Journal of Future Artificial Intelligence and Technologies

E-ISSN: 3048-3719

Research Article

Optimizing Rice Production Forecasting Through Integrating Multiple Linear Regression with Recursive Feature Elimination

Joseph Abunimye Ingio [1], Augustine Shey Nsang [1], and Aamo Iorliam [2,*]

[1] Computer Science Department, School of Information Technology and Computing, American University of Nigeria, Yola, Nigeria; e-mail : joseph.ingio@aun.edu.ng, augustines.nsang@aun.edu.ng
[2] Data Science Department, School of Information Technology and Computing, American University of Nigeria, Yola, Nigeria; e-mail: aamo.iorliam@aun.edu.ng
* Corresponding Author: Aamo Iorliam

Abstract: Rice is a staple food for most Nigerians, making accurate yield prediction is crucial for food security. This study addresses the limitations of traditional forecasting methods by employing Multiple Linear Regression (MLR) coupled with Recursive Feature Elimination (RFE) to predict rice yield in Adamawa and Cross River states, characterized by distinct agroclimatic conditions. Utilizing climatic data and historical yield records from 1990 to 2022, we trained and evaluated MLR and compared the MLR results with two other machine learning models (XGBoost, and K Nearest Neighbours). RFE-optimized feature selection identified All-sky Photosynthetically Active Radiation (PAR) as a key factor. MLR demonstrated a very stable prediction performance with R^2 values of 0.90 and 0.92 for Adamawa and Cross River, respectively, after RFE. This research contributes to developing advanced Agro-information systems, supporting informed agricultural decision-making, and enhancing Nigeria's food security.

Keywords: Adamawa; Cross River; Food Security; Prediction; Multiple Linear Regression.

1. Introduction

Agriculture and food production are very important not only to Nigeria but also to the entire world. The importance of food production in achieving one of the major sustainable development goals of the United Nations (UN) has made it a major topic of discussion on a global scale with a focus on improving food security and decreasing hunger to a considerable extent by 2030[1]. The startling surge in the number of people facing food crises and hunger is the basis for this goal. There were 691–783 million hungry people in the world in 2022 alone, which is approximately 122 million more than the figures in 2019[2]. This shows an obvious need to produce more food, particularly the most important and widely eaten ones, to meet the global demand, which is increasing rapidly.

Rice is a food crop that is consumed by a great number of people, constituting over half of the world's population [3], [4] and it has been termed "the world's most important food crop" [5]. To raise awareness of the role of rice in reducing poverty and malnutrition, the United Nations declared 2004 to be the "International Year of Rice". This further registered its importance as a food source and widespread global consumption [4]. In addition, rice is seen as a commodity that can boost a nation's economic growth as it is a major export commodity for countries like China, India, The Philippines, etc.

In Nigeria, rice has emerged as a staple food over the past few decades, enjoyed in every part of the country[6], [7]. Rice production amounts to about 8.3 million metric tonnes of unmilled rice per year and about 5.4 million metric tonnes of milled rice per year, constituting 46% of the total rice produced in Africa[8]. Nevertheless, this production rate is insufficient to meet the nation's rising rice demand, which has increased reliance on rice importation to

Received: July, 15th 2024
Revised: August, 3rd 2024
Accepted: August, 8th 2024
Published: August, 10th 2024
Curr. Ver.: August, 10th 2024

DOI : 10.62411/faith.2024-17

satisfy the country's teeming population of rice eaters. In 2014, half the quantity of rice consumed in Nigeria was imported[6], and in 2018, over 7 trillion Naira was spent importing rice into Nigeria[9]. The surge in demand can be attributed to various factors, including shifts in consumer preferences, population growth, growing incomes, and a swift urbanization process[7]. The rice produced in Nigeria is cultivated in about 21 states, with eight states producing over 50% of the total amount of rice produced in Nigeria. Most of the states with the highest quantity of rice produced are in the country's Northern region, including Kebbi, Kaduna, Kano, and Borno, while a few are in the Southern region, including Cross River and Ebonyi. Cultivation of rice is usually done in rainfed lowland fields and rainfed highland fields during the rainy season, typically between May and August. However, this spells some challenges for rice farmers in northern Nigeria as about 1200mm to 1600mm of rainfall is needed for optimum rice growth, and this volume of rainfall does not occur in the North. In addition, pest infestation and poor soil fertility are challenged due to increased pressure on land resources due to population expansion [7].

The Agricultural data available can provide valuable insights into trends and patterns that can be used in analysis and prediction. Using data mining techniques is one way to accomplish this. Data mining is a process in which large datasets are searched to uncover new patterns and relationships [10], [11] to extract knowledge from the data and convert it to a human-understandable format. This constitutes a major preliminary step towards applying machine learning methods to forecast or take action based on the knowledge found in the data.

The predominant technique for predicting crop yield among farmers in Nigeria mostly employs a crude method of estimating the yield of a particular crop based on previous yield with very little consideration given to possible climatic and environmental factors that may have changed after the previous yield. Data mining and Machine learning techniques can help increase the prediction accuracy as those factors are considered when building machine learning models for crop prediction, thereby increasing the predictability and accuracy of the predicted yield.

Integrating machine learning into agriculture holds promise as it can bring advantages. One major benefit is the ability to make predictions, which helps reduce errors when relying on manual forecasting, enabling informed decision-making processes and promoting further growth in the agricultural sector. This technological advancement has the potential to address the challenges previously mentioned, such as bridging the gap between rice demand and production in Nigeria. Our research employs Multiple Linear Regression (MLR) as a machine learning technique to forecast rice production in Adamawa and Cross River States in Nigeria. We incorporate Recursive Feature Elimination (RFE) to enhance predictive accuracy for optimal feature selection. The overarching goal is to contribute to increased rice production and improved national food security. Specifically, this study aims to achieve the following objectives:

1. Develop a rice yield prediction model using Multiple Linear Regression.
2. Implement Feature selection using RFE and F-regression to identify the most influential factors affecting rice yield.
3. Compare the performance of the MLR model and two other Machine Learning Algorithms with and without feature selection.
4. Evaluate the generalizability of the developed model across different Agro-ecological zones in Nigeria (e.g. Adamawa and Cross River states).

The remainder of this paper is organized as follows: Section Two presents a Literature Review, Section Three explains the methodology employed in the research, Section Four discusses the results and findings, and Section Five presents the conclusion and future work.

2. Review of Related Studies

Agricultural processes have long been carried out manually, and much of them are still done that way in most developing countries, including Nigeria. In sub-Saharan Africa, up to 65% of farming is done manually, about 25% uses animal traction (donkeys, bulls' carts etc.), and about 10% is mechanized [10]. As a result, farming is seen to be a laborious task. This notion continued until mechanized farming and tractors were introduced into land processing. This was necessitated by the shortage of food, workers, and draft animals caused by the World War [11]. With this new development came the advantages of large-scale farming and an increased efficiency in food production. However, the introduction of modern technologies for agricultural mechanization encountered some hindrances in many developing countries due to factors such as compatibility with the environment, availability of resources, cost, government policies, adequacy, and appropriateness. Consequently, farmers in these countries have inadequately used available resources, resulting in low productivity and high production costs[10]. These hindrances are not the only factors responsible for the low agricultural productivity. Challenges such as climate changes, urban encroachment, and a lack of qualified farmers have brought about new practices for sustainable agriculture and food supply[12]. Precision agriculture, also referred to as smart farming, has arisen as a cutting-edge approach to tackle these existing challenges threatening the sustainability of agricultural practices[13]. Sometimes shortened to digital agriculture, it utilizes modern information technologies, software, and smart devices to enable data-driven, sustainable farm management. Essentially, it employs technology-enabled tools to assist decision-making in agricultural operations [13], [14]. This is ultimately aimed at reducing the cost of food production and the environmental impact of agricultural practices while maintaining an optimum yield and profitability.

Precision agriculture technologies can be categorized into five groups according to Pierce & Nowak[14] – Geographic Information Systems (GIS), Global Positioning Systems (GPS), sensors, computers, and application control tools.

Yield Prediction appears to be one of the most challenging tasks in Precision Agriculture [15] because several parameters contribute to the optimum yield of a particular crop species, and these parameters vary from one species to another. As a result, many models have been proposed so far. Conventional approaches to predicting rice yields prior to harvest have predominantly consisted of statistical regression models[16], process-based crop simulation models grounded in agronomic principles like the CERES model[17], and traditional farmer knowledge and observations.

While valuable, these traditional statistical and simulation modelling techniques face several limitations in accurately capturing the multitude of complex, often non-linear interactions between the diverse factors that influence rice yield in the real world[18]. Crop models are data-hungry, requiring extensive inputs that may not be available, and they make use of assumptions that restrict their generalizability[19].

Traditional farmer knowledge is grounded in local experience but can lack quantitative rigor and predictive precision[20]. It may also fail to holistically integrate the array of biotic and abiotic stresses across the crop cycle that cumulatively shape final yields.

These limitations have motivated increasing research to leverage machine learning techniques as an alternative, data-driven approach for developing more accurate and robust yield prediction models.

Van Klompenburg et al.[15] conducted a Systematic literature review on crop yield prediction using Machine learning and deep learning over the span of more than a decade, and their findings revealed the most used machine learning algorithms, the most preferred features for crop yield prediction, and which evaluation parameters have been used in literature for crop yield prediction. The research concluded that Convolutional Neural Networks (CNN) is the most widely used deep learning algorithm, followed by linear regression, which is commonly used as a benchmark but not necessarily the best-performing algorithm. They identified the following as the most preferred features for crop yield prediction: Temperature, soil type, rainfall, and crop information (weight, growth rate, species of plants, and crop density). And the most used evaluation parameters include Root Mean Square Error (RMSE), R^2, and Mean Absolute error.

Another major contributor to this research domain is Paudel et al. [21]. In their research, they developed a machine learning workflow that can be used for large-scale crop yield prediction. Having identified that the methods and data used in predicting the yield of a particular crop may not be transferable to another crop or location, their workflow focuses on a modular application of machine learning that ensures correctness and reusability and can be applied in different countries with minimal configuration changes.

Also notable is the work of Patrio et al. [22], who compared the performance of Random Forest Regression, Gradient Boosting, SVR, K-Nearest Neighbours Regression, and Decision Tree Regression in predicting rice yield using climatic and yield data from the Sumatra island. Their study identified Linear regression as the best-performing model with an R^2 score of 85.53%.

In Nigeria, Iorliam et al.[23] utilised machine learning techniques like Support Vector Machine, Naïve Bayes, Decision Tree, Logistic Regression, and K-Nearest Neighbour for Okra shelf life prediction and observed that Support Vector Machine, Naïve Bayes, and Decision Tree predicted the shelf life of Okra much better as compared to the other machine learning techniques they used.

Jiya et al. [24] performed a study in Nigeria using rice yield and climatic data from Katsina state between 1970 and 2017 in which they employed various models such as Random Forest, Artificial Neural Network, Random Trees, Logistic Regression, and Naïve Bayes in predicting rice yield in Katsina State and compared the performance of each machine learning technique. Their result showed that random forests and random trees demonstrated better performance in predicting rice yields than the other techniques listed above, offering a tool for proactive measures to ensure food security in the region. Even though this research is closely related to ours, it focused on a different location (Katsina State), and the machine learning algorithm we utilized differs from Jiya et al. [24]. This research is therefore motivated by Iorliam et al. [23] and Jiya et al. [24] with a focus on predicting rice yield in Adamawa State and Cross River State of Nigeria using Multiple Linear Regression.

3. Methodology

Our methodology consists of five phases and is described below and summarized in Figure 1:

1. Data Exploration phase – An initial inspection and preprocessing of the rice dataset to understand features, distribution, and missing values was done.
2. Data Preprocessing phase – This phase involved transforming the raw data into prepared model input through techniques like encoding, normalization, and handling missing values.
3. Model Development phase - The Multiple Linear Regression model was implemented using appropriate libraries and tools in Python.
4. Model Evaluation phase – Systematically evaluating model performance using metrics like RMSE, MAE, and R-squared based on train/test splits.
5. Model Optimization phase – In this phase, hyperparameters were tuned, and the results were analyzed to select the optimal model regarding predictive accuracy on rice yield.

This multi-phase methodology provides a rigorous framework for testing the machine learning regression model based on its ability to predict rice yield from the available dataset features accurately.

3.1. Study Area and Data Collection

Adamawa State is located in northeastern Nigeria within the savannah vegetation zone. It has an area of about 36,917 km2 and an estimated population of 4.9 million [25], [26]. The tropical climate in the state experiences distinct wet and dry seasons. Average annual rainfall ranges from 75 -103 mm, concentrated in the wet season months of May to September. Mean annual temperatures vary from 22°C to 31°C[26], [27]. The vegetative landscape consists primarily of short grasses, scattered trees, and shrubs. Major cash crops grown in the state include maize, rice, cotton, sorghum, and sugarcane.

Cross River State is located in the southern coastal region of Nigeria within the tropical rainforest vegetation zone. It covers an area of 20,156 km2 and has a population of approximately 4.2 million[25]. The state has abundant rainfall exceeding 3036 mm annually and high relative humidity. Temperatures remain relatively constant throughout the year, averaging between 15°C to 30°C. The natural vegetation is dense rainforest rich in timber resources. Major crops grown include rice, cassava, oil palm, cocoa, rubber, and plantains. The Cross River basin provides favorable conditions for wetland rice cultivation.

3.1.1. Justification for Study Area Selection

Adamawa and Cross River states were strategically selected for this paper due to their importance for rice production in Nigeria combined with their distinct geo-climatic characteristics. Both states contribute substantially to the total quantity of rice produced in Nigeria. Implementing model performance between these two Agro-ecological zones with different climates, soil conditions, and farming practices provides insights into the transferability of the machine learning algorithm (Multiple Linear Regression). Any model that consistently performs well in both locations is likely to be effective in generalizing to other rice-growing regions of Nigeria. The multi-year time-series data from the two states also enables the training of sophisticated machine-learning models for yield forecasting, particularly the deep-learning model. This paper provides a template for expanding prediction efforts to more rice-producing states in the future.

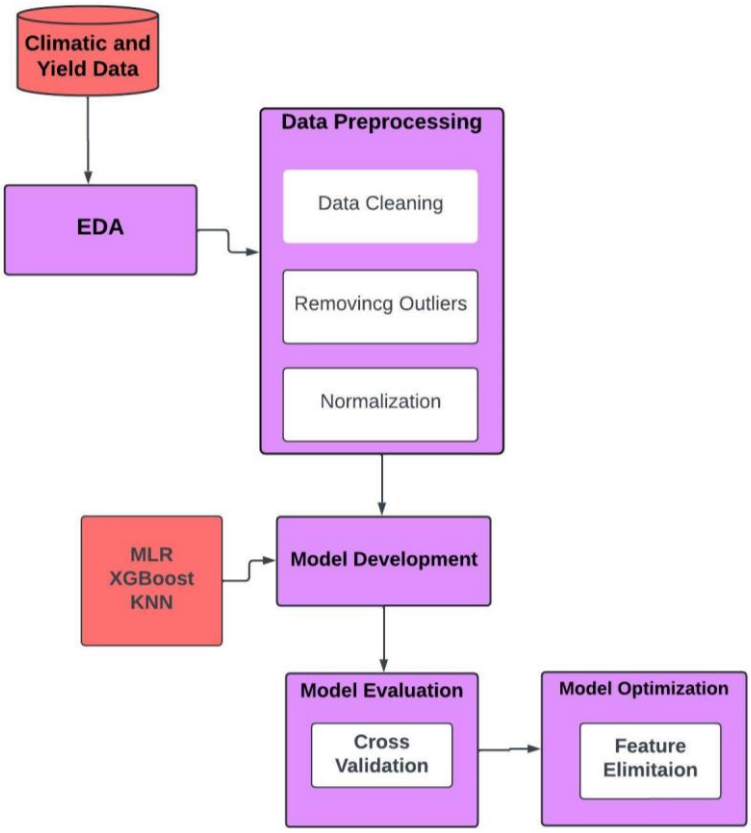

Figure 1. Visualization of the Proposed methodology.

3.1.2. Data Collection Method and Datasets Used.

Annual rice yield data (tons/hectare) from 1997 to 2020 was collected from the National Bureau of Statistics database (NBS) for each state. The NBS data is compiled from state-level agricultural production surveys and provides authoritative aggregated statistics on crop yields. These datasets are made available to the general public through NBS's web-based portal.

The datasets for Adamawa state contain a larger number of rows than those for Cross River because some rows with missing values in the Cross River state dataset were removed. After preprocessing, the final dataset for Adamawa contained 32 rows, while Cross River had 22 rows. These datasets were chosen for this study because they contain the most comprehensive data on rice yield in both regions. However, it's important to note that the relatively small sample size may limit the model's ability to capture complex patterns or generalize to future years. Access to more data is one of the challenges encountered during the course of this study.

Corresponding climatic data was obtained from the NASA Prediction of Worldwide Energy Resource (POWER) project, which provides global meteorological data derived from satellite observations and numerical weather prediction models. Specific location coordinates within each state were used to retrieve POWER API data: Long. 11.41° (+2.02°), Lat. 8.02° (+2.72°) for Adamawa and Long. 8.39° (+0.51°), Lat. 4.99° (+1.77°) for Cross River.

The NASA POWER data parameters include:

- Precipitation - Total monthly rainfall (mm)
- Minimum temperature - Monthly minimum temps (°C)
- Maximum temperature - Monthly maximum temps (°C)
- Specific humidity - Monthly average specific humidity (kg/kg)
- Photosynthetically active radiation - Monthly average downward surface shortwave flux (W/m^2)

- Wind speed – the average wind at 2 metres above the ground (m/s)
- Average Temperature – Monthly average temperature
- Relative Humidity – Monthly average relative Humidity

The climatic data was initially retrieved at a monthly resolution. We calculated annual averages for each climatic variable to integrate it with the annual yield data, ensuring temporal alignment with the yield data. The NBS yield data was combined with each state's 18-year POWER climatic data to compile the input dataset for training and testing the machine learning model (Multiple Linear Regression). The dataset was screened for any missing values and outliers. Rows with missing values were removed. We retained these data points for anomalies that appeared to be valid extreme events (e.g., years with unusual weather patterns) to maintain the dataset's ability to capture rare but important events.

It is worth noting that while the NBS and NASA POWER datasets are generally considered reliable, they may have limitations. The NBS data, being survey-based, could be subject to reporting errors or biases. The NASA POWER data, derived from satellite observations and models, may not always perfectly represent ground-level conditions. These potential limitations in accuracy and representativeness were considered during our analysis and interpretation of results. We agree that future studies could benefit from incorporating additional datasets or refining data collection methods to enhance data quality and generalizability.

By integrating these diverse datasets and applying careful preprocessing techniques, we aimed to create a robust foundation for our rice yield prediction model, while acknowledging and addressing the inherent challenges in working with agricultural and climatic data. The processed datasets will be made available upon reasonable request to facilitate transparency and reproducibility.

3.1.3. Justification of Model Selection

Multiple Linear Regression (MLR) is a straightforward and easily interpretable model, making it suitable for initial analysis and establishing baseline performance. The dataset used in this study includes various climatic factors (temperature, humidity, precipitation, etc.) that could influence rice yield. MLR holds the potential to incorporate multiple independent variables to predict the dependent variable (rice yield).

In addition, this study's exploratory approach in the chosen regions suggests that commencing with a basic model is advisable. This study marks the initial exploration into predicting rice yields in these Nigerian regions. Therefore, opting for a straightforward model such as MLR is a sensible strategy.

Recent studies that applied linear regression as a benchmark in crop yield prediction recorded success with data from a different region. For instance, Patrio et al.[22] linear regression is the best-performing model for rice yield prediction in Sumatra.

The initial exploratory data analysis suggested a linear relationship between rice yield and the selected independent variables (climatic factors). This assumption is fundamental to the application of MLR. Rice yield, being a continuous numerical value, aligns with the requirements of linear regression models. As a well-established statistical technique, MLR is a strong baseline for comparison with other, potentially more complex models in future research."

However, for the purpose of this study, MLR will be compared with two other Machine learning algorithms, namely, eXtreme Gradient Boost (XGBoost) and K Nearest Neighbours(KNN) for their ability to accurately predict the yield of rice in Adamawa and Cross River states of Nigeria.

3.2. Multiple Linear Regression

Multiple Linear Regression, also known as Multilinear Regression, is a machine learning algorithm that utilizes statistical regression analysis to predict the value of a dependent variable based on a set of independent variables. It is an extension of Linear Regression, which is a multivariate technique. Regression analysis aims to construct mathematical models that describe or explain the relationships that may exist between variables. The simplest case is Simple Linear Regression, where there is only one dependent variable and one independent variable. In contrast, Multiple Linear Regression involves more than one independent variable to predict one or more dependent variables. Machine Learning algorithms based on regression analysis are commonly applied in forecasting and, in some cases, to determine the causal relationship between the dependent and independent variables[28]. Forecasting in regression analysis occurs using Equation (1).

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon = \beta_0 + \sum_{i=1}^{p} X_i . \beta_i + \varepsilon \tag{1}$$

where $X_1, X_2, X_3, \dots, X_p$ are the independent variables or features used to predict the dependent or target variable y, and ε is an unobservable random variable (the error component) with mean 0 and variance σ^2 . The relationship described by (1) is known as a multiple linear regression model. β_0 is the intercept, $\beta_1 \dots \beta_p$ are the slope coefficients for each independent variable and $\sigma^2 > 0$ is an unknown error variance[29].

4. Results and Discussion

In this section, we present the results derived from applying the phased methodology described in the previous section on the datasets obtained from Adamawa and Cross River states.

4.1. EDA and Preprocessing

During the Exploratory Data Analysis phase, the datasets from both states were visualized to view its properties and distribution. Several missing values were observed in the Cross-river dataset. These missing values were removed by removing the rows containing them.

4.1.1. Correlation Matrix between Variables

To better understand how the features relate with each other as well as with the target variable in terms of correlation and which features are most important for prediction, a correlation matrix for each of the states was generated, as seen in Figures 2 and 3.

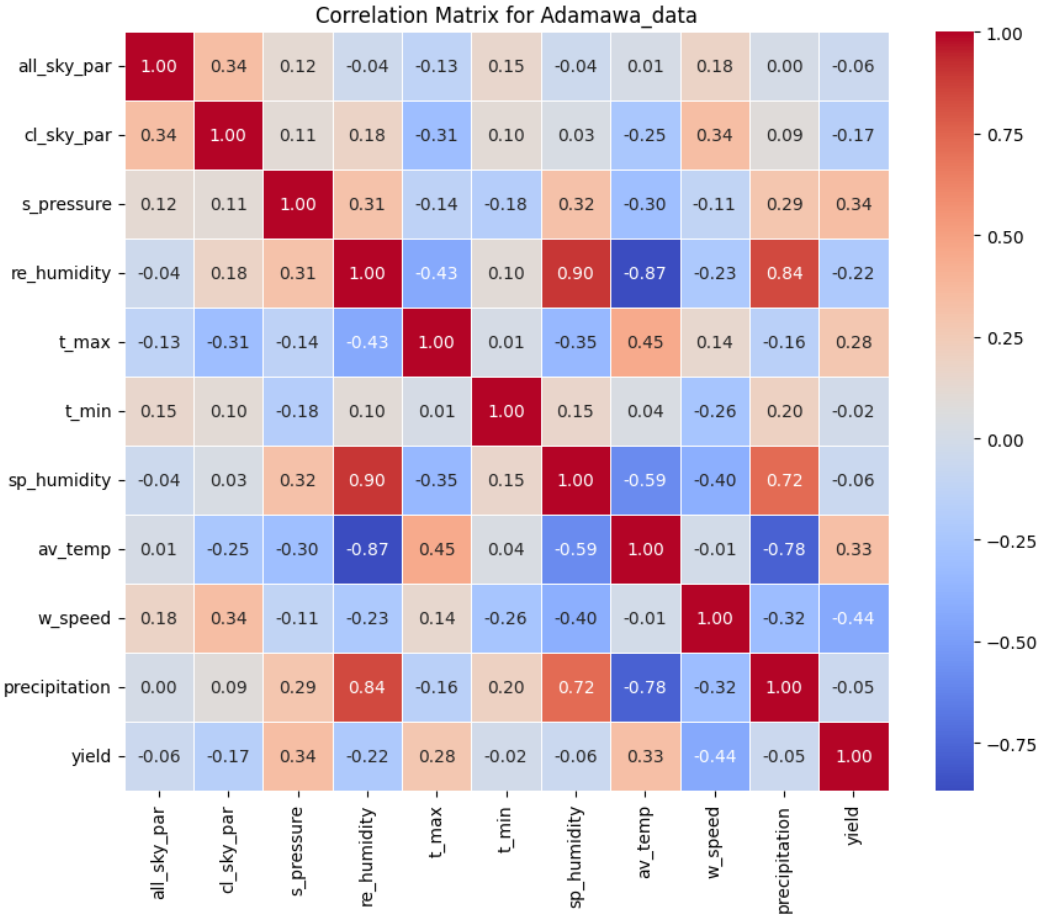

Figure 2. Correlation Matrix for Adamawa State Data

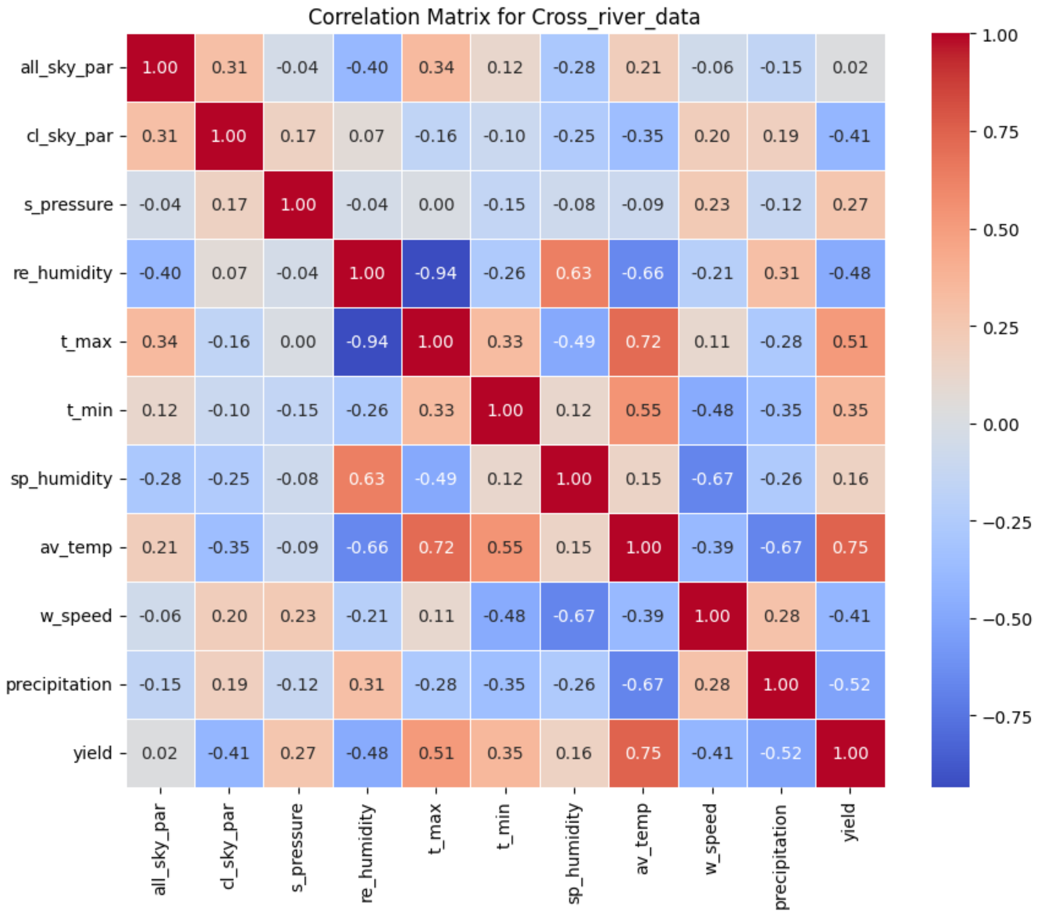

Figure 3. Correlation Matrix for Cross River State Data

Figure 2 shows a positive correlation between specific humidity (sp_humidity) and precipitation (precipitation). The correlation coefficient between these two variables is 0.72, suggesting an increase in sp_humidity whenever precipitation increases.

A strong positive correlation between re_humidity and sp_humidity (0.90) and precipitation (0.84) suggests multicollinearity among these variables. This means that these variables contain redundant information. The model might struggle to distinguish the independent effect of each on yield, leading to inaccurate coefficient estimates and increased variance. Hence, such variables are not ideal predictor choices.

The negative correlation between temperature and yield can be helpful as it clarifies the relationship between temperature and yield (indirectly through precipitation).

From Figure 3, some variables are observed to have relatively high correlations with each other, indicating potential multicollinearity issues. For instance, "t_max" (maximum temperature) and "av_temp" have a correlation of 0.72. The variables "cl_sky_par" (clear sky radiation) and all_sky_par (all sky radiation) have a correlation of 0.31. Other variables like "t_min" (minimum temperature), sp_humidity(specific humidity), all_sky_par (all sky radiation) have low or near-zero correlations with the "yield" variable, suggesting they may have little predictive power for the yield. Features with high multicollinearity were omitted from the training and testing set during feature selection.

4.1.2. Feature Selection

Feature selection is an ideal technique that is usually employed in reducing dimensionality in datasets by selecting only important features for prediction and leaving out the rest[30], [31]. This significantly reduces training time while improving prediction performance and a better understanding of the data[32]. Several Feature selection methods have been presented in various studies within the domain of machine learning[32], with a majority of them best suited to classification machine learning techniques. This study employs recursive feature elimination (RFE). RFE was chosen because it is a wrapper method that works directly with the chosen model (in this case, Multiple Linear Regression). This means that it selects features based on their importance specifically for the MLR model, which can lead to better performance compared to filter methods like Feature Selection by Omitting Redundant Features (FSOR) or Chi-squared. Unlike these filter methods, or other correlation-based selection methods, RFE can capture potential interactions between features. This is particularly important in agricultural systems where various climatic factors may have complex interrelationships affecting crop yield. In considering the datasets used, with its size and dimensionality, RFE offers a good balance between thoroughness and computational efficiency compared to some more exhaustive search methods that may require a larger dataset for a reliable statistical inference.

RFE was carried out to select the features that will best predict the target variable and the features that were selected are shown in Table 1 below. A different feature selection technique, F-regression, was also used to validate the results of the RFE process to present the most relevant features selected using RFE and F-regression. The F-regression selection technique reduces the dimensionality of data by selecting a subset containing the most relevant features for our regression model based on their F-value scores or scores from variance analysis as shown in Table 1.

Table 1. Selected features.

No	Using RFE	Using F-Regression
1	s_pressure	s_pressure
2	t_max	re_humidity
3	av_temp	sp_humidity
4	w_speed	av_temp
5	all_sky_par	w_speed
6	cl_sky_par	precipitation
7	t_min	cl_sky_par

Several iterations of the training and testing of the selected algorithms were carried out, and the performances recorded weren't so good. The number of features to select was reduced to 5 (for RFE: n_features_to_select= 5, for F-regression: k = 5) and the following features were selected: all_sky_par, re_humidity, t_max, w_speed, s_pressure.

4.2. Model Evaluation

The following performance metrics were applied to evaluate the performance of the models built using the machine learning algorithms mentioned in previous sections. Below is a brief description of each matric and its significance.

- MSE: This stands for Mean Squared Error. It measures the average squared difference between predicted and actual rice yields. Lower MSE indicates a better fit between predictions and actual values.
- R^2 Score: This is the R-squared coefficient of determination. It represents the proportion of variance (squared correlation) in the dependent variable (rice yield) that can be explained by the independent variables (features used in the model). R-squared values closer to 1 indicate a better fit.
- MAE: This stands for Mean Absolute Error. It represents the average absolute difference between predicted and actual rice yields. Lower MAE indicates better model performance.
- RMSE: This stands for Root Mean Squared Error. It's the square root of the MSE and represents the standard deviation of the prediction errors. Lower RMSE indicates better performance.
- MAPE: This stands for Mean Absolute Percentage Error. It represents the average absolute percentage difference between predicted and actual rice yields. Lower MAPE indicates better performance.

These metrics were used to measure the accuracy of the model's performance in predicting the yield for both states, and the results are analyzed in the following section.

4.3. Analysis of Evaluation Results

4.3.1 Feature Selection

Table 2 summarizes values for the model performances captured using the metrics mentioned in the previous section. It provides insights on the accuracy of each model built during this study.

Table 2. Results from the Adamawa dataset before and after feature selection.

Metric	Before feature selection			After feature selection		
	MLR	**XGBoost**	**KNN**	**MLR**	**XGBoost**	**KNN**
MSE	3.252120e+08	2.898339e+08	2.113927e+08	5.781650e+07	3.416119e+08	8.881611e+07
R^2	0.455188	0.514455	0.645864	0.903143	0.427714	0.851211
MAE	16748.488682	11570.315714	10719.874286	6211.029384	11623.070179	6534.076571
RSME	18033.635899	17024.508230	14539.349596	7603.715936	18482.746754	9424.229780
MAPE	7.648347	4.983372	4.579970	2.804602	4.965539	2.819843

The results from Table 2 suggest that Multiple Linear Regression remains the top-performing model after feature selection. There is a slight improvement in R^2 score and a reduction in MAE and RMSE compared to the results before feature selection. This indicates that feature selection has helped refine the model's predictive power. XGBoost continues to underperform compared to other models. The significant drop in R^2 score and increase in MAE and RMSE suggest that feature selection did not improve its performance. This model might benefit from further hyperparameter tuning. KNN Shows a slight improvement in R^2 score and a reduction in MAE and RMSE after feature selection. However, it still lags behind Multiple Linear Regression regarding overall performance. Overall, Multiple Linear Regression emerges as the most suitable model for predicting rice yield in this context, even after feature selection.

Table 3 presents a slightly different performance output of the selected models compared to the Adamawa datasets. Before Feature selection, Multiple Linear Regression displays a moderate performance with an R^2 score of 0.634583 and some high errors in MSE, MAE, RMSE, and MAPE. XGBoost, is a more superior performance model with an R^2 score of 0.775510 and lower errors compared to Multiple Linear Regression and KNN in MSE, MAE, RMSE, and MAPE. KNN has the least performance with the lowest R^2 score of 0.530093 and the highest errors in MSE, MAE, RMSE, and MAPE.

Table 3. Results from Cross River dataset before and after feature selection.

Metric	Before feature selection			After feature selection		
	MLR	**XGBoost**	**KNN**	**MLR**	**XGBoost**	**KNN**
MSE	2.090523e+08	1.284290e+08	2.688304e+08	4.719138e+07	2.872232e+07	2.934313e+08
R^2	0.634583	0.775510	0.530093	0.917511	0.949794	0.487091
MAE	10603.209665	6942.637500	13448.040000	5693.183176	4061.531250	14991.240000
RSME	14458.639624	11332.653561	16396.049340	6869.597977	5359.320662	17129.835278
MAPE	4.542748	2.879102	5.846778	2.495869	1.813581	6.435512

After Feature Selection Multiple Linear Regression shows a significant improvement in all metrics. R^2 score improved to 0.917511, indicating a better fit. There is also a drastic reduction in errors: MSE, MAE, RMSE, and MAPE all decreased. XGBoost is further improved with the highest R^2 score of 0.949794 and the Lowest errors in MSE, MAE, RMSE, and MAPE among the three models. However, the performance of KNN deteriorated further with a lower R^2 score of 0.487091 coupled with increased errors in MSE, MAE, RMSE, and MAPE compared to before feature selection.

While XGBoost demonstrated superior predictive accuracy for the Cross River dataset, Multiple Linear Regression exhibited greater potential for generalized yield prediction across both regions. This is evidenced by its more substantial performance improvement following feature selection and its consistently strong performance making it to rank as the top model in the Adamawa dataset and a close second in the Cross River dataset.

4.3.2 Cross Validation

The previous subsection throws more light on the importance of feature selection and its impact on the performance of the selected models. To further assess the generalizability of the models [33], 4-fold cross-validation was carried out on datasets from both states (Adamawa and Cross River state) after feature selection, and the following results were recorded in Table 4.

Table 4. Results from Adamawa and Cross River dataset after feature selection and cross-validation.

Metric	Adamawa			Cross river		
	MLR	**XGBoost**	**KNN**	**MLR**	**XGBoost**	**KNN**
MSE	4.253638e+08	4.740212e+08	3.014979e+08	4.155391e+08	5.527699e+08	3.634009e+08
R^2	0.556743	0.111303	0.436671	0.170834	-0.132085	0.270465
MAE	10912.813003	13922.740057	13351.274061	16758.298144	15481.745573	13775.181667
RSME	15138.170878	21179.623856	17178.531950	18966.310700	21449.729686	18894.272229
MAPE	4.967946	6.243436	5.981455	7.602415	7.713542	7.220279

The result in Table 4 shows a decrease in performance across all models after cross-validation compared to their performances before cross-validation, as seen in Tables 1 and 2 for both datasets, respectively. On the Adamawa dataset, Multiple Linear Regression showed the most consistent performance across both scenarios. This model maintained the highest R^2 score after cross-validation, suggesting it generalizes unseen data better for the Adamawa state dataset. As such, it demonstrates the best balance between fitting the training data and generalizing to new data. XGBoost performed best before cross-validation but showed the most significant drop in performance after cross-validation. This substantial decrease suggests that XGBoost was overfitting to the training data and struggled to generalize well to the Adamawa state dataset. KNN showed the most stable performance regarding the R^2 score, with the smallest decrease after cross-validation, which is a major improvement relative to other models after cross-validation, especially regarding MSE and RMSE. Multiple Linear Regression appears to be the most stable and reliable model for this specific dataset, followed by KNN. XGBoost, despite its initial strong performance, seems least suitable for generalizing to new data in this case.

The significant drop in performance metrics after cross-validation on the Cross-River state dataset suggests that all models overfitted the training data before cross-validation, as shown in Table 4. This is particularly evident in the Multiple Linear Regression and XGBoost models. KNN appears to be the most stable model across both scenarios, maintaining a relatively consistent performance. However, its performance is still not ideal. XGBoost performed best before cross-validation but performed poorly after cross-validation, even yielding a negative R^2 score. This indicates severe overfitting and poor generalization. While it performed best before cross-validation, it dropped significantly after, suggesting it was also overfitting to the training data. Based on the cross-validation results, KNN appears to be the best-performing model, with the highest R^2 score and lowest error metrics overall, followed by the Multiple Linear Regression.

The decrease in R^2 scores and increase in error metrics after cross-validation indicate suboptimal performance across all models on both datasets. This underperformance can be largely attributed to the limited number of observations, which represents a significant limitation of this study. The results underscore the importance of larger, more comprehensive datasets for future research. A broader data collection would provide a more robust foundation for model training, potentially leading to improved predictive accuracy and generalizability. Future studies should prioritize acquiring larger samples to overcome this limitation and enhance the reliability of their findings.

5. Conclusions and Future Work

This paper aimed to utilize the Multiple Linear Regression machine learning algorithm in predicting rice yield in two distinct geo-climatic regions in Nigeria, namely: Adamawa State and Cross River State, while optimizing its performance using recursive feature elimination.

Extensive data on rice yields and weather patterns were obtained. These datasets underwent preprocessing, cleaning, and separation into training as well as testing sets. The data for each region was trained and tested using the Multiple Linear Regression Algorithm. A complete range of model evaluation metrics like mean squared error, R-squared and mean absolute error were computed to evaluate the predictive accuracy. XGBoost and K nearest Neighbours algorithms were also trained with the same datasets, and their performances were evaluated with and without feature selection. Multiple Linear Regression performed excellently well across both geographic regions when it came to yield prediction precision.

As climate change continues to impact agricultural systems globally, applying machine learning algorithms offers valuable insights and tools to address challenges in food production. Rice, a staple crop worldwide and in Nigeria, is crucial in tackling food insecurity and hunger crises. Timely and accurately predicting rice yields across different regions of Nigeria can provide invaluable information to improve overall rice production and ensure food security.

However, to fully harness the potential of machine learning in agricultural modeling, there is a pressing need for systematic and continuous data collection and storage of relevant agricultural, climatic, and socio-economic variables. Establishing robust and comprehensive databases will be a valuable resource for future studies in this domain, enabling more advanced analyses and developing even more sophisticated predictive models.

Sustained efforts in data gathering, coupled with ongoing research in machine learning techniques tailored for agricultural applications, will not only enhance our understanding of the complex interplay between various factors influencing crop yields but also empower stakeholders with actionable insights to make informed decisions and implement effective strategies for sustainable and resilient food production systems.

Author Contributions: Conceptualization: JAI, ASN, and AI; methodology, AI and JAI; software: JAI and AI; validation: AI and ASN.; investigation: AI and JAI.; Datasets and other resources: JAI, ASN and AI; writing—original draft preparation: JAI and AI.; writing—review and editing: JAI, ASN, and AI.; visualization: AI and JAI.; supervision: ASN and AI; project administration: JAI, ASN, and AI.

Funding: This research received no external funding**.**

Data Availability Statement: The datasets and codes can be found in the following GitHub link: https://github.com/Diczen/ML_Rice_Yield_Project.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] W. Rosa, Ed., "Transforming Our World: The 2030 Agenda for Sustainable Development," in *A New Era in Global Health*, New York, NY: Springer Publishing Company, 2017, pp. 529–567. doi: 10.1891/9780826190123.ap02.

[2] FAO, IFAD, UNICEF, WFP, and WHO, *The State of Food Security and Nutrition in the World 2023*. FAO; IFAD; UNICEF; WFP; WHO;, 2023. doi: 10.4060/cc3017en.

[3] B. Das, B. Nair, V. K. Reddy, and P. Venkatesh, "Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India," *Int. J. Biometeorol.*, vol. 62, no. 10, pp. 1809–1822, Oct. 2018, doi: 10.1007/s00484-018-1583-6.

[4] S. S. Gnanamanickam, "Rice and Its Importance to Human Life," in *Biological Control of Rice Diseases*, Dordrecht: Springer Netherlands, 2009, pp. 1–11. doi: 10.1007/978-90-481-2465-7_1.

[5] R. S. Zeigler and A. Barclay, "The Relevance of Rice," *Rice*, vol. 1, no. 1, pp. 3–10, Sep. 2008, doi: 10.1007/s12284-008-9001-z.

[6] K. Gyimah-Brempong, M. Johnson, and H. Takeshima, "Chapter 1. Rice in the Nigerian Economy and Agricultural Policies," in *The Nigerian Rice Economy*, K. Gyimah-Brempong, M. Johnson, and H. Takeshima, Eds. Philadelphia: University of Pennsylvania Press, 2016, pp. 1–20. doi: 10.9783/9780812293753-005.

[7] N. Kamai, L. O. Omoigui, A. Y. Kamara, and F. Ekeleme, "Guide to Maize Production in Northern Nigeria," International Institute of Tropical Agriculture, 2020.

[8] D. D. Sasu, "Production of milled rice in Nigeria 2010-2023," *Statista*, 2023. https://www.statista.com/statistics/1134510/production-of-milled-rice-in-nigeria/ (accessed Sep. 26, 2023).

[9] U. U. Okonkwo, V. Ukaogo, D. Kenechukwu, V. Nwanshindu, and G. Okeagu, "The politics of rice production in Nigeria: The Abakaliki example, 1942-2020," *Cogent Arts Humanit.*, vol. 8, no. 1, Jan. 2021, doi: 10.1080/23311983.2021.1880680.

[10] D. I. Onwude *et al.*, "Mechanization of Agricultural Production in Developing Countries," in *Advances in Agricultural Machinery and Technologies*, CRC Press, 2018, pp. 3–26. doi: 10.1201/9781351132398-1.

[11] A. Karasev, "Excursıon to The Hıstory of Tractor Buıldıng and The Introductıon of Tractors in Agrıculture," *Tekhnicheskiy Serv. mashin*, vol. 1, 2023, doi: 10.22314/2618-8287-2023-61-1-155-163.

[12] N. ElBeheiry and R. S. Balog, "Technologies Driving the Shift to Smart Farming: A Review," *IEEE Sens. J.*, vol. 23, no. 3, pp. 1752–1769, Feb. 2023, doi: 10.1109/JSEN.2022.3225183.

[13] A. Sharma, A. Jain, P. Gupta, and V. Chowdary, "Machine Learning Applications for Precision Agriculture: A Comprehensive Review," *IEEE Access*, vol. 9, pp. 4843–4873, 2021, doi: 10.1109/ACCESS.2020.3048415.

[14] F. J. Pierce and P. Nowak, "Aspects of Precision Agriculture," in *Advances in Agronomy*, D. L. Sparks, Ed. 1999, pp. 1–85. doi: 10.1016/S0065-2113(08)60513-1.

[15] T. van Klompenburg, A. Kassahun, and C. Catal, "Crop yield prediction using machine learning: A systematic literature review," *Comput. Electron. Agric.*, vol. 177, no. August, p. 105709, Oct. 2020, doi: 10.1016/j.compag.2020.105709.

[16] A. K. Mariappan and J. A. Ben Das, "A paradigm for rice yield prediction in Tamilnadu," in *2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR)*, Apr. 2017, pp. 18–21. doi: 10.1109/TIAR.2017.8273679.

[17] J. T. Ritchie, U. Singh, D. C. Godwin, and W. T. Bowen, "Cereal growth, development and yield," in *Understanding Options for Agricultural Production*, 1998, pp. 79–98. doi: 10.1007/978-94-017-3624-4_5.

[18] S. Khaki and L. Wang, "Crop Yield Prediction Using Deep Neural Networks," in *Smart Service Systems, Operations Management, and Analytics*, 2020, pp. 139–147. doi: 10.1007/978-3-030-30967-1_13.

[19] J. van Wart, K. C. Kersebaum, S. Peng, M. Milner, and K. G. Cassman, "Estimating crop yield potential at regional to national scales," *F. Crop. Res.*, vol. 143, pp. 34–43, Mar. 2013, doi: 10.1016/j.fcr.2012.11.018.

[20] P. J. A. van Asten, S. Kaaria, A. M. Fermont, and R. J. Delve, "Challenges and lessons when using farmer knowledge in agricultural research and development projects in Africa," *Exp. Agric.*, vol. 45, no. 1, pp. 1–14, Jan. 2009, doi: 10.1017/S0014479708006984.

[21] D. Paudel *et al.*, "Machine learning for large-scale crop yield forecasting," *Agric. Syst.*, vol. 187, p. 103016, Feb. 2021, doi: 10.1016/j.agsy.2020.103016.

[22] U. Patrio, Y. Yuliska, and Y. Lulu Widyasari, "Predicting Rice Production In Sumatra Island Using Linear Regression," in Proceedings of the 11th International Applied Business and Engineering Conference, ABEC 2023, September 21st, 2023, Bengkalis, Riau, Indonesia, 2024. doi: 10.4108/eai.21-9-2023.2342997.

[23] I. B. Iorliam, B. A. Ikyo, A. Iorliam, E. O. Okube, K. D. Kwaghtyo, and Y. I. Shehu, "Application of Machine Learning Techniques for Okra Shelf Life Prediction," *J. Data Anal. Inf. Process.*, vol. 09, no. 03, pp. 136–150, 2021, doi: 10.4236/jdaip.2021.93009.

[24] E. A. Jiya, U. Illiyasu, and M. Akinyemi, "Rice Yield Forecasting: A Comparative Analysis of Multiple Machine Learning Algorithms," *J. Inf. Syst. Informatics*, vol. 5, no. 2, pp. 785–799, Jun. 2023, doi: 10.51519/journalisi.v5i2.506.

[25] National Bureau of Statistics, "Demographic Statistics Bulletin 2013," 2013. [Online]. Available: https://nigerianstat.gov.ng/download/1241121

[26] Adamawa State Planning Commission, "Adamawa State – Adamawa State Planning Commission," 2024. https://adspc.ad.gov.ng/adamawa-state/ (accessed Feb. 29, 2024).

[27] A. A. Adebayo and A. L. Tukur, *Adamawa State in Maps*. Paraclete Publishers, 1999. [Online]. Available: https://books.google.co.id/books?id=eFTVAAAACAAJ

[28] D. Maulud and A. M. Abdulazeez, "A Review on Linear Regression Comprehensive in Machine Learning," *J. Appl. Sci. Technol. Trends*, vol. 1, no. 2, pp. 140–147, Dec. 2020, doi: 10.38094/jastt1457.

[29] A. Peˇckov, "A Machıne Learnıng Approach to Polynomıal Regressıon," Joˇzef Stefan International Postgraduate School, 2012. [Online]. Available: https://slais.ijs.si/wp-content/uploads/2021/07/Doktorske/phd_aleksandar_peckov.pdf

[30] D. R. I. M. Setiadi, S. Widiono, A. N. Safriandono, and S. Budi, "Phishing Website Detection Using Bidirectional Gated Recurrent Unit Model and Feature Selection," *J. Futur. Artif. Intell. Technol.*, vol. 2, no. 1, pp. 75–83, 2024, doi: 10.62411/faith.2024-15.

[31] F. Omoruwou, A. A. Ojugo, and S. E. Ilodigwe, "Strategic Feature Selection for Enhanced Scorch Prediction in Flexible Polyurethane Form Manufacturing," *J. Comput. Theor. Appl.*, vol. 1, no. 3, pp. 346–357, Feb. 2024, doi: 10.62411/jcta.9539.

[32] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Electr. Eng.*, vol. 40, no. 1, pp. 16–28, Jan. 2014, doi: 10.1016/j.compeleceng.2013.11.024.

[33] E. B. Wijayanti, D. R. I. M. Setiadi, and B. H. Setyoko, "Dataset Analysis and Feature Characteristics to Predict Rice Production based on eXtreme Gradient Boosting," *J. Comput. Theor. Appl.*, vol. 1, no. 3, pp. 299–310, Feb. 2024, doi: 10.62411/jcta.10057.