Degree project

UNIVERSITY OF SKÖVDE

Wine Quality Prediction Model Using Random Forest, Support Vector Machine & Extreme Gradient Boosting

Master Degree Project in Informatics with a specialization in Data Science

Second Cycle 15 credits

Spring term 2026

Student: Mikias Andemariam

Supervisor: Tove Helldin

Examiner: Joe Steinhauer

Abstract

Wine quality is traditionally assessed through sensory evaluation, which depends on human judgement and can vary between tasters. This study evaluates Random Forest (RF), Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost) for multiclass wine quality classification using physicochemical characteristics. The UCI Vinho Verde dataset containing 6,497 red and white wine samples was analysed separately by wine type. Preprocessing included outlier handling, feature standardisation and SMOTE, while the models were tuned using five-fold cross-validation. The results showed that RF achieved the strongest overall performance for both red and white wine followed closely by XGBoost, while SVM produced lower results. All three models performed better on the more common middle quality scores and had more difficulty with less frequent classes. Feature analysis showed that alcohol was the most consistent important attribute for red wine, while the important features for white wine varied more between the models. Overall, the findings suggest that machine learning can provide useful support for wine quality assessment based on physicochemical measurements, but should be used as a complement to rather than a replacement for sensory evaluation.

Contents

List of Images 1
List of Abbreviations 2
1 Introduction 3
1.1 Problem Definition 3
1.2 Aim and Objectives 4
2 Background 6
2.1 Preliminaries 6
2.1.1 Classification 6
2.1.2 Random Forest 7
2.1.3 Support Vector Machine 7
2.1.4 Extreme Gradient Boosting 8
3 Related Work 9
4 Method 11
4.1 Research Method 11
4.2 Dataset Description 11
4.3 Data Preprocessing 12
4.3.1 Synthetic Minority Oversampling Technique (SMOTE) 12
4.4 Model Selection 13
4.5 Model Training and Evaluation 13
5 Implementation 15
5.1 Environment and Tools 15
5.2 Dataset 15
5.3 Data Preprocessing 16
5.4 Model Training and Evaluation 17
6 Results 20
6.1 Dataset and Distribution 20
6.2 Model Performance Evaluation 20
6.2.1 Red Wine Model Performance 20
6.2.2 White Wine Model Performance 23
6.3 Physicochemical Feature Importance 26
6.4 Observations and Findings 28
7 Discussion 29
7.1 Model Performance and Feature Interpretation 29
7.2 Previous Studies and Practical Implications 29

7.3 Ethical and Societal Aspects 30
7.4 Validity and Limitations of the Study 31
7.4.1 Internal Validity 31
7.4.2 External Validity 31
8 Conclusion and Future Work 33
8.1 Conclusion 33
8.2 Future work 33
References 35

List of Images

Figure 1. Distribution of wine quality scores in the red and white wine datasets.

Figure 2. Correlation heatmaps for red and white wine features.

Figure 3. Class distribution of red and white wine quality scores before and after SMOTE.

Figure 4. Hyperparameter search space for the Random Forest classifier.

Figure 5. Hyperparameter search space for the Support Vector Machine classifier.

Figure 6. Hyperparameter search space for XGBoost classifier.

Figure 7. Performance comparison of RF, SVM and XGBoost on the red wine dataset.

Figure 8. Random Forest confusion matrix for the red wine dataset.

Figure 9. SVM confusion matrix for the red wine dataset.

Figure 10. XGBoost confusion matrix for the red wine dataset.

Figure 11. Performance comparison of RF, SVM and XGBoost on the white wine dataset.

Figure 12. Random Forest confusion matrix for the white wine dataset.

Figure 13. SVM confusion matrix for the white wine dataset.

Figure 14. XGBoost confusion matrix for the white wine dataset.

Figure 15. Random Forest feature importance for white wine and red wine datasets.

Figure 16. SVM permutation feature importance for white wine and red wine datasets.

Figure 17. XGBoost feature importance for white wine and red wine datasets.

List of Abbreviations

GA- Genetic Algorithm

GB- Gradient Boosting

IQR- Interquartile Range

KNN- K-Nearest Neighbors

LDA- Linear Discriminant Analysis

NNRA- Neural Network Regression Analysis

NPV- Negative Predictive Value

PCA- Principal Component Analysis

PPV- Positive Predictive Value

RBF- Radial Basis Function

RF-Random Forest

SA- Simulated Annealing

SMOTE- Synthetic Minority Oversampling Technique

SVM- Support Vector Machine

XGBoost- Extreme Gradient Boosting

1 Introduction

Wine quality is an important factor because it influences consumer perception, product recognition and market value (Bisson et al., 2002). Wine quality is generally evaluated through physicochemical analysis and sensory evaluation. Physicochemical analysis measures characteristics such as acidity, alcohol content, sugar levels and other chemical properties, while sensory evaluation depends on human experts who assess characteristics such as taste, aroma and overall quality. Although sensory evaluation is important for assessing qualities that cannot be measured directly through chemical analysis, the results may vary between tasters because human perception is subjective (Hodgson, 2008). Physicochemical analysis on the other hand provides objective and reproducible measurements that can be used to describe the chemical composition of wine. However, the relationship between physicochemical properties and sensory quality is complex and is still not fully understood (Cortez et al., 2009). This creates a challenge in determining how measurable chemical characteristics are related to the quality scores assigned through sensory evaluation.

Different types of wine vary in their chemical composition, which makes understanding the role of physicochemical characteristics important for quality assessment. Properties such as acidity, alcohol content, residual sugar, density and sulphates can vary between wines and may contribute differently to the final quality rating. Examining these characteristics can therefore provide useful information about the factors that are associated with wine quality. With the development of machine learning, it has become possible to analyse these measurable characteristics at the same time and identify patterns associated with wine quality (Preedy and Mendez, 2016). This is useful because the relationship between individual physicochemical properties and quality is not always simple and several characteristics may influence the quality score together. Machine learning models can learn from these relationships and use them to classify wines based on their measured properties. In this way machine learning can provide a more consistent and scalable approach to support traditional wine quality assessment (Gupta, 2018).

This study examines three supervised machine learning classification techniques: Random Forest (RF), Support Vector Machine (SVM) and Extreme Gradient Boosting (XGBoost). These algorithms were selected because previous studies have shown that SVM and RF can perform effectively in wine quality prediction, while boosting-based methods such as XGBoost have also demonstrated strong performance (Aich et al., 2019; Kumar et al., 2020; Beri et al., 2024). Neural Networks were not included because previous research using the same Vinho Verde dataset found that SVM performed better than Neural Networks, while Neural Network models generally require more complex model design, computational resources and hyperparameter tuning (Anami et al., 2022). The red and white wine datasets are analysed separately in order to compare the performance of the three models and identify the physicochemical attributes that are most important for wine quality classification.

1.1 Problem Definition

Traditional sensory methods such as difference tests, descriptive analysis and consumer tests are run with trained or consumer panels using blinded, randomized coded samples in controlled settings with standardized procedures and appropriate statistical analysis. In wine evaluation these trained panels may include expert tasters such as sommeliers, whose professional experience is valuable for assessing aroma, flavor and overall quality. These methods can be expensive and their results depend on human perception, which differs from one person to another and tasters can only evaluate a limited number of wines per session to avoid fatigue.

These limitations have encouraged the use of machine learning methods based on measurable physicochemical characteristics. Many algorithms achieve high predictive performance but provide little insight into the decision-making process raising ongoing concerns about model interpretability and transparency (Chen & Guestrin, 2016). Transparency and interpretability are important in wine classification because transparency refers to how clearly the model, the data and the prediction process can be examined and traced, whereas interpretability refers to how well a human can understand why the model made a prediction and how the input features contributed to that outcome. In this study, transparency means clearly showing which physicochemical variables were used and how the classification model produced its output, while interpretability means explaining how specific chemical features contribute to the predicted wine quality class.

Cortez et al. (2009) documents that the quality of wine has traditionally been difficult to predict particularly based on chemical data exclusively. Building on this limitation, the present study does not assume that wine quality can be predicted perfectly from chemical data alone, but instead investigates whether more recent machine learning methods can still provide useful classification results. Since red and white wines differ in physicochemical characteristics, they are analysed separately in order to reduce complexity. The aim is therefore not to replace sensory evaluation completely, but to examine whether machine learning can serve as a complementary tool for faster and more consistent quality assessment. A successful approach could potentially support winemakers in quality control activities by providing a consistent data driven assessment based on physicochemical measurements.

Previous studies have applied different machine learning methods to wine quality prediction, but they often use different preprocessing steps evaluation methods and model combinations. This makes direct comparison difficult. Therefore, this study compares Random Forest, Support Vector Machine and XGBoost using the same experimental procedure for red and white wine while also examining the important physicochemical features.

In order to address these challenges, the following research questions are formulated.

1. “Which physicochemical attributes are ranked as most important in the quality classification of red and white wines?”
2. “How do Random Forest, Support Vector Machine and XGBoost compare in classifying wine quality based on physicochemical properties in terms of predictive performance and feature based interpretability?”

1.2 Aim and Objectives

The aim of this study is to evaluate and compare Random Forest, Support Vector Machine and XGBoost for multiclass wine quality classification using physicochemical characteristics. The study also aims to identify the physicochemical attributes that are most important for wine quality classification and to examine whether their importance differs between red and white wines.

To achieve this aim, the following objectives are formulated:

- To analyse the red and white wine datasets separately.

- To prepare the datasets by checking data quality and handling outliers.
- To implement the RF, SVM and XGBoost within a unified pipeline and tune hyperparameters with five-fold cross validation.
- To evaluate performance using accuracy, weighted precision, weighted recall and weighted F1-score.
- To compare how RF, SVM and XGBoost rank physicochemical attributes when predicting wine quality in the red and white wine datasets.

2 Background

Wine quality has traditionally been assessed through sensory evaluation by trained human experts. These experts are often referred to as sommeliers, oenologists and sensory tasters, and they assess characteristics such as the wines appearance, aroma, flavour and overall balance (Lesschaeve, 2007). While this method is widely used in the wine industry, it also has some limitations. Human perceptions differ from one person to another and wine quality assessments may vary between tasters and across tasting sessions, making it difficult to ensure that the same wine will always receive the same score (Hodgson, 2008). Physicochemical analysis provides another source of information for wine quality assessment by measuring characteristics such as acidity, residual sugar, alcohol content and other chemical properties. Machine learning makes it possible to analyse several of these characteristics together and identify patterns associated with wine quality. This is particularly useful because the relationship between physicochemical characteristics and perceived quality can be complex and may involve several variables acting together (Cortez et al., 2009).

Researchers and producers may now use machine learning algorithms to analyse datasets and identify patterns between physicochemical attributes and wine quality ratings. The UCI Machine Learning Repository has a dataset that is often used for prediction studies including those by Cortez et al. (2009), Aich et al. (2019), and Kumar et al. (2020). The Portuguese Vinho Verde dataset contains separate red and white wine data and has been used to evaluate different machine learning approaches to wine quality prediction. The dataset contains physicochemical measurements together with sensory quality ratings, making it suitable for examining the relationship between measurable wine characteristics and quality scores.

While machine learning has produced promising results in terms of predicting wine quality there are some significant challenges. One of the challenges is that the quality scores are unevenly distributed with some quality classes occurring much more frequently than others. This class imbalance can influence model training because models may become more strongly influenced by the majority classes. Because quality ratings are based on human sensory evaluation, some variation between ratings may also be present. This can lead to noisy data that impacts model's performance and learning ability (Cortez et al., 2009).

2.1 Preliminaries

2.1.1 Classification

Classification is a supervised learning method that sorts data into fixed categories. It uses each item features to assign it to a predefined class (Chakraborty & Dey 2024). The applications for classification are numerous and include healthcare diagnosis, image recognition, identification of spam, sentiment analysis etc. Practitioners define the problem and target labels, preprocess the data, learn meaningful representations, train competing models and measure generalization on unseen examples. Classification can be applied to both structured and unstructured modalities and the output variable is usually referred to as the target, label or category. The field is organized in four core task types: binary classification, where examples belong to one of two mutually exclusive classes; multiclass classification, its generalization to more than two exclusive classes; multilabel classification, where an example can be associated with more

than one class simultaneously and imbalanced classification, where one or more minority classes are rare and require special handling in the form of resampling, class weighted losses and metrics beyond accuracy to ensure reliable performance (Chakraborty & Dey, 2024). In this study wine quality prediction was formulated as a multiclass classification problem. The original quality scores were retained as separate class labels rather than being grouped into binary low and high quality categories.

2.1.2 Random Forest

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve predictive accuracy while reducing overfitting (Breiman, 2001). A decision tree is a model that makes predictions by splitting data into smaller groups according to the values of input features. Rather than relying on a single tree, RF constructs many trees and combines their outputs, which makes the model more stable and reliable. It is widely used in supervised machine learning and can be applied to both classification and regression problems (Cutler et al., 2012). Each tree is trained on a bootstrap sample of the training data. A bootstrap sample is created by randomly selecting observations with replacement, so some observations may appear more than once while others may be left out. In addition, at each split in the tree RF considers only a random subset of the available features. These two sources of randomness help produce diverse trees and reduce the correlation between them, which improves the model's ability to generalize to unseen data (Breiman, 2001; Cutler et al., 2012).

For classification tasks, the final prediction is determined by majority voting among all trees, whereas for regression tasks it is obtained by averaging the outputs of the trees. One of the main strengths of RF is that it can capture nonlinear relationships and interactions between variables without requiring strong assumptions about the data. It is also relatively robust to noise and can provide estimates of feature importance, which are useful for identifying the variables that contribute most to the prediction. In addition, RF can perform well on large and high-dimensional datasets, making it suitable for problems with many input variables (Cutler et al., 2012). Another useful property of Random Forest is the out-of-bag error estimate, which provides an internal measure of model performance during training without requiring a separate validation set. Important hyperparameters include the number of trees, the maximum depth of each tree, and the number of features considered at each split. Overall, RF is widely used because it offers strong predictive performance, robustness and relatively low risk of overfitting (Breiman, 2001; Cutler et al., 2012).

2.1.3 Support Vector Machine

Technological innovation in new data storage technologies and the widespread use of the internet have been the primary drivers of the actual data explosion experienced over the past decades. Extracting meaningful information from this vast quantity of data is no longer possible for humans alone and therefore calls for highly advanced statistical and data mining methods (Taylor-Sakyi, 2016). SVM is suitable for high dimensional data because it can separate classes by finding the best decision boundary between them. Compared with neural network models, SVM has a more clearly defined training process, which can make the model more stable and predictable **(**Mammone Turchi & Cristianini, 2009).

SVM is a supervised learning model which can be utilized for regression and classification. However, it is largely utilized for classification tasks in machine learning (Gandhi, 2018).

SVM learns a decision boundary that separates classes with the widest possible margin. It solves a convex optimization that balances margin size with classification errors using the hinge loss and a regularization parameter C. When classes aren't linearly separable, SVM uses a kernel to compute similarities that act like mapping data into a higher dimensional space without doing it explicitly. The SVM method seeks to draw the optimal decision boundary that divides n-dimensional space into classes, allowing to efficiently assign data that are new to the appropriate groups. A hyperplane may divide the data into two groups with a suitable nonlinear mapping to a higher dimensional space. This hyperplane is determined by the support vectors critical elements of the training set that lie closest to the decision boundary and they define the position and orientation of the hyperplane. This makes SVM efficient since only a subset of the training data is used (Han, Kamber & Pei, 2012).

SVM introduces the kernel technique. It maps the input points into a high dimensional feature space using a nonlinear transformation and then computes the optimal separating hyperplane. However, two main challenges remain in the SVM process. These are how can the binary classification problem be effectively generalized to multiclass scenarios and how can the optimal hyperplane be made less sensitive to individual sample points and noise. Binary classification is straightforward with SVM but adapting them to multiclass tasks requires methods like one vs one or one vs all. It is also important to develop techniques for optimization that decreases the influence mislabelled data points (Mammone Turchi & Cristianini, 2009).

2.1.4 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a highly optimized gradient boosting framework that provides speed and scalability for large datasets. XGBoost works by building many small decision trees one after another. It starts with a basic prediction, then each new tree tries to fix the mistakes the model just made by learning the direction of those errors. Each tree's impact is kept small using a learning rate and the algorithm can sample rows and columns to avoid overfitting. It also uses built-in penalties to keep the trees from getting too complex. Training is fast because it searches for good splits in parallel and uses caching to handle large data efficiently. XGBoost has been widely used in statistics data mining and machine learning communities since its initial release in 2015 due to its high generalization ability, scalability and fast computing speed**.** It's a composite algorithm that constructs elementary functions with weights to form a highly effective fit for the data. By combining the gradient boosting algorithms with strengthens of decision tree, XGBoost improves prediction accuracy by training and combining outcomes in an ensemble of decision trees (Chen and Guestrin 2016).

The XGBoost algorithm utilizes parallel computation to accelerate the prediction model training, utilizes the caching technology to reduce memory consumption, and provides tuning parameters for users to gain more control over the complexity and model performance. XGBoost iteratively trains a series of weak learners using decision trees and ensemble them to create a strong learner to output accurate predictions for patterns within data that is complex. It develops a group of trees one after the other, where every new tree attempts to minimize to residual errors of the current trees (Chen and Guestrin 2016).

3 Related Work

Gupta (2018) carried out an extensive study in wine quality prediction using machine learning techniques with regard to using Linear Regression, Neural Networks and SVM based on both red and white wine datasets. The study used a dataset of physicochemical measurements for each sample of wine used to predict sensory quality scores. The approach was divided into two main sections. The first section is analysing the way independent variables such as alcohol, pH, sulphates and volatile acidity influence the quality score and the second section is using those variables to develop predictive models. Instead of applying the whole dataset to train models, the study applied only the most important features by identifying those features with most statistical significance in decision making regarding wine quality.

The results showed that some qualities particularly volatile acidity, alcohol concentration and sulphates had a strong correlation with wine quality than others. SVM and Neural Network models achieved better performance with less training and validation errors as a result of the feature set reduction. The best of the methods evaluated according to predictive accuracy was SVM, which validates its suitability for regression problems involving intricate, nonlinear relations. Gupta concluded that adding all the available features does not always enhance model performance, and that careful feature selection can lead to more efficient and interpretable models. The importance of both model optimization and statistical analysis for machine learning model building for wine quality prediction is established in this work. This work is good stepping stone for further research into more advanced algorithms or larger and more diverse datasets (Gupta 2018).

Beri et al. (2024) provides a comparative study of machine learning algorithms K-Nearest Neighbors (KNN), Gradient Boosting (GB), and Extreme Gradient Boosting (XGBoost) for physicochemical depend on wine quality prediction. The compare and contrast were made on an openly accessible Kaggle dataset including red wines and white wines with attributes like acidity, pH, sugar content, sulphates and alcohol content. The study also emphasizes how crucial data preprocessing is including handling of missing values, feature scaling, and dimensionality reduction by Principal Component Analysis (PCA), which increased model performance and efficiency. The metrics applied during evaluation include accuracy, recall, precision and F1-score.

Among the models that were compared, XGBoost was the best performing model with an average precision, recall and F1-score of 0.96 and thus was very effective in wine quality classification. The Gradient Boosting came a close second but KNN was low on performance. The study concludes that boosting based ensemble methods like XGBoost are important algorithms for wine quality prediction. It also suggests future development through the inclusion of sensory or geographic information and more advanced hyperparameter optimization techniques (Beri et al., 2024).

Anami et al. (2022) discussed the use of machine learning models for automatic prediction of wine quality, in particular how to determine whether wine is good or bad. Using the Portuguese Vinho Verde dataset that includes red and white wine samples, they implemented three different algorithms. These are Logistic Regression, Support Vector Machine (SVM) and Neural Networks. The study aimed to decrease the input features dimension using principal component Analysis (PCA) and Neural Network Regression Analysis (NNRA) and selected important characteristics such as residual sugar, volatile acidity, citric acid, sulphur dioxide content and alcohol content. Their results showed that the best performance was achieved by SVM with a mean error of 0.003 and a predicted quality scored of almost 8, outperforming Neural Networks and Logistic Regression. Alcohol and sulphates were more significantly associated with perceived quality than with attributes like density or pH. This verifies the hypothesis that predictive power will be more elevated if only relevant attributes are chosen. The

experiment also illustrated the importance of statistical validation using regression coefficients and p-values to determine feature significance. It found that machine learning in this case SVM, is a reliable and scalable solution to assist winemakers in quality control and assurance processes (Anami et al., 2022).

Aich et al. (2019) examined the effectiveness of feature selection in order to improve wine quality prediction with supervised machine learning techniques using red wine and white wine from the UCI Repository. The authors applied two optimization algorithms which are Genetic Algorithm (GA) and Simulated Annealing (SA), in order to identify the most important chemical descriptors for classification. The study experimented with different classifiers like Support Vector Machines (SVM), Random Forest, Naïve Bayes and Linear Discriminant Analysis (LDA) a classification method which projects data in such a way as to most effectively separate different quality categories reducing variation within a group. Performance indicators were used to compare these models such as accuracy, sensitivity, specificity and Positive Predictive Value (PPV). The Negative Predictive Value (NPV), which determines the proportion of accurate negative forecasts to all negative predictions and the ratio of correct positive predictions to all positive predictions produced using the model are both measured. The results they obtained demonstrated that SVM outperformed other models, especially when combined with SA based feature selection. The models result shows that up to 98.81% accuracy in white wine and 95.23% accuracy in red wine, outperforming other classifiers. The study concludes that the use of SA to select important features significantly enhances model accuracy and saves computational resources. This process provides the wine sector with practical advantage in terms of better-quality analysis, speed and cost savings (Aich et al. 2019).

Kumar et al. (2020) performed an experimental study to predict red wine quality using various well known machine learning models. Since the intake of red wine has been rising, the authors suggested there is a need for quality control not merely for taste but also for consumers health. The study was based on a dataset from the UCI Machine Learning Repository with 1,599 red wine samples that described by 12 chemical attributes. To test the predictive potential of machine learning, they tried out three most widely used classification model which are Naïve Bayes, Support Vector Machine (SVM) and Random Forest. The experiment was carried out in RStudio, where models were trained on 70% of the dataset and the remaining 30% was kept for testing. To compare model performances, they made use of shared performance measures such as accuracy, recall, precision, specificity, F1-score and misclassification error. Among all the models, SVM gave the maximum accuracy of 67.25% on training and 68.64% on test sets. Random Forest performed very well, while Naïve Bayes performed less accurately than they did. The study not only confirmed that machine learning would prove to be beneficial in simplifying and automating wine quality testing, but also postulated that aggregating many algorithms or adjusting their parameters can be utilized to make it work even better in future applications (Kumar et al. 2020).

Previous studies have shown that machine learning methods such as SVM, RF and XGBoost can be useful for predicting wine quality. However, the studies often differ in data preparation, prediction approach and evaluation procedure. This makes it difficult to compare their results directly. For this reason, the present study compares RF, SVM and XGBoost using the same preprocessing, training and evaluation procedure for red and white wines separately.

4 Method

This chapter describes the methodological procedure used in the study. The analysis is based on the UCI Vinho Verde red and white wine datasets, which are analysed separately. The chapter presents the research method, dataset, preprocessing steps, model selection, training procedure and evaluation approach used to compare RF, SVM and XGBoost. Model performance is evaluated using accuracy, weighted precision, weighted recall, weighted F1-score and confusion matrices while feature importance analysis is used to examine the physicochemical attributes that contribute most strongly to the predictions.

4.1 Research Method

This study follows a quantitative comparative experimental approach. A quantitative approach is appropriate because the study is based on numerical data and measurable outcomes. Quantitative research focuses on the systematic collection and analysis of numerical data in order to examine relationships, compare outcomes and test research questions (Creswell & Creswell, 2018). The purpose is to systematically compare machine learning algorithms and examine differences in their predictive performance and feature importance. The study is also experimental because the selected algorithms are evaluated under consistent conditions using a defined and repeatable procedure. Wohlin et al. (2012) describe experimental research as a systematic process involving the planning, execution, analysis and interpretation of an experiment in order to evaluate and compare different alternatives.

The research method is closely connected to the research questions. The first research question examines which physicochemical attributes are most important for the quality classification of red and white wines whereas the second examines how RF, SVM and XGBoost compare in terms of predictive performance and feature-based interpretability. A quantitative comparative approach is appropriate for answering these questions because it allows the selected models to be evaluated systematically using consistent procedures and measurable numerical results. Analysing the red and white wine datasets separately also makes it possible to investigate whether model performance and feature importance differ between the two wine types. The experimental procedure is designed to support systematic and reproducible comparisons between the selected machine learning models. Because the study focuses on measurable model performance and feature importance rather than opinions, experiences or other qualitative phenomena, a quantitative approach provides an appropriate basis for answering the research questions using numerical evidence.

4.2 Dataset Description

The wine quality dataset used in this study was obtained from the UCI Machine Learning Repository. It represents Portuguese Vinho Verde wines and is provided as two separate datasets, one containing red wine samples and the other containing white wine samples (Cortez et al., 2009). The datasets contain a total of 6,497 observations, including 1,599 red wine samples and 4,898 white wine samples (Sirivanth et al., 2021). The 11 physicochemical attributes recorded for each observation are fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates and alcohol. These attributes are used as the predictor variables, while the wine quality score is used as the target variable for classification. The quality score represents the sensory evaluation assigned to each wine sample.

The distribution of quality scores however is not balanced. Most observations are concentrated around the middle quality levels, while relatively few samples occur at the lower and higher quality scores. This imbalance is important for the modelling process because classes with fewer observations may be more difficult for the classification algorithms to learn. The numerical structure of the datasets makes them suitable for supervised machine learning because each wine sample contains a defined set of physicochemical measurements together with an assigned quality score. Retaining the original quality scores also allows the problem to be treated as a multiclass classification task rather than reducing the ratings into broader quality groups. This provides a more detailed basis for examining how the selected machine learning models distinguish between different wine quality levels. Since the red and white wine datasets differ in both sample size and physicochemical characteristics, they are analysed separately throughout the study.

4.3 Data Preprocessing

Data preprocessing was carried out before training the machine learning models. The red and white wine datasets were handled separately during this stage. Both datasets were checked for missing values and duplicate observations. Upon inspection, no missing values were detected in either dataset. As a result, no additional procedure was required to replace or estimate missing values. Duplicate removal was carried out so that repeated observations did not unnecessarily influence the learning process. These initial checks helped ensure that the data used for model development were consistent before further preprocessing was applied. Each dataset was then divided into training and test sets before outlier removal. This was done so that the outlier related preprocessing decisions were based only on the training data and did not influence the final test results. In this study, outliers in alcohol, sulphates and volatile acidity were identified in the training data using the Interquartile Range (IQR) method. Alcohol, sulphates and volatile acidity were selected for outlier handling because they were important in the exploratory analysis and were also identified as relevant features in previous wine quality studies. Values below the first quartile minus 1.5 times the IQR or above the third quartile plus 1.5 times the IQR were treated as outliers. Outlier handling was included because extreme values can have a disproportionate effect on model training, particularly when variables differ in their distribution or scale. These extreme values were removed from the training set to reduce their possible effect on model training. The test data were not subjected to outlier removal and were reserved for the final model evaluation.

4.3.1 Synthetic Minority Oversampling Technique (SMOTE)

The Synthetic Minority Oversampling Technique (SMOTE) was used to address imbalance among the wine quality score classes**.** Its main purpose is to generate synthetic observations for underrepresented classes by interpolating between existing minority-class samples thereby improving the model's ability to learn patterns from classes with fewer observations. In both red and white wine datasets, middle quality scores occurred much more frequently than the extreme quality scores. Since the two datasets were analysed separately, the difference in the total number of red and white wine samples was not treated as a class imbalance. SMOTE was applied only to the training data after the train-test split in order to reduce class imbalance without introducing synthetic observations into the test set. During cross-validation SMOTE was included within the training pipeline so that synthetic samples were generated only from the training folds. SMOTE was applied with k_neighbors = 2 and a random

state of 42. This preprocessing step provided a more balanced training distribution and reduced the risk that the models would be biased toward the most frequent quality classes (Chawla et al., 2002).

4.4 Model Selection

The selection of machine learning algorithms was based on the objective of classifying wine quality from physicochemical measurements and on findings from previous studies. Although wine quality is initially quantified by means of a numerical range, the scores were retained as separate classes in this study allowing the problem to be treated as a multiclass classification task (Cortez et al., 2009). Three machine learning algorithms, RF, SVM, and XGBoost were selected for comparison. Previous studies have shown that these approaches can perform effectively in wine quality prediction, which provided the basis for their inclusion in this study.

Random Forest was selected because it is suitable for structured data and can model nonlinear relationships between physicochemical attributes and wine quality. It also provides feature-importance estimates which are useful for examining which attributes contribute most to the predictions (Breiman, 2001). SVM was selected because it can model nonlinear decision boundaries through the use of kernel functions and has previously shown good performance in wine quality prediction (Cortes and Vapnik, 1995**).** The third model is XGBoost which is an ensemble gradient boosting decision tree algorithm that was included as a boosting based ensemble method. It builds decision trees sequentially and uses regularization to control model complexity (Chen and Guestrin 2016). Using these three algorithms allows different machine learning approaches to be compared under the same preprocessing and evaluation procedure.

4.5 Model Training and Evaluation

Each dataset was divided into training and test sets using an 80%, 20% ratio. Stratified sampling was used to preserve the distribution of the quality classes and a random state of 42 was used to support reproducibility. The training data were used for model development and hyperparameter tuning, while the test data were reserved for the final evaluation. Standardization was included within the modelling procedure to transform the physicochemical variables to a common scale with a mean of zero and a standard deviation of one. StandardScaler and SMOTE were included within the model pipelines so that, during cross-validation, these preprocessing steps were applied only to the training folds. For the final evaluation, the scaler was fitted using the training data and the same transformation was applied to the test data to prevent information leakage (Geron, 2019). The RF, SVM and XGBoost models were trained separately for the red and white wine datasets using the same modelling procedure. RF, SVM and XGBoost models were tuned using GridSearchCV with five-fold cross validation to identify the best combination of hyperparameters. During five-fold cross-validation, the training data were divided into five folds with four folds used for model training and the remaining fold used for validation. This process was repeated until each fold had been used for validation once. WeightedF1-score was used to select the best hyperparameter combination because the wine quality classes were not equally distributed.

The hyperparameter search covered model specific parameters. For RF these included the number of trees, maximum tree depth, minimum samples required for splitting and leaf nodes and the number of features considered at each split. For SVM the C and gamma parameters were tuned using the RBF kernel. For XGBoost, the parameters included the number of trees, maximum tree depth, learning rate, subsample and the proportion of features used when constructing each tree. The exact parameter

values used in the searches are presented in the implementation section. After hyperparameter tuning, the best performing model configuration for each algorithm was evaluated on its respective test set. Model performance was evaluated using accuracy, weighted precision, weighted recall and weighted F1-score. Since this study involves multiple wine quality classes, precision, recall and F1-score were calculated for each class and then combined using a weighted average. Classes with more observations therefore had a greater contribution to the final weighted scores. Weighted F1-score was given particular attention because the distribution of the wine quality classes was imbalanced. Because the prediction task is multiclass, evaluation was carried out across the individual wine quality classes rather than treating the problem as a single positive versus negative classification task.

5 Implementation

5.1 Environment and Tools

All model development and evaluation were performed within a Google Colab notebook. Manipulation of data and numerical computation were performed through the pandas and NumPy libraries, while exploratory plots were created using matplotlib and seaborn. The foundation for the machine learning pipeline was provided by scikit learn library, which consisted of the StandardScaler for feature standardisation to unit variance and zero mean and the evaluation tool used along the way. SMOTE from the imbalanced learn library was used to address imbalance among the wine quality score classes. Random Forest and Support Vector Machine models were implemented and trained using scikit learn while XGBoost was implemented using XGBoost package.

5.2 Dataset

The dataset used in this study was obtained from the UCI Machine Learning Repository and consists of two separate files representing red and white Vinho Verde wines (Cortez et al., 2009). The datasets contain 6,497 samples in total including 1,599 red wine samples and 4,898 white wine samples. Each sample contains 11 physicochemical attributes and one quality score. The two datasets were cleaned and analysed separately. To avoid data redundancy and maintain stable model training, duplicate entries were removed and no missing values were discovered. Preliminary data analysis showed a significant imbalance in the distribution of wine quality scores. As illustrated in Figure 1 the majority of wines were scored between 5 and 7, with relatively few samples rated below 5 or above 7.

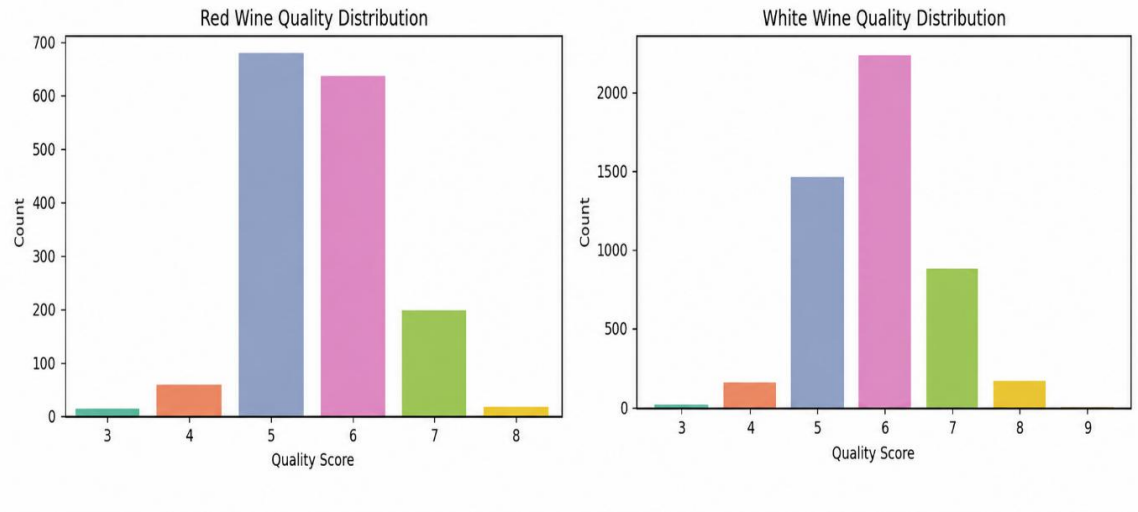


Figure 1. Distribution of wine quality scores in the red and white wine datasets.

In order to better understand the relationships between the characteristics, correlation matrices were created separately for the red and white wine datasets. The relationships between the physicochemical variables and wine quality are shown as heatmaps in Figure 2.

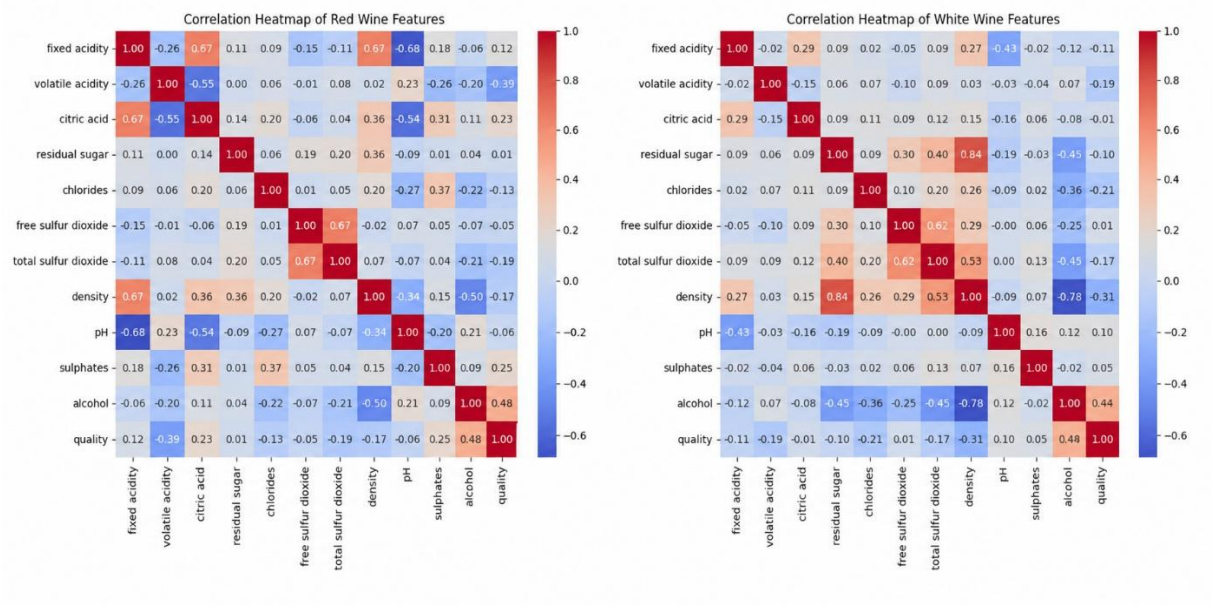


Figure 2. Correlation heatmaps for red and white wine features.

5.3 Data Preprocessing

The red and white wine datasets were pre-processed separately to preserve the differences between the two wine types. The original quality scores were retained as multiple classes, allowing the models to classify wines across the available quality levels. This approach also made it possible to compare model performance and feature importance separately for red and white wines. After the train-test split, outliers among the selected training features, particularly alcohol, sulphates and volatile acidity were detected and removed using the IQR method. Values outside 1.5 times the IQR below the first quartile or above the third quartile were treated as outliers and removed. The purpose of this step was to reduce the influence of extreme observations that could affect the model training process and predictive performance.

Each dataset was divided into training and test sets using an 80%, 20% ratio. Stratified sampling was used to preserve the original distribution of the quality classes, while a random state of 42 was used to support reproducibility. Due to the previously identified class imbalance, SMOTE was applied only to the training data to generate synthetic observations for the minority classes (Chawla et al., 2002). This reduced the risk of the machine learning models becoming biased toward the majority classes, while the test set remained unchanged for evaluation. Figure 3 presents the class distributions before and after the application of SMOTE for the red and white wine training datasets. The resampled data shown in Figure 3 were created only to illustrate the effect of SMOTE, while SMOTE was applied within each model pipeline during model training.

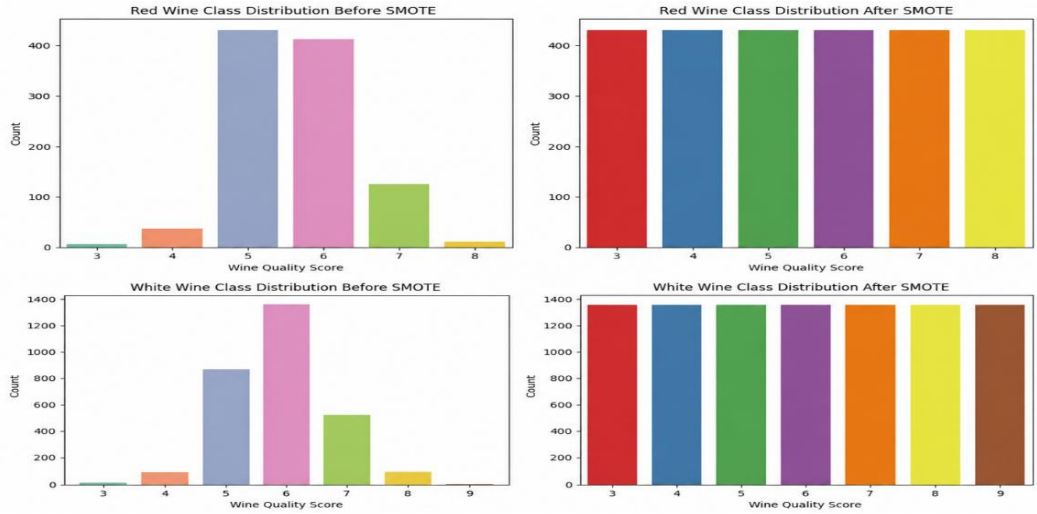


Figure 3. Class distribution of red and white wine quality scores before and after SMOTE.

All physicochemical features were standardised using StandardScaler, which transformed the variables to a common scale with a mean of zero and a standard deviation of one. StandardScaler was included within the model pipelines so that, during cross-validation, it was fitted only on the training folds and applied to the corresponding validation folds. Standardisation was particularly important for Support Vector Machine because its performance is sensitive to differences in feature magnitude (Han et al., 2012).

5.4 Model Training and Evaluation

Once duplicate removal train-test splitting and outlier handling on the training data had been completed, the next step was to train and evaluate the predictive models for multiclass wine quality classification. Three classifiers were used RF, SVM and XGBoost. The models were trained separately on the red and white wine datasets using the same modelling procedure. Standardization and SMOTE were included within each model pipeline, meaning that these preprocessing steps were applied only to the training folds during cross-validation. The models were evaluated using accuracy, weighted precision, weighted recall and weighted F1-score, with particular emphasis on weighted F1-score because of the unequal distribution of wine quality classes.

The Random Forest model was first trained using GridSearchCV with five-fold cross-validation to identify the hyperparameter combination that produced the highest weighted F1-score. Hyperparameter tuning was performed separately for the red and white wine training datasets. For RF the pipeline consisted of StandardScaler followed by SMOTE and the Random Forest classifier. A random state of 42 was used to support reproducibility. The hyperparameter search space included different values for the number of trees, maximum tree depth, minimum samples required for splitting and leaf nodes and the number of features considered at each split as presented in Figure 4. The final RF models were then evaluated on their respective test sets using the previously defined performance metrics. A confusion matrix was also generated for each optimized RF model to examine classification performance across the individual quality-score classes. Feature importance scores were obtained from the built-in feature importance measure of the optimized RF models. The features were ranked separately for the red and white wine datasets to show which physicochemical attributes the models relied on most when making predictions.

Hyperparameter	Values tested
n_estimators	100, 200, 300
max_depth	None, 10, 20
min_samples_split	2, 5
min_samples_leaf	1, 2
max_features	sqrt, log2

Figure 4. Hyperparameter search space for the Random Forest classifier.

The Support Vector Machine model was trained using the Radial Basis Function (RBF) kernel. The model was trained separately for the red and white wine datasets. As with Random Forest, StandardScaler and SMOTE were included in the pipeline before the SVM classifier. GridSearchCV with five-fold corss-validation was used to find the best values for the parameters C and gamma. Weighted F1-score was used to select the best parameter combination because the quality classes were not equally distributed. The tested parameter values are shown in Figure 5. After tuning the best SVM model for each wine dataset was evaluated using accuracy, weighted precision, weighted recall and weighted F1-score. Confusion matrices and classification reports were also produced to show how well the models classified the different quality scores.

Hyperparameter	Values tested
C	0.1, 1, 10
gamma	scale, 0.01, 0.001
kernel	rbf

Figure 5. Hyperparameter search space for the Support Vector Machine classifier.

Unlike Random Forest, SVM does not provide a built-in feature importance score. Therefore, permutation feature importance was used to examine which physicochemical features the SVM model relied on most when making predictions. Each feature was randomly shuffled 30 times and the change in weighted F1-score was measured. A larger decrease in the weighted F1-score indicates that the model relied more strongly on that feature. The feature importance values were calculated separately for the red and white wine datasets so that the important features for each wine type could later be compared.

The third model was Extreme Gradient Boosting (XGBoost). The model was trained separately for the red and white wine datasets using the same general procedure as RF and SVM. Before training, the original quality scores were converted into numerical class labels using LabelEncoder because XGBoost requires the class to start from zero. The pipeline consisted of StandardScaler followed by SMOTE and the XGBoost classifier. GridSearchCV with five-fold cross-validation was used to find the best combination of hyperparameters. The parameters tested included the number of trees, maximum tree depth, learning rate, subsample and the proportion of features used when building each tree, as shown in Figure 6. Weighted F1-score was used to select the best parameter combination.

Hyperparameter	Values tested
n_estimators	100, 200
max_depth	3, 6, 10
learning_rate	0.01, 0.1
subsample	0.8, 1.0
colsample_bytree	0.8, 1.0

Figure 6. Hyperparameter search space for XGBoost classifier.

After tuning, the best XGBoost model for each wine dataset was used to make predictions on the test set. The predicted labels were then converted back to the original wine quality scores. Model performance was evaluated using accuracy, weighted precision, weighted recall and weighted F1-score. Confusion matrices and classification reports were also produced to examine how the models classified the different quality scores. Feature importance scores were obtained from the built-in feature importance measure of the optimized XGBoost models. The features were ranked according to their importance separately for the red and white wine datasets. This made it possible to examine which physicochemical features the XGBoost models relied on most when predicting wine quality and to later compare the feature rankings between the two wine types.

6 Results

6.1 Dataset and Distribution

The original dataset used in this study contained a total of 6,497 wine samples including 1,599 red wine samples and 4,898 white wine samples. Each sample was described by 11 physicochemical attributes together with a quality score. Since red and white wines differ in their chemical characteristics the two datasets were analysed separately throughout the experiment. This made it possible to examine the distribution of quality scores and the relationships between the variables for each wine type without mixing the two groups together.

The distribution of the quality scores was not balanced in either dataset. Most of the samples were concentrated around the middle quality scores especially between scores 5 and 7 while the lower and higher scores appeared much less frequently. This imbalance was present in both red and white wine although the white wine dataset contained considerably more samples overall. The uneven class distribution was important because models trained on this type of data can become more influenced by the classes that appear most often. For this reason, the original class distributions were examined before model training and Figure 1 shows how the quality scores were distributed for the red and white wine datasets.

The exploratory analysis also showed several relationships between the physicochemical attributes and wine quality as shown in Figure 2. Alcohol had the strongest positive relationship with quality in both datasets, meaning that higher alcohol values were generally associated with higher quality scores. Volatile acidity and density showed negative relationships with quality, although the strength of these relationships differed between the two wine types. Volatile acidity showed a stronger negative relationship in the red wine dataset, while density showed a stronger negative relationship in the white wine dataset. These differences suggest that some physicochemical properties may have a different level of importance depending on the type of wine.

Overall, the initial analysis showed that the red and white wine datasets shared some similar patterns but also had noticeable differences in both sample distribution and feature relationships. These observations provided a useful starting point for the later comparison of model performance and feature importance across the two wine types.

6.2 Model Performance Evaluation

6.2.1 Red Wine Model Performance

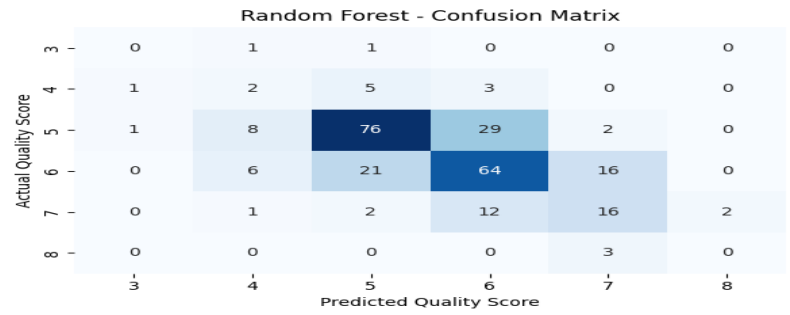
The performance of RF, SVM and XGBoost on the red wine test set was compared using accuracy, weighted precision, weighted recall and weighted F1-score. RF achieved the highest overall performance with an accuracy of 0.5809, weighted precision of 0.5988, weighted recall of 0.5809 and weighted F1-score of 0.5886. XGBoost produced similar results, with an accuracy of 0.5735, weighted precision of 0.5796, weighted recall of 0.5735 and weighted F1-score of 0.5751. SVM gave lower results, reaching an accuracy of 0.4706, weighted precision of 0.5439, weighted recall of 0.4706 and weighted F1-score of 0.4950. Figure 7 shows the performance comparison of RF, SVM and XGBoost on the red wine dataset.

	Model	Accuracy	Precision	Recall	F1 Score
0	Random Forest	0.5809	0.5988	0.5809	0.5886
1	SVM	0.4706	0.5439	0.4706	0.4950
2	XGBoost	0.5735	0.5796	0.5735	0.5751

Figure 7. Performance comparison of RF, SVM and XGBoost on the red wine dataset.

The results show that RF and XGBoost were quite close in performance, with only a small difference between them across the four-evaluation metrics. RF still achieved slightly better scores overall. SVM, on the other hand showed a clearer drop in performance compared with RF and XGBoost. Based on the weighted F1-score, which was used as the main measure for comparing the models, RF gave the strongest result for the red wine dataset, followed by XGBoost and then SVM.

The RF confusion matrix in Figure 8 shows that the model performed best on the quality scores with the largest number of observations. It correctly classified 76 wines with quality score 5 and 64 wines with quality score 6. For quality score 7, 16 wines were correctly classified. The model had more difficulty with the less frequent quality scores. Only 2 wines with quality score 4 were correctly classified, while none of the wines with quality scores 3 and 8 were classified correctly. Most of the errors occurred between neighbouring quality scores, especially between scores 5 and 6 and between scores 6 and 7. For example 29 wines with an actual quality score of 5 were predicted as score 6, while 21 wines with an actual score of 6 were predicted as score 5. A further 16 wines with quality score 6 were predicted as score 7. Overall, the confusion matrix shows that RF was more successful in classifying the middle quality scores than the quality scores at the lower and upper ends of the scale.



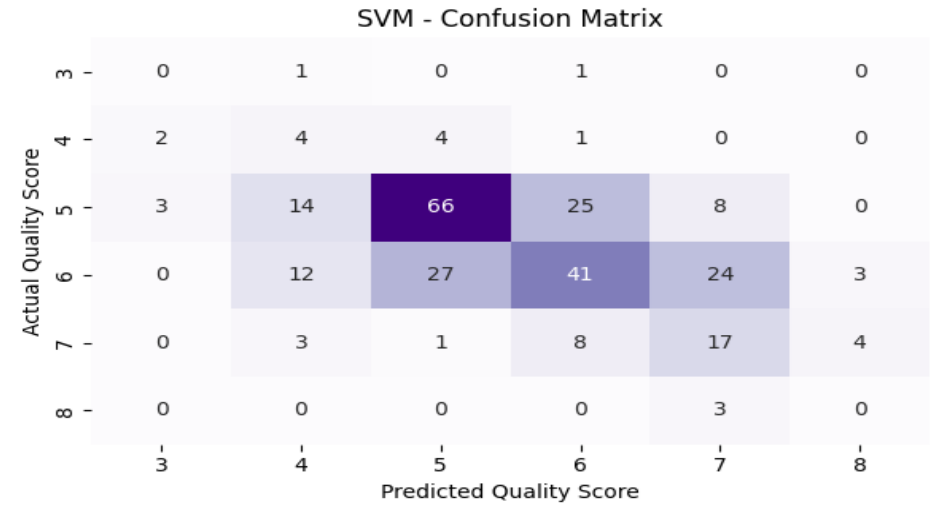
Classification Report:

	precision	recall	f1-score	support
3	0.00	0.00	0.00	2
4	0.11	0.18	0.14	11
5	0.72	0.66	0.69	116
6	0.59	0.60	0.60	107
7	0.43	0.48	0.46	33
8	0.00	0.00	0.00	3
accuracy			0.58	272
macro avg	0.31	0.32	0.31	272
weighted avg	0.60	0.58	0.59	272

Figure 8. Random Forest confusion matrix for the red wine dataset.

SVM correctly classified 66 wines with quality score 5, 41 wines with quality score 6 and 17 wines with quality score 7. The model also correctly classified 4 wines with quality score 4, while none of the wines with quality scores 3 and 8 were classified correctly. Compared with RF the model produced fewer correct predictions for the two most common quality scores, particularly quality score 6. This is consistent with the lower overall accuracy and weighted F1-score obtained by SVM. Overall, the

confusion matrix shows that SVM had more difficulty separating the middle quality classes than RF even though it correctly classified slightly more wines with quality score 7.

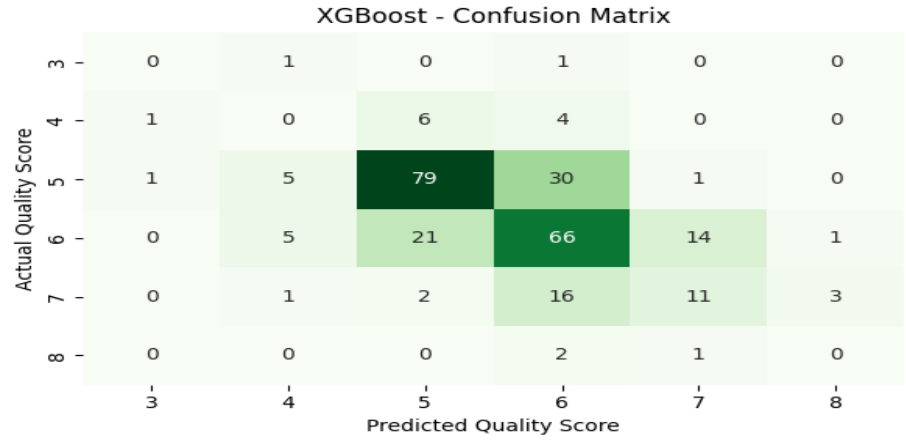


Classification Report:

	precision	recall	f1-score	support
3	0.00	0.00	0.00	2
4	0.12	0.36	0.18	11
5	0.67	0.57	0.62	116
6	0.54	0.38	0.45	107
7	0.33	0.52	0.40	33
8	0.00	0.00	0.00	3
accuracy			0.47	272
macro avg	0.28	0.31	0.27	272
weighted avg	0.54	0.47	0.50	272

Figure 9. SVM confusion matrix for the red wine dataset

Figure 10 presents the confusion matrix for XGBoost. The model correctly classified 79 wines with quality score 5 and 66 wines with quality score 6, slightly more than RF for these two classes. However, for quality score 7, XGBoost correctly classified 11 wines, which was lower than both RF and SVM. The model did not correctly classify any wines with quality scores 3 and 4 and none of the quality score 8 wines were classified correctly. Similar to the other models, most of the incorrect predictions occurred between nearby quality scores. Overall XGBoost performed well on the two most common quality scores but had more difficulty with less frequent classes.



Classification Report:

	precision	recall	f1-score	support
3	0.00	0.00	0.00	2
4	0.00	0.00	0.00	11
5	0.73	0.68	0.71	116
6	0.55	0.62	0.58	107
7	0.41	0.33	0.37	33
8	0.00	0.00	0.00	3
accuracy			0.57	272
macro avg	0.28	0.27	0.28	272
weighted avg	0.58	0.57	0.58	272

Figure 10. XGBoost confusion matrix for the red wine dataset.

6.2.2 White Wine Model Performance

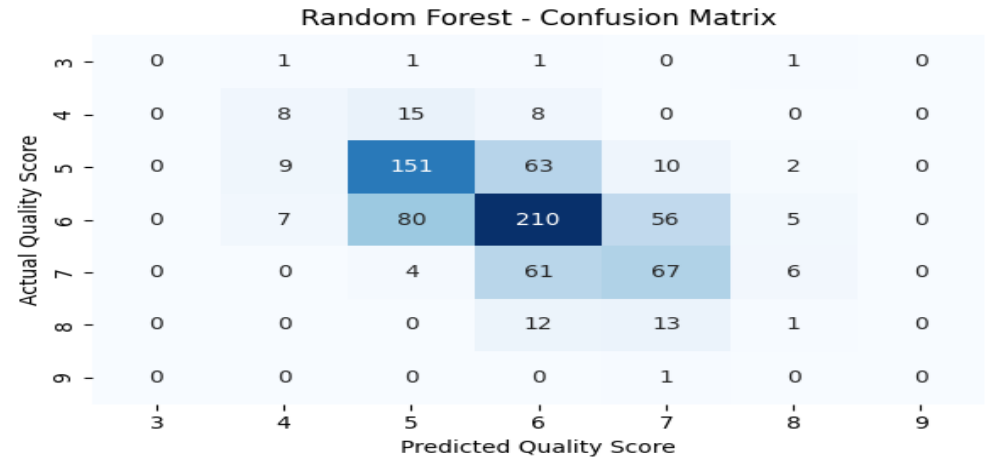
The performance of RF, SVM and XGBoost on the white wine test set was evaluated using accuracy, weighted precision, weighted recall and weighted F1-score. RF achieved the highest overall performance with an accuracy of 0.5511, weighted precision of 0.5393, weighted recall of 0.5511 and weighted F1-score of 0.5447. XGBoost followed closely, reaching an accuracy of 0.5422, weighted precision of 0.5233, weighted recall of 0.5422 and weighted F1-score of 0.5293. SVM showed lower performance than the two tree-based models with an accuracy of 0.4527, weighted precision of 0.4803, weighted recall of 0.4527 and weighted F1-score of 0.4579.

	Model	Accuracy	Precision	Recall	F1 Score
0	Random Forest	0.5511	0.5393	0.5511	0.5447
1	SVM	0.4527	0.4803	0.4527	0.4579
2	XGBoost	0.5422	0.5233	0.5422	0.5293

Figure 11. Performance comparison of RF, SVM and XGBoost on the white wine dataset

The results show a similar pattern to the red wine dataset, with RF and XGBoost producing relatively close results while SVM achieved lower scores. RF had the highest value across all four-evaluation metrics although the difference between RF and XGBoost was fairly small. Based on the weighted F1-score RF gave the strongest overall result for the white wine dataset with 0.5447 followed by XGBoost with 0.5293 and SVM with 0.4579.

The RF confusion matrix in Figure 12 shows that the model classified the middle quality scores more successfully than the less frequent classes. It correctly classified 151 wines with quality score 5, 210 wines with quality score 6 and 67 wines with quality score 7. The model also correctly classified 8 wines with quality score 4 and 1 wine with quality score 8 while none of the wines with quality scores 3 and 9 were classified correctly. Most of the errors occurred between neighbouring quality scores. For example, a number of wines with quality score 6 were predicted as 5 or 7. Overall RF performed better on the more common quality scores while the classes at the lower and upper ends of the scale were more difficult to identify correctly.

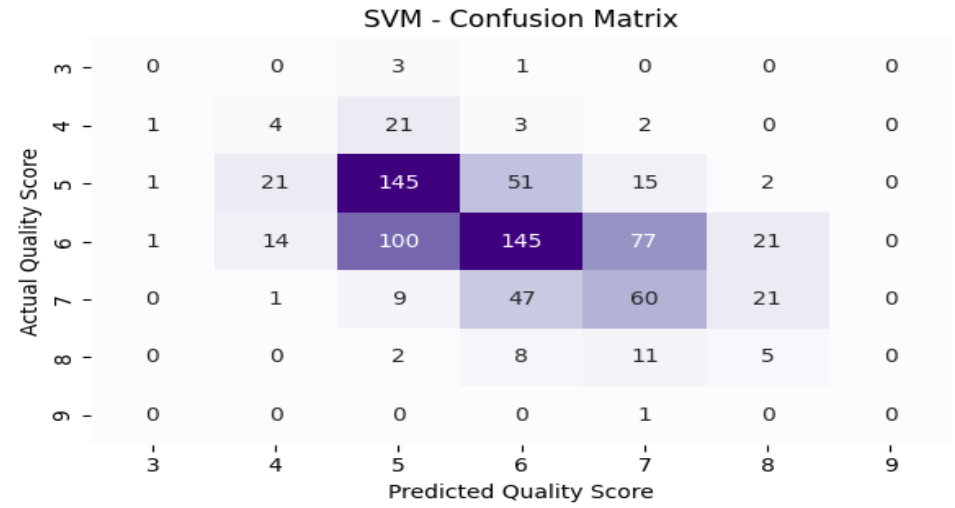


Classification Report:

	precision	recall	f1-score	support
3	0.00	0.00	0.00	4
4	0.32	0.26	0.29	31
5	0.60	0.64	0.62	235
6	0.59	0.59	0.59	358
7	0.46	0.49	0.47	138
8	0.07	0.04	0.05	26
9	0.00	0.00	0.00	1
accuracy			0.55	793
macro avg	0.29	0.29	0.29	793
weighted avg	0.54	0.55	0.54	793

Figure 12. Random Forest confusion matrix for the white wine dataset.

The SVM confusion matrix in Figure 13 shows that the model performed best on the middle quality scores. It correctly classified 145 wines with quality score 5, 145 wines with quality score 6 and 60 wines with quality score 7. For quality score 4, 4 wines were classified correctly while 5 wines with quality score 8 were correctly identified. None of the wines with quality scores 3 and 9 were classified correctly. Compared with its lower overall accuracy and weighted F1-score. Overall, the model had more difficulty distinguishing between the neighbouring middle quality classes.

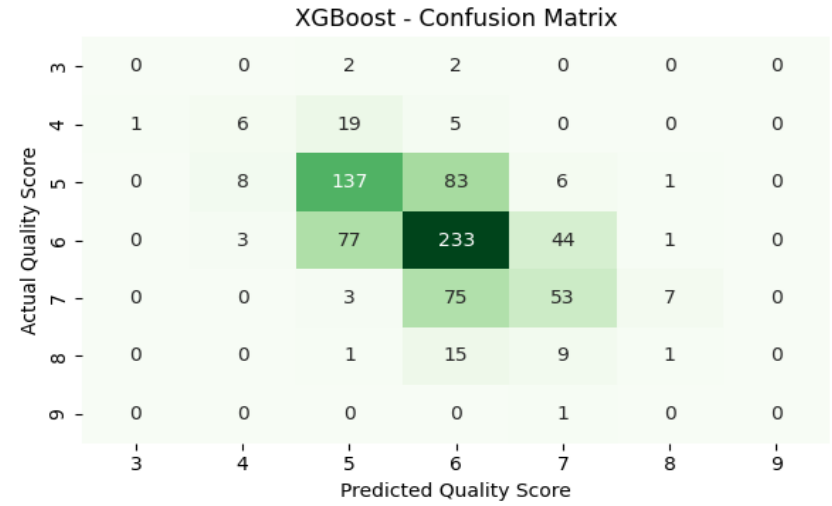


Classification Report:

	precision	recall	f1-score	support
3	0.00	0.00	0.00	4
4	0.10	0.13	0.11	31
5	0.52	0.62	0.56	235
6	0.57	0.41	0.47	358
7	0.36	0.43	0.39	138
8	0.10	0.19	0.13	26
9	0.00	0.00	0.00	1
accuracy			0.45	793
macro avg	0.24	0.25	0.24	793
weighted avg	0.48	0.45	0.46	793

Figure 13. SVM confusion matrix for the white wine dataset

The XGBoost confusion matrix in Figure 14 shows that the model performed best on the middle quality scores. It correctly classified 137 wines with quality score 5, 233 wines with quality score 6 and 53 wines with quality score 7. It also correctly classified 6 wines with quality score 4 and 1 wine with quality score 8, while none of the wines with quality scores 3 and 9 were classified correctly. XGBoost correctly identified more wines with quality score 6 than RF and SVM although it produced fewer correct predictions for scores 5 and 7 than RF. Most of the incorrect predictions were between neighbouring quality scores, particularly between scores 5, 6 and 7. Overall XGBoost performed well on the most common quality classes but showed weaker results for the less frequent scores.



Classification Report:

	precision	recall	f1-score	support
3	0.00	0.00	0.00	4
4	0.35	0.19	0.25	31
5	0.57	0.58	0.58	235
6	0.56	0.65	0.60	358
7	0.47	0.38	0.42	138
8	0.10	0.04	0.06	26
9	0.00	0.00	0.00	1
accuracy			0.54	793
macro avg	0.29	0.26	0.27	793
weighted avg	0.52	0.54	0.53	793

Figure 14. XGBoost confusion matrix for the white wine dataset.

6.3 Physicochemical Feature Importance

Feature importance was examined to identify which physicochemical attributes contributed most to the classification of wine quality. The Random Forest feature-importance results for the white and red wine datasets are presented in Figure 15. For the white wine dataset, free sulfur dioxide was ranked as the most important feature, followed by alcohol and fixed acidity. In the red wine dataset, alcohol had the highest importance, followed by sulphates and volatile acidity. The rankings therefore differed between the two wine types, although alcohol was among the most important features in both datasets. This shows that the Random Forest models relied on somewhat different physicochemical characteristics when classifying the quality of red and white wines.

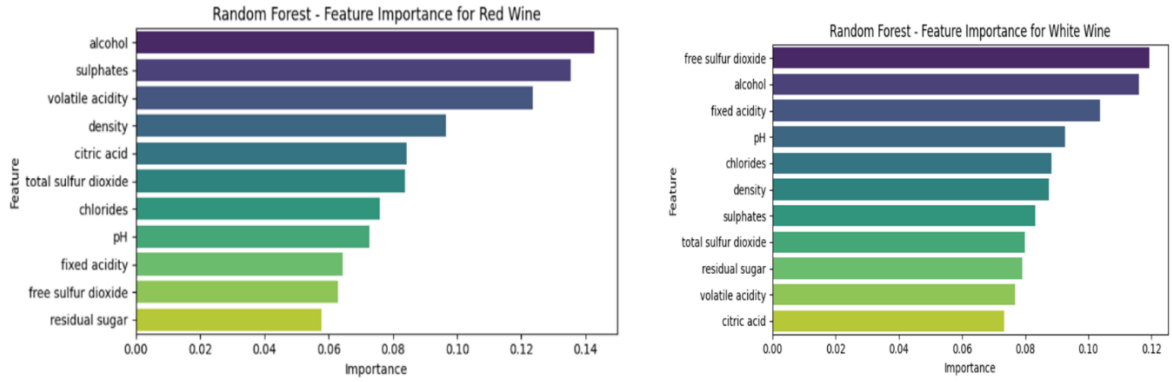


Figure 15. Random Forest feature importance for white wine and red wine datasets.

The SVM permutation-importance results showed a different pattern between the two wine types. For white wine density had the highest importance by a clear margin followed by residual sugar and volatile acidity. For red wine, alcohol was the dominant feature, while sulphates showed only a small positive importance. Most of the remaining red wine features had values close to zero or below zero indicating limited or unstable contribution to the models weighted F1-score when permuted.

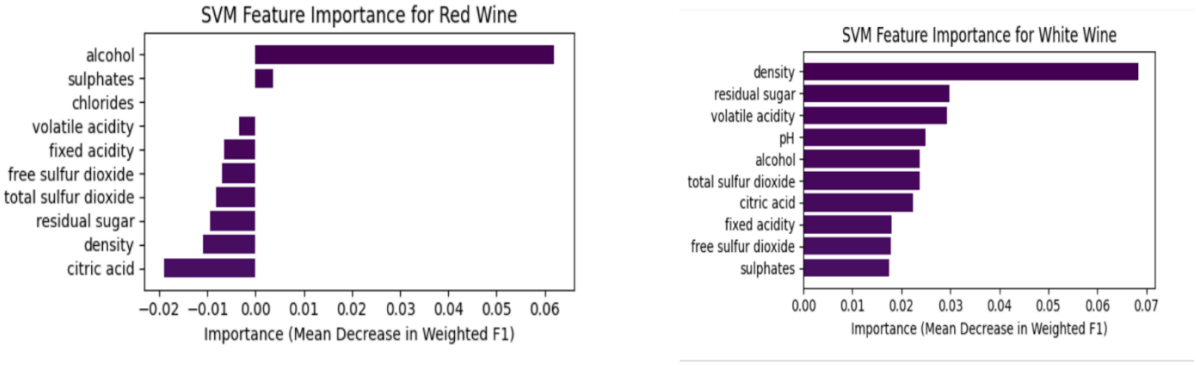


Figure 16. SVM permutation feature importance for white wine and red wine datasets.

The XGBoost feature-importance results also showed differences between the two wine types. For white wine, alcohol was the highest ranked feature, followed by free sulfur dioxide and chlorides. For red wine alcohol was again ranked highest followed by volatile acidity and sulphates. Alcohol was therefore the most important XGBoost feature for both datasets while the remaining top ranked features differed between red and white wine.

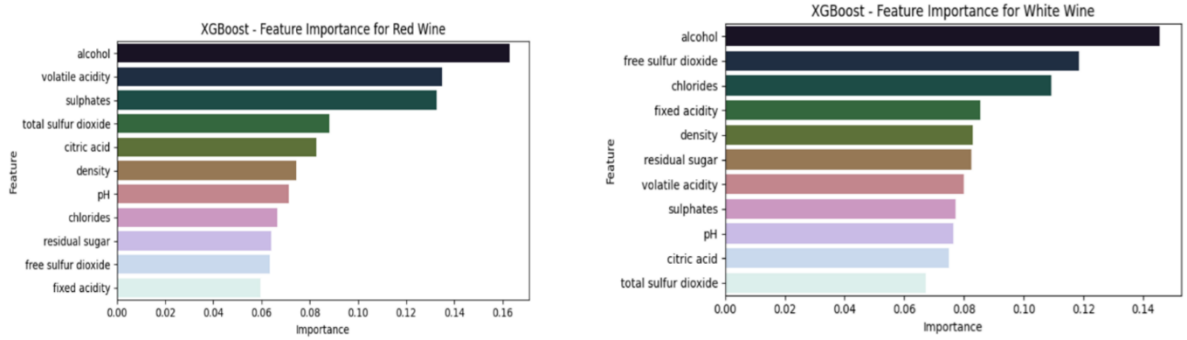


Figure 17. XGBoost feature importance for white wine and red wine datasets.

Overall, the feature rankings varied across the three models and between the two wine types. Alcohol appeared consistently among the highest ranked features and was especially important for the red wines models. For white wine free sulfur dioxide was highly ranked by RF and XGBoost while density was the most important feature for SVM. For red wine alcohol, sulphates and volatile acidity appeared repeatedly among the most influential features. Since the models use different feature importance methods the comparison focused on the relative rankings rather than the absolute importance values.

6.4 Observations and Findings

The results showed a similar pattern in model performance for both the red and white wine datasets. Random Forest achieved the strongest overall performance in both cases while XGBoost followed closely and SVM produced lower results across the main evaluation metrics. Based on the weighted F1-score, RF achieved 0.5886 for red wine and 0.5447 for white wine. XGBoost reached 0.5751 and 0.5293 respectively while SVM obtained 0.4950 for red wine and 0.4579 for white wine. This indicates that the two tree-based models performed more consistently than SVM under the same preprocessing and evaluation procedure. A common pattern was also observed in the confusion matrices. All three models performed better on the more frequently occurring middle quality scores especially scores 5 and 6 while the less frequent scores at the lower and upper ends of the scale were more difficult to classify correctly. Most of the incorrect predictions occurred between neighbouring quality classes particularly among scores 5, 6 and 7. This pattern was present in both datasets showing that the models had more difficulty distinguishing between wines with similar quality ratings than between more clearly separated classes.

The feature importance results showed clearer differences between red and white wine. For red wine alcohol was the most consistent feature across the three models. It was ranked highest by RF and XGBoost and was also the dominant feature in the SVM permutation importance results. Sulphates and volatile acidity also appeared repeatedly among the more important red wine features. These findings are in line with the exploratory analysis where alcohol showed the strongest positive relationship with quality and volatile acidity showed a negative relationship with quality. For white wine the feature rankings varied more between the models. RF ranked free sulfur dioxide as the most important feature followed by alcohol and fixed acidity. SVM ranked density highest followed by residual sugar and volatile acidity while XGBoost ranked alcohol first followed by free sulfur dioxide and chlorides. This shows that no single physicochemical attribute dominated the white wine models in the same way that alcohol did for red wine. It also suggests that the models relied on different combinations of physicochemical characteristics when classifying white wine quality.

Overall, the findings show that the relative performance of the three algorithms was similar for red and white wine with RF giving the strongest overall results followed by XGBoost and then SVM. However, the physicochemical features ranked as most important differed more clearly between the two wine types. Alcohol was consistently important for red wine while the white wine models placed greater importance on a wider range of features including free sulfur dioxide, density, residual sugar and alcohol. These results support the decision to analyse red and white wine separately and show that both model performance and feature importance need to be considered when evaluating machine learning approaches for multiclass wine quality classification.

7 Discussion

7.1 Model Performance and Feature Interpretation

The results showed a similar model performance pattern for both red and white wine. Random Forest achieved the highest weighted F1-score in both datasets with 0.5886 for red wine and 0.5447 for white wine. XGBoost followed closely with 0.5751 and 0.5293 while SVM produced lower scores of 0.4950 and 0.4579. Since the models were trained and evaluated using the same general procedure the results show that RF performed best under the experimental setup used in this study although XGBoost produced relatively close results. The confusion matrices provide a more detailed view of the model performance. All three models classified the more common middle quality scores especially 5 and 6 more successfully than the less frequent scores at the lower and upper ends of the scale. Many errors also occurred between neighbouring scores particularly 5, 6 and 7. This pattern was present in both wine datasets. Although SMOTE was applied during training to reduce class imbalance the rare quality classes were still difficult to classify correctly. This shows why the weighted F1-score should be considered together with the confusion matrices rather than as the only measure of performance.

The feature analysis showed a clearer pattern for red wine. Alcohol was the most consistent feature across the three algorithms. It was ranked highest by RF and XGBoost and was also the dominant feature in the SVM permutation-importance results. Sulphates and volatile acidity were also among the higher ranked red wine features. This was partly consistent with the exploratory analysis where alcohol showed the strongest positive relationship with quality and volatile acidity showed a negative relationship with quality. These results show association and model importance but they do not show that these features directly cause changes in wine quality. For white wine the feature rankings differed more between the models. RF ranked free sulfur dioxide highest followed by alcohol and fixed acidity. XGBoost ranked alcohol first followed by free sulfur dioxide and chlorides while SVM ranked density highest followed by residual sugar and volatile acidity. The exploratory analysis also showed a stronger negative relationship between density and quality for white wine. The different rankings suggest that the three models used the physicochemical information differently when classifying white wine quality.

The feature-importance results should also be interpreted carefully because RF and XGBoost use built in importance measures while SVM was interpreted using permutation importance based on changes in weighted F1-score. Therefore, the absolute importance values should not be compared directly across the models. The relative ranking of the features provides a more appropriate comparison. Overall RF produced the strongest predictive results for both wine types while the feature analysis showed that alcohol was especially consistent for red wine and that the important features for white wine varied more between the three algorithms.

7.2 Previous Studies and Practical Implications

The results of this study differ from several previous studies where SVM was reported as the strongest model. Gupta (2018) found SVM to perform well after feature reduction, while Anami et al. (2022) also reported strong SVM results when predicting wine quality using a different modelling approach. Aich et al. (2019) achieved particularly high performance when SVM was combined with simulated annealing-based feature selection and Kumar et al. (2020) reported a test accuracy of 68.64% for SVM on red wine. In contrast SVM produced the lowest performance in the present study, while RF achieved the strongest results. These differences should be interpreted carefully because the studies used different

preprocessing methods, feature-selection procedures, prediction settings and evaluation approaches. The present study retained the original quality scores as separate multiclass labels and compared all three models using the same general training and evaluation procedure. The XGBoost results also differ from the findings of Beri et al. (2024), where XGBoost was reported as the best performing model with considerably higher evaluation scores. In the present study XGBoost performed well but remained slightly below RF for both wine types. Again, a direct comparison of numerical scores is difficult because the datasets preprocessing steps and modelling procedures were not identical. This supports the view that the performance of a machine learning model in wine quality classification can depend strongly on how the prediction problem and experimental procedure are designed.

There were also similarities between the feature findings and previous research. Gupta (2018) identified alcohol, volatile acidity and sulphates as important variables related to wine quality while Anami et al. (2022) also highlighted alcohol and sulphates. These findings are similar to the red wine results in this study where alcohol was the most consistent feature and sulphates and volatile acidity were also highly ranked. For white wine however the important features varied more across the three models including free sulfur dioxide, alcohol, density and residual sugar. This suggests that some physicochemical characteristics appear repeatedly across studies but their importance can vary depending on the wine type and the algorithm used.

From a practical perspective the results suggest that machine learning could be used as a supporting tool in wine quality assessment. Physicochemical measurements are objective and can be collected consistently which means that models such as RF and XGBoost could provide an additional data driven assessment during quality control activities. However, the moderate overall performance and the difficulty in correctly classifying less frequent quality scores show that the models should not be treated as replacements for sensory evaluation. A more realistic use would be to support experts by providing an additional classification result or helping identify wines that may require further examination. Since the important features also differed between red and white wine separate models for the two wine types may be more appropriate than using one general model for all wines. This is consistent with the aim of the study which is to examine machine learning as a complementary approach rather than replacing sensory assessment. However, these findings are based on the Vinho Verde dataset so the performance of the models may differ when applied to other wine datasets.

7.3 Ethical and Societal Aspects

While the technical result is promising applying prediction models to wine quality evaluation must consider ethical and societal aspects. Models can reflect biases or limitations present in their data. Because the dataset used in this study represents Portuguese Vinho Verde wines the trained models may not represent wines from other regions grape varieties or production conditions equally well. If such a model were applied more broadly this could lead to less reliable assessments for wines that differ from those represented in the training data. A second concern is interpretability. Even though RF can provide feature importance rankings, models such as SVM are less directly interpretable which may make it more difficult for stakeholders to understand how predictions are produced. The use of additional explainability techniques or the preference for models with greater inherent interpretability can help make model decisions easier to understand for producers and other stakeholders. However, feature importance should not be treated as a complete explanation of individual predictions.

Privacy concerns are limited in this study because the dataset contains physicochemical measurements and wine quality scores rather than personal information. No interviews or personal data were collected, so issues such as informed consent and GDPR-related processing of personal data were not central to

the study. If machine learning based quality assessment were widely adopted, there is a risk that wines matching particular physicochemical profiles could be favoured while other styles may be undervalued. This could contribute to greater standardisation in how wine quality is evaluated. The risk can be reduced by regularly testing and validating prediction algorithms using different datasets, relying on human sensory evaluation and informing users of their limits.

Machine learning should be considered as an enhancement rather than a substitute for human expertise and sensory evaluation in winemaking. Wine quality also includes sensory and stylistic characteristics that may not be fully captured by physicochemical measurements alone. Therefore, automated predictions should support rather than replace human judgement. The application of predictive modelling should therefore emphasize transparency appropriate human oversight and clear communication of model limitations.

7.4 Validity and Limitations of the Study

7.4.1 Internal Validity

Internal validity concerns whether the observed results may have been influenced by factors within the experimental design (Wohlin et al.,2012). Several methodological choices may have influenced the results of this study. The use of an 80/20 train-test split with a fixed random state of 42 supported reproducibility, but a different split or seed could have produced somewhat different results, especially for the less frequent quality classes. Similarly, five-fold cross-validation reduced dependence on a single validation split, but the selected number of folds and the predefined GridSearchCV ranges may have influenced which hyperparameters were chosen. Since RF, SVM and XGBoost require different model-specific parameters, it cannot be guaranteed that all three models were tuned equally well. Although the same overall evaluation procedure was applied to all three algorithms, the comparison may not be completely equivalent because RF, SVM and XGBoost have different model structures, sensitivities to preprocessing and hyperparameter spaces. A broader hyperparameter search or alternative optimization method such as randomized or Bayesian optimization could potentially have identified better configurations and provided a more comprehensive comparison.

Preprocessing choices may also have affected the results. Outliers were removed only from alcohol, sulphates and volatile acidity, which may have removed unusual but valid observations. SMOTE was used to address class imbalance, but the synthetic samples depend on the selected settings and may influence how the models learn minority classes. In addition, RF and XGBoost used built-in feature importance while SVM used permutation importance, so differences in feature rankings may partly result from the different importance methods rather than only from differences between the algorithms. Weighted F1-score was emphasized because the classes were imbalanced, but weighted measures give more influence to the common classes. This means that the overall performance results may appear stronger even when the models perform poorly on rare quality scores. For this reason, the confusion matrices and class-level results are important when interpreting the final comparison.

7.4.2 External Validity

External validity concerns the extent to which the findings can be generalized beyond the specific conditions in which the study was carried out (Wohlin et al., 2012). The main limitation to external validity is that the study is based on a single data source, the Portuguese Vinho Verde datasets. The data contain 1,599 red wine samples and 4,898 white wine samples and represent specific types of Portuguese wine. Because the models were developed and evaluated using these data, the findings

cannot automatically be generalized to wines from other countries, regions or production environments. Differences in grape varieties, climate, production methods and chemical composition may influence the relationship between physicochemical attributes and wine quality. Testing the models on independent datasets from other wine regions would provide stronger evidence about how well the findings generalize.

The study is also limited to the eleven physicochemical attributes available in the dataset. These variables provide measurable information about the chemical characteristics of the wines, but wine quality is not determined only by physicochemical measurements. The quality scores in the dataset are based on sensory evaluation, which involves human judgement and may contain some variation between tasters. Other information such as grape variety, geographical origin, production process or additional sensory characteristics was not available in the dataset. The findings should therefore be understood as showing how well wine quality can be classified using the physicochemical variables included in these datasets, rather than as a complete representation of wine quality.

The unequal number of observations across the quality classes also affects the generalizability of the findings. Most wines in the datasets are concentrated around quality scores 5-7 while the lower and higher quality scores contain relatively few observations. As a result, the conclusions for the more common quality levels are based on more observations than those for rare quality levels. Additional data containing more examples of very low and very high-quality wines would be needed to determine whether the observed performance for these classes would remain similar in a broader population. For this reason, the results mainly reflect the behaviour of RF, SVM and XGBoost under the dataset, preprocessing choices and experimental conditions used in this study. Further evaluation using independent and more diverse wine datasets would be needed before making broader conclusions about their performance in other wine quality classification settings. External validation using independently collected wine datasets would therefore be necessary before assuming that the observed model performance and feature rankings also apply to other wine populations.

8 Conclusion and Future Work

8.1 Conclusion

This study investigated the use of Random Forest, Support Vector Machine and XGBoost for multiclass wine quality classification based on physicochemical characteristics. The red and white Vinho Verde wine datasets were analysed separately using the same preprocessing, training and evaluation procedure. The models were compared using accuracy, weighted precision, weighted recall and weighted F1-score while feature importance was used to examine which physicochemical attributes contributed most to the classification of wine quality. The results showed that Random Forest achieved the strongest overall performance for both wine types. For red wine, RF achieved a weighted F1-score of 0.5886 followed by XGBoost with 0.5751 and SVM with 0.4950. A similar pattern was observed for white wine where RF achieved a weighted F1-score of 0.5447, XGBoost achieved 0.5293 and SVM achieved 0.4579. This shows that under the experimental setup used in this study RF provided the best overall predictive performance, although XGBoost produced relatively close results. The confusion matrices also showed that all three models performed better on the more common middle quality scores, especially scores 5 and 6 while the less frequent scores at the lower and upper ends of the scale were more difficult to classify correctly.

The feature importance results also showed differences between red and white wine. For red wine alcohol was the most consistent feature across the three models. It was ranked highest by RF and XGBoost and was also the dominant feature in the SVM permutation importance results. Sulphates and volatile acidity were also repeatedly identified among the more important red wine features. For white wine the feature rankings varied more between the models. RF ranked free sulfur dioxide highest, SVM ranked density highest and XGBoost ranked alcohol highest. These findings suggest that the importance of physicochemical characteristics can differ depending on both the wine type and the machine learning algorithm used.

Overall, the study shows that machine learning can provide useful support for wine quality assessment based on physicochemical measurements. However, the moderate overall performance and the difficulty in classifying less frequent quality scores show that these models should not be viewed as replacements for sensory evaluation. A more appropriate use is as a complementary tool that can support experts with additional data driven information during quality assessment. Since the study is based on Portuguese Vinho Verde wines, the performance of the models may also differ when applied to wines from other regions or production conditions.

8.2 Future work

A broader range of variables and a larger dataset can be utilized in future iterations of this study. Although the physicochemical properties were in focus for this study, others like grape variety, vineyard location, detailed sensory characteristics such as aroma, taste and body and seasonal climatic conditions can help provide a better understanding of factors affecting wine quality. Data from different regions and years would reduce dependence on the characteristics of a single dataset and render the models more generalizable. Such advancement would allow the models to be evaluated more broadly for different wine varieties, production methods and regions and thus make them more useful in a wider range of wine production settings. Future work should compare the current multiclass classification approach with ordinal classification or regression and pursue external validation across regions and vintages. These approaches could preserve the ordered nature of the original wine quality scores and provide

another perspective on model performance. Since red and white wines were modelled separately in this study, future research could also compare this approach with a combined model in which wine type is included as an explicit feature.

The second important direction is to evaluate hybrid or ensemble techniques that combine the strengths of more than one algorithm. For example, future research could investigate whether combining RF and XGBoost which were the two strongest models in this study can improve predictive performance. The incorporation of deep learning models where data volume permits, could also be investigated to identify more complex nonlinear patterns. A common explainability method such as permutation importance or SHAP, could also be applied across the different models to make feature interpretation more consistent and easier to compare. Long term deployment should also involve validating such models using pilot studies and longitudinal testing in actual winery operations. Real world assessment in such applications would in addition to predictive accuracy take into account processing speed, cost and winemaker acceptance providing a more complete picture of their practical usefulness.

References

Aich, S., Al-Absi, A. A., Hui, K. L., & Sain, M. (2019, February 17–20). Prediction of quality for different type of wine based on different feature sets using supervised machine learning techniques. *2019 21st International Conference on Advanced Communication Technology (ICACT)*, 1122–1127. IEEE. https://doi.org/10.23919/ICACT.2019.8702017

Anami, B. S., Mainalli, K., Kallur, S., & Patil, V. (2022, August 26–28). A machine learning based approach for wine quality prediction. *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, 1–6. IEEE. https://doi.org/10.1109/ASIANCON55314.2022.9908870

Beri, M., Gill, K. S., & Sharma, N. (2024, August). Predictive modeling of wine quality using machine learning techniques. *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*, 1017–1022. IEEE. https://doi.org/10.1109/ICoICI62503.2024.10696690

Bisson, L. F., Waterhouse, A. L., Ebeler, S. E., Walker, M. A., & Lapsley, J. T. (2002, August 8). The present and future of the international wine industry. *Nature, 418*(6898), 696–699. https://doi.org/10.1038/nature01018

Breiman, L. (2001, October). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chakraborty, S., & Dey, L. (2024). Multi-objective, multi-class and multi-label data classification with class imbalance: Theory and practices. Springer Nature Singapore. https://doi.org/10.1007/978981-97-9622-9

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002, June). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research, 16*, 321–357. https://doi.org/10.1613/jair.953

Chen, T., & Guestrin, C. (2016, August 13–17). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and* Data *Mining*, 785–794. ACM. https://doi.org/10.1145/2939672.2939785

Cortes, C., & Vapnik, V. (1995, September). Support-vector networks. *Machine Learning, 20*(3), 273– 297. https://doi.org/10.1007/BF00994018

Cortez, P., Cerdeira, A., Almeida, F., Matos, T., & Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems, 47*(4), 547–553. https://doi.org/10.1016/j.dss.2009.05.016

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In C. Zhang & Y. Q. Ma (Eds.), *Ensemble machine learning: Methods and applications* (pp. 157–175). Springer. https://doi.org/10.1007/978-1-4419-9326-7_5

Gandhi, R. (2018, July 5). Support vector machine — introduction to machine learning algorithms. *Medium*. Retrieved May 6, 2025, from https://towardsdatascience.com/support-vector-machineintroduction-to-machine-learning-algorithms-934a444fca47

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly Media.

Gupta, Y. (2018). Selection of important features and predicting wine quality using machine learning techniques. *Procedia Computer Science, 125*, 305–312. https://doi.org/10.1016/j.procs.2017.12.041

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Elsevier.

Hodgson, R. T. (2008). An examination of judge reliability at a major U.S. wine competition. Journal of Wine Economics, 3(2), 105–113.

Kumar, S., Agrawal, K., & Mandan, N. (2020, January 22–24). Red wine quality prediction using machine learning techniques. *2020 International Conference on Computer Communication and Informatics (ICCCI)*, 1–6. IEEE. https://doi.org/10.1109/ICCCI48352.2020.9104095

Lesschaeve, I. (2007). Sensory evaluation of wine and commercial realities: Review of current practices and perspectives. *American Journal of Enology and Viticulture, 58*(2), 252–257.

Mammone, A., Turchi, M., & Cristianini, N. (2009, November). Support vector machines. *WIREs* Computational *Statistics, 1*(3), 283–289. https://doi.org/10.1002/wics.49

Preedy, V., & Mendez, M. L. R. (2016). Wine applications with electronic noses. In V. Preedy (Ed.), *Electronic noses and tongues in food science* (pp. 137–151). Academic Press.

Sirivanth, P., Krishna Rao, N. V., Manduva, J., Sekhar, G. C., Tajeswi, M., Veeresh, C., & Kaushik, J. V. (2021, December). A SVM-based wine superiority estimation using advanced ML techniques. *2021 3rd International Conference on Advances in Computing, Communication Control and* Networking *(ICAC3N)*, 207–211. IEEE. https://doi.org/10.1109/ICAC3N53548.2021.9725492

Taylor-Sakyi, K. (2016). *Big data: Understanding big data* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.1601.04602

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., & Wesslén, A. (2012). *Experimentation in software engineering*. Springer. https://doi.org/10.1007/978-3-642-29044-2