National College of Ireland

Wine Quality Prediction using Machine Learning and Hybrid Modeling

MSc Research Project
Msc in Data Analytics

Avinash Sanjay Gawale
Student ID: x20247303

School of Computing
National College of Ireland

Supervisor: Jorge Basilio

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Avinash Sanjay Gawale
Student ID:	x20247303
Programme:	Msc in Data Analytics
Year:	2022
Module:	MSc Research Project
Supervisor:	Jorge Basilio
Submission Due Date:	15/08/2022
Project Title:	Wine Quality Prediction using Machine Learning and Hybrid Modeling
Word Count:	6475
Page Count:	20

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:	
Date:	19th September 2022

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):	

Wine Quality Prediction using Machine Learning and Hybrid Modeling

Avinash Sanjay Gawale
x20247303

Abstract

Globally, there has been an upsurge in wine consumption. 31 million metric tons of wine are delivered globally, which is a significant quantity. Considering how extremely competitive the wine market is, the wine industry is investing in innovative technologies for both wine-producing and selling processes. Technology has made it possible for businesses to provide consumers with high-quality wine by introducing machine learning and hybrid modeling techniques for wine quality prediction. The study is being carried out to implement the Decision Tree (DT), Random Forest (RF) , and Extreme Gradient Boosting (XGBoost) in wine quality prediction and to identify machine learning techniques' role as hybrid models in wine quality prediction. The dataset for wine quality is available publicly on the repository of UCI machine learning and dataset from said database has been used in the study. Data interpretation has been performed based on accuracy, precision, recall and f1 Score. A comparison of developed models carried out. The models tested include Decision Tree Classifier, Random Forest Classifier, XGboost Classifier and a Hybrid Model. The results indicate that the most accurate and precise model is that of Random Forest with the highest accuracy, precision, recall and f1 score.

Keywords— Wine Prediction, Decision Tree, Random Forest, XGBoost, Hybrid Model

1 Introduction

The production of wine due to excessive consumption has increased in modern times. In most parts of the world wine consumption is a trend due to social and normative factors Kumar et al. (2020). The increased production of wine has been the key factor behind the huge market competition. Most wine producing companies are facing enormous challenges to justify their wine quality. To maintain wine quality, companies are bound to focus on wine cortification to assess wine quality within the market Cardoso Schwindt et al. (2022). The presence of the wine cortification process has allowed companies to improve wine quality post-assessment. The wine business and wine production in the industry massively rely on wine certification. Wine certification is aligned with wine quality. Therefore, wine quality has been one major issue because wine quality is determined by quality experts regardless of deep understanding. Bhardwaj et al. (2022). The stress on wine quality indicates its importance in wine production and wine consumption. It is directly aligned with the health percussions of consumers. Georgieva and Rocha (n.d.) argued that wine quality prediction has uplifted wine production, and consumers are certain about the quality of wine consumption in modern times. In both international and domestic markets, wine consumption increased because of quality wine and high competition.

This indicates that wine consumption is mainly related to its quality instead of taste, seeds or grape variety Zhang et al. (2020).

Wine quality assessment and prediction is a challenging process Lukić et al. (2020). To ensure quality wine production, wine quality prediction becomes an essential factor of consideration. In the conventional ways, wine quality prediction mostly relied on simple mechanisms i.e., post-production quality checks. Traditionally, there was no technical or advanced mechanism such as technology with wine production companies to assess and predict wine quality Aich et al. (2018). In the wine industry, the assessment and check of wine quality improvement and prediction accuracy ratio are crucial Liu (2021). Keeping this into consideration, it is pertinent to discuss that wine quality prediction justifies wine quality post-production. In both pre and post-assessment, the induction of technology is crucial for an efficient quality prediction approach. Both physical and chemical features of wine production must be assessed through quality prediction tools and techniques.

The induction of technology has brought machine learning, deep learning, and hybrid techniques to predict wine quality and increase the accuracy ratio. Wine quality prediction is using machine learning and deep learning techniques for accurate results Gupta et al. (2020). The use of both machine learning and deep learning techniques has revolutionized the wine quality prediction process. Wine-producing companies have shifted towards the use of ML, DL and hybrid learning techniques to overcome the quality assessment challenge. In comparison with manual quality prediction, ML and DL learning techniques has a high-efficiency ratio Bhardwaj et al. (2022).

Keeping this into consideration, the study has focused on the precision of wine quality prediction through machine learning and hybrid modeling techniques. The use of machine learning and hybrid modelling techniques is the focal point of the entire study. Here, Decision Tree Classifier (DTC), Random Forest Classifier (RFC) and Extreme Gradient Boosting (XGBoost) machine learning techniques are focused. In the Hybrid Model all these three machine learning techniques are combined to predict the wine quality.

1.1 Background and Motivation

According to Bhardwaj et al. (2022) wine quality is an essential factor in the consumption and production process in New Zealand. The production of Pinot noir wines is globally accepted therefore, pinot noir wine quality prediction is crucial. Pinot noir is a complex grape for wine production. For this, adaptive boosting, and random forest as key classifiers were used. They used synthetic data to form a machine learning model in New Zealand. All the work was done as a classifier. Liu (2021) has used the gradient boosting method to check and evaluate wine quality by focusing on different parameters in Taiyuan, China. His study formed different datasets of red and white wine to target different values to increase quality wine consumption. According to Yang et al. (2022), the wine rating system is highly relevant to wine consumers. Most consumers in the world focus on a wine rating system which is 100-rating. Consumers refuse to purchase wines rating below 80 out of 100 ratings. Therefore, companies are using machine learning techniques to earn ratings above 80.

Georgieva and Rocha (n.d.) argued that global economic impact has challenged wine's constant growth every year. In both international and domestic markets, wine consumption has been reduced due to health issues. Based on the Viniportugal wine report, the value of wine in 2019 was 820 million euros, in which, domestic counting was around 44.5%. In this regard, wine quality was a major boosting factor in Portugal. Therefore, it is pertinent to argue that wine quality prediction is crucial for higher industrial growth. Additionally, Sirivanth et al. (2021) argued that wine quality prediction has used both machine learning and deep learning algorithms that are generated over years. The use of machine learning and deep learning algorithms has provided an effective communication podium to bring interaction between the abstract wine

environment and its compounds Trivedi and Sehrawat (2018). The use of machine learning and deep learning techniques has an excellent role in wine quality assessment and forecasting. Dahal et al. (2021) added that the success of machine learning and deep learning in different sectors including businesses, pharmacy, astrophysics etc. has brought its use in the wine industry. Supported by Gupta (2018).

Based on the motivational study background, the study has focused on the induction of machine learning techniques to understand its precise impact on wine quality prediction in the wine industry.

1.2 Research Question

The research question of this study is "How effectively a hybrid machine learning model and machine learning can predict a wine's quality?"

1.3 Research Objectives

Following are key research objectives.

- Conduct research using combined dataset of red and white wine.
- To implement the Random Forest (RF), Decision Tree, and XGBoost in wine quality prediction.
- To identify machine learning techniques' role as Hybrid model in wine quality prediction.

1.4 Research Outlines

In the research outlines following information has been provided to the reader.

- In the 1st section, the study has discussed a basic introduction of the study topic followed by motivation and background. Research objectives and research question has been added to this section.
- In the 2nd section, a critical review of previous related work has been explained. All the related work has been critically reviewed to identify relevant study gaps.
- In the 3rd section, the study methodology and overall research design have been explained with proper justification.
- In the 4th section, design specifications and algorithms are explained.
- In the 5th section, the implementation and evaluation of applied techniques are explained.
- The 6th section is based on result comparison to justify the effectiveness of machine and deep learning techniques in wine quality prediction
- In the 7th section, the conclusion and discussion of the conclusion have been explained.

2 Related Work

In the related work section, all related work with machine learning and deep learning's role in wine quality prediction has been critically reviewed. For the critical review, the study has targeted previously done work to justify the validity of the study. The related work has been divided into three each section i.e., machine learning, deep learning, and hybrid modelling.

2.1 Wine Quality Prediction using Machine Learning

Canizo et al. (2019) focused on wine origin by taking grape skin samples from different countries. In analysis, the study relied on MLR, SVM, K-NN and RF techniques to find grape origin using grape skin. Grape skin samples result from accuracy is higher in only SVM and RF models because it classifies through certain parameters. The strength of this study was that it used ICP-MS techniques by inducting 29 elements in the testing process. In this regard, the study results in accuracy can be high through ML techniques. However, one issue is that ICP-MS can have higher costs during implementation. Also, using the ICP-MS technique can mostly rely on heavy elements, the lighter elements are ignored which leads to interference in the process. Below is a graphical representation of the study process.

Furthermore, Kumar et al. (2020) study mainly red wine quality prediction using ML techniques such as support vector machines and Naïve Bayes algorithms. Their study used red wine datasets to test and predict wine quality. To evaluate, the use of ML learning techniques such as SVM mainly focused on red wine, therefore, it can be inaccurate to generalize overall results on both red wine and white wine. Similarly, Naïve Bayes mainly focuses on speculations, the algorithms used are speculated and speculations can be wrong. Additionally, a key limitation of this algorithm is that it requires certain parameters. To further analyze, SVM can be a weak technique because it cannot provide accurate results in large datasets. The study used only red wine which imbalanced datasets. The SVM can have poor results in imbalanced datasets which is an issue. Similarly, one both probability levels of 0.7 and 0.3, they used red wine, therefore, the imbalance in results can be high in the absence of white wine.

Liu (2021) study used gradient boosting machine learning technique to predict wine quality. Their study used classifiers as red and white samples to identify its outliners. To critically evaluate, the use of gradient boosting as a classifier can be an effective technique to predict wine quality. However, gradient boosting can have several limitations. One major limitation of gradient boosting is that it relays on different parameters to classify datasets. Also, the results in GB machine learning can be highly sensitive because it relies on outliers to generate classifiers. Secondly, the gradient boosting method in machine learning can be difficult to scale up. In this regard, wine prediction results can be a challenge.

Sirivanth et al. (2021) study used ML and AL algorithms to predict wine quality. Their study has used correct mines space to provide access to the entire process and makes the process more cost-efficient and more valid. To evaluate their approach toward wine quality prediction, they focused on different aspects by electing them to justify their role in the process. All these aspects were used through ML learning techniques to classify information related to wine quality. Moreover, their study mainly relied on RF algorithms to use information and project score. Similarly, Dahal et al. (2021) study used RR, support vector machine, gradient boosting and ANNs for wine quality prediction. To evaluate their work, they have used publically available wine quality datasets. Moreover, even their study used SVM and GB and RR as ML classifiers, however, among two key categories of red and white wine, they only used red wine data due to its wide acceptance over the white wine group. Therefore, its results can be less reliable because it ignored the results of the white wine group. Also, most of their work supported the use of ML techniques such as gradient boosting to have higher results than ANNs due to their reliability on parameters. Supported by Gupta and Vanmathi (n.d.).

Similarly, Oreški et al. (2021) study mostly used technology for automation and innovation of agricultural processes. To evaluate their study, they used IoT and other machine learning techniques for wine quality prediction. Their study also used decision trees and SVM as major classification factors to predict wine quality. However, their study directly focused on public datasets. The issue with public datasets is that they are large in number. The classification of data becomes highly challenging. Although, the use of machine learning techniques such as decision trees has effectively classified red and white wines. This was possible due to the predictive model intelligence system. Moreover, Caissie et al. (2021) study formulated an in-

tegrated framework for wine quality assessment. To evaluate, the study is unique because it used global testing and unimodal testing that includes human senses and bimodal testing that integrates all the senses with each other. On the other hand, during the testing, their study also used psychological testing predictors to integrate all the senses with these predictors. Most of their work relied on senses and sensory-based predictions of wine quality. To further argue, their study used a highly complex and uncertain method because senses are subjective factors. In both unimodal and bimodal senses, the predictors and sensory features can vary from person to person. Focus on the wine seeds, plants, and grapes can have high prediction accuracy than sensory predictors.

Bhardwaj et al. (2022) study focused on wine quality prediction targeting pinot noir grape. Their study used synthetic data to form a machine learning model, the data has been collected from different regions in New Zealand. A total of 18 pinot noir samples along both physiochemical and chemical features. To critically evaluate Bhardwaj et al. (2022) study, it is pertinent to mention their study has only targeted Pinot noir samples. The data and results generated were all related to Pinot Noir wine grapes. Also, the 7 physiochemical and 47 chemical features were also related to Pinot noir grapes. In this regard, it is too early to say that wine quality prediction can generate effective results through machine and deep learning techniques. Further, their study used SMOTE method to generate 1381 samples. Using SMOTE method can have multiple disadvantages. One primary disadvantage of SMOTE method is that oversamples unnecessary information within the samples. Therefore, it raises questions about the validity of the results. Secondly, using SMOTE method can oversample noisy samples in the sample space. In the SMOTE method, it is relatively problematic to manage results. To further evaluate, their study used classifiers such as random forest and gradient booster to classify and evaluate information related to pinot noir wine. However, it cannot be generalised that both machine learning techniques can accurately predict wine quality. Below is the descriptive analysis of their synthetic data that shows oversampled data.

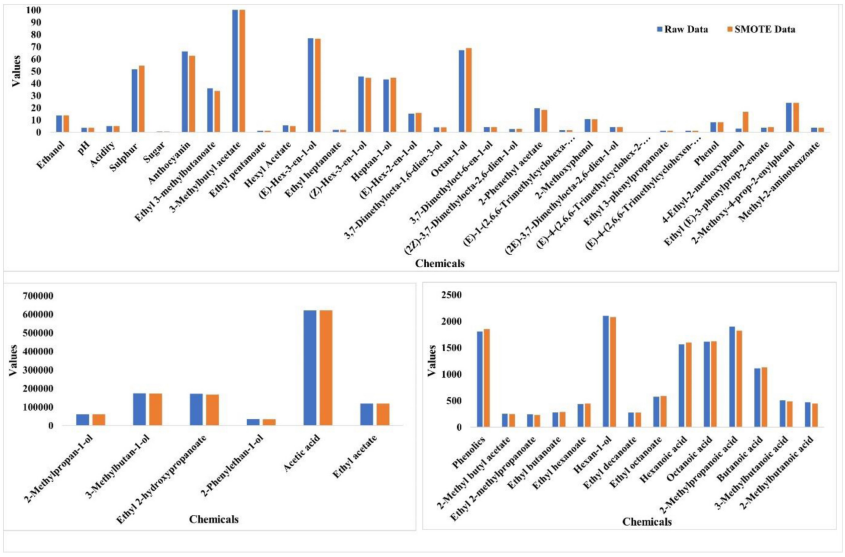

Figure 1: Main Comparison of Raw Data and SMOTE Data Bhardwaj et al. (2022)

Furthermore, Georgieva and Rocha (n.d.)) study used a decision tree classifier, support vector machine and neural network to predict wine quality. Their study has also used a hidden Markov chain to identify and categorize all the classified datasets to bring accuracy to the work. In analysis, their study has drawn multiple conclusions to predict wine quality. The key issue with their study is that they have used numeric results and the approach is the black box method. Using the black box method or approach can be tricky. Sometimes, it can be

problematic to interpret numerical data that can provide better predictions of datasets. To further evaluate, their study has compared both machine and deep learning techniques in wine quality prediction. For instance, they concluded that the Decision tree classifier has generated more effective results than Neural Networks because it classifies information. However, they ignored that these classifiers also rely on different parameters while neural networks are related to different nodes. Overall, the study results are acceptable considering the study objective i.e. machine learning role in wine prediction. The below table shows the comparison between results using both ML and DL techniques to predict wine quality.

2.2 Wine Quality Prediction using Hybrid Modeling

Wine quality predictions are also done through hybrid modelling. The use of both machine learning and deep learning technology/algorithms in hybrid modelling has increased in modern times. For instance, Policastro et al. (2007) worked on a hybrid case-based system to monitor and predict wine quality. To evaluate, they used SVM algorithms as classifiers to manage all data. The issue is that they need to have a single algorithm at each base. It can have implications in terms of numerical data. Below is an example of how they used hybrid case-based system automation to predict wine quality.

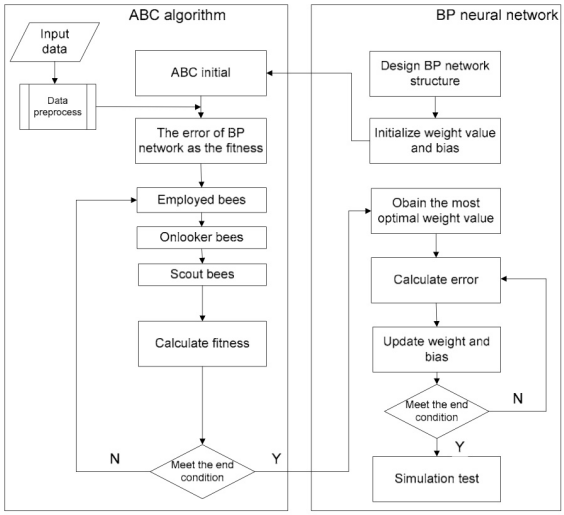

Figure 2: ABC Flowchart Based on BP Network Qiongshuai and Shiqing (2011)

Qiongshuai and Shiqing (2011) study used Artificial Bee Colony Algorithm (ABC) to create a hybrid model for better wine prediction by focusing on weightage. To evaluate, the BP neural network is a multilayer neural network. In most cases, it focuses on error backpropagation, however, the issue with the study is that they did not discuss the input data. In BP neural network formation, data input is crucial.

2.3 Research Gap

The research gap is based on different points that are missing in the previously reviewed literature. These points are the reason for this study to achieve its objectives.

Machine learning techniques used in different studies mostly focused on wine quality prediction by targeting different features of wines such as origin, adore and variety. Wine consumption and the quality of wine consumption are interconnected. The studies have ignored the major role of consumer expertise. From the year 2012, none of any studies has used hybrid modeling

techniques to predict and assess wine quality. The study sees this as the major gap to achieve its objectives.

From the literature it is evident that prior research has not been carried out by combining both the red and white wine datasets. In some research works where both the datasets of red and white wine have been used, the model implementation has been performed separately over the both datasets. Therefore, our study covers this gap. Moreover, hybrid model of the machine learning techniques has not been carried out before and only deep learning algorithms' hybrid models have been made.

3 Methodology

There are many data sources in the world, which creates a huge amount of data available for use, however, there is a need to extract meaningful information from such unorganized and raw data. Raw data is of no use if it cannot be utilized to create meaningful insights. Therefore, data scientists are always on the run to create many possible ways in order to process raw data into such information sets which can be used to take informed decisions and bring the accuracy of processes towards the maximum possible extent Barnaghi et al. (2013).

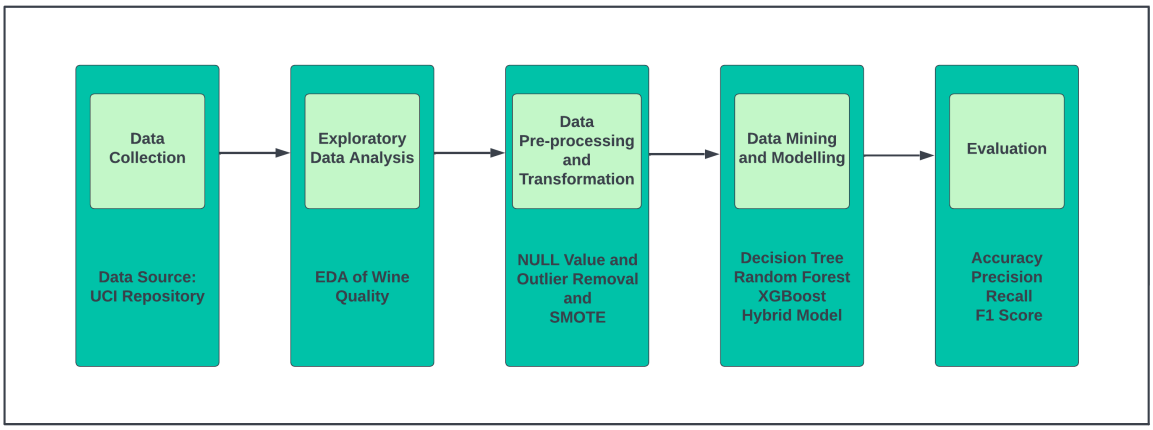

Figure 3: Knowledge Discovery in Databases (KDD)

3.1 Data Collection

The dataset for wine quality is available publicly on the repository of UCI machine learning Cortez et al. (2009). There are two files in the dataset which has data for variants of white wine as well as red wine of the "Vinho Verde" which is a specific wine of Portuguese origin. A large collection of datasets is present in this data based which was collected by the machine learning community so that research studies can be conducted on it. There are 1599 instances in the red wine dataset and 4898 instances in the white wine data sets Sirivanth et al. (2021). 11 input features have been taken into account and there is only one output feature put in place. Physiochemical tests were taken as the basis for collection of input data features and sensory data was the basis for output features which was scaled in an 11 step quality features thus varying from 0 as very bad to 10 as very good. All the other values lied between the two points, and they were taken as the extreme values; whereas very bad was the extremely negative value and very good was taken as the extremely positive value of the data collection by the community.

Furthermore, the databases as taken as the classification of regression. The classes inside the datasets are taken in a manner which is not balanced and order less. There are many other

values than normal which fall inside the very good quality and very bad quality wine. It is suggested that algorithms which can take out or eliminate outliers should be run on the data so that no outliers and left and the data can generate meaningful insights rather than misleading information Aurit et al. (2021). The community has collected this data without considering the relevance of the input features, therefore, it is also suggested for future researchers to test feature selection methods on the datasets and observe their behavior to such tests. The two datasets have been combined for this research study and some random values have been removed in order to increase credibility of the findings of the research.

Following are the physiochemical input features

1. Fixed Acidity of the wine
2. Volatile Acidity of the wine
3. Citric Acid Content
4. Residual Sugar
5. Presence of Chlorides
6. Total Sulphur Oxides
7. Free Sulphur Oxides
8. Density of the liquid
9. pH
10. Alcohol
11. Sulphates

Following is the sensory output feature:

1. Quality (0 to 10)

3.2 Exploratory Data Analysis

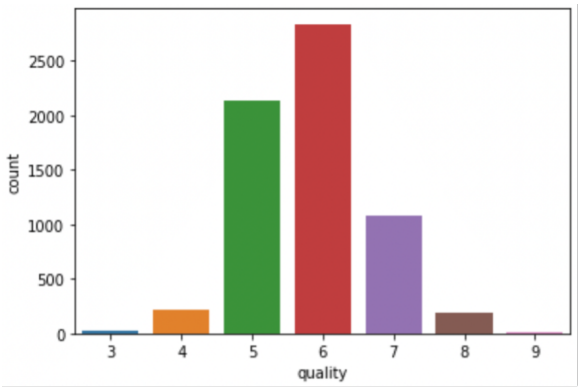

Figure 4: Count Plot of Target Variable

In this step, the research will observe the datasets for any possible bias and discrepancies which can be caused by using the current datasets. Such observation leads to the possibility of eliminating data biases and shortcomings of the findings if an untreated dataset is analyzed

for extracting meaningful information. The range for quality measure has been taken from 0 to 10, however, it has been noted that most of the dataset values fall in the range of 3 to 9. There is a high frequency and count for the middle-class qualities of the wine samples taken in the current datasets. Thus, the model will be more biased towards the middle-class qualities in the current study. The data is therefore imbalanced as not equal representation is there from all sorts of quality measures which are taken into account. As a result, the researcher will have to class-balancing of data so that of the current research can be balanced, and the bias removed. Removing the bias will allow the research to state the findings without any discrepancies, misinterpretation and misleading statements.

3.3 Data Preprocessing and Transformation

3.3.1 Outlier and Null Value Removal

Outliers are the datapoints in a dataset which are unusual and can change and alter the meaning of statistical inference if not removed. It also often violates the assumptions of data sets and statistical analysis. In reality, all sorts of datasets have the possibility of having outliers. Outliers are bound to be removed or they will create misinformation in the analysis of statistical data Bakker and Wicherts (2014). In this study, outliers were present in all the 11 output variables, all of them were removed as they can be problematic and thus their removal is necessary. There were few null values present in the fixed acidity, pH, volatile acidity, sulphates, citric acid, residual sugar and chlorides all of them were replaced with their mean value.

3.3.2 Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is a technique used by researched in order to remove the data imbalance. In the given research data, there are more instances of white wine samples and lesser instances of red wine samples, thus the data is imbalanced and needs to be balanced in order to remove the bias. Furthermore, there are more instances found between the quality value of 3 to 9 and no instances are found beyond these values. In another setting since the values above 6 are deemed to be wine with good quality and those below the value of 6 are deemed to be wine with low quality, therefore, there are more instances are values above 6 and less of those which have values below less. Thus, therefore, SMOTE has been used to balance the imbalanced data and following are the results Chawla et al. (2002).

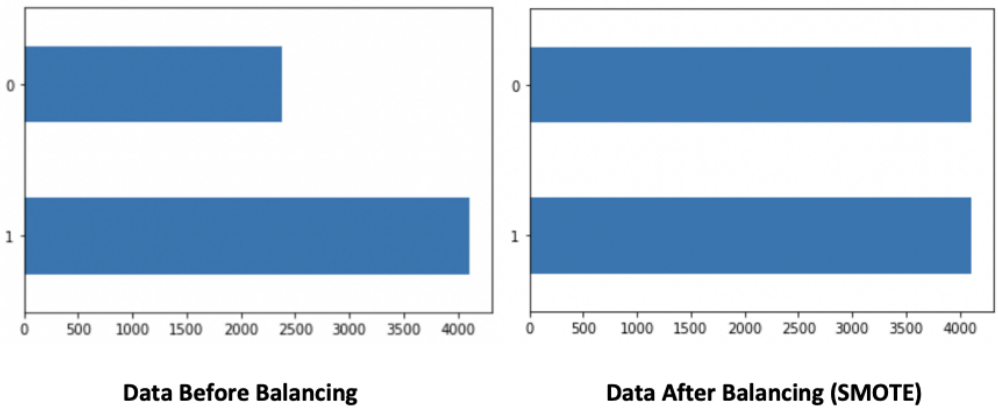

Figure 5: Data Balancing
0 indicates bad quality wine whose value is less than 6
1 indicates good quality wine whose value is greater than 6

3.4 Designing Model and Data Mining

Data Mining models are very important for data mining concept. They are the virtual structures based on which data is group in order to carry out predictive analysis. From the initial look, data mining models resemble the structure of data tables, however, they are fundamentally different from data tables. Tables serve the purpose of representing actual data sets, however, data mining models on the other hand are used for the interpretation of data which are known as cases. In this research, we will implement the Decision Tree, Random Forest, and XGBoost as machine learning models. And combination of all these three models to form a hybrid model.

3.5 Data Interpretation and Evaluation

In section 6, the machine learning models' results are presented and discussed. The results of the machine learning approaches are shown using a confusion matrix and a classification report. Each algorithm's implementation section computes and provides the pertinent f1 scores, accuracy, precision, and recall for measuring model performance.

4 Design Specification

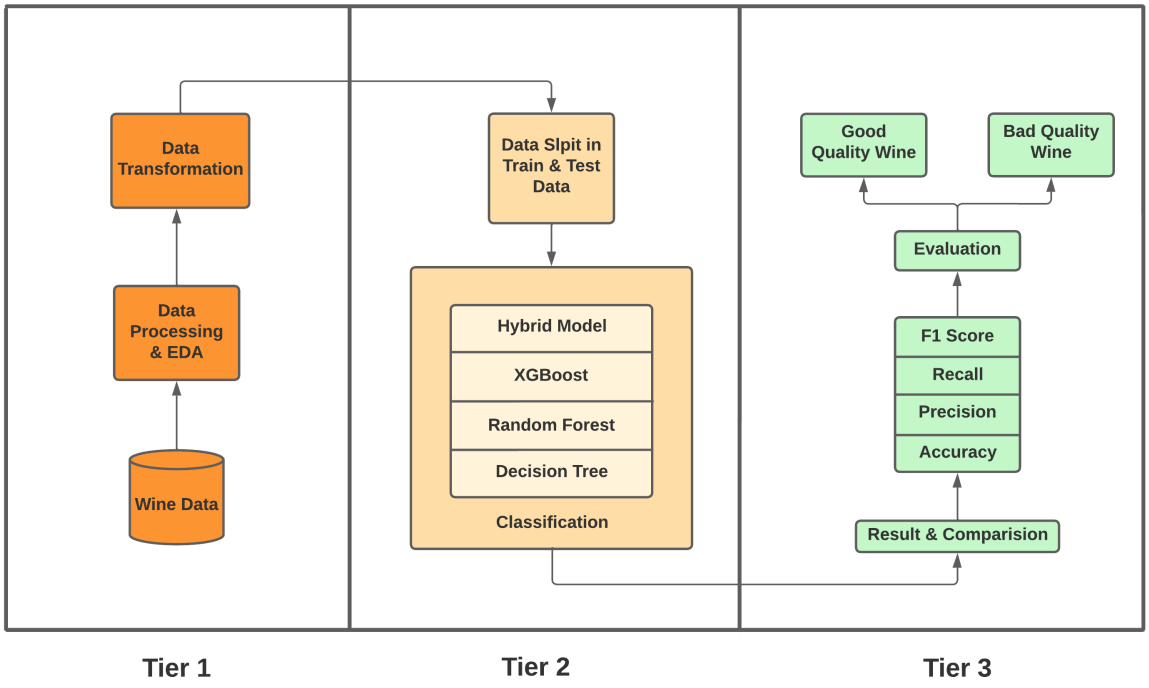

Figure 6: Project Workflow

4.1 TIER 3: Data Persistence Tier

At this stage, the data is identified and collected. The project's dataset was accessible through the Uci machine learning repository. White and red wine data are chosen and combined from the data received for use in this study. In order to generate data frames and conduct operations, the collected data is afterwards fetched in a Google Colab. The next step is pre-processing, which includes eliminating outliers and dropping null values. SMOTE was then utilized to balance

the data. Clean data was required for the classification algorithms as the project's objective is to predict wine quality. Clean data were thus produced.

4.2 TIER 2: Implementation Tier

The clean data in this layer is then split into 70:30 ratio for training and testing. Following that, classification is carried out using different machine learning algorithms, including Decision Tree, Radom forests, and Xgboost. A hybrid model of all these three machine learning algorithms is then created to compare the results and recommend the best algorithm for predicting wine quality.

4.3 TIER 1: Presentation Tier

In this stage, the result from the preceding phase is visualized. The visualizations were created using python tools like matplotlib and seaborn and take the shape of a classification report, confusion matrix, result comparison table, etc. Finally, the wine quality is predicted as either good or bad.

5 Implementation

The wine quality prediction is performed by taking sample of Vinho Verde which is Portuguese wine. The data was arranged in a way that labels were given in terms of good or bad wine quality. The quality of wine were classified as it was not distributed normally. "1" and "0" has been used to label as good or bad wine quality. The section discusses the methods used. The classification reports and the confusion matrix have also been plotted for performance evaluation purposes in section 6. The performance of the techniques are evaluated by the accuracy, precision, recall and f1 score. Finally, the comparison of the models have been made. Out of all the models, the model with best performance is selected.

5.1 Machine Learning Implementation

5.1.1 Decision Tree Classifier Implementation

Decision Tree is a supervised machine learning technique which can be utilized in order to both classify and regress problems. The most common use of decision tree is for the classification of problems Song and Ying (2015). As the name indicates, decision tree is a tree-structured classified where features of a dataset are represented by the internal nodes, decision rules are represented by branches and outcomes are represented by leaves. Nodes are of two kinds in a decision tree model Myles et al. (2004). The decision nodes are the ones which are used to make any decision which can have multiple branches. The leaf nodes are the ones which do not have any further branches and thus ends at a particular outcome (leaf).

5.1.2 Random Forest Classifier Implementation

Random Forest is the name given to a very popular algorithm of machine learning which is a supervised machine learning technique. Classification problems and regression problems are both treated by this algorithm in case of machine learning Rigatti (2017). There is a concept known as Ensemble Learning on which the working of random forest technique is based. This concept combines multiple classifiers in order to find solution to a problem and improve the model performance Brokamp et al. (2018). The name of this model indicates that there are a large number of decision trees in this concept and thus various subsets of the given datasets can

be utilized at once in order to take an average outcome from all and improve the ability of the model to prediction and give final output.

5.1.3 XGBoost Classifier Implementation

The term XGBoost stands for Extreme Gradient Boosting. It is a gradient-boosted decision tree which is scalable in nature and helps in machine learning libraries. It provides parallel tree boost capabilities to the model put in place and leads the machine learning library towards the problems solving of regression, ranking and classification problems Chen et al. (2015). XGBoost is an advanced algorithm which requires the use and understanding of other algorithms such as supervised machine learning, gradient boosting, ensemble learning and decision trees.

Gradient Boosting is a powerful algorithm of machine learning which is used to enhance the accuracy of a large number of operations such as ranking, classification and regression etc Zhou et al. (2021). The algorithm has won every competition and benchmark in the category of structured data. In cases where deep neural networks are not used or required for problem solving, there is a great chance that gradient boosting will be used in such cases.

5.1.4 Hybrid Machine Learning Model Implementation

The reason for taking a hybrid model was the fact that since all the three models have their own pros and cons and each of them has its own specialities and features. The researcher wanted to combine all the models and blend into one in order to see if it performs better than either or all of them. In the recent era, the use of hybrid machine learning models has increased considerably because of their wide application and increase efficiency in many cases. The conventional machine learning models are based on the method of presenting input data to a trained model which is based on target and predictor variables which in other terms are known as dependent and independent variables. The purpose of training is to come up with a model parameter set via an iterative procedure which enhances the relationship between the input and the target variables. As new data is fed to the model, it gets further trained and as a result more patterns are recognized so that more accurate predictions can be made. A hybrid model on the other hand is the combination of two or more conventional machine learning methods. It aims to combine the features of these machine learning models so that a hybrid model could be created which can perform better than the conventional models, however, it depends upon the case for which it is created. There are endless ways in which hybrid models can be generated from conventional models, however, in the current research study, 5 instance of each of the selection models such as Random Forest Model, Decision Tree Model and XGBoost were taken and the best of the 15 instances were selected based on which the hybrid model was generated.

6 Evaluation

As previously mentioned in section 3.5, the classification report and confusion matrix are taken into account while evaluating the model. The factors accuracy, precision, recall, and f1 score are taken into consideration while choosing the optimum model for predicting wine quality. The dataset consists of 6497 samples in total of red and white wine. Keeping a 70:30 ratio, the dataset was split into train and test sets. All of the experiments used the same train and test set. There are four experiments were performed for Decison Tree, Random Forest, Extreme Gradient Boosting and Hybrid Machine Learning Model which are demonstrated below.

6.1 Experiment 1: Decision Tree

```
Model: DecisionTreeClassifier
Accuracy Score: 0.7925445705024311
Precision: 0.7832278481012658
Recall: 0.8061889250814332
F1 Score: 0.7945425361155698
Confusion Matrix:
 [[966 274]
 [238 990]]
Classification Report:
              precision    recall  f1-score   support

           0       0.80      0.78      0.79      1240
           1       0.78      0.81      0.79      1228

    accuracy                           0.79      2468
   macro avg       0.79      0.79      0.79      2468
weighted avg       0.79      0.79      0.79      2468
```

Figure 7: Classification Report of Decision Tree

A confusion matrix and classification report were used to assess the model's performance. As it can be observed in Figure 7, the accuracy of the Decision Tree was 79.25%. Precision, recall, and f1 score are, respectively, 78.32%, 80.61%, and 79.45%.

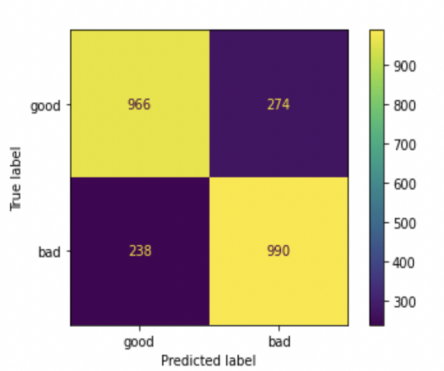

Figure 8: Confusion Matrix of Decision Tree

The confusion matrix in Figure 8 compares the True label against the expected label. Using the Decision Tree model, 966 were determined to be of good quality (TP), while 990 were determined to be of bad quality (TN) correctly. The incorrect wine quality estimate placed good wine at 274 and bad wine at 238.

6.2 Experiment 2: Random Forest

```
Model: RandomForestClassifier
Accuracy Score: 0.8557536466774717
Precision: 0.8682432432432432
Recall: 0.8371335504885994
F1 Score: 0.8524046434494196
Confusion Matrix:
 [[1084  156]
 [ 200 1028]]
Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.87      0.86      1240
           1       0.87      0.84      0.85      1228

    accuracy                           0.86      2468
   macro avg       0.86      0.86      0.86      2468
weighted avg       0.86      0.86      0.86      2468
```

Figure 9: Classification Report of Random Forest

As it can be seen in Figure 9, the accuracy of the Random Forest was 85.57%. Precision, recall, and f1 score are, respectively, 86.82%, 83.71%, and 85.24%.

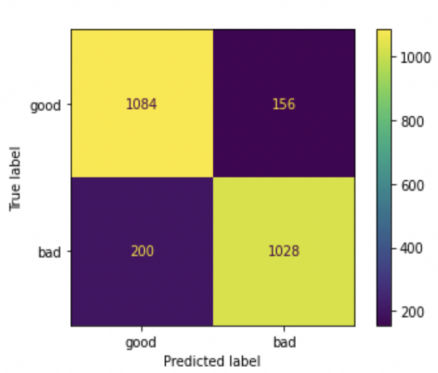

Figure 10: Confusion Matrix of Random Forest

The confusion matrix in Figure 10 compares the True label against the expected label. Using the Random Forest model, 1084 were determined to be of good quality (TP), while 1028 were determined to be of bad quality (TN) correctly. The incorrect wine quality estimate placed good wine at 156 and bad wine at 200. Random Forest performed better than the decision tree.

6.3 Experiment 3: Extreme Gradient Boosting (XGBoost)

```
Model: xgboost
Accuracy Score: 0.7807941653160454
Precision: 0.7953568357695615
Recall: 0.753257328990228
F1 Score: 0.7737348389795065
Confusion Matrix:
 [[1002  238]
 [ 303  925]]
Classification Report:
              precision    recall  f1-score   support

           0       0.77      0.81      0.79      1240
           1       0.80      0.75      0.77      1228

    accuracy                           0.78      2468
   macro avg       0.78      0.78      0.78      2468
weighted avg       0.78      0.78      0.78      2468
```

Figure 11: Classification Report of XGBoost

As it can be observed in Figure 11, the accuracy of the XGBoost was 78.07%. Precision, recall, and f1 score are, respectively, 79.53%, 75.32%, and 77.37%.

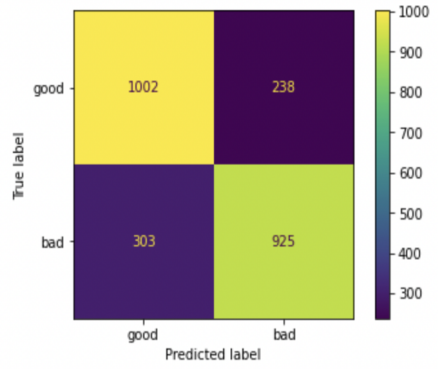

Figure 12: Confusion Matrix of XGBoost

The confusion matrix in Figure 12 compares the True label against the expected label. Using the XGBoost model, 1002 were determined to be of good quality (TP), while 925 were

determined to be of bad quality (TN) correctly. The incorrect wine quality estimate placed good wine at 238 and bad wine at 303. Random Forest performed better than the decision tree.

6.4 Experiment 4: Hybrid Machine Learning Model

```
Model: HybridModel
Accuracy Score: 0.7771474878444085
Precision: 0.79073756432247
Recall: 0.750814332247557
F1 Score: 0.7702589807852965
Confusion Matrix:
 [[996 244]
 [306 922]]
Classification Report:
              precision    recall  f1-score   support

           0       0.76      0.80      0.78      1240
           1       0.79      0.75      0.77      1228

    accuracy                           0.78      2468
   macro avg       0.78      0.78      0.78      2468
weighted avg       0.78      0.78      0.78      2468
```

Figure 13: Classification Report of Hybrid Machine Learning Model

As it can be seen in Figure 13, the accuracy of the Hybrid Model was 77.71%. Precision, recall, and f1 score are, respectively, 79.07%, 75.08%, and 77.02%.

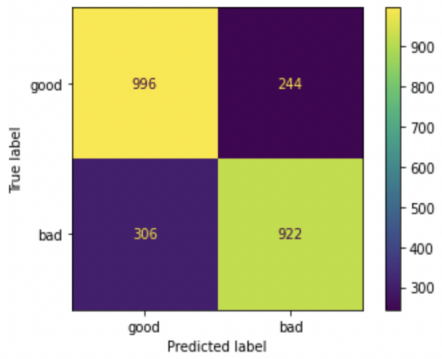

Figure 14: Confusion Matrix of Hybrid Machine Learning Model

The confusion matrix in Figure 14 compares the True label against the expected label. Using the Hybrid model, 996 were determined to be of good quality (TP), while 922 were determined to be of bad quality (TN) correctly. The incorrect wine quality estimate placed good wine at 244 and bad wine at 306.

6.5 Result Comparison and Discussion

6.5.1 Comparison of Results

Model Name	Accuracy	Precision	Recall	F1 Score
Random Forest	85.57%	86.82%	83.71%	85.24%
Decision Tree	79.25%	78.32%	80.61%	79.45%
XGBoost	78.07%	79.53%	75.32%	77.37%
Hybrid Model	77.71%	79.07%	75.08%	77.02%

Table 1: Comparison of Implemented Machine Learning and Hybrid Models

As can be seen in Table 1 comparison of models show that Random Forest model is the best suitable for this study as it has the highest scores in all four criteria namely, Accuracy (85.57%), Precision (86.82%), Recall (83.71%) and F1 Score (85.24%). On second rank we have Decision Tree model in Accuracy (79.25%), Recall (80.61%) and F1 Score (79.45%) and fourth in terms of Precision (78.32%). XGBoost Model comes on third rank in Accuracy (78.07%), Recall (75.32%) and F1 Score (77.37%), however in Precision it ranks on number two (79.53%). The Hybrid Model contrary to expectations ranks last in Accuracy (77.71%), Recall (75.08%), and F1 Score (77.02%). In Precision it ranks on third position being (79.07%).

6.5.2 Comparison of Result with Previous Research

Dataset Type	Dataset	Classifier	Accuracy	Author
Red & White wine	Same	Gradient Boosting	Red Wine:69.20% White Wine: 66.20%	Liu (2021)
Red & White wine	Same	Random Forest	Red Wine:73.25% White Wine: 76.39%	Gupta and Vanmathi (n.d.)
Red wine	Same	SVM	67.25%	Sirivanth et al. (2021)
Red wine	Same	Random Forest	68.83%	Kumar et al. (2020)
Red wine	Same	Logistic Regression	76%	A. Trivedi and Sehrawat (2018)
Red & White wine(Merged Dataset)	Same	Random Forest	85.57%	This Research

Table 2: Comparison of Result with Previous Research

As shown in Table 2 by comparing the results of the previous researchers with values achieved in the current research study, it can be said that the current research study has achieved far better values than the previous ones. Liu (2021) has reported 69.2% accuracy for red wine and 66.2% for white wine using gradient boosting classifier which are lesser than accuracy of all fours model used in the current study. Gupta et al. (2020) has also reported accuracy of red wine (73.25%) and white wine (76.39%) using Random Forest Model using the same data set. However, the accuracy of the current research study for Random Forest Model for both red and white wine combined is 85.57% which is clearly showing better results. Sirivanth et al. (2021) utilized SVM model and reported accuracy of 67.25% for red wine data set which is again pretty less than all accuracy value of all models in the current study. Kumar et al. (2020) has reported accuracy of Random Forest Model for red wine being 65.83% and Trivedi and Sehrawat (2018) reported 76% accuracy for red wine data seeing using LRM which shows that both has lesser accuracy than the current research study.

6.5.3 Result Discussion

The models performed well as they have been shown by numerous researchers to work effectively for these classification problems, providing a complete solution to the research question posed in section 1.2. The primary objective of the research was to perform wine quality predication by building models that provide accurate and efficient outcome. After a thorough literature review,

a basic understanding of current limitations and gaps were discovered. Most of the researchers either worked on white wine dataset or red wine dataset. And some of them have used both dataset in their study but they have implemented machine learning techniques separately over these two datasets. In this research project, we have combined white wine and red wine dataset as mentioned in the research objective. The best results among all were obtained by Random Forest Classifier (RFC) with the overall highest accuracy, precision, recall and f1 Score. The goal of the research included building the best suited classification model for wine quality predication is achieved as mentioned in section 1.3. In terms of performance, all the classification models performed reasonably well.

7 Conclusion and Future Work

7.1 Conclusion

The current research, achieves the research objective of the study and answers the research question of how machine learning and hybrid techniques can be used for the prediction of wine quality. The evaluation criteria used include accuracy, recall, precision and FI score. The comparative analysis has precisely answered the research question by fulfilling the research objectives of the study by implementing Decision Tree Model, Random Forest Model, XGBoost Model and a Hybrid Model implemented using Decision Tree Model, Random Forest Model, and XGBoost Model. The Portugese Vinho Verde wine was used for the prediction purposes. Two types of wine were used i.e. red wine and white wine. For dataset balancing purposes in the data pre-processing stage the Synthetic Minority Oversampling Technique (SMOTE) was used. This was done to optimize the model's performance. In the next step, those features were looked into, that could provide better prediction results. As SMOTE is applied, the performance of the model is more efficient. As the outliers and null values were removed, the dataset performance also enhanced. To conclude that the minority classes of a dataset will not get a good representation on a classifier and representation for each class can be solved by oversampling and under sampling to balance the representation classes over datasets. The accuracy of the Random Forest Classifier (RFC) algorithm is 85.57%, the Decision Tree (DT) algorithm is 79.25%, and the Extreme Gradient Boosting (XGBoost) is 78.07%. And finally Hybrid Model of all these three machine learning techniques is implemented which achieved an overall accuracy of 77.71%. As a result, choosing the right features and balancing the data in the classification algorithms will enhance the model's performance. Accuracy levels of the current research study are higher than the accuracy levels of the previously done research studies even if they used the same machine learning models. One of the reasons for this variance is fact that data for red and white wine were collected in the current research study, whereas previous research studies have taken and treated them as separated data sets.

7.2 Future Work

Following are the future recommendation based on the current research study:

- Future researchers should devise hybrid models based on more than one configuration, which has a greater chance to find a hybrid model which can produce higher accuracy and precision values than the standalone models.
- Future researchers should take into account more than currently used machine learning models such as logistic regression, multiple regression, and SVM etc, so that better comparisons between all of the various options could be made.

- Future researchers should consider deep learning models such as Artificial Neural Network, multilayer perceptron etc, take the datasets both separately as well as combined so that proper comparison could be drawn between the results of different research studies.
- Future researchers should introduce more factors than just accuracy, precision, recall and F1-score that applicability and implementation of the findings of the research study could be overall improved to a great extent.

8 Acknowledgement

My profound appreciation to National College of Ireland, our college, and the MSc in Data Analytics department for making it possible for me to accomplish this research project. Prof. Jorge Basilio, my mentor and supervisor, for his help with the research. His persistent assistance allowed me to complete and deliver this thesis. I appreciate his time, and especially his perceptive views. I think highly of him as a mentor.

References

Aich, S., Al-Absi, A. A., Hui, K. L., Lee, J. T. and Sain, M. (2018). A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques, *2018 20th International conference on advanced communication technology (ICACT)*, IEEE, pp. 139–143.

Aurit, S., Kleffner, A. and Robinson, E. (2021). Final project proposal: Statistical learning (unl stat 983) imbalanced classification and prediction of wine quality, *red* **94**: 92–6.

Bakker, M. and Wicherts, J. M. (2014). Outlier removal, sum scores, and the inflation of the type i error rate in independent samples t tests: the power of alternatives and recommendations., *Psychological methods* **19**(3): 409.

Barnaghi, P., Sheth, A. and Henson, C. (2013). From data to actionable knowledge: Big data challenges in the web of things [guest editors' introduction], *IEEE Intelligent Systems* **28**(6): 6–11.

Bhardwaj, P., Tiwari, P., Olejar Jr, K., Parr, W. and Kulasiri, D. (2022). A machine learning application in wine quality prediction, *Machine Learning with Applications* **8**: 100261.

Brokamp, C., Jandarov, R., Hossain, M. and Ryan, P. (2018). Predicting daily urban fine particulate matter concentrations using a random forest model, *Environmental science & technology* **52**(7): 4173–4179.

Caissie, A. F., Riquier, L., De Revel, G. and Tempere, S. (2021). Representational and sensory cues as drivers of individual differences in expert quality assessment of red wines, *Food Quality and Preference* **87**: 104032.

Canizo, B. V., Escudero, L. B., Pellerano, R. G. and Wuilloud, R. G. (2019). Data mining approach based on chemical composition of grape skin for quality evaluation and traceability prediction of grapes, *Computers and Electronics in Agriculture* **162**: 514–522.

Cardoso Schwindt, V., Coletto, M. M., Díaz, M. F. and Ponzoni, I. (2022). Could qsor modelling and machine learning techniques be useful to predict wine aroma?, *Food and Bioprocess Technology* pp. 1–19.

Chawla, N. V., Bowyer, K. W., Hall, L. O. and Kegelmeyer, W. P. (2002). Smote: synthetic minority over-sampling technique, *Journal of artificial intelligence research* **16**: 321–357.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K. et al. (2015). Xgboost: extreme gradient boosting, *R package version 0.4-2* **1**(4): 1–4.

Cortez, P., Cerdeira, A., Almeida, F., Matos, T. and Reis, J. (2009). Modeling wine preferences by data mining from physicochemical properties, *Decision support systems* **47**(4): 547–553.

Dahal, K., Dahal, J., Banjade, H. and Gaire, S. (2021). Prediction of wine quality using machine learning algorithms, *Open Journal of Statistics* **11**(2): 278–289.

Georgieva, P. and Rocha, E. (n.d.). Machine learning in wine classification.

Gupta, M. and Vanmathi, C. (n.d.). A study and analysis of machine learning techniques in predicting wine quality, *International Journal of Recent Technology and Engineering* **10**.

Gupta, U., Patidar, Y., Agarwal, A., Singh, K. P. et al. (2020). Wine quality analysis using machine learning algorithms, *Micro-Electronics and Telecommunication Engineering*, Springer, pp. 11–18.

Gupta, Y. (2018). Selection of important features and predicting wine quality using machine learning techniques, *Procedia Computer Science* **125**: 305–312.

Kumar, S., Agrawal, K. and Mandan, N. (2020). Red wine quality prediction using machine learning techniques, *2020 International Conference on Computer Communication and Informatics (ICCCI)*, IEEE, pp. 1–6.

Liu, Y. (2021). Optimization of gradient boosting model for wine quality evaluation, *2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI)*, IEEE, pp. 128–132.

Lukić, K., Brnčić, M., Ćurko, N., Tomašević, M., Tušek, A. J. and Ganić, K. K. (2020). Quality characteristics of white wine: The short-and long-term impact of high power ultrasound processing, *Ultrasonics Sonochemistry* **68**: 105194.

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A. and Brown, S. D. (2004). An introduction to decision tree modeling, *Journal of Chemometrics: A Journal of the Chemometrics Society* **18**(6): 275–285.

Oreški, D., Pihir, I. and Cajzek, K. (2021). Smart agriculture and digital transformation on case of intelligent system for wine quality prediction, *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, IEEE, pp. 1370–1375.

Policastro, C. A., Carvalho, A., Delbem, A. C., Mattoso, L., Minatti, E., Ferreira, E. J., Borato, C. E. and Zanus, M. (2007). A hybrid case based reasoning approach for wine classification, *Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007)*, IEEE, pp. 395–400.

Qiongshuai, L. and Shiqing, W. (2011). A hybrid model of neural network and classification in wine, *2011 3rd International Conference on Computer Research and Development*, Vol. 3, IEEE, pp. 58–61.

Rigatti, S. J. (2017). Random forest, *Journal of Insurance Medicine* **47**(1): 31–39.

Sirivanth, P., Rao, N. K., Manduva, J., Sekhar, G. C., Tajeswi, M., Veeresh, C. and Kaushik, J. (2021). A svm based wine superiority estimatation using advanced ml techniques, *2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, IEEE, pp. 207–211.

Song, Y.-Y. and Ying, L. (2015). Decision tree methods: applications for classification and prediction, *Shanghai archives of psychiatry* **27**(2): 130.

Trivedi, A. and Sehrawat, R. (2018). Wine quality detection through machine learning algorithms, *2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE)*, IEEE, pp. 1756–1760.

Yang, C., Barth, J., Katumullage, D. and Cao, J. (2022). Wine review descriptors as quality predictors: Evidence from language processing techniques, *Journal of Wine Economics* pp. 1–17.

Zhang, S., Shao, C. and Xiao, W. (2020). Research on red wine quality based on data visualization, *2020 3rd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, IEEE, pp. 128–132.

Zhou, J., Qiu, Y., Zhu, S., Armaghani, D. J., Khandelwal, M. and Mohamad, E. T. (2021). Estimation of the tbm advance rate under hard rock conditions using xgboost and bayesian optimization, *Underground Space* **6**(5): 506–515.