Comparative Performance Evaluation of Random Forest, Extreme Gradient Boosting and Linear Regression Algorithms Using Nigeria's Gross Domestic Products

M. D. Adewale[1(✉)], D. U. Ebem[2], O. Awodele[3], A. Azeta[4], E. M. Aggrey[1], E. A. Okechalu[1], K. A. Olayanju[1], A. F. Owolabi[1], J. Oju[1], O. C. Ubadike[1], G. A. Otu[1], U. I. Muhammed[1], and O. P. Oluyide[1]

[1] African Centre of Excellence on Technology Enhanced Learning, National Open University of Nigeria, Abuja, Nigeria
ace22140007@noun.edu.ng
[2] Department of Computer Science, University of Nigeria, Nsukka, Nigeria
[3] Department of Computer Science, Babcock University, Ilishan-Remo, Ogun, Nigeria
[4] Department of Software Engineering, Namibia University of Science and Technology, Windhoek, Namibia

Abstract. Statistical methods like linear regression analysis are frequently used to create predictive analytic models. However, these methods have limitations that may affect the accuracy of the models. Using a typical dataset, this study seeks to accomplish two main goals. First, we fitted three predictive models, including linear regression analysis and two ensemble machine learning algorithms: Random Forest Regressor and Extreme Gradient Boosting Regressor. Secondly, we compared the performance of the models using a 5-fold cross-validation technique. The Random Forest Regressor outperformed the other models, with a Mean Absolute Error (MAE) of 10.138, Mean Square Error (MSE) of 139.729, Mean Absolute Percentage Error (MAPE) of 0.071, Root Mean Square Error (RMSE) of 11.821, and Normalised Mean Square Error (NMSE) of 13.782. These results suggest that the Random Forest Regressor is optimal for developing predictive models with similar datasets.

Keywords: Machine learning · Random Forest Regressor · XGboost Regressor · Linear Regression · Gross Domestic Product · 5-fold cross-validation

1 Introduction

In today's world, predictive modelling has taken the forefront in many industries, such as finance and healthcare. It is now a vital tool for companies to gain insights and make informed decisions. The advent of machine learning algorithms has played a significant role in this development, enabling organisations to handle vast volumes of data efficiently and precisely. According to Wang & Lee (2021), machine learning algorithms

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2024
Published by Springer Nature Switzerland AG 2024. All Rights Reserved
A. Seeam et al. (Eds.): InterSol 2023, LNICST 541, pp. 131–150, 2024.
https://doi.org/10.1007/978-3-031-51849-2_9

have been widely adopted in various applications and have gained significant popularity. Among the machine learning algorithms, the Random Forest Regressor, Extreme Gradient Boosting (XGBoost) Regressor, and Linear Regression are some of the most commonly used algorithms. These algorithms have proven to be efficient and accurate in predicting outcomes. However, they have significant differences in predictive performance, computational efficiency, and interpretability. It is imperative to comprehend these differences to select the appropriate model for a given prediction problem. For instance, the Random Forest Regressor is a versatile algorithm that can handle a large number of variables and nonlinear relationships. On the other hand, the XGBoost Regressor is a gradient-boosted model that can provide superior predictive performance and scalability. Lastly, Linear Regression is a simple algorithm that is easy to interpret and is suitable for predicting outcomes when the relationships between variables are linear.

Machine learning algorithms have revolutionised the field of predictive modelling by enabling accurate predictions of outcomes in regression tasks such as predicting stock prices, house prices, or medical diagnoses (Kaliappan et al., 2021; Xu et al., 2022). However, selecting the best algorithm for a given problem can be challenging due to several factors, such as the data distribution, dataset size, and feature number (Raju et al., 1998). Thus, it is crucial to compare the relative performance of algorithms in terms of accuracy. The Random Forest Regressor is an ensemble learning technique that creates multiple decision trees and combines their predictions to generate a more precise forecast (Pandey et al., 2019; Garg & Poornalatha, 2019). It is a flexible algorithm that can handle numerous variables and nonlinear relationships, making various applications possible.

On the other hand, the XGBoost Regressor is a gradient-boosting algorithm (Li, 2021) that also uses decision trees to make predictions (Allawala et al., 2022). It has gained popularity due to its superior predictive performance and scalability. Lastly, Linear Regression is a simple but powerful algorithm widely used due to its ease of implementation and interpretation (Chien et al., 2023). It is suitable for predicting outcomes when the relationships between variables are linear.

This paper aims to present a comparative performance evaluation of the Random Forest Regressor, XGBoost Regressor, and Linear Regression algorithms based on metrics such as Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) using a typical dataset. The objective is to conduct experiments on a given dataset and evaluate the performance of each algorithm using different metrics to provide insights into their relative performances and identify the most suitable algorithm for specific applications. By comparing the performance of these three algorithms, we hope to provide a comprehensive understanding of their strengths and limitations, aiding decision-makers in selecting the most appropriate algorithm for a given prediction problem.

The findings of our study are expected to be highly valuable to researchers and practitioners in machine learning and those who use regression analysis in their work. By providing a comprehensive comparative evaluation of the Random Forest Regressor, XGBoost Regressor, and Linear Regression algorithms, we aim to contribute to the growing body of knowledge on the performance of regression algorithms. The insights gained from our study will enable users to make enlightened choices when picking the

most appropriate algorithm for their specific application, leading to more accurate and reliable predictions.

The paper is structured as follows to clearly and concisely present the evaluation and findings. The three algorithms and their properties are briefly described in Sect. 2, enabling readers to understand the underlying concepts and differences between them. Section 3 delves into the methodology used to evaluate the performance of the algorithms, including the dataset used, the evaluation metrics employed, and the experimental design. In Sect. 4, we present the results of the comparative evaluation, including the performance of each algorithm on various metrics, enabling readers to compare and contrast their relative strengths and weaknesses. Finally, in Sect. 5, we summarise our findings and conclusions, highlighting the most significant insights gained from our study and their implications for those working in the field of machine learning as researchers and practitioners.

2 Literature Review

According to Schonlau and Zou (2020), the question of whether linear regression should be considered a machine learning algorithm or just a statistical method for prediction has been a matter of debate. While some argue that it is a statistical method, others suggest it can also be classified as a machine learning algorithm. Regardless of the conflicting views, it is crucial to evaluate the effectiveness of linear regression and contrast it with that of other machine learning algorithms. This comparison can shed light on the strengths and limitations of each method, allowing us to choose the most appropriate approach for a given problem. Although linear regression is frequently considered a straightforward and simple technique, it has proven helpful in many applications, including forecasting stock prices, examining consumer behaviour, and predicting sales trends (Tan & Al-Barakati, 2022; Wang et al., 2022). The popularity of machine learning algorithms, on the other hand, is growing due to their capacity to learn from data and make precise predictions about challenging issues (Sandra et al., 2021). In order to select the approach that best meets the needs of the current problem, it is crucial to assess the effectiveness of both. Ultimately, the choice should be based on a thorough understanding of the strengths and weaknesses of each technique as well as the application's unique requirements.

Machine learning algorithms have gained popularity in recent years for their ability to predict future trends based on historical data. Regression analysis stands as a frequently employed technique in predictive analytics, and various algorithms have been developed to achieve accurate predictions. This section reviews the literature on three algorithms - Random Forest Regressor, XGBoost Regressor, and Linear Regression - and their performance evaluation in various applications.

Various applications, including economics, social sciences, and environmental studies, use linear regression, a straightforward but effective technique. It establishes a linear relationship between the dependent and independent variables and is widely used for predictive analytics (Mislick & Nussbaum, 2015). For instance, Li et al. (2013) used Linear Regression to predict the energy consumption of buildings, with the ventilation energy consumption predicted at high accuracies of over 99%. Similarly, Mirugwe (2021) used Linear Regression to predict the average dollar tip a waiter can expect from the restaurant given several predictor variables, achieving a minimum Root Mean Square Error

(RMSE) of 1.1815. While statistical methods like Linear Regression have been widely used in economics, they have been criticised for their inability to capture complex non-linear relationships between economic variables. Karen and Louise (2018) and Gareth et al. (2017) have highlighted these limitations and pointed out that relying solely on linear relationships can result in underestimating gross domestic product (GDP).

Moreover, several factors can affect the model's accuracy, such as multicollinearity, heteroscedasticity, and outliers. Therefore, researchers must adopt more advanced techniques to capture complex and dynamic relationships between economic variables. First, including irrelevant variables can reduce the effectiveness of the model. Second, a significantly higher or lower number of observations than predictors may cause the Linear Regression approach to produce inaccurate predictions. As a result, this issue can lead to overestimation or underestimation of the model, limiting its accuracy. Third, relying solely on R^2 statistics and correlation to measure the model's fitness may not be suitable for predicting future data. Finally, the lack of a support system for tuning parameters and cost function prevents applying a bias-variance trade-off to minimise the test Mean Square Error. Thus, alternative approaches, such as machine learning algorithms, may be better suited for accurately predicting economic variables (Agu et al., 2022).

An approach to ensemble learning is the Random Forest Regressor which constructs multiple decision trees and combines their predictions to improve accuracy. It has been used in various applications, including finance, healthcare, and marketing. For instance, Lawrence et al. (2021) used Random Forest Regressor to predict patient survival in healthcare, achieving an accuracy rate of 86.4%. Random Forest has been identified as the best technique for prediction for small data sets (Ameer et al., 2019).

XGBoost Regressor is another ensemble learning method that uses a gradient-boosting algorithm to improve accuracy (Sainikhileaswar & Parthasarathy, 2020). It has been used in various applications, including remote sensing (Öztürk & Colkesen, 2021), healthcare or medical informatics (Huang et al., 2022). A study by Raheja et al. (2021) used an XGBoost Regressor to evaluate groundwater indices over a Haryana state (India) study area. Similarly, Nguyen et al. (2021) used XGBoost Regressor to predict high-performance concrete's compressive and tensile strength. Extreme gradient boosting (XGBoost) is a powerful and efficient algorithm that can accurately handle large datasets (Xu et al., 2022).

In various applications, Random Forest Regressor, XGBoost Regressor, and Linear Regression are widely used algorithms that have proven effective. However, their performance varies depending on the application and dataset. Therefore, conducting a comparative performance evaluation of these algorithms is crucial to identify the most accurate and efficient algorithm for specific applications.

This study aims to overcome the shortcomings inherent in traditional statistical methods in predicting related projects by utilising two ensemble machine learning models, namely the Random Forest Regressor and XGBoost Regressor, along with a Linear Regression statistical model. Our objective is to determine the most effective approach for prediction and underscore the importance of integrating machine learning techniques in data science. While we recognise the limitations of machine learning, our approach aims to highlight its potential in prediction by demonstrating its superiority over traditional statistical methods. As interest in machine learning for prediction grows, our

study contributes to the literature emphasising the necessity for data scientists to harness this technology while comprehending its limitations.

This study aims to develop precise predictive models for Nigeria's GDP by employing various relevant economic and non-economic indicators. The data set is carefully curated and refined to encompass healthcare spending, net migration, population, life expectancy, electricity access, and internet usage. In order to construct and compare predictive models, machine learning methods like Random Forest Regressor, XGboost Regressor, and statistical Linear Regression analysis are employed. The primary objective is to pinpoint the most efficient approach for forecasting Nigerian GDP, considering many factors that affect economic growth. Our research has significant implications for practitioners and scholars in selecting the most suitable algorithm for similar data sets.

3 Methods and Techniques

The relevant dataset consists of 22 instances from 2000 to 2021, has five attributes, and is focused on economic and non-economic parameters such as Nigeria's gross domestic product (gdp) in billions of dollars, healthcare spending (hs) in billions of dollars, population (p), life expectancy (le) in years, and index of economic freedom (ief). The research utilised a secondary dataset from the World Bank (World Bank, 2021) and Nigeria Corruption Perceptions Index, 2001–2022 - knoema.com, (2023). The experiment was carried out using the relevant Sklearn Python libraries at https://colab.research.google.com/. Table 1 shows the first five rows of the sample dataset.

Table 1. First five rows of the sample dataset

Year	GDP	HS	P	LE	IEF
2021	440.78	11.13	213401323	55.12	58.70
2020	432.29	14.7	208327405	54.81	57.20
2019	448.12	13.58	203304492	54.49	57.30
2018	397.19	12.27	198387623	54.18	58.50
2017	375.75	14.09	193495907	53.73	57.10

Hotz (2023) highlighted the prevalence of the CRoss Industry Standard Process for Data Mining (CRISP-DM) adopted methodology in data science projects, as evidenced by Fig. 1. The CRISP-DM approach comprises six key phases, as shown in Fig. 2. The study followed the six stages of the CRISP-DM methodology, which are crucial in ensuring the success of any data mining project.

In the initial business understanding stage, our objective was aimed to contrast the efficacy of the Random Forest Regressor and XGBoost Regressor against Linear Regression Analysis. In the subsequent data understanding stage, we focused on deeply comprehending our data, including pinpointing any quality issues and recognising critical attributes valuable for the modelling process.

Our datasets were sourced from the World Bank (World Bank, 2021) and the Nigeria Corruption Perceptions Index, 2001–2022 - knoema.com, (2023) for the data preparation phase. These datasets encompassed variables such as healthcare spending (hs), population (p), life expectancy (le), and the index of economic freedom (ief), paired with GDP data. After consolidating this information, we identified no missing values, facilitating the designation of independent and dependent variables. A division into training and testing subsets followed this.

During the modelling stage, our dataset was trained, and the data was tailored to fit the three selected algorithms: Random Forest Regressor, XGBoost Regressor, and Linear Regression Analysis. The evaluation stage saw the application of a 5-fold Cross Validation method, aiding in determining key metrics like MAE, MAPE, MSE, RMSE, and NMSE across our modelling strategies. The final deployment phase, intrinsic to the CRISP-DM methodology, varied in its approach based on our set objectives. This ranged from a succinct report of findings to a comprehensive implementation of the developed models. It's paramount to note that our research's initial and final phases were oriented towards GDP prediction. Our exploration and evaluation of varied predictive models culminated in detailed findings in the final phase. Section 4 delves deeper into the nuances of the intermediary phases. Typically, a subset of the data, termed the training set, is utilised during modelling, while the residual dataset, the test set, is reserved for performance evaluation during the evaluation phase. A detailed breakdown of the algorithms employed in our analysis is provided in the following sections.

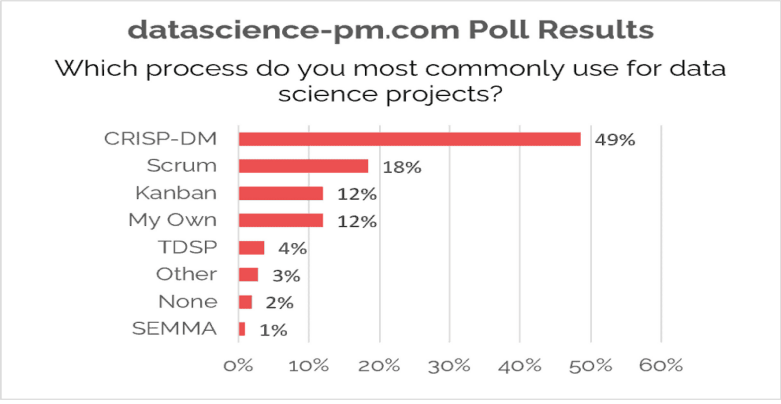

Fig. 1. Most commonly used process for data science projects. Source: Hotz (2023)

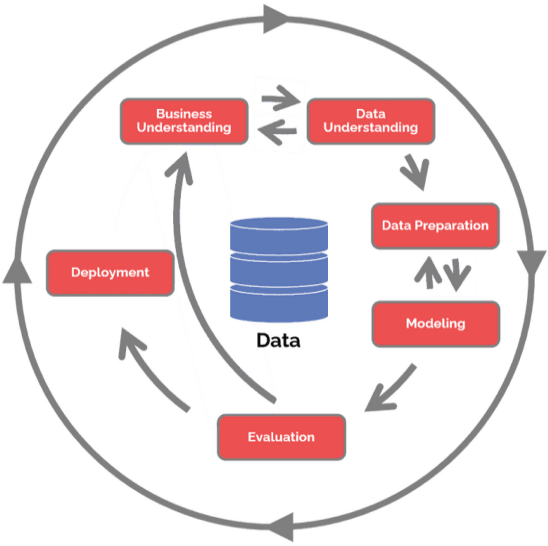

Fig. 2. The Six stages of the CRISP-DM methodology. Source: Hotz (2023)

3.1 Theory/Calculation

This study made use of five features, viz., gross domestic product (gdp), healthcare spending (hs), population (p), life expectancy (le), and Index of economic freedom (ief), that made up the study's conceptual model, with GDP standing in as the dependent variable as shown in Eq. 1 (Gareth et al., 2017). Here, we consider the GDP, which represents economic growth for this period, as a function of healthcare spending, population, life expectancy, and index of economic freedom. In this study, we analyse various machine learning models that take the following variables: healthcare spending, population, life expectancy, and index of economic freedom as the independent variables and gross domestic product (gdp) as the target variable. This is done to create accurate parameter estimates for the models. The research model employed can be described as follows:

$$gdp = f(hs, p, le, ief) + e \tag{1}$$

where the random variable e stands for the error term, independent of the predictors and has a mean of zero, the fixed but unknown function f represents the information the predictors provide about the GDP.

3.2 The Linear Regression Model

The Linear Regression method assumes that the function f in Eq. 1 is linear in (*hs, p, le, ief*) as shown in Eq. 2.

$$gdp = \beta_0 + \beta_1 hs + \beta_2 p + \beta_3 le + \beta_4 ief \tag{2}$$

The Linear Regression method assumes that only five coefficients $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$ need to be estimated, instead of having to estimate a completely arbitrary 5-dimensional

function, making it easier to estimate f (hs, p, le, ief) (Gareth et al., 2017). These coefficients, represented by β_0 to β_4, define the relationships between the selected economic and non-economic variables and the GDP. β_0 represents the constant (intercept) of the equation, while β_1 to β_4 represent the coefficients of the macroeconomic variables. In particular, β_0 is the expected value of the GDP when all variables are zero. At the same time, β_1 to β_4 represent the average effect of a one-unit increase in each of the economic and non-economic independent variables on the GDP, holding all other predictors fixed.

Estimation of the Coefficients

To ascertain the link between the predictor variables and the GDP, the coefficients of Eq. 2 must be estimated, as seen in Eq. 3.

$$\widehat{gdp} = \hat{\beta}_0 + \hat{\beta}_1 hs + \hat{\beta}_2 p + \hat{\beta}_3 le + \hat{\beta}_4 ief \tag{3}$$

In this equation, $\widehat{gdp}$ represents the predicted GDP while $\hat{\beta}_0$, $\hat{\beta}_1$, and so on, up to $\hat{\beta}_4$, represent the estimated coefficients.

Least squares compute the estimated coefficients $\hat{\beta}_0$, $\hat{\beta}_1 \ldots$, $\hat{\beta}_4$ in Eq. 3 by utilising some calculus to diminish the residual Sum of Squares (RSS) given in Eq. 4 (Wikipedia, 2021).

$$\begin{aligned} RSS &= \sum_{i=1}^{n} \left(gdp_i - \widehat{gdp_i} \right)^2 \\ &= \sum_{i=1}^{n} (gdp_i - \hat{\beta}_0 - \hat{\beta}_1 hs - \hat{\beta}_2 p - \hat{\beta}_3 le - \hat{\beta}_4 ief)^2 \end{aligned} \tag{4}$$

where ***i*** denotes a single yearly observation and ***n*** denotes the total number of years.

3.3 The Random Forest Regressor Model

Random Forest Regression is a robust supervised learning algorithm that leverages the ensemble learning approach to perform regression tasks. Combining predictions from multiple machine learning algorithms can generate highly accurate predictions that outperform those from any individual model. The Random Forest model is structured as a collection of decision trees that operate in parallel, as illustrated in Fig. 3. Each tree is constructed independently during the training phase, and the predicted value from most trees is used as the final output (Chaya, 2022). This unique approach ensures that the model is highly robust, even when dealing with complex and noisy datasets, making it an excellent choice for various machine learning and data science applications.

The Random Forest Regressor generates predictions by averaging the forecasts made by the forest's trees. Averaging is a key factor in the superior performance of the random forest over a single decision tree. This increases its accuracy and prevents it from becoming overly effective at its job. The average of the forecasts made by the forest's trees is what the Random Forest Regressor produces (Mwiti, 2022).

Saabas (2014) explained that a decision tree consists of a series of paths from the tree's root to the leaf. Each path represents a series of decisions based on specific features contributing to the final prediction. To define the prediction function of a decision tree,

the feature space is partitioned into ***M*** regions represented by the ***M*** leaves of the tree, denoted as ***Rm*** where ***1 ≤ m ≤ M***. This partitioning is done through a series of decisions made at each internal node of the tree, based on specific features. According to Saabas (2014), the prediction function of a decision tree is defined based on the standard criteria presented in Eqs. 5, 6 and 7:

$$f(x) = \sum_{m=1}^{M} C_m I(x, R_m) \tag{5}$$

In a decision tree, the feature space is divided into ***M*** regions, where ***M*** is the number of leaves in the tree. Each region, denoted by ***Rm, 1 ≤ m ≤ M***, is guarded by a specific feature. The prediction function of the tree is defined as follows: the value of ***Cm*** is established during the tree's training phase, which corresponds to the mean of the response variables for samples that fall under region ***Rm*** in the case of regression trees (or the ratio(s) for classification trees). The indicator function ***I*** returns ***1 if x*** $\in$ ***Rm and 0*** otherwise. Although this definition provides a clear and concise understanding of a tree, it ignores the operational aspect of decision trees, including the informative decision nodes and the path through them. Predictions made by individual trees in a forest are averaged to form the forest's overall prediction.

$$F(x) = \frac{1}{J} \sum_{j=1}^{J} f_j(x), \tag{6}$$

The variable ***J*** represents the total number of trees included in the forest. It is clear from this that the predictions made by individual trees in a forest are averaged to form the forest's overall prediction:

$$\frac{1}{J} \sum_{j=1}^{J} c_{j\,full} + \sum_{k=1}^{K} \left(\frac{1}{J} \sum_{j=1}^{J} contrib_j(x, k) \right). \tag{7}$$

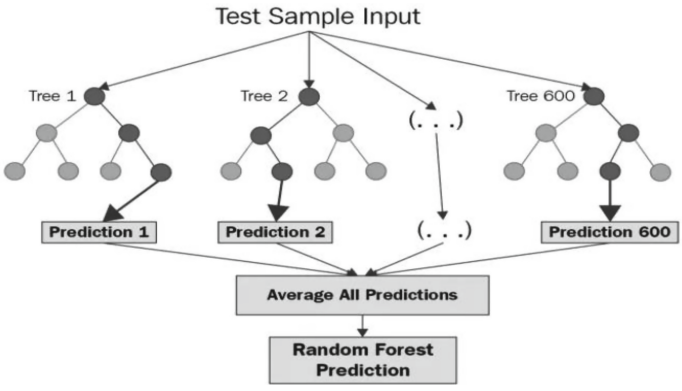

Fig. 3. Random Forest Tree Source: (Chaya, 2022)

3.4 The XGBoost Regressor Model

According to GeeksforGeeks (2023), XGBoost is an optimised distributed gradient boosting library designed for rapid and scalable machine learning model training. This ensemble learning technique combines the predictions from several weak models to produce a stronger prediction. Due to its capacity to manage massive datasets and produce ground-breaking results in numerous machine learning tasks like classification and Regression, XGBoost is one of the most well-known and popular machine learning algorithms. XGBoost is a Gradient Boosted decision tree implementation. XGBoost is a powerful machine-learning algorithm that generates decision trees in sequence while considering the importance of each independent variable. Each variable is assigned a weight, which is used to predict the output of the decision tree. If a variable is mispredicted, its weight is increased and used as input for the next decision tree. By combining multiple classifiers/predictors, XGBoost creates a more accurate and robust model that can handle various problems, including regression, classification, ranking, and user-defined prediction. For instance, let's consider a CART that predicts whether an individual would enjoy a hypothetical computer game X. The final prediction score is calculated by adding the prediction scores from each decision tree. The mathematical representation of the model is detailed in Eqs. 8–24, as cited by GeeksforGeeks (2023):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), f_k \in \mathcal{F} \tag{8}$$

The model's objective function can be expressed mathematically as follows, where $\boldsymbol{K}$ represents the total number of trees in the model, $\boldsymbol{f}$ represents the functional space of the set $\boldsymbol{F}$, and $\boldsymbol{F}$ represents the set of all possible decision trees:

$$obj(\theta) = \sum_{i}^{n} l\left(y_i, \hat{y}_i\right) + \sum_{k=1}^{K} \Omega\left(f_k\right) \tag{9}$$

Instead of attempting to optimise the learning of the tree all at once, which is a complicated process, an additive strategy is used, where the loss of what has been learned is minimised, and a new tree is added, as shown below. The first term in the equation represents the loss function, while the second represents the regularisation parameter.

$$\begin{aligned} \hat{y}_i^{(0)} &= 0 \\ \hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\ \hat{y}_i^{(2)} &= f_1(x_i) = f_2(x_i) + f_2(x_i) \\ &\cdots \\ \hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i) \end{aligned} \tag{10}$$

The model described above has the following objective function:

$$\begin{aligned} obj^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_t) \\ &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)}) + f_t(x_i)) + \Omega(f_t) + constant \\ obj^{(t)} &= \sum_{i=1}^{n} (y_i - (\hat{y}_i^{(t-1)} + f_t(x_i)))^2 + \sum_{i=1}^{t} \Omega(f_i) \end{aligned}$$

$$= \sum_{i=1}^{n} [2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2] + \Omega(f_t) + constant \tag{11}$$

Now, let's expand the Taylor series up to the second order:

$$obj^{(t)} = \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) + constant \tag{12}$$

With $\boldsymbol{g_i}$ and $\boldsymbol{h_i}$ defined as:

$$g_i = \partial_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{13}$$

$$h_i = \partial^2_{\hat{y}_i^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \tag{14}$$

Streamlining and getting rid of the constant:

$$\sum_{i=1}^{n} [g_i f_i(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{15}$$

The model must first be defined before we can determine the regularisation term:

$$f_t(x) = \omega_{q(x)}, \omega \in R^T, \; q : R^d \rightarrow \{1, 2, \ldots, T\} \tag{16}$$

In the XGBoost model, the regularisation term is determined by a combination of factors, including a function that maps each data point to the corresponding leaf *(q),* a vector of scores on tree leaves *(w),* and the total number of leaves *(T).* The regularisation term can be expressed as a function of these components, which helps control the model's complexity and reduce the risk of overfitting. The regularisation term is then mathematically represented by:

$$\Omega(f) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} \omega_j^2 \tag{17}$$

Our objective function is now:

$$\begin{aligned} \text{obj}^{(t)} &\approx \sum_{i=1}^{n} \left[g_i w_{q(x_i)} + \frac{1}{2} h_i w^2_{q(x_i)} \right] + \gamma T + \frac{\lambda}{2} \sum_{j=1}^{T} w_j^2 \\ &= \sum_{j=1}^{T} \left[(\sum_{i \in I_j} g_i) w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2 \right] + \gamma T \end{aligned} \tag{18}$$

The above expression is now simplified:

$$obj^{(t)} = \sum_{j=1}^{T} [G_j \omega_j + \frac{1}{2} (H_j + \lambda) \omega_j^2] + \gamma T \tag{19}$$

where,

$$G_j = \sum_{i \in I_j} g_i \tag{20}$$

$$H_j = \sum_{i \in I_j} h_i \tag{21}$$

Given a particular structure of ***q(x),*** the best $\boldsymbol{w_j}$ in this equation, which is independent of one another, is the greatest achievable objective reduction:

$$\omega_j^* = -\frac{G_j^2}{H_j + \lambda} \tag{22}$$

$$\text{obj}^* = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{23}$$

This algorithm uses a pruning parameter γ, which determines the minimum information gain required to perform a split. To measure the effectiveness of the tree, we optimise one level at a time instead of optimising the entire tree. In this, a leaf is divided into two leaves, and the score it receives is calculated. The score is then used to determine if the split should be accepted. The score it gains is:

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{24}$$

3.5 Error Costs and Estimation

The goal of the study is to compare the results obtained using each model and identify which is the most effective. Aftarczuk (2007) highlights several popular error metrics, as illustrated in Eqs. 25–29, which are frequently incorporated into various machine learning tools:

Mean Absolute Error (MAE): Eq. 25 is the average of individual errors while neglecting the signs to diminish the negative effects of outliers.

$$= \frac{1}{n} \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{25}$$

$$\text{Mean Absolute Percentage Error (MAPE)} = \frac{1}{n} \sum_{t=1}^{n} \frac{|y_t - \hat{y}_t|}{y_t} \tag{26}$$

$$\text{Mean Square Error(MSE)} = \frac{1}{n} \sum_{t=1}^{n} (y_t - \hat{y}_t)^2 \tag{27}$$

$$\text{Root Mean Square Error(RMSE)} = \sqrt{\sum_{i=1}^{n} \frac{(y_t - \hat{y}_t)^2}{n}} \tag{28}$$

$$\text{Normalised Mean Square Error (NMSE)} = \frac{\sum_{t=1}^{n} (y_t - \hat{y}_t)^2}{\sum_{t=1}^{n} (y_t - \overline{y}_t)^2} \tag{29}$$

where y_t and $\widehat{y_t}$ are the actual and predicted values $\overline{y}_t$ is the mean value of y_t. The smaller the error values, the nearer the estimated figures align with the true values.

3.6 5-fold Cross-Validation (5-fold CV) Technique

Since we are considering three prediction models: Random Forest Regressor, XGBoost Regressor, and Linear Regression, we compare the outcomes of their predictions and decide which model is more effective by comparing the results. The 5-fold cross-validation (5-fold CV) method is used for this. The data is split into five randomly chosen folds in this method, which is advantageous as it aids in preventing overfitting and provides a more precise estimation of the test error. Every iteration uses the remaining four folds as the model's training set while treating one-fold as a validation set. This procedure is repeated five times using a different fold as the validation set in each iteration. The MAE, MSE, MAPE, RMSE, and NMSE are computed for each fold. The final 5-fold CV estimate is obtained by averaging the MAEs, MSEs, MAPEs, RMSEs, and NMSEs. Each metric provides different insights into the model's performance and is helpful in different contexts. Thus, it is crucial to consider a diverse set of evaluation metrics to make informed decisions and comparisons when selecting models. Using the 5-fold cross-validation approach, we assess the performance of the models across various data subsets. This provides a more robust gauge of the model's generalisation capabilities. The equations for this validation, pertaining to each error metric, are presented in Eqs. 30–34, as adapted from Pandian (2023).

$$CV_{(5)mae} = \frac{1}{5}\sum_{i=1}^{5} MAE_i \tag{30}$$

$$CV_{(5)mse} = \frac{1}{5}\sum_{i=1}^{5} MSE_i \tag{31}$$

$$CV_{(5)mape} = \frac{1}{5}\sum_{i=1}^{5} MAPE_i \tag{32}$$

$$CV_{(5)rmse} = \frac{1}{5}\sum_{i=1}^{5} RMSE_i \tag{33}$$

$$CV_{(5)nmse} = \frac{1}{5}\sum_{i=1}^{5} NMSE_i \tag{34}$$

In addition to the benefits of evaluating MAE, MSE, MAPE, RMSE, and NMSE using a 5-fold CV, it also provides a way to balance between bias and variance in model selection by allowing us to choose the optimal cost function. Using a 5-fold CV; we can evaluate various models' efficacy and choose the most suitable one with the lowest test error. This helps to avoid overfitting, where a model adheres too tightly to the training data, compromising its ability to generalise well to new data. In this way, a 5-fold CV provides a way to optimise model performance and ensure accurate predictions.

4 Results and Discussion

4.1 Predictive Accuracy

Developing predictive accuracy models to aid data scientists in related research projects is critical. The study's contribution is to build predictive models using machine learning approaches. In this respect, Fig. 4, 5, 6, 7 and Fig. 8 present the 5-fold Cross-Validation of MAE, MSE, MAPE, RMSE, and NMSE of the 22 observations from Nigeria's dataset. The plotted small squares on the line correspond to the MSE values associated with each method on the x-axis. According to the plots, the Random Forest Regressor method resulted in the lowest MAE, MSE, MAPE, RMSE, and NMSE. The last square on the line indicates Var(e), the irreducible error corresponding to the minimum achievable MAE, MSE, MAPE, RMSE, and NMSE among all methods. Therefore, the values obtained from the Random Forest Regressor method are the closest to the optimal value and are recommended for developing a predictive model for Nigeria's GDP.

The study implies that a Random Forest Regressor approach to modelling decision-making in GDP behaviour in this context is more accurate than Linear Regression and XGboost regressor models. These findings are consistent with Giovanni et al. (2021) study, which shows that machine learning techniques perform better in predictive accuracy than conventional ordinary least squares (OLS). Achieving high accuracy and high explainability in forecasting is a critical best practice in developing trust between machine learning models and decision-makers, as Bellotti et al. (2021) highlighted. The concept is that decision-makers should embrace machine learning as a potent instrument and utilise it with mindfulness instead of treating it as an opaque "black box".

Numerous machine learning techniques exist, each with specific applications and inherent limitations, as noted by experts like Katrina (2021), Shaobo (2021), and Brownlee (2019). For this study, we selected algorithms designed for quantitative, continuous numerical data: Linear Regression, XGBoost Regressor, and Random Forest Regressor. In a comparative analysis of GDP prediction studies presented in Table 2, our research, employing the Random Forest Regressor, logged an MSE of 139.729. Agu et al. (2022), with the Principal Component Regression (PCR), reported an MSE of $-7.552007365635066e + 21$. Maccarrone et al. (2021) achieved an MSE of 173e-03 using the K-Nearest Neighbour (KNN), while Flannery (2020) documented an MSE of 2946980.9 using the Artificial Neural Network (specifically, the Multilayer Perceptron). Our model showcased robust performance. However, when interpreting these results, it's essential to account for potential dataset discrepancies and recognise that the selection of independent variables may vary across these research studies.

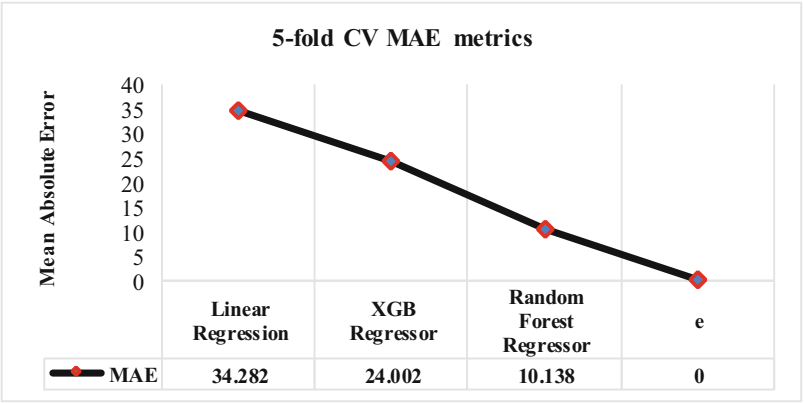

Fig. 4. 5-fold Cross validation MAE plot for Linear Regression, XGBoost Regressor, Random Forest Regressor and *e*

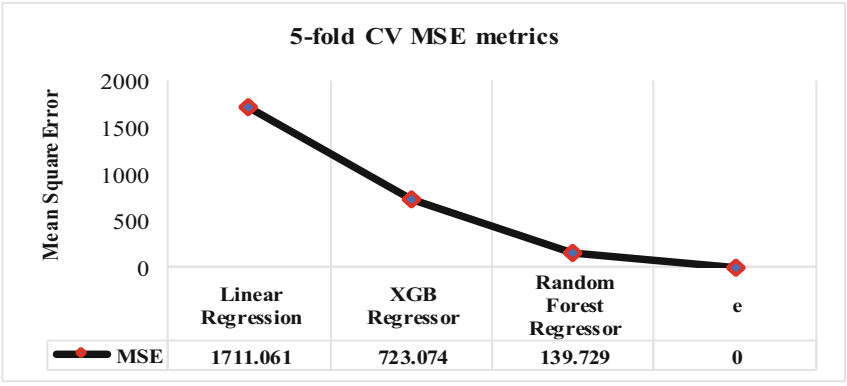

Fig. 5. 5-fold Cross validation MSE plot for Linear Regression, XGBoost Regressor, Random Forest Regressor and *e*

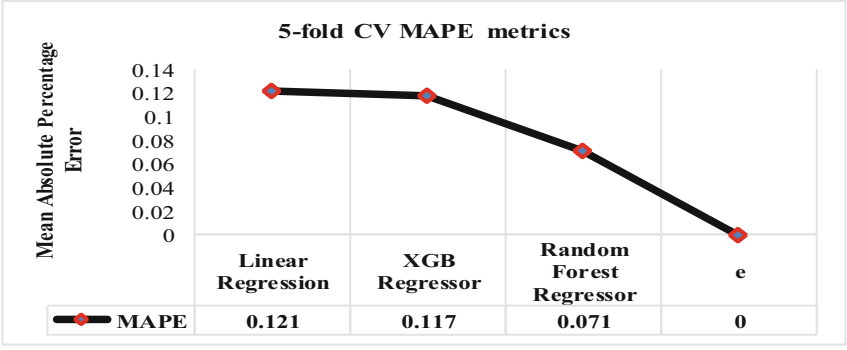

Fig. 6. 5-fold Cross validation MAPE plot for Linear Regression, XGBoost Regressor, Random Forest Regressor and *e*

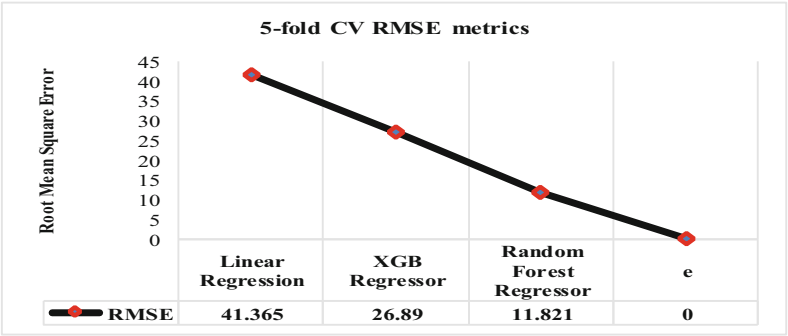

Fig. 7. 5-fold Cross validation RMSE plot for Linear Regression, XGBoost Regressor, Random Forest Regressor and *e*

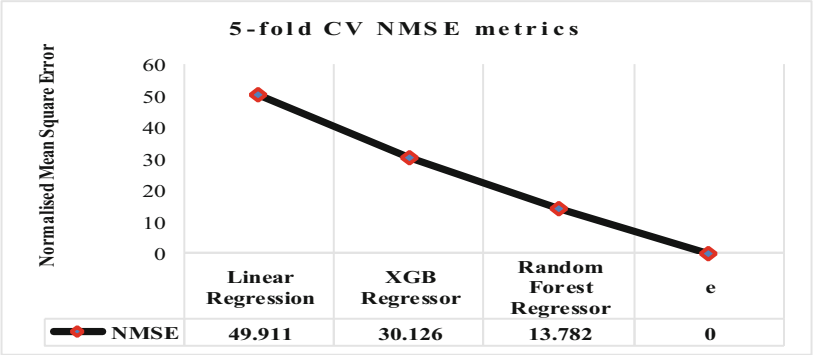

Fig. 8. 5-fold Cross validation NMSE plot for Linear Regression, XGBoost Regressor, Random Forest Regressor and *e*

Table 2. Comparing our best result with other models used to predict GDP

S/N	Study	Best Model	MSE
1	Our Research	Random Forest Regressor	139.729
2	Agu et al. (2022)	Principal Component Regression (PCR)	-7.552007365635066e + 21
3	Maccarrone et al. (2021)	K-Nearest Neighbour (KNN)	173e-03
3	Flannery (2020)	Artificial Neural Network (Multilayer Perceptron)	2946980.9

5 Conclusion

In this paper, we compared and evaluated the performance of three algorithms - Random Forest Regressor, XGBoost Regressor, and Linear Regression - in predicting the target variable. Our analysis was based on metrics, including MAE, MSE, MAPE, RMSE, and NMSE values. Our findings revealed that Random Forest Regressor outperformed the other two algorithms regarding accuracy. The results showed that Random Forest Regressor achieved the lowest MAE, MSE, MAPE, RMSE, and NMSE. On the other hand, Linear Regression was the most straightforward algorithm and performed reasonably well, while XGBoost showed high variability in performance. The rigid and high assumption of linear relationships between predictor variables and response variables, which are the limitations of the Linear Regression method, have prompted the adoption of ensemble machine learning techniques. Several models are trained using the Random Forest Regressor technique using various subsets of the training dataset. Predictions are made using an average of the predictions made by the forest's trees. Parameters must be tuned to achieve the best prediction accuracy when creating these models. Even when dealing with complex, high-dimensional data that defies linear relationships, more precise predictions can be made using this technique.

In our research, we explored the comparative performance of three algorithms: The Random Forest Regressor, XGBoost Regressor, and Linear Regression, especially in predicting GDP using notable variables like healthcare spending (hs), population (p), life expectancy (le), and the index of economic freedom (ief). Such a comprehensive approach offers policymakers detailed insight into societal and economic interrelations and bears profound practical significance. Through our analysis, we intended to guide practitioners and researchers in pinpointing the most suitable algorithm for their specific scenarios. Our findings were inclined favourably towards the Random Forest Regressor, which appeared to be especially adept for datasets characterised by complex inter-variable relationships and smaller sizes. Nonetheless, it's vital to underline the caveats in our study. Our conclusions were based on a single dataset, indicating that different datasets might yield different outcomes, especially with varied variables and aims.

Furthermore, our study was confined to these three algorithms, not accounting for potentially more apt algorithms for other unique applications. The apparent preeminence of the Random Forest Regressor within our studied dataset prompts the reference to the "No Free Lunch" theorem. This theorem emphasises that no singular algorithm is a definitive best across all conceivable contexts (Sterkenburg & Grünwald, 2021). As such, while our dataset-driven findings are significant, broad generalisations would be precipitant. It's always imperative to weigh various algorithms against distinct datasets to discern the optimal one, facilitating the creation of more nuanced and evidence-based economic policies.

Predictive models, especially those developed using statistical methods such as linear regression analysis, remain crucial in several fields. While the study presented here shows promising results for ensemble machine learning algorithms compared to linear regression analysis with a typical dataset and cross-validation technique, it is important to recognise their limitations. Traditional statistical models like linear regression analysis are often simpler and more interpretable while having fewer assumptions than some of these newer techniques.

In conclusion, our research offers insightful perspectives into the comparative performance evaluation of regression algorithms and demonstrates the importance of selecting the appropriate algorithm for specific applications. Future research can explore other algorithms and datasets to advance the field of comparative performance evaluation of algorithms in predictive analytics further.

5.1 Recommendations

To improve the accuracy of predictions, we suggest that future research explore additional non-parametric ensemble methods and Artificial Neural Networks (ANN) using related datasets and compare their predictive abilities to the methods utilised in this study. Furthermore, including more predictor variables and implementing feature selection techniques would be beneficial to determine which variables have the most significant effect on the target variable.

Disclosure of Potential Conflicts of Interest. The authors have stated that any financial or other conflicts of interest did not influence the results and writing of the paper.

References

Aftarczuk, K.: Evaluation of selected data mining algorithms implemented in Medical Decision Support Systems, Blekinge Institute of Technology School of Engineering, Blekinge (2007)

Agu, S., Onu, F., Ezemagu, U., Oden, D.: Predicting gross domestic product to macroeconomic indicators. Intell. Syst. Appl. **14**, 200082 (2022). https://doi.org/10.1016/j.iswa.2022.200082

Allawala, A., Ramteke, A., Wadhwa, P.: Performance impact of minority class reweighting on XGBoost-based anomaly detection. Int. J. Mach. Learn. Comput. **12**(4) (2022). https://doi.org/10.18178/ijmlc.2022.12.4.1093

Ameer, S., et al.: Comparative analysis of machine learning techniques for predicting air quality in smart cities. IEEE Access **7**, 128325–128338 (2019). https://doi.org/10.1109/access.2019.2925082

Bellotti, A., Brigo, D., Gambetti, P., Vrins, F.: Forecasting recovery rates on non-performing loans with machine learning. Int. J. Forecast. **37**, 428–444 (2021). https://doi.org/10.1016/j.ijforecast.2020.06.009

Brownlee, J. (2019). Why One-Hot Encode Data in Machine Learning? https://machinelearningmastery.com/why-one-hot-encode-data-in-machine-learning/. Accessed 6 Feb 2020

Chaya. Random Forest Regression - Level Up Coding. Medium, 14 April 2022. https://levelup.gitconnected.com/random-forest-regression-209c0f354c84

Chien, Y., Zhou, H., Hanson, T., Lystig, T.: Informative g-priors for mixed models. Stats **6**(1), 169–191 (2023). https://doi.org/10.3390/stats6010011

Flannery, R.: A Machine Learning Approach to Predicting Gross Domestic Product. National College of Ireland College in Dublin, Ireland (2020). https://norma.ncirl.ie/4441/1/ronanflannery.pdf. Accessed 11 Aug 2023

Gareth, J., Daniela, W., Trevor, H., Robert, T.: An Introduction to Statistical Learning. Springer, New York (ISL) (2017). https://doi.org/10.1007/978-1-0716-1418-1

Garg, A., Poornalatha, G.: Online news feed data mining and prediction. Int. J. Innov. Technol. Explor. Eng. **8**(11), 409–414 (2019). https://doi.org/10.35940/ijitee.k1381.0981119

GeeksforGeeks. (2023). XGBOOST. https://www.geeksforgeeks.org/xgboost/

Giovanni, M., Giacomo, M., Sara, S.: GDP forecasting: machine learning, linear or autoregression? Front. Artif. Intell. (2021). https://doi.org/10.3389/frai.2021.757864

Hotz, N.: What is CRISP DM? Data Science Process Alliance (2023). https://www.datascience-pm.com/crisp-dm-2/. Accessed 2 Mar 2023

Huang, L., et al.: Comparing multiple linear regression and machine learning in predicting diabetic urine albumin-creatinine ratio in a 4-year follow-up study. J. Clin. Med. **11**(13), 3661 (2022). https://doi.org/10.3390/jcm11133661

Kaliappan, J., Srinivasan, K., Qaisar, S.M., Sundararajan, K., Chang, C.C.S.: Performance evaluation of regression models for the prediction of the COVID-19 reproduction Rate. Front. Public Health **9** (2021). https://doi.org/10.3389/fpubh.2021.729795

Karen, D., Louise, S.: GDP as a Measure of Economic Well-being (2018). https://www.brookings.edu/research/gdp-as-a-measure-of-economic-well-being/. Accessed 6 Mar 2021

Katrina, W.: A Guide to the Types of Machine Learning Algorithms and their Applications (2021). https://www.sas.com/en_gb/insights/articles/analytics/machine-learning-algorithms.html. Accessed 29 Jan 2021

Lawrence, A., Sinha, R.A., Mitrasinovic, S., Price, S.J.: Clinical Features at Presentation for Glioblastoma Patients Impact Survival Predictions in a Machine Learning Model. Neuro-Oncology, 23(Supplement_4), iv18 (2021). https://doi.org/10.1093/neuonc/noab195.046

Li, H.: Responses to RC1, RC2 and RC3 – essd-2021-201. Earth System Science Data (2021). https://doi.org/10.5194/essd-2021-201-ac1

Li, N., Kwak, J., Becerik-Gerber, B., Tambe, M.: Predicting HVAC energy consumption in commercial buildings using multiagent systems. In: 30th International Symposium on Automation and Robotics in Construction and Mining; Held in Conjunction with the 23rd World Mining Congress (2013). https://doi.org/10.22260/isarc2013/0108

Maccarrone, G., Morelli, G., Spadaccini, S.: GDP Forecasting: machine learning, linear or autoregression? Front. Artif. Intell. **4** (2021).https://doi.org/10.3389/frai.2021.757864

Mirugwe, A.: Restaurant tip prediction using linear regression. Int. J. Data Sci. Big Data Anal. **1**(2), 31 (2021). https://doi.org/10.51483/ijdsbda.1.2.2021.31-38

Mislick, G.K., Nussbaum, D.P.: Linear regression analysis. J. Eval. Educ. (JEE) (2015).https://doi.org/10.1002/9781118802342.ch7

Mwiti, D.: Random Forest Regression: When Does It Fail and Why? neptune.ai, 14 November 2022. https://neptune.ai/blog/random-forest-regression-when-does-it-fail-and-why. Accessed 15 Jan 2023

Nigeria Corruption perceptions index, 2001–2022 - knoema.com. (2023). Knoema. https://knoema.com/atlas/Nigeria/Corruption-perceptions-index. Accessed 9 Apr 2023

Nguyen, H., Vu, T.H., Vo, T.P., Thai, H.: Efficient machine learning models for prediction of concrete strengths. Constr. Build. Mater. **266**, 120950 (2021). https://doi.org/10.1016/j.conbuildmat.2020.120950

Öztürk, M.Z., Colkesen, I.: Investigation of the effects of vegetation indices derived from UAV-based RGB imagery on land cover classification accuracy using advanced ensemble learning methods. Mersin Photogramm. J. (2021). https://doi.org/10.53093/mephoj.943347

Pandey, V.B., Choudhary, K., Murthy, C.S.R., Poddar, M.K.: Improved in-season crop classification performance using ensemble learning technique: a case study of lekoda insurance unit, ujjain, madhya pradesh. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XLII-3/W6, pp. 477–481 (2019). https://doi.org/10.5194/isprs-archives-xlii-3-w6-477-2019

Pandian, S.: K-Fold Cross Validation Technique and its Essentials. Analytics Vidhya (2023). https://www.analyticsvidhya.com/blog/2022/02/k-fold-cross-validation-technique-and-its-essentials/

Raheja, H., Goel, A., Pal, M.: Prediction of groundwater quality indices using machine learning algorithms. Water Pract. Technol. **17**(1), 336–351 (2021). https://doi.org/10.2166/wpt.2021.120

Raju, K.A., Sikdar, P.K., Dhingra, S.L.: Micro-simulation of residential location choice and its variation. Comput. Environ. Urban Syst. **22**(3), 203–218 (1998). https://doi.org/10.1016/s0198-9715(98)00043-x

Saabas, A.: Interpreting random forests | Diving into data, 19 October 2014. https://blog.datadive.net/interpreting-random-forests/. Accessed 15 Feb 2023

Sainikhileaswar, S., Parthasarathy, G.: Early detection of breast cancer using ensemble machine learning algorithm. Adv. Parallel Comput. (2020).https://doi.org/10.3233/apc200204

Sandra, L., Lumbangaol, F., Matsuo, T.: Machine learning algorithm to predict student's performance: a systematic literature review. TEM J. 1919–1927 (2021). https://doi.org/10.18421/tem104-56

Schonlau, M., Zou, R.Y.: The random forest algorithm for statistical learning. Stata J. **20**(1), 3–29 (2020). https://doi.org/10.1177/1536867x20909688

Shaobo, L.: Research on GDP forecast analysis combining B.P. neural network and ARIMA model. Comput. Intell. Neurosci. **2021**(Article ID 1026978) (2021). https://doi.org/10.1155/2021/1026978

Sterkenburg, T.F., Grunwald, P.: The no-free-lunch theorems of super vised learning. Synthese, 4 June 2021. https://doi.org/10.1007/s11229-021-03233-1

Tan, Z., Al-Barakati, A.: Application of Sobolev-Volterra projection and finite element numerical analysis of integral differential equations in modern art design. Appl. Math. Nonlinear Sci. (2022). https://doi.org/10.2478/amns.2021.2.00054

Wang, F., Chen, W., Fakieh, B., Alhamami, M.A.: Stock price analysis based on the research of multiple linear regression macroeconomic variables. Appl. Math. Nonlinear Sci. **7**(1), 267–274 (2022). https://doi.org/10.2478/amns.2021.2.00097

Wang, J., Lee, R.Y.: Chaotic recurrent neural networks for financial forecast. Am. J. Neural Netw. Appl. (2021) https://doi.org/10.11648/j.ajnna.20210701.12

Wikipedia. (2021). Residual sum of squares. https://en.wikipedia.org/wiki/Residual_sum_of_squares. Accessed 11 Nov 2021

World Bank. (2021). Indicators. https://data.worldbank.org/indicator/. Accessed 26 Apr 2021

Xu, Y., Cao, Z., Wang, M.: Analysis of factors influencing regional economic expansion based on OOB coefficients under RF algorithm. BCP Bus. Manag. **33**, 242–249 (2022). https://doi.org/10.54691/bcpbm.v33i.2753