Development of a Crop-Based Predictive System for Optimizing Crop Yield for Farmers

KINGA MARY TEMIDAYO[1], OLUTAYO KEHINDE BOYINBODE[2], OLAWALE SOLOMON AKINTOLA[3]

[1, 2, 3]*Department of Information Technology, The Federal University of Technology, Akure, Ondo State, Nigeria.*

Abstract- Crop yield prediction is essential for helping farmers make informed decisions, optimize resource allocation, and enhance productivity. This study presents a machine learning-based predictive system that leverages an ensemble approach, combining Bagging Regressor, XGBoost, and Random Forest models to forecast crop yields. Utilizing a comprehensive dataset that includes vital crop parameters such as crop type, area, temperature, rainfall, and soil type, the system provides valuable insights into the factors influencing predicted yields. By identifying key determinants, the system enables farmers to optimize crop management strategies, leading to improved productivity, reduced uncertainty, and enhanced sustainability. The proposed system empowers farmers with data-driven decision-making tools, contributing to food security and promoting sustainable agricultural practices.

Indexed Terms- Crop yield Prediction, Ensemble Algorithm, Farmers, Machine Learning, Productivity

I. INTRODUCTION

In agriculture, predicting and optimizing crop yield is essential for ensuring food security and economic stability for farmers. Leveraging machine learning techniques, particularly ensemble algorithms such as Random Forest, Bagging Regressor, and XGBoost, can significantly enhance the accuracy of predictive models for crop yield optimization. This document presents a crop-based predictive system that integrates these ensemble algorithms to provide farmers with actionable insights and recommendations for maximizing crop yield.

Ensemble algorithms enhance prediction accuracy and stability by combining the outputs of multiple base models. Random Forest creates multiple decision trees and averages their outputs, which helps reduce overfitting and makes the model applicable across various datasets. Bagging Regressor builds models on distinct subsets of data, reducing variance and preventing dependence on specific features or noisy data. XGBoost (Extreme Gradient Boosting) is another powerful technique that uses gradient boosting to reduce errors by learning from prior models' mistakes, making it effective for large datasets and managing missing values, making it a favorite for agricultural yield prediction [1],[2].

Despite their strengths, ensemble methods have downsides. They can be computationally expensive, especially with large datasets or complex hyperparameter tuning. Random Forest and XGBoost can also struggle with extreme outliers or rare cases, resulting in less accurate predictions for these situations [3]. Additionally, their complexity makes interpretability challenging, as it is difficult to trace how specific predictions are made within an ensemble of trees or models [4]. Nevertheless, they remain effective tools for improving prediction accuracy, providing useful insights for farmers in agricultural decision-making.

II. LITERATURE REVIEW

A. Requirement of integrating the ensemble algorithm in agriculture

Integrating ensemble algorithms such as Random Forest, XGBoost, and Bagging Regressor into agriculture holds great potential for enhancing crop yield predictions. These advanced algorithms are designed to handle complex datasets, capturing the multifaceted relationships between variables such as soil properties, rainfall, temperature, and crop type. Random Forest aggregates decisions from multiple trees, effectively capturing interactions among features [5]. XGBoost, known for its efficiency and accuracy, iteratively improves predictions by focusing on minimizing errors in previous predictions [6]. Bagging Regressor stabilizes performance by reducing variance through averaging predictions from multiple models [Islam]. By combining these models in an ensemble, the system can reduce prediction errors and enhances robustness, resulting in more reliable yield forecasts [4]. This helps farmers make informed, data-driven decisions about resource allocation, crop management, and risk mitigation in response to environmental variables like weather or soil conditions. The integration of ensemble algorithms also supports sustainable farming practices, allowing farmers to optimize productivity while minimizing uncertainties in yield outcomes [7]. Ultimately, such predictive systems contribute to food security by providing actionable insights that improve agricultural efficiency, ensuring that farmers can sustainably meet growing demands [1].

B. Importance of ensemble algorithms in crop yield prediction

The importance of ensemble algorithms in crop yield prediction lies in its ability to enhance prediction accuracy and robustness by combining the strengths of multiple models. Ensemble methods such as Random Forest, XGBoost, and Bagging Regressor are particularly useful in agriculture, where the interplay of various factors—such as soil properties, rainfall, temperature, and crop types—can influence yield outcomes [8].

By integrating multiple algorithms, ensemble methods reduce the risk of overfitting, improve generalization, and minimize prediction errors. For instance, Random Forest captures complex interactions between features [5]. XGBoost improves prediction accuracy through iterative learning [6], and Bagging Regressor reduces variance by aggregating predictions from different models [Islam]. These features make ensemble algorithms ideal for analyzing large agricultural datasets that include diverse and fluctuating variables. The application of ensemble algorithms in agriculture provides farmers with more reliable and actionable insights into yield predictions. This helps in optimizing resource allocation, improving crop management, and mitigating risks from environmental factors such as adverse weather conditions [9]. As noted in this research, such predictive systems empower farmers to make informed decisions that lead to increased productivity and sustainability, contributing to food security and better economic outcomes for small-scale farmers [3].

III. METHODOLOGY

A. RANDOM FOREST ALGORITHM

Random Forest constructs multiple decision trees during training and averages the predictions for more accurate results. Each tree is built using a random subset of the training data, which helps capture the intricate interactions between various factors such as soil type, rainfall, and temperature [5].

Random Forest excels in managing high-dimensional data and can handle missing values, making it particularly suitable for agricultural applications where data variability is common [10]. By aggregating the predictions of multiple trees, Random Forest reduces variance and improves model robustness, minimizing the risk of overfitting [11].

Farmers can rely on these accurate predictions to make informed decisions about crop management and resource allocation, leading to enhanced productivity. The ability of Random Forest to provide reliable insights into expected crop yields helps farmers navigate the uncertainties posed by environmental factors, ultimately contributing to more sustainable farming practices and improved food security [12].

B. BAGGING REGRESSOR ALGORITHM

Bagging Regressor enhances prediction accuracy through the bagging technique, or Bootstrap Aggregating. This method creates multiple subsets of the training data using random sampling with replacement, training separate regression models—often decision trees—on each subset [Rashid]. By averaging the predictions of these models, the Bagging Regressor effectively reduces variance and improves model stability [2].

This is particularly advantageous in agriculture, where datasets can be noisy and contain outliers due to environmental variability. Bagging mitigates the risk of overfitting that can occur when relying on individual models, providing a more generalized prediction that accounts for fluctuations in factors like soil quality and weather conditions [4].

Farmers benefit from the ability of Bagging Regressor to produce accurate insights into expected crop yields, allowing for better resource management and planning [Raja].This leads to improved productivity and sustainability in farming practices, as farmers can adapt their strategies based on reliable forecasts, ultimately enhancing food security and economic stability [7].

C. XGBOOOST ALGORITHM

XGBOOST, or Extreme Gradient Boosting, is an advanced ensemble technique that builds models sequentially, focusing on correcting the errors of previous models. This iterative approach captures complex relationships between variables such as soil characteristics, weather, and crop [6].

By minimizing a specified loss function using gradient descent, XGBoost efficiently optimizes predictions while incorporating regularization techniques to prevent overfitting [2]. This is particularly beneficial in agriculture, where data variability and noise can affect the reliability of yield predictions.

XGBoost's ability to handle missing values and its robustness against overfitting make it suitable for diverse agricultural datasets [8]. The model provides farmers with reliable insights, empowering them to make data-driven decisions regarding crop management and resource allocation [9].

Ultimately, the integration of XGBoost into crop yield prediction systems enhances the accuracy of forecasts, allowing farmers to adapt their practices to maximize productivity and sustainability, thereby contributing to food security and economic viability in agriculture [4].

D. PROPOSED WORKFLOW OF THE MODEL

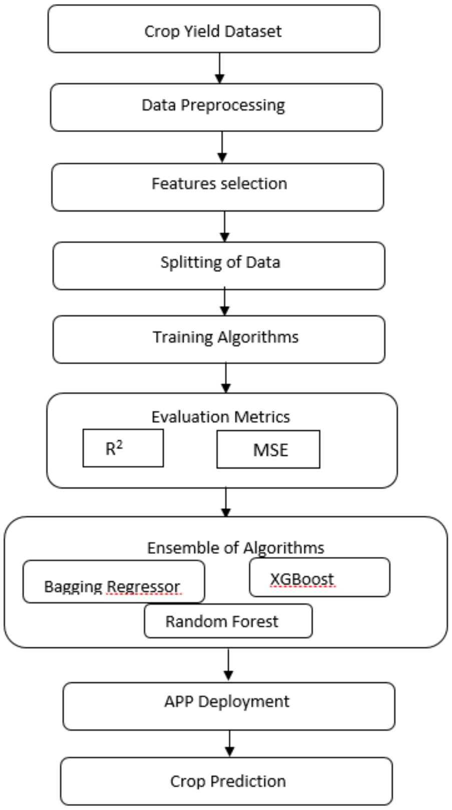


Fig. 1. Working flow of the module

From the Fig.1 flowchart, it can be seen that the data gotten from Kaggle (crop yield dataset) is preprocessed at the next level. This implies that missing values has be filled, datatypes has to be changed to the original datatype for each column and these ensure data integrity and cleanliness.

After preprocessing, features has to be selected to ensure that appropriate inputs are given to the respective modules.

Furthermore, data is split into training and testing datasets. In the next step, machines are to be trained through models. The performance of each model is evaluated by R-squared and Mean Square Error metrics. In the next step, I ensemble the models, (Random forest, Bagging Regressor, XGBoost) because of their high accuracy and performance to estimate the crop yield.

In the next step, the ensemble algorithm is deployed into an app and finally, a crop yield prediction is done in real life.

E. SYSTEM ARCHITECTURE

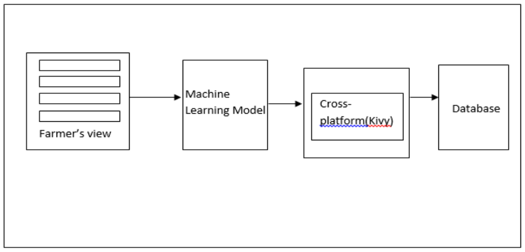


Fig.2. System architecture of the crop yield predictive system

From Figure 2, it is clear that the farmer has a user interface where he or she type into the inputs given by the system. Then after the input collection, the machine learning model is loaded by the app from its database to make the prediction. Then the output is displayed to the farmer (the end user).

IV. RESULTS AND VISUALIZATION

R² measures how well the independent variables explain the variance in the dependent variable, which, in this case, would be crop yield. It ranges from 0 to 1, where values closer to 1 indicate a strong explanatory power of the model, meaning the model captures more variance in the data. Models with a high R² are more reliable for making crop yield predictions, helping farmers optimize decision-making [8], [5].

On the other hand, MSE quantifies the average squared difference between the actual and predicted values, offering a measure of the model's prediction accuracy. Lower MSE values signify a better fit between the model's predictions and real-world crop outcomes, indicating higher precision. MSE is particularly important when working with noisy or variable agricultural data, where even small inaccuracies can lead to significant deviations in crop yield estimates [9], [2].

These metrics are critical for assessing the performance of various machine learning models such as Random Forest, XGBoost, and Bagging Regressor, which are frequently used in agriculture to handle complex and large datasets [4], [1]. Both metrics provide insight into the model's strengths and limitations, allowing for improvements and refinements in future crop yield prediction systems.

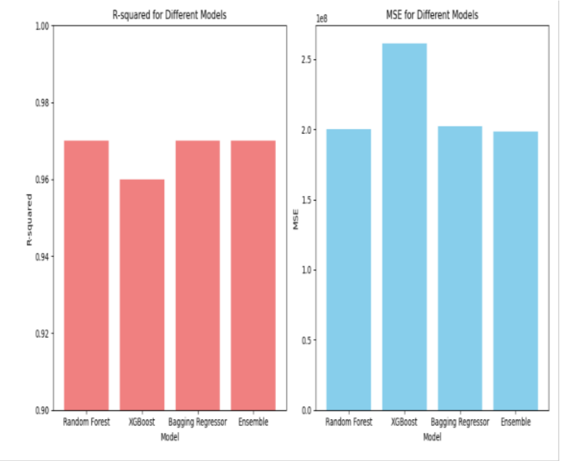
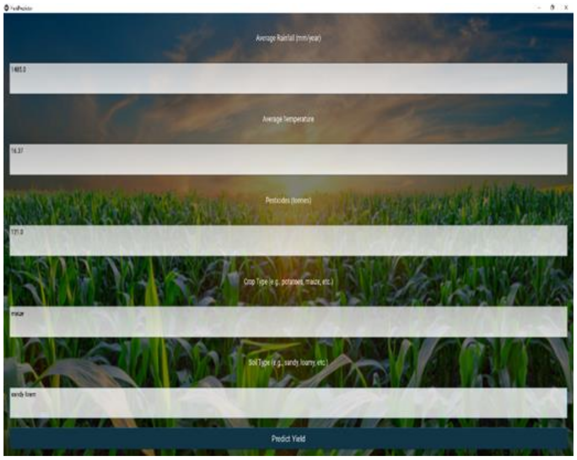


Fig.3. The evaluation metrics (R^2 and MSE) result of the trained models and their ensemble model

Figure 3 compares the performance of four machine learning models (Random Forest, XGBoost, Bagging Regressor, and Ensemble) using two metrics: R-squared and Mean Squared Error (MSE). While all models show high R-squared values between 0.96 and 0.97, XGBoost has a slightly lower R-squared and a notably higher MSE compared to the other models, which perform similarly across both metrics.



Average Rainfall (mm/year)

1485.0

Average Temperature

16.37

Pesticides (tonnes)

121.0

Crop Type (e.g., potatoes, maize, etc.)

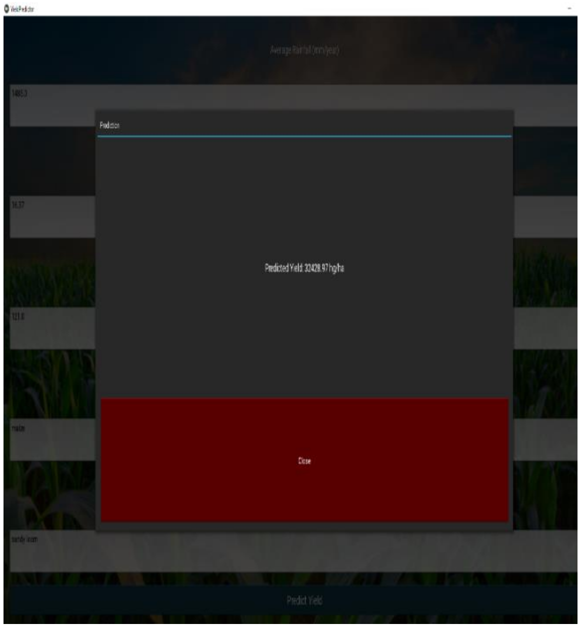
maize

Soil Type (e.g., sandy, loamy, etc.)

sandy loam

Fig.4 Crop Prediction interface with different parameters

Figure 4 shows the App interface of the crop based predictive system where different parameter of bio system are collected as inputs. The inputs contain Average Rainfall (in mm/year), Average Temperature (in Celsius), Pesticides (in Tonnes), Crop Type, Soil Type.



1485.0

16.37

121.0

maize

sandy loam

Predicted Yield: 32428.97 hg/ha

Fig.5 Crop Prediction interface with the predicted yield

Figure 5 shows a pop up interface displaying that 32428.97 hg/ha amount of maize will be produced if harvesting is done under the provided circumstances of bio system.

CONCLUSION

Through the use of ensemble approaches, this study successfully built a crop production prediction system that farmers in various nations may use to maximize agricultural productivity. Improving lives and food security worldwide need. Farmers no longer have to rely only on their own experience to anticipate crop production because the study fills in the technical knowledge gap about crop selection and farm productivity.

Furthermore, by taking into consideration the different biosystem factors at different sites, the machine learning ensemble method addresses a number of issues with crop production prediction. The system addresses the drawbacks of using a single algorithm method by integrating numerous algorithms, which increases its adaptability to various environmental situations.

REFERENCES

[1] S. Sikandar, R. Mahum, & S. Aladhadh, (2022). Automatic Crop Expert System using an Improved LSTM with Attention Block, CSSE, 2023, 47(2), 2017-2022. https://doi.org/10.32604/CSSE.2023.037723

[2] A. Islam, R.A. Ifty, M.N. Arefin, I. Khair, S. Hossain, & J.A. Muhammed, (2023). Ensemble Machine Learning Approach for agricultural crop selection, International Conference on Electrical, Computer and Communication Engineering. https://doi.org/10.1109/ECCE57851.2023.10101585

[3] S. Gowda, & S. Reddy, (2020) Design and implementation of crop yield prediction model in agriculture, International Journal of Scientific & Technology Research, 8(1), 545 – 547.

[4] V. Suraparaju, S. Ujjainia, & P. Gautam, (2021). A crop recommendation System to improve crop productivity using an ensemble technique, Internatonal Journal of Innovative Technology and Exploring Engineering, 10(4), 102-104. https://doi.org/10.35940/IJITEE.D8507.0210421

[5] S.P. Raja, B. Sawicka, Z. Stamenkovic and M. Ganesan, (2022). Crop prediction based on agricultural environmental characteristics using various feature selection techniques and classifiers, 10, 23630. https://doi.org/10.1109/ACCESS.2022.3154350

[6] D. Garg, M. Alam, (2023). An effective crop recommendation method using machine learning techniques. Internatonal Journal of Advanced Technology and Engineering Exploration, 10(102), 501 – 504. http://dx.doi.org/10.19101/IJATEE.2022.10100456

[7] H.M. Monisha, J. K. Dhanush, S.M. Adarsha, & M. Dalli,(2024). A web-based crop recommendation system using various machine learning algorithms. International Journal of

Novel Research and Development, 9(2), C162 – C164. https://doi.org/IJNRD.2024.378469022

[8] M. Rashid, B.S. Bari, Y.B. Yusup, M.A. Kamaruddin, & N. Khan, (2021) A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction, IEEE Transactions and Journals, 4(1), 8-12. https://doi.org/10.1109/ACCESS.2021.3075159

[9] A. Ahmed, S.A. Adewunmi, V. Yemi-Peters, (2023). Crop yield prediction in Nigeria using machine learning techniques: (A case study of Southern part of Nigeria), UMYU Scientifica, 2(4), 31-38. https://doi.org/10.56919/usci.2324.004

[10] T.G. Liliane, & S.C. Mutengwa, (2020) Factors affecting yield of crops, Agronomy – Climate change

[11] and Food Security, pg.1. https://dx.doi.org/10.5772/intechopen.90672

[12] M. Harris, (2023, September 19). *Understanding Random Forest Methods*. Stats with R. Retrieved October 16, 2024, from https://www.statswithr.com

[13] A. Ahmed, S.A. Adewunmi, V. Yemi-Peters, (2023). Seasonal crop yield prediction using machine learning techniques (A case study of Northern Nigeria), FUW Trends in Science and Technology Journal, 8(2), 332-336. https://doi.org/10.48185/jaai.v4i1.728