Available online at www.qu.edu.iq/journalcm

JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS

ISSN:2521-3504(online) ISSN:2074-0204(print)

Oil and Gas Production Forecasting Using Decision Trees, Random Forst, and XGBoost

Mays A. Al shabaan [a], *Zainab N. Nemer* [b*]

a College of Computer Science & Information Technology, University of Basrah, Iraq. Email: itpg.mays.ali@uobasrah.edu.iq

b College of Computer Science & Information Technology, University of Basrah, Iraq. Email: zainab.nemer@uobasrah.edu.iq

ARTICLE INFO

Article history:
Received: 10 /1/2024
Rrevised form: 4 /2/2024
Accepted : 18 /2/2024
Available online: 30 /3/2024

Keywords:

Machine Learning (ML)
Oil and Gas production
Random forest (RFR)
Decision tree(DTR)
XGBoost

ABSTRACT

Oil and gas production forecasting has always been a hot topic in the petroleum industry. Production forecasting in this sector aims to estimate future production rates, facilitating operational planning, production optimization, and resource allocation for companies. Scientists have traditionally attempted to forecast oil and gas production using methods such as Numerical Reservoir Simulation (NRS) and Decline Curve Analysis (DCA). However, these methods present challenges including time-consuming processes lasting hours or even days, uncertain accuracy, reliance on accurate static models, and uncertainty in dynamic model parameters. In this research, aim to address these limitations by leveraging machine learning models for production forecasting. These models enable faster and more precise decision-making by accurately predicting future outcomes based on historical data. Our study employs three models: Decision Trees (DTR), Random Forest (RFR), and XGBoost. In this reserch utilize the Python programming language and a dataset from wells in New York State, USA. Experimental results demonstrate that the RFR model achieves the highest accuracy (99%) for oil and gas production compared to the XGBoost and DTR models.

MSC.

https://doi.org/10.29304/jqcsm.2024.16.11431

1. Introduction

Oil is a significant component in numerous industrial processes and the principal fuel source. It is critical to operating countless vehicles, aircraft, trucks, ships, and machinery [1]. Oil production refers to extracting Oil from wells and transforming these raw materials into final petroleum products that customers can purchase and utilize. Oil

[*] Corresponding author

Email addresses: *itpg.mays.ali@uobasrah.edu.iq*

Communicated by 'sub etitor'

production involves a series of systematic stages, beginning with exploring the site, continuing extraction, and distributing goods to enterprises and the general public [2]. During the process of creating a field development plan, the prediction is crucial because it provides data on production that may be applied to facility capacity design, makes a drilling timetable or sequence, and evaluates economics. Considerable demand exists for a production estimate based on past data from active and nonactive wells [3]. Production forecasting is a crucial process for governmental and organizational entities, enabling them to formulate requisite economic strategies [4]. Forecasting output in the oil and gas sector requires a complicated numerical simulation of reservoirs and their engineering study [5]. The precise prediction is a substantial and different undertaking aimed at monitoring and enhancing Oil reservoirs. When it comes to estimating Oil reservoir production, the petroleum industry makes use of a variety of traditional techniques, Such as NRS and DCA [6].

The NRS approach has conventionally been employed for Oil production forecasts. However, NRS models possess limitations, including their labor-intensive and time-consuming nature, as well as their requirement for an accurate static model and several dynamic model parameters [7]. The DCA is a traditional method to predict future Oil and Gas output [8]. This method takes a long time because much computing power is needed. Propose an alternative way by making this step much more accessible and using less computing power. ML is the solution to forecast Oil and Gas production with high precision and speed; it can use the dataset's features to estimate how much Oil and Gas will be produced [9]. According to the World Energy Report, an estimated 30% of the world's energy came from fossil fuels and natural Gas in 2020. The topic's significance has drawn numerous writers who have studied Oil producing processes using Machine Learning[10]. ML techniques have recently garnered interest in the Oil and Gas business, particularly in the areas of rapid evaluation and production forecasting. During the past several years, researchers have utilized ML [11].

AI is the subfield of computer science that combines computers' processing power using human Intellect to provide intelligent and dependable solutions to extremely nonlinear and highly complex issues. AI allows computers to think and decide for themselves[12]. Machine learning aims to find solutions to real-world problems using large datasets and statistical models built using algorithms [13]. Machines could improve and learn their abilities through repetition and error, much as people do. Prediction, grouping, extraction, and decision-making are ML's primary outputs from a particular set of data [14]. The versatility of machine learning extends to classification and regression [15]. Machine learning methodology has emerged as a promising approach for modeling production methods, owing to the availability of high-performance computing resources and comprehensive datasets[16]. There are many potential benefits to using machine learning techniques for Oil and Gas production forecasting. These techniques can help to improve accuracy, reduce uncertainty, and provide insights into the factors driving production. As the energy industry evolves and quality improves, machine learning plays an increasingly important role in production forecasting[17]. Machine learning can be carried out in several different ways. It employs a supervised learning technique and an unsupervised one, where each sample in the dataset has been labeled to produce results that can be used to infer a label for newly obtained data[13].

2. Related works

The Oil and Gas production industry has started utilizing a range of algorithms that leverage machine learning. This article will review several recent studies that use ML in the Oil and Gas industry.

M. A. M. Fadzil *et al*. (2021)[18] In this research, an alternative strategy is proposed for reducing Oil production losses by employing ML supervised regression models “XGBoost, RFR, DTR, and SVR”, which are constructed and examined to make predictions regarding the plant's operating circumstances. The highest effectiveness and consistency throughout the validation process have been demonstrated by the XGBoost model. Limitations of the research include less frequent lab data and fewer samples for analysis. Presumably, consistent feed data qualities will be maintained, and continual communication within the project upheld. Due to the possibility of lost historical data due to intermittent transmitter failures, operating condition values are determined by linear interpolation.

G. Hui *et al.* (2021) [19] In this study, a comprehensive approach based on machine learning was developed to evaluate shale gas production in the Fox Creek area of Alberta. Four methods were utilized: “linear regression, neural networks, XGBoost, and DTR.” The Extra Trees strategy emerged as the top performer, with a coefficient of determination of 0.809, the highest among the methods evaluated. However, according to a case study the limitations of the research, it is difficult to definitively state that these are the best results achieved after running experiments with the dataset and examining the outcomes for every approach used. Trying new approaches or combining existing ones certainly improves upon the current outcomes. Nevertheless, given the complexity of the problem, continued exploration of new methods is deemed crucial.

C. Tan *et al* (2021) [20] In this research, various techniques were employed by the authors, including six machine learning algorithms: “random forest (RF), back propagation (BP) neural network, support vector regression (SVR), extreme gradient boosting (XGBoost), light gradient boosting machine (Light GBM), and multivariable linear regression (MLR).” Evaluation results indicate that the production prediction model of the XGBoost algorithm was found to be the most effective, with an R^2 value of 0.90. The authors utilized data from the 'WY shale gas block in Sichuan, China.' The limitations of this study pertain to its applicability, as it is based solely on a single shale gas block in Sichuan, China, and its objective is to optimize fracturing productivity in shale gas wells. Additionally, the limitations of the dataset, which includes only 137 wells, could affect the robustness of the model.

N. M. Ibrahim *et al.* (2022)[9] In this study, the process of estimating Oil and Gas production was attempted to be expedited by the authors. Eight experiments involving machine learning and deep learning techniques were conducted: DTR, PLR, SVR, MLR, RFR, RNN, XGBoost, and ANN. The dataset was supplied by Saudi Aramco, and it was found that the best results were yielded by RNN, XGBoost, and ANN, with R^2 values of 0.926, 0.9012, and 0.9627 for Oil, Gas, and water, respectively. However, the potential ethical implications of employing machine learning and deep learning models in the Oil and Gas industry, such as their environmental impact or potential displacement of workers, were not addressed in the study.

X.-y. Wang *el al.*(2023) [21] In this study used a daily Oil production data from 62 Oil wells over ten years. The authors propsed a method involving two models Multiple polynomial regression technique, and random forest , were chosen based on their minimal inaccuracy in predicting outcomes compared to other models. The prediction results using random forests showed a lower error margin than Multiple polynomial regression. The limitations include the study's narrow data scope, potential data quality issues, and potential oversights of external factors for data.

3. Machine learning models (ML)

Machine learning is a subfield of AI that has become an integral part of popular digitalization solutions [22]. “Supervised learning, unsupervised learning, and reinforcement learning” are three main classes of machine learning issues that can be usefully described to gain a general notion of the problems that machine learning can handle [23]. In this research used three different types of supervised machine learning included:

3.3.1 Decision Tree Regressor (DTR)

A strong ML model for classification and regression tasks was named DTR. Decision trees show decisions and their outcomes as a tree. The edges represent decision rules, while the nodes represent events or choices. Trees have nodes and branches. Each node represents qualities in a group to be categorized, and each branch represents a value for the node [24]. First, the training data is used to build a tree. Then, using a binary split method, the original data is split into two parts. The separation process is done on the new growth branches, and it keeps going until a new branch cannot be separated and the node next to it reaches the minimum size and changes into an end node [25]. DTR uses MSE to subdivide nodes. The technique chooses the value and splits the data with a binary tree. Any subgroup can be measured individually for MSE. The tree picks a low MSE value. The MSE for nodes n with N observations and y as the expected value follows this relation in the equations(1)

$$E(n) = \frac{1}{N}\sum_{i=1}^{N}[y_i(n) - \bar{y}\,(n)]^2 \qquad (1)$$

The ideal split is found by maximizing the MSE difference between "the root node E(n) and the left E(nl) and right child E(nr)" nodes. The mean of all node data is used for prediction [26]. Fig. 1 shows the structure of DTR. The first node in the tree serves as a root node, defining the total sample and allowing for further subdivision into other nodes.

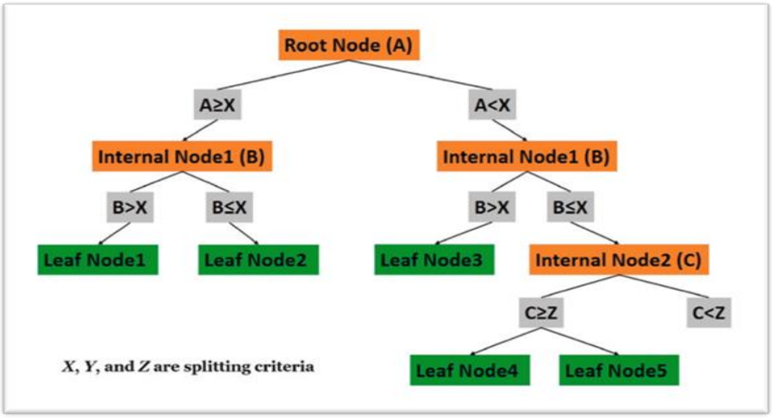

Fig. 1. Illustrate the structure of the Decision Tree Regressor[26].

3.3.2 Random Forest Regressor (RFR)

The Random Forest algorithm has demonstrated efficacy in addressing classification and regression problems. It starts with input data, trains many models, gathers predictions from each model, and finally employs a voting process to choose the best solution[27]. This method depends on the decision tree. It takes the average of results from multiple DTR to arrive at the final forecast. The prediction outcome is determined by computing the mean of the outputs generated by each tree [28]. It was first introduced in 2001 by University of California, Berkeley professor Leo Breiman. Random Decision Forests is another name for this method [29]. The model is constructed by splitting the input into several samples based on how many trees there are, then constructing an easy forecasting model inside each section, and finally merging the results of these models using a bagging method to arrive at the final forecast [30]. There is no requirement to reduce processing speed in the Random Forest because each decision tree has fully matured. With more trees, it can avoid overfitting the data and get more accurate results[31]. Fig.2 shows Random forest architecture.

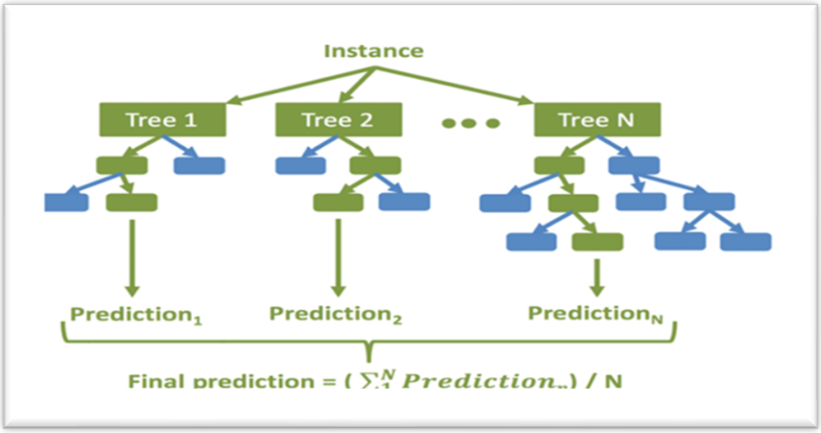

Fig.2. Random forest architecture[28].

Important parameters for the RFR model include the expected number of trees (N), the number of predictor variables to be tested at each splitting node, the node size, or the bare minimum of samples present at a leaf node [32]. A random forest of simple trees estimates a dependent variable in a regression issue. The method generates K separate regression trees $h_K(x)$ based on the input variable x. The model forecasts the average of the predictions made by each tree in the forest for the inputs (x), (k = 1,.., K). It can improve the variety of the trees so that their agregated findings are less likely to be correlated with those of other trees by using a method called bootstrapping [33],[34], displayed in Equation (2) :

$$\textbf{RFR prediction } = \frac{1}{K}\sum_{K=1}^{K} h_K(x) \qquad (2)$$

3.3.3 Extreme Gradient Boosting (XGBoost)

XGBoost is an ML methodology for regression and classification tasks, producing a predictive model represented as a decision tree [35]. XGBoost, developed by" Chen and Guestrin in 2016", is a widely used technique for solving ML problems among the different implementations of Tree Gradient Boosting. [36]. XGBoost is a boosting technique that falls under the category of supervised learning. It is an ensemble approach based on gradient-boosted trees [37]. The purpose of the boosting method is to train subsequent learners using updated versions of the training specimen that have been adjusted based on the training effect of the preceding learner. In this way, we may drastically reduce how much the model's forecast deviates from the actual value, and the model's final forecast is just the weighted vote of all the analysts. Negative gradients quantify errors from the previous iteration, and subsequent corrections are made by gradient descent in the Gradient Boosting Model [38]. The XGBoost is illustrated in Fig. 3 and consists of branches, internal nodes, several root nodes, and leaf nodes.

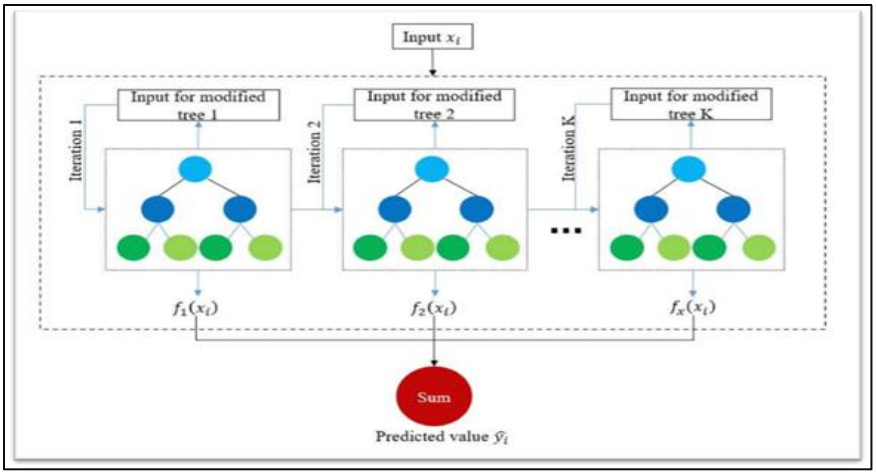
Input x_i
Input for modified tree 1
Input for modified tree 2
Input for modified tree K
Iteration 1
Iteration 2
Iteration K
$f_1(x_i)$
$f_2(x_i)$
$f_x(x_i)$
Sum
Predicted value $\hat{y}_i$

Fig.3: Shows the topology of XGBoost[39].

The i-th parameter, x_i, is inputted into Classification and Regression Trees (CARTs) for initial decision-making. Internal nodes make subsequent decisions, while branch points indicate future decisions. Leaf nodes contain single CART model predictions aggregated for XGBoost model prediction. The mathematical expression y_i of the XGBoost represents the actual value of the experiment. Equation (3) demonstrates that the estimated score $\hat{y}_i$ is obtained by summing all f_k values.

$$\hat{\mathbf{y}}_{\mathbf{i}} = \propto \sum_{\mathbf{k=1}}^{\mathbf{k}} \mathbf{f_k(x_i)} \qquad (3)$$

In this context, $\hat{y}_i$ represents the predicted value associated with the input x_i. The symbol ∝ denotes the learning rate of the individual regression tree, while K represents the total number of Regression Trees. Additionally, f_k signifies the output of the k-th regression tree. [39].

4. Evaluation metrics

The performance of a regression model is assessed using data, often known as testing data, using the following measures.

4.1 Mean Absolute Error (MAE)

The usage of MAE is applicable in cases when outliers within the data are indicative of tainted values. Indeed, the mean absolute error does not excessively penalize training outliers, thereby offering a comprehensive and limited performance metric for the model. Conversely, the model's performance will improve if the test set contains many outliers. Represented in the Equation (4). The best value is 0, while the worst value is positive infinity [40]. Using Equation:

$$\mathbf{MAE} = \frac{\mathbf{1}}{\mathbf{N}} \sum_{\mathbf{i=1}}^{\mathbf{N}} \left| \mathbf{y_{real_i}} - \mathbf{y_{pred_i}} \right| \qquad (4)$$

4.2 Mean Square Error (MSE)

The utilization of MSE is applicable in cases where the identification of outliers is necessary. Indeed, mean squared error is highly effective at assigning greater importance to data points. When the model produces an abysmal forecast, the MSE function's squaring component amplifies the error's magnitude [40] by using Equation (5).

$$\mathbf{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_{real_i} - y_{pred_i}\right)^2 \qquad (5)$$

4.3 Coefficient of determination (R-squared or R^2)

Linear regression models employ a well-defined statistic called the coefficient of determination (R^2) to quantify the extent to which the observed variation in the dependent variable can be attributed to known predictors [41] with the following Equation (6).

$$\mathbf{R}^{2} = 1 - \frac{\sum_{i=1}^{N}\left(y_{real_i} - y_{pred_i}\right)^2}{\sum_{i=1}^{N}\left(y_{real_i} - \bar{y}\right)^2} \qquad (6)$$

5. The proposed system

In this research used three machine learning models. DTR, RFR, and XGBoost are widely utilized in forecasting due to their adaptability, capability to model non-linear relationships, and ensemble learning technique. These algorithms provide valuable insights into feature importance and can efficiently handle complex datasets, including those in the petroleum sector. Their capacity to deliver precise predictions and interpret results enhances decision-making regarding production optimization and resource allocation. Fig. 4 illustrates the architecture of the fundamental system construction procedures. The suggested framework consists of a series of stages, which include Dataset, data preprocessing (data cleaning, data processing, data normalizing), Machine Learning models, and Evaluation stage.

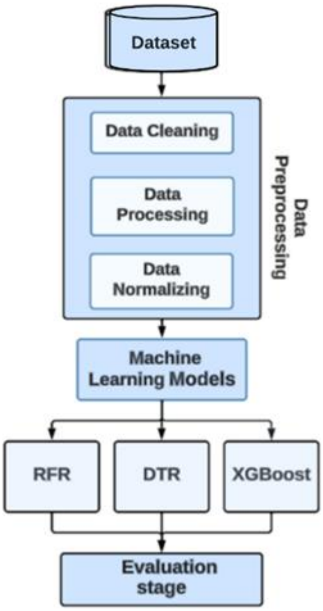

Fig 4. Flowchart for the General Proposed System.

5.1 Datasets Description

The first and most crucial stage of the research is collecting data. The dataset is from USA wells in New York State and contains production information from (1967 -1999)[42]. The dataset included county, town, field, and location operators. It is represented in 30.1K rows and 20 columns. Figure (5) provides part of the Oil and Gas dataset, specifically showcasing the data trends for active oil wells, inactive oil wells, active Gas wells, injection wells, and disposal wells.

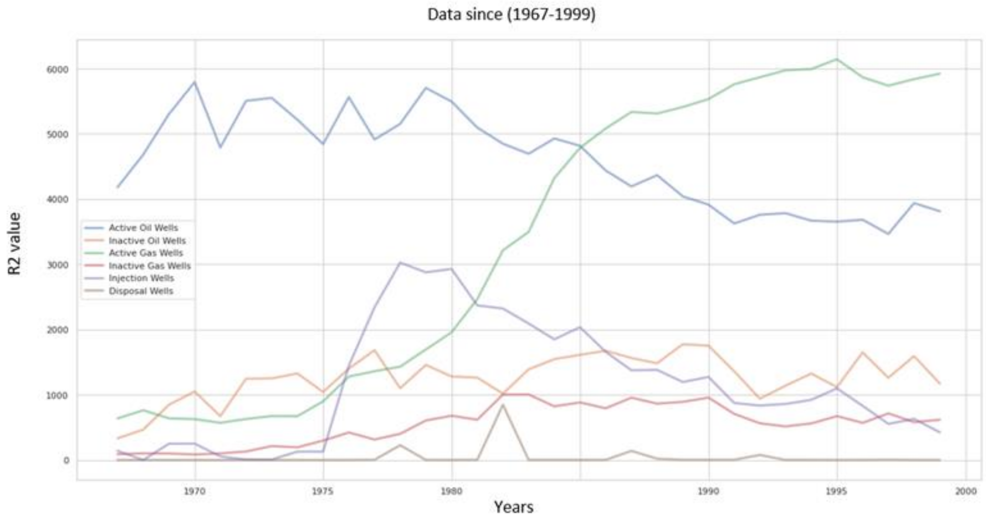

Fig. 5. Illustrated the data trends in the Period time (1967 -1999).

5.2 DATASET PREPROCESSING

The data needs to be preprocessed before manipulation. This research used Google Collab online as an environment for programming in Python. This stage contains three steps for preprocessing data: cleaning, processing, and normalizing. One of the most essential steps in data preprocessing is data cleaning, which is the first step. It is essential to check that the dataset contains all correct information. Second, process the data features like location, each well location will be assigned an index that indicates and split it into two individual sets, X and Y. Finally, Normalizing the dataset. The data normalized by used L2 scaler to transforms numerical values into a range from (0 to 1).

5.3 METHODOLOGY

The most critical and initial step before training a model is determining the parameters for ML models to provide the greatest results. Due to the difficulty of the parameter selection, we attempted all the possible values for each model until we found the optimal values, as shown in Table (1).

Table (1): Machine Learning models parameters

Model	Parameters
DTR	max_depth = 500, random_state = 33
RFR	n_estimators = 50, max_depth =15, random_state = 33
XGBoost	objective =”reg:linear”, n_estimators = 30

After determining the oil and gas parameters, the dataset separated into two independent sets: the training and testing sets.

Training set: This is the most important part of learning datasets to teach machine learning models. It takes up the most space, 75% of data, and (partitioned into a validation and training set). When a model has been trained, it needs a distinct dataset to evaluate performance and adjust hyperparameters, then hyperparameter tuning with the assistance of the validation data; this is called a validation set).

Testing Data: This dataset estimates how the model will perform on fresh data that it has not experienced before in the real world. It takes 25% of the dataset.

6. Experimental results

The same producer was used for three machine learning models with different parameters. The model's results were obtained after the data had been cleaned, normalized, and processed. To determine the most effective model, using metrics such as MAE, R2, and MSE. Table (2) presents performance metrics for three machine learning models (RFR, DTR, XGBoost) across two types of data (Oil and Gas). RFR consistently performs strongly, with slightly higher overall scores than DTR and XGBoost across both Oil and Gas datasets. Meanwhile, XGBoost and DTR exhibit similar mean performance scores, albeit with slightly higher variability in XGBoost's scores Interestingly, Gas data consistently yields higher mean and overall performance scores than Oil data for all models, suggesting potential differences in predictive accuracy between the two datasets. Overall, the findings suggest that RFR is the most consistent performer among the models evaluated. At the same time, Gas data offers slightly better predictive performance than Oil data across all models. The best performance for Oil and Gas production forecasting was achieved with the RFR model, with an average R2 value of 0.9936 for Oil and 0.9982 for Gas.

Tabel (2): illustrated Machine Learning Models result.

ML MODELS	OUTPUT	MAE	MSE	R^2
RFR	Oil	0.0014	0.0002	0.9936
	Gas	0.0017	0.0002	0.9982
DTR	Oil	0.0017	0.0003	0.9892
	Gas	0.0016	0.0003	0.9972
XGBoost	Oil	0.0020	0.0002	0.9921
	Gas	0.0020	0.0004	0.9969

In the Fig. 6 shows the result of the correlation coefficient score (R^2 *value*) for RFR, DTR, and XGBoost.

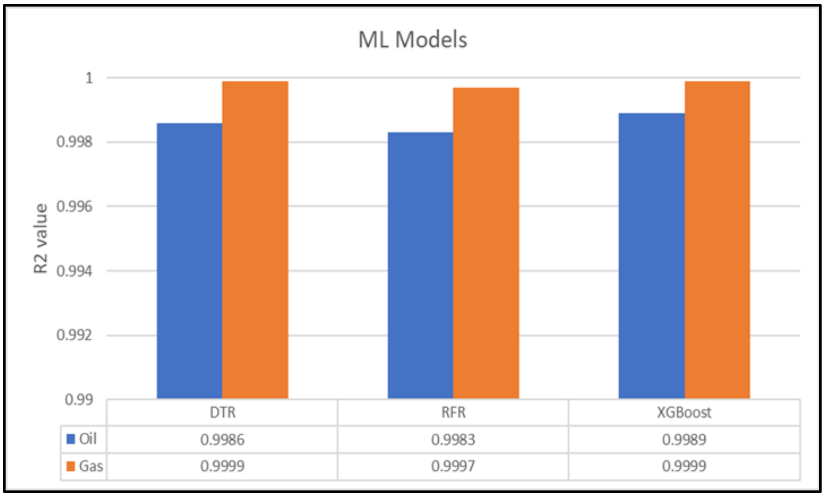

Fig. 6. Illustrated R^2 values for Oil and Gas in ML models.

Fig. 7 shows the matching between the result of predicted Oil produced using machine learning models (RFR, DTR, and XGBoost) and the actual Oil produced. The results indicate perfect alignment between the predicted oil production and the actual oil production. Also, Fig. 8 shows the matching between the result of predicted Gas produced using machine learning models (RFR, DTR, and XGBoost) and the actual Gas produced.

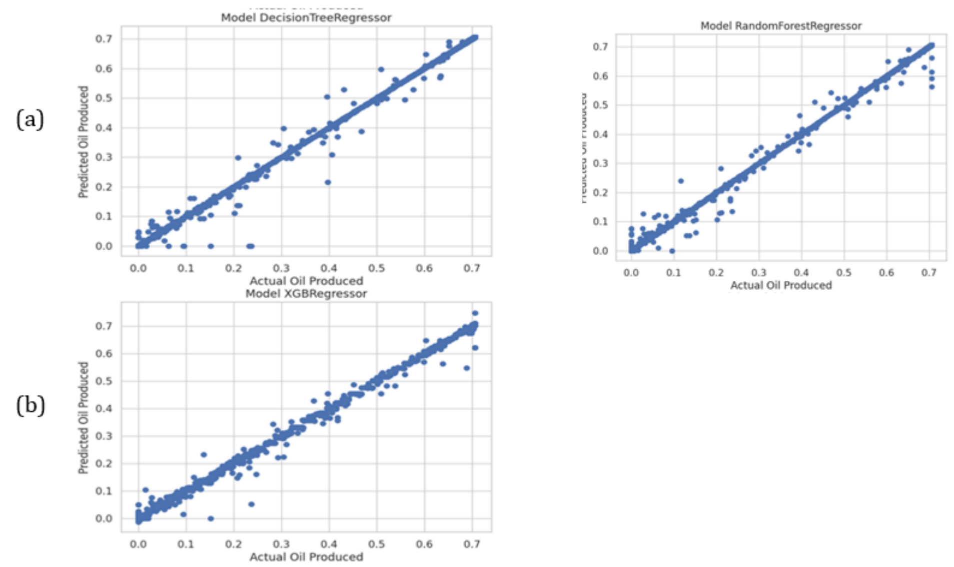

Fig.7. illustrates the matching between the predicte actual Oil production for machine learning models (a) using DTR models, (b) using the XGBoost models, and (c) using the RFR model.

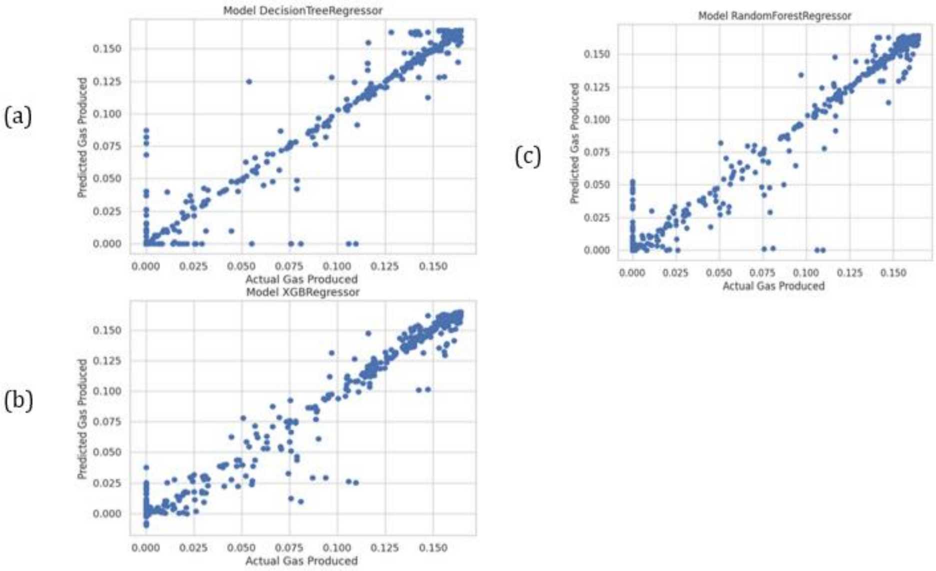

Fig.8. illustrates the matching between the predicted and actual Gas production for machine learning models (a) using DTR models, (b) using the XGBoost models, and (c) using the RFR model.

Conclusion

Precisely predicting how much oil and gas will be extracted is crucial in the petroleum sector, as it enables businesses to better organize their resources, maximize output, and confirm the benefits of forecasting production of oil and gas. Methods and models are utilized to estimate the amount of oil and gas that may be recovered from existing and prospective reserves over a given time period. This research aims to provide a prediction model for oil and gas production in the petroleum sector, facilitating improved resource organization and output maximization for businesses. Oil and gas production forecasts use machine learning models like DTR, RFR, and XGBoost. These models trained and tested the data, then assessed with metrics MSE, MAE, and R2. The experimental result shows the highest accuracy, with the RFR model achieving the highest R^2value of 99%. The objective is to improve and evaluate different sets of procedures. In future projects, develop a system that integrates Machine Learning and deep learning models. Then, compare and select the best model based on dataset type, characteristics, and other relevant criteria.

Reference

[1] S. A. Al-Hilfi and M. A. U. Naser, "Cascade networks model to predict the crude oil prices in Iraq," *International Journal of Electrical and Computer Engineering,* vol. 12, no. 6, p. 6697, 2022, doi: https://doi.org/10.11591/ijece.v12i6.pp6697-6706.

[2] s. culture. "What is Oil and Gas Production?" https://safetyculture.com/topics/oil-and-gas-production/ (accessed.

[3] T. Doan and M. Van Vo, "Using machine learning techniques for enhancing production forecast in north Malay Basin," in *Proceedings of the International Field Exploration and Development Conference 2020*, 2021: Springer, pp. 114-121, doi: https://doi.org/10.1007/978-981-16-0761-5_11.

[4] A. M. AlRassas *et al.*, "Optimized ANFIS model using Aquila Optimizer for oil production forecasting," *Processes,* vol. 9, no. 7, p. 1194, 2021, doi: https://doi.org/10.3390/pr9071194.

[5] I. Makhotin, D. Orlov, and D. Koroteev, "Machine Learning to Rate and Predict the Efficiency of Waterflooding for Oil Production," *Energies,* vol. 15, no. 3, p. 1199, 2022, doi: https://doi.org/10.3390/en15031199.

[6] M. A. Al-Qaness, A. A. Ewees, L. Abualigah, A. M. AlRassas, H. V. Thanh, and M. Abd Elaziz, "Evaluating the applications of dendritic neuron model with metaheuristic optimization algorithms for crude-oil-production forecasting," *Entropy,* vol. 24, no. 11, p. 1674, 2022, doi: https://doi.org/10.3390/e24111674.

[7] E. H. Alkhammash, "An Optimized Gradient Boosting Model by Genetic Algorithm for Forecasting Crude Oil Production," *Energies,* vol. 15, no. 17, p. 6416, 2022, doi: https://doi.org/10.3390/en15176416.

[8] C. S. W. Ng, A. J. Ghahfarokhi, and M. N. Amar, "Well production forecast in Volve field: Application of rigorous machine learning techniques and metaheuristic algorithm," *Journal of Petroleum Science and Engineering,* vol. 208, p. 109468, 2022, doi: https://doi.org/10.1016/j.petrol.2021.109468.

[9] N. M. Ibrahim *et al.*, "Well Performance Classification and Prediction: Deep Learning and Machine Learning Long Term Regression Experiments on Oil, Gas, and Water Production," *Sensors,* vol. 22, no. 14, p. 5326, 2022, doi: https://doi.org/10.3390/s22145326.

[10] M. H. Abed, W. A. Wali, and M. Alaziz, "Machine Learning Approach Based on Smart Ball COMSOL Multiphysics Simulation for Pipe Leak Detection," 2023, doi: https://doi.org/10.37917/ijeee.19.1.13.

[11] B. M. Negash and A. D. Yaw, "Artificial neural network based production forecasting for a hydrocarbon reservoir under water injection," *Petroleum Exploration and Development,* vol. 47, no. 2, pp. 383-392, 2020, doi: https://doi.org/10.1016/s1876-3804(20)60055-6.

[12] Z. Tariq *et al.*, "A systematic review of data science and machine learning applications to the oil and gas industry," *Journal of Petroleum Exploration and Production Technology,* pp. 1-36, 2021, doi: https://doi.org/10.1007/s13202-021-01302-2.

[13] P. F. Orrù, A. Zoccheddu, L. Sassu, C. Mattia, R. Cozza, and S. Arena, "Machine learning approach using MLP and SVM algorithms for the fault prediction of a centrifugal pump in the oil and gas industry," *Sustainability,* vol. 12, no. 11, p. 4776, 2020, doi: https://doi.org/10.3390/su12114776.

[14] K. M. Hanga and Y. Kovalchuk, "Machine learning and multi-agent systems in oil and gas industry applications: A survey," *Computer Science Review,* vol. 34, p. 100191, 2019, doi: https://doi.org/10.1016/j.cosrev.2019.08.002.

[15] K. A. Karoon and Z. N. Nemer, "A Review of Methods of Removing Haze from An Image," doi: https://doi.org/10.37391/ijeer.100354.

[16] Z. Guo, H. Wang, X. Kong, L. Shen, and Y. Jia, "Machine learning-based production prediction model and its application in Duvernay Formation," *Energies,* vol. 14, no. 17, p. 5509, 2021, doi: https://doi.org/10.3390/en14175509.

[17] N. A. Sami and D. S. Ibrahim, "Forecasting multiphase flowing bottom-hole pressure of vertical oil wells using three machine learning techniques," *Petroleum Research,* vol. 6, no. 4, pp. 417-422, 2021, doi: https://doi.org/10.1016/j.ptlrs.2021.05.004.

[18] M. A. M. Fadzil, H. Zabiri, A. A. Razali, J. Basar, and M. Syamzari Rafeen, "Base Oil Process Modelling Using Machine Learning," *Energies,* vol. 14, no. 20, p. 6527, 2021, doi: https://doi.org/10.3390/en14206527.

[19] G. Hui, S. Chen, Y. He, H. Wang, and F. Gu, "Machine learning-based production forecast for shale gas in unconventional reservoirs via integration of geological and operational factors," *Journal of Natural Gas Science and Engineering,* vol. 94, p. 104045, 2021, doi: https://doi.org/10.1016/j.jngse.2021.104045.

[20] C. Tan *et al.*, "Fracturing productivity prediction model and optimization of the operation parameters of shale gas well based on machine learning," *Lithosphere,* vol. 2021, no. Special 4, p. 2884679, 2021, doi: https://doi.org/10.2113/2021/2884679.

[21] X.-y. Wang, Y.-j. Ma, E.-z. Fei, and Y.-f. Gao, "Daily production prediction of oil wells based on machine learning," in *International Conference on Automation Control, Algorithm, and Intelligent Bionics (ACAIB 2023)*, 2023, vol. 12759: SPIE, pp. 516-520, doi: https://doi.org/10.1117/12.2686768.

[22] S. Ray, "A quick review of machine learning algorithms," in *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*, 2019: IEEE, pp. 35-39, doi: https://doi.org/10.1109/comitcon.2019.8862451.

[23] G. Carleo *et al.*, "Machine learning and the physical sciences," *Reviews of Modern Physics,* vol. 91, no. 4, p. 045002, 2019, doi: https://doi.org/10.1103/revmodphys.91.045002.

[24] C. Gkerekos, I. Lazakis, and G. Theotokatos, "Machine learning models for predicting ship main engine Fuel Oil Consumption: A comparative study," *Ocean Engineering,* vol. 188, p. 106282, 2019, doi: https://doi.org/10.1016/j.oceaneng.2019.106282.

[25] M. M. Shawkat, A. R. B. Risal, N. J. Mahdi, Z. Safari, M. H. Naser, and A. W. Al Zand, "Fluid Flow Behavior Prediction in Naturally Fractured Reservoirs Using Machine Learning Models," *Complexity,* vol. 2023, 2023, doi: https://doi.org/10.1155/2023/7953967.

[26] P. Sharma, K. Ramesh, R. Parameshwaran, and S. S. Deshmukh, "Thermal conductivity prediction of titania-water nanofluid: A case study using different machine learning algorithms," *Case Studies in Thermal Engineering,* vol. 30, p. 101658, 2022, doi: https://doi.org/10.1016/j.csite.2021.101658.

[27] J. A. Alhijaj and R. S. Khudeyer, "Integration of EfficientNetB0 and Machine Learning for Fingerprint Classification," *Informatica,* vol. 47, no. 5, 2023, doi: https://doi.org/10.31449/inf.v47i5.4527.

[28] A. Al-Fakih, A. F. Ibrahim, S. Elkatatny, and A. Abdulraheem, "Estimating electrical resistivity from logging data for oil wells using machine learning," *Journal of Petroleum Exploration and Production Technology,* vol. 13, no. 6, pp. 1453-1461, 2023, doi: https://doi.org/10.1007/s13202-023-01617-2.

[29] G. S. Ohannesian and E. J. Harfash, "Epileptic Seizures Detection from EEG Recordings Based on a Hybrid system of Gaussian Mixture Model and Random Forest Classifier," *Informatica,* vol. 46, no. 6, 2022, doi: https://doi.org/10.31449/inf.v46i6.4203.

[30] A. K. Ali and A. M. Abdullah, "Fake accounts detection on social media using stack ensemble system," *International Journal of Electrical and Computer Engineering (IJECE),* vol. 12, no. 3, pp. 3013-3022, 2022.

[31] R. I. Kazim and E. F. Abdullah, "Preprocessing of Drugs Reviews and Classification Techniques," *Journal of Al-Qadisiyah for computer science and mathematics,* vol. 15, no. 3, pp. Page 1-10, 2023, doi: https://doi.org/10.29304/jqcm.2023.15.3.1261.

[32] B. Lu and Y. He, "Evaluating empirical regression, machine learning, and radiative transfer modelling for estimating vegetation chlorophyll content using bi-seasonal hyperspectral images," *Remote Sensing,* vol. 11, no. 17, p. 1979, 2019, doi: https://doi.org/10.3390/rs11171979.

[33] D. K. Seo, Y. H. Kim, Y. D. Eo, W. Y. Park, and H. C. Park, "Generation of radiometric, phenological normalized image based on random forest regression for change detection," *Remote Sensing,* vol. 9, no. 11, p. 1163, 2017, doi: https://doi.org/10.3390/rs9111163.

[34] P. Jain, A. Choudhury, P. Dutta, K. Kalita, and P. Barsocchi, "Random forest regression-based machine learning model for accurate estimation of fluid flow in curved pipes," *Processes,* vol. 9, no. 11, p. 2095, 2021, doi: https://doi.org/10.3390/pr9112095.

[35] N. U. Moroff, E. Kurt, and J. Kamphues, "Machine Learning and statistics: A Study for assessing innovative demand forecasting models," *Procedia Computer Science,* vol. 180, pp. 40-49, 2021, doi: https://doi.org/10.1016/j.procs.2021.01.127.

[36] T. Bikmukhametov and J. Jäschke, "Oil production monitoring using gradient boosting machine learning algorithm," *Ifac-Papersonline,* vol. 52, no. 1, pp. 514-519, 2019, doi: https://doi.org/10.1016/j.ifacol.2019.06.114.

[37] H. Mo, H. Sun, J. Liu, and S. Wei, "Developing window behavior models for residential buildings using XGBoost algorithm," *Energy and Buildings,* vol. 205, p. 109564, 2019, doi: https://doi.org/10.1016/j.enbuild.2019.109564.

[38] C. Cao, P. Jia, L. Cheng, Q. Jin, and S. Qi, "A review on application of data-driven models in hydrocarbon production forecast," *Journal of Petroleum Science and Engineering,* vol. 212, p. 110296, 2022, doi: https://doi.org/10.1016/j.petrol.2022.110296.

[39] M. Zou, W.-G. Jiang, Q.-H. Qin, Y.-C. Liu, and M.-L. Li, "Optimized XGBoost model with small dataset for predicting relative density of Ti-6Al-4V parts manufactured by selective laser melting," *Materials,* vol. 15, no. 15, p. 5298, 2022, doi: https://doi.org/10.3390/ma15155298.

[40] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science,* vol. 7, p. e623, 2021, doi: https://doi.org/10.7717/peerj-cs.623.

[41] D. Zhang, "A coefficient of determination for generalized linear models," *The American Statistician,* vol. 71, no. 4, pp. 310-316, 2017, doi: https://doi.org/10.1080/00031305.2016.1256839.

[42] d. world. "Oil and Gas Summary Production Data: 1967-1999." https://data.world/data-ny-gov/8y5c-ebxg (accessed.