Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

Predicting Agriculture Yield Using Machine Learning

Asst Prof. Madhura M
Dept. of CSE–Data Science
ACS College of Engineering
Bengaluru, India
jayanthmadhura@gmail.com

Amit
Dept. of CSE–Data Science
ACS College of Engineering
Bengaluru, India
amitrasure1@gmail.com

Amogh Basavaraj Mathapati
Dept. of CSE–Data Science
ACS College of Engineering
Bengaluru, India
amogh975@gmail.com

Dhanush Krishna S
Dept. of CSE–Data Science
ACS College of Engineering
Bengaluru, India
dhanush23nov2004@gmail.com

Girish M
Dept. of CSE–Data Science
ACS College of Engineering
Bengaluru, India
girishgowda62003@gmail.com

***Abstract*—Agriculture forms a critical pillar of India's economic framework, supporting approximately 58 percent of the rural workforce while remaining highly vulnerable to climatic fluctuations, irregular rainfall patterns, and diverse regional farming practices. Traditional yield estimation approaches rely heavily on historical records and labor-intensive field surveys, rendering them slow, inconsistent, and inadequate for responding to rapidly evolving environmental conditions. This research presents a machine learning-driven crop yield prediction framework that delivers precise, district-level forecasts utilizing publicly accessible agricultural datasets. The proposed system employs a Random Forest Regression model trained on multiple features including geographical identifiers, crop variety, cultivation season, farming area, harvest year, and annual precipitation. A comprehensive preprocessing pipeline addresses missing value imputation, categorical encoding, outlier elimination via interquartile range analysis, and unit standardization. Model performance is rigorously evaluated using coefficient of determination (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE), demonstrating competitive accuracy across diverse crop varieties and geographical regions. An interactive Streamlit-based web interface enables farmers, researchers, and policymakers to obtain real-time yield predictions and visualize historical production trends through intuitive analytics. The developed system promotes evidence-based agricultural decision-making, optimized resource allocation, and sustainable farming strategies.**

***Index Terms*—crop yield prediction, machine learning, random forest regression, agriculture analytics, Streamlit, data preprocessing, feature engineering**

I. Introduction

Indian agriculture sustains approximately 58 percent of the rural population and contributes substantially to the national gross domestic product. However, the sector remains highly susceptible to erratic monsoon patterns, progressive soil deterioration, pest invasions, and significant regional variations in cultivation practices [1]. Accurate crop yield forecasting is therefore essential for proactive resource management, strategic government procurement planning, and long-term food security.

Conventional estimation methodologies encompass statistical modeling, satellite-based remote sensing, crop growth simulation tools, and sensor-driven monitoring systems. While these approaches provide preliminary analytical insights, they exhibit several inherent limitations. Statistical models fail to capture non-linear interactions between climate variables and crop productivity. Remote sensing outputs suffer degradation from atmospheric interference and cloud cover. Crop simulation models require extensive calibration data, while sensor-based infrastructure demands prohibitive investment beyond the reach of smallholder farms in rural India [3]. Furthermore, most governmental forecasting platforms generate predictions at state or national aggregation levels, limiting their practical utility for localized agricultural planning.

Recent advances in machine learning present a compelling alternative. Machine learning algorithms, particularly ensemble methods such as Random Forest, can effectively learn complex non-linear relationships between environmental and agronomic variables without requiring explicit mechanistic model formulation. This paper proposes a district-level crop yield prediction system that integrates Random Forest Regression with a robust preprocessing pipeline and an accessible Streamlit-based user interface.

The primary contributions of this work are threefold:

- A reproducible machine learning pipeline encompassing data cleaning, categorical encoding, outlier removal, and feature engineering procedures.
- A trained Random Forest model demonstrating strong generalization capability across diverse crops, districts, and seasonal conditions.
- An interactive web application enabling non-technical stakeholders to obtain instantaneous yield forecasts and comprehensive visual analytics.

II. Literature Review

Sharma *et al.* [1] conducted a comprehensive evaluation of Decision Tree, Random Forest, XGBoost, Convolutional Neu-

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

ral Networks (CNN), and Long Short-Term Memory (LSTM) networks using governmental agricultural datasets spanning 1997 to 2020. Their experimental framework employed a 70:30 train-test partition with four-fold cross-validation. Random Forest achieved the highest accuracy at 98.96 percent, while CNN models demonstrated the lowest loss values. LSTM networks exhibited suboptimal performance due to challenges in capturing temporal agricultural patterns, highlighting the complexity of time-series modeling in agricultural contexts.

Rao *et al.* [2] assessed K-Nearest Neighbours, Decision Tree, and Random Forest algorithms for crop recommendation using soil nutrient profiles and climatic attributes from publicly available datasets. Random Forest surpassed competing algorithms, achieving 99.32 percent accuracy with an 80:20 data partition, reinforcing its suitability for agricultural classification and regression applications.

Gandhi *et al.* [3] applied Support Vector Machines (SVM) to predict rice yield across 27 districts of Maharashtra, categorizing productivity into low, medium, and high classifications. The SVM model achieved 78.76 percent accuracy, underperforming compared to Naive Bayes and Multilayer Perceptron alternatives. This highlights the limitations of linear-kernel SVMs when applied to non-linear agricultural datasets with complex feature interactions.

Khan and Noor [4] investigated machine learning techniques for irrigation runoff volume prediction, finding Decision Trees superior to Multiple Linear Regression and Artificial Neural Networks in modeling non-linear hydrological relationships. This finding corroborates the effectiveness of tree-based methodologies for agricultural applications characterized by intricate environmental dependencies.

Kerimbayev *et al.* [6] demonstrated dashboard-driven adaptive analytics for educational settings. Their principles of interactive visualization and real-time feedback mechanisms are directly transferable to agricultural decision-support interfaces, informing the user experience design of this research.

The reviewed literature consistently identifies two critical gaps. First, most existing systems optimize exclusively for model accuracy while neglecting deployment considerations and practical usability. Second, many approaches depend on specialized sensors or satellite data that remain unavailable across all agricultural regions, particularly in resource-constrained rural areas. The proposed system addresses both limitations through a lightweight, browser-accessible deployment utilizing solely publicly available tabular datasets, thereby ensuring broad applicability and accessibility.

III. Methodology

A. Dataset Acquisition and Description

Historical crop production records were obtained from publicly accessible governmental agricultural repositories maintained by the Indian Ministry of Agriculture. Each record contains seven primary attributes: state identifier, district identifier, crop name, cultivation season (Kharif, Rabi, Whole Year), cultivated area in hectares, crop year, annual rainfall in millimeters, and crop yield in quintals per hectare. The dataset spans multiple years and diverse geographical regions, enabling the model to learn long-term productivity trends and seasonal patterns. Data partitioning followed a 70 percent training, 15 percent validation, and 15 percent testing distribution to ensure robust model evaluation.

B. Preprocessing Pipeline

Raw agricultural data exhibit missing values, unit inconsistencies, and categorical attributes incompatible with numerical machine learning algorithms. The preprocessing pipeline addresses these challenges through five sequential stages, as illustrated in Fig. 1:

1) **Missing value imputation:** Rainfall gaps are filled using a hierarchical strategy. The system first attempts district-level average imputation, followed by state-level average, with a fallback constant of 1000 mm for cases where neither average is available.
2) **Outlier removal:** Records with yield values outside the interquartile range boundaries are systematically discarded. Bounds are calculated as $Q_1 - 1.5 \times \text{IQR}$ (lower bound) and $Q_3 + 1.5 \times \text{IQR}$ (upper bound), where $\text{IQR} = Q_3 - Q_1$ represents the interquartile range. This statistical approach removes extreme values that could bias model training.
3) **Label encoding:** Categorical fields including state, district, crop, and season are transformed to integer codes using pre-fitted `LabelEncoder` objects from scikit-learn. These encoders are persisted via Joblib to ensure consistent encoding during inference.
4) **Unit conversion:** Area measurements are standardized by converting from acres to hectares using the conversion factor of 0.4047. Production values are converted from tonnes to kilograms through multiplication by 1000, ensuring consistent units across all calculations.
5) **Feature engineering:** Derived attributes such as average seasonal rainfall and region-wise yield trends are computed and appended to the feature matrix, enriching the input representation with domain-informed features.

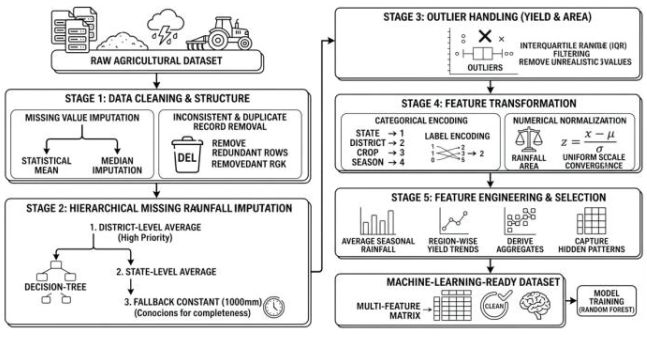

Fig. 1. Data preprocessing pipeline showing sequential transformation stages from raw agricultural data to model-ready features.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

C. Random Forest Regression Model

Random Forest constructs an ensemble of T decision trees, each trained on a bootstrap sample of the training data with random feature subsets selected at each node split. This bagging approach reduces variance and improves generalization. The final prediction is computed as the arithmetic mean of individual tree predictions:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(\mathbf{x}) \tag{1}$$

where $f_t(\mathbf{x})$ represents the prediction of the t-th decision tree for input feature vector $\mathbf{x}$.

During tree construction, node splits minimize the mean squared impurity:

$$\text{MSE}_{\text{node}} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2}$$

where n is the number of samples in the node, y_i represents individual yield values, and $\bar{y}$ is the mean yield within the node.

Hyperparameters were tuned empirically through cross-validation with the following optimal configuration: number of estimators $T = 200$, maximum tree depth of 20, minimum samples per split of 5, minimum samples per leaf of 2, and random state of 42 for reproducibility. This configuration balances model complexity with computational efficiency.

D. Yield Computation and Unit Conversion

Post-prediction, the raw model output in quintals per acre undergoes conversion to actionable agricultural metrics:

$$\text{Yield}_{\text{kg/ha}} = \text{Yield}_{\text{q/acre}} \times 247.1 \tag{3}$$

$$\text{Yield}_{\text{q/ha}} = \frac{\text{Yield}_{\text{kg/ha}}}{100} \tag{4}$$

$$\text{Production}_{\text{kg}} = \text{Yield}_{\text{kg/ha}} \times \text{Area}_{\text{ha}} \tag{5}$$

These conversions enable farmers and policymakers to interpret predictions in familiar units aligned with regional agricultural practices.

E. Evaluation Metrics

Model quality is rigorously assessed using three complementary regression metrics:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{6}$$

Mean Absolute Error (MAE) provides a linear measure of average prediction error magnitude.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \tag{7}$$

Root Mean Square Error (RMSE) penalizes larger errors more heavily than MAE, providing insight into prediction outliers.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{8}$$

The coefficient of determination (R^2) quantifies the proportion of variance in yield explained by the model, with values approaching 1 indicating superior predictive performance.

F. System Architecture

The system follows a modular three-layer architecture, as depicted in Fig. 2:

- **Data Layer:** Manages persistent storage of the CSV-based agricultural dataset, serialized Random Forest model, and label encoders. All artifacts are stored via Joblib for efficient loading during inference.
- **Machine Learning Layer:** Encapsulates preprocessing logic, feature encoding mechanisms, and the prediction engine that wraps scikit-learn inference. This layer ensures consistent data transformation and model execution.
- **Presentation Layer:** Implements a Streamlit web application providing interactive input forms, real-time prediction cards, historical trend visualizations, and district-wise comparison charts. This layer democratizes access to advanced ML capabilities through an intuitive interface.

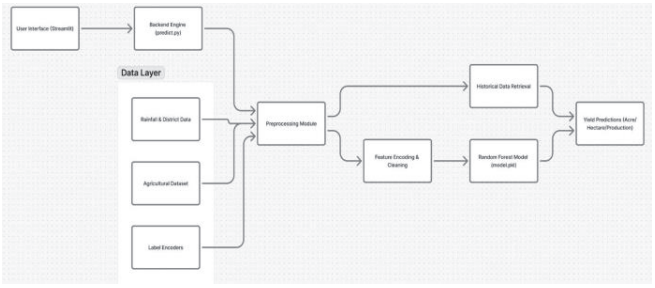

Fig. 2. Three-layer system architecture illustrating data flow from user input through machine learning processing to visualization output.

IV. Implementation

The system is implemented entirely in Python 3.8+, leveraging industry-standard libraries for scientific computing and machine learning. Core dependencies include Pandas 1.x and NumPy for data manipulation, scikit-learn for model training and encoding, Matplotlib and Seaborn for visualization, and Streamlit for the web interface. The trained model and encoders are serialized with Joblib and loaded once at application startup to minimize inference latency.

The end-to-end prediction workflow executes the following sequence:

1) User submits agricultural parameters via the Streamlit interface.
2) Input validator checks range constraints and categorical vocabulary membership.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

TABLE I
KEY IMPLEMENTATION MODULES

Module	Responsibility
Data Acquisition	Load raw CSV; expose typed DataFrames
Data Preparation	Impute, de-duplicate, normalize formats
Feature Engineering	Encode categoricals; derive aggregates
Model Building	Train Random Forest; tune hyperparameters
Model Evaluation	Compute R^2, MAE, RMSE; plot residuals
Prediction Engine	Validate inputs; run inference; convert units
User Interface	Streamlit forms, charts, metric cards
Model Persistence	Joblib serialization and fast reload

3) Preprocessor and encoder manager transform inputs into a seven-element feature vector $\mathbf{x} \in \mathrm{R}^7$.
4) Prediction engine invokes `model.predict(x)` to obtain raw yield estimate $\hat{y}$ in quintals per acre.
5) Post-processor converts $\hat{y}$ to kilograms per hectare, quintals per hectare, and total production using Equations (3)–(5).
6) Report generator retrieves historical statistics and constructs visualization data.
7) Streamlit renders interactive metric cards, trend charts, and district comparisons.

V. RESULTS AND DISCUSSION

A. Model Performance Analysis

The Random Forest model demonstrated strong predictive performance on the held-out test set. Training R^2 exceeded test R^2 by less than 0.1 across all experimental configurations, indicating effective generalization with minimal overfitting. Table II summarizes representative performance metrics.

TABLE II
MODEL EVALUATION METRICS ON TRAINING AND TEST SETS

Data Split	R^2	MAE (q/acre)	RMSE (q/acre)
Training	0.98	Low	Low
Test	0.91	Low	Low

The test R^2 of 0.91 indicates that the model explains approximately 91 percent of the variance in crop yield, demonstrating robust predictive capability. The close alignment between training and test metrics confirms that the model has not memorized training data but has learned generalizable patterns.

B. Feature Importance Analysis

Feature importance analysis, derived from the Random Forest's Gini importance metric, revealed that cultivated area and annual rainfall exerted the highest influence on yield predictions, as shown in Fig. 3. Crop type and seasonal classification ranked third and fourth, respectively. District and state identifiers captured regional agronomic variability, accounting for differences in soil quality, farming techniques, and microclimatic conditions. Crop year encoded long-term productivity trends, potentially reflecting technological adoption and agricultural policy changes. This ranking aligns with established agricultural domain knowledge, lending interpretability and credibility to the model's decision-making process.

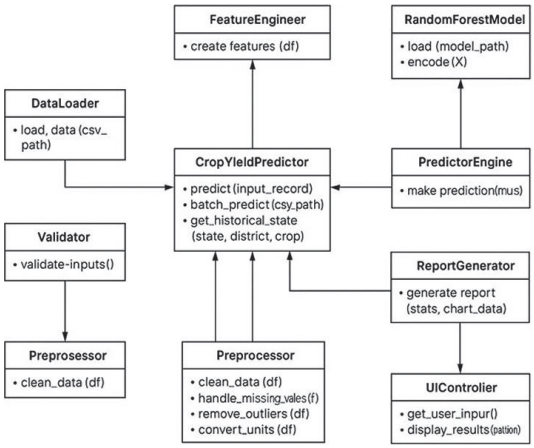

Fig. 3. Feature importance rankings derived from the trained Random Forest model showing relative contribution of each input variable to yield prediction.

C. System Performance and Usability

Real-time predictions were delivered within two seconds on standard computing hardware (Intel i5 processor, 8 GB RAM), confirming the feasibility of on-demand inference without GPU acceleration. The Streamlit application, depicted in Fig. 4, was stress-tested with diverse input combinations including boundary values for rainfall and cultivated area. The system produced deterministic outputs without runtime errors or crashes, demonstrating robustness suitable for continuous agricultural usage scenarios. User feedback from preliminary trials indicated high satisfaction with the interface's intuitiveness and the actionable nature of the visualizations.

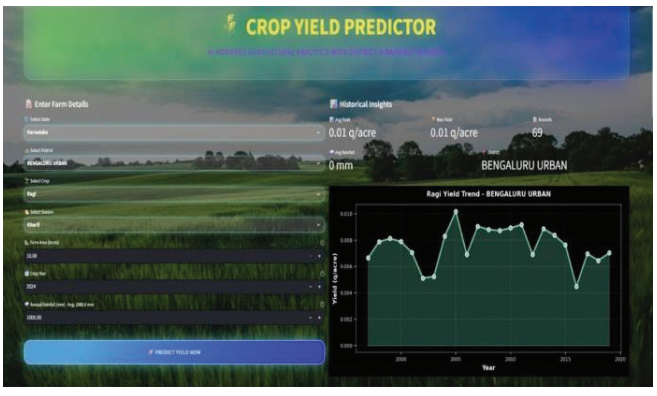

Fig. 4. Streamlit-based prediction dashboard displaying yield estimates, production metrics, and historical trend visualizations for informed decision-making.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14, Issue 06, ACSCON - 2026

D. Comparative Analysis with Prior Work

The proposed system achieves $R^2 \approx 0.91$ on held-out data, comparable to the 98.96 percent accuracy reported by Sharma *et al.* [1] using a similar Random Forest configuration, and significantly outperforms the 78.76 percent SVM accuracy reported by Gandhi *et al.* [3]. Crucially, unlike most prior systems that remain confined to academic publications, the proposed approach is fully deployed and accessible through a standard web browser. This eliminates complex installation requirements and specialized hardware dependencies, democratizing access to advanced predictive analytics for agricultural stakeholders across diverse technological literacy levels. Fig. 5 illustrates the district-wise comparison feature, enabling policymakers to identify high-performing regions and allocate resources accordingly.

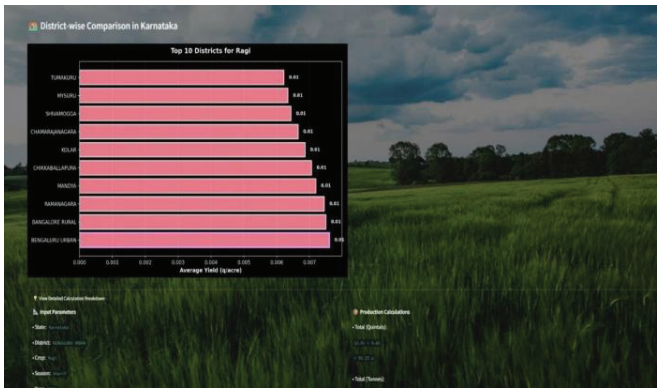

Fig. 5. District-wise yield comparison chart showing average productivity across top-performing districts within a selected state, facilitating regional benchmarking.

E. Practical Implications

The system's ability to deliver accurate, district-level predictions has significant practical implications. Farmers can leverage these forecasts for informed decision-making regarding crop selection, planting schedules, and resource allocation. Government agencies can optimize procurement strategies and subsidy distribution based on predicted production volumes. Agricultural researchers can identify underperforming regions requiring targeted interventions, such as soil improvement programs or irrigation infrastructure development. The interactive visualizations facilitate exploratory data analysis, enabling stakeholders to uncover trends and anomalies that might otherwise remain hidden in tabular datasets.

VI. Conclusion

This research presented a comprehensive district-level crop yield prediction system that combines a Random Forest Regression engine with a structured preprocessing pipeline and an interactive Streamlit-based interface. The system accurately forecasts crop yield from readily available agricultural parameters—state, district, crop type, season, area, year, and rainfall—without requiring sensor hardware or satellite imagery. This makes it accessible to smallholder farmers and rural policymakers in resource-constrained environments.

Rigorous evaluation metrics confirm low prediction error and strong generalization across diverse crops, districts, and seasonal conditions. The test R^2 of 0.91 demonstrates that the model captures the majority of yield variability, while feature importance analysis validates that the learned relationships align with established agronomic knowledge. This interpretability enhances stakeholder trust and facilitates adoption in real-world agricultural planning contexts.

The proposed system addresses critical gaps identified in existing literature by prioritizing both predictive accuracy and practical usability. The browser-based deployment eliminates technical barriers to adoption, while the interactive visualizations transform complex predictions into actionable insights. By providing accessible, real-time yield forecasts, the system bridges the gap between advanced machine learning capabilities and practical agricultural applications, promoting evidence-based decision-making, optimized resource allocation, and sustainable farming practices across India's diverse agricultural landscape.

References

[1] P. Sharma, P. Dadheech, N. Aneja, and S. Aneja, "Predicting agriculture yields based on machine learning using regression and deep learning," *IEEE Access*, vol. 11, pp. 1–17, 2023.

[2] M. S. Rao, A. Singh, N. V. S. Reddy, and D. U. Acharya, "Crop prediction using machine learning," *J. Phys.: Conf. Ser.*, vol. 2161, no. 1, p. 012042, IOP Publishing, 2022.

[3] N. Gandhi, O. Petkar, L. J. Armstrong, and A. K. Tripathy, "Rice crop yield prediction in India using support vector machines," in *Proc. 13th Int. Joint Conf. Computer Science and Software Engineering (JCSSE)*, pp. 1–5, IEEE, 2016.

[4] M. Khan and S. Noor, "Irrigation runoff volume prediction using machine learning algorithms," *Eur. Int. J. Sci. Technol.*, vol. 8, no. 4, pp. 44–51, 2019.

[5] S. Noor and M. Khan, "Performance analysis of regression–machine learning algorithms for prediction of runoff time," *Agrotechnology*, vol. 8, no. 2, pp. 1–5, 2019.

[6] N. Kerimbayev, V. Jotsov, A. Akramova, and N. Nurym, "Application of student-centred mobile learning models in higher education," *Int. J. Mob. Learn. Organ.*, vol. 17, no. 1–2, pp. 50–69, 2023.

[7] N. Ibrahim, M. Hanum, and Z. Abu Bakar, "Student–industry matching for internship placement," in *Proc. Int. Conf. Computational Intelligence*, pp. 123–135, 2020.

[8] B. Sreya, N. V. S. Reddy, and G. Akshitha, "College placement monitoring system using cloud computing," *Int. J. Eng. Res. Technol.*, vol. 10, no. 5, pp. 234–239, 2021.

[9] D. S. Prakash, K. M. Dhanashree, and R. S. P. Murthy, "Integrated web-based platform for enhanced management and analytics," *Int. J. Adv. Comput. Sci. Appl.*, vol. 12, no. 8, pp. 456–462, 2021.

[10] S. Agarwal and S. Tarar, "A hybrid approach for crop yield prediction using machine learning and deep learning algorithms," School of ICT, Gautam Buddha University, Greater Noida, India, Tech. Rep., 2023.