Research Article GU J Sci, Part A, 11(3): 481-496 (2024) 10.54287/gujsa.1503494

Gazi University
Journal of Science
PART A: ENGINEERING AND INNOVATION
http://dergipark.org.tr/gujsa

Comparing Machine Learning Algorithms for Rice Yield Prediction in Adamawa and Cross Rivers States of Nigeria

Joseph Abunimye INGIO[1] Augustine Shey NSANG[1] Aamo IORLIAM[2*]

[1] Department of Computer Science, SITC, AUN, Yola, Nigeria
[2] Data Science Department, SITC, AUN, Yola, Nigeria

Keywords

Machine Learning Algorithms

Cross River

Adamawa

Prediction

Abstract

Rice production is critical for global food security, and accurate yield prediction empowers informed decision-making. This paper investigates machine learning (ML) techniques for rice yield prediction in Adamawa and Cross River states, with distinct agroclimatic conditions. Traditional yield prediction methods that are commonly used often have limitations such as less insights into the available data and reduced accuracy. Hence, this research explores the potential of machine learning for improved prediction accuracy. We leverage climatic data and historical rice yields to train and evaluate Decision Trees, Random Forest, Support Vector Regressor, Polynomial Regressor, Multiple Linear Regression and Long Short-Term Memory (LSTM) models. Performance is compared using Mean Squared Error, Root Mean Squared Error, Coefficient of Determination, Mean Absolute Error, and Mean Absolute Percentage Error. Feature selection identifies All-sky Photosynthetically Active Radiation (PAR) as the most influential factor. Linear Regression emerges as the superior model, achieving an R² of 0.90 (Adamawa) and 0.91 (Cross River), demonstrating robust generalizability across regions. This research contributes to the development of ML-powered Agro-information systems for two Nigerian regions, enhancing agricultural practices and food security.

Cite

Ingio, J. A., Nsang, A. S., & Iorliam, A. (2024). Comparing Machine Learning Algorithms for Rice Yield Prediction in Adamawa and Cross Rivers States of Nigeria. *GU J Sci, Part A*, *11*(3), 481-496. doi:10.54287/gujsa.1503494

Author ID (ORCID Number)		Article Process	
0009-0007-3574-8560	Joseph Abunimye INGIO	**Submission Date**	23.06.2024
0000-0002-6466-9032	Augustine Shey NSANG	**Revision Date**	11.07.2024
0000-0001-8238-9686	Aamo IORLIAM	**Accepted Date**	06.08.2024
		Published Date	28.09.2024

1. INTRODUCTION

The role of agriculture and 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 (Rosa, 2017). 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 (FAO et al., 2023). This shows an obvious need for the production of more food, particularly the most important and widely eaten ones, to meet the global demand which is increasing at a rapid pace.

Rice is a food crop that is consumed by a great number of people constituting over half of the world's population (Gnanamanickam, 2009; Das et al., 2018) and it has been termed "the world's most important food crop" (Zeigler & Barclay, 2008). 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 its widespread consumption globally (Gnanamanickam, 2009). 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.

*Corresponding Author, e-mail: aamo.iorliam@aun.edu.ng

In Nigeria, rice has emerged as a staple food over the past few decades, enjoyed in every part of the country (Gyimah-Brempong et al., 2016; Kamai et al., 2020). 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 which constitutes 46% of the total rice produced in Africa (Sasu, 2023). Nevertheless, this rate of production is insufficient to meet the nation's rising rice demand, which has increased reliance on rice importation to satisfy the teeming population of rice eaters in the country. In 2014 half the quantity of rice consumed in Nigeria was imported (Gyimah-Brempong et al., 2016) and in 2018 over 7 trillion Naira was spent importing rice into Nigeria (Okonkwo et al., 2021). The surge in demand can be attributed to various factors, including shifts in consumer preferences, population growth, growing incomes and a swift urbanization process (Kamai et al., 2020). The rice being produced in Nigeria is cultivated in about 21 states in Nigeria with 8 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 Northern region of the country and they include Kebbi, Kaduna, Kano, and Borno while a few are in the Southern region, and they include Cross River, and Ebonyi. Cultivation of rice is usually done in rainfed lowland fields as well as rainfed highland fields during the raining season which typically spans between May and August. This, however, spells some challenges for rice farmers in northern Nigeria as about 1200mm to 1600mm of rainfall is needed for optimum growth of rice and this volume of rainfall does not occur in the North. In addition, there is the challenge of pest infestation and poor soil fertility as a result of increased pressure on land resources due to population expansion (Kamai et al., 2020). Farmers in the southern region of Nigeria who cultivate rice also face some challenges despite the increased volume of rainfall in the region when compared to the north. This is because of the general inconsistency of rainfall as in times past hence there is a risk of planting and not having enough rain for proper growth of seeds. These challenges and trends in agriculture have been captured in comprehensive datasets that are available for the 36 states in Nigeria.

The Agricultural data available can provide valuable insights on 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 through in an attempt to uncover new patterns and relationships (Geetha, 2015; Vanitha et al., 2019) with a goal to extract knowledge from the data and convert it to a human understandable format. This constitutes a major preliminary step towards the application of 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 taken into account when building machine learning models for crop prediction thereby increasing the predictability and accuracy of the predicted yield.

The integration of machine learning into agriculture holds promise as it can bring advantages. One major benefit is the ability to make predictions, which helps reduce errors that occur 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 challenges previously mentioned, such as bridging the gap, between rice demand and production in Nigeria.

This research aimed at comparing and evaluating the performance of different machine learning algorithms for rice yield prediction in two distinct geo-climatic regions of Nigeria (Adamawa and Cross River States). Our specific objectives were to:

(i) identify key features and potential discrepancies in the rice yield and climate datasets from the two regions through exploratory data analysis. This helped in understanding the data's characteristics and any potential biases.

(ii) build and compare the performance of six machine learning models namely: Decision Trees, Random Forest, Support Vector Regressor, Polynomial Regressor, Multiple Linear Regression and Long Short-Term Memory (LSTM) which were implemented in Python using a Jupyter notebook.

(iii) evaluate the performance of models implemented in (ii) using standard metrics like the Coefficient of Determination (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). This allowed us to compare the effectiveness of these different algorithms in predicting rice yields.

By achieving these objectives, this research contributed valuable insights into the feasibility and effectiveness of machine learning teachniques for rice yield prediction in Adamawa and Cross River States of Nigeria for the first time. Furthermore, this study's findings lays the groundwork for a crop yield prediction system utilizing the best-performing machine learning model, in our case Linear Regression. This system can empower farmers and decision-makers to optimize resource allocation and improve agricultural planning. As such, increasing rice production and enhancing food security across the nation. The rest of this paper is organized into these sections; Section two presents a Literature Review, Section three explains the methodology employed in the research, section four deals with the results and discussions, and section five presents the conclusion, implication and future direction.

2. REVIEW OF RELATED STUDIES

Agricultural processes have long been carried out manually and much of it is still done that way in most developing countries including Nigeria. In sub-Saharan Africa, up to 65% of farming is done manually, about 25% is done using animal traction (donkeys, bulls' carts etc.) and about 10% is mechanized (Onwude et al., 2018). As a result, farming is seen to be a laborious task. This notion continued until the introduction of mechanized farming and a widespread introduction of tractors into land processing, facilitated by the shortage of food, workers, and draft animals caused by the World War (Karasev, 2023). 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 (Onwude et al., 2018). 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 (Elbeheiry & Balog, 2022). 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 (Sharma et al., 2021). 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 (Pierce & Nowak, 1999; Sharma et al., 2021). 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 (Figure 1) can be categorized into five groups according to Pierce and Nowak (1999) – 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 (van Klompenburg et al., 2020) because there are several parameters that contribute to the optimum yield of a particular crop specie 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 (Mariappan & Ben Das, 2017), process-based crop simulation models grounded in agronomic principles like the CERES model (Ritchie et al., 1998), 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 (Khaki & Wang, 2020). Crop models are data-hungry, requiring extensive inputs that may not be available, and make assumptions that restrict their generalizability (Wart et al., 2013).

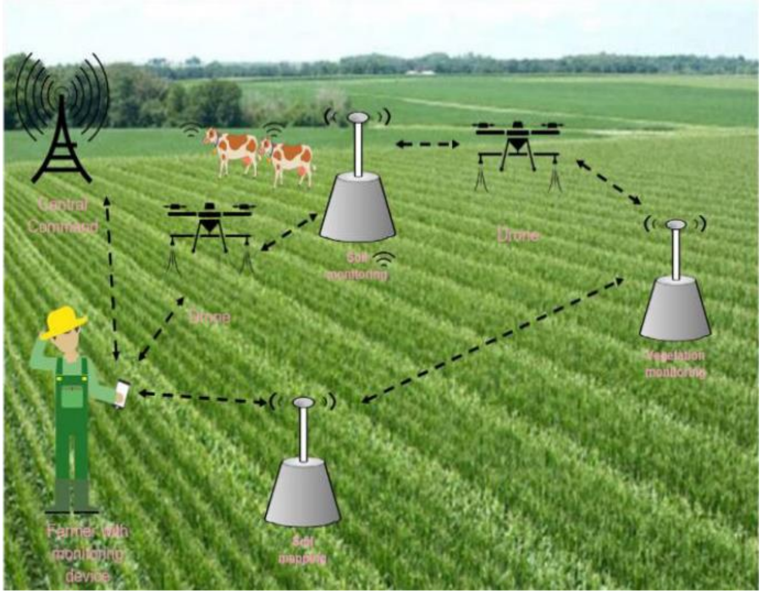


Figure 1. *Precision Agriculture source: (Sharma et al., 2021)*

Traditional farmer knowledge is grounded in local experience but can lack quantitative rigor and predictive precision (Van Asten et al., 2009). 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 into leveraging machine learning techniques as an alternative, data-driven approach for developing more accurate and robust yield prediction models.

van Klompenburg et al. (2020) performed a detailed review of literature based 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 occur in literature relating crop yield prediction. The research concluded that Deep learning algorithms such as Convolutional Neural Networks (CNN) was widely used 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. And the most used evaluation parameters include R^2 (Coefficient of Determination), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

Another major contributor to this research domain is Paudel et al. (2021). In their research, they developed a machine learning workflow that can be used for large scale prediction of crop yield. 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. (2024**)** 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 R2-score of 85.53%.

In Nigeria, Iorliam et al. (2021) utilised machine learning techniques like Logistic Regression, Support Vector Machine, K-Nearest Neighbour, Naïve Bayes, and Decision Tree, for Okra shelf life prediction and observed that Support Vector Machine, Naïve Bayes, and Decision Tree excellently predicted the shelf life of Okra better as compared to the other machine learning techniques they used.

Jiya et al. (2023) 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 Forest and Random Trees demonstrated better performance in predicting rice yields, 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 most of the machine learning algorithms we utilised were different from Jiya et al. (2023). This research is therefore motivated by Iorliam et al. (2021) and Jiya et al. (2023) with a focus on predicting rice yield in Adamawa State and Cross River State of Nigeria using machine learning techniques (Decision Trees, Random Forest, Support Vector Regressor, Polynomial Regressor, Multiple Linear Regression and LSTM).

3. METHODOLOGY

Our methodology consists of six phases and is described below and summarized in Figure 2:

i. Data Collection – Secondary data from NASA POWER and the National Bureau of Statistics database (NBS) was utilised in this study.

ii. Data Exploration phase – This stage involved visualizing the data using charts to comprehend the data.

iii. Data Preprocessing phase – This phase involved data cleaning (removing missing values/outliers) and getting the right features for our proposed models.

iv. Model Development phase - The Decision Trees, Random Forest, Support Vector Regressor, Polynomial Regressor, Multiple Linear Regression and LSTM were implemented using appropriate libraries and tools in python.

v. Model Evaluation phase – Systematically evaluating model performance using metrics like RMSE, MAE, R-squared based on train/test splits.

vi. Model Optimization phase – In this phase hyperparameters were tuned to achieve the best performing model for rice prediction.

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

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 (NBS, 2020; ADSPC, 2022). The tropical climate in the state experiences wet and dry seasons with an average annual rainfall ranging from 75 -103 mm, concentrated in the wet season months of May to September. Mean annual temperatures vary from 22°C to 31°C (Adebayo, 1999; ADSPC, 2022). 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 (NBS, 2020). The state has abundant rainfall exceeding 3036 mm annually, along with 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 favourable 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 a substantial amount to the total quantity of rice produced in Nigeria. Comparing 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 algorithms. Any model that consistently performs well in both locations is likely to generalize effectively to other rice-growing regions of Nigeria. The multi-year timeseries data from the two states also enables training 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.

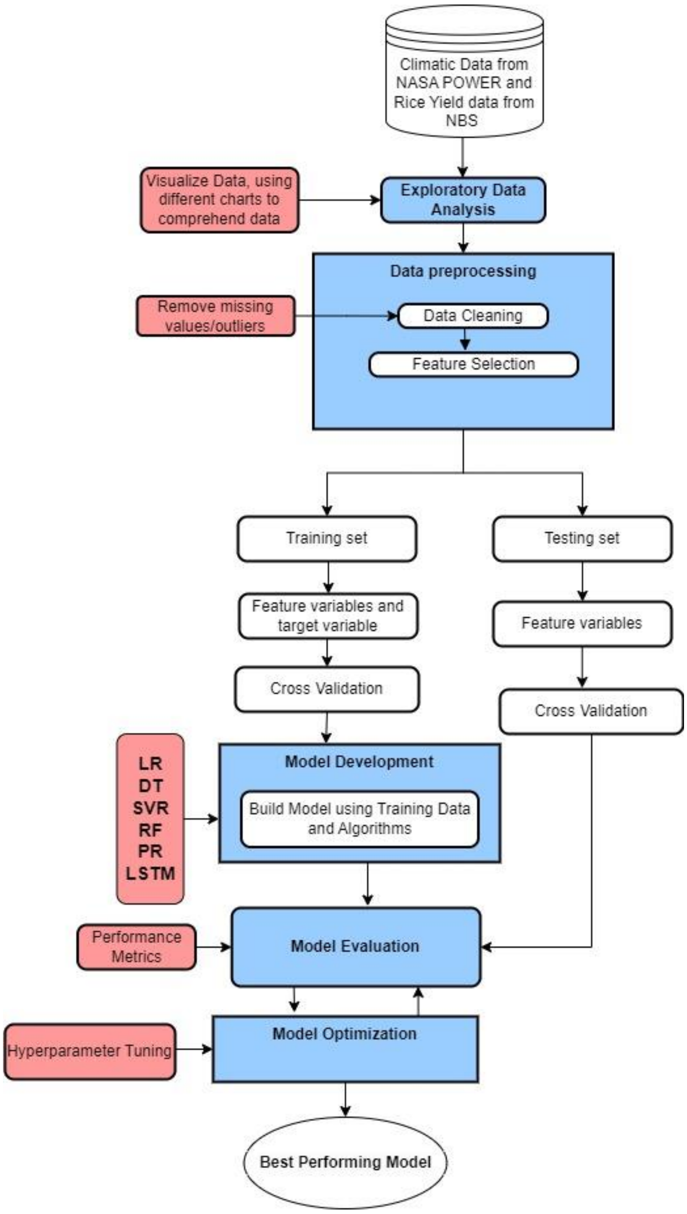

Climatic Data from NASA POWER and Rice Yield data from NBS
Visualize Data, using different charts to comprehend data
Exploratory Data Analysis
Data preprocessing
Remove missing values/outliers
Data Cleaning
Feature Selection
Training set
Testing set
Feature variables and target variable
Feature variables
Cross Validation
Cross Validation
LR
DT
SVR
RF
PR
LSTM
Model Development
Build Model using Training Data and Algorithms
Performance Metrics
Model Evaluation
Hyperparameter Tuning
Model Optimization
Best Performing Model

***Figure 2**. Proposed Methodology workflow.*

3.1.2. Data Collection

Annual rice yield data (tons/hectare) for the period 1997 to 2020 was collected for each state from the National Bureau of Statistics database. The NBS data is compiled from state-level agricultural production surveys and provides authoritative aggregated statistics on crop yields.

Corresponding climatic data was obtained from the National Aeronautics and Space Agency(NASA)’s POWER (Prediction of Worldwide Energy Resource) 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 NBS yield data was combined with the 18-year POWER climatic data for each state to compile the input dataset that was used in training machine learning models and also used in the testing too. The dataset was screened for any missing values and outliers. Rows with missing values were removed.

3.2. Machine Learning Algorithms Utilised

3.2.1. Decision Trees (Regression trees)

A decision tree is a binary tree that separates data into pure leaf nodes, or data that belong to a single class (homogeneous class), repeatedly. The decision node (parent) and the leaf node(child) are the main components in a decision tree. Leaf nodes determine the class of a new data point, while decision nodes carry a condition to split data into them. For regression analysis, this process continues until each class has just one leaf. The information gain of a node is measured by the Variance Reduction, and the trees use this information to determine which decision node to select (Chauhan, 2022). We adapt the Regression decision trees approach presented by Veenadhari et al. (2014) for rice yield dataset as follows:

1. For all rice datapoints, examine potential base cases.

   - Case 1: If all rice data points have the same value for the target variable, Return a leaf node with the predicted value as the average of the target attribute in the data.
   - Case 2: If no attributes remain, return a leaf node with the predicted value as the average of the target attribute in the data

2. For each attribute a,

   - Calculate the normalized information gain for splitting the data based on a.

3. Choose the attribute best_a with the highest normalized information gain.

4. Create a decision node that splits the data based on the value of best_a.

5. For each possible value v of a_best:

   - Create a child node by recursively calling BuildRegressionTree on the subset of data where best_a has the value v.
   - Set the child node as a branch of the current decision node.

6. Return the root node

3.2.2. Support Vector Regressor (SVR)

SVR is a machine learning algorithm applied in regression analysis in a similar manner to how Support Vector Machines are used in Classification tasks. SVR has the capacity to carry out nonlinear multivariate regression with remarkable robustness and efficiency. (Cortes & Vapnik, 1995). The fundamental idea is that a linear relationship in a higher dimensional space can describe a complex nonlinear relationship between some variables. This is possible through the application of linear optimization techniques to the projection of variables of interest into a high-dimensional space. The regression function evaluated, is then applied back to the low-dimensional phase space of the variables that were first observed (Oguntunde et al., 2018). The support vector regressor (Drucker et al., 1996) is given by.

$$y = f(x) + \epsilon. \tag{1}$$

The function $f(x)$ is estimated by:

$$f(x) = \sum_{i=1}^{n} y_i \alpha_i K(x_i, x) + b, \tag{2}$$

where K (xi, xj) = kernel function,

α_i = vector of weight of i^{th} point (Lagrange Multiplier)

b = constant scalar, and ϵ = error term.

3.2.3. Polynomial Regression

This is performed by regressing the dependent variable on the powers of the independent variable (Ostertagová, 2012).

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \cdots + \beta_k x_i^k + e_i \text{ , for } i = 1, 2, \ldots, n \tag{3}$$

the polynomial’s degree is denoted by k and it is the order of the model. This is basically the same as having multiple linear regression models with $X_1 = X$, $X_2 = X^2$, $X_3 = X^3$, etc

3.2.4. Random Forest

Random Forest Regressor is an algorithm used for regression machine learning tasks. It belongs to the ensemble learning family, specifically based on the Random Forest algorithm. It is constructed using multiple decision trees on different subsets of a given dataset, averaging their outcomes to raise the dataset's estimated accuracy (Mwiti, 2022). Mwiti (2022) explained that:

- A distinct sample of rows is used to create each tree, and another sample of features is chosen for splitting at each node.
- Every single tree makes a unique prediction.

To arrive at a single result, these predictions are then averaged.

3.2.5. Multiple Linear Regression

Multiple Linear Regression, also known as Multilinear Regression, is a machine learning algorithm that utilizes statistical regression analysis in predicting the value of an output variable based on a set of input 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 (Seber & Lee, 2003). 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 (Maulud & Abdulazeez, 2020). Forecasting in regression analysis occurs when an equation of the form:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon = \beta_0 + \sum_{i=1}^{p} X_i . \beta_i + \varepsilon \tag{4}$$

Where $X_1, X_2, X_3, \dots, X_p$ are the independent variables or features used to predict the dependent variables or target variable: y is evaluated, where ε is an random variable that cannot be observed, and is also referred to as the error component, with mean 0 and variance σ^2 .

The relationship described by (4) is known as a multiple linear regression model, β_0 is the intercept, $\beta_1 \dots \beta_p$ are the slope coefficents for each independent variable and $\sigma^2 > 0$ is an unknown error variance (Pečkov, 2012).

3.2.6. Long short-term Memory (LSTM)

LSTM is a deep learning algorithm that is ideally used in regression analysis and works well with time series data. It is a variant of recurrent neural network (RNN) that has the ability to learn long-term dependencies in time series or sequence data and is mostly used for forecasting (Arras et al., 2019; Özdoğan-Sarıkoç et al., 2023). This is because standard RNNs have limited memory capacity and struggle to learn dependencies between sequence elements that are separated by long gaps. To address this limitation, the long short-term memory (LSTM) network was developed as an extension of RNNs, constructed from specialized memory blocks or cells. These LSTM cells act as memory units, with the specific purpose of retaining information over extended periods. By maintaining an internal cell state, LSTMs can preserve knowledge of past context to better link widely spaced events and make predictions informed by long-range sequential correlations. The explicit memory of LSTM cells provides the ability to learn temporal dependencies that conventional RNNs lack (Imani, 2019).

4. RESULTS AND DISCUSSION

The results derived from applying the phased methodology described in the previous section on the datasets gotten from Adamawa and Cross river states respectively are presented in this section.

4.1. EDA and Pre-processing

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 the relationship between the features and the variable in terms of correlation and which features are most important for prediction, a correlation matrix for each of the states was generated as shown in Figure 3 and 4.

Figure 3 shows a positive correlation between specific humidity (sp_humidity) and precipitation (precipitation). The coefficient of correlation between these variables is 0.72 suggesting an increase in sp_humidity whenever precipitation increases.

A strong positive correlation also exists between re_humidity and sp_humidity (0.90) and with precipitation (0.84) suggesting 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 choices as predictors.

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

From Figure 4, 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" with 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.

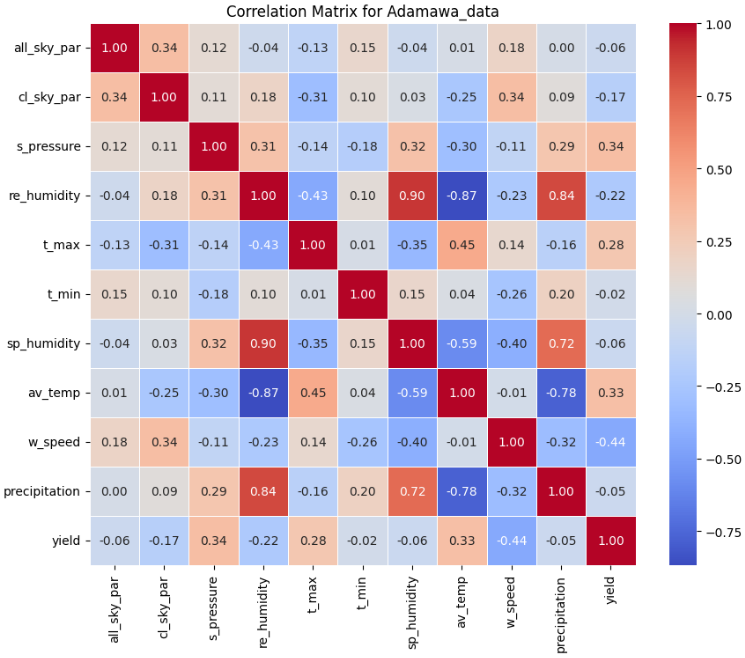


***Figure 3**. Correlation Matrix for Adamawa State Data*

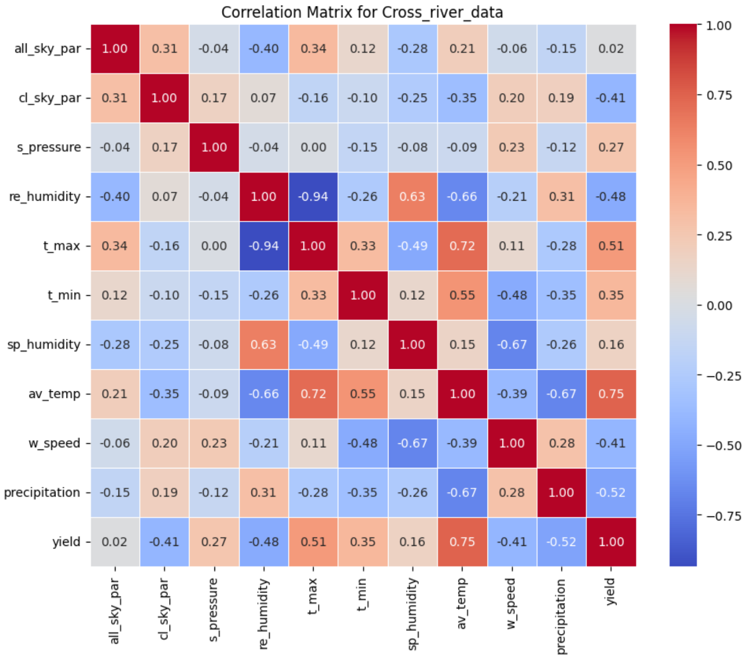


***Figure 4**. Correlation Matrix for Cross River State Data*

4.1.2. Feature Selection

A Recursive feature elimination process was carried out to select the most relevant features that will best predict the target variable and the features selected were; s_pressure, t_max, av_temp, w_speed, all_sky_par, t_min. A similar set of features were selected when a different feature selection (F-regression) technique was used. The F-regression selection technique reduces the dimensionality of data by selecting a subset containing the most relevant features for our regression models based on their F-value scores or scores from analysis of variance. The following features were selected: s_pressure, re_humidity, sp_humidity, av_temp, w_speed, precipitation, cl_sky_par, t_min.

However, several iterations of the training and testing of the selected algorithms did not yield a good performance using the features listed above. Hence, the number of features to select was reduced to 5 (for RFE: n_features_to_select= 5, and k = 5 F-regression). Both feature selection techniques presented the following features as the most relevant: all_sky_par, re_humidity, t_max, w_speed, s_pressure. Therefore, these features where used in the building the models for rice yield prediction.

4.2. Model Evaluation

The following performance metrics were applied in evaluating the performance of the Models built using the machine learning algorithms mentioned in previous sections. Below is a brief description of each metric and its significance.

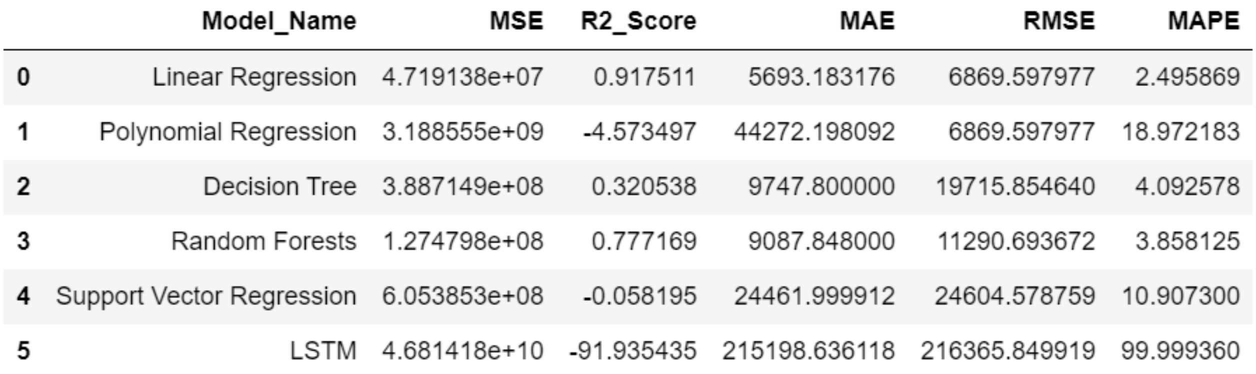
- **MSE:** This stands for Mean Squared Error. It is a measure of the average squared difference between the predicted rice yields and the actual yields. Lower MSE indicates a better fit between predictions and actual values.
- **R2_Score:** This is R-squared also referred to as the coefficient of determination. It is the variance (squared correlation) in the dependent variable (rice yield) that can be explained by the independent variables (features used in the model). Values closer to 1 generally 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.

The metrics mentioned above were used to measure the accuracy of each model's performance in predicting the yield for both states and the following section contains an analysis of the results.

4.3. Analysis of Evaluation Results

Table 1 and 2 contains a concise summary of the values for each model's performance capture using the metrics mention in the previous section. Providing insights on the accuracy of each model built during this study.

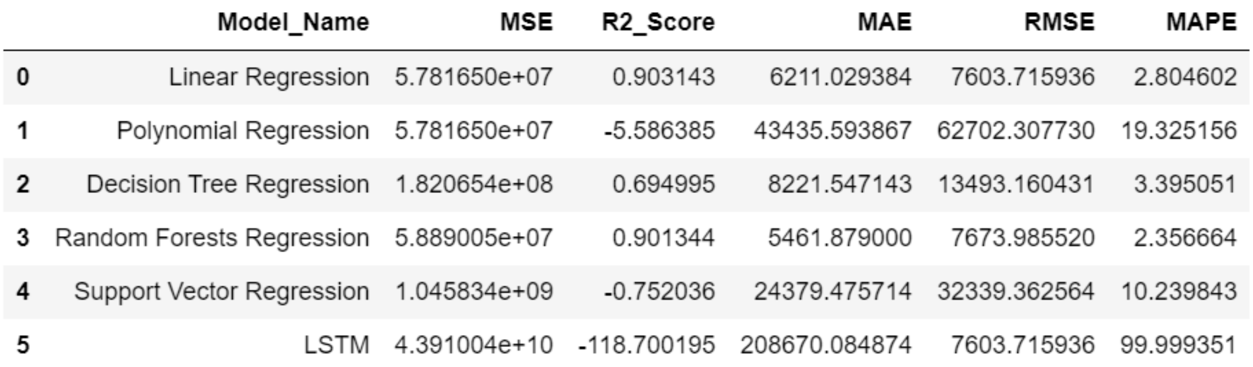
Table 1. *Results from Cross River Dataset*



	Model_Name	MSE	R2_Score	MAE	RMSE	MAPE
0	Linear Regression	4.719138e+07	0.917511	5693.183176	6869.597977	2.495869
1	Polynomial Regression	3.188555e+09	-4.573497	44272.198092	6869.597977	18.972183
2	Decision Tree	3.887149e+08	0.320538	9747.800000	19715.854640	4.092578
3	Random Forests	1.274798e+08	0.777169	9087.848000	11290.693672	3.858125
4	Support Vector Regression	6.053853e+08	-0.058195	24461.999912	24604.578759	10.907300
5	LSTM	4.681418e+10	-91.935435	215198.636118	216365.849919	99.999360

Based on the results displayed in Table 1, Multiple Linear Regression outperforms the other models in predicting the yield using the Cross River state dataset, with an MSE value of 4.72e+07 which appears to be the lowest among the listed models. It also has a relatively high R2_Score (0.91), indicating a good fit between predictions and actual yields. Additionally, Multiple Linear Regression has a lower MAE (5,693.18) and RMSE (6,869.59) compared to other models, suggesting a good balance between underestimation and overestimation of rice yields. The MAPE value for Multiple Linear Regression is also relatively low (2.49%) showing a low average percentage deviation between the predicted and actual values. Its performance is closely followed by that that of Random Forest which demonstrated a moderately high predictive accuracy with and MSE value of 1.274E+8, an R2-score of 0.77 and relatively low values for MAE, RMSE and MAPE.

Table 2. *Results from Adamawa Dataset*



	Model_Name	MSE	R2_Score	MAE	RMSE	MAPE
0	Linear Regression	5.781650e+07	0.903143	6211.029384	7603.715936	2.804602
1	Polynomial Regression	5.781650e+07	-5.586385	43435.593867	62702.307730	19.325156
2	Decision Tree Regression	1.820654e+08	0.694995	8221.547143	13493.160431	3.395051
3	Random Forests Regression	5.889005e+07	0.901344	5461.879000	7673.985520	2.356664
4	Support Vector Regression	1.045834e+09	-0.752036	24379.475714	32339.362564	10.239843
5	LSTM	4.391004e+10	-118.700195	208670.084874	7603.715936	99.999351

From the results in Table 2 Multiple Linear Regression also demonstrated the best performance on the Adamawa dataset when compared with the other algorithms. It has the lowest MSE (5.78e+07), and a relatively high R2_Score (0.903), indicating a very good fit between predictions and actual yields. Additionally, Multiple Linear Regression has a lower RMSE (7,673) compared to other models, suggesting a good balance between underestimation and overestimation of rice yields.

However, the MAPE and MAE value for Multiple Linear Regression seems to be slightly higher than that of Random Forest which has the lowest values of 5461.87 for MAE and 2.35% for MAPE, showing a low average percentage deviation between the predicted and actual values. Multiple Linear regression achieved the best performance in 3 out of the five metrics hence it is considered the best performing algorithm on the Adamawa dataset closely followed by Random Forest algorithm.

5. CONCLUSION, IMPLICATION AND FUTURE DIRECTION

This paper aimed at comparing the performance of five different machine learning algorithms in predicting rice yield in two distinct geo-climatic regions in Nigeria. The machine learning algorithms include Decision Trees, Random Forests, Multiple Linear Regression, Polynomial Regression, Support Vector Regression and Long Short-Term Memory (LSTM) neural networks.

Extensive data on rice yields, and weather pattern, were obtained. These datasets underwent preprocessing, cleaning, and separation into training as well as testing sets. In each of the five machine learning models, data for each region was trained and tested. To evaluate their predictive accuracy a complete range of model evaluation metrics like mean squared error, R-squared and mean absolute errors were computed. Multiple Linear Regression turned out to be superior to all other algorithms across both geographic regions when it came to yield prediction precision. It has an advantage over the other algorithms since it can capture long-range dependencies in time-series data. The results also showed that Random forests displayed a relatively high predictive capacity while the remaining Models exhibited poor performances due to their inability to handle non-linear relationships between features and the target variable.

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

This study contributes significantly to the field of agricultural machine learning by providing a comprehensive comparison of multiple algorithms for rice yield prediction across diverse geo-climatic regions in Nigeria. Our findings, particularly the superior performance of Multiple Linear Regression, offer valuable insights for researchers developing crop yield prediction models in similar contexts as there is still limited work in this domain within nigeria. This work also underscores the importance of considering regional variations in climate and agricultural practices when developing predictive models.

In terms of practical implications for rice farming in Nigeria, this work presents the following: 1. Improved yield forecasting: Farmers and agricultural planners can use our model to make more accurate predictions of rice yields, enabling better resource allocation and planning. 2. Climate adaptation: The insights into the relationship between climatic variables and rice yields can help farmers adapt their practices to changing climate conditions in their respective regions. 3. Policy support: Government agencies can use these models to inform agricultural policies and support programs tailored to different regions.

In the future, we propose the development of a Rice Yield Prediction Support System (RYPSS) based on our best-performing Multiple Linear Regression model as its backend. This system could include: 1. A user-friendly mobile application that allows farmers to input local weather data and receive yield predictions. 2. Integration with weather APIs to automatically fetch relevant climatic data for the farmer's location. 3. Customizable features that allow farmers to adjust inputs based on their specific farming practices. 4. Regular updates to the underlying model as more data becomes available, ensuring continued accuracy. 5. An interface that allows farmers to enter current yield data that will be stored in a database that is accessible by government parastatals like the Ministry of Agriculture which will in turn facilitate informed decision-making at the government level.

Furthermore, we propose the following directions for future work: 1. Expand the study to include more regions in Nigeria, capturing a wider range of geo-climatic conditions. 2. Incorporate additional variables such as soil quality, fertilizer use, and pest incidence to enhance model accuracy. 3. Explore the integration of remote sensing data to improve yield predictions over larger areas. 4. Investigate other advanced machine learning techniques, such as ensemble methods or deep learning models, to further improve the predictive accuracy. 5. Establish robust and comprehensive agricultural databases to enable more advanced analyses and facilitating the development of 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, J.A.I. A.S.N. and A.I.; methodology, A.I. and J.A.I.; fieldwork, J.A.I. and A.I.; software, J.A.I. and A.I.; title, J.A.I. A.S.N. and A.I.; validation, A.I. and A.S.N.; laboratory work, J.A.I. and A.I.; formal analysis, A.I. and A.S.N.; research, J.A.I. A.S.N. and A.I.; sources, J.A.I. A.S.N. and A.I.; data curation, J.A.I. and A.I.; manuscript-original draft, J.A.I. and A.I.; manuscript-review and editing, J.A.I. A.S.N. and A.I.; visualization, A.I. and J.A.I.; supervision, A.S.N. and A.I.; project management, J.A.I. A.S.N. and A.I.; funding, J.A.I. All authors have read and legally accepted the final version of the article published in the journal.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

REFERENCES

Adebayo, A. A. (1999). *Adamawa State in Maps*. Paraclete Publishers.

ADSPC (Adamawa State Planning Commission) (2022). Adamawa State At A Glance. https://adspc.ad.gov.ng/adamawa-state/

Arras, L., Arjona-Medina, J., Widrich, M., Montavon, G., Gillhofer, M., Müller, K.-R., Hochreiter, S., & Samek, W. (2019). Explaining and Interpreting LSTMs. In: W. Samek, G. Montavon, A. Vedaldi, L. K.

Hansen, & K.-R. Müller (Eds.), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (pp. 211-238). Springer-VerlagBerlin, Heidelberg. https://doi.org/10.1007/978-3-030-28954-6_11

Chauhan, N. S. (2022, February 9). Decision Tree Algorithm, Explained. KDnuggets. https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273-297. https://doi.org/10.1007/BF00994018

Das, B., Nair, B., Reddy, V. K., & Venkatesh, P. (2018). Evaluation of multiple linear, neural network and penalised regression models for prediction of rice yield based on weather parameters for west coast of India. International *Journal of Biometeorology*, *62*(10), 1809-1822. https://doi.org/10.1007/s00484-018-1583-6

Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. J., & Vapnik, V. (1996, December 3-5). *Support Vector Regression Machines*. In: M. C. Mozer, M. Jordan, & T. Petsche (Eds.) Proceedings of the Advances in Neural Information Processing Systems 9 (NIPS 1996) (pp. 155-161), Denver Colorado.

Elbeheiry, N., & Balog, R. S. (2022). Technologies Driving the Shift to Smart Farming: A Review. *IEEE Sensors Journal*, *23*(3), 1752-1769. https://doi.org/10.1109/JSEN.2022.3225183

FAO, IFAD, UNICEF, WFP, & WHO. (2023). The State of Food Security and Nutrition in the World 2023. Urbanization, agrifood systems transformation and healthy diets across the rural–urban continuum. Rome, FAO. https://doi.org/10.4060/cc3017en

Geetha, M. C. S. (2015). A Survey on Data Mining Techniques in Agriculture. *International Journal of Innovative Research in Computer and Communication Engineering*, *3*(2), 887-892.

Gnanamanickam, S. S. (2009). Rice and Its Importance to Human Life. In: S. S. Gnanamanickam (Eds.), *Biological Control of Rice Diseases* (pp. 1-11). Springer Netherlands. https://doi.org/10.1007/978-90-481-2465-7_1

Gyimah-Brempong, K., Johnson, M., & Takeshima, H. (2016). Chapter 1. Rice in the Nigerian Economy and Agricultural Policies. In: K. Gyimah-Brempong, M. Johnson, & H. Takeshima (Eds.), *The Nigerian Rice Economy* (pp. 1-20). University of Pennsylvania Press. https://doi.org/10.9783/9780812293753-005

Imani, M. (2019, August 26-27). *Long Short-Term Memory Network and Support Vector Regression for Electrical Load Forecasting*. In: Proceedings of the 2019 International Conference on Power Generation Systems and Renewable Energy Technologies (PGSRET), Istanbul, Türkiye. https://doi.org/10.1109/PGSRET.2019.8882730

Iorliam, I. B., Ikyo, B. A., Iorliam, A., Okube, E. O., Kwaghtyo, K. D., & Shehu, Y. I. (2021). Application of Machine Learning Techniques for Okra Shelf Life Prediction. *Journal of Data Analysis and Information Processing*, *9*(3), 136-150. https://doi.org/10.4236/jdaip.2021.93009

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. https://doi.org/10.51519/journalisi.v5i2.506

Kamai, N., Omoigui, L. O., Kamara, A. Y., & Ekeleme, F. (2020). *Guide to rice production in Northern Nigeria*. International Institute of Tropical Agriculture (IITA).

Karasev, A. (2023). Excursion to the History of Tractor Building and the Introduction of Tractors in Agriculture. *Tekhnicheskiy Servis Mashin*, *61*(1), 155-163. https://doi.org/10.22314/2618-8287-2023-61-1-155-163

Khaki, S., & Wang, L. (2020). *Crop Yield Prediction Using Deep Neural Networks*. In: H. Yang, R. Qiu, & W. Chen (Eds.), Proceedings of the 2019 INFORMS International Conference on Service Science (pp. 139-147). https://doi.org/10.1007/978-3-030-30967-1_13

Mariappan, A. K., & Ben Das, J. A. (2017, April 07-08). *A paradigm for rice yield prediction in Tamilnadu*. In: Proceedings of the 2017 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), (pp. 18-21), Chennai, India. https://doi.org/10.1109/TIAR.2017.8273679

Maulud, D. H., & Abdulazeez, A. M. (2020). A Review on Linear Regression Comprehensive in Machine Learning. *Journal of Applied Science and Technology Trends*, *1*(2), 140-147. https://doi.org/10.38094/jastt1457

Mwiti, D. (2022, July 21). Random Forest Regression: When Does It Fail and Why? Neptune.Ai. https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why

NBS (National Bureau of Statistics) (2020). Demographic Statistics Bulletin. https://nigerianstat.gov.ng/download/1241121

Oguntunde, P. G., Lischeid, G., & Dietrich, O. (2018). Relationship between rice yield and climate variables in southwest Nigeria using multiple linear regression and support vector machine analysis. *International Journal of Biometeorology*, *62*(3), 459-469. https://doi.org/10.1007/s00484-017-1454-6

Okonkwo, U. U., Ukaogo, V., Kenechukwu, D., Nwanshindu, V., & Okeagu, G. (2021). The politics of rice production in Nigeria: The Abakaliki example, 1942-2020. *Cogent Arts & Humanities*, *8*(1), 1880680. https://doi.org/10.1080/23311983.2021.1880680

Onwude, D. I., Chen, G., Hashim, N., Esdaile, J. R., Gomes, C., Khaled, A. Y., Alonge, A. F., & Ikrang, E. (2018). Mechanization of Agricultural Production in Developing Countries. In: G. Chen (Eds.), *Advances in Agricultural Machinery and Technologies* (pp. 3-26). CRC Press. https://doi.org/10.1201/9781351132398-1

Ostertagová, E. (2012). Modelling using Polynomial Regression. *Procedia Engineering*, *48*, 500-506. https://doi.org/10.1016/j.proeng.2012.09.545

Özdoğan-Sarıkoç, G., Sarıkoç, M., Celik, M., & Dadaser-Celik, F. (2023). Reservoir volume forecasting using artificial intelligence-based models: Artificial Neural Networks, Support Vector Regression, and Long Short-Term Memory. *Journal of Hydrology*, *616*, 128766. https://doi.org/10.1016/j.jhydrol.2022.128766

Patrio, U., Yuliska, Y., & Widyasari, Y. D. L. (2024). Predicting Rice Production In Sumatra Island Using Linear Regression. In: B. Santoso, B. Bustami & A. Satria (Eds.), Proceedings of the 11th International Applied Business and Engineering Conference (ABEC 2023), (2023, September 21). Bengkalis, Riau, Indonesia. http://doi.org/10.4108/eai.21-9-2023.2342997

Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., & Athanasiadis, I. N. (2021). Machine learning for large-scale crop yield forecasting. *Agricultural Systems*, *187*, 103016. https://doi.org/10.1016/j.agsy.2020.103016

Pečkov, A. (2012). *A Machine Learning Approach to Polynomial Regression*. PhD Thesis. Jožef Stefan International Postgraduate School.

Pierce, F. J., & Nowak, P. (1999). Aspects of Precision Agriculture. *Advances in Agronomy*, *67*, 1-85. https://doi.org/10.1016/S0065-2113(08)60513-1

Ritchie, J. T., Singh, U., Godwin, D. C., & Bowen, W. T. (1998). Cereal growth, development and yield. In: G. Y. Tsuji, G. Hoogenboom, & P. K. Thornton (Eds.), *Understanding Options for Agricultural Production* (pp. 79-98). Springer Netherlands. https://doi.org/10.1007/978-94-017-3624-4_5

Rosa, W. (Eds.) (2017). Transforming Our World: The 2030 Agenda for Sustainable Development. In: *A New Era in Global Health* (pp. 529-567). Springer Publishing Company. https://doi.org/10.1891/9780826190123.ap02

Sasu, D. D. (2023, November 9). Nigeria: Production of milled rice 2010-2023. Statista. https://www.statista.com/statistics/1134510/production-of-milled-rice-in-nigeria/

Seber, G. A. F., & Lee, A. J. (2003). *Linear Regression Analysis* (2nd Ed.). John Wiley & Sons. https://doi.org/10.1002/9780471722199

Sharma, A., Jain, A., Gupta, P., & Chowdary, V. (2021). Machine Learning Applications for Precision Agriculture: A Comprehensive Review. *IEEE Access*, *9*, 4843-4873. https://doi.org/10.1109/ACCESS.2020.3048415

Van Asten, P. J. A., Kaaria, S., Fermont, A. M., & Delve, R. J. (2009). Challenges and lessons when using farmer knowledge in agricultural research and development projects in Africa. *Experimental Agriculture*, *45*(1), 1-14. https://doi.org/10.1017/S0014479708006984

van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, *177*, 105709. https://doi.org/10.1016/j.compag.2020.105709

Vanitha, C. N., Archana, N., & Sowmiya, R. (2019, March 15-16). *Agriculture Analysis Using Data Mining And Machine Learning Techniques*. In: Proceedings of the 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS), (pp. 984-990), Coimbatore, India. https://doi.org/10.1109/ICACCS.2019.8728382

Veenadhari, S., Misra, B., & Singh, C. (2014, January 03-05). *Machine learning approach for forecasting crop yield based on climatic parameters*. In: Proceedings of the 2014 International Conference on Computer Communication and Informatics, (pp. 1-5), Coimbatore, India. https://doi.org/10.1109/ICCCI.2014.6921718

Wart, J. V., Kersebaum, K. C., Peng, S., Milner, M., & Cassman, K. G. (2013). Estimating crop yield potential at regional to national scales. *Field Crops Research*, *143*, 34-43. https://doi.org/10.1016/j.fcr.2012.11.018

Zeigler, R. S., & Barclay, A. (2008). The Relevance of Rice. *Rice*, *1*(1), 3-10. https://doi.org/10.1007/s12284-008-9001-z