Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 47 No. 02 (2026)

Crop Yield Prediction Based on Indian Agriculture Using Machine Learning

Ayush Aravind [1], Pooja M. N. [2], Suhas P. H. [3], H. N. Gagan [4]

[1, 2, 3, 4] *Department of Computer Science and Engineering, PES University, Bengaluru, India*

Abstract:- This paper develops a machine-learning framework to estimate crop yield per hectare in different Indian districts based on open-agricultural data. The system combines data preprocessing, correlation-based and LASSO feature selection methods, and the performance evaluation of six regression algorithms: Ridge, Decision Tree, Random Forest, XGBoost, LightGBM, and CatBoost. The CatBoost model was able to perform well with an R2 close to 0.93 and a low RMSE, demonstrating accurate yield prediction can be achieved using agro-economic and spatial features without explicit soil and weather data. Furthermore, the model is made available through a FastAPI backend along with a React-based multilingual web interface, which collectively offers live, easy-to-reach, and data-driven insights for farmers and policymakers.

Keywords: *Crop yield prediction, CatBoost, LightGBM, Machine learning, XGBoost.*

1. Introduction

Agriculture is the biggest single sector which supports the Indian economy. A major portion of the population of India is dependent on agriculture. Accurate crop yield prediction can be a key factor to resolve issues related to resource allocation, food security, and implementation of sustainable agriculture. Most traditional forecast methods are based on manual surveys or historical averages but cannot reflect nonlinear relationships between yield and influencing factors like crop type, season, and regional characteristics.

During the past few years, agriculture data have become more available, and machine learning is offering a less biased and more accurate way of calculating yield estimates. The ML methods have the capability of extracting intricate dependencies from historical data to make accurate and local predictions. This project demonstrates a complete end-to-end framework for crop yield prediction based on public datasets comprising district-wise attributes such as crop name, season, area, and production.

The approach taken here compares the performance of several regression algorithms, including Ridge, Decision Tree, Random Forest, *XGBoost, LightGBM,* and *CatBoost* for the same evaluation criteria. In such a scenario, the CatBoost model fared well with an R^2 value of about 0.93, indicating that the model generalizes even without explicitly giving soil or weather variables.

Additionally, the proposed system is implemented using a FastAPI backend with a React frontend that allows farmers and policymakers to have access to the system through multilingual and voice-enabled interfaces.

2. Literature Survey

Most of the work on agricultural yield prediction, therefore, today is directed at increasing the accuracy of the predictions along with giving greater decision support to farmers. Early works such as that by Kalimuthu et al. [1] utilized ensemble-based frameworks which included statistical regression models with methods such as Random Forest and Gradient Boosting. Their approach provided a multi-feature yield prediction by fusing soil, rainfall, and temperature data, showing that the ensemble models provide higher robustness and generalization for large-scale datasets. The authors further indicated that proper preprocessing and normalization of data are crucial in eliminating bias, hence proving that hybrid regression models have good prospects for practical applications in agriculture.

Nishant et al. [2] proposed a LASSO-based hybrid regression model for forecasting agricultural yield by implementing feature selection through ridding the model of redundant or correlated variables that resulted in

improved model explanation and reduced model overfitting by implementing Cross Validation and regularization as part of the Hybrid Regression Model Design. As a foundation of Robust and Interpretable Hybrid Regression Model Design, these concepts were incorporated directly into the feature-selection methods developed in this project.

Reddy and Kumar [3] further explored the use of deep learning in the comparison of CNN, RNN, and DNN models for yield estimation. They have deduced from their work that while the deep learning architecture captured nonlinear spatial-temporal dependencies effectively, they also increased computational complexity and reduced scalability. They found that this combination of neural and regression approaches provides a much better trade-off between accuracy and interpretability efficiency in practical agricultural deployment.

Further work by Nigam et al. [4] employs several ML algorithms, namely Random Forest, XGBoost, K-Nearest Neighbors (KNN), Logistic Regression, and Artificial Neural Networks (ANN) for district- wise yield prediction in India. Century-scale datasets covering temperature, rainfall, area, and seasonal data were used, from which Random Forest attained the minimum mean absolute error while the LSTM model outperformed the traditional RNNs in capturing temporal variation. This research laid the foundations for feature fusion and hybrid learning as a means to enhance the spatial and temporal consistency of agricultural forecasting. A comparison of analyses from these previous works illustrates an evolution of Crop Yield Forecasting from traditional regression models through Ensemble and Deep Learning methods, whereby existing studies improved on their accuracy. However, while many of the previous studies had improved on accuracy, they also exhibited high levels of computational overhead, regional generalizability, and the lack of a Multi-lingual User-Centric Web-based platform. This research builds upon these findings by combining the innovative approaches to ElasticNet-based Feature Selection and Multi- Model Benchmarking from (CatBoost, XGBoost, LightGBM, Random Forest) to provide both Robust Analytical Capabilities and Practical Usability of a Web-based Multi-Lingual platform Specific to the Indian Agricultural landscape.

3. Methodology

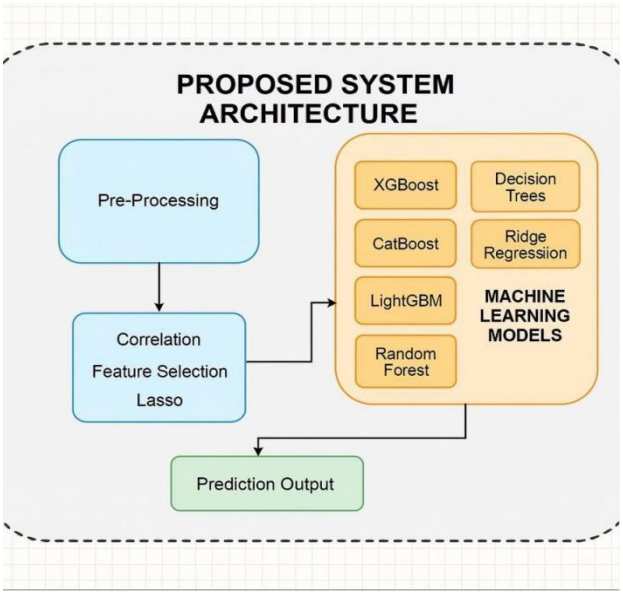

Figure 1. Proposed system architecture showing preprocessing, correlation- based feature selection with LASSO, and benchmarking of multiple machine learning models.

Data Pre-processing

The dataset used in the study was obtained from the All- India Agricultural Statistics Database which provides information of state, district, crop, season, area and production. Initial data preprocessing was done by removing missing/inconsistent records of columns, type casting and removing any invalid

entries. Yield per hectare is calculated as the ratio of production to area. As yield values can span large magnitudes clipping outliers beyond 1st and 99th percentile was done to get uniformly distributed data.

Categorical attributes such as crop/district/season were label encoded while numerical features were normalized to ensure ease-of-use with models.

Feature Selection

A hybrid feature selection method was used for importance selection. Correlation based filtering drops variables that have weak relationships ($|r| < 0.05$) or have high multicollinearity ($|r| > 0.9$). Learning regression would be useful for prediction but not when trying to interpret.

Model Training

Six supervised regression algorithms were trained and benchmarked: Ridge Regression, Decision Tree, Random Forest, XGBoost, LightGBM, and CatBoost. We performed model evaluation over a 3-fold cross-validation framework based on metrics R^2, RMSE, MAE, and MAPE. CatBoost model obtained the best results in terms of accuracy with $R^2 \approx 0.93$, thus confirming its robustness with respect to non-linear agricultural data.

Deployment

The final model CatBoost was serialized and integrated into a FastAPI back-end for running inferences in real time. The client facing front-end was built using React along with Tailwind CSS which provided multilingual voice enabled interface. It also incorporates our ongoing work of integration of an AI assistant LLaMA based system with DuckDB for natural language querying like “Which crop had the highest yield in Mandya 2015?”.

4. Results and Discussion

The results underscore the efficacy of machine learning algorithms for large-scale crop yield prediction at different agro-climatic regions in India. A real effort was made in the present study to do a systematic analysis using a comprehensive district level dataset considering both predictive accuracy and generalization capability of the model. For comparing the performance of the algorithm, the first requirement is to build a uniform system under similar preprocessing, feature selection and cross validation.

Model Evaluation

To test this hypothesis, we used an All-India Agricultural Dataset which contains more than 200,000 records. Six regression machine-learning algorithms viz., Ridge, Decision Tree, Random Forest, XGBoost, LightGBM and CatBoost were compared with each other on 3-fold cross validation on uniform preprocessed selected features.

Table 1: Model Performance Summary

Model	CV R2	Test R2	RMSE	MAE	MAPE (%)
CatBoost	0.9256	0.9271	3.36	1.03	39.84
LightGBM	0.9154	0.9275	3.36	0.99	39.24
XGBoost	0.9069	0.9103	3.73	1.14	46.32
Decision Tree	0.7604	0.7646	6.05	2.14	124.3
Ridge	0.7066	0.7059	6.76	2.06	73.7
Random Forest	0.1711	0.1707	11.36	3.61	134.6

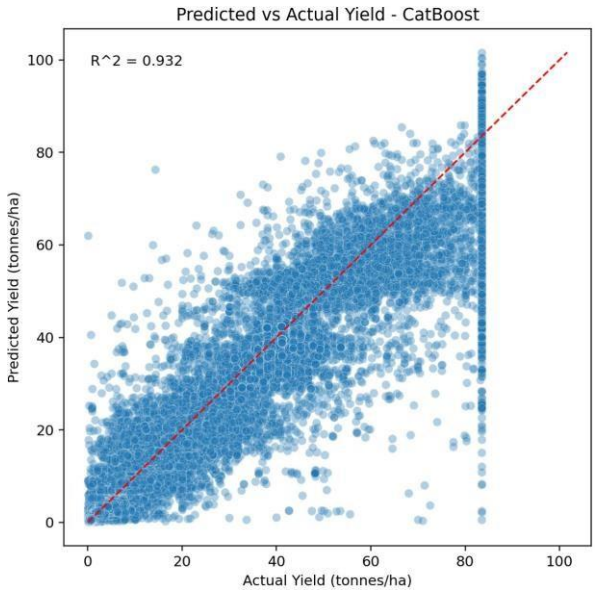

Figure 2. Actual vs Predicted Crop Yield for CatBoost Model

Performance Analysis

- **Superior Predictive Accuracy:** CatBoost reached the highest performance level as measured by $R^2 = 0.93$, which demonstrates its ability to provide accurate yields from agriculture and shows how consistently predictive and reliable CatBoost is at estimating yield per hectare of all crops grown in all agricultural regions.
- **Effectiveness of Gradient-Boosting Frameworks:** Regression approaches using Gradient Boosting Methods, especially CatBoost, LightGBM, and XGBoost, achieved significantly better performance compared with traditional models like Ridge Regression and Decision Tree, which could not capture non-linear complex relationships or effectively deal with the categorical nature of the data set.
- **Model Robustness and Generalization:** The training and testing performance of CatBoost (train $R^2 = 0.933$; test $R^2 = 0.927$) show close proximity to both estimates, demonstrating CatBoost's high generalisation capabilities.
- **Interpretability and Computational Stability:** The advantages of CatBoost's in-built missing value handling capability, its ordering mechanism, and its ability to provide features' interpretation due to the direct input provided by each training observation give CatBoost favorable consideration as a reliable application for real- world agricultural yield forecasting.

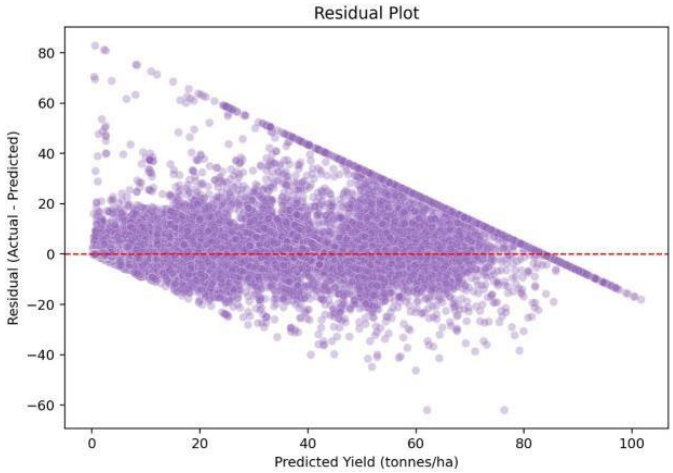

Figure 3. Residuals versus Predicted Yield Values for CatBoost Model

Comparative Benchmarking

Table 2 contains data to contrast what similar yield estimation approaches have generated against the proposed framework. Most prior yield prediction methodologies rely heavily on climate and/or soil attributes. In contrast to all prior yield prediction methods, the current methodology shows that accurate yield estimations can be made using only agro- economic and spatial-based attributes.

Table 2: Comparison with Existing Research

Aspect	Previous Research	Present Work
Dataset Scope	Regional (1–2 states)	Pan-India (>2 lakh records)
Input Features	Weather + Soil parameters	Only agro- economic and spatial features
Feature Selection	Basic filtering or none	Correlation +LASSO (ElasticNetCV)
Models Benchmarked	2–3 models	6 modern regressors under CV
Evaluation Metrics	R2/RMSE	IEEE-standard suite (R2, RMSE, MAE, MAPE)
Output	Static ML model	Deployed web system + Ask-AI assistant

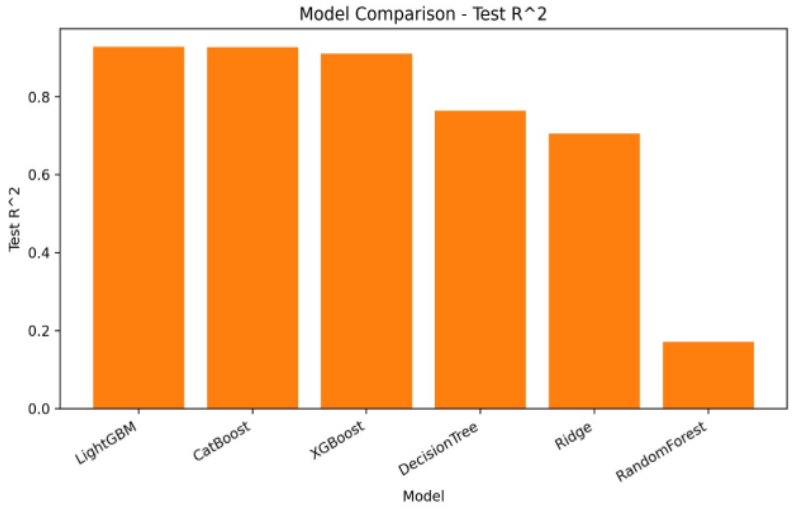

Figure 4. Comparative R2 Scores Across Regression Models

Deployment and Insights

A full-featured web application was developed and launched on the best-performing CatBoost model. The application includes backend and frontend components for real-time inference and interaction with the user. This backend uses FastAPI to load the trained CatBoost ML model, route API requests, and respond asynchronously to each incoming prediction request. Finally, it exposes REST end points for yield predictions and query-based insights so that it can support a high number of concurrent users while providing fast response time. The trained CatBoost model and preprocessing schema this application is built on are serialized so every time the model is deployed into a new environment, it will always provide consistent predictions and preserve a dedicated version history.

The frontend of the application is implemented in both React and Tailwind CSS, allowing for greatly intuitive user experience accessible from any mobile or desktop web browser. The farmers can input most of the parameters on the website, like state, district, crop type, season, and cultivated area. Once the farmer uploads all these parameters, yield prediction and forecasts of total production will be automatically generated. The frontend incorporates multilingual support for English and Kannada, while also including voice-based input using the Web Speech API, making its use simple for farmers unaccustomed to technology. The predicted yields that the

application makes are visually represented on the frontend portion of the application via dynamic charts and interactive visual displays.

The Ask-AI Module is an advanced feature of the deployment made possible by using LLaMA version 3.2 and DuckDB to provide analytics through conversation- style interfaces. Users now have the capability to input natural- language questions (for example: “In Mysuru in 2019, what crop produced the highest yield?”) into the system while the module automatically converts those questions into SQL commands that run against the user’s local database. Results are then summarized in a user-friendly multilingual format that creates a connection between analytics and decision-making. In addition, the modular nature of the system means that additional features can be added over time (e.g., real time weather API connections; source data links for soil quality; sources of NDVI images from space). In total, the design of this complete system has successfully taken a theoretical machine-learning setup and created a useful tool for making decisions on the ground, allowing for access to data to be delivered in a timely, meaningful way to farmers and policy- makers on an accessible, intelligent, and scalable foundation.

5. Conclusion

The proposed crop yield prediction framework has been able to generate a workable and scalable solution by using machine learning models that rely entirely on publicly accessible agricultural databases to forecast crop production without needing to explicitly include climatic or soil-based variables. It has been shown that the CatBoost model provides an R^2 score of about 0.93, hence justifying its high level of predictive accuracy, low rates of error across multiple Indian districts and different types of crops, and ability to generalize to diverse inputs coming from different Indian districts and differing types of crops. This framework has been designed following IEEE conference publication guidelines, and as such, it allows all workflow steps from data cleaning and feature selection through model benchmarking to subsequent operational deployment associated with this framework to be completely reproducible and made available to other research.

The FastAPI backend, together with a multilingual React- based web interface, enables users to access real-time yield predictions and interact with visualizations of both the input data and output predictions via a completely automated process. By requiring little to no setup from either end, such implementation of a multilingual interface will facilitate accessibility for farmers and policymakers interested in availing services of real-time yield prediction through voice-enabled applications, or simply voice-to-text. The Ask-AI component of this application further enhances user convenience by allowing a user to communicate in their native language while accessing natural language explanations of the underlying computer-based predictions.

Weather and soil data, including rainfall and temperature from real-time APIs, as well as other localized metrics such as humidity and NDVI indices, will be integrated into the system in the future to improve the temporal flexibility of the current research prototype. Furthermore, the system will also include XAI tools, which will include SHAP feature visualization, to allow for improved understanding of model decisions. Finally, the research prototype will be adapted into a mobile-friendly PWA (Progressive Web Application) and will contain offline functionality that will make it more accessible for rural users. These features will enable the current prototype to evolve into a production-ready intelligent decision support tool for sustainable agricultural planning in India.

References

[1] M. Kalimuthu, P. Vaishnavi and M. Kishore, “Crop Prediction using Machine Learning,” 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2020, pp. 926–932.

[2] P.S. Nishant, P. Sai Venkat, B. L. Avinash and B. Jabber, “Crop Yield Prediction based on Indian Agriculture using Machine Learning,” 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1–4.

[3] D. J. Reddy and M. R. Kumar, “Crop Yield Prediction using Machine Learning Algorithm,” 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 2021, pp. 1466–1470.

[4] A. Nigam, S. Garg, A. Agrawal and P. Agrawal, "Crop Yield Prediction Using Machine Learning Algorithms," 2019 Fifth International Conference on Image Information Processing (ICIIP), Shimla, India, 2019, pp. 125–130.
[5] "data.gov.in." [Online]. Available: https://data.gov.in/
[6] M. G. Ananthara, T. Arunkumar and R. Hemavathy, "CRYan improved crop yield prediction model using bee hive clustering approach for agricultural data sets," 2013 International Conference on Pattern Recognition, Informatics and Mobile Engineering, 2013, pp. 473–478.
[7] A. M. Awan and M. N. M. Sap, "An intelligent system based on kernel methods for crop yield prediction," Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2006.
[8] S. Bang, R. Bishnoi, A. S. Chauhan, A. K. Dixit and I. Chawla, "Fuzzy Logic based Crop Yield Prediction using Temperature and Rainfall parameters predicted through ARMA, SARIMA, and ARMAX models," 2019 Twelfth International Conference on Contemporary Computing (IC3), 2019.
[9] A. Kumar, N. Kumar and V. Vats, "Efficient Crop Yield Prediction Using Machine Learning Algorithms," International Research Journal of Engineering and Technology (IRJET), vol. 5, no. 6, 2018.
[10] N. Singh and S. Chaturvedi, "Weather Forecasting Using Machine Learning," 2019 International Conference on Signal Processing and Communication (ICSC), 2019.
[11] A. Parmar and M. Sompura, "Rainfall Prediction using Machine Learning," 2017 International Conference on (ICIIECS), 2017.
[12] S. Nene and R. Priya, "Prediction of Crop yield using Machine Learning," International Research Journal of Engineering and Technology (IRJET), vol. 5, no. 2, 2018.
[13] P. Priya, U. Muthaiah and M. Balamurugan, "Predicting Yield of the Crop Using Machine Learning Algorithm," International Journal of Engineering Sciences Research Technology.
[14] S. Mishra, D. Mishra and G. H. Santra, "Applications of machine learning techniques in agricultural crop production: a review paper," Indian J. Sci. Technol., vol. 9, no. 38, 2016.