iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** II **Month of publication:** February 2023

DOI: https://doi.org/10.22214/ijraset.2023.49230

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 11 Issue II Feb 2023- Available at www.ijraset.com

Analysis Of Agriculture Data Using Machine Learning

Mr. N. Harikrishna[1], Ch. Ramanjaneyulu[2], A. Karthikeya[3], B. Somasekhar[4], Ch. Mahesh[5]
[1, 2, 3, 4, 5]*Department of Computer Science & Engineering, KKR & KSR Institute of Technology and Sciences (A), Guntur, India*

Abstract: ***Without a question, agriculture provides the majority of livelihood opportunities in India and significantly boosts the national economy. Practices and managerial choices are two examples of technological elements influencing crop productivity. Hence, forecasting crop production in advance of harvest would aid farmers in making the right decisions. By creating a user-friendly prediction system, we try to find a solution. The outcome of the forecast is suggested to the farmer so that appropriate adjustments can be made to enhance the yield. Crop yield can be predicted using a variety of methods or algorithms. A viable remedy for the issue farmers is facing can be found by assessing all the factors involved, including location, soil nutrients, pH value, rainfall, and moisture. In order to provide insight prior to the actual crop production, this research employs machine learning algorithms to analyse agricultural data and discover the optimal yield.***
Keywords: Crop, Yield, Radom Forest Regressor, Prediction

I. INTRODUCTION

India is currently one of the world's top producers of agricultural goods. The largest economic sector, horticulture plays a significant role in India's economy. Horticulture is a unique form of crop production that is influenced by a variety of economic and environmental factors. With an economy that is primarily based on agriculture, Andhra Pradesh contributes more than 29% of the nation's GDP as opposed to 17% nationally. The state's horticulture industry may be strengthened by providing regular assistance to the ranchers regarding improved farming practises or advancements in aspects impacting the creation of harvests. One of the advancements in rural areas is yield forecast. These kinds of advancements are what are piquing modern man's interest in farming. Ranchers used to predict their yield based on previous experiences. Even young ranchers benefit from mindfulness about the development of yields at the ideal time and location thanks to digitalization in farming. The use of information analytics is required for these kinds of advancements. This approach is one that can be applied to deal with yield forecasts.

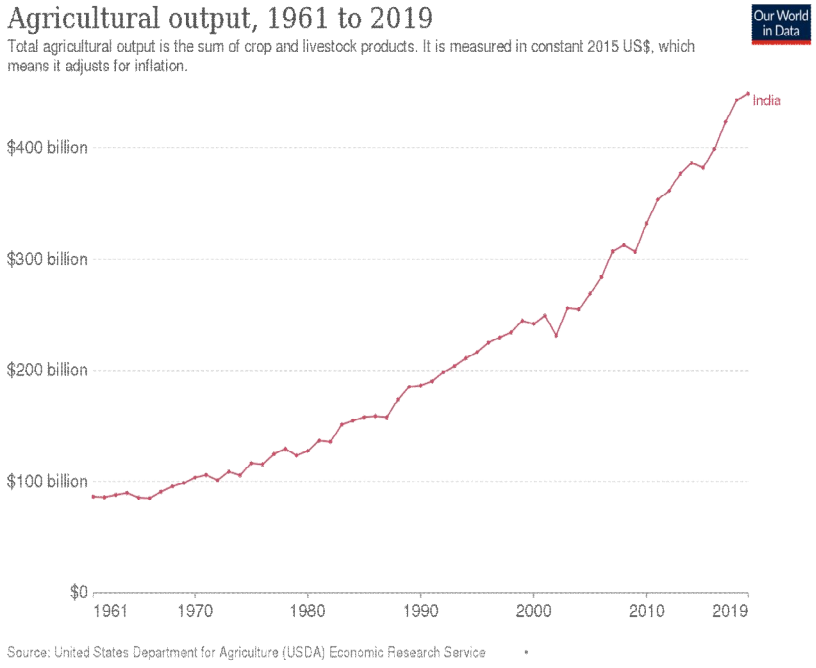


FIG1: AGRICULTURE IN INDIA

©IJRASET: All Rights are Reserved

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue II Feb 2023- Available at www.ijraset.com

A. Motivation Of Work

The main source of employment in India and a substantial contributor to the national economy is agriculture. Yet over time, it was ignored, and farmers' efforts became unappreciated. Many international conventions have acknowledged farming, and nations are now concentrating on the growth of their individual agricultural sectors. Farmers are urged to incorporate digital techniques into their farming methods as part of the digital India initiative. The administrative choices made as well as the procedures employed are technological elements that affect crop productivity. Crop production to meet dependable and timely needs for various agricultural marketing decisions. With data on agriculture, predictions are particularly helpful.

Data on farmers' purchasing habits can be completely utilised by the government by using data mining tools, which also help to better understand farmers' lands and increase farmer profit. Hence, forecasting crop production before it is harvested would help farmers take the necessary action. We create a user-friendly prediction system in an effort to solve the problem. The farmer is informed of the predicted outcomes so that appropriate adjustments can be made to enhance the crop.

II. PROBLEM STATEMENT

This project will use the Random Forest algorithm, one of the regression techniques, to analyse the agricultural data and choose the ideal parameters to maximise crop production. The dataset includes information about the year, the district, the crop, the season, the area, the production (in tonnes), the nitrogen (kg/Ha), the phosphorus (Kg/Ha), the potassium (Kg/Ha), and other elements. Understanding machine learning methods and using them on the dataset is the system's main objective.

III. LITERATURE REVIEW

1) Cluster analysis or clustering is a challenge in unsupervised learning. It is frequently applied as a data analysis tool for identifying intriguing data patterns, such as groupings of clients based on their behaviour. For a variety of uses, many clustering algorithms have been created. Partitioning clustering, which iteratively reallocates objects to enhance the quality of clustering results, hierarchical clustering algorithms, which assign objects to tree-structured clusters, and density-based clustering algorithms, which require that each point in a cluster have at least a certain number of neighbours within a given unit of distance. Grid-based approaches and model-based clustering methods are further types.

2) In the subject of agriculture, various forecasting approaches have been created and tested by researchers worldwide. Some of these studies include: Lakkana Ruekkasaem and Montalee Sasanian worked on the data from January 2013 to December 2017 for a total of 50 months in the Department of Industrial Engineering, Faculty of Engineering, Thammasat University, Pathumthani, Thailand. The Least Square Method, Moving Average Method (3 months, 5 months, and 7 months), Single Exponential Method, Double Exponential Method, and Winter's Method were the seven techniques used to evaluate the data utilising times series analysis.

3) Thompson (1985) used a statistical type model to assess how weather variability and climate change affected corn production in five Midwestern US states. He discovered strong correlations between changes in maize yield from the trend and pre-season precipitation (September to June), June temperature, and temperature and rainfall in July and August. This method greatly enhanced historical corn and soybean yield projections.

4) Lobell et al. (2011, 2013) used statistical models to analyse the impacts of temperature rises on maize production in the USA and came to the conclusion that yield decline will be significantly influenced by temperature increase under climate change. Many yield forecasting studies and programmes frequently use statistical regression to predict yield using agrometeorological inputs (NASS, 2005; Lobell et al., 2009).

5) Sunil Dixit, Manish Mahant, Abhishek Shukla, Dileshwer Patel (2012) Information and communication technology (ICT) is being used more and more in agriculture. E-agricultural is the idea, design, development, testing, and implementation of novel information and communication technology (ICT) applications in rural areas, with a primary emphasis on agriculture. Due to the fact that it comprises of three basic technologies, information and communication technology (ICT) can significantly contribute to maintaining information properties. These technologies are used for managing, processing, and transferring data, knowledge, and information.

6) V. Lohr, E. Cervenkova, HavliCek, J. Vanek (2010) Information and communications technologies (ICTs) have advanced quickly, enabling new uses that were not viable a few years ago. The majority of the rural population in emerging countries depends on agriculture, making it a significant sector. The industry must overcome significant obstacles to increase production in the face of depleting natural resources needed for that production.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |

Using an agricultural computer-based system, ICT is crucial for improving and challenging the livelihoods of the rural population.

7) N. Monica Agu (2013) Most third world economies are based on agriculture, which plays a crucial role in the growth of these nations. Considering the significance of agriculture, progress in this field has generally been inconsistent and underwhelming. It's crucial to acknowledge the diverse responsibilities played by women in farming systems.

8) P. Benda, Z. HavlCek, V. Lohr, and M. (2011) The so-called "digital divide" is a result of technical advancement in ICT (Information and Communication Technologies). Some people are unable to respond to this development on their own, but with the right ICT use, they can get over this obstacle. Depending on the nature and severity of the condition, one option is to develop useful and accessible software. E-learning resources are employed in accordance with the European CertiAgri project to help integrate people with impairments into the horticulture industry.

9) J. Doerflinger and T. Gross (2012) Information and communication technologies for development (ICTD) must be built with scalability and reusability in mind if they are to be long-term sustainable. A technical ICTD design, the Sustainable Bottom Billion Architecture has been successfully replicated in two ICTD projects in Africa's cashew and shea nut farming value chains.

10) Andrew, T.N., and Joseph, M.K. (2008) By offering services to farmers in rural regions, digital ICT created through participatory learning and action research can promote development and end poverty. The employment of a variety of ICTs in agriculture can improve the livelihood of farmers in rural areas and aid in their socioeconomic development, even though no single ICT will be adequate for farmers. In order to effectively employ ICTs in the agricultural domain, the study focuses on a variety of participatory methodologies, such as participatory communication and participatory learning. It emphasises how the development of Dasia's participatory information and communication technologies for rural farming communities might benefit from participatory techniques.

IV. PROPOSED SYSTEM

Despite the fact that there are numerous yield prediction models, neither their functionality nor their implementation in the real world are complete. So, we considered how to make our suggested system both completely functional and easy to design.

Our project's system architecture is shown in the diagram below. The entire system may be broken down into two modules, one of which forecasts the ideal yield and the other of which examines the patterns in the dataset. The above diagram makes it obvious how these model's function.

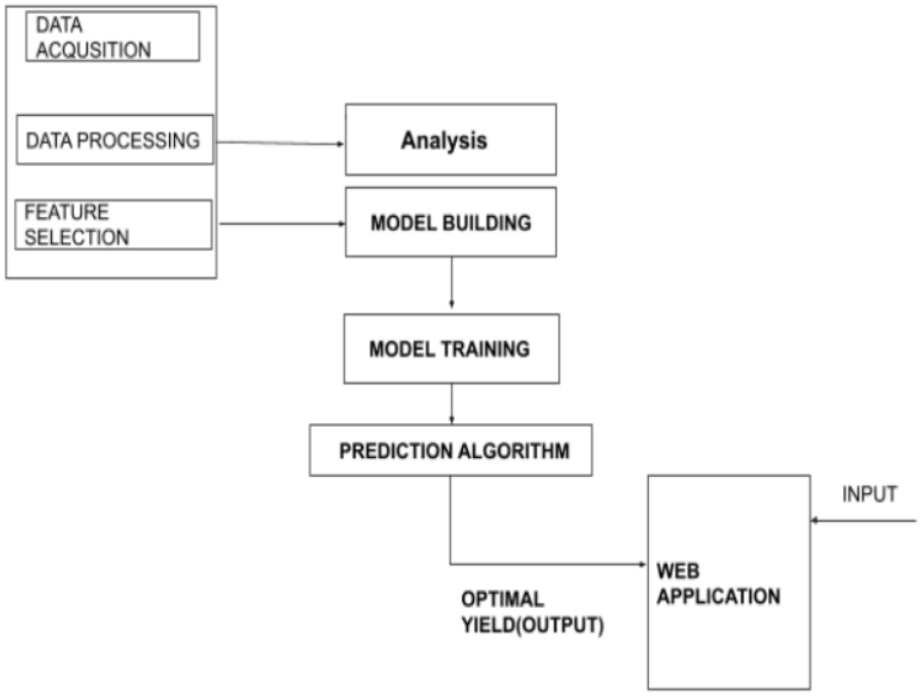


FIG2: BLUEPRINT OF PROPOSED SYSTEM

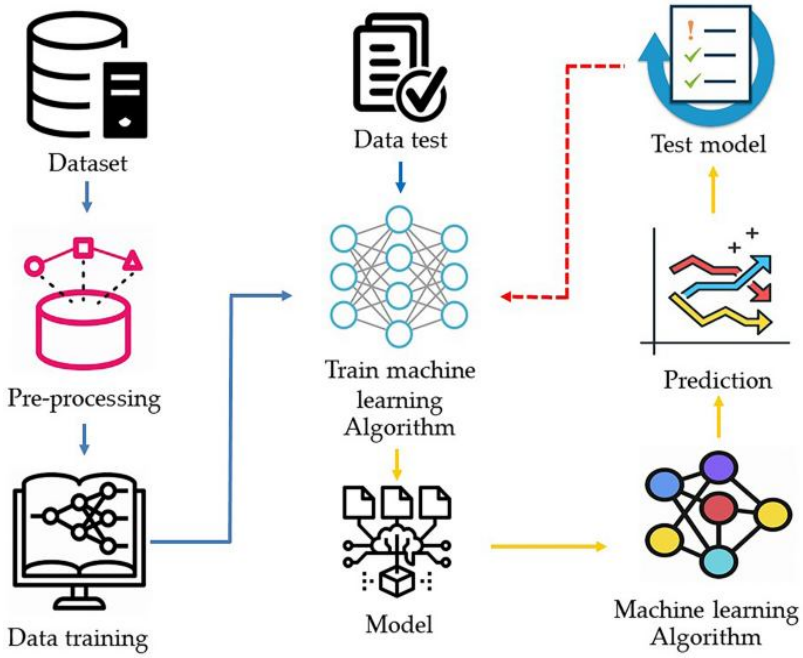


FIG3: SYSTEM ARCHITECTURE

V. REQUIREMENTS

A. Software Requirements

1) Software:

a) Python Version 3.0 or above

b) Django Framework

c) Jupyter Notebook

2) Operating System: Windows 10

3) Tools: Microsoft Visual Studio, Google Collab, Web Browser(Google Chrome/Firefox)

4) Python Libraries: NumPy, pandas, sklearn, matplotlib, seaborn, pickle

B. Hardware Requirements

1) CPU - 8 to 15 with each octa core processor in a distributed network.

2) RAM - 128 to 255 GB

3) Storage – 30 to 50 GB

VI. ANALYSIS

A. Regression Alaysis

1) Random Forest Regression: The basic steps involved in Random Forest algorithm is as follows:

a) Step 1: Start selecting the random samples from the given training dataset.

b) Step 2: Next, this algorithm will construct a decision tree for each sample using the decision tree algorithm. Then for each decision tree an outcome is obtained.

c) Step 3: Next voting will be performed for every result that is predicted.

d) Step 4: Now select the most voted result as the final prediction result.

2) Decision Tree Regression: Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values. 17 In scikit-learn python library, sklearn. Decision Tree Regressor module is used for carrying out Decision Tree regression.

B. Experimental Analysis

In science and engineering, experimental data is information obtained through a measurement, test technique, experimental design, or quasi-experimental design. Researchers of all stripes can replicate experimental data, and these data can then be subjected to mathematical analysis.

C. Cross Validation Score

One statistical method for determining the competence of machine learning models is cross-validation. Because it is simple to understand, simple to use, and produces skill estimates that typically have a lower bias than other methods, it is frequently used in applied machine learning to match and select a model for a given predictive modelling issue. It is also referred to as a resampling technique used to assess machine learning algorithms on a small sample of data. A more accurate measure of model quality is provided by cross-validation, which is crucial if you are making numerous modelling choices. Because it estimates numerous models, it occasionally takes longer to run. It is a well-liked technique because it is easy to comprehend and typically yields a less biased or overly optimistic assessment of the model skill than other techniques, like a straightforward train/test split.

D. Performance Measures

Prediction tool users should be able to comprehend the process of evaluation and how to interpret the findings. There are six primary performance evaluation metrics presented.

1) Sensitivity
2) Specificity
3) positive predictive value
4) negative predictive value
5) accuracy
6) and Matthew's correlation coefficient are a few of them.

a) *Accuracy:* The most logical performance metric is accuracy, which is just the proportion of accurately predicted observations to all observations. One might believe that our model is the finest if it has a high level of accuracy. Yes, accuracy is an excellent indicator, but only when the values of false positives and false negatives are nearly equal in symmetric datasets. As a result, you must consider other factors when assessing the success of your model. Our model's result was 0.803, which indicates that it is about 80% correct.

Accuracy = TP+TN/TP+FP+FN+TN.

b) *True Positives (TP):* These are the accurately predicted positive values, indicating that both the actual and predicted class values are true. For instance, if both the anticipated class and the actual class result show that this passenger survived, you would know the same thing.

c) *True Negatives (TN):* These are the accurately predicted negative values, indicating that both the actual and predicted class values are negative. E.g., If both the actual class and the anticipated class indicate that this passenger did not survive, the information is consistent.

d) *False Positives (FP):* are when the expected class is present but the actual class is absent. For instance, if the predicted class informs you that a passenger would survive but the actual class reports that the person did not survive.

e) *False Negatives (FN):* are when the expected class is no when the actual class is yes. For instance, if the predicted class predicts that the passenger would die but the actual class value shows that the passenger survived.

E. Performance Metrics

We can use the following indicators to gauge how effective our regression model is:

1) *R-squared:* This statistic shows how many variables, in relation to all the variables, the model predicted. Any potential biases in the data are not taken into account by R-squared. As a result, a good model may have a low R-squared value or a high R-squared value for a model that does not fit the data.
2) *Average Error:* The average error is the difference in numbers between the actual value and the value predicted.
3) *Mean Square Error (MSE):* useful if your data has a lot of outliers.
4) *Median Error:* The average of all discrepancies between projected and actual values is known as the median error.

5) *Average Absolute Error:* Similar to average error, but instead of using the difference's relative value to account for outliers in the data, use the absolute value instead.
6) *Median Absolute Error:* The mean of the absolute deviations between prediction and actual observation is known as the median absolute error. Large outliers may consequently have an impact on how the model is ultimately judged because all individual deviations are given the same weight.

F. Experimental Analysis

The comparison of the aforementioned models has led to the identification of the model that fits our system the best. So let's examine the performance of our model in relation to some sample data now. A sample of data points, sample ID, actual rating, and model-predicted rating are all displayed in the table below.

The system's performance for a sample of 5 data points is compared in the table below. The Random Forest Regressor Model, which fits our data the best, makes the predictions. We can see from our sample that the predicted and actual production values are not significantly different from one another.

Test Case Number	Actual Value	Predicted Value
01	18000	18000
02	7100	7100
03	9400	9400
04	7100	7310
05	500	500

VII. USER INTERFACE

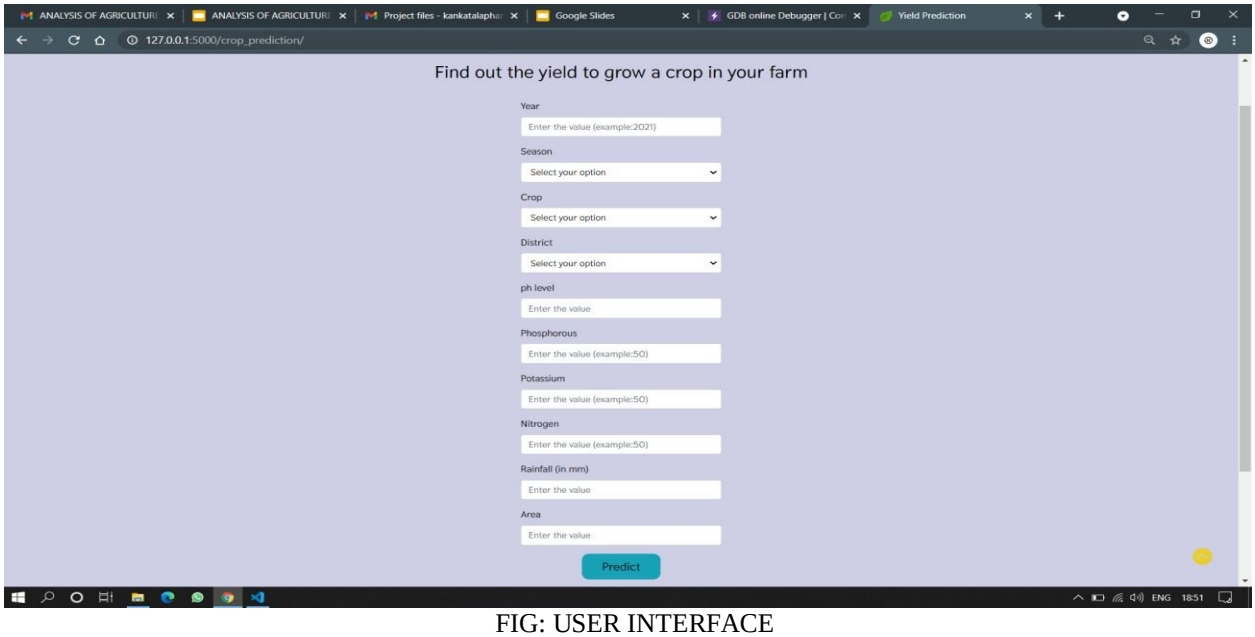


FIG: USER INTERFACE

VIII. CONCLUSION

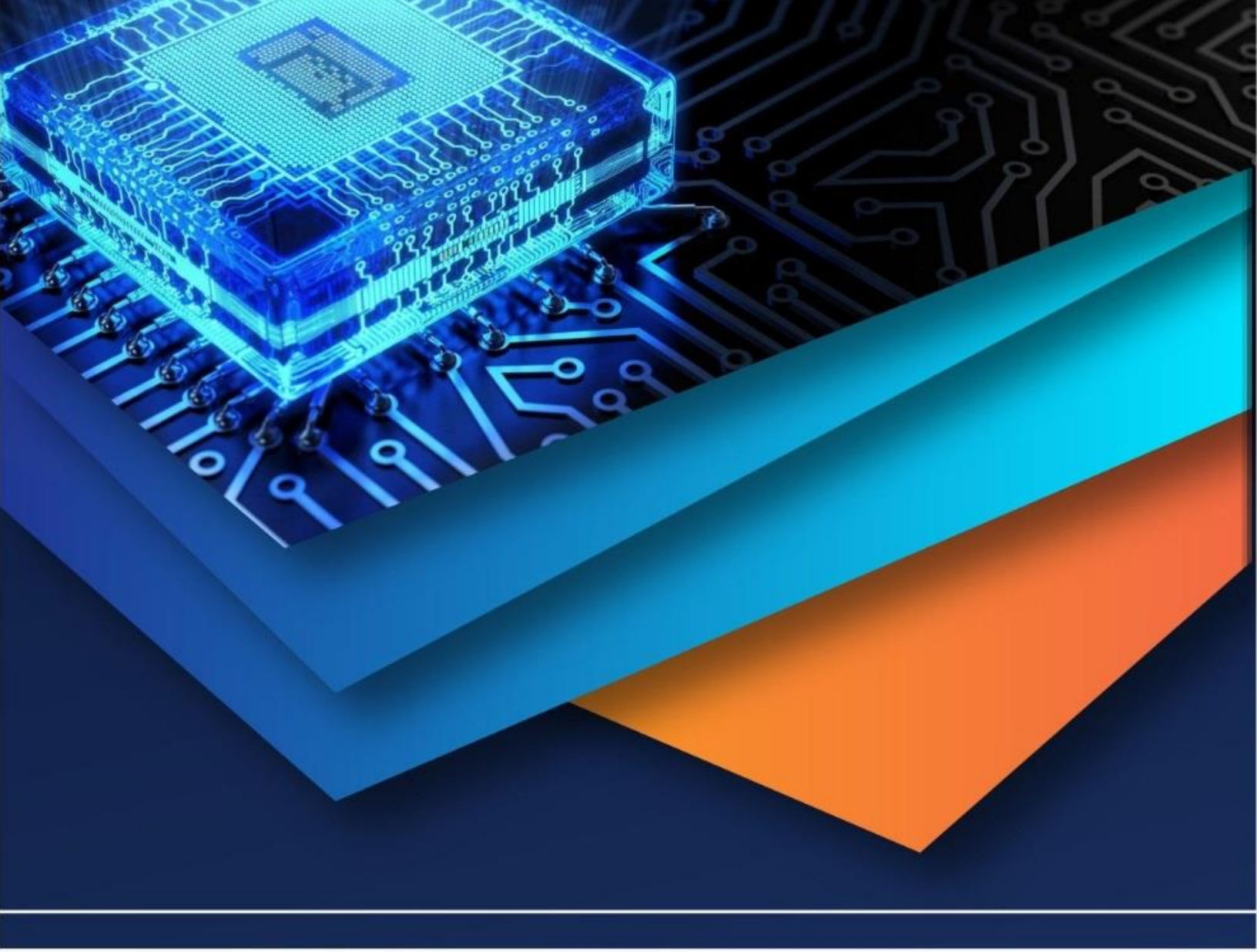
In order to choose the strategy that would produce the highest results, decision tree regression and random forest regression techniques are both applied to the input data. Performance indicators are used to compare these strategies. Both techniques appear to be effective based on metrics studies, but Random Forest regression provides a higher accuracy score on test data than Decision tree regression. The suggested study can be expanded to analyse the crop's climatic circumstances and other aspects in order to boost crop productivity.

IX. FUTURE WORK

Our model can be further trained in the future using new data points from various states. This system can also be expanded to accommodate various climate conditions. If we give the proposed model more precise data using satellite and sensor data, it can be used not only for our state but also for other states.

REFERENCES

[1] Liakos, Konstantinos G., Patrizia Busato, Dimitrios Moshou, Simon Pearson, and Dionysis Bochtis. "Machine learning in agriculture: A review." Sensors 18, no. 8 (2018): 2674.
[2] Benos, Lefteris, Aristotelis C. Tagarakis, Georgios Dolias, Remigio Berruto, Dimitrios Kateris, and Dionysis Bochtis. "Machine learning in agriculture: A comprehensive updated review." Sensors 21, no. 11 (2021): 3758.
[3] Vanitha, C. N., N. Archana, and R. Sowmiya. "Agriculture analysis using data mining and machine learning techniques." In 2019 5th international conference on advanced computing & communication systems (ICACCS), pp. 984-990. IEEE, 2019.
[4] Santos, Luís, Filipe N. Santos, Paulo Moura Oliveira, and Pranjali Shinde. "Deep learning applications in agriculture: A short review." In Robot 2019: Fourth Iberian Robotics Conference: Advances in Robotics, Volume 1, pp. 139-151. Springer International Publishing, 2020.
[5] Kamilaris, A. and Prenafeta-Boldú, F.X., 2018. Deep learning in agriculture: A survey. Computers and electronics in agriculture, 147, pp.70-90.
[6] Jagtap, Santosh T., Khongdet Phasinam, Thanwamas Kassanuk, Subhesh Saurabh Jha, Tanmay Ghosh, and Chetan M. Thakar. "Towards application of various machine learning techniques in agriculture." Materials Today: Proceedings 51 (2022): 793-797.
[7] https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it- 443b4a191c80
[8] https://link.springer.com/chapter/10.1007/978-3-319-14142-8_1
[9] https://link.springer.com/chapter/10.1007/978-3-319-14142-8_5
[10] https://scikit-learn.org/stable/modules/clustering.html#dbscan
[11] http://troindia.in/journal/ijcesr/vol4iss3/55-70.pdf
[12] http://www.apagrisnet.gov.in/
[13] http://www.apagrisnet.gov.in/2018/weekly/October/weekly_report_(Rabi)_05_21-11- 18.pdf
[14] https://desap.in/jsp/social/AGRICULTURALSTATISTICSATAGLANCE201819.pdf
[15] https://desap.in/jsp/social/SEASONANDCROPREPORT201819.pdf
[16] https://stackoverflow.com/questions/58983528/how-to-find-optimal-parametrs-for- dbscan
[17] https://medium.com/@tarammullin/dbscan-parameter-estimation-ff8330e3a3bd
[18] https://blog.exploratory.io/visualizing-k-means-clustering-results-to-understand-the- characteristics-of-clusters-better-b0225fb3d
[19] https://scikit-learn.org/stable/modules/model_evaluation.html



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)