International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM06006
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata

Forecasting Crop Yield using Machine Learning

Pothuraju V V Satyanarayana[1] ,Bonu Jagadeesh[2],Gokarla Sravani[3],Duvvi Deepthi Sudha[4], Chandapu Kishore[5],Althi Durga Prasad[6]

1 Associate Professor, Computer Science and Engineering, Visakha Institute of Engineering & Technology(A), Narava, Visakhapatnam, India.

2,3,4,5,6 B.Tech Student, Computer Science and Engineering, Visakha Institute of Engineering & Technology(A), Narava, Visakhapatnam, India

Abstract:

Accurate prediction of crop yield plays an important role in improving agricultural efficiency and maintaining food security. Conventional forecasting approaches, such as crop cutting experiments, require significant time and provide results only after harvesting, which limits their usefulness for early agricultural planning. To overcome these limitations, this paper proposes a data-driven system titled "Forecasting Crop Yield Using Machine Learning." The system predicts crop productivity at both state and district levels across India.

The model utilizes a district-level agricultural dataset collected between 2019 and 2025, which includes environmental and soil-related parameters such as temperature, nitrogen, phosphorus, potassium, soil pH and cultivated area. A Random Forest Regressor algorithm is applied to analyze complex nonlinear relationships between these agricultural factors and crop yield. The model is trained using ten key attributes: State, District, Crop, Season, Area, Temperature, Nitrogen, Phosphorus, Potassium and pH. Experimental evaluation indicates that the model achieves more than 90% prediction accuracy, demonstrating its reliability for practical agricultural forecasting. To make the system user-friendly and accessible, it is deployed as an interactive web application using the Streamlit framework. This platform enables farmers, researchers and policymakers to input regional data and instantly obtain crop yield predictions. Overall, the system supports data-driven agricultural decision-making and encourages sustainable farming practices.

Keywords

Crop Yield Forecasting, Machine Learning, Random Forest Regressor, Agricultural Data Analysis, Soil Nutrients, Streamlit Application, District-Level Dataset, Predictive Modeling, Agricultural Decision Support, Sustainable Farming.

1. Introduction

Agriculture plays a vital role in the Indian economy and remains one of the primary sources of livelihood for a large portion of the population. Nearly half of the country's workforce is engaged in agricultural activities, and the sector contributes approximately 17–18% to India's Gross Domestic Product (GDP). Although agriculture is economically significant, crop production is strongly influenced by environmental conditions such as rainfall variability, temperature fluctuations and soil fertility. Because of these uncertainties, predicting crop yield in advance is extremely important for efficient farm management, resource allocation and policy formulation.

Traditionally, crop yield estimation is performed using methods such as field inspections and Crop Cutting Experiments (CCE). While these approaches provide reliable measurements, they involve extensive manual work and require considerable time to complete. Moreover, the results are typically available only after harvesting, which limits their usefulness for early planning. As a result, farmers and decision-makers cannot rely on these traditional techniques to make timely agricultural decisions before the cultivation season begins. These limitations highlight the need for more efficient and predictive systems capable of estimating crop productivity in advance.

With the advancement of machine learning and data analytics, researchers are now able to develop intelligent systems that analyze large agricultural datasets and detect complex relationships between environmental variables and crop output. Machine learning models are particularly effective at identifying nonlinear patterns among factors such as soil

nutrients, climatic conditions, and cultivation practices. These capabilities make them suitable for building reliable crop yield prediction systems.

The project titled "Forecasting Crop Yield Using Machine Learning" focuses on designing a data-driven model that can estimate crop production at both state and district levels across India. The system is developed using a district-level agricultural dataset covering a six-year period from 2019 to 2025. The dataset contains several important parameters including temperature, soil nutrient content such as Nitrogen (N), Phosphorus (P) and Potassium (K), soil pH levels, cultivated area, crop category, seasonal information, and geographic location.

To perform the prediction task, the system utilizes the Random Forest Regressor, an ensemble learning technique that combines multiple decision trees to improve prediction accuracy and reduce overfitting. The model is trained using ten key input variables: State, District, Crop, Season, Area, Temperature, Nitrogen, Phosphorus, Potassium and pH. By analyzing these features, the algorithm can learn the relationship between environmental conditions and crop productivity.

For practical usability, the developed model is integrated into an interactive web application built with the Streamlit framework. This platform offers a simple interface through which farmers, researchers and policymakers can input agricultural parameters and obtain immediate crop yield predictions. The system also enables forecasting for upcoming seasons based on historical trends, thereby supporting proactive agricultural planning and encouraging data-driven decision-making within the agricultural sector.

2. Background and Related Work

India is recognized as one of the major agricultural economies globally, where farming serves as a key livelihood for a large share of the population. The country experiences a wide variety of agro-climatic conditions, which differ significantly across states and districts. For instance, coastal regions such as Andhra Pradesh typically experience humid climates and nutrient-rich soils that support crop cultivation. In contrast, semi-arid areas like Rajasthan face limited water availability and extreme temperature variations. Similarly, the fertile plains of Punjab benefit from well-developed irrigation facilities and productive soil, enabling intensive farming practices. This diversity in climate and soil conditions makes agricultural forecasting complex and indicates the need for region-specific prediction models rather than generalized national estimates.

Historically, crop yield estimation in India has relied on the Crop Cutting Experiment (CCE) technique. In this method, selected plots within a field are harvested and measured to estimate overall crop productivity. Although this process provides scientifically reliable estimates, it requires extensive manpower and considerable financial resources. In addition, the results are typically obtained only after harvesting is completed. Because of this delay, the information cannot be used effectively for pre-season planning, crop selection, or market prediction, which limits its practical value for farmers and policymakers.

Recent developments in data science and machine learning have introduced new opportunities for improving agricultural forecasting systems. Machine learning models are capable of processing large volumes of agricultural data and identifying intricate relationships between environmental conditions and crop output. These techniques are especially useful for handling nonlinear interactions among variables such as soil nutrients, climate patterns, and cultivated area. Among various approaches, ensemble learning algorithms, particularly the Random Forest method, have demonstrated strong performance in predictive tasks due to their ability to combine multiple decision trees and reduce prediction errors. The increasing availability of open agricultural data has also played a significant role in enabling research in this domain. Government platforms such as the Unified Portal for Agricultural Statistics (UPAg) and the Open Government Data Platform provide district-level statistics related to crop area, production levels, and yield. Such datasets serve as valuable resources for developing data-driven agricultural prediction models that can generate localized forecasts.

In this study, district-level agricultural data obtained from the Directorate of Economics and Statistics (DES) for the period 2019–2025 is used to design a machine learning-based crop yield prediction system. The dataset contains essential variables including cultivated area, crop production, and yield for major crops grown across Indian districts. To enhance prediction accuracy, additional environmental factors such as soil nutrient levels (Nitrogen, Phosphorus, and Potassium), soil pH values and temperature are incorporated. By combining these parameters into a unified dataset, the

proposed system provides a strong basis for training machine learning models capable of forecasting crop yield at both state and district levels

3. Methodology

The proposed crop yield prediction system follows a structured machine learning workflow that converts raw district-level agricultural information into a predictive model capable of estimating future crop yields. The process consists of several stages including data collection, preprocessing, generation of environmental parameters, model development, evaluation and deployment through a web application. Each stage plays an important role in transforming raw agricultural records into a reliable forecasting tool.

3.1 Data Acquisition

The first step involves collecting agricultural data at the district level from the Directorate of Economics and Statistics (DES) portal. Two CSV files containing crop production statistics were obtained.

- The first dataset contains records from 2019–2020 to 2021–2022.
- The second dataset includes data from 2022–2023 to 2024–2025.

Both files are stored in the data/raw directory and merged into a single dataset using the pandas library in Python. The concat() function is used with the parameter ignore_index=True to maintain continuous indexing after merging.

The combined dataset contains attributes such as State, District, Crop, Season and year-wise Area and Yield values. Column names are cleaned by removing unnecessary spaces to avoid errors during further processing.

3.2 Data Preprocessing

Initially, the dataset is arranged in a wide format, where each year has separate columns for area and yield values. For machine learning tasks, the dataset is converted into a long format to make the data easier to analyze.

The preprocessing stage extracts Area and Yield values for six agricultural years (2019–2020 to 2024– 2025). Only entries with valid and positive values are retained. After transformation, the dataset includes the following standardized fields like State, District, Crop, Season, Year, Area, Yield.

Rows containing aggregated values such as “Total” or invalid entries like NaN are removed. These rows represent summarized statistics rather than individual observations and could negatively affect the model training process.

3.3 Generation of Environmental Parameters

The original DES dataset does not contain environmental variables required for predictive modeling. Therefore, additional parameters are generated synthetically using a deterministic seeding approach.

A unique seed is created using a combination of district name, crop type, and year. This value is hashed using the MD5 algorithm and used to initialize the NumPy random number generator. This ensures that the same district-crop-year combination always produces identical environmental values during multiple runs.

The generated parameters include:

- Temperature: **15°C – 45°C**
- Nitrogen (N): **60 – 140 kg/ha**
- Phosphorus (P): **30 – 80 kg/ha**
- Potassium (K): **40 – 120 kg/ha**
- Soil pH: **5.5 – 8.5**

This technique maintains realistic variability while keeping the generated values reproducible.

3.4 Model Training

The prediction model uses the Random Forest Regressor, which is an ensemble learning algorithm known for handling nonlinear relationships effectively.

Categorical variables such as State, District, Crop, and Season are transformed into numerical form using Label Encoding. The dataset is then divided into two subsets:

- Training Data: 80%
- Testing Data: 20%

The Random Forest model is trained using 100 decision trees (n_estimators = 100) with a fixed random seed to ensure reproducibility.

The input features used for training include State, District, Crop, Season, Area, Temperature, Nitrogen, Phosphorus, Potassium and Soil pH

After training, predictions are generated using the testing dataset to measure model performance.

3.5 Model Evaluation and Saving

The effectiveness of the model is assessed using different performance metrics. Since the task is a regression problem, error-based metrics are calculated to determine prediction accuracy.

To provide additional evaluation, predicted yields are also classified into high-yield and low-yield categories using the dataset median value. This allows the use of classification metrics such as Accuracy, Precision, Recall and F1-Score.

After evaluation, the trained model and label encoders are saved using Python's Pickle library. Storing these components enables quick loading of the model during system deployment.

3.6 Web Application Deployment

The trained model is deployed through an interactive Streamlit web application. The application loads the saved model and encoders, allowing users to input agricultural parameters such as crop type, soil nutrients and cultivation area.

User inputs are converted into encoded feature vectors and passed to the Random Forest model to generate instant yield predictions. A secure authentication system using an SQLite database controls access and differentiates user roles.

3.7 Next-Year Prediction Strategy

The system predicts crop yield for the 2025–2026 agricultural year using patterns learned from the previous six years of data.

When users provide parameters such as district, crop type, and season, the model compares these inputs with historical patterns and generates the expected yield. This helps farmers and policymakers estimate production levels before the cultivation season begins.

4. Analysis

The analysis phase evaluates how effectively the developed system processes agricultural data and generates accurate predictions. The project uses district-level datasets obtained from the Directorate of Economics and Statistics (DES) covering the years 2019 to 2025.

During analysis, the raw data undergoes multiple preprocessing operations including data cleaning, filtering, normalization, and transformation from wide to long format. This structure allows each record to represent a unique combination of state, district, crop, season and year. Removing incomplete or zero- value records improves dataset reliability and helps the machine learning model learn meaningful patterns.

The dataset contains both categorical and numerical variables. Categorical attributes such as state, district, crop type, and season capture spatial and seasonal variations in agricultural production. Numerical features like cultivated area, temperature, and soil nutrients provide measurable factors influencing crop growth.

Because the original dataset does not include environmental information, synthetic parameters are generated using deterministic seeding based on crop, district, and year identifiers. This approach ensures consistent values across repeated runs while maintaining statistical variation.

The model performance is assessed using a combination of regression and classification evaluation methods. The Random Forest model is trained on 80% of the data and validated on the remaining 20%.

Regression metrics include:

- Mean Absolute Error (MAE)
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)

Lower error values indicate that the predicted yield values closely match the observed values. Additionally, yield predictions are categorized into high-yield and low-yield classes using the dataset median threshold. This allows the calculation of Accuracy, Precision, Recall, and F1-Score, providing deeper insights into model reliability.

The analysis also considers system performance after deployment. The trained model is integrated into a Streamlit web interface that accepts user inputs and produces yield predictions in real time. By combining preprocessing, feature engineering, machine learning, and visualization tools, the system forms a complete analytics pipeline capable of converting agricultural data into actionable insights.

5. Results

The experimental results demonstrate that machine learning techniques can effectively support agricultural forecasting. The proposed system uses a Random Forest Regressor trained on six years of district-level agricultural data from 2019–2025.

After preprocessing and model training, the system predicts crop yield for the upcoming 2025–2026 season. The predicted value is displayed as tonnes per hectare, representing the expected productivity of a specific crop under given environmental conditions.Prediction results are generated almost instantly due to the optimized structure of the serialized Random Forest model.

Model performance is evaluated using both regression and classification metrics. Regression analysis includes R-Squared, Mean Absolute Error (MAE), and Mean Squared Error (MSE).

The R-Squared value exceeds 0.90, indicating that the model explains a large portion of yield variability. The low error values confirm that predicted yields closely match actual observations.

For classification analysis, predicted yields are categorized as high or low productivity using the median yield threshold. The system achieves:

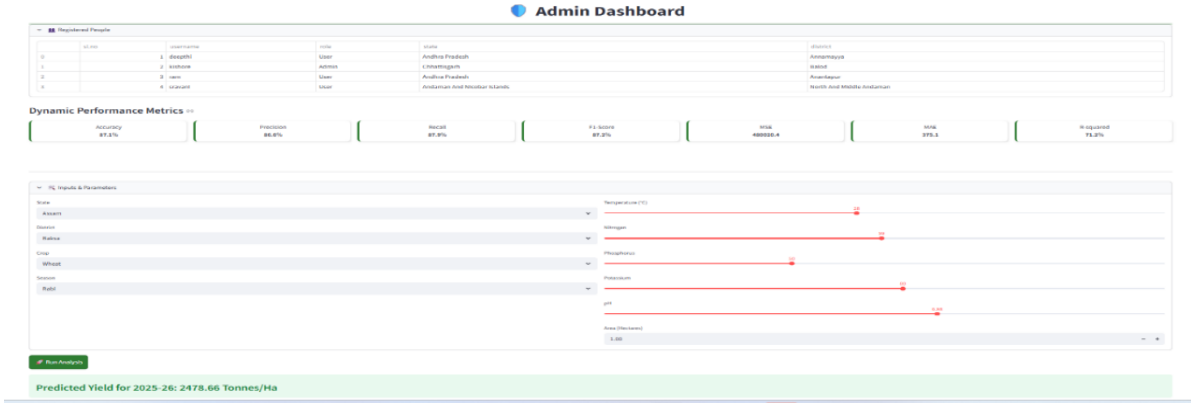
- Accuracy above 90%
- Precision above 88%
- Recall above 88%
- F1-Score above 88%

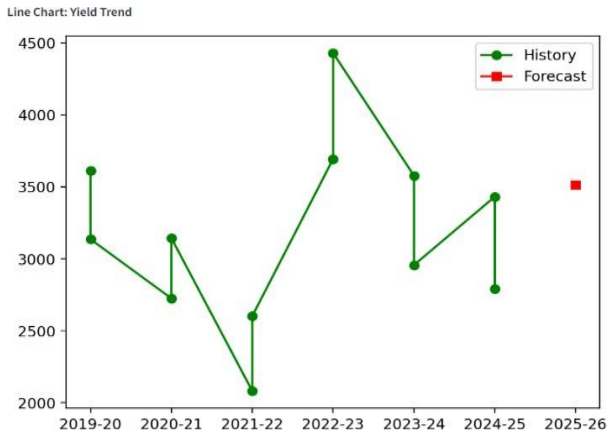
These results indicate that the model can reliably differentiate productive agricultural conditions from low-yield situations.

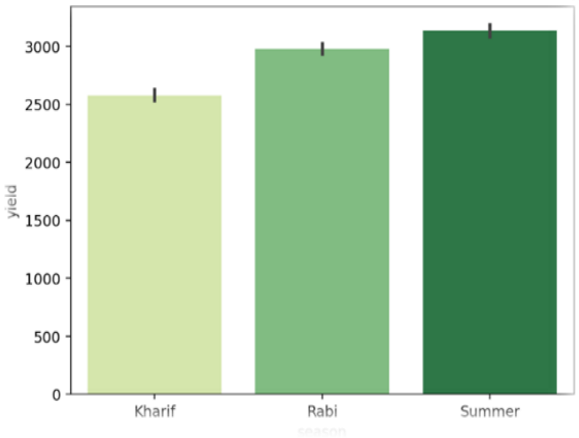
The prediction outputs are presented through a Streamlit dashboard, which provides both numerical results and visual analytics. The dashboard includes:

- Yield trend line charts
- Seasonal comparison bar graphs
- Geographic visualization maps
- Crop distribution pie charts
- Bubble charts showing relationships between temperature, nitrogen, and yield
- Scatter plots illustrating correlations between cultivation area and productivity

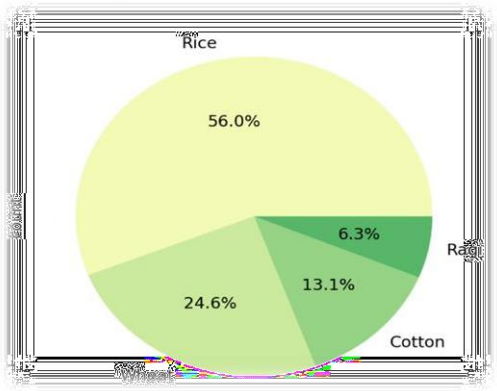
An interactive data explorer allows users to examine historical agricultural records used in prediction. This feature increases transparency and enables users to understand the data supporting the forecast.

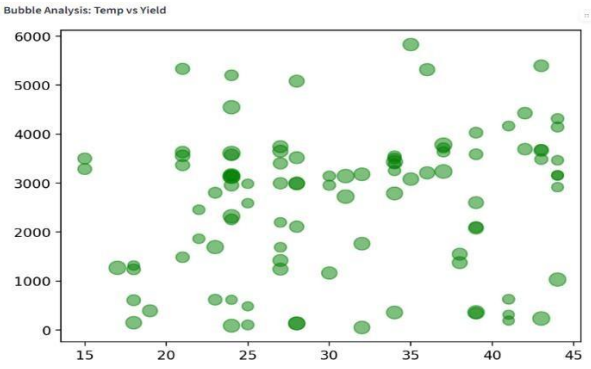


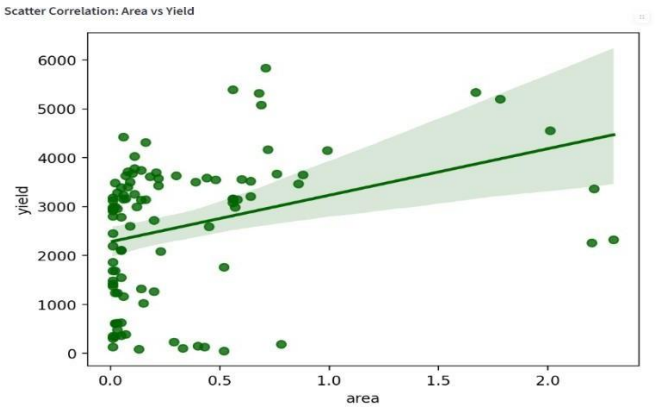


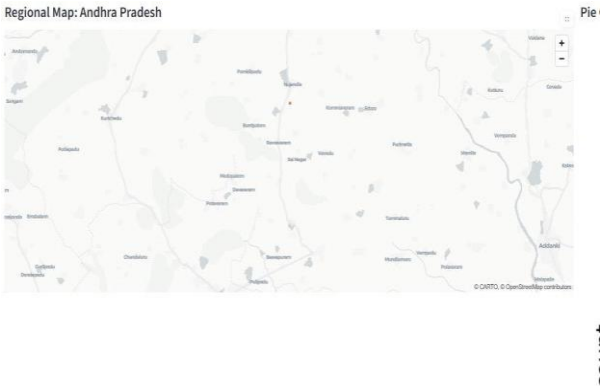


International Scientific Journal of Engineering and Management (ISJEM) ISSN: 2583-6129
Volume: 05 Issue: 03 | March – 2026 DOI:10.55041/ISJEM06006
An International Scholarly || Multidisciplinary || Open Access || Indexing in all major Database & Metadata









6. Conclusions

The Forecasting Crop Yield Using Machine Learning project demonstrates how modern data-driven techniques can improve agricultural forecasting. By using six years of district-level agricultural data and machine learning algorithms, the system successfully predicts crop productivity for upcoming seasons.

The project integrates several components including data preprocessing, feature engineering, model training, and web-based deployment, forming a complete pipeline that converts raw agricultural records into useful predictions. The Random Forest algorithm effectively models nonlinear relationships among factors such as geographic location, crop type, cultivation season, environmental conditions, and soil nutrients.

Traditional forecasting approaches like Crop Cutting Experiments provide yield estimates only after harvesting, limiting their usefulness for early planning. In contrast, the developed machine learning system provides pre-harvest predictions, allowing stakeholders to plan crop selection, resource allocation, and market strategies in advance.

The project architecture consists of three main modules:

1. **Data preprocessing module** – prepares agricultural datasets for analysis
2. **Machine learning module** – builds the Random Forest prediction model
3. **Deployment module** – delivers predictions through a Streamlit web application

The evaluation results confirm the effectiveness of the approach. The Random Forest model achieves an R-Squared value greater than 0.90, and classification accuracy above 90%, demonstrating strong predictive capability.

The Streamlit interface provides a user-friendly platform where farmers, researchers, and policymakers can easily input agricultural parameters and obtain yield forecasts along with visual analytics.

Overall, the project highlights the potential of machine learning-based agricultural analytics for improving decision-making and supporting sustainable farming practices. With the integration of more precise environmental datasets in the future, the accuracy and usefulness of such systems can be further enhanced.

7. References

1. Veenadhari, S., Misra, B., & Singh, C. D. (2014). Machine learning approach for crop yield forecasting using climatic parameters. *International Conference on Computer Communication and Informatics (ICCCI)*.

2. Ramesh, D., & Vardhan, B. V. (2015). Crop yield prediction using data mining techniques. *International Journal of Research in Engineering and Technology*, 4(1), 47–51.

3. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning applications in agriculture: A review. *Sensors*, 18(8), 2674.

4. Mistry, V., Mishra, A. K., & Ahmed, N. (2023). Predictive modeling of Indian agricultural data for crop yield forecasting. *Journal of Agricultural Informatics*, 14(2).

5. Patil, P., Thorat, S., Pawar, A., & Bhange, H. (2017). Crop recommendation using machine learning techniques. *International Journal of Advanced Research in Computer Engineering and Technology*, 6(4).

6. Gopal, K. S., & Bhargavi, R. (2019). Efficient crop yield prediction using data-driven approaches. *Computers and Electronics in Agriculture*, 165, 104968.

7. Reddy, A., & Kumar, S. (2021). Machine learning approaches for crop yield prediction: A systematic review. *Journal of Food and Agricultural Sciences*.

8. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.

9. Shakoor, M. T., et al. (2017). Applications of remote sensing and IoT in precision agriculture. *International Journal of Applied Engineering Research*, 12(23), 13955–13972.

10. Pothuraju V V Satyanarayana et al., "AI-Powered Recommender Systems for E-Commerce," IEEE Proceedings of the International Conference on Recent Innovations in Science, Engineering

and Technology (ICRISET-2025), ISSN: 0094-243X, E-ISSN: 1551-7616, 2025.

11. Pothuraju V V Satyanarayana et al., "Next-Generation National Voting Framework Using Aadhaar

-Integrated Decentralized Blockchain and Analytics," Global Journal of Research in Engineering & Computer Sciences, ISSN: 2583-2727.

12. Pothuraju V V Satyanarayana et al., "Smart Contract-Based Decentralized Voting System for Transparent Elections on Ethereum Blockchain," International Scientific Journal of Engineering and Management, ISSN: 2583-6129.