JOURNAL OF MULTIDIMENSIONAL RESEARCH & REVIEW

http://www.jmrr.org Volume: 6, Issue: 1, April 2025 | ISSN: 2708-9452

Enhancing Indian Agriculture with Data Insights Using Machine Learning

Ahalya M
PG Student, Department of Computer Science and Applications
Vivekanandha College of Arts and Sciences for Women [Autonomous],Tiruchengode
Namakkal, Tamilnadu, India.
Rajagopal D
Assistant Professor & Deputy Controller of Examinations,
Vivekanandha College of Arts and Sciences for Women [Autonomous],Tiruchengode,
Namakkal, Tamilnadu, India.

Abstract

Agriculture is the backbone of the Indian economy, but farmers face problems related to unpredictable weather patterns, soil degradation, and inefficient resource use. The "Enhancing Indian Agriuclure with Data Insights using Machine Learning" project aims to boost agricultural productivity in India by leveraging data analysis. Using data such as temperature, soil nutrient, rainfall, humidity, pH, area, and crop production, we analyze trends and patterns to provide actionable insights. Machine learning techniques are used to determine corp yield and recommend suitable crop types, enabling better decision-making for farmers and policymakers. This study focuses on predicting crop yield, identifying suitable crop types according to the soil, optimizing resource allocation, and reducing risks associated with climate variability.

Keywords: Agriculture, Data Analysis, Machine Learning, Crop Yield Prediction, Soil Nutrients, Rainfall Patterns, Weather Data, Agricultural Sustainability, Indian Farming.

1 Introduction

Agriculture plays a important role in India's economy with large population and contributing significant GDB. However, challenges such as unpredictable weather, soil degradation, inefficient resource utilization, and declining crop yields hinder agricultural productivity. Traditional farming practices often fail to optimize yield due to a lack of data-driven insights. With the increasing availability of agricultural datasets, machine learning (ML) techniques have emerged as powerful tools to analyze historical data and provide actionable insights for better decision-making.

This survey explores various ML-based approaches used in agriculture, specially for predicting crop yields and determining suitable crop types. ML models can enhance productivity, reduce risks, and promote sustainable farming. We review recent studies that apply ML techniques such as Random Forest, Neural Networks, and Regression models to improve yield predictions and resource management. The goal of this paper is to highlight the significance of data-driven agriculture and discuss how ML models can contribute to more efficient and resilient farming practices in India.

1.1 Introduction to Machine Learning

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that allows systems to learn from data, recognize patterns, and make decisions with minimal human intervention. Unlike traditional programming, where explicit rules are defined, ML models improve their performance over time by analyzing large datasets. ML is broadly categorized into three types:

1. Supervised Learning – The model learns from labeled data, mapping inputs to known outputs. Common applications include classification and regression.

2. Unsupervised Learning – The model identifies hidden patterns or relationships in data without predefined labels, commonly used for clustering and anomaly detection.

3. Reinforcement Learning – The model interacts with an environment, learning through trial and error using rewards and penalties.

In agriculture, ML plays a crucial role in optimizing farming practices, predicting crop yields, analyzing weather patterns, and improving resource management. By leveraging ML techniques, data-driven insights can enhance agricultural productivity, sustainability, and decision-making for farmers and policymakers.

1.2 Importance of Machine Learning in Agriculture

Machine learning plays a crucial role in modernizing agriculture by providing data-driven insights that enhance productivity, sustainability, and efficiency. Some key applications include:
Crop Yield Prediction – ML models analyze historical data, soil conditions, and weather patterns to forecast crop yields, helping farmers make informed decisions.
Soil and Crop Health Monitoring – ML-powered image recognition and remote sensing techniques detect diseases, nutrient deficiencies, and pest infestations at early stages.
Weather Forecasting – Predictive analytics helps farmers prepare for climate variations,

reducing risks related to droughts, floods, and extreme weather conditions.
Precision Agriculture – Real-time data from sensors and drones allow for precise irrigation, fertilization, and pest control, optimizing resource usage.
Market Price Prediction – ML models analyze demand-supply trends to predict crop prices, helping farmers maximize profitability.

1.3 Objective of the Paper

The objective of this paper is to explore the application of machine learning techniques in agriculture to enhance crop yield prediction and crop type selection, contributing to data-driven agriculture in India. Specifically, this study aims to: Analyze Existing Research – Review a literature survey on various ML algorithms used in agriculture for crop yield prediction and crop selection.
Evaluate ML Algorithms – Compare the performance of different ML models, such as Random Forest, KNN, Naïve Bayes, Decision Tree, and SVM in agricultural applications.
Identify Challenges – Highlight the limitations and challenges in implementing ML-based solutions in agriculture, including data availability, accuracy, and scalability.
Explore Future Trends – Discuss emerging technologies like AI, deep learning, IoT, and remote sensing, and their potential impact on precision farming and real-time analytics.
This study aims to imporve agricultural efficiency and sustainability through data-driven solutions.

2 Literature Survey

Dr. G. Suresh [1] analyzed machine learning models for predicting Indian agricultural crop yields across major seasons from 1997 to 2020. The study found that ensemble models, particularly Random Forest, outperformed others, achieving the highest accuracy with the lowest errors (MAE, RMSE, RAE). Traditional models like Linear Regression performed poorly due to their inability to handle non-linearity, while Multilayer Perceptron showed mixed results with high computation time. Random Tree and REP Tree offered a balance between speed and accuracy. Overall, Random Forest was identified as the most effective model for crop yield prediction. Pritesh Patil et al.[2] proposed a Crop Selection and Yield Prediction System using machine learning techniques based on weather and soil parameters. The study focused on districts in Maharashtra and implemented Random Forest Regression for yield prediction, achieving an R^2 score of 0.96 and a Mean Absolute Error (MAE) of 0.64. For crop classification, the Naïve Bayes classifier demonstrated the highest accuracy at 99.39%. The study suggests enhancements through real-time IoT data collection, inclusion of irrigation and fertilizer parameters, and mobile app integration for practical use. These insights contribute to data-driven decision-making in Indian agriculture, aligning with efforts to optimize crop selection and yield estimation. Madhuri Shripathi Rao et al. [3] conducted a comparative study on crop prediction using machine learning, evaluating K-Nearest Neighbors (KNN), Decision Tree, and Random Forest. Their results showed that Random Forest (Gini and Entropy) achieved the highest accuracy of 99.32%, while Decision Tree (Gini: 98.86%) performed better than KNN (97.04%). The study highlights the effectiveness of ensemble methods for agricultural predictions and suggests future work incorporating deep learning models like ANN and CNN for improved classification. This aligns with the goal of leveraging machine learning

for enhanced crop prediction in Indian agriculture. Akshay Kumar Gajula et al. [4] proposed a crop and yield prediction model using the K-Nearest Neighbors (KNN) algorithm for feature extraction and classification. The model considers soil properties (N, P, K, pH, temperature) to determine the most suitable crop and expected yield. A web-based GUI was developed to facilitate user input and display results. The approach provides insights into crop selection, yield estimation, and fertilizer requirements, aiding farmers in optimizing agricultural decisions. However, the model is limited by data constraints and does not account for climatic disasters. Future enhancements include geospatial analysis for improved accuracy. Anakha Venugopal et al. [5] conducted a study on crop yield prediction using machine learning, comparing Logistic Regression, Naïve Bayes, and Random Forest algorithms. The study found that Random Forest provided the highest accuracy (92.81%) due to its ability to analyze crop growth in relation to climatic and biophysical changes. The algorithm leveraged bagging techniques to enhance prediction performance. Naïve Bayes (91.50%) and Logistic Regression (87.8%) were also tested but showed lower accuracy. The research highlighted Random Forest as the most effective model for crop yield prediction, offering a reliable decision-support tool for farmers. Mayank Champaneri et al. [6] explored Crop Yield Prediction Using Machine Learning, utilizing Random Forest Classifier as a primary model due to its superior performance in both classification and regression tasks. The study highlights how supervised learning techniques can be applied to labeled agricultural data to predict crop yield with high accuracy. The Random Forest algorithm, known for its ensemble learning approach, was found to provide robust and reliable predictions. The authors suggest that increasing the number of trees in the model enhances prediction accuracy, and further improvements can be achieved by integrating real-time data collection. Nischitha K [7] developed a machine learning-based system for predicting rainfall and crop selection. The study utilized the Support Vector Machine (SVM) algorithm for rainfall forecasting and a Decision Tree algorithm for crop prediction. By analyzing historical weather data, the system aimed to improve prediction accuracy, offering valuable insights to help farmers make informed decisions. Kumar Rajak et al. [8] proposed a crop recommendation system using ensemble machine learning techniques to improve prediction accuracy. The study utilized the Majority Voting technique, combining multiple models such as Support Vector Machine (SVM), Naïve Bayes, Multi-Layer Perceptron (MLP), and Random Forest. These models helped classify and predict suitable crops based on soil parameters. A rule-based approach was also implemented, where specific soil conditions (e.g., pH, depth, and water-holding capacity) determined the recommended crop. The system aimed to enhance agricultural productivity by guiding farmers in crop selection. Future improvements include integrating larger datasets and incorporating yield prediction models for better decision-making. S. Pudumalar et al. [9] proposed a Crop Recommendation System for Precision Agriculture using machine learning models, including Random Tree, CHAID, K-Nearest Neighbor (KNN), and Naïve Bayes. The model achieved an accuracy of 88%, utilizing a rule-based approach where if-then rules were generated for crop recommendations. A web-based GUI recommendation system was developed to provide real-time suggestions based on soil and environmental conditions. This study highlights the potential of ensemble techniques and rule-based learning in aiding farmers with informed crop selection, aligning with the objective of enhancing Indian agriculture through data-driven insights. The literature survey highlights the effectiveness of machine learning models in agricultural prediction, particularly in crop classification and yield estimation. Random Forest emerges as the most accurate model due to its ability to handle complex,

non-linear relationships in agricultural data. Other models like KNN, Naïve Bayes, and Neural Networks also provide valuable insights, depending on the dataset and objectives. While these models significantly improve decision-making in farming, challenges such as data availability, climatic uncertainties, and real-time adaptability remain. Future work should focus on integrating more diverse datasets, geospatial analysis, and real-time environmental factors to enhance model reliability and practical applicability in agriculture.

3 Research Methodology

Machine learning algorithms play a crucial role in analyzing agricultural data and making accurate predictions. Various algorithms are used to predict crop yield and crop type. The following are some commonly used ML algorithms in agriculture:

3.1 Random Forest

Random Forest is an ensemble learning technique that constructs multiple decision trees and aggregates their outputs for more accurate predictions. It is highly effective in crop yield prediction and soil classification due to its ability to handle large datasets and non-linear relationships.

Mathematical Formula:

- For classification: $\hat{y} = \text{mode}(y_1, y_2, \ldots, y_n)$
- For regression: $\hat{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$

3.2 K-Nearest Neighbors (KNN)

KNN is a simple, non-parametric algorithm that classifies data based on the nearest training samples.

Algorithm Steps:

1. Choose the number of neighbors (K).
2. Compute the distance using Euclidean distance: $d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
3. Identify the K nearest points.
4. Assign the most common class (classification) or average (regression).

3.3 Naïve Bayes

A probabilistic classifier based on Bayes' theorem, assuming independence between features.

Steps:

- Prior probability: $P(C) = \frac{\text{samples in class C}}{\text{total samples}}$
- Likelihood: $P(X|C) = P(x_1|C) \cdot P(x_2|C) \cdots P(x_n|C)$
- Posterior: $P(C|X) = \frac{P(X|C)P(C)}{P(X)}$

3.4 Decision Tree

Decision Trees split data into hierarchies based on feature values.

Formulas:

- Gini Index: $\text{Gini} = 1 - \sum_{i=1}^{n} P_i^2$
- Information Gain: $IG = H(\text{parent}) - \sum_i \frac{N_i}{N} H(i)$

3.5 Support Vector Machines (SVM)

A supervised algorithm for classification and regression.

Equations:

- Hyperplane: $w \cdot x + b = 0$
- Optimization: $\min \frac{1}{2}\|w\|^2$ subject to $y_i(w \cdot x_i + b) \geq 1$

3.6 Linear Regression

Models relationships between variables.

Equations:

- $y = mx + c$
- $m = \frac{N\sum xy - \sum x \sum y}{N\sum x^2 - (\sum x)^2}$
- $c = \frac{\sum y - m\sum x}{N}$

3.7 REP Tree (Reduced Error Pruning Tree)

A decision tree that prunes branches to reduce overfitting.

Error Calculation: $\text{Error}_{pruned} = \text{Error}_{unpruned} - \text{Reduction}$

3.8 Accuracy Metrics

3.8.1 Regression Metrics

- Mean Absolute Error (MAE): $\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$
- Root Mean Squared Error (RMSE): $\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
- Relative Absolute Error (RAE): $\text{RAE} = \frac{\sum |y_i - \hat{y}_i|}{\sum |y_i - \bar{y}|} \times 100$
- R-squared: $R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$

3.8.2 Classification Metrics

For crop classification, models are evaluated using the following metrics:

- Accuracy: The proportion of correctly classified instances out of the total instances. Higher accuracy indicates better model performance.
- Precision: Measures how many of the predicted positive cases are actually correct, important when false positives need to be minimized.
- Recall (Sensitivity): Indicates how many actual positive cases the model correctly identified, crucial when false negatives are costly.
- F1-score: The harmonic mean of precision and recall, providing a balanced measure when class distribution is imbalanced.
- Confusion Matrix: A detailed table showing the number of true positives, true negatives, false positives, and false negatives, helping in model evaluation.

4 Proposed Method

4.1 Data Collection

The research utilizes diverse datasets to facilitate both yield prediction and crop prediction. The Yield Prediction Dataset contains key agricultural parameters essential for estimating crop yield, including State, District, Crop, Crop_Year, Season, Area, Production, and Yield. On the other hand, the Crop Prediction Dataset incorporates soil and environmental factors that influence crop selection, with features such as N (Nitrogen), P (Phosphorus), K (Potassium), Temperature, Humidity, pH, Rainfall, and Label (Crop Type)

4.2 Data Preprocessing

To ensure data quality and reliability, preprocessing steps are applied to both the yield prediction and crop prediction datasets.

- Handling Missing Values: Missing entries are addressed using statistical imputation techniques such as mean, median, or mode replacement. Even we can drop the rows and columns.
- Feature Scaling: Continuous variables (e.g., temperature, rainfall, and soil nutrients) are normalized to maintain consistency in data distribution.
- Encoding Categorical Variables: Non-numeric attributes like State, District, Crop, and Season are transformed into numerical values using label encoding or one-hot encoding.
- Outlier Detection and Removal: Extreme values that could distort the model's performance are identified and handled appropriately.
- Data Splitting: The dataset is divided into training(80%) and testing(20%) sets to evaluate model performance effectively.

4.3 Model Selection

Selecting the right machine learning model is important for achieving accurate predictions in both yield prediction and crop prediction tasks. Crop Prediction: A classification model is required to determine the best crop based on soil and environmental factors. Random Forest Classifier or Decision Tree Classifier are preferred due to their ability to capture intricate patterns in data while maintaining high interpretability.
Yield Prediction: A Random Forest Regressor is chosen due to its ability to handle complex, non-linear relationships between agricultural variables. It is robust against overfitting and provides feature importance insights, making it suitable for predicting crop yield.

4.4 Model Training and Validation

Once the machine learning models are selected, they undergo training and validation to ensure optimal performance in both yield prediction and crop prediction tasks. For crop prediction, the Random Forest Classifier (or Decision Tree Classifier) is trained on soil and environmental data to determine the most suitable crop. The model's effectiveness is assessed using accuracy, precision, recall, F1-score, and a confusion matrix to analyze misclassifications. To improve generalization, cross-validation is applied, ensuring the models perform well on unseen data. These trained models provide valuable insights, supporting data-driven decision-making in agriculture for better productivity and resource management.

4.5 Crop Prediction Accuracy:

- Naive Bayes: 0.9955
- Random Forest: 0.9932
- Gradient Boosting: 0.9886
- Decision Tree: 0.9841
- KNN: 0.9818
- SVM: 0.9636

4.6 Yield Prediction Metrics (Random Forest):

- MAE: 3.01
- RMSE: 94.43
- RAE: 1.97%
- R^2: 0.9896

The Random Forest Regressor is trained using historical agricultural data, with hyperparameter tuning applied for optimization. Its performance is evaluated using key metrics: Mean Absolute Error (MAE) of 3.01 yield units, Root Mean Squared Error (RMSE) of

94.43 yield units, Relative Absolute Error (RAE) of 1.97%, and an R-squared (R^2) value of 0.9896, indicating that the model explains 98.96% of the variance in yield.

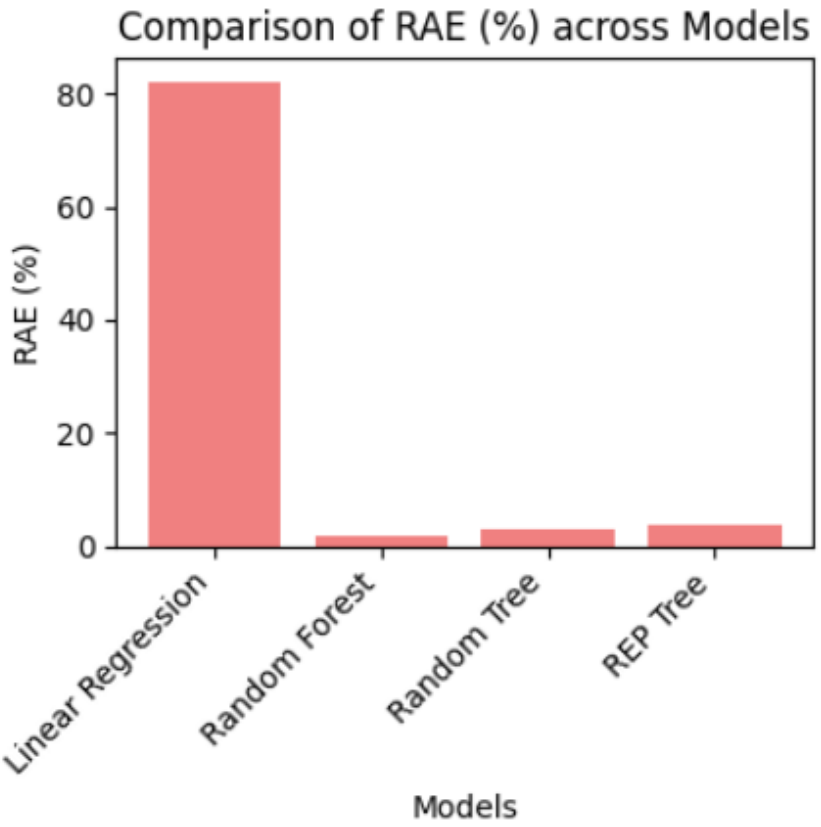

Figure 1: Comparison across four models for yield prediction

Model	MAE	RMSE	RAE (%)	R-squared
Linear Regression	125.50	834.16	82.06	0.187
Random Forest	3.01	94.43	1.97	0.9896
Random Tree	4.81	106.31	3.14	0.9868
REP Tree	5.55	106.31	3.63	0.9868

Table 1: Accuracy comparison of four algorithms across models

5 Results and Discussion

5.1 Crop Prediction

Naive Bayes emerged as the top-performing model with an impressive accuracy of 99.55%, correctly classifying approximately 99.55% of the test instances as shown in the Table:1. Close behind, Random Forest demonstrated similarly strong performance with an accuracy of 99.32%. Gradient Boosting also delivered excellent results, achieving an accuracy of 98.86%. The Decision Tree model followed with a solid accuracy of 98.41%, highlighting its strong predictive capability. K-Nearest Neighbors performed well with an accuracy of 98.18%, slightly trailing the Decision Tree. Meanwhile, the Support Vector

Machine recorded the lowest accuracy among the evaluated models at 96.36%, though it still maintained a respectable level of performance.

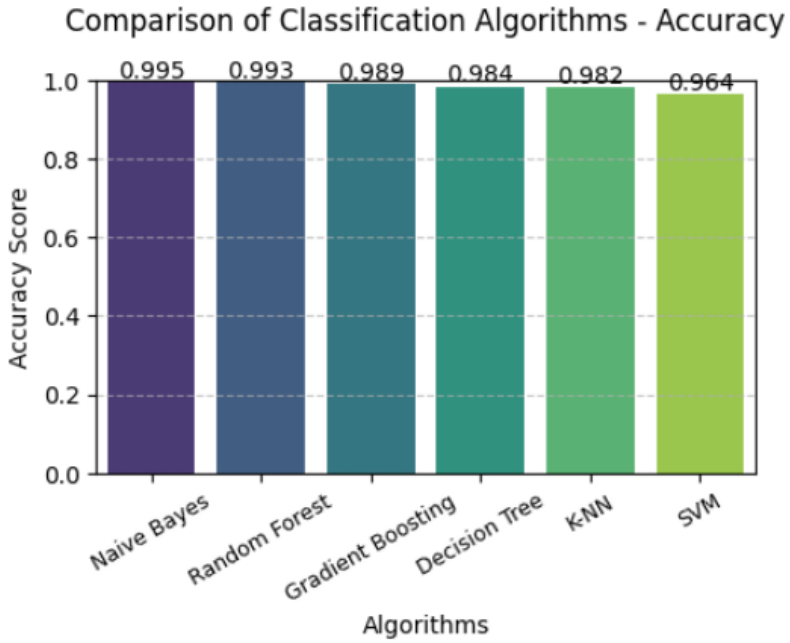

Figure 2: Comparison across MAE across Models

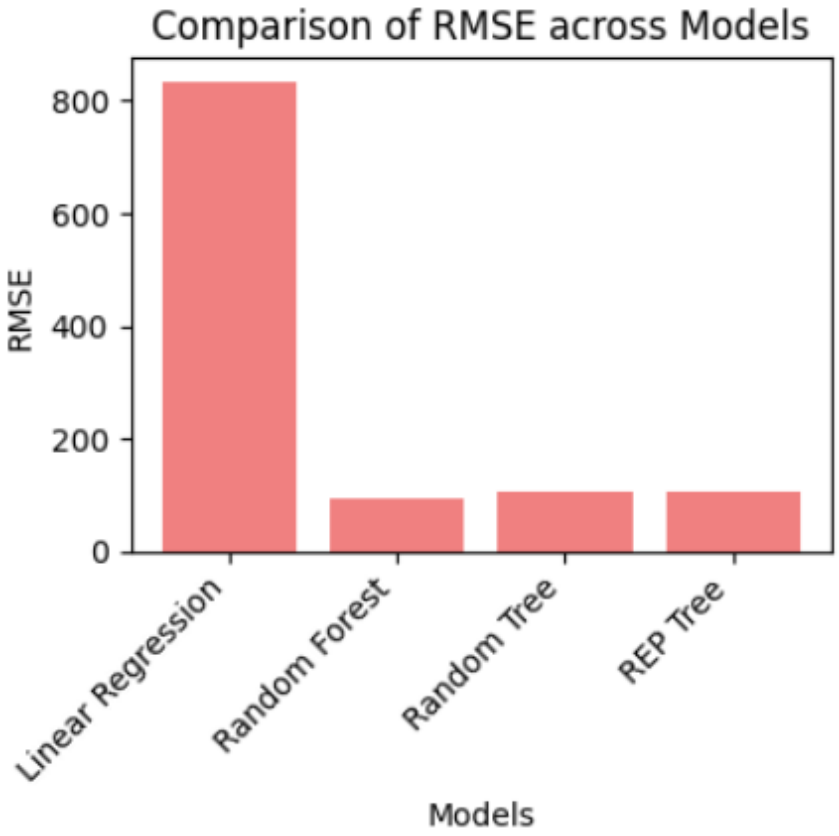

Figure 3: Comparison of RMSE across Models

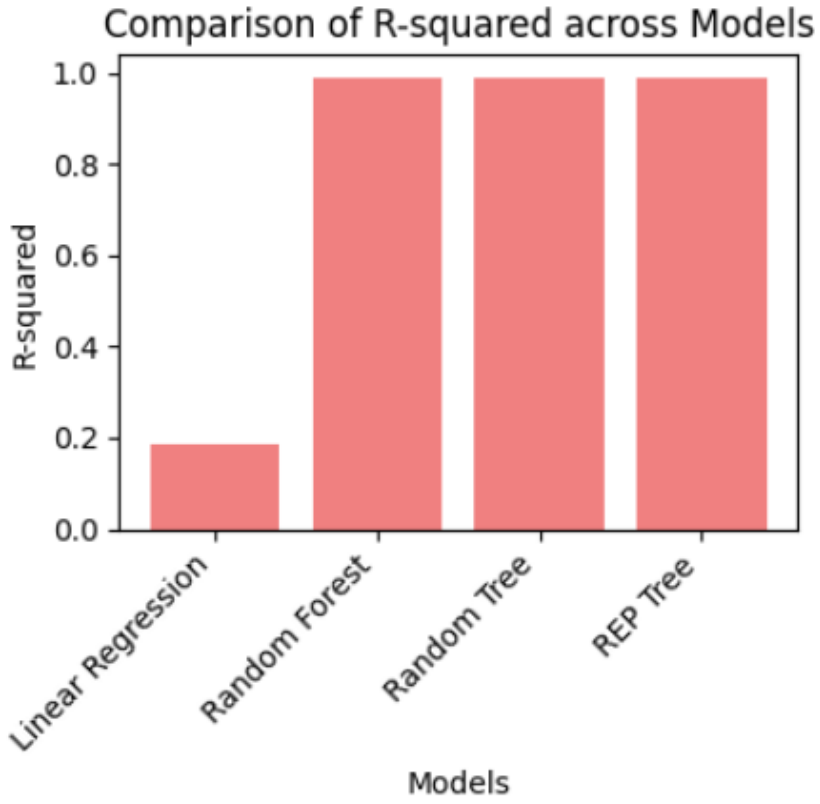

Figure 4: Comparison of RAE(%)across Models

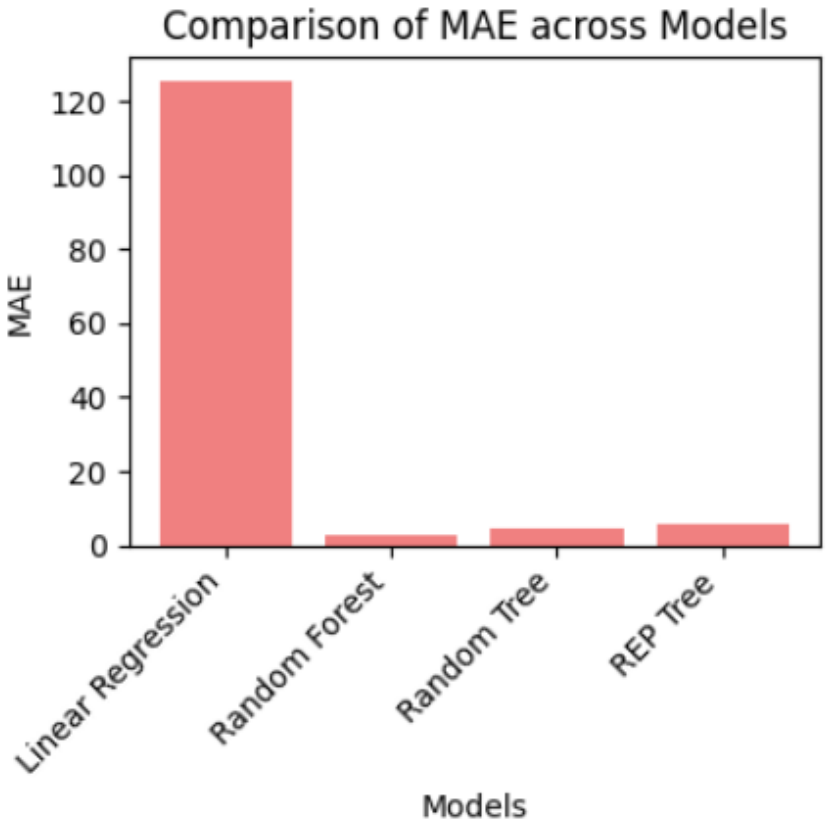

Figure 5: Comparison of R-squared across Models

5.2 Yield Prediction

The performance of four different machine learning models used for predicting crop yield Linear Regression, Random Forest, Random Tree, and REP Tree was evaluated using four key metrics. The Random Forest model, in particular, shows outstanding accuracy: its predictions have an average deviation of approximately 3.01 yield units (MAE), and the typical error, which gives more weight to larger deviations, is around 94.43 yield units (RMSE). The Relative Absolute Error (RAE) is about 1.97%, indicating minimal error relative to the average yield. Additionally, the model achieves an R-squared value of 0.9896 as shown in the Table:2, meaning it explains roughly 98.96% of the variance in yield, demonstrating an excellent fit.

6 Future Scope

The future of agriculture will be driven by AI, IoT, and real-time analytics for smarter and more efficient farming. AI and deep learning will enhance crop yield predictions, automate tasks, and improve pest and disease detection. Real-time analytics will optimize resource use, soil health monitoring, and predictive crop planning based on weather and market trends. IoT and remote sensing using sensors, drones, and satellite imagery will enable real-time monitoring and predictive insights. These advancements will ensure higher productivity, sustainability, and smarter agricultural practices.

7 Conclusion

This research demonstrates how data-driven insights and machine learning can enhance Indian agriculture by improving crop yield prediction and crop selection. By leveraging historical agricultural data, environmental factors, and advanced predictive models, farmers can make informed decisions to optimize productivity. The Random Forest model has shown high accuracy in predicting yields, while classification models effectively identify the most suitable crops based on soil and climate conditions.

References

[1] G. Suresh, "Data Analysis of Indian Agricultural Crop Yield Using Season, Area, and Production Through Machine Learning Models," Nanotechnology Perceptions, ISSN: 1660-6795, Nov. 4, 2024.

[2] P. Patil, P. Athavale, M. Bothara, S. Tambolkar, and A. More, "Crop Selection and Yield Prediction using Machine Learning Approach," Current Agriculture Research Journal, vol. 11, no. 3, pp. 968–980, 2023, ISSN: 2347-4688.

[3] M. S. Rao, A. Singh, N. V. S. Reddy, and D. U. Acharya, "Crop prediction using machine learning," Journal of Physics: Conference Series, vol. 2161, p. 012033, 2022, doi:10.1088/1742-6596/2161/1/012033.

[4] A. K. Gajula, V. C. Dodda, J. Singamsetty, and L. Kuruguntla, "Prediction of crop and yield in agriculture using machine learning technique," in

Proc. 2021 Int. Conf. Comput. Commun. Netw. Technol. (ICCCNT), July 2021, doi:10.1109/ICCCNT51525.2021.9579843.

[5] A. Venugopal, A. S., J. Mani, R. Mathew, and V. Williams, "Crop Yield Prediction using Machine Learning Algorithms," in Proc. Int. Conf. NCREIS, Int. J. Eng. Res. Technol. (IJERT), vol. NCREIS - 2021, 2021, ISSN: 2278-0181.

[6] M. Champaneri, D. Chachpara, C. Chandvidkar, and M. Rathod, "Crop Yield Prediction Using Machine Learning," Int. J. Sci. Res. (IJSR), vol. 9, no. 4, Apr. 2020, ISSN: 2319-7064.

[7] N. K. Nischitha, "Crop Prediction using Machine Learning Approaches," Int. J. Eng. Res. Technol. (IJERT), vol. 9, no. 8, Aug. 2020, ISSN: 2278-0181.

[8] R. K. Rajak, A. Pawar, M. Pendke, P. Shinde, S. Rathod, and A. Devare, "Crop Recommendation System to Maximize Crop Yield using Machine Learning Technique," Int. Res. J. Eng. Technol. (IRJET), vol. 4, no. 12, Dec. 2017.

[9] S. Pudumalar, E. Ramanujam, R. H. Rajashreen, C. Kavyan, T. Kiruthikan, and J. Nishan, "Crop Recommendation System for Precision Agriculture," in Proc. IEEE 8th Int. Conf. Adv. Comput. (ICoAC), 2016.

Cite this article:

Ahalya M & Rajagopal D, "Enhancing Indian Agriculture with Data Insights Using Machine Learning", Journal of Multidimensional Research and Review (JMRR), Vol.6, Iss.2, pp.29-41, 2025