Josue Huitznahui Bolaños Mejía A01207438

Decision Trees for predicting alcohol in Wine.

Introduction

Being able to predict the level of alcohol before producing a beverage can be a really important factor, by doing it one can have an approximation of the consistency and taste about a product even before making any of it.

In this study we will be taking the basic characteristics used to identify the quality of wine to try and predict the alcohol quantity, this to identify if there is a strong relation between the alcohol percentage and the following characteristics:

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

The label we will be using is an identification if the selected wine posses at least 10% of alcohol or more, the value of 10% was selected as it was found to be approximately the average of alcohol the wines in the database possessed, making it so if it is true, then the wine belongs to the superior half of wines with strongest alcohol concentration.

For the method of analysis, a decision tree was selected, as one of the intentions is to identify the most important variables to take into account that could affect the level of alcohol in the beverage. A Decision Tree works by obtaining the Entropy of a System:

$$Entropy = -\sum_{n}^{i=1} Pi\ log_2(Pi)$$

This formula means that entropy is the negative summatory of all the possibilities inside the system, this information reveals us the number of “unknown” factors present in the system and let us have an idea of the correlation between the variables and the expected result.

Once the Entropy of the system is calculated, the next step is to find the most important variable, the one with the highest correlation with the result. This is calculated by obtaining the “Gain” of each variable. The formula for the Gain is as Follows:

$$Gain = EntropySystem - EntropyVariable$$

The Entropy of the variable is a little complex to obtain because one must divide the variable by subgroups first, each subgroup is composed of all the same values of the variable, meaning that if one of our variables is shape, one of our subgroups is circle, another could be triangle, and another could be square. Once divided in subgroups the Entropy evaluated to the label of the data is obtained, multiplied by its probability of happening in the whole system and added, finally obtaining the Entropy of the Variable.

When the Variable with the highest Gain is found, a sub system is created for every of its subgroups and the process is repeated until all variables are used or the entropy of the system reaches 0, meaning we have an absolute certainty that we can predict the result.

Hypothesis

We are working with the hypothesis:

> "It is possible to predict the quantity of alcohol present in wine by looking at the factors which decide the quality"

Meaning that there is an indirect relationship between the quality of a wine and its alcohol percentage.

Experiment

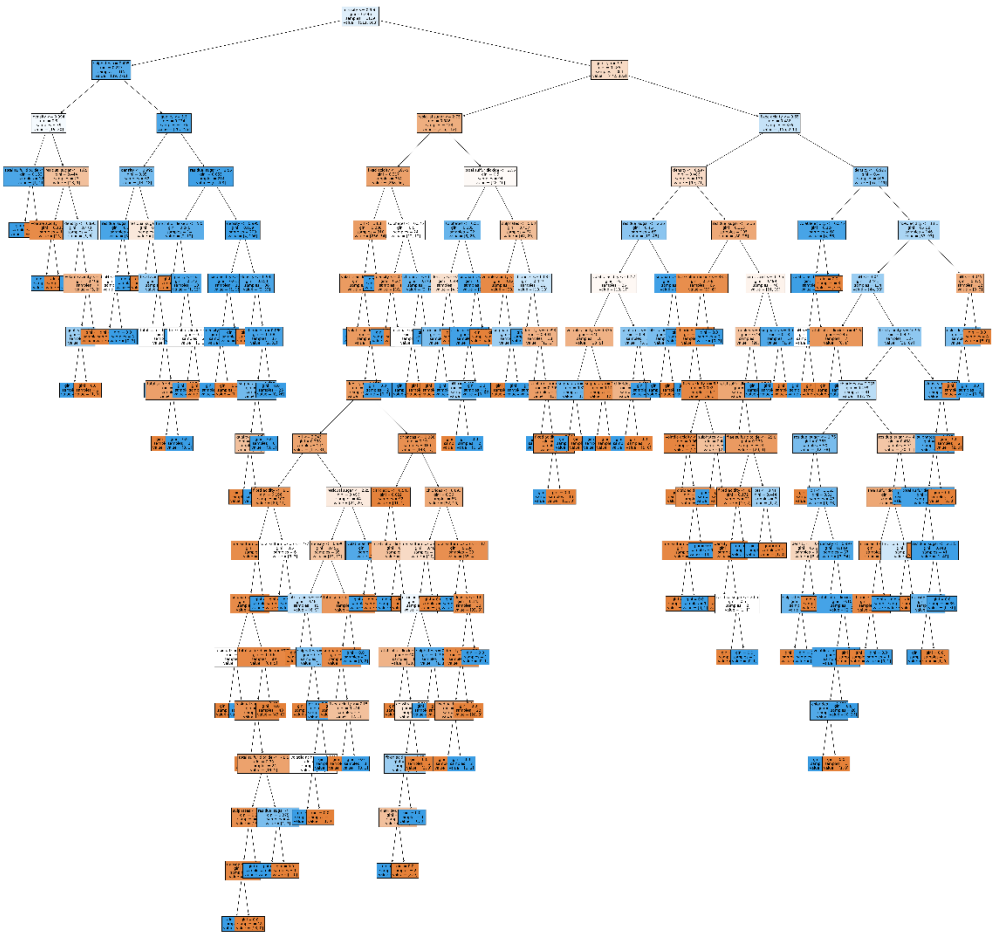
For the experiment, a program in python was developed, it was created using Pandas to extract and manage the data, and sklearn framework to process it and fit it into an already existing model.

The data was obtained from a public source.

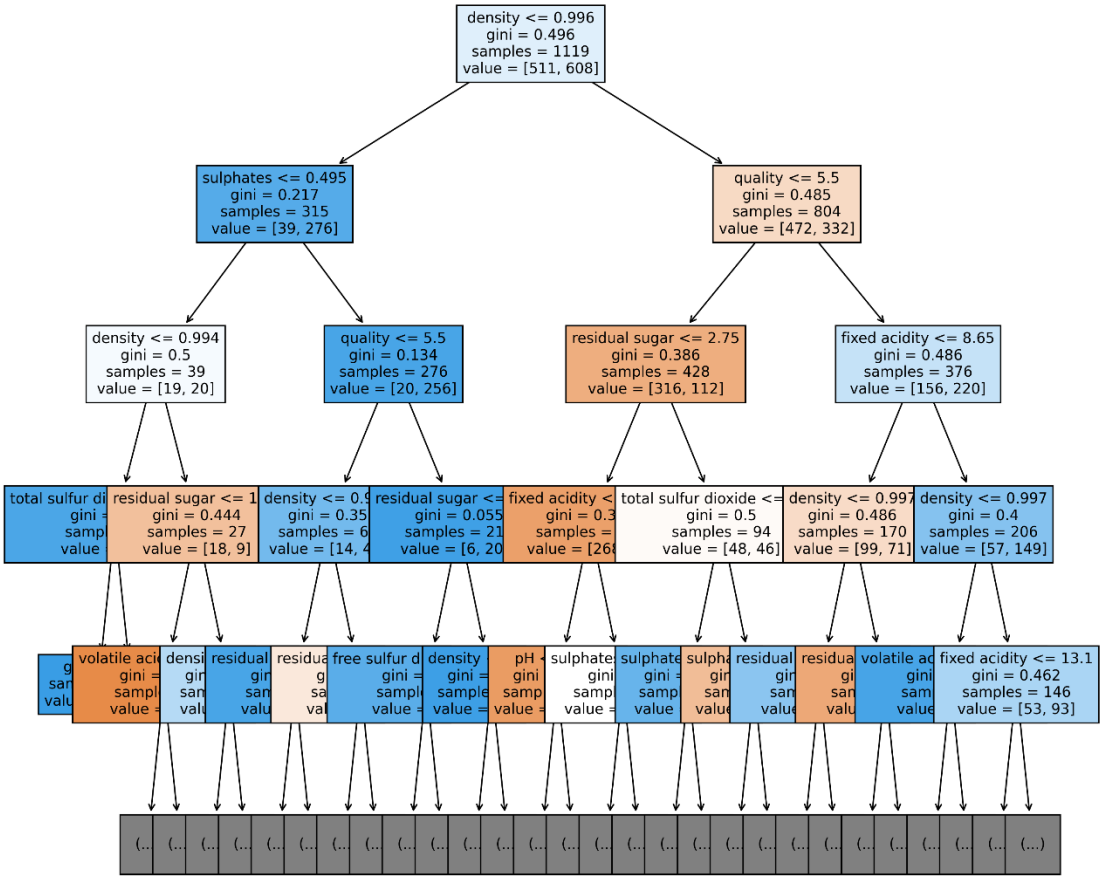
The experiment consisted of trying to compare results by using the data raw and create a tree and a random forest and preprocess de data and train a new Tre and a new Random Forest, then comparing the accuracy score for every one of the iterations.

For the first iteration, the Tree created with raw data the accuracy score obtained from the Evaluation was in average 78.58% higher than chance, but not having a strong enough correlation, for the produced tree the next image represents it.

Josue Huitznahui Bolaños Mejía A01207438



Josue Huitznahui Bolaños Mejía A01207438

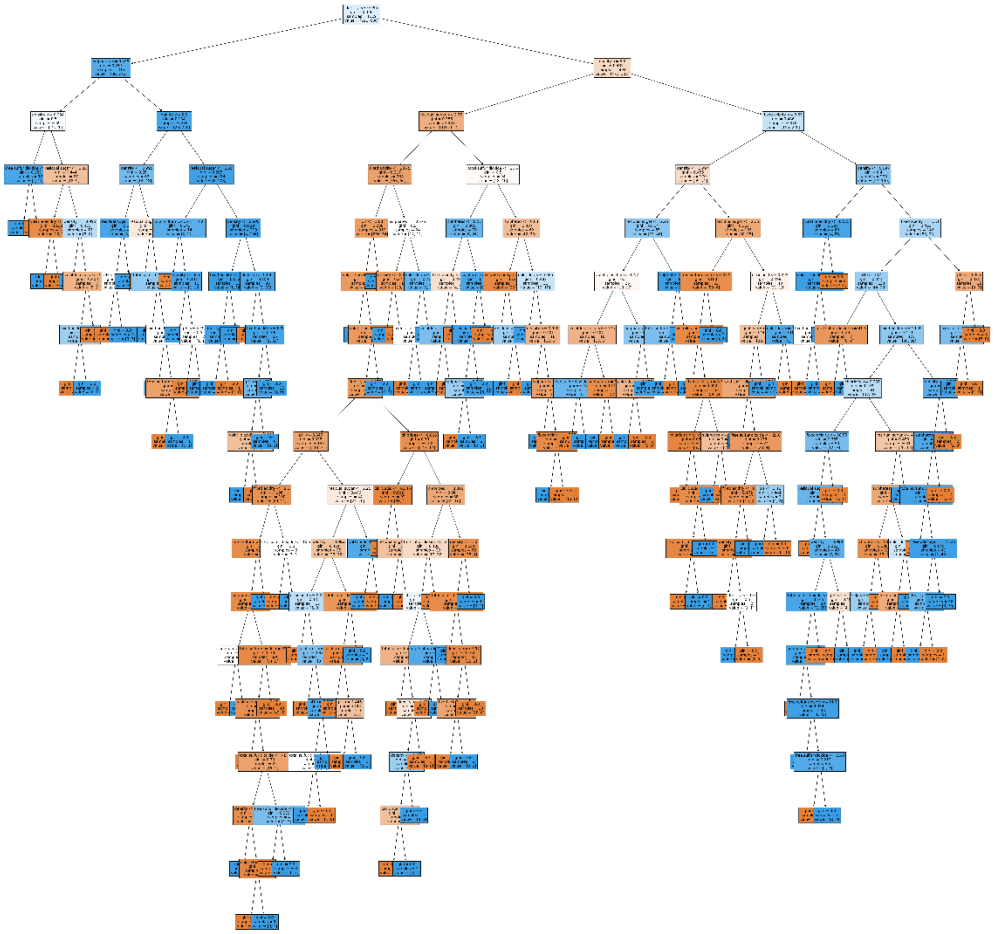


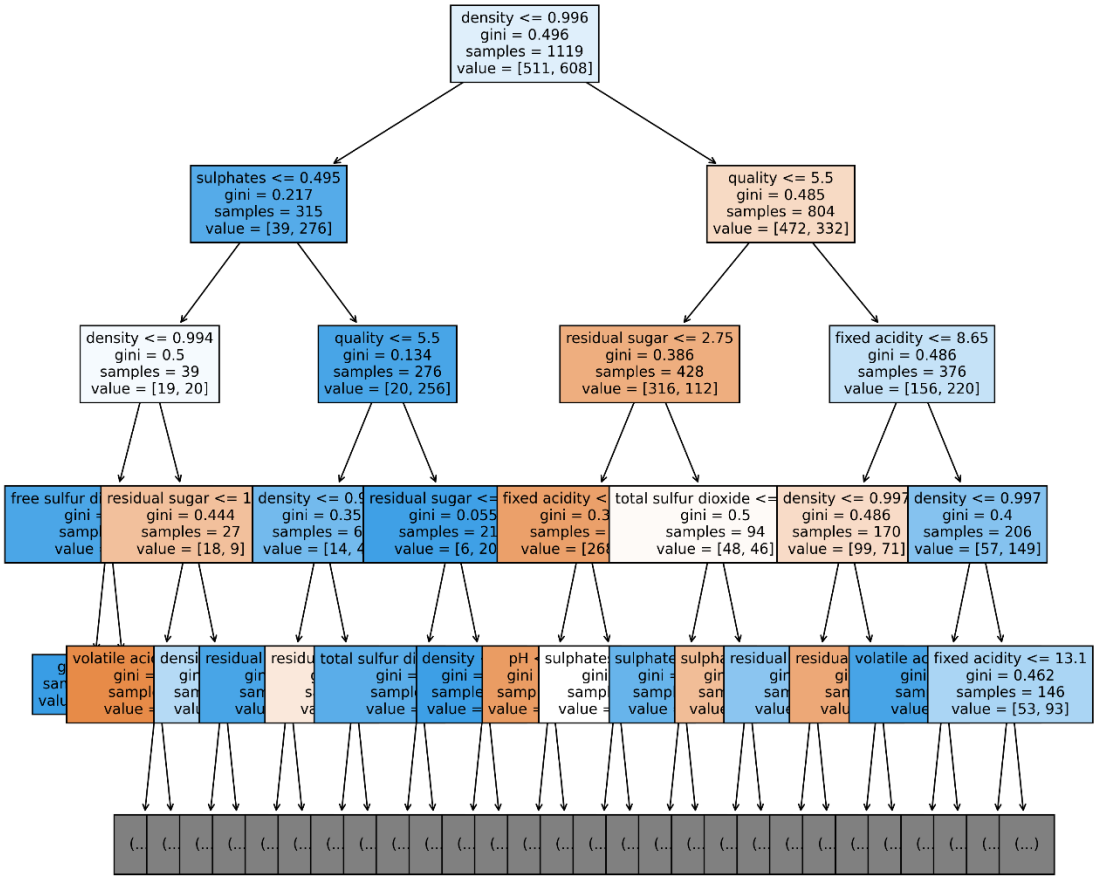
As seen in the pictures, the density and quality are two of the main variables when considering the correlation to alcohol levels.

The random forest produced with raw data has a slightly greater average accuracy of 79.54%, but still not that strong.

The tree with transformed data has an average accuracy of 78.70% and the tree looks as follows:

Josue Huitznahui Bolaños Mejía A01207438





We can see a similar correlation in the variables as we saw on the first tree. For the forest produced by this preprocessed data we can see a decrease in the average accuracy to 79.25% compared to the one with raw data.

Conclusion

With the results obtained from the experiment phase we can conclude that there is a slight correlation (around 78.5%) between the quality of the wine and the percentage of alcohol it contains, the density being the strongest of them all. We can also see that the preprocessing method chosen for the experiment is not significant enough to produce a real change in the results.