Vinho Verde Wine Quality Classification

Indra Rani Araujo[1]

[1]Computer Engineering - Universidade Federal do Pampa (UNIPAMPA)
96460-000 – Bagé – RS – Brazil

indrasantos.aluno@unipampa.edu.br

Abstract. *In an age where there is a vast amount of data being produced daily, it is necessary to create, improve and apply techniques which will make us understand better the situations in hand. Therefore, this project uses a Vinho Verde wine dataset to apply preprocess methods, classifications algorithms and models evaluations, for educational purpose upon data mining.*

1. Introduction

Even though data mining has not a very established meaning, it is possible to define it as a process that gathers some kind of pattern or information from a large amount of data [Han et al. 2022]. From the results of the data mining, it is possible to acquire knowledge upon the situation where the data was extracted. The data itself can be analyzed through different types of models such as statics techniques and machine learning algorithms [Han et al. 2022]. In most cases, the raw dataset is not analyzed, because it can contain values that are not stable, or they are simply not correct. Therefore, before data mining, the dataset goes through a cleaning and adjusting process, named preprocess. Because there are so many ways to analyze the data, after using a model there is a step of evaluation of the model used, to understand if that was the best model to use for the dataset or if the dataset itself is adequate to be analyzed.

To comprehend better the data mining process, this project goal is to preprocess, classify the data with two different classifier algorithms and compare and discuss both models with evaluation metrics. For this matter, the Google Colab environment was used with Python programming language.

This report consists of the following sections: i) description of the dataset used, ii) preprocess techniques applied, iii) classification algorithms implemented, iv) model evaluations, v) discussion of the evaluation results.

2. Dataset

To apply preprocess methods, classifications algorithms and models evaluations, the Vinho Verde red wine quality dataset was used. It was provided by Viticulture Commission of the Vinho Verde Region(CVRVV), Porto, Portugal and University of Minho, Guimarães, Portugal. The dataset consist of 1599 samples with 11 inputs and one output. The input are a variety of chemical wine attributes gathered from objective test, such as Ph values, and the output is a sensory-based value from professionals wine tasters. As it is determined, the goal with the data is to model the wine quality based on physicochemical tests.

The quality (output) of the wine was graded from 1 to 10 according to the sensory test. The range can be classified as poor (1-4), normal (5-6) and excellent (7-10).

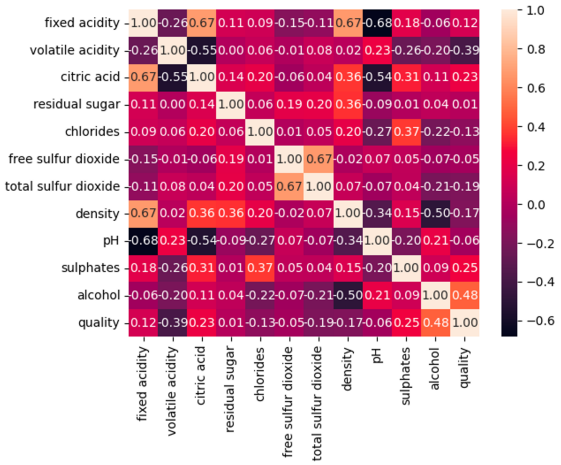
The dataset used can be found at UCI Machine Learning Repository: `https://archive.ics.uci.edu/ml/datasets/Wine+Quality`

3. Preprocess

According to the UCI Machine Learning Repository dataset page, the data is unbalanced, and some attributes may not correlate so much with each other. Thus, the dataset needed some adjustments, so the classification models could work better.

To figure out what were the attributes that had strong (negative or positive) correlation with the quality output, a correlogram was plotted (Figure 1). The attributes that showed weak correlation (between 0.2 and -0.2) were removed from the dataset, leaving only 6 attributes left.

Figure 1. Attributes Correlogram

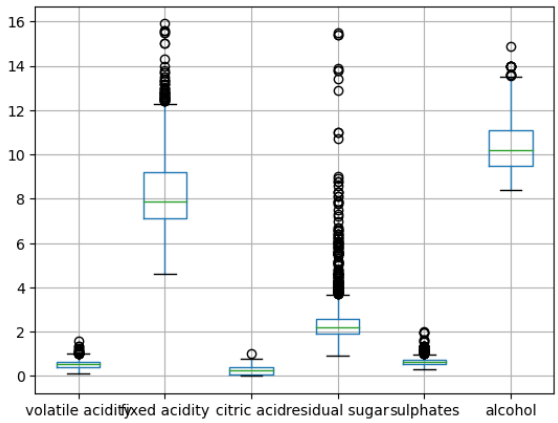


After some analyses, it was uncovered that the dataset had circa of 82.49% of normal wine samples, 13.57% were excellent and only 3.94% were poor quality. To work around the unbalanced samples, was plotted a boxplot for each of the attributes to analyse the samples with outliers. The plot shows that every attribute has a considerable amount of outliers (Figure 2), so the most outliers of it all were chosen to be dropped off of the dataset. With a loop through the dataset comparing each sample to a defined maximum value of the attributes (Table 1), the dataset got reduced to only 1565 samples.

Table 1. Attributes Maximum Values

Attribute	Maximum value
Volatile acidity	1.5
Fixed acidity	15
Citric acid	1
Residual sugar	10
Sulphates	1.5
Alcohol	14

Figure 2. Attributes Boxplot



In order for the classification algorithms to produce models that worked more correctly, the quality range was reduced to an interval of 1 to 3, meaning 1-poor, 2-normal and 3-excellent. This reduction was made with a loop thorough the data and changing every value in the quality output that was in the interval for poor (1-4), normal (4-6) and excellent (1-7) for the corresponding value from the new interval.

4. Classification

The algorithms chosen for the classification stage aimed to compare between a more traditional approach and a more modern one. Thus, the Decision Tree and Neural Network algorithms were chosen, and both were implemented with the SciKit Learn Python library. Before getting the classification models, the dataset was divided into 70% for training and 30% for testing.

Foremost, the model of the decision tree with no settings (Figure 3) was polluted and there were no apparent good decisions made. In order for the model to be better trained, the parameters of the SciKit Learn function were tested, always analyzing the accuracy and confusion matrix results for each round of test. The class weight, criterion and max depth were the most impactful parameters found. In the end, the best values for a better gain were the class weight set to unbalanced, a maximum tree depth of 4 and using the entropy criterion (Figure 4).

Figure 3. Decision Tree Model with no settings

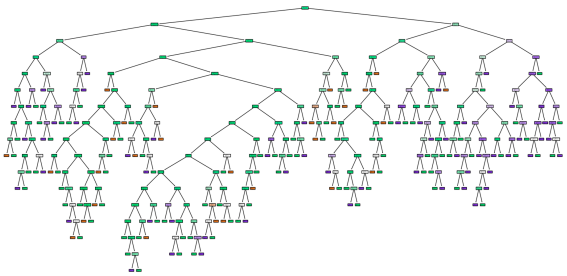
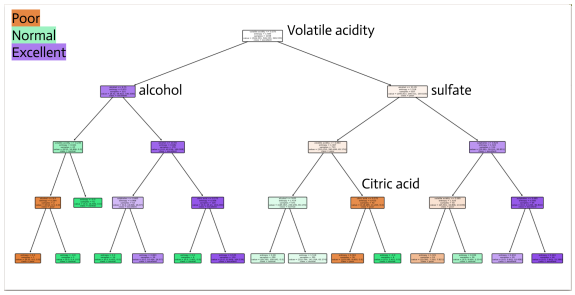


Figure 4. Final Decision Tree Model



The neural network algorithm chosen was the Multi-Layer Perceptron, which has one hidden layer with 100 neurons as default. For getting a better accuracy and confusion matrix, it was added a second layer with 700 neurons and the first layer was changed to 500 neurons.

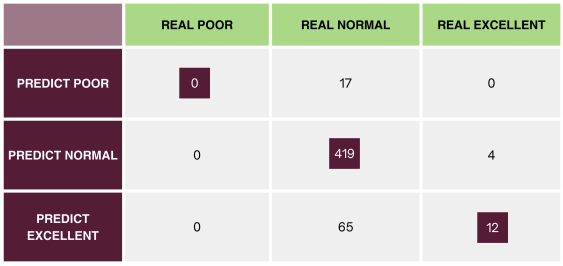
5. Results

The accuracy and confusion matrix metrics upon the testing set were chosen to evaluate each model. For the final decision tree model, the accuracy was 78% and the confusion matrix (Figure 5) showed the difficulty of the model to classify the poor quality wine. In the other hand, the MLP classifier had a better accuracy, resulting in an 83.36%, but the worst confusion matrix 6. The MLP confusion matrix reveals the model's inability of classifying the poor quality wine samples correctly.

Figure 5. Decision Tree Model's Confusion Matrix

	REAL POOR	REAL NORMAL	REAL EXCELLENT
PREDICT POOR	4	15	0
PREDICT NORMAL	24	344	8
PREDICT EXCELLENT	4	18	55

Figure 6. MLP's Confusion Matrix



The code can be found on the GitHub Repository: `https://github.com/indraAraujo/WineQualityClassification`

6. Conclusion

The knowledge gathered from a data mining process can be impacted by a variety of reasons. From the results obtained with the decision tree and the multi-layer perceptron models, it is explicit that the impact made by the attempt to mine an unbalanced dataset is very strong. Even though the MLP classifier got a better accuracy, both models had trouble classifying the excellent quality wines, but mainly the poor ones. This happens because there are only 63 samples of poor quality of the 1599 total samples.

In future studies, an expert could be brought to the discussion to analyze if the patterns recognized by the models were the right ones. Also, it would be enlightening to use a different and more balanced data frame and apply others classification algorithms.

References

Han, J., Pei, J., and Tong, H. (2022). *Data mining: concepts and techniques*. Morgan kaufmann.