International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 12 | Dec 2021 www.irjet.net p-ISSN: 2395-0072

Analysis of Machine Learning Algorithm to predict Wine Quality

Nitin Khilari[1], Pravin Hadawale [2], Hasan Shaikh[3], Sachine Kolase[4]

[1]BE Student, Jaihind College of Engineerin, Kuran, Junnar, Pune, nitinkhilari176@gmail.com
[2]BE Student, Jaihind College of Engineering, Kuran, Junnar, Pune, pravinhadawale95@gmail.com
[3]BE Student, Jaihind College of Engineering, Kuran, Junnar, Pune, hasanshaikh83532@gmail.com
[4]Assistant Professor, Jaihind College of Engineering, Kuran, Junnar, Pune, skolase@gmail.com

---

Abstract - *Product quality certification is used by industries to sell or advertise their products. The quality of wine is assessed by a human specialist, which is a time-consuming process that makes it quite expensive. Several machine learning techniques have already been applied to evaluate wine qualities such as quality and class on wine quality datasets. The quality of wine is determined not only by the amount of alcohol in it, but also by many traits, which change through time and therefore refine the wine's quality. It is critical to establish the wine's quality and categorise it into several categories based on a quality assessment. This study employs a variety of machine learning algorithms to predict wine quality. This research gives a comparison of fundamental and technical analysis based on many characteristics. This research compares and contrasts several prediction algorithms used to predict wine quality. Technical analysis such as time series analysis and machine learning algorithms like Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Ada Boost Classifier, and Gradient Boosting Classifier are examples of these methodologies. With the use of visualisation, several techniques are evaluated based on methodologies, datasets, and efficiency.*

Key Words: **Machine Learning, Wine Quality, Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Precision, Recall, F1-Score, Accuracy**.

1.INTRODUCTION

To improve product quality, testing is a key element that ensures product quality. Today, different types of companies are embracing and implementing new technology to verify and assess product quality. Testing the quality of a product with human expertise is an expensive and time-consuming operation that takes time to complete. For wine quality assurance, this study investigates various machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Ada Boost Classifier, and Gradient Boosting Classifier. These strategies automate the quality assurance process by minimising human interference and utilising accessible product attributes. The research also highlights the key characteristics that can be used to forecast the values of dependent variables. One of the major factors that can be utilised for certification is wine quality assessment, and this sort of quality certification helps to ensure wine quality in the market. Fixed acidity, volatile acidity, citric acid, residual sugar chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, and alcohol are the input variables in the red wine data set. The quality is measured on a scale of one to ten, with a greater value indicating higher wine quality.

The following is how the paper is structured: The second section describes similar work in this field. The proposed technique, the dataset used, and the machine learning algorithm are all discussed in detail in Section 3. Section 4 explains the experimental data and analysis. In section 5, the conclusion is drawn. Characteristics to predict the values of dependent variables.

1.1 Objectives

The first goal is to test various classification systems to discover which one provides the best results. The second goal is to figure out which characteristics are most indicative of a high-quality wine.

2. RELATED WORK

Although the majority of studies employed machine learning algorithms to assess wine quality, there is still a lot of room for improvement. Gupta, Yogesh [1] The study looks at how to use linear regression to find significant features for prediction, as well as how to use neural networks and support vector machines to forecast values. The use of machine learning techniques is explored in two ways in this study. To begin, consider how linear regression determines key features for prediction. Second, neural networks and support vector machines are used to forecast values. Terence Shin is a writer and director.[2] built multiple classification models using Kaggle's red wine quality dataset to predict whether a given red wine is "excellent quality" or not. Devika Pawar and her colleagues.[3] To predict wine quality, researchers employed Logistic Regression, Stochastic Gradient Descent, Support Vector Classifier, and Random Forest. The analysis demonstrates that quality improves as residual sugar is moderate and does not alter dramatically, indicating that this trait is not as important as others such as alcohol and citric acid. A. Saini and colleagues.[4] Long Short Memory Neural Network (LSTM NN) produced superior results than other strategies after doing fundamental and technical examination of numerous algorithms used for predicting future stock market prices. Yunhui Zeng and colleagues.[5] This study examines the impact of physical and

chemical indicators of wine grapes and wine on wine quality, constructs a wine quality analysis and assessment model, and investigates the influence of physical and chemical indicators of wine grapes and wine on wine quality. The results of this study's multiple linear regression analysis of wine quality show that there is a positive correlation linear association between the scores of wine quality aroma and C2H6O, C6H12O2, and other compounds. Jambhulkar and his colleagues. [6] Using a wireless sensor network, researchers employed a variety of strategies to predict cardiac disease. They gathered data from the Cleveland dataset and extracted key traits to predict heart disease. Zaveri and colleagues. [7] Using data mining techniques, several diseases such as cancer, tuberculosis, diabetes, and others were predicted.

2. PROPOSED METHODOLOGY

Machine learning techniques are employed to predict wine quality in this study. The processes in the suggested methodology is depicted in Figure 1. Pre-processing is done on the first wine dataset. The data is further divided into training (80%) and testing (20%) sets, with the training set being utilised to train the model utilising Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine algorithms. The testing set is used to determine the accuracy of several models, and then conclusions are generated to choose the optimal model for predicting wine quality. The trained model is used to determine the testing set's correctness. The accuracy of various algorithms is assessed and compared in order to determine the optimum algorithm for predicting wine quality.

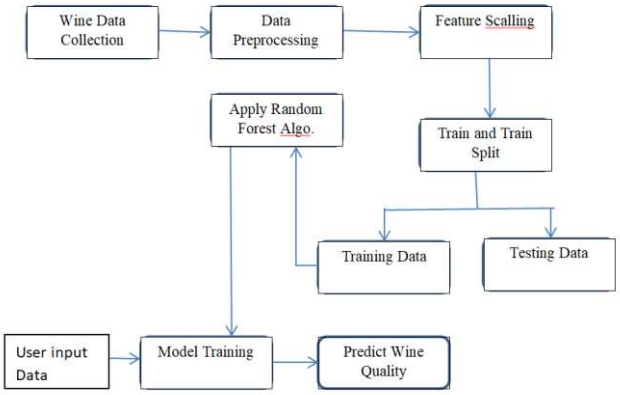

Fig -1: **Proposed System Architecture**

2.1 DataSet

The dataset for this study is a collection of red wines. For Red Wine, there are a total of 1599 samples. Fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol, and quality rating are all included in each sample. The quality classes run from 0 to 10, with 0 being the worst and 10 being the best. Wine collections cannot be used without preprocessing due to several flaws in the dataset. The huge amplitude of variable values, such as sulphates (0.3–2) vs. sulphur dioxide (1–72), is one of the primary flaws, and there are some missing numbers. The mean is used to fill in the missing numbers. The inconsistency in the dataset has an impact on predictions due to the influence of particular variables; this inconsistency is resolved using a linear transformation, which divides all the input values by the maximum variable value. The quality of the wine is translated to a binary output. '1' denotes an excellent quality wine with a score of 7 or above, while '0' denotes a poor quality wine with a score of less than 7.

2.2 Machine Learning Technique

2.2.1 Logistic Regression is a supervised learning technique that uses a machine learning algorithm. It's a method for predicting a categorical dependent variable from a set of independent variables. The output of a categorical dependent variable is predicted by logistic regression, and the outcome must be a categorical value. It can be 0 or 1, Yes or No, True or False, and so on, but instead of giving exact values, it delivers probabilistic values that fall between 0 and 1. The equation for the straight line can be written as (1).

$$y=b_0 + b_1 x_1+b_2 x_2+\ldots b_n x_n \quad (1)$$

Because y in logistic regression can only be between 0 and 1, we can divide the above equation by (1-y) we will get equation (2).

$$y/(1-y);\ 0 \text{ for } y=0 \text{ and infinity for } y=1 \quad (2)$$

However, we require a range of -[infinity] to +[infinity], in which case the logarithm of the equation becomes (3).

$$log\frac{y}{1-y} = b0 + b1\ x1 + b2\ x2 + \cdots bn\ xn \quad (3)$$

The above equation is the final equation for Logistic Regression.

3.2.2 Support Vector Machine is a supervised machine learning technique that can be used to solve problems in classification and regression. It is, however, mostly employed to solve categorization difficulties. In a high or infinite dimensional space, SVM creates a hyperplane that can be utilised for classification, regression, or other tasks.

The SVM algorithm's goal is to find the optimum line or decision boundary for categorising n-dimensional space into classes so that additional data points can be readily placed in the correct category in the future. A hyperplane is the name for the optimal choice boundary. To segregate the classes in n-dimensional space, a hyperplane can be several lines or decision boundaries, however to categorise the data points, we must pick the optimum decision boundary. The hyperplane of SVM refers to the best boundary. Support Vectors are the data points or vectors that are closest to the hyperplane and affect the hyperplane's position. The extreme points or vectors that assist create the hyperplane are chosen via SVM. Support vectors are the extreme instances, and the algorithm is called a Support Vector Machine. Consider the diagram below, which shows how a decision boundary or hyperplane is used to classify two different groups. The Support Vector Machine Hyperplane is shown in Figure 2. Positive hyperplane and negative hyperplane are the two types of data points that the Support Vector divides.

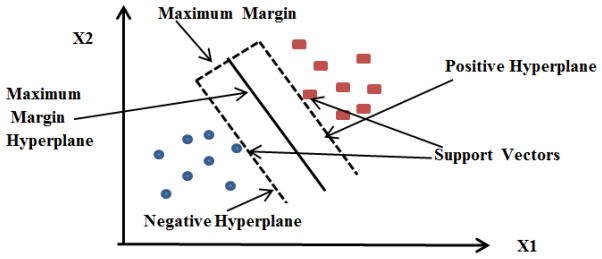

Fig -2: Support Vector Machine Classifier

3.2.3 Decision Tree is a supervised learning technique that can be used to solve problems in classification and regression. It is, however, mostly employed to solve categorization difficulties. It's a tree-structured classifier, with leaf nodes representing outcomes, interior nodes representing dataset attributes, and branches representing decision rules. Decision trees need less effort for data preparation during pre-processing than other methods. Data normalisation and scaling are not required when using a decision tree. In addition, missing values in the data have little impact on the decision tree-building process. A slight change in the data, on the other hand, can result in a significant change in the structure of the optimal decision tree. Calculations can become quite complicated, especially when multiple values are uncertain and/or multiple outcomes are related. The biggest problem with implementing a decision tree is figuring out how to make it work(4).

Information Gain= Entropy(S)-
[(Weighted Avg) *Entropy(each feature) (4)

Entropy is a metric for determining the degree of impurity in a particular property. It denotes the randomness of data. The Gini Index equation is as follows (5).

$$\text{Gini Index} = 1 - \sum_j P_j^2 \quad (5)$$

The Gini index is a measure of impurity or purity used in the Classification and Regression Tree (CART) technique to create a decision tree.

Figure 3 depicts a decision tree, with internal nodes representing dataset attributes and branches representing decision rules.

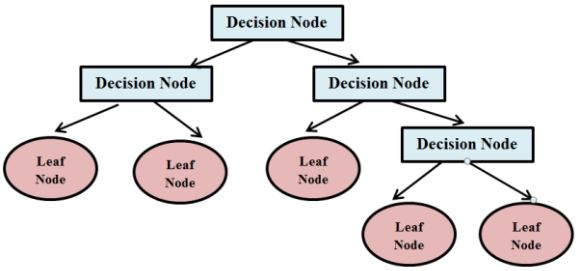

Fig-3: Decision Tree Classifier

3.2.4 Random forest is a supervised learning method that can be used to solve problems in classification and regression. It creates a "forest" out of an ensemble of decision trees, which are commonly trained using the "bagging" method. The bagging method combines several learning models to improve the final outcome. Random Forest is a learning method that employs the construction of many decision trees to achieve its results. The random forest makes the final selection, which is based on the majority of the trees. Random Forest produces n number of decision trees by randomly selecting records from a dataset, as shown in Figure 4. Instead than relying on a single decision tree, the random forest collects the forecasts from each tree and predicts the final output based on the majority votes of predictions. The greater the number of trees in the forest, the higher the accuracy and the less chance of overfitting.

The Random Forest algorithm has two stages: the first is to generate the random forest, and the second is to produce a prediction using the random forest classifier that was created in the first step.

1. Randomly select "K" features from total "m" features where k << m.

2. Among the "K" features, calculate the node "d" using the best split point.

3. Split the node into daughter nodes using the best split.

4. Repeat the a to c steps until "l" number of nodes has been reached.

5. Build forest by repeating steps a to d for "n" number times to create "n" number of trees.

In the next stage, with the random forest classifier created, we will make the prediction.

1. Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome.

2. Calculate the votes for each predicted target.

3. Consider the high voted predicted target as the final prediction from the random forest algorithm.

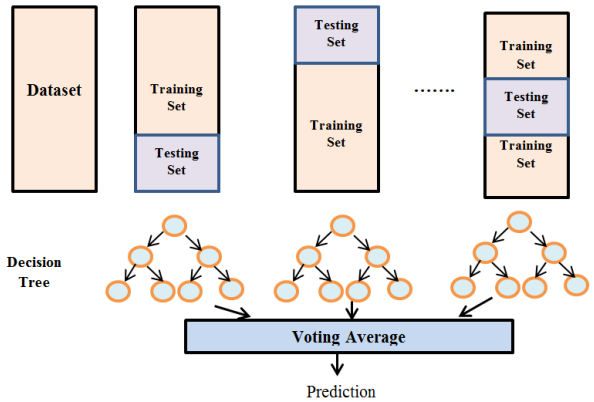

Fig-4: Random Forest Algorithm Classifier

4. EXPERIMENTAL RESULT AND ANALYSIS

There are total *12* variables in red wine collections as discussed in above section. The variable quality rating is considered as dependent variable and other *11* variables are assumed as predictors or independent

variables in this work. The distribution of the quality variable is shown in figure 5. The accuracy of different machine learning algorithm is shown in below tables. The performance of the classification models for a given set of test data is drawn by using confusion matrix. It can only be determined if the true values for test data are known.In, information retrieval and classification in machine learning, precision is also called positive predictive value which is the fraction of relevant instances among the retrieved instances, while recall is also known as sensitivity which is the fraction of relevant instances that were retrieved. Both precision and recall are therefore based on relevance. In statistical hypothesis testing, a type-I error is the rejection of a true null hypothesis is also known as a "false positive" finding or conclusion for example an innocent person is convicted, while a type-II error is the non-rejection of a false null hypothesis is also known as a "false negative" finding or conclusion for example a guilty person is not convicted. The different term used are described below:

Classification Accuracy: It defines how often the model predicts the correct output. It is one of the important parameters to determine the accuracy of the classification problems. The classification accuracy can be calculated as the ratio of the number of correct predictions made by the classifier to all number of predictions made by the classifiers. The formula for classification accuracyis given below (12).

$$Classification\ Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (4)$$

Precision: It can be defined as the number of correct outputs supplied by the model or the percentage of all positive classes that the model correctly predicted being true. The formula for precision is given below (13).

$$Precision = \frac{TP}{TP+FP} \quad (5)$$

Recall: It is defined as the percentage of positive classes that our model accurately predicted. The number of people who must be recalled must be as great as possible.The formula for recall is given below (14).

$$Recall = \frac{TP}{TP+FN} \quad (6)$$

F1-score: When two models have low precision but great recall, or vice versa, it's difficult to compare them. The F-score allows us to assess both recall and precision simultaneously. If the recall equals the precision, the F-score is maximum. The formula for F1-score is given below (15).

$$F1-score = \frac{2*Recall*Precision}{Recal+Precision} \quad (7)$$

True Negative (TN) means that the model predicted No, and the real or actual value likewise predicted No. True Positive (TP) indicates that the model correctly predicted yes and that the actual value was correct as well. False Negative (FN) is a Type-II error in which the model predicted no but the actual value was Yes. False Positive (FP) occurs when the model predicts Yes but the actual result is No. A Type-I mistake is another name for it. Chart-1 depicts the distribution of quality attributes and their counts.

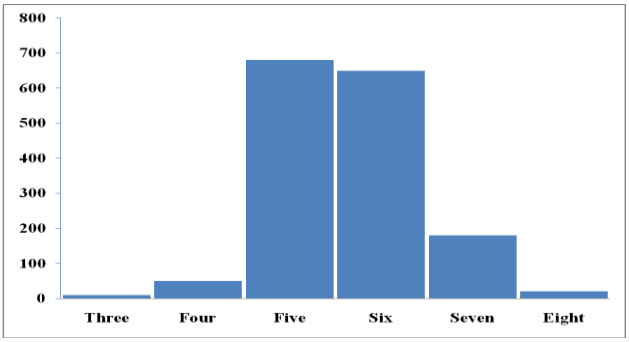

Chart -1: Distribution of the quality variable

4.1 Performance Matrix and Accuracy

The following table shows the performance matrix and accuracy for Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine.

We were able to get a 90% accuracy rate using Logistic Regression. Table 1 shows the Logistic Regression Algorithm Performance Matrix.

Table1: Performance Matrix for Logistic Regression Algorithm

	precision	recall	f1-score	support
0	0.92	0.96	0.94	355
1	0.55	0.38	0.45	45
accuracy			0.90	400
macro avg	0.74	0.67	0.69	400
weighted avg	0.88	0.90	0.89	400

The Decision Tree Classifier had a 90% accuracy rate. Table 2 shows the Performance Matrix for the Decision Tree Classifier Algorithm.

Table2: Performance Matrix for Decision Tree Classifier Algorithm

	precision	recall	f1-score	support
0	0.96	0.92	0.94	355

1	0.53	0.73	0.62	45
accuracy			0.90	400
macro avg	0.75	0.83	0.78	400
weighted avg	0.92	0.90	0.90	400

The Random Forest Classifier had a 92 percent accuracy rate. Table 3 shows the Random Forest Classifier Algorithm Performance Matrix.

Table3; Performance Matrix for Random Forest Classifier Algorithm

	precision	**recall**	**f1-score**	**support**
0	0.95	0.97	0.96	355
1	0.68	0.58	0.63	45
accuracy			0.92	400
macro avg	0.82	0.77	0.79	400
weighted avg	0.92	0.92	0.92	400

The accuracy of the Support Vector Machine was 90%. Table 4 shows the Support Vector Machine Algorithm Performance Matrix.

Table4; Performance Matrix for Support Vector Machine Algorithm

	precision	**recall**	**f1-score**	**support**
0	0.92	0.97	0.94	355
1	0.58	0.31	0.41	45
accuracy			0.90	400
macro avg	0.75	0.64	0.67	400
weighted avg	0.88	0.90	0.88	400

4.2 Comparison of Accuracy by different Algorithm

Physiochemical variables such as fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, and alcohol variables are used to train various machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Ada Boost Classifier, and Gradient Boosting Classifier. In this study, the variable quality rating is treated as a dependent variable, whereas the other 11 variables are treated as predictors or independent variables. This study uses two types of analysis: first, the value of each algorithm in predicting wine quality is determined, and second, the feature section is completed using the best predictors. The accuracy provided by various algorithms is displayed in the table below. The accuracy of the Logistic Regression is 90%, that of the Decision Tree Classifier is 90%, and so on. The Random Forest Classifier Algorithm has the best accuracy (92%) of all the algorithms.

Table7. Comparison of Accuracy obtained by different Algorithm

Sr. No.	**Algorithm**	**Accuracy**
1	Logistic Regression Algorithm	90%
2	Decision Tree Classifier Algorithm	90%
3	Random Forest Classifier Algorithm	92%
4	Support Vector Machine Algorithm	90%

The Random Forest algorithm generates a large number of decision trees, each of which is trained with a distinct dataset. Data is sent to each decision tree to anticipate some output in order to forecast wine quality. To anticipate the eventual outcome, the majority of votes are calculated. Finally, the Random Forest Classifier Algorithm can be described as a superior machine learning technique for predicting wine quality based on outcomes. At the same time, the Random Forest Classifier Algorithm can provide more exact predictions by employing selected predictors rather than all predictors.

3. CONCLUSIONS

In recent years, there has been an increase in interest in the wine sector, necessitating its expansion. As a result, companies are investing in innovative technologies to boost wine production and sales. Wine quality certification is crucial for a product's marketability, and it necessitates human wine testing. This research looks into several machine learning techniques for predicting wine quality. This study shows how the results alter when the test mode is changed for each categorization model. The analysis of classifiers on red wine datasets is part of the research. The percentage of correctly identified cases, precision, recall, and F measure are all used to explain the results. Different classifiers are tested on datasets, including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, Ada Boost Classifier, and Gradient Boosting Classifier. The results of the studies lead us to believe that the Random Forests Algorithm outperforms other classifiers in classification tasks. The Random Forest Algorithm predicts wine quality with a maximum accuracy of 92 percent. We can see that good quality wines have higher alcohol levels on average, higher sulphate levels on average, lower volatile acidity on average, and higher residual sugar levels on average. The study reveals that instead of evaluating all aspects, just essential features can be used to predict the value of the dependent variable with more accuracy. In the future, a huge dataset may be used for research, and various machine learning algorithms for wine quality prediction can be investigated.

REFERENCES

[1] Gupta, Y. (2018), Selection of important features and predicting wine quality using machine learning techniques, Procedia Computer Science, 125, 305–312, doi:10.1016/j.procs.2017.12.041.
[2] Shin, T. (2020, May 8), Predicting Wine Quality with Several Classification Techniques Medium, https://towardsdatascience.com/predicting-wine-quality-with-several-classification-techniques-179038ea6434.
[3] Devika Pawar, Aakanksha Mahajan, Sachin Bhoithe (2019), Wine Quality Prediction using Machine Learning Algorithms, International Journal of Computer Applications Technology and Research, Volume 8–Issue 09, 385-388, ISSN:-2319–8656.
[4] Saini, A., & Sharma, A. (2019), Predicting the Unpredictable: An Application of Machine Learning Algorithms in Indian Stock Market, Annals of Data Science, doi:10.1007/s40745-019-00230-7.
[5]Zeng, Y., Liu, Y., Wu, L., Dong, H., Zhang, Y., Guo, H., Guo, Z., Wang, S.,Lan, Y. (2018). Evaluation and Analysis Model of Wine Quality Based on Mathematical Model. Studies in Engineering and Technology, 6(1), 6, doi:10.11114/set.v6i1.3626.
[6] Jambhulkar and Baporikar. (2015), Review on Prediction of Heart Disease Using Data Mining Technique with Wireless Sensor Network,International Journal of Computer Science and Applications 8 (1) 55-59.
[7] Zaveri, and Joshi. (2017), Comparative Study of Data Analysis Techniques in the domain of medicative care for Disease Predication,International Journal of Advanced Research in Computer Science 8 (3) 564-566.
[8] Er, Y. (2016), The Classification of White Wine and Red Wine According to Their Physicochemical Qualities, International Journal of Intelligent Systems and Applications in Engineering, 4(Special Issue-1), 23–26, doi:10.18201/ijisae.265954.