

Optimization of Oil Well Operations with Machine Learning Model to predict Well Production at Company "XYZ"

Yasser Arapat¹, Juni Nurma Sari², Satria Perdana Arifin³, Yohana Dewi Lulu Widyasari⁴

{yasser22mttk@mahasiswa.pcr.ac.id¹, juni@pcr.ac.id², satria@pcr.ac.id³, yohana@pcr.ac.id⁴}

Politeknik Caltex Riau, Pekanbaru, Indonesia^{1,2,3,4}

Abstract. PT XYZ is a national oil company located in Minas, Siak Regency, Riau, Indonesia. Over time, the oil production process has decreased due to several reasons, one of which is the lack of pressure from the production well. So that at this time the wells that have decreased production are treated by injecting produced water into the reservoir using injection wells with the aim of increasing reservoir pressure so that the oil contained at the bottom of the earth will rise to the surface of the earth. Water that is produced with oil will be treated and injected back into the reservoir. Produced water is treated in accordance with the characteristics of the water contained in the reservoir, which, if different, will damage the reservoir. This research will discuss the application of data mining methods in the well production prediction process. The methods used are ARIMA and ANN, and the testing is done using RMSE and MAE. The results of this study show that the ANN method has higher accuracy results. This can be seen from the average accuracy results of each production with several treatments.

Keywords: petroleum, reservoir, ARIMA, ANN

1 Introduction

Petroleum is a mixture of hydrocarbon compounds composed of mostly carbon and hydrogen with small amounts of sulfur, nitrogen, and other elements [1]. According to the Ministry of Energy and Mineral Resources, the need for petroleum continues to increase along with its considerable use and the population that continues to grow every year [2]. The situation is best compared with the time the petroleum production process is decreasing. Low well production in the oil industry is a major problem, whether in new wells, wells that have been producing for a long time, or wells that are re-worked (workover). Low production is caused by the pressure in the reservoir having decreased, so that the remaining pressure at the bottom of the earth can lift the fluid up to the earth's surface. In this case, the right step to maintain the pressure contained in the reservoir is to inject water into the reservoir [3].

This step has been anticipated by PT XYZ. The company is a national oil industry located in the Minas area, Siak Regency, Riau, Indonesia. PT XYZ injects produced water, which is not ordinary water but water that can be produced with oil. To get the desired produced water, treatment is first carried out, including water treatment injection. In order for the company to evaluate the performance of the water treatment injection at the well, it is necessary to know the

amount of oil production for the next period. When oil production has decreased a lot, treatment is needed for the well [4]. However, technology can actually be utilized to determine the amount of oil production for the next period, namely by using machine learning techniques.

The purpose of this study is to optimize well operations for wells where the amount of production for the next period will be known so that if the well has decreased production, treatment needs to be carried out on the well. This research is expected to provide information about the amount of oil production for the next period, thus helping the company make new decisions when the amount of oil production decreases during the next period.

2 Related Work

Several studies related to predicting the amount of petroleum production have been conducted in the last decade. Windi [1] conducted research on the application of ARIMA models, neural networks, and ARIMA-neural network hybrids. The ARIMA-neural network model is the best model of the amount of petroleum production. With a forecast for the next 14 days, the amount of petroleum production tends to be constant.

In research conducted by Omekara [4], she conducted research on the application of ARIMA models to crude oil production in Nigeria. The dataset used comes from the Central Bank of Nigeria (CBN) website. This research uses the Minitab application with parameter sets, namely $\phi = 0.3813$, $\theta_1 = 0.6931$, and $\theta_2 = 0.8649$.

In research conducted by Augustine [6], she conducted research on crude oil production prediction using the quadratic regression and layer recurrent neural network methods. It was found that the Layer Recurrent Neural Network model was better at forecasting, with better RMSE and MEA for 50–200 days and 400–800 days of data.

From all these previous studies, this research will conduct experimental research by comparing machine learning techniques using the autoregressive integrated moving average (ARIMA) and artificial neural network (ANN) methods. These two methods are used to compare which method is more suitable for use on production data sets. This study uses production data for August 9, 2021, to August 8, 2023, for a total of 214,864 data points. This dataset will compare the performance and accuracy of the three methods for predicting the amount of petroleum production.

3. Theoretical Foundation

3.1 Definition of Petroleum

Petroleum is a unique and complex mixture of thousands of compounds. Most of the compounds in crude oil are hydrocarbons (organic compounds composed of carbon and hydrogen atoms). The other compounds contained in crude oil are not only carbon and hydrogen but also other smaller compounds consisting of elements, mainly sulfur, nitrogen, and metals (for example, nickel, vanadium, and other metals). The compounds formed in various crude oils consist of the smallest and simplest hydrocarbon molecules (CH₄-methane) as well as complex molecules containing up to 50 or more carbon atoms (well, hydrogen and hetero-elements) [7].

3.2 Forecasting

Forecasting is a method to estimate a future value using past data. Good forecasting is a directed preparation step so that it will determine the quality or forecasting carried out by following the procedure or the quality of the compiled forecasting results. Meanwhile, the forecasting principles to consider are that forecasting involves errors, forecasting will only reduce uncertainty but not eliminate it. The characteristics of good forecasting are accurate, require less cost, do not have many conditions, are not affected by excessive computer systems, and re disseminated [7].

3.3 Result Evaluation

The results of the predicted data will be tested by model testing and assessing forecasting with indicators of percentage accuracy, RMSE and MAE.

1. RMSE

Root Mean Square Error (RMSE) is a method for measuring the bias or difference in the prediction value of the model.

$$RMSE = \sqrt{MSE} = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n}} \quad (1)$$

2. MAE

Mean Absolute Error (MAE) is the average value of the absolute difference between the actual (actual) value and the predicted (forecasting) value

$$MAE = \frac{1}{n} \sum_{i=1}^n |A_i - F_i| \quad (2)$$

4 Analysis and Comparison of Methods

From the evaluation results of RMSE and MAE for each method using several kinds of partitioning methods, then analysis and comparison are carried out to see the advantages and disadvantages of each method.

5 Autoregressive Integrated Moving Average (ARIMA)

The ARIMA model dominates in the field of time series forecasting. ARIMA stands for AutoRegressive (AR), Integrated (I), Moving Average (MA). Each of these phrases describes a different part of the mathematical model. Here are the equations ARIMA(p, d, q)(P, D, Q) where (p, d, q) is the non-seasonal part of the model and (P, D, Q) adalah bagian seasonal dari model is the seasonal part of the model [8].

$$\Phi_p(B)\Phi_p(B^s)\nabla^d\nabla_s^D Y_t = \theta(B)\theta(B^s)e_t \quad (3)$$

6 Artificial Neural Network (ANN)

Artificial Neural Network (ANN) is a multilayer neural network used to build the ANN model in this research [9]. The ANN architecture used includes, the number of input neurons in the input layer is 10 neurons, one hidden layer with 30 hidden neurons, and the output layer contains

one neuron to determine the class of hypertension (1) or not hypertension (0) with epoch or iteration 100. This ANN architecture can be seen in Figure 1.

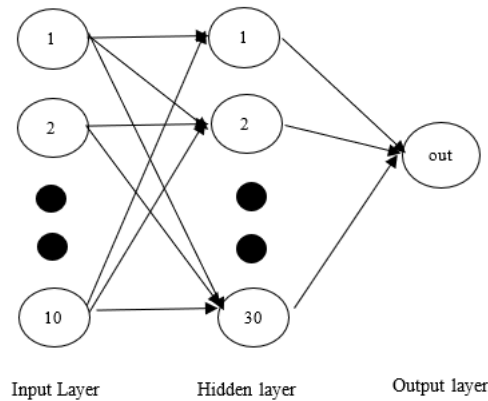


Fig. 1. Architecture ANN

4. Research Methods

The stages to complete this research are shown in Figure 2.

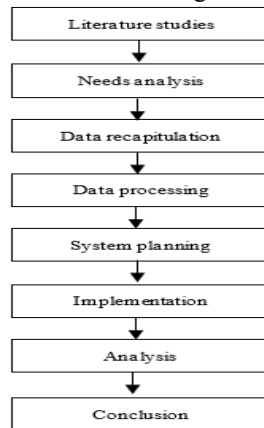


Fig. 2. Methodology flow diagram

Figure 2 explains that this research begins with collecting and then studying the literature used, the next step is to analyze the necessary needs, collect and then process the data used, design the system to be made, implement the design that has been made, test the system that has been made, analyze the results obtained based on the test results, the last step is to make conclusions from the research that has been done.

4.1 Data sources dan research variable

The data in this study are secondary data, namely data on the amount of petroleum production from August 9, 2021 to August 8, 2023, sourced from the PPDM (Public Petroleum Data Model)

database at PT "XYZ". The amount of crude oil production sourced from the PPDM database every day in barrels is the variable used.

4.2 Process Flow Diagram

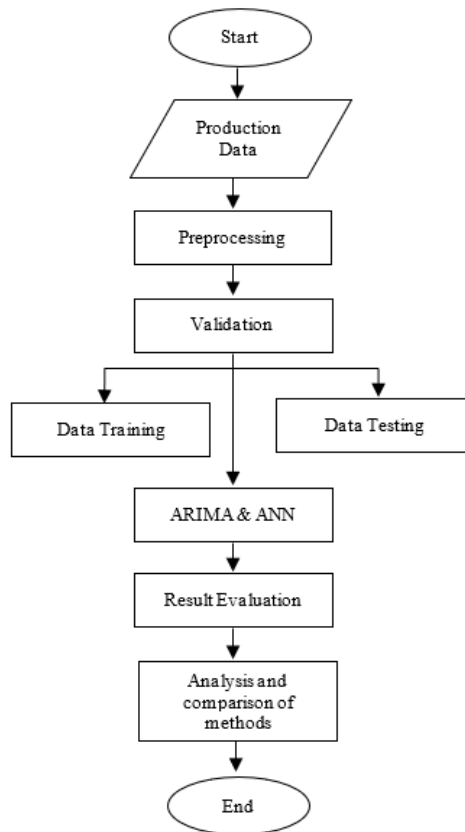


Fig. 3. Process flow diagram

Figure 3 explains the stages of the system, starting from preprocessing, validation, processing with the main algorithms, namely ARIMA, ANN and assessment accuracy.

5. Result and Discussions

From the dataset, data cleaning is then carried out by deleting the critical column because it is irrelevant to the prediction results and the data type is not in accordance with the method to be used. After performing data cleaning, data transformation is then carried out to change the string data type to a date and an integer. After passing the data preprocessing stage, data partitioning is carried out using 3 scenarios, namely 80:20, 70:30, and 60:40. After the data partitioning, each method is implemented with tools, namely KNIME. Figure 4 illustrates the flow of ARIMA and ANN modeling using KNIME. The nodes used in this workflow are also described in Table 1.

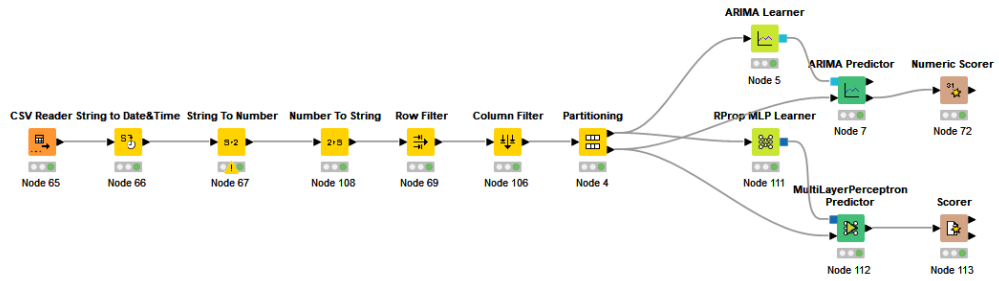












Fig. 4. Workflow of ARIMA and ANN using KNIME

Figure 4 above shows the workflow in ARIMA and ANN modeling using KNIME. Of the several nodes used [11], the description of each node will be explained in Table 1 below.

Table 1. Nodes used in ARIMA and ANN

Node	Node Description
 CSV Reader	Used to import a file that contains data to be predicted
 String to Date&Time	Used to convert a String to Date.
 String To Number	Used to convert a String to Number
 Row Filter	Used to filter the rows to be predicted
 Partitioning	Split the data per row to train and test data
 ARIMA Learner	Estimating the time series parameters of the ARIMA model
 ARIMA Predictor	Calculating the prediction of the estimated ARIMA model (forecast and production in sample)
 RProp MLP Learner	Forward partitioned input data using the kernel and parameters in the form of input layer, hidden layer, output layer
 MultiLayerPerceptron Predictor	Predicting the output value issued by the ANN learner
 Numeric Scorer	Used to display and calculate statistics between numerical column values (r_i) and predicted values (\hat{p}_i).

5.1 Autoregressive Integrated Moving Average (ARIMA) Experiment Results

In this ARIMA modeling, the best model selection is carried out. The ARIMA (p,d,q) models formed are as follows ARIMA (0,2,1), ARIMA (1,2,0), ARIMA (1,2,1), ARIMA (2,2,0), ARIMA (2,2,1), ARIMA (3,2,0) and ARIMA (3,2,1). The results of these parameters are obtained using the KNIME tool. From the modeling results of the Partition Method with Random Sampling, forecasting is then carried out and the accuracy results are shown in Table 2.

Table 2. The level of accuracy in the ARIMA method using error analysis

Well	Ratio	ARIMA (1, 1, 1)		ARIMA (0, 2, 1)		ARIMA (1,2,0)		ARIMA (1,2,1)		ARIMA (2,2,0)	
		MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE
MN1	80:20	1.604	2.629	1.740	3.166	2.343	3.766	1.653	3.010	2.187	3.539
	70:30	1.423	2.350	5.932	6.695	2.187	3.788	3.205	4.034	2.800	3.515
	60:40	1.414	2.432	5.774	6.288	2.099	3.614	1.591	2.767	1.928	3.305
MN2	80:20	3.406	5.298	4.194	7.835	5.119	9.759	3.896	7.327	4.666	8.620
	70:30	3.198	5.643	3.668	6.261	4.527	7.966	3.388	6.035	4.311	7.242
	60:40	2.943	5.860	3.197	5.678	4.260	7.090	3.084	5.513	3.944	6.73
MN3	80:20	1.112	2.591	1.100	2.542	1.752	3.132	1.682	3.038	1.602	2.909
	70:30	0.964	2.750	1.947	4.272	1.822	3.857	1.262	2.932	1.594	3.453
	60:40	0.849	2.400	1.570	3.380	1.462	2.850	1.122	2.275	1.401	2.766
MN4	80:20	2.353	4.509	2.388	4.999	3.218	6.193	14.378	15.217	2.857	5.495
	70:30	2.015	3.6657	2.261	4.469	3.245	5.834	4.641	6.001	2.975	5.419
	60:40	1.829	3.351	1.935	3.862	2.681	4.821	2.193	4.148	2.470	4.546
MN5	80:20	19.608	32.144	13.183	25.285	20.109	30.256	13.180	25.157	18.208	24.745
	70:30	14.986	32.361	10.741	21.445	16.277	26.218	10.619	21.356	13.085	20.155
	60:40	12.411	28.861	12.911	30.196	19.176	39.544	12.826	28.939	17.694	34.705

Based on the results of the best model selection, the smallest RMSE value is obtained in the ARIMA (1,1,1) model in the use of random sampling partitions with a comparison of 60:40.

5.2 Artificial Neural Network (ANN) Experiment Results

For this artificial neural network (ANN) modeler, using random sampling with the number of input neurons in the input layer of 10 neurons, one hidden layer with 30 hidden neurons, and the output layer contains one neuron to determine the class of hypertension (1) or not hypertension (0) with epoch or iteration 100, Figures 5 and 6 show the results of data classification based on well production after training and testing data.

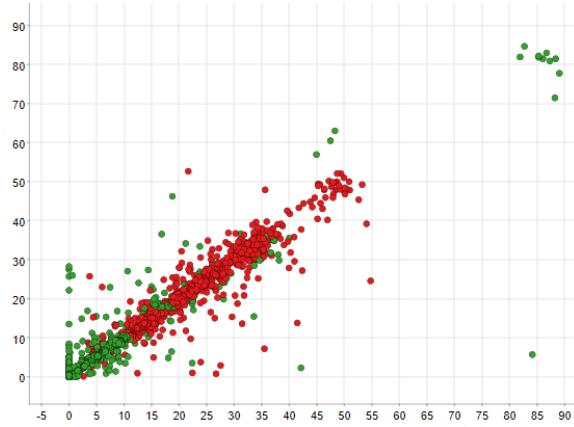


Fig. 5. Classification result of dataset training

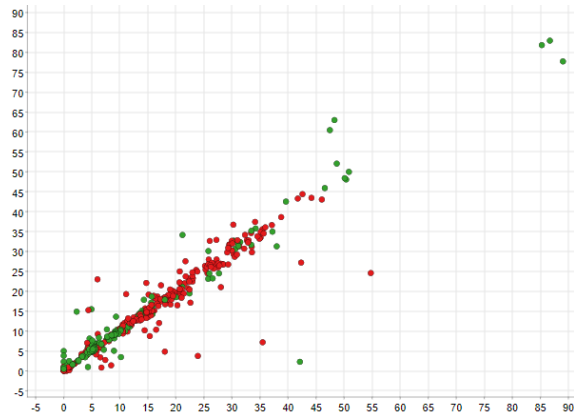


Fig. 6. Classification result of dataset testing

In this partition stage (training and testing data) using random sampling. Next, make an ANN classifier. The following ANN method accuracy results are shown in Table 3.

Table 3. The level of accuracy in the ANN method using error analysis

Well	Ratio	ANN	
		MAE	RMSE
MN1	80:20	0.910	0.975
	70:30	0.881	0.943
	60:40	0.869	0.936
MN2	80:20	0.335	0.806
	70:30	0.363	0.836
	60:40	0.428	0.908
MN3	80:20	0.006	0.082
	70:30	0.090	0.095
	60:40	0.013	0.110
MN4	80:20	0.027	0.165
	70:30	0.959	0.979

Well	Ratio	ANN	
		MAE	RMSE
MN5	60:40	0.948	0.973
	80:20	0.076	0.277
	70:30	0.026	0.162
	60:40	0.944	0.970

6. Conclusion

From the results of the experiment, some conclusions can be drawn:

1. In testing the dataset using ANN using random sampling partitions produces different levels of accuracy for each production well as described above, that random sampling partitions with a ratio of 80:20 have the best accuracy.
2. In testing the dataset using ARIMA using a random sampling partition ratio of 60:40 with parameters (1,1,1) produces the best accuracy level of each production well.
3. From the ARIMA and ANN tests, it can be said that the ANN test has higher accuracy. This can be seen from the average accuracy results with several treatments.

References

- [1] Prameswari, W. C., Susilaningrum, D., & Suhartono, S. (2016). Pemodelan Produksi Minyak dan Gas Bumi Pada Platform "MK" di PT "X" Menggunakan Metode ARIMA, Neural Network, dan Hibrida ARIMA-Neural Network. *Jurnal Sains dan Seni ITS*, 5(2).
- [2] Nasional, D. E. (2014). *Outlook Energi Indonesia 2014*. Kementerian Energi dan Sumber Daya Mineral. Jakarta. ISSN 2527-3000
- [3] Yazid, E., Yusuf, M., & Herlina, W. (2018). Evaluasi Kinerja Water Treatment Injection Plant Untuk Pressure Maintenance Pada Sumur X Struktur Y Di PT Pertamina EP Asset 2 Pendopo Field. *Jurnal Pertambangan*, 2(4), 15-23.
- [4] Samperuru, D. (2007). *Dari Mana Datangnya Minyak Bumi*. Buku Pintar Migas Indonesia, 1-17
- [5] A. Pwasong and S. Sathasivam, "Forecasting crude oil production using quadratic regression and layer recurrent neural network models" vol. 20001, p. 20001, 2016
- [6] Santoni, M. M., Chamidah, N., & Matondang, N. (2020). Prediksi Hipertensi menggunakan Decision Tree, Naïve Bayes dan Artificial Neural Network pada software KNIME. *Techno. Com*, 19(4), 353-363.
- [7] Omekara, C. O., Okereke, O. E., Ire, K. I., & Okangba, C. O. (2015). ARIMA modeling of Nigeria crude oil production. *Journal of Energy Technologies and Policy*, 5(10), 1-5.
- [8] Wiryawanto, T. M. P., Hawani, Z., & Ramadhani, M. A. (2023). Comparison of Support Vector Machine (SVM) and Autoregressive Integrated Moving Average (ARIMA) Methods for Predicting Air Quality Using Python and KNIME. *J Statistika: Jurnal Ilmiah Teori dan Aplikasi Statistika*, 16(1), 384-394.
- [9] Santoni, M. M., Chamidah, N., & Matondang, N. (2020). Prediksi Hipertensi menggunakan Decision Tree, Naïve Bayes dan Artificial Neural Network pada software KNIME. *Techno. Com*, 19(4), 353-363.