Support Vector Regression (SVR) Model for Seasonal Time Series Data

Hanifah Muthiah, Umu Sa'adah, Achmad Efendi
Department of Statistics, Brawijaya University Malang, Indonesia
hanifahmuthiah93@gmail.com, u.saadah@ub.ac.id, a_efendi@ub.ac.id

Abstract

Support Vector Regression (SVR) is one of the methods used in the supervised learning process for regression cases that comes from the Support Vector Machine (SVM). SVR is a method to solve forecasting cases that can overcome overfitting so that it will produce a good performance and has an advantage in optimization with good generalizability and accuracy results. There are several choices of kernel functions that can be seen from its ability to work on the SVR method, one of the most popular and often considered the best is the Radial Basis Function (RBF) because based on some previous research this kernel shows the lowest error value compared to other kernels so that the purpose of this study is to compare the RBF kernel with other kernels to see the performance of the model and the accuracy of forecasting produced by using the kernel on time series data which has a seasonal pattern.

Keywords
Kernels, Seasonal Time Series, Support Vector Regression.

1. Introduction

Data Mining is a work process using one or more computer learning techniques (machine learning) to convert large amounts of data into information or knowledge (Vijayakumar & Nedunchezhian, 2012). The main purpose of data mining is to make it easier to extract data and make decisions in an analysis (Han et al., 2012). There are several data mining methods, including predictive ones, namely classification and, regression which aim to predict the value of a variable based on the value of other variables. The regression technique is a data mining technique that is used to predict the value of a variable based on the value of another variable, where the predictor variable is a known attribute, while the response variable is the value to predict (Scholkopf & Smola, 2018).

One of the methods used in the supervised learning process for regression is Support Vector Regression (SVR). Support Vector Regression is a Support Vector Machine (SVM) method use in regression cases with a large number of data inputs. SVM is a very popular artificial intelligence-based method. SVM tries to find the best hyperplane (dividing line) between classes. Various researches have been carried out by applying the SVR method to solve forecasting cases with good accuracy compared to other methods. SVR is a method that can overcome overfitting so that it will produce a good performance and have advantages in optimizing the pattern recognition system with good generalization and accuracy results (Bharati & Ramageri, 2010).

Time series modeling has become popular in recent years. Using historical information to characterize and forecast the series (Caraka et al., 2017). Prediction or forecasting is an important method in carrying out effective and efficient planning to obtain future predictions based on the past, controlling the process of producing data series, and understanding the mechanism of producing the series. Forecasting and planning need to be done because of the grace period between an event and an upcoming event.

Data is an important thing that supports the test of a predicting or forecasting model. The type of data selected is determined based on the method to be used. In the prediction or forecasting method, time-series data are generally used. In time series, there are several types of data patterns, one of which is seasonal patterns. This pattern is a fluctuation of data that occurs periodically within a certain time, such as in one year, quarter, quarterly, monthly, weekly, or daily. In general, seasonal patterns happen in recurring conditions that tend to be regular, for example, an increase in the number of tourists near the end of the year or new year (Percival & Walden, 2000).

Support Vector Regression (SVR) as a development method for SVM has a reliable performance in predicting or forecasting time series data (Hanke & Wichern, 2009). SVR is successfully used in various research fields, such as prediction on stock price movement data (Henrique et al., 2018) (Patriya, 2020), prediction of sales data (Wu, 2017) (Nava et al., 2018), and many more. Several studies have shown that the performance of the SVR is better than the artificial neural network (ANN) model in prediction (Barbour et al., 2018) (Mustakim et al., 2016) and the reliability of SVR performance is largely determined by the kernel function used and the characteristics of the data used in constructing the SVR (Caraka et al., 2017) (Alida & Mustikasari, 2020). The best kernel according to several studies that have been done is the radial basis function (RBF) because it shows the smallest error value compared to other kernels (Caraka et al., 2017) (Alida & Mustikasari, 2020).

2. Literature Review

2.1 Previous Research

Research (Barbour et al., 2018), which predicts travel time on an urban arterial road in Chennai, India along 2.8 km by determining 8 optimal input values, a C value of 25, and an error of 0.1. The results obtained show that the SVR capability is better than the artificial neural network (ANN) model and the moving average with a mean absolute percentage error (MAPE) of 10 and a root mean squared error (RMSE) of 95%. Research (Alida & Mustikasari, 2020), regarding the prediction of the rupiah exchange rate towards the dollar, the prediction was done by one of the machine learning methods, namely the Support Vector Regression (SVR) algorithm. The prediction model uses 3 kernels. Each kernel is calculated the best model accuracy value and error value for later comparison. In terms of accuracy and error, the RBF kernel has advantages over linear and polynomial kernels with the coefficient of determination (R-squared) of 95.94% and the root mean squared error (RMSE) of 1.25%. Research (Bagheri & Rezaei, 2019), predicts permeability (the ability of fluids to pass through porous media without changing the rock structure that plays a major role in the production rate of a hydrocarbon reservoir) in reservoir rock.

The method used is Support Vector Regression (SVR) with a kernel radial basis function (RBF) to estimate permeability in Iran's South Pars gas field. The scores for the four electrofacies appointed were 88.2, 78.51, 84.73, and 77.54 percent, respectively. The high value of the coefficient of determination (R-Squared) obtained for each electrofacies shows reliable accuracy in the model for predicting reservoir permeability. Research (Chawsheen & Broom, 2017), analyzes the performance of various models constructed using linear kernel SVR and trained on historical bid data for high-frequency currency trading. This model generates a simulation of currency trading in the following year. This model is used to execute record profit, hit ratio, and the number of trades, with the result that it is possible to get good profit as well as hit ratio from the trained linear model. Based on descriptions of some of the previous research literature that became the reference for this study, the application of the Support Vector Regression (SVR) algorithm with the radial basis function (RBF) and Linear kernel was carried out in prediction and forecasting on the seasonal time series Support Vector Regression (SVR) data. Data with seasonal patterns have special characteristics that can be applied to several predictions or forecasting methods to find out how accurate the model is.

2.2. Theoretical Basis

2.2.1 Time Series

Time series is a set of observations of ordered data in time. The time series method is a method of forecasting using an analysis of the pattern of the relationship between the variables to be estimated and the time variable. Forecasting a time series data needs to pay attention to the type or pattern of data. In general, there are four kinds of time series data patterns, namely horizontal, trend, seasonal and cyclical (Hanke & Wichern, 2009). The four patterns in time series are the trends, cyclic, seasonal, and random variation (irregular). Seasonal data patterns occur when time-series data is influenced by seasonal factors. This pattern has a seasonal pattern that repeats itself from period to period. For example, there are repeated fluctuations at certain times, such as what is often found in the quarterly, semester, monthly, or weekly data.

2.2.2 Support Vector Regression

Support Vector Regression (SVR) is an SVM development for regression cases. In the case of regression, the output is either a real number or a continuous number. SVR is a method that can overcome overfitting so that it will produce good performance (Scholkopf & Smola, 2018). If the SVM aims to divide the dataset (classification) into two zones (clusters), then SVR is the opposite, which is to make the entire dataset into one zone, while maximizing the epsilon distance (ε) (small value). You can see the SVR illustration in Figure 1.

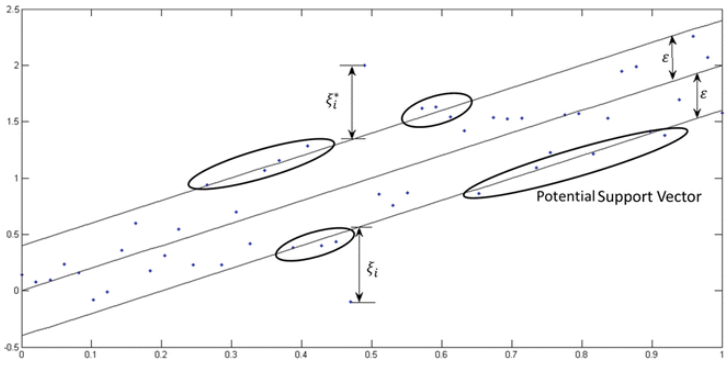

Figure 1. SVR Illustration (Babu & Mohan, 2017)

Figure 1 shows a hyperplane (diagonal line in the middle) that is flanked by two lines + and a boundary -. It also can be seen that there is ε as the distance between the hyperplane and the two boundary lines. There are several data points circled which become potential support vectors, meaning that these data points are data points that can become potential boundaries. The purpose of the SVR is to find a function as a hyperplane (dividing line) in the form of a regression function that matches all data input with the smallest possible error (Scholkopf & Smola, 2018). The purpose of this SVR is to map the input vector into a higher dimension. Suppose a function (4) is a regression line as the optimal hyperplane (Abe, 2005).

$$f(x) = \boldsymbol{w}^T \varphi(x) + b \quad (1)$$

w is the weight vector l dimension, $\varphi(x)$ is a function that maps x to the feature space with l dimensions, b is biased. The function of $\varphi(x)$ shows a point in the higher dimensional feature space, the result of the mapping of the input vector in the lower dimensional input space. The coefficients w and b are estimated by minimizing the risk function. Therefore, to maximize the margin δ, a minimum $\|w\|$ is needed. Optimization of problem-solving as shown in function (2).

$$min\frac{1}{2}\|w\|^2 \quad (2)$$

with the condition, $y_i - w^T\varphi(x_i) - b \leq \varepsilon, for\ i = 1, \dots, l$ and $w^T\varphi(x_i) - y_i + b \leq \varepsilon, for\ i = 1, \dots, l$.
Where y_i is the actual value of i period, and $\varphi(x_i)$ is the estimated value of i period.

The $\|w\|^2$ factor is called regulation. Minimizing $\|w\|^2$ will make a function as thin (flat) as possible so that it can control the functional capacity. All points outside the margin/limit ε will be penalized. Furthermore, the optimization problem above can be formulated into a function (3) (Scholkopf & Smola, 2018).

$$min\frac{1}{2}\|w\|^2 + C\sum_{i=1}^{\ell}(\xi_i + \xi_i^*) \quad (3)$$

with the condition, $y_i - w^T\varphi(x_i) - b - \xi_i \leq \varepsilon, \quad i = 1, \dots, l$
$w^T\varphi(x_i) - y_i + b - \xi_i^* \leq \varepsilon, \quad i = 1, \dots, l$
$\xi_i, \xi_i^* \geq 0$

$C > 0$ constant determines how much the error deviation is from the tolerable limit ε. The formula above is a Convex Linear Programming NLP Optimization Problem which functions to minimize the quadratic function to be converted into a constraint. This limitation can be solved by using the Lagrange Multiplier function. The process of deriving formulas is very long and complicated. After going through mathematical stages, a new equation is obtained with the function:

$$f(x) = \sum_{i=1}^{l}(a_i - a_i^*).(x_i.x) + b \quad (4)$$

Where x_i is the support vector and x is the test vector. The above functions can be used to solve linear problems. Whereas for non-linear problems the values of x_i, and x are first transformed into a high-dimensional feature space by mapping the vectors x_i and x into the kernel function so that the final function becomes:

$$f(x) = \sum_{i=1}^{l}(a_i - a_i^*).K(x_i.x) + b \quad (5)$$

The function $K(x_i.x)$ is the Kernel. The table 1 below shows the kernels used in the SVR calculation (Scholkopf & Smola, 2018).

Table 1. Kernel Function

Kernel	Function
Linear	$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i, \vec{x}_j)$
Radial Basis Function	$K(\vec{x}_i, \vec{x}_j) = exp(-\frac{\|\vec{x}_i - \vec{x}_j\|^2}{2\sigma^2})$

2.2.3 Goodness of Fit Models

The measure of model quality is used to evaluate the model that has been formed as well as to see the effectiveness of a forecast. The measure used in this study is the root mean squared error (RMSE).

$$\boldsymbol{RMSE} = \sqrt{\boldsymbol{MSE}} = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \dot{Y}_i)^2}{n}} \quad (6)$$

Where Y_i is the actual data, $\dot{Y}_i$ is the forecast result data, and n is the number of observed data (Voulgaraki, 2013).

3. Methods

This paper uses two processes, namely pre-processing and post-processing. pre-processing is meant here is to form input $\boldsymbol{x}$ using lags from SARIMA to then carry out Support Vector Regression (SVR) analysis as a post-process, or it can be called the main analysis. Based on Figure 2, this is the explanation detail of the research:
Check data training pattern.
Stationary test with ADF, if data is not stationary, do a difference for data training, and then check again with ADF test.
Check ACF and PACF patterns.
SARIMA parameter estimation for training data.
Testing the suitability of the SARIMA model on the residuals, namely the white noise and normality.
If testing the suitability residual data is not significant, forming the SARIMA model with input data using stepwise and no-stepwise mode for Linear and Radial Basis Function (RBF) kernels.
Calculate Root Mean Squared Error (RMSE) for each SVR model.
Validation of each model with testing data.
Calculate Root Mean Squared Error (RMSE) for testing data.
Compare the RMSE of each SVR model to choose the best model.

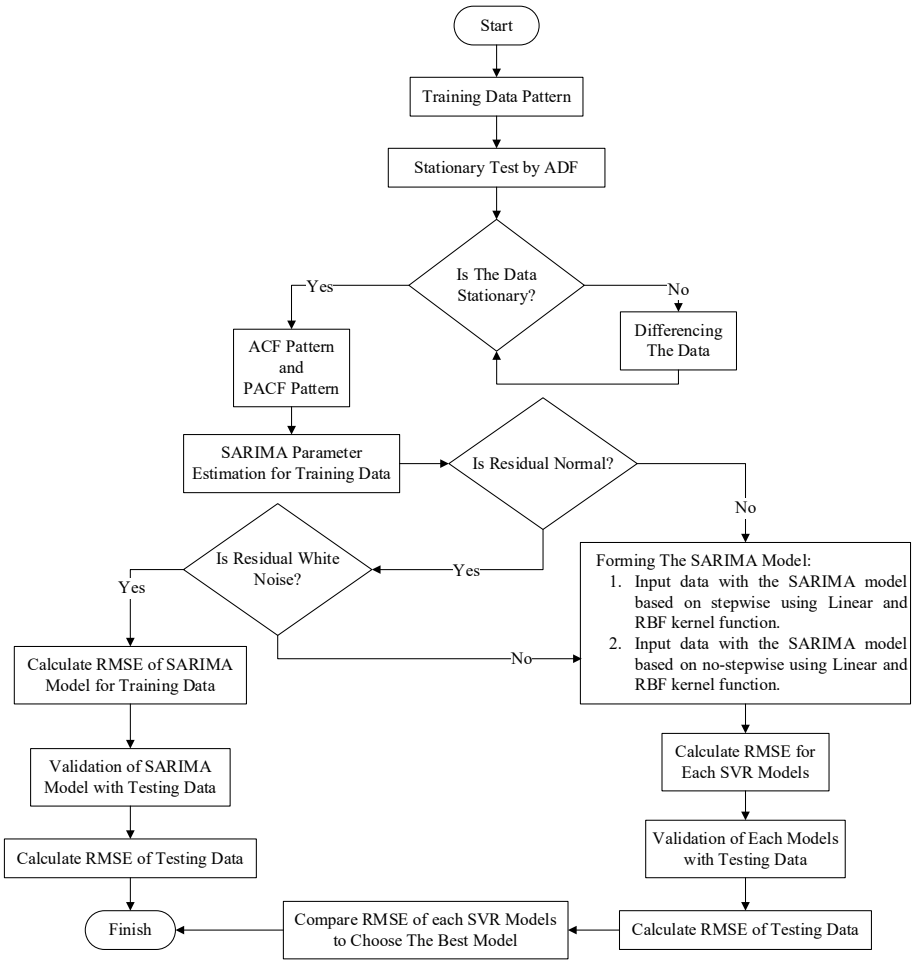

Figure 2. Research Flow Chart

4. Data Collection

In this case, secondary data were used. Secondary data uses a sample of data on the realization of daily electricity loads in the East Java region in units of megawatts (MW) in the period October 07, 2019, to April 05, 2020, obtained from the Java-Bali State Perusahaan Listrik Negara (PLN). The data length $n = 182$ data divided into training data as much as 161 data from the number of samples used which aims to form a model and testing data as many as 21 data from the number of samples used for model testing to predict and forecast data. The data type is weekly seasonal it can be seen in Figure 3.

5. Result and Discussion

5.1 Numerical Results

5.1.1 Support Vector Regression (SVR) with Stepwise Input

In Table 2, there is RMSE as a measure of the goodness of the SVR model with stepwise input for two kernels were chosen, namely Linear function and Radial Basis Function (RBF).

Table 2. Summary SVR Models with Stepwise Input

	Kernel	RMSE
Training	Linear Function	204.0512
	Radial Basis Function	209.3073
Testing	Linear Function	153.8064
	Radial Basis Function	159.9443

Based on Table 2, the linear function kernel has the smallest RMSE value for training data is 204.0512, and testing data is 153.8064, while RBF kernel value for training data is 209.3073 and testing data is 159.9443. The value obtained does not differ much between these two kernels.

5.1.2 Support Vector Regression (SVR) with No-Stepwise Input

In Table 3, there is RMSE as a measure of the goodness of the SVR model with no-stepwise input for two kernels were chosen, namely Linear function and Radial Basis Function (RBF).

Table 3. Summary SVR Models with No-Stepwise Input

	Kernel	RMSE
Training	Linear Function	284.2707
	Radial Basis Function	286.9821
Testing	Linear Function	234.0197
	Radial Basis Function	232.1223

Based on Table 3, the linear function kernel has the smallest RMSE value for training data is 284.2707, and RBF for testing data is 232.1223, while RBF kernel value for training data is 286.9821 and linear function for testing data is 234.0197. Same with SVR with stepwise input, the value obtained does not differ much between these two kernels with no-stepwise input.

5.2 Graphical Results

5.2.1 Secondary Data

At this stage, the data pattern is checked using a line chart. The expected data criterion is a seasonal pattern. An overview of the data patterns used in this study can be seen in Figure 3. Electricity load realization data for October 07, 2019, to April 05, 2020, tends to be stationary with a seasonal trend at the end of the week, namely from Saturday to Sunday, and then it will rise again on Monday. Generally, this happens because Monday to Friday are working days where more activity is carried out so that the load on electricity consumption increases more than on weekends.

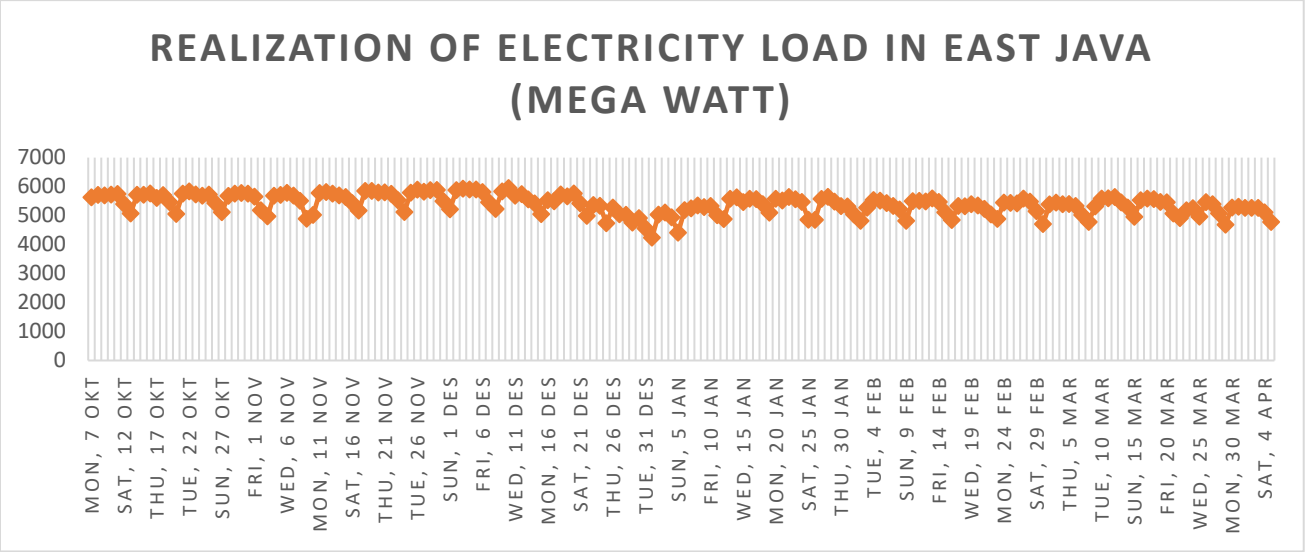

Figure 3. Realization of Electricity Load in East Java

5.2.2 Seasonal Autoregressive Integrated Moving Average (SARIMA)

Significant SARIMA lags are used for SVR input, and seasonal lag enters the input. In this case, using stepwise and no-stepwise mode. In Figure 4 is the ACF pattern for SARIMA, lag shows that are a seasonal pattern, which is similar to Figure 3.

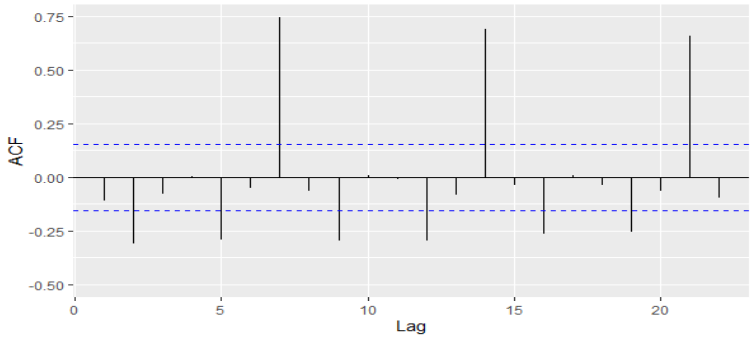

Figure 4. ACF of SARIMA

Based on ACF pattern, SARIMA using stepwise is SARIMA (1,0,0)(2,0,0)[7], and for no-stepwise is SARIMA (3,0,0)(2,0,0)[7].

5.2.3 Result of SVR Linear Kernel with Stepwise Input

The results of SVR forecasting on testing data using the linear kernel function with stepwise input can be seen in Figure 5. It can be seen that the forecasting results and testing / actual data have a pattern that tends to be the same. Testing data is used to test the resulting model from the training data. Training and testing can see the similarity or suitability of their characteristics based on the closeness of the results of the goodness of the models. testing data used starts from March 16, 2020, to April 05, 2020. Forecasting is carried out for 3 weeks or 3 seasonal periods, from April 06 to April 26, 2020.

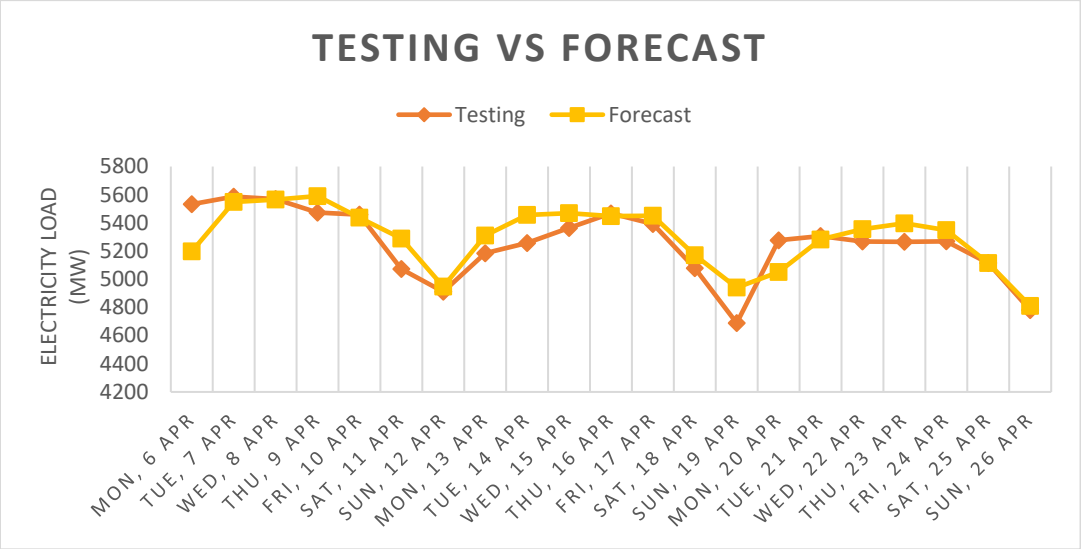

Figure 5. Testing vs Forecast

5.4 Validation

5.4.1 SVR Model with Stepwise Input

Linear function and RBF kernels generate parameter values of $C = 1, \gamma = 0.3333333, and\ \varepsilon = 0.1$ with $\sigma = 335.0145$, produce $\beta_{LIN} = -0.004917671$, and $\beta_{RBF} = 0.1108223$.

1. Linear Kernel Function

$$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i, \vec{x}_j)$$
$$f(x) = -0.004917671(\vec{x}_i, \vec{x}_j)$$

2. Radial Basis Function

$$K(\vec{x}_i, \vec{x}_j) = \exp(-\frac{\| \vec{x}_i - \vec{x}_j \|^2}{2\sigma^2})$$
$$K(\vec{x}_i, \vec{x}_j) = \exp(-\frac{\| \vec{x}_i - \vec{x}_j \|^2}{2(335.0145^2)})$$

$$f(x) = 0.1108223(\exp(-\frac{\|\vec{x}_i - \vec{x}_j\|^2}{2(335.0145^2)}))$$

5.4.2 SVR Model with No-Stepwise Input

Linear function and RBF kernels generate parameter values of $C = 1, \gamma = 0.2, and\ \varepsilon = 0.1$ with $\sigma = 343.0039$, produce $\beta_{LIN} = -0.01964134$, and $\beta_{RBF} = 0.03323177$.

1. Linear Kernel Function

$$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i, \vec{x}_j)$$
$$f(x) = -0.01964134(\vec{x}_i, \vec{x}_j)$$

2. Radial Basis Function

$$K(\vec{x}_i, \vec{x}_j) = exp(-\frac{\| \vec{x}_i - \vec{x}_j \|^2}{2\sigma^2})$$
$$K(\vec{x}_i, \vec{x}_j) = exp(-\frac{\| \vec{x}_i - \vec{x}_j \|^2}{2(343.0039^2)})$$

$$f(x) = 0.03323177(exp(-\frac{\| \vec{x}_i - \vec{x}_j \|^2}{2(343.0039^2)}))$$

5.4.3 Goodness of Fit SVR Model

Based on the previous Table 2 and Table 3, it is found that the linear kernel function with stepwise input has the best model accuracy for RMSE, which means 153.8064 for testing and 204.0512 for training invariance of the Realization of Electricity Load in East Java.

6. Conclusion

Based on the analysis that has been done, the conclusions are obtained, namely:

1. Based on the RMSE value obtained from the linear kernel function and RBF, it was found that the linear kernel function using stepwise for data testing had the smallest RMSE value, namely 153.8064.
2. Based on the RMSE value obtained, linear kernel using stepwise has the best performance and accuracy in prediction, so that it can be said that the SVR model is simple for Realization of Electricity Load in East Java data.

References

Abe, S. (2005). *Support vector machines for pattern classification* (Vol. 2). Springer.

Alida, M., & Mustikasari, M. (2020). Rupiah Exchange Prediction of US Dollar Using Linear, Polynomial, and Radial Basis Function Kernel in Support Vector Regression. *Jurnal Online Informatika*, *5*(1), 53–60.

Babu, N. R., & Mohan, B. J. (2017). Fault classification in power systems using EMD and SVM. *Ain Shams Engineering Journal*, *8*(2), 103–111.

Bagheri, M., & Rezaei, H. (2019). Reservoir rock permeability prediction using SVR based on radial basis function kernel. *Carbonates and Evaporites*, *34*(3), 699–707.

Barbour, W., Mori, J. C. M., Kuppa, S., & Work, D. B. (2018). Prediction of arrival times of freight traffic on US railroads using support vector regression. *Transportation Research Part C: Emerging Technologies*, *93*, 211–227.

Bharati, M., & Ramageri, M. (2010). *Data mining techniques and applications*.

Caraka, R. E., Yasin, H., & Basyiruddin, A. W. (2017). Peramalan Crude Palm Oil (CPO) Menggunakan Support Vector Regression Kernel Radial Basis. *Jurnal Matematika*, *7*(1), 43–57.

Chawsheen, T. A., & Broom, M. (2017). Seasonal time-series modeling and forecasting of monthly mean temperature for decision making in the Kurdistan Region of Iraq. *Journal of Statistical Theory and Practice*, *11*(4), 604–633.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: concepts and techniques, Waltham, MA. *Morgan Kaufman Publishers*, *10*, 971–978.

Hanke, J. E., & Wichern, D. W. (2009). Business forecasting 9th ed. *New Jersey*.

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2018). Stock price prediction using support vector regression on daily and up to the minute prices. *The Journal of Finance and Data Science*, *4*(3), 183–201.

Mustakim, M., Buono, A., & Hermadi, I. (2016). Performance comparison between support vector regression and artificial neural network for prediction of oil palm production. *Jurnal Ilmu Komputer Dan Informasi*, *9*(1), 1–8.

Nava, N., Di Matteo, T., & Aste, T. (2018). Financial time series forecasting using empirical mode decomposition and support vector regression. *Risks*, *6*(1), 7.

Patriya, E. (2020). Implementasi Support Vector Machine Pada Prediksi Harga Saham Gabungan (IHSG). *Jurnal Ilmiah Teknologi Dan Rekayasa*, *25*(1), 24–38.

Percival, D. B., & Walden, A. T. (2000). *Wavelet methods for time series analysis* (Vol. 4). Cambridge university press.

Scholkopf, B., & Smola, A. J. (2018). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. Adaptive Computation and Machine Learning series.

Vijayakumar, V., & Nedunchezhian, R. (2012). A study on video data mining. *International Journal of Multimedia Information Retrieval*, *1*(3), 153–172.

Voulgaraki, M. (2013). Forecasting sales and intervention analysis of durable products in the Greek market. *Empirical Evidence from the New Car Retail Sector. London School of Economics and Political Science*.

Wu, J. Y. (2017). *Housing Price prediction Using Support Vector Regression*.

Biographies

Hanifah Muthiah is a Student of the Magister Program in Statistics Department at Brawijaya University Malang. She received a bachelor's degree in statistics from the Islamic University of Indonesia Yogyakarta. She is interested in the scientific field of statistics, especially in Regression Analysis.

Umu Sa'adah is Lecturer of Mathematics Department at Brawijaya University Malang. She has expertise in Statistical Computation.

Achmad Efendi is A Lecturer and Head of the Undergraduate Program of Statistics Department at Brawijaya University Malang. He has expertise in Statistical Modeling.