AN IMPLEMENTATION OF XGBOOST AND RANDOM FOREST ALGORITHM TO ESTIMATE EFFECTIVE POROSITY AND PERMEABILITY ON WELL LOG DATA AT FAJAR FIELD, SOUTH SUMATERA BASIN, INDONESIA

Ilham Diaz Rahmat Nugroho[1], Muhammad Destrayuda Trisna[2] , Sudarmaji[1*]
[1] Departement of Physics, Universitas Gadjah Mada, Yogyakarta, Indonesia
[2] PT. Pertamina Hulu Rokan, Jakarta, Indonesia
[*]ajisaroji@ugm.ac.id

Received 04-01-2024, Revised 21-04-2024, Accepted 13-09-2024,
Available Online 13-09-2024, Published Regularly October 2024

ABSTRACT

New approaches and methodologies have been developed to petrophysical analysis from well logs data using machine learning. Through this method, a machine learning algorithm is applied to predict the accuracy of the model on effective porosity (Øe) and permeability (K) which implemented using Random Forest and Xtreme Gradient Boosting (XGBoost) algorithm. The dataset used is obtained from well logs data that have been calculated petrophysical analysis. This study proposes the algorithm which is known to be effective in providing accurate predictions in a short time in estimating effective porosity and permeability. The results of the prediction model is optimized by GridSearchCV (GS), validated by the k-fold cross-validation, and evaluated using R2 score and Root Mean Square Error (RMSE). Model is applied to 5 research wells in Fajar Field of south Sumatra Basin, Indonesia with 4 variations of well training and well testing data split. The best evaluation results obtained with evaluation metrics were up to 0.90 (R2 score) and 0.01 (RMSE) for effective porosity and permeability by Random Forest, while evaluation metrics are 0.90 (R2 score) and under 0.68 (RMSE) for effective porosity and permeability by XGBoost. There is no decrease in accuracy until the last variation so that it can be concluded that these algorithm models can effectively estimate reservoir porosity and permeability in the field and contributed an alternative for the problem of many incomplete and dissimilar well logs data.

Keywords: Machine Learning; Petrophysical Analysis; Effective Porosity; Permeability; Random Forest; XGBoost

Cite this as: Nugroho, I. D. R., Trisna, M. D., & Sudarmaji. 2024. An Implementation Of Xgboost And Random Forest Algorithm To Estimate Effective Porosity And Permeability On Well Log Data At Fajar Field, South Sumatera Basin, Indonesia. *IJAP: Indonesian Journal of Applied Physics, 14*(2), 271-280. doi: https://doi.org/10.13057/ijap.v14i2.82901

INTRODUCTION

The formatter will need to create these components, incorporating the applicable criteria that follow. As time progresses, the older the fields that have been discovered before, so that production has entered a decline phase, it is necessary to search for new zones below the surface that allow the prospect of producing oil and gas. Therefore, there is a need for an alternative to review existing and old fields through the study of petrophysical data. Petrophysical analysis in the reservoir characterization study was conducted to analyze geophysical data of well logs and petrophysical parameters such as effective porosity. The analytical method known as model-based deterministic analysis is then converted into a probabilistic analysis by considering the random nature of the data source, called machine learning. This analytical method can play a role in bypassing this, so that machine learning algorithms can learn from well log databases and petrophysical analysis calculations [1].

The existence of a machine learning approach in determining petrophysical parameters is able to estimate the predicted value of accuracy in interpreting reservoir characterization by maximizing processing time compared to conventional methods. Therefore, this research was conducted in order to obtain machine learning prediction results that can estimate petrophysical parameters well. The growth in volume, variety and complexity of porosity data provides an opportunity for researchers to be able to comprehensively analyze data quickly and accurately [2]. Generally, to obtain porosity values through the drilling and coring process by direct analysis and testing, this method is less efficient, and is limited in data that is not too large. In general, this technology develops a machine that can learn by itself without direction from the user. This method uses a computational model to gain knowledge from the model's experience studying a data set. The goal of this technology is to increase productivity by automating time-consuming tasks. Technological advances from machine learning are driven by the development of algorithms and methodologies that are supported by the availability of large-scale data (big data) and low-cost computing. One model that has proven to be effective for producing accurate predictions with fast training time is XGBoost and best accuracy is Random Forest [3]. Fajar field is an oil field which is geologically located in the Air Benakat Formation, South Sumatra Basin [4].

METHOD

Data Collection and Preprocessing

In this study, the data consists of 5 types of logs in each well log data (GR, NPHI, RHOB, ILD, and VSH). The data that is input to the machine learning model is first applied to the one hot encoding process from the skit-learn model. Then before applying the machine learning model, it is also first analyzed the correlation of each feature between the petrophysical parameters and the Pearson coefficient parameters. The model is applied with 5 variations of the separation between the training and test data as shown in Table 1. The data for training and testing come from 5 wells with the name: ID-04, ID-01, ID-02, ID-03, ID-05. Each wells have GR, NPHI, RHOB, ILD, and VSH log as input feature. This variation is made to find the minimum percentage of the amount of data in the entire input sample that must be trained on machine learning to get optimal results when tested using other data that was not previously trained or a blind test.

Table 1. The variation of sample data distribution based on training and testing data.

<table>
<tr><th>Variantion</th><th>ID-04</th><th>ID-01</th><th>ID-02</th><th>ID-03</th><th>ID-05</th><th>Traning</th><th>Testing</th></tr>
<tr><td>var 1</td><td><b>25%</b></td><td colspan="4">75%</td><td>4920</td><td>24632</td></tr>
<tr><td>var 2</td><td colspan="2"><b>48%</b></td><td colspan="3">52%</td><td>4481</td><td>20603</td></tr>
<tr><td>var 3</td><td colspan="3"><b>71%</b></td><td colspan="2">29%</td><td>11468</td><td>17715</td></tr>
<tr><td>var 4</td><td colspan="4"><b>80%</b></td><td>20%</td><td>17399</td><td>13616</td></tr>
<tr><td><b>n_data</b></td><td>4920</td><td>4687</td><td>4826</td><td>7189</td><td>2285</td><td></td><td></td></tr>
</table>

Then the stage of evaluating the model's ability to estimate petrophysical parameters was also carried out using several parameters, namely the R^2 score metric and the Root Mean Square Error (RMSE). This metric evaluation is used to determine which performance is the best of the three algorithms used in this study.

Random Forest Algorithm

It is an algorithm that was first introducesd by Tin Kam Ho in 1995. This algorithm is an ensemble construction using bagging to produce trees with high correlation. This algorithm consists of a combination of CART (Classification and Regression Tree) with randomization in sample data and predictor data (bootstrap samples). The description of the Random Forest algorithm is simplified in Table 2.

Table 2. Random Forest Algorithm

Algorithm. Random Forest for Regression Model

```
1. For $b$ = 1 to B:
   a. Draw a boostrap sample Z* of size N from traning data.
   b. Grow a random-forest tree $T_b$ to the boostrapped data, by recursively repeating the
      following steps for terminal node of the tree, until the minimum node size $n_{min}$ is
      reached.
         i. Select $m$ variables at random forest from the $p$ variables.
        ii. Pick the best variable/split-point among the $m$.
       iii. Split the node into two daughter nodes.
2. Output the ensemble of tress $\{T_b\}_1^B$
```

XGBoost Algorithm

The basic idea of boosting in the XGBoost algorithm is to initiate a simple CART with low accuracy which is iterated repeatedly with a model that evaluates the previous error to form a more accurate model. In evaluating the error in each tree structure, this algorithm derives the gradient of the loss function which can be defined by the following equation.

$$g_i = \partial_{y_i^{(t-1)}} l\left(y_i, p_i^{(t-1)}\right);\ G_j = \sum i \in I_j g_i \qquad (1)$$

$$h_i = \partial^2_{y_i^{(t-1)}} l\left(y_i, p_i^{(t-1)}\right);\ H_j = \sum i \in I_j h_i \qquad (2)$$

Where g_i and h_i are the gradients obtained from the first and second derivatives based on the error of the previous structure, and G_j and H_j are the gradient values in each j-th leaf by adding the gradient in each set of index data points. y_i is the predictor of each sample and p_i is the predicted result. The best tree structure that has a minimum objective function value which is generally divided into 2 forms (training loss and regularization).

$$f_{obj} = L + \Omega \qquad (3)$$

$$L = \sum_{i} (y_i, p_i)^2 \tag{4}$$

$$\Omega = \gamma \mathrm{T} + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \tag{5}$$

Where L is the training loss which is one measure of the model's performance in the training data, generally using the mean squared error metrics. And is the regularization term, which functions to control the complexity of the model to avoid overfitting, where are the gamma parameter, the lambda parameter (for the regression model using reg_lambda), w is the weight scores, and T is the number of leaves.

The tree boosting algorithm uses the approach of adding a new tree in each iteration which generally utilizes the Taylor expansion function with a 2nd order loss function. So that the simplification of the objective function with the best error reduction that can be obtained in each tree structure can be defined by equation 6 and the XGBoost algorithm flow in Table 2 below.

$$f_{obj}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma \mathrm{T} \tag{6}$$

Table 3. XGBoost Algorithm

```
Algorithm. XGBoost for Regression Model
  1. For i = 1 to N:
     a. Initial model prediction for the entire sample data, then apply gradient boosting with the
        flow:
          i. Form a simple tree from training data.
         ii. Break down each ensemble with gradients gi and hi
        iii. Calculate the gradient for each leaf.
         iv. Calculate the weight scores in each leaf.
     b. Get the next prediction of the whole leaf.
     c. Evaluate tree structure by reducing the objective function if an error is still high repeat to
        stage (i).
  2. The best model is represented by a CART with a minimum objective function.
```

RESULTS AND DISCUSSION

Tunning Hyperparameter

The XGBoost and random forest algorithm has a variety of hyperparameters that can be initialized before studying the data. The determination of these hyperparameters is categorized in the stage of model development. The purpose of initiating this hyperparameter is to adjust the model made based on the variation of the data studied by the model, which in this study is the well log data. This adjustment can improve the model's performance in estimating effective porosity. Generally, determining hyperparameters is done manually by understanding the role of each hyperparameter and determining its own value. One solution to automatically initialize hyperparameters is to use the GridSearchCV (GS) module. This module evaluates each selection of the grid / grid of hyperparameter values, then determines the value that has the best performance. The results of several XGBoost algorithm hyperparameters are shown in Table 4.

Table 4. Hyperparameter value after tunning using GridSearchCV (GS)

Hyperparameter	Default	GS-XGBoost	GS-Ramdom Forest
n_estimators	100	60	20
max_depth	6	5	10
reg_lambda	1	0.1	0.1
gamma	0	0	-
tree_method	auto	approx	-

Performance Algorithm in All Variantions

The performace evaluation stage of the model's ability to estimate effective porosity using the R^2 score metric and the Root Mean Square Error (RMSE) as error metric. In this study, the evaluation is applied with variations in the number of wells that have been divided between well training and testing data, namely there are 4 variations shown in Table 1. These two indicators are based on the following equation.

$$R^2(y,p) = 1 - \frac{\sum_{i=1}^{n}(y_i - p_i)^2}{\sum_{i=1}^{n}(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i)^2} \tag{7}$$

$$e_{rmse}(y,p) = \sqrt{\frac{1}{n}\sum_{i=0}^{n-1}(y_i - p_i)^2} \tag{8}$$

In the metrics analysis, if the R^2 score value shows the maximum value, it means that the prediction model built is getting better, otherwise if the value is getting smaller, the prediction model built is not good. In RMSE, if the RMSE value is maximum then the prediction model is not good, and vice versa if the RMSE value is getting smaller than the prediction model is getting better. The evaluation metric for all data variations is shown in Table 5 for effective porosity prediction and Table 6 for permeability prediction.

Table 5. Evaluation metrics of all data variations in effective porosity prediction

Variation			Metrics	RF	XGB
Var 1	1 Train	4 Test	R^2 score	0.82	0.59
n_data	4920	18987	RMSE	0.03	0.05
Var 2	2 Train	3 Test	R^2 score	0.97	0.95
n_data	9670	14300	RMSE	0.01	0.02
Var 3	3 Train	2 Test	R^2 score	0.98	0.96
n data	14433	9474	RMSE	0.56	0.01
Var 4	4 Train	1 Test	R^2 score	0.99	0.96
n data	21622	2285	RMSE	0.01	0.01

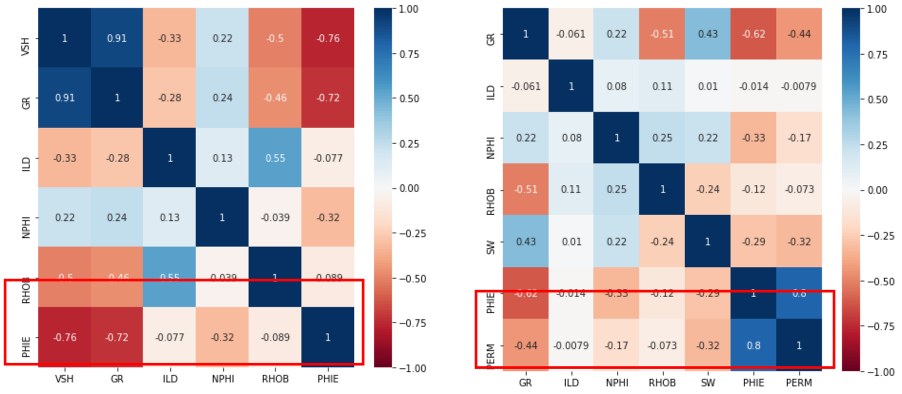
Table 6. Evaluation metrics of all data variations in permeability prediction

Variation			Metrics	RF	XGB
Var 1	1 Train	4 Test	R^2 score	0.01	0.01
n_data	4920	18987	RMSE	114	114
Var 2	2 Train	3 Test	R^2 score	0.01	0.01
n_data	9670	14300	RMSE	135	135
Var 3	3 Train	2 Test	R^2 score	0.98	0.55
n data	14433	9474	RMSE	0.56	3.37
Var 4	4 Train	1 Test	R^2 score	0.99	0.98
n data	21622	2285	RMSE	0.37	0.68

Based on Table 5 and Table 6, the algorithm with the best performance is the Random Forest algorithm. The variation used to get the best performance is variation 4 with 1 well testing data and 4 well training data used.

Correlation of Well Log Data

Each petrophysical parameter has a different ratio to the number and type of logs used. To correlate well log data, Pearson product moment correlation (PPMC) is used in this study. The Pearson correlation coefficient value aims to represent how well the relationship between features can be learned by machine learning models. In this correlation, the correlation coefficient (r) is obtained and produces three values that can be interpreted as a correlation with a value of 0 being a white column, a value of +1 being blue, and a value of -1 being red.



(a)

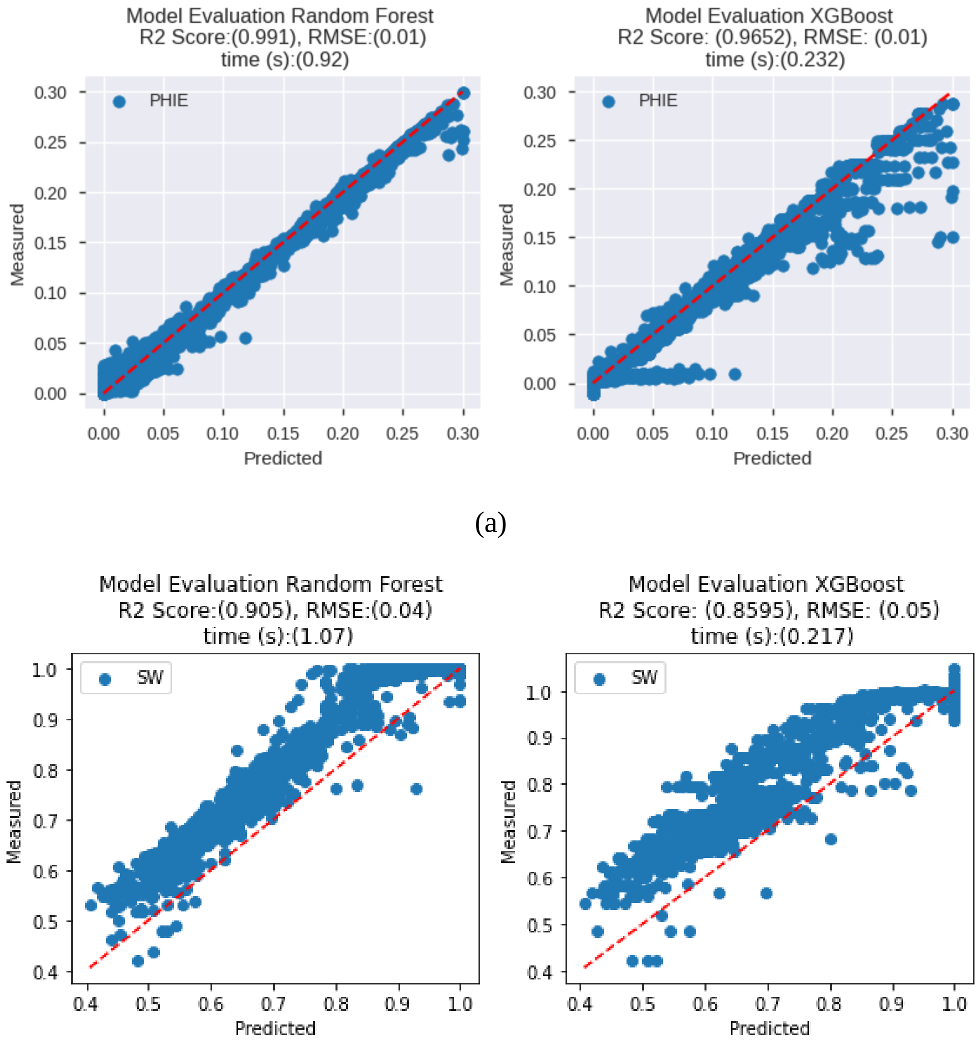
(b)

Figure 1. Correlation Matrix between features (a) effective porosity and (b) permeability

In Figure 1 it is shown that the correlation with good results should stay away from white or as much as possible not resemble white because the value of 0, this indicates a value with a very weak correlation. Conversely, the best correlation should be red, blue, or close to these two colors. This means that red is a strong and unidirectional correlation value, while blue is a correlation value that is as strong as red but only in the opposite direction.

The Proper Algorithm Model Performance

Qualitatively, the performance of the model can be analyzed by comparing the actual effective porosity or permeability data with the predicted results using the distribution and blind well data display. For variations with the best model performance (4 training wells and 1 test well).



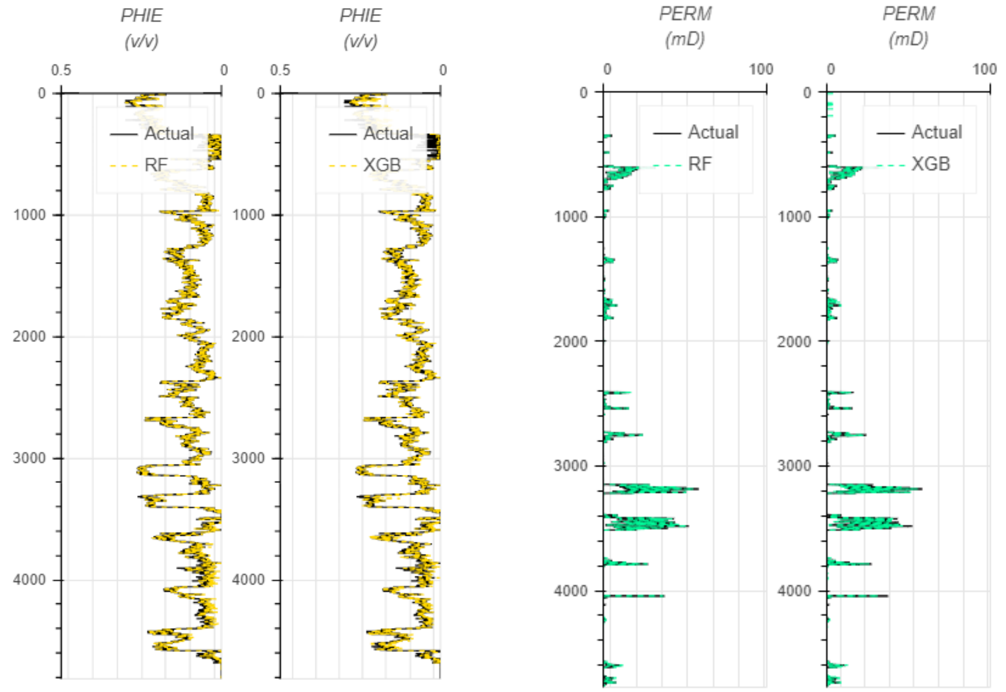
(a)

(b)

Figure 2. Predicted-actual data distribution for (a) effective porosity and (b) permeability

Good prediction results will be distributed on the red line which shows where the prediction results are very close to the value of the actual data (Figure 2). The evaluation results of the two algorithms show that the distribution of the set of effective porosity values (Øe) is well-satisfied and in line with the estimation line. Meanwhile, the distribution of the collection of permeability (K) values is distributed parallel to the estimation line. However, there are some data that spread apart in the 40-60 value range. Model evaluation in both algorithms has obtained a large accuracy above 90%

but needs to be reviewed in terms of error value. The error value obtained from the XGBoost algorithm is 0.68 and the RF algorithm has the smallest value of 0.37.

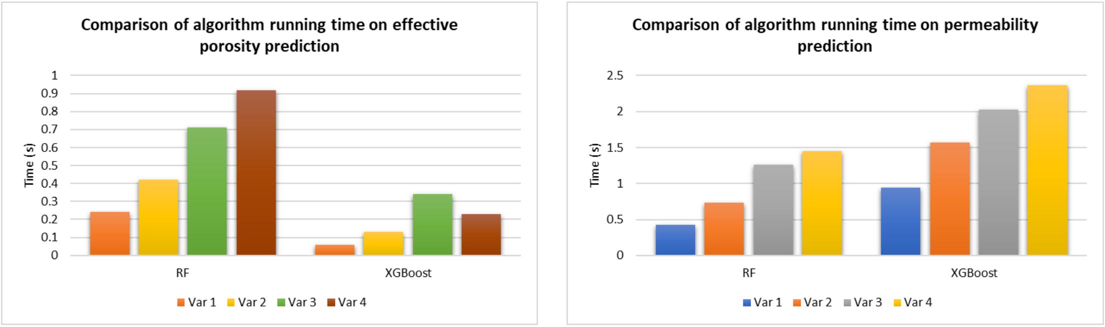


(a)

(b)

Figure 3. Blind Well test on (a) effective porosity and (b) permeability prediction

This result is also in accordance with the appearance of the blind well test (Picture 3), which visualizes the prediction results of the two algorithms for the yellow curve represents the prediction results of effective porosity (Øe) and the black curve is the actual well log value, while the green curve represents the prediction results of permeability. The higher the accuracy value, the yellow and green curves will be plotted with the same value as the black curve.



(a)

(b)

Figure 4. Comparison of algorithm running time on (a) effective porosity and (b) permeability prediction

This means that the prediction is consistent with the actual data, or the prediction results are accurate. The running time results show that the XGBoost algorithm is the most effective in processing effective porosity prediction data, while for permeability prediction, the Random Forest algorithm is shown in Figure 4.

CONCLUSION

The XGBoost and Random Forest algorithm machine learning model using the GridSearchCV module can be used to estimate the effective porosity and permeability with an accuracy up to 89% and a relatively fast time of under 2.5 seconds. This is the best result of 4 variations of training and test data. An increase in the amount of training data results in an increase in model performance. All variations of training and testing did not show any indication of overfitting. This model can estimate porosity values in the range of 0.05 - 0.25 very well but the model still cannot estimate all data with porosity values above 0.30. Then, the model also can estimate permeability in the range of 0.4-0.8 very well. The best running time for effective porosity prediction is XGBoost and permeability with Random Forest algorithm.

ACKNOWLEDGMENTS

We would like to thank to PT. Pertamina Hulu Rokan and Geophysics Laboratory of Faculty of Mathematics and Natural Sciences of Gadjah Mada University for data and facility support.

REFERENCES

1 Ahmad, M. W., Reynolds, J. dan Rezgui, Y., 2018, Predictive modelling for solar thermal energy systems: A comparison of support vector regression, Random Forest, extra trees and regression trees, Journal of cleaner production, 203, 810-821

2 Al-Mudhafar, W. J., 2020, Integrating electrofacies and well logging data into regression and machine learning approaches for improved permeability estimation in a carbonate reservoir in a giant southern iraqi oil field, In Offshore Technology Conference, OnePetro.

3 Asquith, G. dan Gibson, C., 1982, Basic Well Log Analysis for Geologist, The American Associtaion of Petroleum Geologists, Tulsa, Oklahoma.

4 Bishop, M. G., 2001, South Sumatra Basin Province, Indonesia: The Lahat/ Talang Akar Cenozoic Total Petroleum System, USGS Open file report, 99-50-S.

5 Chen, T. dan Guestrin, C., 2016, Xgboost: A scalable tree boosting system, In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, 785-794.

6 Erofeev, A., Orlov, D., Ryzhov, A. dan Koroteev, D., 2019, Prediction of porosity and permeability alteration based on machine learning algorithms, Transport in Porous Media, 128(2), 677-700.

7 Hadavimoghaddam, F., Ostadhassan, M., Sadri, M. A., Bondarenko, T., Chebyshev, I. dan Semnani, A., 2021, Prediction of Water Saturation from Well Log Data by Machine Learning Algorithms: Boosting and Super Learner, Journal of Marine Science and Engineering, 9(6), 666

8 Ma, Y. Z., 2019, Introduction to Geoscience Data Analytics Using Machine Learning, In Quantitative Geosciences: Data Analytics, Geostatistics, Reservoir Characterization and Modeling, 151-171.

9 Moghadasi, L., Ranaee, E., Inzoli, F. dan Guadagnini, A., 2018, Petrophysical well log analysis through intelligent methods, SPE Bergen One Day Seminar, OnePetro.

10 Nugroho, I. d., 2021. Estimation of Petrophysical Parameter from Well Log Data Using Random Forest, XGBoost, and Support Vector Regression (SVR) Algorithm Approach in Kenali Asam (KAS) Field of Sub-Jambi Basin, Jambi. Thesis. Gadjah Mada University.

11 Pandey, Y. N., Kainkaryam, S., Saputelli, L., Rastogi, A. dan Bhattacharya, S., 2020, Machine Learning in the Oil and Gas Industry, New York.

12 Ren, Q., Han, S., Li, M., 2019, Tectonic discrimination of olivine in basalt using data mining techniques based on major elements: a comparative study from multiple perspectives, Big Earth Data, 8-25.

13 Samuel, A., 1959, Some Studies in Machine Learning Using the Game of Checkers, IBM Journal of Research and Development, 3(3), 210-229.

14 Shirangi, Mehrdad G., Durlofsky. dan Louis J., 2016, A general method to select representative models for decision making and optimization under uncertainty, Computers and Geosciences, 96, 109-123.

15 Timur, A., 1968, an Investigation of Permeability, Porosity, and Residual Water Saturation Relationships for Sandstone Reservoirs, The Log Analyst

16 Yadav, S. dan Shukla, S., 2016, Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification, In 2016 IEEE 6th International conference on advanced computing (IACC), 78-83, IEEE.