Predicting Vihno Verde Wine Quality Using Machine Learning on Physiochemical Attributes

Leo Gaunt – 230450260

Word Count – 2472

Abstract

This report investigates predictive modelling of wine quality using two datasets containing physiochemical attributes and quality scores. Initially, data exploration was performed through visualisation techniques, so trends and correlations between each of the variables and the final quality of the wine could be identified. After analysis of the correlations, a subset of the attributes was identified that would be most useful for training machine models, both classification and regression. The classification models involved using binary labels to distinguish between low- and high-quality wines, however the regression models aimed to predict a more accurate, continuous, quality score. The models included in this investigation are: Logistic Regression, Random Forest Classifier, Linear Regression and Random Forest Regressor. The models created were evaluated using k-fold cross validation, and the results demonstrated that the Random Forest models outperformed linear approaches within both the classification and regression tasks. Model performance was assessed using accuracy, AUC and F1 score for classification and MSE and RMSE scores for regression. The findings of the investigation are presented as visualisations and the cross-validation results. The Random Forest Regressor produced the best performance in my investigation, this indicates that in future work non-linear models would be the most suitable.

What was done and how

Dataset overview

In this investigation, I used publicly available Wine Quality datasets (Cortex, et al., 2009). One dataset gave red wine data (1599 records) and the other, white wine data (4898 records). These datasets included 11 physiochemical attributes such as acidity, density, pH, sulphates and alcohol.

Each record represents a single wine sample with measured chemical properties and the quality. The quality in this dataset was the median of at least 3 evaluations made by wine experts, who were asked to grade each sample between 0 (very bad) and 10 (very excellent) (Cortex, et al., 2009). For the purposes of this project, the red and white datasets were initially analysed separately, and then combined to allow for a more generalised predictive machine learning model.

Exploring and Analysing the Datasets

Initial Inspection and Usage

As this data was gathered by an external source, I thought it would be best to check over the data to make sure it was in a usable form for my investigation. This involved: checking for missing values and checking the Data types of each column. After the data was validated, it was imported using pandas into DataFrames for analysis.

Distribution of Wine Quality

To analyse the data, I printed out the number of wines in each dataset to a table to show numerically how many I had for each quality. This showed that every quality lied between 3 and 9. Knowing the spread of quality scores helped me to choose appropriate axis scaling for my graphs.

Quality	Red	White	Total
3	10	20	30
4	53	163	216
5	681	1457	2138
6	638	2198	2836
7	199	880	1079
8	18	175	193
9	0	5	5

Table 1 - Quality Figures for Red and White Wine

I used the pyplot submodule of the matplotlib module to create both a histogram and box plot for the datasets.

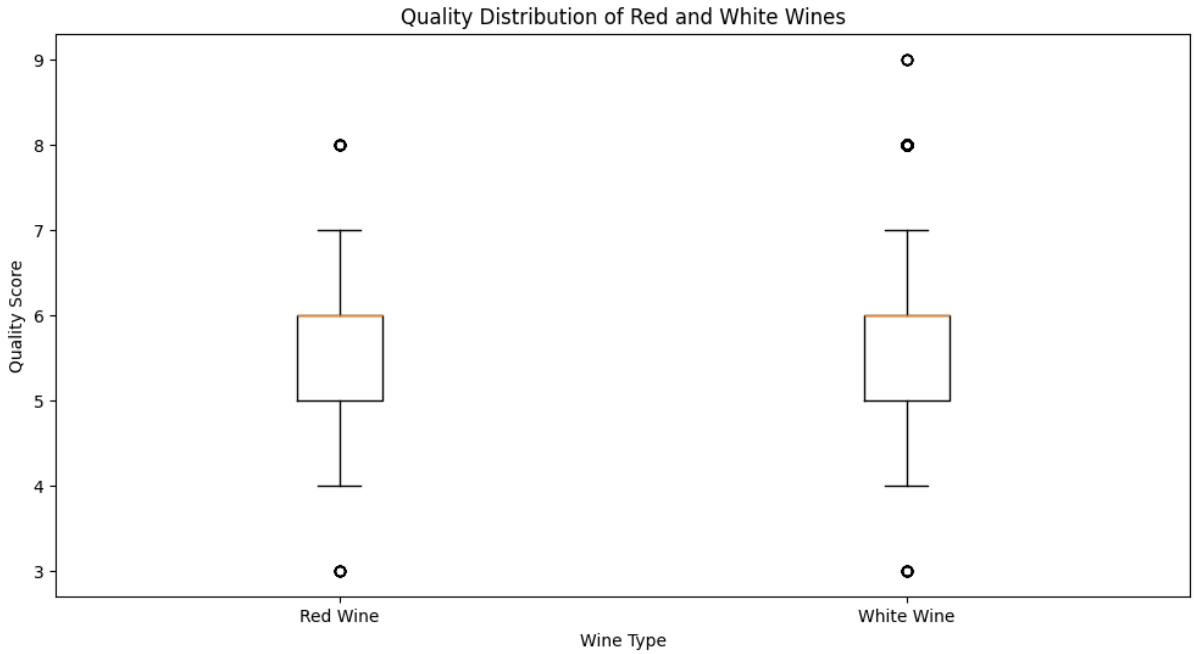

Figure 1 – Boxplot Showing Quality Distribution of Red and White Wines

While the medians and spreads are identical between both red and white wines, white wines appear to have some more high-quality outliers. This suggests that although quality may not strictly relate to the colour of the wine, white wines may occasionally reach a higher level of quality than red wine.

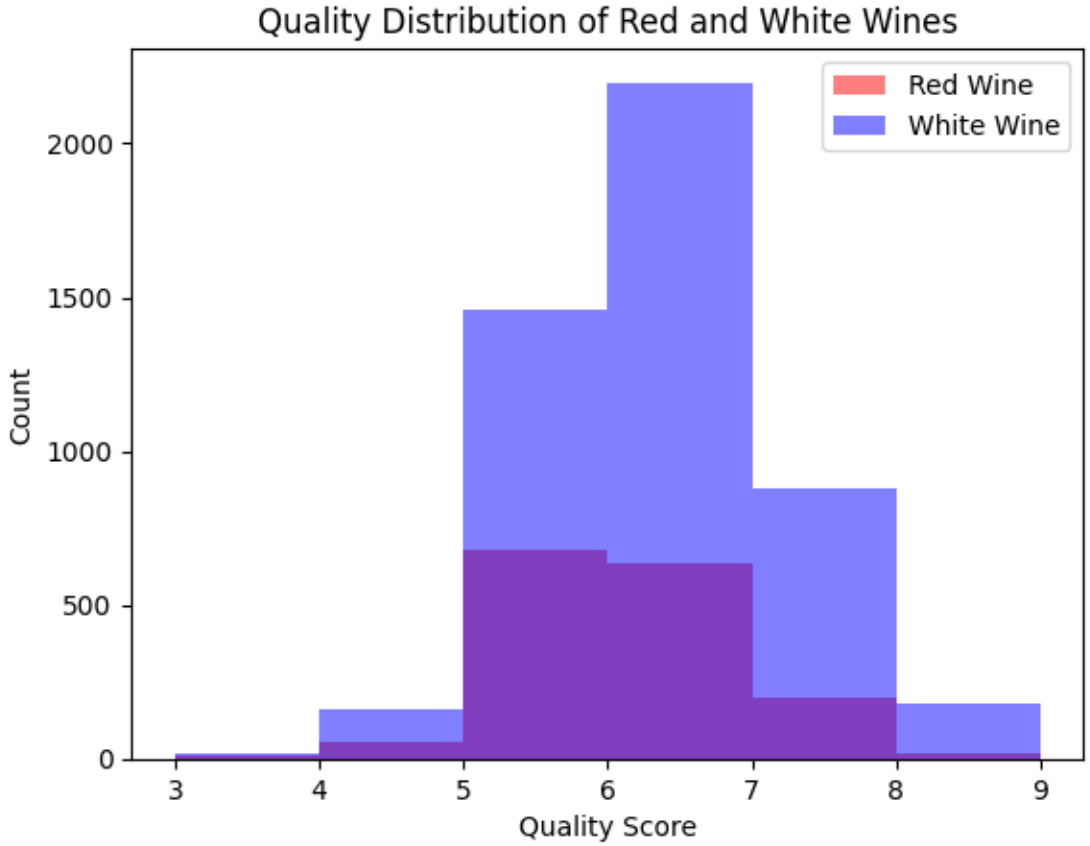

Figure 2 - Quality Distribution of Red and White Wines histogram

Although the boxplot and histogram have provided me with insights into the quality distribution of the wines, the was histogram harder to interpret accurately due to the unbalanced sample sizes. I believe that this imbalance can skew visual interpretation and make it seem as if some quality scores are more common just since one is a larger dataset. Therefore, I normalised the data and created a new histogram.

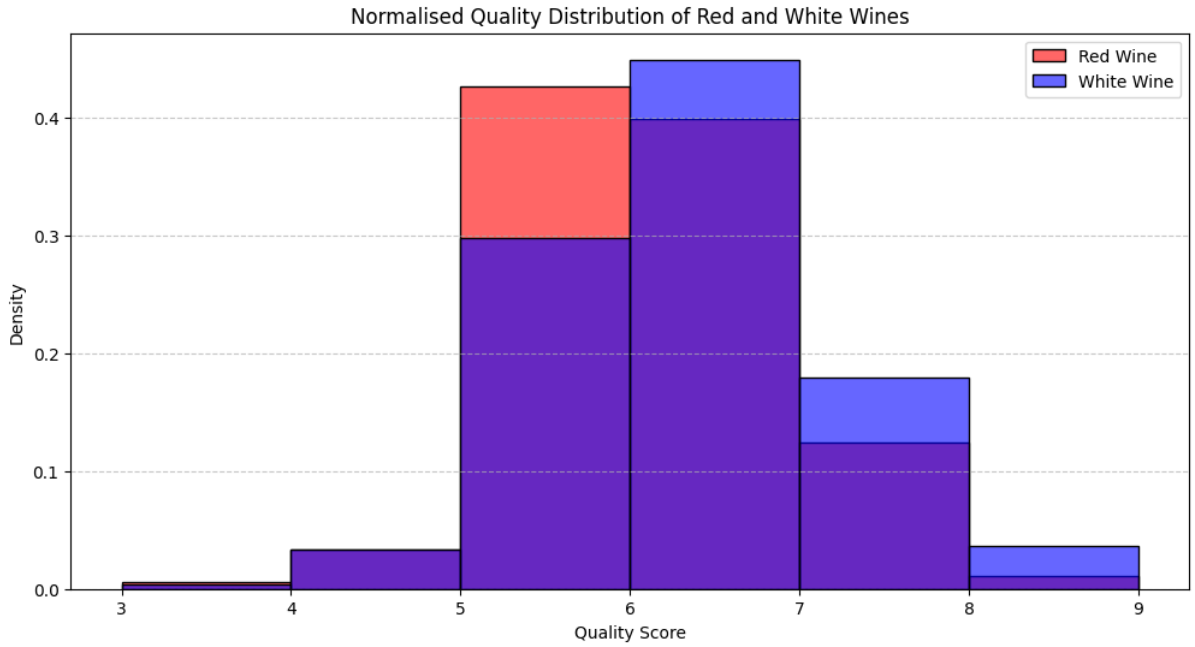

Figure 3 - Normalized Quality Distribution Red and White Wines histogram

We can now easily see the differences in their distributions. For example, white wines show a higher density at the upper end of the quality scale, particularly at scores of 7 and above. This confirms the boxplot and suggests that a greater proportion of white wines are a higher quality than red wines.

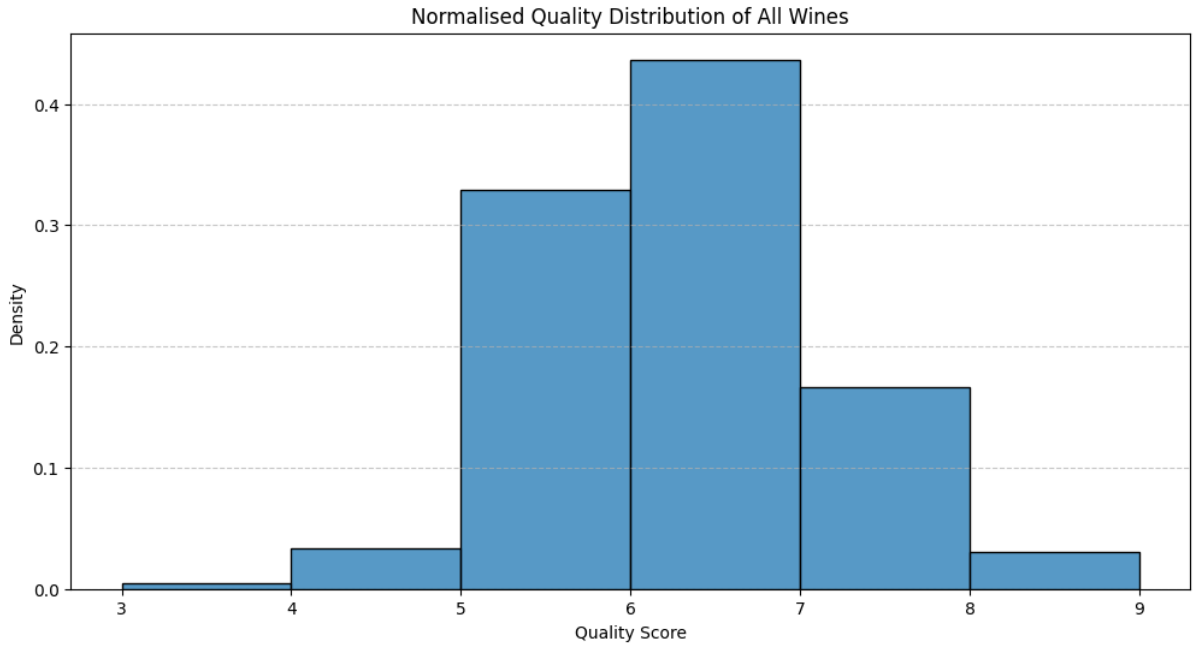

Figure 4 - Normalized Quality Distribution of All Wines

Combining the red and white wine datasets allowed me to compare wine as a whole. There isn't too much difference between this graph and both the red and white wine graph, as both types predominantly scored between 5 and 7 in quality. This indicated to me that independent of the wine colour there were similar physiochemical factors influencing the wine types.

Comparing Alcohol Content against Quality

I discretised the alcohol content variables into low, medium and high based on its distribution. These were decided by:

- low < (average – standard deviation)
- (average – standard deviation) < med < (average + standard deviation)
- (average + standard deviation) < high

I used the cut function of pandas to split the DataFrames into the pre-calculated bins. These values were:

Category	Red	White
low	194	845
med	1125	3121
high	280	932

Table 2 - Table of Red and White Wines Separated by Alcohol Category

I created a boxplot to show wine quality by alcohol content for all wines. I chose a boxplot instead of a histogram because it is better suited for group-to-group comparisons, clearly showing the medians and spreads within the categories.

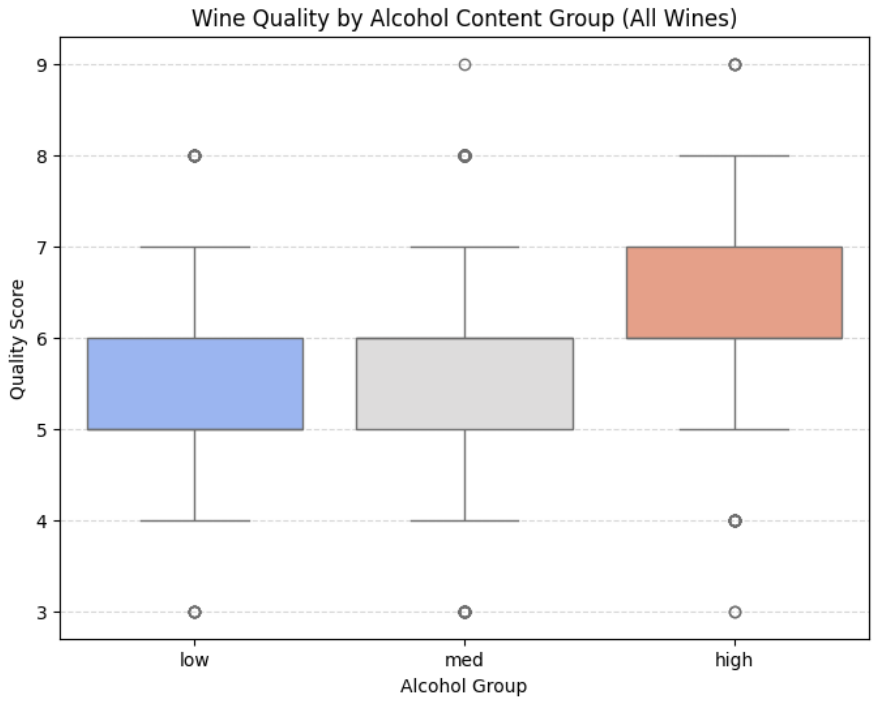

Figure 5 - Wine Quality by Alcohol Content Group boxplot

The box plot shows a clear positive relationship between alcohol content and quality. Low and medium alcohol wines have similar distributions, with median quality scores around 5 to 6, however medium alcohol wines slightly higher outliers. High alcohol wines also show a higher median and a broader upper range of quality scores.

Using Residual sugar to identify sweet and dry wines

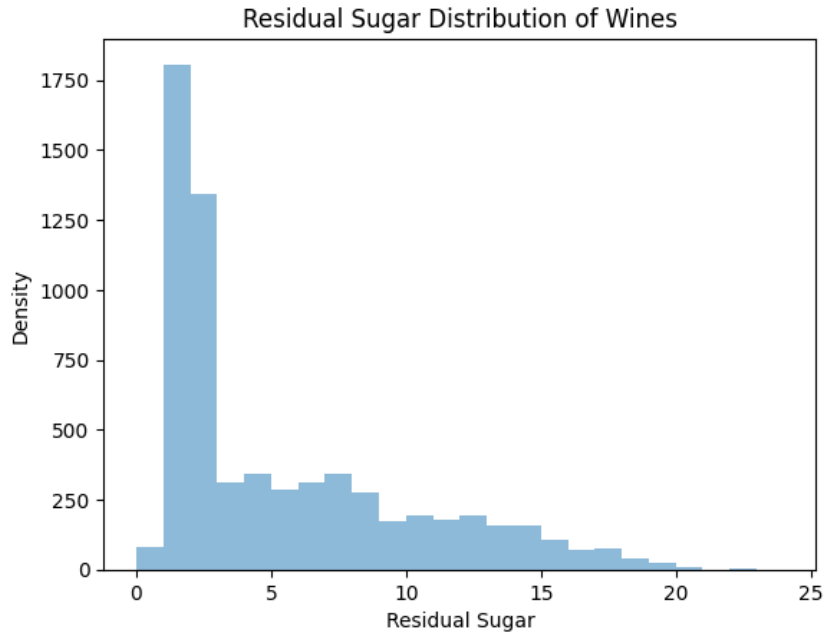

Figure 6 - Residual Sugar Distribution of Wines histogram

I plotted the residual sugar distribution and found that the data was heavily skewed for both red and white wines. I decided to use the median value of the residual sugar variable as a threshold for sweet and dry wines. This turned out to be 3.

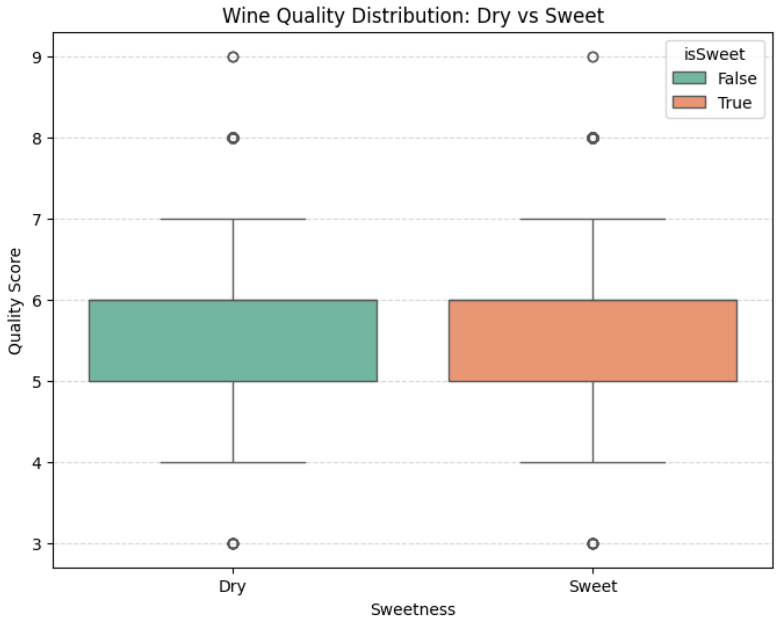

Figure 7 - Wine Quality Distribution of Dry and Sweet Wines boxplot

From the boxplots, I concluded that sweetness does not significantly impact the wine quality, as the two categories have identical plots. This suggests using residual sugar would not be useful in my quality prediction models.

Exploration Summary

After exploring the data, I found that most wines, regardless of colour, were rated between 5 and 7 in quality, but white wines had more high-quality outliers. Higher alcohol content was associated with higher wine quality with a clear trend visible through the boxplots. However, residual sugar showed little impact on wine quality, suggesting that sweetness does not impact quality. Overall, these initial findings highlighted key trends and helped spot some key correlated variables that I could use for model creation.

Determining useful attributes for Machine Learning Models

Now that the data has been explored and understood, I then moved on to analysing which variables would be most useful for creating machine learning models.

To help me make my decision, I decided to use the seaborn package to create a correlation matrix, as a heatmap, of each of the wines features against each other. I decided to use the Pearson ranking method as I could identify both feature-quality relationships as well as feature-feature to check for redundancy.

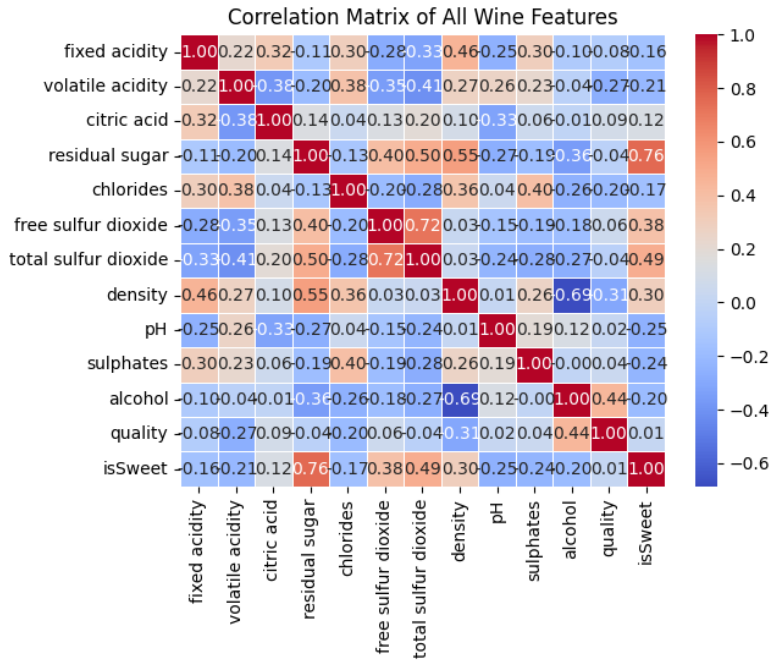

Figure 8 - Attribute Correlation Matrix

From the correlation matrix, several pairs of attributes were seen to have very high correlation with each other. This could initially seem desirable; however, this often implicates redundancy between variables that capture similar data. A good example of this is free sulfur dioxide and total sulfur dioxide. There are strongly correlated (which is expected given what they represent) however if I were to choose both these variables for my models, this would be redundant.

To select the most useful, predictive subset of features, I chose the top five attributes correlated by quality, ensuring minimal redundancy. These were:

Feature	Correlation
Alcohol	+0.44
Sulphates	+0.25
Volatile Acidity	-0.27
Density	-0.26
Chlorides	-0.20

Table 3 - Top five features and their correlated values

Model Design

Binary Classification

Now the features have been chosen, models can begin to be created. I started with binary classification as I thought it would be easier to create and understand data with classifiers before regressors.

I created an extra column on the DataFrame for a quality label where 1 would be high quality and 0 would be low quality. My initial threshold for this would be 6 as all the quality datapoints lie within 3 and 9 so 6 would be the perfect midpoint to start with.

Continuous Regression

For regression, instead of predicting a binary outcome, the models were trained to predict a continuous wine quality score. This allowed for a more fine-grained assessment of wine quality, further than a simple classification of high or low quality.

Model Choice

An 80-20 test-train split was applied, followed by scaling using StandardScaler. I selected this split to ensure that a substantial portion of the data was available for training while retaining enough data to reliably test the model. This balance is common in machine learning practices and helps prevent overfitting to the training set. Scaling was necessary because some models, such as Logistic Regression, perform better when all features are on a similar scale. Using the StandardScaler ensure features are centred and have unit variance, making better model performance.

When selecting my models for both classification and regression, I chose ones that were simple and interpretable, Logistic Regression for classification, and Linear Regression for regression. Then I chose to use Random Forest Models for both classification and regression for its ability to model non-linear correlations and resistance to overfitting.

The models selected allowed comparison between linear and non-linear approaches. This would help me to determine whether linearity assumptions are appropriate for wine quality prediction.

Results and Evaluation

Model Evaluation Approach

I used mean accuracy as a basic measure of the proportion of correct predictions for classification models. However, accuracy alone can be misleading in datasets which are imbalanced (Ghanem, et al., 2023), therefore I calculated the F1 score to account for the balance between precision and recall, this gives a more balanced view of the performance. Additionally, AUC (Area Under the ROC Curve) was measured to evaluate the model's ability to distinguish between classes across all threshold values, offering a threshold-independent metric.

For regression models, MSE (Mean Squared Error) was used to calculate the average squared difference between the predicted and actual values, highlighting the magnitude of the prediction errors, while RMSE (Root Mean Squared Error) was also calculated, showing the average error in the same units as the target variable (in our investigation this is wine quality), making more interpretable results.

Result Metrics

The results of the cross-validation process for both classification and regression models are presented below. These metrics provide a summary of model performance and can be used for a comparison between the different approaches.

Model	Mean F1 Score	Standard Deviation F1 Score	Mean Accuracy	Mean AUC	Mean MSE	Mean RMSE	Standard Deviation RMSE
Logistic Regression	0.775	0.001	0.632	0.505	---	---	---
Random Forest Classifier	0.702	0.001	0.581	0.507	---	---	---
Linear Regression	---	---	---	---	0.555	0.745	0.013
Random Forest Regressor	---	---	---	---	0.445	0.667	0.013

Table 4 – Cross-Validation Results for Classification and Regression Models

Interpretation of Results

Using the results, found in Table 4, the Random Forest Regressor has the best overall performance, as it achieved the lowest MSE and RMSE with the regression tasks. This suggests that the Random Forest model was able to capture complex, non-linear relationships within the dataset. Linear regression ranked below, it still performed reasonably well however this model is limited by the assumption of linearity between the features and the target variable (wine quality).

Among the classification models, Logistic Regression slightly outperformed the Random Forest Classifier, achieving higher mean F1 and accuracy scores. However, both models produced low AUC values. This indicates they struggled to distinguish between high- and low-quality wines across all threshold values. Logistic regression may have benefitted from its' simple decision boundary, while the Random Forest Classifier may have been limited by class imbalance. Given the low AUC values, I felt adjusting the classification threshold would be redundant, suggesting that the model limitations, not the threshold, were the cause.

Overfitting was monitored by evaluating the model performance during this k-fold cross-validation. The performance between folds was consistent, shown through the low standard deviations in F1 scores and RMSEs, therefore there were no strong indications of overfitting.

Overall, the regression models outperformed the classification models in this investigation, and the Random Forest approach provided stronger predictive power compared to linear models.

Conclusions

This project focussed on predicting wine quality based on physiochemical attributes using machine learning techniques. After exploring feature distributions and correlations, new variables were created and predictive features were selected. Classification and regression models, both linear and non-linear, were evaluated

through k-fold cross-validation. The Random Forest Regressor achieved the best performance, suggesting non-linear models would be most effective for predicting wine quality from this dataset.

Future improvements could involve using more advanced models, like XGBoost, which has been shown to outperform Random Forest in structured data predictions (Shao, et al., 2024). Expanding the dataset, particularly by collecting samples from higher quality wine, could help address class imbalance, and adding additional features, such as grape types or average temperature while growing, could improve the accuracy of the models.

Reflecting on this project, the exploration and modelling processes strengthened my skills in data preprocessing, evaluation. Although the F1 scores were solid, the relatively low AUC scores highlighted that using a broader range of metrics can provide a better of performance and told me my models struggled in distinguishing between quality classes. Although redundancy cause issues in this investigation, interpreting correlation matrices selecting features was a new experience and has strengthened my ability to build more effective models. Overall, this project was a valuable experience in applying machine learning to real-world data.

Bibliography

Cortex, P. et al., 2009. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems,* 47(4), pp. 547-553.
Ghanem, M. et al., 2023. Limitations in Evaluating Machine Learning Models for Imbalanced Binary Outcome Classification in Spine Surgery: A Systematic Review. *Brain sciences,* 13(12), p. 1723.
Shao, Z., Ahmad, M. N. & Javed, A., 2024. Comparison of Random Forest and XGBoost Classifiers Using Integrated Optical and SAR Features for Mapping Urban Impervious Surface. *Remote Sensing,* 16(4), p. 665.

Appendix 1 – Table of Figures

Appendix 2 – Table of Tables