Scientific Contributions Oil & Gas, Vol. 49. No. 2, June: 203 - 217

SCIENTIFIC CONTRIBUTIONS OIL AND GAS

Testing Center for Oil and Gas
LEMIGAS

Journal Homepage:http://journal.lemigas.esdm.go.id
ISSN: 2089-3361, e-ISSN: 2541-0520

Artificial Intelligence-Based Reservoir Quality Clustering to Determine New Drilling Well Locations

Jeffier Winarta[1,2], Amega Yasutra[1] and Fajril Ambia[2]

[1]Petroleum Engineering, Faculty of Mining and Petroleum Engineering, Institut Teknologi Bandung, Ganesa Street No. 10, Bandung, West Java, 40132, Indonesia.

[2]SKK Migas,
Wisma Mulia Building, 35th floor, Gatot Subroto Street No.42, Jakarta, Indonesia.

Corresponding author: Jeffier Winarta (Jwinarta@skkmigas.go.id)

Manuscript received: March 3th, 2026; Revised: March 16th, 2026
Approved: March 30th, 2026; Available online: June 04th, 2026; Published: June 04th, 2026.

ABSTRACT - Indonesia's oil and gas industry faces ongoing challenges in maintaining production due to the maturity and decline of many fields. Success in drilling new wells is vital for sustaining output, but performance forecasting is hampered by reservoir variability, geological complexity, and operational differences. These challenges underscore the need for adaptive, data-driven predictive methods in technical planning. This study develops a machine-learning-based model for predicting Estimated Ultimate Recovery (EUR) using a KNIME workflow. K-Means clustering groups wells by reservoir characteristics, and three regression algorithms (Simple Regression, Gradient Boosting, and Random Forest) are compared for predicting EUR. The combined workflow evaluates how reservoir segmentation improves EUR prediction accuracy. The methodology consists of four main stages: (1) collecting and preprocessing historical well data from long-producing oil fields located in the Pekanbaru region, (2) applying K-Means clustering using reservoir features such as porosity, permeability, Net Pay, Water Saturation, and well coordinates, (3) constructing EUR regression models using Simple Regression, Gradient Boosting, and Random Forest with an 80% training and 20% testing scheme, and (4) validating model performance using evaluation metrics such as R^2 and RMSE. All processes were performed using the KNIME platform to ensure a standardiszed, transparent, and easily replicable workflow. This study is expected to produce an EUR prediction model that is accurate and stable, while also identifying the most suitable regression algorithm that is most suitable for mature Indonesian reservoirs. Furthermore, the integrated KNIME workflow can serve as a foundation for a decision support system to assist in drilling planning, investment optimiszation, and the reduction of production uncertainties. Furthermore in addition, the results of the clustering and modeling can be used to identify the mobest prospective reservoir zones, thereby supporting the selection of new well drilling locations in areas with the highest production potential.

Keywords: gradient boosting, K-means clustering, KNIME, machine learning, random forest, simple regression

How to cite this article:

Jeffier Winarta, Amega Yasutra and Fajril Ambia, 2026, Artificial Intelligence-Based Reservoir Quality Clustering to Determine New Drilling Well Locations, Scientific Contributions Oil and Gas, 49 (2) pp. 205-219. DOI org/10.29017/scog.v49i2.2075.

INTRODUCTION

The oil and gas industry is facing significant challenges due to declining production from mature fields, limited discoveries of new reserves, and increasing subsurface complexity in reservoirs that have been producing for decades. In Indonesia, more than 70% of national production comes from fields that have been in operation for more than 25 years, resulting in higher reservoir depletion rates and increased technical uncertainty in drilling planning. In this situation, development drilling and infill drilling are key strategies for maintaining and increasing production, as well as optimizing the recovery of remaining oil.

The success of drilling in old fields is highly dependent on the accuracy of predicting the potential of new wells, especially in estimating the estimated ultimate recovery (EUR). However, the process of evaluating drilling candidates is still largely carried out using conventional approaches such as manual geological analysis, separate assessment of reservoir parameters, or simple spatial interpolation.

These traditional approaches are often time-consuming, rely heavily on the subjectivity of engineer interpretation, and cannot capture the non-linear relationships between reservoir parameters, subsurface conditions, and well performance. With the availability of historical production data, reservoir properties, and operational data in increasingly large volumes, data-driven analytics and machine learning approaches offer significant opportunities to improve the accuracy and consistency of drilling predictions (Candra et al., 2024; Mahfudhoh & Welayaturromadhona 2026).

Previous studies have shown the great potential of machine learning techniques in well performance analysis and production prediction. Chahar et al. (2022) proved that algorithms such as Random Forest, Gradient Boosting, and Artificial Neural Network can achieve high accuracy in predicting daily production rates even in fields with high variability. In an industrial setting, several studies have reported that ensemble learning provides more stable performance than single models, especially on distributed and heterogeneous datasets. In addition, data mining methods such as clustering and association rule mining have demonstrated their effectiveness in identifying complex patterns between reservoir parameters and well operation performance (Maulana et al., 2025) and are widely used in the evaluation of workover candidates and other subsurface work.

A field study at Pertamina Hulu Sanga Sanga (Satriaji et al., 2024) also showed that decision tree models could predict subsurface work performance with an accuracy of more than 90%, indicating that supervised learning can capture multivariable patterns that manual analysis cannot access. Machine learning models are also known to be sensitive to data quality and feature selection, particularly in heterogeneous reservoir conditions (Zainuri et al., 2023; Rohmana et al., 2026).

However, most previous studies have focused on predicting production rates or subsurface work success, and few have developed a comprehensive approach to predicting EUR as a basis for selecting new drilling candidates in mature fields. In fact, the ability to estimate EUR accurately and consistently is an important component in determining the technical and economic feasibility of a well location. Furthermore, there is no integrated approach that combines preprocessing data, reservoir segmentation, feature selection, and regression modeling in a single structured analytical workflow.

In industrial practice, selecting new well locations requires not only an understanding of geological conditions and reservoir parameters but also the ability to assess the potential of a zone based on the performance of surrounding wells already in production. This spatial performance comparison becomes increasingly important in mature fields, where reservoir heterogeneity increases, fluid contact boundaries change, and production responses between wells can vary significantly even within the same zone. Therefore, an analytical method is needed that not only predicts the EUR for candidate wells but also systematically integrates information from existing wells to identify zones with the highest potential. Data-driven approaches have also been successfully applied in supporting engineering decision-making processes, including well candidate selection and performance evaluation (Prayitno et al., 2025; Muhendra et al., 2025).

The integration of historical production information, reservoir contribution patterns, and EUR estimates of surrounding wells can provide an objective basis for engineers in determining the best well locations, while reducing reliance on intuition and subjective interpretation. This need was the main motivation for the development of a machine-learning-based workflow in this study.

In response to these needs, this study aims to develop an integrated KNIME-based analytical workflow that combines K-Means clustering techniques with three regression algorithms (Random Forest, Gradient Boosting, and Simple Regression) to predict EUR in mature fields. This approach integrates data exploration, reservoir segmentation based on geological uniformity, and supervised learning modeling into a single structured system. In addition to producing more accurate EUR predictions, this workflow is intended to support the drilling decision-making process by providing a quantitative basis that helps engineers determine the location of new drilling wells in reservoir zones that have the most potential. The main objective of this study also includes evaluating and comparing the accuracy levels of several regression methods used in conjunction with K-Means clustering results. Through this comparison, it is possible to identify the algorithm that provides the most optimal performance in modeling reservoir variability and producing more precise EUR predictions. Thus, this research is expected to make a significant contribution to improving the accuracy of drilling strategies in long-producing oil fields.

METHODOLOGY

This study was designed to develop a machine learning model capable of predicting the potential EUR in new drilling wells in old oil fields. This approach integrates unsupervised learning techniques (K-Means clustering) and three supervised regression algorithms (Random Forest, Gradient Boosting, and Simple Regression) into a structured analytical workflow based on the KNIME Analytics Platform. The research stages include:

- Data collection and compilation,
- Data preprocessing,
- Reservoir segmentation using K-Means,
- Regression model development, and
- Model performance evaluation.

The following figure 1 shows the flow of the research method used.

Data source

The dataset used in this study comes from well realization data reported by cooperation contract contractors (KKKS). The data include more than 1000 well entries taken from the 2021 to 2024 time span. This dataset represents the conditions of old oil fields in the Pekanbaru region that have been in production for several decades. The dataset includes four major groups of variables:

- Subsurface coordinates, such as: a.) X-coordinate; b.) Y-coordinate
- Reservoir parameters, including: a.)Porosity (Ö); b.) Permeability (k) c.) Net Pay (ft); d.) Water Saturation (Sw) using the initial Sw value obtained from the evaluation results when the well was first drilled (open-hole log interpretation). This value represents the reservoir's initial condition before changes due to production.

- Well design and mechanical factors, such as perforation depth (TOPPRF, BOTPRF)
- EUR Realization Data (actual EUR value used as a target variable in regression modeling). This EUR value comes from analytical evaluation by the contractor (contractor analytical EUR evaluation), using the decline curve analysis or material balance method according to actual production data.

Feature selection follows the approach in previous studies (Satriaji et al., 2024; Chahar et al., 2022), in which a combination of reservoir parameters, well conditions, and production data provides the most significant information for predicting the success of the work.

Methodology

The research methodology is designed to build a machine-learning-based analytical workflow capable of predicting EUR to support the selection of new well drilling candidates in old oil fields. The research framework was developed by integrating data preparation, exploratory data analysis (EDA), reservoir segmentation using K-Means clustering, and regression modeling based on ensemble learning algorithms, which have been widely applied to capture complex non-linear relationships in reservoir data (Sugiyanto & Utama 2025; Mahfudhoh & Welayaturromadhona 2026). The entire process was built using the KNIME Analytics Platform so that the workflow could be visualized in a modular, reproducible, and easily evaluable manner, consistent

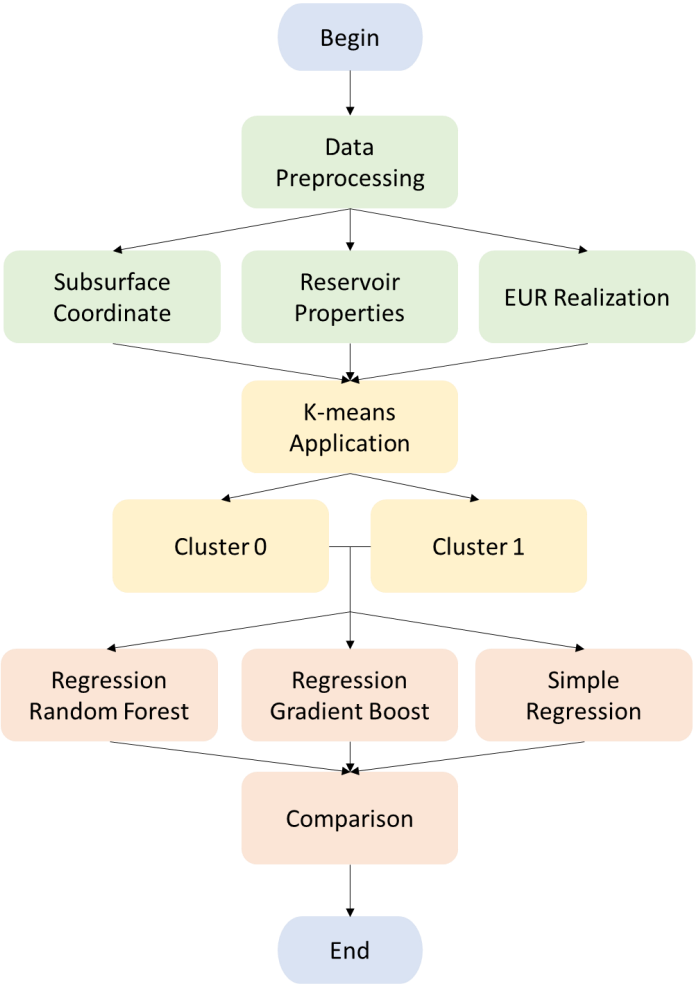


Figure 1. Flowchart of clustering and EUR prediction process.

with previous data-driven workflow implementations in reservoir engineering studies (Muhendra et al., 2025; Rafsanjani et al., 2025).

Data preparation

This stage includes the following processes:

- Data cleaning: addressing missing values, outliers, format inconsistencies, and data duplication
- Data transformation: converting categories into numerical data using label encoding or one-hot encoding as needed
- Data normalization or standardization: performed to ensure that features are on a uniform scale so that machine learning algorithms can learn optimally. Specifically for the permeability parameter (k), transformation is performed using a logarithmic equation, namely:

$$log(Permeability(mD) + 0.1) \quad (1)$$

- This transformation is used to address the very wide range of permeability values (skewed distribution) so that the data distribution becomes more stable and easier for the model to learn.

This process follows the approach of Chahar et al. (2022), which emphasizes the importance of handling noise, missing values, and scaling before entering the machine learning modeling stage.

Exploratory data analysis (EDA)

EDA was conducted to understand the patterns, correlations, and characteristics of workover data. The analysis included: 1.) Distribution of key parameters (porosity, permeability, Net Pay, EUR, Water Saturation); 2.) Pearson and Spearman correlations to examine linear and non -linear relationships between variables; 3.) Identification of dominant parameters that could potentially predict success; 4.) Analysis of historical well work patterns.

This pattern exploration approach follows the analytical research principles used by Maulana et al. (2025), which confirm that historical pattern exploration can provide an overview of the parameters that best determine well performance before the modeling stage is carried out.

Clustering using K-MEANS

After the data went through the cleaning and exploration stages, segmentation was performed using the K-Means clustering algorithm. The purpose of clustering was to group wells based on reservoir characteristic similarities so that the regression model could work on a more homogeneous dataset. The number of clusters was set at three based on the evaluation of the within-cluster sum of squares and field geological considerations, which showed that three groups could provide clear separation of reservoir characteristics. In addition, the clustering results were used to identify reservoir zones with greater potential, thereby serving as a basis for determining the location of new drilling wells.

The features used in clustering include XY coordinates and reservoir properties (porosity, permeability, Net Pay, and Water Saturation). The segmentation results produced two clusters with the following profiles: 1.) high reservoir quality cluster; 2.) low reservoir quality cluster.

To ensure that the clustering process would produce valid reservoir segmentation that could be used in the regression modeling stage, cluster quality testing was performed using the Silhouette Coefficient. This coefficient is used as the main indicator to assess whether the data have been grouped properly. The silhouette value measures two aspects simultaneously, namely the degree of closeness of the data to its own cluster (cohesion) and the degree of separation of a cluster from other clusters (separation). This division shows that groups of reservoirs share similar conditions and production responses, which can help improve the accuracy of the regression model used for EUR prediction. This stage follows an approach commonly used in research related to reservoir segmentation, as described by Liu et al. (2021), which utilized reservoir property parameters and spatial coordinates as the basis for cluster formation.

The clustering results are then used as additional input in the regression modeling process by including cluster labels as new variables (cluster features). The addition of these variables helps the

model recognize differences in characteristics between reservoir zones, thereby improving the model's ability to predict EUR values in a more segmented and accurate manner.

Regression modeling

The machine learning model was built using a supervised learning approach with the target being the EUR value. The regression algorithms evaluated in this study include; 1.) Random Forest Regression; 2.) Gradient Boosting Regression; 3.) Simple Regression.

These three algorithms were chosen because they are ensemble models widely used for oil and gas production prediction, and they can capture non -linear relationships between reservoir parameters (Chahar et al., 2022). The stages include:

- Split the data into train and test sets, 80% - 20%.
- Train the model using hyperparameter tuning to improve accuracy.
- Evaluate the model using:
- R-Squared
- RMSE (Root Mean Square Error)
- MAE (Mean Absolute Error)
- MSE (Mean Squared Error)
- Mean Signed Error (bias)
- Model interpretation, using feature importance or SHAP-based methods.

This approach refers to the workflow used in Chahar et al. (2022), Satriaji et al. (2024), and Ghodsi et al. (2025).

Estimated output

The methodology developed in this study is expected to produce several key outputs to support the process of evaluating drilling candidates in mature oil fields. First, it will yield a highly accurate EUR prediction model that provides more reliable estimates than conventional approaches. Second, this research is expected to identify the reservoir parameters that most influence EUR variation, thereby strengthening technical understanding of the factors that control drilling well performance in mature fields. In addition, this methodology is intended to provide an analytical basis for the development of a decision support system that can help engineers determine drilling priorities or infill drilling based on predicted EUR potential. Finally, the clustering and regression-based approach enables a comprehensive understanding of the relationship between multiple reservoir

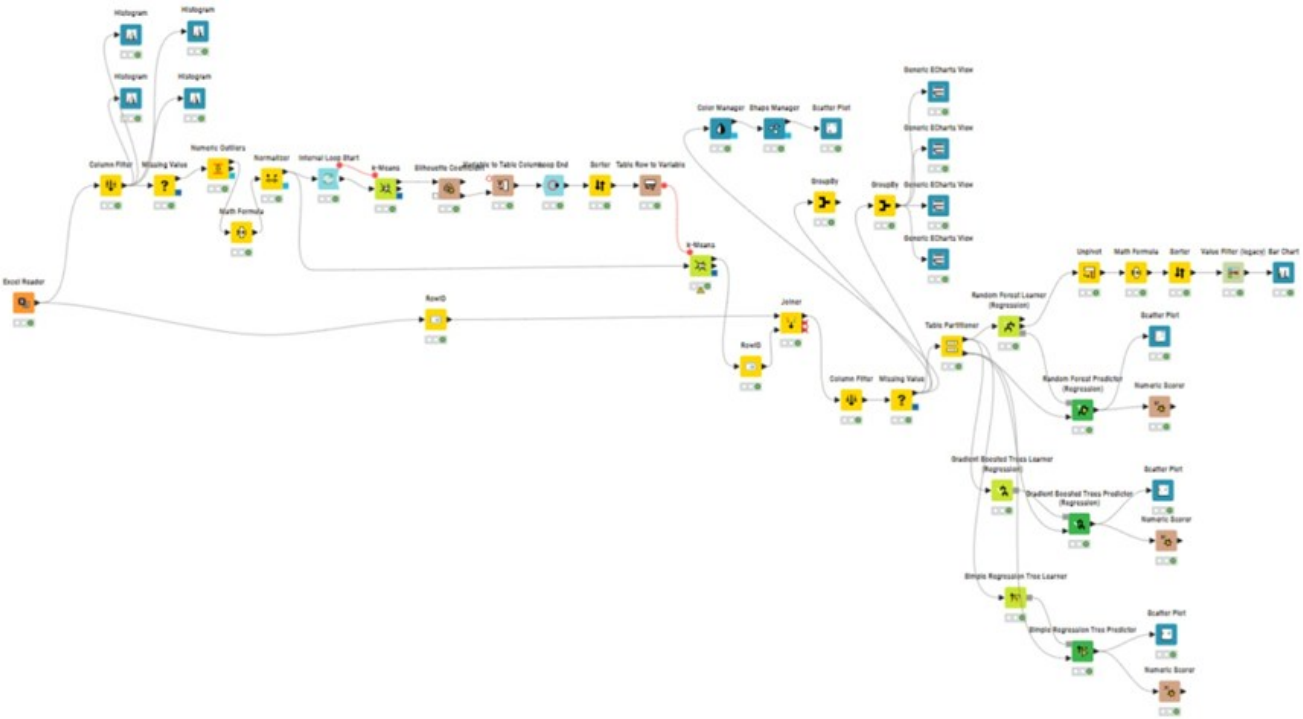


Figure 2. Modeling workflow in KNIME Analytics Platform.

parameters, both spatially and petrophysically, which can be used as an important reference for future field development studies.

RESULT AND DISCUSSION

The evaluation flow used in this study is shown in the figure below. The figure shows the initial part of the KNIME workflow used in this study. The workflow includes data merging, data cleaning, normalization, division of the dataset into training and testing data, and regression model building using the KNIME built-in node. This flow ensures that each modeling stage is carried out systematically and structurally.

Clustering K-MEANS result

The reservoir segmentation process using the K-Means algorithm produced three main clusters representing variations in reservoir quality in the old oil field. Cluster formation was based on a combination of petrophysical parameters of porosity, permeability, Net Pay, and Water Saturation, in addition to well location variables (X and Y coordinates). This combination of features allows the model to separate reservoir characteristics more comprehensively, in terms of both petrophysics and spatial distribution.

In summary, the characteristics of each cluster can be described as follows:

- Cluster 0: This cluster shows relatively low values of porosity, permeability, and Net Pay. These characteristics indicate a more heterogeneous reservoir zone, with limited fluid flow continuity and low productivity potential. This is reflected in the Silhouette Coefficient value of 0.429, which indicates that the separation boundary of this cluster is relatively weak and there is a higher level of overlap with other clusters.
- Cluster 1: This cluster has higher porosity, permeability, and Net Pay values than Cluster 0. These conditions indicate a more consistent reservoir zone with better productivity potential. The Silhouette Coefficient value of 0.73 indicates that the cluster structure is quite strong, with clear boundaries between clusters and a high level of homogeneity.

The Silhouette Coefficient values for each cluster are:

- Cluster 0: 0.429 → relatively weak separation; indicates high heterogeneity.
- Cluster 1: 0.73 → strong separation; indicates clear group boundaries and homogeneity.

These values indicate that the clustering structure in most of the data is well-formed, although one group (Cluster 0) shows higher similarity between data points, so its segmentation is not as strong as in the other clusters. The table 1 summarizes the range of values for the main parameters for each cluster.

The clustering results show that the two clusters have significantly different reservoir characteristics. This separation makes the dataset more homogeneous within each group, allowing for a more targeted EUR regression modeling process. With higher homogeneity, data variation within each cluster becomes more controlled, which improves model stability and EUR prediction accuracy.

Figure 3 shows the distribution of well positions based on subsurface coordinates (X and Y) with colors distinguishing each K-Means cluster. It can be seen that the two clusters tend to form groups in different areas, indicating a clear spatial pattern within the reservoir.

This pattern confirms that the formed segmentation is influenced not only by petrophysical parameters but also by the spatial distribution of wells. Thus, the clustering results

Table 1. Summary of parameters for each cluster.

Cluster	Cluster_0	Cluster_1
Porosity	0 – 0.4	0 – 0.37
Net Pay	0 – 1.139	0 – 1.399
Permeability	0 – 6.439	0 – 4.313
SW	0 – 0.88	0 – 0.76
Silhouette Coefficient	0.429	0.73

Figure 3. Spatial distribution of clustering results.

capture a combination of rock characteristics and well location proximity, resulting in a more geologically representative separation of reservoir zones. This spatial segmentation is an important basis for the EUR modeling process and the selection of new drilling locations, as clusters with better reservoir quality can be prioritized in exploration and infill drilling strategies.

Feature importance analysis in the model

Through feature importance analysis, an overview of the extent to which each reservoir parameter contributes to generating EUR predictions is obtained, allowing dominant factors to be clearly identified. Figure 4 shows the relative importance (feature importance) of each reservoir parameter in the EUR prediction model. Feature importance values are calculated based on the total contribution of each variable to the node separation process (total splits) in the Random Forest algorithm. The higher the total splits value of a feature, the greater the role of that feature in improving the model's ability to explain EUR variations.

From the sensitivity analysis results, several important findings were obtained:

- Net Pay emerged as the parameter with the highest feature importance value. This indicates that the thickness of the productive zone is a dominant factor in determining EUR potential, in line with the basic principle that the thicker the productive layer, the greater the volume of hydrocarbons that can be produced.
- Porosity ranks second, confirming that the fluid storage capacity within the rock is one of the main controllers of EUR. The higher the porosity, the greater the chance that the reservoir will store oil.
- Permeability ranks third. Although slightly below porosity and Net Pay, permeability still makes a significant contribution because it determines the ease of fluid flow to the production well. Variations in k directly affect reservoir deliverability and production rates.
- Water Saturation (Sw) has the lowest feature importance value compared to other parameters. However, this parameter remains relevant because high Water Saturation can reduce effective oil recovery, especially in mature fields with high water cuts.

Overall, these results indicate that the combination of layer thickness (Net Pay), rock

quality (porosity and permeability), and fluid saturation conditions are the main controllers in EUR estimation. Such findings are consistent with the basic concept of reservoir evaluation in old oil fields, where petrophysical heterogeneity is a critical factor in long-term production performance.

Regression model result

After the preprocessing and data segmentation stages, three regression models were applied to predict the EUR value, namely Random Forest Regression, Gradient Boosting , and Simple Regression Tree. Each model was trained using 80% of the data as a training set and tested using the remaining 20% as a testing set. Model performance was evaluated using several metrics, namely R^2, RMSE, MAE, and MSE. Based on the table 2, there are several key findings:

- Random Forest showed the best performance among the three models, with the highest R^2 value (0.677) and the lowest error values across all metrics (RMSE = 20.085; MAE = 14.021; MSE = 403.412). These results indicate that Random Forest is better able to capture the non-linear and complex relationships between reservoir parameters and EUR, thereby providing more stable and accurate predictions.
- Gradient Boosting came in second. This model performed quite well, with an R^2 value of 0.555, but its accuracy was still

Table 2. Comparison of regression model accuracy.

Metric	Random Forest	Gradient Boost	Simple Regression Tree
R^2	0.677	0.555	0.402
RMSE	20.085	23.682	27.346
MAE	14.021	14.956	17.33
MSE	403.412	556.097	747.821

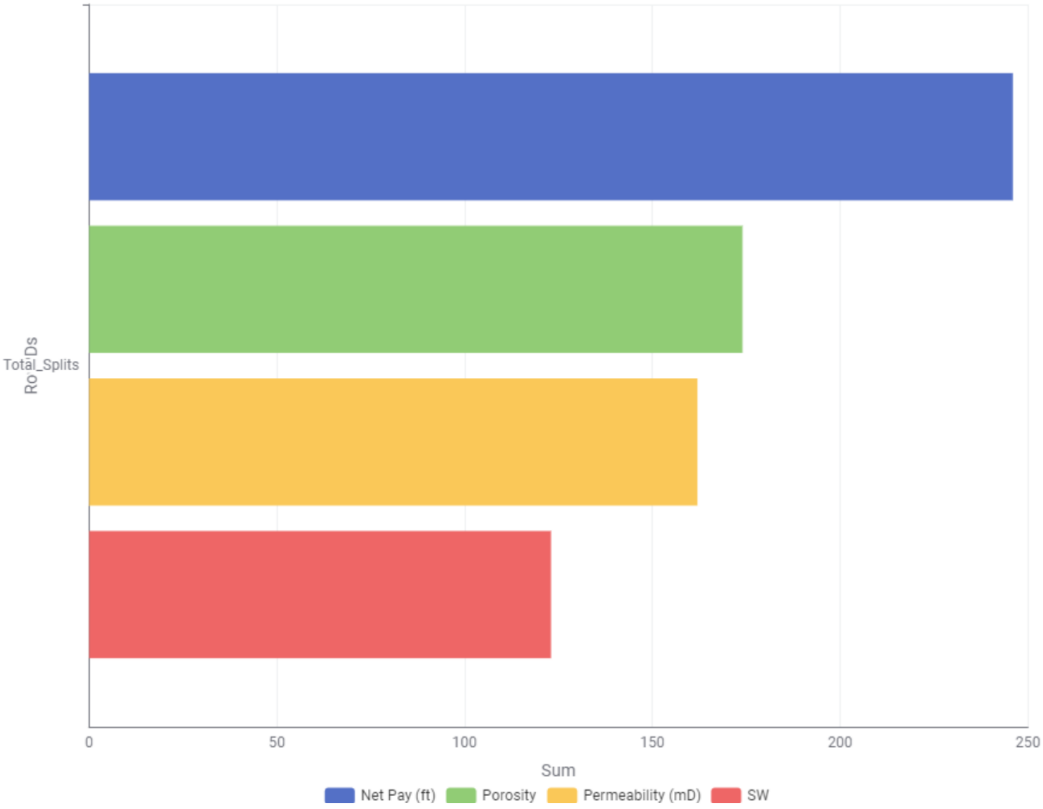


Figure 4. Sensitivity of reservoir feature importance parameter.

below that of Random Forest, as seen from its higher RMSE and MSE values. This shows that Gradient Boosting can learn fairly complex patterns, but its prediction stability is still lower than that of Random Forest in this dataset.

- Simple Regression Tree showed the lowest performance, with an R^2 value of only 0.402 and the largest error (RMSE = 27.346; MAE = 17.33; MSE = 747.821). This low performance is due to the simpler model structure and excessive sensitivity to data variation, making it less effective at capturing non-linear patterns and reservoir heterogeneity.

Overall, these results confirm that Random Forest is the most effective model for predicting EUR in old oil fields with high heterogeneity. Meanwhile, Gradient Boosting can serve as a fairly strong comparative model, while Simple Regression Tree acts as a baseline for assessing performance improvements from more complex models.

EUR prediction result

At this stage, the ability of the three regression models to predict the EUR value is assessed. A scatter plot is used to compare the actual EUR value with the prediction results from each model, revealing how well the prediction pattern follows the actual value. Figure 5 shows a comparison between actual EUR values and predicted results from three regression models, namely Simple Regression Tree, Gradient Boosting, and Random Forest. Each point on the graph represents an actual–predicted value pair (actual vs predicted). In general, the observed patterns can be interpreted as follows:

- Simple regression tree (SR)

  The SR model shows the most scattered points and has high variability relative to the diagonal line. Many points are far from the ideal line, especially in the medium- to high- EUR range. This pattern indicates that the model has not captured the non-linear relationship between reservoir variables, resulting in a relatively lower

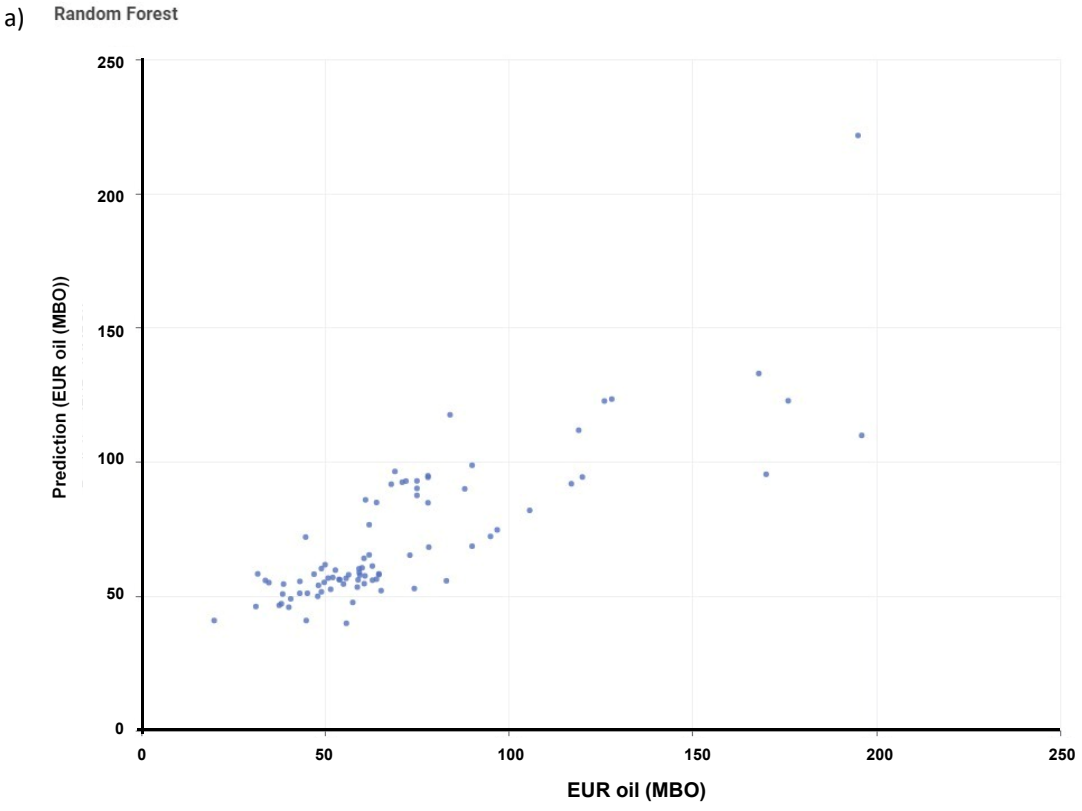
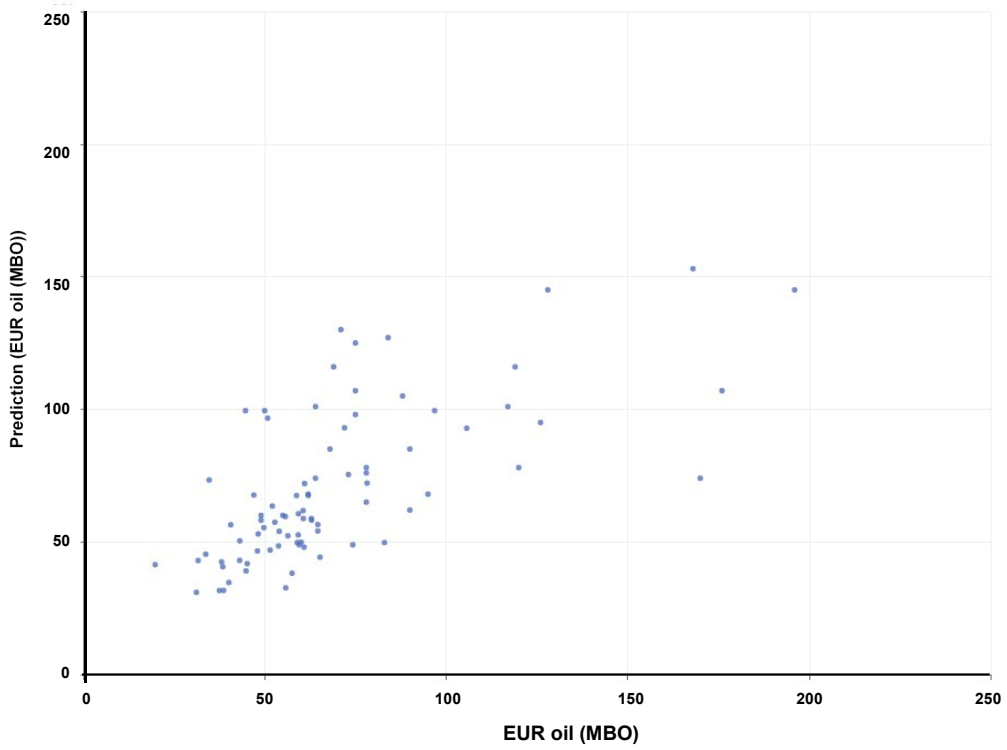


Figure 5. Visualization of EUR prediction accuracy across various models. (continued)

b) Simple Regression



c) XGBoost

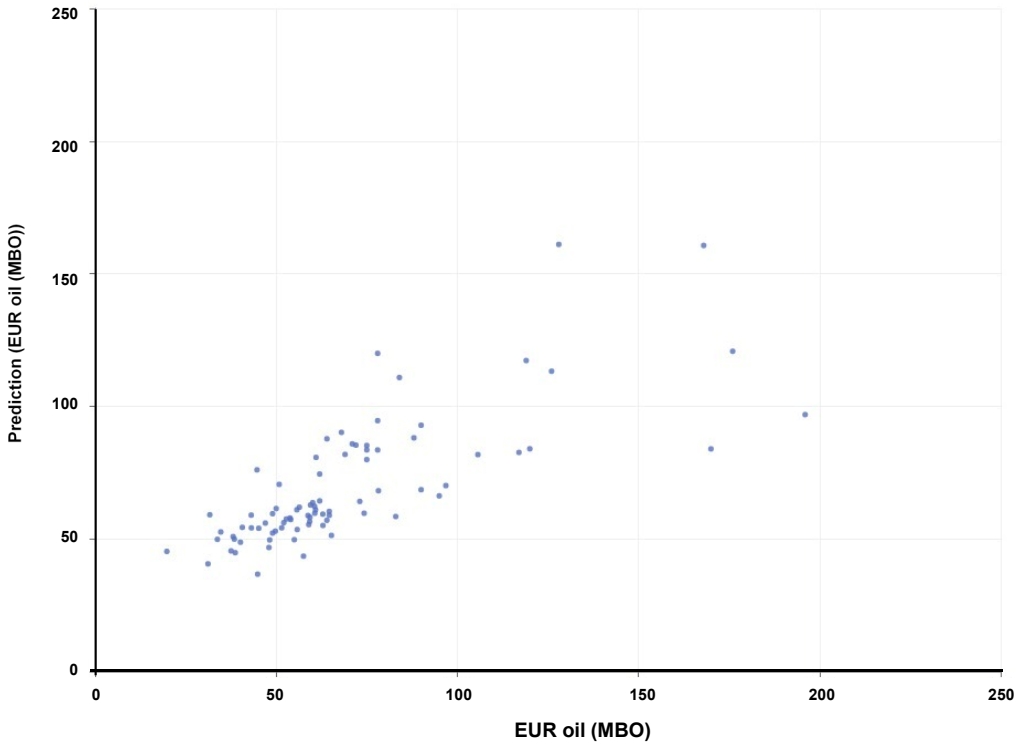


Figure 5. Visualization of EUR prediction accuracy across various models.

level of accuracy compared to the other two models.

- Gradient boosting regression

  The Gradient Boosting model provides predictions that are closer to the ideal line than SR. The distribution of points appears more centered, although some deviations are still visible in the EUR values at the top. Overall, this shows that Gradient Boosting is quite effective at learning non-linear relationships and can produce more stable predictions across the entire EUR range, although not as well as Random Forest.

- Random forest regression (RF)

  The RF model provides the closest and most consistent prediction results to the diagonal line. The distribution of points appears denser and more structured, and it does not show extreme outliers as in SR. This confirms that Random Forest has the best ability to learn complex reservoir patterns, especially in datasets with high petrophysical variability.

Overall, these results show that tree ensemble-based models, particularly Random Forest, are more suitable for modeling petrophysical variability and production characteristics in old oil fields that tend to be heterogeneous.

Discussion

The results show that integrating reservoir segmentation using K-Means and ensemble-based regression modeling improves the accuracy of EUR predictions. Segmentation using K-Means successfully separated well data into two main clusters based on similar petrophysical characteristics. This separation proved effective at reducing internal heterogeneity in the dataset, allowing the regression model to be trained on a more uniform data group, which is consistent with previous studies showing that clustering can improve model performance in heterogeneous reservoir systems (Rohmana et al., 2026).

At the modeling stage, the three regression algorithms performed differently. Random Forest consistently provided the best performance, as indicated by the highest R^2 value and the lowest error values (RMSE and MSE). The actual-prediction graph also showed that Random Forest had the closest distribution of points to the diagonal line, indicating high consistency and prediction accuracy. This finding aligns with previous research highlighting the robustness of Random Forest in handling complex non-linear relationships in reservoir data (Mahfudhoh & Welayaturromadhona 2026; Septiano et al., 2021).

Gradient Boosting ranks in the middle with fairly good accuracy, but still shows deviation at some prediction points. Meanwhile, Simple Regression Tree achieves the lowest accuracy due to its limitations in capturing non-linear relationships in reservoir data, as also observed in machine learning applications where simpler models tend to underperform under complex reservoir conditions (Zainuri et al., 2023).

In general, this study confirms that the combination of data segmentation using clustering and the use of ensemble learning models is highly effective for old oil field data. High petrophysical heterogeneity can be overcome through initial segmentation, while ensemble models such as Random Forest can better capture the complexity of inter-variable relationships. Thus, this approach can be used as a basis for supporting technical decision-making, including in the selection of drilling candidates or infill drilling in the future, in line with previous studies emphasizing the role of data-driven methods in supporting reservoir management and decision-making processes (Muhendra et al., 2025; Rizal et al., 2025).

Furthermore, the reliability of prediction results is highly dependent on how well the model accounts for operational variability and data conditions, as conventional forecasting methods are known to be sensitive to changes in production systems (Maurenza et al., 2023). In addition, data-driven approaches are a practical alternative to conventional reservoir simulation, particularly when dealing with complex datasets and the need for faster decision-making processes (Iskandar & Kurihara 2022).

CONCLUSION

Based on the results of the research and analysis, several main conclusions have been reached as follows: 1). K-Means clustering successfully divided

the well data into two reservoir groups with clearly different petrophysical characteristics. The Silhouette Coefficient values of 0.429 and 0.73 indicate that the cluster structure is quite good and can be used as a basis for segmentation prior to EUR modeling; 2). Random Forest is the best model for predicting EUR, as indicated by the highest R^2 value and lowest error rate compared to Gradient Boosting and Simple Regression Tree. The other two models tend to produce greater deviations and are less stable in predicting EUR; 3). The integrative workflow based on clustering and regression developed in this study has great potential as a decision support system for predicting EUR values in new wells. This approach helps engineers identify more prospective reservoir zones and supports a more targeted and data-driven field development strategy.

Based on the research findings, several recommendations that can be further developed are: 1). Develop clustering with additional parameters, such as Flow Zone Indicator (FZI), Hydraulic Flow Unit (HFU), or reservoir pressure data, so that the reservoir structure can be mapped more accurately and representatively; 2). Test additional regression models, such as LightGBM, CatBoost, or neural network-based models, to evaluate the potential for improving EUR prediction accuracy; 3. Integrate clustering results and EUR predictions into reservoir maps through spatial mapping. This spatial visualization can help identify high-potential drilling zones more intuitively; 4). Develop a drilling point recommendation workflow by combining EUR predictions, reservoir clustering results, and spatial well distribution. This approach has the potential to generate recommendations for new well locations that are optimal from both technical and economic standpoints.

ACKNOWLEDGMENT

The authors would like to thank SKK Migas for its support and data facilities, which enabled the successful completion of this research.

DECLARATIONS

Author contribution

Jeffier Winarta: Writing – Original Draft, Conceptualization, Methodology, Formal Analysis, Investigation, Data Curation, Visualization, Software, Validation. Amega Yasutra: Review & Editing, Supervision, Conceptualization, Methodology. Fajril Ambia: Review & Editing, Investigation, Resources, Data Curation, Validation.

Funding statement

None

Competing interest

All authors declare that there are no competing of interest.

The use of AI or AI-assisted technologies

None

GLOSSARY OF TERMS & SYMBOLS

Terms & Symbols	Definition	Unit
Net Pay	The thickness of reservoir rock that contains economically producible hydrocarbons	ft
k	Permeability	mD
Φ	Porosity	fraction or %
Sw	Water Saturation	fraction

REFERENCES

Candra, A. D., Rahalintar, P., Sulistiyono & Prabowo, U. N., (2024), Comparison of facies estimation of well log data using machine learning. Scientific Contributions Oil and Gas, 47(1), 21–30. https://doi.org/10.29017/SCOG.47.1.1593.

Chahar, J., Verma, J., Vyas, D., & Goyal, M., (2022), Data driven approach for hydrocarbon

production forecasting using machine learning techniques. Journal of Petroleum Science and Engineering, 217, 110757.

Iskandar, U. P., & Kurihara, M., (2022), Long short -term memory (LSTM) networks for forecasting reservoir performances in carbon capture, utilisation, and storage (CCUS) operations. Scientific Contributions Oil and Gas, 45(1), 35–50. https://doi.org/10.29017/SCOG.45.1.943

Liu, X., Sang, X., Chang, J., Zheng, Y., & Han, Y., (2021), The water supply association analysis method in Shenzhen based on k-means clustering discretization and apriori algorithm. PLOS ONE, 16(8), e0255684.

Mahfudhoh, S. A., & Welayaturromadhona., (2026), Machine learning-based prediction of formation, facies, porosity, and permeability in a carbonate reservoir. Scientific Contributions Oil and Gas, 49(1), 31–47. https:// doi.org/10.29017/scog.v49i1.1969.

Maulana, F., Yasutra, A., & Syihab, Z., (2025), Case study of heterogeneity index's effect on the successful workover based on the apriori algorithm. Scientific Contributions Oil and Gas (Lemigas), 48(1), 135–144. https:// doi.org/10.29017/scog.v48i1.1658.

Maurenza, F., Yasutra, A., & Tungkup, I. L., (2023), Production forecasting using Arps decline curve model with the effect of artificial lift installation. Scientific Contributions Oil and Gas, 46(1), 9–18. https://doi.org/10.29017/ SCOG.46.1.1310

Prayitno, S. H., Swadesi, B., Hariyadi, Nandiwardhana, D., Jayadianti, H., Widiyaningsih, I., Cahyaningtyas, N., & Imasuly, G. B., (2025), The application of machine learning in determining idle well reactivation candidates at PT. Pertamina EP. Scientific Contributions Oil and Gas, 48(2), 69–94. https://doi.org/10.29017/scog.v48i2.1657

Rafsanjani, R., Dahlia, A., Ambia, F., Rita, N., & Husbani, A., (2025), Comparative analysis of capacitance-resistance models and machine learning for CO_2-EOR forecasting. Scientific Contributions Oil and Gas, 48(4), 433–462. https://doi.org/10.29017/scog.v48i4.1930

Rizal, A. A., Ambia, F., Rita, N., & Herawati, I., (2025), Comparative study of capacitance resistance model and machine learning for polymer injection performance. Scientific Contributions Oil and Gas, 48(4), 251–279. https://doi.org/10.29017.scog.v48i4.1929

Rohmana, R. C., Ariadji, T., Yasutra, A., & Irawan, D., (2026), A conceptual framework for cross-basin reservoir characterization: Integrating unsupervised-supervised learning to address geological heterogeneity. Scientific Contributions Oil and Gas, 49(1), 375–393. https://doi.org/10.29017/scog.v49i1.2001

Satriaji, G., Wijaya, I. P., Pasaribu, H., Mambu, N., Adinda, P., Munawar, A., & Amar, R., (2024), Application of machine learning algorithm approach to enhance the workover program success ratio in Pertamina Hulu Sanga Sanga Field, Sanga Sanga Block of East Kalimantan, Indonesia. SPE-221349-MS, Society of Petroleum Engineers.

Septiano, J., Yasutra, A., & Rahmawati, S. D., (2021), Build of machine learning proxy model for prediction of wax deposition rate in two-phase flow. Scientific Contributions Oil and Gas, 45(1), 27–41. https://doi.org/10.29017/ SCOG.45.1.922.

Sugiyanto, & Utama, D. N., (2025), Estimation of well flowing bottomhole pressure (FBHP) using machine learning. Scientific Contributions Oil and Gas, 48(3), 37–51. https://doi.org/10.29017/ scog.v48i3.1851

Syahrial, E., & Purnomo, H., (2009), Peningkatan produksi minyak dengan injeksi CO2 pada Lapangan Minyak Tua Sangatta Kalimantan Timur. Lembaran Publikasi Lemigas, 43(2), 166–175. https://doi.org/10.29017/LPMGB.43.2.142.

Susantoro, T. M., Wikantika, K., Suliantara, S., Setiawan, H. L., Harto, A. B., & Sakti, A. D., (2023), Applying random forest to oil and gas exploration in Central Sumatra Basin Indonesia based on surface and subsurface data. Remote Sensing Applications: Society and Environment, 32, 101039.

Wardhana, S. G., Pakpahan, H. J., Simarmata, K., Pranowo, W., & Purba, H., (2021), Algoritma komputasi machine learning untuk aplikasi prediksi nilai total organic carbon (TOC). Lembaran Publikasi Minyak dan Gas Bumi, 55 (2), 75–87. https://doi.org/10.29017/ LPMGB.55.2.606.

Zainuri, A. P. P., Sinurat, P. D., Irawan, D., & Sasongko, H., (2023), Trap prevention in machine learning in prediction of petrophysical parameters: A case study in field X. Scientific Contributions Oil and Gas, 46(3), 115–127. https://doi.org/10.29017/SCOG.46.3.1586.