Red Wine Quality Prediction

Riccardo Bellio

Abstract

This **Red Wine Dataset** is a public and open dataset in .csv format from the famous site **Kaggle**. It relates to the red variant of the Portuguese "Vinho Verde" wine and includes **1599 instances** and **12 attributes**:

My main question in this project is: *Is it possible to create a model to predict wine quality based on its physicochemical values and be able to purchase wine online with confidence, without having to read any wine reviews?*

My method follows the steps outlined in the **data science process** presented in this course. Starting from data collection/exploration, it continues with data preparation and model creation/evaluation. The project ends with the presentation of my conclusions. This is a reiterative process in which visualisation was used to explore data/present findings.

After testing various models, my Decision Tree Classifier predicted red wine quality with an **accuracy of app. 86%** (more insights are included in this presentation).

Motivation

This dataset got my attention because I worked as a restaurant manager for various years in Sydney (Australia). Selecting wine for my venues was one of my duties and wine selection in restaurant environments was the result of wine tasting. I often wondered: *how is it possible to choose a good wine without trying at least a sip of it and buy wine online or in store with confidence*?

This dataset gives me the chance to put data science into practice and create a model to predict wine quality based on its physicochemical values and quality evaluations: *can I do it? Is it possible to find any correlation between pairs of variables? Is wine quality associated with any specific attributes*?

How amazing would it be to possess a model for wine prediction and be able to purchase wine without having to read any reviews? Do we really need wasting our time looking at the screens of our mobiles in search for a reliable review… while we are holding a bottle of wine in hour hand? How confusing can it be looking at all those bottles? So many options!

Follow me on this journey and let's see if we can pour some good data science into our tall wine glasses!

Dataset Overview

This public and open dataset in .csv format relates to the red variant of the Portuguese "Vinho Verde" wine and includes **1599 instances** and **12 attributes**:

- 1 **Output variable**: the **Wine Quality**, represented by the average score given by at least 3 wine experts, with a score between 3 (very bad) to 8 (excellent).
- 11 **Input Variables**, containing the results of different physicochemical tests:
    - **Fixed Acidity** (Tartaric Acid, g/dm^3): A measure of wine acidity and the term fixed refers to the fact these acids do not evaporate readily.
    - **Volatile Acidity** (Acetic Acid, g/dm^3): The amount of acetic acid found in wine. If it is too high, it can lead to an unpleasant, vinegar taste.
    - **Citric Acid** (g/dm^3): In small quantities, it can add a fresh flavour to wines.
    - **Residual Sugar** (g/dm^3): What remains after grapes have gone through the winemaking process. It ranges between 1 and 45 gram/ litre (sweet wine).
    - **Chlorides** (Sodium Chloride, g/dm^3): The amount of salt in wine.
    - **Free Sulfur Dioxide** (mg/dm^3): It is used to prevent microbial growth and the oxidation of wine.
    - **Total Sulfur Dioxide** (mg/dm^3): SO2 is used as a preservative because of its anti-oxidative and anti-microbial properties in wine, but also as a cleaning agent for barrels and winery facilities.
    - **Density** (g/cm^3): In wine, density is affected by alcohol, sugar, glycerol, and other dissolved solids.
    - **Ph**: It describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic). Most wines range from 2.5 to about 4.5 pH, and 7 is neutral.
    - **Sulphates** (Potassium Sulphate, g/dm3): An additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant.
    - **Alcohol** (percentage by volume): The percent of alcohol content in wine.

For more information check the notebook and the “Wine Knowledge” link on the Reference section.

Data Preparation & Cleaning

After importing the libraries needed for this project, data was ingested into a Pandas dataframe and explored. The application of different methods (including statistical analysis and visualisation techniques) provided information for data preparation:

- **Null Values**: The dataset did not include any null values.
- **Unnecessary Columns**: The column “Unnamed: 0” was removed because it was a copy of our index, thus unnecessary.
- **Column Headers Format**: The column headers were improved to make them easier to read. They were capitalised and the “.” used to separate words in the headers was changed with an empty space for clarity.
- **"Quality" column**: A quality ranging between 3 and 8 was confusing. It was turned into a 0-5 range (0=very bad, 5=Excellent). After collecting information about the dataset, in order to prepare the dataset for the modelling phase, the **numerical attribute “Quality”** (ranging from 0 to 5) was turned into a **categorical attribute “Quality Cat”**, with the values: Bad (quality = 0,1), Average (quality = 2,3), and Good (quality = 4,5).
- **Outliers**: After analysing them using scatterplots and checking if their values were in range, they were not removed because it appeared that they represented legitimate observations.
- **Features**: The dataset had a limited number of features (12). Some co-dependencies between pair of features were observed. The first **Decision Tree model** was tested using all the features, then some of the features were removed to see if the observed co-dependencies were limiting its accuracy. The features importance scores were calculated and since they had a limited range (min = 0.075, max = 0.155), removing these features from the model was not expected to highly improve its accuracy. The **K Nearest Neighbours model** was created using all the features because it uses a normalisation process.

My Research Questions

As stated in the introductory section, my main research question is:

Can I create a model to predict wine quality based on its physicochemical values and be able to purchase wine online with confidence, without having to read any wine reviews?

Methods

In order to analyse the dataset and collect important information about it I used the following methods:

- **.info( ) and .head( )**: To print information about my dataframe (including the index dtype, columns, non-null values and memory usage) and to visualise the first 5 rows of the dataset.
- **.unique( )**: To visualise the unique values of our output, the column “quality”.
- .**describe( )**: To view some basic statistical information (such as min, max, std, mean, 25%, 50%, 75%) about the numeric values in my dataframe.
- **.corr( ) and .plot(kind=“hist”)**: To compute pairwise correlation of columns, excluding NA/null values and visualise the data distribution of the attribute “quality”.
- **sns.heatmap( )**: To plot a summary of the histograms / visualise a summary of the data distribution of our attributes (except “quality”).
- **sns.regplot( ):** To visualise correlations and co-dependencies between variables using scatterplots, and pay attention to the presence of outliers.
- **sns.barplot( )**: To plot a barplot and visualise the relationship between Alcohol and Quantity.
- **sns.pairplot( ):** To visualise a summary of all our variable correlations (using histograms) and show trends and the presence of outliers.
- **pie.plot( )**: To visualise using pie charts a comparison of the percentages of numerical values and categorical values of “Quality".
- **model.feature_importances_**: To get the value of the importance of each feature in our dataset (the higher the score, the more relevant is the feature towards our output variable).

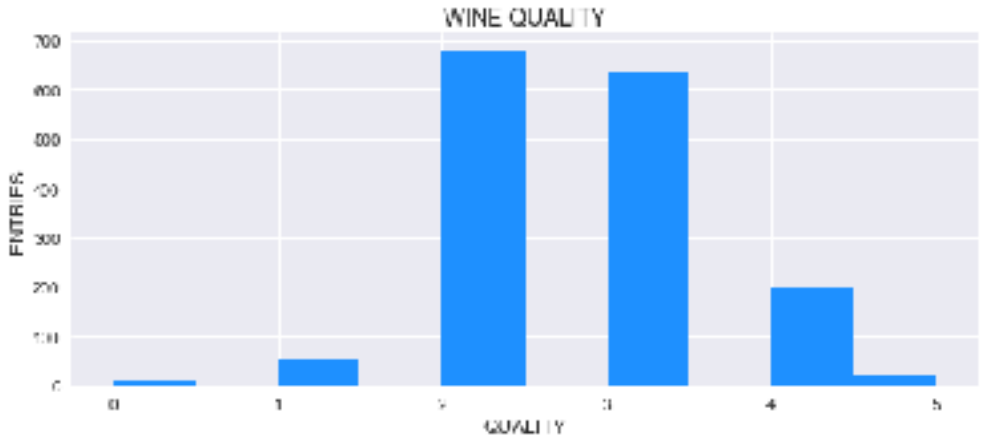
Models were created using **Decision Tree** and **K Nearest Neighbours** classifiers. The data set was split into **Train** and **Test sets** to train the data (test set) and make predictions (test set). The final predictions were compared with our original output, and model accuracy was calculated. In this process, the model’s parameters were optimised.

Findings

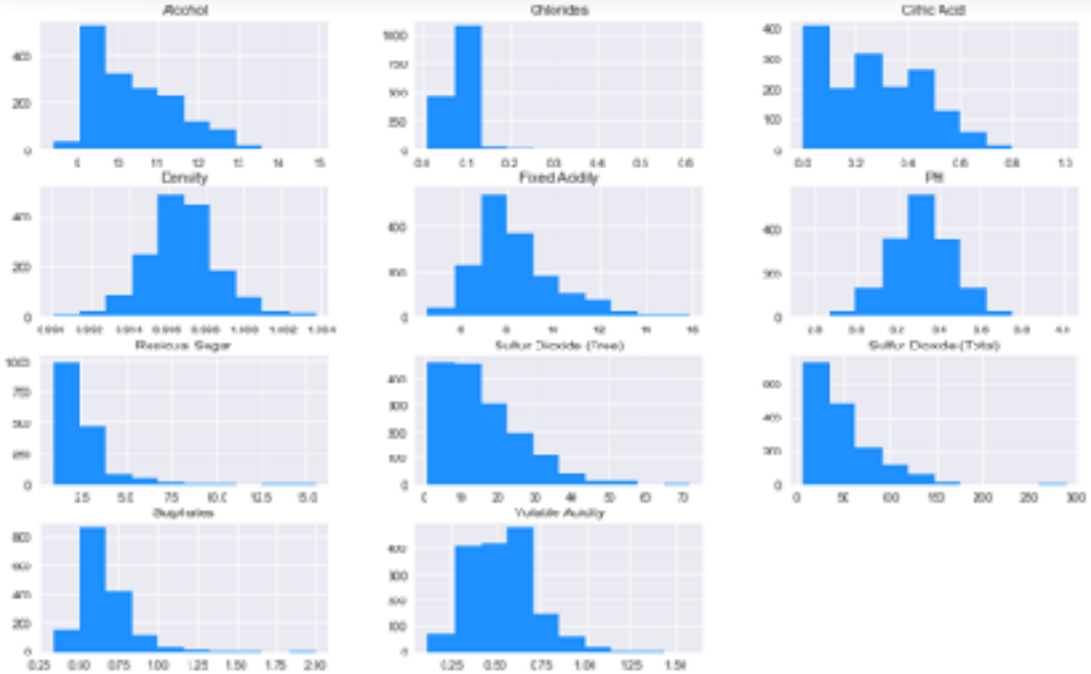
The quality distribution of our data shown in this histogram indicates that our dataset mainly contains samples of red wine rated 2 or 3 (“average” wine). Only a small quantity of data relates to “Bad wine” (value: 0) or “Excellent wine” (value: 5).

The dataset is “skewed”, and its margins may act as outliers for our model and negatively affect its accuracy.

Decision Trees and **K Nearest Neighbours models** include data transformation, thus limiting the negative impact of this distribution.



Findings



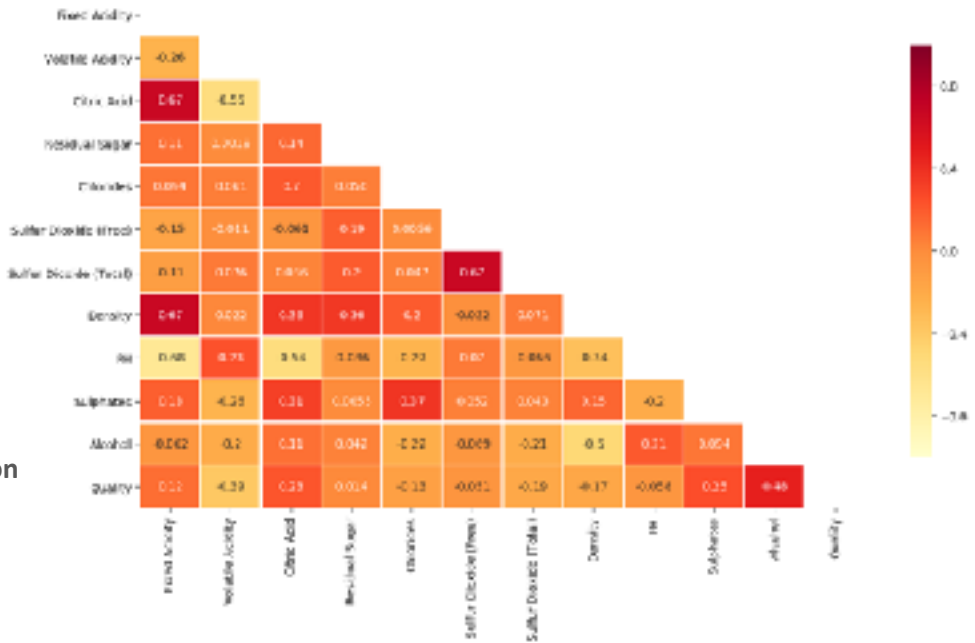
Some attributes show a similar distribution and probably co-dependencies.

Their potential correlations will be further explored numerically and visually.

The data distribution of these attributes shows the presence of outliers.

Findings

This **correlation matrix** was used to better visualise the co-dependencies between attributes: the **darker** the square **colour**, the **higher** the **value**, and the **stronger** the **positive relationship** between features.
Vice versa, the **lighter** the **colour**, the **lower** the **value**, and the **stronger** the **negative relationship** between features.
A value close to zero show almost no co-dependency between features.

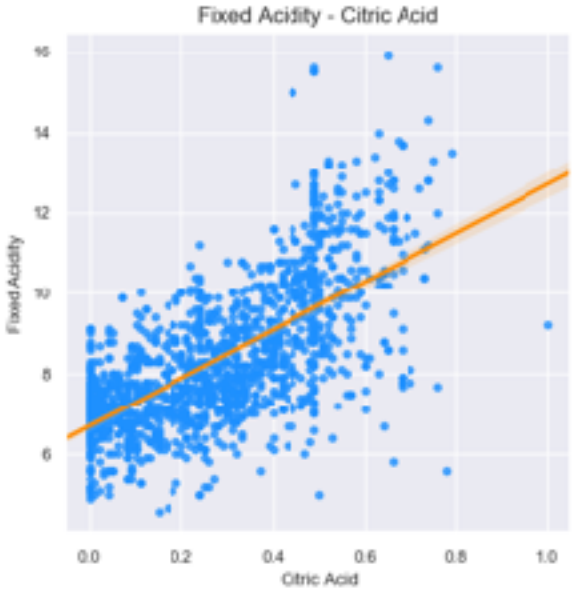


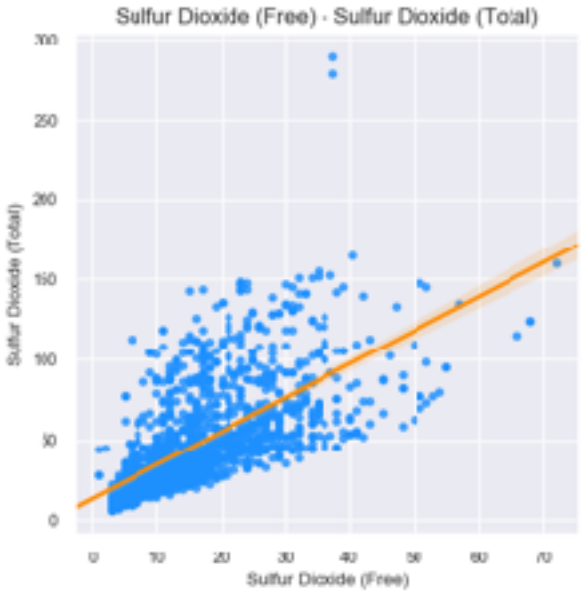
The heat-map shows that:

- **Fixed Acidity** and **PH** show a **relative negative correlation** (-0.68).
- **Fixed Acidity**, **Citric Acid** and **Density** show a **relative positive correlation** (both +0.67).
- **Sulfur Dioxide** (Free) and **Sulfur Dioxide** (Total) show a **relative positive correlation** (+0.67).
- **Citric Acid** and **Volatile Acidity** show a **relative negative correlation** (-0.55).
- **Citric Acid** and **PH** show a **relative negative correlation** (-0.54).
- **Alcohol** and **Quality** show a **relative positive correlation** (+0.48).

Findings

The correlations identified plotting the heatmap were visualised using scatterplots.





Here we have two examples of relative positive linear relationship between pairs of variables.

The outliers visible in these scatterplots refers to the wine that was rated as “bad” or “good”.

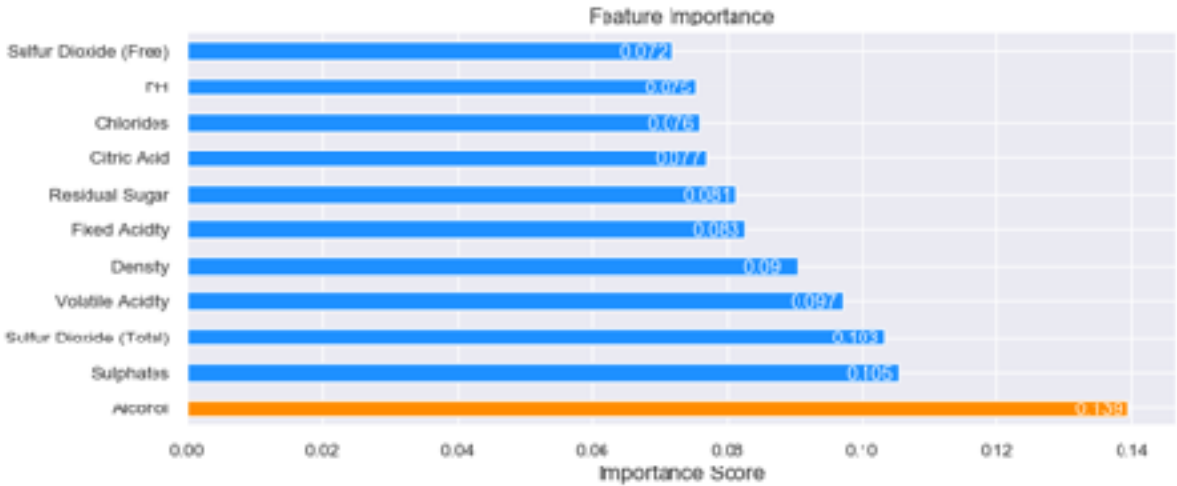
Findings

The importance of each feature towards the output variable in our dataset was obtained using the method **feature_importance_ property** (available on the model **ExtraTreesClassifier**).

The scores returned by the function were plotted to visualise a score for each feature.

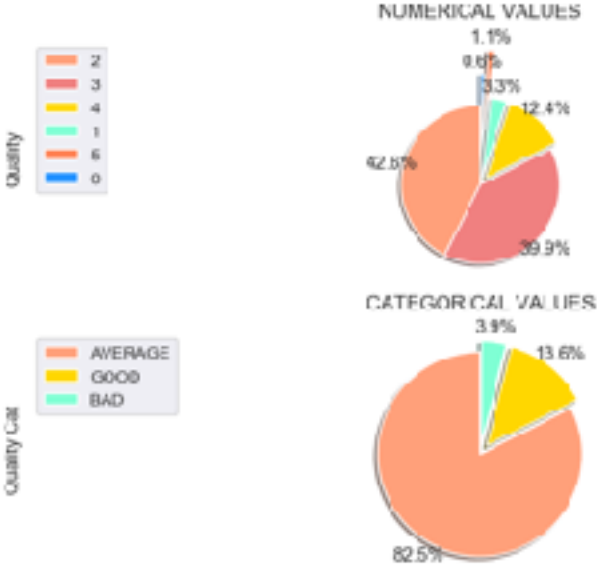
Alcohol content is the most relevant feature towards our output variable.

However, the value range for feature importance is small (from 0.072 to 0.139), an indication that the features posses a similar importance.



Findings

The output variable “Quality” was **numerical** and had to be turned into a **categorical** variable in order to continue with our classification problem.



Since I am interested in purchasing ”good” and “average” wines (if they are cheap), the categories that I chose are:

- **Bad** = Quality values for 0,1.
- **Average** = Quality values for 2,3.
- **Good** = Quality values for 4,5.

These pie charts compare the quality values before and after the conversion. As we can deduct by comparing the percentages of the two charts, the conversion was successful.

Findings

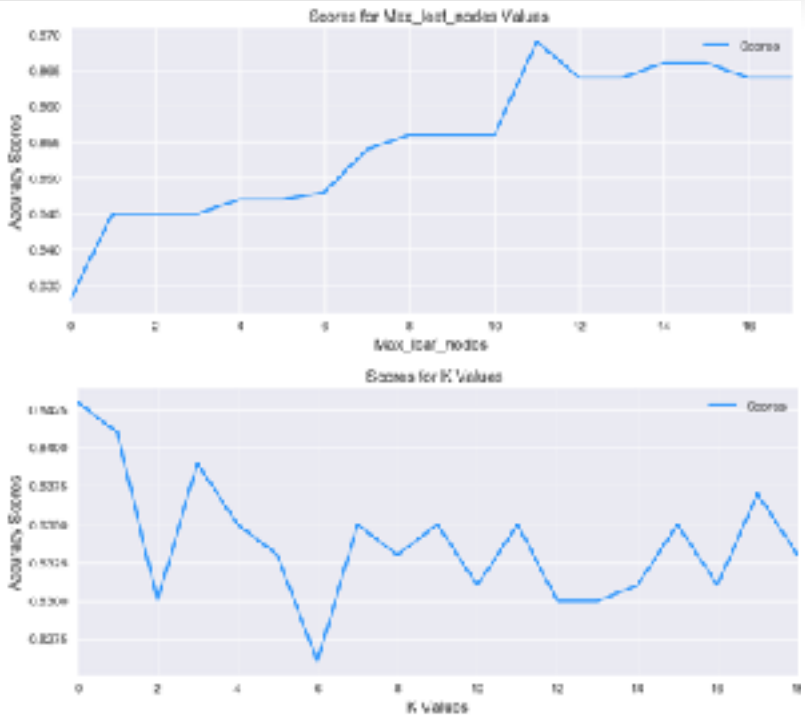
In our **Decision Tree Model**, the parameter **max_leaf_nodes** has an impact on its final accuracy. To find the bast value for this parameter I:

- Increased the values of max_leaf_nodes (in a for loop) for the same model and the resulting accuracies were stored in a list.
- This list was converted into a dataframe, which was used to plot the values for max_leaf_nodes in relation to the the accuracy score.

As we can see, **the first chart** shows that the **model's accuracy peaks when the parameter max_leaf_node value is equal to 11.**

The algorithm **K-Nearest Neighbours** is based on distance of cases and data standardisation was recommended. In this case, the parameter **K**, which is equal to the number of nearest points to the data that has to be predicted, can impact the quality of our model. In order to find its optimal value, we followed the same steps to find the best value for max_leaf_nodes outlined before.

As we can see, **the second chart** shows that the **model's accuracy peaks when the parameter k is equal to 1.**



Limitations

This dataset presents some limitations:

- It does not include grape types and wine brand, and it make it harder researching wine online and choosing the grape variety that we like the most.
- Its number of samples is limited (1599) and their distribution is skewed.
- It only refers to the red variant of the Portuguese "Vinho Verde" wine. For this reason, the final model can be used to only predict this type of wine.

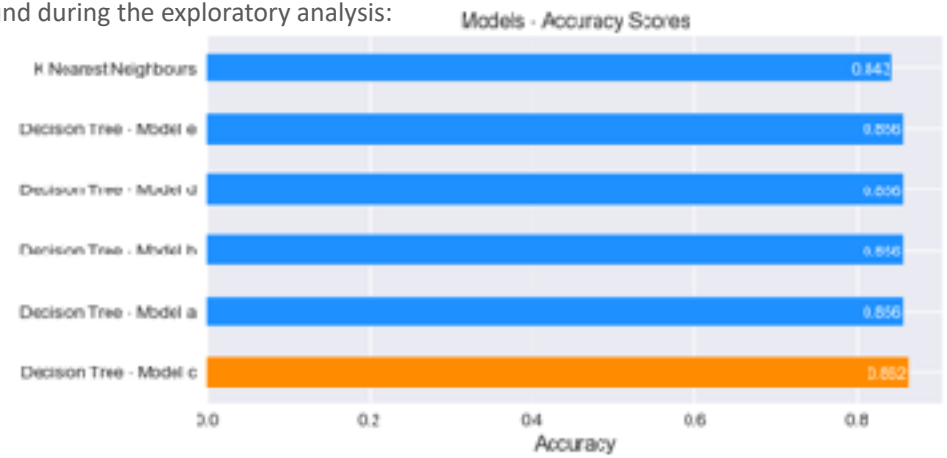
The only algorithms that I used in this project are: Decision Tree and K Nearest Neighbours Classifiers:

- I chose to use these algorithms because they are the ones that I learnt on this course on Edx.
- Future projects will include other algorithms that I will learn soon, to check if they perform better in this situation. Other algorithms are available for classification problems (such as: Naive Bayes, Support Vector Machine, and Logistic regression).

Conclusions

In this project I tested 1 model using a **K Nearest Neighbour Classifier** and 5 variations of the same model using a **Decision Trees Classifier,** while changing its attributes in relation to the **co-dependencies between variables** found during the exploratory analysis:

- Model a: It used all the available features.
- Model b: Removed Fixed Acidy.
- Model c: Removed Sulfur Dioxide (Total).
- Model d: Removed Citric Acid.
- Model e: Removed Fixed Acidity, Sulfure Dioxide (Total), and Citric Acidity.



As shown on the right inside the plot containing the **accuracy scores** for our different models, the final predictive model is the one created using a **Decision Tree Classifier without the feature Sulfur Dioxide (Total)**. This is the model that allows me to choose the red variant of the Portuguese "Vinho Verde" wine with an accuracy of app. 86%.

I am satisfied with this result, and I intend to review this process in the future. A new and more complex red wine dataset, together with an advanced knowledge of different algorithms and data science processes, will allow me to improve this result and to be able to predict wine quality for a wider variety of wines, choose the grape variety that I prefer and address specifying wine brands.

Acknowledgements

- The data was collected from the website kaggle: **Red Wine Dataset**
- Relevant Publication: 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: 0167-9236.2009. Available at: **[@Elsevier]**

References

- Data Science Process: **UC San Diego - Python for Data Science** - **IBM - Data Science Professional Certificate**
- Wine Knowledge: **Understanding acidity in Wine**
- Feature Selection: **Feature Selection with real categorical data** - **How to Choose a Feature Selection Method For Machine Learning**
- Skewed Data (Medium articles): **Skewed Data**
- Outliers Analysis: **Outliers to drop or not to drop**
- Plots(Medium articles): **How to create a seaborn correlation heatmap in Python** - **Pairplot VIsualisation**