Application and Analysis of Machine Learning Algorithms into Predicting Wine Quality and Type

Unsupervised Learning, Regression and Classification Analyses on a Real-World Wine Dataset

Abstract

This report presents an analysis of a dataset focusing on wine quality and type prediction, encompassing exploratory data analysis, unsupervised learning techniques, regression, and classification analysis. Through exploratory data analysis, insights into the dataset's characteristics and distributions were uncovered. Regression and classification analysis, including multi-linear and random forest, were utilized to predict wine quality and type using features derived from unsupervised learning techniques of Principal Component Analysis and K-means clustering. Various parameter tuning methods were explored to enhance predictive performance. These findings offer insights for further research and application into wine.

Candidate ID:

Student ID: 200615628

Table of Contents

Introduction

Wine Quality Dataset

Economically, the wine industry relies on consumer perception of quality and taste to influence market demand and trade. Culturally, wine quality serves as a reflection of the craftsmanship and dedication of winemakers, embodying the unique terroir and traditions of specific regions (Werdelmann, 2014).

The wine quality dataset, sourced from the UC Irvine (UCI) Machine Learning Repository, presents a comprehensive exploration into the common characteristics of red and white wines, particularly wines of the Vinho Verde region in Portugal. This dataset is split into two distinct sets, one for red and one for white wines, offering an array of features associated with the physiochemical testing and sensory quality assessment of each respective wine types. It encapsulates the common physiochemical attributes that could contribute to the quality of Vinho Verde wines. These attributes consist of eleven measurable and two non-measurable variables. The dataset caters to a spectrum of analytical methodologies, making it suitable for regression analysis, unsupervised learning and classification analysis with the following variables. (Cortez, Almeida, Matos, & Reis, 2009)

Variable	Data Type	Role	Short Description
Fixed Acidity	Continuous	Feature	Tartaric acid concentration
Volatile Acidity	Continuous	Feature	Acetic acid concentration
Citric Acid	Continuous	Feature	Citric acid concentration
Residual Sugar	Continuous	Feature	Remaining sugar concentration
Chlorides	Continuous	Feature	Sodium chlorides concentration
Free Sulfur Dioxide	Continuous	Feature	Free sulfur dioxide concentration
Total Sulfur Dioxide	Continuous	Feature	Total sulfur dioxide concentration
Density	Continuous	Feature	Mass/unit volume
pH	Continuous	Feature	pH level
Sulphates	Continuous	Feature	Sulphate concentration
Alcohol	Continuous	Feature	Alcohol concentration
Quality	Integer	Target	Sensory evaluation of quality
Type	Categorical	Target	Red or white wine

Report Objective

This report aims to address specific substantial issues and showcase completed objectives of using regression analysis to predict the numerical quality ratings of Vinho Verde wines, classification analysis to predict wine type and unsupervised learning techniques as feature selection for regression and classification analyses.

Key Insights

Data Preprocessing

Raw data typically include inaccurate, missing, duplicated or irrelevant entries. Hence, data cleaning is often essential during data preprocessing to ensure quality data (Elgabry, 2019). Besides data cleaning, appropriate data manipulation is also done to make sure the data is suitable for the necessary processes afterwards. These processes include Exploratory Data Analysis, Machine Learning or Deep Learning etc. From the UCI Machine Learning Repository, the datasets are noted to be clean, but it is safer to ensure that by running through some code on R. To suit the objectives set, a variable of wine type (red or white) is added to the original variables within the datasets and both datasets are merged into one big dataset labelled as wine. Next, to check for any missing or duplicated values within the wine dataset, simple commands per figure below are used on R.

```
> sum(is.na(wine))
[1] 0
```

```
> sum(duplicated(wine))
[1] 1177
```

There are no missing values found but there are 1177 duplicate entries noted. However, referencing from the original paper, distinct wine samples were collected per row to make up the dataset (Cortez, Cerdeira, Almeida, Matos, & Reis, 2009). Hence, there is no need to remove these duplicates entries.

Exploratory Data Analysis (EDA)

Descriptive Summary Statistics

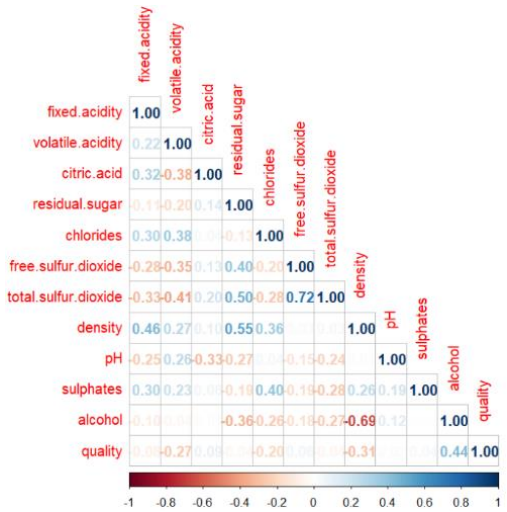
Variables	Minimum	Q1	Median	Mean	Q3	Max
Fixed acidity	3.800	6.400	7.000	7.215	7.700	15.900
Volatile acidity	0.0800	0.2300	0.2900	0.3397	0.4000	1.5800
Citric acid	0	0.2500	0.3100	0.3186	0.3900	1.6600
Residual sugar	0.600	1.800	3.000	5.443	8.100	65.800
Chlorides	0.00900	0.03800	0.04700	0.05603	0.06500	0.61100
Free sulfur dioxide	1.00	17.00	29.00	30.53	41.00	289.00
Total sulfur dioxide	6.0	77.0	118.0	115.7	156.0	440.0
Density	0.9871	0.9923	0.9949	0.9947	0.9970	1.0390
pH	2.720	3.110	3.210	3.219	3.320	4.010
Sulphates	0.2200	0.4300	0.5100	0.5313	0.6000	2.000
Alcohol	8.00	9.50	10.30	10.49	11.30	14.90
Quality	3	5	6	5.818	6	9
Type	**Red:** 1599			**White:** 4898		

From the above table, the following can be found. This gives an idea on data characteristics.

Variables	Range	Variance	Standard Deviation	Interquartile Range
Fixed acidity	12.100	1.681	1.296	1.300
Volatile acidity	1.5000	0.0271	0.1646	0.17000
Citric acid	1.6600	0.0211	0.1453	0.1400
Residual sugar	65.200	22.636	4.757	6.300
Chlorides	0.60200	0.00122	0.03503	0.02700
Free sulfur dioxide	288.00	315.04	17.74	24.00
Total sulfur dioxide	436.0	319.5	56.5	79.0
Density	0.0519	0.0000089	0.00299	0.0047
pH	1.290	0.0258	0.1607	0.210
Sulphates	1.7800	0.0258	0.1488	0.1700
Alcohol	6.90	1.42	1.19	1.80
Quality	6	0.76	0.87	1

Correlation

Correlation quantifies the strength and direction of relationships between variables in datasets. It assesses how changes in one variable correspond to changes in another. From the plot below, most variables seem to be neutral. Although, total and free sulfur dioxide are strongly positive related while alcohol and density are strongly negative related. However, correlation does not imply causation. It is necessary to determine underlying relationships through additional analysis.



Variable Importance

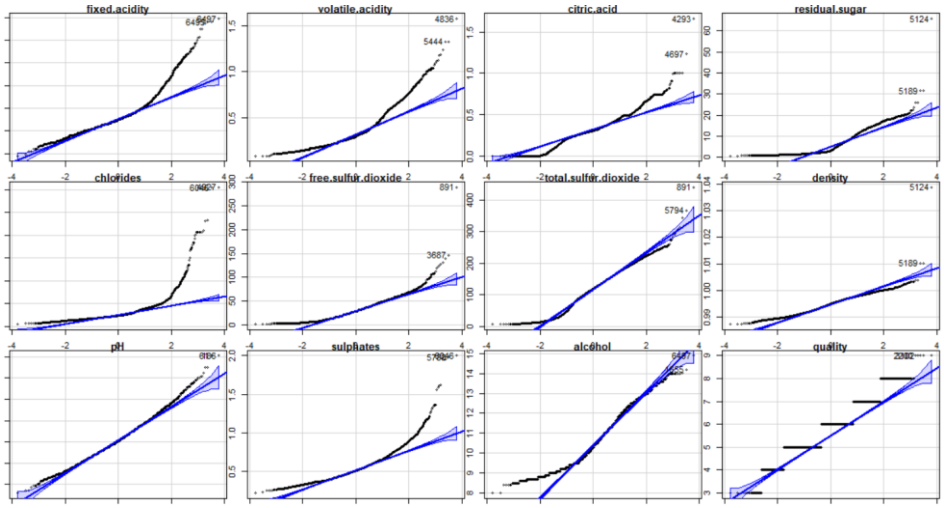
Variable importance aims to identify the relative significance of predictor variables in explaining variation in the response variable and provide insights into which features have substantial impact on the outcome of interest. For quality, it seems that alcohol has the highest impact, followed by density and volatile acidity. For wine type, there is no difference in importance. More analysis should be done.

	Overall
fixed.acidity	6.203149
volatile.acidity	22.211509
citric.acid	6.918488
residual.sugar	2.982355
chlorides	16.507716
free.sulfur.dioxide	4.476745
total.sulfur.dioxide	3.338178
density	25.890303
pH	1.572294
sulphates	3.103902
alcohol	39.970496

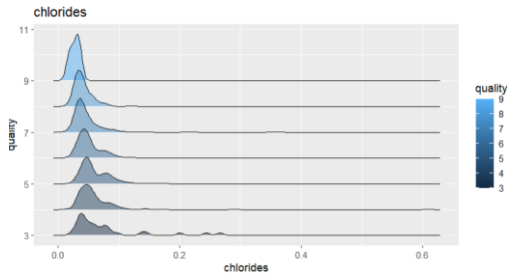
	red	white
fixed.acidity	0.7837824	0.7837824
volatile.acidity	0.9013932	0.9013932
citric.acid	0.6080022	0.6080022
residual.sugar	0.6718949	0.6718949
chlorides	0.9457493	0.9457493
free.sulfur.dioxide	0.8485174	0.8485174
total.sulfur.dioxide	0.9531864	0.9531864
density	0.7736670	0.7736670
pH	0.7254738	0.7254738
sulphates	0.8312107	0.8312107
alcohol	0.5110966	0.5110966

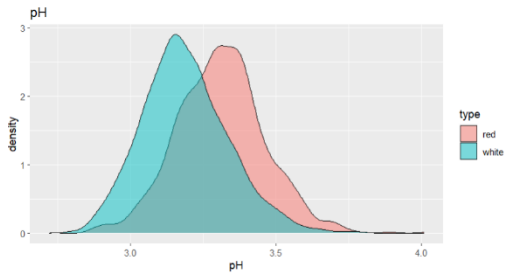
Plots and Graphs

Quantile-Quantile (QQ) plots visually assess the similarity between the distribution of a dataset and a theoretical distribution. If the points on QQ plots approximately fall on a straight line, the dataset follows a normal distribution. Deviations from this line indicates departure from normality in dataset. From the plot below, some QQ plots are observed to be approximately normally distributed, while others such as acidity and chlorides suggest right-skewness.

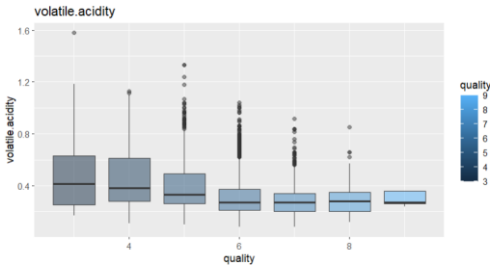


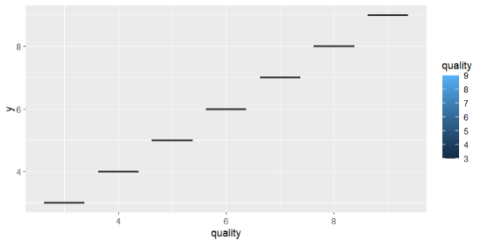
Density plots visualise the distribution of continuous variables within a dataset by shape of curve and shows where data tends to cluster. From the following selected plots representing wine quality and type, it confirms that some variables are more right-skewed while some are normally distributed.





Boxplots give a visual representation of summary statistics and helps identify outliers. From selected plots below, the plots are aligned with the table in descriptive summary. However, it is difficult to subjectively determine the relationship between variables.





To conclude, EDA is a foundational step in the data analysis process, offering insights into the characteristics of datasets. Through various techniques such as summary statistics and visualizations, EDA enables a deeper understanding of the data, identifying potential relationships between variables. (Vanawat, 2023) However, it is not guaranteed that EDA would produce conclusive results.

Unsupervised Learning

As EDA was inconclusive in providing clear insights for determining feature importance and selection, unsupervised learning algorithms offer an approach to extract meaningful information from data. By uncovering hidden patterns and relationships, unsupervised learning can help overcome the limitations of inconclusive EDA and guide subsequent analyses, including assessment of feature importance and selection (Field, 2017). In this report, features and variables are used interchangeably.

Dimension reduction techniques like Principal Component Analysis (PCA) can reveal intrinsic structures and relationships with the data. By analysing the variance explained by each principal component, the importance of different features can be inferred by capturing the underlying variability or patterns in the data and theoretically aid in identifying important variables for analyses (Tayade, Patil, Phalle, Kazi, & Powar, 2019). In this report, PCA will be used to select features for regression analysis of wine quality prediction.

Clustering techniques such as K-means can identify homogenous population groups that contribute most to data structure and variation. Features exhibiting high and low within-cluster similarity may be important for distinguishing different clusters. In theory, these groups of features can be used to build predictive models for machine learning (Azhar & Thomas, 2019). In this report, K-means is used to guide feature selection for classification analysis of wine type prediction. Both PCA and K-means require standardisation of numerical variables (Field, 2017).

Algorithms

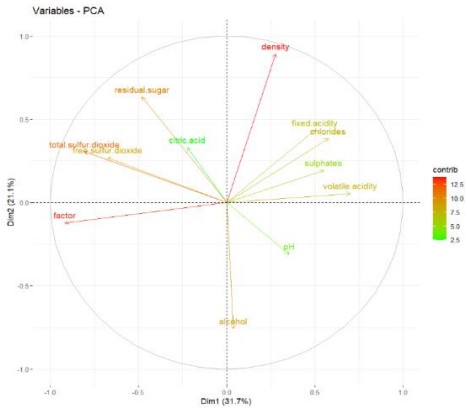
1. Dimension Reduction (PCA) for Wine Quality Prediction

By definition, PCA transforms the original features into a new set of uncorrelated variables (principal components, PC) capturing the maximum variance in the data (Zakaria, Wan Yusoff, & Muhammed, 2024). These PCs aid in deciding how many variables to retain and which are important based on the explained variance, guiding feature selection for wine quality prediction.

PCA is performed on standardised numerical data and from the summary table and plots, 4 PCs contain 73% of total variance. For purposes of this report, 4 variables can be chosen, covering 73% of variance.

	PC1	PC2	PC3	PC4
Standard deviation	1.9518	1.5902	1.2496	0.9853
Proportion of Variance	0.3175	0.2107	0.1301	0.0809
Cumulative Proportion	0.3175	0.5282	0.6583	0.7392

Feature importance can be inferred from variable contribution towards PCs with the following plot, with the colour corresponding the higher contribution being red and lower contribution being black.

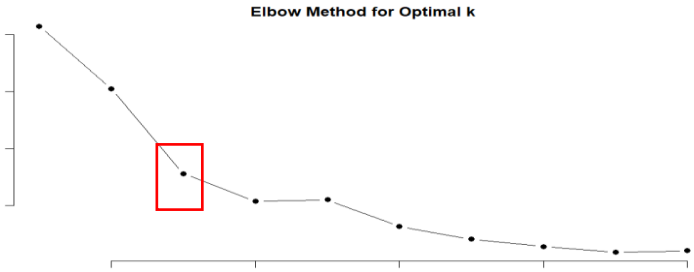


The 4 variables with highest contribution are Density, Type, Total sulfur dioxide and Residual Sugar.

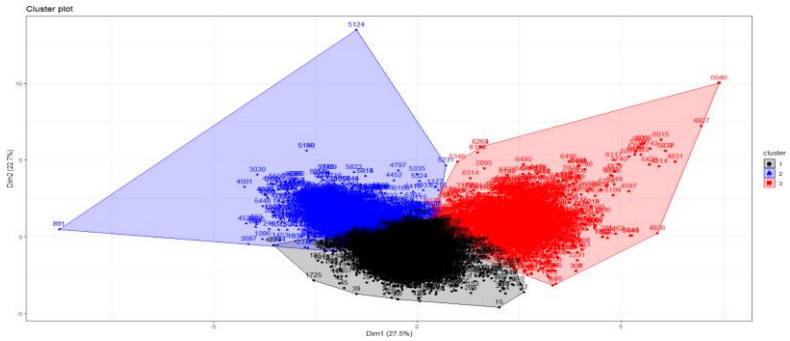
2. Identifying population groups (K-means Clustering) for Wine Type Prediction

By definition, K-means clustering separates data into clusters based on variable similarities. As such, features contributing significantly to cluster separation can be considered as important, which aids in feature selection (Azhar & Thomas, 2019).

The optimal number of clusters (K) is first determined using the elbow method which is commonly used. It runs the k-means algorithm on the dataset for a range of k values (1:10) and plots the sum of squared errors (SSE) between data points and their assigned cluster centroids. The 'elbow' or bend in the plot represents the point where the rate of decrease in SSE levels out, suggesting that increasing cluster numbers does not significantly decrease SSE (Nie, Li, Wang, & Li, 2023). Optimal K is 3.



K-means algorithm is run with the optimal K=3, where the dataset will be separated into 3 clusters as seen with the following plot. The whole process is done on standardised numerical data.



In K-means clustering, centroid values play a role in identifying homogenous population groups for feature selection by providing insights into how variables contribute to the formation of clusters. The centroid of each cluster represents the average value of variables for data points assigned to that cluster. As such, consider the maximum centroid values of each cluster for each feature as a measure of feature importance in K-means clustering. Variables with higher maximum centroid values are considered more important for cluster differentiation and play a significant role in defining cluster boundaries. This makes for easier interpretation of the relative importance of each feature and its

contribution to the clustering process (Umargono, Suseno, & Gunawan, 2020). The output below provides information on feature importance based on their association with k-clusters identified.

```
> print(sort(features.type, decreasing = FALSE))
              alcohol          citric.acid       residual.sugar  free.sulfur.dioxide
                  "1"                  "2"                  "2"                  "2"
total.sulfur.dioxide              density        fixed.acidity     volatile.acidity
                  "2"                  "2"                  "3"                  "3"
           chlorides                   pH            sulphates
                  "3"                  "3"                  "3"
```

The values represent the cluster number to which each variable is most strongly associated based on the maximum centroid value within that cluster. For example, alcohol is most strongly associated with cluster 1. Citric acid, residual sugar, free sulfur dioxide, total sulfur dioxide and density are most strongly associated with cluster 2 and the other variables are most strongly associated with cluster 3. These features are identified as being most representative of each cluster based on the maximum centroid values within each cluster. Therefore, they are likely to be important for distinguishing between different clusters of wine types in classification analysis.

Results and Contrast

Referencing research (Mangubat, 2022), several distinctions exist between the methodologies and outcomes of the 2 analyses. Firstly, both PCAs were conducted on different datasets but are related to wine. The PCA in this report was done on variants of Vinho Verde wine with 13 physicochemical and sensory features while the referenced PCA was done on wine samples from 3 different cultivars with 13 chemical concentrations. Hence there will be variations in the dimensionality and composition of feature space analysed. Secondly, the interpretation of PCs and extraction of meaningful insights differ between the 2 analyses. PCA in this report aims to identify the most influential features contributing to variance in dataset, the referenced PCA aims to find the best low-dimensional representation of variance within its dataset. As such, there are differences to interpretation of feature importance and identification of relevant patterns. Lastly, the objectives of PCA analyses vary. PCA in this report is conducted as part of a broader investigation into wine quality while the referenced PCA may have focused on different applications within the wine industry. It was not specified. Differing research aims could influence the choice of PC and interpretation of results, leading to disparity in outcomes of both PCA analyses. This is evident from differing results in choice of PCs, 4 for this report and 3 in referenced.

Referencing research (Mano, 2021), some similarities and differences exist between methods and outcomes of analyses. Although the same dataset is used, the choice of k differs with k = 3 in this report using elbow method and referenced research, k = 2 using domain knowledge (red & white wine). The variations in parameter selection impacts the granularity of clusters, influencing the interpretation of clustering results. Moreover, interpretation and analysis of clustering outcomes differs between both k-means. In this report, k-means clustering focused on identifying homogenous group for subsequent classification analysis, while referenced k-means could have emphasized different unspecified objectives. Evidently, divergent research contexts and methodology influences differing clusters and its interpretation.

Analysis Conclusion

Variables selected based on highest contribution in PCA are Density, Type, Total sulfur dioxide and Residual Sugar. Variables selected based most representative in k-means clusters are alcohol, citric acid, residual sugar, free sulfur dioxide, total sulfur dioxide, density, fixed acidity, volatile acidity, chlorides, pH and sulphates. Using domain knowledge for k-means could yield better feature selection. This did not line up with variable importance performed during EDA.

Regression Analysis

Market demand is crucial for the industry and wine quality is an important factor influencing consumer behaviour and market dynamics. As consumers increasingly expect for quality wines, it is prudent to

produce wines that meet consumer expectations. Wine quality is a sensory evaluation performed by human tasters, who are prone to subjective opinions. (Werdelmann, 2014) As such, using regression analysis could give a potential gauge of wine quality for wineries by building a model, linking common physiochemical indicators, like alcohol and acidity levels to respective quality rankings.

In this report, various regression analysis models were built with parameter tuning to find the best possible model with the objective of predicting wine quality. The whole dataset split into 70:30 train data to test data. Error measures for model performance will be Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). MSE calculates average squared difference between actual and predicted values, providing a measure for overall model fit and dispersion of prediction errors. RMSE is the square root of MSE, providing a measure of average deviation from the mean. Lower MSE and RMSE values typically indicate better model and predictive accuracy. (Frost, 2024)

Algorithms

1. Multi-Linear Regression

Multi-linear regression is a common statistic technique for prediction by analysing the linear relationship between multiple predictor variables and a continuous dependent variable. In this approach, various physiochemical indicators serve as independent predictor variables and wine quality rating serve as dependent predicted variable. The 1st model had the 4 variables selected from PCA dimension reduction, however the MSE (0.6578) and RMSE (0.81108) of test data was not ideal. The 2nd model used all variables as independent predictors, with lower MSE (0.5412) and RMSE (0.7356) of test data. The 3rd model removed less and insignificant variables, but RMSE with train data (0.7323) was higher than with those variables, hence there was no need to remove them when building subsequent regression models. The final model used all variables and 10-fold cross validation (cv) as parameter tuning and taking the average of MSE and RMSE over 10 iterations of model training and testing. Multicollinearity was observed in some variables, typically with values of more than 10 (Statistics Solutions, 2024) .

```
MSE.test.mean <- mean(MSE.test)
RMSE.test.mean <- mean(RMSE.test)
## MSE = 0.5395028
## RMSE = 0.7338793
```

```
residual.sugar
     10.450772
       density
     26.059556
```

2. Ridge Regression

Considering the presence of multicollinearity, a decision had to be made between employing ridge regression or lasso regression. However, given the many significant predictor variables, ridge regression was chosen. Ridge regression incorporates regularisation parameter (lambda, λ) to penalise large coefficients and improve generalisation performance when dealing with high dimensional datasets with correlated independent variables. (Shubham, 2024) Cross validation was utilized as parameter tuning to find the best option for lambda. Choosing lambda value with minimum cross-validated error tends to produce models prone to overfitting while choosing lambda value within 1 standard error of minimum cross-validated error produces simpler models, prioritizing generalisation and lessen overfitting. (Sneiderman, 2020) 2 models were tested with 2 lambda values. Lower MSE and RMSE values of test data will determine the final model.

```
MSE.ridge <- mean(ridge.error^2)
RMSE.ridge <- sqrt(mean(ridge.error^2))
## MSE = 0.5999168
## RMSE = 0.774543
```

```
MSE.ridge2 <- mean(ridge2.error^2)
RMSE.ridge2 <- sqrt(mean(ridge2.error^2))
## MSE = 0.5900543
## RMSE = 0.7681499
```

Final ridge model chosen has minimum cv error lambda value with lower MSE and RMSE value.

3. Classification And Regression Tree (CART)

CART regression constructs a decision tree model that recursively splits data into homogeneous subsets, with each split optimising a selected criterion and minimizing variance of dependent variable within each subset. Through a process called 'growing the tree', the model identifies the most

influential predictor variables and their optimal thresholds for data partition, allowing the creation of intuitive decision rules to predict selected dependent variables based in the values of independent variables. However, this process leads to model overfit, which does not work well with new data. Using cross validation to calculate errors and pruning the tree model at the point where error is minimum lessens model overfitting. (Hoare, 2024) The 1st CART model was not pruned, hence the MSE and RMSE error were much higher than the 2nd (final) model which was pruned at minimum cross-validated error.

```
MSE.cart <- mean(cart.error^2)
RMSE.cart <- sqrt(mean(cart.error^2))
## MSE = 0.7596302
## RMSE = 0.8715677
```

```
MSE.cart2 <- mean(cart2.error^2)
RMSE.cart2 <- sqrt(mean(cart2.error^2))
## MSE = 0.5985617
## RMSE = 0.7736677
```

Results and Contrast

Multi-Linear Regression slightly outperformed other models in predicting wine quality by lowest error.

Final_Model	MSE	RMSE
Linear Regression	0.5395028	0.7338793
Ridge Regression	0.5900543	0.7681499
CART	0.5985617	0.7736677

Referencing research (Nguyen, 2020), some similarities and differences exist between methods and outcomes of analyses. Although the same source of data, only red wine was used for regression analysis while this report additionally utilised white wine. Lasso regression and random forest models were the slight difference from ridge and CART models within this report. By comparison of RMSE, best model in reference was random forest (0.5843) while best model in report was multi-linear regression (0.7339). This disparity could stem from several factors. Different regression algorithms bring about different modelling assumptions, complexity and flexibility. Random forest effectively captured complex non-linear relationships while multi-linear regression emphasizes linear relationships, potentially better fitted to specific characteristics of dataset. Additionally, variations in feature selection and parameter tuning methods contribute to differences in model performance and RMSE results. Also, considering the inherent variability and randomness in the dataset and differences in sample size and feature representation, this could influence the performance and RMSE of models.

Analysis Conclusion

To conclude, multi-linear regression emerged as the better performing model for predicting wine quality. Most variables except citric acid and chlorides are significant. Despite its simplicity compared to other models, it demonstrated marginally superior performance by MSE and RMSE measures. However, not discounting the potential utility of ridge and CART, which may be better models from the perspective of marginal impact. To further improve model performance, several avenues could be explored. Conducting more extensive data-preprocessing, like outlier removal and feature scaling, could mitigate the impact of data irregularities on model performance. Further exploration of different hyperparameters and regularisation techniques for ridge and CART models may also reveal superior configurations.

Classification Analysis

Quality control and assurance represent significant challenges in the wine industry, where maintaining consistency and authenticity across wine varieties is paramount (Hartwell, 2021). Usage of classification analysis for wine type prediction provides a systematic, data-driven approach to verify and ensure accuracy of wine classifications. Leveraging classification models built on a dataset with physiochemical properties of red and white wine, wineries can classify wines as red or white based on unique characteristics without actively looking. This enables stringent quality control measures.

In this report, various classification analysis models were built with suitable parameter tuning to find the best possible model with the objective of predicting wine type. The whole dataset split into 70:30 train data to test data with baseline dependent variable as ‘white’ as the majority quantity. Model performance will be measured by accuracy, precision and sensitivity using confusion matrix. Accuracy

is defined as the ratio of correct predictions to overall predictions, precision is the ratio of true positive (correct) predictions to all positive predictions and sensitivity measures the proportion of actual positive predictions that are correctly identified (Agrawal, 2024). Higher values indicate better performance.

Algorithms

1. Logistic Regression

Logistic Regression serves as a valuable tool for predicting categorical outcomes by modelling the relationship between predictor and response variables. (Jurafsky & Martin, 2023) In this approach, the probability of wine belonging to each type is predicted by leveraging on common physiochemical features and the model estimates likelihood of each wine being classified as red or white based on weighted combination of variables. The 1st model used all features selected from k-means. With confusion matrix outputted, performance metrics indicate high accuracy, precision and sensitivity of over 99%. The 2nd model removed insignificant variables, yet not affecting performance. Given the same performance metrics yielded, suggesting model performance is not influenced by presence of insignificant features. The 2nd model would be preferred for its simplicity and interpretability of new data. High values of performance metrics suggest the model is effective in correctly classifying instances, with 99.5% overall correctness, 99.7% precision in positive predictions and 99.6% sensitivity to positive instances.

```
          observed
predicted white  red
    red       5  476
    white  1464    4
```

```
## Accuracy = 0.9953822
## Precision = 0.9972752
## Sensitivity = 0.9965963
```

2. Naïve Bayes

Naïve Bayes classification provides a method for predicting categorical outcomes based on distinctive characteristics. It leverages the conditional probability of observing specific features given each wine type to make predictions and calculates class probabilities based on joint probabilities of individual features. (Yang, 2019) Similarly, the 1st model used all variables. Performance metrics derived from confusion matrix indicate high performance of over 97%. The 2nd model attempted parameter tuning using Laplace smoothing (α) in case of zero probability. The optimal value was 0, the default value. High values of performance metrics suggest the Naïve Bayes model is also effective in correctly classifying instances, with 97.7% accuracy, 99% precision and 97.9% sensitivity.

```
          observed
predicted white  red
    white  1438   14
    red      31  466
```

```
## Accuracy = 0.9769112
## Precision = 0.9903581
## Sensitivity = 0.9788972
```

3. Random Forest

Random Forest is an algorithm that constructs multiple decision trees and combines each prediction to produce a final output, useful where skewed data is present. In this approach, random forest leverages the collective iterations of decision trees to classify wine types based on physiochemical variables while mitigating overfitting by aggregating predictions of multiple trees and improving generalisation performance. (Yiu, 2019) The 1st model used all variables and generating confusion matrix. Performance metrics indicate high performance of over 99%. The 2nd model utilised tuning for 2 parameters with cv, number of randomly selected predictor variables considered at each split when constructing individual decision trees (mtry) and number of trees to generate (ntree). Best parameters while minimizing cv error was mtry=2 and inconclusive ntree. Performance metrics outputted from confusion matrix indicate high performance of over 99%, marginally better than 1st model. This suggests the Random Forest model is also effective in correctly classifying instances, with 99.7% accuracy, 99.7% precision and 99.9% sensitivity. Total sulfur dioxide and chlorides are most significant.

```
          observed
predicted white  red
    white  1468    5
    red       1  475
```

```
## Accuracy = 0.9969215
## Precision = 0.9966056
## Sensitivity = 0.9993193
```

Results and Contrast

Random forest slightly outperformed other models in predicting wine type by highest performance.

Model	Accuracy	Precision	Sensitivity
Logistic Regression	0.995	0.997	0.997
Naive Bayes Classifier	0.977	0.990	0.979
Random Forest	0.997	0.997	0.999

Referencing research on Naïve Bayes (Simonovikj, 2020), some similarities and differences exist between methods and outcomes of analyses. Although the same source of data, train-test-split were assumed different proportions and sample sizes. By comparison of metrics, the referenced values ranging 95-96% and report values ranging 97-99%. The disparity could stem from variations in pre-processing methods and randomness within train-test-split. The sample size and variables selection difference would likely be a factor. It underlines the nuanced nature of predictive modelling and impact of subtle variations in factors leading to outcomes in model performance.

Referencing research for other models (Ivamoto, 2020), some similarities and differences exist between methods and outcomes of analyses. Similarly, from the same source of data, train and test set were of different proportions and sample sizes. For random forest being best performer in both instances, performance values were similar despite variations in pre-processing methods and randomness in train-test-split. This suggests the model's predictive reliability transcends variations, providing consistent dependable performance and utility. For logistic regression, the slight difference from reference using linear regression with binary categorical variable shows in worse performance metrics, highlighting the distinct characteristics and assumption difference present in linear and logistic regression. Because linear regression assumes linear relationship between predictor variables and response, possible potential oversimplification of underlying data structure and deviation in predictive accuracy of categorical outcomes. While logistic regression models the probability of categorical outcomes and is specifically designed for binary classification.

Analysis Conclusion

To conclude, random forest model emerged as the superior performer amongst other models evaluated for wine type prediction. Most significant features are total sulfur dioxide and chlorides. With its ability to handle complex relationships within the data, random forest demonstrates robust predictive power and yields highest performance metrics. To further enhance effectiveness, several strategies can be considered. Optimising parameters of maximum depth of trees and minimum node size can fine-tune model performance and mitigate overfitting. Further exploration of different hyperparameters and regularisation methods for other models may reveal superior configurations.

Conclusion

In conclusion, this report explored various aspects of wine quality and type prediction through a comprehensive analysis of the Vinho Verde wine dataset. Through EDA, insights into the characteristics and distributional properties of the dataset were identified. However, potential relationships between variables and feature selection were difficult to identify. Hence unsupervised learning techniques were employed to help with feature selection using PCA and K-means. Regression and classification analyses were conducted on selected features to predict wine quality and type respectively. Multi-linear regression and random forest models were found to be best performing with use of parameter tuning and noting which variables are most significant. However, there are limitations in these analyses. Despite being best performing models, multi-linear regression model still did not perform well with 0.733 RMSE and random forest model with 99% accuracy may be biased towards white wine. Presumably due to unbalanced data, majority of white over red wine and majority of 5 and 6 for quality may have affected model performance and interpretation. Moreover, further exploration into balancing data and additional data collection, more than just common physiochemical tests, could enhance the robustness and performance of predictive models. Ultimately, this report serves as a foundation for continued research and exploration for Vinho Verde wine.

References

Agrawal, S. K. (2024, 02 15). *Metrics to Evaluate your Classification Model to take the right decisions*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/

Azhar, M. A., & Thomas, P. A. (2019). Comparative review of feature selection and classification modeling. *International conference on advances in computing, communication and control (ICAC3)* (pp. 1-9). IEEE.

Cortez, P., Almeida, F., Matos, T., & Reis, J. (2009). *Wine Quality Dataset*. Retrieved from UCI Machine Learning Repository: https://doi.org/10.24432/C56S3T

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 547-553.

Elgabry, O. (2019, March 1). *The Ultimate Guide to Data Cleaning*. Retrieved from towards data science: https://towardsdatascience.com/the-ultimate-guide-to-data-cleaning-3969843991d4

Field, C. (2017). *Unsupervised Learning: Clustering and Dimensionality Reduction.* Hoboken: John Wiley & Sons, Inc.

Frost, J. (2024). *Mean Squared Error (MSE)*. Retrieved from Statistics By Jim: https://statisticsbyjim.com/regression/mean-squared-error-mse/

Hartwell, R. (2021). Analytical Testing in Wine Making: Transform Quality Control Into Quality Design. *Thermo Fisher Scientific*.

Hoare, J. (2024). *Machine Learning: Pruning Decision Trees*. Retrieved from Displayr: https://www.displayr.com/machine-learning-pruning-decision-trees/

Ivamoto, V. (2020). *Wine Type and Quality Prediction With Machine Learning.* Massachusetts: HarvardX PH125 Data Science.

Jurafsky, D., & Martin, J. H. (2023). Speech and Language Processing. Stamford, California, USA.

Mangubat, D. (2022, 8 2). *PCA on Wine Data.* Retrieved from RPubs by RStudio: https://rpubs.com/DarStats_123/867659

Mano, R. (2021). *Kmeans.* Retrieved from RPubs by RStudio: https://rpubs.com/mano_r/649764

Nguyen, D. (2020, 11 23). *Red Wine Quality Prediction Using Regression Modeling and Machine Learning*. Retrieved from towardsdatascience: https://towardsdatascience.com/red-wine-quality-prediction-using-regression-modeling-and-machine-learning-7a3e2c3e1f46

Nie, F., Li, Z., Wang, R., & Li, X. (2023). An Effective and Efficient Algorithm for K-Means Clustering With New Formulation. *IEEE Transactions on Knowledge and Data Engineering*, 3433-3443.

Shubham, J. (2024, 02 07). *Lasso & Ridge Regression | A Comprehensive Guide in Python & R (Updated 2024)*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2017/06/a-comprehensive-guide-for-linear-ridge-and-lasso-regression/

Simonovikj, B. (2020, 05 06). *Wine Type Prediction With Supervised And Unsupervised Learning.* Retrieved from RPubs by RStudio: https://rpubs.com/Billie/703946

Sneiderman, R. (2020, 11 06). *From Linear Regression to Ridge Regression, the Lasso, and the Elastic Net*. Retrieved from towardsdatascience: https://towardsdatascience.com/from-linear-regression-to-ridge-regression-the-lasso-and-the-elastic-net-4eaecaf5f7e6

Statistics Solutions. (2024). *Multicollinearity*. Retrieved from Complete Dissertation by Statistics Solutions: https://www.statisticssolutions.com/multicollinearity/

Tayade, A., Patil, S., Phalle, V., Kazi, F., & Powar, S. (2019). Remaining useful life (RUL) prediction of bearing by using regression model and principal component analysis (PCA) technique. *Vibroengineering Procedia*, 30-36.

Umargono, E., Suseno, J. E., & Gunawan, S. V. (2020). K-means clustering optimization using the elbow method and early centroid determination based on mean and median formula. *In The 2nd international seminar on science and technology (ISSTEC 2019)* (pp. 121-129). Atlantis Press.

Vanawat, N. (2023, 05 05). *How to perform exploratory data analysis - a guide for beginners*. Retrieved from Analytics Vidhya: https://www.analyticsvidhya.com/blog/2021/08/how-to-perform-exploratory-data-analysis-a-guide-for-beginners/

Werdelmann, T. (2014). Quality and Value Creation on the Premium Wine Market. *Journal of Applied Leadership and Management*, 47-72.

Yang. (2019, 9 9). *An Introduction to Naïve Bayes Classifier.* Retrieved from towardsdatascience: https://towardsdatascience.com/introduction-to-na%C3%AFve-bayes-classifier-fa59e3e24aaf

Yiu, T. (2019, 06 12). *Understanding Random Forest.* Retrieved from towardsdatascience: https://towardsdatascience.com/understanding-random-forest-58381e0602d2

Zakaria, N., Wan Yusoff, N. W., & Muhammed, N. (2024). A comparative study of classical and robust principal component analysis in historical multivariate data. *AIP Conference Proceedings.* Pahang: ACB-ISBE.