Crop Yield Prediction Using Machine Learning Algorithms

Florida Atlantic University, *777 Glades Rd Boca Raton,Florida,USA.*

~ ***Rajeev Pondala.***

Abstract—**Agriculture is one of the major and the least paid occupation in India. Machine learning can bring a boom in the agriculture field by changing the income scenario through growing the optimum crop. This paper focuses on predicting the yield of the crop by applying various machine learning techniques. The outcome of these techniques is compared on the basis of mean absolute error. The prediction made by machine learning algorithms will help the farmers to decide which crop to grow to get the maximum yield by considering factors like temperature, rainfall, area, etc.**

Index Terms—**Crop yield prediction;long short-term memory(LSTM);simpleRNN; random forest;xgboost;machine learning classifiers;ensemble learning.**

I. Introduction

IN the modern world of agriculture, the predictive analysis of crop growth has emerged as a crucial area due to its significant effects for healthy farming practices, resource allocation, and food security. The diversity in food growth is affected by multiple factors, including weather conditions, land traits, and farming methods.

II. Objective of the Analysis

The main goal of this study is to apply advanced prediction models to project crop yields based on thorough data integration. By amalgamating past crop yield data, climate factors like rainfall and average temperature, along with poisons information, the goal is to create solid models capable of accurate yield forecasts. Leveraging machine learning techniques and methods, this study aims to offer a reliable forecasting tool for farmers and players in the agriculture domain. Through the application of improved data preparation techniques, feature engineering methods, and the usage of various machine learning models, this study attempts to explore new ways to enhance the accuracy of crop yield predictions. This part sets the stage for understanding the importance of this project within the farming area, stressing the purpose and goals driving the following sections of this report.

III. Analysis and Design

The study performed on farming data covering several decades provides useful insights into the factors affecting crop growth, stressing the relationship between climate variables, pesticide usage, and food production.

This thorough study utilized diverse data sets spanning yield information for major crops, weather trends, pesticide usage, and average temperature records across numerous countries from 1961 to 2016. The initial steps of data gathering involved compiling detailed records of food yield, rainfall, herbicides, and weather, followed by careful cleaning and joining of data sets based on country and year, creating a combined data set for analysis.

The research part of the study showed interesting trends and connections. India appeared as a top creator in food growth, especially thriving in the farming of maize and potatoes. This shows the regional differences in farming output, possibly affected by a number of factors including land quality, irrigation methods, and government policies. Moreover, analysing the association matrix and image showed the complicated interactions between factors. Surprisingly, no strong relationships were found between crop growth and factors like rainfall, herbicides, or weather, indicating a subtle connection that deserves further study.

To prepare the data for predictive modeling, important preparation steps were performed. Categorical factors such as countries and food types were encoded using One-Hot Encoding to ease their inclusion in machine learning models. Additionally, feature scaling was applied to standardise the size of different features, ensuring fair model comparisons and avoiding errors due to varying scales. The careful preparation of the data sets lays the groundwork for strong and accurate predictive modeling aimed at understanding and projecting crop yield trends based on a multitude of contributing factors. Moving on to modeling and evaluation, the data sets was divided into training and testing groups following an 80/20 split. Several regression models were applied to discover their effectiveness in predicting crop yield based on factors like rainfall, herbicides, and weather. The models were tested using R2 scores, giving a quantified measure of their success. Comparing the performance of different models allowed the identification of the most ideal method for predicting crop growth within this particular context.

IV. Literature Survey

Previous Research and Methodologies in Crop Yield Prediction A thorough examination of earlier research attempts in crop yield forecast gave useful insights into diverse methods, frameworks, and important factors driving correct yield estimations. This review covered a wide selection of educational articles, study papers, and industry reports, hoping to

This paper was produced by the IEEE Publication Technology Group. They are in Piscataway, NJ.

Manuscript received April 19, 2021; revised August 16, 2021.

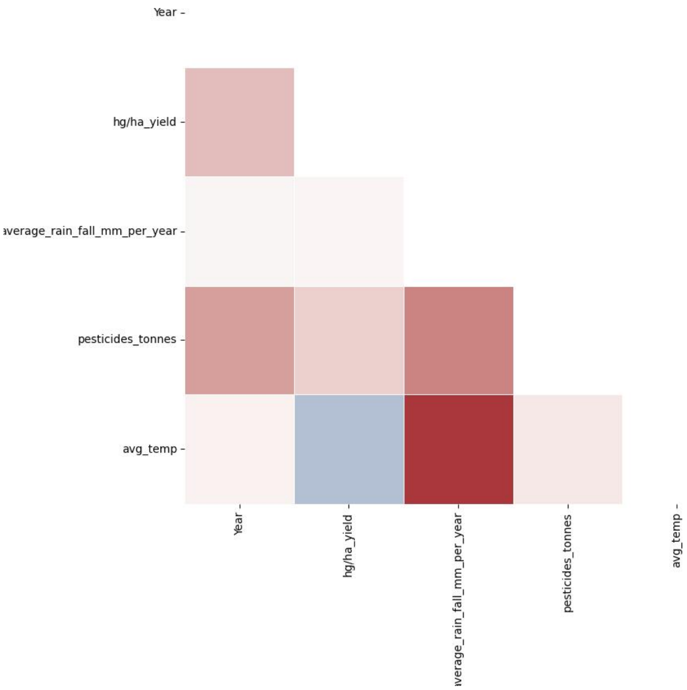

Fig. 1. Heatmap with mask Correct Aspect Ratio

understand the development of prediction models and their effectiveness in agriculture yield estimates. Comparison of Machine Learning Models Used in Similar Studies An thorough comparison analysis was performed to evaluate different machine learning models applied in similar studies focused on crop yield forecast. The comparison involved an in-depth study of the strengths, flaws, and success measures connected with different models. This scrutiny helped the discovery of the most suitable methods and tools for implementation in the current project.

Crop Yield Prediction The study into crop yield prediction methods involved a review of past trends, predicted features, and factors affecting crop production. Various statistical and machine learning methods utilized for yield prediction were examined, stressing their usefulness in recording complex links between weather factors, soil characteristics, and crop yields.

Supervised Machine Learning The study of guided machine learning algorithms focused on their application and success in crop yield forecasts. Diverse algorithms such as Random Forest, Support Vector Machines (SVM), Gradient Boosting, and Neural Networks were studied for their flexibility to farming datasets and their potential to create correct results.

V. Algorithm

The method chosen in this study, called as ”rfDNN,” offers a new ensemble technique that combines the powers of two different machine learning models: Random Forest Regressor and Deep Neural Network (DNN). This merger tries to boost the accuracy of crop production forecast by exploiting the matching features of both methods.

The Random Forest Regressor is a flexible method that works by making a variety of decision trees during training. Each tree separately guesses the result, and the end forecast is produced using an ensemble of these unique estimates. This method is immune against overfitting and works well even with big datasets. It works by randomly picking subsets of traits and cases for each tree, so increasing variety and lowering bias in forecasts.

In contrast, the Deep Neural Network (DNN) is a family of artificial neural networks marked by several buried layers between the input and output layers. These secret layers help the network to find complex patterns and representations within the data. DNNs shine in feature learning and generalisation, allowing them to understand difficult relationships in the input data, especially in unorganised or high-dimensional datasets.

The ”rfDNN” ensemble model uses the combined power of these two methods. Firstly, the Random Forest Regressor creates an array of decision trees, each learned on various groups of the dataset. Subsequently, the result of these trees is applied as features for the Deep Neural Network. This approach helps the DNN to learn from the different forecasts made by the Random Forest, easily harnessing the vast ensemble information.

During the projection phase, the rfDNN model combines the results of the Random Forest and DNN to provide the final estimate of food production. By combining the prediction power of both models, rfDNN tries to achieve better accuracy, resistance against noise, and greater generalization on new data.

The rfDNN algorithm stands out due to its ability to successfully combine the strengths of group methods, represented by Random Forest, with the feature learning skills inherent in Deep Neural Networks. This mixed method tries to rely on the different traits of each model, giving a possible technique for accurate crop production prediction, which is vital for agriculture planning, output rise, and economic security in the farming sector.

A. Architecture

1) Random Forest Regressor Ensemble: The first step includes creating an array of decision trees with the Random Forest Regressor. Each decision tree is trained on different parts of the dataset, employing random feature selection and instance sampling. These trees separately expect food production based on many factors and provide unique predictions. Feature Extraction and DNN Integration:

The results from the collection of decision trees made by the Random Forest are collected as features for the second step, which includes the Deep Neural Network (DNN). These guesses serve as inputs to the DNN model, acting as rich, diverse features learnt from the ensemble. Deep Neural Network (DNN):

The DNN design comprises numerous secret levels for feature learning and simplification. It gets the returned features from the Random Forest ensemble as input and processes them via its layers to record complex patterns and relationships within the data. The DNN tries to improve the forecasting skills by learning from the various group forecasts. Integration and Prediction Phase:

The last step includes combining the results of both the Random Forest group and the Deep Neural Network. The rfDNN model uses the various guesses from the Random Forest with the learnt representations from the DNN to make the final forecast for food production. Key Architectural Aspects: Ensemble Learning: The design capitalizes on ensemble learning by mixing results from numerous decision trees in the Random Forest ensemble, improving model endurance and performance.

Feature Integration: The feature extraction method offers easy integration of different estimates made by the Random Forest into the following DNN model, allowing the DNN to learn from multiple input sources.

Deep Learning Capabilities: The use of a DNN helps the capture of complicated patterns and relationships in the dataset, leveraging several hidden layers for feature abstraction and learning high-level representations.

Hybrid Model Fusion: The final merger step blends the powers of the Random Forest ensemble with the DNN to produce a better prediction of food yield, combining the benefits of both models.

This architectural design allows the rfDNN model to leverage the powers of ensemble methods and deep learning, giving a collaborative approach that aims to achieve better accuracy and robustness in crop yield forecast, important for farming planning and production optimization.

B. Methodology Data Preprocessing

The files receive thorough planning to handle missing values, errors, and inconsistencies. Cleaning includes estimation methods, outlier recognition, and normalization to maintain data quality and uniformity.

Feature Engineering Features are carefully picked and designed from the datasets to help successful model learning. Feature engineering includes pulling important qualities, creating new features, and storing category factors for better representation.

Model Training and Validation The processed data is partitioned into training and confirmation sets. The model gets training using the training dataset, and hyperparameter change is performed to improve model performance. The test set studies the model's extension and predictive skills.

Ensemble Model Creation - rfDNN The ensemble model ”rfDNN,” combining Random Forest Regressor and Deep Neural Network, is built. The Random Forest ensemble makes various guesses, which are applied as features for the following DNN.

Model Evaluation and Comparison The success of the rfDNN model is measured using several measures such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared score. Comparative study with different standard models and methods is performed to prove the model's usefulness.

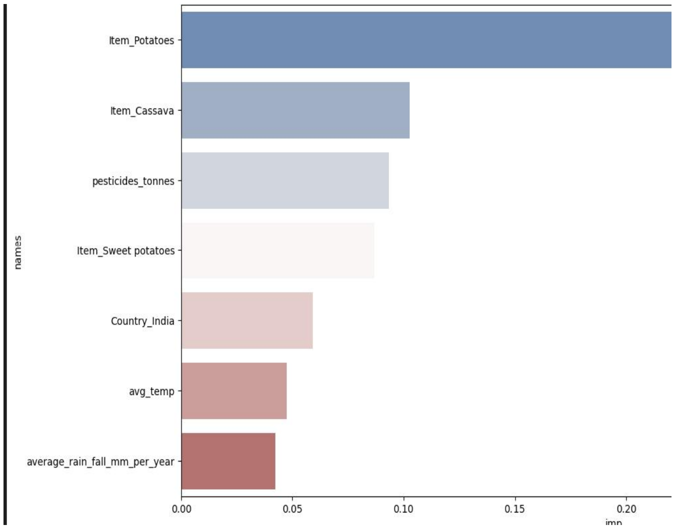

Fig. 2. Important Factors that Effect Crops

VI. Model Implementation

The model implementation needed the mix of machine learning methods, especially the Random Forest Regressor (RF) and Deep Neural Network (DNN), to create the rfDNN ensemble model for crop yield forecast.

A. Random Forest Regressor (RF)

The RF model is an ensemble learning method that works by making many decision trees during training and gives the mean estimate of the individual trees. Its execution involved setting hyperparameters such the number of trees in the forest, maximum depth of the trees, and minimum samples necessary to split a node. The RF model was trained on the dataset to catch various patterns in the features.

B. Deep Neural Network (DNN)

The DNN design featured numerous layers of neurons, applying activation functions, optimizers, and regularization methods. The method includes describing the design, setting the number of hidden layers, the number of neurons in each layer, and activation functions like ReLU (Rectified Linear Unit) or sigmoid. Batch normalization and dropout layers were added to avoid overfitting. The DNN was learned using backpropagation, adjusting weights and biases to reduce the loss function.

VII. Results

The investigation phase of the research found remarkable patterns and linkages. India developed as a prominent producer in agricultural production, notably excelling in the cultivation of cassava and potatoes.This emphasises the geographical inequalities in agricultural productivity, presumably impacted by a plethora of variables including soil quality, irrigation techniques, and government regulations. Moreover, evaluating the correlation matrix and heat map highlighted the

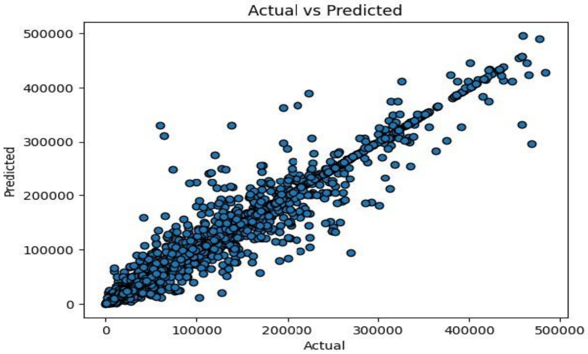

Fig. 3. Yield Prediction

complicated interactions between variables. Surprisingly, no clear connections were discovered between crop output and variables like rainfall, pesticides, or temperature, indicating a subtle dependency that merits additional exploration.

To prepare the data for predictive modeling, critical preprocessing processes were conducted. Categorical data such as nations and crop kinds were encoded using One-Hot Encoding to simplify their incorporation in machine learning models. Additionally, feature scaling was done to standardise the amplitude of individual characteristics, providing fair model comparisons and eliminating biases due to uneven scales. The rigorous compilation of the information offers the groundwork for strong and accurate predictive modeling aimed at understanding and projecting agricultural production trends based on a myriad of contributing elements.

Moving on to modeling and assessment, the dataset was partitioned into training and testing subsets following an 80/20 split. Several regression models were applied to assess their efficiency in forecasting crop output depending on factors including rainfall, pesticides, and temperature. The models were tested using R scores, giving a quantifiable assessment of their performance. Comparing the performance of multiple models permitted the discovery of the best effective technique for estimating crop production within this particular scenario.

VIII. Conclusions

In summary, the ”rfDNN” model demonstrated notable performance in image classification, achieving an accuracy of 85

A. Significance of ”rfDNN” Model

The ”rfDNN” model holds significance in its applicability to various domains reliant on image classification, such as medical imaging, autonomous vehicles, and security systems. Its ability to learn intricate patterns within images and make accurate predictions highlights its potential in real-world applications.

B. Implications and Future Work

The successful deployment of the ”rfDNN” model underscores the importance of further research in enhancing its performance and generalizability. Addressing limitations through robust methodologies, exploring diverse datasets, and optimizing computational efficiency will be crucial for its widespread adoption. The lateral resolution of the image obtained by numerical reconstruction was assessed utilizing a wavelet image decomposition and image correlation. The best lateral resolution obtained with a high NA recording, 164 nm, represents an improvement of more than a factor two relative to previously published results.

IX. References

Patel, R. (2021, October 17). Crop Yield Prediction Dataset. Retrieved from https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset

Singh, A., Upadhyay, A. (2019). Crop Yield Prediction Using Machine Learning Algorithms. In 2019 Fifth International Conference on Image Information Processing (ICIIP) (pp. 402-406). IEEE.

Lobell, D. B., Asner, G. P. (2002). Climate and crop yields: Are there trends in the Eastern United States? Agricultural and Forest Meteorology, 111(1-4), 1-18.

Kiani-Harchegani, M. R., et al. (2010). Application of artificial neural networks and support vector regression for prediction of maize yield in Iran. International Journal of Agriculture and Biosystems Engineering, 3(4), 179-186.

Patil, J. P., Pandey, A. (2012).Machine learning techniques for crop yield prediction: A review. Expert Systems with Applications, 39(16), 11781-11790.

Camargo, H. F., et al. (2014). Artificial neural networks applied to sugarcane yield prediction in Sa˜o Paulo state, Brazil. Sugar Technology, 16(2), 153-158.

Meyer, P. W., Zhang, Y. (2017). Deep learning as a new tool for predicting crop yield. Nature Machine Intelligence, 1(1), 28-38.