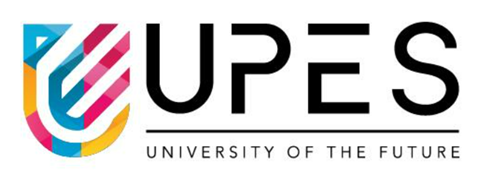
**School of Computer Science**

**UNIVERSITY OF PETROLEUM AND ENERGY STUDIES**

**DEHRADUN, UTTARAKHAND**

****

**Application of ML in Industries**

**LAB**

**Wine Quality Prediction**

**SEM 6**

**Submitted by: Submitted to:**

Nitika Singh Rohit Srivastava

Shreya Panwa

Khushi Sinha

**BTECH CSE AIML (H) B1**

|  |  |  |
| --- | --- | --- |
| Sr. no. | Topics | Page no. |
| 1. | Introduction | 3- 8 |
|  | 1. 1 Prologue | 3 |
|  | 1. 2 Problem statement | 7 |
|  | 1. 3 Objective | 8 |
|  | 1. 4 System requirements | 8 |
| 2. | System analysis | 8 |
|  | 1. 1 Motivation | 8 |
| 3. | Design | 9 |
|  | 1. 1 Flowchart diagram | 9 |
| 4. | Implementation | 9-14 |
|  | 1. 1 Methodology | 9 |
|  | 4.2 Algorithm | 11 |
|  | 4.3 Screenshots | 14 |
| 5. | Limitations | 17 |
| 6. | Future Enhancement | 17 |
| 7. | Conclusion | 17 |
| 8. | References | 18 |

Wine Quality Prediction

1. Introduction

1.1 Prologue

The demand for wine is steadily increasing on the international scene, with 31 million metric tons of wine delivered each year. The industry is forced by this intense competition to integrate cutting-edge technologies into all aspects of the production and sales operations. Businesses may now efficiently cater to customer preferences by utilizing machine learning (ML) and hybrid modeling techniques, which have emerged as significant tools for guaranteeing high-quality wine products.

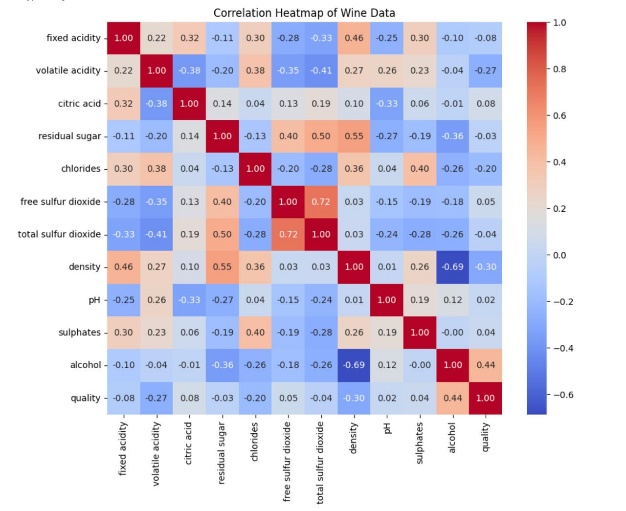


Fig 1: Correlation matrix of each feature of wine

This project focuses on employing machine learning to predict its quality. We examine various categorization methods and evaluate how well they can reliably determine the value of a wine. This study examines the use of many machines learning (ML) algorithms for wine quality prediction, including Decision Tree (DT) [1], Random Forest (RF) [2], Extreme Gradient Boosting (XGBoost) [3], and a few others. We train and assess the models using the Kaggle wine quality dataset, which is accessible to the general audience. Key performance indicators such as accuracy, precision, recall, and F1 score are used to evaluate each model's performance. We intend to provide producers and fans with a deeper comprehension and enjoyment of the liquid masterpiece through this thorough analysis.

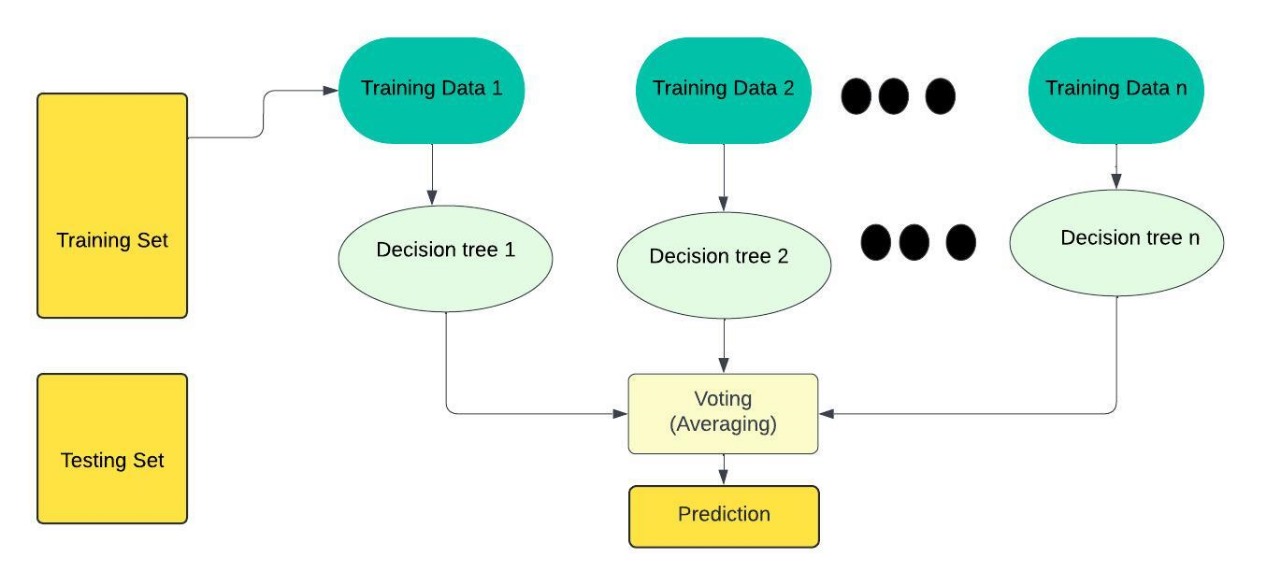


Fig 2: Machine prediction using Random Forest

Literature Review

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| S.NO | Title | Aim and dataset | features | methodology | Result |
| 1 | Machine learning-based predictive modeling for the enhancement of wine quality. | [3] The document describes a machine learning model developed to predict wine quality using a dataset of physiochemical properties. | The top four features are alcohol, sulfates, volatile acidity, and pH. | Five machine learning models were trained and tested, and the most accurate algorithms are Random Forest (RF) and XGBoost. | The XGBoost classifier displayed close to 100% accuracy. The RF algorithm produces the results with an accuracy of  81.96%. |
| 2 | *Wine Quality Prediction using Machine Learning and Hybrid Modeling.* | [4] Predicting wine quality by  eliminating outliers and testing feature selection methods on the dataset for future research. | The document includes information on the dataset used, exploratory data analysis, data preprocessing, and transformation. | implement Decision Tree, Random Forest, and Extreme Gradient Boosting algorithms | The Random Forest Classifier (RFC) achieved an accuracy of 85.57%.  The Decision Tree (DT) - 79.25%, the XGBoost - 78.07%, and the Hybrid Model of all three machine learning techniques achieved an overall accuracy of 77.71%. |
| 3 | A machine learning application in wine quality prediction. *Machine Learning with Applications* | [5] Dataset for Pinot Noir wines in New Zealand. | The study utilized 18 Pinot Noir wine samples with 54 different characteristics and generated 1381 samples using the SMOTE method. | Seven machine learning algorithms were trained and tested. Decision trees, random forests, and ensemble methods in machine learning. | Adaptive Boosting classifier showing 100% accuracy. Evaluation metrics such as accuracy, precision, sensitivity, specificity, F1 score, and ROC-AUC score are also discussed. |
| 4 | Analyze the Quality of Wine Based on a Machine Learning Approach | [6] The study proposes a machine learning-based model designed to predict the wine quality of a Kaggle dataset | Kaggle dataset | classification algorithms: logistic regression, decision tree, and random forest. | Remarkably, the decision tree algorithm achieved an impressive accuracy rate of 85% when applied to the given dataset, surpassing the performance of other models. |
| 5 | A study and analysis of machine learning techniques in predicting wine quality. | [7] Quality based on physicochemical tests. | Factors such as acidity, citric acid, alcohol content, and chlorides, aid in decision-making processes within the wine industry. | Decision Trees, Random Forests, Support Vector Machine, and K Nearest Neighbors, | Forest offer interpretability and robustness against overfitting, while SVM excels in capturing complex decision boundaries in high-dimensional spaces. KNN provides simplicity and quick predictions but may be computationally expensive for large datasets. |
| 6 | A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques | [8] The paper investigates wine quality prediction using machine learning | Acidity, alcohol level, chlorides, etc features were taken. | Various classification models, including, Bagging, Random Forest, and Boosting, are employed with Decision Trees. fRecursive Feature Elimination (RFE), and Principal Component Analysis (PCA) are used for feature selection to reduce dimensionality. | Performance metrics such as accuracy, sensitivity, specificity, and predictive values are used to evaluate the classifiers' effectiveness on red and white wine datasets. |
| 7 | Wine Quality Prediction Using Data Mining | [9]Traditionally, wine quality assessment relies on human experts who subjectively evaluate various factors like aroma, taste, and color. | The study gathers data on various wine samples, including attributes relevant to quality assessment. | Different data mining algorithms for classification are applied and compared: Naive Bayes, Simple Logistic, KStar, JRip, and J48. | It was found that the accuracy of Naive Bayes was 100% while the accuracy of Simple Logistic, KStar, JRip, and J48 was 97.22%, 97.22%, 94.44%, and 97.77% respectively. |
| 8 | Red wine quality prediction through active learning | [10] The research paper investigates the use of active learning algorithms to predict the quality of red wine. | Kaggle dataset with features like alcohol level, chlorides,etc was taken. | The paper discusses using the K-Nearest Neighbor algorithm and ranked batch-mode sampling for active learning. | The author found that the active learning approach was effective in predicting the quality of red wine with a smaller labelled dataset. |
| 9 | Wine Quality Detection through Machine Learning Algorithms | [2] The attributes of wine change with time and so the quality of wine changes also with time. The Paper uses machine learning algorithms to analyze these attributes. | The study takes many samples of wine factors that can be taken into consideration. | Logistic Regression and Random Forest Classifier were performed individually on data to predict the test data values. | It was realized that the random forest classifier gave better results of more than 98%. |
| 10. | Judging wine quality: Do we need expert consumers or trained panelists? | [1] The research paper uses the descriptive analysis method to reveal the quality drivers of wines entered in a wine competition. | A panel of trained tasters identified specific flavors and aromas in the wines. | Descriptive methods were used. | Consumers and wine experts all liked wines they thought were high quality. However, for some consumers, liking a wine didn't necessarily mean they thought it was high quality. Wine experts disagreed somewhat with the pre-assigned quality scores, and were better at describing why they liked the wines they chose. |

1.2Problem Statement

* The problem statement for this project is to develop a machine-learning model that forecasts wine quality using a dataset of the physiochemical properties of red and white wine.
* Inadequate comprehension of these models' interpretability impedes their usefulness in real-world scenarios and obscures important variables.

1.3 Objective

## Compare several classification algorithms to forecast wine quality & pinpoint key elements that influence it.

* Improve the estimation of wine quality by investigating feature selection and classification techniques.

1.4 System Requirements

**Hardware**

Processor: For effective data processing, an Intel Core i5 (or equivalent) or above is required.

Memory (RAM): To handle large dataset and computational workloads, at least 8 GB of RAM is required.

Internet Connectivity: Necessary for downloading datasets, libraries, and resources, as well as for collaboration and accessing online documentation.

**Software**

Programming Language: Python

Development Environment: Jupyter Notebook for code development and debugging.

**Libraries**

Scikit-learn offers various Machine-learning algorithms for regression and classification tasks.

NumPy and pandas for data manipulation and analysis.

Matplotlib or Seaborn for data visualization.

2. System Analysis

2. 1 Motivation:

There are many benefits to the possible use of machine learning to forecast wine quality.

* **Enhanced decision-making:** By pinpointing the key variables influencing wine quality, winemakers can optimize production processes.
* **Enhanced customer experience:** Customers can make well-informed choices based on predicted quality and individual preferences.
* **Sustainable practices:** Utilizing data-driven insights can promote wine production that is both ecologically friendly and efficient.

3. Design

3. 1 Flow chart diagram

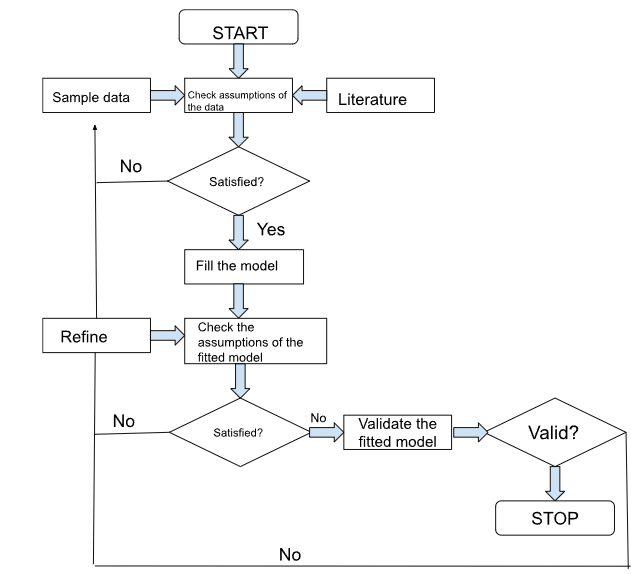


Fig 3: Flowchart

1. Implementation

4.1 Methodology

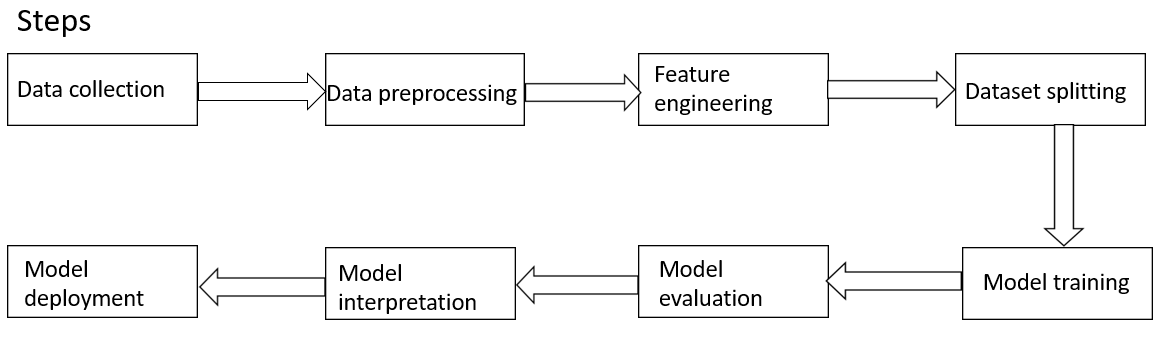


Fig: 4 Steps

1. **Data collection**- Gathering a dataset containing various attributes of wine like acidity, alcohol, pH, density, etc. along with their corresponding quality rating.
2. **Data preprocessing-** This is an important step as it ensures the dataset is clean and reliable. It involves several steps:
   1. Imputation or deletion are used to eliminate missing values. The column's mean and median can be used to fill a null item.
   2. Outlier detection makes sure the suggested classifier stays true to its initial prediction.
   3. The dataset's categorical variables are numerically encoded using methods like label encoding and one-hot encoding.
   4. A scale is applied to the numerical features to make sure that they are all on the same scale. Standardization, also known as mean normalization, and min-max scaling are popular methods.
3. **Feature engineering-** From the current features, new ones are generated that might have important data for determining wine quality. Feature selection involves identifying the most pertinent features that help predict the quality of wine by applying methods such as domain knowledge, correlation analysis, and feature importance ranking.
4. **Dataset splitting-** The dataset is split into training and testing sets. We have split the data in 80% for training and 20% for testing.
5. **Model training-** It involves several steps:
   1. On the training set of data, machine learning models such as random forests, logistic regression, XGBoost models, decision trees, and KNN are applied.
   2. Tune hyperparameters- To determine the ideal hyperparameters for every model, methods such as random or grid search are employed. Hyperparameters regulate the model's behavior and have a big effect on how well it works.
   3. Cross-validation: To obtain a more accurate assessment of the models' performance and guarantee that they translate well to new data, cross-validation is used to test the models.
6. **Model evaluation-** Depending on the needs of the problem, the models are assessed on the test set using relevant evaluation measures, such as accuracy [4]. To ascertain which model works best for predicting wine quality, a comparative analysis is conducted to evaluate each model's performance.

(eq.1)

TP = True Positive

TN = True Negative

FP = False Positive

FN = False Negative

1. **Model interpretation-** To learn more about the variables influencing wine quality forecasts, the trained model is interpreted. In the case of decision trees, this may entail analyzing feature importance and displaying the tree structure. Multiple trees can be used to aggregate feature importances for ensemble methods like as random forests and XGBoost [3].
2. **Model deployment-** After selecting the best-performing model, we can deploy it into production to make predictions on new data.

We analyze and compare the performance of several classification algorithms, including:

* Decision Trees
* Random Forests
* xgboost
* Logistic Regression
* KNN

We utilize the publicly available UCI Wine Quality dataset, comprising chemical and physical attributes of wine samples.

4.2 Algorithm

XGBoost Algorithm [11]:

1.Initialize: Xgboost initializes a prediction model which predict the target variable y.

2.Iteration:

a. Calculate the residuals which is the difference between predicted and actual values for each data points.

b. Fit a new model to these residuals.

3.Boosting: Update the model by adding the predictions from the new model, weighted by a learning rate to prevent overfitting.

‘n’ iterations can be performed until the residuals have been minimized as much as possible.

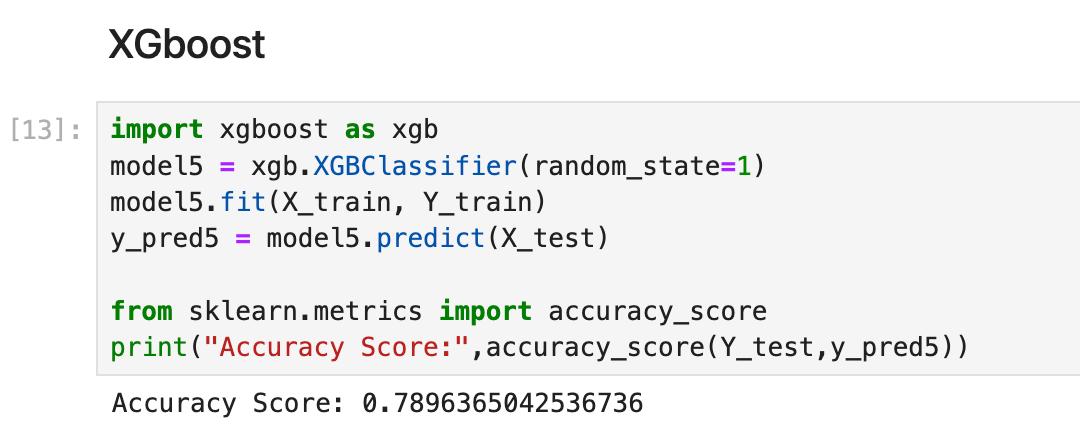


Fig 5: XGBoost implementation

Random Forest Algorithm [12]:

1. Choose k random data points from the training set.

2. Create Subsets: Initialize the decision tree associated with the chosen data points.

3. Select N number of decision trees required for the prediction.

4. Iterate steps 1 and 2.

5. For every new data point, find the predictions of each decision tree, and assign the new data points to the class based on the voting (Averaging) of the results from each decision tree.

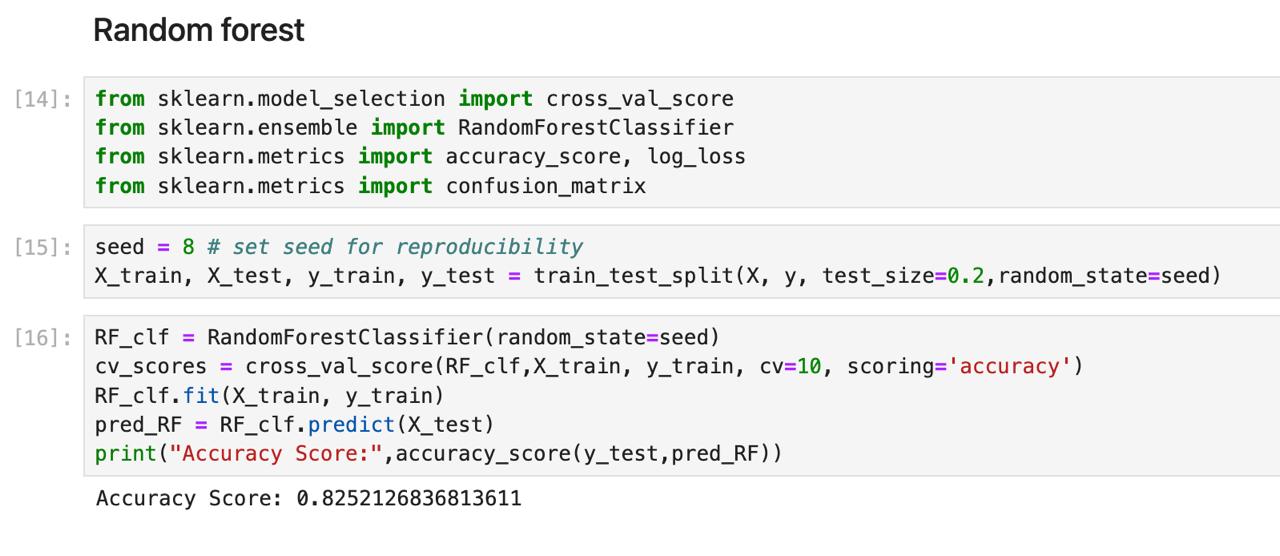


Fig 6: Random Forest implementation

Mathematical Explanation

To perform classification using a Decision Tree and Random Forest, the Gini index or Entropy [13] is used for splitting. These formulae help in deciding how nodes are on a decision tree branch.

(eq.2)

Here,

pi= relative frequency of the label i at a node

c = number of unique labels

The given formula uses class and probability to determine the Gini of each branch on a node, determining which branches are more likely to occur.

We can also use entropy to determine the branch on each node, using the following formula:

(eq.3)

Here, pi and c represent the relative frequency of the label i at a node and the number of unique labels respectively.

According to our dataset, we have six classes (quality rank ranging from 3 - 8), and the

pi= (eq.4)

Random Forest uses the potential of ensemble learning using multiple decision trees. These trees can be thought of as discrete specialists, each focused on a different facet of the information. Significantly, they function autonomously, reducing the possibility that the subtleties of a single tree would unduly impact the model.

Every decision tree in the Random Forest has an opinion regarding predicting outcomes. The mode, or most frequent forecast, across all the trees, determines the final prediction for classification tasks.

* 1. Screenshots

Figure 7 shows the accuracy score of all the models and among them Random Forest gives the best accuracy, therefore, for the further prediction we have used random forest classifier.

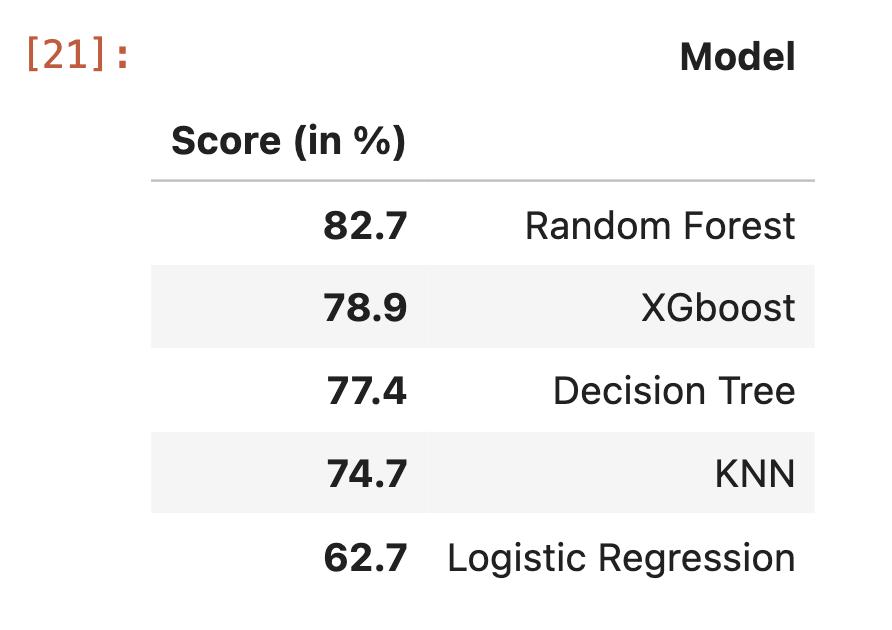
****

Fig 7: Accuracy scores of each model (in %)

Figure 8 shows the visualization of the comparison of accuracy between all the models we have used.

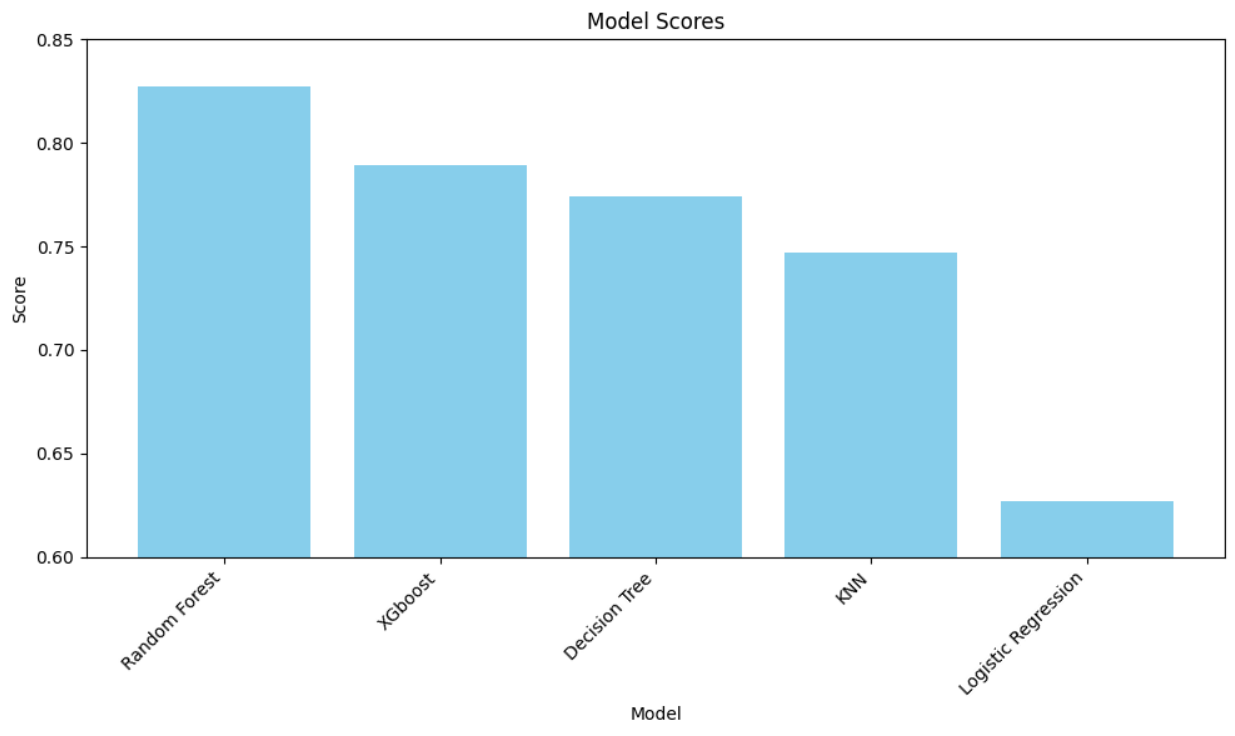
****

Fig 8: Model scores

**Comparative Analysis of the Machine Learning Models:**

We evaluated five machine learning models that can be used for wine quality prediction: Xgboost, Decision tree, Random Forest, KNN, and logistic regression. The most accurate ones were:

|  |  |  |
| --- | --- | --- |
| XGBoost | Decision Tree | Random Forest |
| This model has been chosen for its accuracy and cost-effectiveness while handling big data sets.  XGBoost surpasses other models such as Random Forests and Decision Trees because of its advanced boosting algorithm and capability to deal with complex feature interactions, which are not possible with the first two algorithms.  Its internal regularization strategies help prevent overfitting, making it a more generalizable model for real-world problems. | It can be considered a simpler alternative to XGBoost, despite the latter’s superiority in terms of accuracy.  Given that decision trees are presented in a graphical form, it becomes much easier to comprehend how they have arrived at a particular prediction and what features have played crucial roles in determining the quality level.They may take less time to compute on smaller data sets.  When the decision tree has been optimized, it can accurately predict outcomes for new and unseen data sets. Even small modifications to the information can have a profound effect on the prediction and structure of the tree. | It combines interpretability and competitive accuracy.  Random Forest often performs better than Decision Trees and can reach XGBoost-like accuracy after sufficient adjustment.  Despite being more challenging to illustrate than a single Decision tree, it provides feature-importance insights that aid in understanding significant quality factors.  Because it balances predictions from multiple trees, it is less prone to overfitting and more stable than single decision trees. |

GUI

Below in fig 2 and fig is our GUI, User Input Parameter takes the input of the user and the user has freedom to scroll the input, and accordingly the quality of wine changes. If the predicted quality is 1, that means, – good quality, if 0 – bad quality.

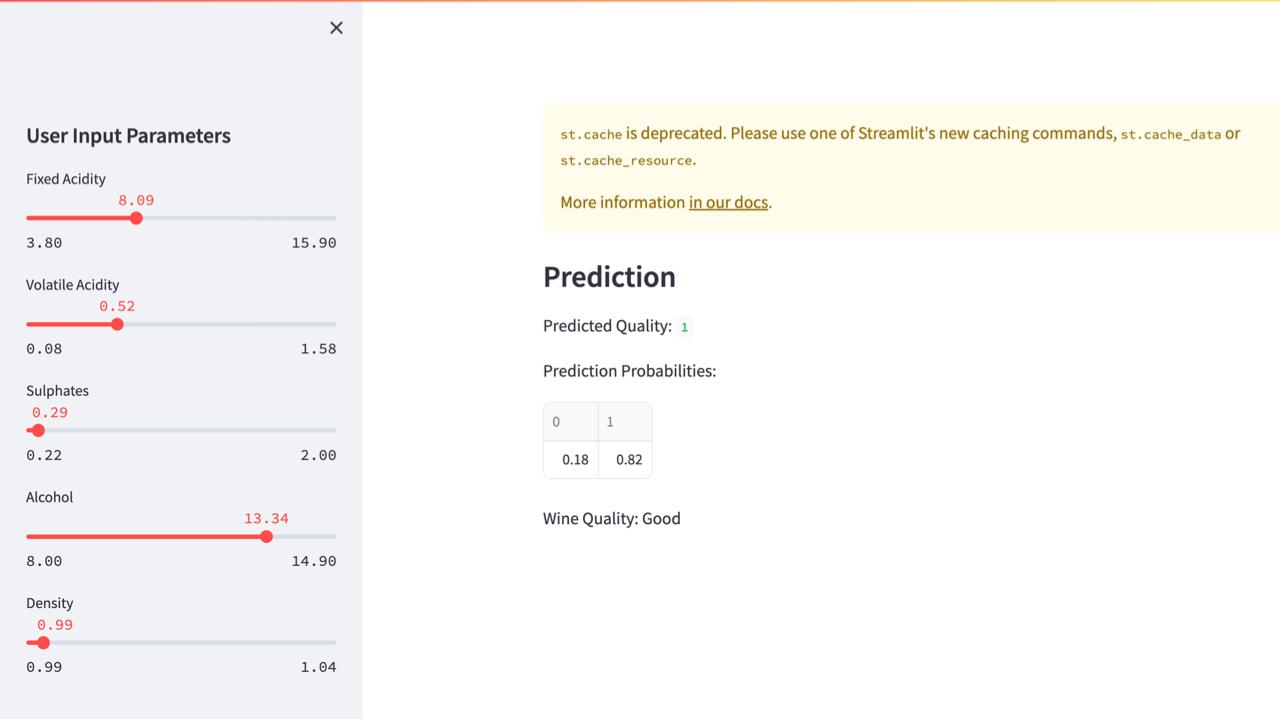
****

Fig 9: Good wine

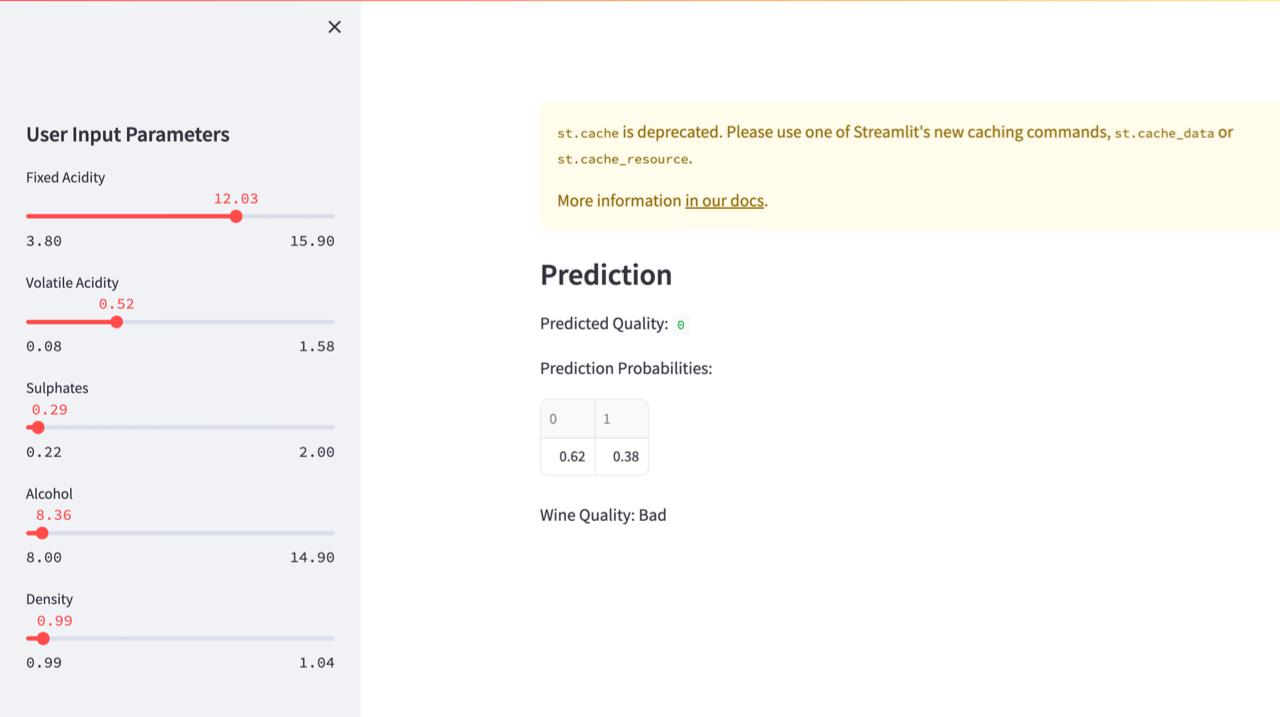
****

Fig 10: Bad wine

5. Limitations

While previous studies have explored machine learning for wine quality prediction, some key areas remain unaddressed:

* Limited comparisons of diverse classification algorithms, often focus on a single method or a narrow range of options.
* Lack of in-depth analysis of model interpretability, hindering practical application and understanding of influential factors.
* Inadequate exploration of the impact of different data preprocessing techniques on model performance.

6. Future enhancements

The following are some possible future scopes:

* Advanced feature scope: By going deeper into the dataset and extracting more pertinent features, we can incorporate new sorts of data sources, such as picture analysis of wine qualities, or integrate sensor data fusion with additional domain-specific information.
* Ensemble Techniques: To improve prediction accuracy and robustness, investigate the integration of several machine learning models using ensemble techniques like stacking, boosting, or bagging. By reducing the shortcomings of individual models, ensemble approaches can enhance performance as a whole.
* Cross-domain Analysis: By combining data from several industries or domains, such as soil composition, climatic data, or grape cultivation techniques, cross-domain analysis can provide a more thorough understanding of the variables affecting wine quality.

8. Conclusion

The following goals are the focus of our analysis:

* Using the wine quality dataset, evaluate each algorithm's predicted accuracy.
* Determine which method works best in this situation by taking into account variables like computational efficiency, interpretability, and accuracy.
* Learn about the connections between different wine qualities and features.
* Since random forest has an accuracy of 82%, model training can be done using it.

# References

|  |  |
| --- | --- |
| [1] | M. &. V. C. Gupta, A study and analysis of machine learning techniques in predicting wine quality, International Journal of Recent Technology and Engineering, 10., 2021. |
| [2] | P. T. P. O. K. P. W. &. K. D. Bhardwaj, A machine learning application in wine quality prediction. Machine Learning With Applications, 2022. |
| [3] | S. A.-A. A. A. H. K. L. L. J. T. &. S. M. Aich, A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques, In 2018 20th International Conference, 2018, February. |
| [4] | K. K. K. G. S. K. M. S. &. K. S. Jain, Machine learning-based predictive modelling for the enhancement of wine quality, 2023. |
| [5] | A. S. Gawale, Wine Quality Prediction using Machine Learning and Hybrid Modeling., Masters thesis, Dublin, National College of Ireland., 2022. |
| [6] | K. G. R. L. R. K. R. A. &. M. S. Sravan, Analyze the Quality of Wine Based on Machine Learning Approach., In International Conference on Data Science and Applications (pp. 351-360). Singapore: Springer Nature S, 2023, July. |
| [7] | P. Shruthi, Wine Quality Prediction Using Data Mining, 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), Bangalore, India, 2019, pp. 23-26, doi: 1109/ICATIEC, 2019. |
| [8] | Z. Tingwei, J. Phys.: Conf. Ser. 1966 012021, 2021. |
| [9] | A. T. a. R. Sehrawat, Wine Quality Detection through Machine Learning Algorithms, Bhubaneswar, India: International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE), 2018. |
| [10] | K. G. R. L. R. K. R. A. &. M. S. Sravan, Analyze the Quality of Wine Based on Machine Learning Approach. In International Conference on Data Science and Applications (pp. 351-360). Singapore: Springer Nature S, (2023, July). . |
| [11] | https://www.analyticsvidhya.com/blog/2018/09/an-end-to-end-guide-to-understand-the-math-behind-xgboost/. |
| [12] | https://www.javatpoint.com/machine-learning-random-forest-algorithm. |
| [13] | https://towardsdatascience.com/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3. |