SAKARYA UNIVERSITY JOURNAL OF COMPUTER AND INFORMATION SCIENCES
http://saucis.sakarya.edu.tr/
e-ISSN: 2636-8129
Publisher: Sakarya University
Vol. 8, No. 2, 212-222, 2025
DOI: 10.35377/saucis...1579599
RESEARCH ARTICLE

Enhanced Oil and Gas Production Forecasting Through Stacked generalization Ensemble Learning Technique

Gülüzar Çit [1], Azhar Alyahya [2,*]

[1]Department of Software Engineering, Sakarya University, Sakarya, Türkiye, ror.org/04ttnw109
[2]Department of Software Engineering, Sakarya University, Sakarya, Türkiye, ror.org/04ttnw109

Corresponding author:
Azhar Alyahya, Sakarya University, Department of Software Engineering
azhar.alyahya@ogr.sakarya.edu.tr

Article History:
Received: 05.11.2024
Revised: 14.04.2025
Accepted: 21.04.2025
Published Online: 13.06.2025

ABSTRACT
Planning a strategy throughout the oil and gas sector depends on production forecasting. Precise projections aid in estimating future output rates, streamlining processes, and effectively allocating resources. Techniques like " Decline Curve Analysis (DCA) and Numerical Reservoir Simulation (NRS) " have been used in the past, but they have drawbacks such reliance on static models and time consumption. A stacked generalization ensemble learning method for predicting oil and gas production is presented in this work. Using Python and data from wells in the state of "New York State", the model contains four machine learning techniques: " Random Forest Regressor (RFR), Extremely Randomized Trees Regressor (ETR), K-Nearest Neighbors (KNN), and Gradient Boosting Regressor (GBR) ". The stacked model works better than separate models, according to the results of experiments, via R2 scores of 0.9709 per oil and 0.9998 per gas.

Keywords: Machine learning models, Random Forest regressor, Extremely Randomized Trees Regressor, K-Nearest Neighbors, Gradient boosting regressor, Stacking model

1. Introduction

Many transportation systems rely on oil as their primary energy source, including vehicles, aircraft, ships, and other machinery [1]. Oil also plays an important role in various industrial uses. Crude oil is extracted from wells and refined into petroleum products suitable for consumption as part of the oil production process. The exploratory, extraction, and distributional phases make up this all-encompassing process [2]. Predicting production is essential in oil field development since it helps with economic evaluations, scheduling drilling operations, and designing facility capacity. Therefore, accurate production predictions using data from both operational and dormant wells are in great demand [3]. In order to create efficient economic planning, production forecasting is crucial for businesses and governments alike [4]. Complex numerical reservoir simulations (NRS) and comprehensive engineering evaluations are usual tools for oil and gas forecasting[5]. If oil reservoirs are to be monitored and optimized efficiently, these forecasts must be quite accurate. When it comes to calculating reservoir output, the petroleum sector typically uses conventional approaches like " decline curve analysis (DCA) and numerical reservoir simulations (NRS) " [6]. However, due to intricate static models and numerous dynamic parameters, numerical reservoir simulation models can be difficult and time-consuming. [7]. Conversely, decline curve analysis has many applications but also requires a lot of time and computing power [8]. One potential solution to these problems is machine learning, which can now anticipate oil and gas production more accurately and more quickly than ever before [9]. To efficiently and accurately predict future output, machine learning models use past data. Worldwide, the use of fossil fuels, including oil and natural gas, produced almost 30% of the world's energy in 2020, according to the World Energy Report [10]. Many academics are drawn to applying machine learning approaches to oil production operations because of the need for accurate production forecasts. More and more, the oil and gas sector is looking to machine learning, especially for quick evaluations and production predictions [11]. The field of computer science known as artificial intelligence (AI) aims to teach computers to think and act like humans by simulating human intelligence and applying it to problems that are very complicated and non-linear [12]. A branch of artificial intelligence known as machine learning leverages massive datasets and statistical models to discover answers, enabling computers to mimic human learning in their ability to learn and adapt [13].

2. Related works

In recent times, " the use of machine learning (ML) and deep learning (DL) methods " has become increasingly effective for forecasting oil and gas production, showcasing their capabilities across different formations and data sets. For instance, a study by " Kim et. al. focused on predicting cumulative gas production (CGP) in Canada's Montney Formation using a

Cite as: G. Çit and A. Alyahya, "Enhanced oil and gas production forecasting through stacked generalization ensemble learning technique," *Sakarya University Journal of Computer and Information Sciences,* vol. 8, no. 2, pp. 212-222, 2025. doi: 10.35377/saucis...1579599

range of models, including artificial neural networks (ANN), 1D convolutional neural networks (1D-CNN), long short-term memory (LSTM) networks, and a combination of 1D-CNN and LSTM models. By utilizing early production data, well details, and fracture treatment parameters, the hybrid model showed outstanding performance in enhancing gas production predictions [14].

Similarly, research by Zanjani et al. assessed the effectiveness of ANN, linear regression (LR), and support vector regression (SVR) using data from the Volve field. Although ANN excelled in forecasting hydrocarbon production for well NO159-F-1C, the research underscored the necessity of customizing model selection for specific datasets, as no single algorithm is universally dominant [15].

Tan et al. utilized six algorithms—MLR, XGBoost, LightGBM, ML, RF, and back propagation—to predict output in the WY shale gas block in China, building on previous work in the field. Despite the research's limitations, such as its constricted dataset from a specific area, XGBoost emerged as the most efficient model via an R^2 value of 0.87 [16]. Instead, extra trees achieved the greatest accuracy ($R^2 = 0.809$) when Hui et al. evaluated shale gas production in the Fox Creek region using four methods: linear regression, neural networks, XGBoost, and extra trees [17].

Eight DL and ML models were investigated in Saudi Arabia by N. M. Ibrahim et al.: ANN, RNN, Decision Tree Regression (DTR), XGBoost, SVR, MLR, Polynomial Linear Regression (PLR), and Random Forest Regression (RFR). Based on the data provided by Saudi Aramco, the top-performing networks were ANN, XGBoost, and RNN, via R^2 values of 0.9627 for oil, 0.9012 for gas, and 0.926 for water, respectively. A disadvantage of the research is that the dataset is limited to wells in Saudi Arabia and does not include any geological variation [18]. For time-series data, S. Hosseini et.al proposed a hybrid LSTM-1D CNN model for predicting oil production in the Volve field, with the LSTM model achieving an R^2 score of 0.98. While promising, the study emphasized the need for further investigation into the model's generalizability across different wells [19].

In another study, Song et al. conducted a comprehensive analysis of the productivity of 394 offshore oil wells in China using various machine learning models, such as Linear Regression (LR), XGBoost, LightGBM, Back Propagation (BP) Neural Network, and Long Short-Term Memory (LSTM). XGBoost emerged as the leading model due to its exceptional generalization ability and stability across diverse datasets. In contrast, LightGBM displayed issues with overfitting, highlighting the critical need for selecting suitable machine learning algorithms that align with the dataset's characteristics and specific application requirements. The study underscores that the choice of model can significantly impact the performance and reliability of production forecasts, especially in complex offshore environments [20].

Lastly, Liu et al. proposed a stacked generalization ensemble model to optimize and predict the rate of penetration (ROP) during gas well drilling in Xinjiang, China. This model integrated six machine learning algorithms: Support Vector Regression (SVR), Random Forest (RF), Extremely Randomized Trees (ET), Gradient Boosting (GB), LightGBM, and Extreme Gradient Boosting (XGB). While the ET model showed the highest individual performance with an R^2 of 0.9268 on the test set, the stacked generalization model surpassed all individual models, achieving an R^2 of 0.9568 on the test set. The study effectively demonstrated the enhanced predictive power of combining multiple models, particularly in complex operations like drilling, where high predictive accuracy is crucial for optimizing operational efficiency [21].

Similarly, F.Ye et al. explored production prediction of tight and shale gas wells using a dataset comprising geological, reservoir, engineering parameters, and production data from over 200 wells in Gas Field A and 207 wells in Shale Gas Field B. The study utilized several machine learning algorithms, including Random Forest (RF), Extremely Randomized Trees (ET), LightGBM, and Gradient Boosting Regression (GBR), alongside a blending ensemble learning model. The blending model demonstrated superior predictive accuracy and generalization, particularly for shale gas production, outperforming individual models. The ensemble learning approach achieved an impressive R^2 of 0.9524 for shale gas production, emphasizing the efficacy of integrating multiple models to enhance prediction accuracy in complex reservoir conditions [22].

3. Machine Learning Models (ML)

The current digital transformation activities depend mainly on machine learning, a subfield of Artificial Intelligence AI [23]. It consists of a variety of approaches that enable systems to learn from data and make decisions accordingly. In order to tackle a wide range of problems, the three primary branches of “ machine learning—supervised, unsupervised, and reinforcement learning ”—offer a general framework [24]. “ This study utilized four different supervised machine learning models and one ensemble learning technique ”.

3.1 Random Forest regressor (RFR)

When it comes to classification and regression, the Random Forest method is the better option. It takes input data as a starting point, builds multiple models, gathers predictions from each, and then utilizes a voting process to choose the best answer [25]. Decision trees are the backbone of this method, which averages the results from various decision tree regressors (DTR) to arrive at the final forecast. The forecast is derived by averaging the results from each individual tree [26]. “ Created in 2001 by Professor Leo Breiman, this method is also known as Random Decision Forests ” [27]. To build the model, we first split the input into several samples according to the number of trees, then generate a basic prediction model for each sample, and finally use a bagging technique to combine the results of all the models to get the final

prediction [28]. Each decision tree in the Random Forest is completely grown, so there's no need to slow down processing. To avoid overfitting and get more accurate findings, it's recommended to increase the number of trees [29]. “As seen in Figure 1”.

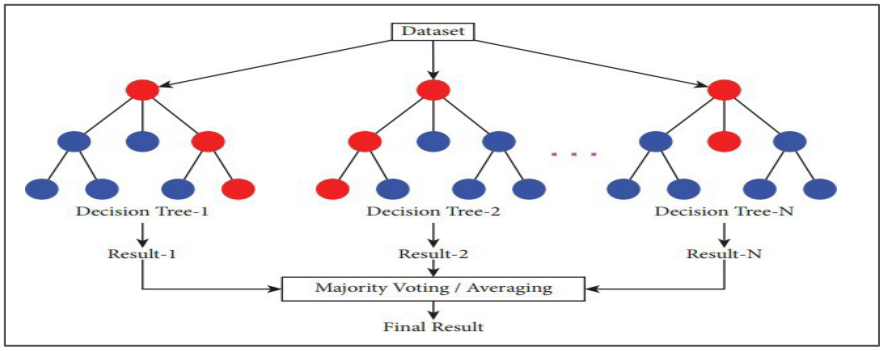

Figure 1. Design of a random forest model [26].

Predicting how many trees to use, how many predictor variables to evaluate at each split, node sizes, and minimum sample counts at leaf nodes are all critical components of the Random Forest Regression (RFR) model [30]. A dependent variable can be predicted by a random forest of simple trees within the framework of a regression model. Using the input variable x, the method generates K separate regression trees $h_{kk}(x)$. The model's forecast is the mean of all the forecasts made by all the trees in the forest using the given set of inputs (x), where k ranges from 1 to K. Using a process called bootstrapping, this methodology can increase the trees' variety, which in turn reduces the possibility that their aggregated results will be similar to other trees [29], “as seen in Equation 1” :

$$\text{RFR prediction } \frac{1}{kk}\sum_{kk=1}^{kk} h_{kk}(x) \qquad (1)$$

3.2 Extremely Randomized Trees Regressor (ETR)

The Extremely Randomized Trees Regressor (ETR) is a robust ensemble machine learning model designed for both regression and classification tasks. Similar to the Decision Tree Regressor, ETR constructs decision trees to model relationships within a dataset. However, it introduces additional randomness to enhance diversity among trees, leading to more robust predictions and reduced overfitting [30].In ETR, the decision trees are built by randomly selecting features and thresholds at each split, as opposed to choosing the optimal split point. This randomness increases the variance among individual trees while maintaining low bias in the overall ensemble. The branches represent the decision-making criteria, while the nodes correspond to options or events. Each node is associated with features, and branches signify potential values of these features [31].

During training, ETR uses a random sampling of data (with replacement, known as bootstrapping) to grow multiple decision trees. The entire dataset is divided into sections using randomly chosen thresholds for the features at each node. This process continues recursively until a stopping criterion, such as the minimum node size, is reached, resulting in terminal nodes [32]. Unlike traditional decision tree algorithms, ETR does not solely rely on the Mean Squared Error (MSE) to determine splits. Instead, splits are chosen by evaluating a random subset of features and thresholds, ensuring faster training and greater model diversity. The model's prediction is the average of the outputs from all trees in the ensemble, calculated, “as seen in Equation 2”.

$$E(m) = \frac{1}{M}\sum_{i=1}^{M} [y_i(m) - \hat{y}(m)]^2 \qquad (2)$$

“As seen in Figure 2”, the initial node in each decision tree acts as the root, representing the entire dataset. Subsequent nodes enable the dataset to be divided into smaller, more homogenous subsets, ensuring accurate predictions while maintaining computational efficiency.

3.3 K -Nearest Neighbors (KNN)

Regression and classification are just two examples of the many applications of the widely used “ K-Nearest Neighbors (KNN) technique”. [33]. It is a simple yet powerful approach in machine learning, where it operates by comparing a data point to its nearest neighbors [34]. The fundamental principle of KNN is to assign an object to a category based on the

characteristics it shares most closely with nearby elements [35]. The value attributed to an object is calculated based on the average of its K nearest neighbors. Applying weights to nearby entities can enhance the method's effectiveness, particularly when closer neighbors have a more substantial influence on the average than those farther away [36]. KNN helps to avoid overfitting by adjusting a parameter, k, which is inversely related to the error rate [35]. “As seen in Equation 3,” One of the most common distance metrics used in KNN is the Euclidean distance, which measures the space between two points, (r, s).

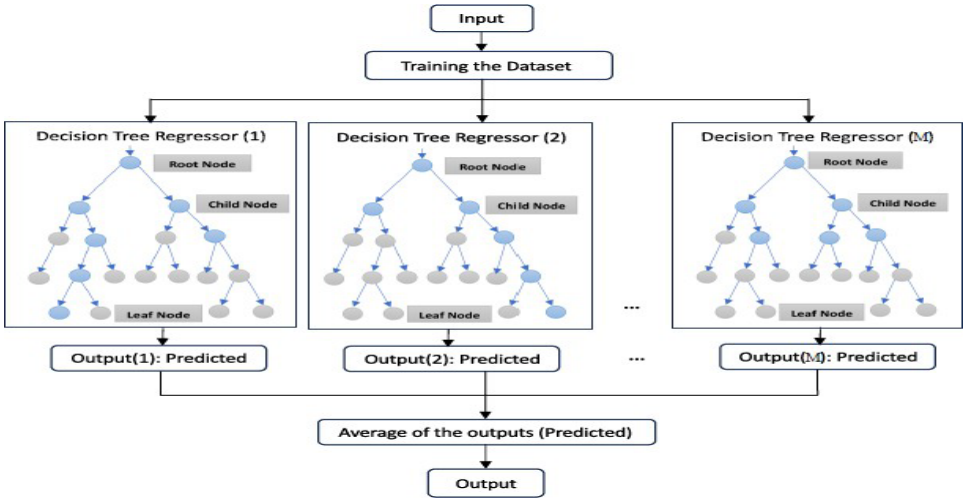

Figure 2. Shows the diagram of the extra-trees regressor [32].

3.3 K -Nearest Neighbors (KNN)

Regression and classification are just two examples of the many applications of the widely used “ K-Nearest Neighbors (KNN) technique”. [33]. It is a simple yet powerful approach in machine learning, where it operates by comparing a data point to its nearest neighbors [34]. The fundamental principle of KNN is to assign an object to a category based on the characteristics it shares most closely with nearby elements [35]. The value attributed to an object is calculated based on the average of its K nearest neighbors. Applying weights to nearby entities can enhance the method's effectiveness, particularly when closer neighbors have a more substantial influence on the average than those farther away [36]. KNN helps to avoid overfitting by adjusting a parameter, k, which is inversely related to the error rate [35]. “As seen in Equation 3,” One of the most common distance metrics used in KNN is the Euclidean distance, which measures the space between two points, (r, s).

$$DDDDDD\,(\mathrm{R,S}) = \overline{\sum_{4=1}^{m} (rr_{ii} - DD_{i})^{2}} \qquad (3)$$

In a space with m dimensions, let R be represented as $rr_1, rr_2, ..., rr_{mm}$, and S as $DD_1, DD_2, ..., DD_{mm}$ [33]. The k-Nearest Neighbors (KNN) Search technique for k equals 3, the three points in dataset D that are nearest to the query point DD_1 “As seen in Figure 3”.

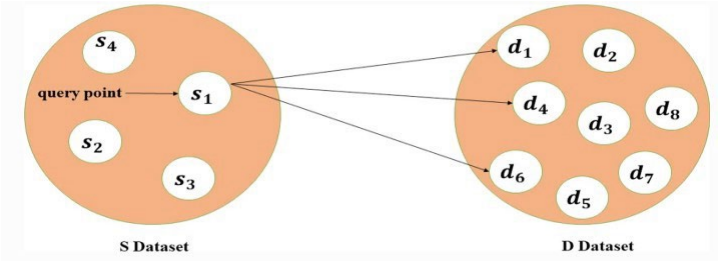

Figure 3. Illustration of the k-Nearest Neighbors (KNN) search method with k set to 3[33].

3.4 Gradient Boosting Regressor (GBR)

Gradient Boosting is a method in ensemble learning that constructs predictive models step by step by integrating the capabilities of weak learners, often decision trees, to develop a robust predictive model [37]. The fundamental principle of Gradient Boosting involves iteratively training new models to address the residual errors of prior models, thereby enhancing the overall accuracy of predictions [38] as seen in Figure 4.

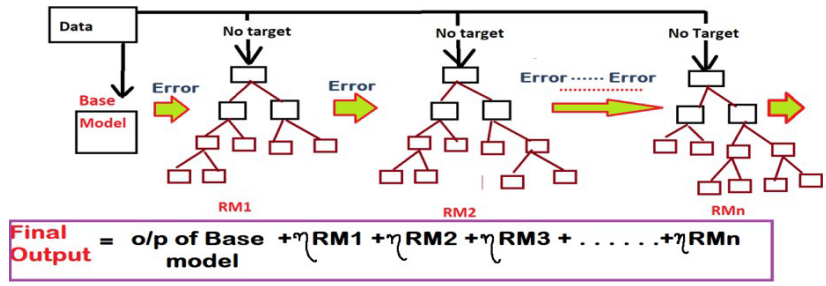

Figure 4. Understanding Gradient Boosting for Regression [38].

3.5 Stacking Regressor

Stacking, also referred to as Stacked Generalization, is an ensemble learning method aimed at boosting predictive accuracy by merging several models. This technique comprises two primary phases: initially, multiple base models (often called level-0 models) are trained using the same dataset; subsequently, a meta-model (level-1) is trained to make predictions from these base models. The outputs of the base models act as input features for the meta-model, which is designed to combine these predictions to enhance the accuracy of the overall prediction, as seen in Figure 5 [39]. In this study, stacking will be applied by integrating four machine learning algorithms. Each base model will analyze the dataset separately, producing predictions that will be utilized by the meta-model.

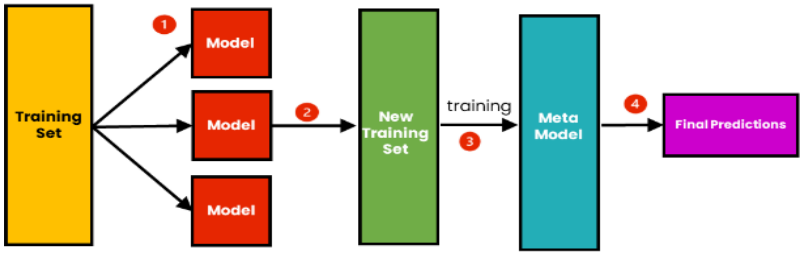

Figure 5. Shows the structure of Stacked Generalization [39].

4. The proposed system

This study utilizes four machine learning techniques, namely Random Forest Regressor (RFR), Extremely Randomized Trees Regressor (ETR), K-Nearest Neighbors (KNN), and Gradient Boosting Regressor (GBR), which are combined through a method known as Stacked Generalization. Prediction challenges frequently make use of these algorithms because of their adaptability, capacity to learn non-linear correlations, and ensemble abilities in learning. Their accuracy in forecasting results and ease of interpretation significantly aid decision-making processes concerning the optimization of the production and distribution of resources. The framework suggested includes several crucial phases, as seen in Figure 6. The procedure starts with the acquisition of the dataset and involves data preprocessing, cleaning and normalization. Subsequently, the machine learning models are applied, utilizing the stacking model to capitalize on the advantages of various algorithms. In the final phase, the system evaluates the models' performance using metrics to ensure accurate and reliable forecasting outcomes.

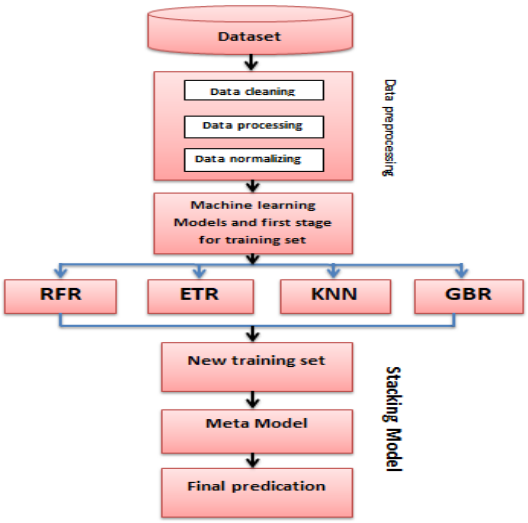

Figure 6. Overview of the Suggested System's Workflow.

4.1 Datasets Description

The main and most important phase of this research is collecting data. This dataset includes production records from oil and gas wells in New York State that were drilled between 2001 and the current day [40]. This data set features details such as county, company name, well status, well type, and producing formation. This dataset comprises 18 columns and approximately 302,000 rows. “As seen in Figures 7 and 8,” showcases a segment of this data set, highlighting oil and gas production for the ten most active wells.

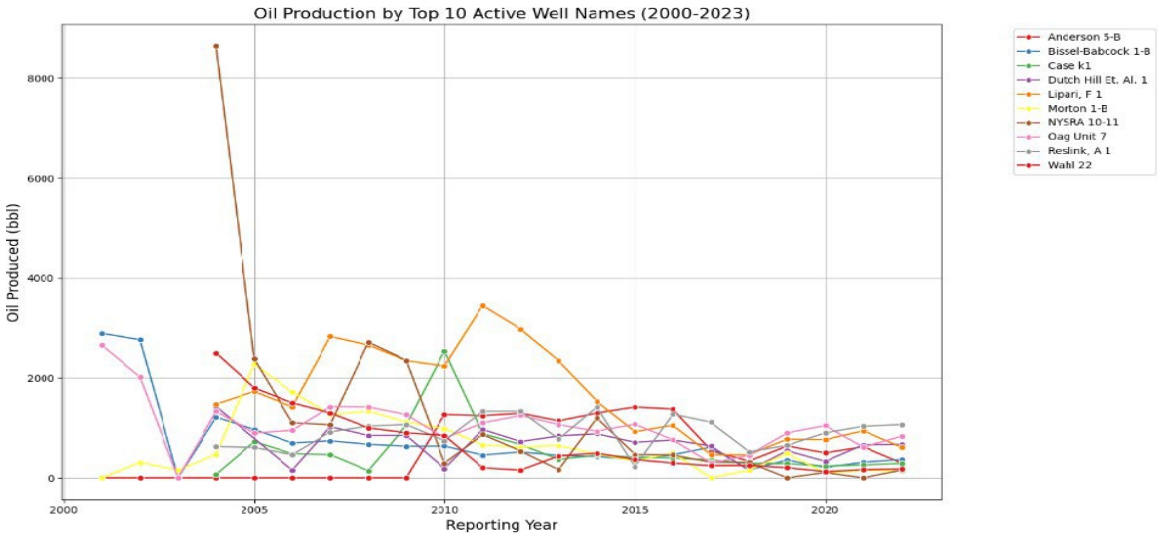

Figure 7.Illustrated oil production: top 10 active wells.

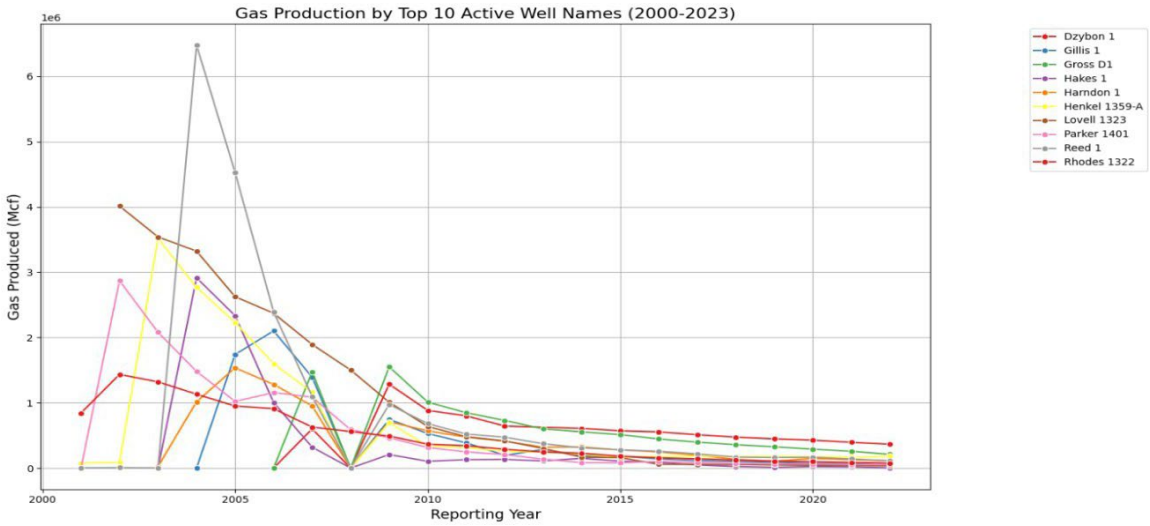

Figure 8. Illustrated gas production of the top 10 active wells.

4.2 Dataset Preprocessing

Before manipulating data, it's essential to prepare the dataset to ensure its quality and suitability for analysis. In this study, Google Colab was employed as the online programming platform for Python. The data preparation phase involved three main steps: cleaning the data, processing the data, and normalizing it. The initial and most crucial step, data cleaning, ensures that the dataset is devoid of errors and inconsistencies, confirming the accuracy and completeness of all information. Next, the data processing step and the normalization step adjust the numerical values using an L2 scaler, which scales the data to a range from 0 to 1.

4.3 Methodology

The essential first step in training a machine learning model involves identifying the parameters that will yield the best outcomes. Given the complexity of selecting these parameters, we explored every possible value until we established the best settings for each model as seen in Table 1.

Table 1. Parameters for machine learning models.

Model	Parameters
RFR	n_estimators = 40, max_depth =13, random_state = 33
ETR	n_estimators = 50, max_depth =15, random_state = 33
KNN	n_neighbors=13, weights='uniform', algorithm='auto
GBR	n_estimators=50, learning_rate=0.1, random_state = 33

When the oil and gas parameters have been defined, the data is split into two sets: the training dataset and the testing dataset. Machine learning models are trained using the training dataset, which is an essential subset of the dataset. It is divided into a training set and a validation set and takes up the most space, 75% of the total data. To evaluate the model's performance and adjust its hyperparameters, a separate dataset is needed after training. Hyperparameter optimization benefits from this dataset, which is called the validation set. The purpose of the testing dataset is to foretell how the model will respond to novel, unseen data in real-world scenarios. It makes up a quarter of all the data.

5. Experimental results

The methodology was consistently applied to four machine learning algorithms: Random Forest Regressor (RFR), Extremely Randomized Trees Regressor (ETR), K-Nearest Neighbors (KNN), and Gradient Boosting Regressor (GBR), each utilizing unique sets of parameters. These algorithms were implemented following the cleaning, normalization, and processing of the dataset. The results generated by these individual models are combined through Stacked Generalization, an ensemble learning model, and used as input features for a final model (known as the meta-model or final estimator), which, by combining the results of the base models, learns to predict the target variable. That captures different patterns and interactions in the data that individual models might miss. In this stacking regressor, the " final estimator is a RandomForestRegressor, with n_estimators=50 and random_state=42". This strategy helps to enhance predictive accuracy. The performance of these models was assessed using essential metrics such as Mean Absolute Error (MAE), R^2, and Mean Squared Error (MSE). "As seen in Table 2", the effectiveness of each single model and the stacked generalization model across two categories of data: Oil and Gas. The stacked model yielded superior results in predicting oil and gas production, achieving the highest R^2 scores, with averages of 0.974750 for oil and 0.999848 for gas. Overall, the models demonstrated slightly better predictive accuracy for gas data compared to oil data, highlighting their superior performance for forecasting in this scenario.

Table 2. Results of machine learning models and the stacking model.

Models	OUTPUT	MAE	MSE	R^2
RFR	Oil	0.000273	0.000011	0.948631
	Gas	0.000273	0.000011	0.999720
ETR	Oil	0.000295	0.000073	0.932277
	Gas	0.000295	0.000007	0.999819
KNN	Oil	0.000324	0.000017	0.938847
	Gas	0.000324	0.000017	0.999580
GBR	Oil	0.00306	0.000276	0.391319
	Gas	0.004396	0.000137	0.996638
Stacking model	Oil	0.000469	0.000013	0.970900
	Gas	0.000221	0.000007	0.999806

As seen in Figure 9, RFR, ETR, KNN, GBR, and the Stacking model correlation coefficient scores (R^2 *vvvvvvvvv*).

As seen in Figure 10, a comparison of the actual oil output with the forecasted oil production using several machine learning methods, such as " RFR, ETR, KNN, GBR, and a Stacking Model". According to the findings, there was an accurate correspondence between the forecasted and actual oil production. Similarly, "As seen in Figure 11", the labeled figure also compares actual gas output with predicted gas production using the same ML models and the Stacking Model. The models' capacity to precisely forecast gas output is demonstrated by the findings, which reveal a high link. The Stacking Model closely matches the real data for both gas and oil production, demonstrating outstanding forecasting capabilities.

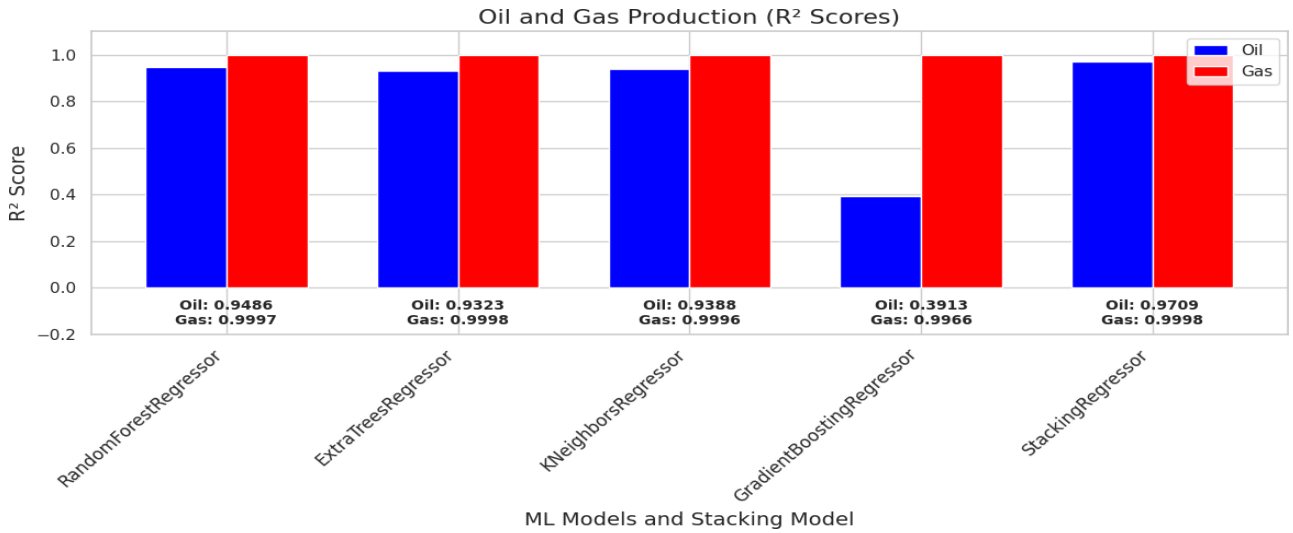

Figure 9. Displays Oil and Gas R² values in ML models and the Stacking Model.

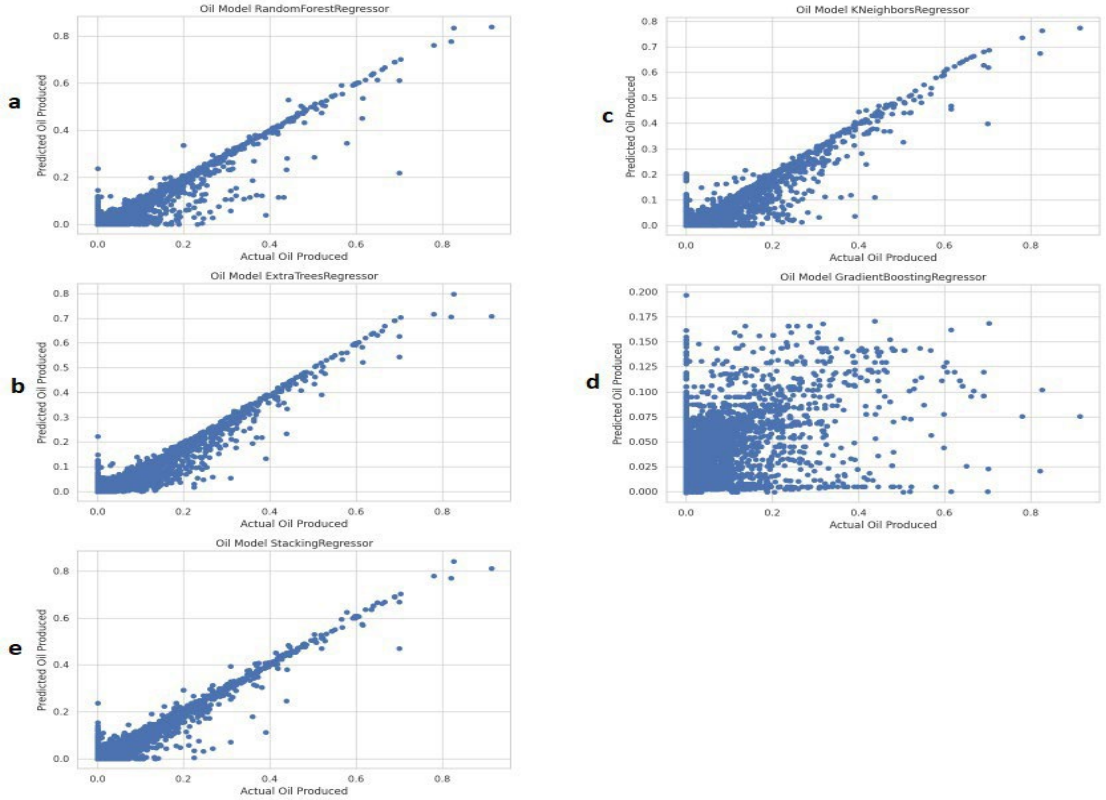

Figure 10. presents a comparison between the forecasted and actual Oil production employing various machine learning models, as well as a stacking model. The results are as follows: (a) Random Forest Regressor (RFR), (b) Extremely Randomized Trees Regressor (ETR), (c) K-Nearest Neighbors (KNN), (d) Gradient Boosting Regressor (GBR), and (e) the Stacking Model.

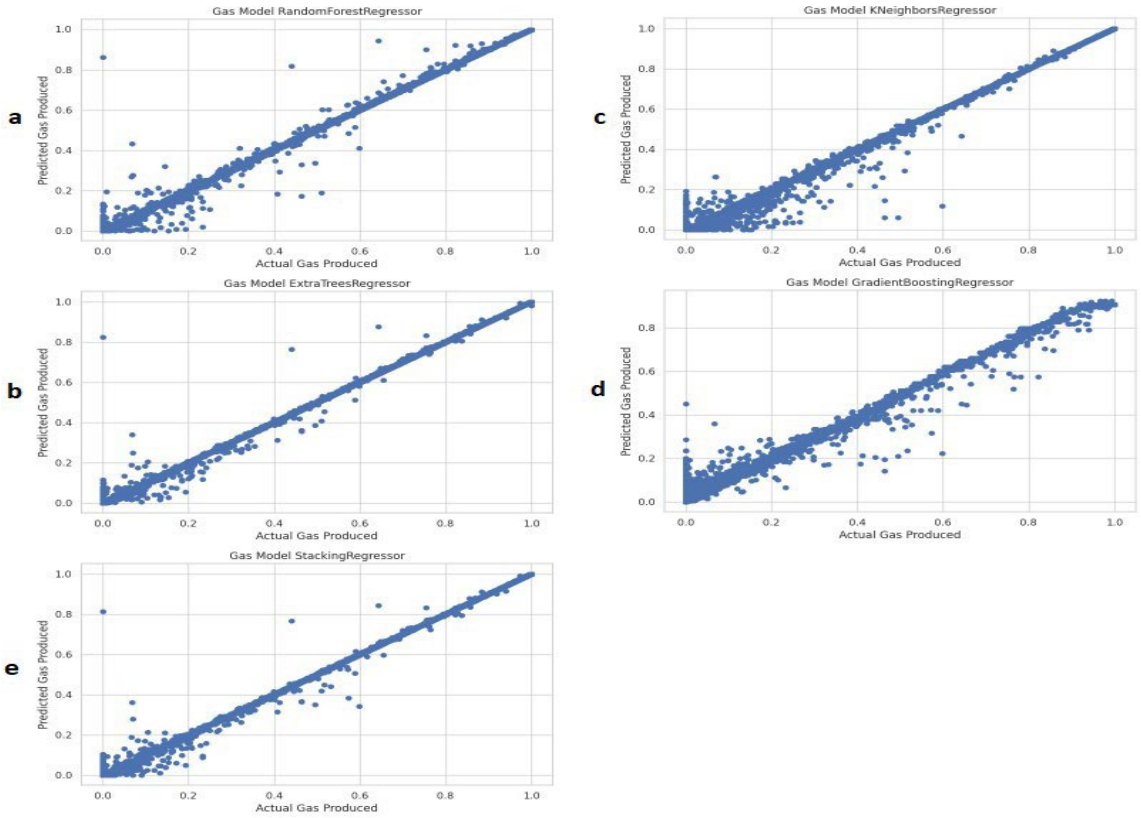

Figure 11. presents a comparison between the forecasted and actual Gas production employing various machine learning models, as well as a stacking model. The results are as follows: (a) Random Forest Regressor (RFR), (b) Extremely Randomized Trees Regressor (ETR), (c) K-Nearest Neighbors (KNN), (d) Gradient Boosting Regressor (GBR), and (e) the Stacking Model.

6. Conclusion

Predicting the amount of achievable oil and gas with any degree of accuracy is necessary for the petroleum sector .This ability allows companies to efficiently allocate resources, optimize production, and validate the advantages of predicting oil and gas output. Various techniques and models are employed to assess the potential recovery from current and future reserves over a specified timeframe. “ The study employs machine learning models such as Random Forest Regressor (RFR), Extremely Randomized Trees Regressor (ETR), K-Nearest Neighbors (KNN), and Gradient Boosting Regressor (GBR) for production prediction ”. These models are trained and tested on production data, and their results are combined using Stacked Generalization, a method of ensemble learning. The performance of the models is assessed using metrics like “ Mean Squared Error (MSE), Mean Absolute Error (MAE), and R^2 ”. The experimental findings show that the Stacking Model delivers the highest accuracy, with an R-squared value of 99%, indicating its superior predictive capability . Improving predicted accuracy is the primary goal of this research, which uses an ensemble learning method. In the future, we want to build a more unified system that integrates ML and DL models with other ensemble learning techniques, evaluates them against each other, and chooses the best model according to dataset type, attributes, and other pertinent factors.

References

[1] British Petroleum, "Statistical Review of World Energy," BP Global, 2021. [Online]. Available: https://www.bp.com

[2] J. G. Speight, Handbook of Petroleum Refining. 2014. [Online]. Available: https://www.academia.edu/63659108/Handbook_of_Petroleum_Refining

[3] D. Orodu, O. F. Aworinde, and A. F. Alayande, "A hybrid machine learning framework for enhanced reservoir characterization," J. Petroleum Sci. Eng., vol. 207, p. 109114, 2021, doi: 10.1016/j.petrol.2021.109114.

[4] A. F. Khan and S. R. Alam, "Adaptive Neuro-Fuzzy Inference System with metaheuristic tuning for petroleum production forecasting," Applied Soft Computing, vol. 114, p. 108050, 2022, doi: 10.1016/j.asoc.2021.108050.

[5] M. A. Ullah, S. M. Khaleque, and S. Sikder, "Prediction of oil production using optimized machine learning models," Energies, vol. 14, no. 16, p. 4923, 2021, doi: 10.3390/en14164923.

[6] M. J. Fetkovich, "Decline Curve Analysis Using Type Curves," J. Petroleum Technol., vol. 32, no. 6, pp. 1065-1077, 1980.

[7] M. J. Abhishek and V. Kumar, "Gradient boosting regression tree model for enhanced oil production prediction," Processes, vol. 10, no. 2, p. 234, 2022, doi: 10.3390/pr10020234.

[8] K. M. Ali and J. Zhang, "Application of metaheuristic optimization algorithms for predictive analysis in petroleum engineering," J. Petroleum Exploration Production Technol., vol. 12, no. 5, pp. 1325–1335, 2022, doi: 10.1007/s13202-021-01402-w.

[9] 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," J. Petroleum Sci. Eng., vol. 208, p. 109468, 2022, doi: 10.1016/j.petrol.2021.109468.

[10] S. D. Mohaghegh, "Machine Learning Applications in Reservoir Engineering: Part 1," J. Petroleum Technol., vol. 69, no. 6, pp. 70-77, 2017, doi: 10.2118/0617-0070-JPT.

[11] J. X. Chen, H. L. Wang, and K. Zhao, "Comparative evaluation of machine learning techniques for hydrocarbon reservoir prediction," Energies, vol. 14, no. 3, p. 806, 2021, doi: 10.3390/en14030806.

[12] A. S. Abou-Sayed, "AI in the Petroleum Industry," Society of Petroleum Engineers AI Newsletter, 2021. [Online]. Available: https://www.spe.org

[13] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed. Pearson, 2021.

[14] I. Goodfellow, Y. Bengio, and A. Courville, Deep Learning. MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[15] M. Kim, "Deep Learning-Based Prediction of the Cumulative Gas Production of the Montney Formation, Canada," GeoConvention, 2020. [Online]. Available: https://geoconvention.com/wp-content/uploads/abstracts/2020/57980-deep-learning-based-prediction-of-the-cumulative-g.pdf

[16] M. S. Zanjani, M. A. Salam, and O. Kandara, "Data-Driven Hydrocarbon Production Forecasting Using Machine Learning Techniques," Int. J. Comput. Sci. Inf. Security, vol. 18, no. 6, pp. 65–72, 2020.

[17] 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: 10.2113/2021/2884679.

[18] 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," J. Natural Gas Sci. Eng., vol. 94, p. 104045, 2021, doi: 10.1016/j.jngse.2021.104045.

[19] 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: 10.3390/s22145326.

[20] S. Hosseini and T. Akilan, "Advanced Deep Regression Models for Forecasting Time Series Oil Production," arXiv preprint arXiv:2308.16105, 2023.

[21] L. Song, C. Wang, C. Lu, S. Yang, and C. Tan, "Machine Learning Model of Oilfield Productivity Prediction and Performance Evaluation," J. Physics: Conference Series, vol. 2468, no. 1, p. 012084, 2022, doi: 10.1088/1742-6596/2468/1/012084.

[22] N. Liu, H. Gao, Z. Zhao, Y. Hu, and L. Duan, "A stacked generalization ensemble model for optimization and prediction of the gas well rate of penetration: a case study in Xinjiang," J. Petroleum Exploration Production Technol., vol. 11, pp. 3533-3546, 2021, doi: 10.1007/s13202-021-01402-z.

[23] F. Ye, X. Li, N. Zhang, and F. Xu, "Prediction of Single-Well Production Rate after Hydraulic Fracturing in Unconventional Gas Reservoirs Based on Ensemble Learning Model," Processes, vol. 12, no. 6, p. 1194, 2024, doi: 10.3390/pr12061194.

[24] S. Ray, "A quick review of machine learning algorithms," in Proc. Int. Conf. Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 2019, pp. 35-39, doi: 10.1109/comitcon.2019.8862451.

[25] M. I. Jordan and T. M. Mitchell, "Machine learning: Trends, perspectives, and prospects," Science, vol. 349, no. 6245, pp. 255-260, 2015, doi: 10.1126/science.aaa8415.

[26] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001, doi: 10.1023/A:1010933404324.

[27] M. Y. Khan, "Automated prediction of Good Dictionary EXamples (GDEX): a comprehensive experiment with distant supervision, machine learning, and word embedding-based deep learning techniques," Complexity, vol. 2021, pp. 1-18, 2021, doi: 10.1155/2021/2553199.

[28] L. Breiman, "Bagging Predictors," Machine Learning, vol. 24, no. 2, pp. 123–140, 1996, doi: 10.1007/BF00058655.

[29] A. K. Ali and A. M. Abdullah, "Fake accounts detection on social media using stack ensemble system," Int. J. Electrical Comput. Eng., vol. 12, no. 3, pp. 3013-3022, 2022.

[30] S. P. Rao and A. V. K. Shetty, "Random forest-based predictive models for enhanced fluid flow estimation in pipelines," J. Petroleum Sci. Eng., vol. 199, p. 108382, 2021, doi: 10.1016/j.petrol.2021.108382.

[31] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," Machine Learning, vol. 63, no. 1, pp. 3–42, 2006, doi: 10.1007/s10994-006-6226-1.

[32] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, 2009.

[33] T. Aziz and M. R. Camana, "REM-Based Indoor Localization with an Extra-Trees Regressor," Electronics, vol. 12, no. 20, p. 4350, 2023, doi: 10.3390/electronics12204350.

[34] R. K. Halder, "Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications," J. Big Data, vol. 11, no. 1, 2024, doi: 10.1186/s40537-024-00973-y.

[35] T. Timbers, T. Campbell, and M. Lee, "Chapter 7 Regression I: K-nearest neighbors," in Data Science: A First Introduction, CRC Press, 2022. [Online]. Available: https://datasciencebook.ca/regression1.html

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

[37] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," The Annals of Statistics, vol. 29, no. 5, pp. 1189-1232, 2001.

[38] A. Ali, "Gradient Boosting Machine Learning Algorithm," Dec. 2023, doi: 10.13140/RG.2.2.31609.65123.

[39] M. Kalirane, "Ensemble Learning in Machine Learning: Bagging, Boosting and Stacking," Analytics Vidhya, Jan. 2024. [Online]. Available: https://www.analyticsvidhya.com/blog/2023/01/ensemble-learning-methods-bagging-boosting-and-stacking/

[40] "Oil and Gas Annual Production: Beginning 2001," Data.gov. [Online]. Available: https://catalog.data.gov/dataset/oil-and-gas-annual-production-beginning-2001

Author Information Form

Author(s) Contributions

Azhar Alyahya was responsible for carrying out the research, analyzing the data, and preparing the first draft of the manuscript. Dr. Gülüzar Çit provided academic supervision throughout the project, contributed to the study design and interpretation of the findings, and offered valuable feedback and revisions on the manuscript. Both authors reviewed and approved the final version of the paper.

Conflict of Interest Notice

No potential conflict of interest was declared by authors.

Ethical Approval

It is declared that during the preparation process of this study, scientific and ethical principles were followed, and all the studies benefited from are stated in the bibliography.

Availability of data and material

https://catalog.data.gov/dataset/oil-and-gas-annual-production-beginning-2001

Artificial Intelligence Statement

No artificial intelligence tools were used while writing this article.

Plagiarism Statement

This article has been scanned by iThenticate ™.