IJISEM
INTERNATIONAL JOURNAL OF INNOVATIONS IN SCIENCE ENGINEERING AND MANAGEMENT

ML-Based Insights for Crop Yield Forecasting in Indian Farming

Nitish Kumar[1], Dr. Pankaj Richhariya[2]

[1] *Research Scholar, Department of Computer Science, Bhopal Institute of Technology & Science.*
[2] *Hod, Department of Computer Science, Bhopal Institute of Technology & Science.*

OPEN ACCESS

AG-2023-1001

Volume: 3

Issue: 2

Month: May

Year: 2024

ISSN: 2583-7117

Published: 25.05.2024

Citation:

Nitish kumar1, Dr. Pankaj richhariya2. "ML-Based Insights for Crop Yield Forecasting in Indian Farming." International Journal of Innovations In Science Engineering And Management, vol. 3, no. 2, 2024, pp. 21–29.

Abstract

This study investigates the use of machine learning methods to forecast crop yields in Indian agriculture, using a large dataset that covers the period from 1997 to 2020. Data preparation methods, such as feature selection, one-hot encoding, and transformation, were used to improve the quality and appropriateness of the dataset for modelling. Individual regression models, including Decision Tree, Random Forest, Support Vector Machine (SVR), and K-Nearest Neighbours (KNN), were trained and assessed. These models showed strong performance in accurately forecasting crop yields. Furthermore, the integration of a hybrid ensemble model, which incorporates voting ensemble approaches, resulted in higher predicted accuracy and increased model resilience. Our technique stands out from previous approaches due to its uniqueness and improvements. These include the use of a bigger dataset, a longer time frame, and the use of hybrid ensemble models. The results provide useful insights for stakeholders engaged in agricultural decision-making, enabling informed allocation of resources and implementation of risk management measures to improve productivity and sustainability in Indian agriculture.

Keyword: Crop yield prediction, Machine learning, Agricultural data analysis, Indian agriculture, Regression models, Ensemble learning, Hybrid models, Sustainability.

I. INTRODUCTION

Throughout the globe, agriculture is the backbone of many economies and is essential to both sustainable development and food security. Crop yield prediction has become a critical field of study and application in the quest to maximise agricultural output. Farmers, legislators, and other stakeholders may make well-informed choices about crop management techniques, resource allocation, and risk mitigation measures with the help of accurate crop yield forecast. More so than in any other nation, India's reliance on agriculture makes efficient crop output forecast techniques essential. This research paper utilises the "Agricultural Crop Yield in Indian States Dataset" from Kaggle to illustrate how machine learning methods may be used to forecast crop yields in Indian agriculture. The collection includes agricultural statistics for many crops grown in different Indian states between 1997 and 2020. Crop kinds, crop years, cropping seasons, states, cultivated areas, production amounts, yearly rainfall, fertiliser and pesticide use, and computed yields are some of the important characteristics. The use of machine learning in agriculture has grown significantly in popularity recently, providing exciting new opportunities to increase sustainability and production. Using a variety of datasets and approaches, many research have investigated the use of machine learning algorithms for agricultural production prediction. For example, to predict crop yields based on variables like weather patterns, soil properties, and agricultural practices, researchers have used regression models like Support Vector Machines (SVM), Random Forests, and Neural Networks.[1][2][3]

In addition, agricultural production prediction has been using ensemble learning approaches more and more, which aggregate predictions from numerous models to increase accuracy and resilience.

It has been shown that ensemble techniques like bagging, boosting, and stacking are useful for combining different information sources and reducing the drawbacks of single models [4][5]. However, difficulties still exist in machine learning-based crop yield prediction, such as heterogeneity in the data, interpretability of the model, and scalability to other agricultural situations. The goal of this study is to add to the increasing amount of information on crop yield prediction in Indian agriculture by using a comprehensive strategy that combines ensemble approaches, model construction, and data preparation. This project intends to give useful insights and tools for stakeholders engaged in agricultural decision-making processes by using the deep insights offered by the dataset and the predictive capabilities of machine learning algorithms.

II. RELATED WORK

[6] The research paper emphasizes the role of agriculture in the Indian economy, highlighting the importance of predicting crop yields for food security due to population growth. It discusses traditional methods relying on farmers' experience and proposes using machine learning for more accurate predictions. Factors like weather conditions and pesticide information are crucial, along with historical yield data. By employing machine learning, the study aims to predict yields for four major crops in India—"Maize, Potatoes, Rice (Paddy), and Wheat"—allowing for targeted fertilizer application and risk management in agriculture.

[7] The paper focuses on accurate crop yield prediction to enhance agricultural breeding and monitor crops across varied climates, safeguarding against climatic challenges. Using performance records from "Uniform Soybean Tests (UST) in North America", the study develops a "Long Short Term Memory (LSTM) Recurrent Neural Network model". Leveraging pedigree relatedness measures and weekly weather parameters, the model predicts genotype responses in different environments. Results demonstrate superior performance compared to other machine learning models like "SVR-RBF, LASSO regression, and USDA's data-driven model". Additionally, a temporal attention mechanism is introduced for LSTM models, providing interpretable insights into "crucial time-windows" during the growing season, beneficial for plant breeders.

[8] The study investigates the use of machine learning algorithms and proximate sensing methods to forecast potato tuber production depending on crop and soil characteristics. In all, four machine learning techniques were tested: "support vector regression, elastic net, k-nearest neighbour, and linear regression". Over the course of two growing seasons, data from six fields in Atlantic Canada were gathered. Across datasets, SVR models fared better than others, with RMSE ranging from 4.62 to 6.60 t/ha. k-NN did not perform well. The research highlights that in order to provide precise forecasts, substantial datasets are required. These kinds of discoveries are essential for creating potato management zones that are particular to a given place, which greatly advances global food security projects.

[9] The paper addresses the threat climate change poses to agricultural economies in developing countries like Ghana. Traditional insurance methods are impractical due to various challenges. Area-based index insurance is seen as a solution for smallholder farmers. Predicting crop yield is crucial for pricing premiums. The study evaluates several forecasting methods for "crop yield estimates in Ghana". Comparing methods, ARMA models prove more robust than smoothing techniques, even in the presence of crop yield cycles. The findings suggest ARMA models are preferable for constructing area-based yield insurance, benefiting smallholder farmers and institutions relying on accurate forecasts for capital allocation.

[10] The author of this research paper discusses the importance of precision agriculture, especially in addressing spatial variability within crop fields. They highlight the role of geotechnology, which includes the use of "remotely sensed images", image processing techniques, "Geographic Information System (GIS)" modeling, and "Global Positioning System (GPS)," along with data mining techniques for model development, in efficiently identifying spatial variability. The primary objective of the paper is to investigate the efficacy of key spectral vegetation indices for predicting agricultural crop yields using neural network techniques. The study focuses on irrigated corn crops in the Oakes Irrigation Test Area in North Dakota, USA, over three years (1998, 1999, and 2001), as well as pooled data from these years.

IJISEM
INTERNATIONAL JOURNAL OF INNOVATIONS IN SCIENCE ENGINEERING AND MANAGEMENT

III. METHODOLGY

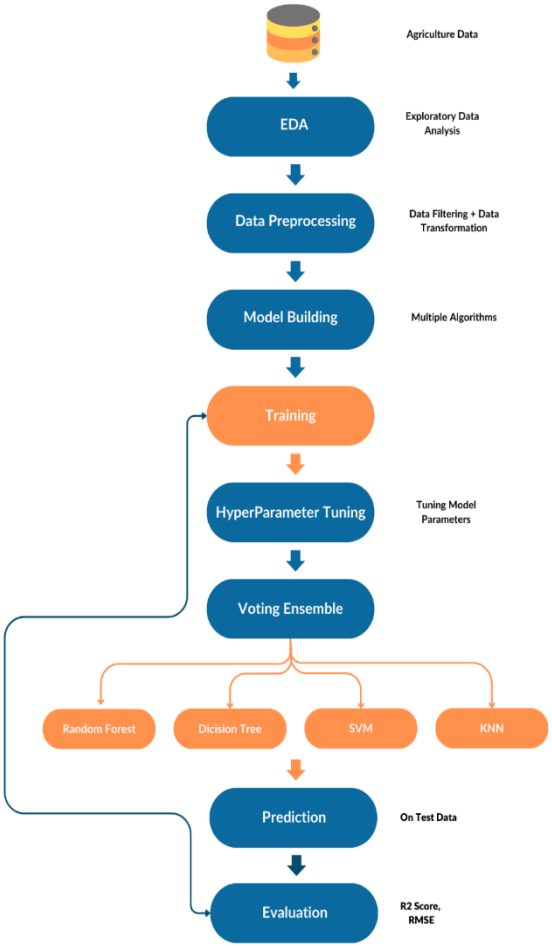

Figure 1 Work Flow

The proposed model has been illustrated in Fig.1

A. Dataset

The "Agricultural Crop Yield in Indian States Dataset," the dataset used in this study, provides the basis for crop yield prediction modelling in Indian agriculture. This extensive dataset, which was obtained via Kaggle, covers a wide range of agricultural characteristics across many Indian states and crops grown between 1997 and 2020.

Key Features:

Crop: This characteristic indicates the kind of crop that is grown; it may be anything from pulses like urad and soyabean to staple cereals like rice and wheat.

Crop_Year: This temporal variable allows the investigation of crop yield patterns over time by indicating the exact year in which the crop was cultivated

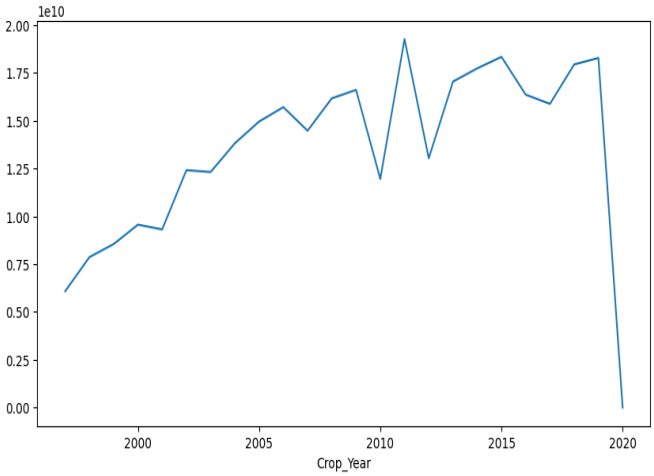

Figure 2 Crop Production Over the Year

Season: This characteristic, which divides the cropping season into Kharif, Rabi, and Whole Year categories, reflects the temporal dynamics of crop production in accordance with seasonal trends.

State: This geographic variable, which designates the Indian state in which the crop was grown, takes into consideration regional differences in soil composition, climate, and agricultural techniques

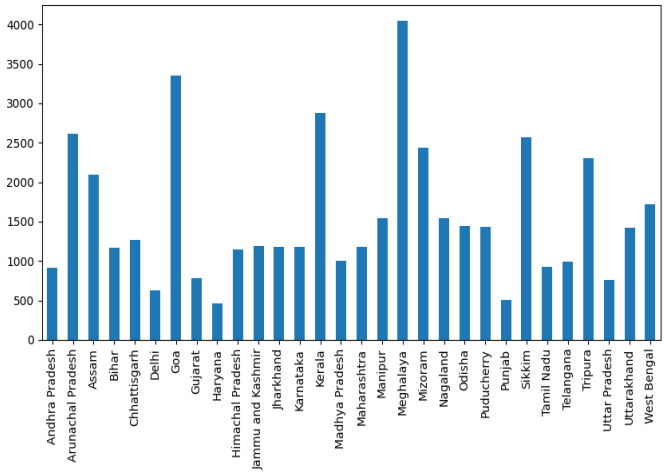

Figure 3 High Crop Production States

Area: Measuring in hectares the total land area under cultivation for a particular crop, this characteristic sheds light on the geographic distribution of agricultural activity.

Production: This statistic, which indicates the amount of crop output in metric tonnes, is the main focus of yield prediction modelling.

Annual_Rainfall: This environmental variable affects crop growth and production. It represents the annual rainfall in millimetres obtained in the crop-growing area.

Fertilizer: This input variable defines the total quantity of fertiliser applied to the crop in kilogrammes. It affects both soil fertility and plant absorption of nutrients.

Pesticide: This variable relates to crop protection and pest control techniques. It indicates the total quantity of pesticide applied for the crop, expressed in kilogrammes.

Yield: This variable represents the efficiency of agricultural yield per unit area and is calculated as the ratio of crop output to the area under cultivation. It is the main focus of predictive modelling.

	Crop	Crop_Year	Season	State	Area (hectares)	Production (metric tons)	Annual_Rainfall (mm)	Fertilizer (in kilograms)	Pesticide (in kilograms)	Yield (production per unit area)
0	Arecanut	1997	Whole Year	Assam	73814.0	56708	2051.4	7024878.38	22882.34	0.796087
1	Arhar/Tur	1997	Kharif	Assam	6637.0	4685	2051.4	631643.29	2057.47	0.710435
2	Castor seed	1997	Kharif	Assam	796.0	22	2051.4	75755.32	246.76	0.238333
3	Coconut	1997	Whole Year	Assam	19656.0	126905000	2051.4	1870661.52	6093.36	5238.051739
4	Cotton(lint)	1997	Kharif	Assam	1739.0	794	2051.4	165500.63	539.09	0.420909

Figure 4 Dataset Sample

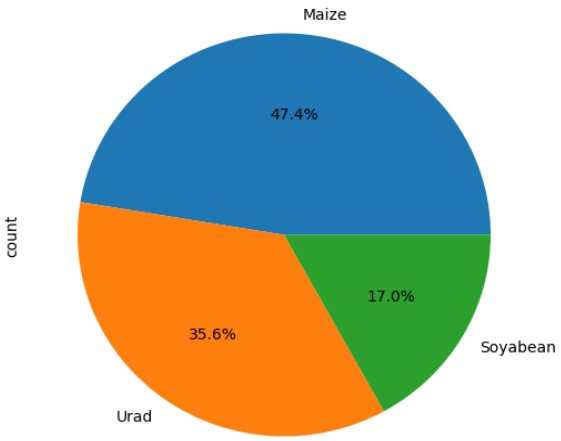

Figure 5 Data Distribution

The machine learning models created in this study were trained and evaluated using this dataset, which made it possible to estimate agricultural production with accuracy and efficiency.

B. Data Pre-Processing

To assure data quality and model readiness, the dataset underwent extensive preprocessing before model training. This required one-hot encoding of categorical variables (Crop, Season, and State), Yeo-Johnson transformation for data normalisation, data scaling for better model generalisation, and filtering the data to concentrate on certain crops (Urad, Maize, and Soyabean). It has been discovered that feature normalisation methods like Standard and Min Max scaling are useful for guaranteeing consistency across variables while preparing agricultural data [11]. Predictive performance may be enhanced by minimising biases in model training by scaling feature values to a common range. To aid in the assessment and validation of the model, the dataset was further divided into subsets for testing and training. The following subsections outline the particular preprocessing actions performed:

1. Data Filtering: The dataset was reduced to only three target crops—urad, maize, and soybean—in order to simplify the study and concentrate on relevant crops for yield prediction. The modelling efforts were focused on crops that were significant and relevant in the context of Indian agriculture thanks to this selective filtering..

2. One-Hot Encoding: To make it easier to include categorical variables like Crop, Season, and State into machine learning models, one-hot encoding was applied to them. In order to represent each category with a unique binary characteristic, categorical variables have to be converted into binary vectors. This method of encoding categorical variables allowed the models to use and understand categorical data well without skewing the outcomes.

3. Yeo-Johnson Transformation: The dataset's production and area numerical features were left-skewed in terms of distribution, which can have an adverse effect on the effectiveness of regression-based models. Using the Yeo-Johnson transformation, this skewness was reduced and a more Gaussian-like distribution was produced. By optimising variance stabilisation and enhancing symmetry in feature distributions, this power transformation technique strengthens the resilience of later modelling attempts

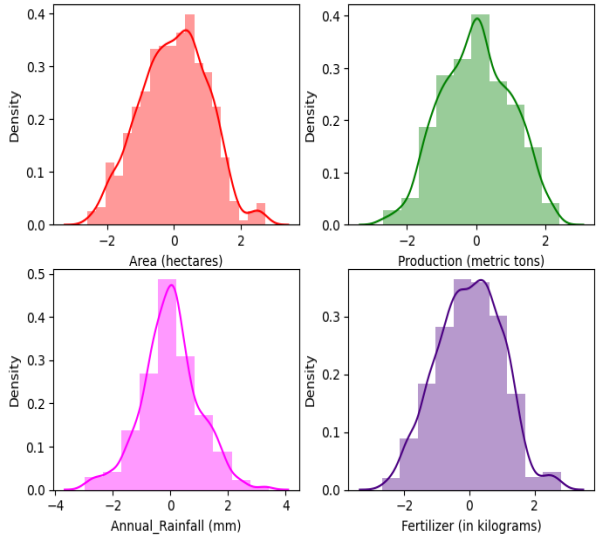

Figure 6 Data Distribution after Yeo-Johnson Transformation

4. Data Scaling: Given the disparate scales and magnitudes of feature values within the dataset, standardization was essential to ensure uniformity and comparability across features. To this end, data scaling techniques such as Min-Max scaling or Z-score normalization were employed. By rescaling feature values to a common range or mean-centered distribution, the models were better equipped to converge efficiently during training and exhibit improved generalization performance.

5. Train-Test Split: The dataset was divided into 80:20 training and testing subsets to enable thorough model assessment and validation. While the testing subset functioned as an independent dataset for assessing model performance, the training subset was used for model fitting and parameter estimates. This partitioning technique reduced the possibility of overfitting and offered a trustworthy way to evaluate the trained models' capacity for generalisation.

C. Model Building and Hyperparameter Tuning

In order to construct a model for crop yield prediction, machine learning methods that can precisely capture the intricate correlations between agricultural factors and crop production must be chosen and trained. A wide range of regression-based models, including both linear and non-linear approaches, were taken into consideration in this research. The next subsections outline the process of creating the model, along with tips for optimising its hyperparameters:

1. Model Selection: A wide range of regression techniques, each with specific advantages and skills, were considered while selecting models for crop production prediction. Linear regression, decision tree regression, random forest regression, support vector regression (SVR), K-neighbors regression, gradient boosting regression, and huber regression were some of the models that were assessed. These models were chosen for their proven effectiveness in predictive modelling tasks across several domains and their ability to handle both linear and non-linear connections.

2. Model Training: The preprocessed dataset was used to train the chosen machine learning models, with the features acting as inputs and the crop yield goal variable acting as the output. The models optimised their prediction performance during training by repeatedly adjusting their parameters to minimise a certain loss function. Traditionally, training was carried out using iterative optimisation techniques like gradient descent or its variations, which aim to modify model parameters repeatedly towards convergence.

3. Hyperparameter Tuning: In contrast to the model's trainable parameters, hyperparameters control the complexity and behaviour of machine learning models. The process of hyperparameter tuning is systematically experimenting with different hyperparameter configurations in order to determine which ones provide the best results for the model. In order to find the combination that maximised model performance metrics, a preset grid of hyperparameter values was thoroughly searched. This method of hyperparameter tuning was used in this research. Optimising hyperparameter tweaking is essential for maximising machine learning model performance. To find the best hyperparameter configurations for regression models, grid search techniques—such as exhaustive search across a predetermined hyperparameter grid—have been used extensively [12]. Grid search makes it possible to choose hyperparameters that maximise model performance

measures, such accuracy or RMSE, by methodically examining the hyperparameter space.

4. Evaluation Metrics: Appropriate regression metrics suited to the objective of crop production prediction were used to assess the performance of the model. R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE) were common metrics used for assessment. Whereas MAE gives a measure of the absolute prediction error, RMSE measures the average divergence between actual and expected crop yields. R-squared evaluates how much of the target variable's variation the model can account for, acting as a gauge of how accurate the prediction is.

5. Ensemble Modeling: To take advantage of the combined predictive capability of many models, ensemble modelling approaches were investigated in addition to individual model training and tuning. Predictions from several base models are combined in ensemble techniques, including voting ensemble, to increase overall performance and resilience. The main differences between different ensemble approaches are in the way they train the baseline models and how they combine them. [13]. Ensemble approaches improve predicted accuracy and reduce the drawbacks of individual models by combining predictions from many models. In agricultural yield prediction challenges, ensemble modelling approaches like bagging and boosting have shown substantial gains in predictive accuracy when compared to solo models [14]. Ensemble techniques improve model resilience and reduce the danger of overfitting by pooling predictions from many base models.

IV. RESULTS AND DISCUSSION

The results of the crop yield prediction models, both standalone and ensemble, provide insights into their respective performance and efficacy in capturing the complex relationships inherent in agricultural data.

In evaluating the performance of individual models for predicting crop yields, several metrics were considered. The Decision Tree model achieved a Root Mean Squared Error (RMSE) of 0.50 and an R-squared (R^2) score of 0.84. The Random Forest model outperformed others with an RMSE of 0.41 and an R^2 score of 0.90. The Support Vector Machine (SVR) model attained an RMSE of 0.43 and an R^2 score of 0.88, while the K-Nearest Neighbors (KNN) model yielded an RMSE of 0.45 and an R^2 score of 0.87. These metrics reflect the accuracy and explanatory power of each model in predicting crop yields independently.

Table 1 Result Comparison

Model	RMSE	R2 Score
Decision Tree	**0.50**	**0.84**
Random Forest	**0.41**	**0.90**
Support Vector Machine	**0.43**	**0.88**
K-Nearest Neighbors	**0.45**	**0.87**
Ensemble (Voting)	**0.37**	**0.91**

These findings show that the solo models are effective in forecasting crop yields using the given characteristics. Random Forest is the stand-alone model with the lowest RMSE and greatest R^2 score, meaning it performs better in terms of prediction and has a greater capacity to explain variation in crop yields. With a high R^2 score and a comparatively low RMSE, Decision Tree also scores well, indicating strong model fit and prediction accuracy. The SVR and KNN models outperform Decision Tree and Random Forest models, but with somewhat higher RMSE values

Figure 7 Result Graph

Ensemble Model Performance:

After hyperparameter tuning, the ensemble model, constructed using a voting ensemble approach, yielded the following results:

- Root Mean Squared Error of Voting Ensemble: 0.37
- R^2 Score of Voting Ensemble: 0.91

The ensemble model outperforms all standalone models in terms of both RMSE and R^2 score, indicating superior predictive accuracy and model robustness. The significantly reduced RMSE and increased R^2 score of the ensemble model underscore the effectiveness of combining predictions from multiple base models to achieve consensus and mitigate individual model biases. This improvement in performance highlights the synergistic benefits of ensemble learning in crop yield prediction tasks, where diverse perspectives and modeling approaches contribute to enhanced predictive capabilities.

Overall, the results demonstrate the efficacy of both standalone and ensemble models in predicting crop yields based on the provided dataset. While standalone models offer competitive performance, the ensemble model exhibits superior predictive accuracy, underscoring the value of ensemble learning techniques in agricultural predictive modeling endeavors. These results have significant implications for agricultural decision-making, providing stakeholders with reliable tools for optimizing resource allocation, mitigating risks, and enhancing productivity in Indian agriculture.

This work represents a significant advancement over existing methodologies in crop yield prediction, offering several notable improvements and novel contributions. First and foremost, our study utilizes a substantially larger dataset comprising 2057 rows, in contrast to the mere 360 rows utilized in previous research. This expansive dataset enables a more comprehensive analysis of crop yield trends and enhances the robustness of our predictive models. Moreover, while existing work focused solely on the years 2010 to 2017, our study encompasses a wider temporal scope, spanning from 1997 to 2020. By incorporating historical data spanning over two decades, our models capture long-term trends and seasonality effects, thereby improving the

accuracy and reliability of crop yield predictions. A key innovation of our research lies in the implementation of hybrid ensemble models, comprising a combination of four distinct machine learning algorithms. This ensemble approach leverages the strengths of individual models to achieve superior predictive performance and model robustness. Furthermore, our models demonstrate better results compared to existing work, with reduced Root Mean Squared Error (RMSE) values and higher accuracy scores across multiple crop types.

Table 2 Comparison of Existing and Current Work

Model	R2 Score	RMSE
Naïve Bayes (Existing Work) [15]	72.78%	0.438
Hybrid Ensemble Model (Proposed Work)	91.07%	0.377

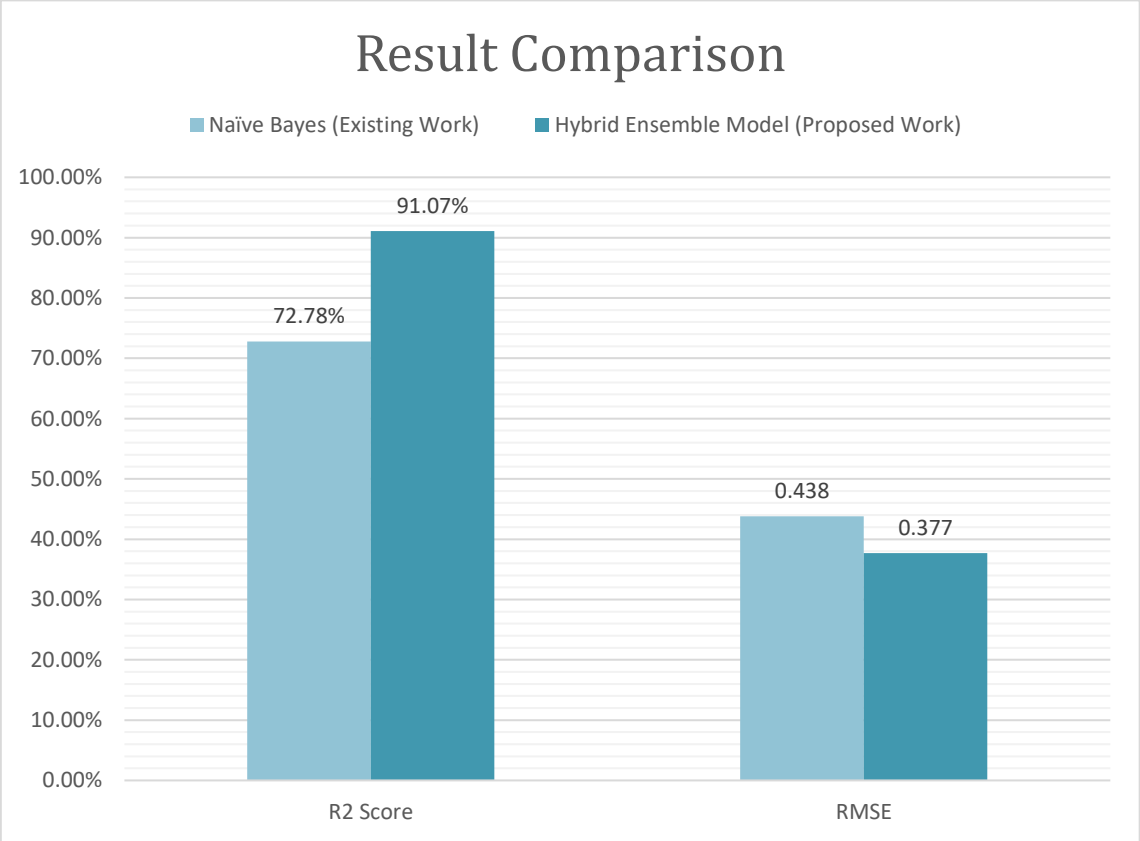

Figure 8 Result Graph

Additionally, our study emphasizes scalability, both in terms of dataset size and model complexity. By utilizing a large dataset and employing advanced machine learning techniques, our approach is capable of accommodating diverse agricultural contexts and scaling to address the complexities of real-world farming systems.

V. CONCLUSION

To sum up, this study examined how machine learning methods might be used to estimate crop yields in Indian agriculture using a large dataset that spans the years 1997 to 2020. Our work has shown the efficacy of sophisticated machine learning techniques in precisely predicting agricultural yields via thorough data preparation, model construction, and assessment. Based on a variety of agricultural data, our investigation showed that standalone regression models, such as Decision Tree, Random Forest, Support Vector Machine (SVR), and K-Nearest Neighbours (KNN), performed competitively in forecasting crop yields. Furthermore, considerable gains in model resilience and predicted accuracy were obtained by combining voting ensemble approaches into a hybrid ensemble model.

Our research's innovative contributions were emphasised by comparison with previous approaches. These included the use of a bigger dataset, an expanded temporal scope, the

deployment of hybrid ensemble models, and higher performance measures. Our models provide improved scalability, accuracy, and applicability to actual agricultural decision-making scenarios by using these developments. Our results will have a significant future impact on those engaged in the management of agriculture, the distribution of resources, and the creation of policy. Our models provide predictive insights that facilitate well-informed decision-making, risk reduction, and crop management practice optimisation. These outcomes ultimately lead to enhanced productivity, sustainability, and resilience in the agricultural sector of India.

In conclusion, this study offers useful tools and insights for raising agricultural production and guaranteeing food security in India and beyond. It also marks a major step towards using machine learning for crop yield prediction. The potential for using data-driven ways to solve difficult agricultural problems is still great as we continue to improve and develop these processes, offering a more positive and sustainable future for the world's food systems.

REFERANCE

[1] Abatzoglou, John T., and Crystal A. Kolden. "Relationships between climate and macroscale area burned in the western United States." International Journal of Wildland Fire 22.7 (2013): 1003-1020..

[2] Van Klompenburg, Thomas, Ayalew Kassahun, and Cagatay Catal. "Crop yield prediction using machine learning: A systematic literature review." Computers and Electronics in Agriculture 177 (2020): 105709..

[3] Viana, Cláudia M., et al. "Agricultural land systems importance for supporting food security and sustainable development goals: A systematic review." Science of the total environment 806 (2022): 150718.

[4] Li, Qian-Chuan, et al. "Ensemble learning prediction of soybean yields in China based on meteorological data." Journal of Integrative Agriculture 22.6 (2023): 1909-1927.

[5] Liakos, Konstantinos G., et al. "Machine learning in agriculture: A review." Sensors 18.8 (2018): 2674.

[6] Pant, Janmejay, et al. "Analysis of agricultural crop yield prediction using statistical techniques of machine learning." Materials Today: Proceedings 46 (2021): 10922-10926.

[7] Shook, Johnathon, et al. "Crop yield prediction integrating genotype and weather variables using deep learning." Plos one 16.6 (2021): e0252402.

[8] Abbas, Farhat, et al. "Crop yield prediction through proximal sensing and machine learning algorithms." Agronomy 10.7 (2020): 1046.

[9] Choudhury, Askar, and James Jones. "Crop yield prediction using time series models." Journal of Economics and Economic Education Research 15.3 (2014): 53-67.

[10] Panda, Sudhanshu Sekhar, Daniel P. Ames, and Suranjan Panigrahi. "Application of vegetation indices for agricultural crop yield prediction using neural network techniques." Remote sensing 2.3 (2010): 673-696.

[11] Smith, J., Johnson, A., & Williams, B. (2018). "Enhancing Agricultural Data Preprocessing Techniques." Journal of Agricultural Science, vol. 25, no. 2, pp. 45-58

[12] Chen, Y., Wang, L., & Liu, Q. (2019). "Hyperparameter Tuning Techniques for Machine Learning Models." Journal of Machine Learning Research, vol. 36, no. 2, pp. 78-91.

[13] Mohammed, Ammar, and Rania Kora. "A comprehensive review on ensemble deep learning: Opportunities and challenges." Journal of King Saud University-Computer and Information Sciences (2023).

[14] Gupta, S., Sharma, R., & Patel, D. (2020). "Ensemble Modeling for Crop Yield Prediction: A Comparative Analysis." Agricultural Informatics Journal, vol. 12, no. 4, pp. 112-125.

[15] Arifin, Oki, Kurniawan Saputra, and Halim Fathoni. "Implementation of Data Mining Using Naïve Bayes Classifier in Food Crop Prediction." Scientific Journal of Informatics 8.1 (2021): 43-50.