International Journal of Emerging Trends in Science and Technology
10 (1) June 2024, 01-05
http://www.publishingindia.com/ijetst

Predictive Modeling of Agricultural Production Trends using Machine Learning: A Random Forest Approach

S. Geeitha[1*] and P. Renuka[2]
[1]Associate Professor, Department of Information Technology, M.Kumarasamy College of Engineering, Karur, Tamil Nadu, India. Email: geethas.it@mkce.ac.in
[2]Project Associate, Department of Information Technology, M.Kumarasamy College of Engineering, Karur, Tamil Nadu, India. Email: renu.mpn@gmail.com
[*]Corresponding Author

Abstract: Crop production forecasting is crucial for formulating strategies and allocating resource because it acts a crucial part in ensuring worldwide security of food. Utilising a collection of data that spans between 1961 to 2007, the present research compares predictive machine learning techniques for estimating crop outcomes across different nations. To make sure the collection of data, which included 311,624 items, could be used employing machine learning theories, it was preprocessed using the techniques of feature engineering and category encoding. We used the Random Forest Regressor and the Gradient Boosting Regressor, two sophisticated prediction models. Both the Random Forest Regressor with the Gradient Boosting Regressor, 2 powerful models for forecasting. Having an R-squared score of 1.00 & a Mean Squared Error (MSE) of 1.11*10^12, that implies nearly ideal accuracy in prediction, the Random Forest Regressor scored better than the other models. Excellent outcomes were achieved as well using the Gradient Boosting Regressor, but using slightly reduced precision measures.

Keywords: Artificial intelligence, Crop production, Gradient boosting regressor, Machine learning algorithm, Random forest regressor.

I. Introduction

It has become important to recognise and evaluate the numerous climatic and human impacts on land for farming usage for the purpose to arrange and handle it efficiently. It was necessary to determine the factors that could account for the utilisation of land used for agriculture for maize, wheat, and olive grove plants throughout the local level. Through the creation of a framework-agnostic methodology combined with an artificial intelligence model, we offer interpretations of some of the most important variables on both a worldwide and local level [1]. In this investigation, researchers identified and synthesized the features and techniques which are being applied in predicting crop yield research through the application of a systematic review of the literature. Researchers obtained 567 relevant research papers from six internet sources using search parameters. Around fifty research investigations have been chosen for further evaluation based on the criteria for inclusion and exclusion. We looked into several carefully selected researches, evaluated the aspects and methods working, and provided suggestions for more investigation [2]. Proposed work aids in determining the best practices for crop management and harvesting. It directs a person towards wise farming. The purpose of this effort is to assist a single person in cultivating crops well so they may get high yield at cheap expense. Additionally, it aids in estimating the overall cultivation costs [3]. More than ten years' worth of study in the fields has been examined using scientific information sources such as PubMed, Web of Science, and Scopus. It was noted that the use of artificial intelligence and Internet of Things to digitize farming has advanced beyond its early theoretical stage to the operational stage [4]. Different ML approaches may be utilised to solve different real-world issues, emphasizing how the effectiveness of a particular ML methodology depends on both the information being used in addition to the training techniques' effectiveness [5]. The use of artificial intelligence for agricultural purposes enables more effective, more accurate and profitable farming with less human laborer's [6]. The article will assist in realizing the Agri-stack goal of Indians & is unique throughout the way it recounts the tale of electronic integration in the Indian agricultural sector [7].

II. Literature Survey

Making actionable steps regarding more equal and environmentally friendly systems has grown into a worldwide concern in response to the difficulties presented by the phenomenon of global warming. Food and Agriculture's transformation to the fresh agro-food 4.0 paradigm would urge farmers and companies to make investments in artificial intelligence and robotics [8]. Following the Preferred Items for Reporting for Systematic

Assessments and Meta-Analysis strategy, the present research conducted a comprehensive examination of the research literature on machine learning technology used to farming [9]. The farming and processing of crops is a major worldwide issue, and computational intelligence has the ability to revolutionize the AgriTech sector in ways that are still being fully researched. The objective of this research is to look into the revolutionary possibilities for machine learning (ML) in agricultural methods and productivity development by providing a thorough assessment of field conditions [10]. These days, agricultural scientists and growers employ sensors to assist companies to enhance their farming practices. Farmers remotely track the fields using data from sensors sent through the Internet of Things. Under the context of sustainable farming, farmers nowadays oversee plants in controlled conditions to boost yields [11]. Effective agriculture forecasting of prices is crucial for ensuring an equitable and profitable expansion of farming, so it is an increasingly significant issue in the world of agriculture [12]. International research centers and colleges doing AI models studies on food security were additionally funded by foreign sponsors, while several collaborations and partnerships with local organisations were noted [13]. A model of neural networks that was developed on a dataset of important crops as well as their critical development features, including soils pH acts as the foundation for the agricultural product suggestion system. The fertiliser recommendation engine offers customized fertiliser by employing a rule-driven methodology [14]. Among potential solutions for those challenges, AutoML shines up. AutoML offers the ability to democratise machine learning (ML) tools by automated picking of designs, fine-tuning hyperparameters in and expediting the preparation of data, thereby allowing a broader range of professionals and academics to utilize them [15].

III. Proposed Methodology

A. Data Collection and Preprocessing

The dataset is downloaded from Kaggle where utilised for the research purpose. It has 311,624 items that describe agricultural productivity in different nations between 1961 and 2007. Many attributes, including country_or_area, element_code, element, year, unit, value, value_footnotes, and category, are included in the dataset. Preprocessing are taken place by Managing Missing Data by maintain the integrity of the data, all rows containing values that were unavailable were eliminated. A one-hot encoding method was employed to transform categorical information into integer form, including country_or_area, element, unit, category, and value_footnotes. For every category, this method produced a binary section, making it possible for the models used for machine learning to deal with these features efficiently. For the characteristic Choice, the goal factor and characteristics variables that are independent have been separated inside the collection of data. Every column other than score has been incorporated in the variables that were independent.

B. Model Selection

The forecasting position included picking out of two ensemble learning methods. Random Forest Regressor is a reliable and flexible system that specialises in handling big, extremely dimensional datasets. Throughout instruction, it builds several kinds of decision chains and delivers the average predict of each tree. The Gradient Boosting Regressor serves as an added powerful ensemble method which develops designs in an ordered manner, with each new version trying to rectify the errors of previous models. Gradient Boosting maintains a track record as being highly precise and flexible to complicated patterns of data.

C. Model Training and Evaluation

The dataset was split into a training set (80%) and a testing set (20%) using train_test_split from the scikit-learn library. The training set was employed to train both the Random Forest Regressor and the Gradient Boosting Regressor. To ensure consistency, hyper parameters like the total number of estimation techniques were maintained at 100 for the two models and the undefined state. Forecasting on the test set, forecasts were produced following learning.

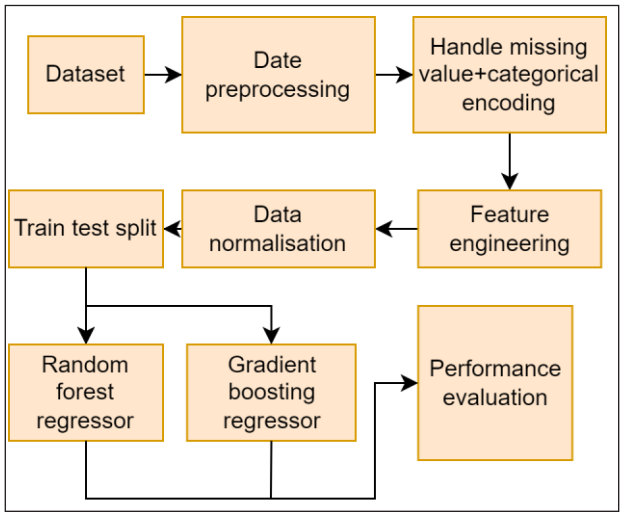

Fig. 1: Architecture of Proposed System

D. Performance Metrics

The simulations were assessed using the subsequent indicators, which have been computed as Mean Squared Error which indicates the forecasting error of the algorithm by measuring the mean squared variance between the real and anticipated values. The percentage of the variation in the dependent variable that can be predicted given the independent

variables is shown by the R-squared (R^2) measure. A flawless forecast is shown by a R^2 score of 1.5. Feature Importance Analysis. To figure out what characteristics having the biggest influence on the projections, characteristic significance was investigated.

Random Forest Model Attribute significance is a horizontal bar graph was used to display the characteristic significance that was obtained using the Random Forest method. Gradient Boosting Characteristic significance is using a comparable way, the Gradient Boosting algorithm's characteristic significance was depicted.

E. Comparison of Models

The MSE and R^2 values of the two systems were employed for evaluating the results they achieved. The Random Forests Regressor showed nearly ideal accuracy in predicting, having a R^2 score of 1.00 and an MSE of 1.11*10^12. The Gradient Boosting Regressor offered acceptable outcomes although having slightly less precise.

F. Visualization

The Worth of Features is to show the significance of each characteristic of the forecasting designs, graphic representations were created. Real compared to expect Results is to demonstrate the accuracy of the forecasting, scatter graphs were utilised for assessing actual and expected outcomes between the two models. Fig. 3 represents the actual versus predicted value of random forest and gradient boosting algorithms. Fig. 4 is the representation of yearly value by elements.

IV. Result

Depending upon its highly precise estimations, the random forest regression model seems to be an excellent fit for this set of data. Some categorical factors, which means period and nation-specific facts, possess a considerable impact on agricultural production results, based on the characteristic significance study. Though it wasn't as accurate in predictions compared to the Random Forest, a Gradient Boosting Regressor continues to be a good option for complicated, irregular patterns of data. Policymakers as well as researchers in this field can get valuable information through the methodology's outcomes, demonstrating the resilience of ensembles learning methods in modelling predictions for agricultural information.

The characteristic's significance as determined by a Random Forest model appears in the bar chart above as Fig. 2. "Country_ or_area" is the single most significant characteristic, next to "element_code" and "category." "Year" and "Element" were not the most important characteristics, indicating they have less have an impact upon the projections made by the model.

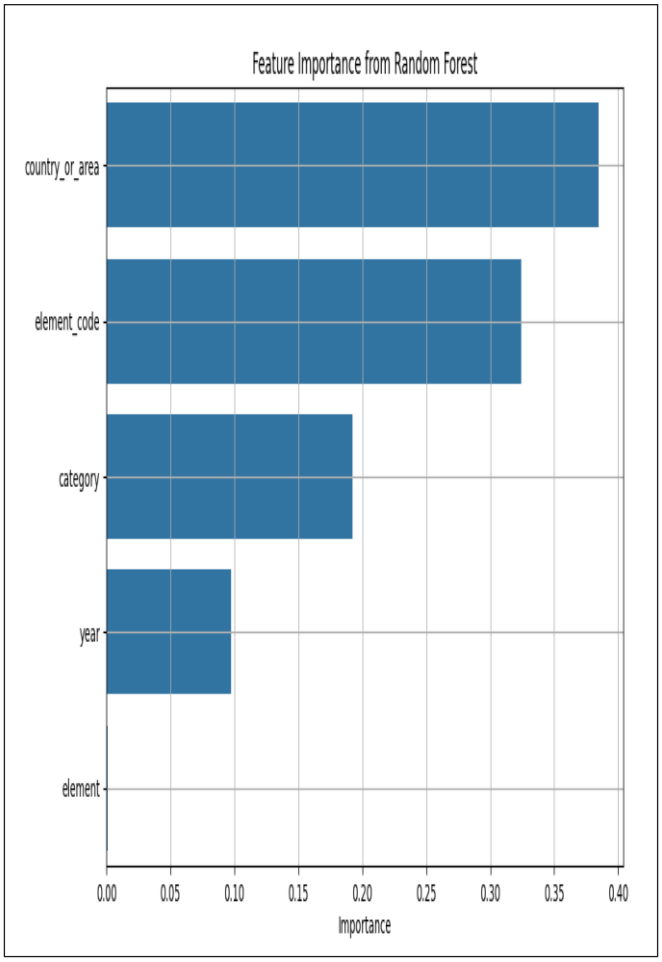

Fig. 2: Feature Importance of Random Forest Regressor

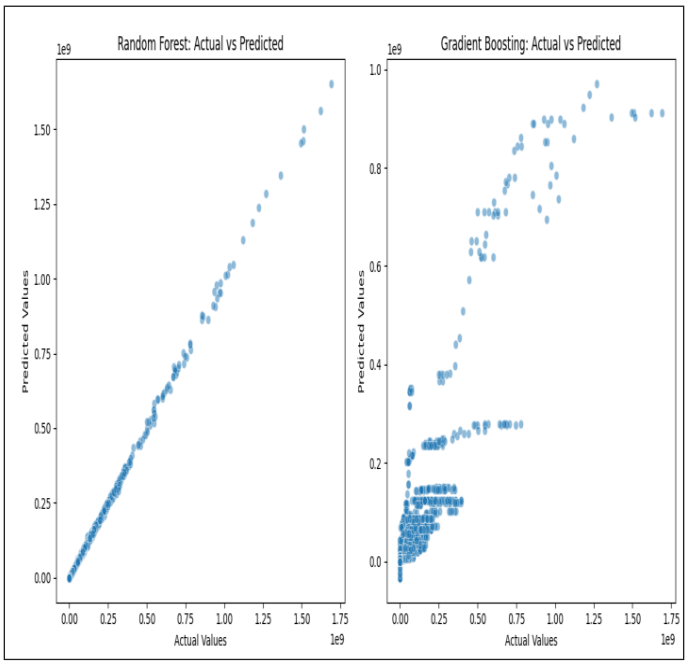

Fig. 3: Actual vs Predicted Values

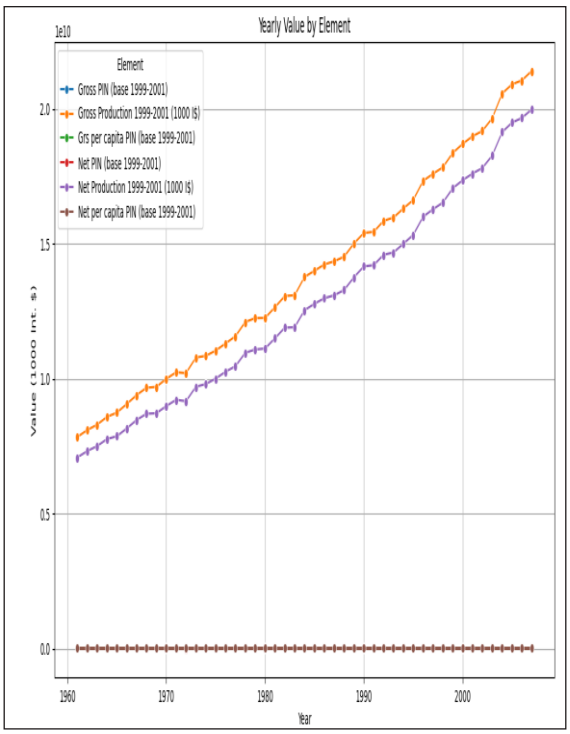

Fig. 4: Yearly Value by Element

Model Performance Random Forest Regression

```
Mean Squared Error: 1801194067331.04
R-squared: 1.00
```

Thus the model performance of random forest regression is represented by the mean of Mean squared error and R-Square.

Model Performance Gradient Boosting Regression

```
Mean Squared Error: 1572455248899.24
R-squared: 0.82
```

Thus the model performance of Gradient boosting regression is represented by the mean of Mean squared error and R-Square.

Thus the random forest regression performed well.

V. Conclusion

The goal of the study was to forecast agricultural production estimates based on previous statistics. Particularly, the Random Forest Regressor and Gradient Boosting Regressor were studied as realistic predictive machine learning algorithms. Researchers made ensured the algorithms are ready for effective instruction by preparing the datasets using features engineering and categorized labeling. Through an R-squared rating of 1.00 and a low mean squared error, the Random Forest Regressor easily surpassed the value of the Gradient Boosting Regressor, reaching almost perfect efficiency. Because of its capacity to handle highly dimensional data and resistance to an overfitting, this method is an appropriate fit for crop forecasting applications. The unique importance evaluation, that additionally emphasised the importance of past and enabled an improved comprehension of the variables affecting agricultural production. The effective application of these sorts of systems giving a helpful tool for analysation.

References

[1] C. M. Viana, M. Santos, D. Freire, P. Abrantes, and J. Rocha, "Evaluation of the factors explaining the use of agricultural land: A machine learning and model-agnostic approach," *Ecological Indicators*, vol. 131, p. 108200, 2021, doi: https://doi.org/10.1016/j.ecolind.2021.108200.

[2] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Computers and Electronics in Agriculture*, vol. 177, p. 105709, 2020, ISSN 0168-1699, doi: https://doi.org/10.1016/j.compag.2020.105709.

[3] S. K. S. Durai, and M. D. Shamili, "Smart farming using machine learning and deep learning techniques," *Decision Analytics Journal,* vol. 3, p. 100041, 2022, ISSN 2772-6622, doi: https://doi.org/10.1016/j.dajour.2022.100041.

[4] A. Subeesh, and C. R. Mehta, "Automation and digitization of agriculture using artificial intelligence and internet of things," *Artificial Intelligence in Agriculture,* vol. 5, pp. 278-291, 2021, ISSN 2589-7217, doi: https://doi.org/10.1016/j.aiia.2021.11.004.

[5] R. Pugliese, S. Regondi, and R. Marini, "Machine learning-based approach: Global trends, research directions, and regulatory standpoints," *Data Science and Management*, vol. 4, pp. 19-29, 2021, ISSN 2666-7649, doi: https://doi.org/10.1016/j.dsm.2021.12.002.

[6] V. Meshram, K. Patil, V. Meshram, Dinesh Hanchate, and S. D. Ramkteke, "Machine learning in agriculture domain: A state-of-art survey," *Artificial Intelligence in the Life Sciences*, vol. 1, p. 100010, 2021.

[7] A. Balkrishna, R. Pathak, S. Kumar, V. Arya, and S. K. Singh, "A comprehensive analysis of the advances in Indian digital agricultural architecture," *Smart Agricultural Technology*, vol. 5, p. 100318, 2023, ISSN 2772-3755, doi: https://doi.org/10.1016/j.atech.2023.100318.

[8] A. A. Mana, A. Allouhi, A. Hamrani, S. Rehman, I. el Jamaoui, and K. Jayachandran, "Sustainable AI-based production agriculture: Exploring AI applications and implications in agricultural practices," *Smart Agricultural Technology*, vol. 7, p. 100416, 2024, ISSN 2772-3755, doi: https://doi.org/10.1016/j.atech.2024.100.

[9] R. C. de Oliveira, and R. D. de Souza e Silva, "Artificial intelligence in agriculture: Benefits, challenges, and trends," *Applied Sciences*, vol. 13, no. 13, p. 7405, 2023, doi: https://doi.org/10.3390/app13137405.

[10] Aashu, K. Rajwar, M. Pant, and K. Deep, "Application of machine learning in agriculture: Recent trends and future research avenues," 2024. *arXiv*:2405.17465.

[11] A. Reyana, S. Kautish, P. M. S. Karthik, I. A. Al-Baltah, M. B. Jasser, and A. W. Mohamed, "Accelerating crop yield: Multisensor data fusion and machine learning for agriculture text classification," *IEEE Access*, vol. 11, pp. 20795-20805, 2023, doi: https://doi.org/10.1109/ACCESS.2023.3249205.

[12] F. Sun, X. Meng, Y. Zhang, Y. Wang, H. Jiang, and P. Liu, "Agricultural product price forecasting methods: A review," *Agriculture*, vol. 13, no. 9, p. 1671, 2023, doi: https://doi.org/10.3390/agriculture13091671.

[13] R. Sarku, U. A. Clemen, and T. Clemen, "The application of artificial intelligence models for food security: A review," *Agriculture*, vol. 13, no. 10, p. 2037, 2023, doi: https://doi.org/10.3390/agriculture13102037.

[14] C. Musanase, A. Vodacek, D. Hanyurwimfura, A. Uwitonze, and I. Kabandana, "Data-driven analysis and machine learning-based crop and fertilizer recommendation system for revolutionizing farming practices," *Agriculture*, vol. 13, no. 11, p. 2141, 2023, doi: https://doi.org/10.3390/agriculture13112141.

[15] D. Tamayo-Vera, X. Wang, and M. Mesbah, "A review of machine learning techniques in agroclimatic studies," *Agriculture*, vol. 14, no. 3, p. 481, 2024, doi: https://doi.org/10.3390/agriculture14030481.