Wine Quality Analysis

by Samuel Venzi Lima Monteiro de Oliveira

Introduction

In the final Lab, we will analyze the Wine Quality dataset (taken from http://archive.ics.uci.edu/ml/datasets/Wine). The dataset is actually divided into two datasets, a white wine and a red wine dataset. As described by the creators of the dataset it has 12 attributes, of which 11 are input attributes such as pH, acidity, residual sugar etc, based on objective physicochemical tests; and 1 is the output attribute of quality based on sensory data of wine experts.

The first step will be to perform the exploratory data analysis (EDA) of the dataset to search for trends in the data that can help us understand better how the attributes are related and see general statistics regarding some of the attributes. After that, we will use the supervised learning classification algorithm k-Nearest-Neighbor to try to predict the quality of the wine.

The analysis of the red wine and white wine is are the same but were done separately.

Methodology

The attributes of the dataset are the following:

```
['fixed acidity', 'volatile acidity', 'citric acid', 'residual sugar', 'chlorides', 'free sulfur dioxide', 'total sulfur dioxide', 'density', 'pH', 'sulphates', 'alcohol', 'quality']
```

The white wine dataset has 4898 instances and the red wine dataset has 1599 instances.

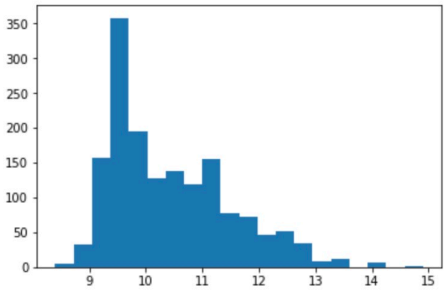
In this project, we performed analysis of histograms of the, arguably, most defining characteristics the may influence the evaluation of the quality of wine. Mainly, the attributes: alcohol, residual sugar, pH, free sulfur dioxide. Also, we see the average value and the median of each of these attributes.

We plotted a correlation matrix between each of the attributes, which gives us a quick visualization of how the attributes are linked to each other in pairs, allowing for a close look at each pair, which will be discussed more in the results section.

The quality attribute ranges from 0 (very bad) to 10 (very excellent), and we categorized them arbitrarily in three categories: bad wines [0,5), normal wines [5,8), good wines [8,10]. The dataset classes are ordered and not balanced (there are much more normal wines than bad or good). Then we defined a classifier model using kNN method so we can predict in which of the categories (bad, normal or good wines) the input wine characteristics are. This makes it a classification task instead of a regression task.

Results

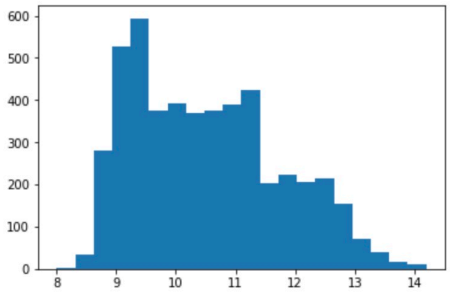
RED WINE



MEAN = 10.42%
MEDIAN = 10.2

ALCOHOL HISTOGRAM

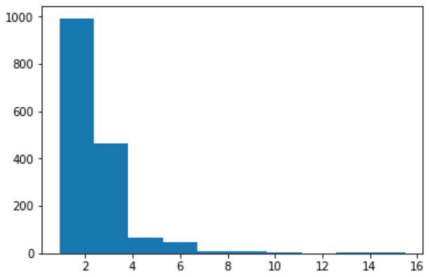
WHITE WINE



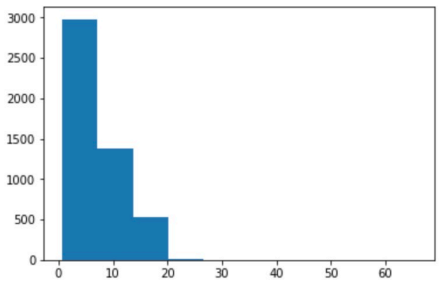
MEAN = 10.51%
MEDIAN = 10.4

RED WINE

WHITE WINE

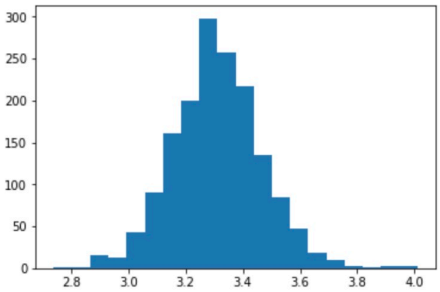


RESIDUAL SUGAR HIST.

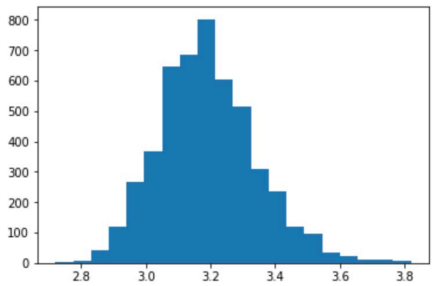


MEAN = 2.53
MEDIAN = 2.2

MEAN = 6.39
MEDIAN = 5.2

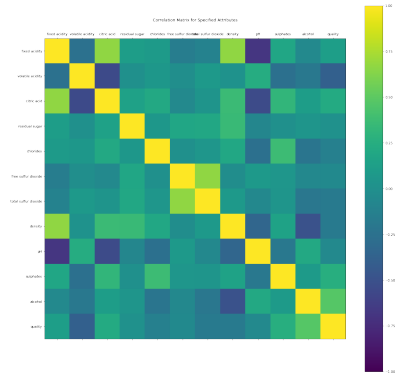


PH HISTOGRAM

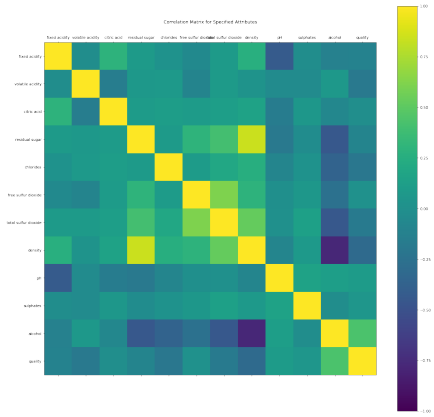


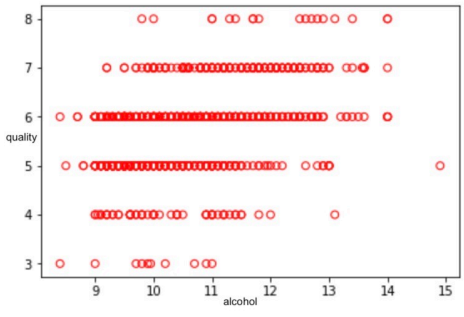
MEAN = 3.31
MEDIAN = 3.31

MEAN = 3.18
MEDIAN = 3.18

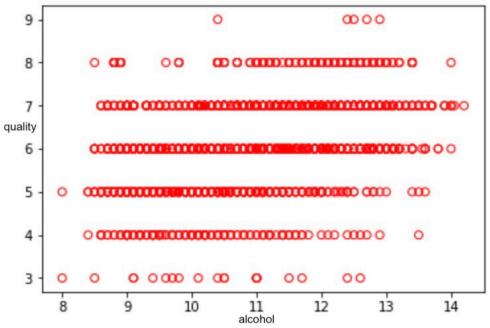


CORRELATION MATRIX





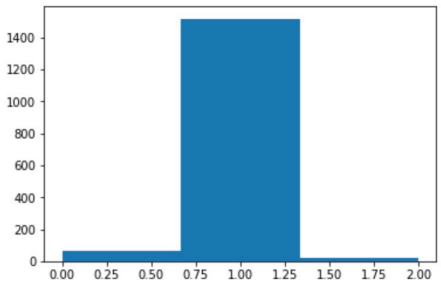
SCATTER PLOT OF
ALCOHOL AGAINST QUALITY



CORR = 0.47

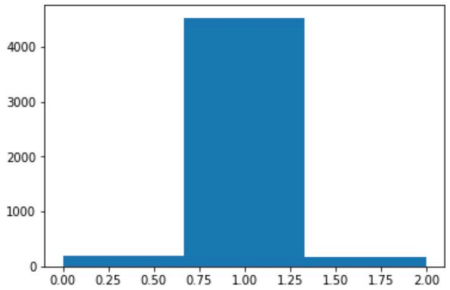
CORR = 0.43

RED WINE



WINE CATEGORIZATION INTO BAD, NORMAL AND GOOD

WHITE WINE



Both wine types have similar values of alcohol levels, but when it comes to residual sugar levels the white wine has almost 3 times more and that might account for a sweeter taste of the wine. Red wine is less acid than white wine in general.

In the correlation matrix we can see in yellow the pairs of attributes that are positively related and in blue the ones that are negatively related. In the file attached to this report it's possible to see some of these pairs analyzed. In the correlation between quality and alcohol we can see that they are moderately correlated, so wines with greater alcohol percentage have a tendency to be better classified.

Finally, the categorical division made reveals that the dataset is indeed unbalanced and that certainly influences results and accuracy of the supervised learning algorithm.

The kNN k parameter for the red wine dataset that showed the best accuracy is 1 and the accuracy of approximately 95%, but that doesn't mean it is a good result because the dataset has a lot of instances in one category. For the white wine dataset the k parameter that showed better accuracy was also 1, with an accuracy of 92%.

Conclusion

In this project we used EDA to to check general characteristics of the dataset and see the statistics of certain attributes. We also used a supervised learning algorithm (kNN) to predict the overall quality of the wine according to its objective characteristics.

Reference

- **https://www.kaggle.com/piyushgoyal443/red-wine-analysis/notebook**
  - I used this as an idea of what to analyze in the dataset
- **http://archive.ics.uci.edu/ml/datasets/Wine** (dataset download page)
- **https://www.kaggle.com/vijayprayagala/correlation-heat-map-and-scatter-matrix**
  - Helped with the correlation matrix heat mapping