Red Wine Quality Prediction Using Machine Learning

Ch. Sai Sri Ram[1], J. Avinash[2], K. Raghu Ram Sri Rishik[3], V. Narendra Reddy[4],

A. Thanuja[5]

[1,2,3,4]Student, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India

[5]Professor, Department of CSE, Narasaraopeta Engineering College, Narasaraopeta, Guntur (D.T) A.P, India

saisriramchunduri@gmail.com[1], jagarlamudirakesh2@gmail.com[2], rishik12457@gmail.com[3], vennann19@gmail.com[4], a.thanuja18@gmail.com[5]

1. ABSTRACT- The goal of this work was to create a model to forecast red wine quality based on its physicochemical characteristics. Several factors influence the accuracy of prediction while analysing the quality of red wine. This paper offers a computational intelligence method using machine learning techniques. In this instance, Random Forest Classifier, Naive Bayes Algorithm, and Support Vector Machine were used. With this information and these machine learning techniques, we can forecast the quality of a sample of red wine.

2. KEYWORDS: Red wine, Naive Bayes algorithm, Support vector machine, quality prediction, and Random forest classifier.

3. INTRODUCTION

Machine learning, a fast growing field of artificial intelligence, allows computers to automatically learn from experience and improve over time without explicit programming. Machine learning allows computers to examine massive volumes of data, spot patterns and trends, and then use that information to predict the future or make decisions. There are many useful uses for machine learning speech recognition, and personalised recommendations. Machine learning is anticipated to have a significant impact on many businesses and facets of daily life as it develops.

Red wine quality prediction using machine learning seeks to increase the precision and effectiveness of processing.

Machine learning models can be trained to recognise patterns and forecast the likelihood that a claim will be approved or refused by utilising historical data and prediction algorithms.

Machine learning, a cutting-edge field of research, enables computers to learn on their own using past data.

In order to build mathematical models and generate predictions based on previously collected data or information, machine learning employs a range of methodologies.

The purpose of utilising machine learning to forecast wine quality is to increase the precision and effectiveness of processing. Some machine learning software packages that can be used to create this system.

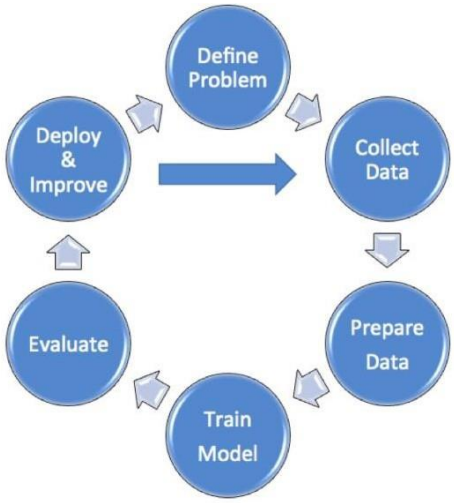

Fig.1 5 steps involved in Model

4. LITERATURE SURVEY

Existing literature was examined in order to gain the necessary knowledge on numerous ideas linked to the current use of our model. Through those, some of the most significant conclusions were drawn.

The research on estimating the quality of red wines utilising a variety of methodologies, including chemical analysis, machine learning, and sensory evaluation. It would go over each method's benefits and drawbacks and point out any gaps in the existing body of knowledge.

A description of red wine's chemical makeup and the elements that affect its quality. It would go over how different substances, including acids, sugars, and phenolics, affect the flavour, fragrance, and colour of red wine.

The numerous machine learning techniques that have been applied to forecast depends on the quality of red wine its physicochemical characteristics. It would emphasise the most important performance measures used to assess each algorithm's performance and go over its advantages and disadvantages.

The techniques used to choose the features that are most useful for forecasting the quality of red wine and to adjust the model's parameters for the greatest performance. It would also go over the difficulties and restrictions of these techniques and make recommendations for the future.

In this study, we attempt to forecast the quality of a sample of red wine using data from the dataset.

The model is trained using the machine learning algorithms Random Forest Classifier, Naive Bayes Algorithm, and Support Vector Machine, where we have achieved accuracy up to 80% using RFC, 72% with Naive Bayes, and nearly 57% with Support Vector Machine.

5. MATERIALS AND METHODOLOGY

Our model is suggested based on the following characteristics.

5.1 Dataset Analysis: We downloaded the datasets from the Kaggle website on the internet, and a dataset of red wine quality is a must if we are to make any predictions.

The dataset includes 12 columns: citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, alcohol, quality, fixed acidity, and volatile acidity.

Column	Description
fixed acidity	Sample's fixed acidity
volatile acidity	Sample's volatile acidity
citric acid	Sample's citric acid
residual sugar	Sample's residual sugar
chlorides	Sample's chloride
free sulfur dioxide	Sample's free sulfur dioxide
total sulfur dioxide	Sample's total sulfur dioxide
density	Sample's density
pH	Sample's pH
sulphates	Sample's sulphates
alcohol	Sample's alcohol
quality	Sample's quality

5.1. Data set

To better comprehend the features, we created a graphical representation of the dataset.

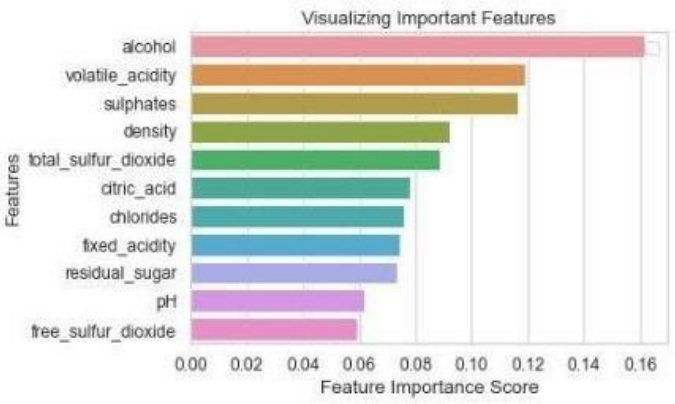

Fig 5.2. Features Visualizations

The dataset above shows several quality values within the specified range.

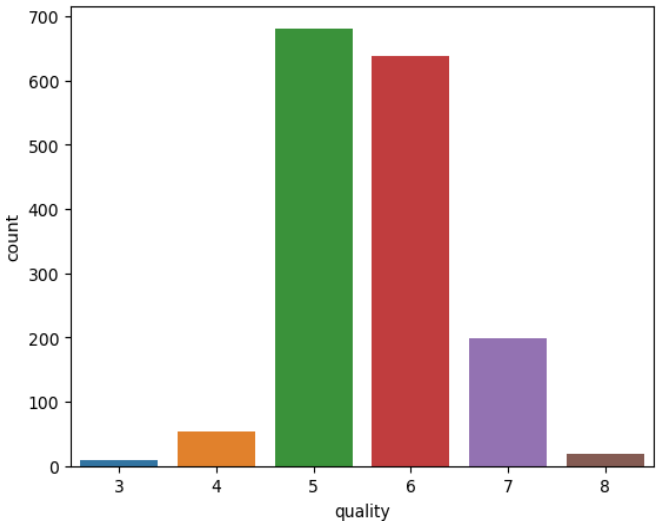

Fig 5.3. Quality

The quality values across the ranges are depicted in the above diagram.

Making adjustments to our data before submitting it to the algorithm is known as pre-processing. a procedure for turning raw data into uncleaned data. In comparable, whenever data are gathered from various sources, they are collected in raw format, making analysis impossible. Data must be in a specific format because of a particular machine learning algorithm. For instance, null values must be handled from the original raw data set in order to apply the Random Forest method because the Random Forest algorithm does not allow null values. A data set should be organised so that many algorithms for machine learning can be utilised to achieve the best outcomes.

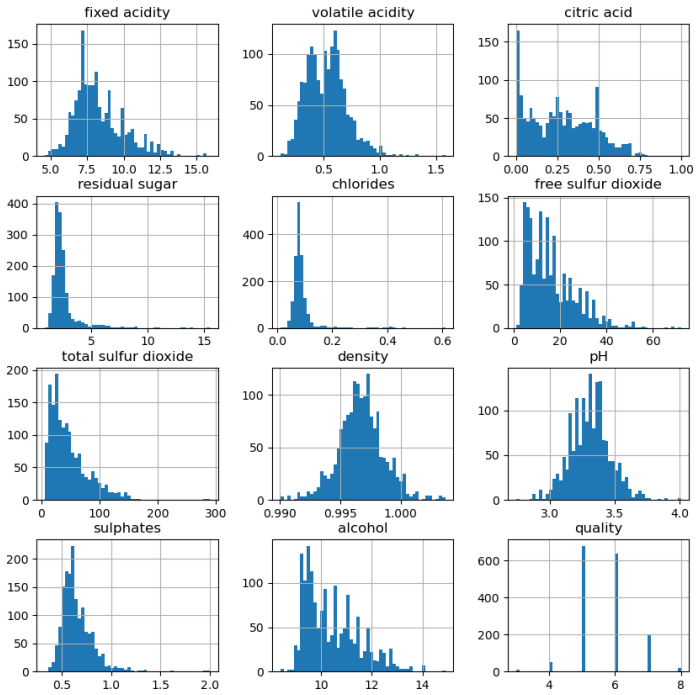

Fig 5.4. Histogram Graph

Eventually, it is discovered that the data contains some outliers. For a clear understanding of the outliers for each column, we employed a boxplot. After utilising percentile to remove the irrelevant data, a cleaned dataset was produced.

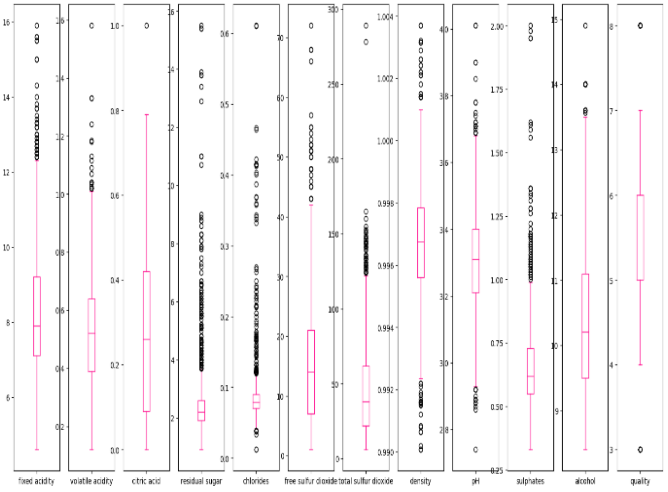

Fig. 5.5 Outliers Detection

We separated the quality feature values from the cleaned datasets into excellent and bad. To convert the excellent and bad values into binary values of 0 and 1, we used LabelEncoder(). Good is found to be transformed to the binary value 1 and bad to the binary value 0.

We divided the dataset for conducting the models into 70% for training and 30% for testing, and we removed the quality attribute because it is thought of as a goal variable.

Random Forest Classifier, Naive Bayes Algorithm, and Support Vector Machine are the tools we're employing for this. For training, a random forest classifier model is employed.

We are measuring both the accuracy of the model and the projected result using the sklearn accuracy score.

The Support Vector Machine and Naive Bayes Algorithm were also used in the same way.

These are the outcomes:

56.8% of support vector machines.

Algorithm of Naive Bayes, 71.7%

80.2% for Random Forest Classifier

Due to this accuracies, we discovered that Random Forest Classifier delivers the best among other two models . Therefore, we believed that this model was the most accurate for assessing the red wine sample's quality.

6. CONCLUSION AND FUTURE SCOPE

The goal of using machine learning methods was the study's goal to predict whether a red wine willbe good or awful. The analysis revealed a considerable improvement in the performance of the models, and we found that, Compared to the support vector machine and naive bayes methods,the random forest classifier has a greater accuracy. We selected a random forest classifier model because our goal was to forecast the quality of redwine.

Future research, however, can focus on exploring a number of additional deep learning applications.The technique that can converge the changes and do multiple frame work can be improved using improved approaches from machine learning and other fields.

7. REFERENCES

[1] Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, &José Reis. Modeling wine preferences by data mining from physicochemical properties. Decision Support Systems, 47(4), 547-553.

[2] Edelmann, Andrea , et al. "Rapid Method for the Discrimination of Red Wine Cultivars Based on Mid- Infrared Spectroscopy of Phenolic Wine Extracts." Journal of Agricultural & Food Chemistry49.3(2001):1139-1145.

[3] Zhang Shiling, Xu Ruimin. Formation and prevention of volatile acidin wine. New Rural Technology, 2008 (06): 81-82.

[4] Dahal, K., Dahal, J., Banjade, H., Gaire, S., 2021. Prediction of Wine Quality Using Machine Learning Algorithms. Open J. Stat. 11, 278–289.

[5] Rish, I., 2001. An Empirical Study of the Naïve Bayes Classifier. IJCAI 2001 Work Empir Methods Artif Intell 3.

[6] Moreno, Gonzalez-Weller, Gutierrez, Marino, Camean, Gonzalez and Hardisson. (2007) “Differentiation of two Canary DO red wines according to their metal content from inductively coupled plasma optical emission spectrometry and graphite furnace atomic absorption spectrometry by using Probabilistic Neural Networks”. Talanta 72 263–268.