



ISSN Print: 2618-060X
 ISSN Online: 2618-0618
 NAAS Rating (2026): 5.20
 IJRA 2026; SP-9(10): 179-186
www.agronomyjournals.com
 Received: 14-08-2026
 Accepted: 24-09-2026
 Published: 08-10-2026

AC Aiwale

PG Scholar, Department of
 Agricultural Meteorology,
 College of Agriculture, Pune,
 Mahatma Phule Krishi
 Vidyapeeth, Rahuri,
 Maharashtra, India

VT Jadhav

Associate Professor of
 Agronomy, College of
 Agriculture, Pune, Mahatma
 Phule Krishi Vidyapeeth,
 Rahuri, Maharashtra, India

AA Bhagat

Assistant Professor of
 Statistics, Division of
 Economics, College of
 Agriculture, Pune, Mahatma
 Phule Krishi Vidyapeeth,
 Rahuri, Maharashtra, India

SK Upadhye

Senior scientist (SWCE),
 AICRPDA, ZARS, Solapur,
 Mahatma Phule Krishi
 Vidyapeeth, Rahuri,
 Maharashtra, India

SV Bagade

Assistant Professor of Agril.
 Meteorology, Department of
 Agril. Meteorology, College of
 Agriculture, Pune, Mahatma
 Phule Krishi Vidyapeeth,
 Rahuri, Maharashtra, India

PV Gaikwad

PG Scholar, Department of
 Agricultural Meteorology,
 College of Agriculture, Pune,
 Mahatma Phule Krishi
 Vidyapeeth, Rahuri,
 Maharashtra, India

Corresponding Author:

AC Aiwale

PG Scholar, Department of
 Agricultural Meteorology,
 College of Agriculture,
 Pune, Mahatma Phule
 Krishi Vidyapeeth, Rahuri,
 Maharashtra, India

Application of machine learning algorithms for prediction of area and production of soybean in Jalgaon district of Maharashtra state

AC Aiwale¹, VT Jadhav², AA Bhagat³, SK Upadhye⁴, SV Bagade⁵, PV Gaikwad⁶

DOI: <https://www.doi.org/10.33545/2618060X.2026.v9.i10Sc.6820>

ABSTRACT

Soybean (*Glycine max*) is an important oilseed crop in India, and its area and production are influenced by weather variability. The present study assessed the performance of machine learning algorithms for weather-based prediction of soybean area and production in Jalgaon district of Maharashtra. A 34-year dataset (1989–90 to 2022–23) comprising soybean area, production and weather parameters, including maximum and minimum temperature, rainfall, relative humidity, wind speed, bright sunshine hours and pan evaporation, was used. The data were obtained from secondary sources and analysed using Python. Ten machine learning algorithms, namely Linear Regression, Ridge Regression, LASSO, Elastic Net, K-Nearest Neighbour, Decision Tree, Random Forest, Extra Trees, Support Vector Regression and Gradient Boosting Regression, were evaluated using the coefficient of determination (R^2) and Root Mean Square Error (RMSE). The results indicated substantial variability in soybean area and production during the study period. Wind speed showed significant negative correlations with soybean area ($r = -0.84^{**}$) and production ($r = -0.72^{**}$), whereas evening relative humidity exhibited significant positive correlations with area ($r = 0.60^{**}$) and production ($r = 0.51^{**}$). Random Forest showed the highest predictive performance for soybean area ($R^2 = 0.9771$, RMSE = 17.54), while Extra Trees performed best for soybean production ($R^2 = 0.9680$, RMSE = 30.91). Forecasts from the selected models indicated an increasing soybean area from 242.09 hundred hectares in 2023 to 356.35 hundred hectares in 2027, while production was projected to fluctuate between 202.51 and 318.01 hundred tonnes during the forecast period.

Keywords: Soybean, Area, Production, Machine learning algorithms, Random Forest, Extra trees

Introduction

Soybean (*Glycine max*) is important oilseed and pulse crop cultivated worldwide. It is known as miracle crop or yellow jewel. Soybean is basically pulse crop but in India mostly grown as oilseed crop. It contains 42% protein, rich in lysine and 20% oil. Soybean originated in China and it was introduced to India centuries ago through Himalayan routes, and also brought in via Burma by traders from Indonesia. (Singh, 2006) [1]. In India soybean is grown in various agro-climatic zones including the tropical, sub-tropical and temperate regions. The major soybean growing states are Madhya Pradesh, Maharashtra, Rajasthan, Karnataka, Andhra Pradesh and Chhattisgarh (Dinesh K. Agarwal *et al.*, 2013) [2]. In India, Soybean occupies nearly 112 to 119 lakh hectares with production ranging from 130 to 151 lakh tonnes and productivity of about 983–1063 kg/ha. (UPAg, SOPA, Research studies, 2025) [3]. The first systematic attempts to develop improved varieties of soybean suitable for Indian environments were made in the early 1900s at the Pusa Agricultural Research Farm in Bihar State, and the work was extended to West Bengal, Uttar Pradesh, Orissa, Punjab, Delhi, Madhya Pradesh, Maharashtra, Tamil Nadu, and Rajasthan (Woodhouse *et al.*, 1913) [4].

In recent years, machine learning has emerged as an effective approach for agricultural forecasting and decision-making by analysing historical data and identifying complex relationships between weather parameters and crop performance. Various machine learning algorithms, including linear regression, decision tree, random forest, support vector regression and other ensemble techniques, have been increasingly applied to predict crop area and production with improved accuracy. Reliable weather-based prediction of soybean area and production can support farmers, researchers and policymakers in crop planning, resource allocation, storage management and market-related decisions. Although several studies have examined the influence of weather parameters on soybean growth and production, relatively limited attention has been given to predicting soybean area and production at the district level using multiple machine learning algorithms. Moreover, the comparative performance of different machine learning algorithms under the agro-climatic conditions of Jalgaon district has not been extensively evaluated. Therefore, assessing suitable machine learning algorithms for weather-based prediction of soybean area and production in Jalgaon district is important for im

proving the reliability of agricultural forecasting and supporting effective crop planning and decision-making.

Materials and methods

The study was conducted in Jalgaon district of Maharashtra State using secondary data. A total of 34 years of soybean data pertaining to the district were utilized for the study, covering the period from 1989–90 to 2022–23. Data on soybean area (ha) and production (tonnes) were obtained from the ICRISAT District Level Database and the Directorate of Economics and Statistics (DES), Department of Agriculture and Farmers Welfare, Government of India. Corresponding meteorological data were collected from the India Meteorological Department (IMD) and included maximum temperature, minimum temperature, rainfall, relative humidity, wind speed, bright sunshine hours and evaporation.

Python was used for data preprocessing, exploratory data analysis, correlation analysis, evaluation of machine learning algorithms, forecasting and visualization, while Microsoft Excel 2016 was used for data organization and preliminary data management. For soybean, separate analyses were carried out for area and production. The dataset was divided into 80 per cent training and 20 per cent testing data, and the evaluated machine learning algorithms were compared using R^2 and RMSE. The best-fit algorithm was identified based on comparative model performance.

The following machine learning algorithms were used

1. Linear regression

Linear Regression (LR) is a supervised learning method used to describe the linear relationship between a dependent variable and one or more independent variables. It estimates model coefficients by minimizing the differences between observed and predicted values and is widely used for prediction and estimation.

$$\hat{Y}_i = \beta_0 + \sum_{j=1}^p \beta_j X_{ij}$$

$$\min_{\beta} \left\{ \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \lambda \left[\alpha \sum_{j=1}^p |\beta_j| + (1 - \alpha) \sum_{j=1}^p \beta_j^2 \right] \right\}$$

where λ is the regularization parameter and α is the mixing parameter ranging from 0 to 1.

5. Extra trees

Extra Trees (Extremely Randomized Trees) is an ensemble learning algorithm that constructs multiple randomized decision trees and combines their predictions. It can capture complex and non-linear relationships between weather parameters and crop area or production.

$$\hat{Y} = \frac{1}{M} \sum_{m=1}^M T_m(X)$$

where M is the number of trees and $T_m(X)$ represents the prediction from the m^{th} tree.

6. K-nearest neighbour

K-Nearest Neighbour (KNN) is a non-parametric algorithm that predicts the response variable based on the values of the K most similar observations. The prediction is obtained by averaging the response values of the nearest neighbours.

$$\hat{Y}(x) = \frac{1}{K} \sum_{i \in N_K(x)} Y_i$$

where $\hat{Y}_i$ is the predicted value, β_0 is the intercept, β_j is the regression coefficient, X_{ij} is the j^{th} predictor, and p is the number of predictors.

2. Ridge regression

Ridge Regression is a regularized linear regression method that uses an L2 penalty to reduce the effect of multicollinearity and overfitting. It shrinks regression coefficients toward zero while retaining all predictors in the model.

$$\min_{\beta} \left\{ \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^p \beta_j^2 \right\}$$

where Y_i is the observed value, $\hat{Y}_i$ is the predicted value, and λ is the regularization parameter controlling the degree of coefficient shrinkage.

3. LASSO regression

Least Absolute Shrinkage and Selection Operator (LASSO) Regression is a regularized linear regression method used for prediction and variable selection. It applies an L1 penalty that can shrink less important regression coefficients to zero, thereby reducing model complexity.

$$\min_{\beta} \left\{ \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

where λ is the regularization parameter controlling the extent of coefficient shrinkage.

4. Elastic net regression

Elastic Net Regression combines the L1 and L2 regularization properties of LASSO and Ridge Regression. It is useful when predictor variables are correlated and variable selection is required.

where $N_K(x)$ represents the set of K nearest neighbours and Y_i is the observed value of the neighbouring observation.

7. Decision tree regression

Decision Tree Regression is a non-parametric supervised learning algorithm that recursively partitions the predictor space into homogeneous groups. It can model non-linear relationships between weather parameters and crop area or production without assuming a predefined functional form.

$$\hat{Y} = f(X)$$

where X represents the input weather parameters and $f(X)$ is the decision-tree-based prediction function.

8. Random forest

Random Forest (RF) is an ensemble learning algorithm that combines predictions from multiple decision trees to improve predictive performance and reduce overfitting. It can effectively capture non-linear relationships and interactions among weather parameters.

$$\hat{Y} = \frac{1}{B} \sum_{b=1}^B T_b(X)$$

where B is the number of trees and $T_b(X)$ is the prediction from the b^{th} decision tree.

9. Support vector regression

Support Vector Regression (SVR) is a supervised learning technique used for predicting continuous response variables. It identifies a function that fits the data within a specified epsilon-insensitive margin and can model non-linear relationships through kernel functions.

$$f(X) = \sum_{i=1}^n (\alpha_i - \alpha_i^*) K(X_i, X) + b$$

where $K(X_i, X)$ is the kernel function, α_i and α_i^* are Lagrange multipliers, and b is the bias term.

10. Gradient boosting regression

Gradient Boosting Regression is an ensemble learning technique that sequentially combines weak learners, generally decision trees, to minimize prediction errors. Each successive tree is fitted to the residual errors of the preceding model.

$$F_M(X) = F_0(X) + \sum_{m=1}^M \gamma_m h_m(X)$$

where $F_M(X)$ is the final prediction, $F_0(X)$ is the initial prediction, $h_m(X)$ is the m^{th} weak learner, γ_m is the learning rate, and M is the number of boosting iterations.

Goodness of fit

The predictive performance of the evaluated machine learning algorithms was assessed using Root Mean Square Error (RMSE) and the coefficient of determination (R^2). Higher R^2 and lower RMSE values

indicated better agreement between observed and predicted values.

1. Root mean square error

RMSE measures the magnitude of prediction errors between observed and predicted values and is expressed in the same units as the response variable.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - \hat{Y}_i)^2}$$

where Y_i is the observed value, $\hat{Y}_i$ is the predicted value, and n is the number of observations.

2. Coefficient of determination

The coefficient of determination (R^2) indicates the proportion of variation in the observed response explained by the model. A value closer to 1 indicates greater agreement between observed and predicted values.

$$R^2 = 1 - \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{\sum_{i=1}^n (Y_i - \bar{Y})^2}$$

where $\bar{Y}$ is the mean of the observed values.

Result and discussion

1. Descriptive statistics

Descriptive statistics were computed for soybean area, production and associated weather parameters during 1989–90 to 2022–23. The analysis included mean, median, standard deviation, minimum, maximum, skewness and kurtosis to characterize the central tendency, variability and distributional properties of the variables.

Table 1. Descriptive statistics for area and production of soybean (1989-90 to 2022-23)

Soybean	Mean	Median	S.D.	Max.	Min.	Skewness	Kurtosis
Area ('00 ha)	142.11	137.50	123.18	341.39	1.00	0.39	-1.43
Production ('00 t)	208.50	178.00	185.15	636.00	1.00	0.63	-0.69

The descriptive statistics of soybean area and production are presented in Table 1. The mean soybean area was 142.11 hundred hectares, with a median of 137.50 hundred hectares and a standard deviation of 123.18 hundred hectares. The area ranged from 1.00 to 341.39 hundred hectares, indicating substantial temporal variability in soybean cultivation. Soybean production recorded a mean of 208.50 hundred tonnes, a median of 178.00 hundred tonnes and a standard deviation of 185.15 hundred tonnes. Production varied from 1.00 to 636.00 hundred tonnes, indicating considerable year-to-year variation. Positive skewness was observed for both area (0.39) and production (0.63), suggesting moderately right-skewed distributions, with production exhibiting relatively greater asymmetry. The negative kurtosis values for area (-1.43) and production (-0.69) indicated relatively platykurtic distributions.

The descriptive statistics of the weather parameters are presented in Table 2. Maximum and minimum temperatures averaged 34.58 °C and 19.84 °C, respectively, with relatively low standard deviations of 0.36 °C and 0.70 °C, indicating limited temporal variability. Average morning and evening relative humidity were 69.39% and 43.13%, respectively. Average wind speed was 9.00 km h⁻¹, while bright sunshine hours and pan evaporation averaged 7.46 hr and 6.99 mm, respectively. Average annual rainfall was 783.58 mm, with a standard deviation of 231.58 mm, indicating comparatively greater interannual variability. Skewness values ranged from -0.34 to 0.56, suggesting approximately symmetric to moderately skewed distributions. Overall, rainfall exhibited the greatest variability among the weather parameters, whereas maximum temperature showed the least variability during the study period.

Table 2. Descriptive statistics for weather parameters of Soybean (1989-90 to 2022-23)

Weather parameters	Mean	Median	S.D.	Max.	Min.	Skewness	Kurtosis
Tmax (°C)	34.58	34.70	0.36	35.40	33.80	-0.31	-0.12
Tmin (°C)	19.84	20.00	0.70	21.00	18.30	-0.27	-0.23
RHI (%)	69.39	69.35	3.45	76.30	61.40	-0.06	0.45
RHII (%)	43.13	43.75	4.33	52.40	35.10	-0.10	-0.35
WS (km/hr)	9.00	9.25	3.00	16.30	4.20	0.56	0.41

(continued on next page)

Weather parameters	Mean	Median	S.D.	Max.	Min.	Skewness	Kurtosis
BSS (hr)	7.46	7.60	0.66	8.50	6.20	-0.34	-0.85
EPAN (mm)	6.99	6.90	0.64	8.60	5.90	0.39	-0.20
Rainfall (mm)	783.58	821.90	231.58	1401.80	365.90	0.32	0.40

Correlation analysis

Correlation analysis for area and production of soybean with weather parameters

Fig.1 indicated the correlation between area under soybean and weather parameters in Jalgaon district. The wind speed exhibited the significant negative correlation with area under soybean ($r = -0.84^{**}$), indicated that years with higher average wind speed were generally associated with a lower area under soybean cultivation. Similarly, bright sunshine hours ($r = -0.70^{**}$) also showed a significant negative correlation indicated that maximum sunshine duration declined significantly area under soybean during study period. In contrast, evening relative humidity (RHII) ($r = 0.60^{**}$) showed the significant positive correlation followed by pan evaporation ($r = 0.32^{*}$) with area under soybean. Maximum temperature ($r = 0.12$), minimum temperature ($r = 0.18$) and minimum relative humidity ($r = 0.20$) exhibited non-significant positive correlations, whereas rainfall ($r = -0.10$) showed a non-significant negative correlation indicated no correlation with area under soybean during study period. Similar results were reported by Bhagat *et al.*, (2020) [5] on green gram in Jalgaon district, where high variability in wind speed (58.4%) and bright sunshine hours (43.9%) was observed, which indicated that important role of wind speed and bright sunshine hours in influencing area and production of green gram in Jalgaon district of Maharashtra state.

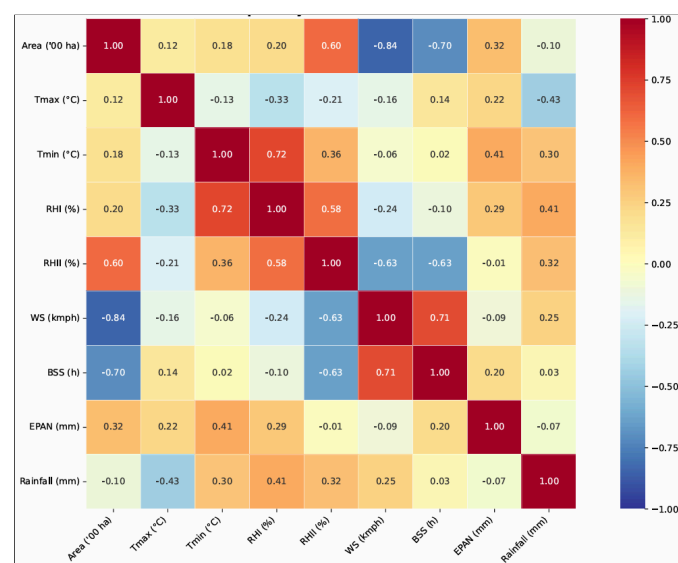


Figure 1. Correlation heatmap for area under soybean with weather parameters

Fig.2 indicated the correlation between production under soybean and weather parameters in Jalgaon district. The wind speed ($r = -0.72^{**}$) exhibited the significant negative correlation with production of soybean followed by bright sunshine hours (-0.57^{**}), indicated that years with higher wind speed and longer sunshine duration were generally associated with lower soybean production. Among the positively correlated weather parameters, evening relative humidity (RHII) ($r = 0.51^{**}$) showed the significant positive correlation with soybean production indicated a moderate positive association between them. The maximum temperature ($r = 0.29$) and pan evaporation ($r = 0.22$) exhibited non-significant positive correlations with production. Minimum temperature ($r = -0.10$), morning relative humidity (RHI) ($r = -0.08$) and rainfall ($r = -0.08$) exhibited non-significant negative correlations with soybean production. Similar results were reported by Salau *et al.*, (2016) [6] in a study on banana production in Ondo

state of Nigeria where weather parameters significantly influenced crop productivity, highlighted the important role of weather parameters on crop area and production.

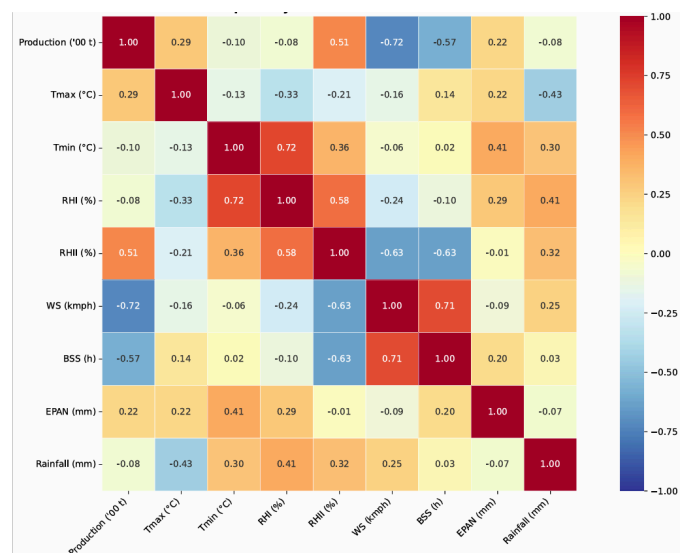


Figure 2. Correlation heatmap for production under soybean with weather parameters

Identification of weather-based prediction model

Different machine learning models were evaluated for predicting the area and production of soybean in Jalgaon district of Maharashtra. The crop area and production were considered as dependent variables, while the selected weather parameters were used as independent variables. The performance of each model was assessed using the coefficient of determination (R^2) and root mean square error (RMSE). A higher R^2 value, closer to 1, and a lower RMSE value, closer to 0, indicated better predictive performance of the model.

Table 3. Results for comparative analysis of Machine Learning models for area of soybean in Jalgaon district.

Model	R^2	RMSE
Random forest	0.9771	17.54
Extra trees	0.9709	19.78
Gradient boosting	0.9016	25.36
Elastic net	0.8948	37.57
Ridge regression	0.8873	38.89
Decision tree	0.8059	51.84
LASSO	0.7834	53.92
Linear regression	0.7666	55.97
KNN	0.1733	98.66
SVR	-0.1044	121.75

The results of comparative performance of machine learning models for area of soybean are presented in table 3. Random forest showed the best predictive performance, achieving the highest R^2 (0.9771) and the minimum RMSE (17.54), indicated that the model explained approximately 97.71% of the variation in soybean area with the lowest prediction error. Extra trees ranked second, with R^2 (0.9709) and RMSE (19.78), also indicated high level of predictive performance. Gradient boosting and Elastic net performed better than ridge regression and Decision tree, with R^2 (0.9016 and 0.8948) and RMSE (2.36 and 37.57). Ridge regression and Decision tree showed comparatively

good performance with R^2 (0.8873 and 0.8059) and RMSE (38.89 and 51.84), while LASSO and linear regression recorded lower predictive performance with R^2 (0.7834 and 0.7666) respectively. KNN showed poorer performance with R^2 (0.1733) and RMSE (98.66). SVR exhibited the weaker predictive performance with negative R^2 (-0.1044) and the maximum RMSE (121.75), indicating limited predictive capability for area under soybean. Similar results were reported by Sharma *et al.* (2023), who found that Random Forest performed better than the other evaluated machine learning models for agricultural yield prediction in India, achieving the highest prediction accuracy of 98.96%.

Table 4. Results for comparative analysis of Machine Learning models for production of soybean in Jalgaon district.

Model	R^2	RMSE
Extra trees	0.9680	30.91
Random forest	0.9432	43.22
Gradient boosting	0.8033	80.69
Linear regression	0.7966	81.79
LASSO	0.7953	82.05
Ridge regression	0.7933	82.44
Elastic net	0.7365	93.10
Decision tree	0.7061	96.35
KNN	0.2011	163.03
SVR	-0.3281	209.00

The results of the comparative performance of different machine learning models of production under soybean prediction is presented in table 4. The results indicated that Extra trees performed best, achieved the highest R^2 (0.9680) and the minimum RMSE (30.91). This indicated that the extra trees model explained approximately 96.8% of the variability in the observed values and produced rela-

tively small prediction error. The random forest model ranked second with, R^2 (0.9432) and RMSE (43.22), indicated that it also provided a strong predictive performance. However, its error was higher than that of extra trees. The Gradient boosting model ranked third, with R^2 (0.8033) and RMSE (80.69), indicated good predictive performance but lower accuracy than Extra trees and Random Forest. Linear regression showed the best performance with, R^2 (0.7966) and RMSE (81.79), followed closely by LASSO (R^2 = 0.7953, RMSE = 82.05) and ridge regression (R^2 = 0.7933, RMSE = 82.44). Elastic net showed comparatively lower performance with, R^2 of 0.7365 and RMSE of 93.10. The Decision tree showed lower predictive performance, with R^2 (0.7061) and RMSE (96.35). The KNN model performed poorly, with R^2 (0.2011) and RMSE (163.03). The SVR model exhibited the poorest predictive performance with a negative R^2 (-0.3281) and the maximum RMSE (209.00).

Validation of best-fit machine learning models

The performance of the best-fit machine learning models was validated by comparing the actual and predicted values of soybean area and production using the training and testing datasets. The dataset was divided into training and testing sets to ensure that both datasets represented the overall data adequately. Furthermore, the validated models were used to forecast soybean area and production for the five-year period from 2023 to 2027. The forecasted values provide an estimate of future soybean area and production based on historical weather conditions and observed patterns.

Validation and forecasting of area under soybean

The original soybean dataset comprised 34 years of observations. After the creation of two lag features, the first two years (1989–90 and 1990–91) could not be assigned lag values because sufficient preceding-year observations were not available. Therefore, these two observations were excluded. The remaining 32 observations were randomly divided into training and testing datasets, with 25 years observations used for training and 7 years observations used for testing.

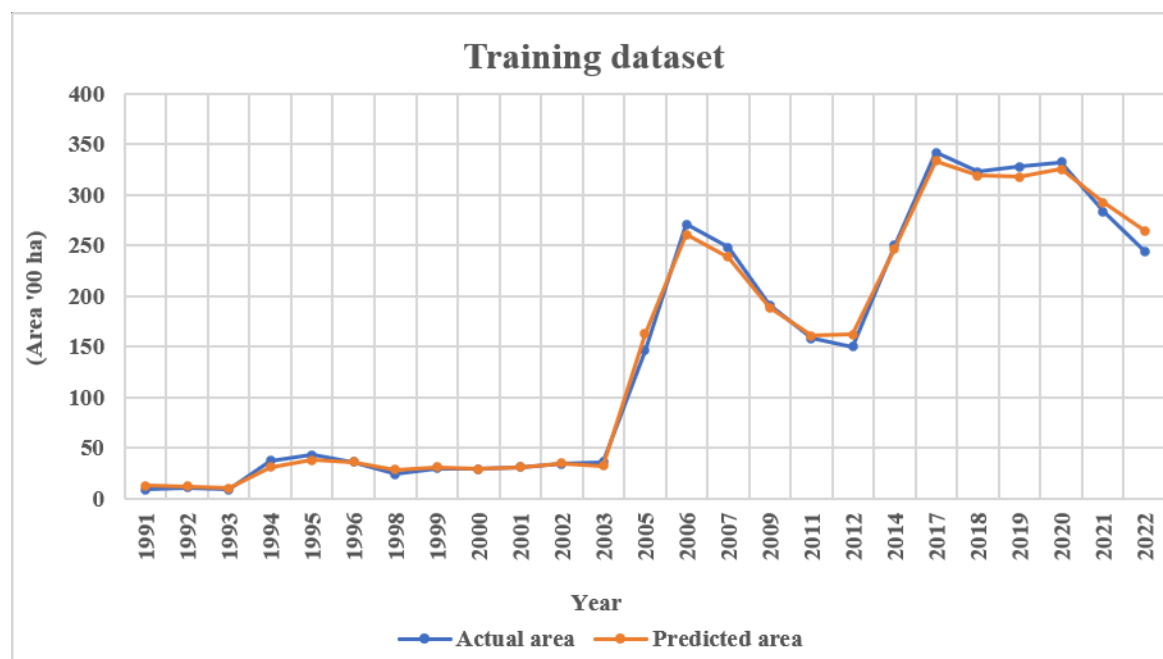


Figure 3. Actual and predicted values of area under Soybean in Jalgaon district using random forest (Training dataset)

Figure 4. represented the testing performance of Random Forest model by comparing the actual and predicted values of area under soybean in Jalgaon district during the study period. The soybean area showed an increased trend from 1997 to 2008, increased from 33 hundred hectares to 181 hundred hectares. The area subsequently declined during 2008 to 2010, followed by a gradual increased up to 2013. A sharp increase in area of soybean was observed between 2013

to 2015 after which the area continued to increased slightly and reached about 338 hundred hectares in the year 2016. From fig it was observed that actual and predicted values are close to each other. It means that Random Forest model showed good predictive accuracy and reliability for estimating the production under soybean in Jalgaon district.

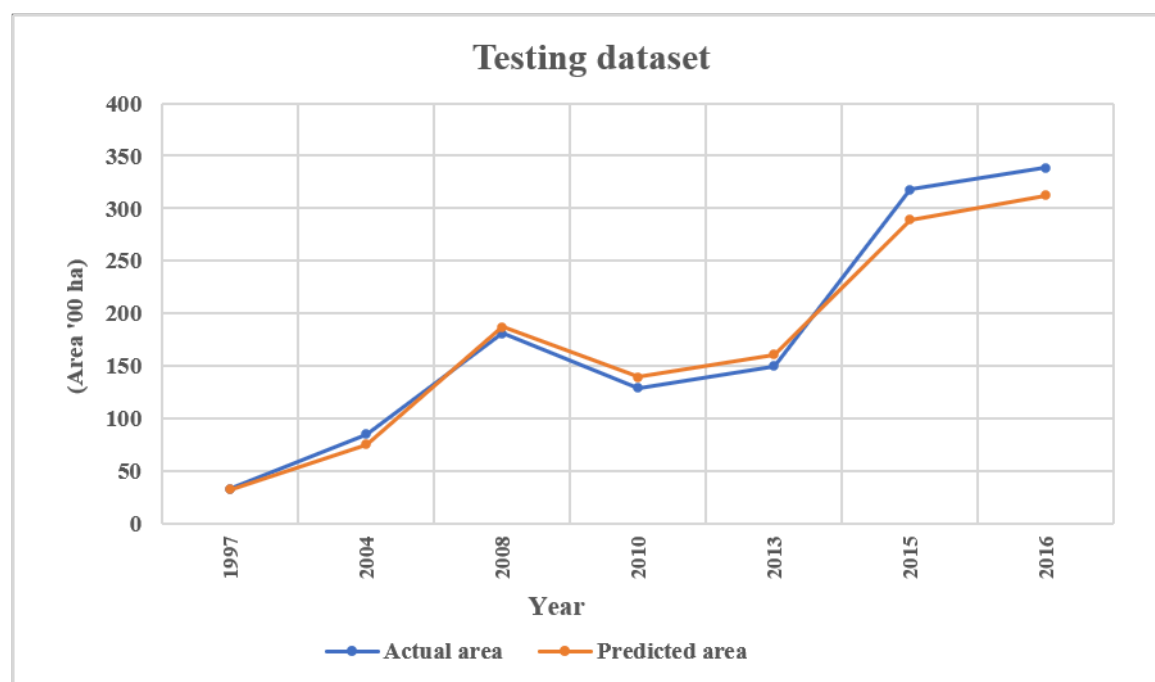


Figure 3. represented the training performance for Random Forest model by comparing the actual and predicted values of area under soybean in Jalgaon district during the study period. The soybean area remained relatively low during the years from 1991 to 2003, followed by a sharp increased from the period 2003 to 2006. Subsequently, the area gradually declined until 2012 and then showed a significant increased from 2014 that was 340 hundred hectares. During 2020 to 2022 both actual and predicted values exhibited a declining trend.

Figure 4. Actual and predicted values of area under Soybean in Jalgaon district using random forest (Testing dataset)

Table 5. Forecasted values for area under soybean in Jalgaon district using Random Forest (2023-2027)

Year	Forecasted area of soybean ('00 ha)
2023	242.09
2024	270.72
2025	309.47
2026	340.17
2027	356.35

Fig. 5 represented the actual, predicted and forecasted area of soybean in Jalgaon district for the period 2023-2027 using the random forest model. The forecast indicated a consistent upward trend in area under soybean throughout the forecast period. The soybean area will be expected to increase from 242.09 hundred hectares in 2023 to 356.3 hundred hectares in 2027, indicated a steady expansion over the five-year forecast period. These findings demonstrated the effectiveness of the Random forest model in forecasting future soybean area and provide valuable information for agricultural planning, policy making and resource management.

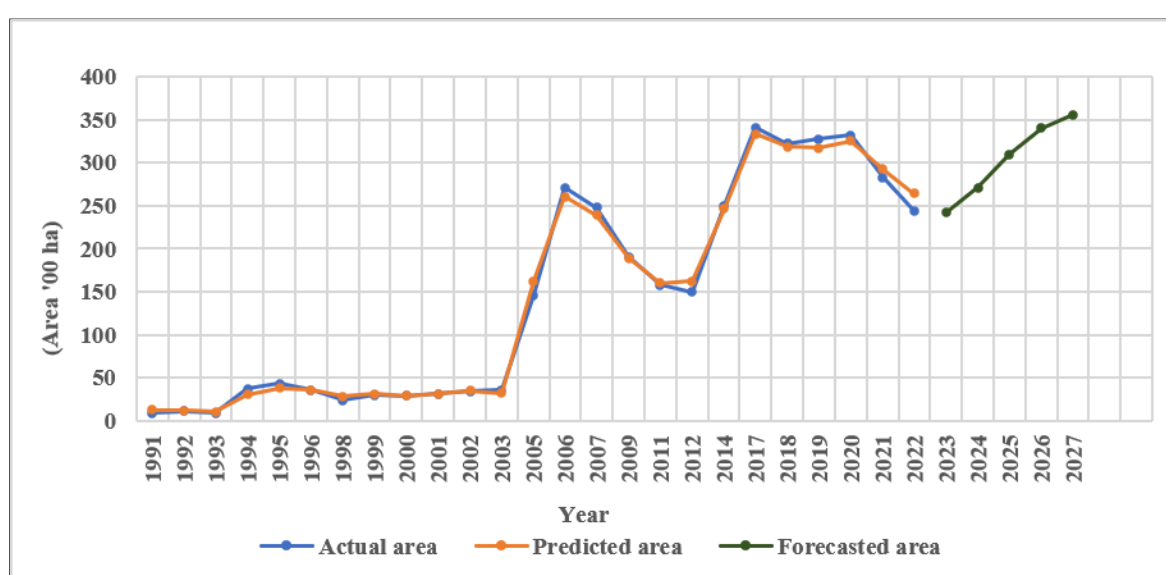


Figure 5. Actual, predicted and forecasted area under soybean in Jalgaon district using Random Forest model

Validation and forecasting of production under soybean

Figure 6. represented the training performance of Extra Trees model by comparing the actual and predicted values of production under soybean in Jalgaon district during the study period. The soybean pro-

duction remained relatively low during 1991 to 2003, with production generally below 100 hundred tonnes. A sharp increase in production was observed between 2003 and 2006, reaching approximately 513 hundred tonnes in 2006. Thereafter, production gradually declined until 2012, followed by an increasing trend from 2014 onwards, reach

ing the highest production of approximately 560 hundred tonnes in 2017. Subsequently, production declined to approximately 180 hundred tonnes in 2019, increased to about 366 hundred tonnes in 2020,

decreased to around 161 hundred tonnes in 2021 and recovered to nearly 355 hundred tonnes in 2022.

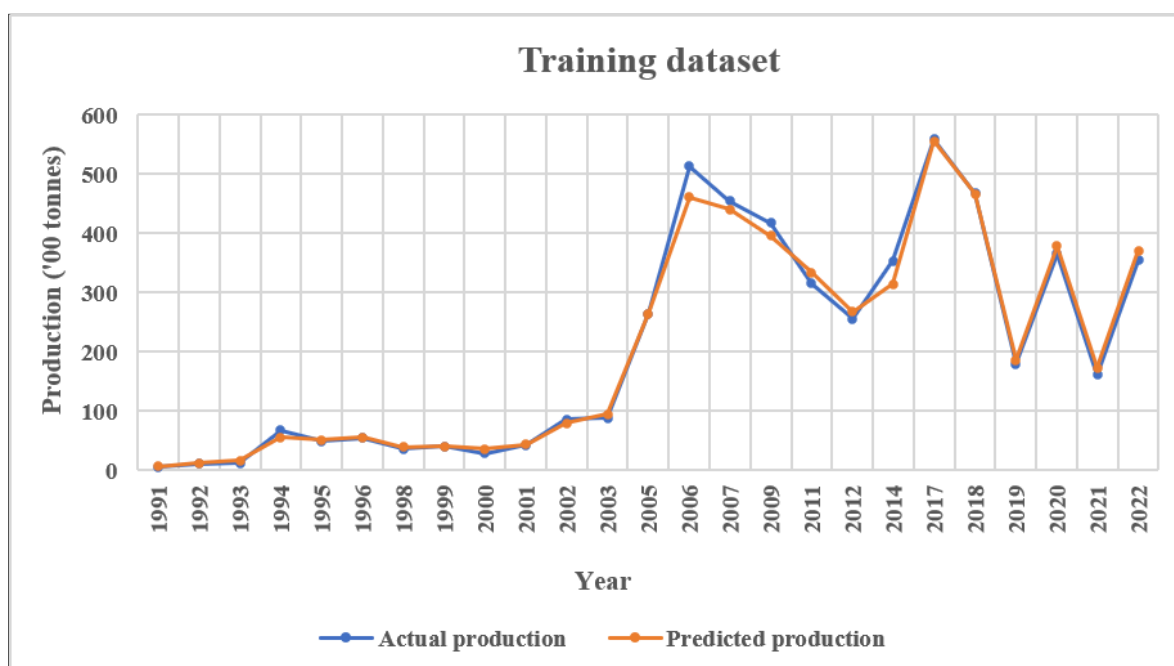


Figure 6. Actual and predicted values of production under Soybean in Jalgaon district using Extra trees model (Training dataset)

Figure 7. represented the testing performance of Extra trees model by comparing actual and predicted value of production under soybean in Jalgaon district. The observed production of soybean showed an increased trend from 1997 to 2008 raised from 39 hundred tonnes to 338 hundred tonnes. Production subsequently declined in 2010, remained almost stable up to 2013, and further decreased to approxi-

mately 176 hundred tonnes in 2015. A substantial increase was observed in 2016 when production reached approximately 636 hundred tonnes. From fig it was observed that actual and predicted values are close to each other. It means that Extra trees model showed good predictive accuracy and reliability for estimating the production under soybean in Jalgaon district.

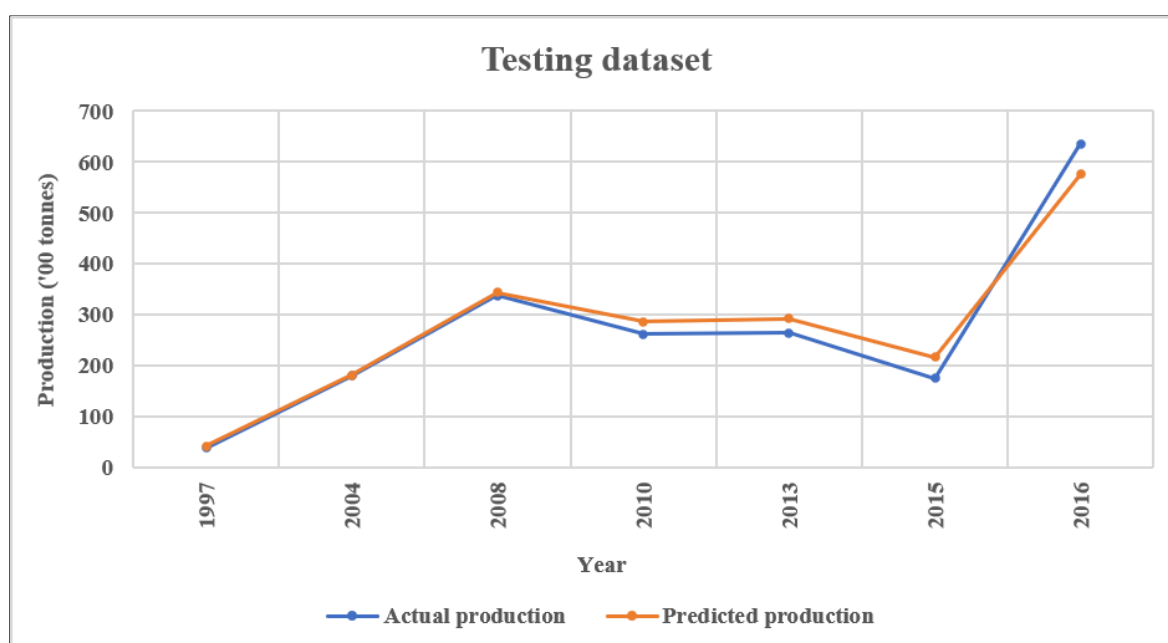


Figure 7. Actual and predicted values of production under Soybean in Jalgaon district using Extra trees model (Testing dataset)

Table 6. Forecasted values for production under soybean in Jalgaon district using Extra trees model (2023 to 2027)

Year	Forecasted production of soybean ('00 tonnes)
2023	202.51
2024	318.01
2025	259.41

2026	277.48
2027	268.74

Figure 8. represented the actual, predicted and forecasted production of soybean in Jalgaon district for the period 2023-27 using extra trees. The forecast indicated that production under soybean will be expected to fluctuate during the forecast period rather than exhibit a continues increasing trend. The forecasted production will be expected 202.51

hundred tonnes in 2023, followed by an increase to 318.01 hundred tonnes in 2024, represented the highest projected production during the forecasted period. Thereafter, production will be expected to decline to 259.41 hundred tonnes in 2025, recover slightly to 277.48

hundred tonnes in 2026, and decline to 268.74 hundred tonnes in 2027. These projected fluctuations suggest that future soybean production in Jalgaon district may continue to vary under changing weather conditions

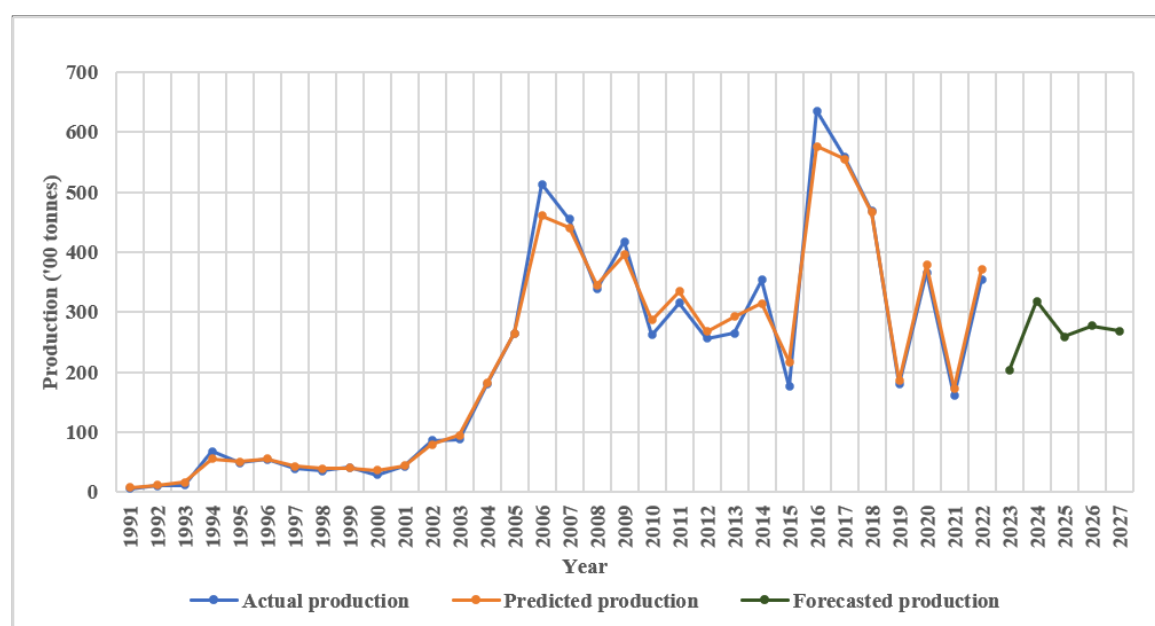


Figure 8. Actual, predicted and forecasted production under soybean in Jalgaon district using Extra trees model

Conclusions

1. The study revealed considerable variation in soybean area and production during 1989–90 to 2022–23. Wind speed and bright sunshine hours showed significant negative associations with soybean area ($r = -0.84^{**}$ and -0.70^{**} , respectively) and production ($r = -0.72^{**}$ and -0.57^{**} , respectively), whereas evening relative humidity showed significant positive associations with area ($r = 0.60^{**}$) and production ($r = 0.51^{**}$).
2. Among the evaluated machine learning algorithms, Random Forest performed best for predicting soybean area ($R^2 = 0.9771$, RMSE = 17.54), while Extra Trees performed best for soybean production ($R^2 = 0.9680$, RMSE = 30.91). The selected models showed comparatively higher R^2 and lower RMSE, indicating satisfactory predictive performance.
3. The selected machine learning models were validated using training and testing datasets, and the close agreement between observed and predicted values indicated satisfactory predictive performance. The Random Forest model forecasted a gradual increase in soybean area from 242.09 hundred ha in 2023 to 356.35 hundred ha in 2027. In contrast, the Extra Trees model projected fluctuations in soybean production, with values ranging from 202.51 to 318.01 hundred tonnes during 2023–2027.

Acknowledgements

The Department Agricultural Meteorology, College of Agriculture Pune and Department of Agricultural Economics and Statistics, College of Agriculture, Pune, Mahatma Phule Krishi Vidyapeeth, Rahuri, provided research facilities to perform this experiment and statistical analysis, so we are thankful to the departments.

References

1. Singh BB. Success of soybean in India: the early challenges and pioneer promoters. *Asian Agri-History*. 2006;10(1):43-53.
2. Agarwal DK, Billore SD, Sharma AN, Dupare BU, Srivastava SK. Soybean: introduction, improvement, and utilization in India problems and prospects. *Agricultural Research*. 2013;2(4):293-300.
3. Soybean Processors Association of India (SOPA). Soybean crop estimates and agricultural statistics, 2025. 2025.
4. Woodhouse EJ, Taylor CS. The varieties of soybeans found in Bengal, Bihar and Orissa and their commercial possibilities. *Memoirs of the Department of Agriculture in India, Botanical Series*. 1913;5(103-176).
5. Bhagat AA, Bhoge RS, Badgujar HC. Variability analysis and impact of weather parameters on productivity of green gram in Jalgaon district of Maharashtra. *International Journal of Farm Sciences*. 2020;10(2):23-27.
6. Salau OR, Momoh M, Olaleye OA, Owolaye RS. Effects of changes in temperature, rainfall and relative humidity on banana production in Ondo State, Nigeria. *World Scientific News*. 2016;44(143-154).
7. Unified Portal for Agricultural Statistics (UPAg). Agricultural statistics and crop estimates. Ministry of Agriculture and Farmers Welfare, Government of India. 2025.

How to cite this article: Aiwale A, Jadhav V, Bhagat A, Upadhye S, Bagade S, Gaikwad P. Application of machine learning algorithms for prediction of area and production of soybean in Jalgaon district of Maharashtra state. *International Journal of Research in Agronomy*. 2026;SP-9(10):179-186. doi: <https://www.doi.org/10.33545/2618060X.2026.v9.i10Sc.6820>