WINE QUALITY PREDICTION

Gheorghița Valentin [1]

[1] Technical University of Moldova, Faculty of Computing, Informatics and Microelectronics, Department of Informatics and Systems Engineering

ABSTRACT

The following research analyses a variety of chemical attributes from different wine samples to determine the quality of new and upcoming products. It uses a public Kaggle dataset with 11 chemical attributes and over a thousand entries. It employs comprehensive data preprocessing, exploratory data analysis, and linear regression models to create a predictive model. Three distinct linear regression models—Simple Linear Regression, Multiple Linear Regression, and a model incorporating All Possible Main Effects—were systematically evaluated using rigorous performance metrics, including Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared. The investigation unveiled that the All Possible Main Effects model demonstrated superior predictive accuracy, underscoring the significance of considering the interactions among multiple chemical features in comprehensively understanding and predicting wine quality. Notably, the use of cross-validation enhanced the robustness of the models, providing internal validation of their performance. While this analysis significantly contributes to the predictive modeling of wine quality, it is essential to acknowledge limitations such as potential dataset biases, the assumption of linearity, and the need for external validation. The findings offer valuable insights for winemakers, researchers, and practitioners seeking to optimize wine quality through informed consideration of chemical attributes, paving the way for more nuanced and accurate predictive modeling in future endeavors.

TODO: finish

Data and code can be found in the following repository: https://github.com/MrPengujn/temp-AD

INTRODUCTION

The calculation of wine quality is a fundamental and imperative aspect of oenological research and industry practices, driven by the need to understand, assess, and enhance the

sensory experience of wine consumption. Wine quality encompasses a complex interplay of chemical, physical, and organoleptic attributes, making it a multidimensional and nuanced phenomenon. The evaluation of wine quality serves several critical purposes, including guiding consumer choices, informing production decisions, and fostering continuous improvement in viticulture and winemaking processes.

For consumers, the quality rating of a wine serves as a valuable cue, aiding in the selection of wines that align with individual preferences and occasions. Understanding the chemical components contributing to quality allows producers to refine and optimize their winemaking techniques, ultimately influencing the market competitiveness of their products. Moreover, quality assessments play a pivotal role in maintaining and elevating the reputation of vineyards and wineries, contributing to the broader cultural and economic significance of the wine industry.

In the context of research, the calculation of wine quality provides a foundation for exploring correlations between chemical composition and sensory perception. This intersection of science and art allows researchers to unravel the complexities of flavor development, identify key contributors to wine excellence, and contribute to advancements in the field of enology. Overall, the necessity to calculate wine quality arises from its pivotal role in shaping both consumer experiences and the ongoing evolution of the wine industry.

The goal of this research is to conduct an educational study, aiming to analyze a dataset that contains 11 chemical attributes and to determine the overall quality of new wine samples. This also includes pinpointing the dataset variable that influence a sample's quality the most. The second goal is developing and training a model that can take brand new chemical properties and process them into a number representing the overall quality.

Initial assumptions

• Does the alcohol level have a big impact on the overall quality of the sample?

• Do acidity levels have any influence on the pH level? What is the correlation between them?

• Are the free_sulfur_dioxide and total_sulfur_dioxide correlated to each other?

Materials & Methods

This analysis uses an open-source Kaggle dataset.

Source: https://www.kaggle.com/datasets/yasserh/wine-quality-dataset [2]

The dataset contains a total of 1144 entries and 13 variables, of which one is the id of each entry, which can be ignored, and the second is the overall quality for each variable set. This research focuses on analyzing the other 11 chemical attributes: fixed acidity, volatile acidity, citric acidity, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol.

Variable	Data Type	Description
fixed_acidity	Numeric	Fixed acidity level of the sample
volatile_acidity	Numeric	Volatile acidity level of the sample
citric_acid	Numeric	Citric acid level of the sample
residual_sugar	Numeric	Residual acidity level of the sample
chlorides	Numeric	Chloride's level of the sample
free_sulfur_dioxide	Numeric	Free sulfur dioxide level of the sample
density	Numeric	Total sulfur dioxide level of the sample
pH	Numeric	Acidity level of the sample
sulphates	Numeric	Sulphate's level of the sample
alcohol	Numeric	Alcohol level of the sample
quality	Numeric	The overall sample quality
id	Numeric	Sample identification number

Data pre-processing

Luckily, the dataset is well organized and processed, thus it doesn't contain any missing or null values. Upon brief analysis of the dataset, it was deduced that the 'id' variable is of no use to the objective of the current research so it was dropped from the list.

Exploratory data analysis (EDA)

During the EDA process tools such as R and the ggplot2 library were used to analyze the wine quality dataset. The process consisted of creating bar charts for visualizing all the existing quality levels and their spread, the mean alcohol level for each wine quality etc. Scatter plots for visualizing the relationship between values, such as free and total sulfur dioxide levels. Correlation heatmaps provide insight into the strength and direction of relationships between variables. Histograms to understand the distribution of individual numerical variables, such as alcohol and pH levels.

VIP (Variable Importance in Projection)

VIP (Variable Importance in Projection) ws used to indetify which chemical attributes are most influential in determining the quality rating of wines. For this stage, two packages were used: caret and pls. In this dataset, it helped identify the most important variables from: fixed acidity, volatile acidity, citric acidity, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol in determining the overall quality.

Linear Regression

In the exploration of the given wine dataset, linear regression serves as a fundamental analytical tool to unveil potential relationships between chemical attributes and the overall quality of wines. Employing the linear regression model, we seek to quantify and understand the linear associations between independent variables, such as acidity, alcohol content, and sulphate concentration, and the dependent variable - wine quality. This methodological approach allows us to model the nuanced interplay of these chemical factors and offers insights into the predictive capacity of specific attributes on the quality rating of wines. Through linear regression analysis, we aim to provide a quantitative framework for discerning key contributors to wine quality, facilitating informed decision-making in viticulture and winemaking practices.

Performance assessment of models

In the evaluation of model performance for the specified wine dataset, we applied a robust approach utilizing the `caret` package in R. Through meticulous cross-validation techniques and metrics such as mean squared error and R-squared, we rigorously assessed the accuracy and goodness-of-fit of models, including linear regression, decision trees, and random forests. The implementation of the `train` function within `caret` facilitated model training and testing, ensuring a comprehensive examination of predictive frameworks. This methodological synergy not only provides valuable insights into model reliability for predicting wine quality but also enhances decision-making processes in the viticulture and winemaking industry.

EDA results

Overall quality tendency

In the graph provided in Figure 1, it ca be seen that most of the samples tend to have a medium to high quality level, with only a few breaking the 4th level. Even if this is normal behaviour, it could prevent us from having clear results for some specific samples.

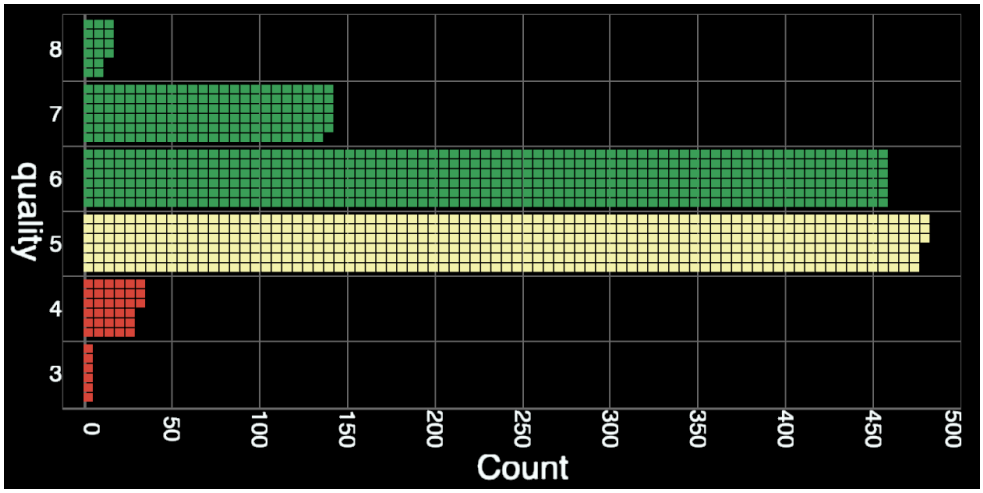

Figure 1. Overall quality distribution

Finding outliers

In order to visualize potential outliers within the wine dataset, a rigorous method was employed using the `ggplot2` and `tidyr` libraries in R. This approach facilitated the reshaping of numerical variables through the `gather` function, enabling the creation of informative boxplots that effectively delineate the distributional characteristics and identify outliers in key chemical attributes. Figure 2 showscases such instances for a couple of variables such as: citric_acid and residual_sugar

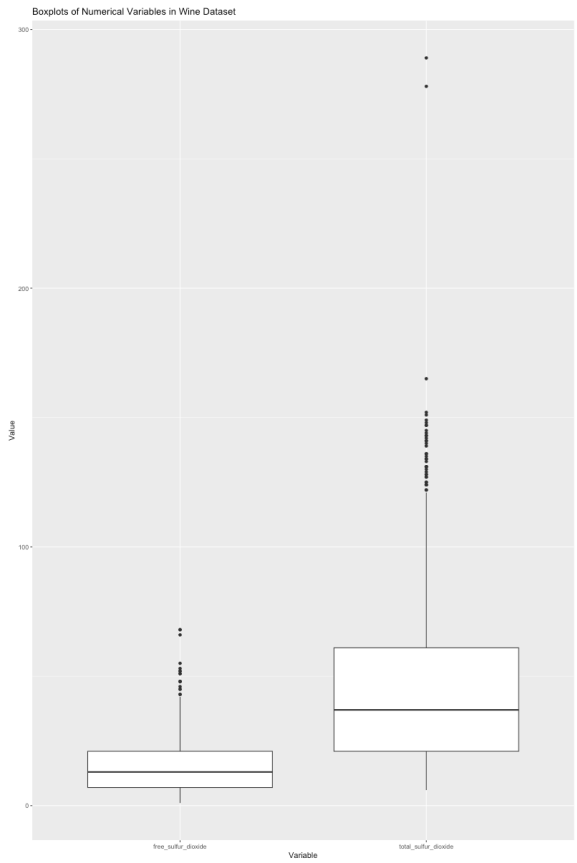

Figure 1. Outliers per class

After visualizing such values, we can apply the IQR method to remove such outliers.

Correlations between variables

Utilizing a correlation heatmap, represented in Figure 3, the intricate interrelationships among chemical attributes in the wine dataset were visually represented, offering a comprehensive understanding of the strength and direction of correlations. The color gradient, ranging from blue to red, effectively conveyed the degree of correlation, with blue hues indicating negative correlations, white representing neutral associations, and red denoting positive correlations. Annotated correlation coefficients, precisely rounded to two decimal places, provided quantitative insights into the numerical strength of these relationships. The strategic rotation of x-axis labels enhanced readability, facilitating the identification of key variables. This scientific visualization serves as a foundational tool for discerning patterns and informing subsequent analyses in the investigation of factors influencing wine quality.

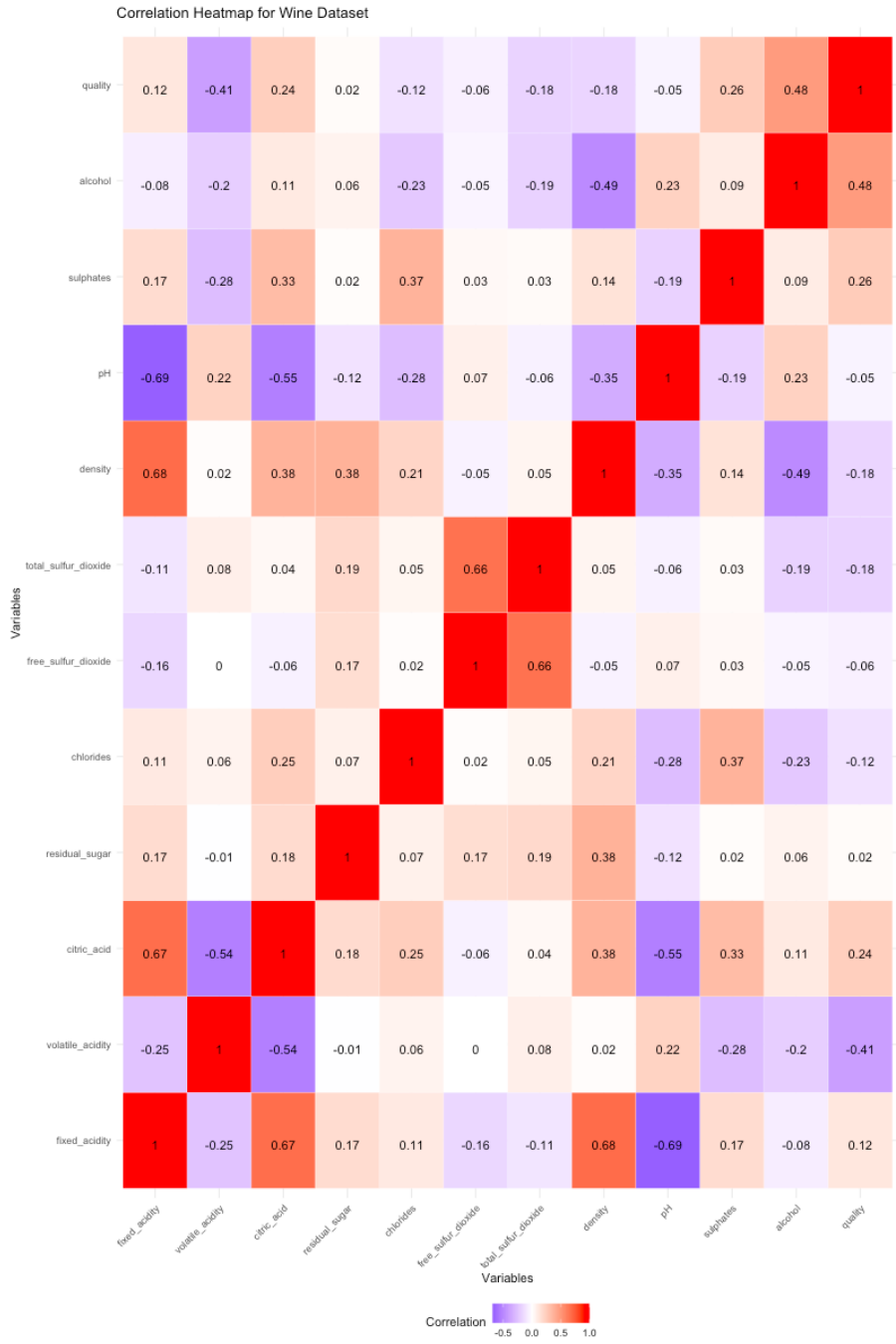

Figure 3. Correlation heatmap

In this heatmap we can observe that the free and total sulfur dioxide, fixed acidity and density, fixed acidity and citric acid variables are closely related to each other.

VIP

During the creation and investigation of the VIP represented in Figure 4 for the cv_aplr model, it can be seen that the sulphates variable has the upmost importance in determining the

quality of the wine, followed by alcohol and volatile_acidity. The lower side is held by the free_sulfur_dioxide variable. This helps in visualizing and determining the quality on which to put upmost importance and precision while elaborating new wine recipes and products.

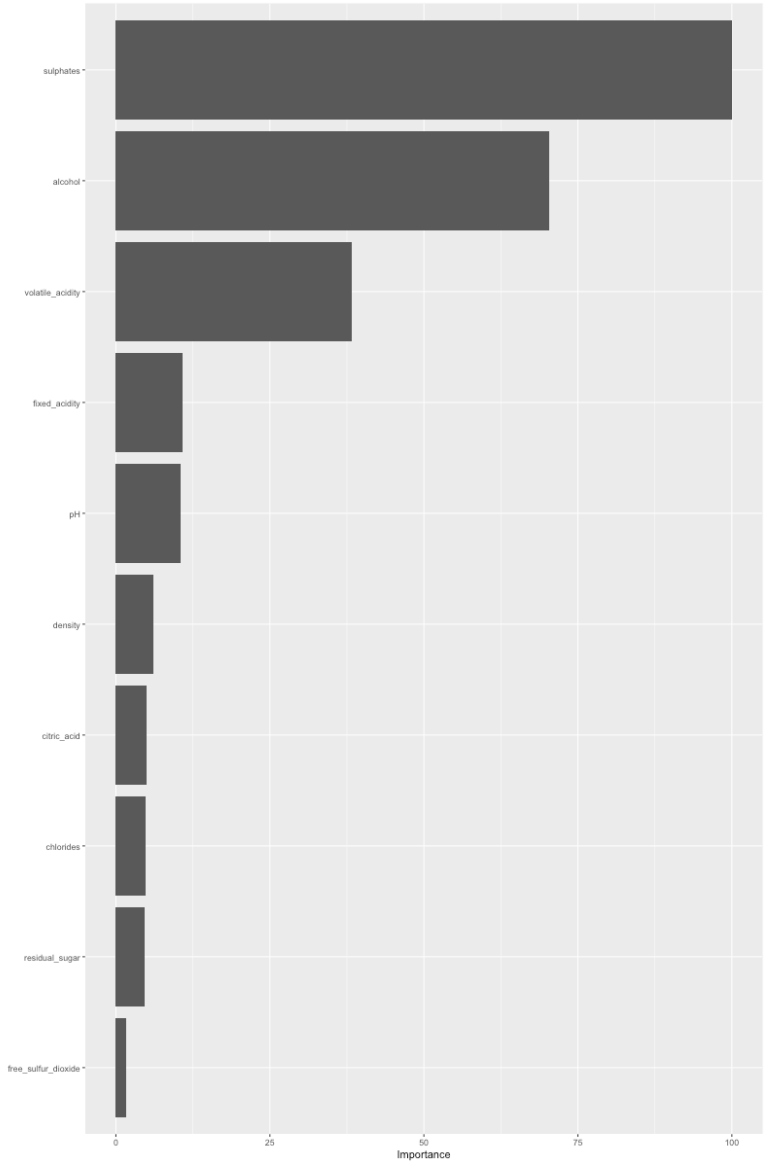

Figure 4. VIP of the cv_aplr model

Model creation and analysis

In the pursuit of understanding the intricate relationships within the wine dataset, three distinct regression models were employed: a simple linear regression, elucidating the influence of individual chemical attributes on wine quality; a multiple linear regression, discerning the collective impact of multiple attributes on the quality rating; and an analysis encompassing all possible main effects, systematically exploring the isolated effects of each variable on the response variable. These modeling approaches serve as invaluable tools in unraveling the

nuanced interplay of chemical factors and offer a quantitative foundation for predictive insights into wine quality determinants.

The three linear regression models applied to the dataset exhibit varying performance metrics, as summarized by Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared, which can be seen in Figure 5.

MAE						
	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
slr_model	0.4769162	0.4942393	0.5173630	0.5119456	0.5246617	0.5495160
mlr_model	0.4263963	0.4676438	0.4889344	0.4802504	0.4948086	0.5354837
aplr_model	0.4067889	0.4232400	0.4453302	0.4464455	0.4691070	0.4916591
RMSE						
slr_model	0.5330384	0.5950184	0.6213896	0.6094696	0.6389369	0.6555972
mlr_model	0.4868573	0.5729749	0.5910809	0.5783396	0.6045281	0.6181669
aplr_model	0.4874986	0.5223249	0.5419076	0.5450068	0.5665100	0.6072436
Rsquared						
slr_model	0.1485757	0.1956357	0.2281502	0.2582414	0.2954272	0.4272595
mlr_model	0.2129830	0.2654424	0.3201591	0.3290941	0.3755377	0.4912873
aplr_model	0.2823209	0.3685003	0.4159264	0.4118592	0.4833854	0.4980332

Figure 5. Performance measure

1. Simple Linear Regression (slr_model):
   - MAE: Mean absolute error ranges from 0.4769 to 0.5495, with a mean of 0.5119.
   - RMSE: Root mean squared error ranges from 0.5330 to 0.6556, with a mean of 0.6095.
   - Rsquared: R-squared ranges from 0.1486 to 0.4273, with a mean of 0.2582.
   - Overall, the model exhibits moderate predictive accuracy with a tendency to underperform on certain observations.
2. Multiple Linear Regression (mlr_model):
   - MAE: Mean absolute error ranges from 0.4264 to 0.5355, with a mean of 0.4803.
   - RMSE: Root mean squared error ranges from 0.4869 to 0.6182, with a mean of 0.5783.
   - Rsquared: R-squared ranges from 0.2130 to 0.4913, with a mean of 0.3291.
   - The multiple linear regression model shows improved performance compared to the simple linear regression, indicating a better fit to the data.
3. All Possible Main Effects (aplr_model):

- MAE: Mean absolute error ranges from 0.4068 to 0.4917, with a mean of 0.4464.
- RMSE: Root mean squared error ranges from 0.4875 to 0.6072, with a mean of 0.5450.
- Rsquared: R-squared ranges from 0.2823 to 0.4980, with a mean of 0.4119.
- The model incorporating all possible main effects demonstrates the best overall performance, reflected in lower MAE and RMSE values and higher R-squared values compared to the other models.

Results and Discussions

In analyzing the wine quality dataset through the lens of three distinct linear regression models, namely Simple Linear Regression, Multiple Linear Regression, and a model incorporating All Possible Main Effects, a comprehensive assessment of predictive performance has been achieved. The models were evaluated using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R-squared. Notably, the All Possible Main Effects model demonstrated superior predictive accuracy, showcasing lower MAE and RMSE values, and higher R-squared values compared to the other models. These findings suggest that considering the interactions among various chemical attributes yields a more robust model for predicting wine quality. The discussion emphasizes the importance of selecting appropriate regression models tailored to the dataset characteristics.

While the analysis of the wine quality dataset and the application of linear regression models have provided valuable insights, it is essential to acknowledge certain limitations inherent in this study. Firstly, the dataset itself may have inherent biases or limitations, as the quality of wine is a complex and multifaceted phenomenon influenced by numerous factors that extend beyond the chemical attributes considered in this analysis. Additionally, the linear regression models employed assume a linear relationship between predictors and the response variable, potentially oversimplifying the intricate interplay of variables. The study's focus on linear models also overlooks potential nonlinear relationships that may exist within the data. Furthermore, the use of cross-validation and metrics like MAE, RMSE, and R-squared provides internal validation, but external validation on an independent dataset would enhance the generalizability of the findings.

The insights gained from this analysis can guide future endeavors in predicting and optimizing wine quality, providing valuable information for winemakers and researchers alike.

This study not only advances the methodology for predicting wine quality but also underscores the significance of model selection tailored to the characteristics of the dataset.

The insights gained from this analysis can potentially inform winemakers and researchers in refining their approaches to quality optimization, fostering a deeper understanding of the underlying factors influencing wine quality and setting the stage for more sophisticated modeling strategies in future studies.

REFERENCES

1. Wine quality factors / JJBuckley Fine wines: https://www.jjbuckley.com/wine-knowledge/blog/the-4-factors-and-4-indicators-of-wine-quality/1009

2. Data science platform / Kaggle: https://www.kaggle.com/

3. An overview to wine quality / Sciencedirect: https://www.sciencedirect.com/topics/food-science/wine-quality

4. Things that affect the wine quality / Liquidline: https://www.liquidline.se/blog/things-that-affect-the-wine-quality/

5. Model basics | R for data science: https://r4ds.had.co.nz/model-basics.html