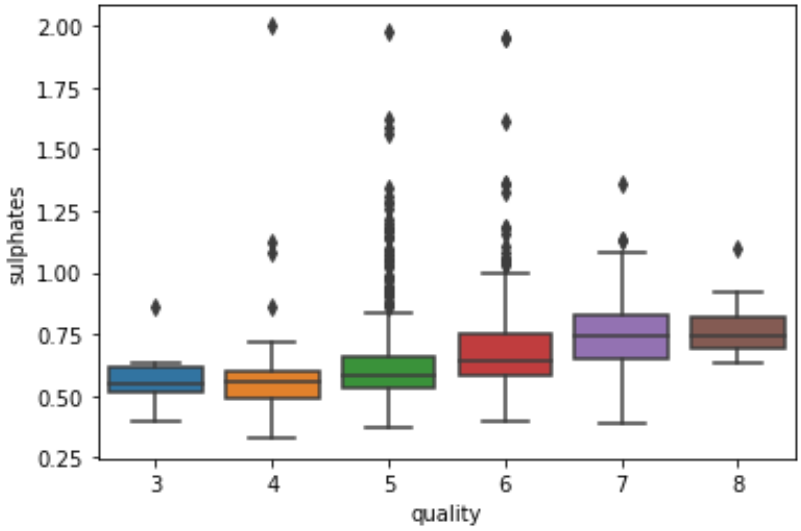

Figure 5 - Boxplot of sulphates e quality

In the manufacture of wine, sulfates are one of the most important additives used, because it acts as antioxidant and antimicrobial, besides maintaining the taste and feeling of freshness. In Figure 5, we have that the higher the amount of sulfates, the higher the quality of the wine.

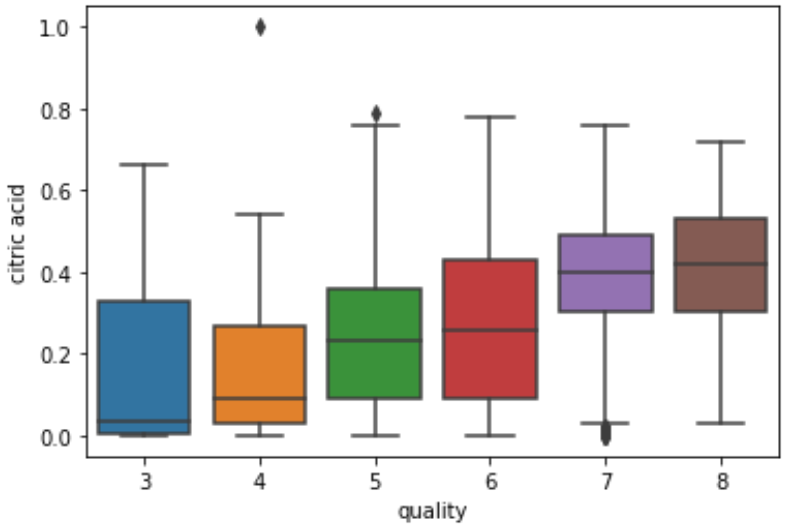

Figure 6 - Boxplot of citric acid and quality

In Figure 6, we have that the higher the amount of citric acid in the wine, the better its quality. This happens, because citric acid provides a feeling of freshness and flavor to the wine.

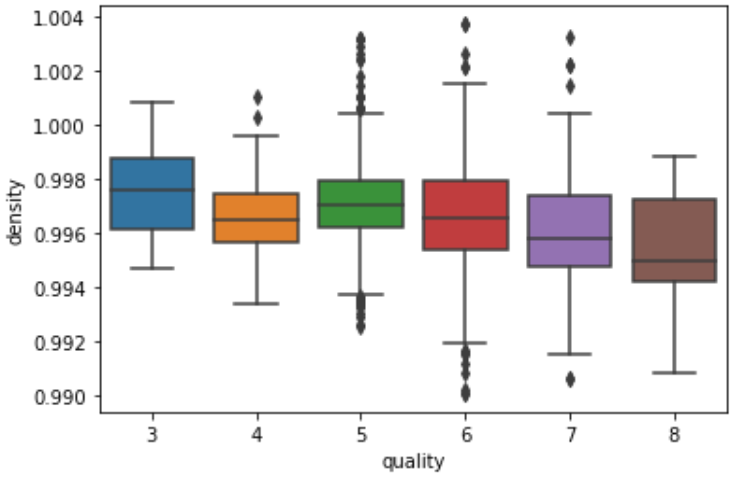

Figure 7 - Boxplot of density and quality

The density does not seem to interfere much in the quality of the wine, as seen in Figure 7, so we can expect its contribution to the model not to be large.

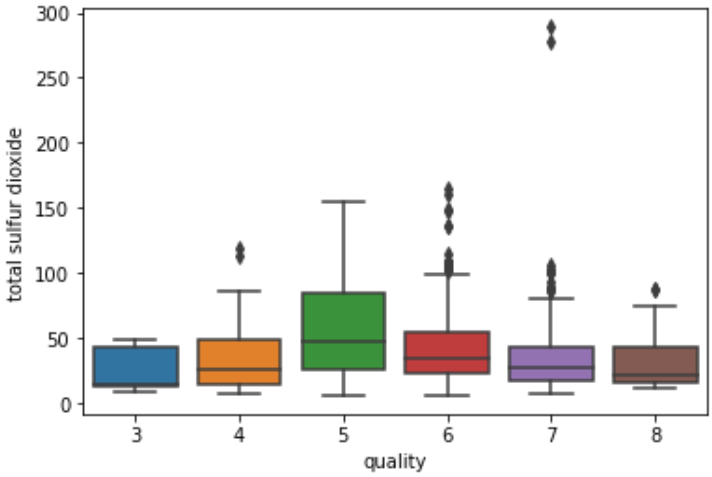

Figure 8 - Boxplot of total sulfur dioxide and quality

The feature total sulfur dioxide represents the amount of total sulfur dioxide. In smaller amounts, free sulfur dioxide does not taste in wine, but in quantities greater than 50 ppm, it can alter the taste of wine, decreasing its quality. That's what was seen in Figure 8.

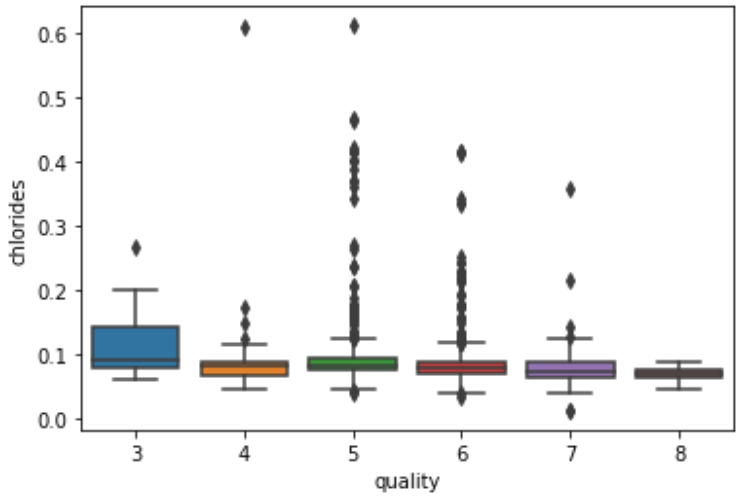

Figure 9 - Boxplot of chlorides and quality

In Figure 9, we have a higher amount of chloride in wine, the lower its quality. This is due to the fact that chloride, in larger quantities, can give the wine a salty taste, impairing its taste.

Table 2 - Model Features

	volatile acidity	citric acid	chlorides	total sulfur dioxide	density	sulphates	alcohol
0	0.700	0.00	0.076	34.0	0.99780	0.56	9.4
1	0.880	0.00	0.098	67.0	0.99680	0.68	9.8
2	0.760	0.04	0.092	54.0	0.99700	0.65	9.8
3	0.280	0.56	0.075	60.0	0.99800	0.58	9.8
4	0.700	0.00	0.076	34.0	0.99780	0.56	9.4
...	...	...	...	...	...	...	...
1594	0.600	0.08	0.090	44.0	0.99490	0.58	10.5
1595	0.550	0.10	0.062	51.0	0.99512	0.76	11.2
1596	0.510	0.13	0.076	40.0	0.99574	0.75	11.0
1597	0.645	0.12	0.075	44.0	0.99547	0.71	10.2
1598	0.310	0.47	0.067	42.0	0.99549	0.66	11.0

With the information of which are the most important features, we have summarized in table 2, the multiple linear regression model was made, in order to predict the quality of the wine.

Table 3 - Model Accuracy

R^2	Bias
39,50%	5,63

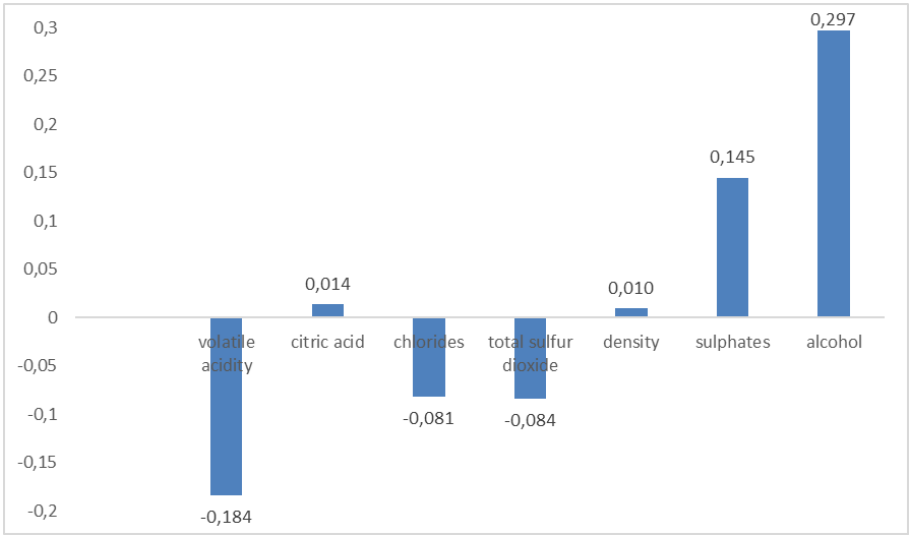

Figure 10 - Weight of each feature of the model

In table 3, we have the result for the accuracy of the model and in Figure 10 we have the values of the weights. In it, we can observe that alcohol is the feature that most impacts the quality of wine, along with sulphates, which acts as antioxidant and antimicrobial. Volatile acidity negatively impacts quality, as a high amount of acetic acid negatively affects the taste of wine. Still in Figure 10, we have density and citric acid are positive for quality, but are not of great importance when compared with sulphates and alcohol. Therefore, a wine factory needs to be aware, mainly, of the amount of acetic acid so as not to affect the quality of its product.

The result seen in table 3 is not ideal, but we can point out some factors that contributed to the poor performance of the model. Information such as year of harvest, time of maturation of the wine, whether it was stored in barrel or tank, quality of raw material, place of the harvest and etc. are fundamental factors to understand the quality of a wine, therefore in possession of this information, we could have a gain in the accuracy of the model. However, the most critical factor for the model was the unbalanced dataset, since there were many quality wines 5 and 6, as seen in Figure 2, compared to the others. This impaired the algorithm when checking what actually increases or decreases the quality of the wine, so with a more balanced dataset, we could have a better model.