Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 13 Issue 08, August-2024

Ensemble–Based Wine Quality Detection using Hybrid Machine Learning Models

An Optimized Approach Combining Random Forest, Logistic Regression, SVM, and GBM for Enhanced Prediction Accuracy

Dodda Abhiram
Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India

Siddharth Mahesh Balijepally
Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India

Yellanki Ekantha Sai Sundar
Computer Science and Engineering
Gokaraju Rangaraju Institute of Engineering and Technology
Hyderabad, India

Abstract— **Quality prediction of wine is one of the most crucial tasks in the wine industry, as it significantly affects its marketing value and consumers. This work provides an ensemble method that combines various machine learning models, including Random Forest, Logistic Regression, Support Vector Machines, and Gradient Boosting Machine, to detect the quality of wines. All these models ensembled via a meta-classifier optimized using Grid Search achieved a high accuracy of 0.884375 on a public wine quality dataset.**

The diversity among these different models is exploited by the ensemble methodology, whereby each model brings a unique perspective to bear on the final prediction. This is facilitated through the effective synthesis of these different perspectives by the meta-classifier implemented as a Neural Network, thereby enhancing the performance of the meta-model. We hereby illustrate that model ensemble technique plays an extremely important role in the attainment of higher accuracy and robustness on more complicated classification tasks, such as wine quality prediction.

It contributes to the domain by proposing a scalable and efficient ensemble learning framework that significantly enhances the accuracy of machine learning models. Results stress that ensemble models are able to outperform single models in quality detection applications, leading the way to more reliable and accurate prediction systems in the wine industry.

Keywords — **Ensemble Learning; Random Forest; Logistic Regression; Support Vector Machines; Gradient Boosting Machine; Meta Classifier; Neural Network; Grid Search Optimization**

I. INTRODUCTION (HEADING 1)

Wine quality prediction plays an important role in the wine industry, affecting everything from production to the way marketing is conducted. Accurate prediction models enable producers to check and improve the quality of their wine, ensuring that only the best products reach consumers. Traditionally, assessing wine quality has been reserved for human experts, but recently, with the help of machine learning, it is possible to implement more automated and consistent methods. Because of the potential ability to combine multiple models to improve prediction accuracy, ensemble learning has received widespread attention in various domains. As such, in this study, we apply ensemble learning techniques to derive an improved robust model for wine quality detection.

Although various machine learning models have been conducted for wine quality prediction, high accuracy and generalization performance still pose a challenge for different types of wines. Single models like Random Forest, SVM, and Gradient Boosting Machines are excellent in their respective strengths but may not fully capture complex relationships between chemical properties and wine quality. That indicates the need to employ an ensemble approach effectively, which can allow these models to complement one another and improve prediction performance. This paper presents the challenge of wine quality prediction based on implementing and optimizing an ensemble learning approach.

The general objective of the paper is to design an ensemble learning method for high-accuracy wine quality prediction. We also evaluate the performance of separate machine learning models comprising Random Forest, Logistic Regression, Support Vector Machines, and Gradient Boosting Machines on the wine quality dataset. This will tune the hyperparameters using Grid Search for optimal performance of these models. We then develop an ensemble methodology that effectively pools the individual model strengths for an improvement in accuracy. Such a comparison of the developed ensemble model to individual models and previous studies will enable us to make inferences about its relative effectiveness.

In this paper, we are introducing a new ensemble learning method that combines Random Forest, Logistic Regression, SVM, and GBM models, optimized through Grid Search. It can be shown that the ensemble model could provide a good prediction accuracy of 88.44%, which is competitive and even higher in comparison with the results of the state-of-the-art benchmark models. It provides insight into the strengths and limitations of using ensemble methods for predicting wine quality that could indicate a potentially useful roadmap for future studies in this area. This provides an effective solution that the wine industry can consider for automation of quality assessments to make them more precise.

II. DATASET DESCRIPTION

A. Wine Quality Dataset Overview

Wine Quality dataset used here, is taken from Kaggle, which is usually employed in the field of predicting wine quality based on physicochemical properties. The dataset is quite extensive, holding 1,599 instances related to red wine. Each example is described by 11 input variables: fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. The rating of the wine quality ranges from 0 to 10. For the purpose of this study, we chose red wine dataset.

B. Data Preprocessing

The following are several preprocessing steps that were done before the training of the machine learning models to ensure the data is of quality and consistency. Perhaps the most obvious checks are on missing values. Although it was detected that the dataset was complete with no missing entries, it's important to standardize the feature variables using StandardScaler so that the feature variable data has normalized values. This way, each feature contributes proportionally to the model training process. Apart from that, the target variable has changed from ordinal to categorical classes for classification tasks. Classes have been defined as "perfect," "good," "average," "bad," and "inedible" according to the quality score.

C. Feature Engineering

Feature engineering was done to achieve better predictive power by adding more features and transformation of the existing features so that the relationship embedded is reflected appropriately. Further feature importance analysis with Random Forests showed that alcohol content and volatile acidity are among the strongest predictors of wine quality. The data is then split into an 80-20 division of training and test sets, respectively, to ensure the model generalizes well on unseen data.

III. METHODOLOGY

In this paper, we propose an ensemble learning method to improve the accuracy of detection in wine quality. Ensemble learning is a way to combine the predictions of multiple models into one and exploits the advantages of each while providing compensation for their weaknesses. In this paper, the ensemble has been constructed using Random Forest, Logistic Regression, SVM, and GBM as base models. These base models were then combined using a meta-classifier - a Neural Network - optimized by grid search for its hyperparameters.

A. Base Models

Random Forest is one of the robust machine learning algorithms and can handle high-dimensional data. It works by constructing several decision trees on training and provides the mode of the classes as output for classification. We tuned the number of trees, maximum depth, and minimum samples split.

$$y_{RF}=\text{majority vote}(T_1(x), T_2(x), \ldots, T_n(x)) \quad (1)$$

Where $T_i(x)$ is the prediction from ith decision tree
n is the total number of trees in the forest

Logistic Regression was chosen because it is simple and easy to interpret. Though linear in nature, it can be surprisingly good if the feature relationships with the target variable are somewhat linear. It also provided a good baseline for our ensemble. Key hyperparameters such as the regularization strength were tuned.

$$P(y=1|x) = 1/(1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p)}) \quad (2)$$

Where $\beta_0, \beta_1, \ldots, \beta_p$ are the model coefficients
$x_1, x_2, \ldots, x_p$ are the input features.

SVMs perform well in high-dimensional feature spaces and have good performances when the number of dimensions is superior to the number of samples. SVM was included due to its nonlinear decision boundaries capability thanks to kernel functions. Optimized hyperparameters were penalty parameter, kernel type, and kernel coefficient.

$$f(x) = \text{sign}(\textstyle\sum_{i=1}^{N} \alpha_i y_i K(x_i, x)+b) \quad (3)$$

Where α_i are the Lagrange multipliers.
y_i are the class labels.
$K(x_i, x)$ is the kernel function.
b is the bias term

Gradient Boosting Machines are amazingly powerful models that build trees greedily - one after another - and each newest tree in the model attempts to fix mistakes of the ones previously generated. GBM was also included because of its knack for improving predictive accuracy. For this model, hyperparameters tuned were over the number of boosting stages, the learning rate, and a maximum depth.

$$y_{GBM}(x) = \textstyle\sum_{m=1}^{M} \nu h_m(x) \quad (4)$$

Where $h_m(x)$ is the prediction from the mth weak learner
ν is the learning rate.
M is the total number of boosting iterations

B. Grid Search for Hyperparameter Optimization

Optimization of the base models' hyper-parameters was conducted by using grid search. Grid search systematically fits the specified grid of parameters and selects the best combination of parameters according to the results of the cross-validation. The best parameters for each model were therefore selected based on maximized accuracy.

C. Meta – Classifier with Neural Network

In this case, the input was the predictions of the base models, while the meta-classifier was a Neural Network. In this case, the design on the Neural Network was simple: it contained two dense layers followed by a softmax output layer, which could predict the final wine quality class. The meta-classifier was trained on the output of base models using a designated separate validation set-even the hyperparameters were fine-tuned by grid searching.

$$Y_{ensemble} = f_{meta}(y_{RF},y_{SVM},y_{LR},y_{GBM}) \quad (5)$$

Where fmeta is the neural network meta classifier
yRF, ySVM, yLR, yGBM

D. Model Ensembling Strategy

It combined the strengths of individual base models using the meta-classifier, based on which the ensemble could capture complex patterns in the data that were missed by an individual model. The performance evaluation of this ensemble on a held-out test set is done with an accuracy of 0.884375, showing the effectiveness of the proposed approach.

IV. EXPERIMENTAL SETUP

A. Training and Testing Split

In this regard, the Wine Quality dataset has been split into training and test sets to assess the performance of the proposed ensemble learning model. In that respect, an 80-20 ratio has been considered-a splitting ratio whereby 80 percent of the data will be used in training the models while the remaining 20 percent will be reserved for testing. Moving further, this training data was used for hyperparameter tuning in the grid search along with the 5-fold cross-validation strategy. The random state was set to 5 for result reproduction.

B. Evaluation Metrics

Since we are using 4 base models of Random Forest, Logistic Regression, Support Vector Machines, and Gradient Boosting Machines, we evaluated these models by their score. We then optimized these models by Grid Search Hyperparameter Optimization and used a Meta – Classifier with Neural Network as an ensemble learning strategy. The ensemble model was trained for 50 epochs and was evaluated by the accuracy score.

V. RESULTS AND DISCUSSIONS

A. Model Performance Comparison

We detail below the performance of the individual base models and the final ensemble model. In this research, we used Random Forest, Logistic Regression, Support Vector Machines (SVM), and Gradient Boosting Machine (GBM) as the different models. Each of them was fine-tuned with grid search for hyperparameter optimization. We give here the tabulated performance metrics like accuracy, precision, recall, and F1-score of each model, which will clearly illustrate their effectiveness in predicting wine quality.

TABLE 1.

Model	Accuracy
Random Forest	0.8670833333333334
Logistic Regression	0.8365962009803921
Support Vector Machines	0.8451838235294117
Gradient Boosting Machine	0.8717769607843138
Ensemble Model	0.884375

VI. CONCLUSION

This paper applied ensemble learning to the task of predicting wine quality and proposed an ensemble Random Forest, Logistic Regression, Support Vector Machine, and Gradient Boosting Machine GBM-based model that utilized Grid Search for hyperparameter tuning. The proposed ensemble model achieved higher accuracy at 0.884375 compared to when using the base models alone. It thus suggests that the ensembling approach works well in such cases by effectively combining the strengths of the base models to improve predictive performance. It shows the potential of ensemble learning in improving the accuracy of classification in wine quality prediction.

The results of this study will go a long way to impact the wine industry. Using an appropriate model in the prediction of wine quality will indeed provide insight for producers in reinforcing quality control analyses, which would enable them to identify the most favorable production methods and make relevant decisions on the pricing of products and the positioning of such products in the market. This ensemble model proposed in the study can be integrated with current quality assessment systems and become a reliable tool in wine quality evaluation before it reaches consumers.

Limitations of the present study are many. First, the dataset was limited to only specific features and the sample size, although decent, was not very large. More general datasets on wine quality may result in different performances. Besides, grid search methods and ensemble methods have some computational complexities that could be a constraint in real-world applications. Another observation is that the study considered a few fixed machine learning models, excluding all the other probably effective algorithm that might have formed part of the committee.

VII. REFERNECES

[1] K. Mittal, K. S. Gill, S. Malhotra and S. Devliyal, "Insights into Red Wine Quality: Utilizing Gradient Boosting for Data Exploration and Prediction Analysis," 2024 International Conference on Innovations and Challenges in Emerging Technologies (ICICET), Nagpur, India, 2024, pp. 1-5, doi: 10.1109/ICICET59348.2024.10616351

[2] M. Nandan, H. Raj Gupta and M. Mondal, "Building a Classification Model based on Feature Engineering for the Prediction of Wine Quality by Employing Supervised Machine Learning and Ensemble Learning Techniques," 2023 International Conference on Computer, Electrical & Communication Engineering (ICCECE), Kolkata, India, 2023, pp. 1-7, doi: 10.1109/ICCECE51049.2023.10085272.

[3] 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.

[4] K. Mittal, K. S. Gill, M. Kumar and R. Rawat, "Predicting the Red Wine Quality using Exploratory Data Analysis and VGG19 Convolutional Neural Network in Deep Learning," 2024 5th International Conference for Emerging Technology (INCET), Belgaum, India, 2024, pp. 1-5, doi: 10.1109/INCET61516.2024.10593311.

[5] S. Kumari, A. Misra, A. Wahi and P. S. Rathore, "Quality of Red Wine: Analysis and Comparative Study of Machine Learning Models," 2023 5th International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2023, pp. 769-772, doi: 10.1109/ICIRCA57980.2023.10220857.

[6] D. J. Garodia, A. Gadad and A. Vikram, "A Pugnacious Comparative Study of Data Analysis Techniques for Wine Quality Prediction," 2024 IEEE 9th International Conference for Convergence in Technology (I2CT), Pune, India, 2024, pp. 1-6, doi: 10.1109/I2CT61223.2024.10543546.

[7] S. Aich, A. A. Al-Absi, K. L. Hui, J. T. Lee and M. Sain, "A classification approach with different feature sets to predict the quality of different types of wine using machine learning techniques," 2018 20th International Conference on Advanced Communication Technology (ICACT), Chuncheon, Korea (South), 2018, pp. 139-143, doi: 10.23919/ICACT.2018.8323674.

[8] D. Yuan, J. Huang, X. Yang and J. Cui, "Improved random forest classification approach based on hybrid clustering selection," 2020 Chinese Automation Congress (CAC), Shanghai, China, 2020, pp. 1559-1563, doi: 10.1109/CAC51589.2020.9326711.

[9] N. R. Likitha and J. T. Nagalakshmi, "Improving Prediction Accuracy in Drift Detection Using Random Forest in Comparing with Modified Light Gradient Boost Model," 2024 Ninth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2024, pp. 1-4, doi: 10.1109/ICONSTEM60960.2024.10568896.

[10] M. -P. Hosseini, M. R. Nazem-Zadeh, F. Mahmoudi, H. Ying and H. Soltanian-Zadeh, "Support Vector Machine with nonlinear-kernel optimization for lateralization of epileptogenic hippocampus in MR images," 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Chicago, IL, USA, 2014, pp. 1047-1050, doi: 10.1109/EMBC.2014.6943773.

[11] K. Chen, H. Yao and Z. Han, "Arithmetic optimization algorithm to optimize support vector machine for chip defect Identification," 2022 28th International Conference on Mechatronics and Machine Vision in Practice (M2VIP), Nanjing, China, 2022, pp. 1-5, doi: 10.1109/M2VIP55626.2022.10041106.

[12] R. Wen and K. Zhang, "Research on Automated Classification Method of Network Attacking Based on Gradient Boosting Decision Tree," 2022 International Conference on Machine Learning and Knowledge Engineering (MLKE), Guilin, China, 2022, pp. 72-76, doi: 10.1109/MLKE55170.2022.00019

[13] K. G, S. C. B. Jaganathan, S. K, A. Mital and S. Awal, "Harmony Gradient Boosting Random Forest Machine Learning Algorithms for Sentiment Classification," 2022 IEEE 2nd International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC), Gunupur, Odisha, India, 2022, pp. 1-5, doi: 10.1109/iSSSC56467.2022.10051210.

[14] N. Ramteke and P. Maidamwar, "Cardiac Patient Data Classification Using Ensemble Machine Learning Technique," 2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT), Delhi, India, 2023, pp. 1-6, doi: 10.1109/ICCCNT56998.2023.10307702.