ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may118

Stacked Ensemble Machine Learning Techniques based Predictive Modelling of Crop Yields

Bhavani R[1]; Abinaya P[2]; Maithili R[3]; Reshma Masutha A[4]; Sivaranjani K[5]

[1,2,3,4,5] Department of Computer Science and Engineering,
Government College of Engineering Srirangam, Tamil Nadu, India

Publication Date: 2025/05/19

Abstract: Agriculture is crucial for food security and economic stability in India, where a most of the population relies on farming. Accurate crop yield prediction is essential for informed planning, efficient resource allocation, and maximizing agricultural productivity. This paper proposes a novel approach for predicting crop yield using an ensemble learning model. The proposed system utilizes historical agricultural data—including district, crop_year, season, area, and production for Tamil Nadu. The stacking ensemble model proposed in this paper integrates K-Nearest Neighbors Regressor and Multiple Linear Regressor as base learner and Decision Tree Regressor as the meta-learner. This ensemble approach enhances prediction performance by leveraging the strengths of each individual model while minimizing their weaknesses. Experimental results, evaluated using R-squared (R^2) metrics, show that the Stacked Ensemble Regressor outperforms standalone models in terms of accuracy. This system offers strong decision-making support for farmers and agricultural stakeholders, helping them make informed, data-driven choices that enhance sustainability and efficiency in farming.

Keywords: *Crop Yield Prediction, Ensemble Learning, Agriculture Data, K-Nearest Neighbors, Decision Tree Regressor, R-Squared Metrics.*

How to Cite: Bhavani R; Abinaya P; Maithili R; Reshma Masutha A; Sivaranjani K (2025) Stacked Ensemble Machine Learning Techniques based Predictive Modelling of Crop Yields. *International Journal of Innovative Science and Research Technology*, 10(5), 580-584. https://doi.org/10.38124/IJISRT/25may118

I. INTRODUCTION

Agriculture is a cornerstone of India's economy, providing a livelihood for a substantial share of the population. Predicting crop yields accurately can help farmers make informed decisions about which crops to grow, when to sow, and how to manage resources efficiently. Yield prediction aids in improving food security by helping government bodies and agricultural organizations forecast production levels and plan imports or exports accordingly.

Traditional yield estimation methods often struggle with the dynamic and nonlinear nature of agricultural systems, which are influenced by multiple interrelated factors such as climate conditions, soil quality, crop type, and cultivation practices. In recent years, Machine Learning (ML) techniques have emerged as powerful tools for analyzing complex datasets and making accurate predictions. Ensemble learning models, in particular, have shown improved performance by combining the strengths of multiple algorithms to reduce bias and variance.

This study proposes an ensemble-based machine learning model for crop yield prediction using historical agricultural data from Tamil Nadu. The dataset includes key attributes such as crop_year, season, area, district name, and crop_type, all of which significantly influence crop production. Models such as Decision Tree Regressor, Multiple Linear Regressor and K-Nearest Neighbor Regressor are trained and evaluated, and a Stacking Ensemble Regressor is used as the final ensemble model to improve overall prediction accuracy and reliability.

To make the system easier to use, a simple web interface has been created. Users like farmers or agricultural officers can enter details such as the year, season, crop name, district, and area. The system then gives an instant prediction of the crop yield. Before making the prediction, the input data is cleaned and converted into a suitable format using preprocessing steps like filling missing values, scaling numbers, and converting text into numerical values. This user-friendly system helps people make better farming decisions and supports effective planning in agriculture.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may118

II. RELATED WORKS

Recent research in crop yield prediction has largely focused on statistical and regression-based modeling techniques. These approaches are frequently applied due to their mathematical simplicity and interpretability. Studies have explored models such as Multiple Linear Regression (MLR), Polynomial Regression, and Time Series Forecasting as foundational tools for estimating seasonal crop productivity using environmental and agronomic variables such as temperature, rainfall, soil type, and crop characteristics [1].

Researchers employing Multiple Linear Regression have attempted to establish linear relationships between multiple independent variables and crop yield. This model has been widely used in agricultural prediction tasks; however, it often exhibits limitations when applied to systems with complex, non-linear interactions [2]. MLR's predictive accuracy tends to decline in heterogeneous or unstable environments, where linear assumptions no longer hold true.

To address non-linearity, some studies have proposed the use of Polynomial Regression, introducing higher-order terms to model more complex patterns. While polynomial models can improve fit in certain scenarios, research has shown that they are particularly susceptible to over fitting when not carefully tuned. This results in poor generalization, especially in cases involving noisy or high-dimensional agricultural data [3].

Logistic Regression applied in binary classification problems, such as predicting whether the yield will exceed a certain threshold. However, it is limited in its applicability for continuous yield prediction [4].Time Series Forecasting methods like ARIMA have been applied to predict yield trends over time. While effective for temporal analysis, these models are sensitive to missing values and cannot easily incorporate multidimensional agro-environmental data [5].

Other research efforts have implemented Time Series Forecasting techniques, particularly ARIMA models, to capture seasonal and temporal patterns in yield data. These models are designed to predict future yields based on past trends and temporal dependencies. However, such models are constrained by their reliance on stationary assumptions and their inability to integrate multiple influencing factors simultaneously [6].

Limited Predictive accuracy these models often underperform when applied to large, heterogeneous datasets. Their assumptions of linearity or fixed relationships restrict their ability to model the complex interactions that exist in agricultural systems [7]. High Sensitivity to Input Variable The accuracy of predictions is heavily influenced by the quality and consistency of input data. In agriculture, datasets are often incomplete, noisy, or inconsistently recorded, which poses significant challenges to traditional methods [8]

Poor Generalization models developed using specific regional or crop data often fail to generalize to other regions or crop types.While higher-order regression models may attempt to capture non-linear patterns, they are prone to overfitting, particularly when applied to small or noisy datasets. Conversely, simpler models such as linear regression tend to underfit the data, failing to capture important variable interactions [9].By observing the works the literature, this paper aims to improve the accuracy of crop yield prediction by proposing an ensemble machine learning model.

III. MATERIALS AND METHODS

The system architecture of the proposed crop yield system is given in Figure 1. The main objective of this system is to predict crop yield using stacked ensemble machine learning models using agricultural data. It provides accurate, data-driven insights to help farmers and officials make appropriate decisions. The system follows a modular design, starting with data collection, data preprocessing, and performing model training and evaluation. Next, the trained model is deployed as a web application using HTML, CSS, JavaScript, and Flask to allow real-time user interaction and prediction.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may118

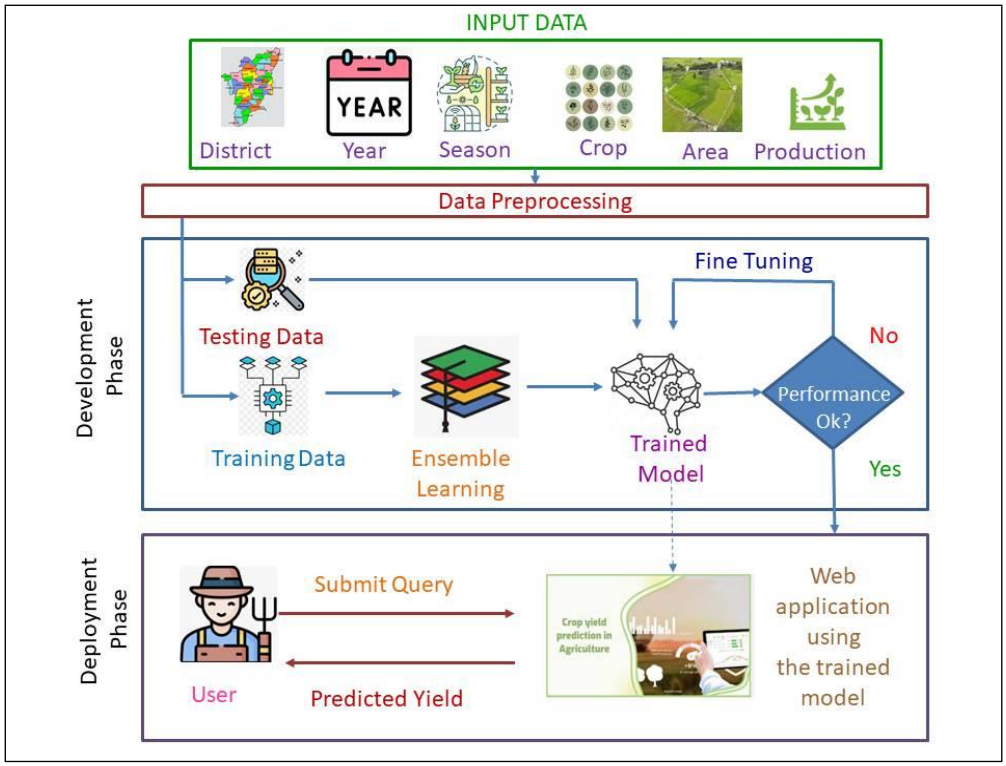

Fig 1.System Architecture for Proposed System

A. Data Collection

The dataset for this project was collected from Kaggle and focuses on agricultural data specific to Tamil Nadu. It includes important features such as District Name, Crop Year, Season, Crop, Area, and Production. This data is imported using Python's pandas library and is initially loaded in CSV format. The dataset serves as the foundation for training and testing machine learning models to predict crop yield, ensuring that the input reflects real-world farming conditions and seasonal variations.

B. Exploratory Data Analysis

Once the data is loaded, exploratory data analysis is performed to understand the characteristics of the features descrbing the data. This step includes checking for missing or duplicated values, understanding the distribution of numeric and categorical variables, and identifying patterns or correlations that may exist between features such as crop type, area, and yield. Visualizations such as correlation matrices and scatter plots are used to uncover trends and relationships that guide feature selection and preprocessing steps.

C. Data Preprocessing

Data preprocessing is an important step that is used to prepare the dataset for modeling. Initially, missing and irrelevant entries are removed to enhance the quality of the data. The column 'State_Name' is excluded as it is not required for prediction. Categorical variables such as District Name, Season, and Crop are transformed using one-hot encoding, while numerical attributes like Crop Year and Area are standardized using a scaling method to ensure consistency across features. These transformations are applied using a ColumnTransformer, resulting in a well-structured dataset suitable for model input.

D. Model Training

Three supervised regression models are employed for training: K-Nearest Neighbor Regressor (KNNR), Decision Tree Regressor (DTR), and Multiple Linear Regression (MLR). The pre-processed data is split in the ratio of 80:20 for training and testing respectively. Each regressor model is trained independently with the input features and the target variable, which is Production. The training process tries to fit the machine learning model to the training data and optimizing it to capture patterns in historical crop yield behavior across different conditions. Next, a stacked ensemble regressor (SER) is built and trained using KNNR and MLR as base learner and DTR as the meta-learner.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may118

E. Model Testing

After training, the models are evaluated using the reserved testing data. The performance of each model is assessed using the R^2 score, which indicates the proportion of variance in the target variable that the model is able to explain. To visualize the predictive performance, actual yield values are plotted against predicted values in scatter plots. These visual comparisons help assess the accuracy and reliability of each model and highlight the effectiveness of the training process.

F. Web Application

To ensure usability and accessibility, the best-performing model is deployed as a web application using the Flask framework. The trained model is serialized using the pickle library and integrated into the backend of the application. Users can enter input parameters such as district, crop, season, area, and year through a web form. The system processes these inputs and returns the predicted crop yield, which is displayed on the webpage. The front-end is developed using HTML and CSS for structure and styling, and JavaScript is used to enhance interactivity.

G. Evaluation Metrics

To assess the effectiveness of the proposed models, regression evaluation metrics were employed. R^2 score measures how well the model's predictions matched the actual data, indicating the proportion of variance explained. This metrics allowed for comprehensive performance analysis and comparison across all implemented models. The R^2 score (coefficient of determination) measures how well the predicted crop yields match the actual values. It indicates the proportion of variance explained by the model and is given by (1).

$$R^2 = 1 - [\ \sum(y_i - \hat{y}_i)^2] / [\sum(y_i - \bar{y})^2] \quad (1)$$

Where:
y_i : Actual value (real crop yield)
$\hat{y}_i$: Predicted value (crop yield predicted by the model
$\bar{y}$: Mean of actual values (average crop yield)

H. Implementation

The implementation of the crop yield prediction system was carried out using Python 3.10, a widely adopted programming language for machine learning applications. All model development and testing were performed on a system running Windows 10, equipped with an Intel Core i5 processor and 4GB RAM. Popular libraries such as scikit-learn, pandas, numpy, matplotlib, and seaborn were employed for model training, data handling, and visualization. The coding was carried out using Jupyter Notebook and Visual Studio Code, providing a user-friendly and interactive development environment. This setup allowed for efficient execution of preprocessing, model building, training, and evaluation processes.

The system begins with importing the dataset, which includes historical crop statistics from Tamil Nadu. Key attributes such as District_Name, Crop_Year, Season, Crop, Area, and Production form the basis for model learning. Data cleaning is performed by removing null entries, duplicates, and irrelevant columns such as State_Name. The cleaned dataset is then structured with appropriate feature selection for machine learning workflows.

To handle heterogeneous data types, the system applies preprocessing techniques. Categorical features like District_Name, Season, and Crop are encoded using OneHotEncoder, while numerical features such as Area and Crop_Year are normalized using StandardScaler. These transformations are performed using a unified pipeline with ColumnTransformer, ensuring consistency across all preprocessing steps.

The processed data is then split into training and testing subsets using an 80:20 ratio. Three supervised regression models KNNR, MLR and DTR are trained on the dataset. Each model is individually evaluated using R^2 Score, and results are visualized through scatter plots comparing actual and predicted yields. Although each model showed promising performance, the individual variations in predictions motivated the use of an ensemble strategy.

To enhance prediction accuracy and robustness, a Stacking Ensemble Regressor is implemented by combining KNNR and MLR as base learner and DTR as the meta-learner. This ensemble approach aggregates the strengths of all base learners, resulting in improved generalization. The Stacking Ensemble Regressor achieved an R^2 Score of approximately 0.85, outperforming all individual models in prediction accuracy and consistency.

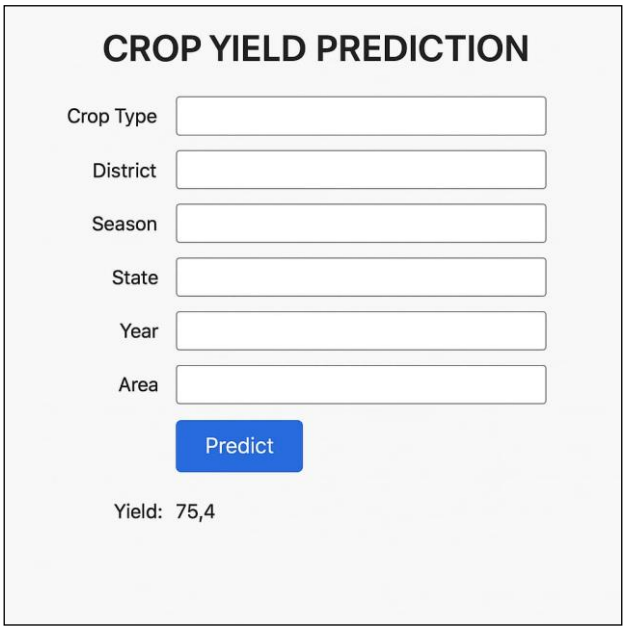

Fig 2.User Interface for Crop Yield Prediction

The final model is serialized using the pickle module and deployed via a Flask framework. The front-end of the web application is built using HTML, CSS, and JavaScript, offering an intuitive form for users to enter input values such as district, crop, season, year, and area. Upon submission, the backend processes the input through the

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/25may118

trained ensemble model and returns the predicted yield in real time. The output is displayed, enabling users such as farmers and agricultural officers to make informed decisions.

IV. RESULTS AND DISCUSSION

The crop yield prediction was conducted using Linear Regression, Decision Tree Regressor, K-Nearest Neighbors Regressor, and a Stacking ensemble Regressor combining the previous three as base learners. The R^2 score of the models during testing is given in Figure 3.

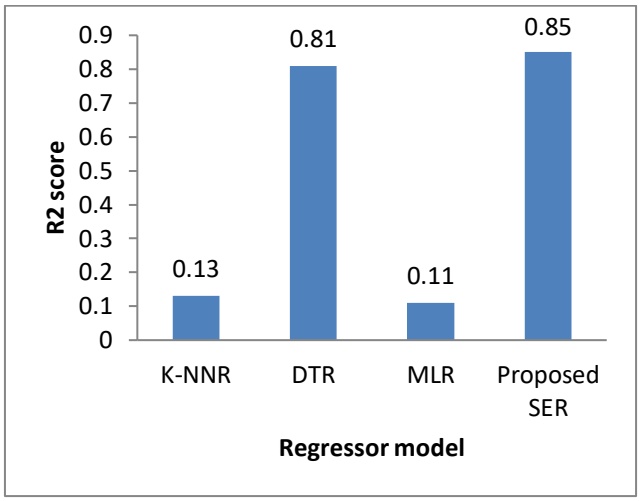

Fig 3. Model Accuracy Comparison

Linear Regression produced decent results by modeling the linear relationship among features, achieving a respectable R^2 score. The Decision Tree Regressor outperformed Linear Regression by capturing non-linear patterns in the dataset, resulting in improved accuracy. K-Nearest Neighbors, being a distance-based algorithm, also performed well after proper scaling, yielding competitive scores. The best results were obtained from the Stacking Ensemble Regressor, which leveraged the strengths of the base models to provide more accurate and stable predictions. It achieved the highest R^2 score and the lowest error values across all metrics, indicating its robustness and effectiveness in handling complex relationships in the data.

V. CONCLUSION AND FUTURE WORK

This paper successfully implemented and compared multiple machine learning models—namely Linear Regression, Decision Tree Regressor, K-Nearest Neighbors Regressor, and a Stacking Regressor—for the purpose of crop yield prediction. The comparative analysis demonstrated that while individual models like Decision Tree and KNN provided good accuracy, the Stacking Regressor significantly improved the overall predictive performance by combining the outputs of these models. The experimental results confirmed the effectiveness of ensemble learning in enhancing model robustness and accuracy.

For future improvements, the system can be expanded by incorporating more granular data such as soil health, humidity, and fertilizer usage, which can potentially improve the prediction quality. Moreover, cross-validation techniques could be applied to enhance model generalization and avoid overfitting. In the long term, developing a user-friendly web or mobile application can make the model accessible to farmers and agricultural planners, thus contributing to informed decision-making in the agricultural sector.

REFERENCES

[1]. Lobell, D.B. and Burke, M.B., 2010. On the use of statistical models to predict crop yield responses to climate change. Agricultural and forest meteorology, 150(11), pp.1443-1452.
[2]. Ramesh, D. and Vardhan, B.V., 2015. Analysis of crop yield prediction using data mining techniques. International Journal of research in engineering and technology, 4(1), pp.47-473.
[3]. Pant, J., Pant, R.P., Singh, M.K., Singh, D.P. and Pant, H., 2021. Analysis of agricultural crop yield prediction using statistical techniques of machine learning. Materials Today: Proceedings, 46, pp.10922-10926.
[4]. Ansarifar, J., Wang, L. and Archontoulis, S.V., 2021. An interaction regression model for crop yield prediction. Scientific reports, 11(1), p.17754.
[5]. Patil, P., Athavale, P., Bothara, M., Tambolkar, S. and More, A., 2023. Crop selection and Yield Prediction using machine learning approach. Current Agriculture Research Journal, 11(3).
[6]. Sadenova, M., Beisekenov, N., Varbanov, P.S. and Pan, T., 2023. Application of machine learning and neural networks to predict the yield of cereals, legumes, oilseeds and forage crops in Kazakhstan. Agriculture, 13(6), p.1195.
[7]. Vijay, N.U., Pandiyan, A.M., Raja, S.P. and Stamenkovic, Z., 2024. Machine learning-based crop yield prediction in south India: performance analysis of various models. Computers 13 (6), 137 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].