Exploratory Analysis of Machine Learning Methods for Total Organic Carbon Prediction Using Well-Log Data of Kolmani Field

Fodio S. Longman, Habeeb Balogun*, Rasheed O. Ojulari, Olaniyi J. Olatomiwa, Husaini J. Balarabe
Ifeanyichukwu S. Edeh, & Olabisi O. Joshua
Department of Innovation Led Research (ILR)
Research Technology and Innovation Division, Nigerian National Petroleum Company (NNPC) Limited
Abuja, Nigeria
Big Data Technologies and Innovation Laboratory*
University of Hertfordshire, Hatfield, United Kingdom*
Email: fodio.longman@nnpcgroup.com, H.balogun@herts.ac.uk, olushola.ojulari@nnpcgroup.com, olaniyi.olatomiwa@nnpcgroup.com
husaini.balarabe@nnpcgroup.com, ifeanyichukwu.edeh@nnpcgroup.com, olabisi.joshua@nnpcgroup.com

***Abstract*—Machine Learning methods have shown significance in automating the prediction of Total organic carbon (TOC) for determining source rock potential in oil and gas exploration. Higher TOC contents could indicate greater potential for oil and gas generation. Making accurate TOC predictions, therefore, is crucial in measuring hydrocarbon deposits for a prospective geological formation. Automating this procedure can save time and resources when compared to conventional geochemical methods. In this study, we explore machine learning methods on a Frontier-Basin well-log data for TOC Prediction. Firstly, we employ feature selection methods to ensure optimal feature usage in regression and classification tasks. Additionally, we compare several supervised Machine Learning methods including Logistic Regression, Decision Tree, K-Nearest Neighbor, Gradient Boosting, and Naive Bayes Methods to categorize the quality of TOC using its defined standard ranges on the well-log data of Kolmani River 2 and 3, respectively. Furthermore, Random Forest method was utilised for the regression task on both data. A 5 fold nested cross validation was utilised for both classification and regression task. We show an exploratory analysis and the prospect of using Machine Learning methods to effectively classify TOC distribution using continuous well log data. Results show that Machine Learning methods are efficient for TOC prediction for non-geochemical approaches. The regression analysis shows an acceptable R^2 value of 0.62 and 0.76 for the respective well-logs with an MSE score of 0.05 and 0.07, respectively. For the classification task, evaluation metrics including F1 score, Precision, Accuracy, and Recall were investigated for the different ML classification methods, and Naive Bayes' was concluded to outperform others using standard metrics.**

***Index Terms*—Machine Learning, Frontier Basin, Total Organic Carbon (TOC), Data Analysis, Oil and Gas, Source Rock Determination**

I. Introduction

The demand for renewable energy has continued to make headlines globally; drawing significant attention to improved solutions and the development of alternative sources for energy generation, conservation, security, and sustainability [1]. Despite the huge success recorded in the expansion of new energy sources, fossil fuel remains an important energy source that guarantees the security and availability of energy resources [7]. However, discovering new fossil fuel reserves requires accurate identification of hydrocarbon source rock, its extensive examination, analysis, and huge capital and human investments [4]. Additionally, the oil and gas exploration endeavor is rigorous and risky consisting among other processes Basin survey, data acquisition, data processing, interpretation, prospect generation and exploratory drilling, production etc. [8]

Today, conventional methods for fossil fuel exploration have continued to experience difficulty in meeting the increasing demand for continuous energy production and the exploration of new energy sources [9]. Recently, a lot of research attention has been geared towards unconventional studies that are domineering of shale gas and oil and their corresponding developmental studies for improvement; and where possible automating the exploration processes and procedures to identify hydrocarbon-bearing source rocks [6]. Consequently, oil and gas-producing countries including Nigeria are venturing into this paradigm shift with efforts to uncover new fossil fuel reserves and to study the exploration and development of these hydrocarbon-bearing shales to expand the production of fossil fuels using a new approach.

Given the above, the Nigerian National Petroleum Company (NNPC) Limited in its bid to grow its hydrocarbon reserves and subsequently its production of oil and gas for sustainable energy security, and explore additional reserves, embarked on a renewed exploration of the Frontier Basins of Nigeria in the late 2000s and in particular the Kolmani oil fields.

The Frontier Basins are sedimentary basins that are gen-

979-8-3503-7565-7/24/$31.00 ©2024 IEEE

erally considered high-risk and under-explored. The Kolmani River is a conventional oil and gas field located inland of Nigeria. It lies in block Oil Prospecting License (OPL) 809 and 810, respectively [15].

Kolmani River is tipped to be a promising exploratory field within the Frontier Basin; uniquely associated with Nigeria's oil and gas exploration considering its projected hydrocarbon deposit prospect. The Exploration activities for hydrocarbons in the Kolmani field in the North-Eastern (NE) part of Nigeria have ignited lots of interest both in research and the oil and gas industry with research efforts pointing towards the determination of its potential for hydrocarbon generation [3]. An important path towards this effort is to develop an integrated approach for a responsive source rock assessment that could be automated to provide the needed efficiency for the task; by doubling research efforts in the exploration studies of the Frontier Basin.

Generally speaking, the exploration procedure for oil and gas can be divided into many steps. However, we would be limiting this to considering the four major procedures in no particular order; namely:

- Prospect Identification: Here geologists analyze geological and geophysical data to identify potential areas with hydrocarbon deposits and study the subsurface structures and rock formations.
- Seismic Surveys: Deploying sound waves to create images of the subsurface and map geological structures to identify potential reservoirs.
- Exploratory Drilling: Drill wells in promising locations to collect rock samples to assess the presence of hydrocarbons.
- Appraisal: conduct further drilling and testing once hydrocarbons are found and evaluate the size, quantity, quality, and productivity of the reservoir.

Ideally, geochemical exploration starts When rock samples are collected, then they undergo specific assessments to determine the presence of hydrocarbons. Two methods are utilized for this purpose namely; geochemical and Non-geochemical methods of source rock determination [10]. The geochemical method for source rock determination involves geochemically sampling the collected rocks and conducting certain chemical tests and procedures to determine source rock potential. This process is often time and labour-consuming. Additionally, it involves studying the chemical composition of the rocks for hydrocarbon exploration. Techniques in this method include Rock-Evaluation pyrolysis, Gas Chromatography-Mass Spectrometry (GC-MS), and elemental analysis [11]. On the other hand, non-geochemical analysis of source rocks involves techniques like petrography, well-log analysis, and seismic data interpretations, respectively [12].

Well-logging also known as wireline log consists of a complete set of logs used in the oil and gas industry to obtain the continuous record of a prospective oil formation's rock properties [13]. It consists of unique features including Gamma Ray (GT), Sonic (DT), MSFL (Resistivity), RHOB (Bulk Density), Neuron Porosity (NPHI), etc. These features or variables provide depth information of the well, the rock and formation type, porosity; which defines the measure of the pore space within a rock, permeability, saturation, and resistivity [14].

In evaluating source rock or shale to determine its hydrocarbon content, Total Organic Carbon (TOC) is a primary feature that determines the mass percentage of organic carbon in the rock and reflects indirectly to the organic matter content within the formation [16]. Accordingly, TOC content usually determines the hydrocarbon generation prospect of a formation. Consequently, finding efficient ways that could accurately predict and quantify TOC is desirable through a systematic and integrated process that would save time, cost, and human resources [17].

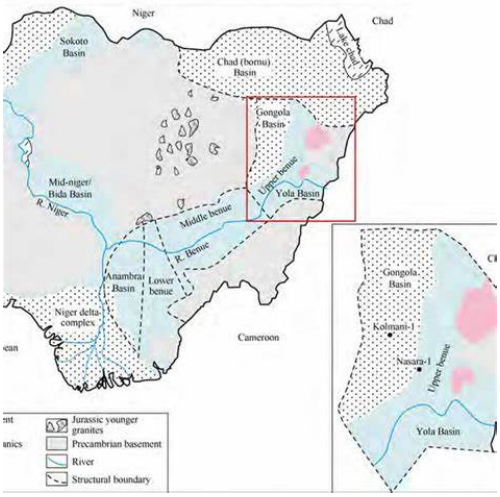

Fig. 1. Location map of Nigeria showing the position of Kolmani River Well-1 [15]

II. Proposed Framework

The proposed framework is shown in Fig 2. It comprises a data pre-processing step to take care of missing data points and perform Exploratory Data Analysis (EDA) to analyze and investigate the Kolmani dataset; summarizing the data characteristics by finding correlations between the data features and variables to determine how best to manipulate the data for best results. The exploratory data analysis includes Q-Q and heat map plots to see which features have a strong correlation.

The framework presented in Fig 2. allows a systematic approach to the proposed study enabling step-wise execution of the processes.

A. Data Description

The dataset utilized for this study is the well log data collected from Kolmani River exploration field located within the Gongola Basin of Upper Benue Trough, Nigeria as shown in Fig. 1. The two fields are named Kolmani River 2 (KR-2) and Kolmani River 3 (KR-3), respectively. The datasets consist of a geophysical well log suite (including Gamma Ray, Density, Neutron, Resistivity, and Caliper), Well Tops/

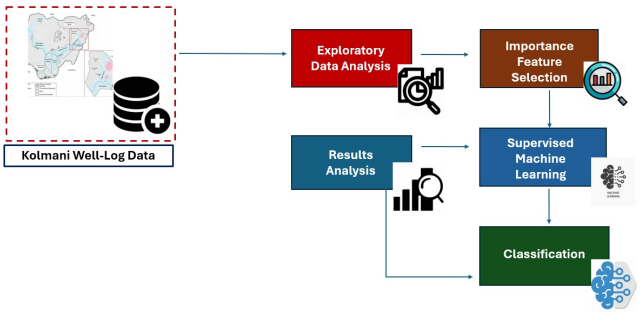

Fig. 2. Framework for Exploratory Analysis of Machine Learning Methods for TOC Prediction Using Well-Log Data

Well Markers, Mud Log, Biostratigraphic data, Geo-chemical (Rock-Eval) data, and the corresponding Geological report. For this study, we utilize the geophysical well log suite and the designed framework presented in Fig. 2 as a systematic approach to guide this work. The dataset consists of 33823 $\times$ 7 rows and columns entries of well-log information.

B. Exploratory Data Analysis

Data Analysis here is to enable us to explore the possibilities in the data and gain insight into the information contained in the two well-logs, respectively. We aim to explore relationships between the different features in the data, check for missing values, characterize the data, describe it, and prepare the data for further analysis. We employ methods including correlation matrix maps and Quantile-Quantile (Q-Q) plots of selected features to check for correlation and normality of the data which allows us to understand the intrinsic relationship between features and variables of the data.

C. Feature Selection

In machine learning predictive modeling, selecting important features is crucial for reducing data dimensionality [21] and improving model performance [22]. We explored embedded and wrapper feature selection methods due to their effectiveness. Embedded methods offer simplicity by using inherent feature importance, while wrapper methods like recursive feature elimination (RFE) provide flexibility by evaluating feature subsets based on model performance. Despite increased computational complexity, wrapper methods can significantly enhance model performance, making them worth considering. To ensure robustness in our predictive model, we decided to investigate both methods.

For the embedded feature selection method; we investigated four (4) different techniques namely; Random Forest, Gradient Boosting, Linear Regression, and Decision Tree methods on both data.

1) Random Forest: Random Forest (RF) feature importance selection is a popular method used for important feature selection for ML analysis. Its major purpose is to discard less significant variables to produce efficient and better performance on the class variables. This process can provide cost-effectiveness and reliability in understanding the data [23]. This method's advantage is that it is highly accurate due to its ensemble learning nature, its ability to generalize, and to be easily interpreted. It is simply defined as:

$$Y - hat = f - hat(X) \tag{1}$$

where X is a random variable with p predictors and Y is the responsive variable. A detailed description of the method for feature selection can be found in [23]–[25]

2) Gradient Boosting: This method is known for its speed and prediction accuracy. It can be used in classification and regression-based ML tasks [26]. Gradient Boosting also known as GBM [27] trains a bunch of models sequentially where each subsequent model learns from the mistakes of a previous model hence explaining and predicting previous errors with subsequent inputs iteratively. It involves two types of models namely; the weak ML model which is usually a decision tree [28] and a strong ML model involving numerous weak models. Consider equation 2 for illustration:

$$\boldsymbol{F}_{i+1} = \boldsymbol{F}_i - \boldsymbol{f}_i \tag{2}$$

where: $\boldsymbol{F}_i$ is the strong model at step i
$\boldsymbol{f}_i$ is the corresponding weak model at step i Additionally, GBM utilizes ensemble learning and comes from the family of Decision Tree models. A thorough explanation of this model is given in [29], [30], respectively.

3) Linear Regression: The goal here is for the feature selection model to predict the dependent variable continuously using the many input variables [31]. We determine important features here by employing the coefficients of the linear regression model defined by:

$$\boldsymbol{y} = \boldsymbol{\beta 0} + \boldsymbol{\beta 1} * \boldsymbol{x} \tag{3}$$

where $\boldsymbol{y}$ is the predicted output, $\boldsymbol{\beta 0}$ and $\boldsymbol{\beta 1}$ represent the Intercept and the coefficient of $\boldsymbol{x}$, respectively to describe how much $\boldsymbol{x}$ is influenced by $\boldsymbol{y}$.

Linear Regression achieves feature selection by identifying the coefficients in (3) for each input feature, and the magnitude of the coefficients calculated represents the relative importance of the features with larger absolute values indicating stronger influence.

4) Decision Tree: In this approach, scores are assigned to features based on their significance in predicting the target variable. The advantage of using this method is its ability to split the data into smaller chunks that are used to predict the target. However, it is likely to suffer from data over-fitting or over-classification when the input data is not significant. This method performs better with adequate data where the scores can be obtained from future importance attributes from

the trained tree model. Decision Tree feature importance can simply be given as:

$$\boldsymbol{K} = \sum(\flat)/\sum(\partial) \tag{4}$$

where $\boldsymbol{K}$ defines the Feature Importance for Feature K, $\flat$ is the Node Importance of Nodes splitting on $\boldsymbol{K}$, ∂ is the Node Importance of all Nodes, respectively.

Continuing with feature selection, we employed a wrapper method called Recursive Feature Elimination with Cross-Validation (RFECV) on both data. The best feature subset was chosen based on cross-validation scores. RFECV allowed us to identify distinct feature subsets, selected using 5-fold cross-validation with the 'accuracy' scoring parameter to ensure optimal accuracy across validation folds.

III. Machine Learning Methods for TOC Prediction

Machine learning algorithms have been widely adopted for TOC prediction including Back Propagation Neural Networks (BPNN), Support Vector Machines (SVM), Gaussian Process Regression (GPR), and Random Forest (RF) methods [12], [18]. Two major applications of machine learning are Regression and Classification. Regression is a supervised learning approach employed to predict the outcome of a continuous outcome. This is achieved when the relationship between two or more variables is established with the predicted outcome usually being numeric. On the other hand, the outcome of classification in Machine Learning is categorical. Classification in ML can be either in a supervised or un-supervised approach [32].

A. Regression Prediction of TOC Values

The prediction of TOC values in a supervised learning process is achieved in this work using the RF algorithm. RF is a classic approach derived from popular decision tree theory that integrates ensemble learning. The prediction takes in multiple decision trees that are modeled with no relationship between each decision tree thus each tree is independently modeled and the average output of each decision tree is the outcome. As stated earlier, the RF algorithm has many advantages including learning through randomness, and a high prediction accuracy and generalization ability [33].

B. Classification Methods

Numerous studies [35]–[37], [45], [46] have investigated the applications of Classification techniques of ML in TOC prediction and categorization including deep-learning methods [47]. In this paper, we take a step further to compare five common techniques including K-Nearest Neighbor (KNN), Naive Bayes, Logistic Regression, Gradient Boosting, and Decision Tree algorithms, and further evaluate them using standard classification evaluation metrics. In the preceding sections, we have defined most of these algorithms as related to ML and TOC prediction. Thus, we will briefly describe only the methods not discussed earlier.

1) K-Nearest Neighbors: Introduced in 1951 by the duo of Evelyn Fix and Joseph Hodges and later extended by Thomas Cover [38], the KNN algorithm for classification uses a supervised learning approach in a non-parametric way; which means it does not make assumptions on the underlying distribution of the data but rather intrinsically explore the pattern of the data using pre-defined training set to make classification usually with an identified feature or attribute. This algorithm is widely used for its versatility, simplicity, and ease of deployment. It can handle both numerical and categorical data.

The KNN algorithm works by first identifying the K neighbors which is a critical task in implementing the algorithm. The distance is determined by a distance metric which is usually the Euclidean distance [39] and finally the classification or categorization of the value of the data point is determined by the average of the K neighbors or the majority vote; hence allowing the algorithm to fit to different classes or patterns making accurate predictions based on the local structure of the data. In the prediction of TOC, [**?**] utilized KNN in predicting the seafloor total organic carbon which showed promising capabilities. Simply KNN can be defined as:

$$\boldsymbol{M}(x) = \underset{m}{\operatorname{argmax}} \sum_{i=1}^{k} I(\boldsymbol{M}(x)_i = m) \tag{5}$$

where $I(.)$ represents the indicator function which returns 1 if the characterization condition inside is true and 0 otherwise, m represents the possible class label to classify data points according to classes of their nearest neighbor.

2) Naive Bayes: The Naive Bayes classification algorithm is based on the popular Bayes' theorem [40]. This approach assumes independence between features of the data. It works by calculating the probability of a given data point belonging to every class of the distribution and then selecting the class with the highest probability as the prediction outcome. This algorithm is popular for its computational efficiency and simplicity when dealing with large datasets. Aside the work of [49] not much of this approach has been utilized in predicting TOC. The Naive Bayes' classification is often used in the classification of text data [41]. We explore this method for the robustness of comparing different classification models. We can simply define the method generated from the popular Bayes' theorem given by:

$$P(\theta|\mathbf{D}) = P(\theta)\frac{P(\mathbf{D}|\theta)}{P(\mathbf{D})} \tag{6}$$

where: $P(\theta)$ represents the prior probability, $P(\mathbf{D})$ is the marginal probability i.e. evidence, $P(\theta|\mathbf{D})$ is the posterior probability of $\mathbf{D}$ and $P(\mathbf{D}|\theta)$ is the likelihood probability that hypothesis will come true based on the evidence. This is a generalized Bayes' from which the classifier can be generated. A detailed explanation and derivation can be found in [42]

IV. Displaying the Results of the Proposed Framework

In this section, we show the results of exploring the various methods on the data described in section II subsection A using the framework presented in Fig.2. We begin with the pre-processing step to determine the correlation between the variables of the data and visually explore the normality of the data by selective features display of Q-Q plots.

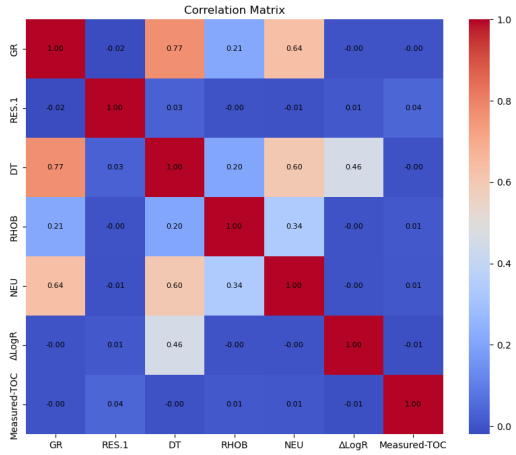

Fig. 3. Correlation Matrix display of the features in KR-2 data showing the correlation of the different features of the data

The heat map presented in Fig 3. shows the correlation matrix of features in KR-2 data. It is observed that there is a strong correlation between features DT and GR, GR and NEU. It is shown [18] that these features show great correlation with the TOC and are the most used in TOC prediction.

In the KR-3 dataset, more features are utilized to explore robustness for the intended ML analysis. Here also, we see an accepted correlation between measured TOC and Passey's ΔLogR method. Additionally, there is a correlation between Measured TOC and Neutron porosity and Neutron porosity with ΔLogR. This coincides with the submission made in [19].

We explore the Q-Q plot to check the normality of the distribution of features or variables in the data. These plots aim to determine if the datasets are from populations with a normal distribution. In Fig. 5-7, It is observed that the data is not normally distributed since the points are not approximately in line with the linearly fit regression model.

We explore different important feature selection methods described in II. In KR-2 data, it is observed that the ΔLogR approach proposed in [19] shows significance when compared to other variables and NEU and Resistivity trailing corresponding to the findings of [18] and the widely used well-log variables utilized for TOC prediction.

V. Model Evaluation

To assess machine learning methods for regression and classification tasks, common evaluation metrics are employed. In regression, R^2 and Mean Squared Error (MSE) measure accuracy and model fit. In classification, metrics such as

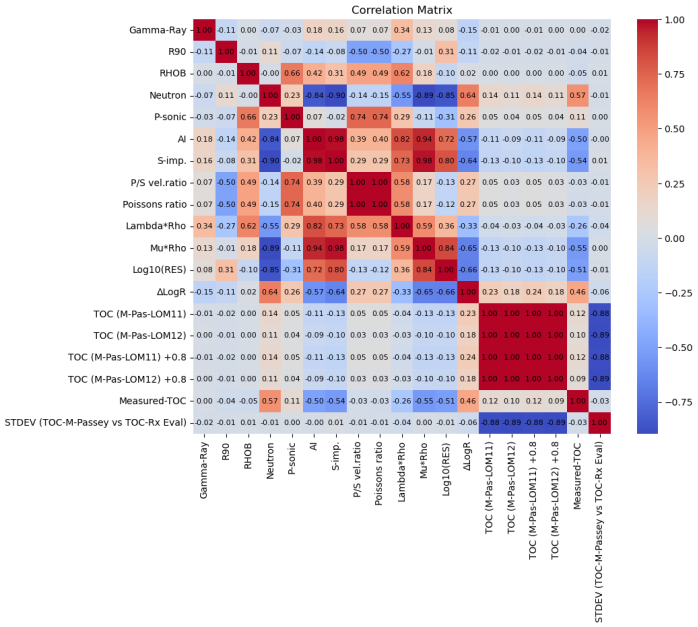

Fig. 4. Correlation Matrix display of the features in KR-3 data showing the correlation of the different features of the data

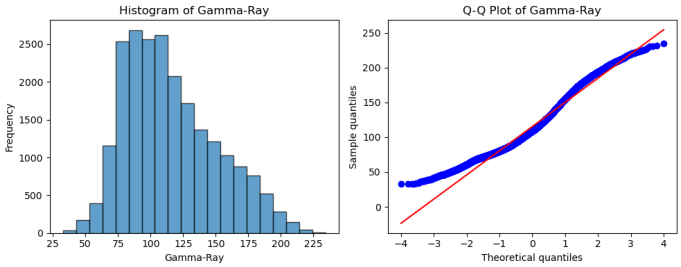

Fig. 5. QQ-Plot and Bar Plot of Gamma Ray showing the Normality of the data distribution in KR-2

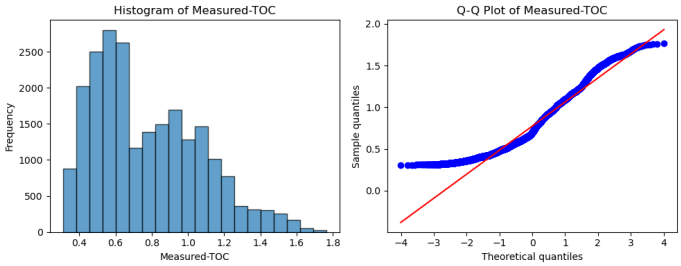

Fig. 6. QQ-Plot and Bar Plot of Measured TOC showing the Normality of the data distribution in KR-2

F1 score, Precision, Recall, Accuracy, and Area Under the Receiver Operator Characteristic (ROC-AUC) curve evaluate classification performance. These metrics provide insights into the model's effectiveness in regression and classification tasks, and they can be computed using the following approaches

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

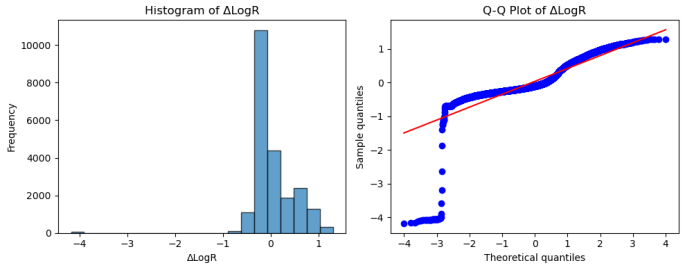

Fig. 7. QQ-Plot and Bar Plot of ΔLogR Method showing the Normality of the data distribution in KR-2

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{8}$$

$$F1 = 2 \times (\frac{Precision \times Recall}{Precision + Recall}) \tag{9}$$

$$Precision = (\frac{TP}{TP + FP}) \tag{10}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{11}$$

$$Recall = (\frac{TP}{TP + FN}) \tag{12}$$

where TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

A good regression model is marked by high R^2 and low Mean Squared Error (MSE). For classification, a good model exhibits high accuracy, Precision, Recall, F1 score and ROC-AUC. However, the least acceptable score for classification is arguably 0.5 [43], [44].

VI. Discussion & Conclusion

This study developed a predictive machine learning model for Total organic carbon (TOC) using well log data of kolmani field (i.e., Kolmani River 2 and Kolmani River 3). The study explored the prediction as a regression and classification problem. In both cases, two common feature selection methods was investigated to check the feature relevancy of the predictors of the TOC. Also, the different feature selection methods were tried to complement each others weakness, as well as to understand and develop a robust predictive model for TOC prediction. In both cases and for both data, we decided to use features commonly selected as the most relevant across both feature selection methods. These features where then used for developing the classification and the regression model. To further emphasize on the robustness, and generalization of the developed model, we implemented the regression and the classification model using 5-fold nested cross-validation. The regression analysis shows a R^2 value of 0.62 and 0.76 for the respective well-logs with an MSE score of 0.05 and 0.07, respectively. Meanwhile, valuation metrics including F1 score, Precision, Accuracy, and Recall were investigated for the different ML classification methods, and Naive Baye's

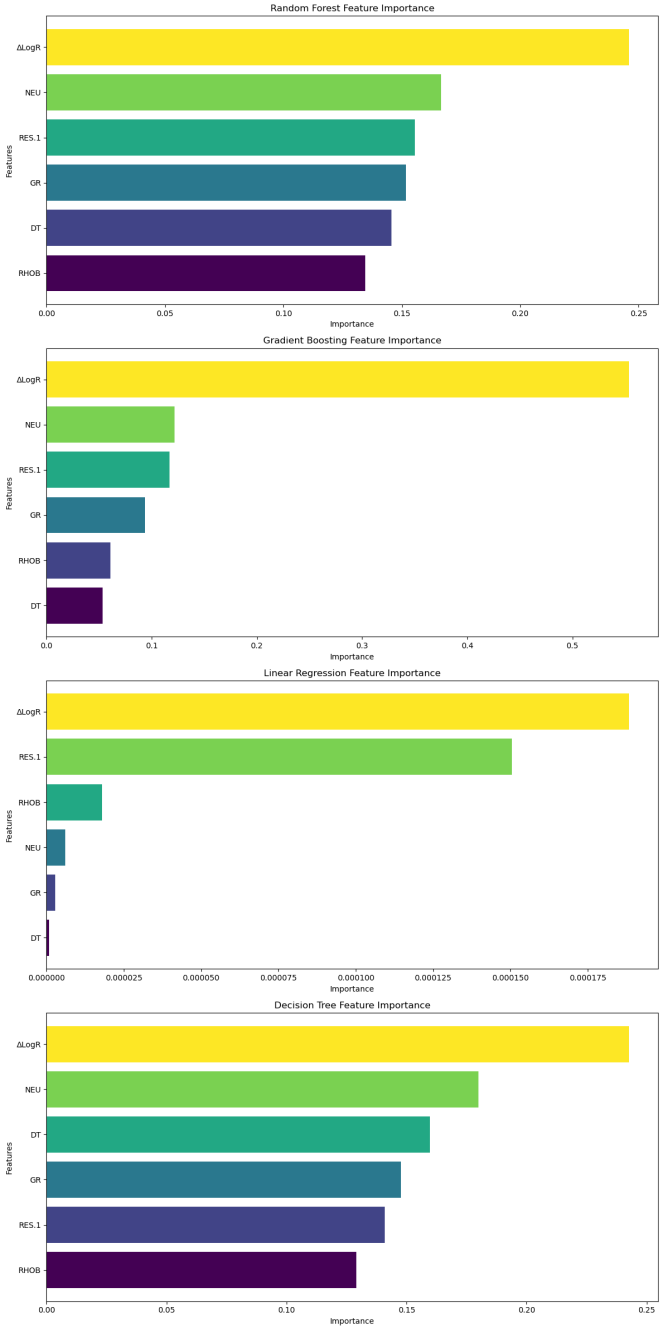

Fig. 8. Determination of Feature Importance in KR-2 using (a) RF (b) Gradient Boosting (c) Linear Regression and (d) Decision Tree

was concluded to outperform others using the defined metrics. Overall, it is observed that using machine learning methods could be efficient in predicting TOC values. In conclusion, it is safe to say that performance can vary between different algorithms to achieve TOC content prediction as illustrated in this study. Additionally, we observe the uniqueness of each algorithm which is the reason for achieving different results even on the same dataset. Lastly, the amount of data and utilized features and variables play a crucial role in the

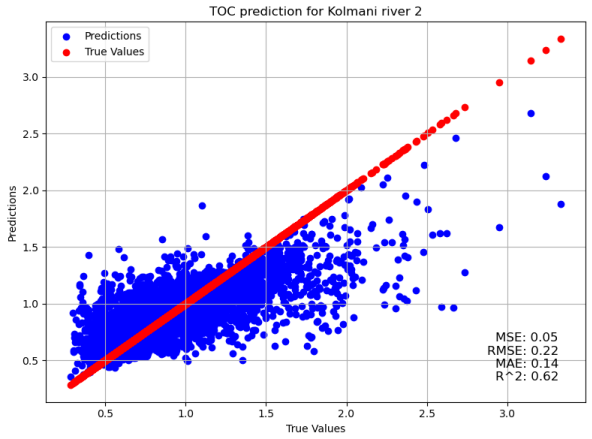

Fig. 9. Regression Analysis on KR-2 to Predict Measured TOC Values

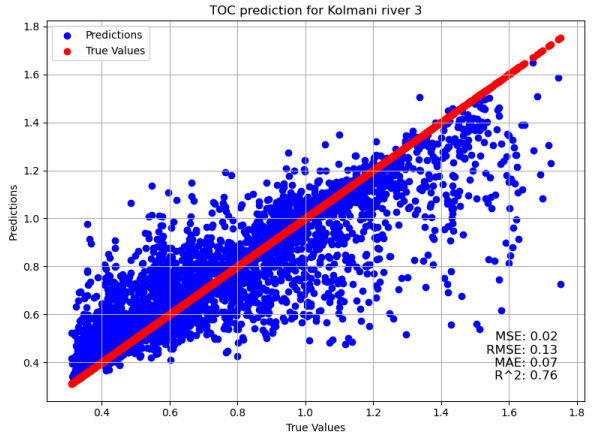

Fig. 10. Regression Analysis on KR-3 to Predict Measured TOC Values

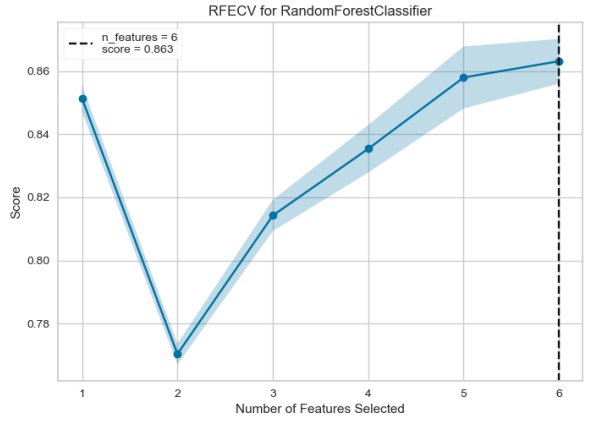

Fig. 11. Recursive Feature Elimination on KR-2 for Classification

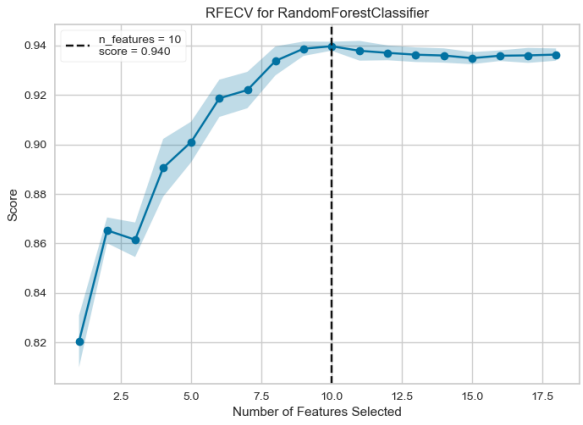

Fig. 12. Recursive Feature Elimination on KR-3 for Classification

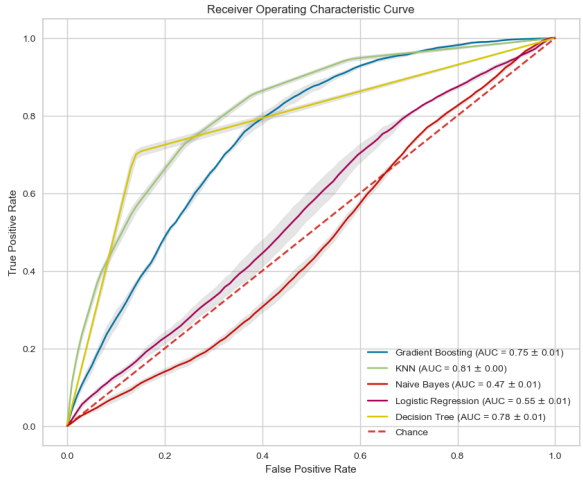

Fig. 13. Receiver Operating Characteristic curve for KR-2

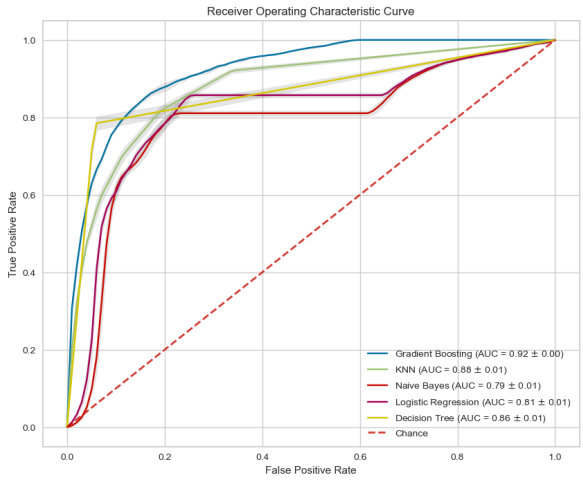

Fig. 14. Receiver Operating Characteristic curve for KR-3

learning process which can further affect performance and outcome.

ACKNOWLEDGMENT

The authors would like to acknowledge the support of the Research Technology and Innovation (RTI) Division of

the Nigerian National Petroleum Company (NNPC) Limited for sponsoring this work under the RTI Research Fellowship Program.

REFERENCES

[1] Ma, X., Wang, H., Zhou, S., Shi, Z. and Zhang, L., 2021. Deep shale gas in China: Geological characteristics and development strategies. Energy Reports, 7, pp.1903-1914.

[2] Didi, C.N., Osinowo, O.O., Akpunonu, O.E. and Nwali, O.I., 2024. Petroleum system and hydrocarbon potential of the Kolmani Basin, Northeast Nigeria. Journal of Sedimentary Environments, pp.1-27.

[3] Energy, B., 2015. Renewable energy sources. Ergon Energy.

[4] Peters, K.E., Curry, D.J. and Kacewicz, M., 2012. An overview of basin and petroleum system modeling: Definitions and concepts.

[5] Waples, D.W., 2013. Geochemistry in petroleum exploration. Springer Science & Business Media.

[6] Holechek, J.L., Geli, H.M., Sawalhah, M.N. and Valdez, R., 2022. A global assessment: can renewable energy replace fossil fuels by 2050?. Sustainability, 14(8), p.4792.interface," IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August 1987 [Digests 9th Annual Conf. Magnetics Japan, p. 301, 1982].

[7] Goodarzi, F., Gentzis, T., Sanei, H. and Pedersen, P.K., 2019. Elemental composition and organic petrology of a Lower Carboniferous-age freshwater oil shale in Nova Scotia, Canada. ACS omega, 4(24), pp.20773-20786.

[8] Yongsheng, M.A., Xunyu, C.A.I. and Peirong, Z.H.A.O., 2018. China's shale gas exploration and development: Understanding and practice. Petroleum Exploration and Development, 45(4), pp.589-603.

[9] Campbell, E.T., 2015. Emergy analysis of emerging methods of fossil fuel production. Ecological modeling, 315, pp.57-68.

[10] Shalaby, M.R., Jumat, N., Lai, D. and Malik, O., 2019. Integrated TOC prediction and source rock characterization using machine learning, well logs and geochemical analysis: case study from the Jurassic source rocks in Shams Field, NW Desert, Egypt. Journal of Petroleum Science and Engineering, 176, pp.369-380.

[11] Ahmed, M.A. and Hassan, M.M., 2019. Hydrocarbon generating-potential and maturity-related changes of the Khatatba Formation, Western Desert, Egypt. Petroleum Research, 4(2), pp.148-163.

[12] Nyakilla, E.E., Silingi, S.N., Shen, C., Jun, G., Mulashani, A.K. and Chibura, P.E., 2022. Evaluation of source rock potentiality and prediction of total organic carbon using well log data and integrated methods of multivariate analysis, machine learning, and geochemical analysis. Natural Resources Research, 31(1), pp.619-641.

[13] Liu, H., 2017. Principles and applications of well logging (pp. 237-269). Berlin, Heidelberg: Springer Berlin Heidelberg.

[14] Shi, N., Li, H.Q. and Luo, W.P., 2015. Data mining and well logging interpretation: Application to a conglomerate reservoir. Applied Geophysics, 12, pp.263-272.

[15] Ajiya, M., 2023. The Implication of Oil Discovery in Northern Nigeria and the Emerging Threats to Grapple With. Available at SSRN 4323459.

[16] Mabitje, M.S. and Opuwari, M., 2023. Determination of total organic carbon content using Passey's method in coals of the central Kalahari Karoo Basin, Botswana. Petroleum Research, 8(2), pp.192-204.

[17] Zhao, P., Ma, H., Rasouli, V., Liu, W., Cai, J. and Huang, Z., 2017. An improved model for estimating the TOC in shale formations. Marine and Petroleum Geology, 83, pp.174-183.

[18] Zhu, L., Zhou, X., Liu, W. and Kong, Z., 2023. Total organic carbon content logging prediction based on machine learning: A brief review Energy Geoscience, 4(2), p.100098.

[19] Passey, Q.R., Bohacs, K.M., Esch, W.L., Klimentidis, R. and Sinha, S., 2010, June. From oil-prone source rock to gas-producing shale reservoir–geologic and petrophysical characterization of unconventional shale-gas reservoirs. In SPE International Oil and Gas Conference and Exhibition in China (pp. SPE-131350). SPE

[20] HU, H., LIU, C. and LU, S., 2015, September. The Method and Application of Using Generalized-Passey Method Technology to Predict the Organic Carbon Content of Continental Deep Source Rocks. In Acta Geologica Sinica-English Edition (Vol. 89, No. s1, pp. 393-394)

[21] Dhal, P. and Azad, C., 2022. A comprehensive survey on feature selection in the various fields of machine learning. Applied Intelligence, 52(4), pp.4543-4581

[22] Balogun, H., Alaka, H. and Egwim, C.N., 2021. Boruta-grid-search least square support vector machine for NO2 pollution prediction using big data analytics and IoT emission sensors. Applied Computing and Informatics.

[23] Jaiswal, J.K. and Samikannu, R., 2017, February. Application of random forest algorithm on feature subset selection and classification and regression. In 2017 World Congress on computing and Communication Technologies (WCCCT) (pp. 65-68). Ieee

[24] Kumar, S.S. and Shaikh, T., 2017, September. Empirical evaluation of the performance of feature selection approaches on random forest. In 2017 International Conference on computer and applications (ICCA) (pp. 227-231). IEEE.

[25] Kursa, M.B. and Rudnicki, W.R., 2011. The all relevant feature selection using random forest. arXiv preprint arXiv:1106.5112.

[26] Xu, Z., Huang, G., Weinberger, K.Q. and Zheng, A.X., 2014, August. Gradient boosted feature selection. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 522-531).

[27] Adler, A.I. and Painsky, A., 2022. Feature importance in gradient boosting trees with cross-validation feature selection. Entropy, 24(5), p.687.

[28] Song, Y.Y. and Ying, L.U., 2015. Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, 27(2), p.130.

[29] Natekin, A. and Knoll, A., 2013. Gradient boosting machines, a tutorial. Frontiers in neurorobotics, 7, p.21.

[30] Upadhyay, D., Manero, J., Zaman, M. and Sampalli, S., 2020. Gradient boosting feature selection with machine learning classifiers for intrusion detection on power grids. IEEE Transactions on Network and Service Management, 18(1), pp.1104-1116.

[31] Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), pp.1157-1182.

[32] Matloff, N., 2017. Statistical regression and classification: from linear models to machine learning. Chapman and Hall/CRC.

[33] Coulston, J.W., Blinn, C.E., Thomas, V.A. and Wynne, R.H., 2016. Approximating prediction uncertainty for random forest regression models. Photogrammetric Engineering & Remote Sensing, 82(3), pp.189-197

[34] Singh, A., Thakur, N. and Sharma, A., 2016, March. A review of supervised machine learning algorithms. In 2016 3rd International Conference on computing for Sustainable Global Development (INDIACom) (pp. 1310-1315). Ieee.

[35] Wibowo, R.C., Dewanto, O. and Sarkowi, M., 2022, October. Total organic carbon (TOC) prediction using machine learning methods based on well logs data. In AIP Conference Proceedings (Vol. 2563, No. 1). AIP Publishing.

[36] Handhal, A.M., Al-Abadi, A.M., Chafeet, H.E. and Ismail, M.J., 2020. Prediction of total organic carbon at Rumaila oil field, Southern Iraq using conventional well logs and machine learning algorithms. Marine and Petroleum Geology, 116, p.104347.

[37] Mandal, P.P., Rezaee, R. and Emelyanova, I., 2021. Ensemble learning for predicting TOC from well-logs of the unconventional goldwyer shale. Energies, 15(1), p.216.

[38] Lopez-Bernal, D., Balderas, D., Ponce, P. and Molina, A., 2021. Education 4.0: teaching the basics of KNN, LDA and simple perceptron algorithms for binary classification problems. Future Internet, 13(8), p.193.

[39] Guo, G., Wang, H., Bell, D., Bi, Y. and Greer, K., 2003. KNN model-based approach in classification. In On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings (pp. 986-996). Springer Berlin Heidelberg.

[40] Swinburne, R., 2004. Bayes' theorem. Revue Philosophique de la France Et de l, 194(2).

[41] Xu, S., 2018. Bayesian Naïve Bayes classifiers to text classification. Journal of Information Science, 44(1), pp.48-59.

[42] Chen, S., Webb, G.I., Liu, L. and Ma, X., 2020. A novel selective naïve Bayes algorithm. Knowledge-Based Systems, 192, p.105361.

[43] Olu-Ajayi, R., Alaka, H., Sulaimon, I., Balogun, H., Wusu, G., Yusuf, W. and Adegoke, M., 2023. Building energy performance prediction: A reliability analysis and evaluation of feature selection methods. Expert Systems with Applications, 225, pp.120109.

[44] Egwim, C.N., Alaka, H., Toriola-Coker, L.O., Balogun, H. and Sunmola, F., 2021. Applied artificial intelligence for predicting construction project delays. Machine Learning with Applications, 6, p.100166.
[45] Tan, M., Song, X., Yang, X. and Wu, Q., 2015. Support-vector-regression machine technology for total organic carbon content prediction from wireline logs in organic shale: A comparative study. Journal of natural gas science and engineering, 26, pp.792-802.
[46] Rui, J., Zhang, H., Zhang, D., Han, F. and Guo, Q., 2019. Total organic carbon content prediction based on support-vector-regression machine with particle swarm optimization. Journal of Petroleum Science and Engineering, 180, pp.699-706.
[47] Zhu, L., Zhang, C., Zhang, C., Zhang, Z., Nie, X., Zhou, X., Liu, W. and Wang, X., 2019. Forming a new small sample deep learning model to predict total organic carbon content by combining unsupervised learning with semisupervised learning. Applied Soft Computing, 83, p.105596.
[48] Lee, T.R., Wood, W.T. and Phrampus, B.J., 2019. A machine learning (kNN) approach to predicting global seafloor total organic carbon. Global Biogeochemical Cycles, 33(1), pp.37-46.
[49] Ganguli, S.S., Kadri, M.M., Debnath, A. and Sen, S., 2022. A Bayesian multivariate model using Hamiltonian Monte Carlo inference to estimate total organic carbon content in shale. Geophysics, 87(5), pp.M163-M177.