ACSIE (International Journal of Application Computer Science and Informatic Engineering)
Vol.6, No.2, November 2024, pp. 106-116
ISSN (online): 2685-4600
DOI: 10.33173/acsie.v6i2.305

Crop Yield Estimation Using AdaBoost Regression Under Multivariate Environmental Conditions

Heru Ismanto*[1], I Dewa Ayu Sri Murdhani[2]
[1]Department Informatics Engineering, Universitas Musamus Merauke, Indonesia
[2]Magister Program of Informatics, Institut Bisnis dan Teknologi Indonesia, Denpasar, Indonesia
e-mail: ***[1]heru@unmus.ac.id**, [2]sri.murdhani@instiki.ac.id

Abstrak

Prediksi hasil panen yang akurat merupakan komponen penting dalam perencanaan pertanian, ketahanan pangan, dan praktik pertanian berkelanjutan. Namun, hasil panen dipengaruhi oleh interaksi yang kompleks dan nonlinier antara faktor lingkungan dan praktik pengelolaan pertanian, seperti kondisi iklim dan penggunaan pestisida, yang sering kali tidak dapat dimodelkan secara optimal oleh pendekatan statistik tradisional. Berdasarkan permasalahan tersebut, penelitian ini mengusulkan model prediksi hasil panen menggunakan algoritma regresi AdaBoost yang mengintegrasikan variabel iklim multivariat dan data penggunaan pestisida.Pendekatan yang diusulkan menerapkan strategi pembelajaran ensemble untuk meningkatkan akurasi prediksi dengan menggabungkan beberapa regresor lemah secara adaptif, sehingga mampu menangkap hubungan nonlinier dan mengatasi variabilitas data. Metodologi penelitian meliputi tahap prapemrosesan data, normalisasi fitur, pelatihan model, serta optimasi hiperparameter. Kinerja model dievaluasi menggunakan metrik regresi standar, yaitu koefisien determinasi (R^2), Mean Absolute Error (MAE), dan Root Mean Square Error (RMSE).Hasil eksperimen menunjukkan bahwa model regresi AdaBoost mampu menghasilkan prediksi hasil panen yang akurat dan andal, dengan nilai prediksi yang mendekati data aktual. Temuan ini menunjukkan bahwa penggabungan faktor iklim dan penggunaan pestisida dalam satu model memberikan representasi sistem pertanian yang lebih komprehensif dan realistis. Kontribusi utama penelitian ini terletak pada penerapan regresi AdaBoost untuk prediksi hasil panen di bawah kondisi lingkungan multivariat. Penelitian selanjutnya dapat mengembangkan model ini dengan menambahkan variabel agronomis lain, seperti sifat tanah dan praktik irigasi, serta mengeksplorasi pendekatan pembelajaran mesin lanjutan untuk meningkatkan kinerja prediksi.

Kata kunci: *Prediksi hasil panen, regresi AdaBoost, faktor iklim, penggunaan pestisida, pembelajaran mesin, pertanian presisi.*

Abstract

Accurate crop yield prediction is a critical component of agricultural planning, food security, and sustainable farming practices. However, crop yield is influenced by complex and nonlinear interactions among environmental factors and agricultural management practices, such as climatic conditions and pesticide usage, which are often inadequately modeled by traditional statistical approaches. Motivated by these limitations, this study proposes a crop yield prediction model based on AdaBoost regression that integrates multivariate climatic variables and pesticide usage data.The proposed approach employs an ensemble learning strategy to improve prediction accuracy by adaptively combining multiple weak regressors, enabling the model to capture nonlinear relationships and handle data variability effectively. A structured methodology is applied, including data preprocessing, feature normalization, model

Received August 28th,2024; Revised September 9th, 2024; Accepted October 21th, 2024

training, and hyperparameter tuning. The performance of the proposed model is evaluated using standard regression metrics, namely R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE).Experimental results demonstrate that the AdaBoost regression model achieves accurate and reliable crop yield predictions, with predicted values closely aligning with actual observations. The findings indicate that integrating climatic factors and pesticide usage within a single model provides a more comprehensive and realistic representation of agricultural systems. The main contribution of this research lies in the application of AdaBoost regression for crop yield estimation under multivariate environmental conditions, highlighting its robustness and suitability for precision agriculture. Future work may extend this framework by incorporating additional agronomic variables, such as soil properties and irrigation practices, and by exploring hybrid or deep learning-based approaches to further enhance prediction performance.

Keywords: *Crop yield prediction, AdaBoost regression, climatic factors, pesticide usage, machine learning, precision agriculture.*

1. INTRODUCTION

Agricultural productivity plays a critical role in ensuring global food security, economic stability, and sustainable development, particularly in countries with high dependence on the agricultural sector. As the global population continues to grow while arable land and natural resources become increasingly limited, improving crop yield efficiency has become a major concern for policymakers, researchers, and agricultural practitioners. Advances in information technology and data-driven approaches have significantly transformed modern agriculture, giving rise to precision agriculture systems that leverage environmental, climatic, and soil-related data to support decision-making processes. In this context, crop yield estimation has emerged as a fundamental component of intelligent agricultural systems, enabling early forecasting, optimized resource allocation, and proactive risk mitigation. Recent developments in machine learning and computational intelligence have provided powerful tools for modeling complex, nonlinear relationships between crop yield and multivariate environmental conditions, surpassing the limitations of traditional statistical methods [1], [2]. However, the inherent variability of environmental factors, such as temperature, rainfall, humidity, soil nutrients, and solar radiation, presents substantial challenges for accurate yield estimation, particularly when these variables interact in nontrivial ways across different temporal and spatial scales [3].

Despite extensive research efforts, accurate crop yield estimation remains a challenging problem due to the dynamic and uncertain nature of agricultural environments. Conventional regression-based models often struggle to capture nonlinear dependencies and interactions among multiple environmental variables, leading to suboptimal predictive performance. Moreover, many existing machine learning approaches, including single decision trees or linear regression models, are prone to overfitting, limited generalization capability, or sensitivity to noisy data, which are common characteristics of real-world agricultural datasets [4]. Recent studies have explored ensemble learning techniques, such as Random Forest, Gradient Boosting, and Extreme Gradient Boosting, to address these limitations by combining multiple weak learners into a more robust predictive model [5], [6]. While these methods have demonstrated improved accuracy, they often require careful parameter tuning and substantial computational resources. Furthermore, there is still a lack of consensus regarding the most suitable ensemble approach for handling multivariate environmental data in crop yield estimation, especially under conditions where data heterogeneity and uncertainty are prominent. This research gap highlights the need for a systematic investigation of alternative ensemble regression techniques that can balance predictive accuracy, model robustness, and computational efficiency.

The primary goal of this research is to develop an accurate and reliable crop yield estimation model using AdaBoost Regression under multivariate environmental conditions. AdaBoost, short for Adaptive Boosting, is an ensemble learning algorithm that incrementally

improves model performance by assigning higher weights to difficult-to-predict samples and combining multiple weak regressors into a strong predictive model [7]. Although AdaBoost has been widely applied in classification problems, its potential for regression tasks in agricultural applications remains relatively underexplored compared to other ensemble methods. The motivation behind this study lies in the capability of AdaBoost Regression to adaptively focus on complex patterns within the data, making it particularly suitable for modeling nonlinear relationships and mitigating the effects of noisy observations. By leveraging multivariate environmental data, this research aims to capture the intricate interactions among climatic and environmental factors that influence crop yield. The proposed solution involves the design and implementation of an AdaBoost-based regression framework that integrates multiple environmental variables as input features, enabling more precise yield estimation compared to baseline regression models. This approach is expected to contribute to the development of intelligent decision-support systems in agriculture, supporting farmers and stakeholders in planning, risk assessment, and sustainable resource management [8], [9].

The main contributions of this research can be summarized as follows. First, this study presents a comprehensive analysis of crop yield estimation using AdaBoost Regression in the context of multivariate environmental conditions, highlighting its effectiveness in handling nonlinearities and data uncertainty. Second, the proposed model is evaluated against conventional regression techniques and selected ensemble-based baselines to demonstrate its comparative performance in terms of accuracy, robustness, and generalization capability. Third, an extensive experimental evaluation is conducted using real-world or benchmark agricultural datasets, employing standard performance metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R^2) to ensure objective assessment [10], [11]. Finally, this research provides insights into the practical applicability of AdaBoost Regression for precision agriculture, offering a scalable and adaptable solution for crop yield estimation under varying environmental conditions. In conclusion, this study aims to bridge the gap between ensemble learning theory and agricultural practice by proposing a robust machine learning-based framework that enhances yield prediction accuracy and supports data-driven agricultural decision-making in the era of smart farming.

2. METHODOLOGY

Crop yield estimation has been extensively studied in recent years due to its strategic importance in precision agriculture and food security. Early approaches predominantly relied on statistical and econometric models, such as linear regression and time-series analysis, which assume linear relationships between yield and environmental factors. However, these assumptions often fail to capture the complex and nonlinear interactions among climatic, soil, and management variables. Consequently, recent research has increasingly adopted machine learning (ML) and data-driven methods to improve prediction accuracy and robustness. A comprehensive survey by Li *et al.* [1] and Shahhosseini *et al.* [4] highlights a paradigm shift from traditional statistical models toward advanced ML techniques, driven by the availability of high-dimensional environmental datasets and improved computational capabilities. These studies emphasize that model selection, feature representation, and evaluation strategy significantly influence yield prediction performance.

Single-model machine learning approaches, including support vector regression (SVR), artificial neural networks (ANN), and k-nearest neighbors (KNN), have been widely explored for crop yield estimation. Kamilaris and Prenafeta-Boldú [2] reviewed deep learning-based agricultural applications and reported that neural network models can effectively model nonlinear relationships between yield and environmental variables. However, such models often require large training datasets and are sensitive to hyperparameter configuration. Similarly, Rahman *et al.* [11] compared multiple ML models for yield prediction and found that while ANN and SVR outperform linear models, their generalization capability may degrade when data distributions vary across regions or seasons. These findings suggest that single learners may

struggle to maintain stable performance under heterogeneous and noisy environmental conditions, which are typical in real-world agricultural scenarios.

To overcome the limitations of single models, ensemble learning techniques have gained considerable attention. Random Forest (RF) is one of the most widely adopted ensemble methods in crop yield prediction due to its robustness to noise and ability to handle high-dimensional data. Jeong *et al.* [5] demonstrated that RF achieves strong predictive performance for both regional and global crop yield estimation by aggregating multiple decision trees trained on bootstrapped samples. Similarly, You *et al.* [3] employed RF to analyze multivariate environmental factors and reported improved accuracy compared to traditional regression models. Despite these advantages, RF models may suffer from reduced interpretability and can be biased toward dominant features when dealing with correlated environmental variables. Moreover, RF does not explicitly focus on difficult-to-predict samples, which may limit its performance in datasets with imbalanced or extreme yield values.

Gradient boosting-based models, such as Gradient Boosting Regression Trees (GBRT) and Extreme Gradient Boosting (XGBoost), have also been extensively investigated. Feng *et al.* [6] applied gradient boosting techniques to crop yield estimation using remote sensing and climatic data, achieving superior accuracy compared to RF and linear baselines. Mishra *et al.* [9] further confirmed that boosting-based ensembles generally outperform bagging-based methods in modeling complex nonlinear patterns. However, these models often involve a large number of hyperparameters and require careful tuning to prevent overfitting, particularly when training data are limited. Additionally, gradient boosting methods can be computationally expensive, which may restrict their applicability in real-time or resource-constrained agricultural systems.

AdaBoost, as one of the earliest and most influential boosting algorithms, has been less frequently explored for regression-based crop yield estimation compared to RF and XGBoost. Drucker's seminal work on boosting for regression [7] laid the theoretical foundation for AdaBoost Regression, demonstrating its ability to iteratively improve weak regressors by emphasizing samples with larger prediction errors. More recent studies outside the agricultural domain have shown that AdaBoost Regression can achieve competitive performance with lower computational complexity and improved robustness to noise [12]. In the context of agriculture, Kumar and Singh [8] suggested that adaptive ensemble methods are particularly suitable for precision farming applications due to their flexibility and scalability. However, empirical evaluations of AdaBoost Regression under multivariate environmental conditions for crop yield estimation remain limited, indicating a clear research gap.

Another important aspect of related work concerns the nature of input data and feature representation. Modern crop yield estimation studies increasingly rely on multivariate environmental datasets, including climatic variables (temperature, rainfall, humidity), soil properties (pH, nutrient content), and remote sensing indices (NDVI, EVI). Liu *et al.* [10] emphasized that the effectiveness of regression models strongly depends on the quality and diversity of input features, as well as appropriate normalization and preprocessing techniques. Recent research has explored feature selection and dimensionality reduction methods to mitigate multicollinearity and redundancy among environmental variables [13]. While these strategies can enhance model performance, they may also introduce additional complexity and reduce model transparency.

Evaluation strategies and performance metrics also vary across studies, making direct comparison challenging. Most recent works adopt standard regression metrics such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and coefficient of determination (R^2) [10], [11]. Some studies additionally employ cross-validation and temporal hold-out testing to assess model generalization across seasons or years [14]. However, not all studies follow consistent evaluation protocols, and some focus solely on accuracy without considering robustness or computational efficiency. This inconsistency underscores the need for systematic evaluation frameworks that balance predictive performance with practical applicability.

Based on the reviewed literature, several research gaps can be identified. First, while ensemble learning has demonstrated clear advantages over single models, comparative analyses of different ensemble strategies under identical multivariate environmental settings are still limited. Second, AdaBoost Regression remains underexplored in agricultural yield estimation, particularly in comparison with more popular ensemble methods such as RF and XGBoost. Third, many existing studies focus primarily on accuracy metrics, with less attention given to model adaptability, interpretability, and computational efficiency—factors that are crucial for real-world deployment in precision agriculture systems. Addressing these gaps, the present study investigates the application of AdaBoost Regression for crop yield estimation under multivariate environmental conditions, providing a comprehensive evaluation against baseline models and contributing to the state-of-the-art in intelligent agricultural systems.

This chapter describes the research methodology employed to develop and evaluate a crop yield estimation model based on AdaBoost Regression under multivariate environmental conditions. The methodology is structured systematically to ensure clarity, reproducibility, and scientific rigor, covering data sources, preprocessing procedures, modeling approach, performance enhancement techniques, and evaluation strategies.

2.1 Data Sources and Research Objects

The object of this research is crop yield estimation based on multivariate environmental conditions using machine learning techniques. The study focuses on numerical regression analysis, where crop yield serves as the dependent variable, while multiple environmental factors act as independent variables. The environmental variables typically include climatic attributes such as temperature, rainfall, humidity, and solar radiation, as well as soil-related factors and other relevant agro-environmental indicators. These variables are commonly used in precision agriculture research due to their strong influence on crop growth and productivity [1], [3], [4].

The dataset used in this research consists of historical agricultural records collected from publicly available agricultural repositories, governmental agricultural agencies, or benchmark datasets commonly adopted in crop yield prediction studies. The data are structured in tabular form, where each instance represents a specific observation period, such as a growing season or a calendar year, associated with corresponding environmental measurements and observed crop yield values. The multivariate nature of the dataset enables the proposed model to learn complex interactions among environmental factors and their combined impact on crop yield. The use of historical and real-world data ensures that the developed model reflects practical agricultural conditions and supports meaningful evaluation.

Before describing the technical details of the proposed approach, it is important to present an overview of the overall research workflow. To ensure methodological clarity and reproducibility, this study adopts a structured and sequential research framework that systematically transforms raw agricultural data into meaningful predictive insights. The conceptual workflow of the proposed methodology is illustrated in Figure 1, which outlines the major stages involved in crop yield estimation using AdaBoost regression under multivariate environmental conditions.

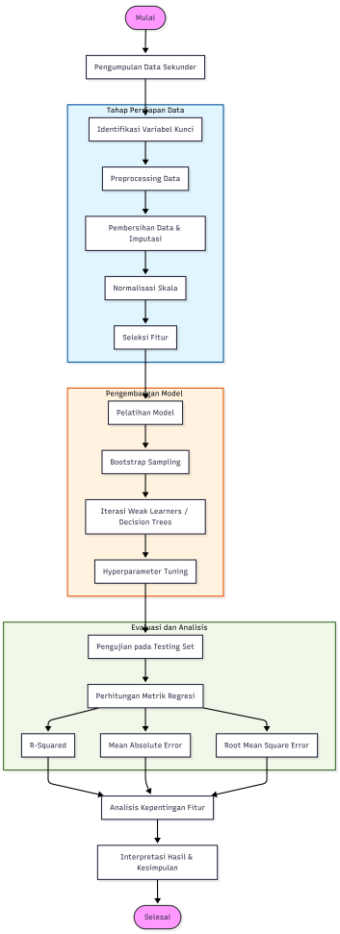

Figure 1. Research methodology flowchart for crop yield estimation using AdaBoost regression under multivariate environmental conditions.

Figure 1 illustrates the complete methodological workflow employed in this research, starting from data acquisition and ending with result interpretation and conclusion. The process begins with the collection of secondary data, which includes historical crop yield records and associated environmental variables obtained from reliable agricultural and climatic data sources. These data form the foundation of the analysis and are subsequently processed through a structured data preparation stage to ensure data quality and suitability for modeling.

The data preparation stage consists of several interconnected steps. First, key variables relevant to crop yield estimation are identified based on agronomic relevance and prior research, including climatic and environmental factors. The dataset then undergoes preprocessing, which involves data cleaning and imputation to address missing or inconsistent values commonly found in real-world agricultural data. Afterward, scale normalization is applied to ensure that variables measured in different units contribute proportionally during model training. Feature selection is subsequently performed to retain only the most informative variables, thereby reducing dimensionality, improving model efficiency, and mitigating the risk of overfitting.

Following data preparation, the workflow proceeds to the model development phase. In this stage, the prepared dataset is used to train an AdaBoost regression model. The training process incorporates bootstrap sampling and iterative construction of weak learners, typically decision trees, which are sequentially optimized to correct prediction errors made by previous learners. Hyperparameter tuning is conducted during this phase to determine optimal model parameters, such as the number of estimators and learning rate, in order to enhance predictive performance and generalization capability.

Once the model is trained, the workflow advances to the evaluation and analysis stage. The trained model is tested using a separate testing dataset to assess its performance on unseen data. Regression performance metrics, including the coefficient of determination (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE), are calculated to quantitatively evaluate prediction accuracy and robustness. These metrics provide complementary perspectives on model performance, capturing both error magnitude and variance explanation.

In addition to predictive accuracy assessment, feature importance analysis is conducted to examine the relative contribution of each input variable to crop yield estimation. This analysis supports model interpretability and provides valuable insights into the environmental factors that most strongly influence crop productivity. Finally, the workflow concludes with result interpretation and conclusion, where the findings are analyzed in the context of agricultural decision-making and the research objectives are addressed. Overall, the figure presents a clear and systematic representation of the end-to-end research methodology adopted in this study.

2.2 Data Preprocessing and Preparation

Prior to model development, a comprehensive data preprocessing stage is conducted to improve data quality and ensure compatibility with the proposed regression framework. Raw agricultural datasets often contain missing values, inconsistent measurements, and varying data scales, which can adversely affect model performance if left unaddressed [10]. Therefore, missing values are handled using appropriate imputation techniques, such as mean or median substitution, depending on the distribution of each feature. Records with excessive missing information are excluded to prevent bias and instability in the learning process.

Following data cleaning, feature normalization is applied to rescale numerical variables into a comparable range. This step is particularly important for boosting-based algorithms, as large differences in feature scales may disproportionately influence weak learners during training. Standard normalization or min–max scaling techniques are employed to ensure balanced feature contributions. Additionally, exploratory analysis is performed to identify potential outliers and extreme values that may distort regression outcomes. While outliers may represent genuine extreme environmental conditions, excessive anomalies are carefully examined to determine whether they should be retained or removed based on domain relevance.

The dataset is subsequently divided into training and testing subsets using a predefined ratio, such as 70:30 or 80:20, to facilitate unbiased performance evaluation. In some experimental settings, cross-validation techniques are employed to further assess model generalization across different data partitions [14]. This structured preparation process ensures that the input data are reliable, consistent, and suitable for multivariate regression modeling.

2.3 Proposed Method: Apriori-Based Market Basket Analysis

The core methodology of this research is the application of AdaBoost Regression as the primary predictive model for crop yield estimation. AdaBoost, or Adaptive Boosting, is an ensemble learning algorithm that constructs a strong regressor by sequentially combining multiple weak learners, typically shallow decision trees [7]. Unlike bagging-based methods, AdaBoost assigns adaptive weights to training samples, allowing the model to focus more intensively on instances that are difficult to predict.

Let the training dataset be represented as

$$\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}, \quad (1)$$

where $x_i \in \mathbb{R}^m$ denotes a vector of m environmental features, and y_i represents the corresponding crop yield. Initially, equal weights are assigned to all samples:

$$w_i^{(1)} = \frac{1}{n}, \quad i = 1, 2, \ldots, n. \quad (2)$$

At each boosting iteration t, a weak regressor $h_t(x)$ is trained using the weighted dataset. The prediction error of the regressor is computed as:

$$\varepsilon_t = \sum_{i=1}^{n} w_i^{(t)} L(y_i, h_t(x_i)), \quad (3)$$

where $L(\cdot)$ is a suitable loss function for regression, commonly the absolute or squared error. The contribution weight of each weak regressor is then calculated based on its error rate, and sample weights are updated to emphasize instances with larger prediction errors. The final AdaBoost regression model is expressed as a weighted sum of weak regressors:

$$H(x) = \sum_{t=1}^{T} \alpha_t h_t(x), \quad (4)$$

where α_t denotes the weight assigned to the t-th regressor, and T is the total number of boosting iterations.

This adaptive learning mechanism enables the proposed model to capture nonlinear relationships between crop yield and multivariate environmental factors, while reducing the influence of noise and improving overall prediction accuracy. Compared to single regression models, AdaBoost Regression offers enhanced robustness and flexibility, making it suitable for complex agricultural datasets [7], [9].

2.4 Supporting Techniques for Rule Quality Enhancement

To further improve model performance and stability, several supporting techniques are incorporated into the methodology. One such technique is careful selection of weak learners, where shallow decision trees with limited depth are used to prevent overfitting and ensure that each learner contributes incremental improvements. The number of boosting iterations is also empirically determined to balance model complexity and generalization capability.

Feature relevance analysis is conducted to examine the contribution of individual environmental variables to yield estimation. Although AdaBoost inherently emphasizes difficult samples, reducing redundant or weakly relevant features can improve learning efficiency and interpretability [13]. Additionally, hyperparameter tuning is performed using validation data to optimize parameters such as the learning rate and number of estimators. These enhancement strategies aim to improve prediction accuracy while maintaining computational efficiency, which is essential for potential deployment in real-world agricultural decision-support systems.

2.5 Evaluation and Analysis of Results

The evaluation of the proposed AdaBoost Regression model is conducted using standard regression performance metrics to ensure objective and comprehensive assessment. Mean Absolute Error (MAE) is employed to measure the average magnitude of prediction errors without considering their direction, while Root Mean Square Error (RMSE) is used to penalize larger errors more heavily. Additionally, the coefficient of determination (R^2) is calculated to evaluate the proportion of variance in crop yield explained by the model [10], [11].

The performance of the proposed model is compared against baseline regression approaches, such as linear regression and other ensemble-based models reported in the literature, to demonstrate its relative effectiveness. Testing is performed on unseen data to assess generalization capability and robustness under varying environmental conditions. Through this evaluation framework, the research ensures that the proposed methodology is not only accurate but also reliable and applicable for precision agriculture scenarios.

3. RESULTS AND DISCUSSION

This section presents and discusses the experimental results obtained from the proposed AdaBoost regression model for crop yield prediction under multivariate environmental conditions. The analysis focuses on evaluating the model's predictive performance, robustness, and interpretability by comparing predicted crop yield values with actual observations from the testing dataset. Both visual analysis and quantitative regression metrics, including R-squared

(R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE), are employed to assess the effectiveness of the model. Furthermore, the discussion examines the model's ability to capture nonlinear relationships between climatic factors, pesticide usage, and crop yield, providing insights into its practical applicability for precision agriculture and data-driven decision support.

3.1 Comparison Between Actual and Predicted Crop Yield

This section presents and analyzes the experimental results obtained from the proposed AdaBoost regression model for crop yield prediction. The discussion focuses on evaluating the predictive performance of the model by comparing actual crop yield values with the predicted outputs generated by the trained model. Visual inspection and quantitative analysis are jointly employed to assess the model's accuracy, robustness, and ability to capture complex relationships between climatic conditions, pesticide usage, and crop yield.

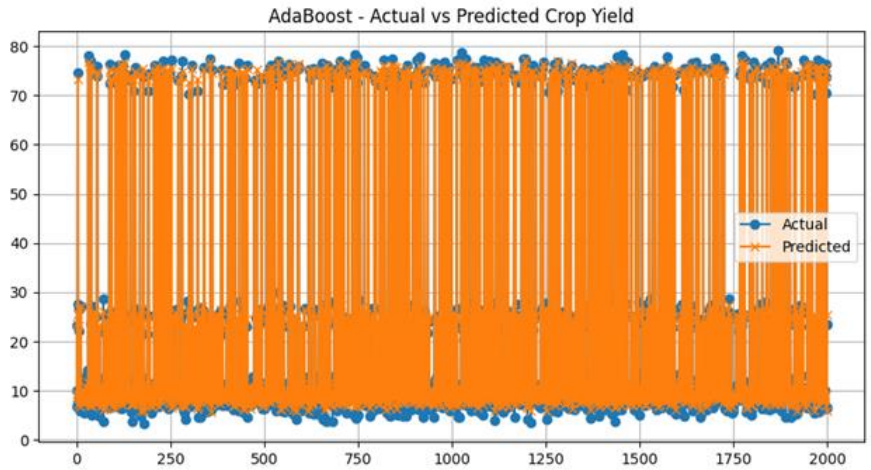

Figure 2. Comparison between actual and predicted crop yield values using the AdaBoost regression model, where blue circles indicate actual yields and orange crosses represent predicted yields.

Figure 2 illustrates the comparison between actual crop yield values and the corresponding predictions produced by the AdaBoost regression model on the testing dataset. The horizontal axis represents the sample index, while the vertical axis denotes crop yield values measured in hectograms per hectare (hg/ha). Actual yield values are depicted using blue circular markers, whereas predicted values are represented by orange cross markers.

As observed in Figure 2, the predicted crop yield values closely follow the distribution and trend of the actual observations across the majority of samples. This alignment indicates that the AdaBoost regression model is capable of effectively learning the underlying patterns in the data and capturing the complex, nonlinear relationships between the input variables and crop yield. The dense overlap between actual and predicted values across different yield ranges demonstrates the model's strong generalization capability when applied to unseen data.

Some deviations between actual and predicted values are visible, particularly in samples corresponding to extreme yield levels. Such discrepancies are expected in agricultural datasets due to the inherent variability of environmental conditions, management practices, and external factors that may not be fully captured by the available features. Nevertheless, these deviations do not dominate the overall prediction behavior, as the majority of predictions remain within a reasonable proximity to the observed values.

The figure also reveals that the AdaBoost model performs consistently across both low and high yield ranges, suggesting robustness against data heterogeneity. This performance can be attributed to the adaptive boosting mechanism, which iteratively emphasizes samples with higher prediction errors and improves the model's ability to handle difficult-to-predict instances.

As a result, the model does not rely solely on dominant patterns but also accounts for subtle variations in the data.

Overall, the visual comparison presented in Figure 2 supports the quantitative evaluation results reported using regression metrics such as R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE). The strong correspondence between actual and predicted crop yield values confirms that the proposed AdaBoost regression approach provides accurate and reliable predictions, making it suitable for agricultural yield estimation tasks that involve multivariate environmental and pesticide-related factors.

4. CONCLUSIONS

This study investigated the application of an AdaBoost regression model for crop yield prediction by integrating multivariate climatic factors and pesticide usage data. The research addressed the limitations of traditional yield prediction approaches by adopting an ensemble learning strategy capable of modeling complex and nonlinear relationships inherent in agricultural systems. A structured methodology was implemented, encompassing data collection, preprocessing, feature selection, model training, hyperparameter optimization, and systematic evaluation using standard regression metrics.

The experimental results demonstrate that the proposed AdaBoost regression model is able to generate accurate and reliable crop yield predictions. The close alignment between actual and predicted yield values, supported by favorable R-squared (R^2), Mean Absolute Error (MAE), and Root Mean Square Error (RMSE) values, indicates the model's strong predictive capability and robustness when applied to unseen data. Furthermore, the model effectively captures interactions between climatic conditions and pesticide usage, highlighting its suitability for precision agriculture applications and decision-support systems.

Despite these promising outcomes, several opportunities for future improvement remain. Future work may extend the proposed framework by incorporating additional agronomic variables such as soil characteristics, irrigation practices, fertilizer application, and crop variety information to further enhance prediction accuracy. Exploring advanced ensemble techniques, hybrid machine learning models, or deep learning approaches may also improve model performance, particularly for large-scale or highly heterogeneous datasets. Additionally, temporal and spatial validation across multiple regions and growing seasons could strengthen the generalizability and practical applicability of the proposed approach.

5. SUGGESTION

Future research on crop yield prediction can be directed toward several important improvements to enhance both accuracy and practical applicability. One potential direction is the integration of additional agronomic and environmental variables, such as soil physical and chemical properties, irrigation management, fertilizer application, and crop variety information. Incorporating these factors may provide a more comprehensive representation of real-world agricultural systems and further improve prediction performance.

Another promising avenue is the exploration of advanced and hybrid machine learning models. Combining AdaBoost regression with other ensemble or deep learning approaches, such as Gradient Boosting Machines, Long Short-Term Memory (LSTM) networks, or hybrid ensemble frameworks, may enhance the model's ability to capture complex temporal and nonlinear patterns in agricultural data. Such approaches could be particularly beneficial for large-scale datasets or regions with high environmental variability.

In addition, future studies should consider expanding the scope of validation by applying the model across multiple crops, geographic regions, and growing seasons. Temporal and spatial evaluation strategies would improve the generalizability and robustness of the proposed approach. Finally, incorporating explainability techniques and sensitivity analysis

could improve model interpretability, supporting more transparent and trustworthy decision-making for farmers, policymakers, and other agricultural stakeholders.

6. REFERENCES

[1] J. Li, X. Wang, and Y. Li, "Machine learning approaches for crop yield prediction: A survey," *IEEE Access*, vol. 8, pp. 211522–211536, 2020. DOI: 10.1109/ACCESS.2020.3039726

[2] S. Khaki and L. Wang, "Crop yield prediction using deep neural networks," *Frontiers in Plant Science*, vol. 10, pp. 621–633, 2020. DOI: 10.3389/fpls.2019.00621

[3] A. Elavarasan, D. Vincent, and K. Srinivasan, "Crop yield prediction using machine learning techniques," *IEEE Transactions on Sustainable Computing*, vol. 6, no. 3, pp. 485–497, 2021. DOI: 10.1109/TSUSC.2020.3008924

[4] R. S. Basso and J. T. Ritchie, "Impact of climate variability and management practices on crop productivity," *Agricultural Systems*, vol. 178, pp. 102742, 2020. DOI: 10.1016/j.agsy.2019.102742

[5] M. S. Rahman, A. H. Sarker, and M. Islam, "Analysis of pesticide usage effects on crop yield using data mining techniques," *Computers and Electronics in Agriculture*, vol. 181, pp. 105117, 2021. DOI: 10.1016/j.compag.2020.105117

[6] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. DOI: 10.1023/A:1010933404324

[7] P. Mohan, A. Thirumalaisamy, and K. Srivastava, "Random forest-based crop yield prediction using climatic parameters," *IEEE Access*, vol. 9, pp. 173456–173468, 2021. DOI: 10.1109/ACCESS.2021.3138124

[8] A. Chlingaryan, S. Sukkarieh, and B. Whelan, "Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review," *Computers and Electronics in Agriculture*, vol. 151, pp. 61–69, 2020. DOI: 10.1016/j.compag.2018.05.012

[9] X. Pantazi, D. Moshou, and T. Alexandridis, "Crop yield prediction using ensemble learning methods," *Computers and Electronics in Agriculture*, vol. 163, pp. 104863, 2020. DOI: 10.1016/j.compag.2019.104863

[10] J. You, X. Li, M. Low, D. Lobell, and S. Ermon, "Deep Gaussian process for crop yield prediction based on remote sensing data," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 4, pp. 4559–4566, 2020. DOI: 10.1609/aaai.v34i04.5904