Final Project Report

Group 4 - Eliza, Haley, Logan, and JennyLou

Introduction

State the Problem

The quality of wine is a critical factor in its market value and consumer preference. Traditionally, assessing wine quality relies on tasting the wine, which is subjective, costly, and time-consuming. With access to detailed physiochemical wine data, there is an opportunity to leverage linear regression methods, alongside other machine learning techniques to predict wine quality. The central problem we plan to address in this project is: Can we accurately predict the quality of wine using its physiochemical characteristics?

Objectives

The primary objective of our project is to predict wine quality based on its physicochemical characteristics. Additionally, we aim to identify the most influential physicochemical features and explore the relationships between these properties and the quality ratings.

Data Description and Preparation

Dataset Overview

https://archive.ics.uci.edu/dataset/186/wine+quality

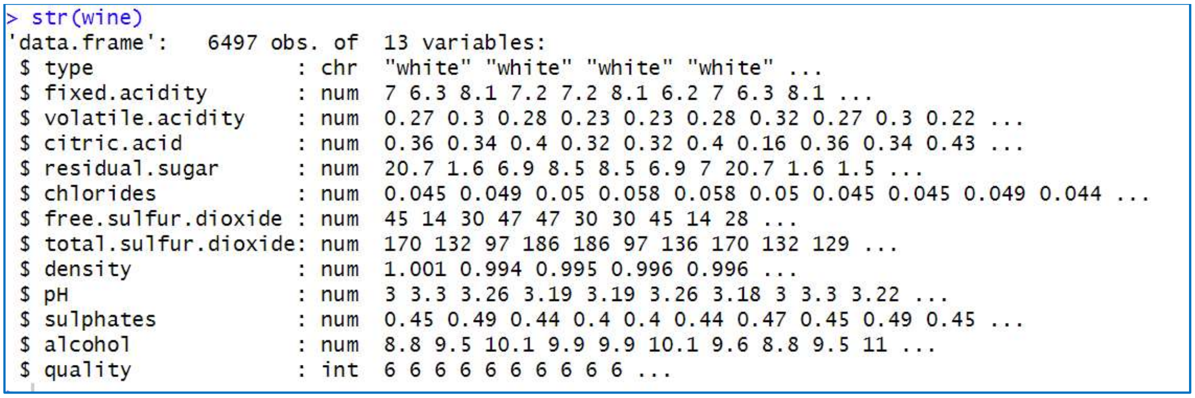
The dataset, sourced from the UCI Machine Learning Repository, comprises 4,898 red and 1,599 white wine samples from northern Portugal (6,497 total observations). It includes 12 physicochemical attributes (e.g., acidity, sugar, density) and 1 sensory output (quality score).

The color of the wine (red or white) is a categorical variable, while other inputs are continuous. The quality score is an integer ranging from 3 to 9. Due to privacy and logistic issues, no data was included about grape types, wine brand, wine selling price, etc.

Data Cleaning and Transformation – Overview

After importing the data, we identified 38 missing values and removed the corresponding observations to ensure the dataset's integrity. We also addressed potential outliers in variables such as sulphates, citric acid, and density. Log transformations were applied to variables with skewed distributions, including volatile acidity, residual sugar, chlorides, and total sulfur dioxide, to stabilize variance and improve model performance. The cleaned dataset contains 6,457 observations and retains all relevant variables for predictive analysis.

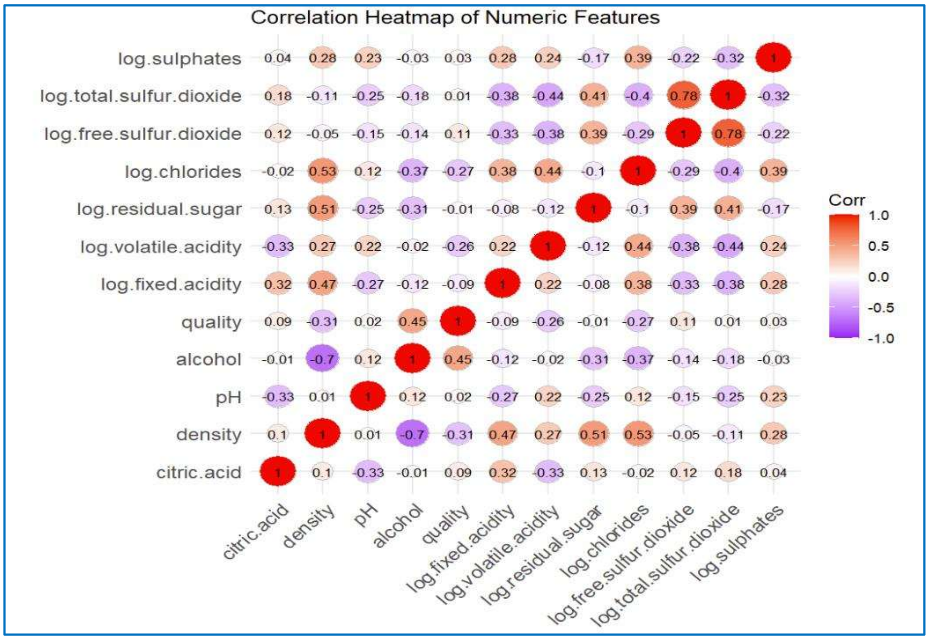
Preliminary Data Analysis/EDA



```
> str(wine)
'data.frame':	6497 obs. of  13 variables:
 $ type                : chr  "white" "white" "white" "white" ...
 $ fixed.acidity       : num  7 6.3 8.1 7.2 7.2 8.1 6.2 7 6.3 8.1 ...
 $ volatile.acidity    : num  0.27 0.3 0.28 0.23 0.23 0.28 0.32 0.27 0.3 0.22 ...
 $ citric.acid         : num  0.36 0.34 0.4 0.32 0.32 0.4 0.16 0.36 0.34 0.43 ...
 $ residual.sugar      : num  20.7 1.6 6.9 8.5 8.5 6.9 7 20.7 1.6 1.5 ...
 $ chlorides           : num  0.045 0.049 0.05 0.058 0.058 0.05 0.045 0.045 0.049 0.044 ...
 $ free.sulfur.dioxide : num  45 14 30 47 47 30 30 45 14 28 ...
 $ total.sulfur.dioxide: num  170 132 97 186 186 97 136 170 132 129 ...
 $ density             : num  1.001 0.994 0.995 0.996 0.996 ...
 $ pH                  : num  3 3.3 3.26 3.19 3.19 3.26 3.18 3 3.3 3.22 ...
 $ sulphates           : num  0.45 0.49 0.44 0.4 0.4 0.44 0.47 0.45 0.49 0.45 ...
 $ alcohol             : num  8.8 9.5 10.1 9.9 9.9 10.1 9.6 8.8 9.5 11 ...
 $ quality             : int  6 6 6 6 6 6 6 6 6 6 ...
```

After loading the dataset, we examined the structure of the data (above). This revealed the need to change the variable ‘type’ to a factor variable. Once this was completed, we decided to examine it for missing or “NA” values. There were 38 missing values in the dataset, so we felt it was appropriate to remove the 34 corresponding observations before doing any predictive analysis. We then examined the distribution of each variable using histograms for a total value look and by boxplots that separated the values into red and white wine type. A preliminary view of the histograms of the variables showed skewed distributions in several of the variables. Fixed acidity, volatile acidity, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, and sulphates levels were log-transformed. Additionally, there were some outliers identified in total sulfur dioxide even after the log transformation, in citric acid, and in density. We removed density values above 1.01, free sulfur dioxide levels greater than 200 (before log transforming), and extreme citric acid values exceeding 1.0.

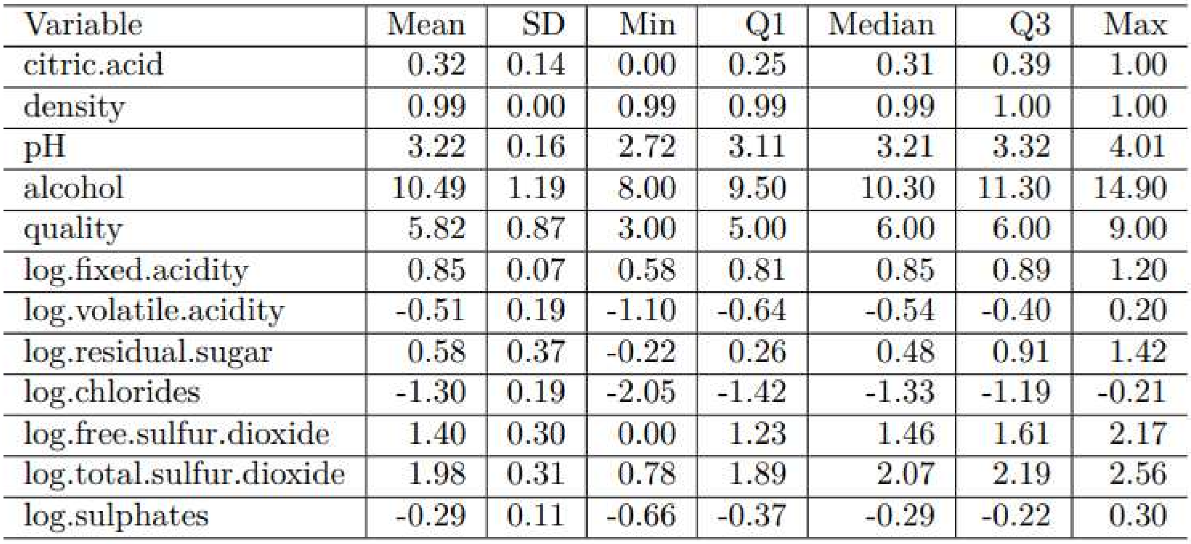
Post-transformation, the visualizations demonstrated more balanced distributions, though not perfectly normal. Differences in distributions by wine type (red vs. white) became clearer, leading us to retain wine type as a predictor. Noticing some potentially correlated variables such as the sulfur dioxide variables and the acidity variables, we conducted a correlation analysis, using a correlation heatmap to show the relationships between the numeric variables (below).



The heatmap illustrates the correlations among numeric features in the wine dataset, highlighting several key relationships. A strong negative correlation (-0.7) is observed between alcohol and density, indicating that wines with higher alcohol content tend to have lower density. Alcohol also shows a positive correlation (0.45) with quality, suggesting that higher alcohol content is associated with better wine quality. Log fixed acidity has a moderate positive correlation (0.33) with quality, pointing to its potential predictive value. On the other hand, citric acid exhibits weak correlations with other variables and quality, suggesting its limited role as a predictor.

The heatmap also reveals multicollinearity concerns, particularly between log total sulfur dioxide and log free sulfur dioxide, which are highly correlated (0.78). This indicates that these variables provide overlapping information and may need to be handled carefully in regression modeling to avoid multicollinearity issues. Interestingly, log volatile acidity and log fixed acidity are not highly correlated, suggesting they capture distinct aspects of wine properties. Overall, the heatmap provides valuable insights into relationships between variables, guiding the decision to retain significant predictors like alcohol and quality while addressing potential redundancy in correlated features such as sulfur dioxide levels.

Table 1: Summary Statistics for Wine Cleaned Data



Variable	Mean	SD	Min	Q1	Median	Q3	Max
citric.acid	0.32	0.14	0.00	0.25	0.31	0.39	1.00
density	0.99	0.00	0.99	0.99	0.99	1.00	1.00
pH	3.22	0.16	2.72	3.11	3.21	3.32	4.01
alcohol	10.49	1.19	8.00	9.50	10.30	11.30	14.90
quality	5.82	0.87	3.00	5.00	6.00	6.00	9.00
log.fixed.acidity	0.85	0.07	0.58	0.81	0.85	0.89	1.20
log.volatile.acidity	-0.51	0.19	-1.10	-0.64	-0.54	-0.40	0.20
log.residual.sugar	0.58	0.37	-0.22	0.26	0.48	0.91	1.42
log.chlorides	-1.30	0.19	-2.05	-1.42	-1.33	-1.19	-0.21
log.free.sulfur.dioxide	1.40	0.30	0.00	1.23	1.46	1.61	2.17
log.total.sulfur.dioxide	1.98	0.31	0.78	1.89	2.07	2.19	2.56
log.sulphates	-0.29	0.11	-0.66	-0.37	-0.29	-0.22	0.30

The summary statistics figure (above) provides an overview of the numeric features in our cleaned dataset. Alcohol has the highest mean, at 10.49, and has a wide range from 8.00 to 14.90, indicating there is significant variation. Conversely, density shows very little variation, with a mean of 0.99 and a standard deviation of 0.003, suggesting it is a stable feature. Our target variable, Quality, has a mean of 5.82 and a median of 6. The log-transformed variables now have more balanced ranges, improving their interpretability. These statistics provide a clear snapshot of our data's central tendency and spread.

Methods and Results

Linear Regression

Linear regression was used as the baseline model to assess the relationships between wine quality and the predictors. The model included all variables from the cleaned dataset, providing a foundation for comparison with more advanced techniques. The analysis showed that variables like alcohol, density, log volatile acidity, and log free sulfur dioxide were significant predictors, while citric acid showed no significant relationship with quality. Alcohol had a positive effect on quality, while density showed a strong negative association. The adjusted R-squared value was 0.31, meaning only 31% of the variance in wine quality could be explained by the predictors. The diagnostic plots showed no major violations of assumptions, but the model was limited in its ability to address any non-linear relationships or multicollinearity between the predictors. This demonstrated the need for more advanced approaches to improve both predictive performance and model fit.

Stepwise Selection

We used stepwise selection as a method for refining the baseline linear model by iteratively adding and removing predictors (using the "both" direction selection) to find the best combination of predictors based on their statistical significance. We optimized our model using AIC to balance complexity and fit. The variables citric acid and log chloride were removed during the process. The stepwise model retained all the other predictors. This approach improved model interpretability without compromising predictive power. The AIC fell from -3248.07 for the model with all predictive variables to -3258.42 for the final model without log chlorides and citric acid. The residual mean square error (RMSE) was also lower in comparison to the linear model. This reflects a better fit than the baseline model and emphasizes the importance of variable selection in regression analysis.

LASSO Regression

LASSO regression was used as an alternative to stepwise selection for variable selection as it simultaneously estimates coefficients and selects variables. This was to see how LASSO addressed potential issues of multicollinearity by applying an L1 penalty to the coefficients. This regularization technique effectively shrank less important coefficients to zero (only citric acid reached zero), simplifying the model while retaining significant predictors. For instance, variables like log total sulfur dioxide, which showed high correlation with log free sulfur dioxide, were penalized to reduce redundancy, though not much. LASSO also provided a means of handling the high dimensionality of the dataset without overfitting. Interestingly, after using cross validation methods to select the optimal lambda penalty, log chlorides remained in the model. This method offered an alternative to traditional regression and stepwise selection, balancing complexity, and predictive accuracy.

Elastic Net Regression

Elastic Net combined the strengths of LASSO with Ridge regression by applying both L1 and L2 penalties to the model coefficients. Ridge Regression is not used for variable selection as coefficients cannot be reduced to exactly zero, but it is used to address multicollinearity in the dataset, that we were unsure had been handled correctly up to this point. Elastic Net Regression retains groups of correlated predictors rather than excluding them entirely as LASSO might. Elastic Net was used to capture the unique and shared contributions of variables like log free sulfur dioxide and log total sulfur dioxide, which are highly correlated. The model provided a balance between simplicity and flexibility, retaining predictors like alcohol, density, and log volatile acidity. By addressing multicollinearity and preventing overfitting, Elastic Net outperformed the baseline linear model, but was almost identical to the LASSO output, which told us that the potential multicollinearity of our variables was not our biggest issue.

Generalized Additive Models (GAMs)

GAMs were utilized to capture non-linear relationships between individual predictors and the target variable while maintaining interpretability. This method fits smooth functions to each predictor, allowing flexibility in modeling their effects on wine quality. GAMs confirmed the non-linear effects of variables like alcohol and log fixed acidity on wine quality, which were less apparent in purely linear models. The model achieved a good balance between flexibility and interpretability, with visualizations of smooth functions providing insights into how predictors influence wine quality. For instance, alcohol showed a threshold effect, where quality improved significantly up to a certain level before plateauing. Seeing a slight improvement in RMSE and considerable improvement in R-squared compared to previous models/methods, we decided to further investigate non-linear relationships between the variables.

Support Vector Regression (SVR)

SVR was used to model wine quality as a non-linear function of the predictors. Unlike traditional regression methods, SVR aims to find a hyperplane that maximizes the margin of tolerance (epsilon) while minimizing errors outside this margin. It is particularly effective in capturing complex, non-linear relationships between predictors and the target variable. In this analysis, SVR performed well in terms of predictive accuracy for wine quality, but its interpretation was more challenging compared to linear models. Tuning hyperparameters such as the kernel type, cost (C), and epsilon was crucial to optimizing the model's performance. SVR highlighted the ability to model relationships that linear methods might overlook, particularly for predictors with non-linear impacts on quality, like log volatile acidity and log residual sugar. The R-squared and RMSE scores for this method improved, suggesting that there were non-linear relationships between the variables in the dataset that were not being appropriately captured in the previous models/methods.

Results

Model	**RSME**	**R^2**
Linear	**0.728**	**0.306**
Stepwise	**0.729**	**0.304**
LASSO	**0.728**	**0.306**
Elastic Net	**0.728**	**0.306**
GAM	**0.714**	**0.333**
SVR	**0.676**	**0.402**

Linear regression models indicated significant predictors of wine quality, including alcohol, density, and pH. However, non-linear methods like support vector regression demonstrated superior performance, with higher R-squared values and lower mean squared errors (MSE).

Key Findings:

- Alcohol content was the most influential predictor, positively correlating with quality.
- Density negatively impacted quality, particularly in white wines.

- Citric acid had varying effects depending on wine type and was often dropped due to insignificant results.
- The Generalized Additive Model performed better than the linear based models, showing that a more complicated model was necessary to better define the relationship between quality and the other variables.
- Support Vector Regression had the best performance of all models, recording the lowest residual mean squared error and highest R-squared values. This may signal a nonlinear relationship between quality and some variables not captured by the linear based regression models.

Conclusion

Our analysis demonstrates that wine quality can be accurately predicted using physicochemical characteristics, with alcohol content, density, and pH as key drivers. Advanced modeling techniques such as Elastic Net, SVR and GAM provided differing performances, highlighting the importance of exploring both linear and non-linear relationships. Further exploration could involve:

- Separate modeling for red and white wines to improve prediction accuracy.
- Incorporating external variables such as vineyard location or climate data to see what drives wine quality beyond its physicochemical properties.
- Expanding to other machine learning methods, such as boosting algorithms or deep learning to capture more nuanced patterns in the data.

This project underscores the potential of data-driven approaches to complement traditional wine quality assessment, offering efficiency and objectivity in the evaluation process. Through our analysis we have shown that objective, reliable predictions of wine quality are feasible and effective. With further refinement and incorporating additional data, our methodology can enhance wine evaluation, providing a more efficient, scalable, and consistent alternative to the traditional sensory assessments.