RED WINE QUALITY PREDICTION USING MACHINE LEARNING

Lakkaraju Datha Sri Laasya[1], Kambham Ritesh[2], Shaik Rehman[3],

Shaik Riyaz[4], S. Anil Kumar[5]

[1,2,3,4] *Student, Department of Computer Science and Engineering, Tirumala Engineering College*
[5] *Professor, Department of Computer Science and Engineering, Tirumala Engineering College*

---***---

Abstract - Wine quality assessment is a critical task in the beverage industry, as it directly impacts consumer satisfaction and market competitiveness. Among various types of wine, red wine stands out for its complexity in flavor, aroma, and texture, making its quality prediction a challenging yet crucial endeavor. This project aims to develop a machine learning model for predicting the quality of red wine based on its physicochemical properties. The dataset utilized for this project consists of various attributes such as acidity, pH, alcohol content, and volatile acidity, among others, collected from red wine samples. These attributes serve as input features for the machine learning model, while the quality of wine, typically rated on a scale, serves as the target variable. The proposed approach involves several stages: data preprocessing, feature selection, model selection, and evaluation. During data preprocessing, techniques like normalization and handling missing values are employed to ensure data quality. Feature selection techniques such as correlation analysis and feature importance are utilized to identify the most relevant attributes for predicting wine quality. Several machine learning algorithms, including but not limited to, linear regression, decision trees, random forests, and support vector machines, are trained and evaluated to determine the most suitable model for the task. Performance metrics such as mean squared error, accuracy, and F1 score are utilized to assess the model's predictive capability.

Key Words: Wine quality, market, aroma, texture, physicochemical, alcohol, acidity.

1.INTRODUCTION

Wine industry shows a recent growth spurt as social drinking is on the rise. The price of wine depends on a rather abstract concept of wine appreciation by wine tasters, opinion among whom may have a high degree of variability. Pricing of wine depends on such a volatile factor to some extent. Another key factor in wine certification and quality assessment is physicochemical tests which are laboratory-based and takes into account factors like acidity, pH level, the presence of sugar and other chemical properties. For the wine market, it would be of interest if human quality of tasting can be related to the chemical properties of wine so that certification and quality assessment and assurance process is more controlled.
ML may generalize the effort or method to formulate the policy. These ML models can be learned by themselves. The model is trained on red wine data. The model can then accurately predict quality by using the necessary elements as inputs. This decreases human effort and resources and improves the company's profitability. Thus, the accuracy can be improved with ML. Our goal is to predict the quality of red wine. Classification is the best choice available to fulfill our needs. We use classification models in this analysis since there are many independent variables used to calculate the dependent(target) variable. For this study, the dataset for quality prediction is used.

1.1 PURPOSE

The purpose of red wine quality prediction using machine learning is multifaceted and serves several important objectives:

Quality Assurance:

One of the primary purposes is to ensure consistency and quality in red wine production. By accurately predicting the quality of red wine based on its physicochemical properties, machine learning models can assist winemakers in identifying potential issues early in the production process and implementing corrective measures to maintain or enhance quality standards.

Optimization of Production Processes:

Machine learning models can provide valuable insights into the relationship between production parameters and wine quality. By analyzing historical data and identifying patterns, these models can help optimize fermentation processes, blending strategies, and aging techniques to produce red wines that meet desired quality specifications.

Cost Reduction and Efficiency Improvement:

Predictive models can help reduce costs associated with quality control measures and minimize wastage by identifying factors that contribute to variations in wine quality. By streamlining production processes and minimizing the need for manual intervention, machine learning can improve operational efficiency and resource utilization in winemaking.

Market Competitiveness:

Accurately predicting red wine quality can enhance the competitiveness of wineries in the market. By consistently producing high-quality wines that meet consumer preferences, wineries can differentiate themselves from competitors and build a strong brand reputation, leading to increased sales and market share.

Consumer Satisfaction:

Ultimately, the purpose of red wine quality prediction is to enhance consumer satisfaction. By producing wines that consistently meet or exceed quality expectations, wineries can foster loyalty among consumers . Machine learning models can also be used to develop personalized recommendations based on individual preferences, further enhancing the consumer experience.

2. LITERATURE REVIEW

- **"Predicting the Quality of Red Wine Using Machine Learning Techniques" by Gabriela Avram, Vlad Boteanu, and Ştefan Daniel Dumitrescu (2019)**

This study explores the application of various machine learning algorithms such as support vector Machines for predicting red wine quality based on its physicochemical attributes. The authors compare the performance of these algorithms and provide insights into the most effective approaches for wine quality prediction.

- **"Wine Quality Prediction Using Machine Learning Techniques" by Ali Ben Salem, Khouloud Boukadi, and Fethi M. Hamdi (2020)**

This paper investigates the use of machine learning algorithms including k-nearest neighbors (KNN), decision trees, and artificial neural networks (ANN) for predicting the quality of red wine. The authors analyze the impact of feature selection methods and hyperparameter tuning on the predictive performance of these models.

- **"Comparison of Machine Learning Techniques for Prediction of Red Wine Quality" by D.G. Bonfim, F.B. Pizzol, and D.M. Silva (2019)**

This study compares the performance of different machine learning algorithms such as linear regression, decision trees, and ensemble methods like gradient boosting and random forests for predicting red wine quality. The authors evaluate the models using metrics such as accuracy, precision, and recall, providing insights into the strengths and weaknesses of each approach.

- **"Quality Prediction of Red Wine by Machine Learning Methods" by Jingying Wang, Xueying Wang, and Pengfei Gao (2018)**

This research investigates the application of machine learning techniques including SVM, decision trees, and KNN for predicting the quality of red wine. The authors explore the impact of feature scaling and parameter optimization on model performance and propose a comprehensive evaluation framework for comparing different algorithms.

- **"Red Wine Quality Prediction Using Machine Learning Algorithms" by Mahima Chawla, Rishabh Kaushik, and S. Sai Satyanarayana Reddy (2021)**

This paper presents a comparative analysis of machine learning algorithms such as logistic regression, decision trees, and gradient boosting for predicting red wine quality. The authors discuss the importance of feature engineering and model evaluation techniques in improving .

- **"Wine Quality Prediction Using Machine Learning Techniques" by Ali Ben Salem, Khouloud Boukadi, and Fethi M. Hamdi (2020)**

This paper investigates the use of machine learning algorithms including k-nearest neighbors (KNN), decision trees, and artificial neural networks (ANN) for predicting the quality of red wine. The authors analyze the impact of feature selection methods and hyperparameter tuning on the predictive performance of these models.

3. METHODILOGY

3.1 EXISTING SYSTEM

Nowadays people try to lead a luxurious life. They tend to use the things either for show off or for their daily basis. These days the consumption of red wine is very common to all. So it became important to analyze the quality of red wine before its consumption to preserve human health. Here some regression algorithms are used like logistic regression, linear regression. However, there are several approaches and models developed by researchers and data scientists that have been proposed and evaluated in academic literature and sometimes in industry applications. These systems typically involve preprocessing data, selecting relevant features, training machine learning models, and evaluating their performance.

Disadvantages

- Doesn't generate accurate and efficient results.
- Computation time is very high.
- Lacking of accuracy may result in lack of efficient further prediction.

3.2 PROPOSED SYSTEM

The proposed system aims to predict the quality of red wine using machine learning algorithms, including Support Vector Machine (SVM), Naive Bayes, and Random Forest. The system follows a systematic approach involving data preprocessing, model training, and evaluation to develop accurate and reliable prediction models.

Data Collection and Preprocessing: Collect a dataset containing physicochemical attributes (e.g., acidity levels, pH, alcohol content) and quality ratings of red wine samples. Preprocess the data by handling missing values, scaling numerical features, and encoding categorical variables if present. Split the dataset into training and testing sets to train and evaluate the machine learning models.

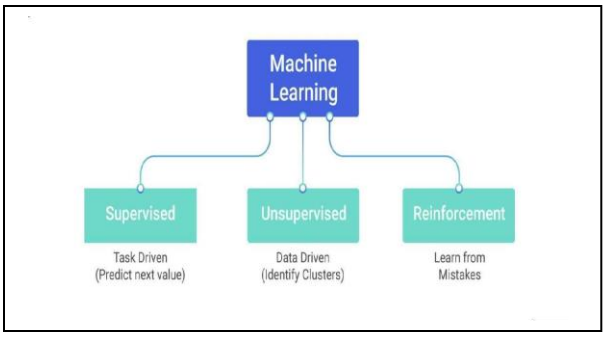
Feature Selection: Perform feature selection to identify the most relevant attributes for predicting red wine quality. Utilize techniques such as correlation analysis, feature importance ranking, or domain knowledge to select informative features.

Model Training: Train three machine learning algorithms: Support Vector Machine (SVM), Naive Bayes, and Random Forest, using the training dataset.

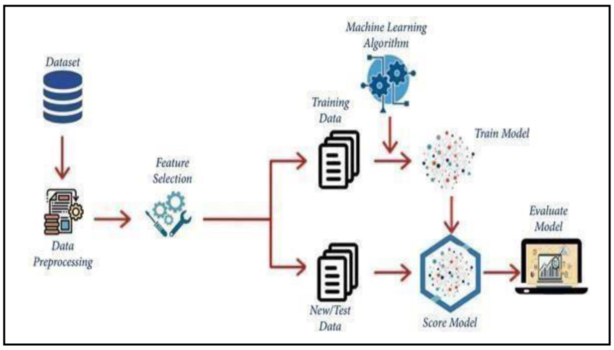
Model Evaluation: Evaluate the performance of each model using appropriate evaluation metrics such as accuracy, precision, recall, F1 score, and ROC curve. Conduct cross-validation to assess the models' generalization ability and mitigate overfitting. Compare the performance of SVM, Naive Bayes, and Random Forest algorithms to determine the most effective approach for red wine quality prediction.

Model Selection and Deployment: Select the best-performing model based on evaluation metrics and deploy it for practical use. Develop a user-friendly interface where users can input the physicochemical attributes of red wine samples, and the deployed model predicts the quality rating. Ensure scalability, reliability, and efficiency of the deployed model to handle real-time prediction requests.

Some Machine Learning Methods

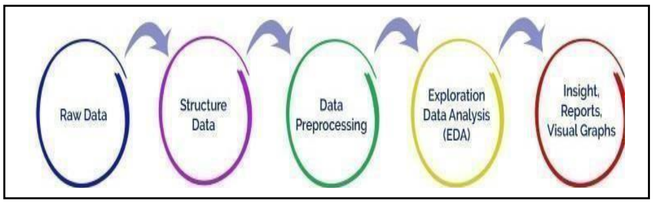


4.SYSTEM DESIGN



DATA PREPROCESSING

Before feeding data to an algorithm, we have to apply transformations to our data which is referred as pre-processing. By performing pre-processing, the raw data which is not feasible for analysis is converted into clean data. In-order to achieve better results using a model in Machine Learning, data format has to be in a proper manner. The data should be in a particular format for different algorithms. For example, if we consider Random Forest algorithm it does not support null values. So that those null values have to be managed using raw data.



Need of Data Preprocessing

For achieving better results from the applied model in Machine Learning projects the format of the data has to be in a proper manner. Some specified Machine Learning model needs information in a specified format. For example, Random Forest algorithm does not support null values, therefore to execute random forest algorithm null values have to be managed from the original raw data set. Another aspect is that data set should be formatted in such a way that more than one Machine Learning and Deep Learning algorithms are executed in one data set, and best out of them is chosen.

5. IMPLEMENTATION

Python is a popular programming language. It was created in 1991 by Guido van Rossum.

It is used for:

1. web development (server-side)
2. software development
3. mathematics

The most recent major version of Python is Python 3. However, Python 2, although not being updated with anything other than security updates, is still quite popular.

It is possible to write Python in an Integrated Development Environment, such as Thonny, PyCharm, NetBeans or Eclipse, Anaconda which are particularly useful when managing larger collections of Python files.

Python was designed for its readability. Python uses new lines to complete a command, as opposed to other programming languages which often use semicolons or parentheses.

Python relies on indentation, using whitespace, to define scope; such as the scope of loops, functions and classes. Other programming languages often use curly-brackets for this purpose. But in the modern days, it is become very much easy and efficient compared to the olden days by various python libraries, frameworks, and modules. Today, Python is one of the most popular programming languages for this task and it has replaced many languages in the industry, one of the reasons is its vast collection of libraries. Python libraries that used in Machine Learning are:

1 . Numpy
2 . Scipy
3. Scikit-learn
4. Pandas
5. Matplotlib

NumPy is a very popular python library for large multi-dimensional array and matrix processing, with the help of a large collection of high-level mathematical functions. It is very useful for fundamental scientific computations in Machine Learning. It is particularly useful for linear algebra, Fourier transform, and random number capabilities. High-end libraries like TensorFlow uses NumPy internally for manipulation of Tensors.

6. TESTING

TESTING DATA

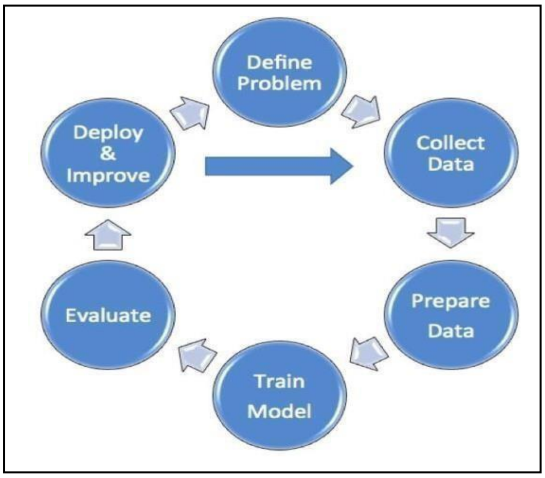


Fig: Prediction model for red wine quality

Testing of data is done based on training model which is classified using supervised learning algorithm. Evaluation of the total responses for every question and determine the polarity of feedback received in context of the given data.

TESTING OBJECTIVE

Software testing is a process used to help identify the correctness, completeness and quality of developed computer software. Software testing is the process used to measure the quality of developed software. Testing is the process of executing a program with the intent of finding errors. Software testing is often referred to as verification & validation.

TYPES OF TESTING

1.White Box Testing
2.Black Box Testing
3.Grey Box Testing

WHITE BOX TESTING

White box testing as the name suggests gives the internal view of the software. This type of testing is also known as structural testing or glass box testing as well, as the interest lies in what lies inside the box.

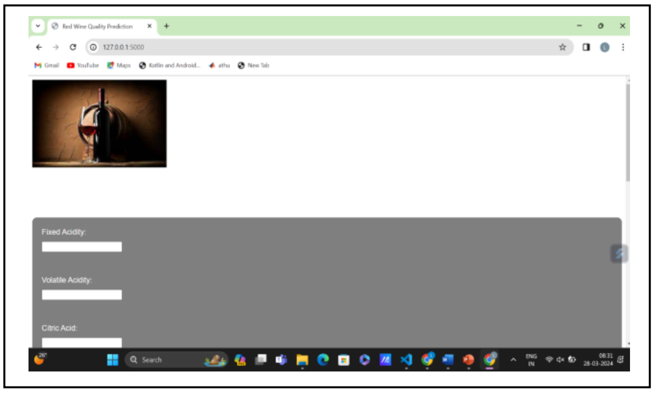
BLACK BOX TESTING

It is also called as behavioral testing. It focuses on the functional requirements of the software. Testing either functional or nonfunctional without reference to the internal structure of the component or system iscalled black box testing

GREY BOX TESTING

Grey Box testing is a software testing method to test the software application with partial knowledge of the internal working structure. It is a combination of black box and white box testing because it involves access to internal coding to design test cases as white box testing and testing practices are done at functionality level as black box testing.

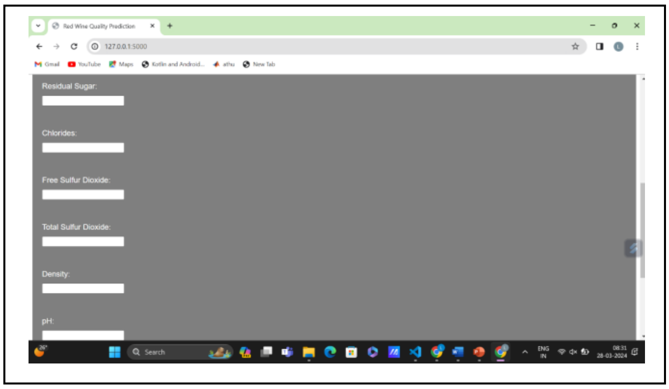
7. RESULTS



Fixed Acidity:

Volatile Acidity:

Citric Acid:



Residual Sugar:

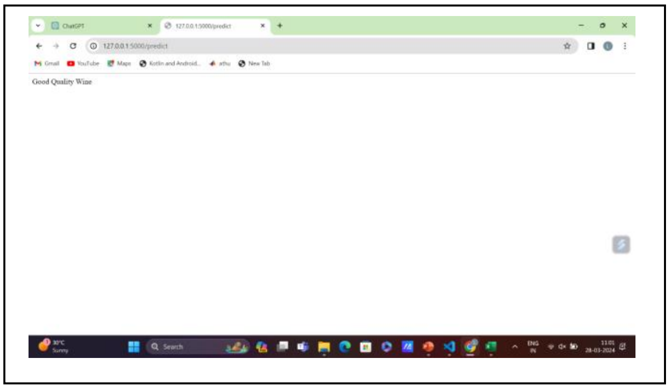
Chlorides:

Free Sulfur Dioxide:

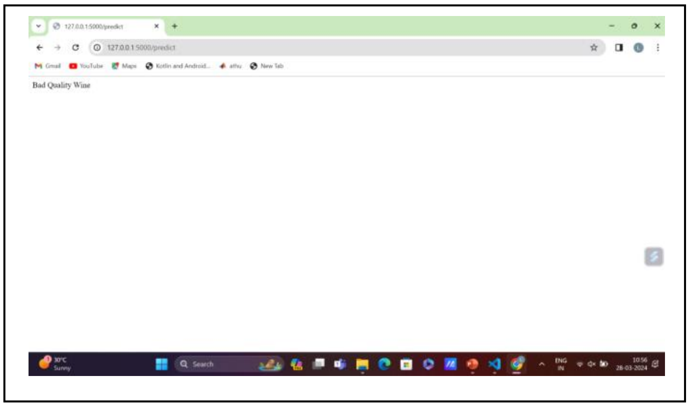
Total Sulfur Dioxide:

Density:

pH:



Good Quality Wine



Bad Quality Wine

8. CONCLUSION

The goal of using machine learning methods was the study's goal to predict whether a red wine will be 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 red wine.

9. 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. S. Ebeler, “Linking Flavour Chemistry to Sensory Analysis of Wine,” in Flavor Chemistry, Thirty Years of Progress, Kluwer Academic Publishers, 1999, pp. 409-422.

3. Asuncion, and D. Newman (2007), UCI Machine Learning Repository, University of California, Irvine, [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

4. 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.

5. N. H. Beltran, M. A. Duarte- MErmoud, V. A. S. Vicencio, S. A. Salah, and M. A. Bustos, “Chilean wine classification using volatile organic compounds data obtained with a fast GC analyzer,” Instrum. Measurement, IEEE Trans., 57: 2421-2436, 2008.

6. S. Shanmuganathan, P. Sallis, and A. Narayanan, “Data mining techniques for modelling seasonal climate effects on grapevine yield and wine quality,” IEEE International Conference on Computational Intelligence Communication Systems and Networks, pp. 82-89, July 2010.

7. B. Chen, C. Rhodes, A. Crawford, and L. Hambuchen, “Wineinformatics: applying data mining on wine sensory reviews processed by the computational wine wheel,” IEEE International Conference on Data Mining Workshop, pp. 142-149, Dec. 2014.

8. K. Agrawal and H. Mohan, "Cardiotocography Analysis for Fetal State Classification Using Machine Learning Algorithms," 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, Tamil Nadu, India, 2019, pp. 1-6.

9. K. Agrawal and H. Mohan, "Text Analysis: Techniques, Applications and Challenges," presented in 2019 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, Tamil Nadu, India, 2019.

10. J. Han, M. Kamber, and J. Pei, “Classification: Advanced Methods,” in Data Mining Concepts and Techniques, 3rd ed., Waltham, MA, USA: Morgan Kaufmann, 2012, pp. 393-443