Research on World Agricultural Economy
https://journals.nasspublishing.com/index.php/rwae

RESEARCH ARTICLE

Comparative Analysis of Machine Learning Models for Predicting Rice Yield: Insights from Agricultural Inputs and Practices in Rwanda

Cyprien Mugemangango [1*] , ***Joseph Nzabanita*** [2], ***Dieudonne Ndaruhuye Muhoza*** [3], ***Nathan D. Cahill*** [4]

[1] *African Center of Excellence in Data Science (ACE-DS), University of Rwanda, Kigali P.O. Box 3900, Rwanda*
[2] *Joseph Nzabanita, College of Science and Technology, University of Rwanda, Kigali P.O. Box 3900, Rwanda*
[3] *College of Business and Economics, University of Rwanda, Kigali P.O. Box 3900, Rwanda*
[4] *College of Science, Rochester Institute of Technology, Rochester, NY 14623, USA*

ABSTRACT

Food security is a global challenge, especially in developing countries like Rwanda. With a growing population and limited agricultural land, Rwanda struggles to meet increasing food demands. Rice is a staple food crop in Rwanda, playing a crucial role in the country's food security. However, factors like climate variability, soil nutrient management, and limited access to high-quality inputs hinder rice yield optimization. This paper investigates the most effective machine learning model for predicting rice crop yield in Rwanda, using agricultural inputs and practices. The study used secondary datasets from the National Institute of Statistics of Rwanda (NISR) for rice yield prediction. Eight supervised machine learning algorithms were used, including Linear Regression, Random Forest, Gradient Boosting, Support Vector Machine, Artificial Neural Network, eXtreme Gradient Boosting Tree, and AdaBoost. The models were evaluated based on their accuracy in predicting rice yields, with RMSE, MAE, and Relative Error as primary metrics. The feature importance analysis was also conducted to identify significant factors influencing yield predictions. The study's findings revealed that the Adaptive Boosting Tree model outperformed

*CORRESPONDING AUTHOR:
Cyprien Mugemangango, African Center of Excellence in Data Science (ACE-DS), University of Rwanda, Kigali P.O. Box 3900, Rwanda; Email: mucypro2@yahoo.fr

ARTICLE INFO
Received: 17 August 2024 | Revised: 18 September 2024 | Accepted: 23 September 2024 | Published Online: 15 November 2024
DOI: https://doi.org/10.36956/rwae.v5i4.1247

CITATION
Mugemangango, C., Nzabanita, J., Muhoza, D.N., et al., 2024. Comparative Analysis of Machine Learning Models for Predicting Rice Yield: Insights from Agricultural Inputs and Practices in Rwanda. Research on World Agricultural Economy. 5(4): 350–366.
DOI: https://doi.org/10.36956/rwae.v5i4.1247

the other machine learning models in predicting rice yield. This model achieved RMSE, MAE and Relative Error of 0.69, 0.46 and 12.4%, respectively, indicating a high level of predictive accuracy. The feature importance analysis further highlighted the key factors that contributed to rice yield predictions, with the quantity of inorganic fertilizer, degree of erosion, season, and seed type emerging as the most influential variables. The study demonstrates the effectiveness of machine learning models, particularly the Adaptive Boosting Tree, in improving rice yield predictions and highlighting the crucial role of agricultural inputs like fertilizer and seed type, in influencing crop yields. The output from this study will help the farmers and stakeholders to make data-driven decisions about resource use and crop management.

Keywords: Agricultural Inputs; Agricultural Practices; Machine Learning; Artificial Intelligence; Precision Farming

1. Introduction

Agriculture is the cornerstone of human civilization, providing not only the food necessary for human nutrition but also a vital source of income across the globe. In developing countries, agriculture plays an even more critical role, being a key driver of national income and employment. In Rwanda, agriculture forms the backbone of the economy, engaging approximately 62% of the population, the majority of whom are smallholder farmers in rural areas. This sector is crucial for both national development and poverty reduction. As of 2023, the agricultural sector in Rwanda contributed 26% to the country's GDP and accounted for 0.7% of the nation's economic growth, underscoring its significance [1]. Additionally, agriculture is a major source of foreign exchange for Rwanda, contributing 37% to the national export value in the 2021–2022 fiscal year [2]. These statistics highlight the vital role of agriculture in Rwanda's economy and the need for ongoing improvements to enhance productivity and sustainability. Increasing agricultural output, therefore, is not just a necessity for ensuring food security but also for driving economic growth, improving livelihoods, and strengthening the national economy.

The Government of Rwanda has made tremendous efforts to increase agricultural productivity toward achieving food security and poverty reduction. It has formulated a coherent strategy for agriculture such as the Strategic Plan for the Transformation of Agriculture (SPTA) in Rwanda. Phase III of the plan (SPTA3) issued in 2013 covered the five years (2013–2017), and phase IV (PSTA4) issued in 2018 covered the five years 2018–2022 in response to the need for an updated strategy for agriculture [3]. One of the goals of the National Strategy for Transformation (NST1), which was established in 2017 to promote transformation throughout the country, was to modernize and increase the productivity of agriculture and livestock by maintaining the agriculture sector's relatively stable average annual growth rate of 5.7% between 2017 and 2024. The PSTA4 predicted that average annual growth in agriculture would reach 10% through 2023 and that the percentage of families facing food insecurity would decrease to 10% by 2023–2024. This plan seeks to facilitate the development of Rwanda's agriculture, through an approach based on resource management, human capacity, and private sector-driven value chains. The agricultural policy in Rwanda continues to promote agriculture intensification to increase productivity, value addition, modernization, and improved quality of livestock to achieve an average annual growth rate of 8.5%. It is in this scope that the government of Rwanda has set up the Crop Intensification Program (CIP) and more recently a Livestock Intensification Program (LIP). The CIP and the Land Use Consolidation Program among others were prioritized programs focusing on specialized crops with a target of increasing agricultural production and food security in Rwanda. However, the lack of enough arable land for agriculture and the population's increase pose a serious challenge to the Government of Rwanda's ability to ensure food security [4].

The 2021 Comprehensive Food Security and Vulnerability Analysis (CFSVA) results indicated that 20.6

percent of Rwandans were generally food insecure, with 1.8 percent being classified as highly food insecure and 18.8 percent classified as having moderate food insecurity. The proportion of stunted children under five years of age in Rwanda decreased significantly from 34.9 percent in 2018 to 32.4 percent in 2021. Of them, 8.4% had severe stunting and 24.0% had moderate stunting. Around 2.4% of children under five suffer from acute malnutrition, also known as wasting. Of these, 1.8% experience moderate acute malnutrition, and 0.6% experience severe acute malnutrition. Between 2018 and 2021, the prevalence of acute malnutrition increased by 0.4 percent to 2.0 percent [5]. The statistics above draw attention to the issue of the lack of sufficient food, which exacerbates ailments associated with malnutrition.

To tackle the pressing issue of food insecurity, enhancing crop yield is of paramount importance. The first critical step in this process is to identify and thoroughly understand the factors that influence crop productivity. These factors are generally categorized into three main groups: technological, biological, and environmental. Technological factors encompass agricultural practices and managerial decisions that directly impact farming outcomes. Biological factors include challenges such as diseases, insects, pests, and weeds that can severely affect crop health. Environmental factors are broad and encompass climatic conditions, soil fertility, topography, and water quality, all of which contribute to the variations in yield observed across different regions of the world [6]. Understanding these multifaceted influences is key to improving crop yields.

In the context of Rwanda, there are different factors that might influence the crop productivity. These factors include agricultural inputs and practices, such as the use of pesticides, the use of inorganic fertilizer, the application of anti-erosion practices, the choice of seed type to be sown, the application of irrigation practices, the type of farmers which affect the quantity and quality of inputs and practices, and the agricultural season, which is subject to change due to variability in climate conditions. Positive or negative effects on agricultural output can result from farmers applying these elements insufficiently or not at all [7]. Helping farmers and other stakeholders understand how to best utilize these inputs to increase agricultural output is therefore essential.

In this context, Machine Learning (ML) has emerged as a promising tool, providing the means to analyze and model these complex interactions, thereby enabling management strategies for enhancing agricultural productivity and better prediction of crop disease, crop yield, weed, and crop health [8]. The following section explores the past research studies incorporating machine learning techniques and reveals their potential gaps and the needs for improvement.

In their study assessing rice yield prediction in Nigeria, Jiya, Illiyasu and Akinyemi [9] examined the effects of climate change on agricultural output by focusing on rice production in Katsina state from 1970 to 2017. The research employed many machine learning models, including Naïve Bayes, Random Forest, Artificial Neural Networks, and Logistic Regression, using climate data from the World Bank Climate Knowledge portal and rice yield data from the Nigeria Bureau of Statistics. The outcomes showed that Random Forest fared better in terms of classification accuracy than the other models. Although the study offers insightful analysis and practical models for yield prediction, it falls short in addressing the impact of non-climatic variables such as soil nutrients and agricultural practices. To improve the precision and usefulness of rice yield prediction models, future studies may find it advantageous to incorporate these extra characteristics.

Li et al. [10] examined the impacts of temperature, solar radiation, and precipitation on rice yields. The study used local polynomial regression (Loess). For quantitative analysis, a linear mixed-effects model was utilized. The results showed that higher average temperatures and precipitation decreases of more than 25% had a negative impact on rice yield, although higher CO_2 concentrations and good management techniques were able to offset these negative effects. A 1 °C increase in average temperature resulted in a 3.85% loss in rice output, but a 100 ppm increase in CO_2 caused a 7.1% increase in yield. The study also showed that there is a lot of variation in rice production projections because of variations in climate models, study sites, and scenario types. While providing extensive insights, the research is constrained by errors in climate models and the omis-

sion of important agricultural elements like soil health and farming practices. In order to close these gaps in knowledge and inform the creation of sustainable agricultural policies, future study could include more variables in order to enhance the precision of rice production forecasts.

Zhou, Xu and Chen [11] conducted a study in Hubei Province of China, to compare three deep learning models: CNN-LSTM, CNN, and ConvLSTM for predicting the yearly rice yield at the county level. In order to account for regional variability, a dummy variable was included in the training process along with ERA5 temperature data and MODIS remote sensing variables such the Soil-Adapted Vegetation Index (SAVI), Gross Primary Productivity (GPP), and Enhanced Vegetation Index (EVI). The labels were rice yield data from 2000 to 2019. Deep learning models were used to train and forecast utilizing remote sensing photos converted into normalized histograms. The results showed that compared to models that only used data from remote sensing, the inclusion of the spatial heterogeneity variable increased predicted accuracy. When it came to prediction performance, CNN-LSTM outperformed both CNN and ConvLSTM among the models that were tested. The study successfully predicted rice output using sophisticated models. However, to improve model performance and robustness, future research might look at adding additional variables such soil characteristics, agricultural inputs and practices, and climate variability.

Satpathi et al. [12] used historical rice yield and meteorological data from three districts namely Raipur, Surguja and Bastar in Chhattisgarh, India, over a 21-year period, to compare five models: Stepwise Multiple Linear Regression (SMLR), Artificial Neural Network (ANN), Least Absolute Shrinkage and Selection Operator (LASSO), Elastic Net (ELNET), and Ridge Regression. Twenty percent of the dataset were used to validate the models while 80% of the data were used for training process. Additionally, ensemble models were built using ELNET, Cubist, Random Forest, and Generalized Linear Model (GLM) techniques. With nearly perfect prediction accuracy (R^2 = 1 for Raipur and R^2 = 0.99 for Surguja), the study results showed that ANN performed best in these two districts. The Random Forest model produced the best results for Bastar (R^2 = 0.85 for training and 0.81 for validation), with ensemble models performing better than individual models. Even though the study produced reliable predictions, it would be advantageous to incorporate more variables, such soil data, agricultural inputs and practices to increase the accuracy of forecasts over a range of agro-climatic zones.

Elbasi et al. [13] explored the prospective advantages of incorporating ML algorithms into contemporary agriculture. The algorithms primarily aim to enhance crop production efficiency and minimize waste by facilitating informed decisions related to crop planting, irrigation, and harvesting. In their research, Nigam et al. [14] focused on using various machine learning algorithms to predict crop yield based on temperature, rainfall, season, and area. It was discovered that the Random Forest Regressor outperformed other ML algorithms in terms of mean absolute error.

P. S. and R. [15] proposed ML algorithms namely Artificial Neural Network, Support Vector Regression, K-Nearest Neighbor, and Random Forest (RF) to evaluate the most needed features for accurate crop yield prediction. The mean square error was used to evaluate the performance of these models. The authors discovered that the Random Forest algorithm obtained the highest accuracy while using the same training agricultural data and a variety of feature subsets.

Kang et al. [16] used six ML algorithms namely Lasso, Support Vector Regressor, Random Forest, XGBoost, Long-Short term memory (LSTM), and Convolutional Neural Network (CNN), and an extensive set of environmental variables derived from satellite observations, weather data, land surface model results, soil maps, and crop progress reports for maize yield prediction in the US Midwest. They found out that the XGBoost algorithm outperforms other algorithms both in accuracy and Stability, while deep neural networks such as LSTM and CNN were not advantageous.

In their study, Kuradusenge et al. [17] used three ML models (Random forest, polynomial regression, and support vector regression) to predict the Irish potato and maize harvests in Rwanda based on rainfall and temperature. The best model was shown to be Random Forest, with root mean square errors for Irish potato and Maize

datasets of 510.8 and 129.9, respectively, and R-square values of 0.875 and 0.817.

In their study, Kumar et al. [18] recommended a technique based on the K-Nearest Neighbors (KNN) algorithm to assess the soil's quality and predict the ideal crop to grow. Temperature and soil quality were considered as inputs to their algorithm. The proposed method suggested the fertilizer based on the crop predicted. The test results showed that the technique accurately predicts crop selection and production. Panigrahi, Kathala and Sujatha [19] fitted an ML model to forecast farm productivity. Through the utilization of supervised learning, they gathered and trained data using six diverse regression models, including Linear Regression, Gradient Boosting Regression, Random Forest Regression, XGboost Regression, and Voting Regression. Notably, the Random Forest Regression demonstrated superior performance, achieving a cross-validation test score of 0.6087 and exhibiting a Mean Absolute Error (MAE) of 468.16, outshining the other models in the comparison. As technology advances, it is expected that ML will play a critical role in solving problems related to agricultural productivity.

Although different machine learning algorithms have been investigated in past studies for predicting crop yield, most of these studies have primarily focused on soil and weather parameters as the main predictors [9–12]. While these factors are indeed crucial, they are often beyond the immediate control of farmers. In contrast, agricultural inputs and practices, such as the use of fertilizers, irrigation techniques, pest management, and crop rotation, are aspects that farmers can directly influence. Despite their significance, few studies have incorporated these inputs and practices as key predictors in crop yield models. Farming inputs and practices play a pivotal role in optimizing agricultural productivity, and their impact can be as significant, if not more so, than soil and weather conditions. Ignoring these factors in predictive models can lead to incomplete or less accurate predictions, which may not fully capture the potential for yield optimization. Therefore, it is essential to conduct further research to evaluate the contribution of these controllable factors to crop yield prediction.

This research aligns with the ongoing efforts to identify the most effective machine learning model for predicting rice crop yield, particularly by emphasizing the importance of agricultural inputs and practices. By integrating these factors into the model, this study aims to provide a more comprehensive understanding of the elements that most significantly influence crop yield. The research focuses on rice, a priority crop within Rwanda's Crop Intensification Program (CIP). The machine learning approach was selected due to its high precision in agricultural crop yield prediction using information gained from historical huge amount of data [9–11].

The findings from this study could offer valuable insights into the improvement of rice yield predictions. Enhancing rice yield prediction can have a big impact on resource management and agricultural productivity because it provides accurate estimates that help decision-makers. The following are some possible effects: (1) Farmers may maximize their usage of inputs like water, fertilizer, and insecticides by using accurate forecasts. They can prevent needless misuse or underuse of resources with a better understanding of future yields, which will save money and promote more ecologically responsible practices. (2) Farmers can modify crop management plans in response to expected variations in weather or soil conditions by using enhanced yield predictions. This allows for timely interventions and modifications during the growing season, which can result in healthier crops and higher yields. (3) Accurate projections offer vital data for organizing food supply networks, assisting in the avoidance of food shortages and price fluctuations. Anticipating output levels helps governments and organizations better manage food distribution and storage, especially in areas where climatic uncertainty is a concern. (4) Precise yield forecasts can assist policymakers in making strategic choices about the distribution of funds and investments in agricultural infrastructure. This also influences agricultural policies by boosting output and assisting farmers in regions where crop yields are anticipated to be difficult.

2. Materials and Methods

2.1. Data Collection

In this study, secondary data on rice crops were obtained from the National Institute of Statistics of Rwanda (NISR) through official datasets from the Seasonal Agricultural Surveys (SAS) conducted from 2020 Season A to 2022 Season B. These surveys provided valuable insights into the yield distribution of rice across different seasons, as illustrated in **Figure 1**, which displays the season-wise rice yield distribution. Rice was selected as the crop of interest due to its significant role in Rwanda's Crop Intensification Program (CIP), which aims to increase agricultural productivity. Rice is one of the most commonly consumed staple foods in Rwanda, contributing significantly to the national food security and economy.

The SAS is a comprehensive agricultural survey conducted regularly in each agricultural season. The primary objective of the survey is to produce reliable, area-based agricultural indicators, including crop area, crop production, and the application of inputs and agricultural practices. The survey covers the entire country and is conducted on sampled plots distributed across all 30 districts of Rwanda, ensuring a representative dataset for agricultural analysis. In this particular study, rice data were collected from 842 sampled plots, providing a robust dataset for analysis.

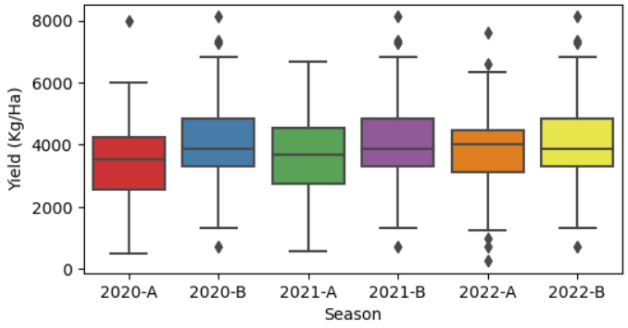

Figure 1. Yield Distribution Per Season.

2.2. Data Cleaning

After collection, the data underwent a meticulous examination and cleansing process to ensure its quality and reliability. This involved a comprehensive review to identify and remove irrelevant entries that did not contribute to the model's objectives. Erroneous data points, which could skew the results, were also carefully identified and corrected or removed. Missing data were addressed using appropriate imputation methods or exclusion, depending on the context and impact on the analysis. Additionally, outliers were detected using the Z-score method, allowing for the identification of data points significantly deviating from the norm, which were then removed to optimize the accuracy and efficiency of the model training process.

2.3. Data Processing

Categorical variables that comprised no more than two categories were straightforwardly encoded into binary values. For categorical variables with more than two categories, these were transformed into a series of binary dummy variables, each representing a unique category. To avoid potential complications with multicollinearity, which occurs when one independent variable in a model is linearly predictable from the others, dummy variables associated with the final category were intentionally omitted from the model. This exclusion is a common technique to prevent the introduction of redundant information, which can skew model performance and lead to unreliable coefficient estimates [20].

2.4. Data Scaling

Data scaling is a crucial preprocessing method in data preparation that ensures numerical features are brought within a similar range, thereby improving the performance and reliability of machine learning models. In this study, data normalization was specifically applied to the quantity of the inorganic variable. This step was necessary because the range of the inorganic variable was significantly larger compared to other variables. By normalizing this variable, we ensured that it did not disproportionately influence the model, allowing for more accurate and balanced predictions. Additionally, this scaling process helped in reducing the impact of potential outliers, leading to a more robust predictive model.

2.5. Feature Selection

As depicted in **Table 1**, a carefully curated selection of 10 key features was identified, each providing substantial information critical for accurate rice crop yield prediction. These features were chosen based on their strong correlations with yield outcomes, ensuring that the model is both robust and reliable in predicting variations in rice production. This selection process involved analyzing numerous potential factors and narrowing them down to those most impactful on the yield.

Table 1. Selected Features.

S/N	Feature
1	Quantity of inorganic fertilizer
2	Severe erosion
3	Season
4	Use of pesticides
5	Type of seed
6	Farmer type
7	Irrigation
8	Anti-erosion
9	Moderate erosion
10	Very low erosion

The following section discusses the theoretical rational behind the choice of features and how they influence crop yield prediction.

2.6. Justification of Feature Selection

2.6.1. The Quantity of Inorganic Fertilizer

It is the continuous variable measured in kilograms and it has a large impact on the rice productivity improvement. Inadequate use of fertilizers by farmers is among the challenges that were highlighted by the report of 2021 of the Ministry of Agriculture and Animal Resources (MINAGRI) of Rwanda. Other issues identified in the report include soil fertility degradation and farmers not applying sufficient organic fertilizers due to the lack of enterprises capable of sustainably supplying organic manure in marshland areas. In response, the Rwandan government prioritizes enhancing the availability and accessibility of inorganic fertilizers for farmers and subsidizes the cost of inorganic fertilizers for the cultivation of priority crops including rice. However, sufficient knowledge about how much fertilizer to apply and how it varies depending on the unique properties of the soil is lacking. Some farmers' crop yield is negatively impacted by improper fertilizer recommendations and inadequate use [7].

2.6.2. Degree of Erosion

It is the ordinal variable, which indicates the level of soil degradation in the plots, and its potential impact on crop productivity. Soil erosion is caused by ineffective catchment management, poor hillside land practices, flash flooding, and sediment accumulation, particularly during extreme weather events, which in turn damage rice crops and irrigation systems [7].

2.6.3. Agricultural Season

The agricultural season, which is a categorical variable, consists of two rice cultivation periods: Agricultural season A, from July to December, and agricultural season B, from January to June of the same year. The two seasons are affected by climatic conditions, particularly rainfall. Variations in rainfall distribution, temperature, and water availability in marshlands impact rice growth and harvest. Even though both seasons include the rainy periods, their effects can differ. Unpredictable rainfall can reduce yields; while consistent rainfall at crucial growth stages can enhance productivity [7].

2.6.4. Use of Pesticides

Rwandan rice farmers heavily use pesticides to prevent pest-related yield losses. However, the unchecked application of these agro-chemicals has raised concerns about potential health risks to humans, animals, and ecosystems due to water pollution. This issue arises from the lack of strict regulation in pesticide distribution and use [7].

2.6.5. Type of Seed Sown

The type of seeds, which indicates whether a farmer uses improved seeds or not, is a categorical variable. Research conducted in Sub-Saharan Africa, including Rwanda, has focused on improving seed genetics. As a result, different rice varieties were released in 10 major rice-producing countries in sub-Saharan Africa by 2020. These improved varieties can be compared with non-improved seeds to assess their effectiveness [21].

2.6.6. Farmer Type

The type of farmers, which includes small-scale and large-scale farmers, is a categorical variable. These two types of farmers differ financially and economically, which may affect their use of farming inputs and practices, potentially influencing crop productivity.

2.6.7. Irrigation

Irrigation, which indicates whether or not the farm was irrigated, is a categorical variable. Competing demands for water resources, such as household use, livestock, and hillside irrigation, limit the availability of water for rice cultivation in marshlands. To address this challenge, some farmers use irrigation, as rice crops require significant amounts of water. The use of irrigation can therefore enhance productivity [7].

2.6.8. Anti-Erosion

This is a categorical variable that indicates whether anti-erosion activities were implemented or not. Erosion control is essential for effective soil nutrient management, which, in turn, contributes to increased crop yield.

2.7. Machine Learning Models Implementation

The study investigated eight supervised machine learning algorithms to develop models for predicting rice crop yield. The algorithms examined include Linear Regression, which serves as a baseline model for comparison; Linear Regression with Interaction, which incorporates interaction terms to capture more complex relationships; Support Vector Machine (SVM), known for its effectiveness in high-dimensional spaces; and Gradient Boosting, which builds an ensemble of trees in a sequential manner to improve prediction accuracy. Additionally, Extreme Gradient Boosting Tree (XGBOOST) was evaluated for its robust performance in handling large datasets and its ability to prevent overfitting. Adaptive Boosting Tree (AdaBoost) was also included for its capability to combine weak classifiers to create a strong predictive model. Random Forest, an ensemble learning method that aggregates multiple decision trees, and Artificial Neural Network (ANN), which mimics the human brain's structure to learn from data, were also part of the study. These models were selected because their popularity and performance in crop yield prediction in the past studies [9, 10, 22–34].

Table 2 provides a comprehensive summary of the Python modules or libraries used for implementing each model, along with the specific hyper parameters tuned during the experimentation. This detailed comparison aims to identify the most effective algorithm for accurately predicting rice yield based on various cultivation resources and methods.

2.7.1. Multiple Linear Regression

A model that uses several predictor variables is known as a Multiple Regression Model. Multiple Linear Regression (MLR) is a statistical approach that helps describe the linear relationship between a dependent variable and one or more independent variables. The dependent variable, often called the target or response variable, is influenced by independent variables, also known as predictors. MLR is based on the least squares method and is commonly applied in fields like agriculture to develop predictive models [23, 35]. In this rice yield prediction study, MLR is employed where the rice yield is the dependent variable, and variables such as quantity of inorganic fertilizer, degree of erosion, season, use of pesticides serve as predictors. In this study the algorithm was implemented by considering the model default parameters including 10-fold cross-validation to avoid model overfitting.

2.7.2. Random Forest

One popular ensemble learning technique for predicting agricultural crop yields is Random Forest. To increase prediction accuracy, it builds many decision trees and aggregates their results. A random subset of the data is used to train each tree in the forest, and the outputs of all the trees are averaged or voted upon to get the model's final forecast. Large datasets including intricate correlations and interactions between variables, such as those pertaining to weather patterns, agricultural inputs, and soil conditions, lend themselves particularly well to this method [9, 24–26]. Because Random Forest can handle both classification and regression problems and is resistant to overfitting, it is a good choice for agricultural

Table 2. Models Implementation Summary.

S/N	Model Type	Python Package Used	Hyperparameters
1	Multiple Linear Regression	Sklearn, Linear Regression	Default parameters of the library, 10-fold cross validation
2	Random Forest	Sklearn, Random Forest Regressor	Default parameters of the library, 10-fold cross validation
3	Gradient Boosting	Sklearn, Gradient Boosting Regressor	Default parameters of the library, 10-fold cross validation
4	Support Vector Machine	Sklearn, SVR	Kernel = rbf, 10-fold cross validation
5	Artificial Neural Network	Tensorflow, Keras	Input layer: dense layer with 64 neurons, first Hidden layer: dense layer with 64 neurons, second Hidden layer: dense layer with 64 neurons, output layer: dense layer with 1 neuron Activation function: ReLu, Number of epochs = 80, Optimizer: Adam, Learning Rate: 0.01, Metric = mean_squared_error, Mean Absolute Error, Batchsize: 32, cross validation: 10-fold
6	Extreme Gradient Boosting Tree	XGBOOST, XGB Regressor	Max depth = 12, Learning Rate=0.4, number of estimators = 60 and 10-fold cross validation
7	Adaptive Boosting Tree	Sklearn, AdaBoost Regressor, Decision Tree Regressor	AdaBoost parameters: • n_estimators = 80 • learning_rate = 0.3 Decision Tree Parameters: • max_depth = 11 10-fold cross validation
8	Linear + Interaction	Sklearn, Linear Regression, Polynomial Features	LinearRegression parameters: • Default parameters Polynomial Features Parameters: • Degree = 2 • Interaction_only = True 10-fold cross validation

yield prediction. To prevent model overfitting, 10-fold cross-validation was included in the algorithm's implementation in this study, along with consideration for the model's default parameters.

2.7.3. Gradient Boosting

A potent machine learning technique for forecasting agricultural crop yield is called gradient boosting. Similar to other boosting strategies, Gradient Boosting builds decision trees one after the other, with each new tree concentrating on fixing the mistakes caused by the preceding ones. The model becomes better over time by minimizing a loss function, which makes it very good at capturing intricate relationships between variables like soil, climate, and crop management techniques. Gradient Boosting is a widely used model for yield outcomes modeling in agriculture due to its versatility in processing different types of data and its ability to generate forecasts with high accuracy. To mitigate overfitting, a 10-fold cross-validation was applied in the algorithm's implementation in this study, while also taking into account the default parameters of the model [27].

2.7.4. Extreme Gradient Boosting (XGBoost)

XGBoost is a high-performance machine learning algorithm that excels in agricultural crop yield prediction due to its ability to model complex, non-linear relationships [28–31]. It sequentially builds decision trees, with each tree correcting the errors of the previous one, making it highly effective for minimizing prediction errors. As shown in **Table 2**, in this study, XGBoost is configured with a maximum tree depth of 12, a learning rate of 0.4, and 60 estimators to balance accuracy and com-

putational efficiency. A 10-fold cross-validation is also employed to ensure the model's generalization and robustness.

2.7.5. Adaptive Boosting Tree (AdaBoost)

Another ensemble learning method that improves crop production prediction models by concentrating on strengthening weak learners is called adaptive boosting, or AdaBoost. Decision stumps, or decision trees with a single split, are constructed sequentially in AdaBoost. Every tree is trained by giving more weight to the observations that were previously incorrectly categorized, which enables the model to perform better and better over time. AdaBoost is helpful in capturing the intricate interactions between predictor variables for agricultural output prediction. It is a useful technique for increasing the precision of yield estimates because of its capacity to strengthen weak predictors through reweighting[27, 32].

As illustrated in **Table 2**, key parameters for the AdaBoost algorithm in this study included a learning rate of 0.3 (learning_rate = 0.3) and 80 estimators (n_estimators = 80). To balance complexity and avoid overfitting, the decision trees employed in AdaBoost were limited to a maximum depth of 11 (max_depth = 11). AdaBoost constructs decision stumps in a stepwise manner, with each new model concentrating on fixing the mistakes caused by the preceding one. Furthermore, 10-fold cross-validation was utilized to guarantee the model's resilience and applicability. By capturing the correlations between agricultural inputs including temperature, precipitation, and fertilizers, this technique helps to improve the accuracy of production forecast.

2.7.6. Support Vector Machine

A supervised learning approach called Support Vector Machine (SVM) is frequently used for regression and classification problems in crop yield prediction. SVM operates by identifying the hyperplane that most effectively divides data into distinct groups or fits a regression function. SVM is helpful in determining how different agricultural factors, such as soil properties, weather, and farming practices, affect crop output when it comes to forecasting crop yield. It is appropriate for modeling complicated agricultural systems because of its capacity to handle high-dimensional data and non-linear interactions through kernel functions. It provides precise forecasts based on a variety of crop-related elements[30, 34, 36].

2.7.7. Artificial Neural Network

An Artificial Neural Network (ANN) is a computational model designed to simulate the way human brains process information, making it effective for capturing complex, non-linear relationships between variables. ANN is mostly used in agricultural for different crops yield predictions including rice crop due to its high performance in prediction accuracy[10, 22, 33, 36].

An Artificial Neural Network (ANN) is used in this study to predict relationships between different agricultural inputs and rice yield. Four layers make up the architecture of the artificial neural network (ANN): an input layer with 64 neurons, two hidden layers with 64 neurons each, and a single neuron in the output layer. To add non-linearity, the ReLU activation function is used across the network. The model is trained over 80 epochs with a batch size of 32 using the Adam optimizer at a learning rate of 0.01. Mean squared error (MSE) and mean absolute error (MAE) are the metrics used to assess the performance, and a 10-fold cross-validation is used to guarantee robustness. This ANN method improves agricultural yield forecast accuracy and offers flexibility in collecting intricate patterns.

2.8. Model Parameters Tuning

The Grid Search approach was employed to optimize the model hyperparameters. This exhaustive search technique is designed to identify the best hyperparameters for a machine learning model. The parameters (max_depth, random_state, learning_rate, number of estimators, for XGBOOST and AdaBoost, Kernel for SVM, activation function, and number of layers and neurons in hidden layers for neural networks) that control the training procedure and model design are known as hyperparameters. To do this, a variety of values for each hyperparameter were tested, and the values that provided the best model prediction accuracy were kept. For the remaining parameters the default parameter values were considered during the model training process.

3. Results and Discussion

3.1. Evaluation Metrics

In the context of yield prediction, evaluating the performance of predictive models is crucial for assessing their accuracy and reliability. This evaluation ensures that the models provide credible forecasts that can guide agricultural decisions. Among the various metrics used for performance assessment, three commonly utilized ones are Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Relative Error. RMSE measures the square root of the average squared differences between predicted and observed values, emphasizing larger errors due to its squaring effect. MAE calculates the average magnitude of errors without considering their direction, providing a straightforward measure of average prediction accuracy. Relative Error, expressed as a percentage of the observed values, allows for a proportional assessment of prediction accuracy in relation to the size of the data. Together, these metrics offer a comprehensive evaluation of model performance, helping researchers and practitioners refine their predictive approaches and improve the reliability of yield forecasts.

3.1.1. Root Mean Square Error Computation

In this paper, the Root Mean Square Error (RMSE) evaluation metric was utilized to assess the accuracy of rice crop yield predictions. RMSE provides a quantitative measure of the average deviation between predicted and actual yield values, reflecting the model's predictive performance. This metric was computed by first determining the squared differences between each predicted value and its corresponding actual observed value. These squared differences are then averaged to obtain the mean squared error. Finally, the square root of this average is taken to obtain the RMSE. This method ensures that larger errors have a proportionally greater impact on the metric, making RMSE a robust measure of prediction accuracy. Mathematically, RMSE is represented as:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \tag{1}$$

With

- n is the total number of observations or samples;
- y_i represents actual yield for observation i;
- $\widehat{y}_i$ represents the predicted yield for sample i.

Upon completion of the training phase for each machine learning model, the Root Mean Square Error (RMSE) was meticulously calculated for both the training and testing subsets to comprehensively evaluate the performance of each model. **Figure 2** illustrates the RMSE values obtained for rice yield prediction across different models. The outcomes derived from the testing phase of the models, utilizing the testing subset, revealed that the Extreme Gradient Boosting Tree model produced an RMSE value of 0.71. This indicates the model's performance in predicting rice yields with a certain degree of error. In comparison, the Adaptive Boosting Tree model demonstrated a slightly lower RMSE value of 0.69, suggesting a marginally better accuracy in rice yield prediction. The comparative analysis of these RMSE values highlights the effectiveness of each model, with the Adaptive Boosting Tree model showing a more refined capability in reducing prediction errors. Furthermore, this evaluation underscores the importance of selecting the most accurate model for practical applications in agricultural forecasting, ultimately aiding in the optimization of rice production practices.

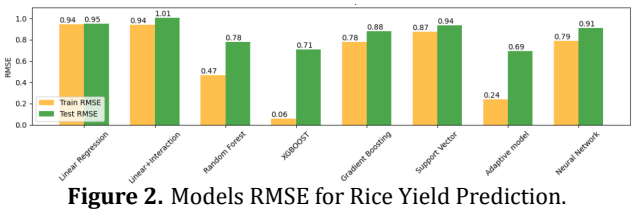

Figure 2. Models RMSE for Rice Yield Prediction.

The findings indicate that both the Extreme Gradient Boosting Tree (XGBoost) and Adaptive Boosting Tree (AdaBoost) models exhibit superior performance compared to other models evaluated in this study. This superiority is evident not only in their accuracy on the testing subset but also in their effectiveness during the training phase. The robustness of these models is highlighted by their ability to generalize well to new, unseen data while maintaining strong predictive power on the training dataset. The consistent performance across both subsets underscores the reliability and efficacy of XGBoost and AdaBoost in predicting rice crop yields, mak-

ing them valuable tools for enhancing agricultural forecasting efforts.

3.1.2. Mean Absolute Error Computation

The Mean Absolute Error (MAE) is a widely utilized metric for evaluating the accuracy of predictive models, as it provides a clear representation of the average prediction error by focusing on the absolute differences between predicted and actual values. MAE is advantageous because it directly measures the average magnitude of errors without taking into account their direction, which simplifies the interpretation of model performance. This metric becomes particularly valuable in scenarios where outliers might skew the results, as MAE does not amplify the impact of large errors through squaring, unlike the Root Mean Squared Error (RMSE).

In this paper, the MAE metric was employed to assess the average deviation of predicted rice yield values from actual observed values. By avoiding the squaring of errors, MAE provides a more robust measure when dealing with datasets that may contain outliers or extreme values. This characteristic makes MAE a preferred choice for evaluating models where an accurate and straightforward measure of prediction error is crucial. The calculation of MAE in this study followed the formula:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i| \qquad (2)$$

Where :

- n is the total number of observations or samples;
- y_i represents actual yield for observation i;
- $\hat{y}_i$represents the predicted yield for sample i.

Throughout the evaluation of predictive models, both the Extreme Gradient Boosting Tree (XGBoost) and Adaptive Boosting Tree (AdaBoost) models consistently demonstrate minimal Mean Absolute Error (MAE) values. This performance is evident not only within the training dataset but also across the testing dataset, indicating robust predictive accuracy and generalizability. These results underscore the models' effectiveness in capturing the complex relationships between agricultural inputs and practices, and rice yield. The accuracy and reliability of these models are visually represented in **Figure 3**, showcasing their strong performance in various testing scenarios. This reinforces the models' potential utility in optimizing rice yield predictions in different agricultural contexts.

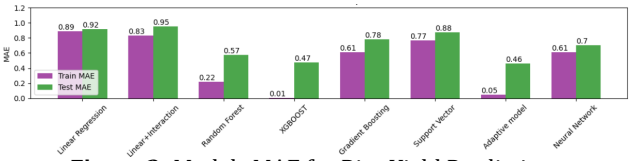

Figure 3. Models MAE for Rice Yield Prediction.

As indicated by **Figure 3**, the Adaptive Boosting (AdaBoost) tree achieved a Mean Absolute Error (MAE) value of 0.46, while the Extreme Gradient Boosting Tree (XGBoost) produced a slightly higher MAE value of 0.47 for rice yield prediction. This result highlights the superior performance of the AdaBoost model over the XGBoost model for predicting rice yield in Rwanda. The lower MAE value of AdaBoost suggests that it is more accurate in minimizing prediction errors, which is crucial for effective agricultural planning and decision-making.

Upon comparing the Root Mean Square Error (RMSE) and MAE values, as illustrated in **Figure 4**, a consistent narrative emerges regarding the optimal selection of machine learning models for rice yield prediction in Rwanda. The validation of these metrics underscores the reliability and effectiveness of both XGBoost and AdaBoost models. The results corroborate the initial findings, confirming that both models exhibit strong performance. However, the marginal advantage of AdaBoost in terms of MAE suggests that it may offer a slight edge in predicting rice yield more accurately, making it a preferable choice for future applications in precision agriculture. This comprehensive analysis highlights the importance of choosing the right model to enhance prediction accuracy and improve yield forecasting in the context of Rwanda's agricultural practices.

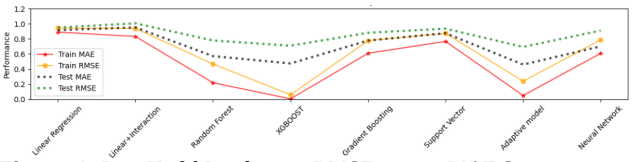

Figure 4. Rice Yield Prediction RMSE versus MAE Comparison.

3.1.3. Relative Error Analysis

The relative error is a crucial metric for assessing the accuracy of predictive models, particularly in the context of agricultural yield forecasting. In this field, pre-

cise predictions are essential for making well-informed decisions about resource allocation, crop management, and economic planning. Relative error measures the proportionate difference between the predicted and actual values, offering a normalized view of the prediction accuracy. This metric is especially valuable as it allows for easy comparison across different scales and datasets, providing a clear understanding of the model's performance. By quantifying how far off the predictions are relative to the actual outcomes, relative error helps in identifying the effectiveness of various modeling approaches and guides improvements. Consequently, it plays a significant role in enhancing the reliability of forecasts, which is vital for optimizing farming practices and achieving better yield outcomes [37].

In this study, we analyzed the relative error for various models trained and tested on rice yield data. The relative error was computed for both training and testing datasets to assess model performance across different conditions. The results are summarized in **Table 3**.

Among the models evaluated, both the XGBoost and Adaptive Boosting (AdaBoost) models demonstrated exceptional predictive performance, with the lowest test relative errors of 12.70% and 12.40% respectively. These figures indicate a strong capability for generalizing well to unseen data, which is crucial for reliable predictions in real-world scenarios. Furthermore, the AdaBoost model exhibited a notably low training error of 1.30%, reflecting a well-balanced fit to the training data without overfitting. This balance between bias and variance enhances the model's robustness and accuracy. In comparison to other models, these results underscore the effectiveness of XGBoost and AdaBoost in capturing the underlying patterns of rice crop yield influenced by various argricultural inputs and practices. Thus, both models represent promising tools for improving yield prediction and informing decision-making in agricultural practices.

3.2. Agricultural Inputs Contribution on Crop Yield

Evaluating the importance of features in yield prediction is crucial for guiding farmers and stakeholders on which parameters or cultivation resources and practices to prioritize for improving agricultural productivity. Understanding these factors is vital as they can have both positive and negative impacts on yield production. By identifying and focusing on the most influential variables, farmers can make informed decisions that enhance their productivity and sustainability. For this purpose, the AdaBoost algorithm, implemented through the scikit-learn Python library, was employed to assess the relative contribution of each agricultural input to yield prediction.

Figure 5 illustrates the results of this analysis, highlighting the contribution of each specific feature in percentage in predicting the rice yield. Among these, the quantity of inorganic fertilizer emerged as a key predictor with contribution of 56% for rice yield prediction, reflecting its significant impact on yield improvement. This is justified by the fact that the most of the rice farmers in Rwanda apply inorganic fertilizers for improving the rice production due to soil degradation. Additionally, the degree of erosion, seasonal variations, and seed types were identified as crucial factors influencing rice yield prediction with contributions of 7%, 6%, 6% respectively. Practices such as anti-erosion measures and organic farming, alongside the use of pesticides, also showed considerable importance in yield prediction. These findings underscore the necessity for farmers to optimize these features in their agricultural practices to achieve better yields. By leveraging such insights, stakeholders can better tailor their strategies to address the most impactful elements, thereby enhancing overall agricultural productivity.

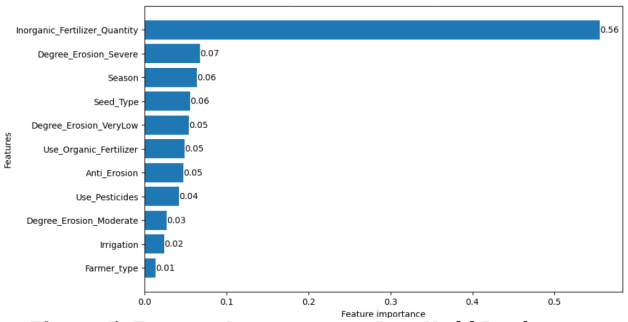

Figure 5. Features Importance in Rice Yield Prediction.

The agricultural inputs may have a beneficial positive or negative impact on rice productivity. As illustrated in the **Figure 6**, quantity of inorganic fertil-

Table 3. Model Relative Error.

Model	Train Relative Error	Test Relative Error
Linear Regression	22.80%	24.90%
Linear + Interaction	21.30%	25.70%
Random Forest	5.60%	15.40%
XGBOOST	0.30%	12.70%
Gradient Boosting	15.60%	21.10%
Support Vector Machine	19.70%	23.80%
Adaptive Model	1.30%	12.40%
Neural Network	15.60%	18.90%

izer, use of pesticides, agricultural season B, use of anti-erosion techniques and apply irrigation practices affect the rice production positively with corresponding correlations of 0.25, 0.25, 0.16, 0.08 and 0.07 respectively. When compared to other study-investigated predictors, these findings show that the amount of inorganic and use of pesticides has a significant favorable impact on rice output. However, traditional seeds type, small scale farmer type and degree of erosion affect the rice crop yield negatively in that order. With the aim of increasing rice agricultural production, these insights may assist farmers and other stakeholders in prioritizing the use of pesticides, improved seed varieties, irrigation techniques, and anti-erosion measures. The findings also imply that rice yield production is more advantageous during agricultural season B than that during agricultural season A.

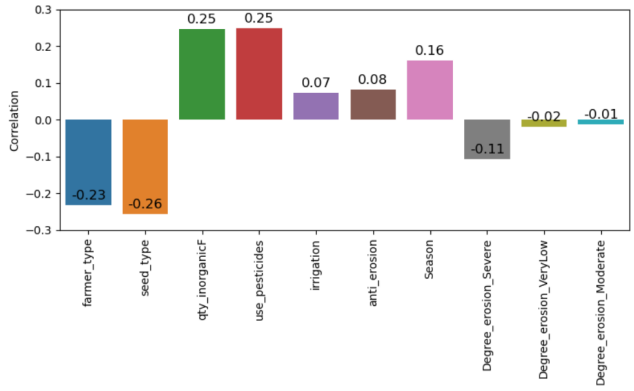

Figure 6. Correlation between Features and Rice Production.

4. Conclusions

In this study, our primary objective was to explore the best machine-learning model for predicting crop yields based on farming inputs and practices. We investigated eight supervised machine learning algorithms, including Linear Regression, Linear Regression with Interaction, Support vector Machine (SVM), Gradient Boosting, Extreme Gradient Boosting Tree (XGBoost), Adaptive Boosting Tree (AdaBoost), Random Forest, and Artificial Neural Network (ANN). The focus was to build models that effectively predict yields for Irish Rice crop.

Secondary data obtained from seasonal agricultural surveys were utilized, specifically 842 plots for rice. Before model construction, feature engineering techniques were applied to enhance the quality of the data. The dataset was then split into training and testing sets for model evaluation.

During the testing phase, utilizing the testing subset, the Extreme Gradient Boosting Tree model demonstrated promising performance, yielding RMSE values of 0.71 for Rice yield prediction. Similarly, the Adaptive Boosting Tree model exhibits competitive results with RMSE 0.69 for Rice yield prediction. The mean absolute error (MAE) depicted also the superiority of Adaptive Boosting tree on Rice yield prediction in Rwanda.

To understand the key factors affecting crop production, the Adaptive Boosting Tree algorithm and correlation matrix were used to investigate feature impact on rice yield production. The analysis highlighted the quantity of inorganic fertilizer and use of pesticides as the most critical factors that have more positive impact on Rice production with correlation of 0.25. However, the use of traditional seeds, erosion and type of farmer (small scale farmers) impact negatively the rice yield productivity. These findings offer valuable guidance for farmers and policymakers on which agricultural practices to prioritize for improving crop productivity.

The current study focused solely on agricultural inputs for predicting rice yield. However, numerous other factors, such as weather and soil conditions, also signifi-

cantly impact rice yield. Therefore, we recommend that future studies evaluate the combined effects of weather and soil parameters on rice yield prediction. Additionally, we suggest incorporating classification models in future research to help farmers assess whether upcoming yield production will be favorable or unfavorable, thereby guiding them in selecting crops that would maximize productivity.

Because Rwandan setting was the main focus of the research, the prediction model outcomes are unique to Rwandan context. However, by retraining the model with the national specific regional parameters that influence the rice yield prediction, the approach can be applied to other nations as well.

Author Contributions

Conceptualization, C.M., N.C., J.N., and D.N.M.; methodology, C.M.; software, C.M.; validation, C.M., N.C., J.N., and D.N.M.; formal analysis, C.M.; investigation, C.M.; resources, C.M.; data curation, C.M.; writing original draft preparation, C.M.; writing review and editing, C.M., N.C., J.N., and D.N.M.; visualization, C.M.; supervision, N.C., J.N., and D.N.M. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable

Data Availability Statement

The dataset is the property of the National Institute of Statistics of Rwanda and can be only accessible upon request.

Acknowledgments

The authors would like to thank the National Institute of Statistics of Rwanda for providing the data used in this research, and African Center of Excellence in Data Science (ACE-DS), University of Rwanda for their guidance and supervision.

Conflicts of Interest

All authors disclosed no any conflict of interest.

References

[1] NISR, 2023. Gross Domestic Product (GDP), Third Quarter 2023. Available from: https://www.statistics.gov.rw/publication/gdp-national-accounts-fourth-quarter-2023#:~:text=In%20the%20fourth%20quarter%20of,estimated%20at%20Frw%204%2C500%20billion (cited 11 March 2024).

[2] MINAGRI, 2023. Rwanda's Agriculture Sector Transformation Journey over the Last 29 Years. 4 July, 2023. Available from: https://www.minagri.gov.rw/updates/news-details/rwandas-agriculture-sector-transformation-journey-over-the-last-29-years (cited 11 March 2024).

[3] MINAGRI, 2023. Strategic Plan for Agriculture Transformation 2018–24. June, 2018. Available from: https://www.minagri.gov.rw/fileadmin/user_upload/Minagri/Publications/Policies_and_strategies/PSTA4__Rwanda_Strategic_Plan_for_Agriculture_Transformation_2018.pdf (cited 11 March 2024).

[4] Imasiku, K., Ntagwirumugara, E., 2020. An impact analysis of population growth on energy-water-food-land nexus for ecological sustainable development in Rwanda. Food and Energy Security. 9(1), e185.

[5] NISR, 2022. Comprehensive Food Security and Vulnerability Analysis. October, 2021. Available from: https://www.statistics.gov.rw/publication/comprehensive-food-security-and-vulnerability-analysis2022 (cited 11 March 2024).

[6] Liliane, T.N., Charles, M.S., 2020. Factors affecting yield of crops. Agronomy-Climate Change & Food Security. 9.

[7] MINAGRI, RWANDA, 2021. National Rice Development Strategy (2021–2030). July 2021. Available from: https://riceforafrica.net/wp-content/uploads/2021/09/rwanda_nrds2.pdf (cited 11 March 2024).

[8] Benos, L., Tagarakis, A.C., Dolias G., et al., 2021. Machine learning in agriculture: A comprehensive updated review. Sensors. 21(11), 3758.

[9] Jiya, E.A., Illiyasu, U., Akinyemi, M., 2023. Rice yield forecasting: A comparative analysis of multiple machine learning algorithms. Journal of Information Systems and Informatics. 5(2), 785–799.

[10] Li, N., Zhao, Y., Han, J., et al., 2024. Impacts of future climate change on rice yield based on crop model simulation—A meta-analysis. Science of The Total Environment. 949, 175038.

[11] Zhou, S., Xu, L., Chen, N., 2023. Rice yield prediction in hubei province based on deep learning and the effect of spatial heterogeneity. Remote Sensing. 15(5), 1361.

[12] Satpathi, A., Setiya, P., Das, B., et al., 2023. Comparative analysis of statistical and machine learning techniques for rice yield forecasting for Chhattisgarh, India. Sustainability. 15(3), 2786.

[13] Elbasi, E., Zaki, C., Topcu, A.E., et al., 2023. Crop prediction model using machine learning algorithms. Applied Sciences. 13(16), 9288.

[14] Nigam, A., Garg, S., Agrawal, A.,et al., 2019. Crop yield prediction using machine learning algorithms. In Proceedings of the 2019 Fifth International Conference on Image Information Processing (ICIIP); November 2019; Shimla, India; pp. 125–130.

[15] P. S., M.G., R., B., 2019. Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms. Applied Artificial Intelligence. 33(7), 621–642.

[16] Kang, Y., Ozdogan, M., Zhu, X., et al., 2020. Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. Environmental Research Letters. 15(6), 064005.

[17] Kuradusenge, M., Hitimana, E., Hanyurwimfura, D., et al., 2023. Crop yield prediction using machine learning models: case of Irish potato and maize. Agriculture. 13(1), 225.

[18] . Kumar Gajula, A., Singamsetty, J., Dodda, V.C., et al., 2021. Prediction of crop and yield in agriculture using machine learning technique. In Proceedings of the 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT); July 2021; Chennai, India; pp. 1–5.

[19] Panigrahi, B., Kathala, K.C.R., Sujatha, M., 2023. A machine learning-based comparative approach to predict the crop yield using supervised learning with regression models. Procedia Computer Science. 218, 2684–2693.

[20] Chan, J.Y.L., Leow, S.M.H., Bea, K.T., et al., 2022. Mitigating the multicollinearity problem and its machine learning approach: A review. Mathematics. 10(8), 1283.

[21] Futakuchi, K., Senthilkumar, K., Arouna, A., et al., 2021.History and progress in genetic improvement for enhancing rice yield in sub-Saharan Africa. Field Crops Research. 267, 108159.

[22] Gopal, P.M., Bhargavi, R., 2019. A novel approach for efficient crop yield prediction. Computers and Electronics in Agriculture. 165, 104968.

[23] Shastry, A., Sanjay, H.A., Bhanusree, E., 2017. Prediction of crop yield using regression techniques. International Journal of Soft Computing. 12(2), 96–102.

[24] Jeong, J.H., Resop, J.P., Mueller, N.D., et al., 2016. Random forests for global and regional crop yield predictions. PloS One. 11(6), e0156571.

[25] Suresh, N., Ramesh, N.V.K., Inthiyaz, S., et al., 2021. Crop yield prediction using random forest algorithm. In Proceedings of the 2021 7th international conference on advanced computing and communication systems (ICACCS); 19–20 March 2021; Coimbatore, India; Volume 1, pp. 279–282.

[26] Moraye, K., Pavate, A., Nikam, S., et al., 2021. Crop yield prediction using random forest algorithm for major cities in Maharashtra State. International Journal of Innovative Research in Computer Science & Technology (IJIRCST). 9(2), 40–44.

[27] Keerthana, M., Meghana, K.J.M., Pravallika, S., et al., 2021. An ensemble algorithm for crop yield prediction. In Proceedings of the 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV); 4–6 February 2021; Tirunelveli, India; pp. 963–970.

[28] Ravi, R., Baranidharan, B., 2020. Crop yield prediction using XG Boost Algorithm. International Journal of Recent Technology and Engineering (IJRTE). 8(5), 3516–3520.

[29] Mariadass, D.A., Moung, E.G., Sufian, M.M., et al., 2022. EXtreme gradient boosting (XGBoost) regressor and shapley additive explanation for crop yield prediction in agriculture. In Proceedings of the 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE); 17–18 November 2022; Mashhad, Iran; pp. 219–224.

[30] Mallikarjuna Rao, G.S., Dangeti, S., Amiripalli, S.S., 2022. An efficient modeling based on XGBoost and SVM algorithms to predict crop yield. In Advances in Data Science and Management: Proceedings of ICDSM 2021; 13 February 2022. Springer Nature Singapore: Singapore. pp. 565–574.

[31] Huber, F., Yushchenko, A., Stratmann, B., et al., 2022. Extreme gradient boosting for yield estimation compared with deep learning approaches. Computers and Electronics in Agriculture. 202, 107346.

[32] Jeevaganesh, R., Harish, D., Priya, B., 2022. A machine learning-based approach for crop yield prediction and fertilizer recommendation. In Proceedings of the 2022 6th International conference on trends in electronics and informatics (ICOEI); 28–30 April 2022; Tirunelveli, India; pp. 1330–1334.

[33] Bondre, D.A., Mahagaonkar, S., 2019. Prediction of crop yield and fertilizer recommendation using machine learning algorithms. International Journal of Engineering Applied Sciences and Technology. 4(5), 371–376.

[34] Dahikar, S.S., Rode, S.V., 2014. Agricultural crop yield prediction using artificial neural network approach. International Journal of Innovative Research in Electrical, Electronics, Instrumentation and Control Engineering. 2(1), 683–686.

[35] Ramesh, D., Vardhan, B.V., 2015. Analysis of crop yield prediction using data mining techniques. International Journal of Research in Engineering and Technology. 4(1), 47–473.

[36] Gandhi, N., Petkar, O., Armstrong, L.J., 2016. Rice crop yield prediction using artificial neural networks. In Proceedings of the 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR); 15–16 July 2016; Chennai, India; pp. 105–110.

[37] Kat, C.J., Els, P.S., 2012. Validation metric based on relative error. Mathematical and Computer Modelling of Dynamical Systems. 18(5), 487–520.