ISSN: 1978-0524

Developing support vector regression model to forcast stock prices of mining companies in Indonesia

Dhanukhresna Hangga Yudhawan [a,1,*], Tuti Purwaningsih [a,2]

[a] Universitas Islam Indonesia , Jl. Kaliurang KM 14,5, Yogyakarta, 55584, Indonesia
[1]16611045@students.uii.ac.id [*]; [2] tuti.purwaningsih@uii.ac.id
** corresponding author*

ABSTRACT

In the modern era, as it is now, the world of stock investment is in great demand by investors, both long-term and short-term stock investments. Stock investment provides many benefits for investors. Investors need to analyze stock investments to predict the shares' price to be purchased to get large profits. Very volatile stock price movements make it difficult for investors to predict stock prices. Investors' main hope is to benefit from each price that changes from time to time or can be referred to as time-series data. Data mining extracts considerable information from data by collecting, using data, historical patterns of data relationships, and relationships in large data sets. Support vector regression has advantages in making accurate stock price predictions and can overcome overfitting by itself. PTBA and ITMG are the leading coal mining companies in Indonesia, so many people want to invest in the company. ADRO, PTBA, and ITMG stock price prediction analysis using support vector regression algorithm has good predictive accuracy values, including. PTBA stock price has an R-square value of 97.9% in the RBF kernel and linear with MAPE, respectively, 2,465 and 2,480. Moreover, for ITMG stock price, it has an R-square accuracy of 94.3% in the RBF kernel and linear with MAPE, respectively 5.874 and 5.875. These results indicate that the SVR method is best used for forecasting stock prices.

Keywords:
Forecasting
Stock price
Support Vector Regression
Time Series

I. Introduction

The capital market (capital market) is an organized financial system in which there are commercial banks and financial institutions as intermediaries for securities such as stocks, bonds, outstanding debt securities. In essence, the capital market means connecting parties with excess funds with those who need funds. As one of the national economies' potentials, capital market activities increasingly place its role in developing the national economy [1].

Stocks are the most popular investment market instruments. Issuing shares is one strategy of a company or business entity to raise funding for the company [2]. In a dynamic and fluctuating movement, the stock investment can also cause a loss for investors. Stock price prediction is an analysis technique to determine future stock prices using historic stock prices in the past. Predictions can be made using several methods, but using the time series model is expected to produce excellent and optimal predictions. The characteristics of stock data are time-series data that move continuously with time [3].

The mining sector in Indonesia supports the national economy and national energy security, both in employment and foreign exchange earnings through exports. The mining sector is further divided into five sub-sectors, one of which is the coal mining sub-sector. The coal mining sub-sector currently accounts for 75 to 80 percent of the total Non-Tax State Revenue (PNBP) in the mineral sector.

SVM algorithm is an algorithm of one of the classification methods that can produce a learning process or learning, separated by a hyperplane line. One of the SVM modifications used for the regression approach is Support Vector Regression. The concept of SVR is to maximize the hyperplane to get vector support data [4]. SVM has been widely used for forecasting stock prices and shows better performance than other algorithms, including ANN. An ANN has already been widely used for forecasting processes, including a promising alternative to predicting stock prices, where ANN finds a solution in the form of local optimal.

DOI:10.26555/jifo.v14i2.a17283

In contrast, SVM finds global optimal [5]. One of the advantages of using SVM can offer the global optimum solution. It can be analyzed theoretically using a concept from computational learning theory and achieve good performance at the same time [6].

II. The Proposed Method

A. Support Vector Regression

Support Vector Regression is a theory adapted from machine learning theory that has been used to solve classification problems, namely Support Vector Machine. SVR itself is an application of the SVM algorithm in the case of regression [7]. The SVR algorithm concept can produce a good forecasting value because SVR can solve overfitting problems [8].

The SVM method concept can be explained simply as a way to find the best hyperplane (Fig. 1). Hyperplane itself functions as a separator in two data classes in the input space.

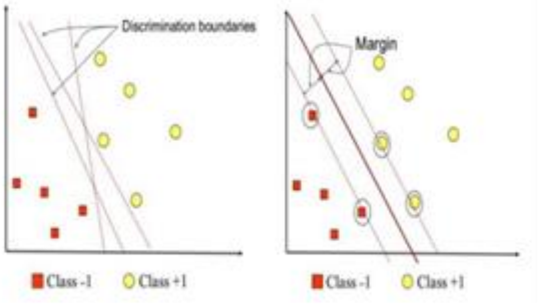

Fig. 1. Hyperplane of Support Vector Machine

Fig. 2 illustrates support vector regression; the picture above shows a hyperplane or center line flanked by two boundary lines (+) and a boundary line (-). It can be seen that there are several circled data points. Those are potential support vectors, or data points can become potential candidates so that all data points can be entered into one zone while minimizing epsilon value (ε).

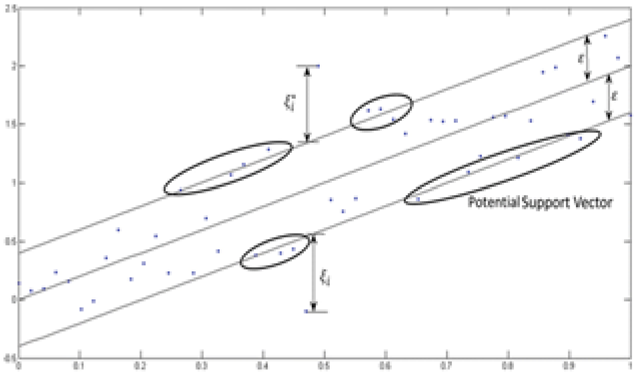

Fig. 2. Illustration of Support Vector Regression

In SVR, which has a low dimension, it will be transformed into a linear regression with high dimensional features. The general form of support vector regression is in (1).

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

Where φ(*x*) is a function that maps *x* in a higher dimension, and *b* is a bias in a constant form. $w^T$is a weighting vector. The coefficients *w* and *b* are estimated by minimizing the risk function.

The loss function is a function that shows the relationship between error, and how this error is charged, the difference in loss function will produce a different SVR formula [9].

Simple loss functions and ε-insensitive loss function as an approach to Huber's loss function that allows support vectors to be obtained [10]. The ε-insensitive loss function formula is in (2).

$$R\big(f(x)\big) = \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} L_\varepsilon(y_i, f(x_i)) \tag{2}$$

The L condition can be defined in (3).

$$L_\varepsilon(y) = \begin{cases} 0, & for\ |f(x) - y| < \varepsilon \\ |f(x) - y| - \varepsilon, & otherwise \end{cases}$$

L_ε is called the ε-insensitive loss function, *C* and ε are prescribed parameters. Concept quadratic programming at equation (2) can be transformed by minimizing.

$$R(w, \xi, \xi^*) = \frac{1}{2}\|w\|^2 + C\left(\sum_{i=1}^{n}(\xi_i + \xi_i^*)\right) \tag{3}$$

White the provision of:

$$y_i - w^T\varphi(x_i) - b - \xi_i \le \varepsilon, \quad i = 1, \dots, l$$
$$w^T\varphi(x_i) - y_i + b - \xi_i^* \le \varepsilon, \quad i = 1, \dots, l$$
$$\xi_i + \xi_i^* \ge 0$$

With a constant C> 0, determine the bargain (trade-off) between the thinness of the function f (x) and the upper limit of the deviation more significant than ε is still tolerated. All deviations greater than ε will be subject to penalties of C [11]. The optimal solution can be solved with the following Lagrange functions (4).

$$\begin{aligned} Q(w, b, \xi, \xi^*, \alpha, \alpha^*, \eta, \eta^*) &= L \\ &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) - \sum_{i=1}^{l}\alpha_i(\varepsilon + \xi_i - y_i + w^T\varphi(x) + b) - \\ &\sum_{i=1}^{l}\alpha_i^*(\varepsilon + \xi_i^* + y_i - w^T\varphi(x) - b) - \sum_{i=1}^{l}(\eta_i\xi_i + \eta_i^*\xi_i^*) \end{aligned} \tag{4}$$

To get the optimal solution, you can do a partial derivative of Q with respect to w, b. From the equation above, w can be written as in (5).

$$w = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\varphi(x_i) \tag{5}$$

Then the optimal hyperplane function is writer as in (6).

$$f(x) = \sum_{i=1}^{l}(\alpha_i - \alpha_i^*)\varphi^T(x_i)\varphi(x_i) + b \tag{6}$$

Suppose beta = $\beta = (\alpha_i - \alpha_i^*)$

$$f(x) = \sum_{i=1}^{l}\beta\varphi^T(x_i)\varphi(x_i) + b \tag{7}$$

The dual optimization problems are:

$$Q(\alpha, \alpha^*) = \frac{1}{2}\sum_{i=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\varphi^T(x_i)\varphi(x_j) - \varepsilon\sum_{i=1}^{l}(\alpha_i - \alpha_i^*) + \sum_{i=1}^{l}y_i(\alpha_i - \alpha_i^*) \tag{8}$$

The optimal solution *b* by using KKT (*karush Kuhn Tucker*):

$$b = y_i - w^T\ \varphi(x_i) - \varepsilon \quad untuk\ 0 < \alpha_i < C \ \ (10)$$
$$b = y_i - w^T\ \varphi(x_i) + \varepsilon \quad untuk\ 0 < \alpha_i^* < C \tag{9}$$

B. *Kernel Function*

The kernel functions used in the Support Vector Regression method are as follows:

1. Linear: XT X
2. Polynomial : (XT X + 1)n
3. Radial Basis Function (RBF) : exp$(-\frac{1}{2\sigma^2}\|x - x_i\|^2)$

Choosing the right kernel function is essential for determining feature space.

C. *Grid Search Algorithm*

Based on the statement for [12], cross-validation is a standardized test carried out to predict error rates. The training data is randomly divided into several parts with the same comparison. The error rate is calculated section by section. The average error rate is calculated to get the overall error rate in cross-validation, known as leave-one-out validation (LOO). In LOO, data is divided into two subsets; subset 1 contains N-1 training data and one remaining data for testing [9].

$$CV = \sum_{i=1}^{n}(y_i - \hat{y}_{\neq i})^2 \quad (10)$$

III. Method

This study uses a population of mining sector companies listed on the Indonesian stock exchange, with samples taken from 2 coal mining sub-sector companies, namely PT Bukit Asam Tbk and PT Indo Tambangraya Megah Tbk. The data used are data of daily stock price for January 1, 2016, to December 31, 2019, with the variable used is close. The method chosen is support vector regression with RBF and linear kernels. The following stages of the analysis are carried out as in Fig 3.

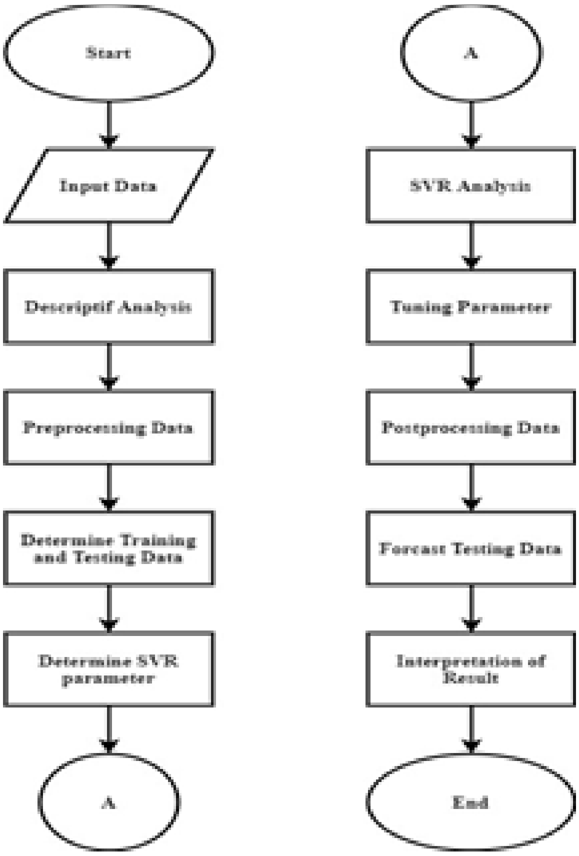

Fig. 3.The stages of the process

The steps Analysis consist of:

1. Preparing daily ADRO, PTBA, and ITMG stock data from January 1, 2015, to December 2019, which was downloaded from Yahoo Finance.
2. Conduct a descriptive analysis of ADRO, PTBA, and ITMG daily stock data.
3. They are preprocessing data, including defining dependent variables (Y) and independent variables (X). then transform the independent and dependent variables.
4. Divide data into two, namely training data and testing data.
5. Determine the kernel to be used and determine the cost (C) and gamma parameters to do the Support Vector Regression analysis.
6. Perform support vector regression analysis by determining parameters and kernels that have been determined by studying the literature first.
7. Tuning parameters to get optimal accuracy and minimum errors.
8. Post-processing, namely by denormalizing data to do forecast on testing data.
9. Forecasting data testing.
10. Interpreting the results of support vector regression analysis, which has obtained the best parameters and kernels.

IV. Results and Discussion

A. Model Evaluation of Support Vector Regression

The parameters used to form the PTBA and ITMG stock data model are linear kernel parameters and radial basis functions. Furthermore, this research's focus is on linear kernel parameters, radial basis functions with parameter C that is 10,100,1000 as tolerance vector support numbers to hyperplane. Gamma parameters for kernel radial basis functions are 0.1, 0.01, 0.001, 0.0001. The performance model that is formed is measured using R-square and MAPE's accuracy value; the more the R-square value approaches 1 (one), the better the model. However, the model must not be overfitting or prediction equal to the actual value. For MAPE to measure the model error in forecasting, the smaller the error value, the better the model formed.

Based on the analysis results using the support vector regression method with the distribution of training data by 80% and testing data by 20%, the accuracy values obtained are excellent (Table 1), with an average of more than 90 percent. Simultaneously, the MAPE or error values are still relatively high, so the model obtained is terrible because it is too high. A hyperparameter tunning is performed using the Grid Search algorithm to get optimal model performance for forecasting stock prices in the testing data.

Table 1. Accuracy values in Testing Data

Kernel	Accuracy Testing Data			
	PTBA		ITMG	
	R-square	MAPE	R-square	MAPE
Linear	0.957	28.98	0.902	40.52
RBF	0.987	30.17	0.981	40.05

Based on the results of tuning parameters that have been done (Table 2), it is found that the performance of the model is optimal. For PTBA shares, the optimal model obtained is in the RBF kernel with an accuracy value of 97.8 percent with a MAPE value or an error of 2.16 percent with an optimal parameter C = 1000, gamma = 0.01. As for the ITMG stock, the optimal model formed is linear kernel and RBF with an accuracy value of 90.6 percent and MAPE value or error of 7.09 percent with the optimal parameter C = 1000, gamma = 0.001 for the RBF kernel, for the linear kernel the optimal parameter C = 10.

Table 2. Accuracy Values in Testing Data After Tunning Hyper Parameter

Kernel	Accuracy Testing Data			
	PTBA		ITMG	
	R-square	MAPE	R-square	MAPE
Linear	0.978	2.17	0.906	7.09
RBF	0.978	2.16	0.906	7.09

B. Forcast result

The next step is to forecast each company's stock prices using the best model that has been formed previously from the experiment parameters result to form a support vector regression model that has been done.

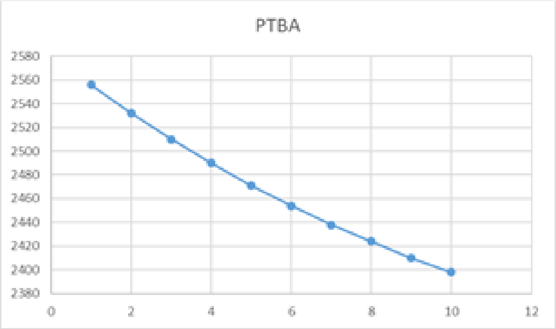

Fig. 4. Forecasting result PTBA stock

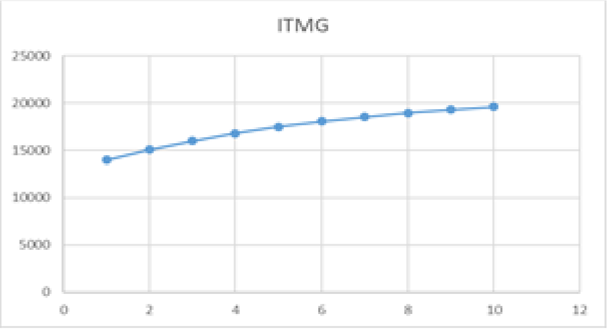

Fig. 5. Forecasting result ITMG stock

The result of the forecasting of PTBA (Fig. 4), and ITMG (Fig. 5) stock prices for the next ten periods. ITMG shares tended to increase while PTBA shares decrease.

V. Conclusion

The SVR method can be applied to forecast PTBA and ITMG stock prices. The optimal SVR model obtained for PTBA stock data with a radial kernel base function with parameters C = 1000 and gamma = 0.01 with an accuracy value of 97.8% and a MAPE value of 2.16. Then for ITMG stock data obtained, an optimal SVR model with a radial kernel base function with parameters C = 1000 and gamma = 0.001, linear kernels with parameter C = 10 with accuracy model of 940.6% in the testing data and a MAPE value of 7.09. Forecasting results of PTBA and ITMG stock prices for the next ten periods in ITMG shares tended to increase while those in PTBA shares experienced a decrease.

References

[1] Jogiyanto, H. (2008). *Teori Portofolio dan Analisis Investasi.* Yogyakarta: BPFE.

[2] BEI. (2018). *Saham.* Retrieved from IDX: https://www.idx.co.id/produk/saham/.

[3] Rahmadayanti, C., Rabbani, H., & Rohmawati, A. (2018). Model GARCH dengan Pendekatan Conditional Maximum Likelihood untuk Prediksi Harga Saham. *Ind. Journal on Computing, 21-28.*

[4] Yasin, H., Prahutama, A., & Utami, T. W. (2014). Prediksi Harga Saham Menggunakan Support VEctor Regression Dengan Algoritma Grid Search. *Media Statistik.*

[5] Ahmad, A. (2017). Mengenal Artificial Intellegence, Machine Learning, Neural Network, dan Deep Learning. *Yayasan Cahaya Islam Jurnal Teknologi Indonesia*.

[6] Chen, F.L, and Li, F.C. (2010). Combination of feature selection approaches with SVM in credit scoring. *Expert System Application 37, 4902-4909.*

[7] Maulana, N. D., Setiawan, B. D., & Dewi, C. (2019). Implementasi Metode Support Vector Regression (SVR) Dalam Peramalan Penjualan Roti. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*.

[8] Furi, R. P., Jondri, & Saepudin, D. (2015). Peramalan Financial Time Series Menggunakan Independent Component Analysis dan Support Vector Regression. *e-Proceeding of Engineering vol.2*.

[9] Santosa, B. (2007). *Data Mining Teknik Pemanfaatan Data untuk Keperluan Bisnis.* Yogyakarta: Graha Ilmu.

[10] Gunn, S. (1998). *Support Vector Machine for Clasification an Regression*.

[11] Amanda, R., Yasin, H., & Prahutama, A. (2014). Analisis Support Vector Regression (SVR) Dalam Memprediksi Kurs Rupiah Terhadap Dollar Amerika Serikat. *JURNAL GAUSSIAN.*

[12] Leidiyana, H. (2013). Penerapan Algoritma K-Nearest Neighbor Untuk Penentuan Risiko Kredit Kempemilikan Kendaraan Bermotor. *Jurnal Penelitian Ilmu Komputer.*

[13] Alan, P.,Prahutama, A. (2015). Prediction of Weekly Rainfall in Semarang City Use Support Vector Regression (SVR) with Quadratic Loss Function. *International Journal Of Science and Engineering.*

[14] Debanjan, P., Chakraborty, M. (2020). A python based support vector regression model for prediction of COVID19 cases in India. *Chaos, Solitons, and Fractals 138.*

[15] Maryam, H., Mekarthy, S. M. R., Cherukuri, H. (2020). Prediction of specific cutting forces and maximum tool temperatures in orthogonal by Support Vector and Gaussian Process Regression Method. *Procedia Manufacturing 48 1000-1008.*

[16] Yaman, H., *et al.* (2020). Two steps hybrid algorithm of support vector regression and K-nearest neighbors. *Alexandria Engineering Journal.*

[17] Ahmed, B., Mihoubi, M. K., Santillan, D. (2019). Seepage and dam deformation analyses with statistical models: support vector regression machine and random forest. *Procedia Structural Integrity.*

[18] He, Y., *et al.* (2019). Uncertainty Forecasting for Streamflow based on Support Vector Regression Method with Fuzzy Information Granulation. *Energy Procedia 158 6189-6194.*

[19] Ma, Z., Ye, C., Ma, W. (2019). Support vector regression for predicting building energy consumption in southern China. *Energy Procedia 158 3433-3438.*

[20] JinXing, C., and JianZ.W. (2014). Short Term Load forecasting using kernel based support vector regression combination model. *Applied Energy,* vol.132 pp 602-609.