Jurnal Presipitasi
Media Komunikasi dan Pengembangan Teknik Lingkungan
e-ISSN: 2550-0023

Vol 20, No 3, 2023, 581-589
Article available at homepage presipitasi

Original Research Article

Water Demand Modeling using Machine Learning Method in Bandung City, Indonesia

Evi Afiatun[1]*, Yonik Meilawati Yustiani[1], Dennis Anugerah Tanra[1]
[1] Department of Environmental Engineering, Faculty of Engineering, Universitas Pasundan, Bandung, 40153, Indonesia
*Corresponding Author, email : eviafiatun@unpas.ac.id

Abstract
This research was conducted at Bandung City with the aim of building a model using machine learning methods so that it can estimated clean water demands in Bandung City, as well as knowing the external factors that are considered to affect the model. Machine learning is a part of Artificial Intelligence (AI) discipline. The modeling is carried out using independent variables in the form of climate parameters which are rainfall, rainy days, and humidity, as well as the dependent variable in the form of drinking water needs which are represented by raw water. Data collection is done through secondary data. The model was built by using the TPOT module, and produces the AdaBoost.R2 algorithm as the most optimal model, by using the model algorithm, the best sub-model is produced with the most influential external factors, namely rainy days and humidity which has an MAE of 326,077.70 and a MAPE of 4.75%. This model is compared with the ARIMA model which has an MAE of 330,672.088 and an MAPE of 5.07%.

Keywords: Bandung city; artificial intelligence; machine learning; climate parameters; raw water

1. Introduction
Bandung city is one of the most densely populated cities in Indonesia. Based on data from the Central Statistics Agency (BPS) and the Tirtawening Regional Drinking Water Company (PDAM) in Bandung city, in 2022 Bandung city has population of 2,530,448 people and the volume of drinking water distributed is 37,334,607 m^3. Since 2012, the main problems with the drinking water supply system, including in Bandung city, are the limited availability of raw water and the high rate of water loss, while the consumption of drinking water is quite large, causing a large gap in meeting the needs of drinking water. This is supported by data for 2022 which shows a water loss rate of 43%. (Andani, 2012; Afiatun et.al., 2018). The quality of surface water as a source of raw water changes from time to time which is influenced by conditions in the upstream, pollution along the river flow, as well as climate and weather conditions. (Afiatun et.al., 2019).

Based on these conditions, modeling the availability of drinking water is deemed necessary. Modeling the availability of drinking water is very dependent on several phenomena, one of which is the influence of climate. Modeling for the short term is very important for the efficient management of the water available in the reservoir and the equipment associated with the reservoir, while modeling for the long term (annual) is very important at the design stage of the distribution pipeline network (Antunes et al, 2018 and Haque et.al., 2017).

Today, artificial intelligence has an important role in modeling and simulation. One part of the scientific field of artificial intelligence that can be used for this research is Machine Learning, in machine learning models it is divided into two groups based on how the computer learns the data provided, namely supervised learning and unsupervised learning (Bishop, 2006; Duerr et.al., 2018). Basically, machine learning is a series of programs created to generate mathematical models.

Research on forecasting the need for clean water for the short term by Antunes et al. (2018), showed that modeling the need for drinking water using machine learning is able to produce a more accurate model when compared to conventional methods.

Modeling was carried out using independent variables in the form of climate parameters which include rainfall, rainy days and air humidity, as well as a dependent variable in the form of drinking water needs represented by raw water. These climate variables greatly influence the availability of raw water, including in the city of Bandung, where raw water availability modeling has never been built using Machine Learning.

The purpose of this research is building a model using machine learning methods so that it can estimated clean water demands in Bandung City, as well as knowing the external factors that are considered to affect the model. Therefore, machine learning is a method that will be used for modeling the water needs of the Bandung City, because this method has never been used in Bandung City, this method is also considered more accurate than conventional methods.

2. Methods

A representative model is needed to represent the actual conditions in order to help answer an existing problem. In this problem the model will be created using a time series data set. The implementation of this research is described as follows.

2.1 Literature Study

Literature study is carried out to get the basics and supporting knowledge in research, includes basic theories regarding modeling using machine learning along with programming support science, as well as covering algorithms that are generally used in similar research.

2.2 Secondary Data Collection

The secondary data used for this modeling are data on climate, as well as drinking water data which includes the total raw water volume, the total distribution volume of drinking water, and the total official consumption of drinking water. The data used is monthly data in the form of a time series with a period from 2014 to 2020.

2.3 Data Processing

Secondary data will be processed and built into a model using the Machine Learning method. The model that will be created using this method is a genetic algorithm model that can automate the process of searching for model algorithms until an optimal model is obtained. The tool that can be used to help process the whole of data is the Python programming language.

2.3.1 Data Preparation

Before building a model using the machine learning method, the data will go through the preparation stages first. Brownlee (2020) defines data preparation as a transformation from data that is still "raw" into data in a form that is more suitable for modeling.

In this study, the data preparation process will only include the process of testing the correlation of external variables to output. Meanwhile, the cleaning process or data cleaning is not carried out, because the data is complete and there is no missing data.

Correlation testing is carried out to determine the effect of each external variable on output, as well as how strong the influence of these external variables is on output. In testing the correlation used linear regression method and Pearson correlation.

2.3.2 Model Building

The model was built using the machine learning method with a tool in the form of the TPOT programming module. In its work, TPOT uses a genetic algorithm, which is a search technique to find an

answer to an existing problem, so that an optimal answer can be produced, where the technique adapts terms found in evolutionary biology, such as population, generation, mutation, and inheritance.

After the algorithm is fit on the data set, the resulting MAE is much different compared to the previous MAE value. The models with rainfall and rainy day variables are also different, this means that the AdaBoost.R2 algorithm has considered the weight of each input. By using the AdaBoost.R2 algorithm, the actual value of R2 can be ignored, because AdaBoost.R2 will use the weight of each input value, as in equation (1) below:

$$w_i = \frac{1}{\Sigma w_i} \quad (1)$$

Where w is the weight value taken from the number of rows in the data, then the initial weight value in the first iteration of the AdaBoost.R2 algorithm, wi = 1/72. The R2 value is only used during the preparation process, to reduce the computer's workload.

Then, the algorithm will make initial predictions based on the median and raw water values, so that the loss function for each prediction is obtained, as shown in equation (2) below, Drucker (1997):

$$L_i = y_i^{(p)}(x_i) - y_i \quad (2)$$

Where Li is the loss function, yi(p)(xi) is the predicted value, and yi is the original raw water value. In Figure 4 above, the prediction is a dotted line. Then, the loss function is averaged using equation (3) as follows:

$$L = \sum_{i=1}^{N_i} L_i p_i \quad (3)$$

After that, based on Drucker (1997), the average loss will be used as a measure of confidence, which will be used to update the initial weight value, as in equation (4) below:

$$\beta = \frac{L}{1-L} \quad (4)$$

Is the equation for the measure of confidence, and:

$$w_i \rightarrow w_i^{\beta(1-L_i)} \quad (5)$$

Equation (5) above is the updated input weight. The process will be repeated continuously until it reaches n_estimation, in this case, the n_estimation generated from genetic programming is 100, then the process will be repeated until it reaches the 100th iteration.

2.3.3 Model Validation

Modeling results must be validated, in order to see how far the accuracy of the model that has been built, by comparing the modeled data with existing data. Validation is carried out on the validation data set using the MAE (Mean Absolute Error) method.

The purpose of using the MAPE method is as a benchmark for the feasibility of modeling values, where a model with MAPE ≤10% - 25% is considered a feasible validation value. Lewis (1982), interprets the MAPE values in Table 1.

Table 1. MAPE value interpretation

MAPE Value	Interpretatiom
≤10	Very Accurate Modeling Results
10 - 20	Good Modeling Results
20 - 50	Reasonable Modeling Results
> 50	Inadequate Modeling Results

2.3.4 Model Accuracy Comparison

The model that has been built will then be validated using the validation data set, then the model with the highest accuracy will be compared with the ARIMA model. The ARIMA model is often used as a comparison model, because it is considered the most consistent model for time series data sets.

Comparison of model accuracy aims to analyze whether modeling using machine learning methods can be more representative compared to conventional models. (Brownlee, 2020; Siami-Namini, et.al., 2018).

The ARIMA model is a model that will be used as a comparison model to models that have been created and developed with the TPOT module. ARIMA is an integration of two different models, namely the Auto Regression (AR) model and the Moving Average (MA) model.(Brownlee, 2020; Putri et.al., 2021).

3. Result and Discussion

3.1. Models Development

The amount of data trained using TPOT is 72 lines of data, and the number of validation data is 12 lines of data. The model was developed using combinations of external variables on raw water. The external variables are rainfall variable, rainy days variable, humidity variable, and combinations between variables

Each data set produces a new algorithm for each generation, as well as the average MAE value for each of these generations. The average MAE is the average value of the entire MAE for the 50 populations in each generation. The resulting average MAE is getting lower with increasing generations. The last generation, namely the 5th generation, produces the lowest average MAE value, where the MAE value is the value that will be compared between each model. Table 2 below describes in detail the average MAE values for the models produced by the last generation.

Table 2. The average MAE value of the best generation model

Variables	R^2	MAE Average
Rainfall	0.156	719,782.94
Rainy Days	0.365	719,782.94
Humidity	0.252	737,240.98
Rainfall + Rainy Days	0.261	744,878.54
Rainfall + Humidity	0.204	731,526.35
Rainy Days + Humidity	0.309	737,696.51
Rainfall + Rainy Days + Humidity	0.258	753,625.53

From the results of model development using a genetic algorithm, all variants produce the same algorithm, namely the AdaBoost.R2 algorithm or Adaptive Boosting Regressor. Based on the development of this model, it was found that the lowest average MAE value was 719,782.94 in the model with Rainfall and Rainy Day variables. However, the two models cannot be used as a reference yet, because the MAE value is still the average value of the entire population. Therefore, the best algorithm that has been produced by TPOT will be fitted to the entire data set, so that the actual MAE value for each variant can be known. Table 3 below presents the actual MAE values.

Table 3. The actual MAE value

Variables	R^2	MAE
Rainfall	0.156	604,466.59
Rainy Days	0.365	613,980.60
Humidity	0.252	631,160.15
Rainfall + Rainy Days	0.261	579,665.63
Rainfall + Humidity	0.204	524,894.96
Rainy Days + Humidity	0.309	549,679.41
Rainfall + Rainy Days + Humidity	0.258	541,252.64

Then, after the algorithm determines the first weight value, the algorithm sorts the data set from the lowest value to the highest value, then the algorithm determines the median for the input. For example, for the temperature variable, the median is 203.3.

Then the algorithm will start to initialize the prediction function ht : x --> y. Figure 3 below is a graph of the prediction function.

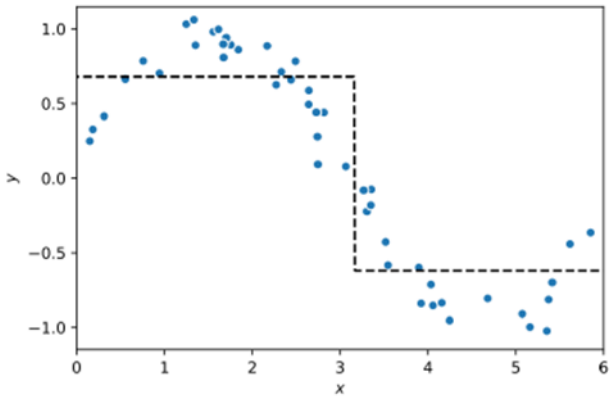


Figure 3. Initial AdaBoost.R2 prediction function

The modeling results must be validated, so that it can be seen to what extent the accuracy of the model that has been built by comparing the modeling data with existing raw water data. After knowing the actual MAE for the training data set, model validation is then carried out for each sub-model. The modeling results must be validated, so that the accuracy of the model that has been built can be seen by comparing the modeling data with existing raw water data. In addition, MAPE was calculated as a benchmark for the feasibility of a model. Validation was carried out on 14.29% of the total data or 12 months of data, and was carried out through the same code description as in the fitting process. Validation was carried out on the validation data set, which is a data set for the last 12 months, and the MAE method was used. Table 4 below is the result of the validation:

Table 4. Validation results on data sets

Variables	MAE
Rainfall	472,035.51
Rainy Days	484,460.42
Humidity	435,703.10
Rainfall + Rainy Days	468,543.21
Rainfall + Humidity	364,889.53
Rainy Days + Humidity	326,077.70
Rainfall + Rainy Days + Humidity	343,151.43

The AdaBoost.R2 model using various climate variables or a combination of two and three climate variables produces actual MAE values for the training data set in the range of 541,252.64 to 631,160.15 with the highest value in the rainfall variable and the lowest value in the combination of rainfall+rainy days+humidity variables . Meanwhile, model validation produced MAE values in the range of 326,077.70 to 484,460.42 with the highest value for the rainy days variable, and the lowest for the combination of rainy days+humidity variables. Based on the results of model validation, the model shows that the combination of rainy days and humidity variables produces the lowest MAE, this can be related to the effect of these variables on raw water. The effect of rainy days on raw water is the number of days it rains, which affects the volume of raw water availability on the surface, where volume is affected by the

intensity of rain per day, while humidity affects the availability of raw water because humidity can affect the occurrence of rain based on the moisture content contained in air. Therefore, in modeling, the combination of rainy days and humidity variables has the lowest MAE because the model assumes that these two variables have a strong relationship with each other, so that the weight of these two variables on total raw water is quite high. Figures 4 to 9 below show validation graphs between the model and the existing data.

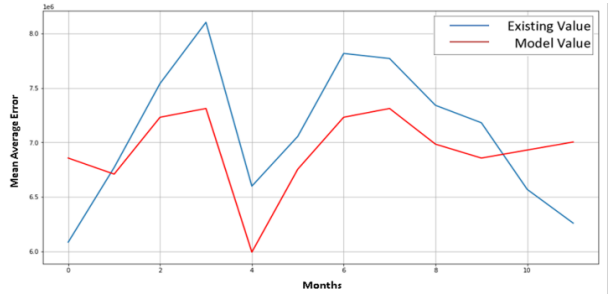


Figure 4. Model validation of rainfall variable

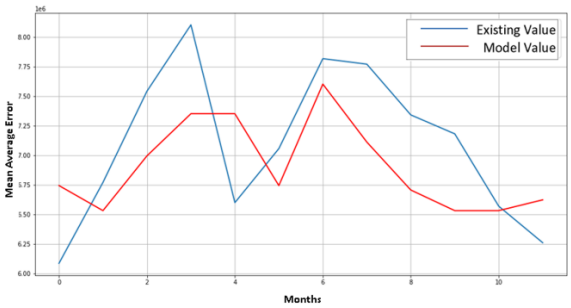


Figure 5. Model validation of rainy days variable

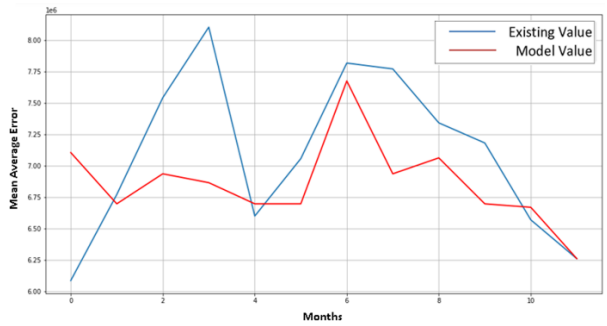


Figure 6. Model validation of humidity variable

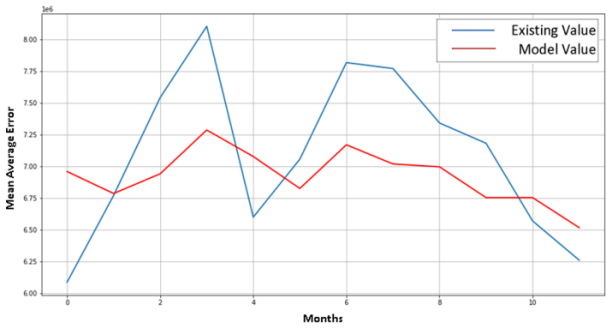


Figure 7. Model validation of rainfall and rainy day variable combinations

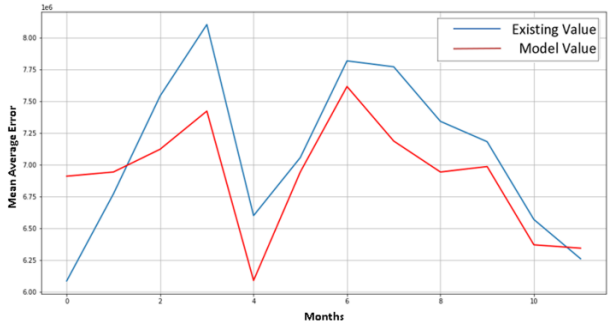


Figure 8. Model validation of rainfall and humidity variable combinations

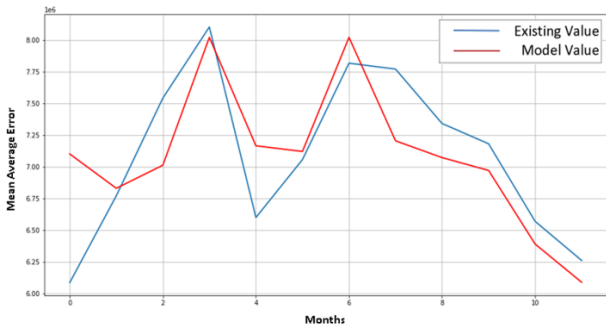


Figure 9. Model validation of rainy days and humidity variable combinations

3.2 Comparison Models

The comparison model used is the ARIMA (Autoregressive Integrated Moving Average) model. In previous studies, the ARIMA model is often used as a comparison model because of its good reliability and consistency for time series data sets. For the research conducted, a variation of the ARIMA model used as a comparison model is the Box-Jenkins ARIMA model. The reason for using this variation as a comparison model is based on the results of the data preparation stage show that the influence of the independent variable on the dependent variable is weak. The model generated from the results of development using TPOT does not consider the strength or weakness of the relationship between the independent variable and the dependent variable at all. Therefore, there is only one variable that is used for the comparison model, namely Raw Water which is the output variable in model development.

The raw water data used for the comparison model received the same treatment as in the creation and development of machine learning models, where the data was first divided into training data sets and validation data sets with the same ratio, namely 72 data for the training data set and 12 data for the validation data set. Like the creation and development of machine learning models, models are validated using the last 12 months of existing data. The raw water training data set needs to be reviewed with the aim of finding out the distribution of the mean data in the data, because in the process, modeling this comparison model is carried out through a process of data differentiation to deviation correction, so that the data distribution will change. The ability of the initial model was tested using 72 data in the training data set, where the data for the n^{th} time period will be used to predict data for the n+1 time period. After testing the capabilities of the initial model, the next step is to test the stationarity of the data. The stationarity test is carried out to obtain stationary data values, meaning that trends in the data must be removed. A good final check for the model is to review the residual error (Brownlee, 2020). The mean residual data after deviation correction is close to 0 (zero), so it can be determined that the comparison model to be used is ARIMA (2,1,2) by considering the deviation correction value. This model was chosen, because although the MAE before bias correction is slightly lower, the difference is not significant.

The validation process is carried out using the same set of codes, model validation produces an MAE value of 330,672.088 with the predicted and existing model values as shown in Table 6.

Table 6. Details of comparison model validation data

Month	Validation Value	Existing Value	MAPE
January 2020	5,402,685.915	6,535,658	17.34 %
February 2020	6,179,503.884	6,376,684	3.09 %
March 2020	6,288,179.188	6,630,117	5.16 %
April 2020	6,776,434.939	6,726,888	0.74 %
May 2020	6,887,332.757	6,966,782	1.14 %
June 2020	6,823,322.977	7,055,663	3.29 %
July 2020	6,773,897.836	6,568,944	3.12 %
August 2020	6,493,293.258	6,801,942	4.54 %
September 2020	6,883,041.879	6,135,998	12.17 %
October 2020	6,317,559.574	6,400,073	1.29 %
November 2020	6,299,022.602	6,452,050	23.7 %
December 2020	6,160,808.442	6,599,260	6.64 %
Average			5,07%

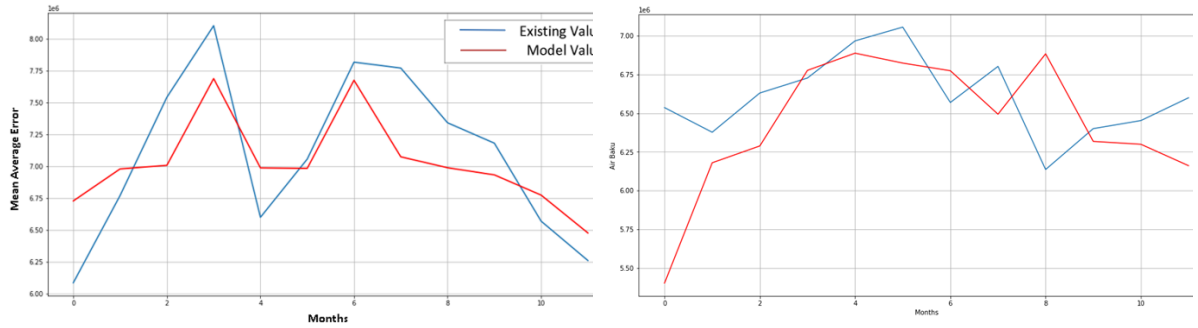
The resulting MAPE value is an average of 5.07%, this is the model's prediction of a miss of 5.07% of the existing value. In addition, based on the MAE value, the ARIMA model misses by 330,672.088 m^3/month from the actual value. Based on the interpretation of MAPE according to Lewis (1982), the ARIMA model produces very accurate modeling values because the average MAPE is at a value of ≤10%.

3.2 Comparison of Model Accuracy

Based on the results of model validation, the model shows that the combination of rainy days and humidity variables produces the lowest MAE, this can be related to the effect of these variables on raw water. The effect of rainy days on raw water is the number of days it rains, which affects the volume of raw water availability on the surface, where volume is affected by the intensity of rain per day, while humidity affects the availability of raw water because humidity can affect the occurrence of rain based on the moisture content contained in air. Therefore, in modeling, the combination of rainy days and humidity variables has the lowest MAE because the model assumes that these two variables have a strong relationship with each other, so that the weight of these two variables on Total Standard Water is quite high.

The other external variables that can be considered for conducting studies on the availability of raw water include landscape, rock composition, and infrastructure, where these variables can help improve model accuracy.(Foster et.al., 2020).

Meanwhile, in comparing the accuracy of the machine learning model with the ARIMA model, the validation results of the machine learning model with the best combination of external variables show a lower MAE value compared to the ARIMA model, even though the difference in MAE values is not significant. This shows that machine learning modeling has quite good potential in making predictions. Comparison of model validation charts using machine learning and ARIMA methods can be seen in Figure 10.



Model Validation Using Machine Learning Methods

ARIMA Model Validation

Figure 10. Comparison of model validation graphs

4. Conclusions

Based on the results and discussion, it can be concluded that the model with the lowest MAE, can be used as a reference for raw water availability, where rainy days and monthly humidity are considered sufficient to influence the amount of available raw water. The resulting model is considered not optimal enough to predict the availability of raw water, even so the model is still representative enough to describe the condition of raw water availability and the factors that influence it. Meanwhile, there is still a lot of room for model development, one of which is the availability of data, because the approach using machine learning is very dependent on the quantity of data.

Referring to the model with the best combination of variables, rainy days and humidity are variables that can be used as a reference for forecasting the availability of raw water, where the model has an MAE value of 326,077.70, and a MAPE of 4.75%. This model is compared with the ARIMA model which has an MAE of 330,672.088 and an MAPE of 5.07%. The machine learning model has better accuracy than the conventional ARIMA model

Acknowledgement

Appreciation for publication support by the Faculty of Engineering of Universitas Pasundan.

References

Afiatun, E., Notodarmojo, S., Effendi, A., & Sidarto, K. 2018. Cost Minimization of Raw Water Source by Integrated Water Supply Systems (A Case Study for Bandung, Indonesia). International Journal of GEOMATE, 14(46), 32-39.

Afiatun, E. , Pradiko, H., & Fabian, E. 2019. Turbidity Reduction for the Development of Pilot Scale Electrocoagulation Devices, International Journal of GEOMATE, 16(56), 123 – 128

Andani, I. G. 2012. Peningkatan Penyediaan Air Bersih Perpipaan Kota Bandung dengan Pendekatan Pemodelan Dinamika Sistem. Jurnal Perencanaan Wilayah dan Kota A SAPPK V1N1, 69 - 78.

Antunes, A., Andrade-Campos, A., & Sardinha-Lourenço, A. a. 2018. Short-term Water Demand Forecasting Using Machine Learning Techniques. Journal of Hydroinformatics, 1343–1366.

Bakker, M., Vreeburg, J., van Schagen, K., & Rietveld, L. 2013. A Fully Adaptive Forecasting Model for Short-term Drinking Water Demand. Environmental Modelling & Software, 141-151.

Bishop, C. M. 2006. Pattern Recognition and Machine Learning. India: Springer.

Brownlee, J. 2020. Data Preparation for Machine Learning. Victoria: Machine Learning Mastery.

Drucker, H. 1997. Improving Regressors Using Boosting Techniques. Retrieved from ResearchGate: https://www.researchgate.net/publication/2424244_Improving_Regressors_Using_Boosting_Techniques

Duerr, I., Merrill, H., Wang, C., Bai, R., Boyer, M., Dukes, M., & Blinzyuk, N. 2018. Forecasting Urban Household Water Demand with Statistical and Machine Learning Methods Using Large Space-time Data: A Comparative study. Environmental Modelling & Software, 29-38.

Foster, T., Mieno, T & Brozović. 2020. Measurement Errors and Their Implications for Agricultural Water Management Policy. Water Resources Research, 56(11): 1-19.

Haque, M. M., de Souze, A., & Rahman, A. 2017. Water Demand Modelling Using Independent Component. Water Resources Management, 299-312.

Lewis, C. D. 1982. International and Business Forecasting Methods. London: Butterworths.

Putri, R.N., Usman, M., Warsono, Widiarti & Virginia, E. 2021. Modeling Autoregressive Integrated Moving Average (ARIMA) and Forecasting of PT Unilever Indonesia Tbk Share Prices During the COVID-19 Pandemic Period. Journal of Physics: Conference Series, 1751: 012027.

Siami-Namini, S., Tavakoli, N., & Namin, A.S. 2017. A Comparison of ARIMA and LSTM in Forecasting Time Series. 2018 17th IEEE International Conference on Machine Learning and Applications, 1394-1401.