Crop Yield Prediction According to Soil and Environmental Factors Using Machine Learning

***Abstract*—This paper investigates how machine learning algorithms predicts crop yields depending on soil and environmental Factors, aiming to optimize agricultural output. Data like temperature,rainfall, nutrient levels and soil pH were analyzed using several algorithms including Support Vector Machine (SVM), Random Forest ,Decision Tree Regressor and XG Boost. Among these, Random Forest demonstrated the best performance with an MAE value 730.353, followed by Decision Tree Regressor, though the latter showed higher error margins. The study reveals that ensemble models like Random Forest are way more successful in executing complex, non-linear relationships in the middle of input variables and crop yields, outperforming simpler models such as SVM, which struggled with low correlation between features and outcomes. Accurate crop yield predictions are critical for improving resource allocation, minimizing waste, and enhancing food security, especially in the face of climate variability. The future will concentrate on enhancing these models and broadening their scope of application to different crops and geographical regions to further support sustainable agricultural practices.**

***Index Terms*—Decision Tree Regressor, XGBoost, non-linear, handling complex.**

I. Introduction

Agriculture is essential to humanity, not just as a source of food but also for employment and economic stability. Though humans have consumed grains and plants for over 100,000 years, the practice of actively managing land and crops began about 11,000 years ago, during the Neolithic period, commonly known as the New Stone Age. In India, agriculture is crucial in supporting the nation's economy and fulfilling most of its food requirements. However, with the country's rapidly growing population and significant climate changes, maintaining a balance between food demand and supply has become critical. To address these challenges, various scientific techniques have been integrated into agriculture to enhance sustainability and productivity. The variability of climatic conditions further complicates decision-making for farmers, as they seek to adapt and remain flexible. In this context, modern technology and innovative farming methods are essential for maximizing output with minimal resources. Accurate crop production estimation is vital for addressing food security concerns, especially in India, where agriculture consumes 70% of the world's available freshwater resources [1].

Machine learning (ML) algorithms have proven to be highly accurate in predicting crop yields. They are particularly well suited for managing complex datasets, which often feature high dimensionality, multicollinearity, outliers, and nonlinear relationships. When combined with high-resolution earth observation data, both spatially and temporally, ML techniques show significant promise in yield prediction. These algorithms are robust, capable of learning patterns from large datasets to generate precise predictions. For instance, algorithms such as Random Forest (RF) are flexible and effective in handling unstructured data, enabling the discovery of complex relationships. However, the predictive power of ML algorithms varies focusing on the characteristics of the input data. While the current study focused on using RF for spatial prediction of maize yield, future research should explore a comparative analysis of different ML algorithms to assess their predictive performance.

In 2018, more than 50% of India's workforce was employed in agriculture, contributing approximately 17-18% to the nation's GDP. Between 1971 and 2020, the country's total land area remained consistent at around 328.7 million hectares [2]. By 2025, India's agricultural sector is expected to reach a value of USD 24 billion. With retail contributing 70% of total sales, India ranks as the sixth-largest food and grocery market in the world. Based on early estimates for the fiscal year 2022-2023, India's total food grain production, including the Kharif season, will reach 149.92 million tons. The sector's growth is largely driven by the country's rapidly expanding population.

Crop yield estimation allows farmers to increase production during favorable conditions and minimize losses during adverse conditions. Several factors influence accurate yield predictions, including farming practices, pest control, fertilizers, weather patterns, and market prices. By analyzing historical yield data alongside variables like climate, regional output, and rainfall, crop yield can be effectively estimated. In recent years, machine learning has made significant advancements across various industries, including agriculture, offering new tools to enhance the accuracy of these predictions.

Various machine learning techniques, including, Artificial Neural Network, Decision Trees, SVR, and Deep Learning [3], have been utilized to estimate agricultural production with greater accuracy and efficiency.

India contributes approximately 40% to the global rice trade, both Basmati and non-Basmati varieties, exporting to over 150 countries. According to data from the Trade Ministry, rice exports increased by 11%, reaching 2.16 million tons [11] in the first half of the 2022-23 fiscal year. Moreover, areas such as classification and fruit recognition have emerged as significant advancements in image classification and computer vision, particularly in the agriculture industry [5].

Identify applicable funding agency here. If none, delete this.

A. Literature Review

Crop yield prediction is an integral part of precision agriculture, enabling farmers and policymakers to establish informed decisions about resource allocation and crop management. Machine Learning models have gained significant attraction in this domain, with several techniques being applied to analyze soil, environmental, and climatic factors to improve yield forecasting. Kamir et al. [6] demonstrated that ensemble learning models, particularly support vector regression (SVR), effectively predict wheat yields by leveraging data such as weather patterns, NDVI, and geographical location, achieving an $R\hat{2}$ score of 0.73. Similarly, Aghighi et al. [7] found that Boosted Regression Trees outperformed other models for predicting silage maize yields using satellite data and surface reflection, underscoring the importance of ensemble models in processing complex environmental data.

Further studies highlight the potential of deep learning models for yield prediction. Kuwata and Shibasaki [8] used deep learning (DL) and SVR models to predict corn yields, with the DL model achieving superior performance by capturing non-linear relationships between vegetation indices and climatic data. Recurrent neural networks (RNNs) have also been employed for rice yield prediction, as shown by Kulkarni et al. [9], who utilized 31 years of soil and rainfall data, achieving an RMSE of 41.497. These studies emphasize that neural network models, particularly deep learning architectures, are well-suited for analyzing temporal and spatial data in agricultural forecasting.

The use of decision tree-based models like Random Forest has also proven effective in crop yield prediction. Multiple studies from India have demonstrated that Random Forest and other tree-based models outperformed traditional regression methods in predicting rice, soybean, and maize yields [10], [**?**]. For instance, in Maharashtra, Random Forest models provided the highest accuracy for rice yield prediction compared to neural networks and SVR [5]. This trend highlights the versatility of decision tree-based methods in handling diverse environmental inputs, including soil properties and climatic variables, making them a reliable tool in precision agriculture. Overall, ML models continue to show promise in optimizing crop yield predictions, particularly when integrating multi-source environmental data.

II. METHODOLOGY

The methodology follows a systematic approach based on the Figure 1 flowchart and specified data features. The process involves several stages, from data collection and prepossessing to model development and evaluation.

A. Data Collection

The primary step involves gathering relevant data for crop yield prediction. The following features are collected from reliable sources such as agricultural departments, meteorological organizations, and soil health labs:

1) State: Geographic region where the crop is grown.

2) Crop: Type of crop under consideration.

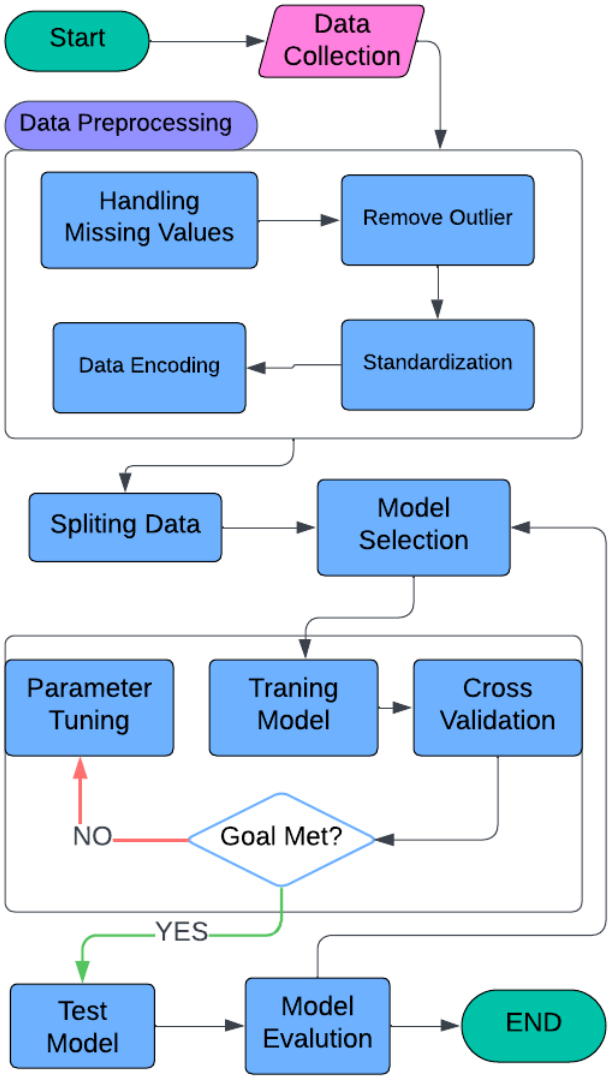

Fig. 1. Flowchart of the methodology

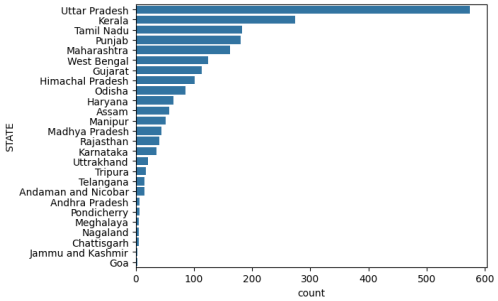

Fig. 2. State count

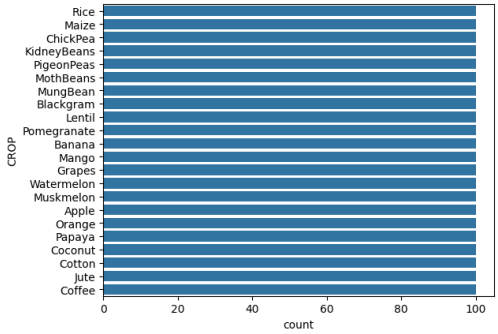

Fig. 3. Croup count

3) Rainfall: The total amount of rainfall (in mm) during the growing season.

4) Temperature: Average temperature (in °C) during the crop's life cycle.

5) Humidity: Average humidity level (in %) throughout the season.

6) Soil pH: Acidity or alkalinity of the soil.

7) Crop Price: Market price of the crop as a proxy for economic yield.

8) Soil Potassium (K_SOIL): **, Soil Nitrogen (N_SOIL), and Soil Phosphorus (P_SOIL)** Levels of essential nutrients in the soil.

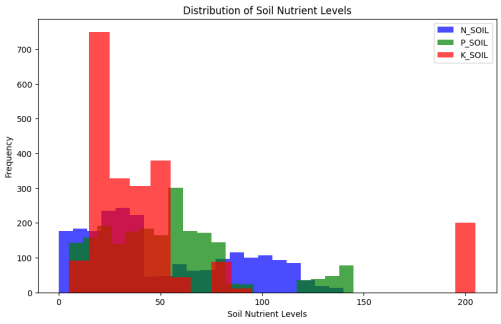

Fig. 4. Distribution of Soil Feature Level

B. Data Preprocessing

Data preprocessing ensures that the dataset is clean and structured for machine learning modeling. Key steps include:

1) Handling Missing Values: Missing data is addressed through imputation techniques, such as filling in missing values with the median or using k-nearest neighbors (KNN) imputation.

2) Outlier Removal: The Interquartile Range (IQR) method is applied to detect and remove extreme outliers in the dataset, especially for numerical variables like soil nutrient levels and weather data.

3) Normalization: Features such as soil nutrient concentrations and environmental factors are normalized using min-max scaling or z-score standardization to ensure that all features are on the same scale.

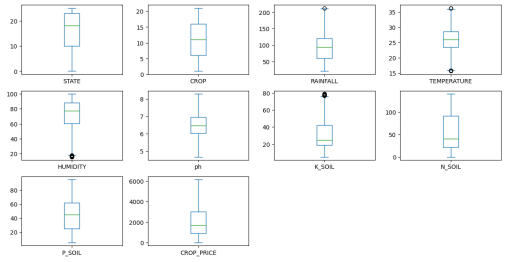

Fig. 5. After Remove Outlier

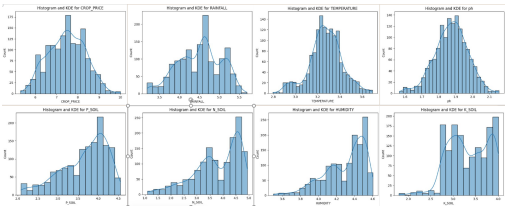

Fig. 6. Standardized Data

4) Categorical Data Encoding: The categorical features 'STATE' and 'CROP' are label-encoded, transforming them into numerical values to be used in machine learning models.

C. Feature Selection

To reduce complexity and improve model accuracy, feature selection techniques are employed:

1) Correlation Analysis: Pearson or Spearman correlation coefficients are used to identify relationships between independent variables (e.g., soil and environmental factors) and the dependent variable (crop yield). Highly correlated variables are further analyzed for multicollinearity.

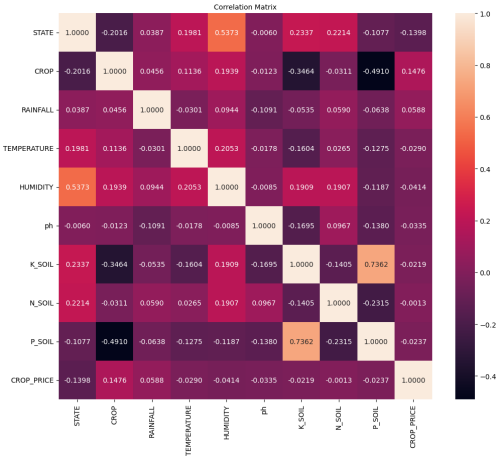

Fig. 7. Correlation Matrix

2) Recursive Feature Elimination (RFE): RFE is applied to iteratively eliminate less important features, retaining the most significant predictors that contribute to the model's accuracy.

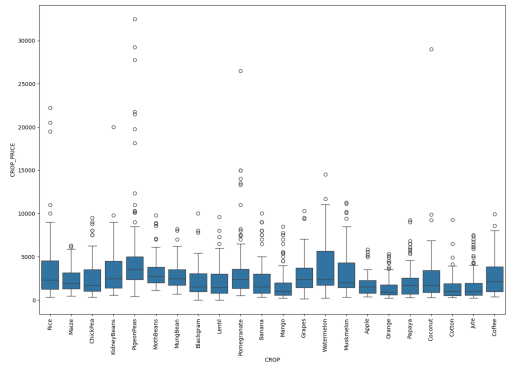

Fig. 8. Standardized Data

D. Model Development

Various Machine Learning algorithms are tested to predict crop yield according to the selected soil and environmental features:

1) Linear Regression: This algorithm is used as a baseline to capture linear relationship between the features and crop yield.

2) Random Forest(RF): As an ensemble method, Random Forest is applied to establish non-linear relationship between features. It builds multiple decision trees and aggregates their predictions to improve accuracy.

3) Support Vector Machine(SVM): SVM is used with a non-linear kernel to find the optimal hyperplane that separates data points for crop yield prediction.

4) XGBoost: The gradient boosting algorithm is employed for its efficiency in handling large datasets and its ability to handle missing data and non-linear patterns effectively.

5) Decision Tree Regressor: A simple tree-based model that splits the data recursively to form a predictive tree for crop yield based on decision rules.

6) KNeighbors Regressor: A distance-based model that predicts crop yield by averaging the outputs of the k-nearest data points in the feature space.

7) Bagging Regressor : An ensemble method that improves model performance by training multiple instances of the same model (e.g., Dicision Trees) on different subsets of the dataset and aggregating their predictions.

8) Extra Trees Regressor: A variant of Random Forest that further reduces variance by splitting nodes randomly in decision trees, providing an even stronger ensemble model for crop yield prediction.

E. Model Evaluation

The performance of each Machine Learning model is evaluated on the test set using the following metrics:

1) Mean Absolute Error(MAE): MAE is used to measure average magnitude of prediction errors.

2) Root Mean Squard Error (RMSE): RMSE is computed to capture the magnitude of larger prediction errors, providing insight into the variability in the model's predictions.

TABLE I
OVERALL RESULT

ML Models	MAE	RMSE	R2 Value
Decision Tree Regressor	2359.186	3777.646	84.24%
Random Forest Regressor	730.353	1132.865	83.43%
XGB Regressor	2063.249	3424.124	51.38%
Extra Trees Regressor	1948.262	2996.36	15.91%
SVR	1653.335	2911.02	9.41%
K Neighbors Regressor	1724.3	2374.177	17.78%
Bagging Regressor	1798.064	2474.87	27.98%

3) R-squared (R^2): R^2 is used to explain the variance in crop yield accounted for by the soil and environmental factors.

F. Testing and Final Model Selection

Once the model reaches the desired performance metrics during training and cross-validation, the final model is tested on a reserved test set. The final evaluation metrics are reported, and the model is prepared for deployment or further validation in real-world agricultural settings.

The developed machine learning model integrates soil properties, weather data, and economic indicators to predict crop yield. The model provides a decision-support tool for farmers and agricultural planners, enabling them to make data-driven decisions to optimize crop prediction and improve economic outcomes. This methodology outlines a robust, systematic approach to crop yield prediction based on interdisciplinary factors, including agronomy, meteorology, and machine learning techniques.

III. RESULTS

From TABLE 1, we can see the results of various machine learning regression models are presented in the table, including their efficiency in terms of Root Mean Squared Error (RMSE), Mean Absolute Error (MAE) and success rate. Among the models evaluated, Random Forest Regressor earned the finest performance with a MAE of 730.353 and RMSE of 1132.865, along with a high success rate of 83.43%. Following closely is the Decision Tree Regressor, which also showed a strong success rate of 84.24%, albeit with higher error values (MAE of 2359.186 and RMSE of 3777.646). On the other hand, models such as the XG Boost Regressor and Bagging Regressor performed moderately, with success rates of 51.38% and 27.98%, respectively. The Extra Trees Regressor and K Neighbors Regressor exhibited poor success rates of 15.91% and 17.78%, with significantly higher error metrics. The SVR (Support Vector Regressor) model displayed the lowest success rate at 9.41%, with corresponding high error values.

In summary, the Random Forest Regressor emerged as the top model with the best combination of accuracy and error metrics, while other models such as SVR and Extra Trees Regressor struggled with low success rates and higher prediction errors.

IV. DISCUSSIONS

The evaluation of different machine learning models shows that the performance is highly influenced by the correlation

between features and the target variable. When correlation is low, simpler models like Linear Regression and Support Vector Regression (SVR) often struggle because they rely on direct, linear relationships between the input features and the target. As a result, these models fail to capture the underlying patterns when the relationships are more complex or non-linear, leading to poor performance. This is evident in our results where SVR had one of the lowest success rates (9.41%) and high error values, indicating its inability to manage weakly correlated data.

In contrast, ensemble models like Random Forest performed significantly better, achieving an 83.43% success rate with low error values (MAE: 730.353). This is because ensemble models, especially those based on decision trees, are well suited for handling complex, non-linear relationships in the data. Random Forest enhances accuracy and reduces variance by aggregating predictions from multiple decision trees, even when individual features have a low correlation with the target variable. Similarly, although the Decision Tree Regressor had a high success rate (84.24%), its higher error values suggest that it is more prone to over-fitting compared to Random Forest, which is mitigated by the averaging of multiple trees in the ensemble.

Models like XG Boost and Bagging Regressor performed moderately well but still lagged behind Random Forest. This suggests that while these ensemble methods can manage low correlation, further tuning may be required to fully optimize their performance. Overall, the results indicate that ensemble methods are more resilient to low correlation, while simpler linear models struggle to adapt in such cases. The key to improving overall performance lies in refining feature selection and engineering to better capture relevant patterns in the data.

V. Conclusion

In this research, we applied machine learning techniques to predict crop yields depending on various environmental and soil parameters. Our study demonstrated that ensemble models, particularly Random Forest, outperformed other models in handling the complexities of crop yield prediction, especially when correlation between features and the target variable was low. By leveraging historical data on factors like humidity ,temperature, pH and soil moisture, we created a predictive model that delivers precise crop yield forecasts, greatly assisting farmers in making well-informed choices regarding crop selection and management.

Our findings align with broader research that highlights the effectiveness of machine learning models like Random Forest in capturing intricate relationships between environmental conditions and crop yields. These models help reduce uncertainty, optimize resource usage, and improve crop management strategies. As climate change continues to impact agriculture, the ability of machine learning models to provide data-driven insights will be essential for promoting sustainable farming practices and ensuring food security.

However, the accuracy of these predictions depends on the quality and availability of data. Continued research in this area is crucial to refining these models and extending their applicability to a wider range of crops and environmental conditions.

References

[1] M. Khan and S. Noor, "Irrigation runoff volume prediction using machine learning algorithms," Eur. Int. J. Sci. Technol., vol. 8, pp. 1–22, Jan. 2019.

[2] NIC IN. "Annual Report 2020–21." Accessed Jan. 20, 2023. [Online]. Available: https://agricoop.nic.in/Documents/annual-report-2020-2

[3] M. Khan and S. Noor, "Performance analysis of regression-machine learning algorithms for predication of runoff time," Agrotechnology, vol. 8, no. 1, pp. 1–12, 2019.

[4] K. Elissa, "Title of paper if known," unpublished.

[5] H. S. Gill, G. Murugesan, B. S. Khehra, G. S. Sajja, G. Gupta, and A. Bhatt, "Fruit recognition from images using deep learning applications," Multimedia Tools Appl., vol. 81, no. 23, pp. 33269–33290, Sep. 2022.

[6] S. Kamir, L. Wang, and P. Khaki, "Crop yield prediction using deep neural networks," Frontiers in Plant Science, vol. 10, p. 621, May 2019.

[7] H. Aghighi, F. Omid, and B. Esmaeilpour, "Maize crop yield prediction using machine learning techniques," Environmental Modeling Software, vol. 22, no. 6, pp. 5687–5708, Aug. 2020.

[8] M. Kuwata and R. Shibasaki, "Machine learning approach for crop yield prediction based on time-series data from satellites and climatic factors," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 7, no. 12, pp. 5070–5081, Dec. 2015.

[9] V. Kulkarni, K. Gopal, and S. S. Naik, "Rice yield prediction using time-series soil and rainfall data," Int. J. Agric. Sci., vol. 31, no. 2, pp. 99-110, 2020.

[10] P. S. M. Gopal and R. Bhargavi, "Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms," *Appl. Artif. Intell.*, vol. 33, no. 7, pp. 621–642, Jun. 2019.

[11] M. S. Rao, A. Singh, N. V. S. Reddy, and D. U. Acharya, "Crop prediction using machine learning," J. Phys.: Conf. Ser., vol. 2161, no. 1, pp. 1–10, 2022.