**Wine Quality Prediction**

**Data and Web Mining Project Report**



**Department of Computer Science and Engineering**

**SRM University-AP, Andhra Pradesh, India**

**G. Bala Krishna – AP20110010033**

# Table of Contents

1. Abstract 3
2. Introduction 3-4
3. Related Work 4-5
4. Proposed scheme 5-7
5. Results 8-9
6. Conclusion 9-10

## 7. References 10-11

**Abstract:**

The quality of wine is extremely important to both consumers and the manufacturing industries. Product quality certification is helping businesses increase their sales. Wine is a widely consumed beverage around the world these days, and businesses rely on product quality certification to increase their market value. Formerly, product quality testing was done at the conclusion of the manufacturing process, which is a time-consuming procedure that necessitates a lot of resources, such as the requirement for multiple human experts to assess product quality, making this process very expensive. Every person has an opinion about the test, so determining the wine's quality based on human specialists is a difficult undertaking. As a result, our effort represents a step toward predicting wine quality based on its numerous features. Dataset is taken from Kaggle. This dataset contains the fundamental characteristics that have an impact on wine quality. We used techniques like Random Forest, Naive Bayes, KNN, Support Vector, and Logistic

Regression to predict wine quality.

**Introduction:**

Wine is generally categorized or differentiated based on color, smell, and taste. However, it is possible to determine whether a wine is of good or low quality. There are various features that were involved in the dataset that we downloaded. Because the qualities are chemical, let's go through them now for a better grasp of the process. The primarily fixed acids present in wine are tartaric, succinic, critical, and malic, which are described by the fixed acidity property. The feature Volatile Acidity represents the gaseous acids present in the wine. The amount of sugar left over after fermentation is known as residual sugar. Citric acid is a weak organic acid that can only be found in citrus fruits. The amount of salt in the wine is measured in chlorides, and free sulfur dioxide is utilized to avoid oxidation and microbial deterioration. Sulphates help to keep wine fresh by protecting it from oxidation and germs. The acidity of a wine is measured using the pH scale.

There is a feature called Type that indicates whether the wine is red or white. The target variable Quality value ranges from 1 to 10. The inputs include fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulphur dioxide, total sulphur dioxide, density, pH, sulphates, and alcohol, while the output is quality.

**Related Work :**

A variety of customers today appreciate wine to an ever-increasing extent. The wine industry is looking into new advances for both winemaking and offering structures in order to keep up with that development [1]. The arrangement of wines is significant for various reasons. Financial estimation of wine items, the protection, and enhancement of the quality of wine, preventing wine corruption, and controlling refreshment preparation are a few of these reasons [2]. Data mining innovations have been applied to forecast wine quality. In order to make wine quality predictions, machine learning techniques like a variety of applications use the information to create models. The UCI store was supplied with an informational index in 1991 with 178 entries with estimated amounts of 13 distinctive chemical compounds, including alcohol and magnesium. This was in order to buy three cultivars from Italy [3]. For new information mining classifiers, this data has been extensively used as a benchmark since it is relatively easy to separate. For wine characterization as indicated by geological area, Principal Component Analysis (PCA) was done and announced [4]. A total of 33 Greek wines with physical and chemical factors were analyzed in their study. Various studies and students have published relevant work in national and international research papers, and chapters to better understand the goal, types of algorithms employed, and various strategies for forecasting wine quality. In [5] random forest, support vector machine, and naive Bayes techniques were used to predict red wine quality based on various attributes [5]. Precision, recall, f1-score, accuracy, specificity, and misclassification error are some of the performance metrics they calculated. When comparing the support vector machine to the random forest and naive Bayes techniques, they found that the support vector machine produced the best results. Using three machine learning algorithms such as k-nearest-neighbourhood, random forest, and support vector machine, Er and Atasoy (2016) proposed a method to classify the quality of red and white wine. For feature selection, they used principal component analysis, and the random forest algorithm yielded the best results. Based on physicochemical data, P. Appalasamy et al. (2012) predicted wine quality. They used the decision tree and naive Bayes algorithms to analyze red and white wine datasets. They compare the classification approach can help improve wine quality during the production process, according to the results of these two algorithms.

**Proposed Scheme :**

Red wine variants of the Portuguese "Vinho Verde" wine are included in this dataset. As a result of privacy and logistic issues, only physicochemical (input) and sensory variables

are available (e.g., grape types, wine brands, wine selling prices, etc.).

These datasets can be considered classification tasks. They are arranged by class and not balanced (e.g., there are more standard wines than excellent or poor wines). Outlier detection algorithms could be a handy way to find those few excellent or poor wines. However, we are unsure whether all input variables are relevant. A feature selection

method may be an interesting experiment to test.

Our goal is to make a machine learning model that can predict the quality of the wine

based on the input variables (features) given above.

From the dataset we have divided the quality of wine into two buckets:

* Bad wine: range 2 - 6.5
* Good wine: range 6.5 - 8

This can be changed as per the requirement of our client. Now we will map the values of bad and good to 0 and 1 respectively, as machine learning models can perform

calculations only on numerical data.

Random forest is a supervised machine learning approach for solving classification and regression issues. The Random Forest model predicts high-quality wines with low overfitting problems and good accuracy thanks to the aggregation of decision trees. The Naive Bayes classifier is built on prior knowledge of possible event situations. The Bayes Theorem underpins it. There is a high level of independence between the assumed attributes. It calculates ratios between events using categorical data. To improve the performance in terms of classification accuracy, a Naive Bayes model was used to increase the diversity of the trees in the forest. Support vector machines discover natural splits in the data that allow the largest consistent margin from the path of the chosen function. Support vector machines are more flexible because they can change the discriminating function type by implementing different kernels. Kernels identify how the support vector machine should separate the data. The KNN algorithm performs classification using the nearest training examples in the feature space. KNN's performance is determined by its K value as well as the distance metric used, such as the Euclidean distance

A confusion matrix is a table that shows how well a classification model (or "classifier") performs on a set of test data for which the true values are known.

The values of the confusion matrix can be defined as :

true positives (TP): Predicted value is yes and actual value is also yes true negatives (TN): Predicted value is no and actual value is also no false positives (FP): Predicted value is yes but actual value is also no false negatives (FN): Predicted value is no but actual value is also yes Confusion matrix of Logistic Regression Classification:

[[276 14]

[ 18 12]]

Confusion matrix of K-Nearest Neighbour Classification:

[[280 10]

[ 14 16]]

Confusion matrix of Support Vector Classification:

[[285 5]

[ 21 9]]

Confusion matrix of Naive Bayes Classification:

[[241 49]

[ 7 23]]

Confusion matrix of Random Forest Classification:

[[283 7]

[ 9 21]]

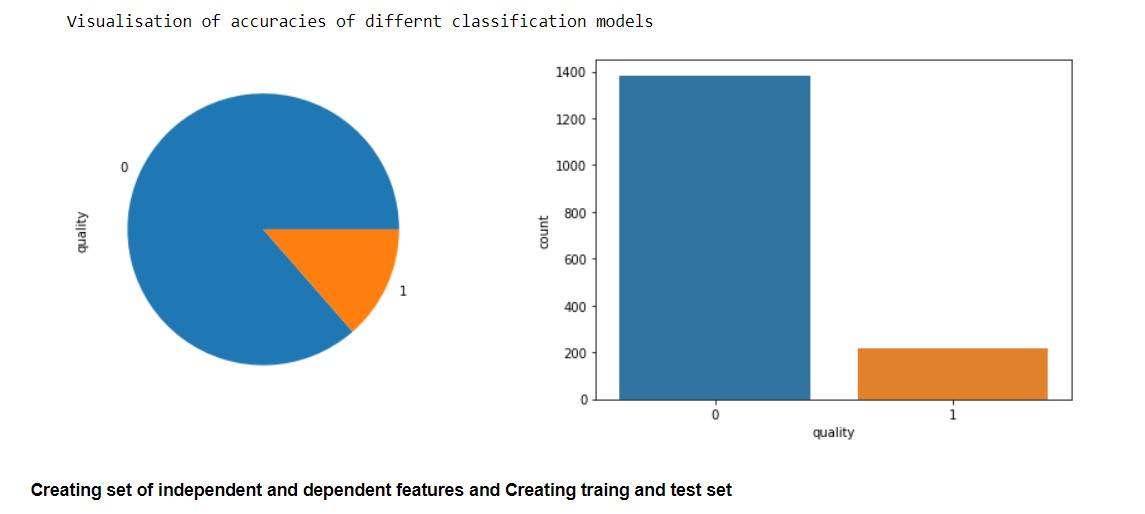
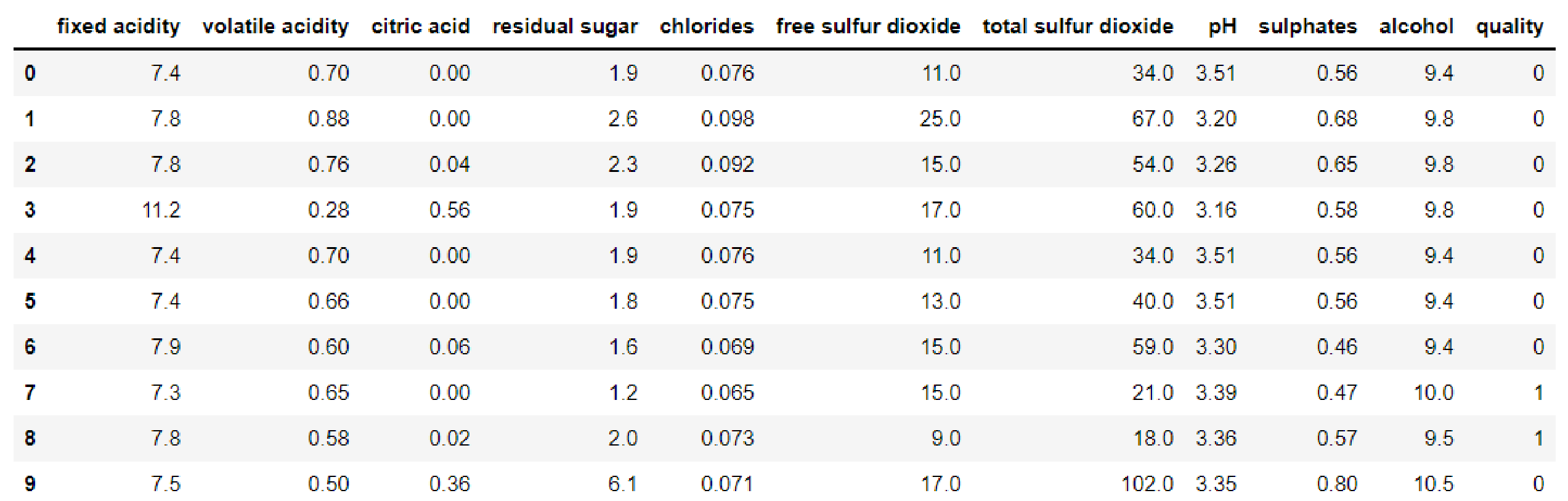
Accuracy of Logistic Regression Classification model is 0.87

Accuracy of K-Nearest Neighbour Classification model is 0.87

Accuracy of Support Vector Classification model is 0.87

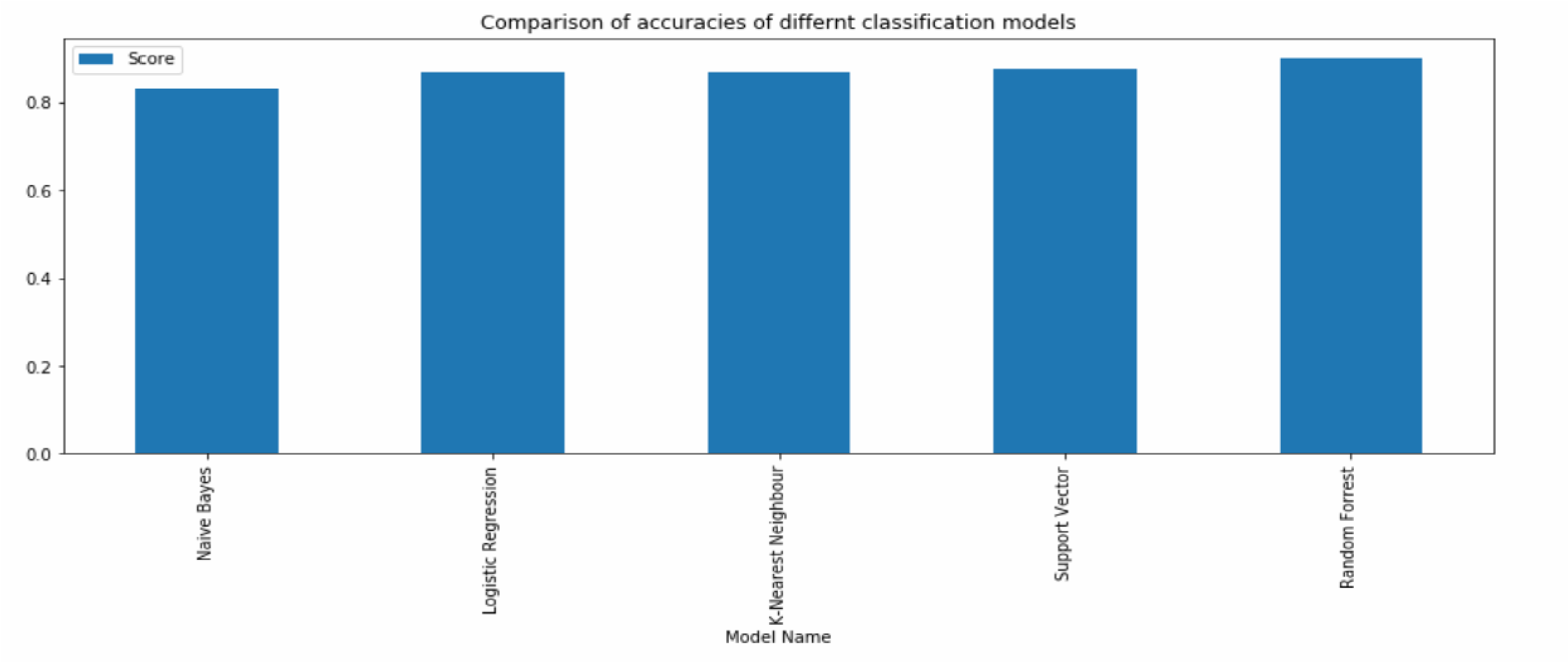
Accuracy of Naive Bayes Classification model is 0.83

Accuracy of Random Forest Classification model is 0.90

**Results :**

From the above scores and visualisations, we can conclude that the Random Forest Classification model gives the most accurate score. We can use it to predict the quality of wine for this particular problem.

Furthermore, other models, such as Logistic Regression, KNN, and SVC, have similar scores to Random Forest and can also help predict the wine quality of the wine.

**Conclusion :**

From the above data engineering and machine learning (classification) techniques we can conclude that:

1. We have chosen not to remove outliers and extract the more relevant features from our dataset - as we were looking for accuracy to minute levels, not just some approximation (high-quality wine may have a very unique composition from other average quality wines)
2. In this problem, the Random Forest Classification model gave the maximum accuracy and can be considered a reliable model for predicting the quality of the wine.
3. However other models like Logistic Regression, KNN, and SVC also have comparable scores to Random Forest and may also be used to predict the quality of the wine.
4. Naive Bayes model gave the least accuracy, which can be considered a bad model to predict the quality of fine.
5. Performance tuning using methods like Grid Search, etc. Can be done to improve the accuracy of these models. So, the accuracy of these models will improve and we might get another suitable model for our problem.
6. We might(will) get different results if we remove outliers and consider feature extraction.

Finally, I would like to end this notebook with the fact that no Data Science technique is perfect, there are many other ways/ models to get better results and there is always scope for improvements.

**References :**

* 1. P. Cortez, A. Cerderia, F. Almeida, T. Matos, and J. Reis, “Modelling wine preferences by data mining from physicochemical properties,” In Decision Support Systems, Elsevier, 47 (4): 547-553. ISSN: 0167-9236
  2. V. Preedy, and M. L. R. Mendez, “Wine Applications with Electronic Noses,” in Electronic Noses and Tongues in Food Science, Cambridge, MA, USA: Academic Press, 2016, pp. 137-151.
  3. A. Asuncion, and D. Newman (2007), UCI Machine Learning Repository, University of

California, Irvine, [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

* 1. S. Kallithraka, IS. Arvanitoyannis, P. Kefalas, A. El-Zajouli, E. Soufleros, and E. Psarra,

“Instrumental and sensory analysis of Greek wines; implementation of principal component analysis (PCA) for classification according to geographical origin,” Food Chemistry, 73(4): 501-514, 2001.

* 1. S. Kumar, K. Agrawal, and N. Mandan, “Red wine quality prediction using machine learning techniques,” International Conference on Computer Communication and Informatics, pp. 1-6, 2020.