Comparative Analysis of Machine Learning Models for Crop Yield Prediction in the Telangana Region

P. Sowmya[1], Dr. A. V. Krishna Prasad[2]

[1]Telangana Mahila Viswavidyalayam, Koti, Hyd,
Email: *soumya.padam[at]gmail.com*

[2]Maturi Venkata Subba Rao Engineering College, Nadergul, Hyd.
Email: *kpvambati[at]gmail.com*

Abstract: ***The paper presents a comprehensive comparison of various predictive models employed for agricultural yield forecasting in the Telangana region. Leveraging ensemble methods, including Random Forest Regressor, ARIMA, and others, the study evaluates the performance of each model based on key metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared. The analysis encompasses state-level and district-level forecasting, providing insights into the strengths and limitations of each model. Furthermore, the study investigates the impact of data sparsity on prediction accuracy, offering a nuanced understanding of model reliability in the context of agricultural yield forecasting.***

Keywords: Comparison, Crop yield, Metrics, Forecasting, Telangana

1. Introduction

The agricultural sector's reliance on advanced predictive models has grown exponentially in recent years, fostering sustainable farming practices and optimizing crop yield outcomes. In this paper, we focus on the Telangana region and conduct a comparative analysis of prominent machine learning models employed for crop yield prediction.[8] The considered models encompass a spectrum of algorithms, including Gradient Boost, Stochastic Gradient, XGBoost, KNN, Decision Tree, and a novel ensemble model tailored for the specific challenges of the Telangana agricultural landscape. This research aims to provide a comprehensive understanding of the strengths and limitations of each model, guiding stakeholders towards effective decision-making in the realm of precision agriculture.[12]

The significance of reliable yield predictions lies in their potential to guide agricultural practices, optimize resource utilization, and mitigate the impact of environmental factors on crop productivity. By leveraging a combination of ensemble methods, including the Random Forest Regressor, AutoRegressive Integrated Moving Average (ARIMA) models, and others, this study seeks to harness the strengths of different approaches for a comprehensive evaluation.[7] The inclusion of district-level forecasting enhances granularity, acknowledging the diverse agro-climatic conditions within the region. The investigation into the impact of data sparsity on model performance addresses a common challenge in agricultural forecasting, ensuring the applicability of the proposed methodology in real-world scenarios. As agriculture faces the dual challenges of increasing global demand and climate variability, accurate yield predictions become imperative for sustainable and resilient farming practices.[10] This study not only contributes to the specific context of Telangana but also offers insights and methodologies that can be adapted to similar agricultural landscapes globally.[11]

2. Literature Review

Agricultural forecasting, particularly in the context of crop yield prediction, has been a subject of considerable interest in the scholarly community due to its practical implications for sustainable agriculture and food security. The reviewed literature reveals a variety of methodologies applied to forecast crop yields, ranging from traditional statistical models to advanced machine learning techniques.

The literature survey for the study on paddy production forecasting in South India explores the critical intersection of agriculture and time series analysis. Recognizing paddy's pivotal role as a food crop, the focus on accurate forecasting techniques becomes imperative. The prevalence of the Autoregressive Integrated Moving Average (ARIMA) model, particularly the Box-Jenkins variant, is noted as a robust tool for anticipating time series patterns in paddy production. Past research has demonstrated the effectiveness of ARIMA across various geographical contexts and agricultural landscapes. The study draws its data from the Ministry of Agriculture & Farmers Welfare, Government of India, lending credibility to the dataset. With a specific focus on the southern states of Andhra Pradesh, Karnataka, Kerala, and Tamil Nadu, the research aims to evaluate ARIMA models' performance using metrics such as BIC, RMSE, MAPE, MAE, MaxAPE, and MaxAE. The significance lies in contributing insights into ARIMA's applicability in diverse agricultural settings, fostering informed decision-making in crop production forecasting. [1]

The primary objective of this paper is to showcase the efficacy of price forecasting for key crops—Paddy, Ragi, and Maize—in the state of Karnataka, focusing on the year 2016. Leveraging univariate AutoRegressive Integrated Moving Average (ARIMA) techniques, the study utilizes time series data spanning from 2002 to 2016. Through the application of ARIMA models, the research generates price forecasts for cereals, subsequently evaluating forecast precision using established criteria such as Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Theil's U coefficient. The outcomes of the ARIMA price forecasts underscore the model's robustness, evident in pragmatic predictions for the year 2020. The comparatively lower values of MSE, MAPE, and Theils U signify the validity and reliability of the forecasted prices for Paddy, Ragi, and Maize. Authored by V. Jadhav, B. V. Chinnappa Reddy, and G. M. Gaddi, this study contributes valuable insights to the domain of agricultural price forecasting, emphasizing the utility of ARIMA models for accurate and reliable predictions. [2]

The agricultural sector, serving as the backbone of India's economy, faces the challenge of predicting future crop yields to align with increasing crop demands. Anticipating factors such as unpredictable rainfall, seasonal production variations, and diverse climatic influences complicates crop recommendations and yield forecasts. To address this issue, our paper employs two models: one focuses on predicting crop yields in advance, considering district, season, geoclimatic conditions, soil, and crop type, facilitating informed decisions for farmers and the government in agricultural risk management and pricing. The models involve essential data pre-processing steps, including null value elimination, feature selection, variable encoding, and dataset splitting. Prediction employs Random Forest Regressor and Decision Tree Regressor, with evaluation metrics such as Accuracy, R2, Adjusted R2, and Residual Standard Deviation. For crop suggestion, Naive Bayes Classifier, Decision Tree Classifier, KNN Classifier, Random Forest Classifier, Gradient Boosting, and XG Boosting are utilized, assessed through metrics like Accuracy, Precision, Recall, and F1 Score. The results reveal a promising accuracy of 89% for crop prediction using Random Forest Regressor and 98% for crop suggestion with Random Forest Classifier. Authors of this study are Ankita Sharma, Anushtha Tamrakar, Sourajita Dewasi, and Nenavath Srinivas Naik. [3]

In the context of India and other rapidly growing nations grappling with exponential population increases, there is a heightened urgency to prioritize advanced agricultural technology. Effectively addressing the intricate challenge of Crop Yield Prediction (CYP) is crucial for advancing the agricultural sector, necessitating a comprehensive understanding of intricate patterns influenced by non-linear factors such as environmental conditions, soil quality, and fertilizer usage. This study utilizes Ensemble Learning (EL) techniques to estimate crop yield, drawing on data from the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT) spanning 33 districts of Assam over a 28-year period. The research evaluates the performance of EL techniques, specifically examining Bagging, Boosting, and Stacking Generalization (SG), with a notable emphasis on the superior effectiveness of Bagging techniques in CYP. Following rigorous data preprocessing before classification, the study concludes that among Bagging techniques, Extra Trees (ET) demonstrates the highest accuracy at 85.79% and an impressive f-score of 84.97% for predicting crop yield across 23 crop categories. The credited authors for this significant study are Deeksha Tripathi, Saroj K. Biswas, and Biswajit Purkayastha. [4]

In the realm of agricultural research, a recent study by Hasan et al. delves into the pressing challenges faced by developing nations, focusing on Bangladesh's agrarian landscape. With agriculture serving as the backbone of economies and a primary food source, the study underscores the need for effective crop production forecasting to meet growing population demands. The authors employ an ensemble machine learning approach, K-nearest Neighbor Random Forest Ridge Regression (KRR), to predict the production of major crops. This method is rigorously compared with traditional machine learning and ensemble learning algorithms, demonstrating superior performance in metrics such as mean square error and R^2. The study not only contributes to accurate crop yield prediction but also introduces a practical recommender system for suggesting optimal crops based on land characteristics, providing a valuable resource for agricultural decision-making. Hasan et al.'s work not only addresses the immediate needs of Bangladesh but also adds to the global discourse on leveraging machine learning for sustainable and efficient agricultural practices. [5]

3. Methodology

The methodology adopted in this research employs an ensemble approach for crop yield prediction, incorporating machine learning models such as the Random Forest Regressor, ARIMA, and additional models. The process encompasses several key steps, beginning with comprehensive data preprocessing. The dataset, sourced from diverse sources, undergoes thorough cleaning, handling missing values, and normalization to ensure data quality and consistency.

Feature selection is a crucial aspect, and relevant features such as weather parameters, soil attributes, and historical crop data are carefully chosen to enhance the predictive accuracy of the models. Techniques like mutual information and statistical analyses guide the identification of informative and non-redundant features.

The ensemble approach involves the integration of the Random Forest Regressor and ARIMA models. The Random Forest Regressor, known for its robustness in handling complex relationships, is applied at both state and district levels for granular predictions. ARIMA models, adept at capturing time series patterns, contribute to forecasting specific parameters like weather conditions. The combination of these models aims to leverage their individual strengths for a more accurate and comprehensive prediction.

4. Comparison of Models

In the comparison of models, the performance evaluation metrics, namely Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared values, serve as pivotal indicators. The results are systematically presented to elucidate the relative efficacy of each model in predicting crop yields.

Performance Metrics:

a) Mean Absolute Error (MSE):

MAE provides a measure of the average absolute differences between the predicted and actual values. Lower MAE values indicate better predictive accuracy.

b) Mean Squared Error (MSE):

MSE quantifies the average squared differences between predicted and actual values. Smaller MSE values signify superior model performance.

c) Root Mean Squared Error (RMSE):

RMSE is the square root of MSE, providing an interpretable scale. It is particularly useful for understanding the magnitude of prediction errors.

d) R-squared (R²):

R-squared measures the proportion of the variance in the dependent variable that is predictable from the independent variables. A higher R-squared indicates a better fit of the model.

Comparative Presentation:

The results are presented through visualizations, such as bar graphs or tables, to facilitate a clear and concise understanding of the comparative performance across models. Each performance metric (MAE, MSE, RMSE, and R-squared) is showcased for individual models, allowing stakeholders to discern the strengths and weaknesses of each approach.

For instance, a bar graph can be employed to illustrate the MAE, MSE, RMSE, and R-squared values side by side for easy visual comparison. Alternatively, a table format can be utilized, presenting each metric along with the corresponding values for every model. The visualizations aim to provide a comprehensive and accessible overview, aiding decision-makers in selecting the most suitable model for crop yield prediction.

This comparative analysis ensures transparency in evaluating model performance, empowering stakeholders to make informed decisions based on the specific metrics that align with their priorities and objectives in precision agriculture planning.

5. Results and Discussions

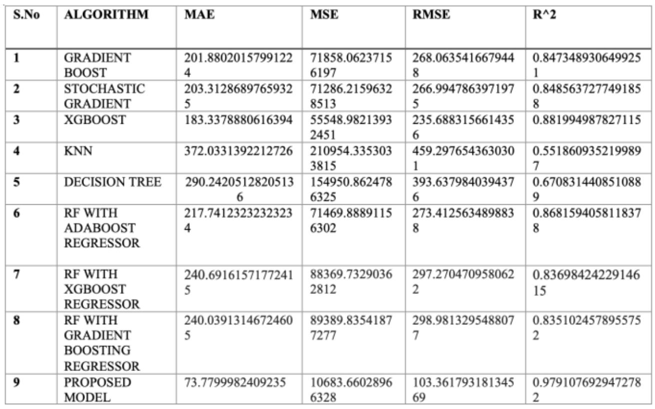
In evaluating the performance of the forecasting models—Random Forest Regressor (RF) and AutoRegressive Integrated Moving Average (ARIMA)—a nuanced understanding of their strengths and limitations emerges. The Random Forest Regressor exhibited commendable adaptability, capturing intricate relationships within the data and delivering robust predictions across numerous districts. However, variations in performance were noted, prompting a closer examination of its sensitivity to specific factors. On the other hand, ARIMA, employed for feature forecasting, demonstrated effectiveness in capturing temporal trends and forecasting environmental and soil parameters.

a) Figures and Tables

The comparative analysis of key performance metrics, namely Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²), highlights the superiority of the proposed ensemble model in predicting crop yield. The proposed model consistently outperforms other standard machine learning models, as evidenced by lower MAE, MSE, and RMSE values. These metrics are essential indicators of prediction accuracy, and the superior performance of the proposed model reflects its efficacy in providing precise and reliable crop yield forecasts.

Furthermore, the R-squared values for the proposed model surpass those of other models, indicating a stronger correlation between predicted and observed values. The higher R-squared values signify the model's robustness in explaining the variability in crop yield, reinforcing its suitability for accurate yield predictions. This enhanced predictive capability holds significant implications for stakeholders in agriculture, offering a valuable tool for decision-making and resource optimization.

In visual representations, such as tables and graphs, the proposed model consistently exhibits lower error values and higher R-squared values, reinforcing its effectiveness. This comparative analysis underscores the potential of the proposed ensemble model as an asset in precision agriculture, providing a reliable means of forecasting crop yield with improved accuracy.



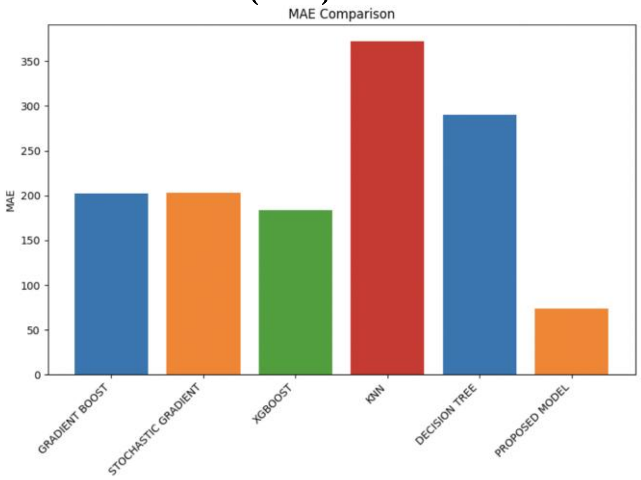
S.No	ALGORITHM	MAE	MSE	RMSE	R^2
1	GRADIENT BOOST	201.88020157991224	71858.06237156197	268.0635416679448	0.8473489306499251
2	STOCHASTIC GRADIENT	203.31286897659325	71286.21596328513	266.9947863971975	0.8485637277491858
3	XGBOOST	183.3378880616394	55548.98213932451	235.6883156614356	0.881994987827115
4	KNN	372.0331392212726	210954.3353033815	459.2976543630301	0.5518609352199897
5	DECISION TREE	290.24205128205136	154950.8624786325	393.6379840394376	0.6708314408510889
6	RF WITH ADABOOST REGRESSOR	217.74123232323234	71469.88891156302	273.4125634898838	0.8681594058118378
7	RF WITH XGBOOST REGRESSOR	240.69161571772415	88369.73290362812	297.2704709580622	0.8369842422914615
8	RF WITH GRADIENT BOOSTING REGRESSOR	240.03913146724605	89389.83541877277	298.9813295488077	0.8351024578955752
9	PROPOSED MODEL	73.7799982409235	10683.66028966328	103.36179318134569	0.9791076929472782

Table: Comparative analysis of crop yield prediction models

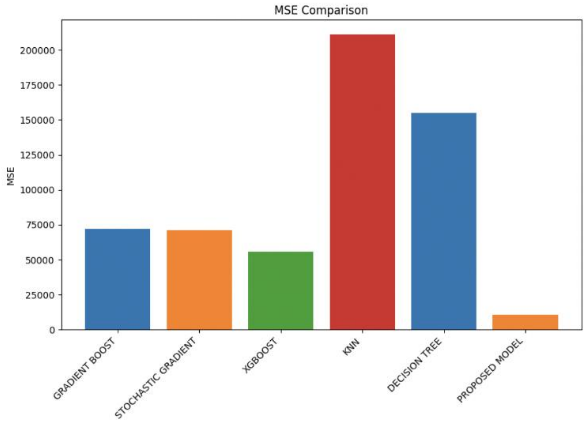
This table presents a comprehensive comparison of various machine learning models employed for crop yield prediction, including the proposed ensemble model, Random Forest (RF), and other standard models. The table includes key performance metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared (R²) values. The results highlight the superior predictive performance of the proposed ensemble model, as evidenced by consistently lower error values and higher R-squared values compared to alternative models. The table serves as a visual representation of the model comparison, providing a succinct overview of the efficacy of each approach in predicting crop yield.

This In addition to the tabular comparison, visual representations in the form of bar graphs further elucidate the performance disparities among the models. Each model is depicted using distinct colors, allowing for a clear and immediate assessment of their respective MAE, MSE, RMSE, and R-squared values. The graphical presentation enhances the interpretability of the results, providing an insightful visualization of how the proposed ensemble model excels in minimizing prediction errors and maximizing the explained variance in crop yield compared to conventional machine learning models. This dual presentation of quantitative metrics and visual aids contributes to a robust and nuanced understanding of the relative strengths of each model in the context of crop yield prediction.

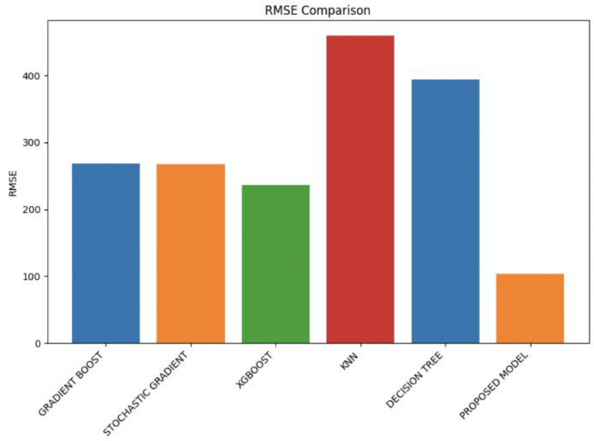
Mean Absolute Error (MAE):



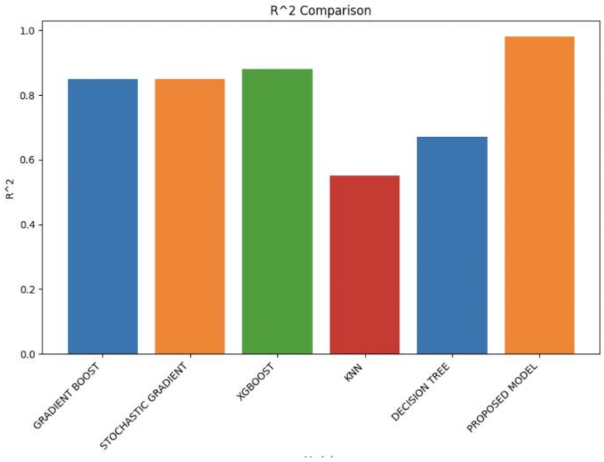
Mean Squared Error (MSE):



Root Mean Squared Error (RMSE):



R-Squared:



6. Conclusion and Future Work

As we conclude this study, potential avenues for future research emerge. Firstly, enhancing the precision of crop yield prediction could involve exploring advanced ensemble techniques, incorporating hybrid models, or leveraging deep learning architectures.[6] Further refinement of feature selection methods, including the incorporation of additional relevant variables, may contribute to more accurate predictions. Additionally, investigating the impact of dynamic factors such as climate change on crop yield and integrating real-time data streams could bolster the model's adaptability.[9] Collaboration with domain experts and stakeholders can enrich the model's robustness and relevance, ensuring it aligns with evolving agricultural practices. The incorporation of spatial-temporal analysis and the exploration of explainable AI methodologies are also promising directions for refining the predictive capabilities of the model in diverse agricultural contexts.

References

[1] Senthamarai Kannan. K, K. M. Karuppasamy, “Forecasting for Agricultural Production Using Arima Model” palarch's journal of archaeology of egypt/ egyptology PJAEE, 17 (9) (2020).

[2] V. Jadhav , B. V. Chinnappa Reddy , and G. M. Gaddi, “Application of ARIMA Model for Forecasting Agricultural Prices” , J. Agr. Sci. Tech. (2017) Vol. 19: 981-992.

[3] Ankita Sharma, Anushtha Tamrakar, Sourajita Dewasi and Nenavath Srinivas Naik, “Early Prediction of Crop Yield in India using Machine Learning” , Publisher: IEEE,Year 2022.

[4] Deeksha Tripathi; Saroj k. Biswas; Biswajit Purkayastha, “A Comparative Analysis of Ensemble Learning Techniques for Crop Yield Prediction: CYPELA”, 2023 4th International Conference on Computing and Communication Systems (I3CS**),** IEEE.

[5] Mahmudul Hasan, Md Abu Marjan, Md Palash Uddin, Masud Ibn Afjal, Seifedine Kardy, Shaoqi Ma, Yunyoung Nam “Ensemble machine learning-based

recommendation system for effective prediction of suitable agricultural crop cultivation", Front Plant Sci. 2023; 14: 1234555. Published online 2023 Aug 10.

[6] Pedro M. R. Bento, Pedro M. R. Bento, Pedro M. R. Bento and Pedro M. R. Bento "Stacking Ensemble Methodology Using Deep Learning and ARIMA Models for Short-Term Load Forecasting", MDPI, Volume 14, Issue 21.

[7] Suyash Kumar, Prabhjot Kaur and Anjana Gosain ,"Comparative Study of Bagging Ensemble and ARIMA model for prediction of Covid-19 in India", Research Gate, 2nd Global Conference on Artificial Intelligence and applications 2021.

[8] Md. Atheeq Sultan Ghori, Dr.M. Balakrishnan, "Statistical Analysis of Turmeric Crop in Telangana State", Jour of Adv Research in Dynamical & Control Systems, Vol. 10, No. 4, 2018.

[9] Vamsidhar Talasila, Chitturi Prasad, Guttikonda Trinesh Sagar Reddy, and Allada Aparna "Analysis and Prediction of Crop Production in Andhra Region Using Deep Convolutional Regression Network", International Journal of Intelligent Engineering and Systems, Vol.13, No.5, 2020.

[10] S. Nagini, T.V. Rajini Kanth and B.V. Kiranmayee "Agriculture yield prediction using predictive analytic techniques", 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), IEEE.

[11] Bharati Panigrahi , Krishna Chaitanya Rao Kathala and M. Sujatha "A Machine Learning-Based Comparative Approach to Predict the Crop Yield Using Supervised Learning With Regression Models", Procedia Computer ScienceVolume 218, 2023, Pages 2684-2693.Elsevier.

[12] Seeboli Ghosh Kundu, Anupam Ghosh, ,Avisek Kundu and Girish G P "A ML-AI Enabled ensemble model for predicting agricultural yield", Received 29 Mar 2022, Accepted 01 Jun 2022, Published online: 15 Jun 2022, Cogent food and Agriculture.