Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.3, May : 2024

WINE QUALITY PREDICTION USING MACHINE LEARNING

G. Saranya Asst.Professor Computer Science and Engineering Narasaraopeta Engineering College Narasaraopet, Andhra Pradesh dasarisaranya4@gmail.com
Siddu Chennamsetty Student Computer Science and Engineering Narasaraopeta Engineering College Narasaraopet, Andhra Pradesh siddualiea@gmail.com
Pavan Kumar Rachupalli Student Computer Science and Engineering Narasaraopeta Engineering College Narasaraopet, Andhra Pradesh dasarisaranya4@gmail.com

ABSTRACT
Wine is a popular drink across the globe and the gender. Older the wine, better is the taste but, expensive. The wine quality is measured based on the important parameters, such as free Sulphur dioxide, Volatile acidity, Citric Acid and Residual sugar. The traditional way of wine quality assessment was time consuming. This project gives an automatic prediction of Wine quality, as good or bad, using machine learning approaches which are Neural Networks, Logistic Regression and Support Vector Machine are implemented on datasets of Portuguese “Vinho Verde” Wine. The results are compared with standard values. The work is useful in Wine industry for quality testing and assurance for customers.

Keywords—
Random Forest Classifier, Machine Learning, Classification, Feature Engineering, Data Analysis**.**

I. INTRODUCTION
One of the most consumed beverages worldwide is wine. Wine quality is determined by the ingredients that go into its production. Its quality depends on several stages, usually the older the better in taste, wine. Customers and manufacturers alike depend on wine quality in the current small market to spur production growth.
All wines can be divided into five basic categories: dessert wine, red wine, rose wine, sparkling wine, and white wine. In order to make red wine, black grapes are utilized. Red wine often ranges in color from light to dark. Wine made from green and black grapes is known as white wine. Figure displays the appearance of the Red and White Wines.
Many approaches must be used from the beginning to improve the quality of the wine if it is of poor quality. Each person has an opinion about the quality of the wine, mostly based on taste. It's noteworthy to note that wine quality is categorized based on individual preferences. Thanks to advancements in a variety of technology, wine producers may now rely on a multitude of methods for ensuring wine quality.
The producers will then have a better understanding of wine quality.These days, all manufacturers use various techniques to maximize output and build a well-organized process throughout the whole process. With time, these techniques become more and more elegant, yet they also become more challenging.
Over the past few years, India's wine industry has expanded. Several of India's leading states, including Maharashtra, Karnataka, Andhra Pradesh, and Himachal Pradesh, have made significant contributions to the development of wine manufacturing and wine-making processes.
Red wine consumption Controlled wine consumption may improve digestive, mental, and cardiac health among other health issues. Because it contains composites that have lipid-improving, anti-inflammatory, and antioxidant qualities.
Chocolate and cosmetics such as cleansers and perfumes can be made using wine.. Stable acidity, total sulfur dioxide, volatile acidity, residual sugar, chlorides, citrus acid free sulphur dioxide, solidity, PH, sulphates, alcohol, etc. are the factors that are most important in determining the quality of wine .

II. LITERATURE SURVEY

1. "Predicting Wine Quality Using Machine Learning Techniques" by Sachin Pawar and P. A. Tijare: This paper explores the application of machine learning techniques such as decision trees, random forests, and support vector machines for predicting wine quality based on physicochemical properties. The study compares the performance of these algorithms and discusses their effectiveness in wine quality prediction.
2. "Wine Quality Prediction Using Artificial Neural Networks" by Chuan Sun et al.: This research investigates the use of artificial neural networks (ANNs) to predict wine quality. The study explores different ANN architectures and training algorithms to identify the most suitable model for wine quality prediction. It also discusses the impact of input features on prediction accuracy.
3. "Comparison of Machine Learning Techniques for Wine Quality Classification" by Luís Torgo et al.: This paper presents a comparative study of various machine learning techniques for wine quality classification. The study evaluates the performance of algorithms such as k-nearest neighbors, decision trees, and ensemble methods on different wine datasets. It discusses the strengths and weaknesses of each approach and provides insights into their suitability for wine quality prediction tasks.
4. "Predicting Wine Preferences: A Comparative Study of Classification Models" by Paulo Cortez et al.: This research investigates the use of classification models to predict wine preferences based on sensory attributes. The study compares the performance of models such as decision trees, neural networks, and logistic regression in predicting wine quality ratings. It discusses the influence of feature selection and data preprocessing techniques on prediction accuracy.
5. Predicting Wine Quality and Varietal from Physicochemical Properties Using Machine Learning Techniques" by Utku Kose et al.: This study explores the application of machine learning techniques for predicting wine quality and varietal based on physicochemical properties. The research evaluates the performance of algorithms such as support vector machines, random forests, and gradient boosting machines on a large wine dataset. It discusses the importance of feature engineering and model selection in improving prediction accuracy.
6. "Wine Quality Prediction: A Machine Learning Approach" by Sarah I. Clopp and Joydeep Ghosh: This paper presents a comprehensive study on wine quality prediction using machine learning techniques. The research explores the application of regression and classification models to predict wine quality ratings based on physicochemical properties. It evaluates the performance of different algorithms and discusses the impact of feature selection and data preprocessing on prediction accuracy.
7. "Machine Learning Techniques for Wine Quality Classification: A Comparative Study" by Marcos E. Nascimento et al.: This study conducts a comparative analysis of machine learning techniques for wine quality classification. The research evaluates the performance of algorithms such as decision trees, support vector machines, and artificial neural networks on multiple wine datasets. It investigates the effect of hyperparameter tuning and ensemble methods on model performance.
8. "Predicting Wine Quality Using Hybrid Machine Learning Models" by Ahmad Taher Azar et al.: This paper proposes the use of hybrid machine learning models for wine quality prediction. The research combines different algorithms such as genetic algorithms, fuzzy logic, and artificial neural networks to enhance prediction accuracy. It explores the synergy between these techniques and evaluates their performance on a wine quality dataset.
9. "Wine Quality Prediction Using Deep Learning Models" by Jie Zhang et al.: This research investigates the application of deep learning models for wine quality prediction. The study explores architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to extract features from wine sensory data. It discusses the advantages and limitations of deep learning approaches compared to traditional machine learning techniques.
10. "Ensemble Learning for Wine Quality Prediction" by Emad Abuelrub and Majdi Mafarja: This paper explores the use of ensemble learning techniques for wine quality prediction. The research combines multiple base learners such as decision trees, support vector machines, and neural networks

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.3, May : 2024

to create a robust predictive model. It investigates the impact of ensemble methods such as bagging, boosting, and stacking on prediction performance.

III. METHODOLOGY

A. DATA ANALYSIS & PREPROCESSING:

Feature Exploration : Understanding the qualities found in the wine data is the process of feature exploration. These characteristics may include things like residual sugar, acidity levels, alcohol concentration, and fixed acidity.

Data cleaning: Inconsistencies or missing values may be present in the data. It may be essential to use methods like imputation or to remove rows that have missing entries.

Feature Scaling: Various scales may be used to quantify distinct features. By ensuring that all features are scaled similarly, scaling keeps features with higher values from taking center stage in the model.

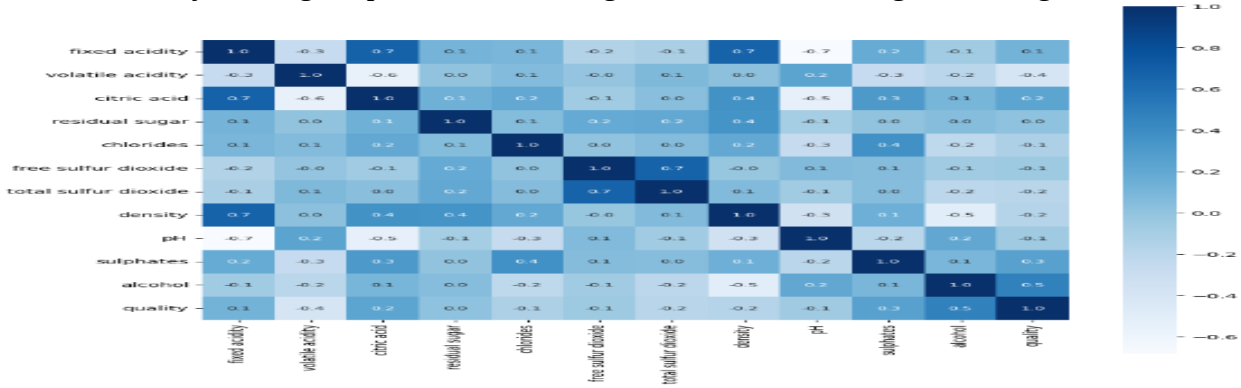


Fig 1: Confusion Matrix

The rows represent the actual labels (predicted quality), and the columns represent the model's predictions. Here's a breakdown of the information in the confusion matrix:

True Positive (TP): These are the data points where the model correctly predicted high-quality wine (quality = good).

False Positive (FP): These are the data points where the model incorrectly predicted high-quality wine (predicted good, but actual quality = bad).

True Negative(TN): These are the data points where the model correctly predicted low-quality wine (quality = bad).

False Negative(FN): These are the data points where the model incorrectly predicted low-quality wine (predicted bad, but actual quality = good).

B. RANDOM FOREST MODEL:

Ensemble Learning: Random Forest is an ensemble learning technique that aggregates predictions from several models (decision trees) to provide a final prediction that is more reliable and accurate.

Building Decision Trees: A random subset of characteristics and data points are used to build each tree in the forest using a process known as bootstrapping. By doing this, overfitting to the training set is avoided.

Making Predictions: Each choice tree in the forest is traversed by a fresh wine sample as it is delivered. Based on the rules it has learnt, each tree predicts the quality of the wine. The final prediction for the new sample is determined by taking the majority vote, which is the most frequently made forecast among all the trees.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.3, May : 2024

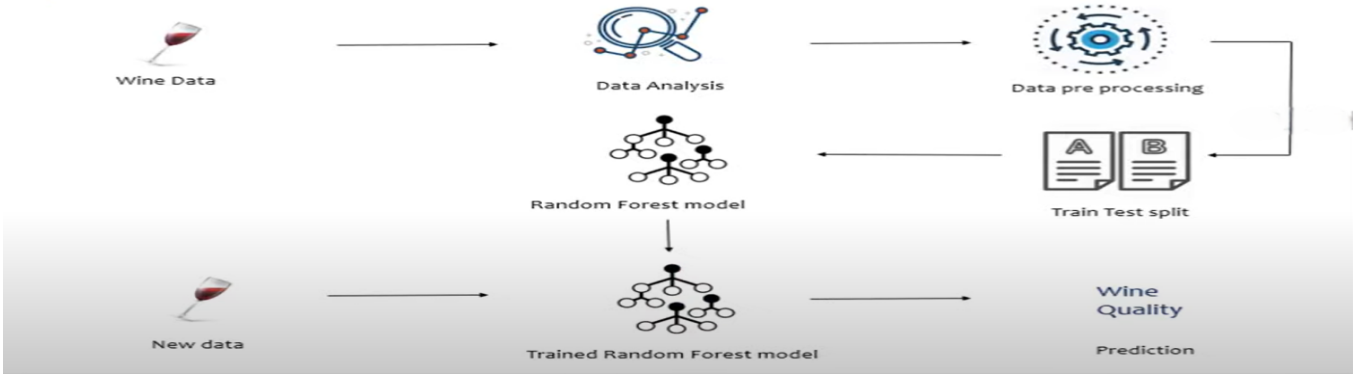


Fig 2: Proposed methodology

C. Train Test Split:

Training Set: The Random Forest model is trained using this subset of the data. Drawing on the features present in the training data, the model derives patterns that allow it to distinguish between wines of varying quality.

Testing Set: After training, the model's performance is assessed using this set of unobserved data. The testing set predictions of the model show that it is generalizable to new data.

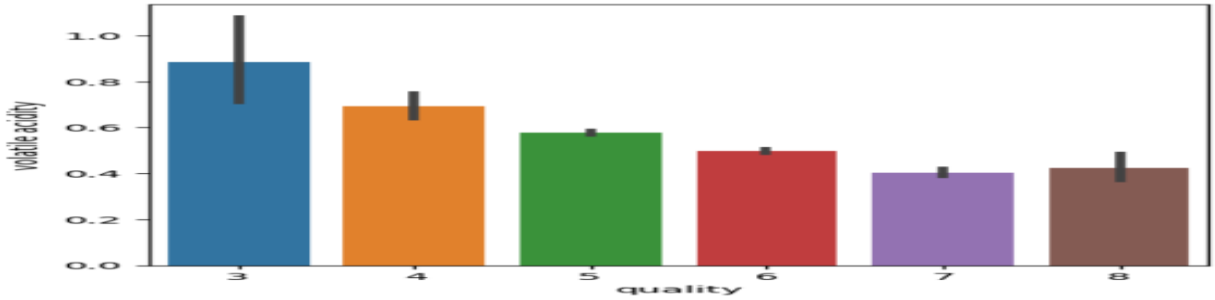


Fig 3: Volatile acidity vs Quality

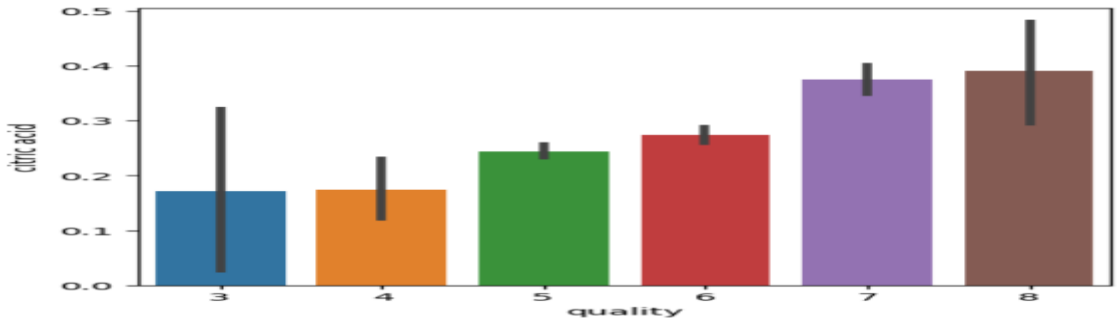


Fig 4: Citric acid vs Quality

D. Evaluation Metrics:

Accuracy: Out of all occurrences classified, the accuracy measure shows the number of cases that are accurately classified. Accuracy is the degree to which a calculated value resembles a real or standard value.

Precision: The percentage of all true positive predictions (TP) that are accurate compared to all false positive predictions and accurate positive predictions yields the precision. Thus, it follows that the precision guarantees that the objects are labelled as such when a model expects a good outcome.

Recall: Recall determines the proportion of true positives to the total of false negatives and true positives. When the expense of a false negative is substantial, this will be helpful. Recall, sometimes

referred to as sensitivity, quantifies the percentage of real spam emails that the models accurately identify.

F1 Measure: The algorithm's total accuracy is determined by combining recall and precision to create the F1 score. Reliability of the model is demonstrated by low false positive and false negative values.

IV. RESULTS AND DISCUSSION

I. **Feature Engineering:** Explore the dataset further to identify potential new features that could improve the model's performance.

II. **Model Evaluation Metrics**: Besides accuracy, consider using other evaluation metrics such as precision, recall, and F1-score to assess the model's performance, especially given the potential class imbalance in the dataset.

III. **Hyperparameter Tuning:** Experiment with different hyperparameters of the RandomForestClassifier and use techniques like GridSearchCV or RandomizedSearchCV to find the optimal combination of hyperparameters for improved performance.

IV. **Ensemble Methods:** Explore ensemble methods such as Bagging, Boosting, or Stacking to further enhance the predictive power of your model.

V. **Cross-Validation:** Implement cross-validation techniques to get a better estimate of the model's performance and ensure its generalizability.

VI. **Feature Importance:** Analyze the feature importance scores provided by the RandomForestClassifier to gain insights into which features are most influential in predicting wine quality.

VII. **Visualizations:** Create visualizations to better understand the relationships between features and the target variable, as well as to interpret the model's predictions and performance.

A. Data Imbalancing Techniques

SMOTE (Over Sampling Technique): By creating synthetic samples for the minority class, SMOTE (Synthetic Minority Over-sampling Technique) is a potent technique for resolving class imbalance in classification tasks.

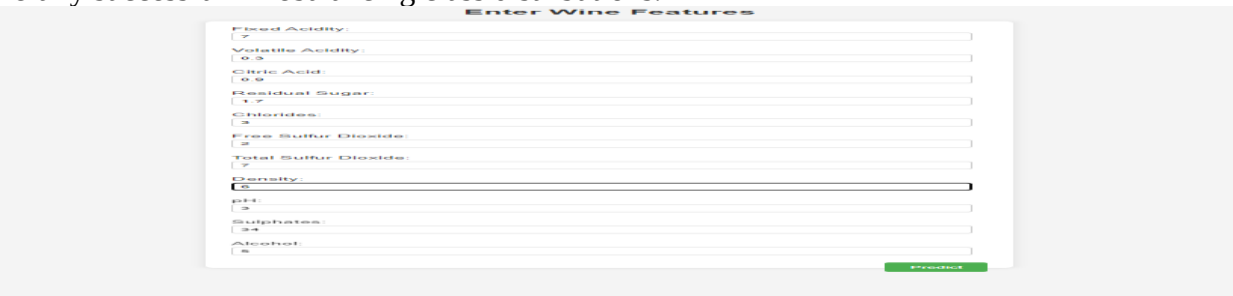
SMOTE effectively boosts the minority class representation in the dataset by interpolating between existing minority class samples. This method finds the k-nearest neighbours of a sample chosen from the minority class in the feature space.

Then, in order to produce a more evenly distributed class distribution, synthetic samples are made along the line segments that link the sample and its neighbors.

B. NearMiss (Under sampling Technique):

Conversely, NearMiss is an under-sampling method intended to decrease the amount of majority class samples while maintaining the crucial data required for classification. In order to guarantee that the samples that are kept are the most similar to the minority class, NearMiss selects samples from the majority class that are closest to samples from the minority class.

Several variants of NearMiss are available, including NearMiss-1, NearMiss-2, and NearMiss-3, each with unique standards for choosing samples from the majority class. By undersampling the majority class and retaining the discriminatory information required for precise classification, NearMiss is generally successful in rebalancing class distributions.



Industrial Engineering Journal
ISSN: 0970-2555
Volume : 53, Issue 5, No.3, May : 2024

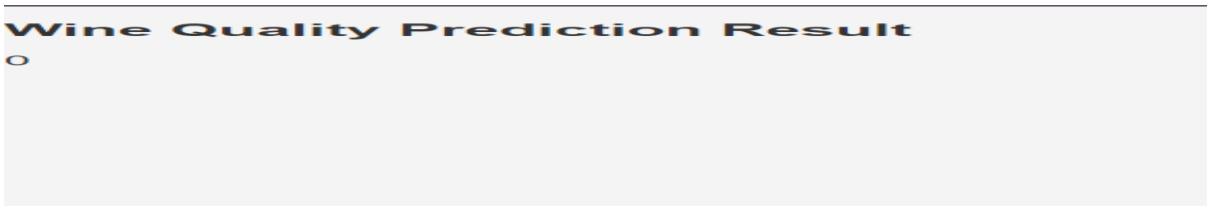


Fig 5: Wine Quality Prediction

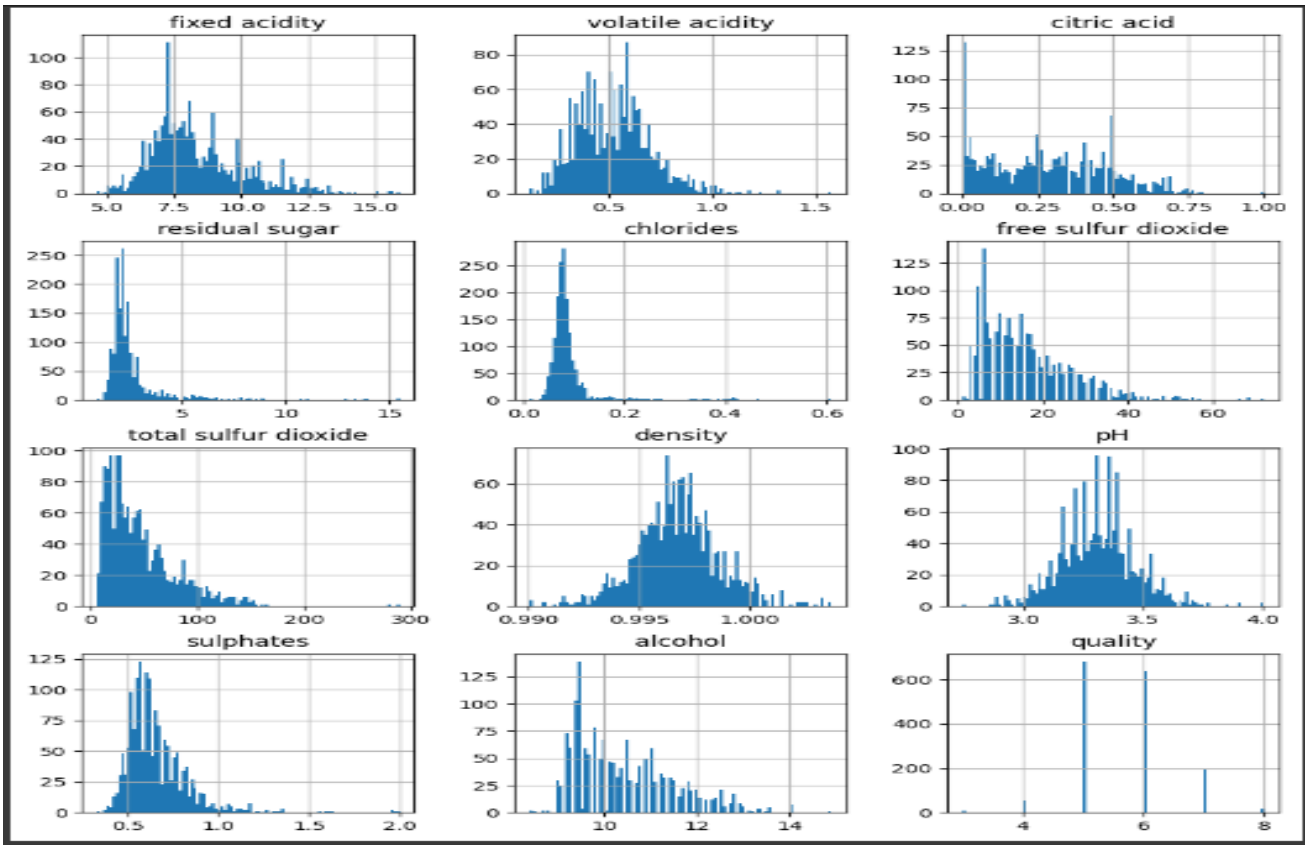


Fig 7: Distribution of various chemical compounds found in wine

V. CONCLUSION

We looked into how well a Random Forest classifier might predict a red wine's quality. We conducted exploratory data analysis using a dataset comprising several physicochemical properties of red wine samples in order to comprehend the properties of the dataset, such as its distribution, correlations, and statistical measures.

Next, we trained and assessed the classifier using machine learning techniques to determine which wines were of higher quality (rated 7 or higher) and which weren't.

We illustrated the efficacy of the Random Forest technique in precisely classifying wine quality through a series of experiments comprising training-test data splits, model fitting, and accuracy assessments.

Additionally, by predicting the quality of fresh wine samples based on their physicochemical characteristics, we demonstrated the usefulness of the trained model in practice.

Our results demonstrate the potential of machine learning algorithms—more especially, Random Forest classifiers—as useful instruments for quality evaluation and decision-making in the viticulture sector.

VI. REFERENCES

[1] Li, H., Zhang Z. and Liu, Z.J. (2017) Application of Artificial Neural Networks for Catalysis: A Review. Catalysts, 7, 306.https://doi.org/10.3390/catal7100306

[2] Shanmuganathan, S. (2016) Artificial Neural Network Modelling: An Introduction. In: Shanmuganathan, S. and Samarasinghe, S. (Eds.), Artificial Neural Network Modelling, Springer, Cham, 1-14.https://doi.org/10.1007/978-3-319-28495-8_1

[3] Jr, R.A., de Sousa, H.C., Malmegrim, R.R., dos Santos Jr., D.S., Carvalho, A.C.P.L.F., Fonseca, F.J., Oliveira Jr., O.N. and Mattoso, L.H.C. (2004) Wine Classification by Taste Sensors Made from Ultra-Thin Films and Using Neural Networks. Sensors and Actuators B: Chemical, 98, 77-82. https://doi.org/10.1016/j.snb.2003.09.025

[4] S. Kumar, K. Agrawal and N. Mandan, "Red Wine Quality Prediction Using Machine Learning Techniques," 2020 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, India, 2020, pp. 1-6,
doi: 10.1109/ICCCI48352.2020.9104095.

[5] P. Shruthi, "Wine Quality Prediction Using Data Mining," 2019 1st International Conference on Advanced Technologies in Intelligent Control, Environment, Computing & Communication Engineering (ICATIECE), Bangalore, India, 2019, pp. 23-26, doi: 10.1109/ICATIECE45860.2019.9063846.

[6] Horowitz, I., & Lockshin, L. (2002). What Price Quality? An Investigation into the Prediction of Wine-quality Ratings. Journal of Wine Research, 13(1), 7–22. https://doi.org/10.1080/0957126022000004020

[7] J. Chen, R. Li and J. Yang, "A Prediction Model for Quality of Red Wine through Explainable Artificial Intelligence," 2022 IEEE International Conference on Sensing, Diagnostics, Prognostics, and Control (SDPC), Chongqing, China, 2022, pp. 268-272, doi: 10.1109/SDPC55702.2022.9916008.

[8] Y. Liu, "Optimization of Gradient Boosting Model for Wine Quality Evaluation," 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), Taiyuan, China, 2021, pp. 128-132, doi: 10.1109/MLBDBI54094.2021.00033.

[9] K. B. Pascua, H. D. Lagura, G. S. Lumacad, A. K. N. Pensona and M. J. I. Jalop, "Combined Synthetic Minority Oversampling Technique and Deep Neural Network for Red Wine Quality Prediction," 2023 International Conference in Advances in Power, Signal, and Information Technology (APSIT), Bhubaneswar, India, 2023, pp. 609-614, doi: 10.1109/APSIT58554.2023.10201733.

[10] M. S. Amzad Basha, K. Desai, S. Christina, M. M. Sucharitha and A. Maheshwari, "Enhancing red wine quality prediction through Machine Learning approaches with Hyperparameters optimization technique," 2023 Second International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT), Trichirappalli, India, 2023, pp. 1-8, doi: 10.1109/ICEEICT56924.2023.10157719.

[11] B. S. Anami, K. Mainalli, S. Kallur and V. Patil, "A Machine Learning Based Approach for Wine Quality Prediction," 2022 2nd Asian Conference on Innovation in Technology (ASIANCON), Ravet, India, 2022, pp. 1-6, doi: 10.1109/ASIANCON55314.2022.9908870.

[12] D. Oreški, I. Pihir and K. Cajzek, "Smart Agriculture and Digital Transformation on Case of Intelligent System for Wine Quality Prediction," 2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO), Opatija, Croatia, 2021, pp. 1370-1375, doi: 10.23919/MIPRO52101.2021.9596979.