



International Journal of Artificial Intelligence and Machine Learning

Publisher's Home Page: <https://www.svedbergopen.com/>



Open Access

Machine Learning Based Prediction of Agricultural Foodgrain Production in Indian States Using Agricultural, Irrigation and Fertilizer Indicators

Akash Baboo¹, Kokab Askari², Rakesh Kumar Mishra³, Akash Rai⁴, Shyam Srivastav⁵, Pawan Kumar Yadav⁶

¹S.M.T. School of Agriculture Science, Sandip University, Sijoul, Madhubani-847235, Email: akashpatel9485@gmail.com

²Faculty of Agriculture, Maharishi Markandeshwar University, (deemed to be University) Mullana, Ambala-133207, Email: dr.kokabaskari@mmumullana.org

³School of Advanced Agriculture Sciences and Technology, Chhatrapati Shahu Ji Maharaj University Kanpur-208024, Email: rakeshkmishra@csjmu.ac.in

⁴Aryabhata College, University of Delhi, New Delhi-110021, Email: drakashrai5@gmail.com

⁵Dr. Rammanohar Lohia Avadh University, Ayodhya-224001 Email: shyamanandfzd@gmail.com

⁶College of Dairy Technology, Etawah, CSAUA&T, Kanpur, Email: pawankyadav14@gmail.com

ABSTRACT

Agricultural foodgrain production is influenced by multiple interacting factors, including cultivated area, irrigation availability, fertilizer use, rainfall, and previous production levels. Reliable prediction of foodgrain production can support agricultural planning, resource allocation, and food-security management. This study develops a machine-learning-based framework for predicting agricultural foodgrain production across Indian states using agricultural, irrigation, fertilizer, climatic, and historical production indicators. A state-year dataset comprising 600 observations from 25 Indian states during 2000–01 to 2023–24 was developed. Seven predictor variables, namely Net Sown Area, Gross Sown Area, Net Irrigated Area, Gross Irrigated Area, Fertilizer Consumption, Annual Rainfall, and Previous-Year Foodgrain Production, were used to predict Foodgrain Production. Two ensemble regression algorithms, Random Forest (RF) and Extreme Gradient Boosting (XGBoost), were developed and evaluated using the correlation coefficient (R), coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and scatter index (SI). On the testing dataset, RF achieved $R = 0.996298$ and $R^2 = 0.987867$, with RMSE = 2003.8967 thousand tonnes, MAE = 957.6846 thousand tonnes, MAPE = 6.919%, and SI = 0.148889. XGBoost achieved $R = 0.996503$ and $R^2 = 0.983717$, with RMSE = 2321.4478 thousand tonnes, MAE = 933.7023 thousand tonnes, MAPE = 11.312%, and SI = 0.172483. Based on the overall testing performance, RF was identified as the better-performing model. Feature-importance analysis further identified Previous-Year Foodgrain Production as the most influential predictor in both RF and XGBoost. The findings demonstrate the potential of ensemble machine-learning models, particularly RF, for data-driven prediction of agricultural foodgrain production across Indian states.

Keywords: Agricultural Foodgrain Production; Machine Learning; Random Forest; XGBoost; Agricultural Prediction

INTRODUCTION

Agriculture is a major component of India's economy and plays a central role in national food security, rural livelihoods, and the availability of essential food commodities. Among the different agricultural indicators, foodgrain production is particularly important because it represents the combined production of major cereal and pulse crops and is directly associated with the country's food-supply system. India has a highly heterogeneous agricultural environment, with substantial differences among states in cultivated area, irrigation infrastructure, fertilizer use, climatic conditions, cropping patterns, and agricultural productivity. Consequently, the magnitude and temporal variability of foodgrain production differ considerably across Indian states. The Directorate of Economics and Statistics (DES), Department of Agriculture and Farmers Welfare, Government of India, provides systematic state- and national-level statistics on agricultural area, production, yield, irrigation, and related indicators. The official statistics report total foodgrain production of approximately 329.69 million tonnes in 2022–23 and 332.30 million tonnes in 2023–24, demonstrating the large scale of India's foodgrain production system [1].

The spatial variation in Indian foodgrain production is substantial. Official Government of India statistics show

pronounced differences among states, with Uttar Pradesh, Madhya Pradesh, Punjab, Rajasthan, Bihar, and several other states contributing substantially to national foodgrain production [1,2]. For example, the reported foodgrain production of Uttar Pradesh and Madhya Pradesh in 2023–24 was approximately 60.29 and 41.47 million tonnes, respectively, whereas production in several smaller states was considerably lower [1]. Such variation indicates that national-level production alone cannot adequately represent the underlying agricultural production processes. State-level analysis is therefore useful for identifying the agricultural and resource-related factors associated with differences in foodgrain production.

Agricultural production is determined by the combined influence of land availability, irrigation, agricultural inputs, climatic conditions, and management practices. The cultivated area represents the land resource available for crop production, while irrigation can reduce dependence on rainfall and support crop cultivation under variable climatic conditions. Official land-use statistics indicate that India's net irrigated area was approximately 82.42 million hectares and gross irrigated area approximately 130.55 million hectares in 2023–24 [3]. The considerable expansion and spatial variation of irrigation infrastructure make irrigation-related indicators important variables for assessing agricultural production. Similarly, fertilizer consumption represents an important agricultural input associated with crop productivity and production intensity. These variables therefore provide useful explanatory information when developing a predictive framework for foodgrain production.

Climatic variability further complicates agricultural production forecasting. Rainfall is particularly important for Indian agriculture because of the strong influence of the southwest monsoon on crop water availability. However, the effect of rainfall is not spatially uniform because irrigation coverage, soil conditions, cropping patterns, and agricultural practices differ substantially among states. Consequently, agricultural production prediction requires consideration of multiple interacting variables rather than relying exclusively on a single climatic indicator. Recent research on Indian agriculture has demonstrated that integrating climate, remote-sensing, agronomic, and historical yield information can improve the ability of machine-learning models to represent spatial and temporal variations in crop yield [4–8]. For example, De Clercq and Mahdi investigated machine-learning-based rice-yield prediction across 247 rice-producing districts in India using two decades of climate, satellite, and yield data and demonstrated the feasibility of machine-learning approaches for district-level yield forecasting [4].

Historical production is another potentially important predictor of agricultural output. Foodgrain production generally exhibits temporal persistence because state-level production is influenced by relatively persistent characteristics such as cultivated area, irrigation infrastructure, cropping patterns, agricultural practices, and production capacity. Incorporating Previous-Year Foodgrain Production can therefore provide the model with information about the historical production level of a state that may not be fully captured by contemporaneous irrigation, fertilizer, cultivated-area, and rainfall variables. Recent Indian agricultural ML studies have similarly incorporated historical yield information and temporal features to improve predictive performance [4,9]. More recent work has demonstrated the value of lagged yield and temporal indicators in multi-crop prediction frameworks for India [9].

Conventional statistical models have traditionally been used for agricultural forecasting; however, the relationships among agricultural inputs, climatic conditions, cultivated area, irrigation, and production can be nonlinear and may involve interactions among several predictors. Machine learning (ML) provides a data-driven alternative capable of identifying complex relationships without requiring a predefined mathematical relationship between predictors and the response variable. ML-based crop-yield prediction has consequently received increasing attention in agricultural research. Studies conducted in India have evaluated Random Forest, support vector methods, gradient-boosting approaches, and other ML techniques for crop-yield prediction and have demonstrated their potential for modelling nonlinear agricultural relationships [5–8]. A study focused on Rajasthan, for example, found Random Forest to be an effective model among the approaches evaluated for crop-yield prediction [5].

Among machine-learning algorithms, Random Forest (RF) and Extreme Gradient Boosting (XGBoost) are particularly suitable for agricultural regression problems. RF is an ensemble learning method that constructs multiple decision trees using randomized subsets of observations and predictor variables and combines their outputs to obtain a robust prediction [10]. XGBoost is a scalable gradient-boosting framework that sequentially constructs decision trees while optimizing an objective function incorporating both prediction loss and regularization [11]. Both methods can capture nonlinear relationships and interactions among predictor variables, while requiring relatively limited assumptions about the underlying distribution of the data. Their successful application to crop-yield prediction in different agricultural settings further supports their use for the present investigation [4–9].

Although numerous studies have investigated crop-yield prediction using machine learning, much of the existing literature has concentrated on individual crops, district-level yield, remote-sensing variables, or combinations of satellite and climatic observations [4,6–8]. Comparatively fewer studies have focused on state-level total foodgrain production using a unified set of agricultural, irrigation, fertilizer, rainfall, and historical-production indicators. This distinction is important because total foodgrain production is an aggregate output measure rather than the yield of an individual crop. Furthermore, state-level production data provide a broader perspective for agricultural planning and resource assessment.

Accordingly, the present study develops a machine-learning framework for predicting agricultural foodgrain production across Indian states using agricultural, irrigation, fertilizer, climatic, and historical production indicators. The dataset covers the period 2000–01 to 2023–24 and includes state-year observations. The response

variable is annual foodgrain production (thousand tonnes), while the predictor variables include Previous-Year Foodgrain Production, Net Sown Area, Gross Sown Area, Net Irrigated Area, Gross Irrigated Area, Fertilizer Consumption, and Annual Rainfall. Two ensemble regression algorithms, Random Forest and XGBoost, are employed and comparatively evaluated using the correlation coefficient (R), coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and scatter index (SI). In addition, feature-importance analysis is conducted to identify the relative contribution of the agricultural and climatic indicators to foodgrain-production prediction.

The principal objective of this study is therefore not only to obtain accurate predictions of state-level foodgrain production but also to determine which agricultural and resource-related variables contain the greatest predictive information. The resulting framework may provide a useful data-driven approach for agricultural production assessment, comparative state-level analysis, and evidence-based agricultural planning in India.

2. Research Methodology and Machine-Learning Framework

2.1 Research Framework

The methodology was designed to develop a machine-learning-based framework for predicting annual agricultural foodgrain production across Indian states. The overall research procedure consisted of six major stages: (i) collection of secondary agricultural data, (ii) development and integration of the state-year dataset, (iii) exploratory data analysis, (iv) data preparation and chronological model partitioning, (v) development and optimization of RF and XGBoost regression models, and (vi) model evaluation and interpretation.

The study was formulated as a supervised regression problem in which annual foodgrain production was considered the dependent variable and agricultural, irrigation, fertilizer, rainfall, and previous-year production indicators were used as explanatory variables. The complete dataset comprised 600 observations corresponding to 25 Indian states and 24 agricultural years from 2000–01 to 2023–24.

The analytical workflow was designed to maintain consistency between the temporal nature of agricultural production and the machine-learning methodology. In particular, the observations were arranged chronologically before separating the training and testing datasets. This approach avoids unnecessarily mixing later-year observations with earlier-year observations during model development and provides a more realistic assessment of predictive performance for subsequent observations.

2.2 Data Sources and Dataset Development

The dataset was developed from secondary agricultural and environmental information reported through authoritative sources of the Government of India. The principal source was the Directorate of Economics and Statistics (DES), Department of Agriculture and Farmers Welfare, Ministry of Agriculture and Farmers Welfare, Government of India, particularly the Agricultural Statistics at a Glance publication. These publications provide officially compiled information on agricultural area, production, yield, irrigation coverage, and related agricultural indicators at national and state levels.

The official Agricultural Statistics at a Glance 2024 includes dedicated tables for foodgrain area, production, yield, and irrigation coverage for major producing states. For example, the 2023–24 data reported in the publication indicate an all-India foodgrain production of approximately 332.29 million tonnes and provide corresponding state-level production and irrigation information. The official Directorate also maintains historical agricultural statistics and land-use datasets, allowing agricultural indicators to be examined over extended periods.

For the present ML analysis, the collected information was organized into a structured panel-like state-year dataset. Each observation represents a particular Indian state during a particular agricultural year. The final dataset contains 600 state-year observations, representing 25 states $\times$ 24 agricultural years (2000–01 to 2023–24). The target variable was defined as Foodgrain Production (thousand tonnes).

The predictor variables were selected to represent four major dimensions of agricultural production: cultivated land, irrigation resources, agricultural inputs, and climatic conditions. Previous-year foodgrain production was additionally included to represent temporal production continuity.

2.3 Variables Used for Machine-Learning Prediction

Seven predictor variables were used to estimate annual state-level foodgrain production. These variables were selected because they represent measurable agricultural and environmental conditions that can influence the scale and productivity of foodgrain cultivation.

Net Sown Area (thousand ha) represents the area sown with crops at least once during the agricultural year and provides an indication of the land base available for cultivation. Gross Sown Area (thousand ha) accounts for the total cropped area, including areas cultivated more than once, and therefore reflects the intensity of land use.

Net Irrigated Area (thousand ha) and Gross Irrigated Area (thousand ha) represent the extent of agricultural land supported by irrigation. These variables are particularly relevant to Indian agriculture because irrigation can reduce dependence on rainfall and influence crop productivity. Official agricultural statistics explicitly report foodgrain production together with irrigation coverage at both national and major-state levels.

Fertilizer Consumption (kg ha^{-1}) represents the intensity of fertilizer use and serves as an indicator of agricultural input availability. Appropriate nutrient availability is an important component of crop productivity, although the relationship between fertilizer use and production may be nonlinear and may vary according to agro-climatic

conditions.

Annual Rainfall (mm) represents the climatic water input received during the agricultural year. Rainfall is particularly relevant to Indian agriculture because of the substantial dependence of agricultural production on monsoon variability. Previous studies have identified rainfall as an important determinant of Indian crop production and have demonstrated the potential of ML methods for modelling rainfall-production relationships. Finally, Previous-Year Foodgrain Production (thousand tonnes) was included as a temporal predictor. Agricultural production often exhibits persistence across successive years because factors such as cropping patterns, agricultural infrastructure, cultivated area, technology adoption, and production systems do not change independently from one year to the next. Including the previous year's production therefore provides the models with information about recent production conditions. The complete input-output structure is summarized in Table 1.

Table 1. Description of variables used for machine-learning prediction

Variable	Role	Unit	Function
Net Sown Area	Input	thousand ha	Land availability
Gross Sown Area	Input	thousand ha	Cropping intensity
Net Irrigated Area	Input	thousand ha	Irrigation availability
Gross Irrigated Area	Input	thousand ha	Total irrigated cropping area
Fertilizer Consumption	Input	kg ha ⁻¹	Agricultural input intensity
Annual Rainfall	Input	mm	Climatic condition
Previous-Year Foodgrain Production	Input	thousand tonnes	Temporal production indicator
Foodgrain Production	Output	thousand tonnes	ML prediction target

2.4 Data Preprocessing and Quality Control

Before model development, the dataset was subjected to a systematic quality-control procedure. The state and agricultural-year identifiers were retained for data organization and subsequent interpretation, whereas the numerical agricultural variables were converted into appropriate numerical formats.

The dataset was examined for missing observations, duplicate records, invalid numerical entries, and inconsistencies in variable types. Observations containing missing values in the variables required for model development were removed after verification. The final dataset contained 600 complete observations.

Descriptive statistical analysis was subsequently performed to determine the mean, standard deviation, minimum, maximum, median, variance, and coefficient of variation of the numerical variables. Distribution plots and boxplots were used to examine the spread of the variables and identify potentially influential observations.

A correlation analysis was also performed to investigate the linear association between the predictor variables and foodgrain production. The correlation analysis was used for exploratory interpretation rather than for eliminating variables solely on the basis of correlation magnitude, because RF and XGBoost can model nonlinear relationships and interactions between predictors.

2.5 Training and Testing Strategy

Because the observations represent successive agricultural years, the dataset was ordered chronologically before model development. A chronological partition was adopted rather than a purely random train-test split. This strategy provides a more realistic representation of a forecasting-oriented application because earlier observations are used for model development while later observations are retained for independent testing.

Approximately 80% of the observations were assigned to the training dataset and the remaining 20% to the testing dataset. The training dataset was used for model fitting and hyperparameter optimization, whereas the testing dataset was kept separate for final evaluation.

Five-fold cross-validation was performed within the training dataset during hyperparameter optimization. The test dataset was not used during parameter selection, thereby reducing the possibility of optimistic performance estimates.

Because RF and XGBoost are tree-based algorithms, feature standardization was not required for their final model development. Tree-based algorithms partition observations according to feature thresholds and are therefore generally insensitive to the measurement scale of individual predictors.

2.6 Random Forest Regression

Random Forest (RF) is an ensemble-learning algorithm that combines the predictions of multiple decision trees to obtain a robust regression output. Each tree is developed using randomly selected training observations and predictor variables, while the final prediction is obtained by averaging the individual tree predictions.

RF was selected because of its ability to model nonlinear relationships and interactions among agricultural, irrigation, fertilizer, and climatic indicators without requiring a predefined functional relationship between the predictors and foodgrain production. The use of multiple decision trees also makes the algorithm comparatively robust to complex patterns in heterogeneous state-level agricultural data.

The RF model was optimized by varying the number of estimators, maximum tree depth, maximum number of features considered during splitting, minimum number of samples required for a split, and minimum number of samples required at a leaf node. A five-fold cross-validation procedure was used for hyperparameter optimization, with the coefficient of determination (R^2) used as the principal optimization criterion.

The final RF configuration was selected according to the best cross-validation performance obtained during the optimization procedure. The optimized model was subsequently refitted using the complete training dataset and used to generate foodgrain-production predictions for both the training and independent testing datasets.

2.7 Extreme Gradient Boosting Regression

Extreme Gradient Boosting (XGBoost) is an ensemble regression technique based on the sequential construction of decision trees. In contrast to RF, in which trees are constructed largely independently, XGBoost develops trees sequentially so that subsequent trees progressively improve the errors of the existing ensemble.

XGBoost was selected because of its ability to capture nonlinear relationships, interactions, and complex patterns within structured tabular datasets. Such characteristics are relevant to agricultural production modelling, where the effects of rainfall, irrigation, cultivated area, fertilizer consumption, and previous production may interact rather than operate independently. Previous agricultural studies have demonstrated the applicability of XGBoost and other tree-based algorithms to crop-production and yield-prediction problems.

The XGBoost model was optimized with respect to the number of estimators, learning rate, maximum tree depth, minimum child weight, subsampling ratio, and column-sampling ratio. Five-fold cross-validation was used during the optimization process, with R^2 serving as the principal selection criterion.

The optimized XGBoost model was then trained using the complete training dataset. Predictions were generated for the training and testing datasets, and the resulting predictions were compared with the observed foodgrain-production values using the statistical performance indicators described in the following section.

2.8 Model Performance Evaluation

The predictive performance of the developed Random Forest (RF) and Extreme Gradient Boosting (XGBoost) models was evaluated using multiple statistical indicators. The selected indicators were intended to assess correlation, goodness of fit, absolute prediction error, relative prediction error, and normalized prediction error.

The evaluation metrics included the Correlation Coefficient (R), Coefficient of Determination (R^2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Scatter Index (SI).

The correlation coefficient, R, measures the strength of association between observed and predicted foodgrain production values, with values closer to unity indicating stronger agreement. R^2 measures the proportion of variation in observed foodgrain production explained by the model. RMSE provides greater weighting to larger prediction errors, whereas MAE represents the average absolute difference between observed and predicted production. MAPE expresses prediction error in percentage terms, while SI normalizes RMSE relative to the mean observed production and facilitates comparison of prediction accuracy across datasets.

These indicators were calculated separately for the training and testing datasets. Particular emphasis was placed on the independent testing performance, since it provides the most appropriate basis for comparing the generalization capability of RF and XGBoost.

2.9 Feature Importance Analysis

Feature importance analysis was performed to identify the relative contribution of the predictor variables to the prediction of agricultural foodgrain production across Indian states. This analysis provides additional interpretability to the developed machine-learning models by identifying the variables that contribute most strongly to their predictions. For the Random Forest (RF) model, feature importance was determined based on the reduction in impurity associated with each predictor across the ensemble of decision trees. For the Extreme Gradient Boosting (XGBoost) model, feature importance was assessed based on the contribution of individual predictors to the improvement of the model during the sequential construction of decision trees. The importance values were normalized and expressed as percentages to facilitate comparison among the input variables. Higher importance values indicate a greater contribution of the corresponding predictor to the model's estimation of foodgrain production. The feature-importance results obtained from RF and XGBoost were subsequently compared to identify predictors that consistently contributed to model performance. This analysis enhances the interpretability of the developed models and provides useful insights into the relative importance of cultivated area, irrigation availability, fertilizer consumption, rainfall, and previous-year foodgrain production in predicting state-level agricultural foodgrain production.

3. Results and Discussion

3.1 Exploratory Data Analysis

Exploratory data analysis (EDA) was performed to examine the statistical characteristics, distribution patterns, spatial variability, temporal trends, and interrelationships among the agricultural, irrigation, fertilizer, rainfall, and foodgrain-production variables. The EDA comprised descriptive statistical analysis, distribution analysis, average state-wise foodgrain production, annual foodgrain-production trends, correlation analysis, and correlation heatmap

analysis. These analyses provide an initial understanding of the variability and structure of the dataset before evaluating the predictive performance of the developed machine-learning models.

3.1.1 Descriptive Statistical Analysis

The descriptive statistics of the six predictor variables and the target variable are presented in Table 2. The dataset comprises 600 state-year observations, representing substantial spatial and temporal variability in agricultural conditions. The Net Sown Area, Gross Sown Area, Net Irrigated Area, and Gross Irrigated Area exhibited relatively high coefficients of variation, ranging from 90.811% to 102.647%. This indicates considerable differences in cultivated and irrigated areas among the state-year observations.

Fertilizer consumption and annual rainfall showed comparatively lower, although still substantial, variability, with CV values of 53.832% and 51.976%, respectively. Foodgrain production exhibited particularly high variability, with a CV of 134.432%, while previous-year foodgrain production showed a similar CV of 134.484%. This large variation reflects differences in agricultural production scale among Indian states and across years.

Overall, the descriptive statistics indicate that the dataset encompasses a broad range of agricultural, irrigation, climatic, and production conditions. Such variability provides an appropriate basis for developing machine-learning models capable of learning relationships across heterogeneous state-year observations.

Table 2. Descriptive statistics of agricultural, irrigation, fertilizer, rainfall, and foodgrain-production variables

Variable	Mean	SD	Min	Median	Max	CV (%)
Net Sown Area (thousand ha)	5275.858	4949.896	113.14	4038.4	17572.04	93.822
Gross Sown Area (thousand ha)	9163.595	8321.534	177.32	7890.465	33505.48	90.811
Net Irrigated Area (thousand ha)	2798.829	2853.112	24.93	2032.285	13719.62	101.939
Gross Irrigated Area (thousand ha)	4589.371	4710.838	39.45	3404.64	22755.24	102.647
FertilizerConsumption (kg/ha)	104.94	56.491	20.07	102.08	307.96	53.832
Annual Rainfall (mm)	1422.823	739.524	411.19	1181.1	3649.52	51.976
Previous-YearFoodgrain Production(thousand tonnes)	11426.17	15366.35	54.67	7760.575	83106.96	134.484
FoodgrainProduction (thousand tonnes)	11599.37	15593.27	56.95	7860.77	83106.96	134.432

3.1.2 Distribution Analysis

The distributions of the predictor variables and foodgrain-production target are illustrated in Fig. 1. The distribution plots provide a visual assessment of the range, concentration, dispersion, and potential asymmetry of the variables. Considerable differences in the distributions are evident because the dataset combines observations from multiple Indian states with substantially different agricultural scales.

The production-related variables display relatively wide distributions, consistent with the high coefficients of variation observed in Table 2. In particular, the foodgrain-production variable exhibits a broad range, reflecting the substantial difference between states with relatively small and large agricultural production. Similar patterns are observed for sown and irrigated areas.

The distribution analysis therefore complements the numerical descriptive statistics and confirms that the dataset contains substantial heterogeneity. This characteristic is important for the subsequent application of ensemble tree-based algorithms, which can accommodate nonlinear relationships and interactions among variables.

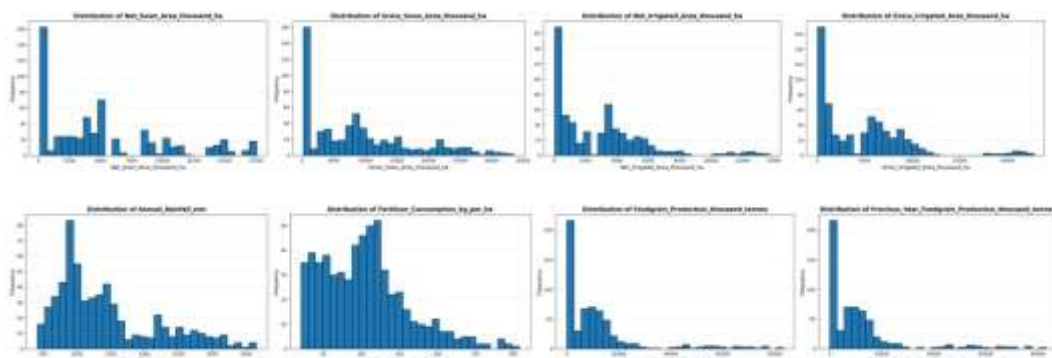
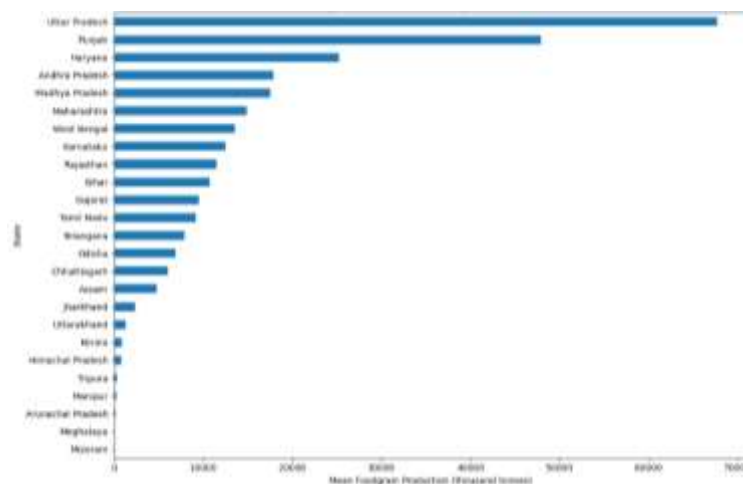


Fig. 1. Distribution plots of agricultural, irrigation, fertilizer, rainfall, and foodgrain-production variables**3.1.3 Average State-Wise Foodgrain Production**

The spatial variation in agricultural production was further examined by calculating the average foodgrain production for each Indian state over the study period. The results are presented in Fig. 2. The figure demonstrates substantial differences in average foodgrain production among the states.

States with larger agricultural areas and relatively greater irrigation availability generally exhibit higher average foodgrain production, whereas states with smaller cultivated areas tend to record lower production levels. However, production differences cannot be attributed to cultivated area alone because agricultural productivity is also influenced by irrigation, fertilizer application, rainfall, cropping intensity, agricultural practices, and other state-specific factors.

The substantial state-wise variation observed in Fig. 5 provides an important justification for developing a state-level machine-learning framework rather than relying solely on an aggregate national production series. By retaining individual state-year observations, the proposed dataset incorporates spatial heterogeneity that would otherwise be lost through national-level aggregation.

**Fig. 2. Average state-wise agricultural foodgrain production during the study period****3.1.4 Annual Foodgrain Production Trend**

The temporal variation in foodgrain production was investigated by examining the annual production trend over the study period. The resulting trend is presented in Fig. 3. The figure illustrates year-to-year variations in aggregate foodgrain production across the states included in the dataset.

The annual trend demonstrates that foodgrain production does not remain constant over time and exhibits temporal fluctuations. Such variations may be associated with changes in cultivated area, irrigation availability, fertilizer use, rainfall conditions, agricultural practices, and other factors affecting agricultural production. The inclusion of Previous-Year Foodgrain Production as a predictor is therefore particularly relevant because production in the preceding year can provide useful information about the production level in the subsequent year.

The temporal trend also demonstrates the importance of preserving the chronological structure of the dataset during model development. A chronological partitioning strategy helps ensure that the testing observations represent later state-year conditions rather than randomly selected observations from the same period as the training data.

**Fig. 3. Annual trend of agricultural foodgrain production across the Indian states during the study period****3.1.5 Correlation Analysis**

The pairwise relationships among the predictors and the target variable were examined using the correlation matrix. The analysis provides an initial indication of the strength and direction of linear associations among the agricultural indicators and foodgrain production. The relationships between foodgrain production and the land-use, irrigation,

fertilizer, rainfall, and previous-year production variables provide useful preliminary information regarding their potential predictive relevance. Relationships among Net Sown Area, Gross Sown Area, Net Irrigated Area, and Gross Irrigated Area are also of interest because these variables represent related dimensions of agricultural land utilization and irrigation.

However, correlation coefficients represent only linear associations and should not be interpreted as evidence of causation. A variable with a relatively weak pairwise correlation may still contribute substantially to prediction when nonlinear relationships or interactions are present. Therefore, all theoretically relevant variables were retained for the RF and XGBoost models and subsequently evaluated through model-based feature-importance analysis.

For visual interpretation of the correlation structure, the correlation coefficients were also represented using a heatmap, as shown in Fig. 4. The heatmap facilitates rapid identification of positive and negative associations among the variables. The visualization is particularly useful for examining the relationships among the sown-area and irrigation variables, as well as their associations with foodgrain production. It also provides an initial indication of whether the predictors contain substantial interrelationships. Nevertheless, the presence of correlated predictors does not automatically require their removal in the present study because RF and XGBoost can model complex nonlinear relationships and interactions among predictors.

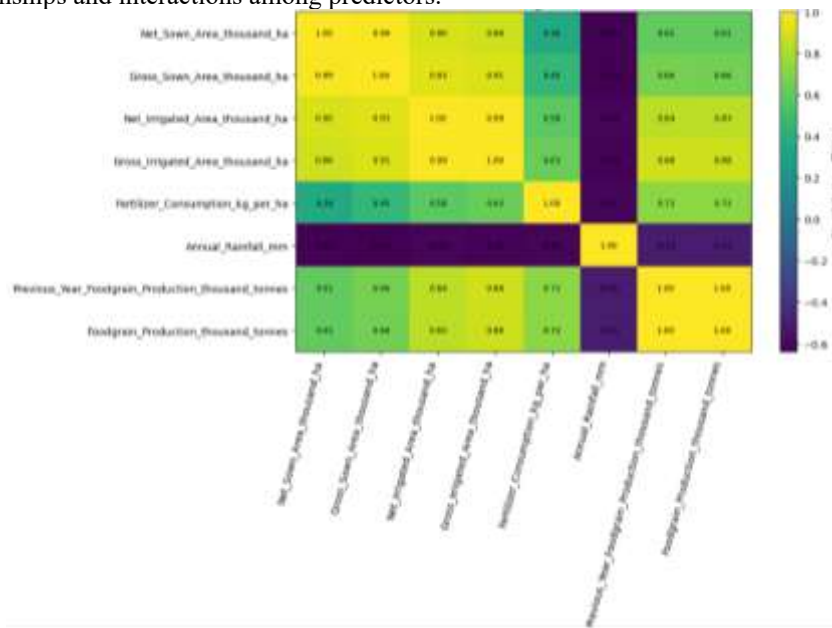


Fig. 4. Correlation heatmap of the predictor variables and agricultural foodgrain production

3.2 Model Performance on Training and Testing Datasets

The predictive performance of the developed Random Forest (RF) and Extreme Gradient Boosting (XGBoost) regression models was evaluated using the correlation coefficient (R), coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and scatter index (SI). The comparative results for the training and testing datasets are presented in Table 3. In addition, the agreement between observed and predicted foodgrain production values can be visually assessed from the observed-versus-predicted plots shown in Fig. 5.

Table 3. Performance of RF and XGBoost models on training and testing datasets

Dataset	Model	R	R^2	RMSE	MAE	MAPE (%)	SI
Training	RF	0.999698	0.999386	367.1008	180.9051	1.616	0.03297
Testing	RF	0.996298	0.987867	2003.897	957.6846	6.919	0.148889
Training	XGBoost	0.999788	0.999575	305.5009	196.9641	8.709	0.027437
Testing	XGBoost	0.996503	0.983717	2321.448	933.7023	11.312	0.172483

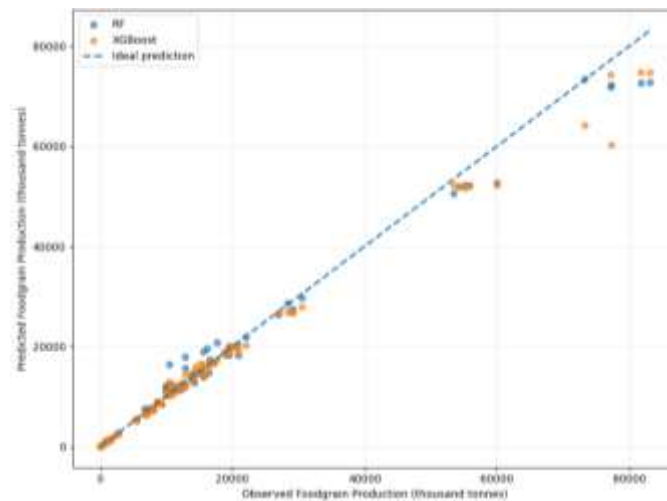


Fig. 5. Observed versus predicted agricultural foodgrain production for the RF and XGBoost models on the training and testing datasets.

Table 3 demonstrates that both RF and XGBoost achieved very high predictive performance on the training dataset. The RF model yielded $R = 0.999698$ and $R^2 = 0.999386$, with RMSE of 367.1008, MAE of 180.9051, MAPE of 1.616%, and SI of 0.032970. These results indicate an almost complete agreement between the observed and predicted foodgrain production values within the training dataset. Similarly, XGBoost produced $R = 0.999788$ and $R^2 = 0.999575$, accompanied by RMSE = 305.5009, MAE = 196.9641, MAPE = 8.709%, and SI = 0.027437. Thus, XGBoost exhibited a marginally higher R and R^2 and lower RMSE and SI than RF during training, although its MAPE was higher.

The testing results provide a more important assessment of the generalization capability of the models. As shown in Table 3, RF achieved $R = 0.996298$ and $R^2 = 0.987867$ on the testing dataset. The corresponding RMSE, MAE, MAPE, and SI were 2003.8967, 957.6846, 6.919%, and 0.148889, respectively. These values demonstrate that the RF model retained a high level of predictive accuracy when applied to observations not used during model training. The high testing R^2 indicates that approximately 98.79% of the variation in the observed foodgrain production values was explained by the RF model.

The XGBoost model also maintained strong testing performance, with $R = 0.996503$ and $R^2 = 0.983717$. Its testing RMSE was 2321.4478, while MAE, MAPE, and SI were 933.7023, 11.312%, and 0.172483, respectively. Although XGBoost obtained a marginally higher R than RF, its testing R^2 was lower and its RMSE, MAPE, and SI were higher. Therefore, the results indicate that the slightly stronger correlation achieved by XGBoost did not translate into superior overall predictive accuracy.

The comparison between training and testing results also provides information regarding model generalization. For RF, R^2 decreased from 0.999386 in training to 0.987867 in testing, whereas the corresponding decrease for XGBoost was from 0.999575 to 0.983717. The reduction in performance between the two datasets is expected because the testing observations were not used during model development. Nevertheless, both models maintained R^2 values above 0.98 on the testing dataset, indicating strong predictive capability for the state-year observations included in the study.

As illustrated in Fig. 5, the predicted values are expected to closely follow the corresponding observed foodgrain production values, particularly for the RF model. The concentration of observations around the ideal prediction line further supports the strong agreement between measured and predicted production values. However, the spread of some observations away from the ideal line indicates that prediction errors remain for certain state-year combinations. Such deviations may be associated with the substantial variability in agricultural production among Indian states and across different years.

3.3 Feature Importance Analysis

Feature importance analysis was performed to identify the relative contribution of the selected agricultural, irrigation, fertilizer, climatic, and historical production indicators to the prediction of agricultural foodgrain production. The analysis was conducted separately for the Random Forest (RF) and Extreme Gradient Boosting (XGBoost) models. The resulting feature-importance values are presented in Table 4. Higher importance values indicate a greater contribution of the corresponding predictor to the model's prediction.

Table 4. Feature importance obtained from RF and XGBoost models

Rank	Predictor	RF Importance (%)	XGBoost Importance (%)
1	Previous-Year Foodgrain Production (thousand tonnes)	34.18	56.052

2	Gross Irrigated Area (thousand ha)	16.671	6.015
3	Net Irrigated Area (thousand ha)	16.185	5.294
4	Net Sown Area (thousand ha)	10.672	21.124
5	Gross Sown Area (thousand ha)	10.541	1.809
6	Fertilizer Consumption (kg/ha)	10.286	9.646
7	Annual Rainfall (mm)	1.464	0.061

As shown in Table 4, the RF model identified Previous-Year Foodgrain Production as the most influential predictor, with an importance of 34.180%. This was followed by Gross Irrigated Area (16.671%), Net Irrigated Area (16.185%), Net Sown Area (10.672%), Gross Sown Area (10.541%), and Fertilizer Consumption (10.286%). Annual Rainfall exhibited the lowest importance in the RF model, contributing 1.464%. The results indicate that historical production and the availability of irrigated agricultural land were particularly influential in determining the predicted foodgrain production.

The XGBoost model showed a somewhat different distribution of feature importance. Previous-Year Foodgrain Production was again the dominant predictor, accounting for 56.052% of the total importance. The second most influential variable was Net Sown Area, with an importance of 21.124%, followed by Fertilizer Consumption (9.646%), Gross Irrigated Area (6.015%), Net Irrigated Area (5.294%), and Gross Sown Area (1.809%). In contrast, Annual Rainfall contributed only 0.061% to the XGBoost prediction.

The consistent identification of Previous-Year Foodgrain Production as the most important predictor in both RF and XGBoost is a significant finding. Its importance was 34.180% in RF and increased to 56.052% in XGBoost. This indicates that historical production contains substantial information about subsequent foodgrain production at the state-year level. This relationship may reflect the persistence of agricultural production systems, including established cropping patterns, cultivated land availability, production capacity, irrigation infrastructure, and other relatively persistent state-level characteristics.

The importance of irrigation-related variables was more pronounced in the RF model. Gross Irrigated Area and Net Irrigated Area contributed 16.671% and 16.185%, respectively, giving a combined importance of approximately 32.856%. This suggests that irrigation availability plays an important role in explaining variations in foodgrain production. In XGBoost, the corresponding importance values were 6.015% and 5.294%, respectively. Although their individual contributions were lower than those obtained from RF, irrigation remained among the relevant predictors.

Similarly, the cultivated-area indicators demonstrated substantial importance. In RF, Net Sown Area and Gross Sown Area contributed 10.672% and 10.541%, respectively. In XGBoost, Net Sown Area had considerably greater importance (21.124%), whereas Gross Sown Area contributed only 1.809%. The difference between the two models indicates that the algorithms captured the relationships among cultivated-area variables differently. Nevertheless, the importance of Net Sown Area in both models confirms the relevance of the agricultural land base in predicting foodgrain production.

Fertilizer Consumption also made a meaningful contribution to the prediction. Its importance was 10.286% in RF and 9.646% in XGBoost, showing relatively consistent importance across the two algorithms. This consistency suggests that fertilizer consumption contains useful predictive information concerning variations in foodgrain production within the developed dataset.

In contrast, Annual Rainfall exhibited relatively low model-derived importance, with values of 1.464% for RF and only 0.061% for XGBoost. This result should not be interpreted as evidence that rainfall is unimportant to agriculture in general. Rather, it indicates that, within the variables and state-year observations used in the present machine-learning framework, rainfall contributed relatively little additional predictive information compared with historical production, cultivated area, irrigation, and fertilizer indicators. The low importance may also reflect interactions and redundancy among predictors captured by the ensemble models.

4. Conclusion

This study developed a machine-learning-based framework for predicting agricultural foodgrain production across Indian states using agricultural, irrigation, fertilizer, climatic, and historical production indicators. A state-year dataset comprising 600 observations, representing 25 Indian states over the period 2000–01 to 2023–24, was developed and analyzed. Two ensemble-learning algorithms, Random Forest (RF) and Extreme Gradient Boosting (XGBoost), were employed to establish the relationship between the selected predictor variables and foodgrain production.

The results demonstrated that both models provided excellent predictive performance on the independent testing dataset. RF achieved $R = 0.996298$, $R^2 = 0.987867$, $RMSE = 2003.8967$ thousand tonnes, $MAE = 957.6846$ thousand tonnes, $MAPE = 6.919\%$, and $SI = 0.148889$. XGBoost achieved $R = 0.996503$, $R^2 = 0.983717$, $RMSE = 2321.4478$ thousand tonnes, $MAE = 933.7023$ thousand tonnes, $MAPE = 11.312\%$, and $SI = 0.172483$. Although XGBoost produced a marginally higher correlation coefficient and lower MAE, RF demonstrated better overall testing performance through its higher R^2 and lower RMSE, MAPE, and SI. Accordingly, RF was identified as the preferred model for the prediction task.

The feature-importance analysis provided further insight into the variables contributing to foodgrain production prediction. Previous-Year Foodgrain Production was the most influential predictor in both models, with importance values of 34.180% for RF and 56.052% for XGBoost. Irrigation-related variables, particularly Gross Irrigated Area and Net Irrigated Area, also made substantial contributions in RF. Net Sown Area and Fertilizer Consumption were important predictors in both modelling approaches. In contrast, Annual Rainfall showed comparatively low model-derived importance, particularly in XGBoost. This finding indicates that, within the present dataset and selected predictors, historical production and agricultural resource indicators provided stronger predictive information than annual rainfall.

The findings demonstrate the potential of ensemble machine-learning techniques for modelling state-level agricultural foodgrain production in India. The developed framework can assist in identifying production patterns and generating data-driven production estimates from readily available agricultural indicators. In particular, the strong performance of RF suggests that it can serve as a useful predictive tool for agricultural planning and resource-management applications.

However, the results should be interpreted within the scope of the available variables and historical state-year dataset. Future studies could improve the framework by incorporating additional factors such as crop-specific production, temperature, soil characteristics, cropping intensity, groundwater availability, extreme weather events, pest and disease incidence, and socioeconomic indicators. External validation using future agricultural years would also provide a stronger assessment of the practical forecasting capability of the developed models.

REFERENCES

1. Directorate of Economics and Statistics (DES), Department of Agriculture and Farmers Welfare, Ministry of Agriculture and Farmers Welfare, Government of India: Agricultural Statistics at a Glance 2024. Government of India, New Delhi (2024).
2. Ministry of Agriculture and Farmers Welfare, Government of India: State-wise production of food grains from year 2018–19 to 2023–24. Government of India (2024).
3. Directorate of Economics and Statistics, Department of Agriculture and Farmers Welfare, Government of India: Land Use Statistics at a Glance 2023–24. Government of India, New Delhi (2025).
4. De Clercq, D., Mahdi, A.: Feasibility of machine learning-based rice yield prediction in India at the district level using climate reanalysis and remote sensing data. *Agric. Syst.* 220, 104099 (2024). <https://doi.org/10.1016/j.agry.2024.104099>
5. Jhajharia, K., Mathur, P.: Machine learning based crop yield prediction model in Rajasthan region of India. *Iraqi J. Sci.* 65(1), 390–400 (2024). <https://doi.org/10.24996/ijs.2024.65.1.32>
6. Jhajharia, K., Mathur, P., Jain, S., Nijhawan, S.: Crop yield prediction using machine learning and deep learning techniques. *Procedia Comput. Sci.* 218, 406–417 (2023). <https://doi.org/10.1016/j.procs.2023.01.023>
7. Vinson Joshua, S., Selwin Mich Priyadharson, A., Kannadasan, R., Khan, A.A., Lawanont, W., Khan, F.A., Ur Rehman, A., Ali, M.J.: Crop yield prediction using machine learning approaches on a wide spectrum. *Comput. Mater. Contin.* 72(3), 5663–5679 (2022). <https://doi.org/10.32604/cmc.2022.027178>
8. Jhajharia, K., Mathur, P.: Machine learning approaches to predict crop yield using integrated satellite and climate data. *Int. J. Ambient Comput. Intell.* 13(1), 1–17 (2022). <https://doi.org/10.4018/IJACI.300799>
9. Shinyclimensa, C.S., Parthiban, A.: Network-enhanced machine learning framework for multi crop yield prediction: a comprehensive analysis of Indian agricultural data. *Front. Agron.* 8, 1767878 (2026). <https://doi.org/10.3389/fagro.2026.1767878>
10. Breiman, L.: Random forests. *Mach. Learn.* 45(1), 5–32 (2001). <https://doi.org/10.1023/A:1010933404324>
11. Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794. ACM, San Francisco, CA, USA (2016). <https://doi.org/10.1145/2939672.2939785>