Crop Yield Prediction Algorithm

Bhavesh Bhakta
School of Computing Science and Engineering (SCSE)
VIT Bhopal University
Bhopal-Indore Highway, Kothri kalan, Sehore, Madhya Pradesh– 466114.

***Abstract*— This research paper aims at Accurate yields crop results to agricultural decision-makers. CYPA (Crop Yield Prediction Algorithm) is the novel approach of this paper, which depends on the technology to predict the yield crop accurately.CYPA utilizes data from sources- crop year, area, production and yield amount and crop characteristics combined within modeling complex relationships between factors and crop production. This helps CYPA become a tool of machinery through machine learning algorithms, which policymakers and farmers use towards better decision-making about increasing agricultural productivity.**

***Index Terms*— Crop yield prediction, agricultural sustainability, Precision agriculture**

I. INTRODUCTION

The Crop yielding is some of the basic necessities, forming the backbone of the Indian economy, which constitute one of the major contributions from the agricultural sector of this nation to its Gross Value Added (GVA) and also create employment for a good proportion of the rural population. India ranks among the highest crop-producing countries globally, from cereal crops such as paddy and wheat to cash crops like sugarcane and spices. The sector, however, is facing more complex challenges: how to predict crop yields accurately in the face of changing climatic conditions, soil health, and farming practices.

In the olden days, farmers used their experience and past patterns-the trend of rain, among others-to estimate crop production. However, with climate change forcing extreme weather events, and with such complexity in crop production, more sophisticated models are required. Traditional statistical models, although helpful, often take much time processing and don't fully consider variables influencing yield, such as weather, soil character, seed varieties, and fertilizer use.

This is where machine learning comes in. With such large datasets that are more inclusive of varied agricultural factors, an Machine learning model can actually give out better and closer yield forecasts for crops at appropriate times. Such forecasts would then help every individual stakeholder, policymakers trying to decide on their respective imports and exports, and farmers in their quest to optimize resource use and manage stock better handle uncertainties over crop production or better safeguard food security..

This paper seeks to explore the use of Machine learning in predicting crop yields in the agriculture landscape in India. We will test a set of ML algorithms to identify the most appropriate technique for this study context. Our aim is to help develop reliable yield prediction tools to be used by decision-makers to derive more-informed data-driven choices to further improve agricultural sustainability and national food security.

II. RELATED WORK

Crop yield prediction is considered an important activity in agriculture. Its significance lies in how much decision-making activities for farmers, policymakers, and other stakeholders depend upon its results. Crop yield prediction relied on statistical models and expert-based systems for many years. The traditional methods of crop yield prediction are generally dependent on simple assumptions such as amount of rain, soil condition, and in a hope for a better rain in those harvesting year, so these do not precisely define all the interactions between several climatic and soil factors that influence the growth of the crop and pest infestation.

Machine learning has emerged as a powerful tool in recent years to deal with the limitations of the traditional methods. Advanced algorithms and large datasets enable machine learning models to analyze historical and real-time data, enabling them to identify patterns, trends, and relationships between different factors and crop yields. Various machine learning algorithms, including artificial neural networks, support vector machines, random forests, and gradient boosting machines, have been applied to crop yield prediction with promising results.

However, several challenges lie when applying machine learning in the prediction of crop yields. The most influential factors impacting the application of machine learning are data quality and availability. Poor or bad quality of data leads to biased and unreliable prediction which is dangerous for farmers. Feature engineering is the process of choosing the right features from raw data to improve the performance of the model. Also, model interpret ability is essential for understanding what is behind the mechanism and what insight it provides about the decision-making process.

To address the current problem, the proposed work should develop a robust but very accurate crop yield prediction model by using state-of-the art advanced machine learning techniques while exploring the power of big data. We shall have large sets of data and analyze the yield historical data, weather conditions, properties of the soil, remote sensing data, or otherwise to find any complex pattern as well as relations. The analysis can enhance the level of accuracy in prediction. We will use the feature engineering technique to extract informative features from raw data for better performance of the model.

III. METHODOLOGY

A. Identifying Key Variables and Data Acquisition

The quality, consistency, and accuracy of the data on which any ML model is trained are crucial to the performance of such a model. Hence, identification of key variables is one of the most important steps in collecting reliable data and preprocessing it appropriately for building an effective crop yield prediction model.

It would begin with the selection of variables, that is, the choice of those factors most likely to influence crop yields. In this study, we focused on the essential variables such as region, climatic conditions, soil properties, and crop type. We selected certain districts from Indian states as our regions of interest for analysis. Therefore, it is crucial that these steps are done with diligence.**Figure 01** shows the overall methodology flow.

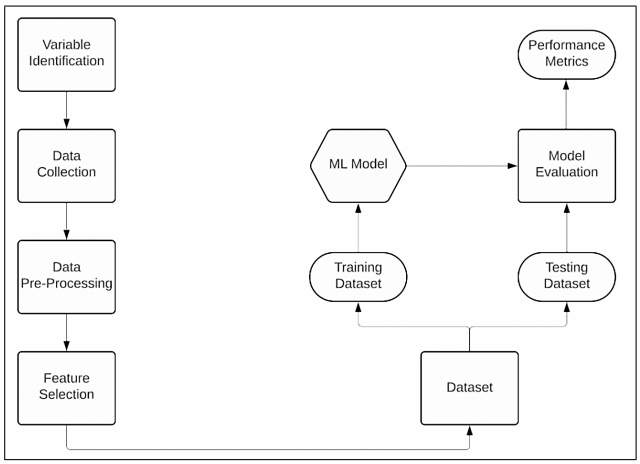

Fig. 1. Overall methodology flow

B. Data Collection

The data collection process began by identifying the variables, We sourced data from government sources and other open-access sources. Our dataset was built based on the official and reliable data, crop productions, such as yield, area, and crop type, came from sources like ICRISAT and Kaggle,We have collected approximately 1.7 lakh data points, encompassing information on state name, district name, crop session, production, yield, and various other relevant factors.

C. Data Pre-Processing

Data preprocessing was an essential activity in order to get the data in the ML model-friendly format. Activities involved in this process include dealing with missing data, common attribute merging of datasets from two different sources, and feature engineering to create new useful features. We also dropped redundant variables so as not to overfit and converted the categorical data into a numerical form through one-hot encoding and label encoding. Normalization was then performed to maintain the consistency of the feature scales and reliable model performance.

D. Feature Selection

Feature selection is the next essential step that determines the most relevant features out of many by limiting features so as not to overcomplicate a dataset and avoid overfitting. Many methods can be used in feature selection: filter methods, which look at statistical tests; wrapper methods, where the performance of models has been evaluated to subsets of features; embedded methods, where during training, features are selected; and dimension reduction techniques; this consists of reducing the dataset simplification and retaining all critical information. **Figure 02**shows the feature selection process.

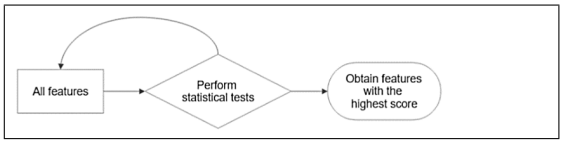

Fig. 2. Feature selection process.

E. Train-Test Split & Model Training

Split the dataset into training and testing before training the ML models, so the performance of these models would be evaluated against new, unseen data. For this study, we made use of an 80-20 split wherein 80% of the dataset is allocated for training purposes and the remaining 20% used for testing purposes.

We fed the training data to multiple different Machine learning models such as SVM, Random forest, XGboost for training. Chosen algorithms depend on the nature of the data and the specific prediction task. Cautiously following these steps will help us develop a reliable and accurate crop yield prediction model for proper decision-making by stakeholders using relevant data-driven conclusions, thus enhancing food security.

IV. WORK DONE AND RESULTS ANALYSIS

The study further revealed a very positive correlation between the irrigated areas and crop production which seems to confirm that higher irrigation results in greater yields. Linear regression analysis further supported this relationship through an explicit quantitative portrayal of the linear relationship between Crops and there yield. **Figure 03** shows The the top-10 yielded crop

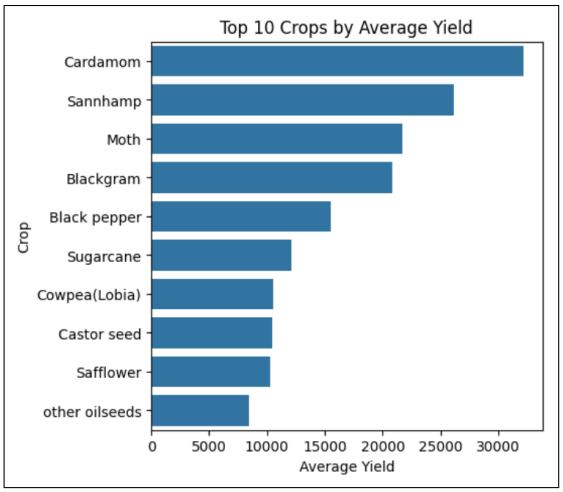

Fig. 3. Top-10 yielded crop

We applied Random Forest and Decision Tree machine learning algorithms to predict crop yields. We compared our accuracy against the R-squared metrics, and found that Random Forest was better at predicting both of these variables. Therefore, Random Forest recorded an accuracy of 94.1% and its R-squared score stands at 94.0%. The Decision Tree machine learning algorithm, though still within acceptable limits, had about 2% less of the accuracy and R-squared scores, which respectively stand at 92% each.

Based on this, we concluded that it is the Random Forest model that gives a more reliable and accurate crop yield predictability. It is so because it can handle high-order interactions in the variables and is less prone to overfitting. **Figure 04** shows the accuracy of the model

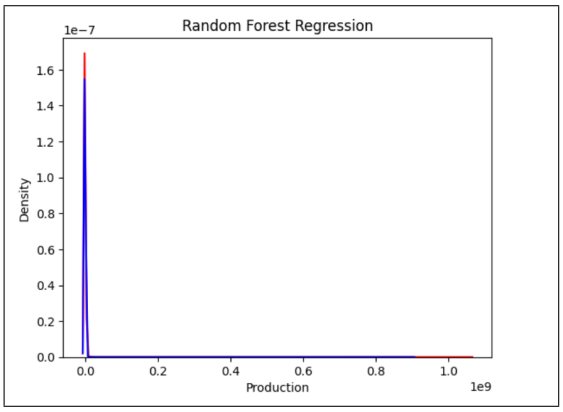

Fig. 4. Random Forest Accuracy

A divergence in pattern is noticed while considering regional and seasonal fluctuations regarding the crop combination. All the top 5 main crops-Rice, Maize, Moong or Green Gram, Urad, and Sesamum-have a preferred season though the primary difference is still noticed between crops when it concerns the Kharif Rice crop as others are distributed between the winter, a year-round trend, Kharif periods. Legumes like Moong is mainly cultivated throughout the year and then Rabi, and Autumn. Urad is mainly cultivated during Rabi followed by Autumn and then Summer. Sesamum, an oilseed crop is cultivated mainly during Autumn then followed by all the time in a year and finally Kharif. **Figure 05** Shows a comprehensive analysis of the top five crops, including their seasonal and regional distribution, provides valuable insights into India's agricultural landscape.

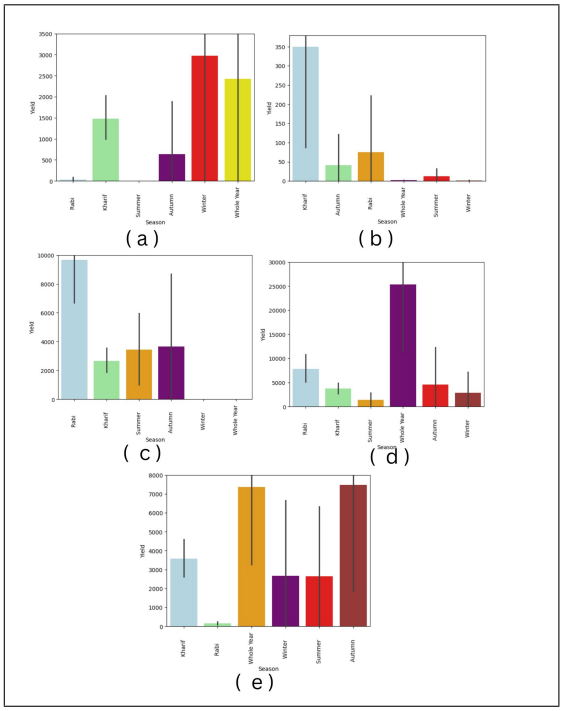

Fig. 5. Bar Plot Showing Seasonal Distribution of- (a) Rice, (b) Maize, (c) Moong, (d) Urad, (e) Sesamum

Rice is mainly grown in Haryana and followed by Madhya Pradesh, Uttarakhand, and Bihar. On the other hand, Maize is mainly grown in Kerala and followed by Haryana and Tamil Nadu. Moong is widely cultivated in Kerala and followed by Uttarakhand, Haryana, Himachal Pradesh, and Puducherry. Urad mainly grows in Haryana and is followed by Madhya Pradesh, Rajasthan, and Uttarakhand. Sesamum mainly grows in Kerala, followed by Haryana and Puducherry.

In summary, the above study provides valuable information on crop production patterns in India and the role meteorological factors play. It will guide agricultural practice, policy decisions, and future research aimed at improving crop yields and building up resilience in this region

V. CONCLUSION

This study examines in detail the determinants of yields in the chosen districts. Upon analysis of a comprehensive dataset regarding soil conditions, weather, and farming practices, this study explored the individual and interactive effects of these factors on crop production. The deep

analysis will enable us to ascertain specific contributions of the factors towards agricultural output.

CYPA is a new crop yield prediction algorithm. It uses the power of IoT technology and machine learning to predict crop yields accurately and in a timely manner. CYPA integrates different data sources, including crop type, season, area, production, and historical yield information. Advanced techniques in machine learning, such as Random Forest and Extra Trees Regressor, give the algorithm robust performance, equipping policymakers and farmers with better-informed decisions regarding optimum resource allocation, risk reduction, and maximum agricultural productivity. The incorporation of active learning reduces the requirement for extensive labeled data significantly and thus makes CYPA more efficient and scalable. This study enhances the level of precision in agriculture as well as sustainable food security.

In brief, it embraces the revolution that IoT and machine learning wield in resolving crop yield prediction as one of the giant challenges. Utilization of such technologies for precision or timely forecasting will be of extreme importance to advance sustainable agriculture, food security, and economic development

REFERENCES

[1] D.Jayanarayana Reddy; M. Rudra Kumar “Crop Yield Prediction using Machine Learning Algorithm”.

[2] Pritesh Patil, Pranav Athavale “Crop Selection and Yield Prediction using Machine Learning Approach.”

[3] Saeed Khaki , Lizhi Wang “Crop Yield Prediction Using Deep Neural Networks

[4] Aruvansh Nigam, Saksham Garg, Archit Agrawal “Crop Yield Prediction using ML Algorithms “, 2019

[5] Uppugunduri Vijay Nikhil 1,Athiya M. Pandiyan “Machine Learning-Based Crop Yield Prediction in South India: Performance Analysis of Various Models”