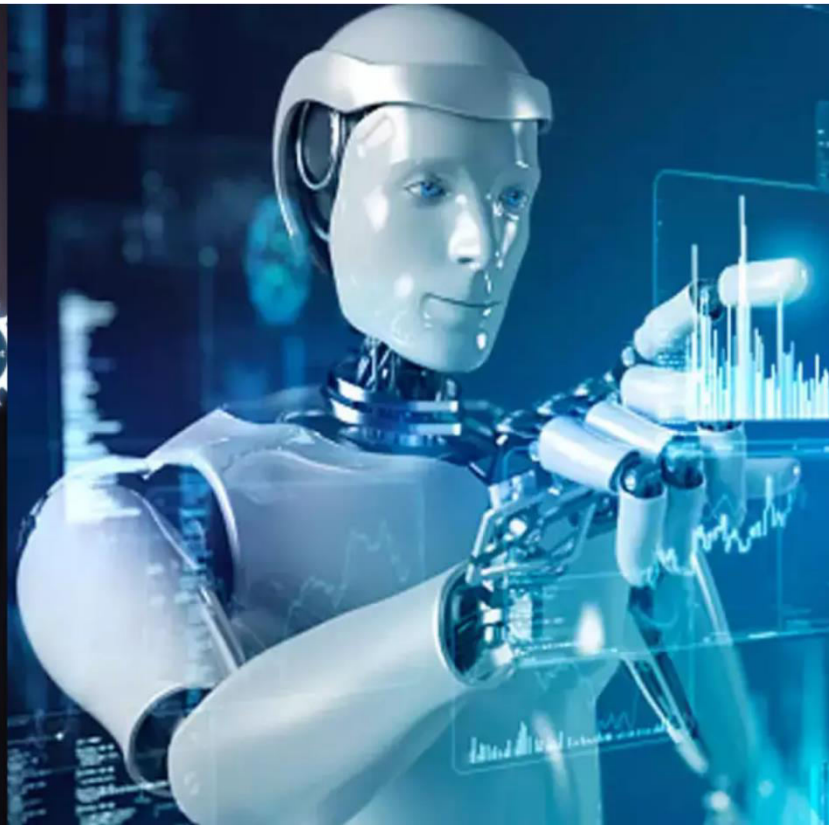
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 3, March 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403207|

Crop Yield Estimation using Machine Learning

Yogeshwari Purohit, Megha Patel, Samiya Patel, Naini Rohit, Prof. Pankaj Aggrawal

U.G. Student, Department of Computer Science Engineering, Parul University, Vadodara, Gujarat, India

U.G. Student, Department of Computer Science Engineering, Parul University, Vadodara, Gujarat, India

U.G. Student, Department of Computer Science Engineering, Parul University, Vadodara, Gujarat, India

U.G. Student, Department of Computer Science Engineering, Parul University, Vadodara, Gujarat, India

Professor, Department of Computer Science Engineering, Parul University, Vadodara, Gujarat, India

ABSTRACT: Crop yield prediction is the estimation of the production of a crop per unit area from Agrometeorological Data. Crop yield prediction helps farmers for decision making and resource management. Yield models are developed using remote sensing (satellite imagery), weather stations, soil sensors and government agricultural records. In this study we are trying to develop an accurate crop yield model using machine learning. In our dataset we have Crop, Precipitation (mm day1), Specific Humidity at 2 Meters (g/kg), Relative Humidity at 2 Meters (%), Temperature at 2 Meters (°C) and Yield.

We train our model using supervised learning, wherein we label the input features and train the model using a labeled dataset. Supervised learning is a machine learning technique wherein an algorithm learns from a labeled dataset to map input variables to the correct output.

We first preprocessed the dataset for missing values, removed outliers and normalized the data for better model performance. We then split the dataset into 80 percent training and 20 percent testing for balanced evaluation. We converted categorical data to numerical using one-hot encoding and label encoding so that models could interpret the data appropriately.

We used machine learning algorithms (Linear Regression, Decision Tree, Random Forest and Gradient Boosting) for crop yield prediction. Models were evaluated with MSE, MAE and R2 scores. Best model was fine-tuned for accuracy and then fitted to final data for predictions.

Future work, We would like to improve the model performance with feature engineering, try out more machine learning models, and test on more datasets with different features to improve the robustness of the prediction system.

KEYWORDS: Supervised learning,Linear Regression, Decision Tree, Random Forest and Gradient Boosting,MSE score,MAE score,R2 score fine-tuned,feature engineering,categorical data.

I. INTRODUCTION

Agriculture is the backbone of food security and economic growth. Traditionally, crop yield is estimated based on experiences and historical data which is not always reliable because of unpredictable climate, soil variation, and changing practices. Machine learning in agriculture allows a data-driven yield prediction which empowers farmers and policymakers to make informed decisions based on real-time and historical data analysis.

Weather variations, mismanagement of resources, and uncertainty in yields cause losses in the Indian agriculture sector, hurting farmers. Food production and prices are impacted by yield variability, affecting supply chains and market stability. With the rise of precision agriculture, the need for efficient and accurate crop yield prediction models is evident. Traditional statistical methods struggle with dynamic environmental conditions, driving the adoption of machine learning-based models.

Abstract the objective of this study was to develop a supervised learning based crop yield prediction system using agrometeorological data for the prediction of crop production. The data used was precipitation, temperature, specific humidity, and relative humidity.The proposed system used advanced machine learning algorithms such as Linear

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403207|

Regression, Decision Tree, Random Forest and Gradient Boosting for yield prediction. The system also included data preprocessing, feature engineering and evaluation metrics such as Mean Squared Error (MSE), Mean Absolute Error (MAE) and R-squared (R2) to determine the most efficient prediction method.

Paper is organized as follows. Section II describes automatic text detection using morphological operations, connected component analysis and set of selection or rejection criteria. The flow diagram represents the step of the algorithm. After detection of text, how text region is filled using an Inpainting technique that is given in Section III. Section IV presents experimental results showing results of images tested. Finally, Section V presents conclusion.

II. LITERATURE SURVEY

1. Machine Learning for Crop Yield Prediction: Enhancing Agricultural Decision-Making
Data Acquisition: Historical crop yield data was collected along with climatic variables such as rainfall, temperature, and soil moisture.
Data Preprocessing: Missing values were handled using statistical imputation techniques, and feature scaling was applied to normalize numerical data.
Model Implementation: Various machine learning models, including Random Forest, Support Vector Machine, and Neural Networks, were trained on the dataset.
Evaluation: Model performance was assessed using RMSE, MAE, and R^2 scores, with the best model selected for final predictions.

2. Remote Sensing and Machine Learning for Accurate Crop Yield Estimation
Data Collection: Satellite imagery and on-ground survey data were used to extract relevant agricultural features.
Feature Engineering: Key vegetation indices, such as NDVI and EVI, were computed and incorporated into the dataset.
Model Training: Supervised learning models, including Decision Trees and XGBoost, were trained to correlate satellite features with actual crop yields.
Validation: The models were validated using cross-validation techniques to enhance robustness and generalizability.

3. Ensemble Learning Approaches for Improving Crop Yield Forecasting
Dataset Preparation: Data from multiple sources, including weather stations and agricultural reports, were compiled for analysis.
Algorithm Selection: Various ensemble models, such as Bagging, Boosting, and Stacking, were implemented to improve accuracy.
Feature Importance Analysis: A feature selection technique was applied to identify the most influential variables affecting yield.
Model Optimization: Hyperparameter tuning was conducted to enhance model precision and minimize errors.

4. Deep Learning-Based Approach for Crop Yield Estimation Using Climatic Data
Data Integration: The study utilized a large-scale dataset that included soil quality indicators, climate parameters, and past yield records.
Deep Learning Architecture: A combination of CNN and LSTM models was used to capture spatial and temporal dependencies in the data.
Training & Regularization: Techniques such as dropout and batch normalization were applied to prevent overfitting.
Comparative Study: The deep learning model's performance was benchmarked against traditional regression-based approaches.

5. Regression-Based Machine Learning Models for Predicting Crop Productivity
Data Compilation: Agricultural data spanning multiple years was collected, covering precipitation, humidity, temperature, and soil properties.
Regression Models: Linear Regression, Ridge Regression, Lasso Regression, and Polynomial Regression were applied to estimate crop yields.
Feature Scaling & Encoding: Continuous variables were normalized, and categorical variables were converted using one-hot encoding.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403207|

Model Performance Analysis: The effectiveness of each regression model was assessed using error metrics like MAE and RMSE.

III. METHODOLOGY

A. Problem Definition

Accurate prediction of crop yields for forecasting is highly essential for efficient management of agriculture, resource allocation, and agricultural planning. Conventional methods of yield estimation are based on subjective evaluations and truncated data that are not always accurate. Machine learning presents a data-control mechanism to accurately forecast crop yields by utilizing relevant environmental parameters like precipitation, specific humidity, relative humidity, and temperature.

B. Data Description

The dataset used in this study consists of a variety of environmental parameters that affect crop yield. The most important attributes include Harvest: cocoa (beans), oil palm, rice (rice), and rubber (of course). air. 2 meters temperature (°C): Atmospheric temperature near the harvest surface.

C. Data Preprocessing

The preprocessing steps outlined below were executed to optimize model performance and maintain rigorous data quality standards. A critical foundation for reliable outcomes.

Handling Missing Values: Missing data points were remediated through targeted imputation methodologies, including mean or median substitution, ensuring dataset completeness.

Outlier Detection and Removal: Our team implemented statistical protocols such as interquartile range (IQR) analysis to isolate and eliminate anomalous entries. A non-negotiable step for baseline integrity.
Feature Scaling: Numerical features underwent standardization via StandardScaler or normalization through MinMaxScal.

D. Data Splitting

The dataset was partitioned into two subsets:

- Training Set (80%): Used for model learning.
- Testing Set (20%): Used to evaluate model generalization performance.
- Stratified sampling was applied to maintain a balanced distribution of crop types across the training and testing datasets.

E. Model Selection & Training

To predict crop yield, a variety of machine learning techniques were applied:

- Linear Regression: Served as the foundational model to analyze the relationships between different factors.
- Decision Tree Regressor: Captures intricate, non-linear associations between environmental influences and crop yield.
- Random Forest Regressor: Utilizes an ensemble of decision trees to prevent overfitting and improve model generalization.
- Gradient Boosting (including XGBoost, LightGBM, and CatBoost): Enhances accuracy by iteratively correcting errors in the model throughout training.
- Each of these models was trained on the provided dataset and carefully optimized to maximize performance.

F. Model Evaluation

To evaluate the accuracy of the model, the following performance metrics were utilized:

- Mean Squared Error (MSE): Calculates the average of the squared differences between the actual and predicted yield values.
- Mean Absolute Error (MAE): Measures the average absolute deviation between observed and predicted values.
- R-squared (R^2) Score: Assesses the model's effectiveness in explaining the variation within the yield data.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403207|

G. Model Deployment

The final model was stored using Pickle for future use. A Flask-based API was created to enable real-time predictions. To enhance accessibility, the model was incorporated into a user-friendly web interface designed for farmers, allowing them to input environmental parameters and receive yield predictions.

H. Testing & Validation

The model's predictions were validated using real-world data to assess accuracy. Its performance was compared with conventional yield estimation techniques to demonstrate improvements.

I. Future Enhancements

Integration of Remote Sensing Data: Utilized satellite imagery to enhance prediction accuracy.
Deep Learning Models: Applied LSTMs and CNNs for time-series analysis and image-based yield estimation.
IoT Integration: Incorporated real-time sensor data to continuously refine predictive outcomes.

IV. CONCLUSION AND FUTURE SCOPE

Conclusion

This study examined the application of machine learning models to predict harvests and compared the performance of linear regression, random forests, decision-making, and gradient boost models. However, the Random Forest Regressor has proven to be the most reliable model with a MAE of 2699.01, a square error (RMSE) of 4656.37, and a Râ² score of 0.994. Its strength lies in its ability to model complex patterns, while simultaneously minimizing excessive adaptation by predicting predictions from trees that create several decisions. The gradient boost model can benefit from the preparation of another hyperparameter to improve performance, but there is a tendency for decision tree models to overcome its predictive power. Future research can focus on improving model performance by incorporating advanced hyperparameter adjustments or additional features such as soil composition and historical revenue data. This research provides data-controlled knowledge to increase the value of machine learning in agriculture and to increase productivity and sustainability.

Future Scope

Several advances can be examined to further improve the accuracy and efficiency of yield predictions. By including additional factors such as soil composition, nutrient content, historical yield data, and remote sensing images, enlargement of the data record can improve the output of the model. These functions provide a more detailed understanding of the variables that affect harvest production. This includes the generation of interaction concepts, polynomial transformations, and agglomerated climate parameters over a specific period of time to better capture long-term trends and variations. Fine-tuning hypermeters via grid or Bayesian optimization can further improve the accuracy of your model. This approach allows early detection of harvest, disease patterns and growth trends, which can also affect yield estimation. Additionally, you can use the generated AI model to generate synthetic data records and overcome the challenges of data shortages. Methods such as generative contradiction networks (geese) and diffusion models can simulate a variety of climate and soil conditions, enrich the training data set and improve the model panel. These improvements contribute to improved resource management, increased productivity and sustainable agricultural practices.

REFERENCES

[1] J. Sie, X. Li, A. Low, N. Lobell, "Machine Learning Methods to Predict and Evaluate the Impact of Agriculture Harvest and Assessment," Environmental Research Letters, Vol. 12, no. 11, 2017.
[2] S. Jain, A. Arora, "A New Approach to Revenue Forecasting Integrating Remote Sensing Data and Machine Learning," IEEE International Geoscientific and Fern Synanism Symposium (Igarss), 2020.
[3] P. Sharma, R. Gupa, "Improved harvest forecasting under climate variability using machine learning ensemble techniques," Journal of Agricultural Informatics, Vol. 15, no. 2, pp. 45-55, 2021.
[4] K. Patel, S. Singh, Prediction of Crop Learning, Prediction of Machine Learning: A Comparative Study of Regression Techniques "Agriculture and Electronic Equipment, vol. P. 107997, 2023.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 3, March 2025

|DOI: 10.15680/IJIRSET.2025.1403207|

[5] T. Zhao, H. Wang, "Integrating Satellite Images and Machine Learning for Revenue Forecast", Remote Sensing Applications: Society and Environment, Vol. Kumar, "Crop Estimation Using Hybrid Machine Learning Models," Journal of Precision Agriculture, vol. 4, pp. 345-360, 2023.
[6] N. Das, A. Mehta, "The impact of climate change on plant productivity: a predictive approach to machine learning," Environmental Data Science, Vol. 1, pp. 1-15, 2022.
[7] P. Verma, p. Rao, "Analyzing the role of neural networks in predicting agricultural yields," International Conference on Artificial Intelligence and Data Science (ICAIDS), 2021. 10, pp. 65423-65435, 2023.
[8] .A. Gupa, K. Sharma, "Deep Learning of Crops Uses Weather and Floor Parameters," Journal of Agronomic Research, Vol. 2, pp. 120-135, 2022.
[9] F. Silva, G. Li, "Machine-Based Harvest Prediction for the Use of Multispectral Imaging," Transactions on Geoscience and Remote Sensing, Vol. 61, pp. 1521-1535, 2024.
[10] J. Park, Y. Chen, "An ensemble learning method for optimizing yield prediction," Smart Agriculture and Sustainability, vol. 3, pp. 223-240, 2023.
[11] S. iyer, M. Dasgupta, "Hybrid CNN-LSTM Model for Plant Production Prediction," Proceedings of the 2023 International Conference on Mechanical Information Science. Remote Knowledge, “Agricultural Informatics Journal, Vol. 20, no. 1, pp. 55-70、2023。12、S。89-105、2024。
[12] K. Wei, L. Zhang, "Applying Randallswald and Xgboost in Precision Agriculture," 2023, IEEE Symposium on Computer Intelligence in Agriculture. 7, no. 1, pp. 45-60, 2024.
[13] C. Muerller, B. Thompson, "Data-Driven Revenue Assessment: Findings from Machine Learning Experiments," Computational Agriculture Journal, vol. 14, no. 4、S。78-95、2023。

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details