Grenze International Journal of Engineering and Technology, June Issue

Enhancing Crop Yield Projections through Machine Learning Models

N. Bhanu Teja[1], S.V.V.D. Jagadeesh[2], K. Geetha Ramya[3] and B. Sai Durga Rao[4]
[1-4]Lakireddy Bali Reddy College of Engineering, Mylavaram, India
Email: bhanutej.official@gmail.com, jagadeesh.s@lbrce.ac.in, kothagundlageetharamya@gmail.com, saidurgarao409@gmail.com

***Abstract*—Our research aims to revolutionize crop yield prediction by combining machine learning (ML) techniques with government datasets and detailed soil information. Traditional prediction methods frequently overlook soil health, leading to inaccuracies. To address this, we incorporate soil data into our predictive model. Using the Random Forest Classifier and extensive datasets that include crop-specific and soil characteristics, our approach achieves high accuracy in forecasting crop yields. Our results highlight the power of data-driven methods in guiding agricultural decision-making and stress the importance of interdisciplinary collaboration to address modern agricultural challenges.**

***Index Terms*— Crop yield prediction, machine learning, random forest, government data.**

I. INTRODUCTION

Indian agriculture, a cornerstone of the nation's economy and livelihood for millions, faces multifaceted challenges amidst a backdrop of diverse geographical and climatic conditions. With a significant portion of the population dependent on agriculture, the sector's productivity and sustainability are paramount for food security and economic growth. However, traditional farming practices often encounter hurdles such as unpredictable weather patterns, soil degradation, and limited access to modern technology.In recent years, the incorporation of technology, especially machine learning, has become a promising approach to overcoming these challenges and enhancing agricultural practices. Leveraging vast repositories of government data, including historical crop yields, weather patterns, soil characteristics, and land use, machine learning algorithms offer the potential to analyze complex interactions and predict crop yields with unprecedented accuracy.This research endeavors to harness the power of machine learning to develop a robust crop yield prediction system tailored to the unique context of Indian agriculture. By synthesizing diverse datasets from government sources, our aim is to create a decision support system that empowers farmers with actionable insights for crop selection and management. The ultimate goal is to enhance agricultural productivity, mitigate risks associated with environmental variability, and promote sustainable farming practices across different regions of India.Through this study, we seek to contribute to the ongoing efforts to modernize and optimize Indian agriculture, ensuring food security, economic prosperity, and environmental sustainability for current and future generations.

II. LITERARY SURVEY

In their research, Aruvansh Nigam, Saksham Garg, and Archit Agrawal have shown that the Random Forest

Grenze ID: 01.GIJET.10.2.645

machine learning algorithm provides the highest accuracy for predicting crop yields using datasets from the Indian government. They also highlighted the effectiveness of Sequential models for rainfall prediction and LSTM models for temperature prediction. [2] Leo Brieman has emphasized the importance of minimizing correlation while maintaining strength in Random Forest algorithm applications, particularly through the utilization of bagging methods. [3] Balamurugan's research focuses solely on Random Forest for crop yield prediction, considering factors like rainfall and temperature, albeit without comparative analysis with other machine learning algorithms. [4] Mishra discusses various machine learning techniques theoretically without implementing any algorithms, thereby lacking practical insights. [5] Dr. Y. Jeevan Nagendra Kumar's work underscores the effectiveness of supervised learning techniques, particularly emphasizing the application of Random Forest in crop yield prediction, highlighting its ability to achieve accurate predictions with minimal model complexity. These studies collectively showcase the potential of machine learning algorithms, especially Random Forest, in enhancing crop yield prediction accuracy, thereby informing agricultural practices and ensuring food security.In their respective studies, [1] Aruvansh Nigam, Saksham Garg, and Archit Agrawal have demonstrated that the Random Forest machine learning algorithm yields the highest accuracy in predicting crop yields using Indian government datasets, while also noting the efficacy of Sequential and LSTM models for rainfall and temperature prediction, respectively. [2] Leo Brieman has emphasized the importance of minimizing correlation while maintaining strength in Random Forest algorithm applications, particularly through the utilization of bagging methods. [3] Balamurugan's research focuses solely on Random Forest for crop yield prediction, considering factors like rainfall and temperature, albeit without comparative analysis with other machine learning algorithms. [4] Mishra discusses various machine learning techniques theoretically without implementing any algorithms, thereby lacking practical insights. [5] Dr. Y. Jeevan Nagendra Kumar's work underscores the effectiveness of supervised learning techniques, particularly emphasizing the application of Random Forest in crop yield prediction, highlighting its ability to achieve accurate predictions with minimal model complexity. These studies collectively showcase the potential of machine learning algorithms, especially Random Forest, in enhancing crop yield prediction accuracy, thereby informing agricultural practices and ensuring food security.

III. METHODOLOGY

A. Data Collection and Compilation

The initial step in our research project on crop yield prediction involved the comprehensive collection of data from various government databases and repositories. Given the diverse sources and formats of these datasets, each individual dataset did not fully meet the requirements of our study due to variations in features and parameters. To address this limitation, we adopted a systematic approach to merge multiple datasets, culminating in the creation of an integrated dataset that encompasses a wide range of relevant features essential for our analysis. The compiled dataset includes essential information such as soil characteristics, water availability, crop types, and historical yield data. By amalgamating these disparate datasets, we aimed to create a comprehensive and representative dataset that captures the multifaceted factors influencing crop yields. This process involved careful validation and harmonization of data to ensure consistency and accuracy across all variables. The resulting ultimate dataset serves as the foundation for our crop yield prediction model, providing a rich and diverse set of features essential for training and evaluation. By incorporating a wide array of factors such as soil quality, water availability, crop varieties, and historical yield trends, our dataset offers a holistic perspective on the complex interactions influencing agricultural productivity. This approach not only enhances the robustness and reliability of our predictive model but also facilitates informed decision-making for farmers and stakeholders in the agricultural sector. The compilation of diverse datasets into a unified framework underscores our commitment to leveraging comprehensive data-driven approaches to address real-world challenges in agriculture and food security. In summary, the methodology of data collection and compilation underscores our rigorous and systematic approach to assembling a comprehensive dataset tailored to the specific requirements of our crop yield prediction research. This integrated dataset serves as a valuable resource for advancing our understanding of the factors influencing crop productivity and informing evidence-based agricultural management strategies.

B. Data Preprocessing

Data preprocessing is a pivotal step in converting raw data into a clean, usable dataset suitable for analysis. It involves handling inconsistencies, missing values, and errors typically found in data collected from diverse sources. Techniques such as imputation and deletion are applied to address missing values, ensuring the dataset's integrity. Additionally, the dataset is partitioned into training and testing subsets, with a larger portion allocated for training (usually 80%) and the rest for testing. This split ensures effective model training while enabling evaluation of model performance. Factors Influencing Crop Yield and Production: Various elements impact crop yield and production,including temperature, rainfall, area under cultivation, and soil nutrients such as nitrogen, potassium, and phosphorus. Temperature affects crop growth and development, while rainfall and soil moisture levels play crucial roles in sustaining plant health. The geographical area under cultivation directly impacts production volumes. Essential soil nutrients like nitrogen, potassium, and phosphorus are vital for plant growth and development, influencing processes such as protein synthesis and root development. Incorporating these key factors into predictive models enables accurate forecasting of crop yields and production trends.

C. Selecting Model

Prior to selecting an algorithm, it is imperative to conduct a thorough evaluation and comparison to determine the most suitable approach for the specific dataset. Machine learning emerges as the optimal technique for addressing crop yield prediction challenges due to its practical applicability and effectiveness. In this study, various machine learning algorithms are considered and evaluated for their predictive performance on the dataset. The following algorithms are included for selection and accuracy comparison:

- Logistic Regression: This supervised learning classification algorithm is designed to predict the probability of a target variable and is particularly effective for datasets with a binary target variable. When applied to our dataset, logistic regression achieves an accuracy of 64.8%.
- Naive Bayes: The Naive Bayes classifier operates on the assumption that the presence of a specific feature in a class is independent of the presence of other features. It is particularly useful for large datasets and exhibits simplicity and efficiency. Despite its simplicity, Naive Bayes often outperforms more complex classification methods. In our study, Naive Bayes achieves an accuracy of 82.50%.
- Random Forest: The Random Forest algorithm is highly effective in analyzing crop growth dynamics in relation to climatic conditions and biophysical changes. It works by constructing multiple decision trees from different data samples and then combining their predictions through voting, which ensures robust outcomes. By employing the bagging method for training data, Random Forest improves result accuracy. In our dataset, Random Forest achieves the highest accuracy of 94.21%.

The analysis shows that Random Forest exhibits higher accuracy when compared to Logistic Regression and Naive Bayes. This emphasizes the effectiveness of Random Forest in precisely predicting crop yields, utilizing the various features available in the dataset.

D. Random Forest Model for Crop Prediction

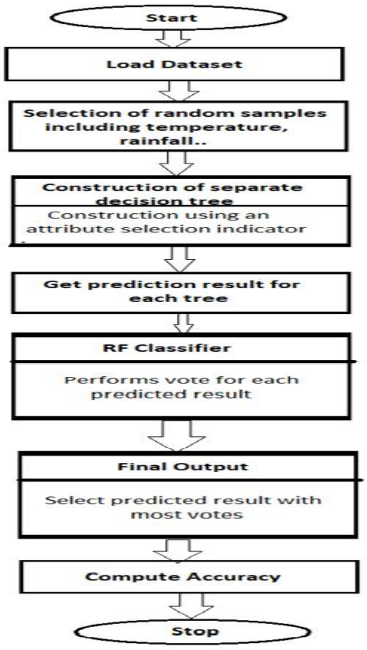

Fig. 1. Flow chart of Randomforest model

Random Forests operate through the aggregation of tree predictors, where each tree's decisions are based on values sampled from a random subset, distributed independently and uniformly across all trees in the forest. This ensemble method utilizes the bagging technique to train the data, enhancing result accuracy. In our study, We utilized the Random Forest algorithm to attain a high level of predictive accuracy, assessing the model's predictions against the actual outcomes within the dataset. The analysis indicates a predicted accuracy of 91.34% for the model. Figure 2 illustrates the flowchart depicting the Random Forest model's implementation for crop yield prediction.

E. Proposed Application

We have developed an application that incorporates all essential data required for crop prediction and utilizes machine learning algorithms for accurate predictions. The application is designed to identify the most suitable crop for farmers based on the prevailing conditions. By leveraging predictive algorithms, the application analyzes various factors such as soil quality, weather patterns, and historical crop performance to recommend the optimal crop for cultivation. This approach enables farmers to make informed decisions regarding crop selection, maximizing productivity and yield potential.

F. Analysis of The System

- Python 3.8.5 (Jupyter Notebook): Python is the primary coding language employed for conducting machine learning analyses. Jupyter Notebooks play a crucial role in visualizing the analysis process and generating essential results.
- Weather_API (Open Weather Map): The Weather API operates as an application programming interface, offering access to real-time weather information for specified locations. The API key sourced from Open Weather Map supplies the necessary current weather forecasts essential for crop prediction.
- Android Studio (Version 3.4.1): Android Studio serves as the designated integrated development environment (IDE) for Android application development. Java is employed as the frontend design framework within this project. The IDE and the application are linked via the USB debugging method.
- Python Flask Framework (Version 2.0.1): Flask is a lightweight web framework in Python, utilizing the WSGI (Web Server Gateway Interface) toolkit and Jinja2 template engine. In this context, Flask functions as the backend framework for building the application. It provides a range of modules and libraries that simplify application development by abstracting away low-level tasks like protocol handling and thread management.
- Heroku: Heroku operates as a container-based cloud platform, enabling developers to develop, deploy, and manage applications entirely in the cloud. In this context, Heroku functions as the server component. By creating an account on Heroku and linking it with a GitHub repository, deploying the application becomes feasible.

IV. RESULT AND DISCUSSIONS

This study emphasizes the improvement of crop production through the utilization of machine learning methodologies. By employing techniques that deliver high accuracy, precise predictions of crops and their respective yields can be achieved. The implementation of machine learning algorithms is conducted within Python 3.8.5, utilizing Jupyter Notebook and integrating essential libraries such as Scikit-Learn, Numpy, Keras, and Pandas. Furthermore, an Android application was created to access the results of the machine learning analysis. Developed using Flutter, this Android app presents the predicted crop names alongside their associated yields.

A. Dataset Used

The datasets were sourced from various official Government websites, including:

- data.gov.in: Providing details pertaining to area, production, and crop names [8].
- indianwaterportal.org: Offering information on rainfall [9].
- power.larc.nasa.in: Furnishing data on temperature, humidity, and wind speed [10].

The combined dataset comprises 99,849 instances, incorporating features such as crop names, area, production, temperature, rainfall, humidity, and wind speed across fourteen districts in Kerala. Through the data preprocessing phase, we ensured the accuracy and integrity of the dataset. The following figures shows the heatmap for our dataset.

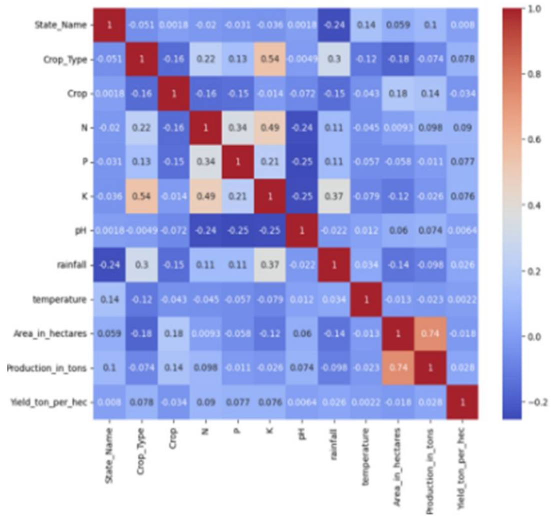

Fig.2. Heatmap

B. Classifier Used

To assess accuracy and predict outcomes, we utilized three machine learning classifiers: Logistic Regression, Random Forest, and Naïve Bayes. These classifiers underwent training on the dataset, and we generated a comparison graph to depict their performance. Figure 5 presents the performance metrics of each model. Among the three classifiers employed, Random Forest demonstrated the highest accuracy level. The figure illustrates the performance comparison of the algorithms.

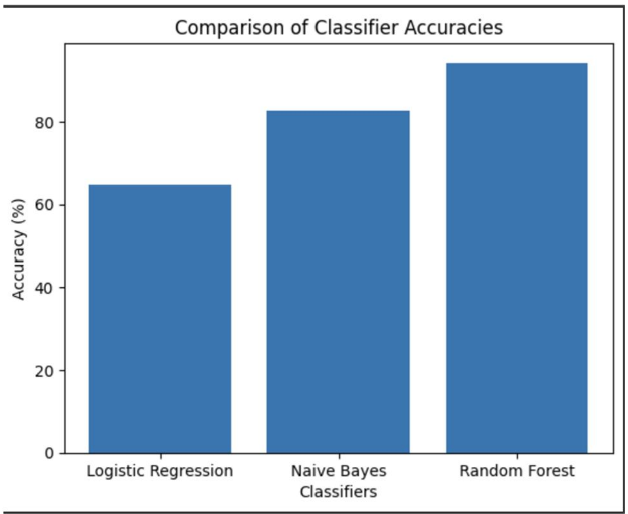

Fig.3. Performances of models

C. Crop Yield Calculation

The Random Forest Classifier forecasts the crop type, which is then linked to its anticipated production. Subsequently, dividing the total production by the user-input area yields the crop's productivity.

Yield= Production/Area

This association between predicted crop names and their corresponding yields aids farmers in determining the optimal timing for cultivating the most suitable crop, thus maximizing agricultural output.

D. Android Application

An Android application has been developed to facilitate access to the outcomes of machine learning analyses. Designed to operate on Android OS version 7 and programmed using Java, the application boasts a user-friendly interface requiring minimal user interaction for desired results. When users input the location and area of their field, the application identifies the most suitable crop for cultivation in that specific area. Upon receiving user-entered details, the application triggers a query to the machine learning analysis. Leveraging location information, the application accesses weather data via an API. The weather data is utilized by the machine learning classifier to predict the optimal crop and estimate its potential yield. The resulting output is fetched from the server and displayed within the application interface. Key functionalities within the application encompass account creation, detail entry, and results retrieval. Account creation facilitates user interaction with the application, enabling seamless navigation through its features. On the home page, users input relevant details to proceed to the results activity. The retrieved data is then forwarded to the machine learning model, which predicts the crop name and computes the corresponding yield value.

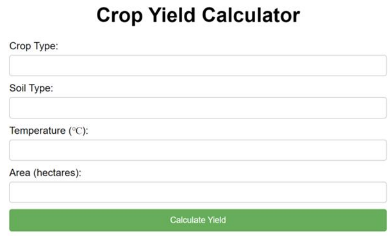

Fig.4. Mobile application interface

V. Conclusion

This study centers on utilizing machine learning techniques for crop prediction and yield estimation. To ensure accuracy, various methodologies are explored, with a specific focus on employing the Random Forest classifier for crop prediction within a selected district.
A systematic approach is developed to predict crops based on historical data analysis, aiding farmers in decision-making regarding crop cultivation. The primary objective is to provide farmers with valuable insights into efficient crop selection, ultimately optimizing harvesting practices. Accurate predictions across different states in India are essential for improving agricultural productivity, and this research aims to contribute to this endeavor by leveraging advanced machine learning techniques to facilitate efficient and productive harvesting practices.

Future Scope

In the future scope of our project, we aim to enhance the accuracy of crop prediction by integrating additional data sources, specifically soil and crop fertilizer data, alongside existing crop data. We envision leveraging advanced technologies such as deep learning algorithms to further refine our predictive models. By incorporating these supplementary datasets and employing sophisticated machine learning techniques, we anticipate achieving higher levels of accuracy in crop prediction. This expansion of our project scope holds significant promise for improving agricultural decision-making and optimizing crop cultivation practices.

References

[1] Crop Yield Prediction using Machine Learning Algorithms al Journal of Engineering Research & Technology (IJERT)ISSN: 2278-0181Published by, www.ijert.org,NCREIS - 2021 Conference Proceedings
[2] Aruvansh Nigam, Saksham Garg, Archit Agrawal “Crop Yield Prediction using ML Algorithms “,2019
[3] Leo Brieman, “Random Forests”, 2001
[4] Priya, P., Muthaiah, U., Balamurugan, M.”Predicting Yield of the Crop Using Machine Learning Algorithm”,2015
[5] Mishra, S., Mishra, D., Santra, G. H.,“Applications of machine learning techniques in agricultural crop production”,2016
[6] Dr.Y Jeevan Kumar,”Supervised Learning Approach for Crop Production”,2020
[7] Ramesh Medar,Vijay S, Shweta, “Crop Yield Prediction using Machine Learning Techniques”, 2019
[8] Ranjini B Guruprasad, Kumar Saurav, Sukanya Randhawa,”Machine Learning Methodologies for Paddy Yield Estimation in India: A CASE STUDY”, 2019
[9] Sangeeta, Shruthi G, “Design And Implementation Of Crop Yield Prediction Model In Agriculture”,2020
[10] https://www.data.gov.in
[11] https://power.larc.nasa.gov/data-access-viewer/