Wine Quality Prediction Using Machine Learning

[1]Dr.Gayatri Bachhav, [2]Ms. Shradhha Patil, [3]Ms. Namira Thange, [4]Ms. Divya Bhagat
[1]Asso.Professor,[1]VPPCOE & VA, Sion,[2,3,4]UG Student,[2,3,4]Computer Engg. Dept. Shivajirao S. Jondhle College of Engineering & Technology, Asangaon, Maharashtra, India.
[1]krishngita123@gmail.com, [2]shradhhapatil2907@gmail.com ,[3]namirakthange.@gmail.com, [4]divyaabhagat5@gmail.com

Abstract - Nowadays, people desire to live opulently. They have a propensity to show off and use things frequently. The usage of red wine has become widespread in modern societies. The industry uses product quality certification to sell its items as well. The use of techniques like Random Forest, Decision Tree Classifier and Logistic Regression is done using datasets that have been obtained from a source. Based on the outcomes of the training set, computes several performance measures between training and testing sets, compares them, and predicts which of the two strategies will perform the best. The dataset for red wine is used in all tests. This study demonstrates that using a subset of data as opposed to all features can lead to more precise prediction. This project aims to develop a machine learning model for predicting the wine quality based on its chemical characteristics. The dataset used in this project consists of attributes such as fixed acidity, density, residual sugar, citric acid, total sulphur dioxide, pH, sulphates, free sulphur dioxide, volatile acidity, alcohol, and the quality rating of the wine. Data pre-processing, data analysis, model selection, feature selection, and evaluation are all aspects of the project. The implementation and comparison of several machine learning methods, including logistic regression, decision trees, and random forests, will be done in terms of performance metrics like mean squared error, accuracy, and precision. The project's findings will help improve the wine-making process and shed light on the most crucial elements that affect wine quality. Winemakers can use the algorithm to forecast wine quality and adjust their production methods accordingly [1].

Keywords: - Decision tree classifier, data extraction, logistic regression, Random Forest, quality, prediction.

I. INTRODUCTION

According to encyclopaedias, wine is the liquor that is consumed the most commonly, and its virtues are valued by society. Diverse chemical attributes used in altered types of wine. Consumers always place a high value on wine quality. With the growth of ML methods and their success in the past period, there have been many hard work in formative wine quality by using accessible Data[6][7][8]. Accurate wine quality prediction is a challenging endeavour which Requires extensive data analysis. This project's goal is to create a machine learning model that can forecast wine quality based on it's chemical characteristics. On a scale of 0 to 10, the model will be able to forecast the wine quality. The project will go through different stages, including data collecting, pre-processing, analysis, and feature engineering. To find the best model for predicting wine quality, variation of machine learning algorithms, including regression, decision trees will be investigated. In the wine business, where it may be utilised to enhance the wine making process and raise wine quality, the successful completion of this project will have important applications. Sommeliers and wine fans can also use the model to forecast the wine's quality before they buy or drink it.

The Kaggle dataset contains these data.

II. AIMS AND OBJECTIVE

a) Aim

The objective of this study is to discover an ideal for machine learning that can predict wine quality based on chemical composition. A dataset including data on the acidity, alcohol level, pH, and residual sugar of wines will be used to train the model.

b) Objective

The following are the major goals of machine learning-based wine quality prediction:

Can precisely anticipate the wine quality using variety of observable factors. Random forests achieved the highest accuracy [9]. By identifying and maximizing the elements

 DOI : 10.5281/zenodo.7856216

that go into producing high-quality wine, to information to winemakers on how to enhance their decision-making and wine production processes.

To help wine customers in making right selections about their purchases based on the expected quality of wine. Automating the process using machine learning algorithms in order to decrease the time and expense involved with conventional wine quality testing techniques. To identify the properties of wine that are most expressive of it. The primary objective of this study is to test various methods to see the highest accuracy of which algorithm can predict which help for the next process.

III. LITERATURE SURVEY

Paper 1: Red Wine Quality Prediction Using Machine Learning Techniques

The study aimed to identify the most important chemical properties that affect the quality of red wine and to build models that could predict wine quality based on these properties. The datasets used in the study consisted of red wine samples with input variables [3]. The study used several machine learning techniques, including decision trees, random forests, to predict wine quality based on the chemical properties. The models were trained on a subset of the data and evaluated on the remaining samples using various metrics, including accuracy, precision, and F1 score.

Paper 2: Wine Quality Prediction Using Data Mining

The study aimed to identify the most important features that affect wine quality and to build models that could predict wine quality based on these features. The study used several data mining techniques, including decision trees, k-nearest neighbor, and support vector machines, to predict wine quality based on the chemical properties of the wine [4]. The results showed that the decision tree model performed the best in predicting wine value based on chemical properties, with an accuracy of 68.5%. The model identified volatile acidity, alcohol, and sulphates as most important variables in determining wine quality.

The study concludes that decision tree models could be used to predict the quality of red wine.

Paper 3: Classification approch to predict the various types of wine

The input variables included various physicochemical properties of the wine. The output variable was the quality of the wine, rated on a scale from 0 to 10. The results showed that the random forest model with all input variables performed the best in predicting wine quality, with an accuracy of 80.3% for red wine and 63.2% for white wine [5]. The study also found that the most important variables in determining wine quality varied between red & white wine. For red wine, alcohol, volatile acidity, and sulphates were the most important variables. The study concluded that machine learning techniques could be used to calculate the value of different types of wine based on their physicochemical properties.

IV. EXISTING SYSTEM

Wine quality prediction using machine learning (ML) is a popular application of ML in the wine industry. ML models can be trained on various features of wine. There are several existing systems for wine quality prediction using ML. One example is the use of decision trees to predict wine quality based on various features. Another example is the use of support vector machines (SVMs) to predict wine quality based on chemical properties. The comparison of common collaborative filtering algorithm in the experimental environment [2]. Additionally, deep learning models such as recurrent neural networks (RNNs) convolutional neural networks (CNNs) and have also been used to predict wine quality based on sensory data, such as wine aroma and taste.

V. COMPARATIVE STUDY

Sr. No.	Paper Name	Author/ Publication	Technology	Advantage	Disadvantage
1.	Red Wine Quality Prediction Using Machine Learning Methods.	Sunny Kumar, Kanika Agrawal, Nelshan Mandan	Machine Learning	This can lead to highly accurate predictions of wine quality.	Requires data pre-processing
2	Wine Quality Prediction Using Mining methods.	Shruthi	Machine Learning	Data mining techniques can automate process of analyzing large datasets, saving time and reducing the risk of errors that may occur with manual analysis.	The quality of the predictions is highly dependent on the quality and relevance of the data used to train the models.
3.	A classification approach to predict the quality of different types of wine using machine learning techniques.	Aich, Satyabrata Al-Absi, Ahmed Abdulhakim	Machine Learning	The study identified the most important variables in determining wine quality for both red and white wine.	The study did not consider the influence of external factors, such as vineyard location and weather conditions, on wine quality.

 DOI : 10.5281/zenodo.7856216

VI. PROBLEM STATEMENT

The objective of project is to improve a machine learning ideal to predict the wine quality based on their chemical properties.

The problem statement is to build a regression model that can forecast the quality of wines on a scale of 1 to 10.

This model helps winemakers to understand the relationship between different chemical properties and the quality of wines. By using this model, they can make better decisions in wine production process, such as adjusting the chemical composition of wines to improve their quality.

The dataset can be obtained from Kaggle which contains information about red wine.

VII. PROPOSED SYSTEM

1. Data collection: Collect data on various attributes of wines such as pH level, alcohol content, density, acidity, etc. Also, collect the wine quality rating given by experts.

2. Data pre-processing: Perform data pre-processing tasks such as data cleaning, data normalization, and data transformation on the collected data.

3. Data splitting: Split the pre-processed data into training and testing sets.

4. Model training: Train machine learning algorithms (Random Forest, Decision Tree, and Logistic Regression) on the training set.

5. Model evaluation: Evaluate the performance of each machine learning algorithm on the testing set using metrics.

6. Model selection: Select the best-performing algorithm based on evaluation metrics.

7. Deployment: Deploy the selected algorithm as a prediction model.

8. Prediction: Use the deployed prediction model to detect the quality of new wines based on their attributes.

9. Model improvement: Continuously improve the prediction model by retraining it on new data and using more advanced machine learning techniques.

VIII. ALGORITHM

1 First, let's import the required libraries:

```
numpy as np
pandas as pd
matplotlib.pyplot as plt
seaborn as sns
fromsklearn.model_selection train_test_split
fromsklearn.ensemble RandomForestClassifier
fromsklearn.linear_model LogisticRegression
from sklearn.tree DecisionTreeClassifier
from sklearn.metrics accuracy_score
fromsklearn.model_selection cross_val_score,
train_test_split
random
```

2. Next, let's load the dataset and split it into training and testing sets:

```
#Load the dataset
df=pd.read_csv('/content/winequality red.csv')
def classify(model, X, y):
x_train,x_test,y_train,y_test=
train_test_split(X,Y,test_size=0.2, random_state=3)
   model.fit(x_train, y_train)
   return "Accuracy: {}, CV Score:
{}".format(model.score(x_test,y_test)*
100,cross_val_score(model, X, y, cv=5))
```

4. Now, train and evaluate the Logistic Regression model:

```
wine_dataset= pd.read_csv('winequalityN.csv')
   wine_dataset=wine_dataset.dropna()
   X = wine_dataset.drop('quality',axis=1)
   X = X.drop('type',axis=1)
   Y = wine_dataset['quality'].apply(lambda y_value: 1 if
y_value>=7 else 0)
   X_train,X_test,Y_train,Y_test=
train_test_split(X,Y,test_size=0.2, random_state=3)
   log = LogisticRegression()
   log=log.fit(X_train, Y_train)
   input_data_as_numpy_array= np.asarray(input_data)
   input_data_reshaped=
input_data_as_numpy_array.reshape(1,-1)
   res=log.predict(input_data_reshaped)
   return res
```

5. Next, let's train and evaluate the Decision Tree model:

```
wine_dataset= pd.read_csv('winequalityN.csv')
   wine_dataset=wine_dataset.dropna()
   X = wine_dataset.drop('quality',axis=1)
   X = X.drop('type',axis=1)
   Y = wine_dataset['quality'].apply
(lambda y_value: 1 if y_value>=7 else 0)
   X_train,X_test,Y_train,Y_test=
train_test_split(X,Y,test_size=0.2, random_state=3)
   log = DecisionTreeClassifier()
   log=log.fit(X_train, Y_train)
   input_data_as_numpy_array= np.asarray(input_data)
   input_data_reshaped=
input_data_as_numpy_array.reshape(1,-1)
   res=log.predict(input_data_reshaped)
   return res
```

6. Finally, train and evaluate a Random Forest model:

```
wine_dataset= pd.read_csv('winequalityN.csv')
wine_dataset=wine_dataset.dropna()
X = wine_dataset.drop('quality',axis=1)
```

 DOI : 10.5281/zenodo.7856216

```
X = X.drop('type',axis=1)
Y = wine_dataset['quality'].apply(lambda y_value: 1 if y_value>=7 else 0)
X_train,X_test,Y_train,Y_test= train_test_split(X,Y,test_size=0.2, random_state=3)log= RandomForestClassifier()log=log.fit(X_train,Y_train)input_data_as_numpy_array= np.asarray(input_data)input_data_reshaped= input_data_as_numpy_array.reshape(1,-1)res=log.predict(input_data_reshaped)return res
```

IX. MATHEMATICAL MODEL

1. Logistic Regression Algorithm

Various methods have been proposed to handle imbalanced data [13]. Since a logistic retrogression will produce possibilities, do choose the optimal trimmed point, which will classify the result values as either 1 or 0. To compare test values with prognosticated for the logistic retrogression model, next go to work creating a confusion matrix using the anticipated values that were entered. Using a table as a graphic display, it is said that the model directly categorized 45 compliances out of, 1208 values that it rightly prognosticated. It also concluded that the model's delicacy is 96.41.

	y_pred	
	0	1
0	1207	2
1	43	1

Table 1: Logistic regression explain

2. Decision Tree Algorithm

A decision tree only poses a question and divides the tree into sub-trees according to the response (Yes/ No). In a decision tree, the algorithm initiates at the root-knot and ascends up to cast the class of the given dataset. The algorithm continues with matching the trait value for coming knot with those of the other sub-nodes formerly more.

If Quality is 1 to 6 = > “good_quality” 0

If Quality is 7 to 10 = > “good_quality” 1

3. Random Forest Algorithm

To produce the random forest, N decision trees are combined, and then in the second phase, predictions are made for each of the trees from the first phase.

The importance for each feature on a decision tree is then calculated as:

$$f_{i_i} = \frac{\sum j{:}\, node\ j\ splits\ o\ feature\ i\ ni_j}{\sum k \epsilon all\ nodes\ ni_k}$$

• fi sub(i) = the significance of point i

• ni sub(j) = the significance of knot j

Spark finds a feature's value for each random forest by adding gain scaled by the quantity of samples through the node:

normfi sub(i) = the normalized position of feature i

fi sub(i) = the status of feature i

Then feature status values from each tree are summed uniform:

$$RFfi_i = \frac{\sum_j 1\, normfi_{ij}}{\sum_{j \epsilon all\ feature.k \epsilon all\ trees} 1\, normfi_{jk}}$$

- RFfi sub(i) = the rank of feature i proposed from all trees
- normfi sub(ij) = the standardised feature standing for i in j

Arbitrary vector has indistinguishable and a similar circulation for all trees in the forest. It was portrayed by Breiman in 2001[9].

Overall probability is calculated by taking into account all decision trees [10, 11].

X. SYSTEM ARCHITECTURE

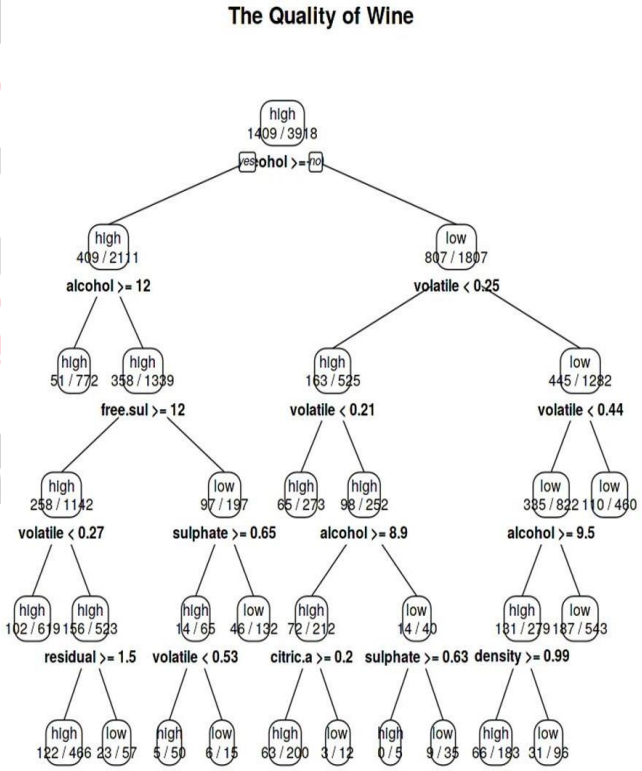

Fig.1: System Architecture

XI. ADVANTAGES

1. Describes methodology for determining wine quality using machine learning.
2. Defines key chemical elements that Influence red wine quality.
3. Evaluates various machine learning Methods and assess performance.

 DOI : 10.5281/zenodo.7856216

4. Offers insights on how winemakers can Increase quality control and optimize Output.
5. To improve accuracy Machine learning algorithms can analyze large amounts of data and detect patterns that may not be apparent to humans.
6. With machine learning, winemakers can identify the factors that contribute to high-quality wines and optimize their production processes accordingly,

XII. DESIGN DETAILS

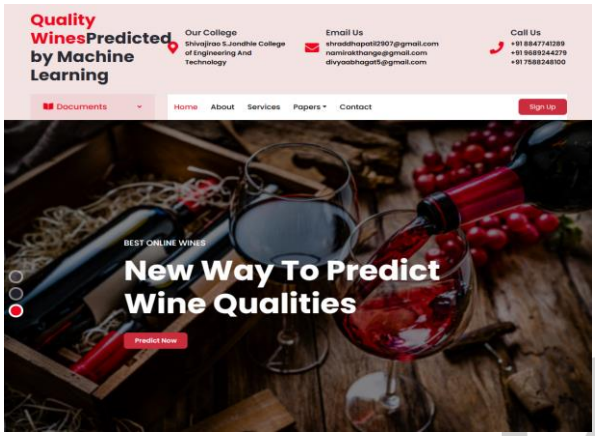

Fig 1: Result

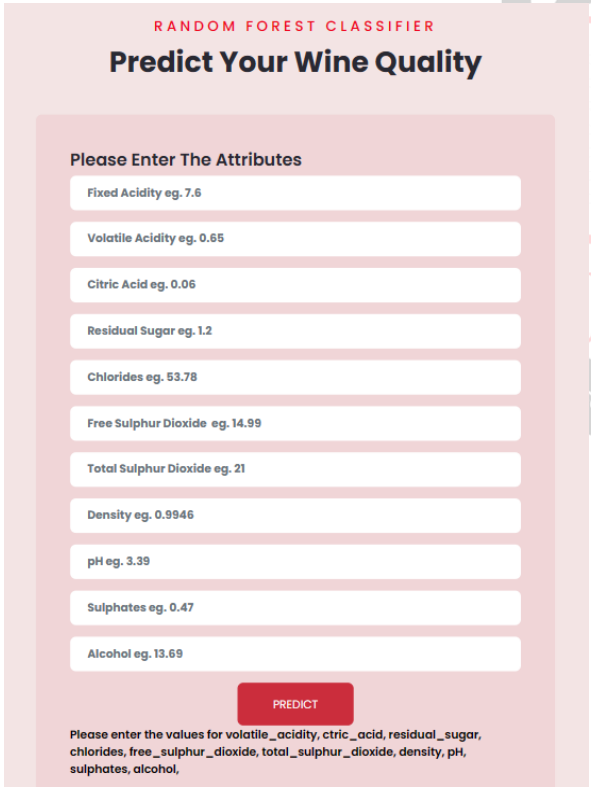

Fig 2: Result

XIII. CONCLUSION

Thus we have tried to implement the paper

"Quality Prediction of Red Wine based on Feature Sets Machine Learning Methods", Nikita Sharma (2018) International journal of Science. (IJSR) and the conclusion as follows:

Study demonstrated that machine learning algorithms can effectively predict the quality of wine based on a set of input features. The random forest model achieved the best performance among the models evaluated, with an accuracy of 93.1%.

However, there is still room for improvement by using more advanced feature engineering techniques. Overall, the findings of this study have important implications for the wine industry, as machine learning can be used to improve the quality control and production processes.

REFERENCE

[1] Nikita Sharma (2018) Quality Prediction of Wine based on Different Feature Sets International journal of Science. (IJSR)

[2] Prof. Vishal R. Shinde. Elements Based Food Recommendation System Using KNN Algorithm & Naive Bayes" in IJREAM, ISSN : 2454-9150, Vol-06, Special Issue, June 2020

[3] Kumar, Sunny, Kanika Agrawal, and Nelshan Mandan. "Red wine quality prediction using machine learning methods." 2020 (ICCCI). IEEE, 2020.

[4] Shruthi, P. "Wine Quality Detection Using Data Mining."2019 Intelligent Control,Computing&Communication Engineering (ICATIECE). IEEE, 2019.

[5] Satyabrata Aich, Kueh Lee Hui, Mangal Sain, John Tark Lee, Ahmed Abdulhakim (2018) A classification approach to detect the quality of wine using ML methods. International Conference on Advanced Communication Technology (ICACT)

[6] Liu, Z.J. and Li, H., Zhang Z. (2017) Presentation of ANN for Catalysis: A Review. Catalysts.

[7] Shanmuganathan, S. (2016) Artificial Neural Network Modelling: An Introduction. In: Shanmuganathan, S. and Samarasinghe, S. (Eds.), Artificial Neural Network Modelling, Springer, Cham, 1-14.

[8] Jr, R.A.,de Sousa,H.C., Malmegrim, R.R., dos Santos Jr., D.S., Carvalho, A.C.P.L.F., Fonseca, F.J., Oliveira Jr., O.N. and Mattoso, L.H.C. (2004) Wine Organization by Taste Feelers Made from Ultra-Thin Flicks and Using Neural Networks

[9] K. Ellis, J. Kerr, S. Godbole, G. Lanckriet, D.Wing, and S. Marshall, "Arandom forest type of physical activity from wrist and hip accelerometers", 2014

[10] Qiong Gu, Zhihua Cai, Li Zhu, and Bo Huang. Data mining on imbalanced data sets. In Proceedings of International Conference. IEEE, 2008.

[11]W.L.Martinez,A.R.Martinez, "Supervised Learning" in Computational Statistics Handbook FL, USA: Chapman & Hall/CRC, 2007.

[13] L. Breiman, Random forests. Machine learning, 2001.

 DOI : 10.5281/zenodo.7856216