Predicting Agriculture Yields Using Machine Learning

Shaik Amrin[1], B. Murali[2]
[1]*PG Student, CSE, Quba College of Engineering & Technology*
[2]*Associate professor, CSE, Quba College of Engineering & Technology*

***Abstract*— Agriculture contributes a significant amount to the economy of India due to the dependence on human beings for their survival. The main obstacle to food security is population expansion leading to rising prediction, technology can assist farmers in producing more. This project main goal is to predict crop yields utilizing the features of amount of rainfall, crop, area, production, pesticides and fertilizers that have posed a serious threat to the long-term viability of agriculture. Crop yield prediction is a decision support tool that uses machine learning algorithms, that can be used to make decisions about which crops to produce. To estimate the agriculture yield, machine learning techniques: Decision Tree, Random Forest, Support Vector Machine, K Nearest Neighbors Regressor and XG-Boost have been used and for performance evaluation accuracy, root mean square error, mean square error and mean absolute error are compared.**

***Index Terms*—. Crop Yield Prediction, Machine Learning, Agriculture, Food Security, Sustainable Agriculture**

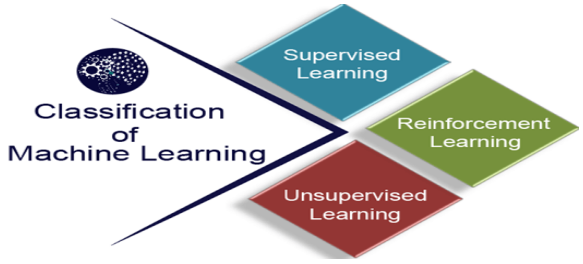
I. INTRODUCTION

ABOUT MACHINE LEARNING

Machine learning is a subset of artificial intelligence that focuses on developing algorithms and statistical models that enable computers to learn and improve performance on a specific task without explicit programming. It involves the use of data to train models, allowing systems to make predictions or decisions based on patterns and trends. The key components of machine learning include input data, a learning algorithm, and a trained model that can generalize to new, unseen data. It finds applications in various fields, such as image recognition, natural language processing, and recommendation systems. The ultimate goal is to create intelligent systems that can adapt and improve their performance over time.

Machine learning is instrumental in revolutionizing industries through its capacity to enable computers to learn from data and autonomously make informed decisions. Its importance lies in analyzing extensive datasets, identifying patterns, and extracting valuable insights, empowering businesses to make data-driven decisions. In healthcare, finance, and cybersecurity, machine learning enhances predictive analytics for early disease detection, risk assessment, and threat identification. It also drives innovation in technologies like self-driving cars and personalized recommendation systems, optimizing efficiency and user experience. With the ever-growing volume of data, machine learning is indispensable for tapping into its potential, fostering innovation and progress across various domains.

Machine learning plays a crucial role in various industries by automating processes, optimizing operations, and enabling data-driven decision-making. Its ability to analyze large volumes of data and uncover hidden patterns helps businesses gain valuable insights into customer behavior, market trends, and operational efficiency. Machine learning algorithms are also used in healthcare for disease diagnosis and personalized treatment plans, in finance for fraud detection and risk management, and in autonomous vehicles for navigation and safety. Furthermore, machine learning contributes to advancements in fields like natural language processing, computer vision, and robotics, driving innovation and shaping the future of technology.



II .LITERATURE SURVEY

In recent years, researchers have explored various machine learning and deep learning techniques for

agricultural applications. H.S. Gill, G. Murugesan, B.S. Khehra, G.S. Sajja, G. Gupta, and A. Bhatt (2022) proposed a fruit recognition system using deep learning applications, demonstrating the effectiveness of convolutional neural networks (CNNs) in identifying fruits from images. A. Suruliandi, G. Mariammal, and S.P. Raja (2021) developed a crop prediction model based on soil and environmental characteristics, employing feature selection techniques to enhance prediction accuracy. F. Raimundo, A. Glória, and P. Sebastião (2021) focused on weather forecasting for smart agriculture using machine learning techniques, emphasizing the role of predictive analytics in supporting decision-making for agricultural management. E. Khosla, R. Dharavath, and R. Priya (2020) utilized aggregated rainfall-based modular artificial neural networks (ANNs) and support vector regression (SVR) to predict crop yields, highlighting the importance of environmental factors in yield estimation. M. Khan and S. Noor (2019) investigated irrigation runoff volume prediction using machine learning algorithms, showcasing the potential of data-driven approaches in optimizing water resource management in agriculture. These studies collectively demonstrate the diverse applications of machine learning in agriculture, offering insights into crop prediction, weather forecasting, and water resource management for sustainable agricultural practices.

III. SYSTEM REQUIREMENTS SPECIFICATIONS

HARDWARE REQUIREMENTS SPECIFICATION

Processor	:	Dual Core 1.6 GHz
RAM	:	8 GB
Hard Disk	:	500 GB

SOFTWARE REQUIREMENTS SPECIFICATION

Operating System	:Windows 10 or above
Programming Language	:Python
Tools	:Jupyter Notebook/Google Coolab

TECHNOLOGY DESCRIPTION

Agriculture yield prediction has seen significant advancements through the integration of various advanced technologies, including Decision Tree Regressor, Random Forest Regressor, XG Boost Regressor, Support Vector Regressor, and K Nearest Neighbors Regressor. Decision Tree Regressor and Random Forest Regressor, both belonging to the ensemble learning family, leverage historical agricultural data to estimate future yields based on observed patterns, effectively capturing complex relationships between various agricultural factors. XG Boost Regressor, a scalable and accurate gradient boosting framework, offers enhanced predictive capabilities by combining the strengths of multiple weak learners into a strong predictive model, making it particularly suitable for handling large and high-dimensional agricultural datasets. Similarly, Support Vector Regressor and K Nearest Neighbors Regressor provide alternative approaches for predicting agricultural yields, each with its own strengths in handling different types of data and capturing diverse patterns. Among these techniques, XG Boost Regressor and Random Forest Regressor emerge as the most prominent in terms of accuracy for test data, offering reliable predictions for optimizing agricultural productivity and informing decision-making in the dynamic agricultural sector.

IV. SYSTEM ANALYSIS

EXISTING SYSTEM AND ITS DISADVANTAGES

1. Decision Tree:

- Identification: Determine the optimal tree depth and other hyperparameters through cross-validation or grid search based on model performance metrics.
- Estimation: Use recursive partitioning to split the data into subsets based on feature values, minimizing the variance of the target variable within each subset.
- Model Fitting: Fit the decision tree regressor to the data, resulting in a hierarchical structure of decision rules for predicting the target variable.
- Diagnostic Checking: Evaluate the model's performance using metrics such as mean squared error or R-squared, and visualize the tree structure for interpretability.

Disadvantages:

- Overfitting: Decision trees are prone to overfitting, especially with deep trees or noisy data, which can lead to poor generalization performance on unseen data.

- Lack of Smoothness: Decision tree predictions can be discontinuous and sensitive to small changes in the training data, making them less suitable for tasks requiring smooth output.
- Instability: Decision trees are sensitive to variations in the training data, resulting in different tree structures for small changes in the training set, which can affect model interpretability and reproducibility.

2.Random Forest:

- Identification: Determine the number of trees in the forest and other hyperparameters through cross-validation or grid search to optimize model performance.
- Estimation: Train multiple decision tree regressors on bootstrap samples of the data, using random subsets of features at each split to reduce correlation between trees.
- Model Fitting: Aggregate the predictions of individual trees to obtain the final prediction, reducing variance and improving generalization performance.
- Diagnostic Checking: Assess the model's performance using evaluation metrics such as mean squared error or R-squared on a validation set.

V. SYSTEM DESIGN

5.1 ARCHITECTURAL DESIGN

Designing an architectural framework for agriculture yields prediction using machine learning regression models like Decision Tree, Random Forest, XG Boost, Support Vector Machine and K Nearest Regressor involves several steps. Below is a high-level architectural design for such a system:

Data Collection and Preprocessing:

- Collect historical agriculture yields data from reliable sources such as Government websites, Kaggle or Surveys.
- Preprocess the data by handling missing values, outliers, and formatting it into a machine readable format.

Feature Engineering:

- Extract relevant features from the raw data that could potentially influence crop yield, such amount of rainfall, pesticides used, fertilizers used, etc.
- Transform the features to make them suitable for model training, such as scaling or normalization.

Model Selection and Training:

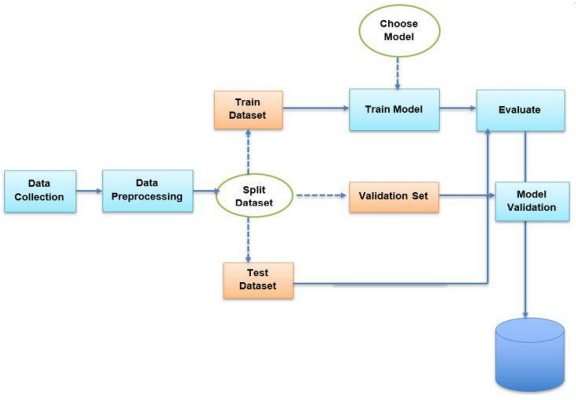
- Implement multiple machine learning models like Decision Tree, Random Forest, XG Boost, Support Vector Machine and K Nearest Regressor.
- Split the data into training, validation, and test sets.
- Train each model using the training data and validate their performance using the validation set.
- Optimize hyperparameters for each model using techniques like grid search or Bayesian optimization.

Model Evaluation:

- Evaluate the performance of each model using appropriate evaluation metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), or Root Mean Squared Error (RMSE) and r2_score.
- Compare the performance of different models to identify the best-performing one.

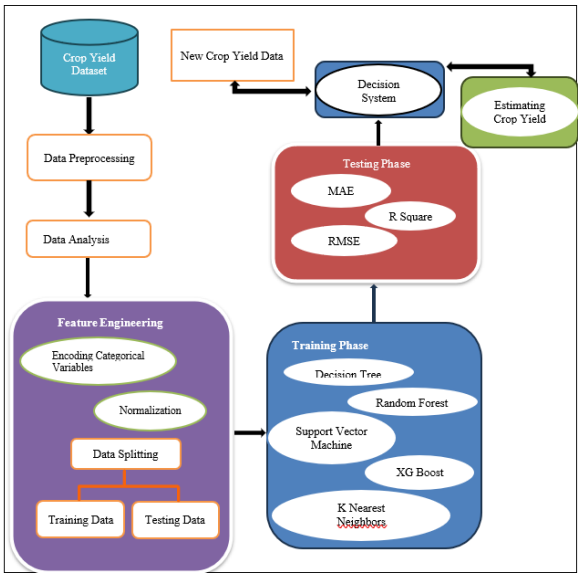
Monitoring and Maintenance:

- Implement monitoring mechanisms to track model performance and drift over time.
- Regularly retrain the model with updated data to ensure its accuracy and relevance.
- Continuously monitor the model's predictions and fine-tune it as needed.



ACTIVITY DIAGRAM

- The activity diagram for agriculture yield prediction using XG Boost begins with data collection, where historical agriculture yields data are retrieved and preprocessed.



VI.SYSTEM TESTING

TESTING

Testing is the systematic process of evaluating software to ensure it meets specified requirements and functions as intended. It involves executing software components under controlled conditions to identify defects and ensure quality.

There are mainly three types of testing:

- Unit Testing
- Integration Testing
- System Testing

Unit Testing:

Unit testing is a foundational practice in software development, focusing on validating individual units or components of code in isolation. It ensures that each unit functions correctly and produces the expected output, helping catch bugs early in the development process. By promoting modular design and providing rapid feedback to developers, unit testing enhances software quality and maintainability.

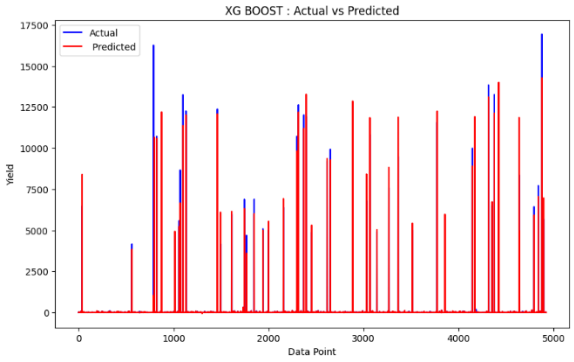
Integration Testing:

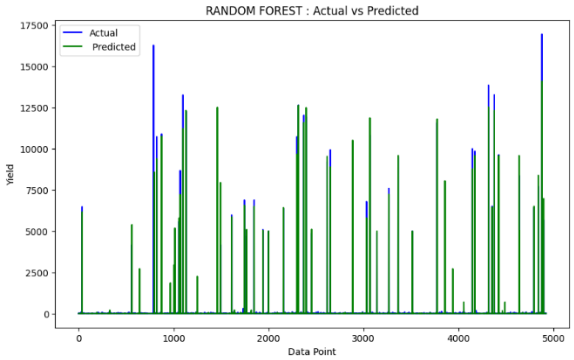
Integration testing evaluates the interaction and collaboration between different modules or units within the software system. It verifies integration points and interfaces, ensuring smooth data flow between components and detecting defects arising from interactions. By facilitating the seamless integration of software modules, integration testing contributes to the development of robust and interoperable systems.

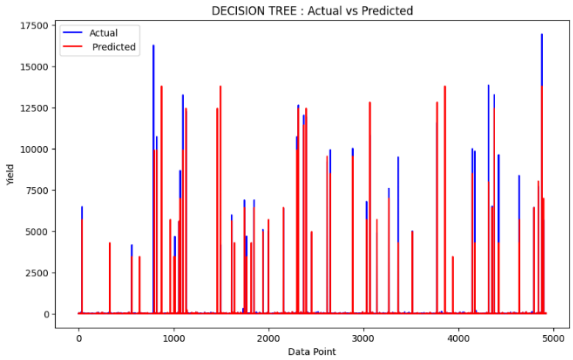
System Testing:

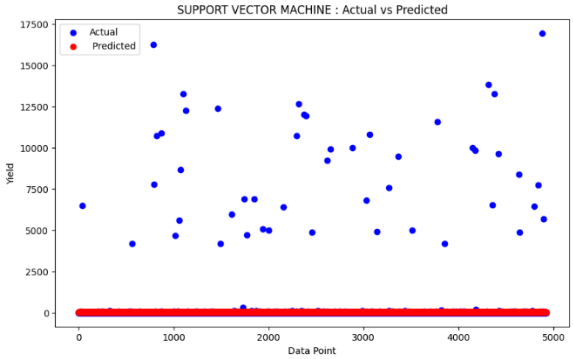
System testing evaluates the entire software system in a holistic manner, validating its compliance with specified requirements and assessing its behavior and performance. It encompasses various types of testing, including functional, performance, and usability testing, ensuring that the system meets user expectations and performs reliably under different conditions. By confirming the system's overall functionality and readiness for deployment, system testing plays a crucial role in delivering high-quality software products.

VII. RESULTS AND SCREENSHOTS









Result Comparison

Model	Train Score	Test Score	MAE	RMSE	R^2 Score
XG Boost	0.999985	0.917019	14.808057	260.889933	0.917019
Decision Tree	0.957025	0.846197	25.351203	355.182860	0.846197
Random Forest	0.990409	0.911665	17.953744	269.175736	0.911665
K Nearest Neighbor	0.937268	0.879917	94.004002	908.960541	0.013878
Support Vector Machine	-741426.553037	-808851.767155	85.985706	899.361867	0.013878

VIII CONCLUSION

In this article, we propose a lightweight CNN-based model to classify the four major cardiac abnormalities using public ECG images dataset of cardiac patients. According to the results of the experiments, the proposed CNN model achieves remarkable results in cardiovascular disease classification and can also be used as a feature extraction tool for the traditional machine learning classifiers. Thus, the proposed CNN model can be used as an assistance tool for clinicians in the medical field to detect cardiac diseases from ECG images and bypass the manual process that leads to inaccurate and time-consuming results

IX FUTURE SCOPE

The future prospects for utilizing XG-Boost in agriculture yield prediction are highly promising, presenting numerous avenues for further advancements and applications. One potential direction involves exploring innovative features or feature engineering methodologies tailored to agricultural datasets, which could enhance the model's ability to capture relevant factors influencing yield variations. Additionally, there's ample opportunity to delve into real-time prediction capabilities of XG-Boost models for agriculture, paving the way for integration into precision agriculture systems or farm management software. Research efforts could focus on developing adaptive models that can dynamically adjust to evolving environmental conditions and crop-specific requirements, thereby offering valuable insights for optimizing agricultural practices and enhancing productivity. Furthermore, the integration of XG-Boost models with remote sensing technologies and IoT devices holds great potential for creating data-driven decision support systems in agriculture, facilitating timely interventions and resource allocation for sustainable crop management. Overall, the future scope for leveraging XG-Boost in agriculture yield prediction is vast, with the potential to revolutionize farming practices and contribute to global food security initiatives.

REFERENCE

[1] 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.

[2] Suruliandi, G. Mariammal, and S. P. Raja, ‘‘Crop prediction based on soil and environmental characteristics using feature selection techniques,’’ Math. Comput. Model. Dyn. Syst., vol. 27, no. 1, pp. 117–140, Jan. 2021.

[3] F. Raimundo, A. Glória, and P. Sebastião, ‘‘Prediction of weather forecast for smart agriculture supported by machine learning,’’ in Proc. IEEE World AI IoT Congr. (AIIoT), May 2021, pp. 160–164.

[4] E. Khosla, R. Dharavath, and R. Priya, ‘‘Crop yield prediction using aggregated rainfall-based modular artificial neural networks and support vector regression,’’ Environ., Develop. Sustainability, vol. 22, no. 6, pp. 5687–5708, Aug. 2020.

[5] 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.

[6] S. Khaki and L. Wang, ‘‘Crop yield prediction using deep neural networks,’’ Frontiers Plant Sci., vol. 10, p. 621, May 2019.

[7] 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.

[8] P. S. Maya Gopal and R. Bhargavi, ‘‘Optimum feature subset for optimizing crop yield prediction using filter and wrapper approaches,’’ Appl. Eng. Agricult., vol. 35, no. 1, pp. 9–14, 2019.

[9] 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.

[10] N. Gandhi, L. J. Armstrong, O. Petkar, and A. K. Tripathy, ‘‘Rice crop yield prediction in India using support vector machines,’’ in Proc. 13th Int. Joint Conf. Comput. Sci. Softw. Eng. (JCSSE), Jul. 2016, pp. 1–5.