A Comparative Study of Machine Learning Classifiers for Binary Wine Quality Prediction

Diaa Salama AbdElminaam[1], Omar Atef Abdelwahaab[2], Khaled Mohamed[3],
Mahmoud khaled mostafa[4], Ahmed Essam Saad[5]
Faculty of Computer Science
Misr International University, Cairo, Egypt
diaa.salama[1], Omar2308664[2],
Khaled2306954[3], Mahmoud2302578[4], Ahmed2308487[5] {@miuegypt.edu.eg}

***Abstract*—Predicting wine quality is very important in the modern wine industry due to the limitations of traditional sensory evaluation methods, including subjectivity,high cost,and lack of scalability.This paper is presenting a comparative study of multiple supervised machine learning classifiers for binary wine quality prediction,focusing on utilizing physicochemical properties to improve accuracy and consistency.This study is targeting the problem of subjective human judgments into a reliable binary classification task.Traditional tasting methods are lacking scalability and consistency,making them unsuitable for large scale industrial environments and motivating the adoption of automated machine learning solutions.In this study we are implementing and comparing several supervised machine learning algorithms,including Logistic Regression,Decision Tree, Random Forest,KNN,CatBoost,and Neural Networks.The methodology is involving dataset preprocessing,removing duplicates and outliers,applying feature scaling,converting quality scores into binary labels,and training models using physicochemical attributes such as acidity,alcohol content,sulphates,pH,density,and sulfur dioxide.The evaluation is that Random Forest and Neural network models are achieving superior performance in terms of accuracy,precision,recall,and F1 score for binary wine quality prediction.Other classifiers are showing competitive results while offering simpler implementation and lower computational cost.At the end machine learning based wine quality prediction is providing an objective,scalable,and cost effective alternative to traditional sensory evaluation methods.The results are supporting the integration of automated quality assessment systems in the wine industry,enabling reliable quality control and data driven decision making with out human intervention.**

Keywords: Wine Quality Prediction, Machine Learning, KNN, Random Forest, Neural Network, CatBoost, Logistic Regression, Decision Tree, Binary Wine Quality prediction, Physicochemical Properties, Classification

I. Introduction

Wine is an alcoholic drink made from fermented grape juice. It is produced and consumed in many regions around the world in a wide variety of styles, which are influenced by different grape varieties, growing environments and production techniques. Wine can be broadly broken down into a few main types: Red wines, White wines, Sparkling and Fortified wines. There are 5 basic characteristics that help you find the best wine. Basic features are: sweetness, acidity, tannin, alcohol and body. It's very essential to understand the basic characteristics of wine, to learn how to taste it. There are several advantages and disadvantaged of wine for human health. Some of its advantages including supporting heart health, promoting mental well-being and Supporting healthy blood vessels. However, its disadvantages including sleep disruption, dental health concerns and possible allergic reactions.

Nowdays Predicting Wine Quality Had Become An Important Issue. in The Market of Wine Industry predicting wine Quality had become an important factor to Enhance Sales and get Customers Satisfaction.Traditional Tasting Methods Became ;non Scalable.this Study aims to Develop Machine learning Models to predict wine quality based on physiochmeical Factors.To help in maintaining and enhancing the Quality of Wine.

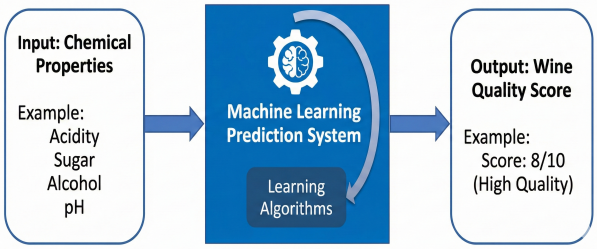

Fig. 1. General System Scope

Machine Learning is a branch of artificial intelligence that focuses on analyzing algorithms by learning from data and generalizing it's knowledge to past unseen samples. By leveraging statistical methods and optimization techniques, machine learning algorithms can enhance their performance through experience and make predictions or decisions depending on unseen data. These algorithms are widely applied in different fields like: image recognition, natural language processing and medical diagnosis. Depending on the nature of the data machine learning methods are divided into supervised, unsupervised and reinforcement learning.

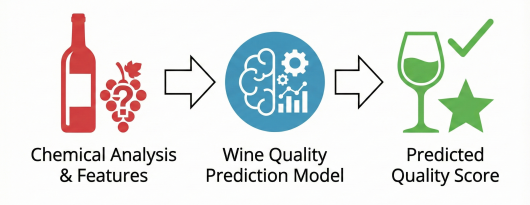

Fig. 2. Objectives

Recent studies approved that wine quality is influenced by a combination of physicochemical properties, such as acidity levels, alcohol content, sulphates, pH, and sulfur dioxide. Traditionally, experts evaluation methods suffer from subjectivity, high cost, limited scalability, human errors, Not everyone has the same taste, making them unsuitable for large-scale industrial. that's why we used machine learning techniques for wine quality prediction. ML is an effective tool to enable accurate classification of wine quality based on measurable physicochemical attributes. previous research show that algorithms such as Logistic Regression , Decision Tree Classifier, Random Forest Classifie, KNN Classifier, CatBoost Classifier and Neural Network when combined achieved strong results and performing well in wine quality prediction.

Furthermore, this study using multiple supervised ML classifiers to evaluate their effect on binary wine quality prediction by checking physicochemical features . this approach aims to provide a scalable, accurate, and cost-effective solution for automated wine quality check , this supporting quality control (No need for human)in addithion to constant , reliable decision-making processes within the wine industry

The main contribution of this paper can be summarized as follows:

- Developing a machine learning based wine quality prediction system using phsicochemical properties such as acidity,alcohol content,ph,sulphates,and sulfur dioxide for binary wine quality classification.
- Implementing and comparing multiple supervised machine learing algorithms,including logistic Regression,Decision tree Classifier, Random Forest Classifier,KNN Classifier,CatBoost Classifier,and Neural Network (MLP),to evaluate their effectiveness in predicting wine quality.
- Analyizing model performance and accuracy to demonstrate the ability of machine learning techniques to overcome the limitations of tradtional sensory evaluation methods, such as subjectivity,high cost,and lack of scalability.
- Providing a scalable and cost effective automated solution for wine quality assesment,supporting reliable quality control and decision making processes without human intervention.
- Highlighting the influence of physicochemical features on wine quality prediction, confirming their siginficance in improving classification results and supporting previous research findings.

The rest of the paper is organized as follows: Section II reviews related works on wine quality prediction; Section III proposed methodology; Section IV results and analysis; finally, Section V conclusion.

II. Related Work

The first study discusses how machine learning algorithms is used to predict red wine quality depending on physicochemical features. In this study the authors used Decision Tree, KNN, SVM, Random Forest and Naive Bayes techniques to predict wine quality. Many supervised learning algorithms were trained and evaluated using standard performance metrics like: accuracy, precision, F1-score and recall. Based on the evaluated algorithms, Random Forest had the highest predictive accuracy while other machine learning algorithms achieved lower accuracy. The results also showed that alcohol content and acidity play a huge role in predicting wine quality. To conclude, this paper confirms that Random Forest is an effective approach for predicting red wine quality. [1].

The second study presents a comparative prediction for wine quality using 3 machine learning techniques: K Nearest Neighbors(KNN), Gradient Boosting (GB) and Extreme Gradient Boosting (XGB) using a Kaggle dataset. The researchers applied data preprocessing, feature selection methods like Recursive Feature Elimination to identify the most relevant chemical attributes. The results showed that XGB has the heights performance over other models, achieving an average precision and F1-score 96% compared to 94% for GB and 87% for KNN. The research concludes that XGB is the optimal and offers superior predictive power for wine quality control. [2].

The research [3] predicts white wine quality using 6 machine learning techniques such as: KNN, Logistic Regression, SVM, Random Forest, Decision tree and Naive Bayes according to 11 chemical features from a Kaggle dataset. Applying preprocessing and Interquartile Range (IQR) methods helps improving data dependability by deleting 925 extreme samples. Experimental results showed that Random Forest had 67.9% accuracy which is the highest one while Decision tree achieved 58.1%. The analysis showed that alcohol was the most significant feature and limiting the feature set from 11 to 9, this leads to 3% drop in accuracy. To conclude the results shows the using feature importance to reduce input variables simplifies the model.

This research paper [4] uses 5 supervised machine learning algorithms including Random Forest, KNN, Decision tree, SVM and Gradient Boosting, depending on physicochemical attributes from Kaggle wine quality dataset. The study

depends on classification problem and evaluate algorithms using accuracy, recall, F1-score and other features. The results show that the most accurate algorithm is Gradient Boosting with 86% accuracy percentage, followed by Random Forest with 85% percentage. The analysis identified alcohol and volatile acidity as the most influential factors affecting wine quality prediction. To conclude this research shows that ensemble learning techniques outperform individual models and effectively support data-driven wine quality prediction.

Another paper [5] explains the performance of a Naive Bayes classifier algorithm for wine quality prediction using physicochemical features. Experimental results showed fantastic classification performance, with 100% accuracy, recall, precision and F1-score features depending on three wine quality classes. Confusion matrix confirmed zero misclassifications. Both macro and weighted averages achieved a score of 1.00, depending on consistent performance instead of class distribution. To conclude, this paper focuses on explaining the perfect performance of Native Bayes on the given dataset but this doesn't mean that this will be effective in other large datasets.

This study worked on the problem of wine quality prediction. it tests the quality using several algorithms such as XGBoost,Decision tree,K nearest neighbor and the random forest. it used a dataset of 1599 wine samples.models were evaluated by reliability,precision and f1 score. the algorithms that got highest accuracy were the Random forest and the XGBoost.this study explained how effectively can the machine learning highly affect the wine quality industry. [6]

this paper demonstrates the wine quality prediction problem by determing the most effective features either chemical or physical factors it uses maching learning algorithms as support vector machine ,naive bayes and artificial neural network to assess the classifaction the results showed that the Ann acheived better performance than the SVM features were evaluated by factors as accuarcy and precision and f1.[7]

the study explains the increase need to automatically predict the quality of wine by the help of machine learning than relying on human testing. the authors used diffrent methods as logistic regression ,AdaBoost,Random forest, and decision tree. the performance of each is scaled by diffrent factors as accuarcy,precision and f1. the random forest has been recorded as best performance. the study summarizes the prediction of quality in terms of features as alcohol content and volatile acidity [8].

this paper classify the wine based on therir phsiochemical properties this aims to distnguish between the wine types clearly. authors applied different algorthims. the dataset has passed throw preprocessing and selecting effective features. performance was compared using accuarcy and classification scores. the results explained that machine learning aims to help in distinguishing the types of wine.[9].

this study proposes wine classifer to enhance the classifying of red wine by combination of multiple machine learning approaches it used dataset containing physiochemical and quality attributes to train and test the models it combined random forest and XGBoost the accuarcy increased to 0.885 which shows that this combination beats the single models.the results indicates that the combinations could lead to enhancing the classifaction of the wine types[10]

In the paper [11], The author is describing comparing study of machine learning techniques for predicting wine quality using physicochemical properties.The study is analyzing red and wine vinho verde wine datasets and applied classifiers such as Decision Tree,Random forest,KNN,SVM,and the M5P model tree.Feature selection results indicated that alchohol content,acidity,chlorides,density,and ph significantly infeluence wine quality,while residual sugar has minimal impact.Expermintal results showed that the M5P model outperformed other classifiers,achiveing the highest accuracy for red and whine datasets.The authors is summarizing that machine learning based quality prediction can assist wine producers and certification in making quality control and decision making processes.

In their 2024 study, the Author [12] is using machine learning for classifing wine quality and taste using physicochemical properties.The research is utilized red and white wine datasets from the UCI repository and applied classification algorithms such as Logistic Regression,Stochastic Gradient Descent, Support Vector Classifier,and Random forest.Feature analysis revealed that alcohol content,volatile acidity,pH,and acidity significantly following wine quality and taste.The results is showing that Random forest achieved the higest accuracy compared to other classifiers, following by logistic Regression and Support Vector Classifier.The author conculded that machine learning models can effictevly support wine quality evaluation by improving classification accuracy and reducing reliance on tradtional manual assessment methods.

The author [13] is using machine learning for analyzing wine type and quality using physicochemical attributes of red and white wine datasets. The study employed supervised learning algorithms such as logistic Regression for wine type classification and Decision Tree and Random Forest for predicting wine quality level catagorized as low,medium,and high.The data set included key attributes such as acidity,residual sugar,chlorides,sulphates,density,pH,and alcohol,which were analyzed to identify their following on wine quality.The results is explaining that tree based models effectively captured relationships between chemical properties and wine quality,achieving higher accuracy than tradtional methods.at the end machine learning techniques provide relible and efficient solution for automated wine identification

and quality prediction,supporting decision making in wine production and quality control.

The paper [14] is disscussing the use of machine learning technique for wine quality assesment and prediction using phsicochemical characteristics of red and white wine datasets.The reasearch applied linear regression to determine the dependency of wine quality on eleven phisicochemical attributes and to identify the most important features.Neural Networks and Support vector machines were predicting wine quality using both all features and reducing set of selected features.The results achieve lower prediction errors compared to those using available features,with Support Vector Machines providing more accurate predictions overall.And finally the machine learning feature selection combined with predictive modeling can importantly improve the efficiency and objectivity of wine quality.

The paper [15] is using machine learning for predicting red wine quality using physicochemical attributes and pythin predictive modeling techniques.The study is explaining comprehsive data preprocessing steps ,including exploratory data analysis,feature selection,normalization,and dataset splitting,to prepare data obtained from the UCI Machine learning Respository.Random forest was selected as the primary classifier due to ability to handle complex and multidimensional,and its performance was compared with support vector machine and decision tree models.The results that Random Forst has achive high performance,with an accuracy of 88 percent,outperforming the other models across prescision and recall metrics.Finally machine learning driven wine quality prediction an objective,scalable,and reliable alternative to tradtional senesory evaluation,offering valuable support for decison making in the wine industry.

This article [16] considered an efficient machine learning algorithm to predict wine quality which is the Light Gradient Boosting Machine (LightGBM) algorithm , the study used a dataset has physicochemical attributes such as alcohol content, aciditylevels, sulphates, and phenolic compounds which are essential for our prediction , data preprocessing was made on dataset: data cleaning , remove outlier ,remove duplicates,future scaling and Synthetic Minority. Oversampling Technique (SMOTE) ,this algorithm produced a significant performance and results : high accurcy ,recall ,F1 score and outperforming classifiers such as Random Forest, Support Vector Machine, and Logistic Regression. feature analysis show that attributes such as Proline, Flavanoids, and Magnesium played an important role in prediction . Authors conclude that LightGBM provide a significant data-driven algorithm for predicting wine quality.

This article [17] Zhang's research introduces a great study on wine quality prediction by using machine learning algorithms with feature selection tools , zhang uses the (WineQT) dataset that contains physicochemical attributes ,he consider this problem as multi-class classification problem , using four classifiers SVM,KNN,RandomForst,ANN, to enhance performance and reduce overheads author uses features selection to predict which features has the big impact . Results shown that all tested models achieved big results between (0.80 - 0.85) , SVM achieved the highest results in accuracy , recall ,F1Score. the author highlighted that feature selection is an important step that make a significant difference in performance . The author concluded that the combination of effective feature selection with suitable machine learning models making the best results in wine quality prediction accuracy and efficiency.

The article [18] investigates ML techniques for wine quality testing using physicochemical properties as predictive features, the author used a dataset from UCI Machine Learning Repository that contains red wine samples from the Vinho Verde region of Portugal , the dataset has 11 physicochemical properties such as acidity, pH, sulphates, and alcohol content. wine quality from 0 to 10 and considered as supervised learning problem . machine learning models such as SVM , decision tree ,AdaBoost and Random forest . data preprocessing includes handle missing values, normalization, data reduction, and discretization.Random forest achieved the highest results in accuracy with (71.6%), recall ,F1Score. The author concluded that Random Forest is reliable and effective solution for wine quality prediction .

The main aim of this research paper [19] is to use binary classification (0,1) which is the most related one to our paper by using many different machine learning algorithms, also they trained the same dataset 1,599 samples with 11 input features such as acidity levels, residual sugar, sulphates, pH, and alcohol content . in the original dataset quality was from (0 to 10) that changed to binary , any number grater than five is good (1) and the rest is bad (0) . supervised classifications such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, AdaBoost, and Gradient Boosting classifiers , but we extend them by adding KNN, CatBoost, and MLPClassifier. performance is evaluated using accuracy, precision, recall, F1-score, and confusion matrices. the results as follow : Logistic Regression,Decision Tree and SVV scores 90% , but Random Forest scores about 92% The authors concluded that Random Forest provide the best performance for wine quality classification tasks.

This paper [20]uses classical linear regression(Ridge and Lasso) for predicting wine quality . the study using white wine dataset from the UCI Machine Learning Repository, the data contains 11 physicochemical attributes ,quality is an integer variable from (0 to 10) , author tests a regression problem . authors used Ridge and Lasso regularization techniques. in addition to cross-validation to find the parameters for regularization.The results of the test showed that the regularized regression models made mistakes when predicting the test set. this indicates that these models are

better at generalizing than regression problems.The authors found that Lasso regression is especially better . It can automatically choose the important features by removing some coefficients. This makes the model easier to understand for ML model . Lasso regression and Ridge regression are tools for improving regression models. The study concluded that regularized regression techniques provide a strong , effective approach for wine quality prediction.

III. Proposed Methodology

A. Datasets Descriptions

1. Wine Type Classification Dataset:

The Wine type classification dataset contains 6497 samples of red and white wines with 13 features describing the chemical properties. the dataset is designed as tasks such as wine types(red or white) and the target is to predict quality each sample contains features as:

- fixed acidity
- volatile acidity
- citric acidity
- residual sugar
- chlorides
- free sulfur dioxide
- ph
- denisty

this dataset is ideal for machine learning tasks such as binary classification(predicting wine type) or relational data to test relationship between chemical properties and wine characteristics it combines datasets from red and white.

TABLE I
FEATURES OF DATASET 1

Feature	Type	Description
Fixed acidity	Numerical	Concentration of fixed acids in wine
Volatile acidity	Numerical	Amount of acetic acid affecting taste
Citric acid	Numerical	Adds freshness and flavor
Residual sugar	Numerical	Sugar remaining after fermentation
Chlorides	Numerical	Salt content of the wine
Free sulfur dioxide	Numerical	Prevents microbial activity
Total sulfur dioxide	Numerical	Total sulfur dioxide content
Density	Numerical	Density related to alcohol and sugar
pH	Numerical	Acidity or alkalinity level
Sulphates	Numerical	Wine preservative contributor
type	Categorical	Red or white wine
Quality	Categorical	Target class (0 = bad, 1 = good)

2.Wine Quality processed Dataset:

This second dataset is a processed version of the dataset from the UCI Machine Learning Repository, and this dataset contains 5321 samples with 11 features plus wine_type and quality_label which is our target output . this dataset contains features such as :

- fixed acidity
- volatile acidity
- citric acidity
- residual sugar
- chlorides
- free sulfur dioxide
- ph
- sulphates
- alcohol content
- density
- Total sulfur dioxide

TABLE II
FEATURES OF DATASET 2

Feature	Type	Description
Fixed acidity	Numerical	Concentration of fixed acids in wine
Volatile acidity	Numerical	Amount of acetic acid affecting taste
Citric acid	Numerical	Adds freshness and flavor
Residual sugar	Numerical	Sugar remaining after fermentation
Chlorides	Numerical	Salt content of the wine
Free sulfur dioxide	Numerical	Prevents microbial activity
Total sulfur dioxide	Numerical	Total sulfur dioxide content
Density	Numerical	Density related to alcohol and sugar
pH	Numerical	Acidity or alkalinity level
Sulphates	Numerical	Wine preservative contributor
Alcohol content	Numerical	Alcohol percentage in wine
Wine_type	Categorical	Red or white wine
Quality_label	Categorical	Target class (0 = bad, 1 = good)

B. Data Preprocessing

A very important step to enhance evaluation and to give accurate results frome its steps is to remove outliers and missing values

Fig. 3. Dataset 1 Preprocessing

*1) **Dataset 1: Preprocessing Steps**:*

1) **Data Loading:**
the Wine Type classification Dataset was loaded as

Csv Format also it was uploaded in google colab and imported in pandas dataframe to enable efficient analysis

2) **Data Preprocessing:**
the dataset was examined and no missing values were found and 1177 duplicates were removed wine quailty were converted into binary 1 for good quality and 0 for bad. this enhanced the prediction for the models.
3) **training and testing :**
our dataset was split into training and testing 80 percent for training and 20 for testing. the trainig set was used to train models while the testing was preserved for measuring performance

4) **Feature Scaling:**
feature scaling was applied using standralization to normalize inputs this step decreases the error percentage and increases algorithms efficeincy,improves the performance of distance-based and gradient-based

5) **model selection:**
six algorithms were selected to test wine quality classification these are: logistic regression,random forest,KNN,XGBoost and desicison tree these are the selected machine learning algorithms
6) **model training:**
each model was trained on the training dataset and there was a diverse in the accuarcy and percision and the f1 score allowing each model to be trained and learn the pattern and relationships between features and target

*2) **Dataset 2: Pre-processing Steps**:* **Understanding the Data:**

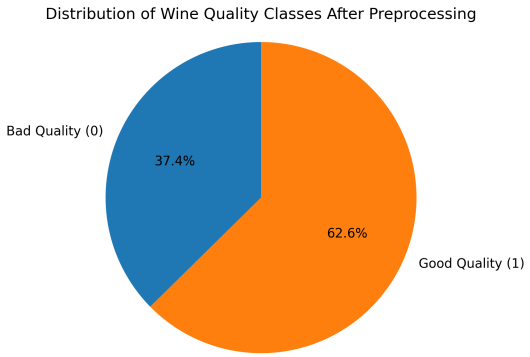

Fig. 4. Distribution of wine quality classes in Dataset 2 after preprocessing

The wine quality processed dataset consists of physico-chemical features describing wine samples. Each represents a single wine sample by numerical chemical attributes, in addition to binary quality label indicating wine quality as good or bad.before training, the dataset was examined to ensure data consistency, reliability, and suitability for supervised machine learning classification.

Data Inspection:
- Load the dataset and check for the structure
- Verify data types of all features
- Examine the target variables class distribution to look for any imbalances.

Handling Missing Values:
- check for any null values or missing values
- confirm that there is no missing values
- ensure that everything is ok for all samples before starting training phase

Filtering and Deduplication:
- remove duplicate records
- ensure that there is only unique samples to improve generalization
- validate dataset before starting training phase

Target Label Transformation:
- convert quality numbers that ranging from 0 to 10 to good and bad
- convert good and bad to binary classification 0 for bad and 1 for good quality

Feature Scaling:
- Apply standardization using the StandardScaler
- Normalize numerical features
- perform scaling only for algorithms such as Logistic Regression, KNN, and Neural Networks.

Train–Test Split:
- split data into 80 training and 20 testing

Final Dataset Preparation:
- ensure all steps applied correctly
- prepare the final processed datase

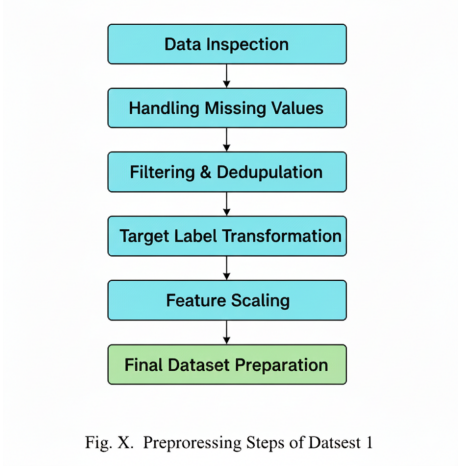

Fig. X. Preproressing Steps of Datsest 1

Fig. 5. Preprocessing Steps of Dataset 2

C. *Data Visualization*

1) Dataset 1 Visualization:

1) **Distribution of Wine Quality:**
the opposite figure illustrates the Distribution of wine

quality across different samples the chart explains an overview for the high Quality(1) wine and the low Quality(0) wine, understanding the Distribution is important for Dataset evaluation and also to select the most suitable model for appropriate performance

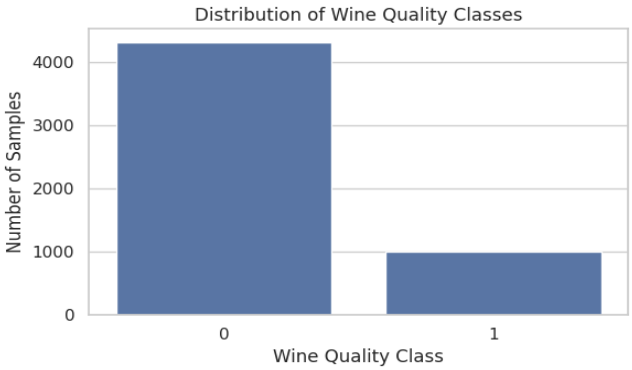

Fig. 6. Distribution of Wine Quality Classes

2) **Distribution of physicochemical features:**
figure 2 represents the Alcohol Distribution of Alcohol content and the high quality wine(1) and the low quality wine(0). there is a difference in the concentration of the alcohol. the higher quality wine tend to have more alcoholic content this figure shows how the increase in the alcoholic content affects and help in enhancing wine quality which helps us in model quality prediction

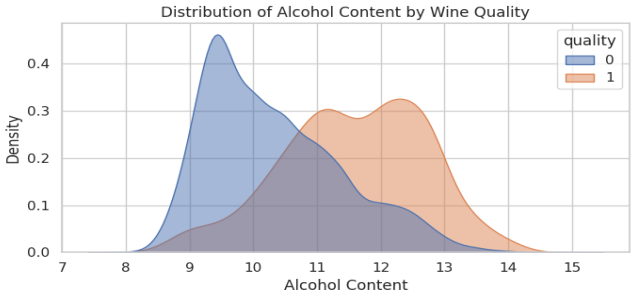

Fig. 7. Distribution of Physicochemical Features

3) **Feature importance**
figure 3 shows us the scores obtained from the bar chart against the features the bar chart is obtained from random forest algorithm. the most contributing feature is alcohol as shown.also in the second place comes the density attribute which also shows a significant effect in the wine quality. this figure helps us in understanding the features more so we could enhance the quality

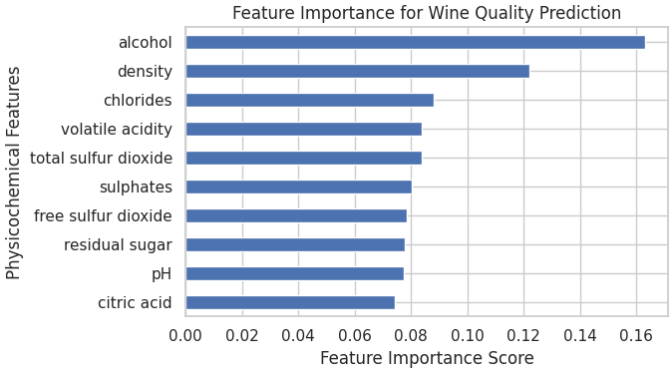

Fig. 8. Feature Importance (Random Forest)

2) Dataset 2 Visualization:

a) **Distribution of Wine Quality Classes :** The histogram shows the distribution of the of Wine Quality Classes as showen in fig we make them two classes low quality (bad) is (0) and high quality (good) is (1) . this figure shows a class imbalance highr portion of good compared to bad wine , we have to consider this evaluation and also to select the most suitable model for generating best results.

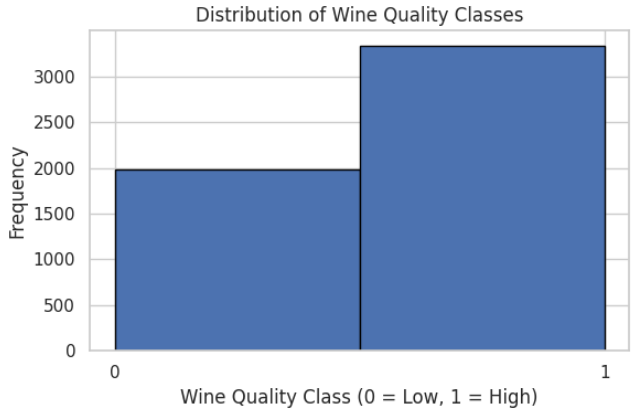

Fig. 9. Distribution of Wine Quality Classes Dataset 2

b) **Distribution of Physicochemical Features :** Two figures contains three features one for Alcohol Content and the other for Acidity and Sulphates. for Alcohol , alcohol is known as a strong indicator of wine quality. fig shows that values are centered with an average range with a little right distribution which indicates that samples have similar levels for alcohol , this difference supports the need for feature scaling when using machine learning algorithms . second fig show the distribution of Acidity and Sulphates , Volatile acidity shows a focus at lower values which is consistent with acceptable quality wines unlike sulphates that has intersection with lower

and higher values ,This distributions of features indicates their different contributions to wine quality prediction .

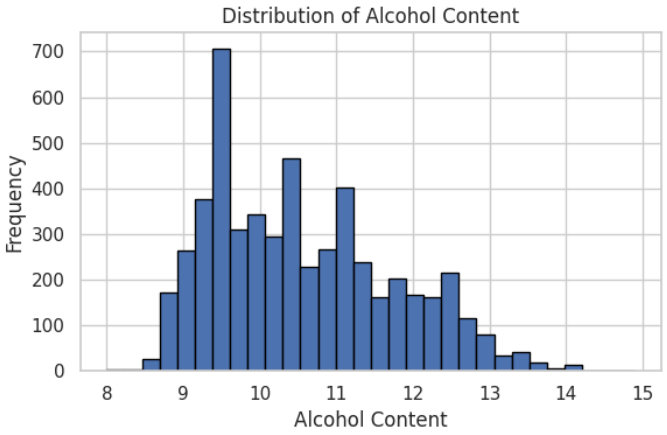

Fig. 10. Distribution of Alcohol

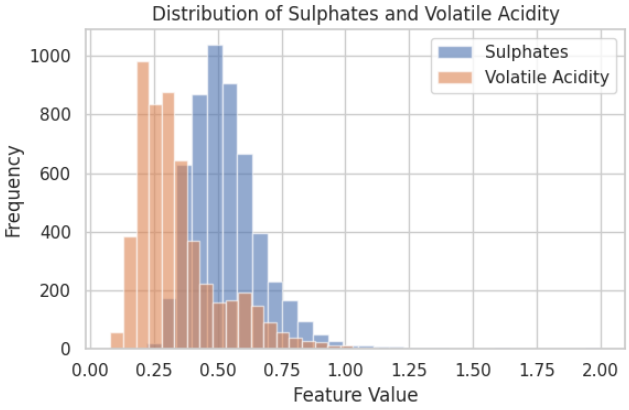

Fig. 11. Distribution of Sulphates and Volatile Acidity

c) **Correlation Analysis of Features:**
The Correlation Analysis was made to show the linear relation between the features and the Label , poositive and negative relationships as shown , A limited number of features show strong positive , while others show negative relation .This distribution confirmed that wine quality is not determined by single feature, but by a combination of multiple physicochemical features.

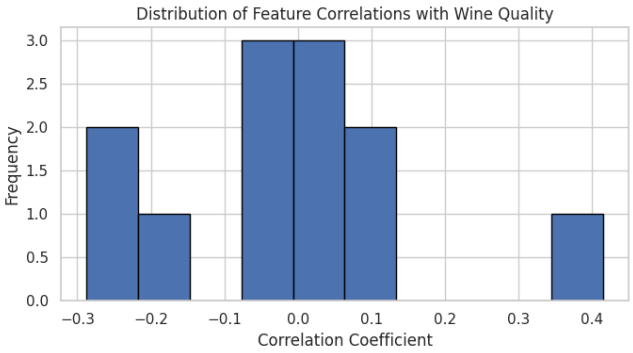

Fig. 12. Correlation Analysis of Features Dataset 2

D. *Used Algorithms*

Wine quality prediction was performed using a combination of machine learning algorithms. This section categorizes the models depending on their underlying methodology and their role predicting red and white wine quality.

***Machine Learning Models**:*

a) **Logistic Regression:**
Logistic Regression is a machine learning classification algorithm that is primarily used for binary classification tasks. It uses the sigmoid function to convert the results of the linear equation into a number between 0 and 1, then we can map it to classes.

$$P(y=1 \mid \mathbf{x}) = \sigma\left(\mathbf{w}^{\top}\mathbf{x}+b\right) = \frac{1}{1+e^{-(\mathbf{w}^{\top}\mathbf{x}+b)}}$$

b) **Decision Tree Classifier:**
Decision Tree Classifier is a powerful machine learning algorithm for sorting data. it divides data into smaller pieces based on features. The Two types of this algorithm are classification which predict categorical outcomes and Regression which predict continuous outcomes.
Entropy: S is a set which contain n training set, S is divided into c classes, each class contain n, then the entropy of S divided by c classes is

$$\mathrm{E}(\mathrm{S}) = -\sum_{i=1}^{c} \frac{n_i}{n} \log_2\left(\frac{n_i}{n}\right) = -\sum_{i=1}^{c} P_i \log_2\left(P_i\right)$$

The information gain of attribute A in according to training set S is:

$$\mathrm{Gain}(S,A) \equiv \mathrm{Entropy}(S) - \sum_{v \in V \text{ alues } (A)} \frac{|S_v|}{|S|} \mathrm{Entropy}\left(S_v\right)$$

c) **Random Forest Classifier:**
Random Forest Classifier is an algorithm that uses

a lot of decision trees for making better predictions. This algorithm is good for: Handling missing data, Showing feature importance, Working well with big and complex data and used for different tasks like classification and regression.

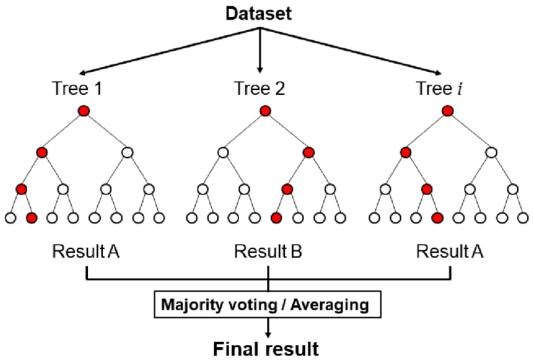

Fig. 13. Random Forest Classifier

d) **k-Nearest Neighbors (KNN):**
KNN Classifier is a non-parametric algorithm that predicts the class of a data point by considering the majority class of its K nearest neighbors. It is useful technique for both classification and regression problems. This algorithm has several advantages like: Simplicity, No training phase and flexibility.

$$\hat{y}(\mathbf{x}) = \text{mode}\left\{y_i \mid \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x})\right\}$$

where neighbors are selected using Euclidean distance

$$d\left(\mathbf{x}, \mathbf{x}_i\right) = \sqrt{\sum_{j=1}^{d} \left(x_j - x_{ij}\right)^2}$$

e) **CatBoost Classifier (Categorical Boosting)**
CatBoost Classifier is a powerful machine learning algorithm that uses gradient boosting over decision trees for solving classification problems. Key Features of this algorithm is: Ordered Boosting, Symmetric Trees, Fast Training and Minimal Parameter Tuning.

$$F_M(\mathbf{x}) = \sum_{m=1}^{M} \eta f_m(\mathbf{x})$$

f) **Neural Network (MLP Classifier)**
MLP Classifier is an algorithm that is used for classifying data into categories automatically by learning patterns from training data. It consists of interconnected layers of neurons. It is used for classification problems. There are main three types of classification in this algorithm: Binary, Multiclass and Multilabel classification.

IV. Results and Analysis

The results collected from Logistic Regression, Decision Tree, KNN, Random Forest, Catboost, Neural Network machine learning algorithms.
The following results are from the first dataset.

1) Results Visualization of Dataset 1

We begin with ROC curves to compare model discrimination across different feature sets and algorithms, then follow with a sample predictions table that highlights model successes and remaining errors.:

a) **Logistic Regression:**

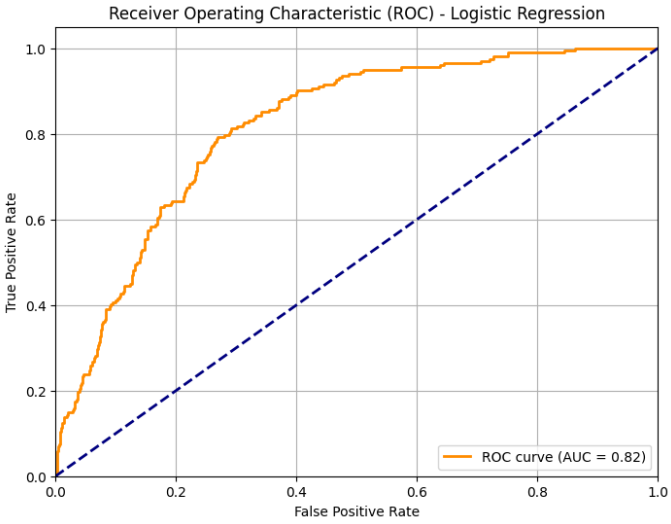

Fig. 14. ROC curve of Logistic Regression on POS columns

b) **Decision Tree:**

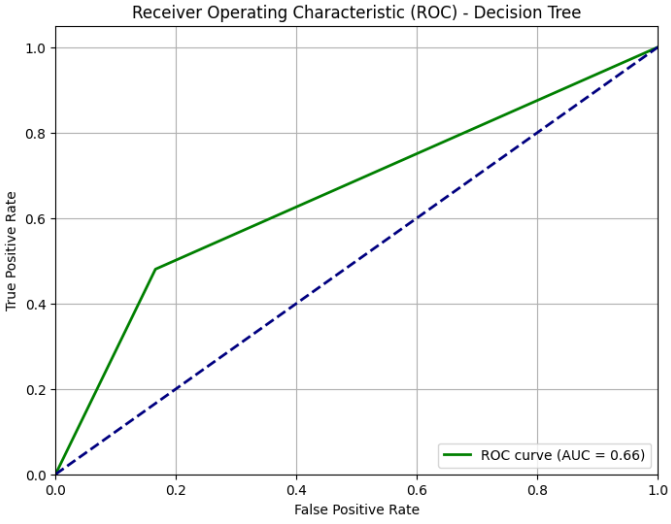

Fig. 15. ROC curve of Decision Tree Classification on POS columns

c) **KNN:**

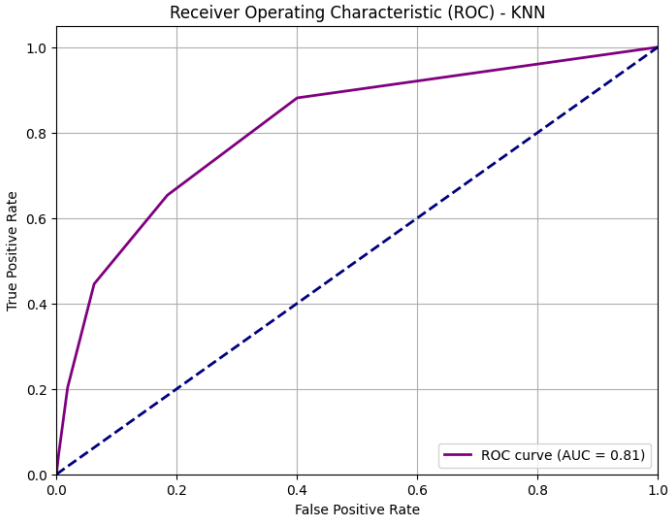

Fig. 16. ROC curve of KNN Classification on POS columns

d) **Random Forest:**

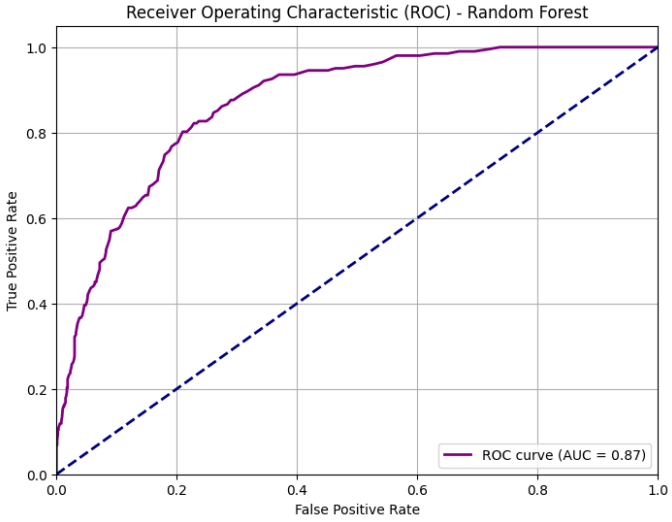

Fig. 17. ROC curve of Random Forest on POS columns

e) **Catboost Classification:**

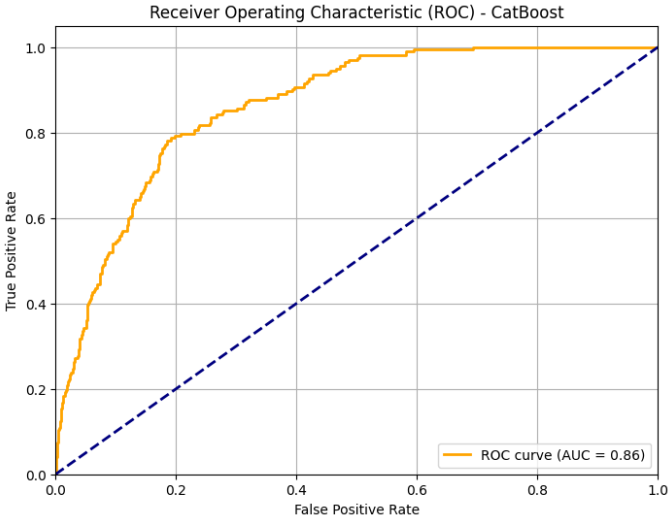

Fig. 18. ROC curve of Catboost Classification on POS columns

f) **Neural-Network:**

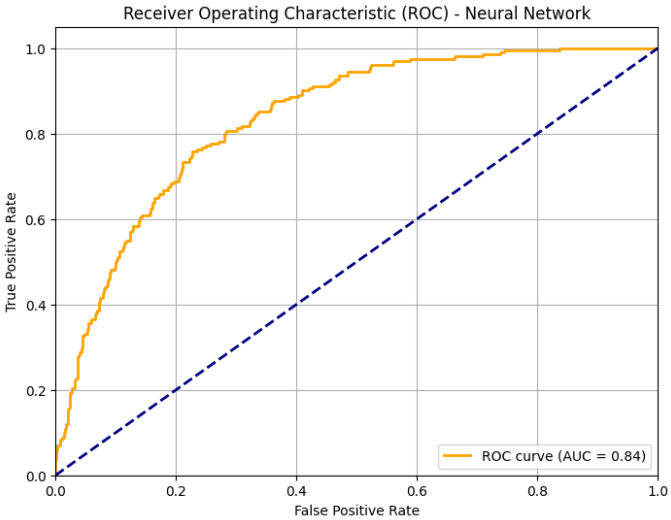

Fig. 19. ROC curve of MLP Classification on POS columns

TABLE III
CLASSIFICATION RESULT FOR LOGISTIC REGRESSION FOR WINE QUALITY DATASET 1.

Quality	Precision	Recall	F1-score	Support
Low Quality	0.8488	0.9316	0.8882	862
High Quality	0.5	0.2920	0.3688	202
Accuracy	0.8102	0.8102	0.8102	0.8102
Macro Avg	0.6744	0.6118	0.6285	1064
Weighted Avg	0.7826	0.8102	0.7896	1064

2) Results Visualization of Dataset 2

a) **Logistic Regression:**

TABLE IV
CLASSIFICATION RESULT FOR DECISION TREE FOR WINE QUALITY DATASET 1.

Quality	Precision	Recall	F1-score	Support
Low Quality	0.8726	0.8341	0.8829	862
High Quality	0.4042	0.4802	0.4389	202
Accuracy	0.7669	0.7669	0.7669	0.7669
Macro Avg	0.6382	0.6572	0.6459	1064
Weighted Avg	0.7836	0.7669	0.7743	1064

TABLE V
CLASSIFICATION RESULT FOR KNN FOR WINE QUALITY DATASET 1.

Quality	Precision	Recall	F1-score	Support
Low Quality	0.8783	0.9374	0.9068	862
High Quality	0.625	0.4455	0.5202	202
Accuracy	0.8439	0.8439	0.8439	0.8439
Macro Avg	0.7516	0.6915	0.7135	1064
Weighted Avg	0.8302	0.8439	0.8334	1064

TABLE VI
CLASSIFICATION RESULT FOR RANDOM FOREST FOR WINE QUALITY DATASET 1.

Quality	Precision	Recall	F1-score	Support
Low Quality	0.8655	0.9629	0.9116	862
High Quality	0.6952	0.3614	0.4756	202
Accuracy	0.8487	0.8487	0.8487	0.8487
Macro Avg	0.7804	0.6621	0.6936	1064
Weighted Avg	0.8331	0.8487	0.8288	1064

TABLE VII
CLASSIFICATION RESULT FOR CATBOOST FOR WINE QUALITY DATASET 1.

Quality	Precision	Recall	F1-score	Support
Low Quality	0.8753	0.9362	0.9047	862
High Quality	0.6127	0.4307	0.5058	202
Accuracy	0.8402	0.8402	0.8402	0.8402
Macro Avg	0.7439	0.6834	0.7053	1064
Weighted Avg	0.8254	0.8402	0.8289	1064

TABLE VIII
CLASSIFICATION RESULT FOR NEURAL NETWORK FOR WINE QUALITY DATASET 1.

Quality	Precision	Recall	F1-score	Support
Low Quality	0.8809	0.9012	0.8911	862
High Quality	0.5329	0.4802	0.5052	202
Accuracy	0.8214	0.8214	0.8214	0.8214
Macro Avg	0.7069	0.6907	0.6981	1064
Weighted Avg	0.8149	0.8214	0.8178	1064

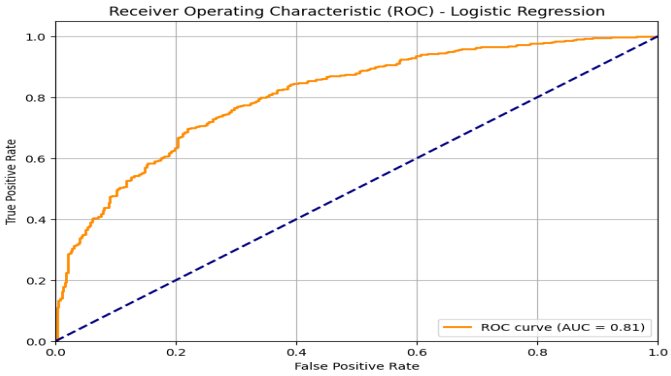

Fig. 20. ROC curve of Logistic Regression on POS columns

b) **Decision Tree:**

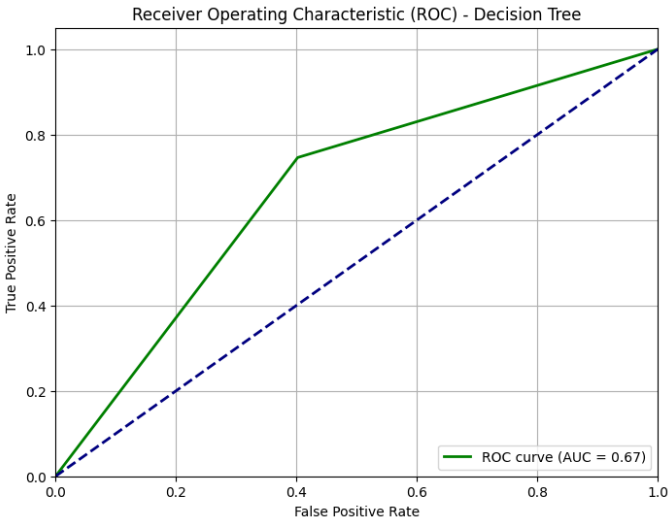

Fig. 21. ROC curve of Decision Tree Classification on POS columns

c) **KNN:**

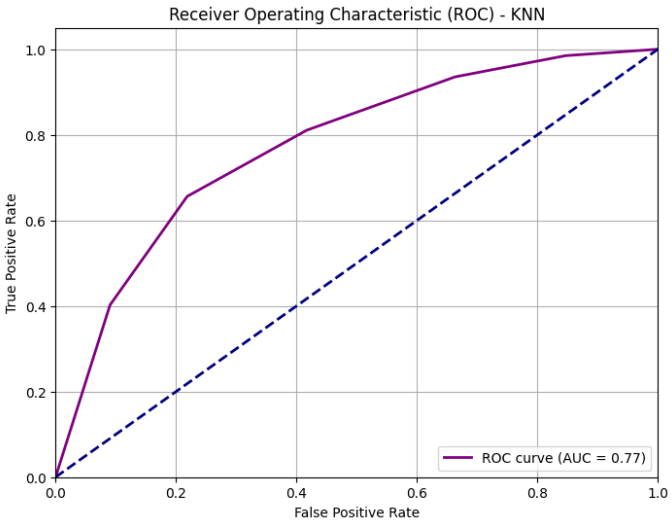

Fig. 22. ROC curve of KNN Classification on POS columns

d) **Random Forest:**

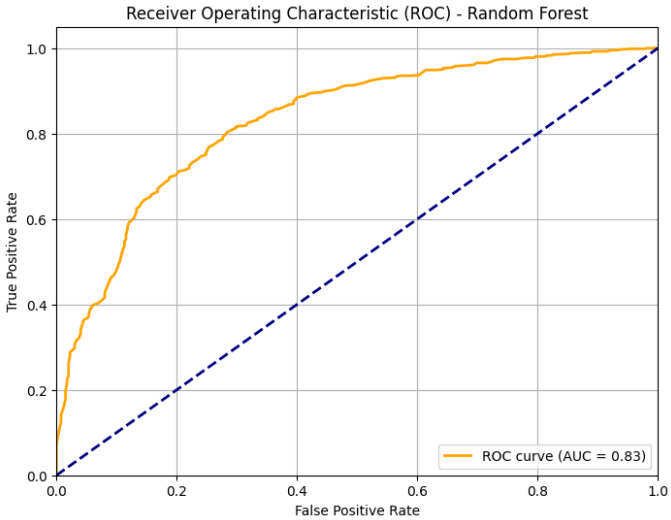

Fig. 23. ROC curve of Random Forest on POS columns

e) **Catboost Classification:**

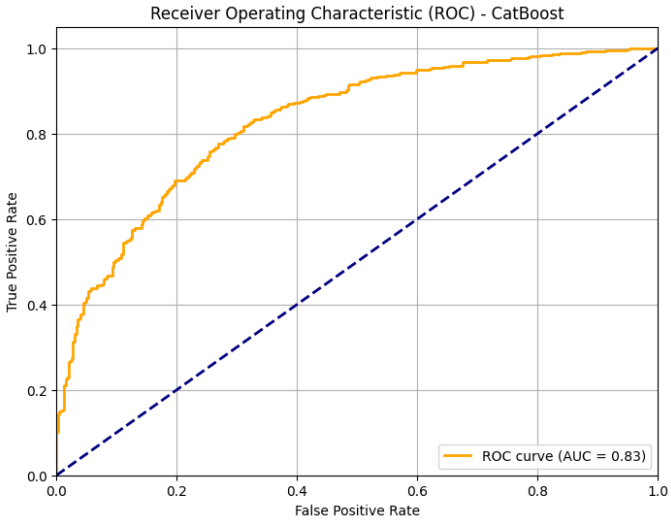

Fig. 24. ROC curve of Catboost Classification on POS columns

f) **Neural-Network:**

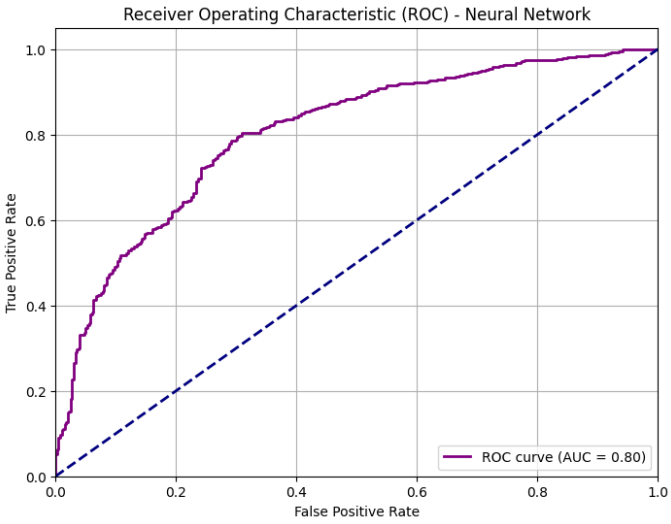

Fig. 25. ROC curve of MLP Classification on POS columns

TABLE IX
CLASSIFICATION RESULT FOR LOGISTIC REGRESSION FOR WINE QUALITY DATASET 2.

Quality	Precision	Recall	F1-score	Support
Low Quality	0.6901	0.6156	0.6507	398
High Quality	0.7842	0.8348	0.8087	666
Accuracy	0.7528	0.7528	0.7528	0.7528
Macro Avg	0.7372	0.7252	0.7297	1064
Weighted Avg	0.7490	0.7528	0.7496	1064

TABLE X
CLASSIFICATION RESULT FOR DECISION TREE FOR WINE QUALITY DATASET 2.

Quality	Precision	Recall	F1-score	Support
Low Quality	0.5848	0.5979	0.5913	398
High Quality	0.7565	0.7462	0.7513	666
Accuracy	0.6908	0.6908	0.6908	0.6908
Macro Avg	0.6706	0.6721	0.6713	1064
Weighted Avg	0.6922	0.6908	0.6915	1064

TABLE XI
CLASSIFICATION RESULT FOR KNN FOR WINE QUALITY DATASET 2.

Quality	Precision	Recall	F1-score	Support
Low Quality	0.6480	0.5829	0.6138	398
High Quality	0.7649	0.8108	0.7872	666
Accuracy	0.7256	0.7256	0.7256	0.7256
Macro Avg	0.7065	0.6969	0.7005	1064
Weighted Avg	0.7212	0.7256	0.7223	1064

TABLE XII
CLASSIFICATION RESULT FOR RANDOM FOREST FOR WINE QUALITY DATASET 2.

Quality	Precision	Recall	F1-score	Support
Low Quality	0.7207	0.6482	0.6825	398
High Quality	0.8017	0.8499	0.8251	666
Accuracy	0.7744	0.7744	0.7744	0.7744
Macro Avg	0.7612	0.7490	0.7538	1064
Weighted Avg	0.7714	0.7744	0.7718	1064

TABLE XIII
CLASSIFICATION RESULT FOR CATBOOST FOR WINE QUALITY DATASET 2.

Quality	Precision	Recall	F1-score	Support
Low Quality	0.7060	0.6457	0.6745	389
High Quality	0.7986	0.8393	0.8184	666
Accuracy	0.7669	0.7669	0.7669	0.7669
Macro Avg	0.7523	0.7425	0.7465	1064
Weighted Avg	0.7639	0.7669	0.7646	1064

TABLE XIV
CLASSIFICATION RESULT FOR NEURAL NETWORK FOR WINE QUALITY DATASET 2.

Quality	Precision	Recall	F1-score	Support
Low Quality	0.6816	0.6508	0.6658	398
High Quality	0.7968	0.8183	0.8074	666
Accuracy	0.7556	0.7556	0.7556	0.7556
Macro Avg	0.7392	0.7345	0.7366	1064
Weighted Avg	0.7537	0.7556	0.7544	1064

V. CONCLUSION

In conclusion,Predicting wine quality using machine learning is already important for ensuring consistent quality control and improving decision making in the wine industry. This research paper have the most popular physicochemical features such as acidity,alcohol content,sulphates,pH,and sulfur dioxide enables effective binary classification of wine quality,with ensemble learning models achieving higher predictive performance than simpler baseline classifiers.Also the study shows that Random Forest has achieve the highest performance comparing to the other classifiers,highlighting their ability to capture relationships between chemical attributes and wine quality.At the time there are models such as Logistic Regression and KNN providing reliable baseline performance with lower computational cost,making them suitable for lightweight or real time application.comparing multiple supervised learning algorithms highlights the strength of tree based and boosting methods in capturing complex relationships within the data,while maintaining scalability and cost efficiency.replacing subjective human evaluation with automated prediction systems is supporting more reliable and objective quality assessment,reducing operational costs, and enhancing production efficiency.continuing advancements in data availability,feature engineering,and model optimizing are expected to further improve prediction accuracy,strengthening the role of machine learning in modern wine quality evaluation systems.

REFERENCES

[1] R. N. Shahriar, A. Muhammad, M. A. Shakib, and M. A. Habib, "A machine learning application to extricate the red wine quality," 2020.

[2] M. Beri, K. S. Gill, and N. Sharma, "Predictive modeling of wine quality using machine learning techniques," in *2024 Second International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)*. IEEE, 2024, pp. 1017–1022.

[3] J. Yan, "White wine quality prediction and feature importance analysis based on chemical composition and machine learning models," *Science, Engineering and Technology CDMMS*, 2023.

[4] R. K. Kaushal, S. VishwaNarma, S. Soman, R. Dani, A. Sharma, and I. Ahamad, "Quality of wine prediction using machine learning algorithms."

[5] A. A. S. Pradhana, K. S. Batubulan, and I. N. D. Kotama, "Predicting wine quality based on features using naive bayes classifier," *Jurnal Sistem Informasi dan Komputer Terapan Indonesia (JSIKTI)*, vol. 7, no. 1, pp. 275–284, 2024.

[6] H. Arshad, "The wine quality prediction using machine learning," *Journal of Innovative Computing and Emerging Technologies*, vol. 4, no. 2, 2024.

[7] R. D. Kothawade, "Wine quality prediction model using machine learning techniques," 2021.

[8] N. Korade and M. Salunke, "Identification of appropriate machine learning algorithm to predict wine quality," *Submitted to International Journal of Scientific Research in Engineering and Management (IJSREM)*, vol. 5, no. 05, 2021.

[9] B. Ahammed and M. M. Abedin, "Predicting wine types with different classification techniques," *Model Assisted Statistics and Applications*, vol. 13, no. 1, pp. 85–93, 2018.

[10] D. J. I. Supriatna, H. Saputra, and K. Hasan, "Enhancing the red wine quality classification using ensemble voting classifiers," *Infolitika Journal of Data Science*, vol. 1, no. 2, pp. 42–47, 2023.

[11] M. Gupta and C. Vanmathi, "A study and analysis of machine learning techniques in predicting wine quality," *International Journal of Recent Technology and Engineering*, vol. 10, pp. 314–321, 2021.

[12] A. Sinha and A. Kumar, "Wine quality and taste classification using machine learning model," *International Journal of Innovative Research in Applied Sciences and Engineering*, vol. 4, no. 4, pp.

715–721, 2020.

[13] S. D. S. Johar, S. JNNCE, M. Ganavi, and S. N. Nayak, "Analyzing wine types and quality using machine learning techniques."

[14] Y. Gupta, "Selection of important features and predicting wine quality using machine learning techniques," *Procedia Computer Science*, vol. 125, pp. 305–312, 2018.

[15] G. Singh, S. J. Quraishi, D. Ather, V. Saxena, T. Z. Baig, and R. Kler, "Machine learning-based wine quality prediction using python: A predictive modeling approach," 2024.

[16] B. Muslimin, "Efficient wine quality prediction and classification using lightgbm model," *Jurnal Sistem Informasi dan Komputer Terapan Indonesia (JSIKTI)*, vol. 5, no. 3, pp. 365–375, 2023.

[17] H. Zhang, "Research of machine learning and feature selection in wine quality prediction."

[18] S. Mani, R. A. Krishnankutty, S. Swaminathan, and P. Theerthagiri, "An investigation of wine quality testing using machine learning techniques," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 2, p. 747, 2023.

[19] N. Khilari, P. Hadawale, H. Shaikh, and S. Kolase, "Analysis of machine learning algorithm to predict wine quality," *International Journal of Scientific Research in Science, Engineering and Technology*, vol. 9, pp. 231–36, 2022.

[20] M. Thevaraja, A. Rahman, and M. Gabirial, "Recent developments in data science: Comparing linear, ridge and lasso regressions techniques using wine data," in *International Conference on Digital Image and Signal Processing*, 2019, pp. 1–6.