

ML Algorithm Selection

Selections on Vinho Verde Red Wine

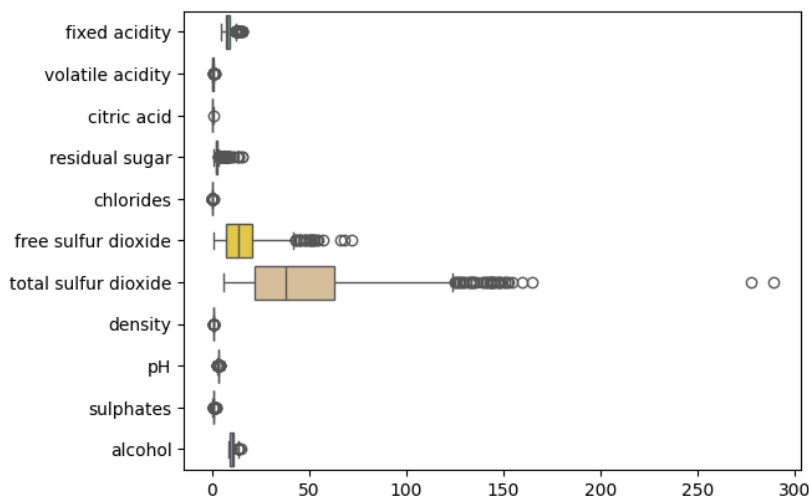
Maxie Machado
University of Wyoming

1 Introduction

The exercise being performed is ML Algorithm Selection on the red wine quality dataset. As it's in the name we will have a collection of red wines from vinho verde which are rooted from Northern Portugal, and will be looking at different factors which will affect the quality of the red wine positively or negatively. Different machine learning algorithms will be done on the dataset, and will have an accuracy score done on them and be visualized using heatmaps. This dataset will be put through three different ML algorithms including, decision tree classifier, random forest classifier, and logistic regression.

2 Description and Cleaning of Red Wine Dataset

The red wine dataset consists of 1599 observations. Another thing which is important to note is the amount of features there are within the dataset and what they are. Within the red wine dataset there are 11 features, which include the following:

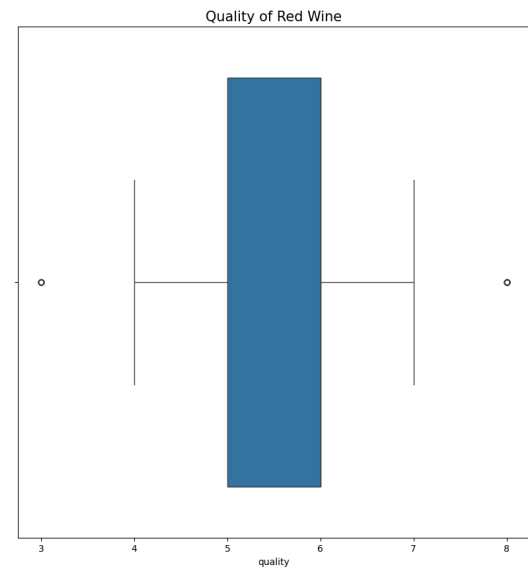


- Fixed Acidity
- Volatile Acidity
- Citric Acid
- Residual Sugar
- Chlorides
- Free Sulfur Dioxide
- Total Sulfur Dioxide
- Density
- pH
- Sulphates
- alcohol

In examining the data type of the red wine it will be shown all of these are float64 or int which mean we will not need to worry about converting objects. Although the dataset still needs to be cleaned. For this the first step is checking to see if the dataset of red wine has any missing values.

For this particular dataset there are no missing values which need to be taken care of. Moving on, the dataset will be checked for duplicates. As shown the dataset does indeed have some duplicates, 240 to be exact. Once the duplicates are handled the untouched data will be visualized using a simple boxplot [fig 1]. Now the dataset will be checked for outliers within the quality by using z-score calculations. As shown there are a lot of outliers that could possibly affect the accuracy of the data. Looking specifically at the quality column of red wine, using a

simple boxplot, it's shown there are outliers on both sides [fig 2]. The dataset will be analyzed to see the effects of outliers that are greater than 7 and less than 4. Starting with outliers greater than 7, we will filter the dataframe. Once completed the calculation of the percentage of the total outliers over 7 will be done. As shown 1.25% of total quality is over 7. Which is significant enough to make the accuracy of the data less. Now looking at outliers less than 4 using the same methods, will show that 0.74% of quality is less than 4. Which is also significant enough to make the accuracy of the data less. Once the duplicates are handled and after analyzing the effects of the outliers we will look at the amount of observations and features again. As suspected, removing the duplicates has the amount of observations changed from 1599 to 1359 observations. This cleaned data is what we will be using to split and train the data. Although before splitting and training the data, list comprehension will be used to create the binarization of the target variable, quality. Once this is completed the dataset will be split and trained.



3 Processes Performed of Red Wine Dataset

Now that the red wine data has been cleaned the different ML algorithms can be performed. The first ML algorithm that will be performed is the decision tree classifier. The accuracy of this will be calculated to see how well this method works on the dataset. Then it will be visualized using a heatmap from a confusion matrix. Next random forest classifier will be performed on the dataset. Like before, the accuracy of this will be calculated to see how well this method works on the dataset. Then it will be visualized using a heatmap from a confusion matrix. Lastly, logistic regression will be performed on the dataset. And just like the previous ML algorithms, the accuracy of this will be calculated to see how well this method works on the dataset. Then it will be visualized using a heatmap from a confusion matrix.

3.1 Decision Tree Classifier on Red Wine Dataset

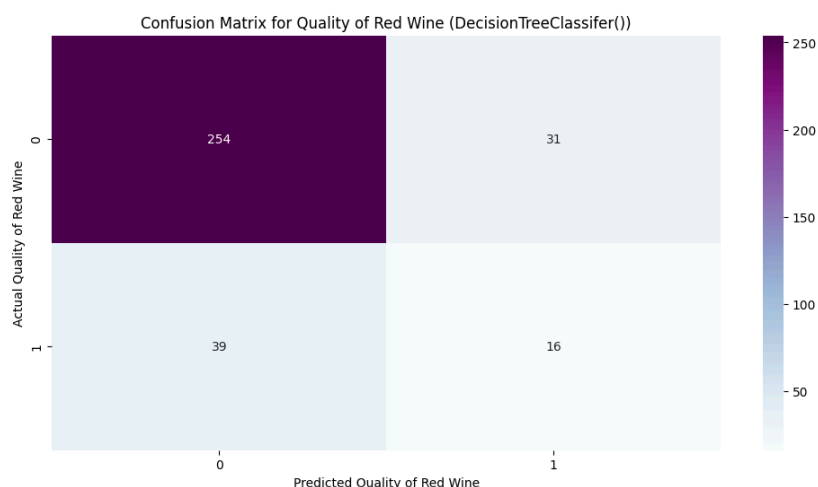
254	31
39	16

The first ML algorithm to be performed is the decision tree

classifier. For this we will be using the cleaned data that is split and trained. We will then do a prediction using the decision tree classifier. Once this is completed the accuracy score of this will be calculated. As shown, the calculation is 79.41%.

Meaning that the decision tree classifier and the predictions from it

is about 79% accurate. Now from this the confusion matrix will be created [fig. 3]. Now to better visualize this a heatmap will be used [fig. 4]. Out of all the ML algorithms being performed on the red wine dataset, this is the weakest when it comes to accuracy.



3.2 Random Forest Classifier on Red Wine Dataset

263	22
42	13

The second ML algorithm to be performed is the random forest

classifier. For this we will be using the cleaned data that is split and trained. We will then do a prediction using the random forest classifier. Once this is completed the accuracy score of this will be calculated. As shown, the calculation is 81.18%.

Meaning that the random forest classifier and the predictions from it

is about 81% accurate. Now from this the confusion matrix will be created [fig. 5]. Now to better visualize this a heatmap will be used [fig. 6]. Out of all the ML algorithms being performed on the red wine dataset, this is the most moderate when it comes to accuracy.



3.3 Logistic Regression on Red Wine Dataset

267	18
39	16

Lastly, the ML algorithm to be performed is logistic

regression. For this we will be using the cleaned data that is split and trained. We will then do a prediction using logistic regression. Once this is completed the accuracy score of this will be calculated. As shown, the calculation is 83.24%. Meaning that logistic regression and the predictions from it is about 83%

accurate. Now from this the confusion matrix will be created [fig. 7] . Now to better visualize this a heatmap will be used [fig. 8]. Out of all the ML algorithms being performed on the red wine dataset, this is the strongest when it comes to accuracy.

