Machine Learning Project

Statistical model for Portuguese Wine quality detection

By Gustavo S Garofalo

1. **Project Motivation**

Imagine you work on a product development department of a wine cellar in Portugal, and you are about to devise the formulation of a brand-new wine bottle that will be launched in the market soon. If only there was a tool or a sequence of methods that could tell you what is the chemical components that contribute to the best quality of product. Luckily for you there is! And the Machine Learning tools and methods are easier than ever to use and implement. In this project we will develop a machine learning model with a series of procedures to help in this particular case. Although, this project contemplates a specific issue (Wine quality), these same procedures or some very similar to these could be used in a lot of different scenarios. A wide variety of products consists of chemical components, could we use machine learning in order to determine the best quality of the product based on its chemical components? That is exactly what this project attempts to accomplish, whether you are developing a new wine product, dish detergent, shampoo, etc. This work can be a good blueprint for the task in hands. Being a wine enthusiast myself, I find it interesting to discover more information about it in order to get more technical knowledge on the subject, as a good conversation or maybe get a good model accuracy score in order to predict the quality of wines I have never tried.

1. **Problem Statement**

What attributes in terms of chemical components contributes for a good wine? What would make the best wine possible? With which precision degree could we classify if a wine (from the test dataset) is good or bad based on its chemical components? What chemical components are more relevant?

Are there any differences between the wine composition, quality perceived and whether the wine is red or white?

Achieve the highest accuracy possible in the White wine and Red Wine dataset, improve the model to its fullest, desired precision would be more than 80%? Try to predict a wine quality from another region!

1. **The Data (EDA)**

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

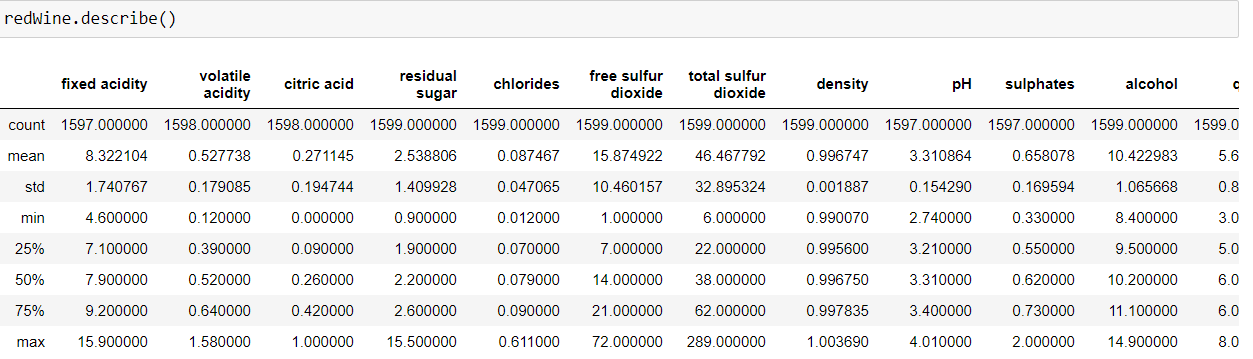
11 - alcohol

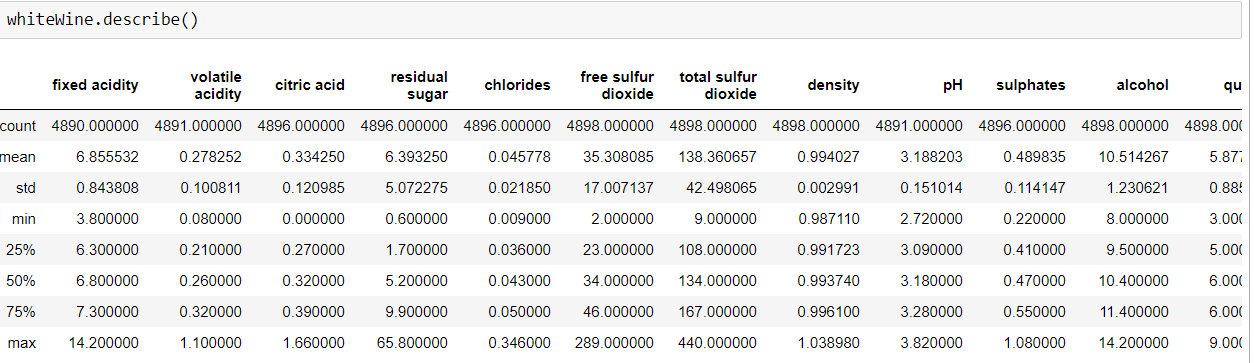
Output variable (based on sensory data):

12 - quality (score between 0 and 10)

The dataset that will be utilized is publicly available for researching purposes, and this version was acquired in Kaggle because it was in the desired CSV format already. This data contemplates a large number of samples for Portuguese red and white wines (“Vinho verde”) evaluation made by wine experts. More specifically, the dataset consists of 6,497 data points, and for each observation we have 11 input variables (predictors) and 1 output (response) variable. The observations collected can be divided into 1,599 red wine scoring and 4,898 white wine. The input variables consist of physicochemical attributes of the wine such as PH, chlorides, density and citric acid. And the response variable is made of the median of at least 3 evaluations of each instance by wine experts, the evaluation ranges from 0 to 10, which would be the highest score possible. The input variables are numerical, and the output, although it is also numerical, was transformed in categorical. The transformation is a binary response given that wines with quality of 6 and above are considered as good and therefore gets a value 1 assigned to it, the remaining scores are bad wines and gets a 0.

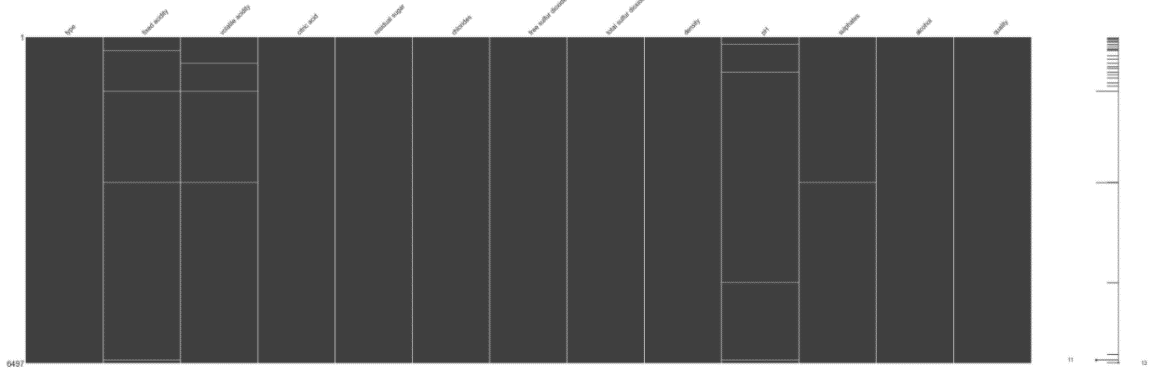
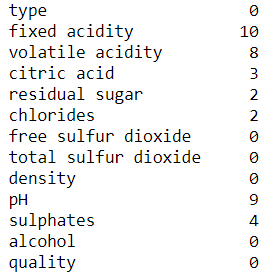
Initially the idea was to compare the results of red and white wines with the complete dataset (that contains both red and white wines combined), but after primary examination, the red and white wines present some differences in its main statistical indicators, so the analysis will be done separately for both. It will be interesting to see if the models will differentiate as the models and its features are improved.

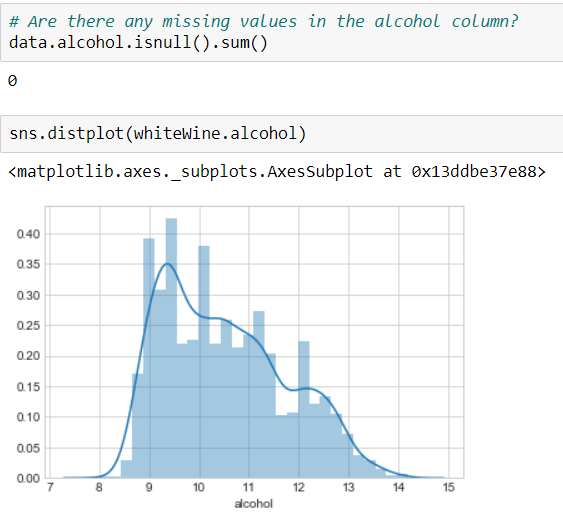




As we can see from the basic statistical description of each variable, there are a lot of differences between attributes of the white wine type when in comparison to red wine. For instance, the average residual sugar found in white wines is almost 3 times higher than the residual sugar of the red wines. The same discrepancies can be found in sulphates and sulfur dioxide.

The limitations include a number of outliers, meaning that there are some really bad and some exceptional wines samples that could cause distortion in the model performance, although most of the wines are normal, these observations could be removed. Another limitation for this and pretty much every dataset used for machine learning is that it does not contemplates every possible predictor to the wine quality perception, instead it focuses only on some metrics. It doesn’t contemplate the type of grape, the price and brand for example that could be some relevant variables to this task. Also, the wines analyzed are from a specific region, this means that the model could only be useful to classify wines in that region.

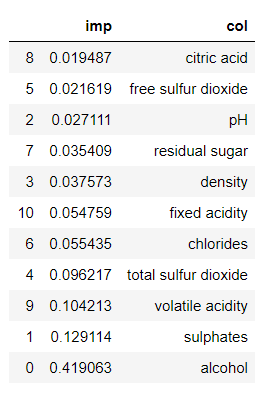
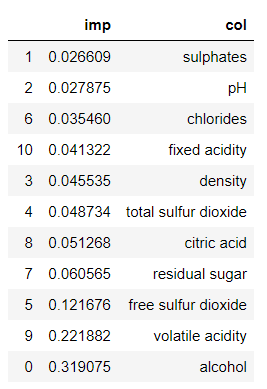


 In the above images it is possible to Visualize a graph with all the missing values by each predictor. In this case the number of missing values is very low and a lot of them were from coincident observations, they were delt with by exclusion. Checking for null values were the second process after the statistical description, and the third and last part of the process I created a distribution graph for each and every feature individually, In order to have a physical representation of ho the data is spread out. In this stage of the process is when the null values were excluded and the categorical value was created for the target variable. Below are some examples of the graphs created.



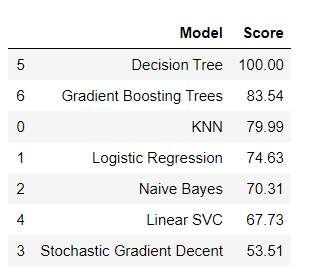
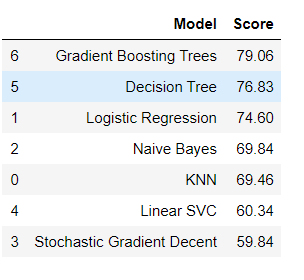
1. **Feature Selection**

Since a lot of the features are correlated, for example total sulfur dioxide and free sulfur dioxide, not all features may be relevant and figuring out which features are more relevant and the differences between the type of wines will provide support to answer to the first question of the project scope. The feature selection could also be used to improve the model accuracy, but it was not the case in this project. The feature selection is as follows for white and red wine (left and right respectively).

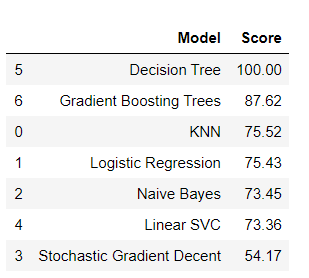
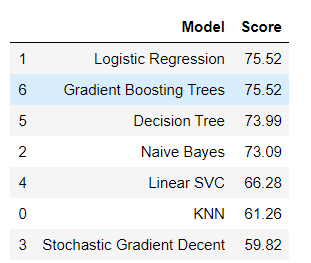
For both types of wines, the Alcohol level is the most important predictor. For white wines, after the alcohol, volatile acidity and free sulfur dioxide are the most relevant features in order to predict whether a wine is good or not. In the other hand, for the red wines, the top predictors after alcohol are sulphates and volatile acidity as well. The alcohol has a way higher relevance for the red wines than white wines and residual sugar is also on the top of the chart for the white wine as well.

1. **Baseline models**

Initially I applied seven different statistical models to the training dataset both for white wines and for red wines. That provided me with the initial feeling on which model would perform better and which one I should focus in order to improve afterwards.

For white wines

For red wines

The tables on the right displays the accuracy scores and in which my decision was made, the reason for that is because they have cross\_validation and they will avoid overfitting on the test dataset as evidnced below. The Gradient Boosting Trees was the chosen model to be improved upon.



1. **Conclusion**

What attributes in terms of chemical components contributes for a good wine? What would make the best wine possible? With which precision degree could we classify if a wine (from the test dataset) is good or bad based on its chemical components? What chemical components are more relevant?

* Refer to section 4, feature selection for the relevant chemical components, or the graphs below;
* Baseline model section;

Achieve the highest accuracy possible in the White wine and Red Wine dataset, improve the model to its fullest, desired precision would be more than 80%? Try to predict a wine quality from another region! Almost 80% for the red wines, after some parameter tuning. The final optimized model result is shown below.