Available online: http://journal.uir.ac.id/index.php/JEEE/index

Journal of Earth Energy Engineering

Publisher: Universitas Islam Riau (UIR) Press

ROP Prediction with Random Forest, Gradient Boosting, and Support Vector Machine; a Case Study

G. R. Darmawan[1*], Dedy Irawan[2]
[1]Department of Petroleum Engineering, Faculty of Design & Technology, Bandung Institute of Science Technology, Kota Deltamas CBD, Ganesha Boulevard, Cikarang Pusat, Bekas, Jawa Barat. Indonesia.
[2]Department of Petroleum Engineering, Faculty of Mining & Petroleum Engineering, Bandung Institute of Technology, jl. Ganesha 10, Bandung, Jawa Barat. Indonesia.
*Corresponding Author: ganesharinkudarmawan@gmail.com

Article History:

Received: September 24, 2021
Receive in Revised Form: March 21, 2022
Accepted: April 4, 2022

Keywords:

ROP prediction, supervised machine learning, drilling parameters

Abstract

Optimum drilling penetration rate, known as the rate of penetration (ROP) has played a big role in drilling operations. Planning the well ROP always becomes a challenge for drilling engineers to calculate the drilling time needed for the section. Optimum ROP is achieved when the time to drill the section is as planned. Many empirical approaches were develop to model the ROP based on the drilling parameters, and might not always match the actual ROP. In some cases, the actual ROP was slower than planned, which may increase the drilling cost, which needs to be avoided. Hence, some approaches using artificial intelligence (AI), and supervised machine learning have been develop to overcome it. Supervised machine learning is used to developed a ROP model and ROP prediction for one of the development fields, based only on two wells drilling parameters data. The model was trained using Gradient Boosting, Random Forest, and Support Vector Machine. Drilling parameter test data then is used to validate the model. The model of Random Forest shows a good or promising result with R^2 of 0.90, Gradient Boosting shows R^2 of 0.86, and Support Vector Machine with R^2 0.72. Based on the models generated, the Random Forest has shown a good trend which could be used for modeling ROP in the future development wells.

INTRODUCTION

Why need to optimize ROP? Over the years, the topic on how to optimize ROP always became a big discussion and research. Prediction ROP has always been of fundamental interest to the drilling industry (Hegde, Chiranth, Wallace, Scott, 2015). Most of well-drillingcost is not product cost dependent, but time-dependent, therefore the main goal of drilling optimization is to reduce the total drilling time, as well as selecting optimum drilling parameters prior drilling (Barbosa et al., 2019). Mitchell & Miska, (2011); Barbosa et al., (2019) stated that understanding how the drilling parameters really affect the ROP is an open question in drilling engineering. The idea is, how to maximize the ROP in the field, hence could significantly reduce the drilling time which led to optimizing the drilling budget. ROP is related to the speed with which drilling is performed and maximizing it, is one of critical factors affecting the commercial success of drilling operation (Chandrasekaran & Govindarajan, 2019).

Barbosa et al., (2019) classify the three possible approaches for ROP predictions, such as traditional models, statistical models, and machine learning models. Many researchers try to find a simple relation between ROP and rock properties because the majority of reservoir mechanical properties can be inferred from well logs (Shi et al., 2016). Traditional models are those ROP models which try to establish a mathematical equation among the drilling parameters (Shi et al., 2016 ; Barbosa et al., 2019). Singh et al., (2019) stated that determining a correlation or linear relationship among the drilling parameters with ROP is very difficult. Statistical models have some similarities with the traditional models, necessity of preselection a

DOI: 10.25299/jeee.2022.7772
eISSN: 2540-9352 pISSN: 2301-8097

Journal of Earth Energy Engineering
Vol. 11 No. 1, March 2022, pp 52-59

model for ROP as a function of drilling parameters, and with main difference was the statistical models did not model the physics of drill bit mechanism and the formation and bit interactions (Barbosa et al., 2019). Machine learning models able to learn complex patterns during the training (or learning) phase, without having to specify a ROP models, afterwards, the trained model can make predictions given novel inputs (Barbosa et al., 2019).

What can we do? During drilling operations, there are so many data gathered and captured, which hardly used to evaluate the drilling performance for the next campaign, or there is no optimization evaluation, mostly engineering planning based on the average ROP or the best ROP for the next well. Basically, there are many data that can be used to predict the optimized ROP based on the previous drilling data. Since there are many wells located in a field, some are close to one to another, data collection from previous wells become cruisial to know the important impact in drilling cost reduction (Shi et al., 2016). The goal is to have maximum ROP to drill the well to save operating time and eventually reduce the drilling cost, with artificial intelligent, machine learning, etc.

Why use supervised machine learning? Since ROP is the main target to be optimized, then the use of supervised machine learning become essential. A supervised machine learning model for ROP prediction was developed that is efficient for use with real data (Singh et al., 2019), supervised machine learning provides a target from series of training dataset with sets of predictor features (Noshi & Schubert, 2018). (Hegde, Chiranth, Wallace, Scott, 2015); Barbosa et al., (2019) compared traditional models of ROP calculations with machine learning techniques concluded that a higher of accuracy could be achieved in ROP prediction with intelligent techniques. Most of the framework to obtain the ROP mostly with a regression problem solving, hence for the ROP prediction should lies in the supervised machine learning.

The most common inputs for the ROP predictions in machine learning were analyzed by Barbosa et al., (2019), from 43 reviewed works. The top six inputs are weight on bit (WOB), rotation (RPM), depth, flow rate, mud weight, and bit diameter. These drilling parameters, some have the similarities to the parameters used in traditional model such as WOB and RPM. Barbosa et al., (2019) mentioned the use of drilling fluid and hydraulics play a role in drilling progress, where some works include flow rate (GPM) and the mud weigth as additional feature. Thus, this research will use additional drilling data such as mud weight, PV (plastic viscosity) and YP (yield point) from traditional data. Since drilling progress is relate to hydraulics, some papers have drilling parameters inputs for mud weight, flow rate and standpipe pressure. In general ROP optimization involves the adjustment of WOB and RPM for efficient drilling, but there are several other parameters to solve the ROP relationship (Mantha & Samuel, 2016).

The data prepared for training based on two wells that has been drilled in the same cluster of the field. Both drilling parameters data will be use, such as:

- Depth (m)
- Rotation (RPM-rotation per minute). Including RPMM (motor rotation, sliding) and RPMT (total rotation).
- Torsion (Klbs-ft)
- Weight on bit (Klbs)
- Standpipe pressure (psi)
- Flow rate (GPM)
- Bit diameter (inch)
- Plastic viscosity (cp)
- Yield point ($lbs.ft^2$)
- Mud weight (ppg)

The drilling parameters used is mostly surface drilling parameters with three drilling fluids parameters such as plastic viscosity, yield point and mud weight. No lithology data available for this project.

Supervised Machine Learning

Singh et al., (2019), in supervised machine learning the goal is to approximate the mapping function so well that that you have new input data you can predict the output variables for that data. Fernandes et al., (2018)had mentioned the ensemble learning family algorithms build a model by training several relatively simple base models and then combine them to create a more predictive model, the most well-known ensemble learning algorithms use bootstrap aggregation, known as Bagging, Random Forest, and Gradient Boosting. Fernandes et al., (2018)) stated that decision tress known as regression trees, are regression methods that consist of partitioning the input parameters spare into distinct and non-overlapping regions following a set of if-then rules, which identify regions that have homogeneous response to the predictor and fitted in the regression framework. The uses of decision trees as a regression technique have advantage of

the splitting rules represent an intuitive and very interpretable way to visualize results. Hegde, Chiranth, Wallace, Scott, (2015) stated the main difference between Bagging and Random Forrest is that random subsets of predictors are used to build trees in Random Forest, which help in de-correlating trees.

Random Forest is an ensemble of decision trees, or it can be thought of as a forest of decision tress which make this model is more accurate than a decision tree because much more knowledge is incorporated from many predictions Belyadi & Haghighat, (2021). For regression problems, Random Forest uses the average of decision tress or final prediction Höhn et al., (2020). Random Forest are highly effective when dealing with noisy and large multi-attribute data (Noshi & Schubert, 2018). Gradient Boosting is an ensemble that will train many models sequent by placing more weights on instances with erroneous predictions, and gradually minimize a loss function ((Belyadi & Haghighat, 2021). Gradient Boosting is similar to Random Forest, is based on combination a larger set of different weaker learner models to reach a new combined model with a significantly higher prediction accuracy, however the difference that the tree in Gradient Boosting is not independent of each other but incrementally improve the decision trees (Höhn et al., 2020). SVM algorithm is set to choose a hyperplane that splits the two classes, and when the classes are linearly classifiable, a plane is chosen which is on the exact center between two classes. SVM perform effectively with high dimensional data but can function poorly with noisy data and can require high processing time (Noshi & Schubert, 2018). Random Forest, Gradient Boosting and SVM could be used for both regression and classification problems (Belyadi & Haghighat, 2021).

This paper will discuss the supervised machine learning (Random Forest, Gradient Boosting and Support Vector Machine), how to cleanse the data, training the data set, building the model and test the data test to know the model accuracy and capabilities (Belyadi & Haghighat, 2021) using Orange Data Mining (Demšar et al., 2013).

METHOD

This paper is an experimental machine learning as outlined in the figure 1. As a start, after setting the objectives, raw data is prepared, cleansed, structured and determining the features and target. Experimental trial on supervised machine learnig such as Gradient Boosting, Random Forest and SVM model. The promising model then tested or uses as prediction.

Raw data prepared from 2 drilling wells, the raw data then structured to meet the software inputs standard. Data cleansing performed under the software after setting the features and target. Continued with experimentation and predictions.

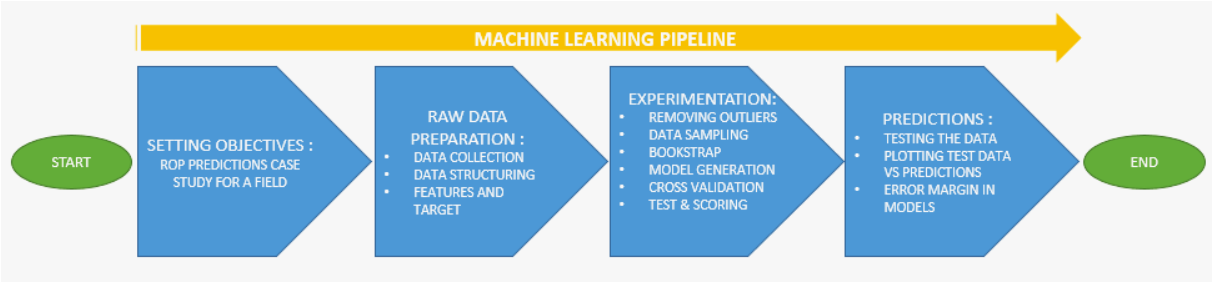


Figure 1. Machine Learning Pipeline to illustrate the experimental steps for ROP prediction.

RESULT

Raw Data Preparation

As mentioned above there are 12 data/features including the surface drilling parameters, drilling fluids parameters and hydraulics parameter will be use as inputs to different model in this paper. Those data need to be verified and analyze prior modelling to clean it from wrong values and to keep only the actual/real rotation/drilling time. Then the data goes thru outlier removals. Outlier is a data that differ from the data set, which could be error data or bad data that might affect the model if not removed. All rotation data is used, including the mud motor RPM, RPM total and rotation RPM.

Only two directional wells data used for this modelling from 17-1/2" drilling section to 8.5" drilling section at total depth. There are 4111 datasets used for these experiments. Both wells will be use as the training data set and remaining data from bookstrapping will be captured and used as blind data set to test the model's accuracy and capabilities. Bookstrap can improve the statistic learning method model such as trees (Hegde, Chiranth, Wallace, Scott, 2015)

Journal of Earth Energy Engineering
Vol. 11 No. 1, March 2022, pp 52-59

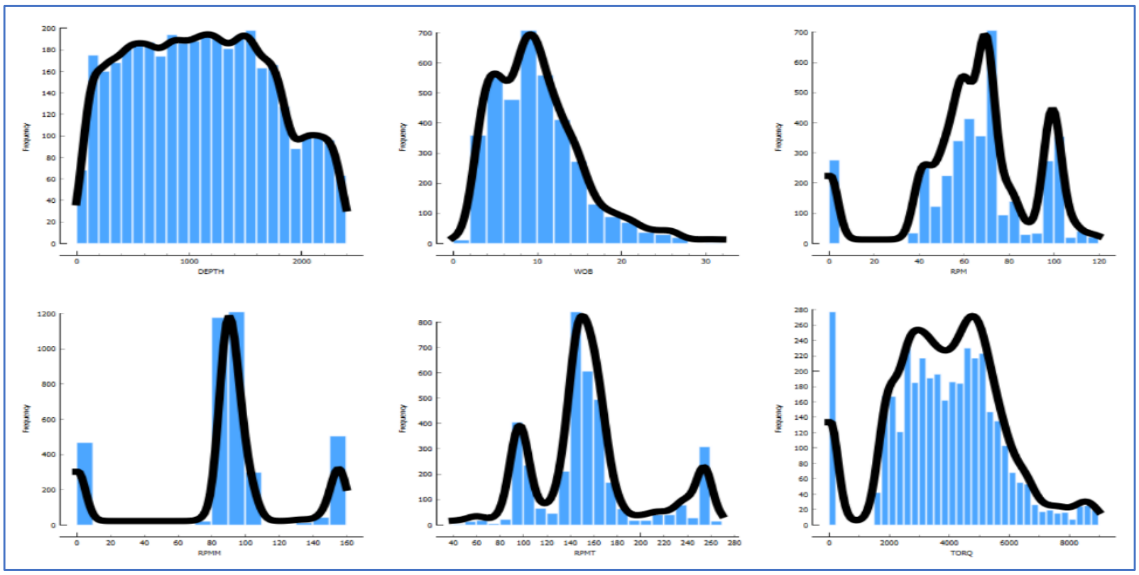
Basic Statistic Parameters

The bookstrapping data shows basic statistic, for verification of range of every parameter after removing outliers as shown in figure 2.

Name	Distribution	Mean	Median	Dispersion	Min.	Max.	Missing
DEPTH		1111.07	1097	0.55	51	2364	0 (0%)
WOB		9.756	9.2	0.495	1.2	30.7	0 (0%)
RPM		64.94	66	0.40	0	120	0 (0%)
RPMM		89.83	92	0.45	0	157	0 (0%)
RPMT		154.77	150	0.29	45	266	0 (0%)
TORQ		3828.80	3844	0.47	0	8934	0 (0%)
SPP		1787.33	1835	0.23	367	2626	0 (0%)
GPM		775.40	804	0.17	220	956	0 (0%)
Bit Diameter		12.2311	12.25	0.2136	8.50	17.50	0 (0%)
PV(cp)		15.12	15	0.29	8	28	0 (0%)
(YP lbs.ft2)		27.19	28	0.14	20	36	0 (0%)
MW ppg		9.44902	9.3296	0.0503095	8.7465	10.3292	0 (0%)
ROP		30.0820	31.02	0.3702	0.17	62.02	0 (0%)

Figure 2. Basic feature statistics of the data.

Afterwards, distribution plots generated to see the parameters distributions. From figure 3, each parameters have a good distribution, even though only with two wells data set. Distribution plots are one of tools in data preparation and analysis. Belyadi & Haghighat, (2021)define the main reason for using a distribution plot is to make sure the distribution of the input and output are normal (Gaussian), this is because most of the Machine Learning algorithm assume that the distribution parameter is normal.



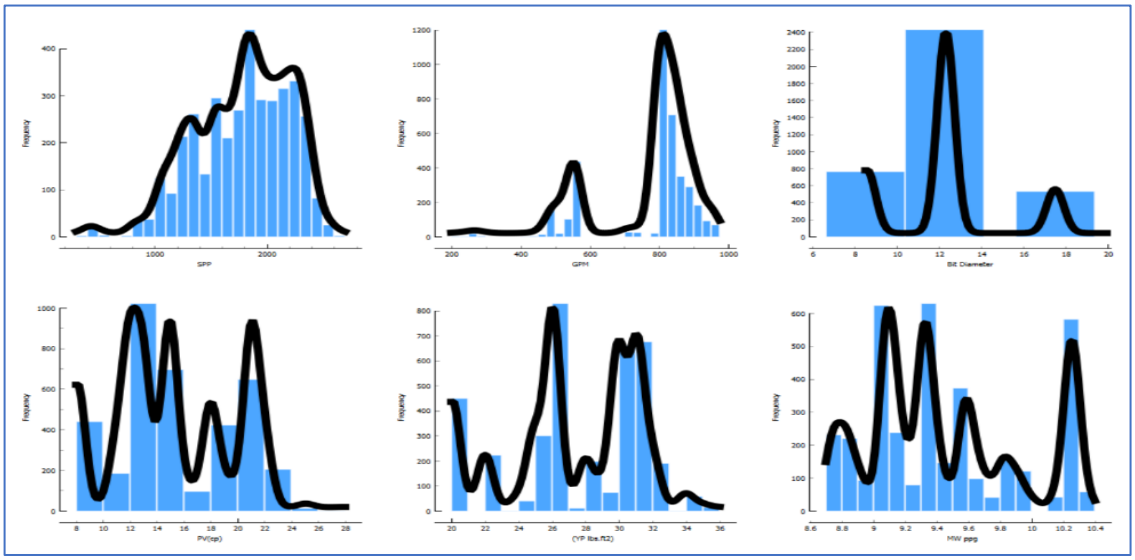


Figure 3. Distribution plots of the drilling data, Depth, WOB, RPM, RPMM, RPMT, TORQ, SPP, GPM, Bit Diameter, PV, YP, MW.

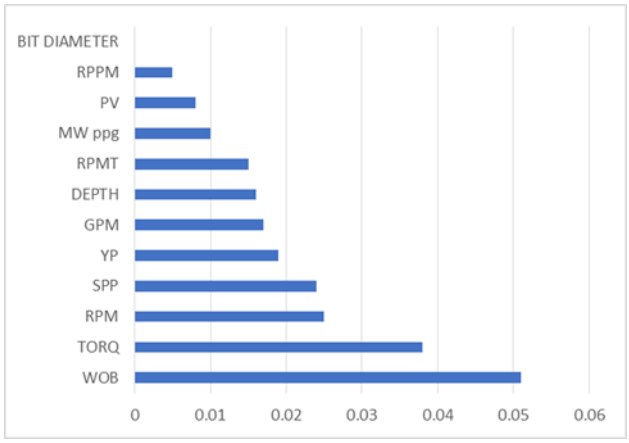


Figure 4. Feature ranking from the drilling data

The feature importance for the models as outlined in figure 4, where the WOB, torsion, RPM and standpipe pressure are the best rank to generate ROP.

Experimentation & Prediction

The data separated, around 70% (randomly sampled) used for training the model and around 30% of the data was sampled to be the data test. Train data is boostraped to the models. The data then inputted to Random Forest, Gradient Boosting and SVM, to train the algorithm tried to predict the target by drawing conclusions from the features (Höhn et al., 2020). Model trained with cross-validation, a statistical method used to estimate accuracy by making partitions of the data, and analyze it on each partition, and averaging it to show the overall error estimate.

Table 1. Modelling results

Model Name	RSME	MAE	R^2
Random Forest	3.68	2.06	0.90
Gradient Boosting	4.07	2.50	0.86
SVM	5.83	4.25	0.72

RMSE is the result error rate prediction, where the smaller or closer to 0, the prediction will be more accurate. RMSE shows the model consistency. MAE represents the average error (error) absolute between the forecast results with the actual value. R^2 is the comparison between the prediction results with the actual value between the independent variables and dependent variables. The metrics can be expressed as follows:

Journal of Earth Energy Engineering
Vol. 11 No. 1, March 2022, pp 52-59

Mean Absolute Error: $$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}| \quad (1)$$

Root Mean Squared Error: $$RMSE = \sqrt{MSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y})^2} \quad (2)$$

R-squared: $$R^2 = 1 - \frac{\sum(yi-\hat{y})^2}{\sum(yi-\bar{y})^2} \quad (3)$$

The trained to show the R^2 and result shows that Random Forest gives 0.90, Gradient Boosting gives 0.86 and SVM gives 0.72.

To know the test data spreads on the prediction, plot of ROP test data versus ROP prediction for Random Forest and Gradient Boosting. The plots represent the R^2 value the ROP prediction in Random Forest shows a good pattern and regression result compared to Gradient Boosting as per Figure 5.

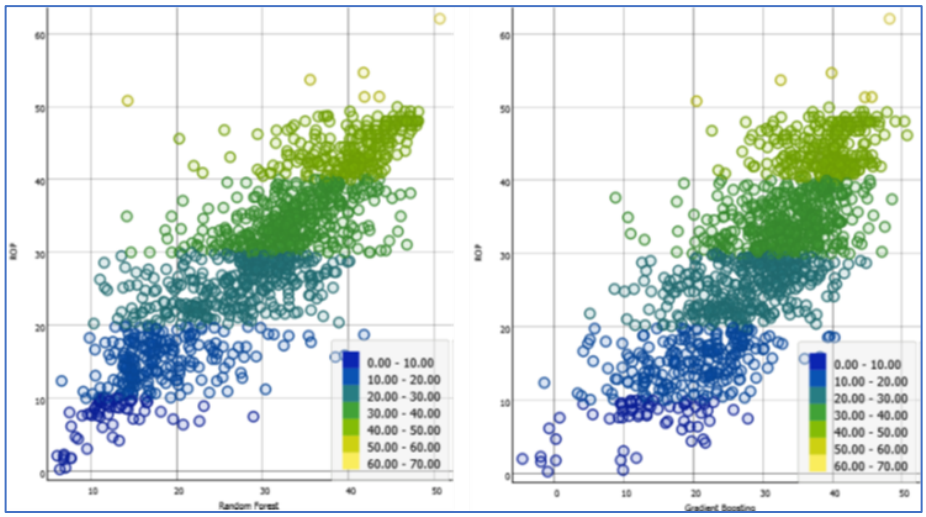


Figure 5. ROP plots between ROP prediction and ROP test data for Random Forest (left) and Gradient Boosting (right) models.

As illustrated in the figure 5, the trend for Random Forest model appears to be the best approach, then other models. The dot in Random Forest is close to the diagonal, i.e. most values are predicted with reasonable accuracy (Höhn et al., 2020).

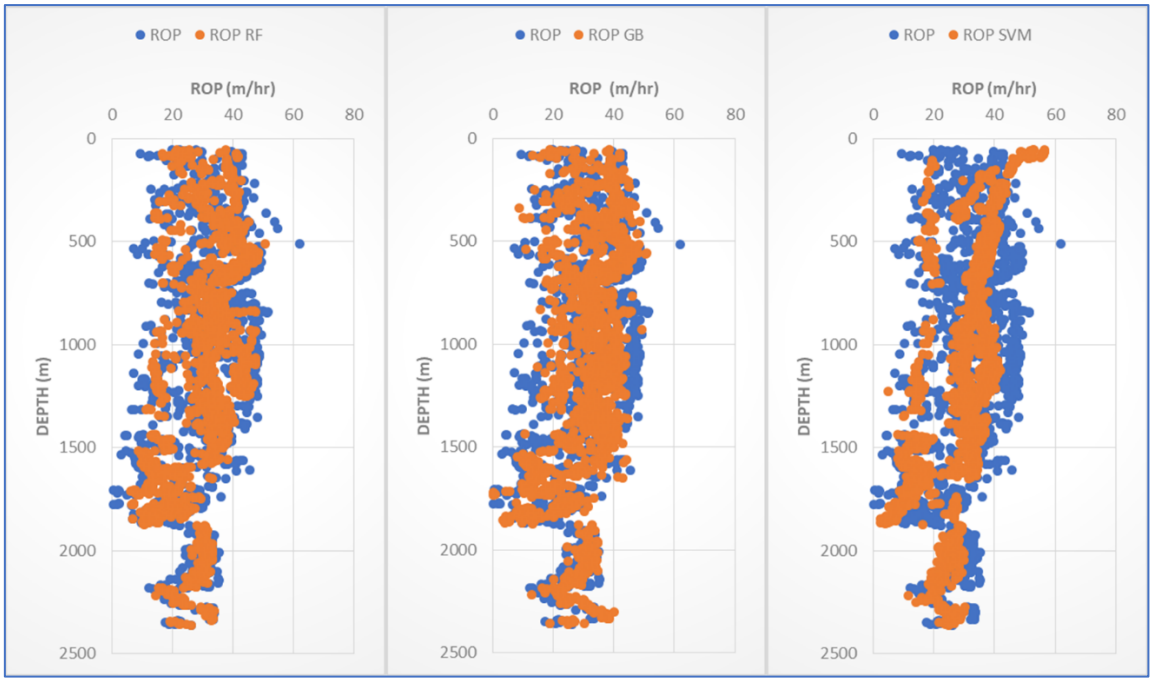
DISCUSSION

This supervised machine learning could be advantageous model to predict ROP in a development field were based on every known layer lithology property, the speed of the drill bit reduces or increases (Hinduja, 2020) could be recognized for proper prediction. Additional data (parameters) needed to make the model to have better prediction, such as lithology, compressive strength, type of bit, etc. In contrary, one of the key aspects of successful in obtaining predictive data-driven model is the process of selecting the inputs (feature engineering) with preference by selecting small sub-sets of inputs (up to eight inputs) as concluded by ((Barbosa et al., 2019)

Random Forest did not show overfitting tendency, and Gradient Boosting shows some overfitting tendency as shown in figure 6. Overfitting is a condition where the analysis from the model matched too closely to data sets, resulting to fit for the blind datasets or when to predict. Futher additional fine-tuning parameters was tried to avoid overfitting but haven't improved anything. Overfitting could be one of weakness in using a decision tree. Random Forest and Gradient Boosting provide the opportunity to retrace which features had the most impact in predicting the target (Höhn et al., 2020). On the other hand, the SVM has shown underperforming conditions when applied to the prediction data sets. It is a challenge to improve the model further with high accuracy by adding more data of the wells to the models to have better ROP prediction in that (specific) field. Another challenge to improve the accuracy by preparing feature engineering or feature augmentation from the domain experts.

Random Forest model perform the best prediction during model testing which shown a good trend with R^2 of 0.90. Random Forest is effective in predicting ROP values with high accuracy as mentioned by previous

research that were trained for specific drilling field. The model could be used for further development within the field with the same lithology.



(a) (b) (c)

Figure 6. ROP prediction versus depth based on the models of Random Forest (a), Gradient Boosting (b) and Support Vector Machine (c).

CONCLUSION

ROP modelling and prediction with supervised learning were performed with Random Forest, Gradient Boosting and SVM with tipical yet traditional data for ROP prediction, combined with flow rate (GPM), mud weight, and mud properties (PV and YP).

Random Forest model perform the best prediction during model testing which shown a good trend with R^2 of 0.90. Random Forest is effective in predicting ROP values with high accuracy as mentioned by previous research that were trained for specific drilling field.

ACKNOWLEDGEMENTS

Author would like to acknowledge the use of Orange Data Mining, a Machine Learning Software and the support of Petroleum Engineering Department, Bandung Institute of Science Technology.

REFERENCES

Barbosa, L. F. F. M., Nascimento, A., Mathias, M. H., & de Carvalho Jr, J. A. (2019). Machine learning methods applied to drilling rate of penetration prediction and optimization-A review. *Journal of Petroleum Science and Engineering*, *183*, 106332.

Belyadi, H., & Haghighat, A. (2021). *Machine Learning Guide for Oil and Gas Using Python: A Step-by-Step Breakdown with Data, Algorithms, Codes, and Applications*. Gulf Professional Publishing.

Chandrasekaran, S., & Govindarajan, S. K. (2019). Optimization of rate of penetration with real time measurements using machine learning and meta-heuristic algorithm. *International Journal of Scientific and Technology Research*, *8*, 1427–1432.

Demšar, J., Curk, T., Erjavec, A., Gorup, Č., Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., & Starič, A. (2013). Orange: data mining toolbox in Python. *The Journal of Machine Learning Research*, *14*(1), 2349–2353.

Journal of Earth Energy Engineering
Vol. 11 No. 1, March 2022, pp 52-59

Fernandes, S. G., Touzani, S., & Granderson, J. (2018). *Gradient boosting machine for modeling the energy consumption of commercial buildings*.

Hegde, Chiranth , Wallace, Scott, and K. G. (2015). Using Trees, Bagging, and Random Forests to Predict Rate of Penetration During Drilling. *Paper Presented at the SPE Middle East Intelligent Oil and Gas Conference and Exhibition, Abu Dhabi, UAE*.

Hinduja, H. (2020). Rate of Penetration Prediction using K-means and Ensembles, a Machine Learning Approach. *International Journal for Research in Applied Science and Engineering Technology*, *8*(7), 843–846. https://doi.org/10.22214/ijraset.2020.30357

Höhn, P., Odebrett, F., Paz, C., & Oppelt, J. (2020). Case Study ROP Modeling Using Random Forest Regression and Gradient Boosting in the Hanover Region in Germany. *International Conference on Offshore Mechanics and Arctic Engineering*, *84430*, V011T11A023.

Mantha, B., & Samuel, R. (2016). ROP optimization using artificial intelligence techniques with statistical regression coupling. *SPE Annual Technical Conference and Exhibition*.

Mitchell, R., & Miska, S. (2011). *Fundamentals of drilling engineering*. Society of Petroleum Engineers.

Noshi, C. I., & Schubert, J. J. (2018). The role of machine learning in drilling operations; a review. *SPE/AAPG Eastern Regional Meeting*.

Shi, X., Liu, G., Gong, X., Zhang, J., Wang, J., & Zhang, H. (2016). An efficient approach for real-time prediction of rate of penetration in offshore drilling. *Mathematical Problems in Engineering*, *2016*.

Singh, K., Yalamarty, S. S., Kamyab, M., & Cheatham, C. (2019). Cloud-Based ROP Prediction and Optimization in Real Time Using Supervised Machine Learning. *SPE/AAPG/SEG Unconventional Resources Technology Conference*.