Media Statistika

p-ISSN 1979 – 3693 e-ISSN 2477 – 0647
MEDIA STATISTIKA 15(2) 2022: 163-174
http://ejournal.undip.ac.id/index.php/media_statistika

MEASUREMENT OF SUPPORT VECTOR REGRESSION PERFORMANCE WITH CLUSTER ANALYSIS FOR STOCK PRICE MODELING

Izza Dinikal Arsy, Dedi Rosadi
Statistics Study Program, Universitas Gadjah Mada, Indonesia

e-mail: arsyizzadinikal@gmail.com

DOI: 10.14710/medstat.15.2.163-174

Article Info:
Received: 29 July 2022
Accepted: 2 January 2023
Available Online: 4 April 2023

Keywords:
Support Vector Regression; Stock; Cluster; Volatility
.

Abstract: Risk-averse investors will seek out stock investments with the minimum risk. One step that can be taken is to develop a model of stock prices and predict their fluctuations in the coming months. Significant studies on the modeling of stock movements have used the ARCH/GARCH method, but this method requires some assumptions. This paper will discuss the performance of stock modeling using Support Vector Regression. The performance is measured using the root mean square error value in two stock clusters based on its volatility value, e.g., stocks with large volatility and stocks with small volatility. This case study makes use of daily closing price data from 10 LQ-45 index shares from October 12, 2018 to October 11, 2019. In conclusion, SVR's performance on stocks with high volatility produces RMSE, which is considerably higher than SVR's performance on stocks with low volatility.

1. INTRODUCTION

Investors conduct investment activities in the form of financial assets or real assets. Financial assets such as stocks are employed as investment objects in this study. Before making investing decisions, investors must carefully analyze the funds to be invested in and the stocks to be chosen. An investor expects maximum profit with minimal risk. However, in its implementation, the profits are always directly proportional to the risks. Investors who are risk-averse will look for the types of stock investments with minimum risk. As a result, the first step is to create a model that predicts stock prices in the future.

However, in practice, stock price movements are extremely volatile. As a result, not all stocks can be modeled and predicted with a low error rate. Intuitively, it can be understood that stocks with stable movements will be easier to model and produce lower errors than stocks with volatile movements. The indicator that reflects fluctuations in stock movements is volatility. Therefore, investors need to sort and choose the stocks that can be predicted with low errors based on their volatility values so that future investment risks can be predicted and minimized.

Various studies on modeling or predicting market movements have used the ARCH/GARCH approach, which needs several assumptions. However, along with the development of statistical science, various new methods are currently being formulated which are commonly referred to as Machine Learning, where one of the methods is Support Vector Regression. The selection of this method is based on the fact that the modeling

concept is quite simple and has a relatively high performance. The concept of this method is to choose the best hyperplane that can cover all the information from the data.

The calculation of the Support Vector Regression method depends on the distribution of the data or, in this case, the volatility of the stock price. Therefore, in this study, the researchers are interested in adding cluster elements based on stock volatility in order to determine the performance of the method on stocks in each cluster.

Other papers or journals used to support stock price modeling using Support Vector Regression in this undergraduate thesis include Basak et al. (2007), Bini and Mathew (2015), Choudhury et al. (2013), Henrique et al. (2018), Maharesi (2013), Mishra and Padhy (2019), and Saputra et al. (2019).

2. LITERATURE REVIEW

2.1. K-Means Clustering Algorithm

The K-Means algorithm is one of the partitional algorithms. K-Means is based on determining the initial number of groups by defining the initial centroid value. The term K-Means comes from the formation of K clusters with the new centroid value being the mean of the data in each cluster. The K-Means algorithm uses an iterative process to obtain a cluster database. The centroid value chosen to be the initial center will be calculated using the Euclidean Distance formula, which is to find the closest distance between the centroid point and the data/object. Data that has the shortest or closest distance to the centroid will create a cluster.

The K-Means algorithm is as follows:

1. Determine *K* as the number of clusters to be formed
2. Randomly determine the initial K Centroid (cluster center point)
3. Calculate the distance of each object to each centroid of each cluster. To calculate the distance between the object and the centroid, the Euclidian Distance formula between two objects can be used

   $$d(x,y) = \sum_{i=1}^{n} (x_i - y_i)^2 ; \qquad i = 1, 2, 3, \dots, n$$

   where x_i is i^{th} value of object x, y_i is i^{th} value of object y, and n is number of objects
4. Allocate each object into a cluster where the distance of the object to the centroid is the closest distance compared to the distance of the object to the centroid of other clusters. If an object is found within the same distance from 2 or more centroids, then the K Centroid is re-determined as in step 2
5. Perform an iteration, then determine the position of the new centroid using the equation

   $$v_k = \frac{\sum_{i=1}^{n_k} x_{ik}}{n_k}; \qquad i = 1, 2, 3, \dots, n_k$$

   where v_k is centroid of the k^{th} cluster, x_{ik} is i^{th} value of object of the k^{th} cluster, and n_k is number of objects that are members of k^{th} cluster

Repeat step (3) to step (5) until the new centroid position is consistently the same as the previous centroid position.

2.2. Simple Moving Average

Moving average is one of the methods used to forecast data that contains a trend. This method is done by taking a group of values, examining the average, and then using the average as a forecast for the next period. This method is called a moving average because

every time new observation data is available, the average number is calculated and used for future use.

The Properties of the Simple Moving Average (Subagyo, 1986) are as follows:

1. To create a forecast value requires historical data over a certain period of time. If there is data for n periods, it can make forecasts for the $n + 1$ period.
2. The longer the moving average, the smoother the moving average will be.

However, this approach has drawbacks in addition to its benefits, including (Subagyo, 1986):

1. Requirement of historical data
   This method requires sufficient historical data. For forecasts with 3 months moving average, historical data for the last 3 months is required.
2. All data are equally weighted
   According to this method all data are equally weighted. The formula of the Simple Moving Average is as follows:
   $$S_{t+1} = \frac{X_t + X_{t-1} + \cdots + X_{t-n+1}}{n}$$
   where S_{t+1} is forecasting for $t + 1$ period, X_t is data in t period, n is moving average timeframe

2.3. Karush Kuhn-Tucker Condition

The Karush Kuhn-Tucker model can be used to solve a linear or non-linear function. In this method, the completed program has inequality constraints. The Karush Kuhn-Tucker method is a development of the solution of a non-linear model constrained by equations by seeking for stationary points, or places with the potential to be optimal points.

There are several Karush Khun-Tucker requirements for constrained optimization problems. Karush and Khun-Tucker formulated the requirement. The following is a theorem that explains the Karush Kuhn-Tucker conditions for the maximum and minimum problems.

Theorem 1 (Winston, 2003) *If $f(X)$ and $g_1(X)$ are a maximizing pattern problem. If $X = (x_1, x_2, \dots, x_n)$ is an optimal solution for $f(X)$ and $g_i(X)$, then $X = (x_1, x_2, \dots, x_n)$ is a non-linear function and there are multipliers $\lambda_1, \lambda_{2,\dots,}\lambda_m$ and slack variables $s_1, s_{2,\dots,}s_n$ so that it satisfies*

1. $\dfrac{\partial f}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \dfrac{\partial g_i}{\partial x_j} + s_j = 0;$ $\quad for\ j = 1, 2, 3, \dots, n$
2. $\lambda_i [b_i - g_i(X)] = 0;$ $\quad for\ i = 1, 2, 3, \dots, m$
3. $\left(\dfrac{\partial f}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \dfrac{\partial g_i}{\partial x_j}\right) x_j = 0;$ $\quad for\ j = 1, 2, 3, \dots, n$
4. $\lambda_i \geq 0;$ $\quad for\ i = 1, 2, 3, \dots, m$
5. $s_j \geq 0;$ $\quad for\ j = 1, 2, 3, \dots, n$

Theorem 2 (Winston, 2003) *If $f(X)$ and $g_1(X)$ are a problem with a minimization pattern. If $X = (x_1, x_2, \dots, x_n)$ is an optimal solution for $f(X)$ and $g_i(X)$, then $X = (x_1, x_2, \dots, x_n)$ is a non-linear function and there are multipliers $\lambda_1, \lambda_{2,\dots,}\lambda_m$ and surplus variables $e_1, e_{2,\dots,}e_n$ so that it satisfies*

1. $\frac{\partial f}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \frac{\partial g_i}{\partial x_j} - e_j = 0;$ $for\ j = 1, 2, 3, \dots, n$
2. $\lambda_i [b_i - g_i(X)] = 0;$ $for\ i = 1, 2, 3, \dots, m$
3. $\left(\frac{\partial f}{\partial x_j} - \sum_{i=1}^{m} \lambda_i \frac{\partial g_i}{\partial x_j}\right) x_j = 0;$ $for\ j = 1, 2, 3, \dots, n$
4. $\lambda_i \geq 0;$ $for\ i = 1, 2, 3, \dots, m$
5. $e_j \geq 0;$ $for\ j = 1, 2, 3, \dots, n$

In general, the condition of complementary slackness in quadratic programming can be expressed in property 1 as follows:

Property 1 (Winston, 2003) *Complementary slackness in quadratic programming*
1. *e_j and s_j in the Kuhn-Tucker condition and x_j neither can be positive.*
2. *Surplus variable (excess) or slack for the i^{th} value of constraint and λ_i neither can be positive.*

2.4. Support Vector Regression

Support Vector Regression (SVR) is an extension of the Support Vector Machine. While the Support Vector Machine aims to classify input data, the Support Vector Regression is a method that seeks to solve the regression problem.

2.4.1.Support Vector Regression Model

This SVR model is able to overcome overfitting because it uses the principle of Structural Risk Minimization (SRM) to estimate a regression function by minimizing the upper limit of the generalization error resulting in a superior performance (Smola and Schölkopf, 2004). The function of the SVR is as follows:

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

where y is the output data matrix, $\boldsymbol{x}$ is the input data matrix, $\boldsymbol{w}$ is the weight matrix, and $\boldsymbol{b}$ is the bias matrix.

SVR is presently becoming the subject of numerous studies that have been developed. Until now, many types of SVR have been produced, one of which is ε-SVR.

2.4.2.$\boldsymbol{\varepsilon}$-SVR

$\boldsymbol{\varepsilon}$-SVR was introduced by Vapnik, who added the concept of $\boldsymbol{\varepsilon}$-insensitive loss function. This concept is represented by ε notation, which reflects the magnitude of the deviation from the actual target of y_i value for all training data. The value of ε is the limit of the range of y_i values that can be estimated by the function. Visually, ε-SVR will form a *tube* with a radius equal to the value of ε. A perfect regression equation can be obtained if $\varepsilon = 0$ because the analysis includes all data without having to use a support vector. Figure 1 will explain the visualization of ε-SVR where the dotted line shows the supporting hyperplane, the straight line shows the main hyperplane, and the red dots that intersect with the dotted line are support vectors.

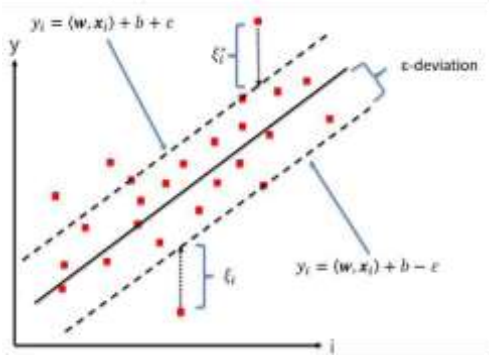

Figure 1. ε-SVR Visualization

Given as many as N training data sets (x_i, y_i) with $i = 1,2,\ldots,N$ where $x \in \mathbb{R}$ is a vector in the input space and $y_i \in \mathbb{R}$ is the output value based on corresponding x_i, the primal form of SVR with a precision of ε is as follows:

$$\min_{w,b} \frac{1}{2}\|\boldsymbol{w}\|^2 \tag{2}$$

condition

$$y_i - (\boldsymbol{w}^T x_i + b) \leq \varepsilon$$
$$(\boldsymbol{w}^T x_i + b) - y_i \leq \varepsilon \qquad i = 1,2,\ldots,N$$

Minimizing $||\boldsymbol{w}||$ will make the function as thin as possible, so as to control the capacity of the function. In the case of this regression, it is assumed that all pairs of points (x_i, y_i) are within the range $f \pm \varepsilon$ (feasible). In an unfeasible situation where there are several points that might be out of range $f \pm \varepsilon$, it is necessary to add ξ, ξ^* slack variable into equation (2) to overcome the problem of infeasible constraints in the optimization problem as shown in Figure 1 at the points colored Red. Furthermore, the optimization problem in equation (2) can be formulated as follows (Cortes & Vapnik, 1995):

$$\min_{w,b,\xi_i,\xi_i^*} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) \tag{3}$$

condition

$$y_i - (\boldsymbol{w}^T x_i + b) \leq \varepsilon + \xi_i$$
$$(\boldsymbol{w}^T x_i + b) - y_i \leq \varepsilon + \xi_i^*$$
$$\xi_i, \xi_i^* \geq 0 \qquad i = 1,2,\ldots,N$$

The constant $C > 0$ determines the bargaining value between the thinness of the function f and the maximum deviation limit greater than ε that can still be tolerated. That means that any deviation with a value greater than ε will be penalized by C.

In SVR, ε is equivalent to the accuracy of the approximation effort to the value of the training data. A small value of ε allows a high level of accuracy for the approximation function and a high value for the slack variable .ξ_i, ξ_i^*. On the other hand, a high value for ε indicates a small accuracy of the approximation function. According to equation (3), a low variable value and a high slack value will make empirical errors and have a considerable influence on the $||\boldsymbol{w}||$ regularization factor. In SVR, the support vector is the training data that lies on and outside of the $\pm\varepsilon$ limit of the decision function. This causes the number of

support vectors to decrease as the value of ε increases. This relationship characterizes the role of ε function, which is denoted by $|\xi|_\varepsilon$ with the following details:

$$|\xi|_\varepsilon = \begin{cases} 0, & \xi \le \varepsilon \\ |\xi| - \varepsilon, & \text{otherwise} \end{cases}$$

Before converting into dual form, first consider for the Lagrange form of the primal form of SVR in equation (3), namely:

$$L(\boldsymbol{w}, b, \xi_i, \xi_i^*) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{N}(\xi_i + \xi_i^*) - \sum_{i=1}^{N}(\eta_i\xi_i + \eta_i^*\xi_i^*) \quad (4)$$

$$-\sum_{i=1}^{N}\alpha_i(\varepsilon + \xi_i - y_i + (\boldsymbol{w}^T x_i) + b) - \sum_{i=1}^{N}\alpha_i^*(\varepsilon + \xi_i^* + y_i - (\boldsymbol{w}^T x_i) - b)$$

where the non-negative variable of $\alpha_i, \alpha_i^*, \eta_i, \eta_i^*$ is a Lagrange multiplier. The Lagrange equation (4) must satisfy the limiting equation of

$$\alpha_i, \alpha_i^*, \eta_i, \eta_i^* \ge 0$$

Furthermore, the partial reduction of the Lagrange function to w, b, ξ_i, ξ_i^* values are as follows:

Condition 1:

$$\frac{\partial L(\boldsymbol{w}, b, \xi_i, \xi_i^*)}{\partial \boldsymbol{w}} = 0$$

$$w - \sum_{i=1}^{N}(\alpha_i x_i) + \sum_{i=1}^{N}(\alpha_i^* x_i) = 0$$

$$w = \sum_{i=1}^{N}(\alpha_i - \alpha_i^*)\, x_i \quad (5)$$

Condition 2:

$$\frac{\partial L(\boldsymbol{w}, b, \xi_i, \xi_i^*)}{\partial b} = 0$$

$$-\sum_{i=1}^{N}\alpha_i + \sum_{i=1}^{N}\alpha_i^* = 0$$

$$\sum_{i=1}^{N}(\alpha_i - \alpha_i^*) = 0 \quad (6)$$

Condition 3:

$$\frac{\partial L(\boldsymbol{w}, b, \xi_i, \xi_i^*)}{\partial \xi_i} = 0$$

$$C - \sum_{i=1}^{N} n_i - \sum_{i=1}^{N}\alpha_i = 0$$

$$C = \sum_{i=1}^{N}(\alpha_i + n_i) \quad (7)$$

Condition 4:

$$\frac{\partial L(\boldsymbol{w}, b, \xi_i, {\xi_i}^*)}{\partial {\xi_i}^*} = 0$$

$$C - \sum_{i=1}^{N} {n_i}^* - \sum_{i=1}^{N} {\alpha_i}^* = 0$$

$$C = \sum_{i=1}^{N} ({\alpha_i}^* + {n_i}^*) \tag{8}$$

By eliminating equations (7) and (8), then substituting into equation (6), equation (9) is obtained as follows:

$$\sum_{i=1}^{N} (\eta_i - \eta_i^*) = 0 \tag{9}$$

By inserting equations (5), (6), (7) and (9) into the Lagrange function in equation (4), it will be obtained

$$L(\boldsymbol{w}, b, a) = -\frac{1}{2}\left[\sum_{i=1}^{N}\sum_{j=1}^{N} (\alpha_i - \alpha_i^*)\,(\alpha_j - \alpha_j^*) x_i x_j\right] - \varepsilon \sum_{i=1}^{N} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i (\alpha_i - \alpha_i^*) \tag{10}$$

Furthermore, the Lagrange equation (10) produces a dual equation for the optimization problem of ε-SVR as follows:

$$\max_{\alpha, \alpha^*} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} (\alpha_i - \alpha_i^*)\,(\alpha_j - \alpha_j^*)(x_i, x_j) - \varepsilon \sum_{i=1}^{N} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{N} y_i (\alpha_i - \alpha_i^*)$$

condition

$$\sum_{i=1}^{N} (\alpha_i - \alpha_i^*) = 0$$

$$0 \le \alpha_{i,} \alpha_i^* \le C \qquad\qquad i = 1,2, \dots, N$$

where C is defined as a penalty, (x_i, x_j) is dot-product where C is defined as $(x_i, x_j) = (x_i) \cdot (x_j)$. Furthermore, equation (5) obtained a new form of $w = \sum_{i=1}^{N} (\alpha_i - \alpha_i^*)\, x_i$ which can be included in equation (1) and formulated as follows:

$$f(x) = \left(\sum_{i=1}^{N} (\alpha_i - \alpha_i^*)\, x_i\right) x + b \tag{11}$$

Equation (11) can be called the Support Vector expansion, which is a weight matrix of $\boldsymbol{w}$ described by a linear combination of training samples. The complexity of the regression function does not depend on the dimensions of the input data, but entirely on the number of support vectors associated with the Lagrange multiplier value of α_i, α_i^*.

The value of b can be found using the Karush Kuhn-Tucker condition stating that at the optimal solution value, the dot product of the dual variable, and the constraint function will cancel each other. If written briefly, the equation notation will be adjusted with the notation listed in the Karush Kuhn-Tucker condition as follows:

$$f(\xi,\xi^*) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + {\xi_i}^*)$$

$$f(\xi,\xi^*) = \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}(\xi_i + {\xi_i}^*)$$

$$g_1(\xi,\xi^*) = y_i - (\boldsymbol{w}^T x_i + b) - \xi_i$$

$$b_1 = \varepsilon$$

$$g_2(\xi,\xi^*) = (\boldsymbol{w}^T x_i + b) - y_i - \xi_i^*$$

$$b_2 = \varepsilon$$

$$\lambda_1 = \alpha_i$$

$$\lambda_2 = {\alpha_i}^*$$

This is due to the fact that the slack variable of s_j from the Karush Khun–Tucker condition is not a general component that has been defined in the optimization problem. The value of s_j will appear according to the existing optimization problem while still following Property 1. In the case of Support Vector Regression optimization, the variable will be defined as s_i. Furthermore, from the conditions of $\frac{\partial f}{\partial \xi_i} - \sum_{i=1}^{m}\lambda_i\frac{\partial g_i}{\partial \xi_i} + s_j = 0$ obtained

$$C - \alpha_i + s_i = 0 \leftrightarrow C - \alpha_i = -s_i \quad (12a)$$

$$C - {\alpha_i}^* + s_i = 0 \leftrightarrow C - {\alpha_i}^* = -s_i \quad (12b)$$

then from the condition of $\lambda_i[b_i - g_i(\xi,\xi^*)] = 0$ obtained

$$\alpha_i(\varepsilon - y_i + (\boldsymbol{w}^T x_i) + b + \xi_i) = 0 \quad (13a)$$

$$\alpha_i{}^*(\varepsilon + y_i - (\boldsymbol{w}^T x_i) - b + {\xi_i}^*) = 0 \quad (13b)$$

and from the condition of $\left(\frac{\partial f}{\partial \xi_i} - \sum_{i=1}^{m}\lambda_i\frac{\partial g_i}{\partial \xi_i}\right)\xi_i = 0$

$$(C - \alpha_i)\xi_i = 0 \quad (14a)$$

$$(C - {\alpha_i}^*){\xi_i}^* = 0 \quad (14b)$$

From the conditions (12a), (12b), (14a), (14b) and Property 1, the value of s_i will be zero when $\xi_i, {\xi_i}^*$ is positive or in other words when the data is outside the boundary. On the other hand, s_i will be positive when $\xi_i, {\xi_i}^*$ is zero or in other words when the data is inside the boundary.

Furthermore, a Support Vector forming *w*, namely training data with a value of $\alpha_i, {\alpha_i}^* > 0$. Then the Support Vector can be found when:

i. $\varepsilon + {\xi_i}^* + y_i - (\boldsymbol{w}^T x_i) - b = 0$ or the training data is located at the upper boundary or above the upper boundary.
ii. $\varepsilon + \xi_i - y_i + (\boldsymbol{w}^T x_i) + b = 0$ or the training data is located at the lower boundary or below the lower boundary.

Furthermore, the value of *b* can be determined by considering (13a), (13b), (14a) and (14b) when:

i. $\alpha_i, {\alpha_i}^* = C$ for $\xi_i, {\xi_i}^* \neq 0$ or i[th] value of data is out of limit $f(x) \pm \varepsilon$.
ii. $\xi_i, {\xi_i}^* = 0$ for $0 \leq \alpha_{i,}\alpha_i^* \leq C$ or i[th] value of data is within the limit $f(x) \pm \varepsilon$.

In order to produce a precise value of *f(x)* on the data, the calculation of the value of b can be done using condition (ii) or in other words involving data that is within the limit, $f(x) \pm \varepsilon$ namely when the value of $\xi_i, \xi_i^* = 0$ and $0 < \alpha_i, \alpha_i^* \leq C$. Therefore, the value of b can be written as

$$b = y_i - (\boldsymbol{w}^T x_i) - \varepsilon; \quad for\ \alpha_i \in (0, C) \tag{15a}$$

$$b = y_i - (\boldsymbol{w}^T x_i) + \varepsilon; \quad for\ \alpha_i{}^* \in (0, C) \tag{15b}$$

3. RESEARCH METHODS

This paper discusses the application of Support Vector Regression optimization with K-means cluster analysis to the stock data variance of 10 stocks. To evaluate the performance of Support Vector Regression in each cluster based on the variance value, each cluster was constructed based on the variance value of the stock data.

3.1. Data Description

The objects in this study were stocks. Secondary data in the form of daily stock price data was employed. The closing price of the stock was utilized, which is the price that displays right before the market closes. The stocks employed in this study were ten equities from the list of LQ-45 stock groups. Furthermore, the stocks chosen had the biggest capitalization in each area. The observation period lasted 261 days, from October 11, 2018 to October 11, 2019. This period represents the total number of active trading days in the stock for one year.

3.2. Research Method

The method employed in this case study was a process of comparing the performance of Support Vector Regression in a cluster constructed using K-means Clustering based on the variance value of stock data. The reference for the performance comparison, in this case, is the root mean square error generated by the Support Vector Regression model.

Considering stock close price data is a univariate time series, additional data is required as input for Support Vector Regression modeling. The input data used for the formation of the model was ith value of lag of the data with *i* being the value of the best order simple moving average when used to model the data for each stock using the simple moving average method. Because the input data is lag of the actual data, there will be as many as *i* last data that do not have input data so that data needs to be removed from the dataset.

Furthermore, the modeling was done by splitting the data into training data and testing data with a ratio of 80:20. The training data set included 208 data points for each stock's first close price from October 12, 2018, to July 31, 2019, whereas the test data set included 52 closing price data points from August 1, 2019 to October 11, 2019. The model that was formed based on the training data was measured for its performance by using the model to predict the test data and calculating the root mean square error value.

Table 1. Data Description

Data	Daily stock price
Source	www.finance.yahoo.com
Period	October 11, 2018 – October 11, 2019
Criteria	LQ-45
Observation frequency	261 days

Table 2. Stock List

No	Stock Code	Stock Name
1	ASII.JK	Astra International Tbk.
2	BBNI.JK	Bank Negara Indonesia (Persero) Tbk.
3	EXCL.JK	XL Axiata Tbk.
4	TPIA.JK	Chandra Asri Petrochemical Tbk.
5	SMGR.JK	Semen Indonesia (Persero) Tbk.
6	WIKA.JK	Wijaya Karya (Persero) Tbk.
7	BMRI.JK	Bank Mandiri (Persero) Tbk.
8	ICBP.JK	Indofood CBP Sukses Makmur Tbk.
9	ADHI.JK	Adhi Karya (Persero) Tbk.
10	BBRI.JK	Bank Rakyat Indonesia (Persero)

4. RESULTS AND DISCUSSION

4.1. Cluster formation based on variance

The Support Vector Regression approach searches for a hyperplane with the smallest thickness as a model equation, where the thickness is affected by the data distribution. Therefore, two clusters were formed using K-means Clustering based on the variance value of each stock. The two clusters formed displayed stock categories with high and small deviations. Table 3 and 4 represent the variance value and the results of the clustering.

Table 3. Stock Variance

Stock	Variance	Stock	Variance
ASII.JK	395,367.87	WIKA.JK	137,281.15
BBNI.JK	561,224.64	BMRI.JK	161,473.64
EXCL.JK	221,077.48	ICBP.JK	953,165.25
TPIA.JK	1,585,411.91	ADHI.JK	25,685.16
SMGR.JK	1,503,236.70	BBRI.JK	142,798.14

Table 4. Stock Clusters

Small Variance Cluster Stock	Large Variance luster Stock
ASII.JK	TPIA.JK
BBNI.JK	SMGR.JK
EXCL.JK	ICBP.JK
WIKA.JK	-
BMRI.JK	-
ADHI.JK	-
BBRI.JK	-

4.2. Stock Price Modeling using SVR

Furthermore, using the Simple Moving Average, it was found that the ten stocks in this case study can be modeled effectively using a lag order 1, which means yesterday's data $(t-1)$ can be used to predict today's data (t). Therefore, the data lag 1 was used as the independent variable and the data at time t, which was used as the dependent variable. The modeling results of each stock are presented in Table 5.

Table 5. Support Vector Regression Model

Stock	Kernel, cost, gamma, epsilon	Number of Support Vectors	W	b	RMSE
ASII.JK	Linear,1,1, 0,1	125	0.9670126	-0.011618	126.059
BBNI.JK	Linear,10,1, 0,1	112	0.9877608	-0.008210	150.409
EXCL.JK	Linear,10,1, 0,3	20	0.9935115	-0.021029	74.428
TPIA.JK	Linear,1,1, 0,1	116	0.9595929	0.012225	210.666
SMGR.JK	Linear,1,1, 0,1	106	0.9758033	0.006688	264.748
WIKA.JK	Linear,1,1, 0,1	81	0.9855069	-0.006720	53.949
BMRI.JK	Linear,1,1, 0,1	146	0.9728070	0.011467	123.751
ICBP.JK	Linear,10,1, 0,3	27	0.9754748	0.000103	170.382
ADHI.JK	Linear,1,1, 0,1	125	0.9640289	0.004790	25.608
BBRI.JK	Linear,10,1, 0,3	8	0.9907833	-0.004442	73.100

As an example of the interpretation of ASII.JK stock with a linear kernel, the value of *cost* = 1 and tolerance of epsilon = 0.1 was considered. As a result, 125 data that form the weights of w = 0.9670126 and b = -0.011618 and produce a root mean square error of 126.059 were observed. From the model and RMSE formed in the previous sub-chapter, the performance of the model in each cluster was compared. For stocks belonging to clusters with small variances, the results obtained are tabulated in Table 6. Meanwhile, for stocks belonging to clusters with large variances, the results are illustrated in Table 7.

It can be seen that the stocks belong to a cluster with small variance and large variance. This is in accordance with the concept of Support Vector Regression that the model will look for the best hyperplane composed of support vectors or data that are close to each other. Therefore, if the variance or data distribution is greater, the hyperplane formed will produce a larger error compared to a dataset that has a smaller variance or data distribution value.

Table 6. Root Mean Square Error of Cluster Stocks

Stock	Cluster	RMSE	Stock	Cluster	RMSE
ASII.JK	Small Variance	126.059	TPIA.JK	Large Variance	210.666
BBNI.JK	Small Variance	150.409	SMGR.JK	Large Variance	264.748
EXCL.JK	Small Variance	74.428	ICBP.JK	Large Variance	170.382
WIKA.JK	Small Variance	53.949			
BMRI.JK	Small Variance	123.751			
ADHI.JK	Small Variance	25.608			
BBRI.JK	Small Variance	73.100			

5. CONCLUSION

Based on the results of this study, it can be seen that from the 10 LQ-45 stocks used in the case study, there are 3 stocks classified as large variance clusters and 7 stocks classified as small variance clusters. Furthermore, Support Vector Regression modeling can be used for stock modeling by using the results of the Simple Moving Average analysis of order 1 as the dependent variable. The results of the modeling show that stocks belonging to the large variance cluster have a larger RMSE than stocks belonging to the small variance cluster when modeled using Support Vector Regression.

REFERENCES

Bain, L. J. and Engelhardt, M. (1992). *Introduction to Probability and Mathematical Statistics*. California: Duxbury Press.

Basak, D., Pal, S., and Patranabis, D. C. (2007). Support Vector Regression. *Neural Information Processing*, Vol. 11, No. 10.

Bini, B. S. and Mathew, T. (2016). Clustering and Regression Techniques for Stock Prediction. *Procedia Technology*, *24*, 1248–1255.

Biri, R., Langi, Y. A., and Paendong, M. S. (2013). Penggunaan Metode Smoothing Eksponensial dalam Meramal Pergerakan Inflasi Kota Palu. *Jurnal Ilmiah Sains*, *13*(1), 68.

Choudhury, S., Ghosh, S., Bhattacharya, A., Fernandes, K. J., and Tiwari, M. K. (2014). A Real Time Clustering and SVM Based Price-Volatility Prediction for Optimal Trading Strategy. *Neurocomputing*, *131*, 419–426.

Cortes, C. and Vapnik, V. (1995). Support-Vector Networks. *Machine Learning 20*, 273–297.

Henrique, B. M., Sobreiro, V. A., and Kimura, H. (2018). Stock Price Prediction Using Support Vector Regression on Daily and up to the Minute Prices. *Journal of Finance and Data Science*, *4*(3), 183–201.

Kowalczyk, A. (2017). Support Vector Machines Succintctly, Syncfusion. *Succinctly E-Book Series*, 114. www.syncfusion.com.

Maharesi, R. (2013). Penggunaan Support Vector Regression (SVR) pada Prediksi Return Saham Syariah BEI. *Proceeding PESAT*, *5*, 8–9.

Mishra, S. and Padhy, S. (2019). An Efficient Portfolio Construction Model using Stock Price Predicted by Support Vector Regression. *North American Journal of Economics and Finance*, *50*(May), 101027.

Prahutama, A., Utami, T. W., and Yasin, H. (2014). Prediksi Harga Saham Menggunakan Support Vector Regression Dengan Algoritma Grid Search. *Media Statistika*, *7*(1), 29–35.

Rosadi, D. (2014). *Analisis Runtun Waktu dan Aplikasinya dengan R*. Yogyakarta. Gadjah Mada University Press.

Saputra, G. H., Wigena, A. H., and Sartono, B. (2019). Penggunaan Support Vector Regression dalam Pemodelan Indeks Saham Syariah Indonesia dengan Algoritme Grid Search. *Indonesian Journal of Statistics and Its Applications*, *3*(2), 148–160. https://doi.org/10.29244/ijsa.v3i2.172

Subagyo, P. (1986). *Forecasting Konsep dan Aplikasi*, Yogyakarta: BPFE Yogyakarta,

Smola, A. J. and Schölkopf, B. (2004). A Tutorial for Support Vector Regression. *Statistics and Computing 14*, 199–222.

Winston, W. L. (2004). *Operations Research: Application*. Boston: Duxbury Press

Yahoo Inc. (2019). *Yahoo! Finance*. URL: http://www.finance.yahoo.com/