Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Data Analysis of Indian Agricultural Crop Yield Using Season, Area, and Production Through Machine Learning Models

Dr. G. Suresh

Assistant Professor, Department of Computer Science, Government Arts College for Women, Salem, (Deputed from Dept. of Computer and Information Science, Annamalai University, Annamalainagar-608 002) Tamil Nadu, India.
Email: suresh2023phd@gmail.com

Crop yield in Indian agriculture varies significantly based on several factors such as season, area, production techniques, and yield per hectare. General insights into crop yields in India like Seasonal Variations. India has three major cropping seasons namely Kharif Season, Rabi Season, and Zaid Season. Machine learning provides a powerful tool to predict and analyze crop yields based on seasonal, regional, and production factors. The choice of model depends on the complexity of the data and the intended application. With accurate forecasts, farmers, policymakers and agricultural planners can better manage resources, increase productivity and minimize the risks associated with unpredictable climate conditions. This paper considers Indian crop production dataset like state, district, crop, crop_year, season, area, production, and yield. The machine learning approaches are used to analyze and predict the dataset using linear regression, multilayer perceptron, random forest, random tree, and REP tree. Numerical illustrations are provided to prove the proposed results with test statistics or accuracy parameters.

Keywords: Machine learning, crop yields, decision tree, correlation coefficient, and test statistics.

1. Introduction

Crop yield in Indian agriculture varies significantly and depends on several factors such as season, area, production techniques and yield per hectare. Here are some general insights into crop yields in India: Seasonal variations: India has three main growing seasons: Kharif season: June to October, including grains like rice, maize, millet, etc. Rabi season: November to March, including grains like wheat, barley, mustard, etc. Zaid season: Shorter duration crops grown between March and June, such as fruits, vegetables, etc.

Area and production: Crop yield varies depending on the region due to differences in climate, soil type, irrigation facilities, etc. and agricultural practices. States like Punjab, Haryana, Uttar Pradesh and Maharashtra are known for their high agricultural production due to fertile soils and extensive irrigation. Yield per hectare: Yield per hectare depends on the crop type and region. For example, wheat yields can range between 2 and 4 tons per hectare, while rice yields can vary between 2 and 3 tons per hectare depending on factors such as irrigation, fertilization and pest control practices. Government Initiatives and Policies: The Government of India plays an important role in promoting agricultural productivity through various schemes, subsidies and support for irrigation and technology adoption. Specific data on crop yields in India may be found in the agricultural statistics of the Ministry of Agriculture and Farmers' Welfare or respective state agriculture ministries, as they regularly publish detailed reports on crop production, area under cultivation and yields across different regions and seasons.

Utilizing machine learning and data mining in this research can enhance agricultural practices, boost crop yields, foster sustainability, and increase the efficiency of farming in India. Furthermore, it can assist farmers and policymakers in making well-informed decisions. Research in data mining has the potential to yield valuable discoveries, enhance decision-making, and foster a deeper comprehension of intricate phenomena across diverse domains. It entails an iterative journey, frequently requiring refinement and revisitation of various stages to attain significant outcomes. Machine learning research extends across diverse fields, encompassing areas like computer vision, natural language processing, healthcare, finance, and beyond. This research frequently entails experimentation, rigorous testing, and iterative processes to pioneer groundbreaking algorithms and applications, expanding the frontiers of machine capabilities and knowledge.

Authors suggest introduces a system designed to forecast crop yields based on historical data. The approach involves the utilization of machine learning algorithms such as Support Vector Machine and Random Forest on agricultural datasets, enabling the recommendation of appropriate fertilizers tailored to specific crops. The primary objective of this research is the development of a predictive model that can be employed for future crop yield predictions. Additionally, the paper offers a concise analysis of crop yield forecasting using machine learning methodologies [1].

Machine learning techniques to forecast the yields of four widely cultivated crops across India. Once the crop yield predictions are made for specific locations, flexible adjustments in fertilizer application can be tailored to match the anticipated crop and soil requirements. The research focuses on the utilization of machine learning methodologies to construct a trained model that identifies patterns within the data, facilitating crop predictions. The study specifically addresses the prediction of the most commonly cultivated crops in India, which encompass Maize, Potatoes, Rice (Paddy), and Wheat [2].

The application of machine learning to classify soils into hydrologic groups. Leveraging characteristics like sand, silt, clay percentages, and saturated hydraulic conductivity values, our machine learning models were trained to categorize soil into four distinct hydrologic groups. We compared the classification results obtained from various algorithms, including k-Nearest Neighbors, Support Vector Machine with Gaussian Kernel, Decision Trees, Classification Bagged Ensembles, and TreeBagger (Random Forest), against those derived from traditional soil texture- based estimation. The models' performance was assessed and compared using per-class metrics as well as micro- and macro-averages. Overall, kNN, Decision Tree, and TreeBagger outperformed SVM-Gaussian Kernel and Classification Bagged Ensemble in terms of performance metrics. Interestingly, among the four hydrologic groups, Group B exhibited the highest rate of false positives [3].

A model designed to assess soil fertility, facilitate the optimal selection of crop seeds for fertile soil, and predict crop yields based on varying soil characteristics. The model's predictions can inform recommendations for crops that are likely to thrive. We leverage a range of Machine Learning algorithms, including Support Vector Machine (SVM), Random Forest, Naive Bayes, Linear Regression, Multilayer Perceptron (MLP), and Artificial Neural Networks (ANN), for soil classification and crop yield predictions. Our test results demonstrate that the proposed ANN approach adopts a deep learning architecture with multiple interconnected layers for enhanced accuracy, surpassing the performance of numerous existing methods [4].

Conducts an examination of various parameters found in the literature that are employed to describe soil characteristics and their utilization as inputs for machine learning algorithms in the prediction of soil fertility. This investigation reveals that employing optimized soil parameters can enhance the accuracy of soil fertility predictions and reduce the need for extensive human intervention, thereby enabling more efficient prediction techniques [5].

Data mining is a valuable tool for the practice of examining large pre-existing databases to generate previously unknown helpful information; in this paper, the input for the weather data set denotes specific days as a row, attributes denote weather conditions on the given day, and the class indicates whether the conditions are conducive to playing golf. Attributes include Outlook, Temperature, Humidity, Windy, and Boolean Play Golf class variables. All the data are considered for training purpose, and it is used in the seven-classification algorithm likes J48, Random Tree (RT), Decision Stump (DS), Logistic Model Tree (LMT), Hoeffding Tree (HT), Reduce Error Pruning (REP) and Random Forest (RF) are used to measure the accuracy. Out of seven

classification algorithms, the Random tree algorithm outperforms other algorithms by yielding an accuracy of 85.714% [6].

A framework for predicting the absolute Crop Growth Rate (CGR) in hydroponic tomato cultivation using machine learning techniques. Key input variables, such as Electric Conductivity (EC) limits, Nutrient Solution (NS), ion concentration uptake, and dry fruit weight, play pivotal roles in ensuring the successful growth of hydroponic tomato crops. Our study reveals both positive and negative correlations between growth parameters, fruit dry weights, and the absolute CGR of the plants. We analyze the dynamics of nutrient ion uptake, including Na, K, Mg, N, and Ca, throughout the tomato fruit growth process and investigate their impact on the target variable, absolute growth. This correlation analysis enables us to identify the critical variables influencing CGR, providing valuable insights into the optimal nutrient supply for robust crop growth and development. The proposed system design offers an intelligent and efficient approach for predicting and achieving optimal absolute CGR, while also aiding in the estimation of the ideal values for essential parameters to ensure high-quality crop yields [7].

The authors introduce an IoT monitoring sensor board designed for horticulture, aimed at establishing an IoT framework within the agricultural industry to monitor soil micro and macronutrients and analyze various soil parameters in Thiruvarur District, Tamil Nadu. This framework facilitates data-driven decision-making by collecting information from IoT sensors and storing it on a server for subsequent analysis using machine learning (ML) algorithms. The ML model categorizes the dataset based on micro and macronutrient threshold values obtained from the National Food Security Mission (NFSM). To assess the effectiveness of this classification, various ML algorithms, including Naive Bayes (NB), Logistic Regression (LR), Random Tree (RT), and K-Nearest Neighborhood (KNN), are employed. The performance of these classifiers is evaluated using metrics such as accuracy, Relative Absolute Error (RAE), Root Mean Square Error (RMSE), Root Relative Squared Error (RRSE), and Mean Absolute Error (MAE). The KNN classifier stands out with lower MAE, RMSE, and RRSE values of 0.2398, 0.3908, and 94.1845, respectively, demonstrating superior performance compared to the other classifiers. However, the RT algorithm achieves a lower RAE value of 66.24 when compared to KNN [8].

Authors suggest an innovative approach for assessing leaf nutrient levels in citrus trees using unmanned aerial vehicles (UAVs) equipped with multispectral imagery and artificial intelligence (AI). The study was conducted across four separate citrus field trials located in

Highlands County and Polk County, Florida, USA. Each trial consisted of 'Hamlin' or 'Valencia' sweet orange scions grafted onto over 30 different rootstocks. Laboratory analysis of collected leaves determined macro- and micronutrient concentrations through traditional chemical methods. Spectral data from tree canopies were captured across five distinct wavelength bands (red, green, blue, red edge, and near-infrared) using a UAV fitted with a multispectral camera. An estimation model was developed utilizing gradient boosting regression trees and evaluated using various metrics, including mean absolute percentage error (MAPE), root mean square error, MAPE- coefficient of variance (CV) ratio, and difference plots. This innovative model achieved high precision in determining macronutrients (nitrogen, phosphorus, potassium, magnesium, calcium, and sulfur) with average errors of less than 9% and 17% for the 'Hamlin' and 'Valencia' trials, respectively, while exhibiting moderate precision in determining micronutrients (average errors of less than 16% and 30% for 'Hamlin' and 'Valencia' trials, respectively). Overall, this UAV and AI-based methodology demonstrated efficiency in assessing nutrient concentrations and generating nutrient maps within commercial citrus orchards, with potential applications in other crop species [9].

Data mining is discovering hiding information that efficiently utilizes the prediction by stochastic sensing concept. This paper proposes an efficient assessment of groundwater level, rainfall, population, food grains, and enterprises dataset by adopting stochastic modeling and data mining approaches. Firstly, the novel data assimilation analysis is proposed to predict the groundwater level effectively. Experimental results are done, and the various expected groundwater level estimations indicate the sternness of the approach [10] and [11].

The input for the chronic disease data denotes a specific location as a row; attributes denote topics, questions, data values, low confidence limit, and high confidence limit. All the data are considered for training and testing using five classification algorithms. In this paper, the authors present the various analysis and accuracy of five different decision tree algorithms; the M5P decision tree approach is the best algorithm to build the model compared with other decision tree approaches [12].

2. Backgrounds and Methodologies

CART is a general-purpose, user-friendly machine learning model that can be applied to regression and classification problems alike. When it comes to predicting crop yields in Indian agriculture, CART can be used to make interpretable decisions by helping to comprehend the intricate relationships that exist between input factors such as area, season, and yield. It is frequently used as a foundational model, though, with larger and more complicated datasets performing better with more sophisticated ensemble approaches [13].

2.1 Linear Regression

By identifying the best straight line to fit the data points, a statistical method known as linear regression is used to understand and predict the relationship between two variables. For the purpose of making predictions and identifying trends, it helps determine the relationship between changes in one variable and changes in another. To find the best-fitting straight line, or "regression line," from a scatterplot of data points is the fundamental idea behind linear regression. A linear equation of the following form is represented by this line.

$$y = mx+b \qquad \dots (1)$$

Where:

□ y is the dependent variable (the one you want to predict or explain).

□ x is the independent variable (the one you're using to make predictions or explanations).

□ m is the slope of the line, representing how much

□ y changes for a unit change in x.

b is the y-intercept, indicating the value of y when x is 0.

2.2 Multilayer Perception

A Multilayer Perceptron (MLP) is an artificial neural network consisting of multiple layers of interconnected nodes or neurons. It's a fundamental architecture in deep learning and is used for various tasks, including classification, regression, and more complex tasks like image recognition and natural language processing. The architecture of an MLP typically includes three types of layers:

i. Input Layer: This layer consists of neurons receiving input data. Each neuron corresponds to a feature in the input data, and the values of these neurons pass through the network.

ii. Hidden Layers: These layers come after the input layer and precede the output layer. They are called "hidden" because their activations are not directly observed in the final output.

iii. Output Layer: This layer produces the network's final output. The number of neurons in the output layer depends on the problem type.

2.3 Random Forest

Random Forest is a popular machine learning ensemble method for classification and regression tasks. It is an extension of decision trees and is known for its high accuracy, robustness, and ability to handle complex datasets. Random Forest is widely used in various domains, including data science, machine learning, and pattern recognition. The main idea behind Random Forest is to create an ensemble (a collection) of decision trees and combine their predictions to make more accurate and stable predictions.

Steps involved in Random Forest

Random Forest is an ensemble learning method combining multiple decision trees to make more accurate and robust predictions for classification and regression tasks. The steps involved in building a Random Forest are as follows:

Step 1. Data Bootstrapping

Step 2. Random Feature Subset Selection Step 3. Decision Tree Construction

Step 4. Ensemble of Decision Trees Step 5. Out-of-Bag (OOB) Evaluation

Step 6. Hyperparameter Tuning (optional)

2.4 Random Tree

In machine learning, a Random Tree is a specific type of decision tree variant that introduces randomness during construction. Random Trees are similar to traditional decision trees but differ in how they select the splitting features and thresholds at each node. The primary goal of introducing randomness is to create a more diverse set of decision trees, which can help reduce overfitting and improve the model's generalization performance. Random Trees are commonly used as building blocks in ensemble methods like Random Forests.

Steps involved in Random Tree

Step 1. Data Bootstrapping:

Step 2. Random Subset Selection for Features:

Step 3. Decision Tree Construction:

Step 4. Voting (Classification) or Averaging (Regression):

2.5 REP Tree

REP (Repeated Incremental Pruning to Produce Error Reduction) Tree is a machine learning algorithm for classification and regression tasks. A decision tree-based algorithm constructs a decision tree using a combination of incremental pruning and error-reduction techniques. The key steps involved in building a REP Tree are as follows:

- Recursive Binary Splitting
- Pruning
- Repeated Pruning and Error Reduction

Steps involved in REP Tree

Below are the steps involved in building a REP Tree.

Step 1. Recursive Binary Splitting Step 2. Pruning

Step 3. Repeated Pruning and Error Reduction Step 4. Model Evaluation

2.6 Accuracy Metrics

The predictive model's error rate can be evaluated by applying several accuracy metrics in machine learning and statistics. The basic concept of accuracy evaluation in regression analysis is comparing the original target with the predicted one and using metrics like R-squared, MAE, MSE, and RMSE to explain the errors and predictive ability of the model [14]. The R-squared, MSE, MAE, and RMSE are metrics used to evaluate the prediction error rates and model performance in analysis and predictions [15] and [16].

The correlation coefficient, often denoted by the symbol "r," is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It is commonly used to assess the degree to which changes in one variable are associated with changes in another. The correlation coefficient takes values between -1 and 1:

$r = \Sigma ((X - \bar{x}) (Y - \bar{y})) / \sqrt{} (\Sigma (X - \bar{x})^2 * \Sigma(Y - \bar{y})^2)$ ... (1)

- ❖ $r \approx +1$: A strong positive correlation (as X increases, Y increases).
- ❖ $r \approx -1$: A strong negative correlation (as X increases, Y decreases).
- ❖ $r \approx 0$: Little to no linear correlation (no consistent relationship between X and Y).

MAE (Mean absolute error) represents the difference between the original and predicted values extracted by averaging the absolute difference over the data set.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}| \qquad (3)$$

RMSE (Root Mean Squared Error) is the error rate by the square root of MSE.

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y})^2} \qquad (4)$$

Relative Absolute Error (RAE) is a metric used in statistics and data analysis to measure the accuracy of a forecasting or predictive model's predictions. It is particularly useful when dealing with numerical data, such as in regression analysis or time series forecasting.

$$RAE = \frac{\Sigma|y_i - \hat{y}|}{\Sigma|y_i - \bar{y}|} \qquad (5)$$

Root Relative Squared Error (RRSE) is another metric used in statistics and data analysis to evaluate the accuracy of predictive models, especially in the context of regression analysis or time series forecasting.

$$RRSE = \sqrt{\frac{\Sigma(y_i - \hat{y})^2}{\Sigma(y_i - \bar{y})^2}} \qquad (6)$$

Equation 3 to 7 are used to find the model accuracy, which is used to find the model performance and error. Where Yi represents the individual observed (actual) values, Ŷi represents the corresponding individual predicted values, Ȳ represents the mean (average) of the observed values and Σ represents the summation symbol, indicating that you should sum the absolute differences for all data points.

3. Numerical Illustrations

The dataset is rich with detailed crop production statistics for India, organized by state and district. It encompasses data for four significant crop seasons - kharif, rabbi, summer, and autumn- spanning from 1997 to 2020. This dataset offers insights into both the yearly crop production and yield across various regions of the nation. This dataset provides extensive agricultural production statistics for India, sourced directly from the Indian government's Area Production Statistics (APS) database. The APS, overseen by the Ministry of Agriculture and Farmers Welfare, offers in-depth information regarding crop production, yield, and cultivation area, spanning various states and districts across India. [17].

Table 1. Crop production statistics India sample dataset

State	District	Crop	Crop_Year	Season	Area	Production	Yield
Andaman and Nicobar Island	NICOBARS	Arecanut	2007	Kharif	2439.6	3415	1.4
Andaman and Nicobar Island	NICOBARS	Arecanut	2007	Rabi	1626.4	2277	1.4
Andaman and Nicobar Island	NICOBARS	Arecanut	2008	Autumn	4147	3060	0.74
Andaman and Nicobar Island	NICOBARS	Arecanut	2008	Summer	4147	2660	0.64
Andaman and Nicobar Island	NICOBARS	Arecanut	2009	Autumn	4153	3120	0.75
Andaman and Nicobar Island	NICOBARS	Arecanut	2009	Summer	4153	2080	0.5
Andaman and Nicobar Island	NICOBARS	Arecanut	2000	Kharif	1254	2000	1.59
Andaman and Nicobar Island	NICOBARS	Arecanut	2001	Kharif	1254	2061	1.64
Andaman and Nicobar Island	NICOBARS	Arecanut	2002	Whole Year	1258	2083	1.66
Andaman and Nicobar Island	NICOBARS	Arecanut	2003	Whole Year	1261	1525	1.21
Andaman and Nicobar Island	NICOBARS	Arecanut	2004	Whole Year	1264.7	806	0.64
Andaman and Nicobar Island	NICOBARS	Arecanut	2006	Whole Year	896	478	0.53
Andaman and Nicobar Island	NICOBARS	Arecanut	2010	Rabi	944	1610	1.71
Andaman and Nicobar Island	NICOBARS	Arecanut	2011	Rabi	957	1090	1.14
Andaman and Nicobar Island	NICOBARS	Arecanut	2012	Rabi	959	1362	1.42
Andaman and Nicobar Island	NICOBARS	Arecanut	2013	Rabi	890.5	846	0.95

Table 2: Machine Learning Models with Correlation coefficient

ML Approaches	Area	Production	Yield
Linear Regression	0.0538	0.4395	0.4369
Multilayer Perceptron	0.0031	0.4637	0.7506
Random Forest	0.9833	0.9970	0.9961
Random Tree	0.9842	0.9931	0.9943
REP Tree	0.9611	0.9890	0.9865

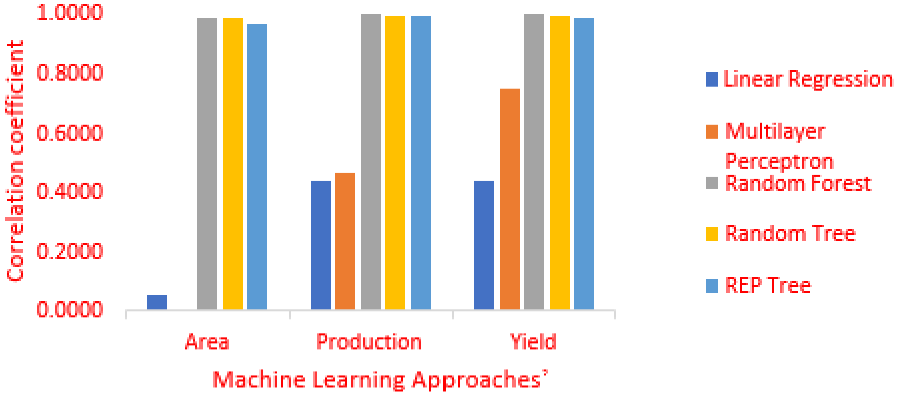

Fig. 1. Correlation coefficient for Machine Learning Approaches

Table 3: Machine Learning Models with Mean Absolute Error

ML Approaches	Area	Production	Yield
Linear Regression	17470.6918	1356734.4153	124.8619
Multilayer Perceptron	24167.2920	1293959.9860	68.6951
Random Forest	520.0557	112393.9307	7.7286
Random Tree	775.0137	136177.9119	9.2886
REP Tree	917.7509	165275.1456	12.2442

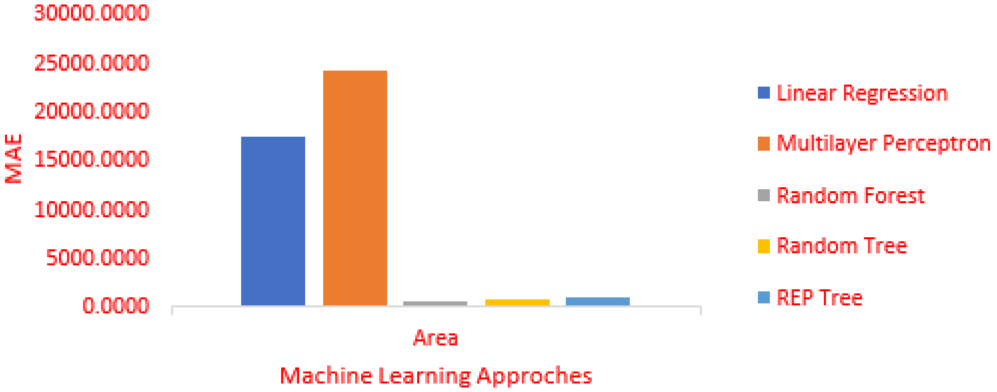

Fig. 2. Machine Learning Models with MAE

Table 4: Machine Learning Models with Root Mean Squared Error

ML Approaches	Area	Production	Yield
Linear Regression	46059.8890	19339651.3698	830.5156
Multilayer Perceptron	51159.144	19078111.64	610.7335
Random Forest	8613.781	1733092.371	83.366
Random Tree	8200.9297	2532332.264	99.4777

REP Tree	12774.208	3214853.217	151.055

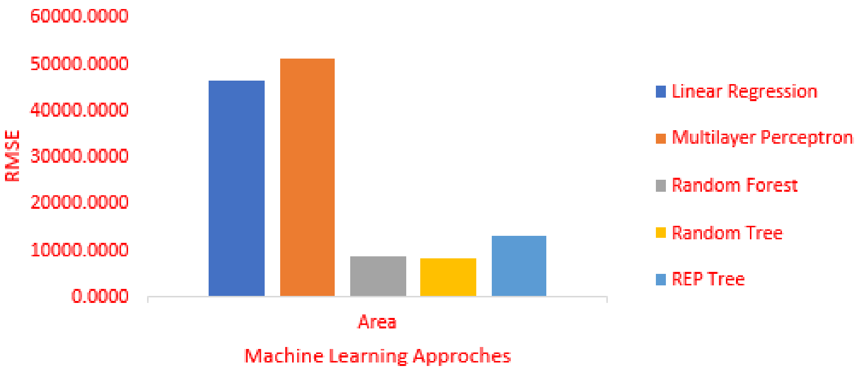

Fig. 3. Machine Learning Models with RMSE

Table 5: Machine Learning Models with Relative Absolute Error (%)

ML Approaches	Area	Production	Yield
Linear Regression	99.3725	73.7382	82.7120
Multilayer Perceptron	137.4624	70.3264	45.5055
Random Forest	2.9581	6.1086	5.1197
Random Tree	4.4082	7.4012	6.1531
REP Tree	5.2201	8.9827	8.1109

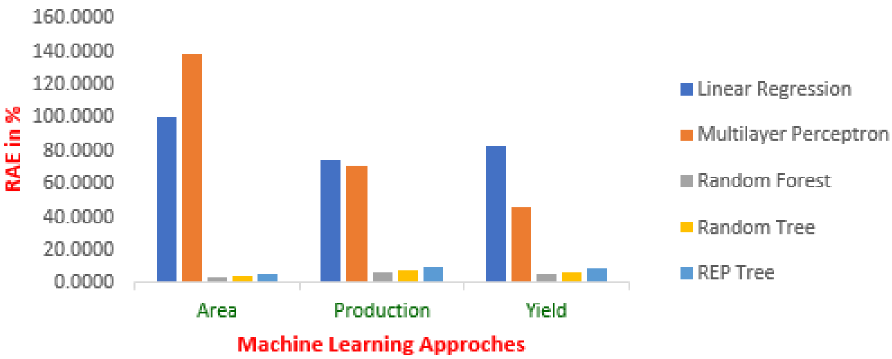

Fig. 4. Machine Learning Models with RAE (%)

Table 6: Machine Learning Models with Root Relative Squared Error (%)

ML Approaches	Area	Production	Yield
Linear Regression	99.8548	89.8229	89.9530
Multilayer Perceptron	110.9097	88.6082	66.1484
Random Forest	18.6741	8.0493	9.0294
Random Tree	17.7791	11.7614	10.7744
REP Tree	27.6936	14.9314	16.3607

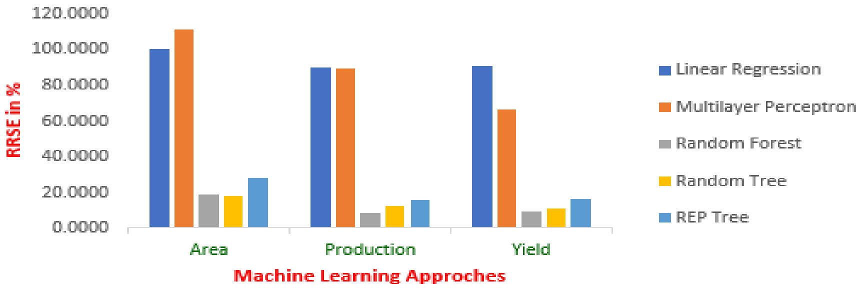

Fig. 5. Machine Learning Models with RRSE (%)

Table 7: Machine Learning Models with Time Taken to Build Model (Seconds)

ML Approaches	Area	Production	Yield
Linear Regression	1.3000	0.3600	0.3200
Multilayer Perceptron	69.6100	127.1700	152.6300
Random Forest	163.4900	73.0900	196.6600
Random Tree	6.2100	1.1400	2.5800
REP Tree	3.4800	1.3300	2.2500

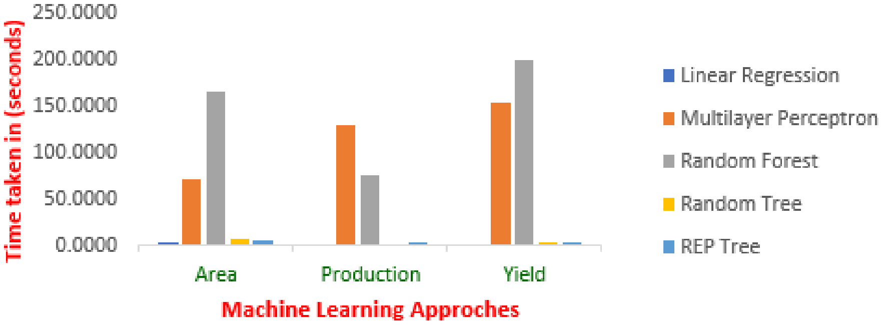

Fig. 6. Machine Learning Models and its Time Taken to Build the Model (Seconds)

4. Results and Discussion

The performance of various machine learning models for predicting area, production and yield of crops in India is analyzed using multiple evaluation metrics. The dataset includes four major crop seasons - Kharif, Rabi, Summer and Autumn - and covers a period from 1997 to 2020. The correlation coefficient measures the strength of the linear relationship between predicted and actual values. Higher values indicate stronger relationships. Random Forest, Random Tree, and REP Tree models performed the best across all categories (Area, Production, and Yield), with correlation coefficients close to 1.0, indicating strong predictive power. Linear Regression and Multilayer Perceptron (MLP) showed significantly weaker correlations, particularly for Area and Yield, where MLP performed poorly in comparison (Area: 0.0031, Yield: 0.7506). The related numerical illustrations are in Table 2 and Figure 1.

MAE indicates the average magnitude of the error in a set of predictions, without considering the direction. Random Forest exhibited the lowest MAE across all variables, meaning it had the smallest absolute prediction errors for Area (520.05), Production (112393.93), and Yield (7.73). In contrast, Multilayer Perceptron had the highest MAE for Area (24167.29), indicating significant deviations from actual values. The related experimental results are shown in Table 3 and Figure 2.

RMSE penalizes larger errors more than MAE, providing a measure of accuracy that is sensitive to large deviations. Random Forest again showed superior performance with the lowest RMSE across all dimensions. Linear Regression had the highest RMSE in all cases, indicating it is not suitable for complex predictions like crop yield where relationships may be nonlinear. The related numerical examples shown in Table 4 and Figure 3.

RAE provides a ratio of the sum of absolute errors of the model to the sum of absolute errors of the baseline model. Random Forest had the lowest RAE, indicating high predictive accuracy concerning the baseline. Multilayer Perceptron showed a poor performance in terms of RAE, especially for Area (137.46%), which means its error relative to the baseline is quite high. The related experimental results are shown in Table 5 and Figure 4.

The time required to train the models is a crucial consideration, especially in large datasets. Random Tree and REP Tree models were the fastest to build, while Random Forest took the longest, likely due to the complex nature of ensemble models. Multilayer Perceptron was also slow due to the complexity of neural network training, particularly when optimizing multiple layers. The numerical illustrations are shown in Table 6 and Figure 5.

5. Conclusion

Random Forest consistently performed the best across all metrics (MAE, RMSE, RAE), demonstrating its ability to model complex, non-linear relationships in crop yield predictions. This makes it an ideal choice for agricultural applications where data characteristics vary significantly. Linear Regression underperformed, especially for non-linear relationships like crop yield and area predictions. Its high error rates and weak correlation coefficients suggest that it may not capture the underlying patterns in the data well. Multilayer Perceptron showed strong results for yield prediction but struggled with area and production. It also had high build times, making it less practical for large datasets or real-time prediction. Random Tree and REP Tree offer a balance between speed and accuracy, providing strong predictive power while requiring much less time to build the models.

Random Forest emerges as the best overall model for predicting Indian agricultural crop yields based on various metrics, though Random Tree and REP Tree also show potential due to their balance of accuracy and speed.

References

1. Bel, Nizar, and Hadj Ali. "A Global Approach to Optimization of Metal Structures with Genetic Algorithms," n.d., 199-206.
2. Bel Hadj Ali, Nizar, Jean-Claude Mangin, et Af Cutting-Decelle. « Optimization of the steel structures design with genetic algorithms », 2002
3. Houck, Christopher, Jeffrey Joines, et Michael Kay. « A Genetic Algorithm for Function

Optimization: A MATLAB implementation ». NCSUIE-TR-95-09. North Carolina State University, Raleigh, NC, USA 22 (mai 1998)

4. Tahar Belarbi, m.; Mehdi Bitam, m. "contribution to the overall conceptual design optimization of steel structures." courrier du savoir, [s.l.], vol. 7, may 2014. Issn.
5. Colson, a., jm Hottier, and a. Moricet. "Simplified models of semi-rigid connections—comparative economic analysis." construction metallique 4 (1996).
6. Nethercot, D.A., et L. Gardner. « The EC3 approach to the design of columns, beams and beam-columns ». Steel and Composite Structures 5, no 2_3 (25 avril 2005): 127 40.https://doi.org/10.12989/SCS.2005.5.2_3.127.
7. Collette, Yann, et Patrick Siarry. Optimisation Multiobjectif. Paris: Eyrolles, 2002.
8. Ravindran, A. Ravi, Kenneth M. Ragsdell, Gintaras V. Reklaitis, K. M. Ragsdell, et G. V. Reklaitis. Engineering Optimization: Methods and Applications. 2. ed. Hoboken, NJ: Wiley, 2006.
9. Talbi, El-Ghazali. Metaheuristics: From Design to Implementation. 1re éd. Wiley, 2009. https://doi.org/10.1002/9780470496916.
10. Kaveh, A., et T. Bakhshpoori. « An efficient multi-objective cuckoo search algorithm for design optimization ». Advances in Computational Design 1, no 1 (25 janvier 2016): 87 103. https://doi.org/10.12989/ACD.2016.1.1.087.
11. Nguyen, T. H., V. D. Le, X. H. Vu, et D. K. Nguyen. « Reliability-Based Design Optimization of Steel-Concrete Composite Beams Using Genetic Algorithm and Monte Carlo Simulation ». Engineering, Technology & Applied Science Research 12, no 6 (1 décembre 2022): 9766 70. https://doi.org/10.48084/etasr.5366.
12. Amir, Ilham Yahya, Abdinasir Mohamed Yusuf, et Ikenna D. Uwanuakwa. « A Metaheuristic Approach of predicting the Dynamic Modulus in Asphalt Concrete ». Engineering, Technology & Applied Science Research 14, no 2 (2 avril 2024): 13106 11. https://doi.org/10.48084/etasr.6808.
13. Noureddine, Sarir, Sebaa Morsli, Allaoui Tayeb, et Denai Mouloud. « Optimal Fractional-Order PI Control Design for a Variable Speed PMSG-Based Wind Turbine ». Journal Européen des Systèmes Automatisés 54, no 6 (29 décembre 2021): 915 22. https://doi.org/10.18280/jesa.540615.
14. Benanane, s., l. Amamra, z.i. Bouzadi, a. Benanane, s.m. Bourdim, and m. Titoum. "Application of Genetic Concepts to the Optimization of a Reinforced Concrete Cantilever Beam." *Academic Journal of Civil Engineering*, February 12, 2021, pp. 276-279. https://doi.org/10.26168/AJCE.37.1.58.
15. Terki Hassaine, M. I. E., S. M. E. A. Bourdim, H. Varum, A. Benanane, et A. Nour. « Push-over Analysis of Optimized Steel Frames ». Engineering, Technology & Applied Science Research 12, no 6 (1 décembre 2022): 9720 25.https://doi.org/10.48084/etasr.5326.
16. Kumar, V., et S. K. Dhull. « Genetic Algorithm Based Optimization of Uniform Circular Array ». Engineering, Technology & Applied Science Research 10, no 6 (20 décembre 2020): 6403 9.https://doi.org/10.48084/etasr.3792.
17. Maskaoui, Z. El. « Genetic Algorithm Parameters Effect on the Optimal Structural Design Search ». IOSR Journal of Mechanical and Civil Engineering 14, no 3 (juin 2017): 124 30. https://doi.org/10.9790/1684-140305124130.
18. Nguyen, Anh-Tuan, Sigrid Reiter, et Philippe Rigo. « A Review on Simulation-Based Optimization Methods Applied to Building Performance Analysis ». Applied Energy 113 (janvier 2014): 1043 58. https://doi.org/10.1016/j.apenergy.2013.08.061
19. Barnier, Nicolas, and Pascal Brisset. "Optimization by Genetic Algorithm under Constraints," n.d.
20. Eurocode 3 in 1993: Examples of Application to Steel Structure Calculations. Brussels: Centre Information Acier, 2007