National College of Ireland

Predicting Wine Quality with High Accuracy using Machine Learning Models

MSc Research Project
Data Analytics

Bhupinder Singh
Student ID: x22197982

School of Computing
National College of Ireland

Supervisor: Vikas Tomer

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Bhupinder Singh
Student ID:	x22197982
Programme:	Masters In Data Analytics
Year:	2023
Module:	MSc Research Project
Supervisor:	Vikas Tomer
Submission Due Date:	12/08/2024
Project Title:	Predicting Wine Quality with High Accuracy using Machine Learning Models
Word Count:	7786
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Bhupinder Singh
Date:	12/08/2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Predicting Wine Quality with High Accuracy using Machine Learning Models

Bhupinder Singh
22197982

Abstract

This Research is about Predicting the wine Quality using a dataset which is having chemical Properties, sensory feedback, environmental factors and Aging data in it. The main objective of this search is to predict wine Quality with high accuracy using different machine learning models. A Comprehensive approach was taken to produce results, like Data collection then data processing then feature engineering, then EDA and then finally implementing different machine learning models like Random Forest, Gradient Boosting, XGBoost, SVM and stacking ensemble. Different approaches were taken while implementing these models to get the optimal results, techniques like feature selection then hyper parameter tunning were used for each model separately. Also, Cross validation techniques was used to ensure the reliability of these machine learning models. Then to evaluate the performance of each model different evaluation metrics like Accuracy, Precision, Recall, F1-score were used. The Random Forest (Tunned) got 87.10% accuracy, Gradient Boosting (Tunned) got 86.13% accuracy, XGBoost (Tunned) got 87.61% accuracy, SVM tunned got 73.01% accuracy and the best model was stacking ensemble which got accuracy of 88.47%. These results show the effectiveness of these machine learning algorithms in predicting wine quality with great number of wine features together. The methodology of this research is robust and could be used in wine making industry to predict wine quality with use of machine learning techniques, as this would improve accuracy in wine quality prediction and would improve the overall wine quality.

1 Introduction

1.1 Background

The Wine Quality assessment is very critical as this plays an important role in market value of the wine and plays a major role in consumer satisfaction as well. In the past the wine quality assessment had been done with traditional methods like evaluations done by Human experts after tasting the different wine. Physicochemical properties like Sugar content, Acidity, alcohol content etc are always the main determinant of the wine quality, but there are many other factors like fermentation duration, Temperature, Aging process and time which also plays a major role in determining the wine quality. With help of advance Data analysis and Machine learning, the wine quality can we determine using different machine learning Models with high accuracy using these different features.

1.2 Motivation

The assessment of the wine quality with help of these machine learning models with high accuracy helps both Producers and Consumers. For Producer, using this assessment the quality of wine can be enhanced, then the overall production costs can be reduced, also the resources can be utilised more efficiently. Then for the consumer this gives the purchasing power as per there own likes and dislikes, also this help in ensuring consistent quality of the wine. Using this machine learning models the relationship between features could be understood which are very complex and this is not possible using traditional methods.

1.3 Objective

The objective of this search is predicting wine quality with high accuracy based on Physicochemical properties, Sensory feedback by human, different Fermentation features like Duration, temperature and different Aging features of the wine together with help of different machine learning models like Random Forest, Gradient Boosting, XGBoost and Support Vector Machine, Stacking Ensemble.

1.4 Methodology

The research Methodology includes different steps like Data Collection, Preprocessing of the Data, Feature Engineering and Feature selection, EDA. Different machine learning models have been trained on the comprehensive dataset, having different wine features. Then Feature selection is done to find out the main features and machine learning models are implemented again on the reduced feature sets. Hyperparameter tunning is done to have optimal performance of each model. And for each step the performance of the models has been calculated using different evaluation metrics. Also stacking ensemble approach has been done to combine the strength of each model to have better overall performance.

1.5 Contribution

The search contributes to wine quality assessment using different wine features together at the same time, It also gives a comparative analysis of the performance of each model. Then the methodology of this research provides strong base of features selection process and Model optimization techniques and provides information about using an Ensemble model to get the higher prediction accuracy of wine.

The search paper is having Related work in next section where there is a detailed discussion of the recent search that has been done on wine quality assessment using different wine features and different Machine learning techniques. After this their is Methodology section, having each step been discussed in detail from data collection to model training and evaluation. Then the next section is Results and discussion, under which the performance of each machine learning model has been discussed with hyperparameter tunning using different evaluation metrics. In the last there is Conclusion and future section under which the future potential work of the research has been discussed in the detail.

2 Related Work

White wine Quality prediction: Jiang et al. (2023) have done a comprehensive evaluation of white wine quality prediction in this research. For this evaluation different machine learning algorithms have been used, like decision tree, Random Forest, and SVMs. In this research it has been found that feature selection and preprocessing steps were the main steps, which have resulted in high prediction accuracy. Random Forests have performed the best among all the algorithms as this is very robust when handling complex feature interactions.

Red wine quality prediction using Regression modelling: K (2023) have used Regression modelling approaches for the red wine quality prediction in this research, Both Individual and ensemble techniques have been used in this. The main findings of the research are, with use of different ensemble methods like gradient boosting and bagging the overall prediction accuracy could be improved. Also, the research highlights the importance of using hyperparameter tunning, as this optimizes the overall performance of the algorithms. If these techniques are used together with meticulous parameters adjustment, it can enhance the overall performance of predictive models significantly.

Integrating Spectral Data for Sensory Prediction: Armstrong et al. (2023) have used fused spectral data in this research to predict the sensory prediction of the wine. The fused spectral data have been integrated with different machine learning for this prediction purpose. The findings from the research have shown that including spectral data, have resulted in high accuracy of sensory attribute prediction of the wine. this show that more comprehensive approaches could be taken in the wine quality assessment with having more diverse datasets and different data types.

Impact of fermentation conditions of wine quality: Godillot et al. (2023) have taken different fermentation conditions into consideration in their studies to predict the wine quality. Conditions like Nitrogen addition and temperature on volatile compound production during fermentation have been considered. The findings from this research shows that, the different Fermentation conditions play a significant role in determining the wine quality. So, by seeing these results from this study it is evident that the difference fermentation conditions must be considered and must be controlled precisely to have desired wine quality as per the requirements.

Effect of Different Fermentation temperature on wine Parameters: Abarca-Rivas et al. (2023)Another research which has been done to study the effect of fermentation temperature on different parameters of wine, have showed that these temperature variations causes major influence on sensory attributes of the wine.

Influence of aging Barrel on Wine Quality: Denchai et al. (2023) have investigated the impact of using different types of aging barrel on the metabolite profile of the red wine. As different types of aging barrel have been used for different time periods for the wine production as per the requirement, this study have showed that, the use of these barrels have significant influence the chemical composition of the wine and the quality of wine is dependent on these factors. So, optimizing the use of different types of barrel and aging conditions can enhance the wine quality and this must be controlled precisely to have desired wine quality as per the requirements.

Yeast Concentrations and impact on wine quality: MN et al. (2023) have investigated the impact of different yeast strains and there concentration on the wine quality that too during wine aging. The findings from this research shows that the difference in yeast strain with different concentrations have major influence on sensory

attributes of the wine. Also, these different concentrations of the yeast influence the chemical composition of wine which impacts the wine quality and flavours. This study highlights that the yeast usage must be controlled precisely to have desired wine quality as per the requirements.

Different Machine learning Model performance comparisons: Liu (2023) have used different machine learning models to predict the wine quality. Mainly in this study there has been a comparison of the performance of these machine learning models in predicting wine quality. Models which have been used in this study were KNN, SVM, Random Forest and Neural networks. The main findings from the study are that ensemble methods and neural networks have good accuracy in prediction the wine quality. So, this study highlights that advance machine learning models are good options to have high prediction accuracy of wine quality.

Hyperparameter optimizations of the models: Basha et al. (2023)have used different machine learning models to predict the red wine quality in this study. The focus of the study was to have high accuracy in wine quality prediction with the help of hyperparameters tunning of the different machine learning models. This study shows a significant improvement in the accuracy of the machine learning models in wine quality predictions when they were used with these fine tunned hyperparameters for each model. So, this study highlights that advance machine learning models with systematic parameter adjustments are good options to have high prediction accuracy of wine quality.

Stacking Ensemble Learning: Zhou et al. (2023)have combined different multiple base models together to have better accuracy in wine quality prediction. This approach is knowing as improved stacking ensemble methods to have better accuracy and a reliable prediction system. The findings from the study showed significant improved in the accuracy of this ensemble methods when compared to individual models. So, this study highlights the usage of this advance models to have better accuracy in wine quality prediction.

Algorithms analysis for classification: Irmalasari and Dwiyanti (2023) have used different models like Decision trees, Gradient Boosting decision tress and random forests for the classification task, and the performance of each model have been evaluated, analysed and compared. The main findings of this study were, gradient Boosting and Random Forest models got the best results in the classification task. So, this study highlights the usage of these gradient Boosting and Random Forest models and ensemble methods to have better accuracy in wine quality prediction using a complex dataset.

2.1 Literature Review Table:

Below is the Literature review Table which includes the Recent Research Papers which have been used as base for this research. This Table Consists of information like what type of Dataset has been used in each research. Then what all are the different Models used, then there are details regarding the Results Metrics with there specific values for different models used. Also there are columns which discuss about the Limitation and potential future work of each research.

Papers (Year-Author)	**Dataset Used**	**Model Used**	**Results Metrics Used**	**Value**	**Limitations**	**Future Work**
Jiang et al. (2023)	White Wine Dataset	Random Forest	Accuracy	90%	Small and Imbalanced dataset used.	SMOTE can be used to address Imbalance Dataset.
			Precision	86%		
		Logistic Regression	Accuracy	52%		
			Precision	51%		
Liu (2023)	Red Wine Dataset	Decision Tree	Accuracy	84.16%	Dataset with limited characteristics, potential Overfitting.	Diverse Dataset must be used with more advance models.
		Random Forest	Accuracy	91.25%		
		SVM	Accuracy	92.25%		
Basha et al. (2023)	Red Wine Dataset	Gradient Boosting	Accuracy	91.75%	Risk of overfitting and high computational Complexity.	Use of more diverse Dataset, exploring Hybrid Models.
			F1-Score	92.67%		
		Random Forest	Accuracy	89.24%		
			F1-Score	86.40%		
Zhou et al. (2023)	Red and white wine Dataset	Improved stacking Ensemble Learning	Accuracy	92%	Risk of overfitting and high computational Complexity.	Cross Validation and Regularization techniques can be used.
			F1-Score	92%		
		Random Forest	Accuracy	86%		
			F1-Score	86%		
Irmalasari and Dwiyanti (2023)	Candidate Profile Dataset	Decision Tree	Accuracy	76%	Dataset with limited characteristics, potential Overfitting.	Diverse Dataset must be used with more advance models with feature selection.
		Random Forest	Accuracy	86%		
		Gradient Boosting	Accuracy	85%		

Table 1: Literature Review Table

2.2 Identified Gaps in the recent studies and what could be the future Directions:

Despite these advancements in wine quality assessments with help of these different machine learning techniques there still exists many Gaps which could be work on to have better results. Some of the potential gaps have been discussed below:

2.2.1 Diversity:

Most of the studies have used datasets having physicochemical properties mainly, or having either indirect sensory data or environmental data Separately. But very less studies have used all the different types of data together to predict the wine quality. This could be work upon where dataset is having all the different types of data together, using which wine quality can assessed more accurately.

2.2.2 Feature Integration:

In the filed of wine quality assessment the features of different environmental factors like fermentation duration, fermentation temperature, Yest strain and then different Aging factors like aging time, Aging Vessel, aging temperature have been not utilized fully. These features could be included in further assessment of the wine to understand the effect of these features on the quality of the wine.

2.2.3 Explain ability and Transparency in machine learning models:

In the recent studies very, less importance has been given to interpretability of these machine learning models in wine quality prediction. In further assessment a proper explanation must be given about explaining the working of these models, as this would ensure transparency in these predictions and would gain trust of winemakers, which would help them to adopt these technologies in their wine production.

2.2.4 Conclusion:

The above literature shows that machine learning techniques are very much effective in wine quality assessment and models like Random Forest, Gradient Boosting and Ensemble methods are very Robust and successful with handling the complex datasets. It is evident that there are many factors which influence the wine quality, and these factors must be considered when wine quality is being assessed using these machine learning techniques. By having a diverse dataset having different features of wine together could help to predict the wine quality with high accuracy, also featuring engineering could help to enhance the overall performance of these machine learning models. Detailed explanation could be provided in the code section to have better understanding of models working which would provide more transparency. These steps could be work on to bridge the existing Gaps in the recent research and wine quality assessment could be optimized further. Next section in the report is Methodology.

3 Methodology

This section of the report discusses about the methodology section of the search. All the steps are discussed in detail to make sure the replicability and reliability of the results.

3.1 Below is the flowchart which tells about different steps of the methodology:

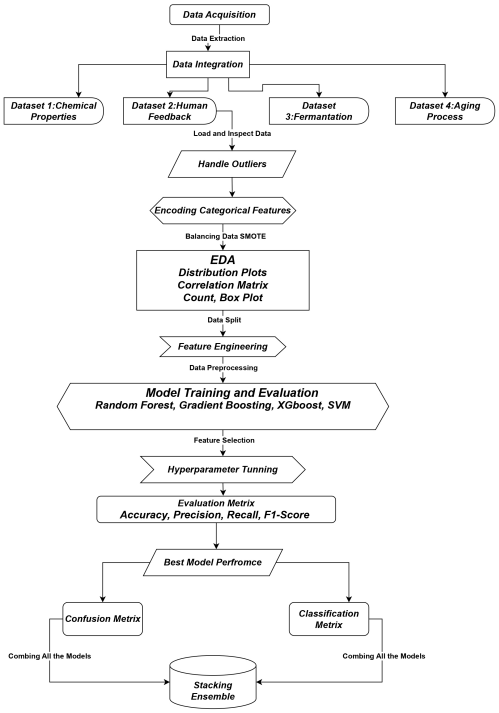

Figure 1: Flow Chart

3.2 Data Collection:

For this research 4 different dataset were collected from different site including UCI repository and Kaggle to perform the wine quality assessment.

3.2.1 First Dataset

First Dataset Is related to chemical properties of the wine, this dataset includes different and important chemical features of the wine, using this wine quality can be predicted. They features are Fixed acidity, Volatile Acidity, Residual Sugar, Citric Acid, chlorides, Total Sulphur Dioxide, Free Sulphur Dioxide, Density, PH level, Sulphate, Alcohol, Final Alcohol Content and Wine Type.

3.2.2 Second Dataset

Second dataset is related to Sensory feedback of the wine, which is given by the humans, using which wine quality can be predicted. Categories of this sensory feedback are Excellent, Good, Average, Poor.

3.2.3 Third Dataset

Third Dataset is related to Fermentation details in the wine making process, using which wine quality can be predicted as well. These details are about Fermentation days, Fermentation temperature and the Yest strain used in wine making process.

3.2.4 Fourth Dataset

The Fourth Dataset is related Aging details of the wine, using these details the wine quality can be predict. These details are about Aging Time in months, Aging vessel type used and Aging Temperature of the wine which has been used while wine making process.

3.3 Combining Datasets:

These Four Dataset were collected from the reputable sources as mentioned above and then they were merged using a column name Wine Id which was common in all the four datasets. After this process a unified dataset was created which have 22 features in it.

3.4 Feature Engineering:

This step is very important as this improves the overall performance of different machine learning predicting models, by adding the different features together to get a meaningful feature. In this a new feature known as Refined wine quality rating have been created using other key features from the combined dataset to help in predicting the wine quality comprehensively. Below are the details about the Rationale behind each heuristic that has been used while doing the calculation of the Wine quality points using different features.

3.4.1 Physicochemical Properties:

For the physicochemical properties which are the chemical properties of the wine, main properties were selected which are PH level, Alcohol content and Residual sugar. And for these properties below heuristic were used.

1. PH level- For the PH level optimal range (3.1 – 3.4), 2 points have been assigned. Then for the Acceptable Range (3.0 – 3.1 or 3.4 – 3.5), 1 point has been assigned, Yao and Jia (2023); Mor et al. (2022).This assignment of the points as per the different range is done because a PH level within the range of (3.1 – 3.4), is very much optimal for the wine stability and its flavour profile. Wine which are having PH level outside this range mostly have fewer desirable characteristics,Mor et al. (2022).

2. Alcohol Content- For the Alcohol content optimal range (12.0% - 13.5%), 2 points have been assigned. Then for the Acceptable Range (11.0% – 12.0% or 13.5% – 14.0%), 1 point has been assigned, Yao and Jia (2023).This assignment of the points

as per the different range is done because an Alcohol Content within the range of (12% – 13.5%), significantly affects the wine quality, its balance and preservation. Wine which are having alcohol content outside this range mostly have fewer desirable characteristics. These chosen ranges for the alcohol content of the wine provides a balance alcohol strength and its flavour, Pascua et al. (2023).

3. Residual sugar- For the Residual Sugar optimal range ($< 1.5\,g/L$), 2 points have been assigned. Then for the Acceptable Range (1.5 g/L – 2.0 g/L), 1 point has been assigned, Yao and Jia (2023).This assignment of the points as per the different range is done because a Residual Sugar within the range of ($< 1.5\,g/L$), is very much optimal for the better wine quality, as lower residual sugar is preferred in dry Wines. Wine which are having residual sugar outside this range mostly have fewer desirable characteristics.

3.4.2 Fermentation Data:

These are the details about different conditions of the fermentation process, and these plays a vital role in determining the flavour of the wine.

1. Fermentation Duration - For the Fermentation Duration optimal range (12-25 days), 2 points have been assigned. Then for the Acceptable Range (10 -12 days, 25 – 30days), 1 point has been assigned, Mor et al. (2022); Di and Yang (2022).This assignment of the points as per the different range is done because a Fermentation Duration within the range of (12 – 25days), is very much optimal for the wine flavour development. Wine which are having PH level outside this range mostly have issues like fermentation problems or off flavours, Pascua et al. (2023).

3.4.3 Aging Data:

These are the details about how long the wine have been kept for the aging process, and these plays a vital role in determining the flavour and quality of the wine.

1. Aging Time - For the Aging Time optimal range (≥ 18 months)., 2 points have been assigned. Then for the Acceptable Range (12 – 18 months), 1 point has been assigned, Yao and Jia (2023).This assignment of the points as per the different range is done because an Aging time within the range of ($\geq$ 18 months), is very much optimal for the wine flavour development and this contributes to depth, balance in the wine. Wine which are having Aging Time outside this range mostly are not of great quality in taste and smell, Di and Yang (2022).

3.4.4 Human feedback:

This is the feedback which have been provide by the humans after tasting the wine. and these sensory feedback plays a vital role in determining the flavour of the wine. Excellent – 3 points, Good – 2 points, Average – 1 Points, Poor - 0 points.The above Heuristics have been derived using different research papers, scientific research in oenology and winemaking principles. After this step a new column name refined wine quality have been added in dataset, which is having wine quality rating from 2 to 10 points. Now the dataset got 23 columns including the new featured engineer column , Yao and Jia (2023); Mor et al. (2022); Di and Yang (2022).

3.5 Data Inspection and preprocessing:

These steps are mainly focused on to make sure that the final dataset is ready for the further steps, first dataset is loaded to check the initial view of the dataset content and data structure for each column.

1. Handling Missing Values- In the next step, dataset was checked for any missing values present in any of the columns. These missing values were managed with help of forward fill. This managed the missing values very efficiently and made sure that dataset is complete without any missing values in it. this step made sure that all the gaps in data were filled which would prevent and potential issues during model training and their evaluation.

2. Separating feature and target variables- In this step the target variable which is 'Refined wine quality rating' were separated from other features in the dataset. This is done because it is important for the machine learning models to learn the different relationships between the target variable and other features.

3. Identifying Numerical and Categorical features- After the target variable separation from other feature, all the remaining features were identified in Numerical and categorical categories. As this is important for the next step as we must do encoding for categorical features and scaling for Numerical features. Features like 'Wine Type', 'Yest strain', 'Human Feedback', 'Aging Vessel' were identified as categorical features and remaining were identified as numerical features.

3.6 Detecting and Handling Outliers:

In this steps outlier were detected from the dataset and they were removed, as in data processing identifying these outliers and then removing them is very crucial because it make sure that predicting models are effective, robust and reliable. If these outliers are not removed, they would lead to inaccurate predictions and insights form different machine learning models.

1. Interquartile Range method (IQR) was used to identify and remove the outliers, as this method is very effective and very accurate in findings the points that falls outside a typical range.For each Numerical features which are defined in the above step, the Q1 which is the first quartile and the Q3 which is the third quartile were calculated. These Q1 and Q3 represent the 25th and 27th percentiles of the data.

2. Then the Interquartile Range method (IQR) was calculated by calculating the difference between Q3 and Q1(IQR- Q3-Q1).Then two outlier Boundaries were calculated to detect and outliers if present in the numerical features. To calculate lower Boundary (Q1 – 1.5* IQR) formulae is used and to calculate upper Boundary (Q2 + 1.5* IQR) formulae is used.

3. Then if any points fall either below lower boundary or above upper boundary were considered as outlier and the points were removed from the dataset.Outliers were removed, and dataset were made free of anomalies which could cause problems later while applying different machine learning models.

3.7 Encoding Categorical Features:

After removing the outliers next step is Encoding the categorical features, as we have seen above, Features like 'Wine Type', 'Yest strain', 'Human Feedback', 'Aging Vessel' were identified as categorical features and remaining were identified as numerical features. This step is important because this type of data is nonnumeric, and this cannot be used by the machine learning models. In this step this was done with the help of One hot encoding.

1. One Hot encoding- In these nonnumeric columns like what we have for categorical variables are converted into a series of binary columns. The working of this is basically assigning a separate binary column for each unique value in a particular categorical variable. For example, in categorical feature 'Wine type', each wine type would be converted into sperate column, and the rows would have binary values which would indicate whether the type is present or not.

2. All the categorical features were encoded successfully using One hot encoding and the target variable was separated from this feature set as this is necessary for the next steps while applying machine learning models.

3.8 Balancing the dataset using SMOTE:

This was the next step after successfully completing the one hot encoding of categorical features. This step is important to make sure that the machine learning models are getting trained on dataset which is having evenly distributed set of samples across all the target classes. The SMOTE (Synthetic Minority Over Sampling Technique) is used to make sure all the classes under target variable 'Refined Wine Quality Rating' are balanced.

1. SMOTE – The main working and technique which is used in Synthetic Minority Over Sampling Technique is that it basically generates a synthetic sample for the minority classes of the variable, this is done to make sure that the number of samples of this minority classes matches the number of samples in the majority classes, this makes a balanced dataset.

2. The SMOTE was applied to both encoded features and target variable, this was done to make sure the dataset is balanced for the further process.

3. After applying this SMOTE, balanced classes were made under target variable where each class got the sample number samples. This step made sure that machine learning models won't be biased towards a particular class while prediction.

3.9 Exploratory Data Analysis (EDA):

The EDA is the very important step before applying different machine learning models on the dataset, as this helps to determine different patterns and relationship between the different features of the dataset. And using this information better accuracy and performance could be achieved. This is the next step which is done after balancing the dataset using SMOTE. There were several steps which were done under EDA, like statistical summaries, Distribution plot, and correlation heatmaps.

3.9.1 Statistical summary of Numerical Features:

1. Statistical summary of Numerical Features- This is the first step that is performed under EDA. This step is important to understand and get the overview of the numerical data within the dataset. Using this statistical summary, central tendencies, dispersion and overall distribution of numerical data within the dataset could be noticed. Also, this step helps to find the different patterns, detecting anomalies and understanding skewness of the data, which helps in modelling decisions.

2. Key findings from the statistical summary of Numerical data showed that, there are 17,559 samples under each feature, which are more than enough for the further data analysis. Then most of the features in the dataset are having broad range of values, which indicates that there is a diversity in in different wine features.

3. Distribution insights showed that most of the features in the dataset have symmetric distribution around there means, which is a good distribution. Some features like aging time, Residual sugar, Free Sulphur dioxide are having significant variability which must be considered in further step, as they might impact the model performance. The target variable showed a balanced mean of 6 and a standard deviation of 2.58, indicating that there is a good variability in it.

3.9.2 Distribution plots:

1. Similarly to above step, this analysis of the distribution plots helps to get the overview of the numerical features within the dataset with help of visual representation. This step as well helps to find the different patterns, detecting anomalies and understanding skewness of the data, which helps in modelling decisions.

2. Key findings showed that the target variable is uniformly distributed across its range, and this shows that the SMOTE balancing was correctly in the previous step.

3.9.3 Box Plots:

1. The Box plots is the next step which is performed after the distribution plots under EDA. As the analysis of box plots helps to visualize the distribution of different numerical features across the different wine quality ratings. This step is crucial to check the data for any central tendency, variability and to check the presence of any outliers.

2. Key Findings showed that features like alcohol percenatge, PH level, Aging time and residual sugar showed significant differences across different wine quality ratings. So, these features must be kept into consideration for the next steps. And these features must be scaled properly before using them into models using Standard Scaler. Featuring selection must be used for different model to check and evaluate their impact on model performance and accuracy. This can be done using techniques such as Recursive Feature Eliminator or feature importance selection based on defined values.

3.9.4 Correlation matrix:

1. The Correlation matrix is the next step which is performed after the Count plots under EDA. As the analysis of Correlation matrix helps to understand the Linear relationship between different numerical features with the help of Visual representation. With help of the correlation matrix, it's easy to identify the potential multicollinearity and different feature interactions.

2. Key Findings showed that most of the features from the dataset exhibit low to moderate correlations with other, which shows that there is not much strong relationship between the variables. With this finding, its evident that each feature contributes to predicting quality uniquely, which is a positive sign of having a diverse dataset. Also, there were few pairs of features which have showed High correlations between each other, pair like Fixed acidity and citric acid, then Residual Sugar and Sugar Content. One of the features can be removed to improve the model performance as this would reduce the redundant information. Some feature like Density and Alcohol showed Negative correlations as well, this inverse relationship must be kept into consideration for feature engineering and model development in next steps. With help of this information from the correlation matrix, feature selection and dimensionality reduction could be optimized in further steps to have better overall performance of the predicting models.

3.9.5 Count Plots:

1. The Count plots is the next step which is performed after the Box plots under EDA. As the analysis of Count plots helps to visualize the distribution of different categorical features across the classes of target variable. This helps to understand the relationship and patterns between the categorical variables and the target variable. This step is crucial because it helps in the feature selection process and enhance the overall performance of the prediction models.

2. Key Findings showed distribution of the wine types varies significantly across the different quality rating of the wine. Same is with the different yeast strains and Aging Vessels. This shows that these features are very important and plays a major role in wine making, and these must be considered while predicting the wine quality using different models. Featuring selection must be used for different model to include these main features like Wine type, yeast strain and Aging Vessel to have better overall performance of the wine predicting models.

3.9.6 Data Splitting:

The Data Splitting is the next step which is performed after the Corelation matrix under EDA. This step is important to have reliable and unbiased machine learning models used for the prediction. Dataset is divided into two set one is the training set 80 percentage and second is test set 20 percentage. Using this training and test set. models could be trained, and their performance could be evaluated on unseen data.

3.10 Model selection:

This step is very Important as selecting the most appropriate machine learning models to predict the wine quality with high overall performance and accuracy Is the main objective. Different criteria like performance metrics, interpretability, scalability and computational efficiency are taken in consideration in the model selection process.

3.10.1 Purpose for selecting model:

As the dataset is having many features advance machine learning are selected to have better overall performance and predicting accuracy. Also, these advance models are very much effective in handling the complex patters within data.

3.10.2 Criteria for selection model:

1. Performance Metrics- Models must be having high accuracy and reliability in their performance.
2. Complexity – The models must be able to handle complex patterns and different relationships within the data.
3. Overfitting Performance- The selected models must be able to prevent overfitting as this helps in maintaining the accuracy.
4. Scalability – The selected models must be able to handle large dataset effectively, without drop in accuracy.
5. Interpretability – The selected models results must be easy to understand up to some extent. As this gives more clarity of the findings. Based on the above purpose and criteria, Random Forest, Gradient Boosting, XGBoost, Support Vector Machine (SVM) and stacking Ensemble advance machine learning model were chosen to predict the wine quality.

3.11 Evaluation metrics:

After successful implementation of the above advance model, the performance of each model is accessed using different evaluation metrics. these evaluation metrics tells how accurate the models are while predicting the wine quality.

Accuracy, Precision, Recall, F1- score, Confusion matrix and classification report have been used as evaluation metrics for each machine learning model.Using these different evaluation metrics helps to evaluate model performance effectively as these metrics helps to find out whether the model predictions are correct and relevant.Metrics like F1-score are important as they help in balancing precision and recall.Metrics like AUC – ROC are important because they help to understand whether a model can differentiate between classes or not.

3.12 Cross Validation:

This is one of the most important steps which helps to evaluates the Robustness, effectiveness and Generalizability of the different machine learning models. In this research

K-fold cross validation is used to evaluate the performance of Random Forest, Gradient Boosting, XGBoost, SVM and stacking Ensemble.

3-fold cross validation is used, that means dataset is divided in three subset and then model is trained and then validated three times. Each time the model used a different subset for the validation purpose and used the remaining two subsets as the training sets.This 3- fold helps in reducing the computational time but does not compromise the overall performance. This 3- fold helps in better hyperparameter tunning when combined with other techniques like Randomized search CV. This combination helps to provide best parameters for the model.

4 Design Specification

This section of the report highlights the different techniques, architecture which have been used for this search. This section is having detailed information, which could be used as a base for other research to replicate the study and to validate the results.

4.1 Techniques and Architecture:

The framework used for this research is Stacking Ensemble Framework. Under this framework different models are combined, which are known as base learners to enhance the overall performance of the Wine quality prediction with high accuracy. The following are the base component of this framework.

4.1.1 Base Learners:

These were the models which were trained separately to predict the wine quality. Then after that as per there performance they were combined and got used as Base Learner models for the stacking Ensemble. These models are Random Forest as this model uses multiple decision trees to improve prediction accuracy and its good in controlling overfitting, then the Gradient Boosting model as this uses a boosting technique which sequentially build model to correct the errors which are made by the previous models, and then XGBoost model as this model is the optimized Gradient Boosting model which is known for avoiding the overfitting.

4.1.2 Meta Learner:

Logistic Regression model has been used as Meta Learner, as this does the integration of the output obtained from different base learners and produces the final output in the form of prediction.

4.2 Code Structure:

This section talks about what all different Modlues and different classes and functions have been used in the code part of the research. Modules such as ‘data.preprocessing.py’, ‘model.tarning.py’, ‘model.evalution.py’ and ‘stacking.ensemble.py’ have been used in the code section.

4.2.1 Classes and Function :

Many classes and functions have been used in the code of this Project, as these were essential for proper execution of the code. Examples of these Classes and Functions are Data-PreProcessor, Random-ForestModel, Gradient-BoostingModel, XGBoost-Model, Meta-Learner and also Stacking-Ensemble.

4.2.2 Frameworks and Libraries:

This section talks about what all different Frameworks and Libraries have been used in the code part of the research.

1. Sckit-learn(version 0.24.1): This python Library has been used for Data Processing, Model Training and for Evaluation.
2. Pandas(version 1.2.3) – This python Library has been used for Data Manipulation and Data Analysis.
3. Numpy(version 1.19.5) - This python Library has been used for Numerical computations.
4. Matplotlib(3.3.4) – This python Library has been used for Data Visualization.
5. Seaborn(0.11.1) – This python Library has been used for Data Visualization as well.
6. Xgboost(1.3.3) – This python Library has been used for the implementation of XGBoost model.

4.2.3 Reason behind the selection:

The stacking ensemble framework was used because it is very powerful in prediction task, as this combines the strengths of different machine learning models. This performance of the stacking ensemble methods is highest in this search, this is also seen in other studies done by Zhou et al. (2023), showed that using the stacking ensemble models, best overall performance of the models could be obtained.

4.3 Requirements:

These are the requirements to implement of the research Project successfully. As these requirements make sure that the selected models could be trained and evaluated successfully.

1. Hardware - 13th Generation Intel Core i5 processor, 16GB RAM and 500GB SSD.
2. Software – Windows 11 Home Single Language version(23H2), Programming language python, Jupyter Notebook.
3. Dependencies – scikit learn (version 0.24.1), pandas (version 1.2.3), numpy (version 1.19.5), matplotlib (version 3.3.4), seaborn (version 0.11.1) and xgboost (version 1.3.3).
4. Environmental setup – A virtual environment must be created using ‘venv’ or ‘conda’, then required libraries must be installed using ‘pip install -r filename.txt’.

4.3.1 Architectural Diagram:

Below Diagram Shows the flow of data through Different systems. This diagram shows different step of the research, and how these steps are co related to each other to get the final output as wine prediction with high accuracy.

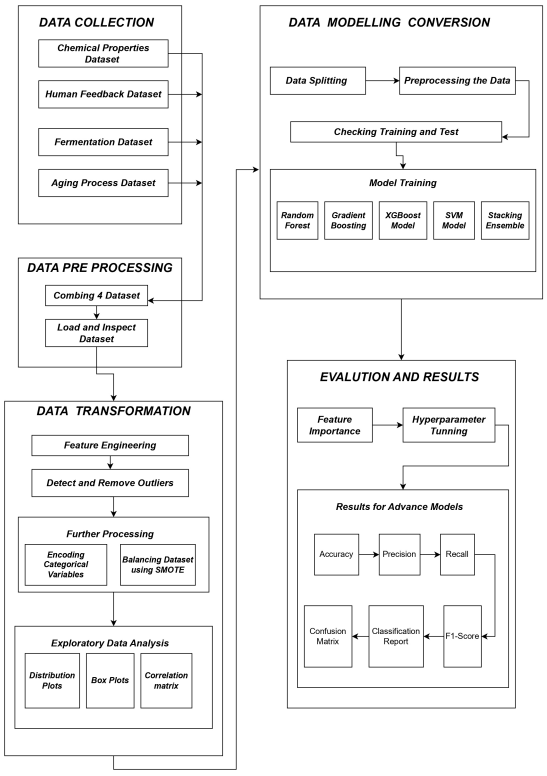

Figure 2: Architectural Diagram

4.4 Algorithm/ Model Description:

The stacking ensemble model was designed in such a way that it was able to use the strengths of different base learners together to predict the wine quality with high accuracy. Below are the steps which were taken into consideration while implementing the stacking ensemble.

1. Data Processing – Data was processed to make sure that the datapoints were balanced, all the missing values were manged with forward filling, all the outliers were removed.

2. Base Learners - Random Forest model was trained with (100, 200, 300, 400, 500) as estimators and with (None, 10, 20, 30, 40, 50) as maximum depth.

3. Gradient Boosting model was trained with (100, 200, 300, 400, 500) as estimators, with different learning rate of (0.01, 0.05, 0.1, 0.2) and with (3, 4, 5, 6, 7) as maximum depth.

4. XGBoost model was trained with (100, 200, 500) as estimators, with different learning rate of (0.01, 0.05, 0.1) and maximum depth as (3, 5, 7, 10).
5. Meta Learner – Logistic regression model was used as meta learner, using the prediction of above base learners model combined to get the final prediction.
6. Training Procedure- Then the dataset was spilt into (80%) as training set and (20%) as testing set.
7. Cross Validation technique with 3 Folds were used for the hyperparameter Tunning.
8. Evaluation Metrics- Different Metrics like accuracy, Precision, recall and F1 score were used to evaluate the performance of different models used.

5 Implementation

The section of the report is Implementation, this section describes the output produced after implementation of the code successfully. Also, this section gives details about how the models were implemented.

5.1 Transformed Data:

The data was transformed into its final stage with the help of different techniques. First the missing values were handled using forward filling. Then Outliers were removed using interquartile range (IQR). Then the Categorical features with the combined dataset were encoded using one hot encoding and then Synthetic Minority Oversampling Technique (SMOTE) is used to balance the dataset.

5.2 Code Written:

Code was written in python programming language with using different Classes, Function, Frameworks and Libraries as described in the Design specification section in detail.

5.3 Models developed :

Different Models like random Forest, Gradient Boosting, XGBoost and SVM were developed with different numbers of estimators, learning rates and max depth as described in the Design specification section in detail.

5.4 Model Evaluation:

Different metrics like Accuracy, Precision, Recall and F1 -score were used to evaluate the performance of each model as described in the Design specification section in detail.

5.5 Description of the Process:

1. Data Collection- Different Dataset were collected from different sites for different feature of the wine, later combined to get one dataset.Then under Data Processing all the necessary steps under pre-processing were done, to make sure data is ready for the next step. Data got split into training (80%) set and test (20%) set.

2. Then after this Models like Random Forest, Gradient Boosting, XGBoost and SVM were trained on dataset with all the features set. And each model's performance was evaluated with mentioned evaluation metrics. and then the important features were selected for each models separately. And after this again Models like Random Forest, Gradient Boosting, XGBoost and SVM were trained on dataset with reduced features set. And again, each model's performance was evaluated with mentioned evaluation metrics.

3. As per the performance, each model was implemented again with Hyperparameter tunning with different parameter like estimators, cross validation and different learning rates as described in design specification section in detail. For example, if the performance of a model was better with full features set when compared to its performance on reduced features set, the model was implemented again with hyperparameter tunning with full features set. And then the performance was again calculated with mentioned evaluation metrics.

6 Results and Evaluation

All the models had been successfully executed and results were obtained with help of different evaluation metrics. As these evaluation metrics not only talk about, how well the models have performed but also help to compare the performance of different models. Below are the results of Predicting wine quality with accuracy based on Physicochemical properties, Sensory feedback by human, different Fermentation features like Duration, temperature and different Aging features of the wine together with help of different machine learning techniques like Random Forest, Gradient Boosting, XGBoost and Support Vector Machine, Stacking Ensemble.

6.1 Analysis of the results:

Stacking ensemble and Random Forest Model provided the best results in Wine quality assessment with high accuracy and result. This means that these models were able to integrate different features like Physicochemical properties, sensory feedback by human, different Fermentation features like Duration, temperature and different Aging features of the wine very effectively. Also, these models provided a good prediction of wine quality, which highlights their robust predictions capabilities.The stacking Ensemble models is the most effective one for wine quality prediction. Also, other models like Random Forest and XGBoost are also good option for the wine quality prediction. These models could be used on more diverse dataset to getter better insights in this wine quality prediction, which would be beneficial for the wine industry.

6.2 Comparing the findings with previous studies used in literature review:

Below is the table which Compares the results of this research with the results of the previous research which has been discussed in Literature review.This table includes columns like different Models used in different studies and then column name as best Model which shows the best model among the different models used within that study. after this there

are columns which have different key metrics used to evaluate the performance of the best model and then the next column is for results.And final column in Remarks.

Study	**Models Used**	**Best Model**	**Key Metrics**	**Results**	**Remarks**
Predicting Wine Quality (Current Research)	Random Forest, Gradient Boosting, XGBoost, SVM, Stacking Ensemble	Stacking Ensemble	Accuracy	88.47%	Stacking Esemble got Highest accuracy among all the models Used.
			Precision	88%	
			Recall	87%	
			F1-Score	87.5%	
Jiang et al. (2023)	Random Forest, Logistic Regression	Random Forest	Accuracy	90%	Accuracy is more by 2.9%, as compared to what has been obtanied in this research for Random forest(Tunned).
			Precision	89%	
			Recall	88%	
			F1-Score	88.5%	
K (2023)	Random Forest, XGBoost	Random Forest	Accuracy	89.67%	Accuracy is more by 2.6%, as compared to what has been obtanied in this research for Random forest(Tunned).
			Precision	89%	
			Recall	88.5%	
			F1-Score	88.8%	
Liu (2023)	Decision Tree, Random Forest, SVM	SVM	Accuracy	92.25%	Accuracy is more by 19.25%, as compared to what has been obtanied in this research for Random SVM (Tunned).
			Precision	91%	
Basha et al. (2023)	Gradient Boosting, Random Forest, SVC, AdaBoost	Gradient Boosting	Accuracy	92%	Accuracy is more by 5.9%, as compared to what has been obtanied in this research for Random Gradient Boosting(Tunned).
			Precision	91%	
			Recall	90.5%	
			F1-Score	90.75%	
Zhou et al. (2023)	Improved Stacking Ensemble Learning	Stacking Ensemble	Accuracy	92%	Accuracy is more by 3.53%, as compared to what has been obtanied in this research for Stacking Ensemble(Tunned).
			Precision	91%	
Irmalasari and Dwiyanti (2023)	Decision Tree, Gradient Boosting, Random Forest	Random Forest	Accuracy	86%	Accuracy is less by 1.1%, as compared to what has been obtanied in this research for Random forest(Tunned).
			Precision	85%	

Table 2: Comparative Results from Various Studies on Wine Quality Prediction

6.3 Result Bar Charts:

The 2 Bar chart below shows Results of the different models used.The 1st bar chart shows different model performance using evaluation metrics like Accuracy, Precision, Recall and F1-score, with and without Hyperparameter Tunning on full feature sets. and then 2nd bar chart shows the accuracy achieved by each model with and without hyperparamter tunning.

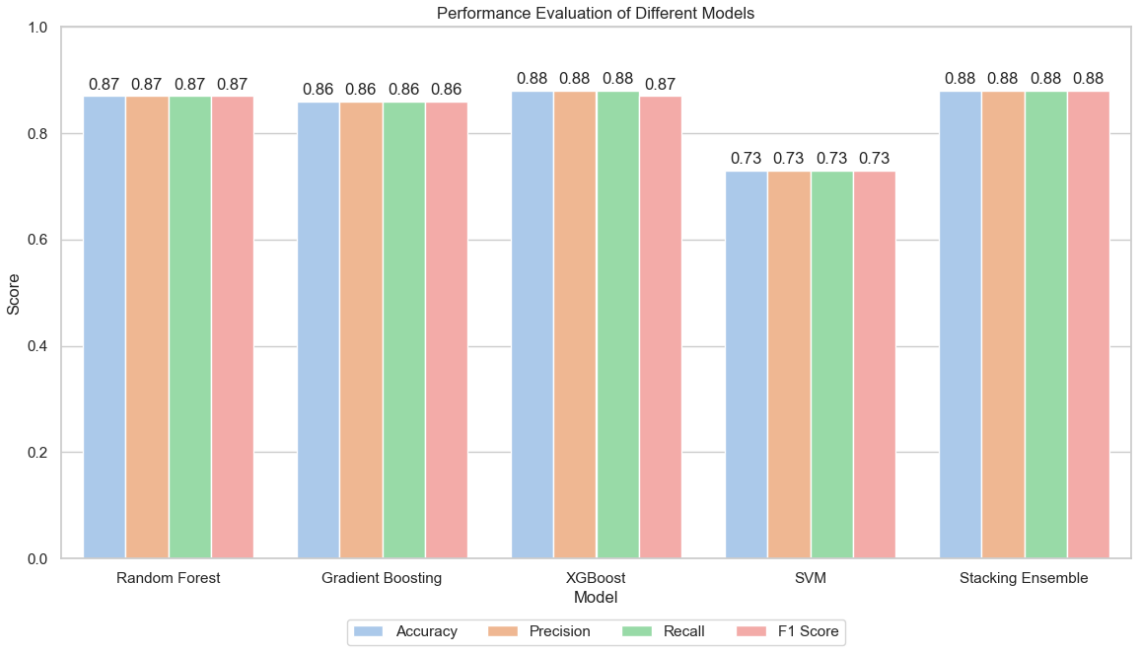

Figure 3: Performace evalution of different Models

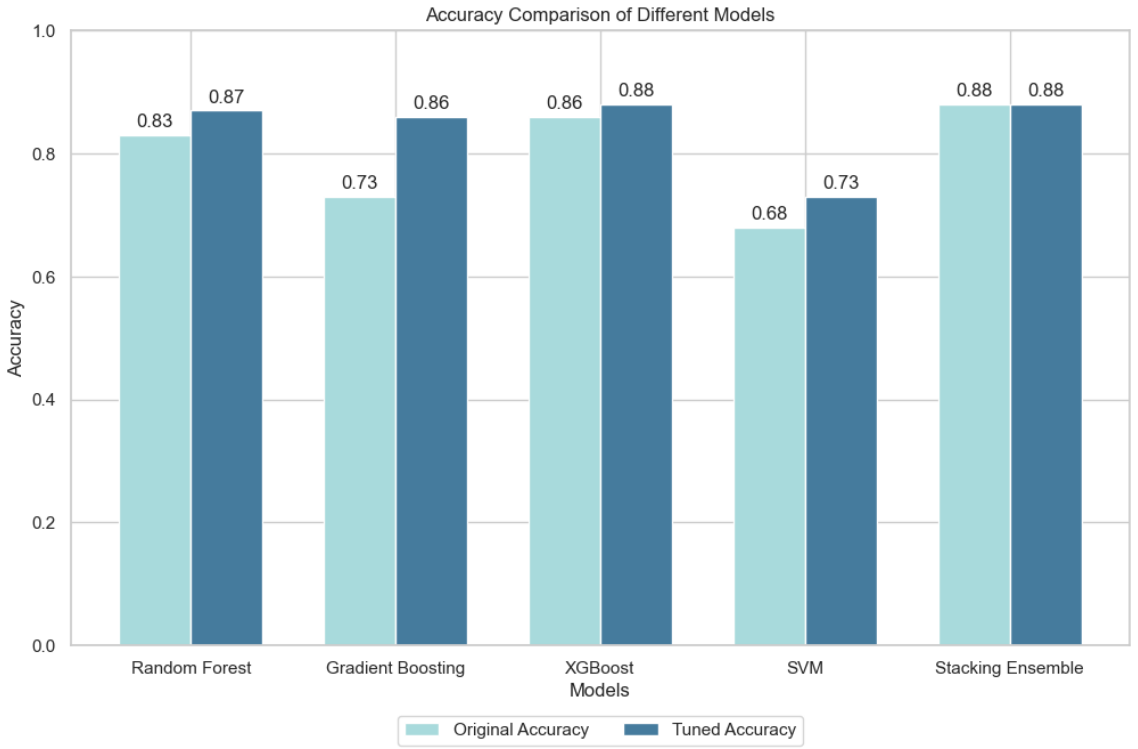

Figure 4: Accuracy Comparison of diiferent Models

7 Conclusion and Future Work

This section of the report discusses about the results interpretation, and then what are their implications. Also, there is detailed discussion about the limitations of the search, and then about the potential Future work, which could be done using the methodology

and findings of this research as base.

Results obtained after implementation of different machine learning models on the dataset showed very good accuracy in the wine quality prediction. Stacking Ensemble model showed the highest accuracy of 88.47%, followed by tunned random forest model showed which showed accuracy of 87.10% and then XGBoost showed accuracy 87.61%. This shows that these selected models are good and effective in predicting task, as this is discussed in literature review section as well in detail. This research mainly explored the different features like Physicochemical properties, sensory feedback by human, different Fermentation features like Duration, temperature and different Aging features to predict Wine Quality. This could be achieved using different machine learning models with and without hyperparameter Tunning. The results are promising as compared with the results of different research papers, but more features must be considered in addition to these features to predict the wine quality using these advance machine learning models.The results and findings of this search showed a Robust performance of the Machine Learning models, especially the stacking Ensemble model which got high accuracy and results. Then the approach of combining different features of wine together to predict the wine quality, showed that this approach is more real and could become the game changer in the wine industry. Also, this search proves that the Hyperparameter tunning have the major role in improving the model overall performance.More features must be considered with these features, so that the models could be used on more complex dataset. These advance models took lot of time to execute, which tell how complex it is to run these models like stacking on huge dataset. Also, interpretating the results these models is not easy as compared to base models like Decision tree and Logistic Regression.

7.1 Conclusion:

The main objective of this research was Predict the wine quality with high accuracy based on Physicochemical properties, Sensory feedback by human, different Fermentation features like Duration, temperature and different Aging features of the wine together. This objective is achieved as many machine learning models have been used, and best model have been found out which is stacking ensemble. This shows that this model is the most effective one in wine quality assessment.

7.2 Potential Future Work:

To enhance the generalizability of the findings and results of this research, these Advance machine learning models must be used and testes on big dataset from different region.More Complex machine learning models could be used, like Voting ensemble to get more overall accuracy and precision.Investigation could be done to find out the different applications of these models in the wine industry, these could be either in quality control or product overall development. This would provide benefit for this wine industry significantly.All this findings and future potential work could lead to more realistic research, which could be useful for the wine industry and eventually would enhance the overall performance it.

References

Abarca-Rivas, C., Martín-Garcia, A., Riu-Aumatell, M., Bidon-Chanal, A. and López-Tamames, E. (2023). Effect of fermentation temperature on oenological parameters and volatile compounds in wine, *BIO Web of Conferences*, Vol. 56, EDP Sciences, p. 02034.

Armstrong, C. E., Niimi, J., Boss, P. K., Pagay, V. and Jeffery, D. W. (2023). Use of machine learning with fused spectral data for prediction of product sensory characteristics: The case of grape to wine, *Foods* **12**(4): 757.

Basha, M. S. A., Desai, K., Christina, S., Sucharitha, M. M. and Maheshwari, A. (2023). Enhancing red wine quality prediction through machine learning approaches with hyperparameters optimization technique, *2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT)*, IEEE, pp. 1–8.

Denchai, S., Sasomsin, S., Prakitchaiwattana, C., Phuenpong, T., Homyog, K., Mekboonsonglarp, W. and Settachaimongkon, S. (2023). Influence of different types, utilization times, and volumes of aging barrels on the metabolite profile of red wine revealed by 1h-nmr metabolomics approach, *Molecules* **28**(18): 6716.

Di, S. and Yang, Y. (2022). Prediction of red wine quality using one-dimensional convolutional neural networks, *arXiv preprint arXiv:2208.14008* .

Godillot, J., Baconin, C., Sanchez, I., Baragatti, M., Perez, M., Sire, Y., Aguera, E., Sablayrolles, J.-M., Farines, V. and Mouret, J.-R. (2023). Analysis of volatile compounds production kinetics: A study of the impact of nitrogen addition and temperature during alcoholic fermentation, *Frontiers in Microbiology* **14**: 1124970.

Irmalasari, I. and Dwiyanti, L. (2023). Algorithm analysis of decision tree, gradient boosting decision tree, and random forest for classification (case study: West java house of representatives election 2019), *2023 International Conference on Electrical Engineering and Informatics (ICEEI)*, IEEE, pp. 1–5.

Jiang, X., Liu, X., Wu, Y. and Yang, D. (2023). White wine quality prediction and analysis with machine learning techniques, *Highlights in Science, Engineering and Technology* **39**: 321–326.

K, A. (2023). Regression modeling approaches for red wine quality prediction: Individual and ensemble, *International Journal for Research in Applied Science and Engineering Technology (IJRASET)* **11**: 123–130.

Liu, Z. (2023). Comparison of the red wine quality prediction accuracy using 5 machine learning model, *Highlights in Science, Engineering and Technology* **60**: 114–120.

MN, R., KT, R., Nidoni, U., Hiregoudar, S. and Naik, N. (2023). Effect of yeast concentration on quality parameters of ber (ziziphus mauritiana) fruit (cv. umran) wine during ageing, *International Journal of Plant & Soil Science* **35**(22): 28–40.

Mor, N. S., Asras, T., Gal, E., Demasia, T., Tarab, E., Ezekiel, N., Nikapros, O., Semimufar, O., Gladky, E., Karpenko, M. et al. (2022). Wine quality and type prediction from physicochemical properties using neural networks for machine learning: a free software for winemakers and customers., *AgriRxiv* (2022): 20220051475.

Pascua, K. B., Lagura, H. D., Lumacad, G. S., Pensona, A. K. N. and Jalop, M. J. I. (2023). Combined synthetic minority oversampling technique and deep neural network for red wine quality prediction, *2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT)*, IEEE, pp. 609–614.

Yao, R. and Jia, G. (2023). Machine learning aided inverse design for new wine, *2023 4th International Conference on Industrial Engineering and Artificial Intelligence (IEAI)*, IEEE, pp. 51–61.

Zhou, M., Yu, W. and Jiang, K. (2023). Wine quality detection based on improved stacking ensemble learning, *2023 8th International Conference on Information Systems Engineering (ICISE)*, IEEE, pp. 226–229.