KOMPUTASI: JURNAL ILMIAH ILMU KOMPUTER DAN MATEMATIKA
Vol.20 (2)(2023), 160-168 p-ISSN: 1693-7554, e-ISSN:2654-3990

Clean Water Demand Prediction Model Using The Long Short Term Memory (LSTM) Method

Delviani Permata Sari[1], **Lita Karlitasari**[2,*], **Fajar Delli Wihartiko**[3]

[1,2,3]Department of Computer Science, Faculty of Mathematics and Natural Science, Pakuan University, Bogor, West Java, 16143, Indonesia

Abstract

Cities or districts as population centers with various service facilities, really need the provision of clean water. The agency that handles clean water in Indonesia is the Regional Drinking Water Company (PDAM). PDAMs were established in every city and district in Indonesia as agencies that serve the community's need for clean water. One of them is the Regional Public Company (Perumda) Tirta Pakuan and as time goes by the number of customers will definitely increase so that the need for clean water will also increase. The purpose of this research is to create a Clean Water Demand Prediction Model using the Long Short Term Memory (LSTM) Method to find the most optimal modeling. The data in this study were obtained from data reports is from Perumda Tirta Pakuan. The prediction model development process is carried out through Visual Studio Code tools. To find a model with the smallest error rate using various ratios, namely 80:20, 70:30, 60:40, and 50:50, then testing is also carried out based on the number of different hyperparameter values in batch sizes 5, 10, 15, 20, 25 and max epoch 50, 100, 150, 200, 250. From all the experiments that have been carried out, the most optimal is batch size 5 and epoch 50 with a ratio of 60:40 for water production to get RMSE 0.4862 and MAPE 2.5252% while for the amount of water use with a ratio of 50:50 get RMSE 0.4674 and MAPE of 2.5163%.

Keywords: *Clean Water Needs; Data Mining; Prediction; LSTM*

1. Introduction

Water is one of the most important assets and is the main thing for human consumption. Cities or districts as population centers with various service facilities, really need the provision of clean water [1]. The agency that handles clean water in Indonesia is the Regional Drinking Water Company (PDAM). PDAMs were established in every city or district in Indonesia as agencies that serve the community's need for clean water [2]. One of them is PDAM Tirta Pakuan which in 2019 has changed its name to the Regional Public Company (Perumda) Tirta Pakuan and as time goes by the number of customers will definitely increase so that the need for clean water will also

[*]**Corresponding author:** ***E-mail adress:*** *lita.karlitasari@unpak.ac.id*
Received: 12 May 2023, Accepted: 15 July 2023 and available online 30 July 2023
https://doi.org/10.33751/komputasi.v20i2.8060

increase. Therefore prediction or forecasting regarding the need for clean water is needed so that Perumda Tirta Pakuan is not excessive and not lacking in providing clean water.

The need for water is categorized into domestic and non-domestic water needs [12]. The use of water among the people has different uses in each region, the use of water in one area is definitely different from another [18]. High water consumption causes the need for clean water supplies to continue to increase while the supply of clean water each year continues to decrease [15]. Population growth must be followed by the availability of healthy and sufficient clean water [13]. Prediction tries to find answers as close as possible and does not have to give a definite answer to what will happen [17]. Forecasting is the process of predicting future events based on past data which is measured periodically and will form a time series of data.

Deep Learning is part of machine learning which is tasked with studying available data through existing algorithms. The learning process can be done through three options, namely directed learning, semi-directed, and not directed. Directed learning allows algorithms to learn based on available and sufficient data, whereas semi-directed learning is based on available but insufficient data, whereas non-directed learning relies on algorithms to learn on their own without any data input [3]. Deep Learning has several algorithms in it, one of which is the Long Short Term Memory (LSTM) algorithm and the best forecast is based on the prediction error rate.

Long Short Term Memory (LSTM) is a variant model of the Recurrent Neural Network (RNN) that can remember long-term information or long term dependency and has memory cells that function to store previous information [4]. The LSTM method can be used for forecasting cases by making accurate predictions of a variable. The smaller the error rate produced, the more precise the method is in predicting both for long and short periods of time [5]. The purpose of this study is to create a Clean Water Demand Prediction Model using the Long Short Term Memory (LSTM) Method to find the most optimal modeling with the smallest error rate using various ratios and then testing based on the number of different hyperparameter values on batch size and max epoch.

2. Methods

The research method used is the Cross Industry Standard Process for Data Mining (CRISP-DM) which can be seen in Figure 1.

Figure 1. RCRISP-DM method [6]

2.1. Business Understanding

In this study, the business goal to be achieved is to create a modeling architecture with the LSTM deep learning algorithm to predict clean water needs. In the LSTM architectural model, several tests will be carried out with various ratios and different initialization of hyperparameters in each test to obtain optimal accuracy results.

2.2. Data Understanding

At the data understanding stage, the dataset acquisition process is carried out which will be used to carry out the modeling process, while the data used in this study is Perumda Tirta Pakuan's daily time series data from January 2022 to December 2022. After that, first create a csv file on Microsoft Excel so that the data can be read by the program, the total data is 365 data and consists of 3 attributes namely date, water production, and amount of water usage.

2.3. Data Preparation

Data preparation is a preprocessing stage to prepare datasets to be used in the modeling process. In this research, the data preparation process is carried out by selecting data using 2 variables, namely water production and the amount of water usage, then carrying out the normalization process is a method by carrying out a linear transformation of the original data which aims to process scaling the attribute values of the data so that can be in a certain range and change the data size to be smaller without having to change the original data [7]. One of the normalization methods that can be used is Min-max normalization with the following formula:

$$X' = \frac{X_i - X_{min}}{X_{max} - X_{min}} \tag{1}$$

Then proceed with the distribution of training data and test data using various ratios which can be seen in Table 1.

Table 1. Comparison of Training Data and Test Data

No	Ratio	Train Data	Test Data
1	80:20	292	73
2	70:30	256	109
3	60:40	219	146
4	50:50	183	182

2.4. Modeling

This stage is the stage of determining data mining techniques, modeling parameters, and data mining algorithms to build the research model architecture. This research conducts forecasting of water production data and the amount of water usage using a deep learning algorithm model, namely Long Short Term Memory (LSTM) was created to overcome long-term memory problems in RNNs [16]. Memory cells are used to overcome the occurrence of vanishing gradients in the RNN when processing long sequential data [14] which can be seen in Figure 2.

The LSTM method consists of two activation functions, namely the sigmoid and tanh functions, there are also three gates that function to control the use and update of previous information, namely the Forget gate, Input gate, and Output gate which are designed to be able to read, store, and update previous information [9]. In the LSTM process there are equations for performing calculations in the layers which can be seen as follows:

$$f_t = \sigma(W_{fx}x_t + W_{fh}h_{t-1} + b_f) \tag{2}$$
$$i_t = \sigma(W_{ix}x_t + W_{ih}h_{t-1} + b_i) \tag{3}$$
$$\bar{C}_t = \tanh(W_{cx}x_t + W_{ch}h_{t-1} + b_c) \tag{4}$$
$$C_t = f_t * C_{t-1} + i_t * \bar{C}_t \tag{5}$$
$$o_t = \sigma(W_{ox}x_t + W_{oh}h_{t-1} + b_o) \tag{6}$$
$$h_t = o_t * \tanh(C_t \tag{7}$$

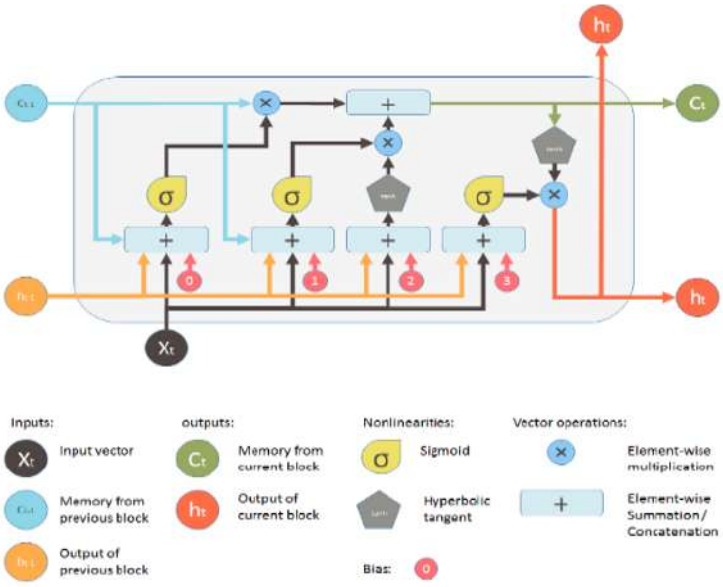

Figure 2. Illustration of LSTM [8]

2.5. Evaluation

At this stage an evaluation of the performance and performance of the algorithm model will be carried out. In this study, the model evaluation process was carried out using the Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE) calculation model errors which aimed to determine the level of forecasting accuracy. The smaller the level of error generated, the more precise a model is in forecasting.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}{n}} \tag{8}$$

$$MAE = \frac{\sum_{i=1}^{n} \frac{|Y_i - \bar{Y}_i|}{Y}}{n} * 100 \tag{9}$$

MAPE has a range of values that can be used as a benchmark in decision making, namely $<$10% for the very good category, 10-20% for the good category, 20-50% for the feasible category, and $>$ 50% for the bad model conditions [10].

3. Result and Discussion

In this study, the data was obtained from the Perumda Tirta Pakuan data report for the period January 2022 to December 2022 with a total of 365 data, the research variables used were water production and total water usage, then in the distribution of training and testing data using various ratios namely 80:20, 70:30, 60:40, and 50:50.

3.1. Modeling Process

In the built model, scenario trials will be carried out based on the number of different batch size and max epoch parameters. Hyperparameters are parameters that are set before the model learning process begins which can be seen in Table 2.

The LSTM method uses Hidden Neurons as a parameter to determine the number of hidden layers in the deep learning model, then Loss Function to measure the performance of the modeling between the actual data and the predicted results, in order to minimize the error function the Adam Optimizer model is used with a default learning rate of 0.001.

Table 2. Hyperparameter Architecture and Initiation

No	Type	Value/Name
1.	Layer	LSTM
2.	Neuron Hidden	50
3.	Optimizer	Adam
4.	Learning Rate	0.001
5.	Batch Size	5, 10, 15, 20, 25
6.	Epoch	50, 100, 150, 200, 250
7.	Loss Function	MSE

Usually large batch sizes are used whenever possible computational acceleration. If use a small batch size, it will take a very long time. Batch sizes that are too large will produce less than optimal results. The larger the batch size, the less accurate the results will be. For this reason, deep learning users will usually weigh between the capabilities of the tools used, the time that must be spent on a long training process and the optimization of results [11].

3.2. Model Testing Accuracy Results

In this study, modeling test scenarios were carried out using various ratios and in each trial the number of parameters was made in a gradual scenario and continued to increase in each trial so that later a graph would be selected based on the model results with the smallest error rate in each ratio.

Table 3. Water production modeling trial result

Batch Size / **Epoch**	**80:20**		**70:30**		**60:40**		**50:50**	
	RMSE	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**
5 50	0,2021	2,6099%	0,8533	2,6475%	**0,4862**	**2,5252%**	0,6421	2,6159%
10 100	0,7091	2,7021%	0,9181	2,6777%	0,5317	2,5776%	0,2646	2,6605%
15 150	0,6532	2,7200%	0,7380	2,6903%	0,3386	2,5378%	0,8963	2,7398%
20 200	0,7150	2,7490%	0,5488	2,6980%	0,683	2,6499%	0,6846	2,8483%
25 250	0,5476	2,7209%	0,5816	2,6961%	0,6314	2,6949%	0,7535	2,9561%
Rata-rata	0,5654	2,7003%	0,7279	2,6819%	0,5341	2,5970%	0,6482	2,7641%

Based on Table 3 the results of modeling trials for water production get the smallest MAPE of 2,5252% which is obtained from the model using a 60:40 ratio then batch size 5 and epoch 50. Figure 3 is a graph of the predicted results for water production where the blue line depicts training data, for data testing at this ratio starting from August. The yellow line which is the predicted value shows the line even though it is not perfect but has followed the up and down flow of the red valid or actual line.

Based on Table 4 the results of modeling trials for amount of water usage get the smallest MAPE of 2,5163% which is obtained from the model using a 50:50 ratio then batch size 5 and

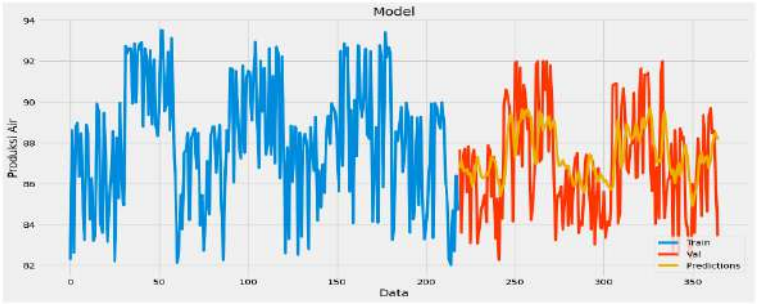

Figure 3. Prediction water production

Table 4. Amount of water usage modeling trial result

Batch Size / Epoch	80:20		70:30		60:40		50:50	
	RMSE	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**	**RMSE**	**MAPE**
5								
	0,2249	2,7251%	0,3035	2,6039%	0,3683	2,5576%	0,4674%	**2,5163%**
50								
10								
	0,3886	2,7450%	0,4437	2,6692%	0,3005	2,6148%	0,5870	2,5567%
100								
15								
	0,4203	2,7609%	0,2527	2,6491%	0,1516	2,5868%	0,3736	2,5429%
150								
20								
	0,2846	2,7305%	0,3080	2,7126%	0,3795	2,6920%	0,6848	2,5673%
200								
25								
	0,2872	2,7597%	0,1195	2,6715%	0,1998	2,7892%	0,5052	2,5766%
250								
Rata-rata	0,3211	2,7442%	0,2854	2,6612%	0,2799	2,6480%	0,5236	2,5519%

epoch 50. Figure 4 is a graph of the predicted results for amount of water usage the same as the previous graph where the blue line represents training data and the yellow line is the predicted value. For valid or actual values that are colored red in this ratio starting from July.

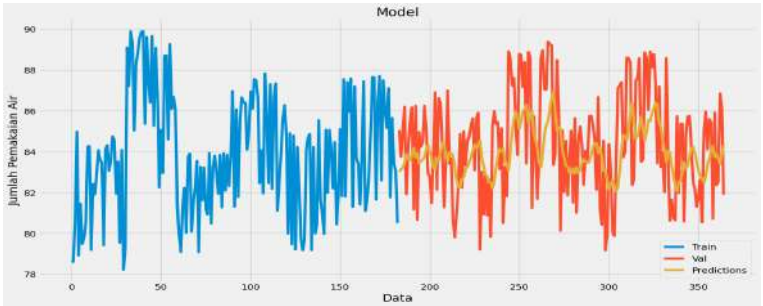

Figure 4. Prediction amount of water usage

Of all the experiments that have been carried out, the model that will be proposed in this study is the model that produces the smallest MAPE values for both variables, namely using batch size 5 and epoch 50 with a ratio of 60:40 for water production while for amount of water usage using a ratio of 50:50.

3.3. Predictions for the Next 7 Days

In this study, after carrying out modeling test scenarios using various ratios and with increasing hyperparameter values in each test. The following results have been obtained from modeling with the best hyperparameters based on the smallest MAPE error rate which can be seen in Table 5.

Table 5. Amount of water usage modeling trial result

No	Date	Prediction Results	
		Water Production (liter/second)	Amount of Water Usage (liter/second)
1.	01/01/2023	87,17	83,96
2.	02/01/2023	87,36	84,71
3.	03/01/2023	87,74	84,36
4.	04/01/2023	87,51	84,16
5.	05/01/2023	88,03	84,28
6.	06/01/2023	88,68	84,10
7.	07/01/2023	88,51	83,77
Total		615,00	589,04
Number of previous weeks		610,43	586,27

The performance level of accuracy obtained in this study is influenced by the hyperparameter value initiated during the trial, especially the batch size and max epoch values. In this study, from all the experiments that have been carried out, the most optimal is batch size 5 and epoch 50 with a ratio of 60:40 for water production while for the most optimal amount of water use is a ratio of 50:50 and any predicted data obtained will be added into new data one by one to become additional memory, then the prediction results for the next week the water production and the amount of water usage show an increase from the previous week but not too significantly, so that Perumda Tirta Pakuan's production facilities are still sufficient to serve its customers.

4. Conclusion

In this study, based on the results of the entire experiment with the distribution of training and testing data using various ratios and testing was also carried out based on the number of different hyperparameter values in the batch size and max epoch, it was found that the most optimal were batch size 5 and epoch 50 with a ratio of 60:40 for water production gets RMSE 0.4862 and MAPE is 2.5252% while for total water use with a ratio of 50:50 gets RMSE 0.4674 and MAPE is 2.5163%. For this reason, it can be concluded that the LSTM deep learning algorithm shows predictive results with an error rate of below 10% which is included in the very good forecasting category range, then based on the prediction results data for water production and the amount of water use the results have increased compared to last week. before but not too significant, so that Perumda Tirta Pakuan can still serve the number of existing customers and does not need to add production facilities.

5. Acknowledgement

Thank you to Allah SWT for all the blessings so that the writer can finish this journal, thanks to the supervisor lecturer, beloved parents and family who have always provided support, prayer, and motivation to the writer. And also to the parties from Perumda Tirta Pakuan who have allowed and helped to do research.

References

[1] Brahmanja, A. Ariyanto, and K. Fahmi, "PREDIKSI JUMLAH KEBUTUHAN AIR BERSIH BPAB UNIT DALU - DALU 5 TAHUN MENDATANG (2018) KECAMATAN TAMBUSAI KAB ROKAN HULU," 2013.

[2] V. N. P. Hasan, W. F. Mahmudy, and M. Z. Sarwani, "PEMODELAN REGRESI NON LINEAR MENGGUNAKAN ALGORITMA GENETIKA UNTUK PREDIKSI KEBUTUHAN AIR PDAM KOTA MALANG," vol. 59, pp. 59–65, 2016.

[3] F. Fadli, S. Suwilo, and M. Zarlis, "Model Prediksi Data Besar Distribusi Produk Farmasi: Analisis Kinerja Model Deep Learning," CSRID (Computer Sci. Res. Its Dev. Journal), vol. 14, no. 1, p. 68, 2022, doi: 10.22303/csrid.14.1.2021.79-91.

[4] E. Supriyadi, "Prediksi Parameter Cuaca Menggunakan Deep Learning Long-Short Term Memory (Lstm)," J. Meteorol. dan Geofis., vol. 21, no. 2, p. 55, 2021, doi: 10.31172/jmg.v21i2.619.

[5] L. Wiranda and M. Sadikin, "Penerapan Long Short Term Memory Pada Data Time Series Untuk Memprediksi Penjualan Produk Pt. Metiska Farma," J. Nas. Pendidik. Tek. Inform., vol. 8, no. 3, pp. 184–196, 2019.

[6] M. Stang, M. Bohme, and E. Sax, "Applied Machine Learning: Reconstruction of Spectral Data for the Classification of Oil-Quality Levels," Technol. Sci., vol. 5, pp. 1–13, 2019, [Online]. Available: www.isres.org.

[7] P. Sugiartawan, A. A. J. Permana, and P. I. Prakoso, "Forecasting Kunjungan Wisatawan Dengan Long Short Term Memory (LSTM)," J. Sist. Inf. dan Komput. Terap. Indones., vol. 1, no. 1, pp. 43–52, 2018, doi: 10.33173/jsikti.5.

[8] S. Yan, "Understanding LSTM and Its Diagrams", 2016, [Online]. Available: https://medium.com/mlreview/understanding-lstm-and-its-diagrams-37e2f46f1714.

[9] A. Faadilah, "Analisis Sentimen Pada Ulasan Aplikasi Tokopedia di Google Play Store Menggunakan Metode Long Short Term Memory", 2020.

[10] W. W. K. Wardani, "Prediksi Harga Saham Syariah Menggunakan Metode Recurrent Neural Network-Long Short Term Memory", 2021.

[11] N. Rochmawati, H. B. Hidayati, Y. Yamasari, H. P. A. Tjahyaningtijas, W. Yustanti, and A. Prihanto, "Analisa Learning Rate dan Batch Size pada Klasifikasi Covid Menggunakan Deep Learning dengan Optimizer Adam," J. Inf. Eng. Educ. Technol., vol. 5, no. 2, pp. 44–48, 2021, doi: 10.26740/jieet.v5n2.p44-48.

[12] Asta, "Analisis Kebutuhan Air Bersih Dan Distribusi Jaringan PDAM Persemaian Kota Tarakan (Studi Kasus Kecamatan Tarakan Barat)," vol. 2, no. 1, pp. 61–68, 2018, [Online]. Available: http://jurnal.borneo.ac.id/index.php/borneoengineering.

[13] Bahar and S. A. Yahya, "Penerapan Algoritma Backpropagation Untuk Prediksi Kebutuhan Air Bersih pada PDAM Intan Banjar," 2019.

[14] M. A. Faishol, Endroyono, and A. N. Irfansyah, "Prediksi Polusi Udara Perkotaan Di Surabaya Menggunakan Recurrent Neural Network - Long Short Term Memory," JUTI J. Ilm. Teknol. Inf., vol. 18, no. 2, p. 102, 2020, doi: 10.12962/j24068535.v18i2.a988.

[15] K. Istiqara, M. T. Furqon, and Indrianti, "Prediksi Kebutuhan Air PDAM Kota Malang Menggunakan Metode Fuzzy Time Series Dengan Algoritma Genetika," vol. 2, no. 1, pp. 133–142, 2018.

[16] M. R. Sopany, D. E. Herwindianti, and J. Hendryli, "Prediksi Kelembapan Tanah Pada Tingkat Kecamatan di Wilayah Bogor Dengan Metode CNN LSTM," Comput. J. Comput. Sci. Inf. Syst., vol. 6, no. 1, p. 1, 2022, doi: 10.24912/computatio.v6i1.15740.

[17] A. L. Putra and A. Kurniawati, "Analisis Prediksi Harga Saham PT. Astra International Tbk Menggunakan Metode Autoregressive Integrated Moving Average (ARIMA) dan Support Vector Regression (SVR)," J. Ilm. Komputasi, vol. 20, no. 3, pp. 417–423, 2021, doi: 10.32409/jikstik.20.3.2732.

[18] B. Putro, "Prediksi Jumlah Kebutuhan Pemakaian Air Menggunakan Metode Exponential Smoothing (Studi Kasus: PDAM Kota Malang)," Biomass Chem Eng, vol. 3, no. 2, p. , 2018.