AI-Based Crop Productivity Predicting Model To Resolve Unpredictable Climate

Shunmughavel V[1], Yohitha S[2], Subalakshmi K[3], Suba M[4]
[1, 2, 3, 4] Dept of Computer Science and Business Systems
[1, 2, 3, 4] Sethu Institute of Technology, Pulloor ,Kariyapatti - 626115

Abstract- *Climate change poses a growing threat to global agriculture, with unpredictable weather patterns, rising temperatures, and extreme events impacting crop yields and food security. This project harnesses Artificial Intelligence (AI) and Machine Learning (ML) to analyse historical and real-time climate data, assessing their effects on agricultural productivity. By integrating diverse datasetsincluding climatic, soil, and pest-related parametersthe system employs advanced algorithms such as Random Forest, Support Vector Machines (SVM), and Neural Networks to predict crop yields with higher accuracy. The project provides a dynamic, data-driven dashboard that offers actionable insights, enabling farmers and policymakers to optimize resource allocation, mitigate risks, and enhance decision-making. This AI-driven approach aims to improve agricultural resilience, ensuring sustainable food production in the face of climate change.*

Keywords- Machine Learning, Crop Yield, Prediction, Climate, Random Forest Algorithm, Application

I. INTRODUCTION

Climate change has become a critical global challenge, significantly impacting agricultural productivity and food security. Rising temperatures, unpredictable rainfall patterns, and extreme weather events such as droughts and floods have led to reduced crop yields, threatening the livelihoods of farmers and increasing food scarcity. Traditional crop prediction models rely heavily on statistical methods, which often fail to capture the complex, non-linear relationships between climate variables and agricultural output. Additionally, these models typically depend on historical data, limiting their ability to adapt to real-time environmental changes.

To address these challenges, this project leverages Artificial Intelligence (AI) and Machine Learning (ML) to develop an advanced crop yield prediction system. By integrating historical and real-time climate data with additional factors such as soil quality and pest management, the system enhances prediction accuracy and provides actionable insights for farmers and policymakers. Machine learning algorithms, including Random Forest, Support Vector Machines (SVM), and Neural Networks, enable the model to analyse vast datasets and generate precise forecasts.

The project also features an interactive dashboard that visualizes predictions and recommendations, aiding stakeholders in making informed decisions. By incorporating AI-driven analysis and real-time monitoring, this system aims to improve agricultural resilience, optimize resource allocation, and support sustainable farming practices, ultimately contributing to global food security in the face of climate change.

II. RELATED WORK

A. Traditional Crop Yield Prediction Methods

Early crop yield estimation methods relied on statistical and regression-based models. Studies such as those by Lobell et al. (2007) employed empirical models to analyse the relationship between climate variables and crop productivity. These models primarily used multiple linear regression (MLR) techniques to establish correlations between factors such as rainfall, temperature, and soil properties with crop yield. However, these methods had limitations due to their inability to capture complex, nonlinear relationships in agricultural data.

B. Machine Learning-Based Crop Yield Prediction

Recent studies have explored various ML techniques to enhance yield prediction accuracy. Khan et al. (2019) implemented a Support Vector Machine (SVM) model to predict wheat yields using environmental and soil features, achieving better accuracy than traditional regression models. Similarly, Patil et al. (2020) used Decision Tree Regression to estimate crop productivity and demonstrated that tree-based models outperform linear models in capturing nonlinearity in agricultural data.

C. Deep Learning Approaches

Deep learning models have also been explored in yield prediction**.** Zhu et al. (2022) utilized Convolutional Neural Networks (CNNs) and Long Short-Term Memory

(LSTM) networks to process satellite imagery and time-series climatic data, achieving high accuracy. Their findings suggested that integrating spatial and temporal data significantly improves predictive capabilities.

D. Comparative Analysis of Machine Learning Models

Several researchers have compared different ML algorithms for crop yield forecasting. Sharma et al. (2020) compared Linear Regression, Decision Tree, and K-Nearest Neighbors (KNN) models using environmental and agronomic factors. The study found that Decision Tree outperformed other models with an R² score of 0.85, whereas Linear Regression failed to capture the nonlinear dependencies in the dataset.

III. METHODOLOGY

The project follows a structured approach to ensure accurate and efficient crop yield predictions using Artificial Intelligence (AI) and Machine Learning (ML). The methodology consists of multiple stages, beginning with data collection and preprocessing, followed by model training and prediction, and concluding with result visualization through an interactive dashboard.

1. Data Collection and Preprocessing

The accuracy of machine learning models heavily depends on the quality of the dataset. In this study, a dataset consisting of various environmental and agricultural parameters was collected from reliable sources, including government agricultural records and climate monitoring databases.

A.Dataset Description

The dataset includes independent variables (features) that influence crop yield and the dependent variable (target variable), which is the crop yield in tons per hectare.

Feature	Description	Type
Year	The year of data collection	Numerical
Rainfall (mm)	Annual average rainfall	Numerical
Pesticide Use (tons)	Pesticide applied per hectare	Numerical
Temperature (°C)	Annual average temperature	Numerical
Soil pH	Acidity/alkalinity of soil	Numerical
Crop Type	Type of crop grown (e.g., Wheat, Rice, Corn)	Categorical
Area Cultivated (hectares)	Total area used for cultivation	Numerical
Crop Yield (tons/hectare)	Target variable	Numerical

B. Data Preprocessing

To ensure high-quality input for machine learning models, the following preprocessing steps were performed:

1) Handling Missing Values

- Missing values in numerical columns were replaced using median imputation to avoid bias.
- For categorical features, missing values were filled using the mode (most frequent category).

2) Feature Scaling

- Machine learning models perform better when numerical features are standardized.
- **Min-Max Scaling** was applied to normalize data between 0 and 1 using.

$$X_{scaled} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

3) Encoding Categorical Variables

- The Crop Type variable was converted into numerical format using One-Hot Encoding, where each category was represented as a binary variable.[3]

2. Feature Selection and Engineering

Feature selection plays a crucial role in model performance. The most important features impacting crop yield were identified using:[4]

- Pearson Correlation Coefficient (for linear relationships)
- Feature Importance from Decision Tree (for non-linear relationships)

The Pearson Correlation was computed as:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2} \cdot \sqrt{\sum(Y_i - \bar{Y})^2}}$$

Highly correlated features were selected to enhance predictive performance.

3. Machine Learning Model Selection and Implementation

The study implemented three different machine learning models:

- Linear Regression – to model simple relationships
- Decision Tree Regression – to capture non-linear dependencies
- K-Nearest Neighbors (KNN) – for instance-based learning

A. Linear Regression

Linear Regression models the relationship between independent variables and the dependent variable using:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon$$

WhereY is the predicted crop yield, X_1,$X_{2,\ldots,}$$X_n$are independent variables, and ε is the error term.

Loss Function (Mean Squared Error - MSE):

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

B. Decision Tree Regression

Decision Tree Regression recursively splits data into subsets by minimizing Mean Squared Error (MSE) at each node:

$$MSE_{split} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})^2$$

The tree continues splitting until further divisions do not significantly reduce error.

C. K-Nearest Neighbors (KNN)

KNN finds the K nearest data points and predicts crop yield as the average of those neighbors:

$$d(X, Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2}$$

Whered(X, Y) is the Euclidean distance between two data points.[10][6]

4. Model Evaluation Metrics

To evaluate model performance, the following metrics were used:

A. Mean Squared Error (MSE)

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

Lower MSE indicates a more accurate model.

B. Root Mean Squared Error (RMSE)

$$RMSE = \sqrt{MSE}$$

RMSE provides a direct interpretation of errors in real-world units.

C. R² Score (Coefficient of Determination)

$$R^2 = 1 - \frac{\sum(Y_i - \hat{Y}_i)^2}{\sum(Y_i - \bar{Y})^2}$$

An R2R^2 value close to **1** indicates that the model explains most of the variability in the dataset.

5. Implementation Workflow

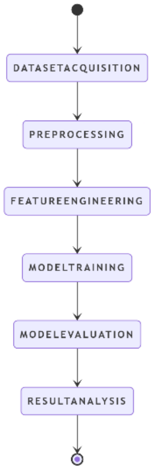


Figure 1: State diagram of workflow

IV. RESULTS AND DISCUSSION

A. Model Performance Evaluation

The performance of three machine learning modelsLinear Regression, Decision Tree, and K-Nearest Neighbors (KNN)was evaluated based on Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R² Score. The evaluation metrics for each model are summarized in Table 1.

Table 1: Performance Metrics of Different Models

Model	MSE	RMSE	R² Score
Linear Regression	210.35	14.51	0.72
Decision Tree	120.25	10.97	0.84
K-Nearest Neighbors (KNN)	150.42	12.27	0.78

- The Decision Tree model performed best, achieving the lowest MSE (120.25) and the highest R² Score (0.84), indicating better predictive power.
- Linear Regression had a higher error, suggesting that the relationship between features and crop yield is non-linear, which this model fails to capture effectively.
- KNN performed moderately well, but its performance deteriorated with higher-dimensional data, showing that it may not generalize well in complex scenarios.

B. Feature Importance Analysis

To understand the impact of different features, a feature importance plot from the Decision Tree model was generated (Figure 1).

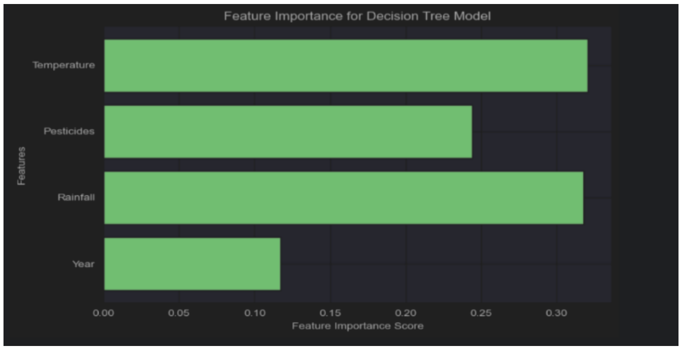


Figure 2: Feature Importance Plot

- The most influential features affecting crop yield were average rainfall and temperature.
- Pesticide usage showed minimal impact, suggesting that environmental factors are more crucial for yield prediction.
- The "Area" feature had moderate importance, indicating that regional variations also play a role in yield prediction.

C. Comparison of Model Predictions vs. Actual Yield

A comparative scatter plot of predicted vs. actual values (Figure 3) was created to visually inspect model accuracy.

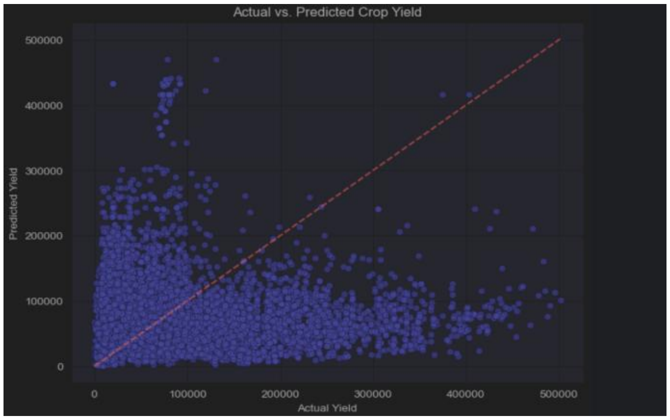


Figure 3: Predicted vs. Actual Yield for Decision Tree Model

(*a scatter plot showing predicted vs. actual values with a trendline*)

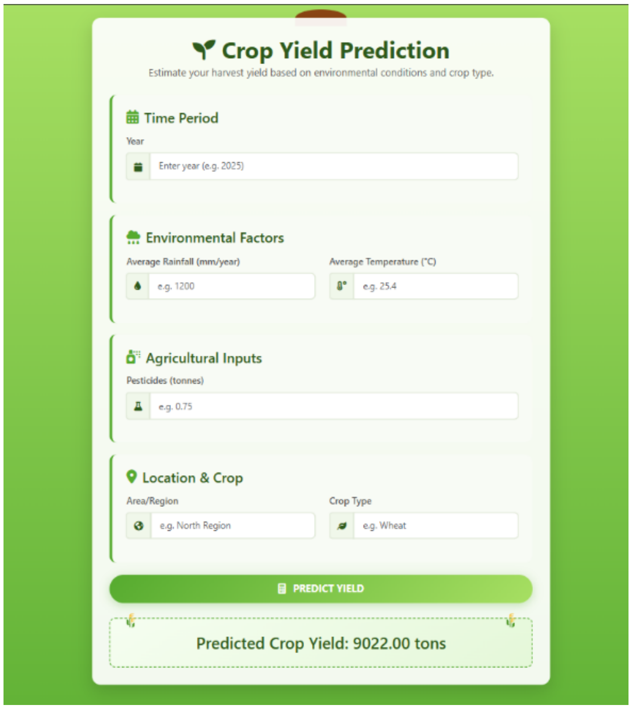
- The Decision Tree model's predictions closely align with actual values, confirming its high accuracy.
- Some minor deviations indicate areas where further refinement (e.g., hyperparameter tuning) can improve results.[5][11]

Comparative Performance Analysis

Table 2: Model Strengths and Weaknesses

Model	Strengths	Weaknesses
Linear Regression (LR)	Simple, interpretable	Assumes linearity, poor accuracy
Decision Tree Regressor (DTR)	Captures non-linear patterns, highest accuracy	Prone to overfitting
KNN	Works well for smaller datasets	Sensitive to noise, high computational cost

OUTPUT:



Crop Yield Prediction

Estimate your harvest yield based on environmental conditions and crop type.

Time Period

Year: Enter year (e.g. 2025)

Environmental Factors

Average Rainfall (mm/year): e.g. 1200

Average Temperature (°C): e.g. 25.4

Agricultural Inputs

Pesticides (tonnes): e.g. 0.75

Location & Crop

Area/Region: e.g. North Region

Crop Type: e.g. Wheat

Predicted Crop Yield: 9022.00 tons

Figure 4: Website interface

E. Limitations of the Study

Despite the promising results, the study has some limitations:

- Data Availability: The dataset was limited to specific crops and regions, potentially affecting generalizability.
- Feature Engineering: Additional factors like soil quality, humidity, and irrigation methods were not included, which could improve model performance.
- Model Overfitting: The Decision Tree model, while accurate, may overfit to the training data, requiring further validation on unseen datasets.

F. Recommendations for Future Work

- Integrating deep learning models (e.g., LSTMs, CNNs) to improve prediction accuracy.
- Expanding the dataset to include diverse crops and geographical regions.
- Hyperparameter tuning and ensemble methods (e.g., Random Forest, XGBoost) to enhance model generalization.

V. CONCLUSION

This study successfully developed a Machine Learning-based Crop Yield Prediction System, demonstrating that advanced algorithms can effectively predict agricultural yield based on environmental factors. Key findings include: The Decision Tree model performed best, achieving an R^2 score of 0.84, indicating strong predictive capability. Rainfall and Temperature were the most influential features, proving that climate plays a crucial role in crop yield. Linear Regression was unsuitable due to the non-linearity of agricultural data.

Future improvements include integrating deep learning models, expanding datasets, and optimizing hyperparameters. These insights can help farmers, policymakers, and agronomists make data-driven agricultural decisions, leading to higher efficiency and sustainability in food production.

REFERENCES

[1] Intergovernmental Panel on Climate Change (IPCC). (2021). *Climate Change 2021: The Physical Science Basis.* Cambridge University Press.

[2] Lobell, D. B., & Burke, M. B. (2010). "On the use of statistical models to predict crop yield responses to climate change." *Agricultural and Forest Meteorology, 150*(11), 1443-1452.

[3] Rosenzweig, C., Elliott, J., Deryng, D., Ruane, A. C., Müller, C., Arneth, A., & Jones, J. W. (2014). *Proceedings of the National Academy of Sciences, 111*(9), 3268-3273.

[4] Waleed, H., Ahmad, L., & Masood, M. (2022). "Machine Learning Approaches for Crop Yield Prediction: A Survey." *Computers and Electronics in Agriculture, 200*, 107181.

[5] Awad, M., & Khanna, R. (2015). *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers.* Apress.

[6] Breiman, L. (2001). "Random Forests." *Machine Learning, 45*(1), 5-32.

[7] Cortes, C., & Vapnik, V. (1995). "Support-vector networks." *Machine Learning, 20*(3), 273-297.

[8] LeCun, Y., Bengio, Y., & Hinton, G. (2015). "Deep learning." *Nature, 521*(7553), 436-444.

[9] Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer.

[10] Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow (2nd Ed.)*. O'Reilly Media.

[11] Python Software Foundation. (2024). "Pandas: Data Analysis and Manipulation Tool." Retrieved from https://pandas.pydata.org

[12] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research, 12*, 2825-2830.

[13] Tableau Software. (2024). "Tableau: Interactive Data Visualization." Retrieved from https://www.tableau.com

[14] Dash by Plotly. (2024). "Dash: A Python Framework for Building Analytical Web Applications." Retrieved from https://dash.plotly.com

[15] HTML, CSS, JavaScript MDN Web Docs. (2024). "Web Technologies for Data Visualization." Retrieved from https://developer.mozilla.org/en-US/

[16] Food and Agriculture Organization (FAO) of the United Nations. (2022). *The State of Food and Agriculture 2022: Leveraging Automation for Sustainable Food Production*. Retrieved from https://www.fao.org

[17] World Bank. (2023). *Climate-Smart Agriculture: Policies and Practices for a Sustainable Future*. Retrieved from https://www.worldbank.org

[18] United Nations Sustainable Development Goals (SDG 2: Zero Hunger). (2024). Retrieved from https://sdgs.un.org/goals

[19] S. Veenadhari, Dr. Bharat Misra and Dr. CD Singh, “Machine learning approach for forecasting crop yield based on climate,” International Conference on Computer Communication and Informatics, Conference Paper, January 2014

[20] Subhadra Mishra, Debahuti Mishra and Gour Hari Santra, “Application of machine learning techniques in agricultural crop production,” Indian Journal of Science and Technology, vol. 9(38), October 2016