Ensemble Machine Learning Methods to Predict Oil Production

M. D. Adewale[1(✉)], I. A. Adeyanju[2], J. Oju[1], O. C. Ubadike[1], U. I. Muhammed[1], and S. T. Omisakin[3]

[1] Africa Centre of Excellence On Technology Enhanced Learning, National Open University of Nigeria, Abuja, Nigeria
mdadewale@gmail.com, ace22140007@noun.edu.ng
[2] Department of Computer Engineering, Federal University of Oye-Ekiti, Oye, Nigeria
[3] DaraTom Consulting Incoporated, Toronto, ON, Canada

Abstract. This research unveils an ensemble prediction model tailored for Nigeria's oil production, addressing the need for accurate forecasting in a critical economic sector. Nigeria, a leading oil producer in Africa, relies heavily on its oil sector, necessitating robust prediction models for economic planning and stability. This study aims to create a predictive model integrating pivotal factors such as oil reserves, oil consumption, oil prices, and political stability. These factors were chosen due to their significant impact on oil production dynamics, encompassing economic, political, and consumption-related influences that critically determine production outcomes. Using data from 1980 to 2016, we employed advanced machine learning algorithms—including Extra Trees Regressor, XGBoost Regressor, and Random Forest Regressor—to enhance prediction accuracy. The Extra Trees Regressor emerged as the superior algorithm, demonstrated by a correlation coefficient of 0.8155, a mean absolute error of 0.2812, and a root mean squared error of 0.3929. Our findings confirm the model's predictive power, highlighting the significant influence of critical variables on Nigeria's oil production trends. This study offers indispensable insights to stakeholders, aiding in informed decision-making and showcasing the significant capabilities of ensemble machine learning in improving oil production forecasts. Ultimately, this work enhances understanding of the factors affecting oil production and supports strategic planning within Nigeria's oil sector, promoting economic stability and growth.

Keywords: Oil Price · Machine Learning · Oil Consumption · Oil Production · Oil Reserve

1 Background of the Study

Developing a predictive model for oil production in Nigeria is essential due to its significant economic implications. Nigeria, a leading oil producer in Africa, relies heavily on its oil sector, which contributes substantially to its GDP. Accurate oil production forecasts enable the government and stakeholders to make informed decisions about fiscal

© ICST Institute for Computer Sciences, Social Informatics and Telecommunications Engineering 2025
Published by Springer Nature Switzerland AG 2025. All Rights Reserved
K. Cheikh M. F. Kebe et al. (Eds.): InterSol 2024, LNICST 610, pp. 338–357, 2025.
https://doi.org/10.1007/978-3-031-86493-3_27

budgeting, infrastructure investments, and long-term economic policies, which are crucial for managing global oil market volatility that affects national revenue and economic health (Obite et al., 2021).

Such a model can enhance operational efficiency by anticipating fluctuations in oil production and allowing for proactive strategies in drilling operations, maintenance, and workforce management. This foresight is essential for maintaining competitiveness in the global oil market, especially amidst growing environmental concerns and the shift towards renewable energy sources (Gbakon et al., 2022).

A predictive model aids in forecasting production trends, facilitating effective environmental protection measures and sustainable resource management and addressing issues like oil spillage and environmental degradation in Nigeria (Asumadu-Sarkodie & Owusu, 2016; Gbakon et al., 2022).

The model also promotes socio-economic stability by informing labour market strategies, bolstering job security, and supporting workforce development. A predictive model for Nigeria's oil production is crucial for economic optimisation, sustainable development, environmental stewardship, and socio-economic stability (Asumadu-Sarkodie and Owusu, 2016; Falode & Udomboso, 2021; Gbakon et al., 2022).

However, the inherent volatility of global oil markets and Nigeria's fluctuating political and economic conditions present significant challenges for accurate oil production forecasting. Traditional models often fail to capture the complex interplay of socio-economic and political factors influencing oil production, leading to suboptimal forecasting accuracy and unreliable policy decisions.

This study aims to develop a more accurate and reliable forecasting model for Nigeria's oil production by incorporating various socio-economic and political factors using ensemble machine learning techniques. The goal is to improve forecasting precision, supporting informed decision-making and strategic planning in Nigeria's oil sector. The central aim is to create a system capable of processing socio-economic data and converting it into meaningful insights for future oil production forecasting using ensemble machine learning algorithms to attain high accuracy levels.

The following research inquiries are addressed to clarify the contributions of this work through the developed model:

(**R.Q.1**) *Which variables directly influence the value of oil production?*

(**R.Q.2**) *Can the Ensemble Machine Learning Model effectively forecast oil production values based on various factors?*

These questions guide the study's focus on creating a predictive model that leverages the latest technological advancements to address the complexities of oil production forecasting in Nigeria.

1.1 Identifying Key Variables Influencing Oil Production Value

(R.Q.1) Which variables directly influence the value of oil production?

This study acknowledges several critical factors in exploring the variables significantly impacting oil production. These include oil consumption levels, oil reserves, global exchange rates, environmental considerations, political dynamics, and oil market speculation. These elements play a vital role in shaping oil supply and demand. For

instance, oil consumption, directly correlated with demand, increases production as consumption levels rise (Karakurt, 2021). Oil reserves, a primary supply determinant, can enhance production capabilities when abundant (Falode & Udomboso, 2021). Exchange rates also influence oil prices and, by extension, supply and demand, particularly in significant oil-importing and exporting countries (Salisu et al., 2020). As the global energy landscape shifts towards sustainability, environmental policies to reduce carbon emissions significantly impact oil demand (Asumadu-Sarkodie & Owusu, 2016). Political factors, including stability and government policies, significantly affect oil supply and demand (Gbakon et al., 2022). Financial market speculations can also cause oil price fluctuations, affecting perceived supply and demand (Obite et al., 2021; Salisu et al., 2020).

The need for a predictive model in Nigeria that encompasses oil reserves, consumption, oil price, and political stability is highlighted by several studies. These models demonstrate the complexity of forecasting oil production and the importance of various interacting factors. Cavallo (2004) notes the role of stable markets, demand growth, import availability, and accurate reserve estimates in predicting oil production, which is particularly relevant given Nigeria's fluctuating political and economic conditions (Cavallo, 2004). Moroney and Berg (1999) show that integrating physical reserves with economic variables like exchange rates yields more accurate forecasts than models based solely on reserves or economic data (Moroney & Berg, 1999). The importance of considering political, economic, and resource-related variables is underscored by Duhu (2019), especially pertinent for Nigeria, where political stability influences both domestic and global market perceptions. Karakurt (2021) emphasises the need to model and forecast oil consumption, a critical factor in oil production, highlighting the importance of understanding consumption patterns for accurate forecasting (Karakurt, 2021).

These studies collectively indicate the necessity of a comprehensive predictive model that includes oil reserves, consumption, oil price, and political stability. Such a model is critical for Nigeria to aid in economic planning and effectively navigate the complexities of the global oil market.

1.2 Efficacy of the Ensemble Machine Learning Model

(*R.Q.2) Can the Ensemble Machine Learning Model effectively forecast oil production values based on various factors?*

This inquiry evaluates the efficacy of ensemble machine learning models, explicitly incorporating XGBoost, Random Forest, and Extra Trees Regressor, in accurately forecasting Nigerian oil production. The strength of these ensemble models lies in their sophisticated ability to process and analyse complex datasets, identifying nuanced patterns crucial for precise prediction tasks. XGBoost is renowned for its efficiency and accuracy across various prediction challenges. It is a scalable ensemble technique that has demonstrated reliability and efficiency in solving complex machine-learning challenges. Random Forest Regressor builds upon bagging multiple decision trees to reduce variance and improve prediction accuracy. It effectively captures nonlinear relationships without the need for extensive parameter tuning. The Extra Trees Regressor introduces additional randomness compared to Random Forest by randomising thresholds for feature splits,

leading to more diversified tree ensembles. This approach enhances the model's generalisation ability, making it suitable for complex regression tasks. This research aims to assess how effectively these models can utilise inputs such as oil reserves, consumption patterns, exchange rates, and political stability to forecast oil production accurately.

The significance of employing ensemble machine learning models, which amalgamate the predictive power of multiple algorithms, has been underscored in recent literature. Studies by Sagi and Rokach (2018) and Shahhosseini et al. (2020) have shown that ensemble models surpass the predictive performance of individual models in complex data environments. The adaptability and robustness of ensemble approaches, highlighted in research by Sagi and Rokach (2018) and further supported by Shahhosseini et al. (2020), advocate for their application in dynamic and complex markets such as the oil sector. These attributes are precious in addressing Nigeria's oil sector's unique and challenging aspects, which may not align with global oil production trends.

Given the distinct characteristics and challenges inherent to Nigeria's oil industry, there is a critical need for a predictive model custom-fitted to the Nigerian context. This model must capture the general predictors of oil production and account for local nuances that influence Nigeria's production outcomes. Therefore, this study proposes using an ensemble machine learning framework that leverages the combined strengths of XGBoost, Random Forest, and Extra Trees Regressor. By integrating these advanced algorithms, the model aims to offer a more refined and context-sensitive tool for forecasting. This approach is expected to yield more accurate, reliable predictions, facilitating better-informed decision-making and strategic planning within Nigeria's oil sector.

In essence, by harnessing the capabilities of these ensemble machine learning models, this research seeks to advance the accuracy and relevance of oil production forecasts in Nigeria, thereby contributing to the optimisation of production strategies and broader economic planning within the country.

1.3 Gaps in the Previous Studies

Previous works on forecasting oil production have left significant gaps that this study aims to address. Firstly, traditional models have often focused narrowly on physical and economic factors, neglecting broader socio-economic and political influences on oil production (Moroney & Berg, 1999). Additionally, these models struggle to adapt to the dynamic nature of the global oil market, especially in regions like Nigeria, known for market volatility (Falode & Udomboso, 2021). Moreover, existing models lack the predictive accuracy and efficiency needed in today's fast-paced energy sector, often failing to capture the nonlinear relationships between factors affecting oil production (Obite et al., 2021). Policy decisions in Nigeria's oil sector have relied on historical data rather than contemporary insights, leading to delayed responses to market changes (Gbakon et al., 2022). Furthermore, traditional forecasting models have overlooked environmental and socio-economic considerations, highlighting the need for a more holistic approach to promote sustainable practices (Asumadu-Sarkodie & Owusu, 2016). Finally, generalised models developed with a global perspective may not adequately address the unique complexities of the Nigerian oil market (Karakurt, 2021).

1.4 Novelty in the Current Work

This study introduces an advanced predictive model for forecasting Nigerian oil production using the Ensemble Machine Learning Approach, including the Extra Trees Regressor, Random Forest Regressor, and XGBoost algorithms. The model integrates socio-economic, political, and environmental factors, providing a holistic analysis of influences on Nigeria's oil production. It adapts to market dynamics through adaptive algorithms like Extra Trees Regressor, capturing the global oil market's complexities. The ensemble machine learning algorithms ensure high predictive accuracy by handling complex data relationships. Additionally, the model offers real-time, data-driven insights for informed decision-making, crucial for strategic planning in Nigeria's oil industry. It assesses oil production's environmental and socio-economic impacts and aligns with sustainable development goals. Custom-tailored for Nigeria, this approach offers practical forecasting solutions specific to the country's challenges. This work significantly advances predictive modelling in the oil industry, addressing gaps and introducing tailored solutions for Nigerian oil production forecasting.

2 Methods and Techniques

In our study, we adopted a comprehensive approach to forecasting Nigeria's oil production using advanced machine-learning techniques. Our initial step involved gathering secondary data from several authoritative sources, including Worldometer, Statista, and TheGlobalEconomy.com, focusing specifically on Nigeria's oil reserves, production, consumption statistics, and political rights. Following the data acquisition phase, we began data preprocessing to ensure compatibility with our machine learning models. Given the diverse scales and units across our dataset's features, we normalised the data to a range of [0, 1]. This normalisation was crucial for accommodating our data's non-Gaussian distribution and optimising the performance of the selected algorithms, particularly for models like Extra Trees Regressor and XGBoost, where feature scale uniformity enhances model training efficiency. Although Random Forest Regressor is inherently scale-invariant, normalisation facilitated a consistent preprocessing pipeline across all models (Huangi et al., 2018; Li, 2019; Mrabet et al., 2022). The dataset was then partitioned into training and test sets (70/30), employing a 5-fold cross-validation method to maximise the use of our data in a robust evaluation framework. This approach is beneficial for datasets of limited size because it ensures comprehensive training and validation across multiple folds while mitigating the risks of overfitting and variance, which can compromise model reliability.

The core of our analysis involved the application of three distinct machine learning models: Extra Trees Regressor, XGBoost, and Random Forest Regressor. Each model was chosen for its unique strengths in predictive accuracy and handling complex datasets. We meticulously trained these models on the normalised data and then conducted an exhaustive evaluation phase. This phase utilised a suite of performance metrics, including Mean Absolute Error, Relative Absolute Error, Root Mean Squared Error, Root Relative Squared Error, and the Correlation Coefficient, to assess and compare each model's predictive accuracy and reliability. The final stage of our research focused on leveraging the trained models to predict Nigeria's oil production, showcasing the potent application

of machine learning in forecasting within the energy sector. The procedural flow of our research, from data collection to oil production forecasting, is detailed in Fig. 1, providing a visual overview of the methodological steps undertaken in our analysis.

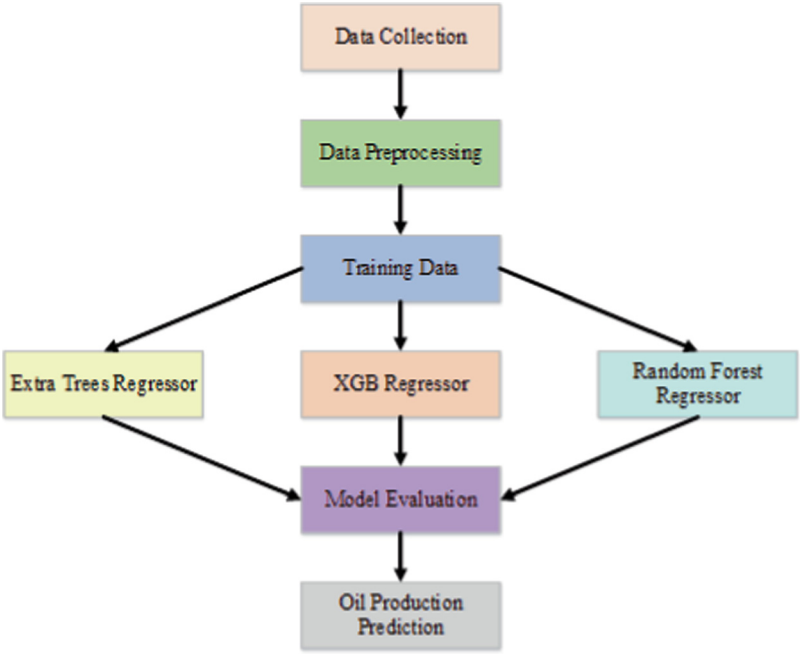

Fig. 1. Steps Followed in Oil Production Prediction

2.1 The Ensemble Algorithms Approach

Predicting Nigerian oil production with advanced ensemble machine learning models offers a nuanced understanding and higher accuracy in prediction tasks. This section elaborates on using XGBoost, Random Forest Regressor, and Extra Trees Regressor, each bringing unique strengths to the ensemble approach.

2.1.1 Extra Trees Regressor (Extremely Randomised Trees Regressor)

The Extra Trees Regressor introduces additional randomness compared to RandomForest by randomising thresholds for feature splits. This approach can lead to more diversified tree ensembles, enhancing the model's generalisation ability. We highlight the effectiveness of the Extra Trees Regressor in feature importance extraction and regressor fitting for predictive modelling, providing valuable insights into influential factors affecting outcomes. The adaptability and efficiency of the Extra Trees Regressor make it an essential tool for predictive modelling across various fields (M. Devi et al., 2019).

The selection of splits in the Extra Trees Regressor can be represented as a randomised process, differing from the optimal split selection in traditional decision tree algorithms. This randomised process described by Geurts, Ernst, and Wehenkel (2006) could be visualised through algorithmic descriptions, highlighting the randomness in split selection. The process can be summarised in pseudo-code, as shown in Fig. 2:

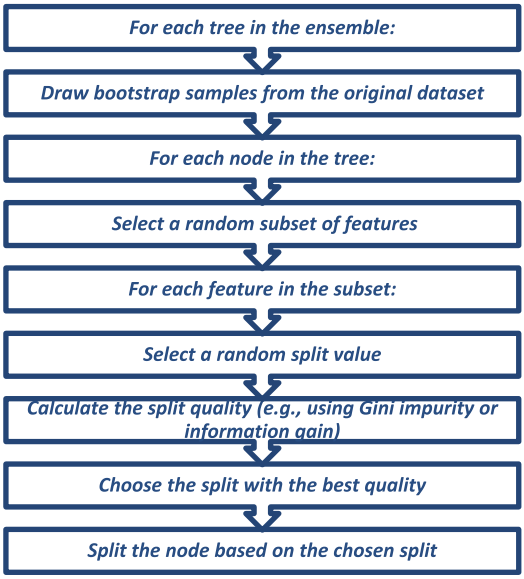

Fig. 2. Pseudo-code for ExtraTrees Regressor

2.1.2 XGBoost (Extreme Gradient Boosting)

XGBoost, standing for Extreme Gradient Boosting, is recognised for being a powerful and efficient ensemble learning framework that excels in a wide range of predictive modelling tasks. It distinguishes itself by simultaneously optimising predictive accuracy and computational efficiency, which makes it exceptionally suitable for handling complex data sets. At the heart of XGBoost's approach is an objective function it seeks to minimise, which artfully combines a loss function with a regularisation term to enhance model performance and robustness. The objective function is formulated as:

$$\text{Obj} = \sum_{(i)} \boldsymbol{L}\left(y_i, \widehat{y_i}\right) + \sum_{(k)} \boldsymbol{\Omega}(f_k) \tag{1}$$

where $\boldsymbol{L}$ is the loss function that measures the difference between the predicted $\hat{\boldsymbol{y}}_{\boldsymbol{i}}$ and actual $\boldsymbol{y}_{\boldsymbol{i}}$ values and Ω represent the regularisation term to prevent overfitting (Chen & Guestrin, 2016). XGBoost is an ensemble method, precisely a type of boosting, where it builds models sequentially, each new model correcting errors made by previous models and combining them to produce a final, more accurate prediction (Adewale et al.,

2024a; Shi et al., 2021). This method effectively improves the model's performance by integrating the strengths of multiple learners.

2.1.3 Random Forest Regressor

Random Forest Regressor builds upon bagging multiple decision trees to reduce variance and improve prediction accuracy. It is particularly effective in capturing nonlinear relationships without extensive parameter tuning. The effectiveness of the Random Forest Regressor in various applications, including the prediction of train arrival delays, demonstrates its reliability and adaptability in handling complex datasets and achieving high accuracy in predictions (Ji et al., 2020).

Incorporating these advanced ensemble models for forecasting Nigerian oil production promises higher accuracy and provides a methodologically sound and innovative approach adaptable to the domain's complexities. Integrating recent studies and mathematical formulations will further strengthen the research, offering a scientifically rigorous foundation for these predictive models.

2.2 Empirical Validation

The metric of accuracy—five well-established statistical evaluation measures (see Eqs. 2 to 6—forms the basis of our validation process. The selection of these metrics is informed by the work of notable researchers in the application of machine learning to solve practical problems, including Adewale et al., 2024a, Adewale et al., 2024b, and Yao et al. (2010). By integrating a comprehensive suite of evaluation criteria, our approach ensures a robust and multidimensional assessment of the hybrid model's forecasting capabilities.

The mean absolute error (MAE) measures the average magnitude of the absolute differences between predicted and actual values, representing the average error magnitude.

$$\text{Mean Absolute Error (MAE)} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right| \tag{2}$$

The root mean squared error (RMSE) calculates the square root of the average of the squared differences between predicted and actual values, indicating the standard deviation of the errors.

$$\text{Root Mean Squared Error (RMSE)} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{3}$$

The relative absolute error (RAE) compares the total absolute error of the predictions to the total absolute error of a naive baseline model, which always predicts the mean of the actual values.

$$\text{Relative Absolute Error (RAE)} = \frac{\sum_{i=1}^{n}\left|y_i - \widehat{y_i}\right|}{\sum_{i=1}^{n}\left|y_i - \overline{y}\right|} \tag{4}$$

The root relative squared error (RRSE) is similar to the RMSE but is normalised by the variance of the actual values.

$$\text{Root Relative Squared Error(RRSE)} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}} \tag{5}$$

The correlation coefficient (Pearson's r) evaluates the linear relationship between predicted and actual values, ranging from −1 to 1.

$$\text{Correlation Coefficient}\left(\text{Pearson}'\text{s } r\right) = \frac{\sum_{i=1}^{n}(y_i - \overline{y})\left(\widehat{y_i} - \overline{\hat{y}}\right)}{\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2 \ \sum_{i=1}^{n}\left(\widehat{y_i} - \overline{\hat{y}}\right)^2}} \tag{6}$$

where:

- n *is the total number of observations.*
- y_i *denotes the actual outcome for the* i^{th} *observation.*
- $\widehat{y_i}$ *is the predicted value for the* i^{th} *observation.*
- $\overline{y}$ *is the mean of the actual values.*
- $\overline{\hat{y}}$ *is the mean of the predicted values.*

The proximity between predicted and actual values is directly proportional to the decrease in error values. These metrics allow us to gauge the accuracy of our predictions against actual outcomes, with lower error values and a correlation coefficient closer to 1, indicating better model performance.

2.3 Research Model

This research crafted a unique model targeting oil production (OPR), weaving together four pivotal predictors: oil reserves (OR), oil consumption (OC), oil prices (OP), and political rights (PR). Central to the model, these factors define OPR as the outcome of interest. Figure 3 shows the conceptual model of the oil production prediction system. The investigation aimed to unravel the nexus between OPR, a metric of the nation's energy sector prowess, and its influencing predictors, such as reserves, consumption rates, exchange rate shifts, and political rights, often seen as a foundation for political stability. When citizens can express their opinions, participate in free and fair elections, and engage in the political process, it can lead to a more satisfied populace. The study employed these predictors to forecast OPR, utilising machine learning techniques to fine-tune the estimation of model parameters. The essence of this research model is succinctly captured in fundamental Eq. 7, which is in line with the strategy delineated by Gareth et al. (2017), as shown in Eq. 7.

$$OPR = f(OR, OC, OP, PR) \tag{7}$$

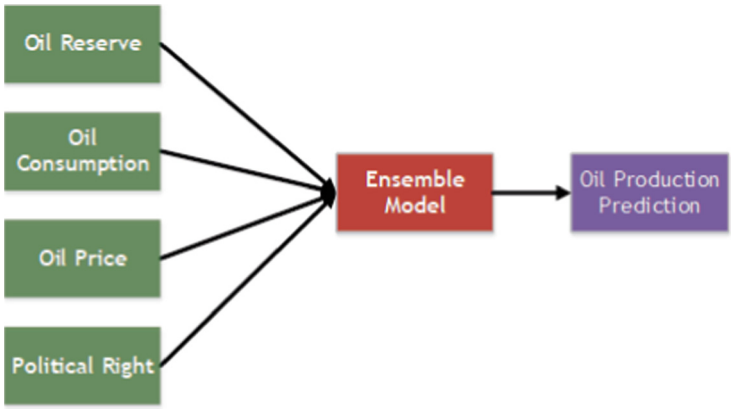

Fig. 3. Conceptual model showing the oil production prediction system

2.4 Hypothetical Case Study Application

This case study applies our research model to predict Nigeria's oil production, utilising historical data to demonstrate its efficiency and practical implications. By focusing on key variables—oil reserve (OR), oil consumption (OC), oil price (OR), and political rights (PR)—the model assesses how these factors collectively influence oil production trends.

The parameters for the study are set as follows:

- **Oil Reserve (OR):** This is measured in billions of barrels.
- **Oil Consumption (OC):** This is recorded in millions of daily barrels.
- **Oil Price (OR):** This is represented in USD.
- **Political Right (PR):** This is scored based on the electoral process, political diversity and engagement, and government effectiveness. Ratings range from 1 (high rights) to 7 (low rights).

Using machine learning algorithms, the model predicts oil production and highlights the potential impact of policy adjustments in these areas. By analysing the practical implications, we demonstrate how adjustments in oil reserves, consumption, pricing policies, and political reforms can enhance Nigeria's oil output. For example, expanding oil reserves through exploration and investment can boost production capacity. Similarly, policies to stabilise oil prices and improve political rights can create a more favourable environment for oil production.

This approach offers policymakers a compact yet comprehensive tool, showcasing the model's utility in guiding strategic decisions. The model helps formulate policies that optimise oil production and stimulate economic growth by providing actionable insights and demonstrating their practical implications in real-world scenarios.

2.5 Data Sources, Definitions and Measurement of Variables

The dataset under examination extends from 1980 to 2016, incorporating a broad spectrum of data across several years to analyse socio-economic factors (Nigeria Oil

Reserves, Production and Consumption Statistics - Worldometer (2024); *Nigeria: Oil Production 2022 | Statista* (2023); *Brent Crude Oil Price Annually 1976–2024 | Statista* (2024); *Nigeria Political Rights - Data, Chart | TheGlobalEconomy.com* (2024). Table 1 offers a glimpse into a segment of this dataset, covering the years 1999 to 2004, and illustrates a selection of key variables relevant to the study. These variables include:

Oil Production (OPR): The country's daily crude oil production, measured in millions of barrels, encapsulating both onshore and offshore activities. This metric is pivotal for assessing a nation's contribution to the global energy sector and economic vitality.

Oil Reserves (OR): This is quantified in billions of barrels, indicating the estimated quantities of crude oil that are anticipated to be commercially recoverable.

Oil Price (OP): The cost of crude oil, expressed in USD, reflects the global market valuation.

Political Rights (PR): Evaluated on a scale from 1 (indicating high levels of political rights) to 7 (signifying low rights), this score is derived from an analysis of the electoral process, political pluralism, participation, and the functionality of government.

Oil Consumption (OC): This is measured in millions of barrels daily, representing the total domestic oil volume.

This subset of the dataset, as outlined in Table 1, provides a concise overview of these critical indicators within the specified timeframe, enabling a focused analysis of the trends and patterns that emerge from these variables.

Table 1. Part of the Dataset used for modelling

Year	OPR	OR	OP	PR	OC
1999	2136263.00	22500000000.00	17.90	4.00	252000.00
2000	2169135.00	22500000000.00	28.40	4.00	227354.00
2001	2261416.00	22500000000.00	24.45	4.00	286822.00
2002	2123323.00	24000000000.00	25.01	4.00	292196.00
2003	2278592.00	24000000000.00	28.83	4.00	280225.00
2004	2331696.00	25000000000.00	38.10	4.00	280188.00

3 Results and Discussion

3.1 Descriptive Statistics

The descriptive statistics in Table 2 detail various indicators related to oil dynamics and governance in Nigeria, covering 37 instances, and provide insights into the critical variables analysed in this study. Indicators include oil production (OPR), oil reserve

(OR), oil price (OP), political rights (PR), and oil consumption (OC). The average oil production (OPR) of 2,001,529 barrels, with a standard deviation of 385,432.20, highlights variability in output, reflecting differing capacities across oil fields and its critical economic role. The oil reserve (OR) data, normalised around a mean of 0 and a standard deviation of 1.01, suggests a wide range of recoverable crude oil, emphasising strategic reserves. Oil Price (OP), averaging 41.70 USD with a standard deviation of 30.63, illustrates global price fluctuations impacting Nigeria, driven by supply-demand dynamics, geopolitical tensions, and economic conditions. Normalised political rights (PR) scores reveal a range of political climates crucial for understanding governance and its effect on oil production policies. Also, normalised oil consumption (OC) shows varied usage levels, indicating domestic energy demand and economic activity. This analysis offers insights into the interplay between oil production, reserves, pricing, consumption, and political governance, underscoring the complexities of managing oil resources and the importance of political stability for sustainable energy production and economic growth.

Table 2. Descriptive Statistics

	OPR	OR	OP	PR	OC
count	37.00	37.00	37.00	37.00	37.00
mean	2001529.00	0.00	41.70	0.00	0.00
std	385432.20	1.01	30.63	1.01	1.01
min	1246000.00	–0.97	12.80	–1.74	–1.80
25%	1817000.00	–0.86	19.10	–0.49	–0.70
50%	2060000.00	–0.40	28.70	–0.49	–0.09
75%	2278592.00	1.34	54.38	1.39	0.40
max	2630860.00	1.45	111.63	1.39	3.14

3.2 Test for Multicollinearity

The Variance Inflation Factor (VIF) is a measure used to identify the presence of multicollinearity among independent variables in a regression analysis. Using the rule of thumb that a VIF greater than 10 indicates significant multicollinearity, as stated by Forthofer et al. (2007), our analysis of the VIF factors for features related to oil production, as shown in Table 3, demonstrates that none of the predictors—oil reserve, oil price, political right, and oil consumption—exceed this threshold. Specifically, oil reserve and oil price, with VIFs of 7.314537 and 5.235808, respectively, indicate moderate multicollinearity, which, under this stricter criterion, does not present a significant concern for the regression model's stability or its interpretability concerning predicting oil production. Similarly, political rights and oil consumption, with VIFs significantly below 10, show minimal to low multicollinearity, reinforcing their suitability as predictors without necessitating immediate corrective measures to address multicollinearity.

As Cripps and Nguyen (2001) note, multicollinearity does not affect the model's predictive capacity within the observed data range, although it complicates identifying individual variable impacts. This aspect, while important, is less critical for our research objectives. In light of this and considering the more conservative approach to managing multicollinearity, our findings suggest that the current variables can remain within the model. The lack of significant multicollinearity, as evidenced by VIF values not exceeding the ten thresholds, suggests that our model maintains sufficient robustness for the intended analysis. This finding is in harmony with our research objectives, which emphasise the model's overall predictive accuracy and understanding of each independent variable's specific impact on Oil Production. Consequently, this reduces the urgency for immediate modifications to diminish multicollinearity within the model.

Table 3. VIF for each feature

S/N	Features	VIF Factor
1	Oil Reserve	7.314537
2	Oil Price	5.235808
3	Political Right	1.196773
4	Oil Consumption	2.035931

.

3.3 Performance Evaluation of Extra Trees, XG Boost, and Random Forest Regressor

In the endeavour to predict Nigeria's Oil Production (OPR) using a combination of economic and non-economic indicators, Table 4 delineates the performance metrics of three distinct machine-learning models. We rigorously evaluated each model's efficacy across a spectrum of metrics, including the relative absolute error (RAE), mean absolute error (MAE), root relative squared error (RRSE), root mean squared error (RMSE), and correlation coefficient. These metrics offer a holistic view of each model's predictive accuracy. Figure 4 graphically contrasts the performance of the models, providing a visual interpretation of their effectiveness.

The Extra Trees Regressor emerged as the most accurate model, as evidenced by its superior correlation coefficient of 0.8155, indicating a solid alignment between the predicted and actual OPR values (See Fig. 5). It also maintained the lowest error margins, with an MAE of 0.2812 and an RMSE of 0.3929, suggesting high precision in its predictions. Conversely, the XGBoost Regressor, while demonstrating reasonable accuracy with a correlation coefficient of 0.6434, showed higher error rates (MAE of 0.4032 and RMSE of 0.5222) compared to the Extra Trees Regressor. Although a robust model in various contexts, the Random Forest Regressor exhibited a relatively lower correlation coefficient of 0.4362 in this scenario. Its error metrics, with an MAE of 0.5607 and RMSE of 0.6211, further indicated its limitations in accurately forecasting Nigeria's OPR.

Table 4. Model's Evaluation

Performance Metrics	Extra Trees Regressor	XGBoost Regressor	Random Forest Regressor
Correlation coefficient	0.8155	0.6434	0.4362
Mean absolute error	0.2812	0.4032	0.5607
Root mean squared error	0.3929	0.5222	0.6211
Relative absolute error	0.3481	0.5055	0.5431
Root relative squared error	0.4044	0.5440	0.5757

Fig. 4. Performance Metrics of Machine Learning Algorithms

These findings underscore the importance of model selection in machine learning tasks. The choice of algorithm has a profound impact on the accuracy and reliability of predictive analyses, as demonstrated by the varied performance of the evaluated models in the context of this study. Regarding R.Q.2, the Extra Trees Regressor has proven to be the most effective model, providing accurate forecasts based on the selected variables.

3.4 Comparative Analysis and Validation

Considering the comparative analysis in Table 5, our findings align with the 'No Free Lunch' theorem, which asserts that no single algorithm excels universally (Sterkenburg & Grünwald, 2021). This principle is evident in the varied performance of algorithms across different studies. Our application of the Extra Trees Regressor yielded a mean absolute error (MAE) of 0.2812, demonstrating notable accuracy within our Nigerian dataset.

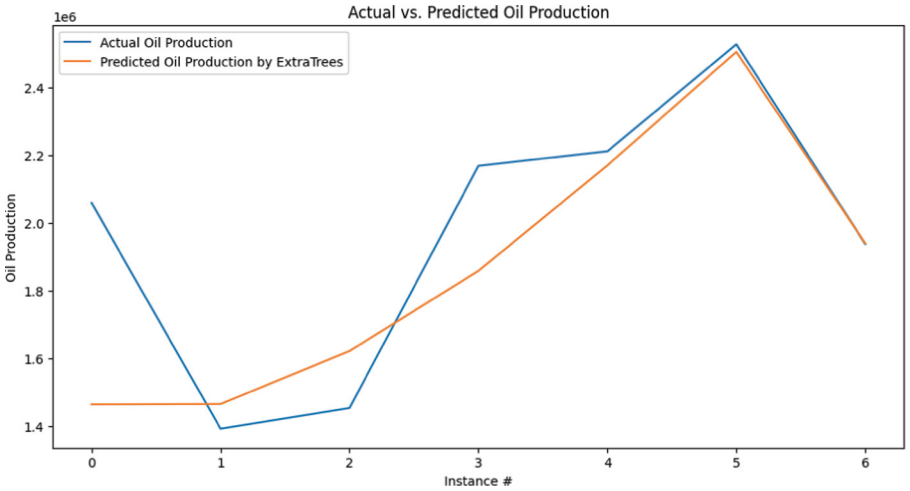

Fig. 5. Actual vs Predicted Oil Production by ExtraTrees Model

However, this success does not guarantee similar results in other economies due to varying sensitivities to socio-economic factors. Therefore, testing and validating these algorithms in diverse contexts is crucial for tailored machine-learning applications in economic forecasting.

Table 5 includes our study using the Extra Trees Regressor, which focuses on variables such as oil reserve, oil price, political rights, and oil consumption over the past 37 years. Another study by Al-qaness et al. (2022) employs the Modified Aquila Optimizer with Opposition-Based Learning, analysing oilfield production data over 11 years. AlRassas et al. (2021) use a neurofuzzy-based slime mold algorithm on similar data. Each study's methodology, independent variables, and dataset size uniquely contribute to understanding oil production prediction. The Extra Trees Regressor's low MSE in our study signifies its effectiveness in capturing socio-economic nuances within our dataset. The Extra Trees Regressor excels at handling high-dimensional data and offers valuable insights into feature importance, making it particularly useful for policymakers.

While the MAE reported by Al-qaness et al. (2022) is lower than our model's, it is important to note several contextual factors. First, the dataset size and composition in Al-qaness et al.'s study differ, focusing on various oilfields' production data over 11 years, which may present different dynamics compared to the broader 37-year dataset used in our study. Additionally, the Modified Aquila Optimizer with Opposition-Based Learning employed by Al-qaness et al. (2022) may be particularly well-suited to their specific dataset characteristics, potentially explaining the lower MAE.

However, our approach has several strengths that extend beyond the MAE metric. The Extra Trees Regressor offers robustness in handling high-dimensional data and provides insights into feature importance, highlighting critical predictors like 'Oil Reserve' and 'Oil Consumption.' This transparency in the decision-making process is particularly valuable for policymakers who require understandable and actionable insights.

Moreover, our model incorporates a broader range of socio-economic and political factors, providing a comprehensive view of the determinants affecting oil production. This enriched model can better inform strategic decisions and policy formulations.

Furthermore, our methodology's application of rigorous cross-validation techniques ensures that the model's performance is consistently reliable across different subsets of the data, minimising the risks of overfitting and enhancing its generalizability. Incorporating diverse predictors such as exchange rates and political stability also broadens the analytical lens, offering a more holistic understanding of the factors driving oil production. While the MAE of our model may not be the lowest compared to some other studies, the Extra Trees Regressor's robustness, transparency, and comprehensive analytical approach provide significant advantages. These strengths highlight the model's potential for practical application in policy development and strategic planning within Nigeria's oil sector.

3.5 Feature Importance

Figure 6 highlights the varying impact of different attributes within the Extra Trees model. The feature importance graph for the Extra Trees model indicates that 'Oil Reserve' and 'Oil Consumption' are the most influential variables when predicting oil production, with 'Oil Reserve' being the most significant. 'Oil Price' and 'Political Right' have less importance in the model, with 'Political Right' being the least influential. This suggests that while market factors like oil price and political stability play a role, the physical availability of oil and consumption patterns are stronger predictors in this model's context. Our findings indicate that 'Oil Reserve' and 'Oil Consumption' are the most significant predictors of oil production, addressing R.Q.1.

Table 5. Comparing our best result with other models used to predict Oil production

S/N	Reference	Top Performing Model	Independent variables used	The size of the dataset	Mean Absolute Error
1	Our Investigation	ExtraTrees Regressor	Oil reserve, oil price, political rights, and oil consumption	37 years	0.281
2	Al-qaness et al. (2022)	Modified Aquila Optimizer (AO) with the Opposition-Based Learning (OBL) technique	Various oilfields' production data	11 years	0.080
3	AlRassas et al. (2021)	Neuro fuzzy-based slime mould algorithm	Various oilfields' production data	11 years	0.693

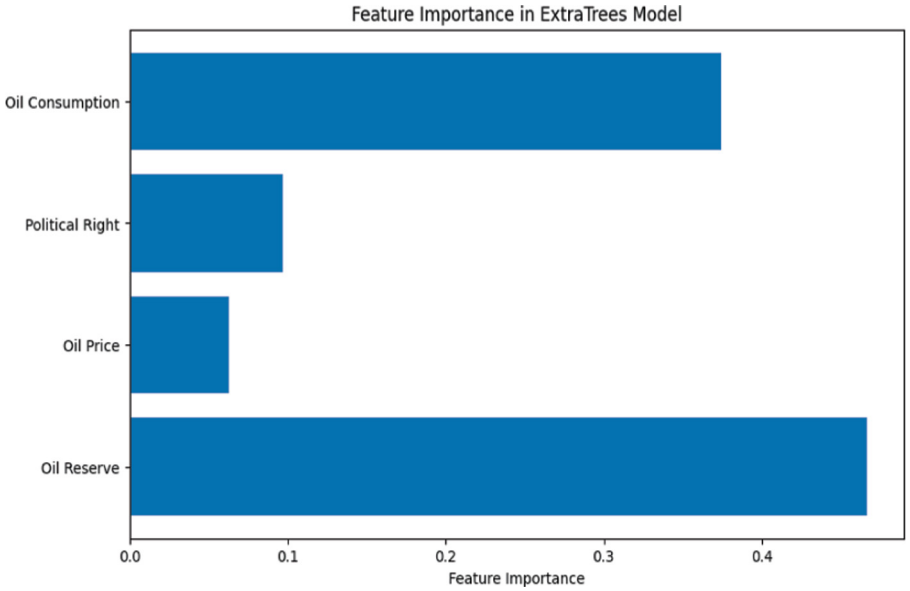

Fig. 6. The Features Importance of the Extra Trees Model.

3.6 Limitations of the Methodologies and Models Used in the Study

While extra trees regressor models effectively handle complex, high-dimensional data, they have certain limitations. High computational demands can limit scalability, and the model's performance may vary significantly across different economic contexts, necessitating careful localisation and data adaptation. The model's accuracy depends heavily on data quality; incomplete or biased datasets can impair performance. Furthermore, the model may not generalise well to other regions without significant modifications, limiting its immediate applicability to broader contexts.

Future work should focus on enhancing the robustness and generalizability of the proposed method. This can be achieved by exploring and integrating other machine learning algorithms to complement the strengths of the Extra Trees Regressor, developing hybrid models to address specific weaknesses, and implementing transfer learning approaches to adapt the model to new regions or contexts with minimal retraining. Additionally, extending the methodology to other economic forecasting tasks beyond oil production and incorporating additional data sources, such as real-time market data, can improve the model's resilience and predictive capabilities. These steps can make the model more versatile and practical for various economic forecasting applications.

4 Conclusion

In conclusion, our study enhances the precision of Nigeria's oil production forecasts by comparing the Extra Trees Regressor, XGB Regressor, and Random Forest Regressor using performance indicators such as MAE, RAE, RMSE, RRSE, and the Correlation

Coefficient. The Extra Trees Regressor emerged as the most accurate model, effectively capturing socio-economic nuances and reducing forecasting errors. Key predictors identified include 'Oil Reserve' and 'Oil Consumption,' highlighting the model's robustness in economic forecasting. By incorporating variables like Oil Reserves, Oil Consumption, Exchange Rates, and Political Stability, our approach comprehensively analyses Nigeria's oil production, emphasising the influence of economic policies, market dynamics, and political stability. This challenges the traditional reliance on direct production and economic indicators alone, proposing a more nuanced prediction model. While acknowledging the study's limitations, such as reliance on a specific dataset and limited algorithm selection, our findings advocate for diversified algorithmic exploration to support insightful and practical policy development in the oil industry, aligning with the "No Free Lunch" theorem (Sterkenburg & Grünwald, 2021).

4.1 Recommendations

Building on the findings of this study, which utilised the Extra Trees Regressor for predicting Nigeria's oil production, future research should embrace a broader spectrum of socio-economic and environmental factors. It is recommended that analyses across global oil-producing countries, comparative model evaluations, and sensitivity tests be conducted under varying market conditions. Additionally, it is crucial to explore other nonlinear machine learning algorithms, such as neural networks and deep learning models. These advanced algorithms can capture complex patterns and interactions within the data, potentially enhancing prediction accuracy. Incorporating case studies from diverse oil economies and examining their unique socio-economic and geopolitical influences will further sharpen forecasting accuracy. This approach aims to deepen insights into global oil production drivers, supporting crafting more refined oil sector policies.

Disclosure of Potential Conflict of Interest. The authors confirm that no financial or personal relationships among them could appear to bias the findings presented in this manuscript

Authors Contributions
M.D. Adewale spearheaded the concept, methodology, and data collection and drafted the initial manuscript. Other authors contributed to validation, project management, and manuscript editing. All authors approve this submission to the journal.

Data Availability Statement. The datasets utilised in this research were obtained from multiple sources, including Nigerian Oil Reserves, Production, and Consumption Statistics from Worldometer (2024), Oil Production in Nigeria as reported by Statista (2023), Annual Brent Crude Oil Prices from Statista (2024), and Nigerian Political Rights data available at TheGlobalEconomy.com (2024). For additional inquiries, please contact the lead author.

Author Statement. The authors of this piece affirm their commitment to the guidelines and stipulations set forth by the Journal of Forecasting publication.

References

Adewale, M.D., et al.: Comparative performance evaluation of random forest, extreme gradient boosting and linear regression algorithms using Nigeria's gross domestic products. In: Seeam,

A., Ramsurrun, V., Juddoo, S., Phokeer, A. (eds.) Innovations and Interdisciplinary Solutions for Underserved Areas. InterSol 2023. LNICS, Social Informatics and Telecommunications Engineering, vol. 541, pp. 131–150. Springer, Cham (2024a). https://doi.org/10.1007/978-3-031-51849-2_9

Adewale, M., et al.: Predicting gross domestic product using the ensemble machine learning method. Syst. Soft Comput. **6**, 200132 (2024b). https://doi.org/10.1016/j.sasc.2024.200132

AlRassas, A.M., et al.: Advance artificial time series forecasting model for oil production using neuro fuzzy-based slime mould algorithm. J. Pet. Explor. Prod. Technol. **12**(2), 383–395 (2021). https://doi.org/10.1007/s13202-021-01405-w

Al-qaness, M.A.A., Ewees, A.A., Fan, H., AlRassas, A.M., Elaziz, M.A.: Modified aquila optimiser for forecasting oil production. Geo-spatial Inf. Sci. **25**(4), 519–535 (2022). https://doi.org/10.1080/10095020.2022.2068385

Asumadu-Sarkodie, S., Owusu, P.A.: Forecasting Nigeria's energy use by 2030, an econometric approach. Energy Sources Part B **11**, 990–997 (2016). https://doi.org/10.1080/15567249.2016.1217287

Brent crude oil price annually 1976–2024 | Statista. (2024, February 12). Statista. https://www.statista.com/statistics/262860/uk-brent-crude-oil-price-changes-since-1976/. Accessed 18 Feb 2024

Cavallo, A.: Hubbert's petroleum production model: an evaluation and implications for world oil production forecasts. Nat. Resour. Res. **13**, 211–221 (2004). https://doi.org/10.1007/S11053-004-0129-2

Chen, T., Guestrin, C.: XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016). https://doi.org/10.1145/2939672.2939785

Devi, M., Mathew, R., Suguna, R.: Regressor fitting of feature importance for customer segment prediction with ensembling schemes using machine learning. Int. J. Eng. Adv. Technol. (2019). https://doi.org/10.35940/ijeat.f8255.088619

Duhu, I.G.: Democratic stability in mitigating the impact of corruption on economic growth in Nigeria. Eur. Sci. J. ESJ **15**(31) (2019). https://doi.org/10.19044/esj.2019.v15n31p106

Falode, O., Udomboso, C.: Efficient Crude Oil Pricing Using a Machine Learning Approach. Day 2 Tue, 03 August 2021. https://doi.org/10.2118/207152-ms

Forthofer, R.N., Lee, E.S., Hernández, M.: Linear regression. Elsevier eBooks, pp. 349–386 (2007). https://doi.org/10.1016/b978-0-12-369492-8.50018-2

Gareth, J., Daniela, W., Trevor, H., Robert, T.: An introduction to statistical learning. Springer, New York (ISL) (2017)

Gbakon, K., Ajienka, J., Gogo, J., Iledare, O.: Oil Production Forecasting Models and Oil End-Use Optimization Framework under Global Energy Transition Dynamics. Day 2 Tue, 02 August 2022. https://doi.org/10.2118/211967-ms

Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomised trees. Mach. Learn. **63**, 3–42 (2006). https://doi.org/10.1007/s10994-006-6226-1

Huangi, L., Yang, D., Liu, B., Deng, J.: Decorrelated batch normalisation. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018). https://doi.org/10.1109/cvpr.2018.00089

Ji, Y., Zheng, W., Dong, H., Gao, P.: Train delays prediction based on feature selection and random forest. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), pp. 1–6 (2020). https://doi.org/10.1109/ITSC45102.2020.9294653

Karakurt, I.: Modelling and forecasting the oil consumptions of the BRICS-T countries. Energy **220**, 119720 (2021). https://doi.org/10.1016/j.energy.2020.119720

Li, C.: Preprocessing Methods and Pipelines of Data Mining: An Overview. arXiv (Cornell University) (2019). https://doi.org/10.48550/arxiv.1906.08510

Moroney, J.R., Berg, M.: An integrated model of oil production. Energy J. **20**, 105–124 (1999). https://doi.org/10.5547/ISSN0195-6574-EJ-VOL20-NO1-6

Mrabet, Z.E., Sugunaraj, N., Ranganathan, P., Abhyankar, S.: Random forest regressor-based approach for detecting fault location and duration in power systems. Sensors (2022). https://doi.org/10.3390/s22020458

Nigeria: Oil Production 2022 | Statista. (2023, August 25). Statista. https://www.statista.com/statistics/265195/oil-production-in-nigeria-in-barrels-per-day/. Accessed 18 Feb 2024

Nigeria oil reserves, production and consumption statistics - Worldometer. (2024). Worldometer.info. https://www.worldometers.info/oil/nigeria-oil/. Accessed 18 Feb 2024

Nigeria Political rights - data, chart | TheGlobalEconomy.com. (2024). TheGlobalEconomy.com. https://www.theglobaleconomy.com/Nigeria/political_rights/. Accessed 19 June 2024

Obite, C.P., Chukwu, A., Bartholomew, D.C., Nwosu, U.I., Esiaba, G.E.: Classical and machine learning modeling of crude oil production in Nigeria: identification of an eminent model for application. Energy Rep. **7**, 3497–3505 (2021). https://doi.org/10.1016/J.EGYR.2021.06.005

Sagi, O., Rokach, L.: Ensemble learning: a survey. WIREs Data Min. Knowl. Discov. **8**(4) (2018). https://doi.org/10.1002/widm.1249

Salisu, A.A., Cuñado, J., Isah, K.O., Gupta, R.: Oil Price and Exchange Rate Behaviour of the BRICS. Emerg. Mark. Financ. Trade **57**, 2042–2051 (2020). https://doi.org/10.1080/1540496X.2020.1850440

Shahhosseini, M., Hu, G., Archontoulis, S.: Forecasting corn yield with machine learning ensembles. Front. Plant Sci. **11** (2020). https://doi.org/10.3389/fpls.2020.01120

Shi, R., Xu, X., Li, J., Li, Y.: Prediction and analysis of train arrival delay based on XGBoost and Bayesian optimisation. Appl. Soft Comput. **109**, 107538 (2021). https://doi.org/10.1016/J.ASOC.2021.107538

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

Yao, W., Chen, X., Van, Tooren, M., Wei, Y.: Euclidean distance and second derivative based widths optimisation of radial basis function neural networks. In: International Joint Conference on Neural Networks (IJCNN) (2010). https://doi.org/10.1109/ijcnn.2010.5596528