Factors of Wine Quality

Authors:

Ethan Chase

Matthew McClure

Payton Parrish

Abstract

This report aims to analyze the various factors that impact wine quality using predictive models. Wine is a complex blend of multiple components, and to achieve quality, understanding the contributing factors is crucial. This study utilizes the Wine Quality dataset from UCI and investigates 11 input variables of white and red variations of the Portuguese “Vinho Verde” wine. The analysis will provide insights into the connection and importance between the input variables and wine quality, thereby helping to improve wine production processes.

Introduction

The production of high-quality wine involves numerous factors that vary significantly between batches and even individual bottles. Several elements, such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol, are crucial for wine quality. This research paper aims to investigate these factors and their impact on the quality of wine produced by a specific brand, Vinho Verde. Through the analysis of these parameters, we aim to gain insights into the characteristics that make a bottle of wine exceptional and to develop a better understanding of the wine production process.

Problem Statement

Wine quality depends on various factors, and identifying the crucial variables that contribute to it is critical for the wine industry. This research paper aims to investigate the factors that impact wine quality and their relative importance. Additionally, we aim to analyze whether these factors differ between red and white wine. Through this research, we seek to contribute to the wine industry's understanding of the factors that impact wine quality and how these factors can be utilized to consistently produce high-quality wines. We also intend to develop a predictive model that can assess the quality of new wines added to the dataset.

Proposed Methodology

The proposed methodology for this research paper involves analyzing different regression and classification models to determine the most accurate model for predicting wine quality. The following methods will be utilized: random forest classification and Poisson regression since wine quality ranges from the discrete values of 1-10. We will also implement standard regression with rounding towards the nearest integer and other classification methods.

To identify the variables that contribute most to wine quality, we will use the forward stepping feature selection process. We will compare the features selected for each dataset to analyze the important variables between red and white wine. Finally, to determine the predictability of our model for new wines added to the dataset, we will analyze the accuracy of our models.

Exploratory Data Analysis

Before we can perform regression or classification analysis on our datasets, we must first understand the data. Beginning with the white wine dataset, there are 4898 observations and 12 features. There are no null values in any of the observations. Every feature except the dependent variable, quality, is of type float64, and quality is of type int64.

Chart, bar chart

Description automatically generated

Figure : White Wine, quality vs. count

As seen above, the quality of white wine only ranges from 3-9, so there are no observations for white wine of quality 1, 2, or 10. The data is not balanced but the quality to count chart shows a distribution similar to normal, with more data points focused near the center. This could cause a problem since the quality variable is supposed to range from 1-10, so our models won’t have any training data to judge new wine added to the dataset of very high or very low quality. There is a strong positive correlation between residual sugar and density, and a strong negative correlation between alcohol and density.

For the red wine dataset, there are 1599 observations and 12 features. Once again, there are no null values in any of the observations, and all the features are the same as the white wine dataset.

Chart, bar chart

Description automatically generated

Figure : Red wine, quality vs. count

As seen above, the quality of red wine only ranges from 3-8. Just like white wine, there are no observations for red wine of quality 1, 2, or 10, but red wine is also lacking any data for quality of 9. The distribution of quality in the red wine dataset does not follow normal distribution as well as the white wine dataset does but is still loosely followed. Again, like the white wine, the lack of balancing in the dataset will likely cause problems for the models if new wines are added of very high or very low quality. There are no strong positive or negative correlations between any of the features of the red wine dataset.

Variable Selection

Some of the questions we had for the dataset were related to which features of the data were the most important in predicting quality, and whether the important features were the same for red and white wines. We used SFS variable selection to investigate this. We found eight significant variables for the white wine data, and seven significant variables for the red wine data.

Out of all the variables, Citric Acid content was the only one that was not significant to either dataset. Chlorides and Total Sulfur Dioxide were significant to white wine but not red wine, and Fixed Acidity, Residual Sugar, and Density were significant to red wine but not white wine.

Regression Methods

We used several regression methods to try to predict the quality of the wine data based on input variables. The two most successful were Linear Regression on variables selected by SFS variable selection and Poisson Regression. Both methods of regression gave us a continuous number as output, so to analyze the predictions, we rounded the output to the nearest integer. The results of the regression techniques can be seen in the charts below.

|  |  |
| --- | --- |
|  |  |
|  |  |

The charts show that, when the results are rounded, normal regression methods perform slightly better than Poisson Regression on the data. On the other hand, when analyzing how close the continuous prediction was to the true value of each wine, we calculated the following Mean Squared Error values for the different models.

|  |  |
| --- | --- |
| White Wine SFS + Regression | 1.038 |
| Red Wine SFS + Regression | 0.446 |
| White Wine Poisson Regression | 0.949 |
| Red Wine Poisson Regression | 0.481 |

Looking at the values of the actual outputs, Poisson regression performed slightly better for the white wine data, and slightly worse for the red wine data.

Classification Methods

Four different types of classifiers were used to find accuracy, precision, and most importantly the recall of the classification: The K-Nearest Neighbor (KNN), the Naïve Bayes, the Support Vector Machine (SVM), and Random Forest. Most of the other methods, especially the Logistic Regression classifier, would not be as accurate because this data is a Multi-Class Classification.

|  |  |
| --- | --- |
|  |  |
|  |  |

As shown on the pie charts, they all had values that were off by 2. Random Forest and Naïve Bayes have that at max (2% and 5%, respectively), SVM has (less than 1% of) values off by 3 at max, and KNN has (less than 1% of) values off by 4. With this in mind, Random Forest and Naïve Bayes are superior. Comparing those two, both classifiers have 37% of values off by 1. But with the number of correct values, Random Forest Classification has a higher percentage (that being 61%) than the Naïve Bayes Classification (that being 58%). The Random Forest classifier seems to be the best method in this case.



Naïve Bayes K-Nearest Neighbors



Support Vector Machine Random Tree

The data for each classifier has been collected. Of all the data received, however, the only important variables are accuracy, weighted average precision, and most importantly weighted average recall. In terms of accuracy, the best classifiers are Random Tree, Naïve Bayes, and a tie between KNN and SVM. In terms of precision, the best classifiers are SVM, Naïve Bayes, Random Tree, and KNN. Finally, in terms of recall, the best classifiers are the exact same as the accuracy. Therefore, Random Tree is once again the superior classifier.

With both comparisons, Random Tree seems to be the best classifier in overall general to get an accurate classification for our wine datasets.

Conclusions

Our analysis of the data showed that we can predict the quality of the wine datasets with reasonable accuracy. While it is difficult to get the result exactly correct, we were able to reliably predict the quality of upwards of 90% of the wine data within one point of the true quality.

We also were able to learn about the factors of the wines that contribute the most to determining quality. These factors were different for red and white wine, which we did not initially expect.

Possible areas of continuing this research moving forward could be trying different types of classification, including breaking down the quality results into fewer categories. This may allow us to make more accurate predictions by minimizing the amount of possible result categories. Another potential area of improvement could be data balancing. Either by oversampling or undersampling, we could try to get our models to more accurately recognize the minority classes of quality.

Lessons Learned

We all learned a lot throughout the project and class as a whole. During the project, the primary topic of focus was teamwork. We did a very good job of delegating different sections of each assignment to every person. This allowed us to work independently from each other towards the same goal, which was crucial to getting this project done, but also limited us from being a part of every section. Also, there were creative differences that commonly had to be discussed and overcome.

In terms of the class, the lessons we learned revolved more around critical thinking. In most classes, the solution to whatever problem we are trying to solve has a predefined solution. In data mining, there’s more room for creative freedom and subjectivity in results. We think this training in more complex problems with undefined definitive solutions is key to the class and should be maintained for future semesters.

Bibliography

<https://archive.ics.uci.edu/ml/datasets/Wine+Quality>