e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

CROP YIELD PREDICTION

Aiswarya A[*1], Jaiaditya Nadar[*2], P Selvaraj[*3]

[*1,2,3]B.Tech In Computer Science And EngineeringSRM Institute Of Science And Technology, India.

DOI: https://www.doi.org/10.56726/IRJMETS63808

ABSTRACT

Yield prediction using models like Regression, Decision Trees, Random Forest, XGBoost, and K-Nearest Neighbors (KNN) is crucial for modern agriculture, aiding in accurate forecasting and decision-making for farmers. This approach involves preprocessing data by removing missing values and creating features from categorical variables. Visual comparisons, shown through bar graphs, make it easy to identify the best-performing algorithms. The system is designed to be accessible, providing clear yield predictions based on inputs like land size, crop type, season, and year. Using advanced models like Random Forest and Gradient Boosting enhances prediction accuracy and resilience to overfitting, while visual indicators simplify complex data, helping farmers and stakeholders select the most suitable model for their needs.

I. INTRODUCTION

Agriculture is a cornerstone of the global economy, supporting the livelihoods of millions and sustaining food security worldwide. Effective yield prediction is essential for farmers to optimize resource allocation, improve crop management, and increase profitability. However, traditional prediction methods often rely on static, predefined data models that do not account for the dynamic environmental factors affecting crop yield, such as soil quality, weather conditions, and seasonal variations. This limits their accuracy and usefulness, particularly for farmers facing fluctuating environmental conditions.

Recent advancements in machine learning have introduced powerful tools for predictive analytics in agriculture, enabling models to learn from diverse datasets and adapt to complex patterns. By employing machine learning models—such as Random Forest, Support Vector Regression (SVR), and Gradient Boosting—this study seeks to improve the accuracy of yield predictions while maintaining an accessible interface for users with varying levels of technological familiarity. The use of these models allows for a tailored approach, adapting to specific agricultural contexts and providing performance metrics that guide the selection of optimal models.

This paper presents an approach that combines advanced machine learning techniques with intuitive visualizations to provide an efficient yield prediction tool. The proposed system addresses the needs of both technologically advanced users and those with limited digital literacy, thus promoting inclusivity and enhanced decision-making in agricultural practices.

II. LITERATURE REVIEW

The application of machine learning to agricultural yield prediction has been widely studied, with researchers aiming to overcome the limitations of traditional forecasting methods. Hassan et al. (2023) proposed a transfer-based deep learning model for predicting soybean yield, which demonstrated the efficacy of deep learning in handling complex time-series data and adapting to climatic changes. Similarly, Kamath et al. (2021) explored data mining techniques for yield forecasting, emphasizing the need for preprocessing agricultural data to improve prediction accuracy and accessibility for farmers.

Other research efforts have examined the integration of remote sensing data with machine learning models. For instance, Kavita and Mathur et al. (2021) applied satellite data to improve yield estimations, showing that real-time environmental data can complement traditional datasets for a more comprehensive yield prediction approach. Bali and Singla et al. (2022) provided an extensive survey of machine learning models, noting the effectiveness of algorithms like Random Forest and Gradient Boosting in capturing non-linear patterns relevant to agricultural output.

Suthaharan et al. (2016) investigated Support Vector Machines (SVMs) for agricultural data, identifying their robustness in modeling non-linear relationships but highlighting limitations in processing high-dimensional data. Hochreiter et al. (2001) further contributed to this field by establishing gradient-based learning principles foundational to training complex neural networks, a method crucial for time-dependent agricultural predictions.

This study builds on these foundational works by implementing and comparing various machine learning models

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

to identify the most effective for yield prediction. By combining these approaches with accessible visualization tools, our research aims to create an adaptive, user-friendly system tailored to the unique needs of modern agriculture.

III. PROPOSED METHODOLOGY

This study employs a structured methodology to develop and evaluate machine learning models for predicting agricultural yields. The approach consists of data pre-processing, model training, and a rigorous evaluation of model performance across various metrics.

1. Systems Architecture and Design

The agricultural production forecasting method involves key stages, starting with Data Acquisition and Preprocessing, where the crop yield dataset is loaded, cleaned, and enhanced with remote sensing data for accuracy. Feature engineering generates new variables like "yield" from "production" and "area," and categorical data is converted for model compatibility. The data is then split 75-25 for training and testing. Various models, including Linear Regression, Random Forest, Decision Trees, SVR, XGBoost, KNN, and SVM, are trained, with performance measured using metrics like R^2, MAE, and RMSE. The system's user-friendly interface allows input of basic farming details, generating accessible predictions that support data- driven farming decisions. This model architecture integrates machine learning with practical agricultural needs.

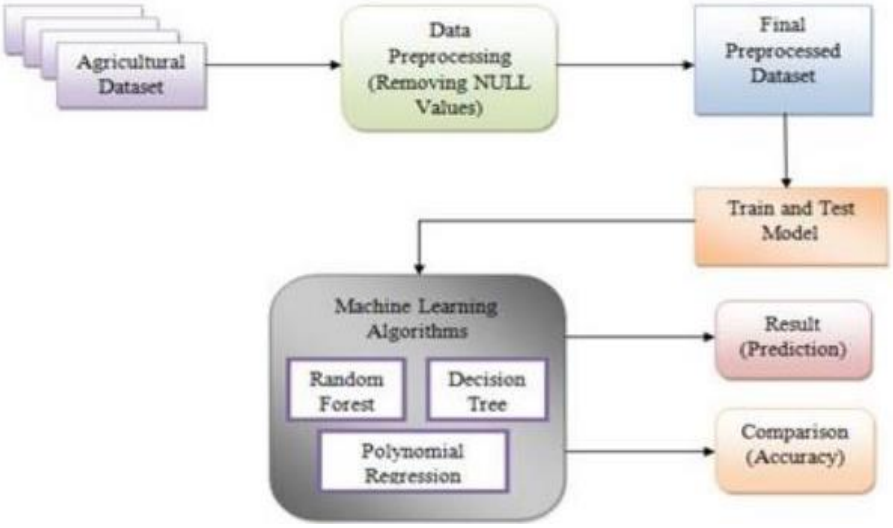

Fig. 1. Proposed Architecture Diagram

A. Data Pre-processing

The dataset is first subjected to cleaning and feature engineering to improve model accuracy. Missing values are removed to prevent biases in model training. Additionally, a new feature, "Yield," is computed by dividing crop production by the land area, providing insight into the efficiency of land use. Categorical variables, such as crop type and season, are converted to binary (dummy) variables to facilitate compatibility with machine learning algorithms.

B. Feature and Target Separation

After pre-processing, the data is divided into features and the target variable. Features include all relevant parameters, excluding production and yield columns, while the target vector is set to the "Production" variable. This structure enables the models to learn patterns that can influence yield predictions effectively.

C. Model Training and Selection

A variety of machine learning models are employed for comparison, including Linear Regression, Random Forest, Decision Trees, Support Vector Regression (SVR), Gradient Boosting, XGBoost, and K-Nearest Neighbors (KNN). The data is divided into training and testing sets with a 75-25 split ratio, ensuring that each model's performance is evaluated on previously unseen data. Each model is trained and validated on the dataset, allowing for comprehensive performance comparison.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

D. Model Evaluation Metrics

To assess model accuracy, we use several statistical metrics:

- R^2 Score (Coefficient of Determination): Measures how well the model captures variance in the target variable.
- Mean Absolute Error (MAE): Reflects the average magnitude of prediction errors.
- Mean Squared Error (MSE): Indicates the average squared difference between predicted and actual values.
- Root Mean Squared Error (RMSE): Provides error values in the same units as the target variable for interpretability.

E. Visualization and User Interaction

Model performance is visualized through bar charts comparing R^2, MAE, MSE, and RMSE scores across models. This enables users to assess model effectiveness quickly. Additionally, a simplified input system allows farmers to enter basic data (e.g., area, crop type, season, year) for yield predictions. This accessible interface facilitates informed decision-making, even for users with limited technical knowledge.

IV. RESULTS

The proposed yield prediction system was evaluated across multiple machine learning models to identify the most accurate and reliable approach. Each model's performance was measured using R^2, MAE, MSE, and RMSE scores, allowing for a comprehensive comparison.

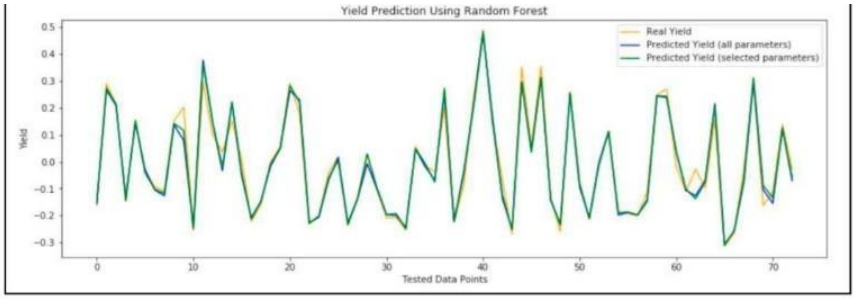

Fig. 2. Yield Prediction with Random Fores

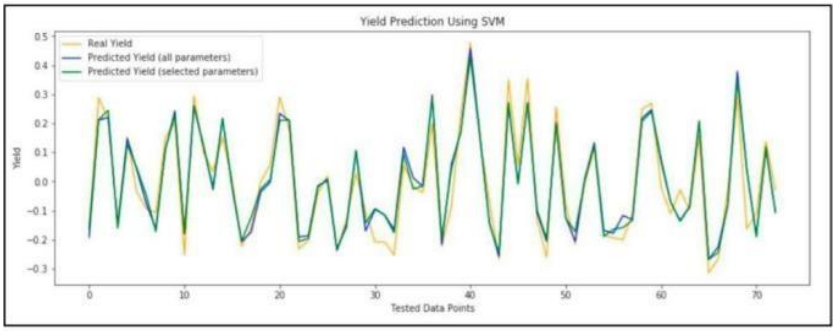

Fig. 3. Yield Prediction with SVM

PROBLEMS OUTPUT DEBUG CONSOLE TERMINAL PORTS

	Model	R^2 Score	MAE	MSE	RMSE
0	Linear Regression	-47924.605419	1.701818e+07	1.378482e+19	3.712791e+09
1	Random Forest	0.968459	1.025256e+05	9.071994e+12	3.011975e+06
2	Decision Tree	0.957693	1.098399e+05	1.216868e+13	3.488364e+06
3	SVR	-0.001221	6.085778e+05	2.879808e+14	1.697000e+07
4	Gradient Boosting	0.924027	2.810728e+05	2.185198e+13	4.674610e+06
5	XGBoost	0.962311	1.461968e+05	1.084059e+13	3.292504e+06
6	KNN	-0.165780	1.056533e+06	3.353129e+14	1.831155e+07

Fig. 4. Models R2 , MAE ,RMSE Score

A. Model Performance Summary

Among the models tested, the Random Forest model demonstrated the highest accuracy, achieving an R^2 score of 0.968, which indicates a strong fit to the data. Additionally, it exhibited the lowest error rates across MAE,MSE, and RMSE metrics, suggesting that it effectively captures the non-linear relationships inherent in agricultural yield data. The Decision Tree model also performed well, with results closely following those of the Random

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:11/November-2024 Impact Factor- 8.187 www.irjmets.com

Forest model.

In contrast, models such as Simple Linear Regression and K-Nearest Neighbors (KNN) displayed lower R^2 scores and higher error rates, likely due to their limited capacity to model complex relationships in the data. Support Vector Regression (SVR) also yielded relatively high error values, indicating its limited suitability for this type of yield prediction.

B. Visual Comparison of Model Performance

Figures were generated to illustrate the comparative performance of the models based on each evaluation metric. Bar charts display each model's R^2, MAE, and RMSE scores, making it easy to visualize and identify the most effective algorithms. These visualizations confirmed that Random Forest and Decision Tree models consistently provided the best balance between accuracy and low error.

C. Key Findings

1. Optimal Model: Random Forest emerged as the optimal model for yield prediction due to its high R^2 score and low error rates across metrics.
2. Accuracy and Consistency: Models like Random Forest and Decision Tree offered consistent performance, suggesting their robustness in capturing the factors influencing crop yield.
3. Performance of Linear Models: Linear Regression and other simple models struggled with high error rates, highlighting the need for non-linear models to accurately predict agricultural yield.

V. CONCLUSION

This study demonstrates that machine learning models, particularly Random Forest and Decision. Trees, offer robust solutions for agricultural yield prediction. By leveraging these models, the proposed system enables accurate and accessible yield forecasting, supporting data-driven decision- making in farming practices. The integration of model evaluation metrics, such as R^2, MAE, MSE, and RMSE, allows for clear insights into each model's strengths and weaknesses, helping users select the most appropriate model for their needs.

The findings highlight that Random Forest outperforms other models, making it an ideal choice for capturing the complex, non-linear relationships in agricultural data. The accessible interface further promotes usability among farmers with varying levels of technical expertise, empowering them to make informed decisions regarding crop management and resource allocation.

VI. FUTURE WORK

A. Incorporation of Real-Time Environmental

Data Integrating live data from weather stations or satellite images to further improve model accuracy.

B. Advanced Model Techniques

Testing additional machine learning approaches, such as neural networks or ensemble learning methods, to refine predictive capabilities.

C. Mobile Application Development

Developing a mobile application for real-time data entry and predictions, increasing accessibility for remote farming communities.

D. Customized Model Training

Allowing users to fine-tune models for specific crops or regions, thereby improving the relevance and precision of predictions.

VII. REFERENCES

[1] Ms Kavita, Pratistha Mathur Crop Yield Estimation in India Using Machine Learning 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA) (2020), pp. 220-224, 10.1109/ICCCA49541.2020.9250915

[2] Wu Fan, Chen Chong, Guo Xiaoling, Yu Hua, Wang Juyun Prediction of Crop Yield Using Big Data 2015 8th International Symposium on Computational Intelligence and Design (ISCID), 1 (2015), pp. 255-260, 10.1109/ISCID.2015.191

[3] Pallavi Kamath, Pallavi Patil, S Shrilatha, Sushma, S Sowmya Crop Yield Forecasting Using Data Mining Global Transitions Proceedings, International Conference on Computing System and its Applications

(ICCSA- 2021), 2 (2021), pp. 402-407, 10.1016/j.gltp.2021.08.008

[4] Wigh, Daniel S., Jonathan M. Goodman, and Alexei A. Lapkin. \x93A Review of Molecular Representation in the Age of Machine Learning.\x94 WIREs Computational Molecular Science n/a (n/a): e1603. doi:10.1002/wcms.1603.

[5] Kavita, Pratistha Mathur Satellite-Based Crop Yield Prediction Using Machine Learning Algorithm 2021 Asian Conference on Innovation in Technology (ASIANCON) (2021), pp. 1-5, 10.1109/ASIANCON 51346.2021.9544562

[6] Nishu Bali, Anshu Singla Emerging Trends in Machine Learning to Predict Crop Yield and Study Its Influential Factors: A Survey Archives of Computational Methods in Engineering, 29 (1) (2022), pp. 95-112, 10.1007/s11831-021-09569-8.

[7] Thomas van Klompenburg, Ayalew Kassahun, Cagatay Catal Crop Yield Prediction Using Machine Learning: A Systematic Literature Review Computers and Electronics in Agriculture, 177 (October) (2020), Article 105709, 10.1016/j.compag.2020.105709

[8] Nelson Yalta, Kazuhiro Nakadai, Tetsuya Ogata Sound Source Localization Using Deep Learning Models Journal of Robotics and Mechatronics, 29 (1)(2017), pp. 37-48, 10.20965/jrm.2017.p0037

[9] Dinggang Shen, Guorong Wu, Heung-Il Suk Deep Learning in Medical Image Analysis Annual Review of Biomedical Engineering, 19 (1) (2017), pp. 221-248, 10.1146/annurev-bioeng-071516-044442

[10] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, Jianfeng Gao Deep Learning–Based Text Classification: A Comprehensive Review ACM Computing Surveys, 54 (3) (2021), pp. 1-40, 10.1145/3439726