iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.83652

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

Data-Driven Empirical Study of Agricultural Production Forecasting in India Using Supervised Machine Learning Models

Ms. Sayali Parab[1], Mr. Chayan Bhattacharjee[2]
[1]*Department of Information Technology, SES's L. S. Raheja College of Arts & Commerce, Mumbai, India*
[2]*Department of Information Technology Chikitsak Samuha's Patkar Varde College, Mumbai, India*

Abstract: ***As we all know in India; agriculture is the country's backbone. This report forecasts the yield of practically every type of crop grown in India. This programme is created by utilising simple characteristics such as state, district, and the agricultural yield can be predicted based on the season, area, and the user. the year he or she wishes to participate in. The paper makes use of machine learning algorithms like linear regression and random forest. I have also discussed why predicting on the basis of the time series algorithm is a bad idea. And with the multiple factors in the picture like average rain, political concerns, current economy, using time series models will be a bad idea since the predictions will be of no use or with minimum accuracy.***
Keywords: Machine learning, Linear Regression, Random Forest, prediction, agriculture, Farming.

I. INTRODUCTION

India's economy is based on agriculture. Its contribution to GDP has fallen from 57 percent in 1950–51 to roughly 28 percent in 1998–99, owing mostly to expansion in other sectors of the economy. However, the agricultural sector's shrinking share has had no impact on the sector's importance in the Indian economy. The trend in agricultural production has largely impacted the growth of India's GDP, owing to both the direct value of agricultural products and agriculture's indirect impact on employment, rural life, and other sectors that use agricultural products. It has a significantly higher impact on the country's welfare than macroeconomic measures suggest. It has a significantly higher impact on the country's well-being than macroeconomic measures suggest: approximately 70% of the working population relies on agricultural operations for a living. Cereal and pulse cultivation provide food for the bulk of India's people. Agriculture also provides a significant amount of raw resources to industry. Cotton and jute for textiles, as well as sugar and vegetable oil, are examples. Agricultural production accounts for over half of the income generated in India's manufacturing sector, either directly or indirectly. Agricultural commodities, as well as products that rely on agriculture, make for about 70% of total export value. Tea, sugar, oilseeds, tobacco, and spices are among the most important export items.

The methodology of the research paper begins with the acquisition of agricultural data from the Government of India's official open data portal. The dataset is systematically cleaned and pre-processed to ensure reliability and analytical consistency. Following data preparation, exploratory analysis is conducted to understand underlying patterns and assess the feasibility of different predictive approaches. Suitable machine learning algorithms are then selected and evaluated based on their performance and relevance to the problem domain. The findings are interpreted through quantitative analysis and supported by clear visualizations generated using Python-based analytical tools.

II. UNDERSTANDING THE DATA

The data that we have is extracted from the government of India's official data portal (https://data.gov.in/). The data has State name, District name, Crop year (Year in which crop was grown), Season (Season of the crop in which it is grown), Crop name, Area (Area under cultivation), production (Number of crops produced). There was no unit for the area and number of crops, so we shall just refer it to unit area and unit crops. Understanding data column by column: State - This column contains names of the states and the union territories in the country, the highest number of appearances in this column is Uttar Pradesh. The next column is the District which has names of the district, it has 646 unique districts. The next column is the year which has years from 1997 to 2015. Moving on to the next column there is season, season has crop seasons, in this we have 5 unique seasons in which the crop grows. They are Kharif, Rabi, winter, summer and the whole year. Lastly we have the crops that are grown there are a variety of crops that are grown throughout the season.

Then we have the area in which farms are located. We don't have any metric given in the data so for sake of simplicity we choose to call it unit area. Last but not the least we have production in the area again these won't have any metric so we again for simplicity sake we use unit measure. This dataset contains a wealth of information about agricultural production in India across a number of years. The eventual goal, based on the data, would be to use strong machine learning algorithms to anticipate crop yield. We have around 2,46,091 rows and 7 columns.

III. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis refers to the critical process of conducting preliminary investigations on data in order to discover patterns, spot anomalies, test hypotheses, and validate assumptions using summary statistics and graphical representations. In statistics, exploratory data analysis is a way of analyzing datasets to summarize their main features, often using statistical graphics and other data visualization methods. So the dataset includes state-by-state crop production statistics, as well as district-by-district data. It contains information from 1997 to 2014. The crop season and field area are two other interesting features that are available to us. In this case, I'm going with the flow. I'll also back it up with some research (along with links) to check if the data and analysis line up. Note that this information only pertains to India. As a result, the scope of the investigation will be limited to India alone. Let's start with the state-by-state production. We'll need to aggregate data to discover patterns, which is both instructive and granular. Let's look at the zone specifics to see if there's any trend there.

Source: https://www.mapsofindia.com/zonal/.

Because there is activity before and after 2005, we believe it is significant. Similarly, we notice minor modifications in the Central, West, and North Zones around 2011, which differ from the East and North East Zones. Let's have a look at 2005 and 2011. Excerpts from the IMD's annual reports for the years in question. The rainfall in 2005 was not evenly distributed. In June, rainfall was below normal (12 percent below LPA) across the country. In July, however, the monsoon was vigorous, resulting in excessive rainfall (14 percent above LPA). In August, the monsoon was subdued, with a high LPA shortage of 28%. The monsoon began again in September (rainfall +17 percent over LPA), assisting in a timely resurrection and improving the seasonal rainfall status across the country. In 2011, 33 of the 36 meteorological subdivisions received excess/normal season rainfall, accounting for 92 percent of the total area of the country, while the remaining three subdivisions (Arunachal Pradesh, Assam & Meghalaya, and NMMT, accounting for 8% of the total area of the country) received deficient season rainfall.

Using the SNS library in python it could be clearly found that the South Zone leads in terms of overall production. Digging deep into the data we find Kerala leads the production in the southern zone by a huge margin and the most grown crop in Kerala is none other than Coconut. Coconut should not be a surprise, since coconut is an all-weather tree and the plantation apparatus around Kerala is suitable for the growth of it.

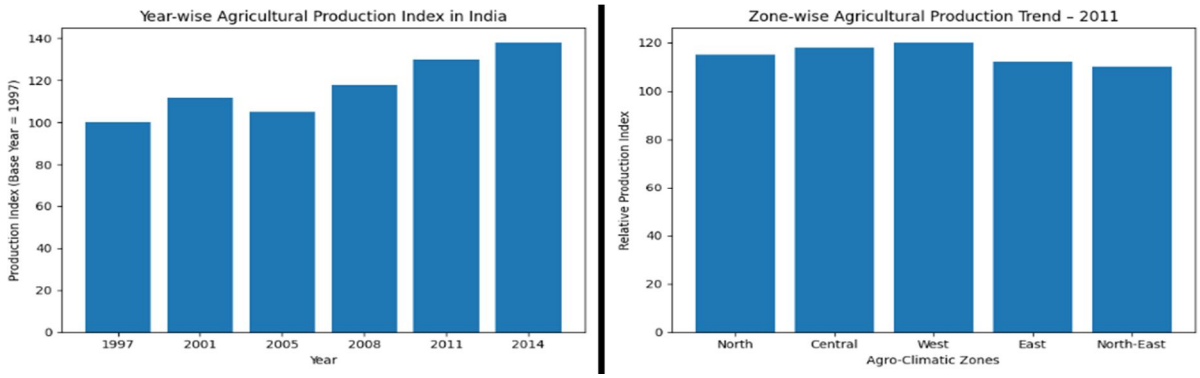
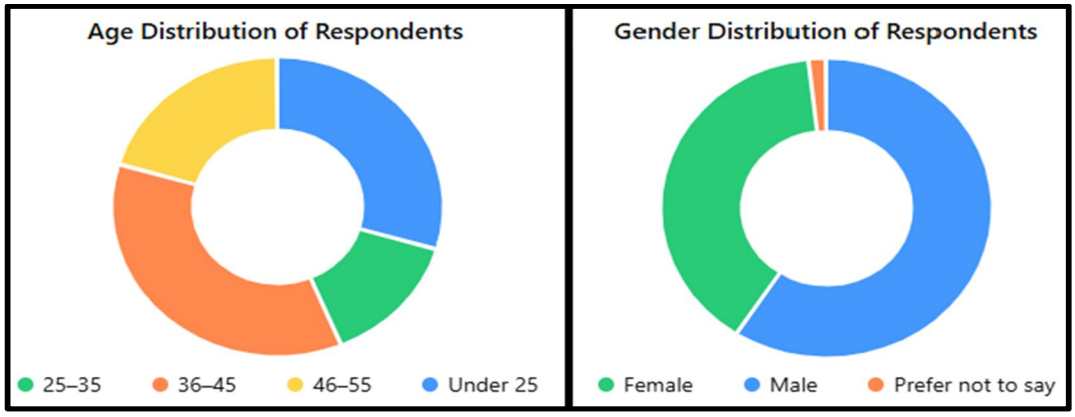


Fig :Trend zone wise

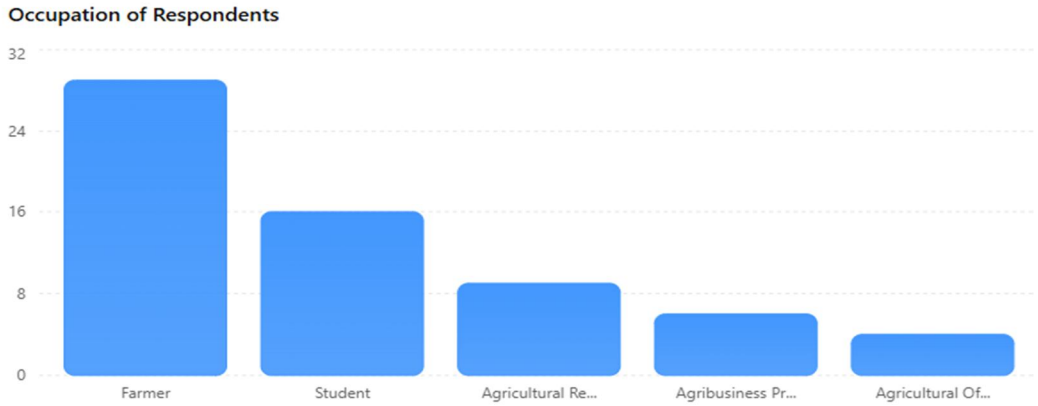
IV. SURVEY ANALYSIS

A. Responder Analysis

Out of the total 64 respondents, the largest age group of respondents was 36–45 years (35.9%), followed by Under 25 years (29.7%). Respondents aged 46–55 years accounted for 20.3%, while 25–35 years represented 14.1% of the sample. The survey was predominantly completed by male respondents (59.4%), while 39.1% were female and 1.6% preferred not to disclose their gender.

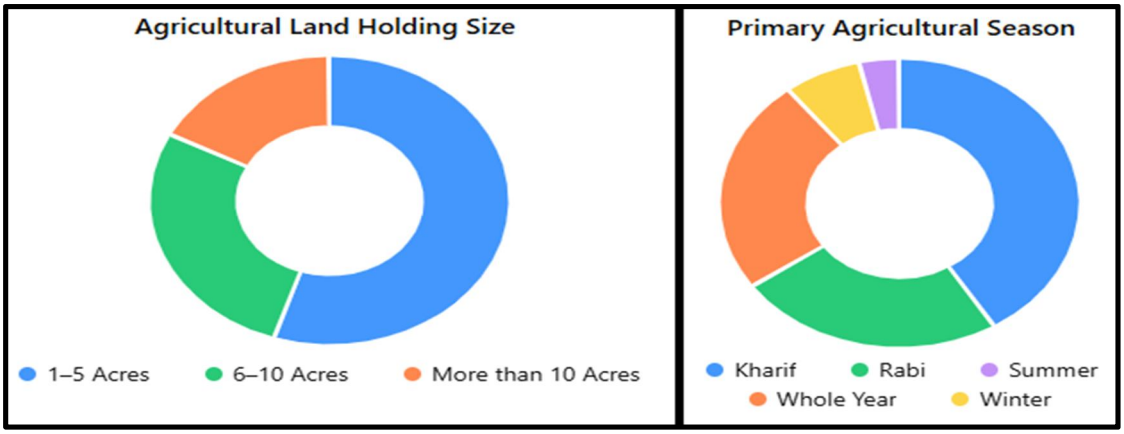


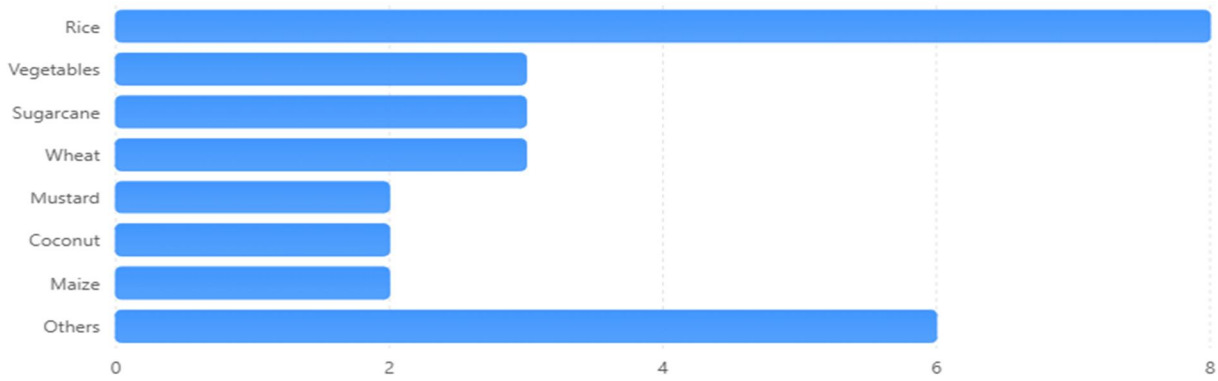
The sample was dominated by farmers (45.3%), followed by students (25.0%). Agricultural researchers (14.1%), agribusiness professionals (9.4%), and agricultural officers (6.3%) formed the remaining respondents. This indicates that the survey captured perspectives from both agricultural practitioners and stakeholders associated with agricultural research and administration.



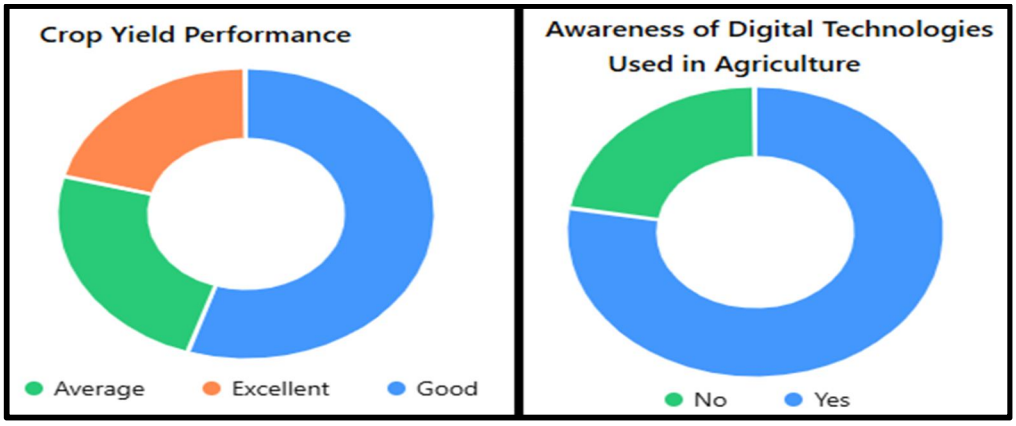
B. Agricultural Analysis

Among respondents involved in farming, small land holdings were predominant. More than half of the farmers reported owning between 1 and 5 acres of agricultural land, while fewer respondents possessed medium-sized (6–10 acres) or large holdings (more than 10 acres). This suggests that agricultural prediction systems should be designed with small and medium-scale farmers as primary beneficiaries. Kharif was identified as the most important production season by respondents. This highlights the strong dependence of agricultural output on monsoon-related climatic conditions and reinforces the importance of weather forecasting and seasonal prediction systems.





Rice emerged as the most frequently cultivated crop, indicating its importance among the surveyed farming population. The presence of diverse crops such as wheat, sugarcane, vegetables, mustard, maize, and coconut reflects the varied agro-climatic conditions represented in the survey.



Most respondents rated their crop yields as good, while a smaller proportion reported average or excellent performance. These findings suggest generally satisfactory agricultural productivity, although there remains potential for improvement through technology-enabled forecasting and decision-support systems. The findings indicate that a substantial majority of respondents (**77.8%**) are aware of digital technologies used in agriculture, while only **22.2%** reported having no awareness of such technologies. This suggests a relatively high level of exposure to agricultural digitalization among the surveyed population. The result reflects increasing dissemination of technological innovations such as mobile applications, weather forecasting systems, IoT-based monitoring tools, and AI-enabled decision-support systems within the agricultural sector.

V. ALGORITHM

Time series forecasting methodologies are extensively employed for the analysis and prediction of phenomena that change over time, as they utilize historical data and the temporal patterns inherent in it. These models demonstrate particular efficacy when the dataset reveals discernible traits such as long-term trends, seasonality, cyclical behaviour, or consistent temporal dependencies. In such scenarios, previous observations yield valuable insights for forecasting future values. Conversely, when these fundamental temporal structures are weak, inconsistent, or non-existent, the dependability of time series forecasting diminishes significantly, irrespective of the seemingly favourable performance suggested by statistical evaluation metrics.

An exploratory examination of the agricultural production dataset uncovered a lack of a distinct and stable temporal trend throughout the study period. Rather than adhering to a predictable trajectory, production levels exhibited irregular fluctuations from one year to the next. These variations were chiefly driven by external and uncontrollable factors, notably changes in rainfall, weather conditions, and other environmental elements. Given that such exogenous variables significantly influence agricultural output, relying exclusively on historical production data proves inadequate for producing precise forecasts. Moreover, widely utilized performance indicators such as the coefficient of determination (R^2), Mean Absolute Percentage Error (MAPE), and similar percentage-based metrics may yield an excessively optimistic evaluation of model performance in time series contexts. Elevated metric values do not inherently signify authentic forecasting proficiency, particularly when the model fails to account for the genuine factors driving variation within the data. Consequently, conclusions derived solely from these metrics can be deceptive.

Given the absence of a clear temporal framework and the significant impact of external factors, the use of conventional time series models was considered inappropriate for this dataset. Consequently, this research emphasizes alternative supervised machine learning methods that can integrate various explanatory variables and more effectively illustrate the intricate relationships that dictate agricultural production. Linear Regression is among the most commonly employed statistical methods for analyzing the connection between a dependent variable and one or more independent variables. The aim of this model is to uncover a linear association that can facilitate the prediction of the target variable's value. Linear Regression can be divided into two categories: Simple Linear Regression, which involves a single predictor variable, and Multiple Linear Regression, which incorporates two or more predictor variables. In contrast to deterministic relationships, where one variable can be precisely calculated from another, linear regression aims to represent statistical relationships that encompass uncertainty and variability. For instance, while the temperature in Fahrenheit can be accurately derived from the temperature in Celsius, the correlation between variables such as height and weight is statistical, as individuals with identical heights may possess varying weights.

In this research, Linear Regression was assessed as a possible predictive model for agricultural production. Nevertheless, the findings revealed that the dataset was not ideally suited for a linear modelling approach. Agricultural production is affected by a multitude of intricate and non-linear factors, including rainfall, climatic conditions, soil characteristics, and farming practices. Consequently, the assumptions underlying linear regression were not sufficiently met. The evaluation of the model yielded an R^2 score of merely **0.02975,** signifying that the model accounted for less than 3% of the variance in the target variable. Such a low coefficient of determination illustrates a very weak linear relationship between the chosen predictors and agricultural production. Therefore, Linear Regression was deemed inappropriate for this dataset, prompting the exploration of more sophisticated machine learning models capable of capturing non-linear patterns. Random Forest is a robust supervised machine learning algorithm that is extensively utilized for both classification and regression tasks. It represents an ensemble learning approach that integrates the predictions of numerous decision trees to enhance overall model performance and predictive accuracy. Instead of depending on a solitary decision tree, the Random Forest algorithm generates a substantial number of trees during the training phase and consolidates their outputs to produce the final prediction. In the case of regression problems, the final prediction is derived by averaging the outputs of all individual trees, whereas for classification tasks, the final class is determined by majority voting.

The algorithm utilizes a method known as bootstrap aggregation, or bagging, which involves creating multiple subsets of the training data through random sampling with replacement. Each decision tree is trained on a distinct subset of data, which diminishes the risk of overfitting and enhances the model's capacity to generalize to new, unseen data. Additionally, Random Forest incorporates randomness in feature selection during the construction of trees, which fosters diversity among the trees and bolsters prediction robustness. In the realm of agricultural production prediction, Random Forest has demonstrated itself to be one of the most effective models assessed in this study. Agricultural datasets frequently exhibit complex, non-linear relationships influenced by variables such as rainfall, temperature, soil conditions, and farming practices. Unlike linear models, Random Forest is capable of capturing these intricate interactions without necessitating stringent assumptions regarding data distribution. Although the algorithm demands a relatively longer training duration due to the creation of multiple decision trees, it provides highly dependable predictions. The evaluation of the model yielded an R^2 score of 0.820364, which signifies that around 82% of the variance in agricultural production can be accounted for by the model. This elevated coefficient of determination reflects a robust predictive ability and implies that Random Forest is exceptionally well-suited for modelling agricultural production data. The findings validate its efficacy in managing intricate datasets and support its choice as one of the favoured prediction methods in this research.

VI. ADVANTAGES AND DISADVANTAGES

The study offers several advantages in the field of agricultural production forecasting. Enhanced predictive accuracy is achieved through the Random Forest model, which attained an R^2 score of roughly 0.82, signifying robust predictive capabilities for forecasting agricultural production. The model also demonstrates a strong capacity to manage non-linear relationships, as agricultural production is influenced by numerous intricate factors, and Random Forest proficiently captures the non-linear interactions among these variables. Furthermore, it provides data-driven decision support that can aid farmers, agricultural officials, and policymakers in strategizing crop cultivation, resource distribution, and production management. The research utilizes government agricultural datasets, ensuring that the methodology is transparent, reproducible, and scalable. Another significant advantage is the minimized risk of overfitting, as Random Forest utilizes bagging and multiple decision trees, enhancing generalization and resilience. Additionally, forecasting models can facilitate agricultural planning by enhancing regional agricultural policies and contributing to food security planning. The study also benefits from extensive dataset coverage, examining over 246,000 records from various states, districts, crops, and seasons throughout India.

The study also presents several limitations that should be considered when interpreting the findings. One major limitation is the limited feature set, as the dataset omits critical variables like rainfall, temperature, soil quality, fertilizer application, pest infestations, and market conditions, all of which have a substantial impact on crop yield. Additionally, data availability constraints exist because the dataset spans only from 1997 to 2015, potentially failing to capture current agricultural trends and shifts in climate. Another limitation is the absence of standardized measurement units, as the area and production figures are presented without standardized units, which could hinder their interpretation and practical use. Agricultural output is also heavily reliant on unpredictable weather patterns, making weather dependency a significant challenge that restricts the reliability of forecasts during extreme climatic events. Furthermore, the Random Forest algorithm involves computational complexity, demanding greater computational power and longer training durations compared to more straightforward models like Linear Regression. The research is also characterized by a limited geographical scope, as it is solely concentrated on India, meaning that the results may not be applicable to other nations with distinct agricultural practices. Finally, although Random Forest achieves high accuracy, model interpretability remains a concern, as it is less interpretable than conventional statistical models, complicating the explanation of individual predictions to end users.

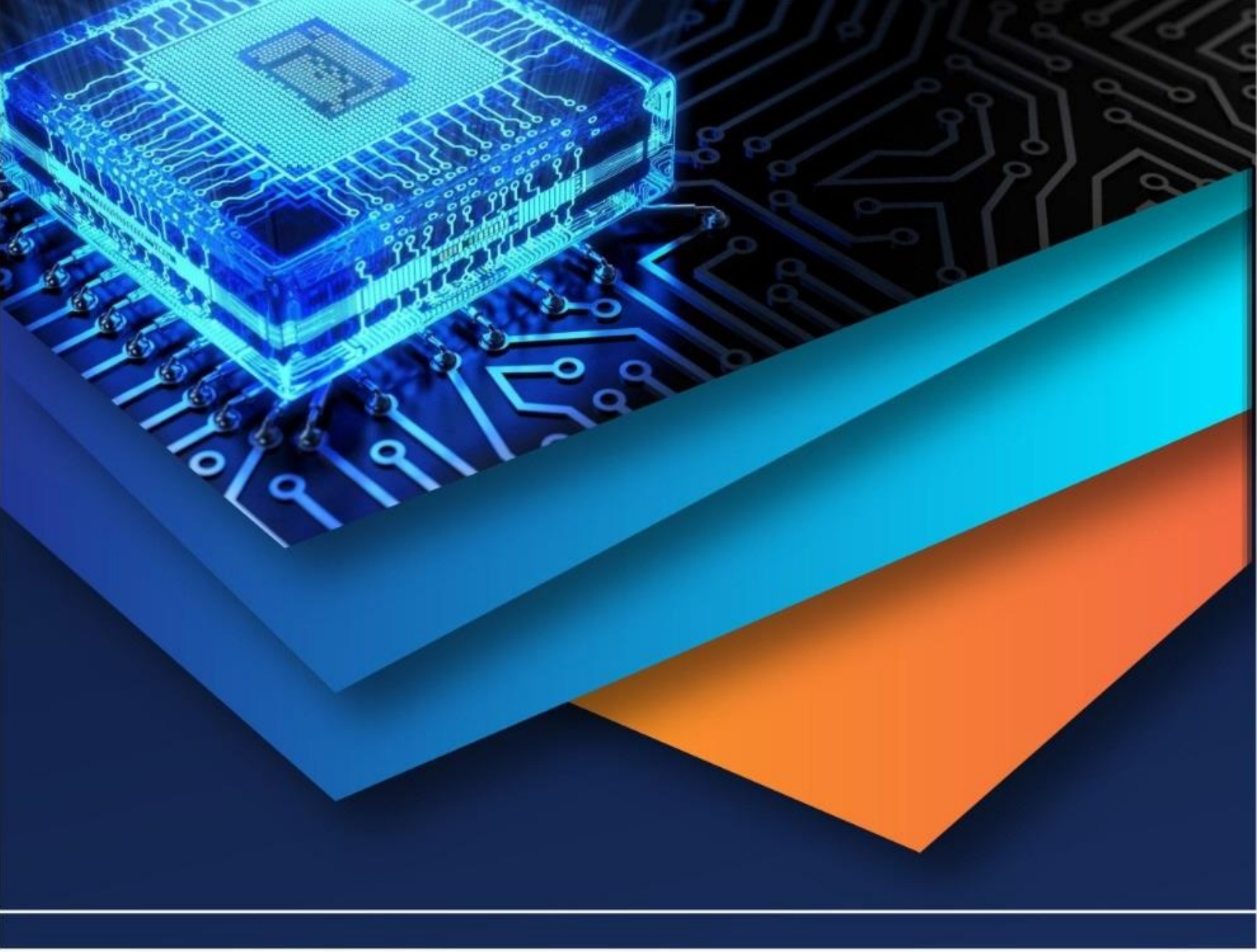
VII. CONCLUSION

The precise prediction of agricultural output continues to pose a significant challenge, primarily due to the considerable impact of uncertain and dynamic elements, including weather patterns, climate fluctuations, and local farming practices. This research underscored the critical role of exploratory data analysis in the preparation and comprehension of agricultural datasets, as it aids in data cleansing, trend detection, and pattern recognition. By utilizing agricultural production data from India, the effectiveness of supervised machine learning models was assessed, with Random Forest identified as the most proficient method. This model adeptly captured intricate, non-linear interactions among the variables and demonstrated enhanced predictive accuracy when compared to conventional techniques such as Linear Regression. The results emphasize the promise of machine learning methodologies in improving agricultural forecasting and facilitating evidence-based decision-making. Additionally, the reliance on publicly accessible government datasets promotes transparency and reproducibility in the research process.

The suggested methodology can support policymakers, agricultural organizations, and farmers in strategic planning and resource distribution, ultimately leading to enhanced agricultural productivity, sustainability, and food security.

REFERENCES

[1] Chand, R., Raju, S. S., & Pandey, L. M. (2017). Growth crisis in Indian agriculture: Severity and options at national and state levels. Economic & Political Weekly, 52(21).
[2] Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). Machine learning approaches for crop yield prediction and nitrogen status estimation. Computers and Electronics in Agriculture, 151, 61–69.
[3] Birthal, P. S., Joshi, P. K., Roy, D., & Thorat, A. (2012). Diversification in Indian agriculture towards high-value crops. Agricultural Economics Research Review, 25(1), 1–12.
[4] Jeong, J. H., Resop, J. P., Mueller, N. D., et al. (2016). Random forests for global and regional crop yield predictions. PLOS ONE, 11(6).
[5] Ministry of Agriculture & Farmers Welfare. State of Indian Agriculture. Government of India, New Delhi.
[6] National Sample Survey Office (NSSO). Situation Assessment Survey of Agricultural Households in India. Ministry of Statistics and Programme Implementation, Government of India



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY