Journal of Engineering, Management and Information Technology

ASPUR PUBLISHING

Vol. 03, No. 02 (2025) 113-120, doi: doi:10.61552/JEMIT.2025.02.005 - http://jemit.aspur.rs

ADVANCING SUSTAINABLE AGRICULTURAL DEVELOPMENT AND FOOD SECURITY THROUGH MACHINE LEARNING: A COMPARATIVE ANALYSIS OF CROP YIELD PREDICTION MODELS IN INDIAN AGRICULTURE

Ritika Kapoor
Shellja Mittal
A. Charan Kumari [1]
K. Srinivas

Received 11.01.2024.
Revised 29.03.2024.
Accepted 02.05.2024.

Keywords:

AdaBoost Regressor, agriculture, CatBoost Regressor, crop yield prediction, Machine Learning.

Original research

A B S T R A C T

This paper discusses the crucial issue of improving decision-making in agriculture and ensuring food security, which are key elements of Sustainable Development Goal (SDG) 2: Zero Hunger. Its objective is to eliminate hunger, attain food security, enhance nutrition, and advance sustainable agriculture. Conventional methods for estimating crop yields based on past historical data, weather patterns, and expert insights often lack the scalability and accuracy necessary for sustainable agricultural progress. This investigation delves into the application of machine learning (ML) strategies to forecast crop yields, utilizing a comprehensive dataset from Indian agriculture encompassing factors like fertilizer usage, seasonal fluctuations, and soil compositions. The performance of ten ML algorithms was assessed, including Linear Regression, LASSO, RIDGE, K-Neighbours Regressor, Decision Tree Regressor, Support Vector Regressor, Extreme Gradient Boosting Regressor, AdaBoost Regressor, CatBoost Regressor, and Gradient Boost Regressor. Our results indicate that CatBoost Regressor, K-Neighbours Regressor, and Gradient Boost Regressor demonstrate superior predictive precision, showcasing their potential for practical implementation in the agricultural domain to facilitate the realization of SDG 2. These models present notable reliability and efficacy in predicting crop yields, which are crucial for informed decision-making aimed at optimizing agricultural yields and contributing to sustainable agricultural growth and food security. The study furnishes valuable insights for farmers, policymakers, and supply chain managers, proposing that the adoption of efficient ML models can greatly enhance decision-making processes in agriculture.

1. INTRODUCTION

The heart of the Indian economy is agriculture. The population of the world is estimated to increase to 4.9 billion by 2030 which will lead to an increase in demand for agricultural products (Reddy & Kumar, 2021). Several farmers use pesticides, fertilizers, etc. to increase the fertility of the soil. In this process, the quality of the soil tends to degrade after a certain period. Similarly,

[1] Corresponding author: A. Charan Kumari
Email: charankumari@dei.ac.in

climate conditions (such as temperature, rain, etc.) have a splendid effect on crop yield. For the same reasons, it is a necessity to have exact knowledge about crop yield history whenever there is a need to make decisions about agricultural risk management or to forecast crop yield (Jhajharia et al., 2023). Predicting crop yield will help the farmers and others who are involved in agriculture, to decide what to plant and when to plant.

Various techniques are available for predicting crop production, but few can match the effectiveness and innovation of contemporary machine learning (Mulla & Quadri, 2020). Machine Learning (ML) provides a practical and dependable method. It takes on the responsibility of forecasting by generating a range of calculated datasets that shed light and offer guidance for the future. Previous challenges or outcomes shape the upcoming trajectory and, in some way, establish the inevitable reality. Our trust in these models stems from the understanding that sooner or later the future will become the past, allowing us to alter or forecast it. Our fixation on the future is accompanied by a fear that the resources of our world may deplete and disappear.

An ML model assumes that the output (crop yield) is a nonlinear function of the input variables. The goals of crop yield prediction go beyond simple forecasting; they involve optimizing resource allocation, reducing environmental impacts, and improving resilience to climate change. By using cutting-edge methodologies and advanced technologies, experts are focused on creating predictive models that not only make accurate predictions about crop yields but also generate effective new plans for agricultural usage and can validate various protocols that influence decision-making processes and provide clear and accessible data for policy makers.

The research in this field elaborates on efficient strategies and systems that can be utilized for predicting crop yields, as well as the methodologies involved. They do not only outline significant frameworks that illustrate the challenges faced in this area and what needs to be achieved soon. By doing so, we are examining the assertions of researchers worldwide, exploring the potential replacement of outdated agricultural strategies to mitigate risks to humanity. This is crucial in preventing a decline in living standards caused by inconsistent outcomes and yield levels in the resilient agricultural sector.

The main contributions of this journal paper are:

1. The utilization of a comprehensive model assessment is employed to juxtapose the effectiveness of ten machine learning algorithms for the prediction of crop yields in the agricultural sector of India, offering a broad viewpoint on the current capabilities of predictive technologies.
2. In this work the crop yield predictions of CatBoost Regressor, K-Neighbours Regressor, and Gradient Boost Regressor machine learning models are most accurate, demonstrating superior performance in terms of mean absolute error (MAE), mean squared error (MSE), and root mean square error (RMSE) metrics.
3. By focusing on the improvement of agricultural productivity and food security, the work directly contributes to Sustainable Development Goal (SDG) 2, whose objective is to achieve Zero Hunger by eliminating hunger, achieving food security, improving nutrition, and promoting sustainable agricultural practices/development.
4. It employs extensive preprocessing methodologies to boost the accuracy of data and optimize model efficiency, ensuring that the outcome is compatible with a variety of machine-learning models.
5. The document imparts practical insights for farmers, policymakers, and supply chain managers in the agricultural domain, proposing that the integration of efficient machine learning models can result in more informed decision-making and streamlining of agricultural methodologies.

These contributions highlight the paper's importance in propelling the realm of agricultural data science forward and its potential influence on sustainable agricultural growth and food security.

The paper is structured in the following manner: A literature review provides an in-depth analysis of pertinent research and works related to crop yield prediction. Subsequently, the study's methodology delineates the data acquisition process, pre-processing steps, data partitioning technique, implemented algorithms, and various performance evaluation metrics. The result section compares the efficacies of the algorithms based on the standard metric scores. Lastly, the conclusion and suggestions for further research provide suggestions for future research. All sources cited are presented in the references section.

2. LITERATURE REVIEW

For the review, various research papers in the field have been examined. The papers include data on crops that are sown on the soil of India and all the factors that affect the yield and fertility of the soil (Reddy et al., 2003). Below is the summary of the research in this area. (Jhajharia et al., 2023), employed various models including Random Forest, Support Vector Machine (SVM), Gradient Descent, Long Short-Term Memory, and Lasso Regression. The researchers conducted a comparative analysis of these models and determined that Random Forest exhibited superior performance over the others. (Mulla & Quadri, 2020), employed the Decision Tree algorithm to predict the prices of agricultural commodities produced by farmers, drawing on historical data. The scope of their investigation encompassed crops such as paddy and bajra, typically cultivated during the kharif and rabi seasons. Moreover, the researchers conducted an evaluation of anticipated profits and expenditures over the subsequent twelve-month period,

juxtaposing these figures against those of the preceding year, and delivered an exhaustive analysis of time series data.

Rashid et al. (2021), used several machine learning as well as deep learning algorithms to predict the palm oil crop yield. They also discussed several areas such as remote sensing, disease recognition, the best features considered, etc. (Banu Priya et al., 2021), used factors such as state, crop, year, etc. for their prediction instead of rainfall, nutrients in the soil, etc. They used models such as Random Forest, Decision Tree, and Gradient Boost to perform their prediction and used ensemble algorithms to reap higher predictions and minimize errors (Brillante et al., 2015). When the models were compared, it was inferred that random forest was the most efficient. Thomas (2023) studied and analyzed the use of various machine learning algorithms as well as deep learning algorithms in crop yield prediction. As per their work, DNN, CNN, and LSTM were identified as the most suitable deep-learning algorithms, and random forest and gradient boost as the most used Machine-learning algorithms. (Veenadhari et al., 2014), developed an application to suggest the crops that work districts of Madhya Pradesh (MP), India. The user must input certain factors as input such as temperature, cloud cover, rainfall, and observed yield in the past and the application will forecast the yield (Basso & Liu, 2019). Furthermore, based on the trigger values set, the results would be obtained along with the crops labeled.

(Bhosale et al., 2018), used three algorithms namely clustering k-means, Apriori and Bayes algorithm, then they hybridized the algorithm for better efficiency of yield prediction and they considered parameters like Area, Rainfall, Soil type. (Gandge, 2017), used various machine-learning algorithms for several crops. They used various models such as K-means, SVR, Decision tree, and Bee-hive clustering and used deep learning algorithms namely, Neural networks, etc. The factors that were taken into consideration were soil nutrients such as nitrogen, potassium, phosphate, and soil pH.

3. METHODOLOGY

This section presents the methodology adopted in this work for crop yield prediction.

3.1. Dataset

The dataset consists of a total of 19000 rows with 11 features such as fertilizer, season, state, area, type of soil, distribution of rainfall, etc. The output variable in the dataset is yield. Table 1 describes all the features along with the type of data values they hold.

Table 1. Description of the dataset

Features	Description	Data Category
Crop	Name of the crop cultivated	String
Crop_Year	Year in which crop was grown	Numerical
Season	Cropping Season	String
State	State where the crop was cultivated	String
Area	The area under cultivation (in sq. m)	Numerical
Production	Crop production (Quantity)	Numerical
Annual_Rainfall	Rainfall received by crop growing region (in mm)	Numerical
Fertilizer	Fertilizer used in crop production (in kg)	Numerical
Pesticide	Pesticide used for crop production (in kg)	Numerical
Yield	Crop yield (Calculated as production per unit area)	Numerical

3.2. Pre-Processing

Pre-processing of the dataset was performed as a second step. Data preprocessing is a method that is used to convert the raw data into a clean data set. Furthermore, because the data set contains numeric data, the method of robust scaling was employed, which is a normalization process with the interquartile range.

3.3. Data Partitioning

Later, the dataset is divided into two parts – Testing data and Training data. The training data comprises more than half of the rows from the dataset. The training dataset is the initial dataset that is used to train ML algorithms to learn and output the correct predictions. Once the model is trained, the remaining testing data is fed to the model. Based on the testing data results, it can be confirmed whether the training of the model has been done correctly or not. The train and test sizes used in this paper are 0.8 and 0.2, respectively.

3.4. Algorithms

This sub section presents the machine learning algorithms that have been implemented for the purpose.

3.4.1. Linear Regression

Linear regression is a statistical method utilized to depict the connection between a dependent variable and one or more independent variables, assuming a linear relationship among them. The goal is to determine the coefficients of the linear equation that best suits the given data points by minimizing the sum of squared differences between predicted and observed values. Following the fitting of the regression model, it can be employed for forecasting, inference, and hypothesis testing, offering valuable insights into the impact of changes in independent variables on the dependent variable and enabling the prediction of future observations based on the acquired data relationships. Linear regression remains a crucial tool due to its simplicity, interpretability, and effectiveness in modeling linear relationships.

3.4.2. LASSO

LASSO, an abbreviation for Least Absolute Shrinkage and Selection Operator, is a regression approach utilized for regularization and variable selection. It introduces a penalty term to the standard linear regression objective function, restricting the absolute values of the coefficients. This penalty promotes sparsity in coefficient estimates by reducing some coefficients to zero, facilitating variable selection. Particularly beneficial for high-dimensional datasets with a surplus of predictors compared to observations, LASSO assists in identifying crucial predictors while discarding irrelevant ones, ultimately leading to simpler and more understandable models. Furthermore, LASSO aids in preventing overfitting by penalizing large coefficients, enhancing the generalization capability of the model. This technique finds applications where feature selection and regularization play a vital role in constructing accurate and interpretable predictive models.

3.4.3. Ridge

Ridge regression, also recognized as Tikhonov regularization, is a form of linear regression employed for regularization purposes. Like LASSO, it incorporates a penalty term into the standard linear regression objective function; however, instead of restricting the absolute size of the coefficients, it penalizes the squared magnitude of the coefficients. The ridge penalty in this technique helps in shrinking the coefficients towards zero, aiming to enhance the model's numerical stability and decrease the variance of parameter estimates. Particularly valuable in scenarios of multicollinearity, where independent variables are highly correlated, ridge regression strikes a balance between bias and variance, enhancing the model's overall performance by exchanging some bias for decreased variance. This regularization technique is extensively utilized when predictive modeling with high-dimensional datasets is prevalent.

3.4.4. K-Neighbors Regressor

The K-Neighbors Regressor is an algorithm used in regression to predict continuous target variables based on the k-nearest neighbors' approach. The prediction for a new data point is determined by averaging the target values of its k nearest neighbors, with "k" being a user-defined hyperparameter. A distance metric such as Euclidean or Manhattan distance is typically utilized to gauge the similarity between data points. This algorithm does not assume any specific data distribution, making it beneficial for modeling intricate relationships or when the distribution is unknown. Nonetheless, it may encounter computational challenges, especially with large datasets, as it necessitates storing all training samples and computing distances for each prediction.

3.4.5. Decision Tree Regressor

The Decision Tree Regressor, a machine learning algorithm for regression tasks, constructs a tree-like decision model based on input features to predict continuous target variables. It divides the feature space into regions recursively, selecting splits that minimize target variable variance within each region. Predictions are made by traversing the tree from the root node to a leaf node, where the predicted value equals the average target variable within that leaf. While intuitive and interpretable, Decision Tree Regressor can overfit, especially with deep trees or noisy data. Techniques such as pruning and limiting tree depth are commonly used to mitigate overfitting and improve generalization performance.

3.4.6. Support Vector Regressor

Support Vector Regressor (SVR) is a robust machine learning algorithm for regression tasks, particularly beneficial when traditional linear regression methods are inadequate. SVR identifies a hyperplane that best fits the data points while maximizing the margin, which is the distance between the hyperplane and the nearest data points known as support vectors. Unlike traditional linear regression, SVR can capture non-linear relationships by transforming input features into a higher-dimensional space using a kernel function. SVR aims to minimize the error between actual and predicted values while penalizing deviations beyond a specified

threshold, known as the epsilon-insensitive tube. This approach enables SVR to emphasize capturing pertinent data patterns while disregarding noise.

3.4.7. Extreme Gradient Boosting Regressor

The Extreme Gradient Boosting Regressor, also known as XGB Regressor, is an effective ensemble learning technique applied in regression scenarios. This algorithm is part of the gradient boosting algorithm family, which involves constructing a series of weak learners (specifically decision trees in the case of XGB Regressor) to form a robust predictive model. XGB Regressor constructs trees within a gradient boosting framework, where each subsequent tree rectifies the mistakes of its predecessors. It utilizes gradient boosting, a method that reduces a loss function by gradually incorporating new trees, with each tree optimized to minimize the residual error from previous iterations. XGB Regressor provides numerous benefits, such as its capacity to handle intricate non-linear relationships, interpret feature importance, and scale effectively to large datasets.

3.4.8. AdaBoost Regressor

AdaBoost Regressor, also known as Adaptive Boosting Regressor, is a prominent ensemble learning technique employed in regression tasks. It falls under the boosting algorithms category, which merges numerous weak learners to construct a robust predictive model. AdaBoost Regressor trains a series of weak regressors on the dataset in a sequential manner, with each subsequent regressor focusing more on rectifying the mispredictions made by the preceding ones. It assigns greater weights to the inaccurately classified data points, enabling subsequent regressors to concentrate more on them during the training process. Consequently, AdaBoost Regressor effectively learns from its errors, progressively enhancing its predictive capability. The final ensemble prediction is generated by aggregating the predictions from all weak regressors through a weighted sum.

3.4.9. CatBoost Regressor

CatBoost Regressor, a machine learning algorithm developed by Yandex, is tailored for regression tasks. It is an extension of gradient boosting algorithms and excels in managing categorical features in tabular data without extensive preprocessing. The algorithm of CatBoost Regressor automatically encodes categorical variables to maintain their original meanings, thereby reducing overfitting risks and enhancing model performance. It integrates techniques like ordered boosting and feature combinations to improve prediction accuracy. With inherent support for numerical features and missing values, CatBoost Regressor is user-friendly and deployable.

3.4.10. Gradient Boost Regressor

Gradient Boost Regressor is a machine learning technique employed in regression scenarios, which falls under the umbrella of ensemble learning methodologies. The principle behind this method involves aggregating the forecasts of numerous weak learners, typically decision trees, in a sequential manner to establish a robust predictive model. In the context of Gradient Boosting Regressor, each subsequent weak learner is trained to reduce the residual errors generated by its predecessors. This optimization is accomplished by training the new learner on the negative gradient of a designated loss function, such as mean squared error. The amalgamation of predictions from all weak learners culminates in the final ensemble prediction. Noteworthy characteristics of Gradient Boost Regressor include its superior predictive accuracy, resilience against overfitting, and capacity to handle diverse data formats.

3.5. Evaluation of Model Performance

To assess the effectiveness of the implemented algorithms, benchmark performance evaluation metrics have been analyzed to evaluate prediction accuracy, including the mean absolute error, mean squared error and root mean square error.

3.5.1. Mean Absolute Error (MAE)

Mean Absolute Error (MAE) is a commonly employed metric for assessing the efficacy of a regression model. It quantifies the mean absolute discrepancy between the anticipated values and the observed values.

Mathematically, MAE can be represented as shown in equation 1.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \bar{y}_i| \quad (1)$$

Where, n is the number of data points, y_i is the actual value of the target variable for the ith data point and $\bar{y}_i$ is the predicted value of the target variable for the ith data point.

MAE offers a singular gauge of the model's forecasting precision, with reduced values indicating superior performance. It is frequently utilized in regression assignments that involve outliers or prioritize the assessment of error magnitudes rather than their squared equivalents.

3.5.2. Mean Squared Error (MSE)

Mean Squared Error (MSE) serves as a prevalent metric for appraising the effectiveness of regression models. It calculates the mean squared deviation between the forecasted values and the real values.

Mathematically, MSE can be represented as shown below in equation 2.

$$MSE = \frac{1}{n}(y_i - \bar{y}_i)^2 \quad (2)$$

Where, n is the number of data points, y_i is the actual value of the target variable for the ith data point and $\bar{y}_i$ is the predicted value of the target variable for the ith data point.

MSE furnishes a singular assessment of the model's predictive precision, with lower values signifying enhanced performance. This metric accentuates significant errors more than minor ones and is sensitive

to outliers. MSE is frequently applied in regression tasks that necessitate reducing the average scale of errors.

3.5.3. Root Mean Square Error (RMSE)

Root Mean Square Error (RMSE) is a widely accepted metric for evaluating the performance of regression models. It computes the square root of the average squared difference between projected values and actual values.

Mathematically, RMSE can be represented as shown below in equation 3.

$$RMSE = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_i)^2\right)} \quad (3)$$

Where, n is the number of data points, y_i is the actual value of the target variable for the i[th] data point and $\bar{y}_i$ is the predicted value of the target variable for the i[th] data point.

RMSE provides a singular evaluation of the model's predictive precision, with lower values indicating superior performance. Its utility lies in the fact that it shares the same units as the target variable, facilitating interpretation. RMSE is commonly employed in regression assignments that emphasize minimizing the average scale of errors.

4. RESULTS

This section presents the results obtained by various implemented algorithms for crop yield prediction.

Table 2 presents the standard performance metric scores obtained by all the models.

Table 2. Comparison of performance of the algorithms based on the standard metric scores

Model	Mean Absolute Error (MAE)	Mean Squared Error (MSE)	Root Mean Square Error (RMSE)
Linear Regression	0.0061	0.0015	0.0389
LASSO	0.0149	0.0080	0.0895
Ridge	0.0061	0.0015	0.0390
K-Neighbour Regressor	0.0012	0.0003	0.0176
Decision Tree Regressor	0.0012	0.0004	0.0217
Support Vector Regressor	0.0648	0.0052	0.0727
XGB Regressor	0.0015	0.0006	0.0256
AdaBoost Regressor	0.0021	0.0006	0.0245
CatBoost Regressor	0.0011	0.0001	0.0137
Gradient Boost Regressor	0.0013	0.0002	0.0161

The examination of machine learning algorithms utilizing standard metric scores for predicting crop yield demonstrates notable variations in their predictive performance and dependability. Linear Regression and Ridge Regression present comparable MAE and MSE values, with a minor difference in RMSE, indicating that the regularization in Ridge Regression has a minimal effect on model accuracy. Nonetheless, their moderate performance in comparison to other models suggests possible limitations in grasping complex, nonlinear relationships within the dataset. In contrast, LASSO Regression shows higher error metrics across all three criteria, suggesting that its feature selection process may not be beneficial for this dataset. This could be attributed to the requirement of all features in predicting crop yield or the model's failure to efficiently capture intricate data relationships. K-Neighbours Regressor and Decision Tree Regressor demonstrate low error metrics, indicating high predictive precision and proficiency in capturing nonlinear patterns without assuming a specific data distribution. The slight disparity in their RMSE values may be due to the inherent characteristics of each method, with K-Neighbours Regressor potentially benefiting from its adaptability to local data structures. Conversely, Support Vector Regressor exhibits the highest error metrics among the tested models, indicating its limited effectiveness for this dataset. This could be linked to SVR's sensitivity to hyperparameters, emphasizing the importance of selecting suitable models for agricultural decision-making and optimization.

The juxtaposition of ensemble models like CatBoost Regressor and Gradient Boost Regressor emphasizes their exceptional effectiveness in crop yield prediction, highlighting their robustness and efficiency in handling complex, nonlinear data structures. These models excel by incorporating weak learners into a powerful predictive framework, enabling them to identify intricate patterns within the dataset, a crucial aspect for accurate crop yield prediction. AdaBoost Regressor and XGB Regressor also demonstrate strong performance, albeit with slightly higher error metrics, indicating that while they are potent tools, their effectiveness may be influenced by the specific characteristics of the data and model setup.

The results underscore the importance of ensemble models, particularly CatBoost Regressor and Gradient Boost Regressor, in crop yield prediction due to their superior performance and reliability. Additionally, the analysis underscores the potential of K-Neighbours Regressor and Decision Tree Regressor as promising alternatives, given their competitive error metrics. On the contrary, traditional regression models and Support Vector Regressor show lower effectiveness, highlighting the need for models that can effectively handle the complexities inherent in agricultural data. This comparison underscores the transformative role of machine learning in enhancing agricultural decision-making processes and productivity, with a clear

preference for models capable of capturing and predicting the intricate nuances of crop yield data.

5. CONCLUSION AND FUTURE WORK

This study conducts a thorough investigation into a wide range of machine learning algorithms to predict crop yields, a critical component of achieving Sustainable Development Goal (SDG) 2: Zero Hunger, which highlights the importance of sustainable agricultural development and food security. Through the analysis of an extensive dataset from Indian agriculture, including factors such as fertilizer usage, seasonal variations, soil compositions, among others, CatBoost Regressor, K-Neighbours Regressor, and Gradient Boost Regressor have emerged as the top-performing models due to their remarkable predictive accuracy and reliability. The findings of this research not only underscore the potential of advanced ML techniques in transforming agricultural practices but also offer guidance to stakeholders in the agricultural sector, such as farmers, policymakers, and supply chain managers, enabling them to make informed decisions that could significantly improve crop yields and sustainability. The comparative assessment presented in this paper stresses the significance of selecting suitable ML models based on specific criteria like Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Square Error (RMSE) to ensure optimal predictive performance. Moreover, this study contributes to the broader discourse on the application of data science in agriculture, suggesting that the strategic use of ML can enhance the resilience of food systems and contribute to alleviating global hunger and food insecurity. The practical implications of this study are clear: by adopting the most effective ML models, agricultural stakeholders can increase productivity, reduce waste, and progress towards fulfilling sustainable development goals.

Future research endeavors in predicting crop yields could entail incorporating satellite imagery and weather forecast data to enhance model accuracy, refining feature engineering methods, exploring multi-modal AI learning models, validating models in diverse real-world agricultural settings, and engaging with stakeholders to ensure practicality and generalizability.

References:

BanuPriya, N., Tejasvi, D., & Vaishnavi, P. (2021). Crop yield prediction based on Indian agriculture using machine *learning. International Journal of Modern Agriculture*, 10(3), 73-82.

Basso, B., & Liu, L. (2019). Seasonal crop yield forecast: *Methods, applications, and accuracies. advances in agronomy*, 154, 201-255.

Bhosale, S. V., Thombare, R. A., Dhemey, P. G., & Chaudhari, A. N. (2018, August). Crop Yield Prediction Using Data Analytics and Hybrid Approach. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-5). IEEE.

Brillante, L., Gaiotti, F., Lovat, L., Vincenzi, S., Giacosa, S., Torchio, F., Segade R. S., Rolle L., & Tomasi, D. (2015). Investigating the use of gradient boosting machine, random forest and their ensemble to predict skin flavonoid content from berry physical–mechanical characteristics in wine grapes. *Computers and Electronics in Agriculture*, 117, 186-193.

Gandge, Y. (2017, December). A study on various data mining techniques for crop yield prediction. In 2017 International Conference on Electrical, Electronics, Communication, Computer, and Optimization Techniques (ICEECCOT) (pp. 420-423). IEEE.

Jhajharia, K., Mathur, P., Jain, S., & Nijhawan, S. (2023). Crop yield prediction using machine learning and deep learning techniques. *Procedia Computer Science*, 218, 406-417.

Mulla, S. A., & Quadri, S. A. (2020). Crop-yield and price forecasting using machine learning. *International journal of analytical and experimental modal analysis*, 12(8), 1731-1737.

Rashid, M., Bari, B. S., Yusup, Y., Kamaruddin, M. A., & Khan, N. (2021). A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction. *IEEE access*, 9, 63406-63439. DOI: 10.1109/ACCESS.2021.3075159.

Reddy, B. V. S., Reddy, P. S., Bidinger, F., & Blümmel, M. (2003). Crop management factors influencing yield and quality of crop residues. *Field Crops Research*, 84(1-2), 57-77.

Reddy, D. J., & Kumar, M. R. (2021, May). Crop yield prediction using machine learning algorithm. In 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 1466-1470). IEEE.

Thomas, D. (2023, May). A review of crop yield prediction based on Indian agriculture sector using machine learning. In AIP Conference Proceedings (Vol. 2773, No. 1). AIP Publishing.

Veenadhari, S., Misra, B., & Singh, C. D. (2014, January). Machine learning approach for forecasting crop yield based on climatic parameters. In 2014 International Conference on Computer Communication and Informatics (pp. 1-5). IEEE.

Ritika Kapoor
Dayalbagh Educational Institute,
Dayalbagh, Agra,
India.
ritikakapoor527@gmail.com
ORCID: 0009-0005-8433-751X

Shellja Mittal
Dayalbagh Educational Institute,
Dayalbagh, Agra,
India.
shelljamittal1305@gmail.com
ORCID: 0009-0004-7261-2628

A. Charan Kumari
Dayalbagh Educational Institute,
Dayalbagh, Agra,
India.
charankumari@dei.ac.in
ORCID: 0000-0002-3160-1912

K. Srinivas
Dayalbagh Educational Institute,
Dayalbagh, Agra,
India.
ksri12@gmail.com
ORCID: 0009-0002-3884-6282