Water Consumption Prediction of Semarang Water Utilities using Support Vector Regression Radial Basic Function Kernel Method

Detri Setiyowati[1], Alamsyah[1], Much Aziz Muslim[1]
[1] Department of Computer Science, Universitas Negeri Semarang, Semarang, Indonesia
*Corresponding author: detri.936@gmail.com

ARTICLE INFO

Article history
Received 15 July 2019
Revised 16 September 2019
Accepted 3 October 2019

Keywords
Prediction
Support vector regression
RBF
Water consumption
PDAM

ABSTRACT

People have various needs that can't be released considering their role as living things. The diversity of human needs requires planning for the future, one of which is the provision of water supply because water needs are increasing. Estimation models can be done using the Support Vector Regression (SVR) method. SVR is a development of the Support Vector Machine for regression cases. SVR has four kernels that are commonly used, and in this study, the kernel used is the Radial Basis Function Kernel because RBF is considered capable of maintaining good predictive accuracy. The purpose of this study is to apply the SVR method to predict water consumption with the Radial Basis Function kernel by getting the best SVR parameter and knowing the error value generated from the SVR method. The data used in this study is Semarang Water Utilities' (PDAM) water consumption data from January 2013 to March 2018. The SVR method test results obtained the best parameters are lambda (λ) = 10, sigma (σ) = 0.001, cLR = 0.01, C (Complexity) = 0.01, epsilon (ε) = 0.00000001, with the number of iterations = 1000, produces the lowest Mean Absolute Percentage Error (MAPE) is 1.751%.

1. Introduction

Humans life has undergone many changes with the rapid development of technology, one of which to obtain any pieces of information and solutions (Muslim & Retno, 2014). Problems in society become easily solved using computer technology with a shorter solution time than before (Pramesti, Arifudin, & Sugiharti, 2016). Problems solving can be done with the help of computer science which is an effective algorithm method for calculating a function (Sampurno, Sugiharti & Alamsyah, 2018). The diversity of human needs requires the planning of human needs for the future, one of which is the supplying of water supply because the need for clean water continues to increase (Ajbar & Ali, 2015). Naturally, water is needed by living things in meeting their needs and sustaining life because water is a natural resource that is very important for living things and ecosystems. The increased demand for water makes people try to find water sources that are guaranteed quality (Istiqara, Furqon & Indriati, 2017). Learning studies on water demand have been conducted since 1960 to increase modeling efforts in estimating water demand (Abushammala & Bawazir, 2017).

Various methodologies have been applied in predicting/estimating water requirements such as regression, ARIMA, fuzzy time series, and artificial neural networks. Prediction is important to know an event in the future by recognizing patterns from past events, then humans can prepare everything that will happen (Hikmawati, Arifudin & Alamsyah, 2017). Prediction can be done with various kinds of calculation techniques, one of the techniques that can be used is Support Vector Machine (SVM). SVM is an artificial intelligence-based method with the ability to generalize well in pattern recognition systems. Nonlinear input data in SVM are separated linearly into several higher-dimensional spaces, to provide classification and regression. The results of SVM development in a regression case are called Support Vector Regression (SVR) which can be used for forecasting like a regression method on a statistical approach. The concept of the SVR method is based on risk minimization which is to estimate a function by minimizing the generalization error limit so that it can overcome overfitting better than the usual regression method and artificial neural network (Raharyani, Putri & Setiawan, 2018). The SVR method has the advantage of performing nonlinear relations modeling which balances the complexity of the model and the accuracy of predictions on training data (Yu, Qi & Zhao, 2013).

The SVR method has been widely applied in the estimation of various aspects including the prediction of newspaper/magazine sales (Yu et al., 2013), the number of tourism visitors (Raharyani et al., 2018), and predicting gold prices (Dubey, 2016). Based on the results of Dubey's (2016) research explained that the comparison of models using the SVR and adaptive neuro-fuzzy inference system (ANFIS) methods in predicting gold prices obtained an error value with SVR superior to ANFIS. The experimental results showed that the value of Root Mean Square Error (RMSE) using ANFIS was 15,931 while the RMSE value of SVR was 14,859. The Mean Absolute Percentage (MAPE) ANFIS value was 0.0083 while the MAPE SVR was 0.0063. These results prove that the SVR algorithm is effective in performing regression analysis by capturing nonlinear relations in the feature space. The generalization ability of the SVR model is determined by kernel space features and kernel parameters that affect the value of kernel matrix elements (M. Xie, Wang & L. Xie, 2018). SVR has several kernel functions that are often used, namely the Linear kernel, Radial Basic Function kernel, Polynomial kernel, and sigmoid kernel. Based on these 4 kernel functions, radial basic functions have better performance than other kernels because RBF can maintain relatively high accuracy and can achieve the lowest mean square error of cross-validation so that the results obtained are optimal (Ibrahim & Wibowo, 2014).

The population of the city of Semarang experienced an average increase of 60.04% per year which affected water use. According to Law No. 32 of 2004 concerning the Regional Government, Regional Drinking Water Companies (Perusahaan Daerah Air Minum/PDAM) have an important task in managing and providing clean water services to improve community welfare. Large amounts of water consumption are the cause of a rapid reduction in the water supply (Putriwijaya & Mahmud, 2018). Often the estimates in the water supply are not optimal, where the amount of water produced is greater or smaller than demand. This problem can be overcome by predicting water consumption so that the water produced meets the needs of the community (Jauhari, Himawan & Dewi, 2016). The purpose of this study is to estimate the monthly water consumption by applying the regression vector support method and the basic radial kernel function with the optimal parameter values obtained as well as knowing the average value of the relative errors generated by this method.

2. Methods

Testing measures of methods in this study were demonstrated in the flowchart of the SVR method. Flowcharts for the Support Vector Regression method are shown in Figure 1.

2.1 Preprocessing Data/Normalization

Normalization is part of data transformation with a data scaling process so that data is within a certain range of values. Normalization data aims to standardize the range of feature values in data by making data in the same range of values. MinMax normalization is known as a scaling feature where the numerical range value of the data feature is reduced to a scale between 0 to 1. Normalization data calculation using Formula 1.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{1}$$

Variables description:

x' : the result of data normalization
x : normalized data
x_{min} : the smallest value of all data
x_{max} : the largest value of all data

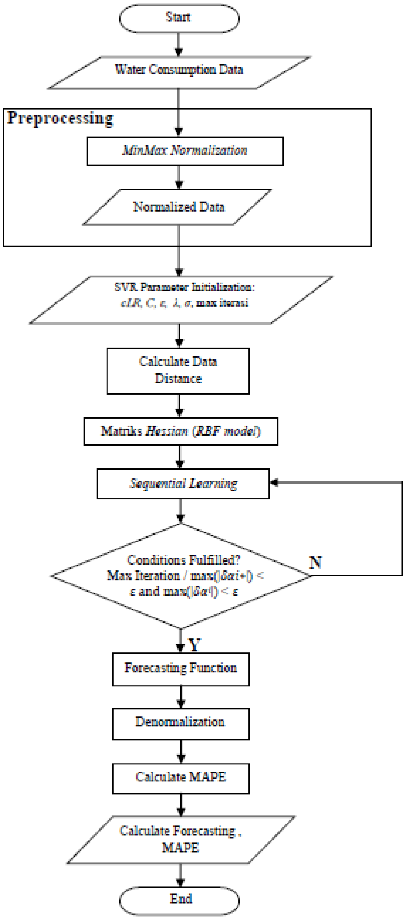

Figure 1. Flowchart of the support vector regression method

2.2 Support Vector Regression (SVR) Method

After the data normalized fourth step sequential learning to perform the computation function in a non-linear SVR is as follows:

1. Initialize complexity (C) parameters, sigma (σ), learning rate constants (cLR), epsilon (ε), lambda (λ), and max iterations.
2. Initialize the values of α_i and α_i^* by giving the value 0. Calculate the hessian matrix using Formula 2 with R_{ij} is the matrix with row *i* and column *j*.

$$R_{ij} = K(x_i, x_j) + \lambda^2 \tag{2}$$

Where $K(x_i, x_j)$ is a kernel function. This study applied the Radial Basis Function kernel as shown in Formula 3.

$$K(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2.\sigma^2}\right) \tag{3}$$

Variables description:

R_{ij}	: Hessian matrix
$K(x_i, x_j)$	: kernel function
x_i	: the *i*th data
x_j	: the *j*th data
λ	: scalar variable

$\|x_i - x_j\|^2$: square of the distance between vectors x_i and x_j
σ : sigma parameter

The scalar variable or parameter λ shows the scalar size in space mapping in the SVR kernel (Vijayakumar & Wu, 1999).

3. For each training data *i, j* = 1, 2, …, n. Perform the following steps:

   a) Calculate the error value with Formula 4.

$$E_i = y_i - \sum_{j=1}^{n}(\alpha_i^* - \alpha_i)\, R_{ij} \quad (4)$$

   b) Calculate $\delta\alpha_i$ and $\delta\alpha_i^*$ using Formula 5 and Formula 6.

$$\delta\alpha_i^* = \min\{\max\,[\gamma(\,E_i - \varepsilon), -\alpha_i^*], C - \alpha_i^*\} \quad (5)$$

$$\delta\alpha_i \;= \min\{\max\,[\gamma(\,-E_i - \varepsilon), -\alpha_i], C - \alpha_i\} \quad (6)$$

   Variables description:
   E_i : error value
   y_i : actual value of training data
   α_i^* : Lagrange multiplier
   α_i : Lagrange multiplier
   R_{ij} : Hessian matrix
   $\delta\alpha_i^*$: single variable, change the value α_i^*
   $\delta\alpha_i$: single variable, change the value α_i
   γ : learning rate
   ε : epsilon parameter
   C : complexity parameter

   The learning rate can be obtained using formula 7.

$$\gamma = \frac{konstanta\ learning\ rate}{\max\,(diagonal\ matriks\ R_{ij})} \quad (7)$$

   c) Change the value of α_i and α_i^* with formula 8 and formula 9.

$$\alpha_i^* = \;\delta\alpha_i^* + \alpha_i^* \quad (8)$$

$$\alpha_i = \;\delta\alpha_i + \alpha_i \quad (9)$$

4. Return to the third process if it has not reached the maximum iteration or max($|\delta\alpha_i|$) < ε and max($|\delta\alpha_i^*|$) < ε.
5. By using Lagrange multiplier and optimality conditions, the regression function is explicitly formulated in formula 10.

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)\, K\big(x_i, x_j\big) + \lambda^2 \quad (10)$$

   Variables description:
   α_i^* : Lagrange multiplier
   α_i : Lagrange multiplier
   x_i : the *i*th data
   x_j : the *j*th data
   λ : scalar variable

6. After the forecast function is formed, a calculation is made until the prediction results are obtained. The output generated from sequential learning processes is the output value in the normalized form. Then the normalization value output must be returned (denormalizing) to the original value to get the output value in the actual range. The formula is shown in formula 11.

$$x = x' * (x_{max} - x_{min}) + x_{min} \tag{11}$$

where:
x : actual data
x' : normalized data
x_{max} : maximum of data
x_{min} : minimum of data

7. Calculates the MAPE value (Mean Absolute Percentage Error) to get the error value of the function based on the formula shown in Formula 12. MAPE is a measure of relative determination used to determine the percentage of deviation from forecasting results.

$$MAPE = \frac{1}{N}\sum_{i=1}^{n}\left|\frac{\hat{y}_i - y_i}{y_i} \times 100\right| \tag{12}$$

where:
n : number of predicted data
y_i : the actual data
$\hat{y}_i$: prediction result

3. Results and Discussion

This research uses MATLAB R2015b tools in the application of the methods. The data used for the prediction process are Semarang city water consumption data in 2013-2018. Data was obtained from PDAM *Tirta Moedal* Semarang city. Before the calculation process is carried out the data that has been obtained is divided into training data and test data.

The next process the data is processed or preprocessing. The preprocessing stage in this study is scaling using MinMax Normalization. Normalized data is carried out as a process sequential learning to find out the alpha and Alpha star values which are then used to predict test data. Parameter testing is needed to evaluate the optimal parameter values used in the prediction process, resulting in good prediction results and low error values.

3.1 Testing SVR Parameter Values

Testing of SVR parameters consists of the number of iterations, lambda, sigma, cLR, Complexity, and epsilon. The test starts from the number of SVR iterations followed by testing lambda, sigma, cLR, C, and the last test for the epsilon parameter. The range of parameter values for each test is adjusted starting from each parameter, for the number of iterations starting from 10-5000. This test is carried out to evaluate the optimal / best parameter values which produce the lowest MAPE value then applied to the prediction process. Figure 2 shows the results of testing the number of iterations that have been done.

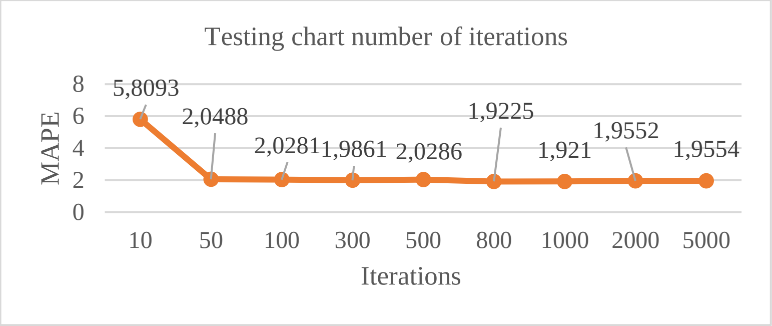

Figure 2. Results of testing the number of SVR iterations

Furthermore, the number of iterations that have been obtained is applied to the next parameter test to obtain the optimal parameters. Figure 3 shows the Lambda parameter test.

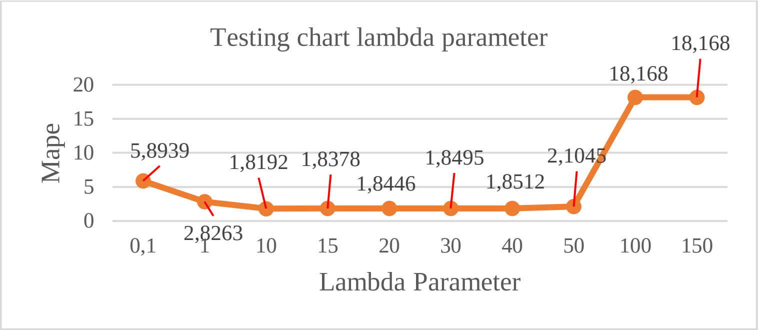

Figure 3. Results of the lambda parameter test

Testing the lambda parameters starts from 1-150 with the optimal parameters found in lambda 10 with a MAPE value of 1.8192%. Next, the sigma parameters are tested as shown in Figure 4.

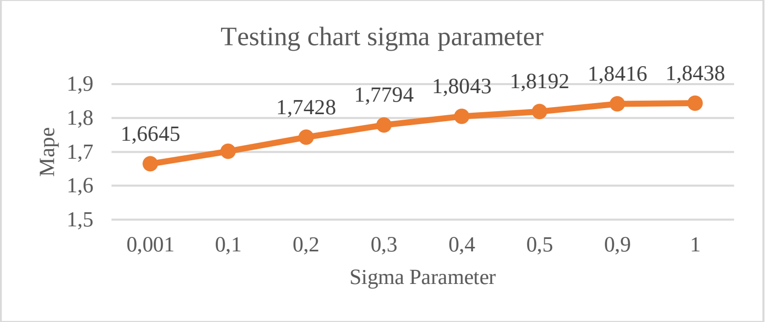

Figure 4. Comparison of MAPE in the sigma parameter

The optimal sigma parameter value is shown at point 0.01 with the MAPE value is 1.6645%. The next test by applying the best parameter results previously obtained is the cLR parameter as shown in Figure 5.

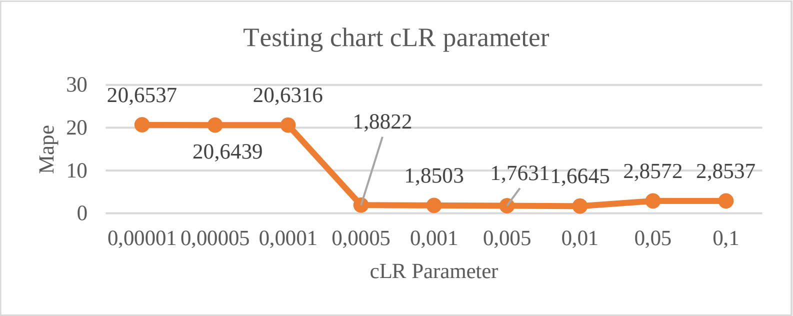

Figure 5. Comparison of MAPE values in the cLR parameter

Based on the testing in Figure 4 shows a decrease in the value of a significant error between 0.00001 to 0.0001. The lowest MAPE results shown in the cLR parameter test are 1.6645% at 0.01. The next parameter testing is parameter C which is the value limit for the regression error shown in Figure 6.

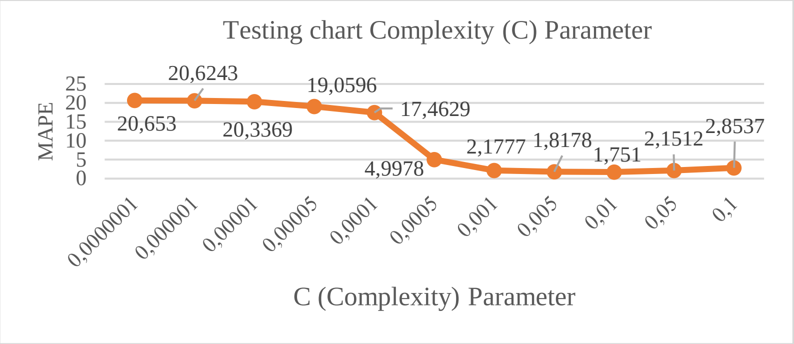

Figure 6. Comparison of MAPE values in parameter C

The result of the parameter C test shows that the MAPE value moves constantly from 0.005 to 0.1. The test results state that the greater the value of C, the smaller the value of the error obtained and the result of the prediction is quite good. This is because the large C parameter values make the prediction model more tolerant of errors. Figure 6 shows the lowest MAPE value and constant motion assessed by parameter 0.01 with the lowest MAPE value which is 1.751%. Testing the last parameter is the epsilon parameter shown in Figure 7.

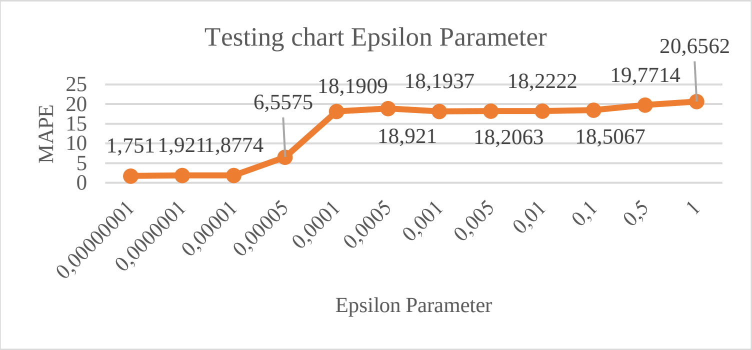

Figure 7. Comparison of MAPE values in epsilon parameters

Based on the test shown in Figure 6, the lowest MAPE value is the parameter value of 0.00000001 with a result of 1.751%. This shows that the higher the epsilon value the higher the error is generated. The value of the large epsilon parameter can cause the search for solutions to come out of the boundary as shown in the value 0.1-1.

3.2 Optimal Application of Parameters

The parameters that have been obtained in the previous test are then applied in testing test data or prediction processes. In this study testing data used as many as 12 data. The smallest predictive error (MAPE) obtained was 1.921%. The prediction results for 12 data testing are shown in Figure 7. Whereas the optimum SVR parameter values that have been obtained based on the previous test are shown in Table 1.

Table 1. The best SVR parameter results

Lambda (λ)	Sigma (σ)	cLR	C	Epsilon (ε)	Iteration
10	0.001	0.01	0.01	0.00000001	1000

The estimated results of water consumption for April 2018 produce a prediction of water consumption of 4.118.321 m^3 with actual data held at 4.124.185 m^3. A comparison of the graph of prediction results to the test data in the test is shown in Figure 7. The test results in Figure 7 shows that the results obtained do not have much difference compared to the actual data, or more clearly the results of the prediction close to the actual data.

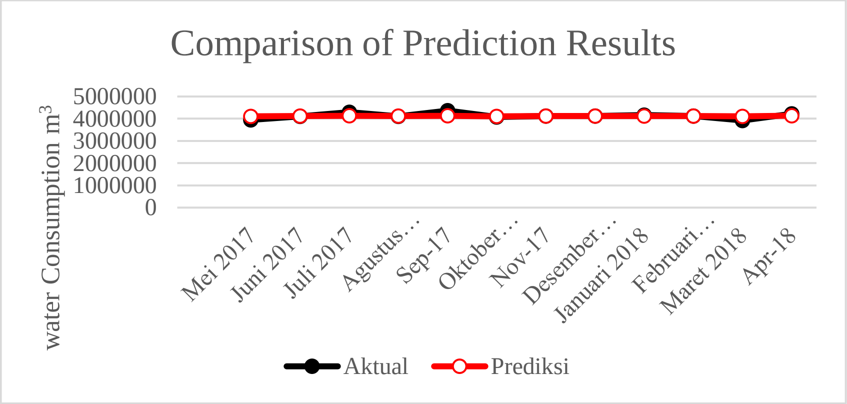

Figure 7. Prediction results using the RBF kernel SVR

4. Conclusion

The application of the support vector regression method for predicting water consumption is carried out in the preprocessing phase, which is data normalization. At this stage, the data will be changed to a range of 0 to 1 to simplify the calculation. After normalizing the data, a calculation is called a sequential learning process wherein the alpha and alpha star values are used in calculating forecasting results. Previously, parameter initialization in SVR was done, which is a lambda, sigma, complexity, cLR, epsilon and the number of iterations. The effect of SVR parameters on forecasting data varies. The small lambda and sigma parameter values show poor forecasting results with high MAPE results shown. The value of parameter C and cLR that are of great value produce good predictions and small MAPE values. Whereas for small value epsilon parameters can produce a small error value. The results of testing the support vector regression method have produced a minimum MAPE value of 1.751% with optimal SVR parameter values obtained, namely lambda = 10, sigma = 0.001, cLR = 0.01, C = 0.01, epsilon = 0.00000001 and iterations = 1000. Results the prediction for April 2018 is 4.118.321 m3 with the actual data held at 4.124.185 m3.

References

Abushammala, M. F. M. & Bawazir, A.K. (2017). Domestic water demand forecasting for Makkah, Saudi Arabia. *European Water*, 58, 481-487.

Ajbar, A., & Ali, E. M. (2015). Prediction of municipal water production in touristic Mecca City in Saudi Arabia using neural networks. *Journal of King Saud University-Engineering Sciences*, 27(1), 83-91. doi:10.1016/j.jksues.2013.01.001

Dubey, A. D. (2016). Gold price prediction using support vector regression and ANFIS models. *International Conference on Computer Communication and Informatics (ICCCI)*, 1(1), 1-6. doi:10.1109/iccci.2016.7479929

Hikmawati, Z. F., Arifudin, R., & Alamsyah. (2017). Prediction the number of dengue hemorhagic fever patients using fuzzy tsukamoto method at public health service of Purbalingga. *Scientific Journal of Informatics*, 4(2), 115-124. doi: 10.15294/sji.v4i2.10342

Ibrahim, N. & Wibowo, Antoni. (2014). Support vector regression with missing data treatment based variables selection for water level prediction of galas river in Kelantan Malaysia. *WSEAS Transactions on Mathematics,* 14(1), 69-78.

Istiqara, K., Furqon, M. T. & Indriati. (2017). Prediksi kebutuhan air PDAM kota Malang menggunakan metode fuzzy time series dengan algoritma genetika [Prediction of water needs of Malang City PDAM uses fuzzy time series methods with genetic algorithms]. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(1), 133-142.

Jauhari, D., Himawan, A. & Dewi, C. (2016). Prediksi distribusi air PDAM menggunakan metode jaringan syaraf tiruan backpropagation di PDAM kota Malang [Prediction of PDAM water distribution using backpropagation neural network method in PDAM Malang city]. *Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK)*, 3(2), 83-87. doi:10.25126/jtiik.201632155

Muslim, M. A. & Retno, N. A. (2014). Implementasi cloud computing menggunakan metode pengembangan sistem agile. *Scientific Journal of Informatics*, 1(1), 29-38. doi:10.15294/sji.v1i1.3639

Pramesti, A. A., Arifudin, R., & Sugiharti, E. (2016). Expert system for determination of type lenses glasses using forward chaining method. *Scientific Journal of Informatics,* 3(2), 177-188. doi:10.15294/sji.v3i2.7914

Putriwijaya, N. N. & Mahmud, W. F. (2018). Peramalan jumlah pemakaian air di PT. Pembangkitan Jawa Bali unit pembangkit Gresik menggunakan support vector regression [Forecasting the amount of water usage in PT. Generation of Java-Bali Gresik generating units using support vector regression]. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(10), 3788-3795.

Raharyani, M. P., Putri, R. R. & Setiawan, B. D. (2018). Implementasi algoritma support vector regression pada prediksi jumlah pengunjung pariwisata [Implementation of support vector regression algorithm in predicting the number of tourism visitors]. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, 2(4), 1501-1509.

Sampurno, G. I., Sugiharti, E. & Alamsyah. (2018). Comparison of dynamic programming algorithm and greedy algorithm on integer knapsack problem in freight transporatation. *Scientific Journal of Informatics,* 5(1), 40-49. doi: 10.15294/sji.v5i1.13360

Vijayakumar, S. & Wu, Si. (1999). Sequential support vector classifiers and regression. *Proceeding International Conference on Soft Computing,* 1(1), 610-619. Italy: Genoa.

Xie, M., Wang, D. & Xie, L. (2018). One SVR modeling method based on kernel space feature. *IEEJ Transactions on Electrical and Electronic Engineering*, 13(1), 168-174. doi:10.1002/tee.22510

Yu, X., Qi, Z. & Zhao, Y. (2013). Support vector regression for newspaper/magazine sales forecasting. *Procedia Computer Science*, 17(1), 1055-1062. doi:10.1016/j.procs.2013.05.134