Crop Yield Prediction Using ML

Honey Joseph
Dept. of Computer Science
Amal Jyothi College of Engineering
Kottayam, India
honeyjoseph@amaljyothi.ac.in

Aaron M Vinod
Dept. of Computer Science
Amal Jyothi College of Engineering
Kottayam, India
aaronmvinod2024@cs.ajce.in

Abin Mathew varghese
Dept. of Computer Science
Amal Jyothi College of Engineering
Kottayam, India
abinmathewvarghese2024@cs.ajce.in

Aby Alex
Dept. of Computer Science
Amal Jyothi College of Engineering
Kottayam, India
abyalex2024@cs.ajce.in

Aleena Sain
Dept. of Computer Science
Amal Jyothi College of Engineering
Kottayam, India
aleenasain2024@cs.ajce.in

***Abstract*—India's agriculture sector is pivotal to the nation's economy and sustains livelihoods for millions. With diverse agro-climatic zones, India boasts a rich agricultural heritage encompassing crops like rice, wheat, sugarcane, and cotton.For farmers, decision-makers, and other stakeholders to allocate resources and ensure food security, accurate crop yield prediction is essential. This study looks into how machine learning algorithms might be used to increase the precision of crop yield forecasts in India.The study looks at how machine learning models can take into account a number of variables that impact crop yields, such as crop type, season, state, area, fertilizer, pesticide, and rainfall. The effectiveness of various algorithms, such as LinearRegression, Lasso, Ridge and DecisionTreeRegressor, is evaluated.Out of the three Machine Learning methods, the DecisionTreeRegressor algorithm demonstrated the best performance, as seen by its lowest MAE (mean absolute error) value and highest R^2 value. These findings imply that machine learning algorithms have the potential to greatly increase agricultural yield projections' accuracy in Morocco, which might enhance food security and maximize farmers' use of available resources.**

I. INTRODUCTION

The agriculture sector in India is not only a cornerstone of the economy but also a lifeline for millions of people whose livelihoods depend on it. With a rich history of diverse crops grown across various agro-climatic zones, the country faces the imperative task of ensuring food security and effective resource allocation. Machine language (ML) techniques have emerged as promising tools to enhance crop yield predictions by meticulously examining factors such as crop type, seasonality, geographic location, cultivated area, pesticide and fertilizer application, and rainfall patterns. Through sophisticated analysis, these ML models provide valuable insights into the complex dynamics of agricultural productivity, empowering stakeholders to make informed decisions and implement targeted interventions to bolster resilience and sustainability in the face of evolving climatic and socio-economic challenges.Incorporating ML-driven insights into agricultural forecasting not only enhances the accuracy of predictions but also equips stakeholders with the foresight needed to navigate uncertainties posed by climate change and shifting market dynamics. By leveraging these advanced analytical tools, policymakers, farmers, and other stakeholders can devise adaptive strategies to optimize resource utilization, mitigate risks, and promote inclusive growth across the agricultural value chain. Ultimately, the integration of ML techniques into agricultural planning and policy formulation holds tremendous potential to drive transformative change, fostering resilience, and ensuring the long-term prosperity of India's agricultural sector and the well-being of its farming communities.

II. LITERATURE SURVEY

[1].This paper investigates the use of Machine Learning (ML) algorithms for predicting crop yields in Morocco, focusing on factors like weather patterns, soil moisture, and rainfall. It compares ML algorithms (Decision Trees, Random Forests, and Neural Networks) with traditional statistical models. The study finds that ML algorithms outperform statistical models, with the Feed Forward Artificial Neural Network performing the best. The methodology involves training and evaluating algorithms using metrics like Mean Squared Error (MSE) and coefficient of determination (R^2). The study also conducts sensitivity analysis to identify key variables impacting crop yields. Overall, the results support the effectiveness of ML algorithms for crop yield prediction in Morocco, offering insights to optimize resource allocation and enhance food security. [2]. This study introduces a novel approach using a federated random forest algorithm for crop variety yield prediction, addressing data sharing limitations among seed companies. Using maize field trial data from 248 sites across China, the federated method outperformed individual models while maintaining data privacy. The methodology involves data collection, federated model development, and experimen-

tal validation, showcasing its potential to revolutionize crop breeding practices and ensure food security. [3]. Accurate in-season crop yield prediction is vital for farmers, governments, and commodity traders due to yield variability influenced by genotype, environment, and management, compounded by limited high-resolution yield data for within-field estimation. To address this, the authors propose a novel hybrid model integrating modeled crop water stress and remotely sensed vegetation indices for subfield maize yield prediction. Using multi-year maize yield data, weather data, and remotely sensed imagery, they train and validate the model, which incorporates a gap-filling algorithm and a cumulative drought index to address missing data and capture in-season crop water deficit impact on yield. Yield stability maps evaluate model accuracy. The study concludes that the hybrid model improves in-season yield prediction accuracy, providing valuable insights for financial decisions and management practices. The model's potential applications extend to other crops and regions, demonstrating the efficacy of integrating remote sensing and process-based crop models for enhanced subfield yield prediction.

[4].Crop yield prediction is a critical aspect of precision agriculture, especially in regions like India facing potential food crises in the future, underscoring the need for innovative technologies. While traditional statistical models like multivariate linear regression have been used, their accuracy is limited. In contrast, machine learning models such as BPNN, SVM, and GRNN have demonstrated improved performance in crop yield prediction. Environmental factors like rainfall, temperature, humidity, and management practices including fertilizers and pesticides significantly influence crop growth and yield. Understanding these relationships is crucial for accurate prediction. The integration of diverse data sources such as climate data, vegetation indices, and satellite imagery can further enhance prediction models. Future research directions should prioritize the assimilation of multiple data sources, ensuring model scalability across different crops and regions, and developing active learning algorithms to enhance prediction accuracy while reducing labeling costs. [5]Crops yield prediction based on machine learning models: Case of West African countries was a research study that discussed the use of machine learning models in Depending on the study situation and research questions, descriptive or predictive. Predictive models are used to forecast future events, whereas descriptive models are used to learn from the data gathered and describe what has happened. It has been applied to numerous fields, including biology, finance, medicine, and most recently, agriculture, to solve issues.In order to estimate crop yields, machine learning is a crucial tool for decision support. It can help with choices about which crops to plant and how to manage those crops during their growth season. The methodology involves collecting and analyzing climate and agricultural data, applying feature engineering techniques, training machine learning models, and evaluating the performance of the models. The authors collected climate and agricultural data for the region and applied feature engineering techniques to the data. They then used logistic regression, decision trees, and k-Nearest neighbor algorithms to build a prediction model for six crops: maize, yam, cotton, cassava, rice, and banana. The authors used hyper-parameter tuning techniques to avoid overfitting and evaluated the performance of the models using the coefficient of determination. [6]The study used a variety of machine learning methods, such as Lasso regression, Gradient Descent, Random Forest, SVM, and LSTM, to estimate crop yield in Rajasthan, India. The Random Forest algorithm proved to be the most successful of these techniques, estimating crop yield with a high degree of accuracy. By locating a hyperplane in N-dimensional space, SVM was used to classify data points separately, and the Gradient Descent optimisation algorithm was used to reduce the cost function and identify the best-fit coefficients for the data. The study emphasised that in order to improve the performance of the models, a larger dataset and more precise historical data regarding the environment and weather during each crop year are required. Furthermore, the study indicated that district-level statistical data and remote sensing data might be combined to enhance the model's capacity to forecast crop yield even more.

III. PROPOSED METHODOLOGY

Data Collection:

The foundation of our study lies in the collection of agricultural data, which serves as the cornerstone for predictive modeling. We gathered a diverse dataset encompassing various parameters such as crop types, seasonal information, geographical attributes, agricultural inputs (such as fertilizer and pesticide usage), and production metrics. The dataset was meticulously curated from reliable sources, ensuring its integrity and relevance to our research objectives.

Data Preprocessing:

Prior to model training, it was imperative to preprocess the raw data to ensure its compatibility with machine learning algorithms. Our preprocessing pipeline consisted of several key steps:

1) **One-Hot Encoding (OHE)**: Categorical variables such as crop types, seasons, and geographical locations were encoded using the OneHotEncoder function from the scikit-learn library. This transformation converted categorical variables into binary representations, enabling the models to interpret and utilize categorical data effectively.
2) **Standard Scaling**: Numerical features such as agricultural area, production metrics, annual rainfall, fertilizer, and pesticide usage were standardized using the StandardScaler function. Standard scaling ensured that numerical features were on a consistent scale, preventing features with larger magnitudes from dominating the modeling process.
3) **Column Transformation**: The ColumnTransformer function facilitated the simultaneous application of multiple preprocessing techniques to different columns within the dataset. This versatile tool allowed us to

drop unnecessary columns, apply One-Hot Encoding to categorical features, and pass through numerical features without any transformation, streamlining our preprocessing pipeline.

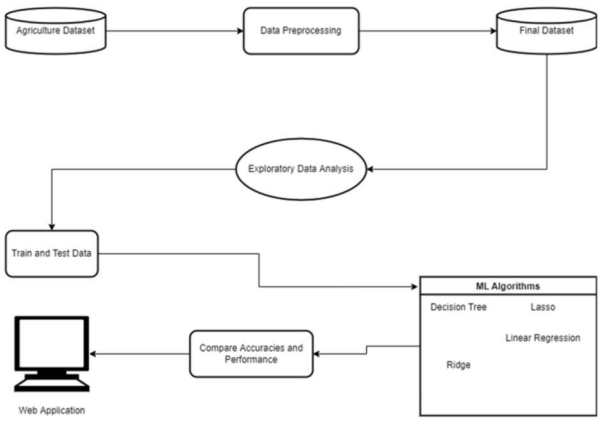

Fig. 1. Methodology

1. Model Training and Evaluation: With the preprocessed dataset in hand, our next step involved training and evaluating multiple regression models to predict agricultural production metrics. We divided the dataset into training and testing sets to assess each model's performance accurately.

2. Model Training: We trained a range of regression models, including Linear Regression, Lasso Regression, Ridge Regression, K Nearest Neighbors Regression (KNN), and Decision Tree Regression. Each model was fitted to the training data using the respective algorithms provided by the scikit-learn library in Python. During training, the models learned the underlying patterns and relationships within the data, adjusting their parameters to minimize the discrepancy between predicted and actual production metrics.

3. Model Evaluation: After training, we evaluated the performance of each model using a set of evaluation metrics. We calculated metrics such as Mean Absolute Error (MAE) and R-squared score (R^2) to assess the models' accuracy and predictive power. MAE measures the average absolute difference between predicted and actual values, providing insights into the models' overall performance. R-squared score quantifies the proportion of variance in the target variable that is explained by the independent variables, with higher values indicating better fit.

4. Cross-Validation: To ensure the robustness of our models, we employed cross-validation techniques such as k-fold cross-validation. This approach partitions the dataset into k equal-sized folds, trains the model on k-1 folds, and evaluates it on the remaining fold. We repeated this process k times, rotating the fold used for evaluation each time. Cross-validation helps in assessing the models' performance across different subsets of the data, reducing the risk of overfitting and providing more reliable performance estimates.

5. Hyperparameter Tuning: Additionally, we performed hyperparameter tuning for model that required it like Decision Tree Regression. Hyperparameters are parameters that are not learned during training and must be set beforehand. We utilized techniques such as grid search or randomized search to systematically explore the hyperparameter space and identify the optimal combination that maximizes the models' performance.

By rigorously training, evaluating, and fine-tuning multiple regression models, we aimed to develop accurate and reliable predictive models for agricultural production forecasting, providing valuable insights for agricultural planning and decision-making.

IV. RESULTS AND DISCUSSIONS

The study's thorough methodology for using machine learning (ML) algorithms to predict crop yields in India produces noteworthy findings with broad ramifications for agricultural practices, the formulation of public policy, and initiatives to ensure food security. Through the validation of machine learning algorithms, specifically the Decision Tree Regressor, and the utilization of rigorous evaluation metrics like R-squared (R^2) and Mean Absolute Error (MAE), the study lays a strong basis for the advancement of agricultural prediction methodologies. Actionable insights are given to agricultural sector stakeholders by this validation, improving resource allocation and management choices.

The decision tree regression's superior crop yield prediction performance highlights its potential as a trustworthy agricultural productivity forecasting tool. Stakeholders can confidently incorporate this approach into decision-making processes, maximizing resource allocation and management strategies, by showcasing its effectiveness in comparison to other algorithms. This validation gives farmers, decision-makers, and other stakeholders the information they need to improve agricultural productivity and address issues related to food security.

Additionally, the sensitivity analysis carried out in the study provides insightful information about the critical elements affecting crop yields for India's major crop varieties. Equipped with this understanding, stakeholders can effectively customize strategies and interventions to maximize agricultural productivity. These insights can be used by policymakers to create focused policies and initiatives that support sustainable farming methods and address particular issues. In order to increase yields while lowering risks, farmers can modify their farming techniques and resource management plans, improving food security and rural communities' standard of living. The study does, however, also recognize the need for more investigation to address its limitations, which include sample size limitations and data scarcity. Future research can improve predictive models and increase their accuracy and applicability in actual agricultural settings by extending datasets and investigating more sophisticated machine learning techniques, like deep learning networks. This acknowledgement of research gaps demonstrates the study's dedication to ongoing innovation and advancement in agricultural prediction techniques. The

development of a user-friendly web platform and the integration of an end-to-end solution augment the research findings' practical utility. This strategy gives farmers the power to make knowledgeable decisions about crop management, resource allocation, and risk mitigation techniques by democratizing access to sophisticated analytical tools and insights. Through the establishment of a link between scientific research and practical agricultural methods, the study bolsters food security, resilience, and sustainability in the farming community.

As a whole, the study's conclusions offer practical advice on how to use machine learning to predict crop yields in India, paving the way for better farming methods and resolving issues with food security. The study advances agricultural prediction methodologies and empowers stakeholders to make informed decisions aimed at improving agricultural productivity and food security by validating ML algorithms, identifying key influencing factors, and integrating an end-to-end solution.

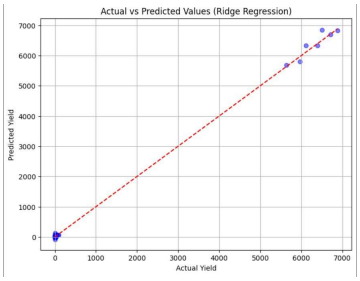

Fig. 2. Ridge algorithm

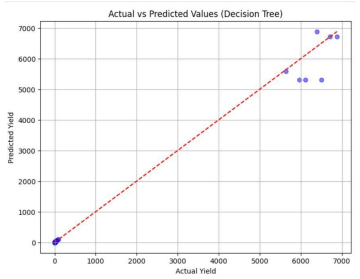

Fig. 3. Decision tree regression

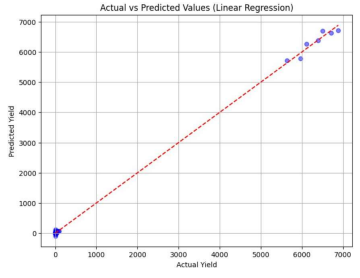

Fig. 4. Linear regression

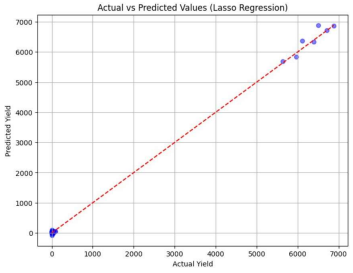

Fig. 5. Lasso regression

V. CONCLUSION

The study's findings show the potential of machine learning algorithms for predicting crop yields in India. Using MEA and R2 metrics, the examination of the Linear Regression, Lasso, Ridge, and DecisionTree Regressor methods revealed that With the best R2 values and the lowest MAE of the three algorithms, the DecisionTreeRegressor algorithm proved to be the most efficient. Important information regarding the factors that most influence crop yields for arecanut, ragi, rice, black pepper, garlic, ginger, groundnut, sesamum, sugarcane, coconut, and numerous other major crops was provided by the sensitivity analysis's results. India's food security can be increased and resource allocation can be optimized with the use of this information. The study's findings support the application of machine learning algorithms for crop production prediction in India and outline the variables that yields from crops. Nevertheless, this study has a number of shortcomings that need be investigated further in subsequent studies. For instance, a larger sample size and more information on the different factors influencing crop yields could increase the prediction's accuracy. Furthermore, more intricate machine learning algorithms, such deep learning networks, might be investigated.

REFERENCES

[1] Rachid Ed-Daoudi, Altaf Alaoui,Badia Ettaki,Jamal Zerouaoui. "Improving Crop Yield Predictions in Morocco Using Machine Learning Algorithms". Journal of Ecological Engineering 2023, 24(6), 392–400

[2] Qiusi Zhang, Xiangyu Zhao, Yanyun Han, Feng Yang, Shouhui Pan, Zhongqiang Liu, Kaiyi Wang b, Chunjiang Zhao. "Maize yield prediction using federated random forest". Computer and Electronics in Agriculture 210 (2023) 107930.

[3] Shuai, Guanyuan, and Bruno Basso. "Subfield maize yield prediction improves when in-season crop water deficit is included in remote sensing imagery-based models." Remote Sensing of Environment 272 (2022): 112938.

[4] Talaat, Fatma M. "Crop yield prediction algorithm (CYPA) in precision agriculture based on IoT techniques and climate changes." Neural Computing and Applications 35.23 (2023): 17281-17292.

[5] Lontsi Saadio Cedric,Wilfried Yves Hamilton Adoni, Rubby Aworka .Crops yield prediction based on machine learning models: Case of West African countries, Smart Agricultural Technology Volume 2, December 2022, 100049

[6] Kavita Jhajhariaa, Pratistha Mathura,Sanchit Jaina,Sukriti Nijhawana. "Crop Yield Prediction using Machine Learning and Deep Learning Techniques". Procedia Computer Science 218 (2023) 406–417.