Forecasting Oil Production of Well 159-F-14H in the Volve Field Using Machine Learning Model

[1]Devy Ayu Rhamadhani, [2]Eriska Eklezia Dwi Saputri, [3]Riska Laksmita Sari
[1,2,3]Departement of Petroleum Engineering, University of Jember
Email: [1]191910801016@unej.ac.id, [2]eriska.eklezia@unej.ac.id, [3]riskalaksmita@unej.ac.id

Article Info

Article history:
Received Oct 19th, 2023
Revised Jan 25th, 2024
Accepted Feb 24th, 2024

Keyword:
Artificial Intelligence
Forecasting
Machine Learning
Production
Regression

ABSTRACT

Petroleum engineers require information about the production performance of a well in order to know when the well is no longer feasible to produce. By using the approachment technique of machine learning, the research was conducted using a tree-based regression model, Random Forest Regressor, Extra Trees Regressor, and Gradient Boosting Regressor. This research was done by predicting the production of an existing well in the Volve field, namely well 159-F-14H using its field data; average downhole pressure, average downhole temperature, average wellhead temperature, average wellhead pressure, on-stream hours, average choke size percentage, gas volume from well, water volume from well. The data used is 1093 days and 70% is used for training and as much as 30% for testing. A comparative study was carried out on the predictive performance of the three models. Random Forest shows the best testing result as well as RMSE 5.134 and R^2 0.974, followed by Gradient Boosting shows RMSE 5.927 and R^2 0.965, and Extra Trees shows RMSE 6.524 and R^2 0,958.

Corresponding Author:
Eriska Eklezia Dwi Saputri,
Departement of Petroleum Engineering,
University of Jember,
37 Kalimantan Tegalboto St., Sumbersari Township, Jember County 68121, East Java, Indonesia.
Email: eriska.eklezia@unej.ac.id

DOI: http://dx.doi.org/10.24014/ijaidm.v7i1.24907

1. INTRODUCTION

The application of Artificial Intelligence (AI) in predicting oil well production addresses the challenges posed by the complexity of reservoir properties, such as pressure, porosity, permeability, saturation, and others. AI's ability to handle non-linear relationships, process diverse datasets, and capture intricate patterns enhances the accuracy of production forecasts, making it a valuable tool in the oil and gas industry. The integration of AI in predicting crude oil production enhances the industry's ability to analyze complex data, optimize operations, and make informed decisions. This, in turn, contributes to increased efficiency, reduced costs, and improved overall performance in the oil and gas sector.

Authors undertook this research endeavor to complement the preceding study conducted by Cuthbert Shang Wui Ng et al. in 2022 titled *Well production forecast in Volve field: Application of rigorous machine learning techniques and metaheuristic algorithm*. In their work, they employed various machine learning models such as SVR, FNN, RNN, and PSO to predict hydrocarbon production in an oil well in Volve field. Each of these models demonstrates excellent outcomes in training, validation, and testing phases, achieving correlation coefficients (R2) surpassing 0.98. Furthermore, their predictive capabilities remain strong, with R2 consistently exceeding 0.94.

While this research was conducted by analyzing the predictive performance of machine learning models in predicting oil production of well 159-F-14H in Volve field as well as Random Forest Regressor, Extra Trees Regressor, and Gradient Boosting Regressor. The Volve field is located at a depth of between 2750 m and 3210 m below sea level. The rock characteristics in this field have a permeability value of around

1000 mD, porosity of 0.21, and net-to-gross ratio 0.93, and the average water saturation of the oil bearing zone is 0.2.

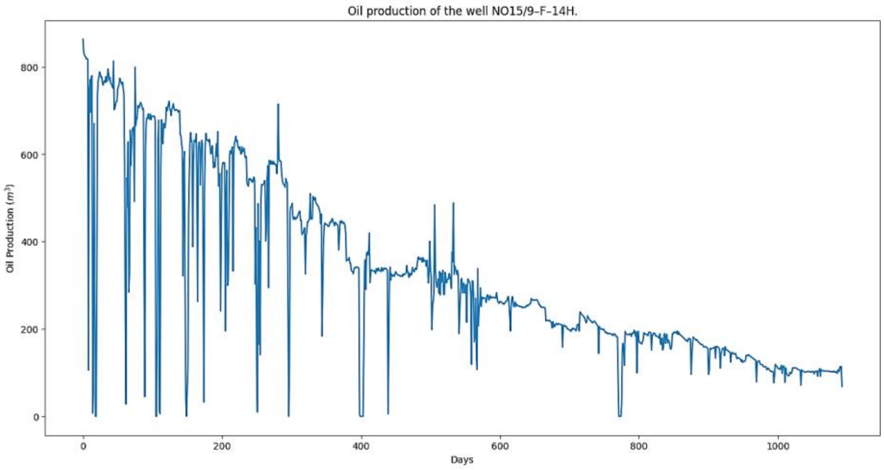


Figure 1. Oil production of well 159-F-14H from July 14, 2013 to July 13, 2016

2. RESEARCH METHOD

This research was carried out using literature studies and the authors considered using production data for well 159-F-14H from an existing field, namely the Volve field which is located in Block 15/9 in the southern part of the Norwegian North Sea. Research is done by doing modelling with several machine learning approachments. The authors will determine whether the data is feasible for further processing or not. Whether or not the data is appropriate is seen from the amount of data that is empty. If there is a lot of empty datas or unfit for use, a literature study will be carried out again to find data that is considered feasible to process.

After finding the feasible data to process, the authors will import the library that will be used for modelling. The authors used production data of the well 159-F-14H as well as 1093 days production data. Modelling is done using Random Forest Regression, Extra Trees, and Gradient Boosting with Python. the The data set will be divided into 2 parts, training data as well as 70% of the total data testing data as well as the rest of it (30%). In the training stage, models are trained using the training data. Each model has parameters that can be used for modelling to improve the performance of the model. Parameters that are used for modelling will greatly affect the value of RMSE and R^2 (output). So, to get a good RMSE and R value, we have to do several experiments to do a trial and error technique to get a fit paramater for each model. If the RMSE and R values^2 the results are considered good, then this stage can be stopped and the model can be used for the next stage (testing) with the selected parameters.

In the testing stage, models that have been trained will be used to learn the testing data set and this is a final stage from modelling. This stage is used to determine the accuracy of each model in predicting oil production in the field that is studied. At this stage, the model that has been trained will be tested for its capabilities with different data. Once the RMSE and R^2 values are appeared then it can be used to analyze its result. Data set that is used of well 159-F-14H; average downhole pressure, average downhole temperature, on stream hours, average choke size percentage, average wellhead pressure, average wellhead temperature, gas volume from well, water volume from well.

The authors used parameters for each model as well as random_state and max_depth. This is based on the output that is produced by the models are considered as good and acceptable.

Table 2. Parameter for each model (Paragraph, after=12, before=13)

Model	Parameter

Random Forest	max_depth=8
	random_state=20

Extra Trees	max_depth=9
	random_state=20

Gradient Boosting	max_depth=3
	random_state=20

random_state is used to ensure that the results produced by the model will be the same every time it runs, as long as the value of random_state is not changed. If the value is changed, the results produced by the model will also be different. max_depth used to determine the depth of the decision tree, this parameter is usually used to control the complexity of the model Small value used in this parameter will be better because it can avoid overfitting. Large values can increase the model's ability to predict new data but can cause overfitting. Overfitting is a condition where a model is too fit with the training data so that it cannot learn other data as well.

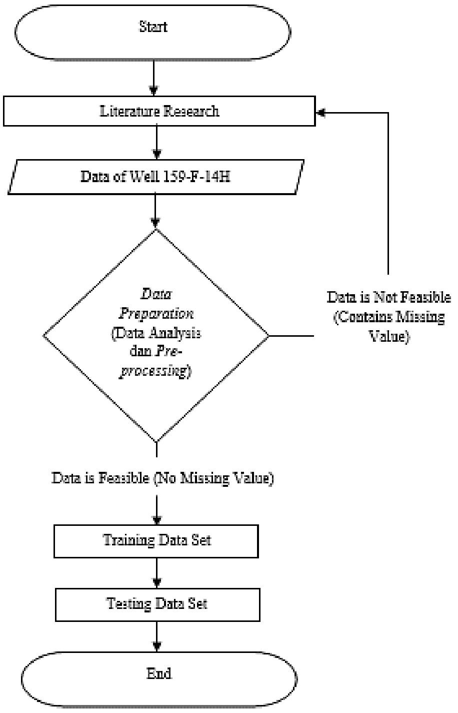

Start
Literature Research
Data of Well 159-F-14H
Data Preparation (Data Analysis dan *Pre-processing*)
Data is Not Feasible (Contains Missing Value)
Data is Feasible (No Missing Value)
Training Data Set
Testing Data Set
End

Figure 5. Research flowchart

2.1. Forecasting

Forecasting is a process to predict the future events. This involves taking past data and applying it to a mathematical model [2]. In the world of oil and gas, engineers do this to find out about the production performance of a well during the desired time span using data of well property and others.

2.2. Random Forest

Random Forest is a combination of several tree predictors or called as decision trees where each tree depending on the value of random vector which is sampled freely and evenly in all trees in its forest. Prediction results of Random Forest obtained through the most results of each individual decision tree (voting for classification and mean for regression).

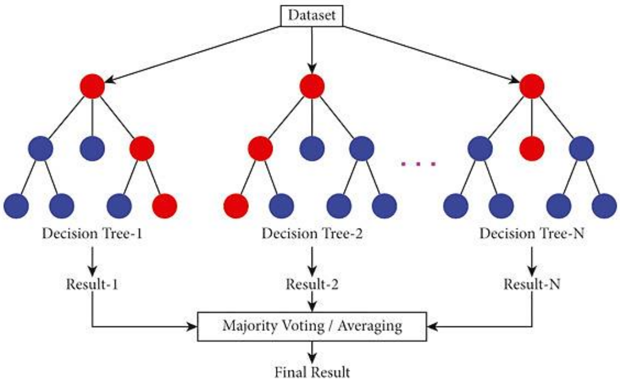


Figure 2. Ilustration of random forest algorithm works

Random Forest has an internal mechanism that provides estimation of its error generalization called out-of-bag (OOB) error estimate. OOB error estimation is the average of the prediction errors for each training case y use tree which does not include y in its bootstrap sample. Then, when the model is created, all the training cases down each tree and the proximity matrix for each case is calculated based on the pair of cases that arrive at the same node terminal.

2.3. Extra Trees

The classification and regression method known as "Extra Trees" or also called "Extremely Randomized Trees" is a form of development of a random decision tree. The dataset is divided into several subsections and an average of the result of the decisions is taken to improve the prediction accuracy and control overfitting. In the algorithm explanation document of extra trees written by Pierre Geurts, this algorithm has a very great resemblance to its predecessor, namely Random Forest. The difference lies only in the way of the constructionof the decision tree. decision tree mechanism involves a tree structure, where each internal node represents testing of certain attributes, each branch represents testing of the test results, and leaf node represents a class or category [6].

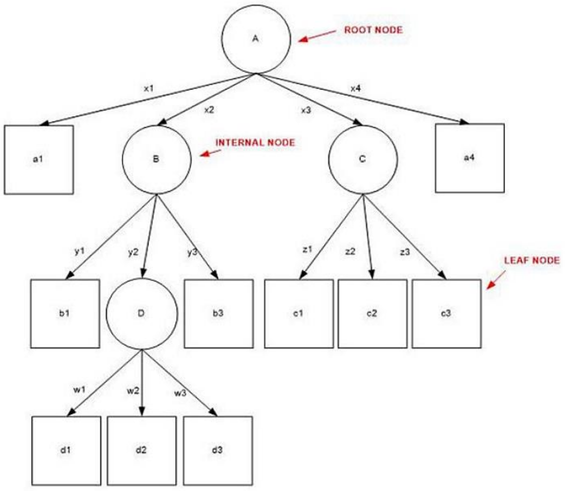


Figure 3. Ilustration of decision tree

2.4. Gradient Boosting

Gradient boosting is decision-tree-based model which can be used for classification. Gradient boosting just like other Boosted family that have the ability to improve the predictive accuracy of the model. Gradient boosting algorithm works sequentially by adding previous predictors that don't match the predictions into ensemble to correct existing errors. Ensemble is a combination of decisions from several machine learnings, where the class that gets the majority of "votes" will be the class that is predicted by ensemble overall. Gradient boosting starts with generating an initial classification tree then adapting the new tree by minimizing the loss fuction.

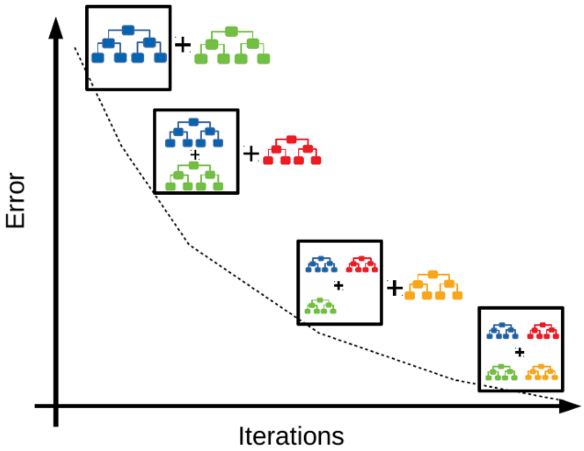


Figure 4. Iterations approachment of gradient boosting

2.5. RMSE and R^2

Root Mean Square Error (RMSE) is a measure of the level of prediction error, where the smaller (closer to 0) the RMSE value, the more accurate the prediction results will. The RMSE value is used to differentiate model performance between the calibration period and the validation period and is used to compare performance between one prediction model and another [9].

$$RMSE = \sqrt{\frac{\Sigma_{j=i}^{n}(y'-y)^2}{n}} \tag{1}$$

The determination coefficient (R^2) basically measures the extent to which the model is able to explain variations in the dependent variable. R^2 has a value range between zero and one. R value^2 which is low indicates the limited ability of the independent variable in explaining the variation of the dependent variable. Meanwhile, a value close to one indicates that the independent variable provides almost all the information needed to predict the variation of the dependent variable. The coefficient of determination serves to determine the percentage of independent variable influence on the dependent variable. For example, if the value of R^2 is 0.6, meaning that 60% of the variation of the dependent variable can be explained by the independent variable. The remaining 40% cannot be explained by independent variables.

$$R^2 = 1 - \frac{\sum_{t=1}^{n}(y_t - \hat{y})^2}{\sum_{t=1}^{n}(y_t - \bar{y})^2} \tag{2}$$

Table 1. Interpretation of R^2 Value

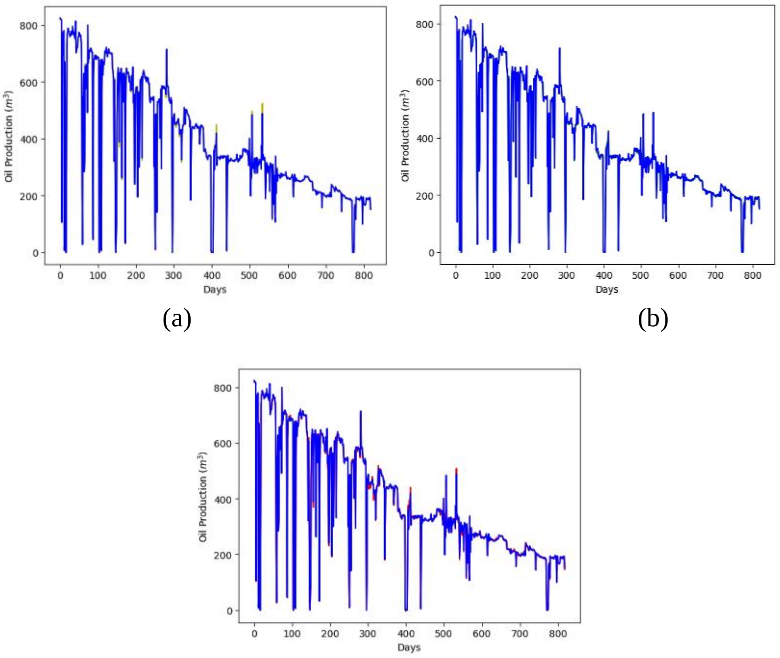
Interval of Coefficient	Category
1 – 0,8	Very Strong
0,6 – 0,79	Strong
0,4 – 0,59	Strong Enough
0,2 – 0,39	Weak
0 – 0,19	Very Weak

3. RESULTS AND ANALYSIS

Based on modelling that has performed with all three models. From testing stage, Random Forest generates RMSE and R^2 values the best, means that the model has the best level of accuracy in predicting compared to the other two models. In the next sequence, Gradient Boosting become the best accuracy level model after Random Forest. Almost all models have good accuracy when they performed in training stage because the model is set as well as possible to be able to study the training data, and it produces lower result in testing stage with different data set and produced higher error as can be seen in Table 3 of This naturally happens because the model has never studied the data and only set it as best as possible for training data set. The smaller the RMSE value (closer to zero) means the error that is produced by the model is smaller, and the R^2 value is higher (close to one) means the model is getting more accurate at studying the data. However, because the parameters in determining predictive performance are not only seen from the RMSE but also the R^2 value then the R^2 value also has a major influence on the results of model performance. Mmodel with a relatively small RMSE value but has an R^2 value which is also small, means that the model cannot be clearly said to be a model with good predictive performance because it means that with an R^2 value small indicates that the distribution of the dependent variable cannot be explained properly by the independent variables.

Table 3. RMSE and R^2 of each model

Model	Training	
	RMSE	R^2
Random Forest	5,590	0,999
Extra Trees	5,391	0,999
Gradient Boosting	6,421	0,998
Model	Testing	
	RMSE	R^2
Random Forest	5,134	0,974
Extra Trees	6,524	0,958
Gradient Boosting	5,927	0,965

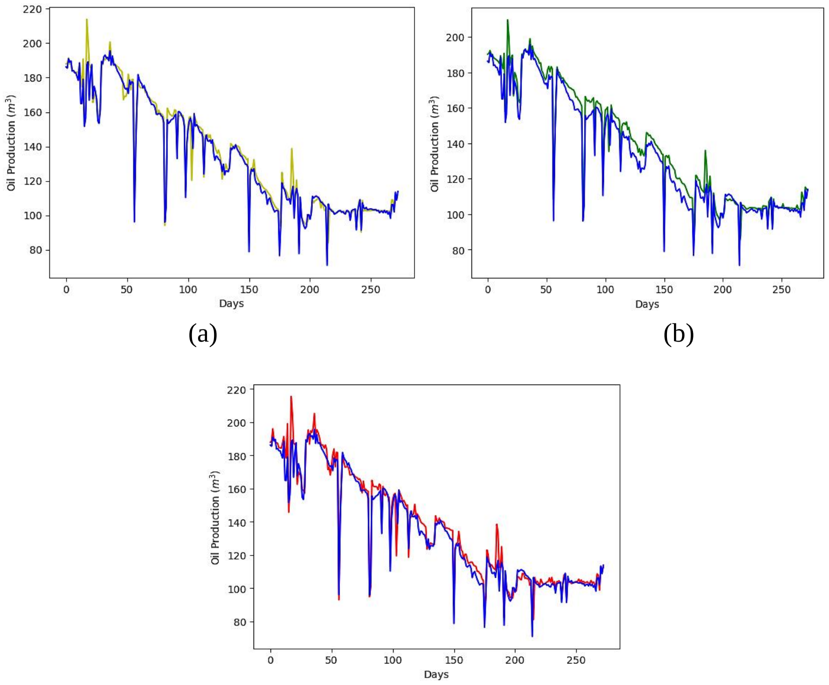


(a) (b)

(c)

Figure 6. Training plot of (a) Random Forest (b) Extra Trees (c) Gradient Boosting

In the training stage, it can be seen clearly through the graph that the three models can learn the data set well. The blue color chart shows the actual production data and the red, yellow, and green ones are the prediction result. The three models show that the predicted results are very close to the actual values of oil production. Although some models produce higher RMSE values in the training stage than in the testing one as well. But the R^2 value which is quite high indicates the predictive performance of the three models is very good means the independent variables can well explain the distribution of the dependent variable.



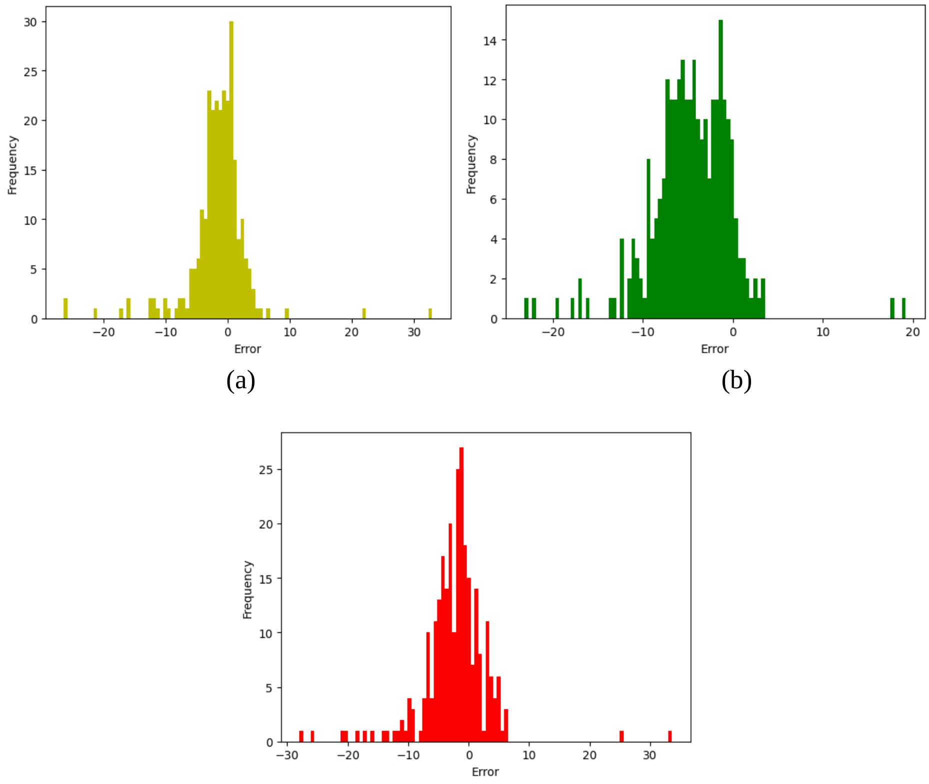
(a) (b)

(c)

Figure 7. Testing plot of (a) Random Forest (b) Extra Trees (c) Gradient Boosting

In the testing stage, Random Forest is the most accurate in predicting because the prediction is closest to the actual values of oil production. Meanwhile, the other two models look not much different. Extra

Trees and Gradient Boosting are are quite good at predicting judging from the distance between the actual graph and the prediction graph which is not too far or close enough. But both of these models have produced more errors than Random Forest did



(a) (b)

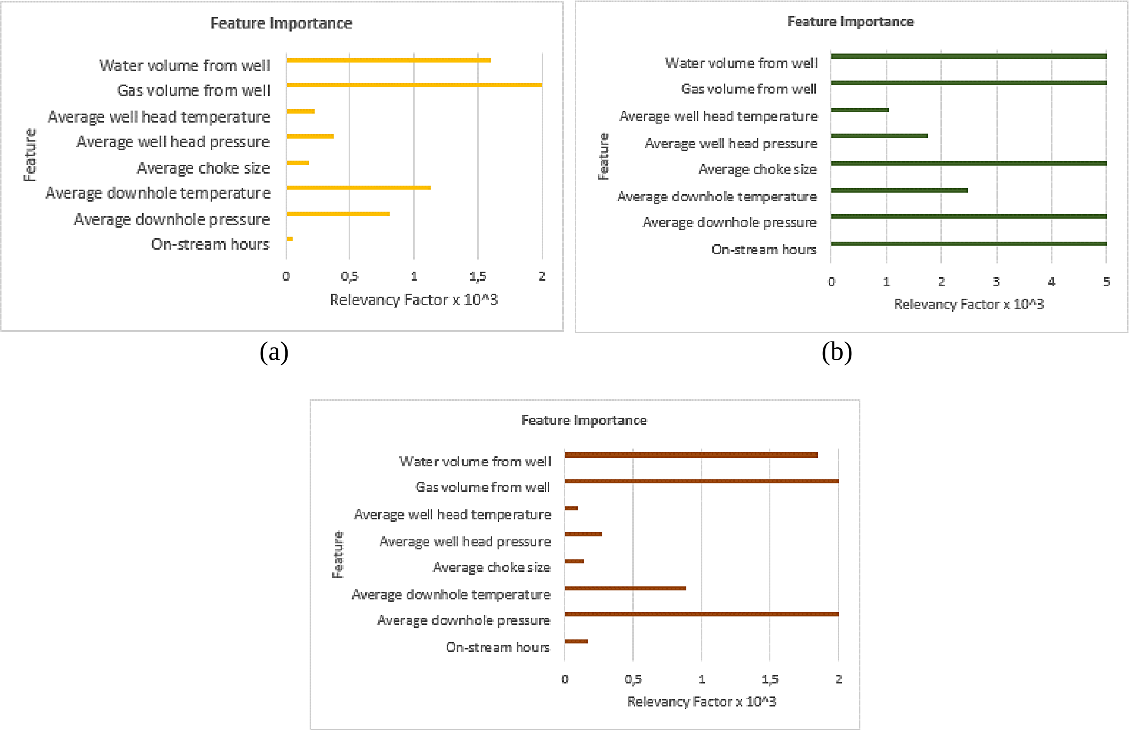
(c)

Figure 8. Error distribution of (a) Random Forest (b) Extra Trees (c) Gradient Boosting

From Figure 8. shows that each model indicates that only Random Forest produced a better error distribution than other two models due to a more bell-shaped distribution. Error mean how big the difference between the predicted results and the actual production data. If model produced graphs as right-side graph or greater than zero indicates that the model produced a prediction value that is smaller than the actual value of production and if model produced is as a lef-side or produced it as less than zero indicates that the model is experiencing overestimation or the model produces a greater predictive value than the actual value of production.

A good distribution graph should have a bell shape that is symmetrical and centered on zero, if the graph does not show this, it means that there is some structure in the model prediction error. The symmetrical distribution indicates that the model makes random and unsystematic errors. Extra Trees and Gradient Boosting have an asymmetrical graphic shape, then the graph strengthens the statement of authors that those models are no better than Random Forest in predicting. The frequency on the graph shows the number of occurrences of values distributed in a certain interval. From the graph of the three models, it shows that the three models tend to produce a greater predictive value than the actual value of the data shown by Fugure 8. which tends to have a greater frequency on the left or negative side or below zero. Random Forest is the best model because it shows the low frequency of the difference in value between the predicted results and the actual data.

From Figure 9 the three models show that gas production has the most influential impact on the output or prediction of oil production. This can happen because gas production coincides with oil production. Gas has a lower density than oil so that the gas phase is above the oil phase and when oil is flowed from the reservoir to the surface, gas will also flow and be produced. This causes the size of the rate of oil production to trend the same as gas production. And because the production trend is the same or linear, then the model produced that gas volume is one of the biggest factors in the size of oil production. Feature importance is a benchmark for the magnitude of the contribution of various data to the performance of the prediction model. Extra Trees and Gradient Boosting, feature that affects the oil production after gas volume is average downhole pressure. Meanwhile on Random Forest, the most influential feature after the gas volume is the volume of formation water. The more water produced by a well, the lower of well’s productivity will become.



(a) (b)

(c)

Figure 9. Feature importance of (a) Random Forest (b) Extra Trees (c) Gradient Boosting

Formation water can also cause production problems such as scale. Formation water contains anions dissolved in it, but these anions can turn into solids as a result of changes in pressure and temperature because their balance is disturbed. These solids will later become a problem, because if they are produced in large quantities they will settle on the production equipment through which water passes. This is what will eventually reduce oil production. Choke is used to regulate the production flow rate in order to maintain oil production to remain at an optimum point, produce reservoirs at the most efficient rate, and prevent the entry of sand during the production process. That way oil production can be regulated as desired. Wellhead pressure has bigger influence on Extra Trees than on other two models. If the wellhead pressure etting bigger then the production flow rate will also be even greater. Not only does the downhole pressure affect the size of the production rate, but so does the pressure at the wellhead. Likewise with the wellhead temperature, high temperatures cause the oil to flow more easily because the viscosity is reduced and low temperatures cause the oil molecules to become stagnant so that the viscosity increases and the ability of the oil to flow decreases.On-stream hours is the time the well is alive or doing production (hours), on-stream hours is influential enough because it can be controlled by the engineer, engineer is able to have control about the time when well will be activated and when it’s not, the oil production is also based on the active time of the well.

In this study, authors found that Random Forest is a model that has lowest error compared to other models, means that this model has the best predictive performance compared to Extra Trees and Gradient Boosting. This is also evidenced in Figure 8. That Random Forest has the lowest frequency of error. By producing RMSE as well as 5.134 and R^2 as well as 0.974, Random Forest considered as a model with the highest accuracy. From all the data obtained from modelling, it can be seen that the RMSE and R^2 the three models are not much different from one model to another. This is because they belong to the same type of regression model as tree-based model. One of the advantages of tree-based model is its ability to predict data with high accuracy compared to other machine learning models. The greater the number of trees, the higher the accuracy value will be. Random Forest looks the most superior or accurate compared to the other two models because Random Forest itself is a model with better accuracy than other tree-based models and can work well on larger data sets. Output diffrence produced by each model is a very natural thing to happen in modelling because each model has a different way of working with its own limitations in predicting. Random Forest works by choosing random sample from existing data set then create a decision tree for each selected sample. The Output from decision tree will be processed by voting for each predicted result. Gradient boosting, in simple terms this model works by correcting errors from weak learner. This model uses an iterative approach to produce error which is quite small. This model has advantages as well as produces high accuracy and has a faster computational speed compared to random forest. But that makes this model no better than random forest is this model is easy to experience overfitting. This can happen if the parameter settings are not done properly. Extra trees is quite similiar random forest where to it uses decision tree but has a different tree structure. This model builds a tree with all samples and chooses a random intersection point for each feature considered. Extra trees does the node separation randomly compared to random forest who chooses best node to split. With random separation node, the algorithm will be less affected by certain features or patterns in the data set. It is this working principle that makes each model able to produce different output.

4. CONCLUSION

Based on the research conducted, the researchers drew several conclusions that all models have good predictive performance, in terms of the RMSE value which is not too high and the R value^2 which is not too low. The model that has the best predictive performance or the most accurate in predicting is Random Forest Regressor with RMSE value of 5.174 and R^2 of 0.974. Whereas Extra Trees produces RMSE values of 6.524 and R^2 0,958. Gradient Boosting is the model with the best predictive performance after Random Forest with RMSE value of 5.927 and R^2 0.965. This is also reinforced by the results of the error distribution which shows Random Forest has lower error frequency. Based on the error distribution, all models tend to produce a prediction value that is greater than the actual value of the oil production. In terms of the effect of the input variables, all models show that gas volume is the most influential on oil production. The three models show some differences in the magnitude of the influence of each feature (relevancy factor).

REFERENCES

[1] Dariato, E. (2022). Analisa dan Perancangan Machine Learning Untuk Mendeteksi Kegagalan Job di Apache Spark. *Arcitech: Journal of Computer Science and Artificial Intelligence*, *2*(1), 1. https://doi.org/10.29240/arcitech.v2i1.4124

[2] Difitria, R., & Cholissodin, I. (2020). *Penerapan Support Vector Regression dan Particle Swarm Optimization untuk Prediksi Jumlah Kunjungan Wisatawan Mancanegara ke Daerah Istimewa Yogyakarta*. *4*(5), 1364–1371. http://j-ptiik.ub.ac.id

[3] Homepage, J., Roihan, A., Abas Sunarya, P., & Rafika, A. S. (2019). IJCIT (Indonesian Journal on Computer and Information Technology) Pemanfaatan Machine Learning dalam Berbagai Bidang: Review paper. In *IJCIT (Indonesian Journal on Computer and Information Technology)* (Vol. 5, Issue 1).

[4] JARINGAN SARAF TIRUAN Studi Kasus, P., DAS Siak Hulu, S., & Suprayogi, I. (n.d.). *MODEL PREDIKSI LIKU KALIBRASI MENGGUNAKAN*. http://ce.unri.ac.id

[5] Mitchell, T. M. (Tom M. (n.d.). *Machine Learning*.

[6] Mostafa, S. M., & Amano, H. (2019). Effect of clustering data in improving machine learning model accuracy. *Journal of Theoretical and Applied Information Technology*, *97*(21), 2973–2981.

[7] Ng, C. S. W., Jahanbani Ghahfarokhi, A., & Nait Amar, M. (2022). Well production forecast in Volve field: Application of rigorous machine learning techniques and metaheuristic algorithm. *Journal of Petroleum Science and Engineering*, *208*(PB), 109468. https://doi.org/10.1016/j.petrol.2021.109468

[8] Nurani, A. T., Setiawan, A., Susanto, B., Salatiga, D., & Tengah, J. (2023). *Perbandingan Kinerja Regresi Decision Tre e dan Regresi Linear Berganda untuk Prediksi BMI pada Dataset Asthma*. *6*(1), 34–43.

[9] Putra, B. P., & Kiono, B. F. T. (2021). Mengenal Enhanced Oil Recovery (EOR) Sebagai Solusi Meningkatkan Produksi Minyak Indonesia. *Jurnal Energi Baru Dan Terbarukan*, *2*(2), 84–100. https://doi.org/10.14710/jebt.2021.11152

[10] Somvanshi, M., & Chavan, P. (n.d.). *A Review of Machine Learning Techniques using Decision Tree and Support Vector Machine*.

[11] Vedapradha, R., Hariharan, R., & Shivakami, R. (2019). Artificial Intelligence: A Technological Prototype in Recruitment. *Journal of Service Science and Management*, *12*(03), 382–390. https://doi.org/10.4236/jssm.2019.123026

[12] Yunita, L. (2019). Penentuan Kehilangan Tekanan dari Wellhead menuju Separator dengan Bantuan Simulator pada Sumur Panas Bumi. *ReTII*, *2019*(November), 496–502. https://journal.itny.ac.id/index.php/ReTII/article/view/1523%0Ahttps://journal.itny.ac.id/index.php/ReTII/article/view/1523/943

[13] Zebua, Y. A., Sitompul, D. R. H., Sinurat, S. H., Situmorang, A., Ruben, R., Ziegel, D. J., & Indra, E. (2022). Prediksi Penetapan Tarif Penerbangan Menggunakan Auto-Ml Dengan Algoritma Random Forest. *Jurnal Teknik Informasi Dan Komputer (Tekinkom)*, *5*(1), 115. https://doi.org/10.37600/tekinkom.v5i1.508

BIBLIOGRAPHY OF AUTHORS

Devy Ayu Rhamadhani, a final year student at University of Jember, majoring in Petroleum Engineering. She has an interest in learning about data science and currently developing her skill by doing research about artificial intelligence approachment. Another research can be found in journal.unej.ac.id/JSED/index

Eriska Eklezia Dwi Saputri, is currently a lecturer from Department of Petroleum Engineering University of Jember. She teaches actively for Fluid and Rock Properties, Well Loging, Well Testing and Drilling. Received Bachelor’s Degree and Master’s Degree from Petroleum Engineering at Institut Teknologi Bandung.

Riska Laksmita Sari is currently a lecturer from Department of Petroleum Engineering University of Jember. She teaches actively for Fluid and Rock Properties, Reservoir Engineering, Well Testing and Well Stimulation. Received Master’s Degree from Petroleum Engineering at Institut Teknologi Bandung.