iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** VI **Month of publication:** June 2023

DOI: https://doi.org/10.22214/ijraset.2023.54363

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue VI Jun 2023- Available at www.ijraset.com

Regression Modeling Approaches for Red Wine Quality Prediction: Individual and Ensemble

Amrutha K
Department of Mathematics, Amrita School of Physical Sciences, Kochi, Amrita Vishwa Vidyapeetham, India

Abstract: ***This paper aims to compare the performance of several regression models and a combination of regression and ensemble models in predicting the quality of red wine using the wine quality dataset from the UCI Machine Learning Repository. The dataset consists of white and red vinho verde wines from northern Portugal, with 6,497 samples. Before training the models, the dataset undergoes appropriate preprocessing steps to ensure data quality and consistency. Five re-gression algorithms, namely Linear Regression (LR), Random Forest Regressor (RF), Support Vector Regression (SVR), Decision Tree Regressor (DT), and Multi-layer Perceptron Regressor (MLP) are trained and tested on the dataset. Additionally, the predictions of these individual regression models are combined with four ensemble models: XGBRegressor (XGB), AdaBoostRegressor (ABR), BaggingRegressor (BR), and GradientBoostingRegressor (GRB). The results indicate that among the individual models, Random Forest (RF) performs the best, exhibiting the lowest MAE, MSE, and RMSE values and the highest R2 score. This suggests that RF better fits the red wine quality dataset compared to the other regression models. However, the combination of Random For-est with Bagging Regressor (RF and BR) outperforms the individual models, demonstrating lower errors and a relatively higher R2 score.***

Keywords: Red Wine Quality Prediction, Regression Models, Ensemble Mod-els, Evaluation Metrics, Wine Quality Dataset.

I. INTRODUCTION

Red wine is a popular and widely consumed beverage that is highly valued for its diverse flavours, aromas, and overall quality. The traditional way to predict red wine quality includes three parts: sight, smell and taste. All of them need to be certificated by people with years of professional training which already cost many resource, time and money, not to mention the wine quality test only can be accomplished after the whole production process ended. What industrial production need is a technology that can perform quality identification at any time[1].

Machine learning techniques, specifically regression models, have emerged as powerful red wine quality prediction tools. The use of regression models for red wine quality prediction offers several advantages. Firstly, it allows for a quantitative and objective assessment of wine quality based on measurable properties, reducing the subjectivity associated with sensory evaluations. Secondly, these models can capture complex patterns and interactions among the numerous physicochemical variables, providing valuable insights into the key factors influencing red wine quality. Lastly, regression models enable wine pro-ducers to optimize their production processes, make informed decisions about grape selection, and improve the overall quality of their wines.

Although machine learning models are very powerful, a single model always has limitations. Ensemble learning as an algorithm that can fuse multiple mod-els that have performed well in the field of ML provides a way to break through these limitations to get a higher accuracy rate[1]. Ensemble models, which com-bine the predictions of multiple regression models, have shown promising results in improving the accuracy and robustness of predictions.

In this context, this research paper aims to explore and compare various regression models and their combinations with ensemble models for red wine quality prediction. By analyzing a comprehensive dataset of red wine samples, we seek to identify the most accurate and effective models for predicting red wine quality based on physicochemical properties. The findings of this study will contribute to understanding the factors influencing red wine quality and provide practical insights for wine producers, sommeliers, and consumers in their decision-making processes.

II. RELATED WORKS

Wine quality prediction using machine learning (ML) has emerged as a popular and effective approach in the wine industry. ML algorithms can analyze the various chemical and sensory attributes of wines and make predictions about their quality.

Numerous research papers have already been published on the implementation of ML techniques to predict wine quality, highlighting major developments and contributions to the field.

Using the wine quality dataset present in the UCI repository, Qingwen Zeng
[1]employed an ensemble learning strategy to forecast the quality of red wines. To build the stacking model in this paper, he chose the models below: SVM, Random Forest, MLPClassifier, Logistic Regression, and XGBClassifier. Com-bining LogisticRegression as a meta-model and the base models MLPClassifier, XGBClassifier, and RandomForest. Stacking, followed by XGBClassifier by about 1% was found to have the best performance; however, XGBClassifier ap-pears to have an overfitting issue. K. R. Dahal, J. N. Dahal, H. Banjade, and S. Gaire [2] presented one of the most recent research papers on wine quality prediction using ML techniques. They evaluate the efficiency of the Ridge Re-gression (RR), Support Vector Machine (SVM), Gradient Boosting Regressor (GBR), and Artificial Neural Network (ANN) among different ML models. The results of the analysis demonstrated that GBR outperformed all other models, with MSE, R, and MAPE values of 0.3741, 0.6057, and 0.0873, respectively. In this study, Aich, Al-Absi, Hui, Lee, and Sain [3] proposed a novel method for predicting wine quality by considering various feature selection algorithms, including Principal Component Analysis (PCA), Recursive Feature Elimina-tion Approach (RFE), and nonlinear decision tree-based classifiers for analysing performance metrics. Nonlinear classifiers like RPART, C4.5, PART, Bagging CART, Random Forest, and Boosted C5.0 have been used by them. While pre-dicting the quality of red wine using RFE-based feature sets, the Random Forest classifier achieves the highest accuracy of 94.51%; however, when predicting the quality of white wine using RFE-based feature sets, the same classifier achieves the highest accuracy of 97.79%. Trivedi and Sehrawat [4] explored the applica-tion of machine learning algorithms for wine quality detection. To predict the values of the test data, both logistic regression and random forest classifiers are applied individually to the data. Compared to logistic regression (LR), which has an accuracy rate of 76%, the random forest (RF) classifier performs better. Later, a new framework that combined XGBoost, LightGBM, and multifractal detrended cross-correlation analysis (MF-DCCA) was proposed by Chao Ye, Ke Li, and Guo-zhu Jia [5]. They believe the proposed approach represents a de-velopment in the classification of red wine quality based on the results of the correlation importance and classification. The most complex factor affecting the quality of red wine is residual sugar, while volatile acidity and chlorides have weaker cross-correlations. The classification accuracy of LightGBM and XGBoost was higher than that of the other machine-learning algorithms.
To address the issue of unbalanced data, Hu, Xi, Mohammed, and Miao
[6]oversampled the minority class using the Synthetic Minority Over-Sampling Technique (SMOTE). Then, using three different classification techniques—decision tree, adaptive boosting, and random forest—it was suggested a data analysis approach to categorise the white wine dataset into three groups: high, normal, and poor quality. Random forest produced the desired results in terms of er-ror rates and ROC values.
The research paper by Andy Liaw and Matthew Wiener [7] serves as an introduction to the Random Forest algorithm and its implementation in R. It highlights the advantages of using random forests for classification and regression tasks and provides practical guidance on how to use the Random Forest package for data analysis and prediction. An overview of ensemble approaches for regression and their potential to enhance predictive performance across a variety of domains is provided by Jo˜ao Mendes Moreira, Carlos Soares, Al´ıpio M´ario Jorge, and Jorge Freire de Sousa [8]. They look at several ensemble techniques used with regression models, including bagging, boosting, and stacking.
This study compares various techniques based on their performance metrics and discusses the benefits and drawbacks of ensemble ap-proaches. Additionally, it makes use of ensemble regression in several fields, such as engineering, finance, and environmental studies. The key research directions and unresolved issues in ensemble regression are highlighted in the survey's conclusion.

III. METHODOLOGY

A. Dataset Description

Researchers P. Cortez, A. Cerdeira, F. Almeida, T. Matos, and J. Reis developed the wine quality dataset. It was made accessible to the public in 2009 and is housed in the UCI machine learning repository (see [9]). The dataset consists of two files that contain information on different white wine varieties as well as red "Vinho Verde" wine, a particular wine with Portuguese origins. These datasets can be used to perform both regression and classification tasks. There are 1,599 examples of red wines and 4,898 occurrences of white wines combined.
There are far more average wines than exceptional or subpar wines, demonstrating how the classes are ordered but unbalanced. For the regression task, the red wine dataset was chosen for my research, which comprises 11 input variables relating to various chemical attributes of wines.
The input variables, their units, and their descriptions are listed in Figure 1.

VARIABLE	UNITS	DESCRIPTION
Fixed Acidity	g(tartaric acid)/dm³	The concentration of nonvolatile acids in the wine.
Volatile Acidity	g(tartaric acid)/dm³	The concentration of acetic acid in the wine.
Citric Acid	g/dm³	The concentration of citric acid in the wine.
Residual Sugar	g/dm³	The amount of sugar left in the wine after fermentation
Chlorides	g(sodium chloride)/dm³	The amount of salt in the wine.
Free Sulphur Dioxide	mg/dm³	The amount of free sulfur dioxide in the wine.
Total Sulphur Dioxide	mg/dm³	The total amount of sulfur dioxide (free and bound forms) in the wine.
Density	g/dm³	The density of the wine.
pH	No Units	The acidity level of the wine.
Sulphates	g/(potassium sulphate)/dm³	The concentration of sulfates in the wine.
Alcohol	volume %	The alcohol content of the wine.

Figure 1: Input Variables (Name, Unit, Description)

The dataset also includes a target variable that, in addition to the input factors, rates the excellence of the red wine on a scale from 0 (poor) to 10 (ex-cellent). The dataset's goal is to study the relationships between the sensory quality(output variable) of wine and its physical features (input variables). Fu-ture researchers are advised to test feature selection methods on the datasets and watch how they respond to such analysis as the community acquired these data without considering the value of the input features.

B. Data Preprocessing and Transformation

Skewness is a measurement of the distortion of symmetrical distribution or asymmetry in a data set. It is demonstrated on a bell curve when data points are not distributed symmetrically to the left and right sides of the median on a bell curve. If the bell curve is shifted to the left or the right, it is said to be skewed [10].

Here, the skewness is calculated, and filtered based on a thresh-old (in this case, 0.5) and a logarithmic transformation is performed on those variables identified as skewed based on their skewness values. The logarithmic transformation can help mitigate the skewness in the distribution of the fea-tures and make them more suitable for certain statistical analyses or modelling techniques that assume normality.

Outliers are those data points that are significantly different from the rest of the dataset. They are often abnormal observations that skew the data distri-bution and arise due to inconsistent data entry, or erroneous observations[11]. In this study, outliers were present in all 11 input variables, all of them were removed as they can be problematic and thus their removal is necessary. The Isolation Forest algorithm is used to detect outliers in this study, which is a popular method for detecting outliers by constructing decision trees and isolat-ing anomalies in the data and then they are removed to ensure data quality and consistency.

C. Regression Models

1) Linear Regression

Linear Regression is a statistical/machine learning technique that attempts to model the linear relationship between the independent predictor variable x and a dependent quantitative response variable y. It is important that the predictor and response variables be numerical values [12]. A general linear regression model can be represented mathematically as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \quad (1)$$

where:

- y is the target (dependent) variable being predicted.
- $x_1, x_2, \ldots, x_n$ are the input features or independent variables.
- $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients or weights associated with each feature.
- ϵ represents the error term, accounting for the difference between the predicted and actual values.

2) Random Forest Regression

Random forest is one of the most powerful supervised learning algorithms which is capable of performing regression as well as classification tasks. Random forest regression is a supervised learning algorithm and bagging technique that uses an ensemble learning method for regression in machine learning. The trees in random forests run in parallel, meaning there is no interaction between these trees while building the trees[13]. The combined decision trees are called base models, and they can be represented more formally as:

$g(x) = f_0(x) + f_1(x) + f_2(x) + ...$

Random forest uses the Bagging or Bootstrap Aggregation technique of en-semble learning in which aggregated decision tree runs in parallel and do not interact with each other. With the help of Random Forest regression, we can prevent Overfitting in the model by creating random subsets of the dataset.

3) Decision Tree Regression

A decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time, an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches, each representing values for the attribute tested. The leaf node represents a decision on the numerical target. The topmost decision node in a tree corresponds to the best predictor called the root node. Decision trees can handle both categorical and numerical data.[14]

4) Support Vector Regression

Support Vector Regression is a regression algorithm that works for continuous variables. Below are some keywords which are used in Support Vector Regres-sion:

- Kernel: It is a function used to map lower-dimensional data into higher-dimensional data.
- Hyperplane: It is a line that helps to predict the continuous variables and covers most of the data points.
- Boundary line: Boundary lines are the two lines apart from the hyper-plane, which creates a margin for data points.
- Support vectors: Support vectors are the data points that are nearest to the hyperplane and opposite class.
- In SVR, we always try to determine a hyperplane with a maximum margin, so that the maximum number of data points are covered in that margin. The main goal of SVR is to consider the maximum data points within the boundary lines and the hyperplane (best-fit line) must contain a maximum number of data points.[15]

5) Multi-Layer Perceptron Regressor

MLPRegressor is a powerful machine-learning algorithm for regression tasks. It provides a high degree of accuracy and can handle complex, non-linear datasets. MLPRegressor is an artificial neural network model that uses backpropagation to adjust the weights between neurons to improve prediction accuracy. ML-PRegressor implements a Multi-Layer Perceptron (MLP) algorithm for train-ing and testing data sets using backpropagation and stochastic gradient de-scent methods. It includes several parameters that can be used to fine-tune the model's performance including several hidden layers, activation functions, solvers (for optimization), etc. It is an efficient method for solving regression problems as it can learn complex non-linear relationships between input and output variables.[16]

D. Ensemble Models

1) XGB Regressor

XGBoost, which stands for Extreme Gradient Boosting, is a scalable, distributed gradient-boosted decision tree (GBDT) machine learning library. It provides parallel tree boosting and is the leading machine-learning library for regression, classification, and ranking problems[17]. The algorithm works by iteratively adding weak regression models to the ensemble, each one attempting to correct the errors made by the previous models. The training process involves optimiz-ing an objective function that quantifies the difference between the predicted and actual values.

2) AdaBoost Regressor

An AdaBoost regressor is a meta-estimator that begins by fitting a regressor on the original dataset and then fits additional copies of the regressor on the same dataset but where the weights of instances are adjusted according to the error of the current prediction.

As such, subsequent regressors focus more on difficult cases[18]. The algorithm works by iteratively training a sequence of weak regression models on differently weighted versions of the training data. In each iteration, the algorithm assigns higher weights to the samples that were in-correctly predicted by the previous models, allowing subsequent models to focus more on those challenging samples. This adaptive process helps the algorithm progressively improve its performance.

3) Bagging Regressor

Bagging regressors are similar to bagging classifiers. They train each regressor model on a random subset of the original training set and aggregate the pre-dictions. Then, the aggregation averages over the iterations because the target variable is numeric[19]. The algorithm is based on the concept of bootstrap aggregating, or bagging, which involves creating multiple subsets of the train-ing data by random sampling with replacement. Each subset is used to train a separate base regression model, and the predictions from these models are combined to make the final prediction.

4) GradientBoosting Regressor

Gradient Boosting is a powerful boosting algorithm that combines several weak learners into strong learners, in which each new model is trained to minimize the loss function such as mean squared error or cross-entropy of the previous model using gradient descent[20]. The algorithm works by sequentially adding weak regression models, usually decision trees, to the ensemble. Each new model is trained to correct the errors made by the previous models. The training process involves optimizing a loss function, such as mean squared error (MSE), to minimize the difference between the predicted and actual values.

5) Model Evaluation

The evaluation of the models is based on various performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R2 score. These metrics provide insights into the accuracy, precision, and goodness of fit of the models.[21]

a) Mean Absolute Error(MAE) is the mean size of the mistakes in collected predictions. We know that an error is the absolute difference between the actual values and the values that are predicted. The absolute difference means that if the result has a negative sign, it is ignored.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |x_i - y_i| \qquad (2)$$

b) The Mean Squared Error (MSE) is the squared mean of the difference between the actual values and predictable values.

$$MSE = \frac{1}{N}\sum_{i=1}^{N} (x_i - y_i)^2 \qquad (3)$$

c) RMSE is the standard deviation of the errors which occur when a predic-tion is made on a dataset. This is the same as MSE (Mean Squared Error) but the root of the value is considered while determining the accuracy of the model.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (x_i - y_i)^2} \qquad (4)$$

d) The coefficient of determination or R-squared (R2) metric indicates how well a model fits a given dataset. It indicates how close the regression line (i.e. the predicted values plotted) is to the actual data values. The R squared value lies between 0 and 1 where 0 indicates that this model doesn't fit the given data and 1 indicates that the model fits perfectly to the dataset provided.

$$R^2 = 1 - \frac{(x_i - y_i)^2}{(x_i - y)^2} \qquad (5)$$

Here, yi represents the predicted value and y represents the mean value. The lower value of MAE, MSE, and RMSE implies higher accuracy of a regression model. However, a higher value of R square is considered desirable.

IV. RESULT ANALYSIS

A. *Performance Of Individual Regression Models*

After training on multiple regression models such as LR, RF, SVR, DT and MLP, to assess their performance evaluation was conducted in the test data using evaluation metrics.

For the individual models:

1) Random Forest (RF) has the lowest MAE and RMSE, indicating better performance in terms of average absolute error and root mean squared error. It also has the highest R2 score, suggesting a better fit to the data compared to the other individual models.

2) Decision Tree (DT) has the highest MAE, RMSE, and MSE values, and the lowest R2 score, indicating poorer performance compared to the other individual models.

Therefore, among the individual models, Random Forest (RF) performs the best based on the provided evaluation metrics.

MODEL	MAE	RMSE	MSE	R2
LR	0.516078	0.657117	0.431803	0.352314
RF	0.499923	0.646179	0.417547	0.373697
SVR	0.511213	0.660384	0.436106	0.345860
DT	0.555985	0.878750	0.772201	-0.158267
MLP	0.520810	0.665534	0.442935	0.335617

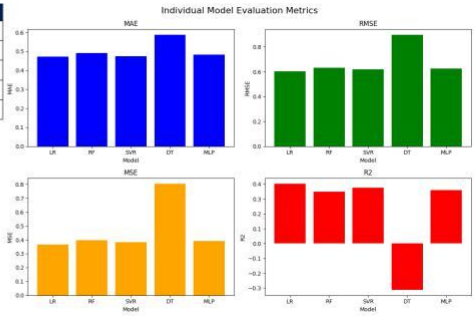


Figure 2: Performance of individual regression model

B. *Performance Of Combination Of Regression and En-semble Models*

Further, the combination of regression models with ensemble models was ex-plored to potentially improve predictive performance. The ensemble models XGB, ABR, BR, and GRB were used to comprehensively evaluate individual models and their ensemble combinations.

For the individual and ensemble models:

1) RF + BR (Random Forest + Bagging Regressor) has the lowest MAE, RMSE, and MSE values among all the combinations. It also has a rela-tively high R2 score compared to the other combinations.

2) SVR + GRB (Support Vector Regression + Gradient Boosting Regressor) and LR + BR (Linear Regression + Bagging Regressor) have nearly the same evaluation metrics as RF + BR, suggesting similar performance.

Hence, the results indicate that the RF + BR combination achieved better prediction accuracy and a stronger fit to the data compared to the other com-binations.

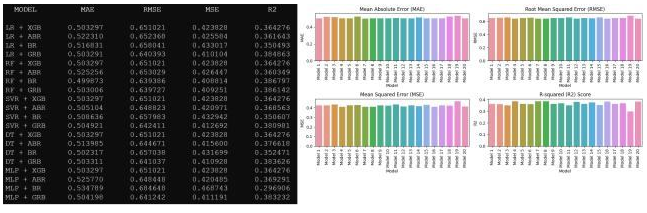


Figure 3: Performance of combination of regression and ensemble models

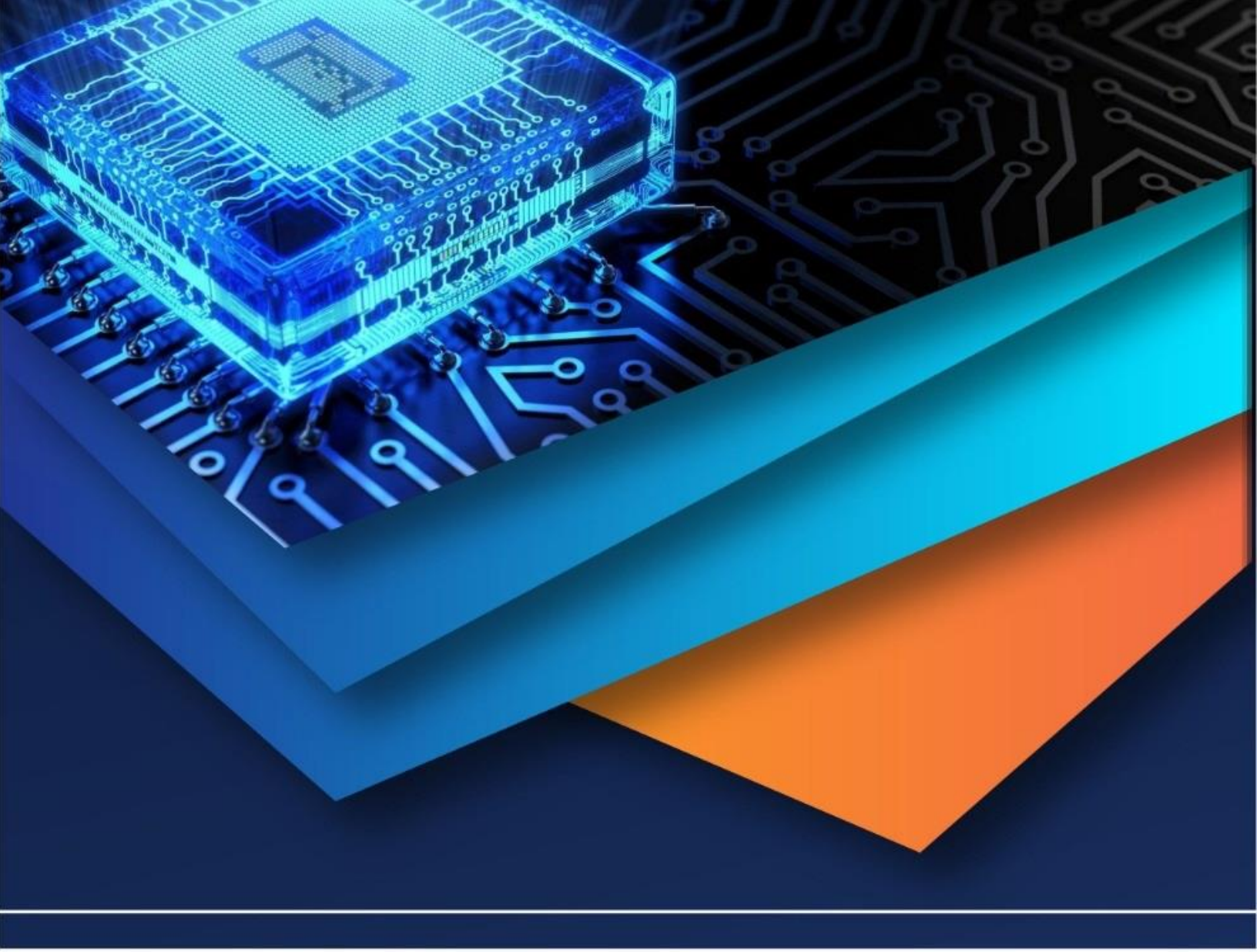
V. CONCLUSION

In conclusion, this research paper aimed to compare the performance of various regression models and their combination with ensemble models for the predic-tion of red wine quality. The evaluation was conducted on a dataset comprising the physicochemical properties of red wines.

The results of the evaluation in-dicated that Random Forest (RF) performed the best among the individual regression models. Furthermore, the combination of Random Forest with Bag-gingRegressor (RF + BR) outperformed the other combinations with ensemble models. It achieved the lowest MAE, RMSE, and highest R2 score, indicating improved prediction accuracy and a stronger fit to the data. These findings highlight the effectiveness of Random Forest in capturing complex relationships in the red wine quality dataset. The combination with BaggingRegressor further enhanced its performance by reducing overfitting and improving generalization. This comparative assessment of regression models and their combination with ensemble models highlights the importance of model selection in accurately predicting red wine quality. The results underscore the superior performance of Random Forest and its combination with Bagging Regressor, offering practical implications for the wine industry and consumers alike.

REFERENCES

[1] Qingwen Zeng. Prediction of wine quality using ensemble learning approach of machine learning. In 2022 International Conference on mathematical statistics and economic analysis (MSEA 2022), pages 770–774. Atlantis Press, 2022.
[2] KR Dahal, JN Dahal, H Banjade, and S Gaire. Prediction of wine quality using machine learning algorithms. Open Journal of Statistics, 11(2):278– 289, 2021.
[3] Satyabrata Aich, Ahmed Abdulhakim Al-Absi, Kueh Lee Hui, John Tark Lee, and Mangal Sain. A classification approach with different feature sets to predict the quality of different types of wine using machine learning tech-niques. In 2018 20th International conference on advanced communication technology (ICACT), pages 139–143. IEEE, 2018.
[4] Akanksha Trivedi and Ruchi Sehrawat. Wine quality detection through machine learning algorithms. In 2018 International Conference on Re-cent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE), pages 1756–1760. IEEE, 2018.
[5] Chao Ye, Ke Li, and Guo-zhu Jia. A new red wine prediction framework using machine learning. In Journal of Physics: Conference Series, volume 1684, page 012067. IOP Publishing, 2020.
[6] Gongzhu Hu, Tan Xi, Faraz Mohammed, and Huaikou Miao. Classification of wine quality with imbalanced data. In 2016 IEEE International Confer-ence on Industrial Technology (ICIT), pages 1712–1217. IEEE, 2016.
[7] Andy Liaw, Matthew Wiener, et al. Classification and regression by ran-domforest. R news, 2(3):18–22, 2002.
[8] Joao Mendes-Moreira, Carlos Soares, Al´ıpio M´ario Jorge, and Jorge Freire De Sousa. Ensemble approaches for regression: A survey. Acm computing surveys (csur), 45(1):1–40, 2012.
[9] Dheeru Dua and Casey Graff. UCI machine learning repository, 2019.
[10] James Chen. Skewness: Positively and negatively skewed defined with formula. Investopedia, 2023.
[11] Bala Priya C. How to detect outliers in machine learning – 4 methods for outlier detection. FreeCodeCamp, 2022.
[12] Vishwa Pardeshi. Linear regression model for machine learning. Towards Data Science, 2020.
[13] Afroz Chakure. Random forest regression in python explained. BuiltIn, 2023.
[14] Saed Sayad. Decision tree regression, 2022.
[15] Tapas Roy. Unlocking the true power of support vector regression. Towards Data Science, 2019.
[16] Ajitesh Kumar. Sklearn neural network example - mlpregressor. Vitalflux, 2023.
[17] NVIDIA. Xgboost. https://www.nvidia.com/en-us/glossary/ data-science/xgboost/#:.:text=XGBoost%2C%20which%20stands% 20for%20Extreme,%2C%20classification%2C%20and%20ranking% 20problems., Unknown. Accessed: June 4, 2023
[18] scikit-learn contributors. AdaBoostRegressor: Machine learning in Python. https://scikit-learn.org/stable/modules/generated/ sklearn.ensemble.AdaBoostRegressor.html, 2021. Accessed: June 4, 2023.
[19] Packt Subscription. Bagging regressors. https://subscription. packtpub.com/book/data/9781789136609/5/ch05lvl1sec26/ bagging-regressors, 2019. Accessed: June 4, 2023.
[20] nikki2398. Gradient boosting in ml. GeeksforGeeks, 2023.
[21] Ibrahim Abayomi Ogunbiyi. Evaluation metrics for regression problems in machine learning. FreeCodeCamp, 2023.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY