Crop Yield Prediction

1 N. Sri Chakradhar, 2 Gopa Pulastya, 3 Rishi Anirudh K  
*Department of Computer Science and Engineering,*  
*Amrita School of Computing,*  
*Amrita Vishwa Vidyapeetham,*   
Bengaluru,Karnataka.  
1 [bl.en.u4aie22132@bl.students.amrita.edu](mailto:bl.en.u4aie22132@bl.students.amrita.edu) , 2 [bl.en.u4aie22116@bl.students.amrita.edu](mailto:bl.en.u4aie22116@bl.students.amrita.edu) , 3 [bl.en.u4aie22139@bl.students.amrita.edu](mailto:bl.en.u4aie22139@bl.students.amrita.edu)

*Abstract*— In contemporary agriculture, accurately forecasting crop yields holds significant importance for optimizing resource allocation, improving decision-making processes, and safeguarding food security. algorithms, data preprocessing techniques, and statistical analysis on historical and real-time agricultural data. The datasets contain the parameters such as weather conditions, pesticides information, rainfall data, temperature, yield value and historical yield data. Data preprocessing encompasses tasks such as merging, cleaning, normalization, and feature engineering. Various machine learning techniques such as clustering, classification, regressions are implemented to analyze the complex relationships between input and predict the crop yields. This aims to provide farmers, agricultural policymakers, and stakeholder predict the crop yield effectively. By facilitating informed decision-making, the Crop Yield Prediction System contributes to increased agricultural productivity and food security.

*Keywords - Crop Yield Prediction, Machine Learning, Data Integration, Feature Engineering, Exploratory Data Analysis, Model Selection, Model Evaluation, Clustering, Regression Models, Machine Learning Techniques, Agricultural Decision-Making, Predictive Model.*

# INTRODUCTION

In the complex relationship between farming and technology, the predictive power of data driven models emerges as a pivotal role. This project explores the realm of crop yield prediction, an essential aspect of modern farming practices. Leveraging a diverse dataset encompassing climate variables, pesticide usage, and rainfall patterns, our endeavor aims to unravel the complex dynamics influencing crop productivity. Navigating through the details of data preprocessing, feature engineering and machine learning algorithms. The integration of different machine learning techniques such as clustering, classification, regression and this analysis will help in the easy analysis of the crop yield prediction based on the graphs that allows us to discern patterns and relationships within the data, paving the way for accurate yield predictions.

To achieve precise predictions and address fluctuations in temperature, rainfall and the pesticide usage, diverse machine learning algorithms can be employed to identify patterns. This approach aims to support agricultural development in India, ultimately contributing to an improved quality of life for farmers. Historically, numerous researchers have utilized machine learning techniques to bolster the agricultural sector's growth in the country. Using diverse machine learning algorithms on agricultural data in India aims to improve predictions, optimize resource allocation, and support sustainable farming, driving economic growth and resilience in the sector.

We have used various datasets taken from Kaggle which will affect the crop yield such as the yield data, temp data, rainfall data, pesticides data in which yield data talks about the yield of the particular area in the respective years. The temp data talks about the temperature of the particular area in the respective years. The rainfall data talks about the average rainfall in a particular area in the respective years and the pesticides data tells us about the usage of the pesticides in the particular area in the respective years. Combining these datasets in agricultural analysis aids in creating new features, exploring correlations among variables, and gaining contextual insights into crop yield determinants. This practice enhances model performance and maintains data integrity by integrating related information seamlessly.

# LITERATURE SURVEY

The previous works in this field indicate prevalent use of Machine Learning algorithms, such as Artificial Neural Networks (ANN), incorporating features like temperature, rainfall, and soil type. In deep learning studies, Convolutional Neural Networks (CNN) emerge as the dominant algorithm for crop yield prediction. However, specific details on evaluation parameters and challenges are not explicitly outlined in the summary. This Schematic Literature Review (SLR) contributes to understanding the landscape of ML and deep learning applications in predicting crop yields [1]. It emphasizes the need to leverage statistical and machine learning methods to analyze and predict crop yields accurately, taking into account factors such as soil quality, weather conditions, and agricultural practices [2]. In support of this work, we can include Working with more useful data set which play a vital role in the dynamic Environment. The study explores the use of ensemble machine learning techniques, such as stack regression (SR) and cascade regression (CR), in combination with computer simulation data to predict wild blueberry yields [3].

In a relevant study, researchers investigate maize hybrids as part of the 2018 Syngenta Crop Challenge. They utilize a deep neural network (DNN) consisting of 21 hidden layers, each containing 50 neurons, and incorporate genetic markers, weather data, and soil variables into their analysis. The DNN outperforms other models, showcasing superior accuracy in predicting yield. The study underscores the significance of deep neural networks in capturing complex relationships, addressing challenges related to network depth, and proposing solutions. This contributes valuable insights for agricultural decision-making and breeding programs [4]. Another comprehensive review concentrates on machine learning algorithms for predicting palm oil yield. The article highlights the importance of early and reliable crop yield estimation, explores widely used features and prediction algorithms, and critically evaluates machine learning applications in palm oil yield prediction. The paper covers the benefits, obstacles, and potential future directions, suggesting a forthcoming framework for predicting palm oil yields using machine learning techniques. [5]. In a different context, the research addresses crop yield prediction for paddy crops. The study introduces a hybrid MLR-ANN model, leveraging both Artificial Neural Network (ANN) and Multiple Linear Regression (MLR) for enhanced prediction accuracy. The hybrid model outperforms conventional models, emphasizing the importance of accurate yield prediction for effective agricultural planning. The study discusses challenges in data-driven models and highlights the significance of proper feature selection and model choice [6].

For predicting wheat and barley yield based on remote sensing data from Unmanned Aerial Vehicles (UAVs), the study employs Convolutional Neural Networks (CNNs). It investigates the impact of CNN parameters on prediction accuracy, showcasing the potential of deep learning in precision agriculture for yield prediction [7]. The machine learning techniques for wild blueberry yield prediction, combining computer simulation data with various machine learning algorithms. The research showcases how synthetic data generated from computer simulations, combined with machine learning, can effectively predict agricultural yields. This presents an exciting direction for further investigation in future studies [8]. Similarly, the study proves the effectiveness of an approach in accurately predicting wheat yield, utilizing advanced machine learning techniques. The findings emphasize the potential of integrating extreme learning machine models with optimization algorithms to increases the accuracy and reliability of crop yield predictions [9].

The significance of smart agriculture in meeting global food needs is emphasized, integrating IoT devices and cloud computing to optimize agricultural processes. The study contributes to the advancement of sustainable and efficient agricultural practices, offering valuable insights for addressing food security challenges on a global scale [10]. Another study underscores the potential of deep learning methodologies and remote sensing data in improving understanding of agricultural output and enhancing crop yield prediction. The complexity of crop yield prediction due to genetic, environmental, and management factors is highlighted, emphasizing the need for innovative approaches to address this ongoing scientific challenge [11]. Machine learning approaches are emphasized for addressing the challenging problem of crop yield prediction in precision agriculture. The review highlights the potential of machine learning models in handling complex research problems and improving the accuracy of crop yield prediction [12]. Another work encompasses a wide spectrum of sources and databases, emphasizing the significance of multi-source environmental data and machine learning techniques in enhancing the accuracy of wheat yield prediction. It underscores the potential of integrating diverse environmental data sources with machine learning algorithms to improve the precision and reliability of crop yield forecasting [13]. Similarly, a wide range of sources and databases, emphasizing the significance of integrating UAV-based data and machine learning algorithms to enhance the accuracy and efficiency of crop yield prediction. The study underscores the potential of UAV technology in capturing high-resolution agricultural data, contributing to precise and timely crop yield predictions [14]. The paper talks about cutting down crop output and making less money for farmers because of things like bugs, sick plants, or bad weather. The farming system based on smart technology wants to make groundnut production better by watching and controlling things that affect how much crop is made. The study's results match with other research about using internet-connected devices in farming. These studies have tested things like predicting peanut crop yields for small farms by using Planet Scope data and better groundnut harvest automation systems using WSN technologies. These studies show how farming with IoT technology can raise crop output and help solve problems for farmers in the fields of agriculture [15].

The paper wants to support farmers in making smart choices about when and what plants they grow by using machine learning tricks. It helps them guess how much food crops will make. The results of this paper agree with other studies that looked at how machine learning methods can help farming. These include using deep learning and Sentinel 2 data to guess potato yield, as well as predicting crop output through the use of remote sensing and advanced computer algorithms. These studies show how machine learning can help increase crop growth and solve problems for farmers in the farming world [16]. The paper points out that these technologies can make crops grow better, lessen waste and support more eco-friendly farming methods. The study's results match with other ones that looked at how deep learning and IoT technologies are used in farming. These studies include comparing smart farming methods using these two things, as well as asking about ways to use them fully in this area. These studies show how deep learning and IoT solutions can increase crop production. They also help farmers to deal with the problems they face in farming industries [17]. The paper looks at how machine learning and IoT can be used in farming. It mainly talks about their use for keeping an eye on plants and smart farming systems. It explores how machine learning can be used in sensor data analysis for the farming environment. The research agrees with other studies on using computer learning and internet-connected farming systems. These technologies have the potential to change agriculture, increase crop output, and allow for precise farm methods [18].

# RESEARCH GAP

From the review of previous works, it's evident that existing machine learning models and frameworks have predominantly focused on utilizing a single dataset comprising crucial factors influencing crop yield, such as rainfall, temperature, and season. However, in our proposed research, we aim to broaden the scope by incorporating multiple datasets that encompass various factors affecting crop yield, including pesticide usage, temperature, rainfall, and the previous year's yield specific to a particular area and crop. By integrating diverse datasets, our research seeks to provide a more comprehensive understanding of the intricate dynamics influencing crop productivity. The inclusion of factors beyond traditional meteorological variables allows for a more nuanced analysis and prediction of crop yields. Leveraging machine learning techniques in conjunction with this integrated dataset enables researchers to gain deeper insights into crop yield patterns and trends. By training and deploying models on this multifaceted dataset, we expect a notable enhancement in the precision of crop yield forecasts. The extensive dataset, paired with sophisticated machine learning algorithms, boosts our capacity to understand and analyze intricate relationships among different factors and how they influence crop production. Ultimately, the integration of different datasets and machine learning techniques in our research endeavors to provide a more precise and reliable means of predicting crop yields. This has the potential to greatly benefit agricultural practices by empowering farmers and stakeholders with actionable insights to optimize productivity and resource management.

# PROPOSED METHODOLGY

Data Collection and Overview

Feature Engineering

Exploratory Data Analysis

Model Selection and Training

Model Evaluation

Predicting the Yield Values

Fig-1: The flowchart describing the proposed Methodology

## Data Collection and Overview:

The datasets are collected on the basis of the parameters that will be affecting the crop yield from Kaggle. The dataset integrates diverse sources, including yield data encompassing area, item, year, and crop yield values. Supplementary datasets feature historical temperature records, average rainfall data, and details on pesticide application, providing a comprehensive foundation for analysis and modeling in agricultural research. These datasets are crucial for understanding the complex interactions between environmental factors, farming practices, and crop productivity, facilitating informed decision-making and sustainable agricultural development strategies.

## Feature Engineering:

In our endeavor for a deeper comprehension, we engineered new features that encompassed the interaction between temperature and rainfall, logarithmic transformations of pesticide values for distribution normalization, and the amalgamation of temperature, rainfall, and pesticide factors to evaluate their combined influence on crop yield. These advancements were designed to capture intricate relationships within the dataset and enhance the predictive abilities of subsequent models.

## Exploratory Data Analysis (EDA):

The EDA involves in meticulous data cleaning which includes handling missing values and standardizing column names. The datasets were harmoniously merged based on shared attributes such as 'Area' and 'Year' to form a comprehensive dataset for subsequent analysis. The initial phase of EDA involved computing descriptive statistics, histograms, and correlation matrices. These insights provided a foundational understanding of the distribution and relationships among key variables, aiding in the identification of potential patterns. After merging the datasets, the yield values are categorized into 5 different types each with different thresholds. Then we have also implemented the label encoding technique to convert the Categorical values such as the area names into numeric types. We have also implemented the outlier’s elimination based on the statistical methods to ensure data quality. We have also included the class balancing for the yield values categorization done and balanced them by using the Machine Learning technique Imputation.

## Model Selection and Training:

Several machine learning models such as classification algorithms to predict crop yield based on standardized features like area, temperature, rainfall, and pesticide usage. Each model was trained on a standardized dataset, where categorical variables were encoded. The target column was defined as the yield category, representing different levels of crop yield. Training and testing sets were used to evaluate the models' performance in accurately predicting crop yield categories, aiming to inform agricultural decision-making and promote sustainable farming practices.

## Model Evaluation:

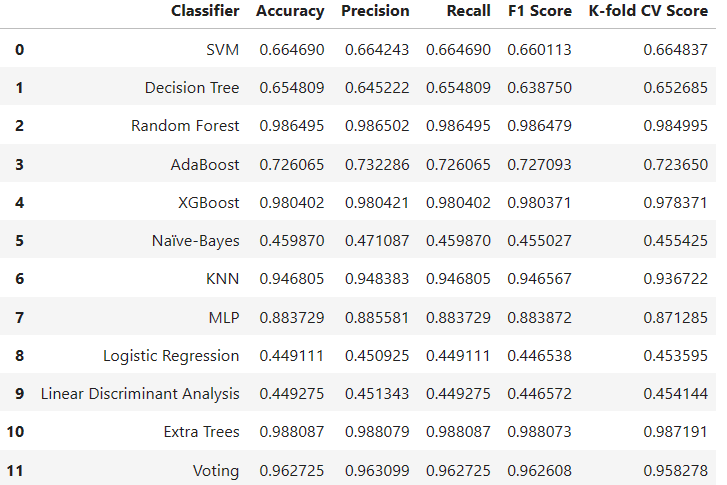
The model evaluation process rigorously assessed performance metrics such as accuracy, precision, recall, and F1 score to gauge the models' predictive accuracy in categorizing crop yield levels. Techniques like cross-validation, hyperparameter tuning, and confusion matrix analysis were employed to ensure robustness and reliability in model assessment. The evaluation aimed to identify model strengths, weaknesses, and areas for improvement, contributing to informed decision-making in agricultural practices and sustainable crop management strategies.

## Predicitng the yield values:

The selected model is deployed to predict crop yield values based on the features of temperature, rainfall, and pesticide application. These models underwent training using the standardized dataset, enabling them to learn patterns and relationships within the data. Subsequent predictions were generated to evaluate the models' efficacy in forecasting crop yields under varying agricultural conditions.

# RESULTS AND DISCUSSION

The performance of various machine learning classifiers in predicting crop yield categories was evaluated using the provided dataset. The classifiers considered included Support Vector Machine (SVM), Decision Tree, Random Forest, AdaBoost, XGBoost, Naïve-Bayes, K-Nearest Neighbors (KNN), and Multi-Layer Perceptron (MLP), Logistic Regression, Linear Discriminant Analysis, Extra Tress, Voting.



*Fig-2: The tabulated form of the classifiers with performance metrices*

From the *Fig-2*, it is observed that the models Random Forest, XG Boost and Extra Trees have achieved exceptional performance, with all the models achieving 98% accuracy, precision, recall, and F1 score. This indicates the robustness and reliability of these tree-based ensemble methods in accurately classifying crop yield categories based on environmental factors. On the other hand, AdaBoost exhibited lower performance metrics, with an accuracy of 72.6%, precision of 73.22%, recall of 72.6%, and F1 score of 72.7%. This indicates that AdaBoost may not be as effective as other classifiers in handling the complexity of the dataset and making accurate predictions for crop yield categories.

Among the boosting algorithms, XGBoost demonstrated high accuracy, precision, recall, and F1 scores, showing their effectiveness in handling complex datasets and making accurate predictions. Voting Classifier also performed very well by giving the accuracy of more than Decision tree with 96.27% accuracy. KNN and MLP classifiers also performed reasonably well, achieving accuracy scores of 94.68% and 88.3%, respectively. This indicates that neural network-based models like MLP can effectively capture nonlinear relationships and patterns in the data, leading to accurate predictions.

The Random Forest, XGBoost, Extra Trees classifiers may exhibit signs of regular fit, while the Logistic Regression, Linear Discrimant Analysis, Naïve-Bayes classifiers may suffer from underfitting. The remaining algorithms are more regular fitting behavior, striking a balance between complexity and generalization.

Overall, the results highlight the diverse performance of machine learning classifiers in predicting crop yield categories. Decision Tree, Random Forest, XGBoost, CatBoost, KNN, and MLP classifiers showed strong performance, while AdaBoost exhibited relatively lower accuracy and precision. These findings can guide decision-makers in selecting appropriate models for crop yield prediction tasks, ultimately contributing to informed agricultural management strategies and sustainable farming practices.

In decision trees, Information Gain is used to determine the best attribute to split the dataset at each node. It measures the reduction in entropy or disorder of the target variable class labels after a dataset is split based on an attribute. A higher Information Gain indicates that splitting on that attribute will result in more organized and predictive subsets, making it a good choice for the root node of the decision tree. From the above dataset we can infer that the best attribute is label as the ‘Item’.

Binning groups continuous data into categories. Equal width makes equal-sized intervals (e.g., 0-10, 10-20), while frequency groups data so each category has similar counts. This simplifies analysis, especially when categories are more meaningful than exact values. The function lets you choose binning type and number of bins for flexibility in data interpretation.

On successful building of the decision tree module we can see that the feature is Area, left and right node values are none. From the provided information, we can infer that 'Item' is identified as the best attribute for the root node of the decision tree, and the accuracy achieved with the decision tree model is 66%.

# CONCLUSION

The study evaluated multiple machine learning classifiers for predicting crop yield categories based on environmental factors. The classifiers exhibited varying levels of performance, with some showcasing exceptional accuracy and precision, while others displayed lower metrics indicating potential overfitting or underfitting. The results underscore the importance of selecting appropriate models and optimizing hyperparameters to achieve accurate predictions in agricultural applications. Moving forward, further research could focus on ensemble methods, deep learning architectures, and advanced feature engineering techniques to enhance predictive capabilities and support sustainable farming practices.

# FUTURE SCOPE

The further improvements that can be done for the extension of the work done are as follows:

## Exploring Deep Learning Architecture:

Delving into the realm of deep learning architectures such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) can unlock insights into capturing intricate spatial and temporal relationships in agricultural data. Leveraging deep learning models has the potential to elevate prediction accuracy and handle large datasets effectively.

## Advancing Feature Engineering Techniques:

Continuous advancements in feature engineering techniques, including interaction features, polynomial features, and domain-specific feature creation, can significantly enhance the predictive capabilities of machine learning models. Employing feature selection methods like recursive feature elimination (RFE) and feature importance analysis can pinpoint crucial variables for precise crop yield prediction.

## Harnessing Data Augmentation and Integration:

Embracing data augmentation techniques such as synthetic data generation, oversampling, and minority class augmentation can address challenges related to class imbalance and improve model performance, especially for rare yield categories. Furthermore, integrating diverse data sources like satellite imagery, soil composition data, and crop health indicators can enrich machine learning models' predictive prowess.

## Implementing Real-time Monitoring and Predictive Maintenance:

Implementing real-time monitoring systems alongside predictive maintenance algorithms can empower proactive management of agricultural equipment, irrigation systems, and pest control measures. Predictive analytics can facilitate early issue identification, optimize resource allocation, and mitigate crop loss risks.

By exploring these, the agricultural sector can harness the full potential of machine learning techniques to boost productivity, sustainability, and resilience in farming practices, ultimately contributing to food security and rural development.

# REFERENCES

1. Van Klompenburg, Thomas, Ayalew Kassahun, and Cagatay Catal. "Crop yield prediction using machine learning: A systematic literature review." Computers and Electronics in Agriculture 177 (2020): 105709.
2. Seireg, Hayam R., et al. "Ensemble machine learning techniques using computer simulation data for wild blueberry yield prediction." IEEE Access 10 (2022): 64671-64687.
3. Khaki, Saeed, and Lizhi Wang. "Crop yield prediction using deep neural networks." Frontiers in plant science 10 (2019): 452963.
4. Rashid, Mamunur, et al. "A comprehensive review of crop yield prediction using machine learning approaches with special emphasis on palm oil yield prediction." IEEE access 9 (2021): 63406-63439.
5. Gopal, PS Maya, and R. Bhargavi. "A novel approach for efficient crop yield prediction." Computers and Electronics in Agriculture 165 (2019): 104968.
6. Nevavuori, Petteri, Nathaniel Narra, and Tarmo Lipping. "Crop yield prediction with deep convolutional neural networks." Computers and electronics in agriculture 163 (2019): 104859.
7. Abbaszadeh, Peyman, et al. "Bayesian multi-modeling of deep neural nets for probabilistic crop yield prediction." Agricultural and Forest Meteorology 314 (2022): 108773.
8. Ali, Mumtaz, et al. "Coupled online sequential extreme learning machine model with ant colony optimization algorithm for wheat yield prediction." Scientific Reports 12.1 (2022): 5488.
9. Muhammad, Khalid Bin, et al. "Iot and cloud based smart agriculture framework to improve crop yield meeting world's food needs." IJCSNS. Vol. 22. No. 6. 2022.
10. Gupta, Monika, et al. "Crop Yield Prediction Techniques Using Machine Learning Algorithms." 2022 8th International Conference on Smart Structures and Systems (ICSSS). IEEE, 2022.
11. Joshua, S. Vinson, et al. "Crop yield prediction using machine learning approaches on a wide spectrum." Computers, Materials & Continua 72.3 (2022): 5663-5679.
12. Li, Linchao, et al. "Developing machine learning models with multi-source environmental data to predict wheat yield in China." Computers and Electronics in Agriculture 194 (2022): 106790.
13. Hussain, Nida, et al. "Predict the crop-yield through uav using machine learning a systematic literature review." 2022 International Conference on IT and Industrial Technologies (ICIT). IEEE, 2022.
14. Rekha, P., et al. "High yield groundnut agronomy: An IoT based precision farming framework." 2017 IEEE Global Humanitarian Technology Conference (GHTC). IEEE, 2017.
15. Prakash, Kalari, et al. "Machine Learning-Based Crop Prediction: A Way Towards Smart Farming." 2022 Seventh International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE, 2022.
16. Varman, S. Aruul Mozhi, et al. "Deep learning and IoT for smart agriculture using WSN." 2017 ieee international conference on computational intelligence and computing research (ICCIC). IEEE, 2017.
17. Bhanu, K. N., H. J. Jasmine, and H. S. Mahadevaswamy. "Machine learning implementation in IoT based intelligent system for agriculture." 2020 International Conference for Emerging Technology (INCET). IEEE, 2020.