UAS DATAMODELLING

Anggota Kelompok 4

- Dava Virgio Kertawijaya - 00000056848
- Nigel Andrian - 00000055946
- Michael Owen Kohar - 00000056755
- Samuel Andrew - 00000056975

Abstract

The stock market has become a popular investment channel in recent years due to low returns on other investment options. Stock price prediction is of interest to private investors and institutional investors. Accurate stock price forecasting is an intriguing yet challenging task in the business world. The forecasting techniques used in the literature can be classified into two categories: linear models and nonlinear models. One of the forecasting techniques in the nonlinear model is Support Vector Regression (SVR). Essentially, SVR adopts the principle of minimizing structural risk to estimate a function by minimizing the upper bound of the generalization error. The optimal parameters for SVR modeling can be determined by finding the optimal values of the parameters C, gamma, and epsilon. In this study, the SVR model used the RBF kernel function. The results showed accurate predictions of Tesla's stock price with an accuracy of 99.94% for the training data and 99.96% for the testing data. The conclusion of this research is that the SVR model with the RBF kernel function and optimized parameters can provide accurate stock price predictions.

Introduction

The capital market is a structured financial system that involves the participation of numerous commercial banks, financial intermediaries, and financial instruments. One of the advantages of the capital market is that it provides individuals with opportunities to invest [1]. One example of an investment instrument is stocks, which are securities issued by companies. The income received by stockholders depends on the performance of the issuing company. If the company generates significant profits, stockholders will also receive substantial returns. However, as the potential returns increase, so does the accompanying risk [2]. Therefore, predicting stock prices based on previous prices is crucial.

Problem Formulation and Scope

1) The focus of this study is on predicting Tesla's stock prices using the Support Vector Regression (SVR) algorithm.
2) Only historical data of Tesla's stock prices from previous years will be used to forecast future stock prices.
3) External factors such as news or market events that may impact Tesla's stock prices will not be considered.
4) The study will only consider internal factors such as previous stock prices and trading volume as variables in the SVR algorithm.
5) The Mean Squared Error (MSE) evaluation metric will be used to measure the accuracy of the Tesla stock price predictions.

Research Objectives and Benefits

Considering the importance of stock price prediction in the capital market, especially in the rapidly growing electric vehicle industry, this research aims to predict the trend of Tesla's stock prices using the Support Vector Regression (SVR) algorithm. The research aims to provide accurate predictions of future Tesla stock prices by developing a prediction model using the SVR algorithm. This will result in more precise forecasts of the trend in Tesla stock prices. The findings of the research will benefit investors and market participants by enabling them to make more informed and rational investment decisions regarding Tesla stocks

Literature Study

Support Vector Regression (SVR)

Support Vector Regression (SVR) is a machine learning technique that is part of the Support Vector Machine (SVM) family and was introduced by Vapnik in 1995. In SVM, a linear function is typically used, which operates in a highdimensional feature space and is trained using optimizationbased learning algorithms that apply the concept of learning bias. SVM utilizes an epsilon-insensitive loss function. In this context, SVM can be utilized as an approach known as SVR [3]. The concept of SVR is based on risk minimization, aiming to minimize the upper bound of the generalization error to estimate a function.

Support Vector Regression (SVR)

The regression equation in SVR can be expressed as follows:

$$f(x) = w^\top \varphi(x) + b$$

Support Vector Regression (SVR)

The following is the loss function that needs to be minimized:

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

Support Vector Regression (SVR)

The following is the loss function that needs to be minimized:

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

Where

$$L_\varepsilon(y_i, f(x_i)) = \left\{ \begin{pmatrix} 0; |y_i - f(x_i)| \leq \varepsilon \\ |y_i - f(x_i)| - \varepsilon; otherwise \end{pmatrix} \right\}$$

Support Vector Regression (SVR)

The following is the loss function that needs to be minimized:

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

Where

$$L_\varepsilon(y_i, f(x_i)) = \left\{ \begin{pmatrix} 0; |y_i - f(x_i)| \leq \varepsilon \\ |y_i - f(x_i)| - \varepsilon; otherwise \end{pmatrix} \right\}$$

Support Vector Regression (SVR)

The underlying concept of the quadratic loss function involves minimizing the following value:

$$R(w, \xi, \xi^{\star}) = \frac{1}{2} \parallel w \parallel^2 + C \left(\sum_{i=1}^{n} (\xi_i + \xi_i^{\star}) \right)$$

Support Vector Regression (SVR)

subject to wφ(Xi) +b-yi ≤ ε+ξ ⋆ i ; yi -wφ(xi)-b ≤ ε+ξi and ξi , ξ⋆ i ≥ 0. By using the Lagrange equation in the form of:

$$
\begin{aligned}
L(w, b, \xi, \xi^\star, \alpha_i, \alpha_i^\star, \beta_i, \beta_i^\star) = \frac{1}{2} \parallel w \parallel^2 + C \left(\sum_{i=1}^{n} (\xi_i + \xi_i^\star) \right) \\
- \sum_{i=1}^{n} \alpha_i \left[w\phi(x_i) + b - y_i + \varepsilon + \xi_i \right] \\
- \sum_{i=1}^{n} \alpha_i^\star \left[y_i - w\varphi(x_i) - b_i + \varepsilon + \xi_i^\star \right] \\
- \sum_{i=1}^{n} (\beta_i \xi_i + \beta_i^\star \xi_i^\star)
\end{aligned}
$$

Support Vector Regression (SVR)

By using the Karush-Kuhn-Tucker method, the obtained result is as follows:

$$Q(\alpha, \alpha^{\star}) = -\frac{1}{2} \sum_{i,j=1}^{l} (\alpha_i - \alpha_i^{\star})(\alpha_j - \alpha_j^{\star}) K(x_i, x_j) - \varepsilon \sum_{i=1}^{l} (\alpha_i - \alpha_i^{\star}) + \sum_{i=1}^{l} y_i (\alpha_i - \alpha_i^{\star})$$

Kernel Function

Many methods in data mining often rely on linear functions, but in many real-world cases, data exhibits nonlinear characteristics. To address this, data transformation into a higherdimensional space is required. SVM can handle nonlinear data by utilizing a kernel approach, which allows for linear separation of data in a new feature space. In the SVR method, the kernel function used can be chosen as follows:

Kernel Function

1) Linier: $\mathbf{x}^\top \mathbf{x}$
2) Polinomial: $\left(\mathbf{x}^\top \mathbf{x} + 1\right)^n$
3) Radial Basis Function (RBF): $\exp\left(-\frac{1}{2\sigma^2} \parallel x - x_i \parallel^2\right)$

Related Works

Time series prediction can be divided into two types: linear prediction and non-linear prediction. For linear prediction, one example is the AutoRegressive Integrated Moving Average (ARIMA) method. On the other hand, non-linear prediction encompasses various techniques, including the Adaptive Neuro Fuzzy Inference System (ANFIS), Fuzzy, Neural Network, and Support Vector Machine (SVM). The SVM algorithm is a machine learning-based approach that can be used for regression tasks. On the other hand, the Support Vector Regression (SVR) algorithm is a variation of SVM specifically designed for regression tasks. One advantage of SVR is its ability to handle overfitting issues.

METHODOLOGY

Business Understanding

The research problem focuses on predicting the trend of Tesla's stock prices using the Support Vector Regression (SVR) algorithm. The objective is to develop a prediction model that utilizes SVR to provide accurate forecasts of future Tesla stock prices. The stakeholders, including investors and market participants, have a requirement for precise and reliable stock price predictions. By meeting this requirement, the research aims to assist investors and market participants in making informed investment decisions and adjusting their trading strategies based on the predicted trends. The accurate predictions generated by the SVR model contribute to enhancing the decision-making process and supporting the needs of stakeholders in the investment and financial industry.

Data Understanding

The initial step in the research process involves gathering historical data of Tesla's stock prices from previous years. This data serves as the foundation for the subsequent analysis. Once collected, the data is thoroughly explored and analyzed to gain insights into its structure, quality, and characteristics. This analysis helps in understanding the patterns and trends within the data. From the analysis, relevant variables and features are identified, with a focus on factors such as previous stock prices and trading volume. These variables play a crucial role in the SVR model, as they provide valuable information for predicting the trend of Tesla's stock prices.

Data Preparation

After gathering the historical data of Tesla’s stock prices, the next step involves cleaning the data by addressing missing values, outliers, and inconsistencies. This process ensures that the data is reliable and accurate for further analysis. Additionally, the data is transformed and preprocessed to meet the requirements of the SVR algorithm. This may involve feature scaling, normalization, or encoding categorical variables, depending on the specific needs of the model. Once the data is prepared, it is divided into training and testing sets. The training set is used to train the SVR model, while the testing set is utilized to evaluate the model’s performance and assess its predictive capabilities.

Modeling

The next phase involves applying the Support Vector Regression (SVR) algorithm with the RBF kernel function to develop a prediction model for Tesla's stock prices. To optimize the SVR model, parameter tuning techniques are employed to find the optimal values of parameters C, gamma, and epsilon. This ensures that the model is fine-tuned for improved accuracy and performance. The SVR model is then trained using the training data, allowing it to learn the underlying patterns and relationships between the input variables (such as previous stock prices and trading volume) and the target variable (Tesla's stock prices). By training the SVR model, it becomes equipped to make predictions based on the learned patterns and relationships in the data.

Evaluation

After training the SVR model, its performance needs to be assessed. This is done by evaluating its predictions against the actual stock prices. The Mean Squared Error (MSE) is used as the evaluation metric to measure the accuracy of the Tesla stock price predictions. By calculating the MSE, the model's ability to minimize the difference between predicted and actual prices can be quantified. It is important to compare the model's performance on both the training and testing datasets to ensure it generalizes well. If the model performs well on the training data but poorly on the testing data, it may indicate overfitting. Conversely, consistent performance across both datasets indicates that the model has learned the underlying patterns and can make accurate predictions on new, unseen data.

Deployment

Once the SVR model has been trained and evaluated, it can be deployed to make future predictions of Tesla's stock prices. These predictions can provide valuable insights for stakeholders, including investors and market participants. The findings and insights can be effectively communicated to stakeholders through reports or presentations, allowing them to stay informed about the predicted trends in Tesla's stock prices. With accurate stock price predictions at their disposal, stakeholders can make more informed investment decisions. Based on these predictions, recommendations can be provided to guide investment strategies and decisions, enabling stakeholders to seize potential opportunities and manage risks effectively. Throughout the CRISP-DM process, iterative cycles may occur, where steps are revisited and refined based on the results and insights gained. This ensures a systematic and iterative approach to the research process, leading to reliable and accurate predictions of Tesla's stock prices.

Data Source

The data used in this study is secondary data consisting of daily stock prices of Tesla from June 29, 2010, to March 17, 2017. This dataset consists of 1692 daily stock price data points. The data will be divided into two parts, namely training data and testing data, with an 8:2 ratio. In this research, it is assumed that today's stock price is only influenced by previous stock prices.

Research Steps

Research Steps

The data used in this study was obtained from an open-source data provider that provides information on Tesla stocks. The data used consists of daily Tesla stock prices, which include two variables: the independent variable x and the dependent variable y. Both variables have been prepared for further analysis.

Research Steps

A series of data preprocessing steps are conducted to ensure that the data is ready for analysis. The preprocessing steps include data cleaning, splitting the data into training and testing data with an 8:2 ratio for SVR model formation, feature normalization, and feature selection. In this study, the ”Open,” ”High,” and ”Low” features are used as independent variables x, while the ”Close” feature is used as the dependent variable y.

Research Steps

SVR Model Formation: Building a stock price prediction model using the SVR method. Determining the best parameters for the model involves several steps, as explained below:

1) Selecting a suitable kernel function and determining the parameters (cost, gamma, and epsilon) to optimize the formation of hyperplanes on the training data.

2) Performing SVR modeling by testing all possible combinations of cost, gamma, and epsilon values with the RBF kernel function using the Python programming language, as follows:

a) Splitting the data into training and testing data with an 8:2 ratio.

b) Optimizing by using combinations of cost, gamma, and epsilon parameters for each hyperplane used in regression model formation.

c) Calculating the error for each regression model using evaluation matrices such as Mean Absolute Error (MAE), Mean Squared Error (MSE), and Root Mean Squared Error (RMSE) to determine the optimal hyperplane.

3) Using the best hyperplane generated from the regression model on the testing data.

4) Evaluating the regression model on the testing data using the coefficient of determination R2 .

RESULT AND ANALYSIS

Finding the Optimal Kernel Function and Best Parameters for the Hyperplane

In this study, we exclusively utilized the RBF (Radial Basis Function) kernel function for the SVR (Support Vector Regression) hyperplane. To determine the best parameters for the kernel function, we conducted experiments with a range of values to construct the hyperplane. The parameters we optimized included the value of C, the value of gamma, and the value of epsilon. We evaluated the performance of different hyperplanes using an error evaluation matrix to identify the combination of parameters that yielded the smallest error. Through this selection process, we determined that the optimal parameters for the hyperplane using the RBF kernel function were C = 10, gamma = 0.001, and epsilon = 0.001, as presented in Table I below.

Table

TABLE I
APPLICATION OF SVR MODEL ERROR TO FIND OPTIMAL PARAMETERS

Parameter hyperplane			MSE	RMSE	MAE	R-Square
C	Gamma	Epsilon				
0,1	0,1	0,1	9778.69	98.88	89.75	-0.04
0,1	0,01	0,1	7576.57	87.04	79.33	0.18
0,1	0,001	0,1	4835.13	69.53	62.25	0.48
0,1	0,0001	0,1	2548.39	50.48	44.61	0.72
0,1	0,1	0,01	9785.31	98.92	89.76	-0.04
0,1	0,01	0,01	7578.46	87.05	79.33	0.18
0,1	0,001	0,01	4831.36	69.50	62.24	0.48
0,1	0,0001	0,01	2549.42	50.49	44.61	0.72
0,1	0,1	0,001	9785.97	98.92	89.76	-0.04
0,1	0,01	0,001	7579.00	87.05	79.33	0.18
0,1	0,001	0,001	4831.36	69.50	62.24	0.48
0,1	0,0001	0,001	2549.42	50.49	44.61	0.72
0,1	0,1	0,0001	9786.04	98.92	89.77	-0.04
0,1	0,01	0,0001	7579.05	87.05	79.33	0.18
0,1	0,001	0,0001	4831.36	69.50	62.24	0.48
0,1	0,0001	0,0001	2549.42	50.49	44.61	0.72
1	0,1	0,1	3115.92	55.82	46.65	0.66
1	0,01	0,1	741.19	27.22	13.98	0.92
1	0,001	0,1	182.76	13.51	5.72	0.98
1	0,0001	0,1	3115.86	55.81	46.64	0.66
1	0,1	0,01	38.83	6.23	3.11	0.99
1	0,01	0,01	741.49	27.23	13.98	0.92
1	0,001	0,01	182.82	13.52	5.72	0.98
1	0,0001	0,01	38.82	6.23	3.10	0.99
1	0,1	0,001	3115.89	55.82	46.64	0.66
1	0,01	0,001	741.45	27.22	13.98	0.92
1	0,001	0,001	182.79	13.52	5.72	0.98
1	0,0001	0,001	38.90	6.23	3.10	0.99
1	0,1	0,0001	3115.90	55.82	46.64	0.66
1	0,01	0,0001	741.44	27.22	13.98	0.92
1	0,001	0,0001	182.79	13.52	5.72	0.98
1	0,0001	0,0001	38.90	6.23	3.11	0.99
10	0,1	0,1	593.55	24.36	10.93	0.93
10	0,01	0,1	38.68	6.21	2.60	0.99
10	0,001	0,1	3.44	1.85	1.10	0.99
10	0,0001	0,1	38.68	6.21	2.60	0.99
10	0,1	0,01	593.57	24.36	10.93	0.93
10	0,01	0,01	38.56	6.21	2.59	0.99
10	0,001	0,01	3.43	1.85	1.10	0.99
10	0,0001	0,01	4.51	2.12	1.32	0.99
10	0,1	0,001	593.58	24.36	10.93	0.93
10	0,01	0,001	38.58	6.21	2.59	0.99
10	0,001	0,001	3.42	1.85	1.10	0.99
10	0,0001	0,001	4.54	2.13	1.32	0.99
10	0,1	0,0001	593.58	24.36	10.93	0.93
10	0,01	0,0001	38.59	6.21	2.59	0.99
10	0,001	0,0001	38.68	6.21	2.60	0.99
10	0,0001	0,0001	4.54	2.13	1.32	0.99

SVR Model Validation

The evaluation was performed on the testing data using the best parameters obtained from the training data. This evaluation resulted in accuracy values that compare the accuracy of the training and testing models using the most optimal parameters as determined above. The comparison of SVR model accuracies can be seen in Table II.

TABLE II
ACCURACY OF SVR MODEL FOR TESLA STOCK PRICES

Data	Coefficient of Determination
Training	0,999436928173908
Testing	0,999632324589655

Visualitation

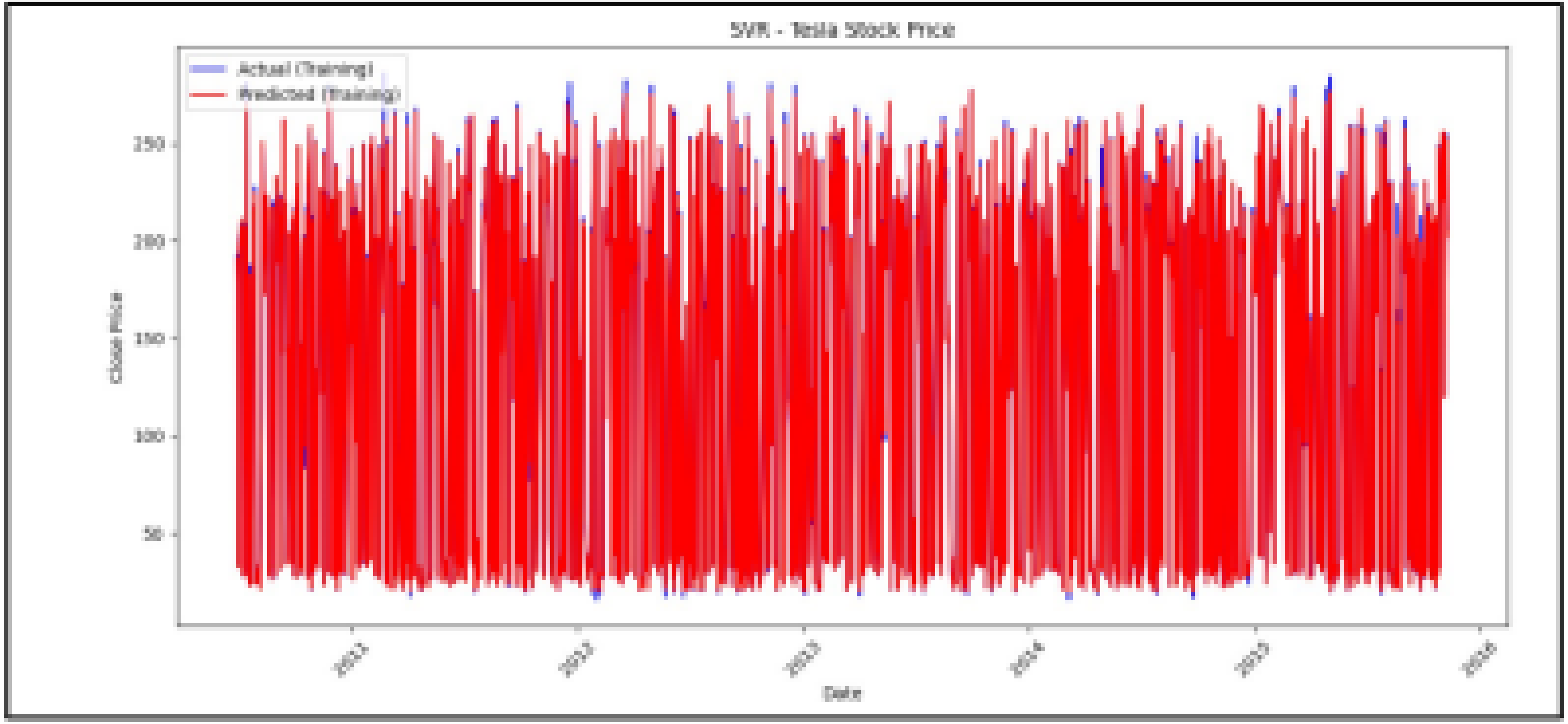

Fig. 1. Actual and Predicted Results Graph from Training Data

Visualitation

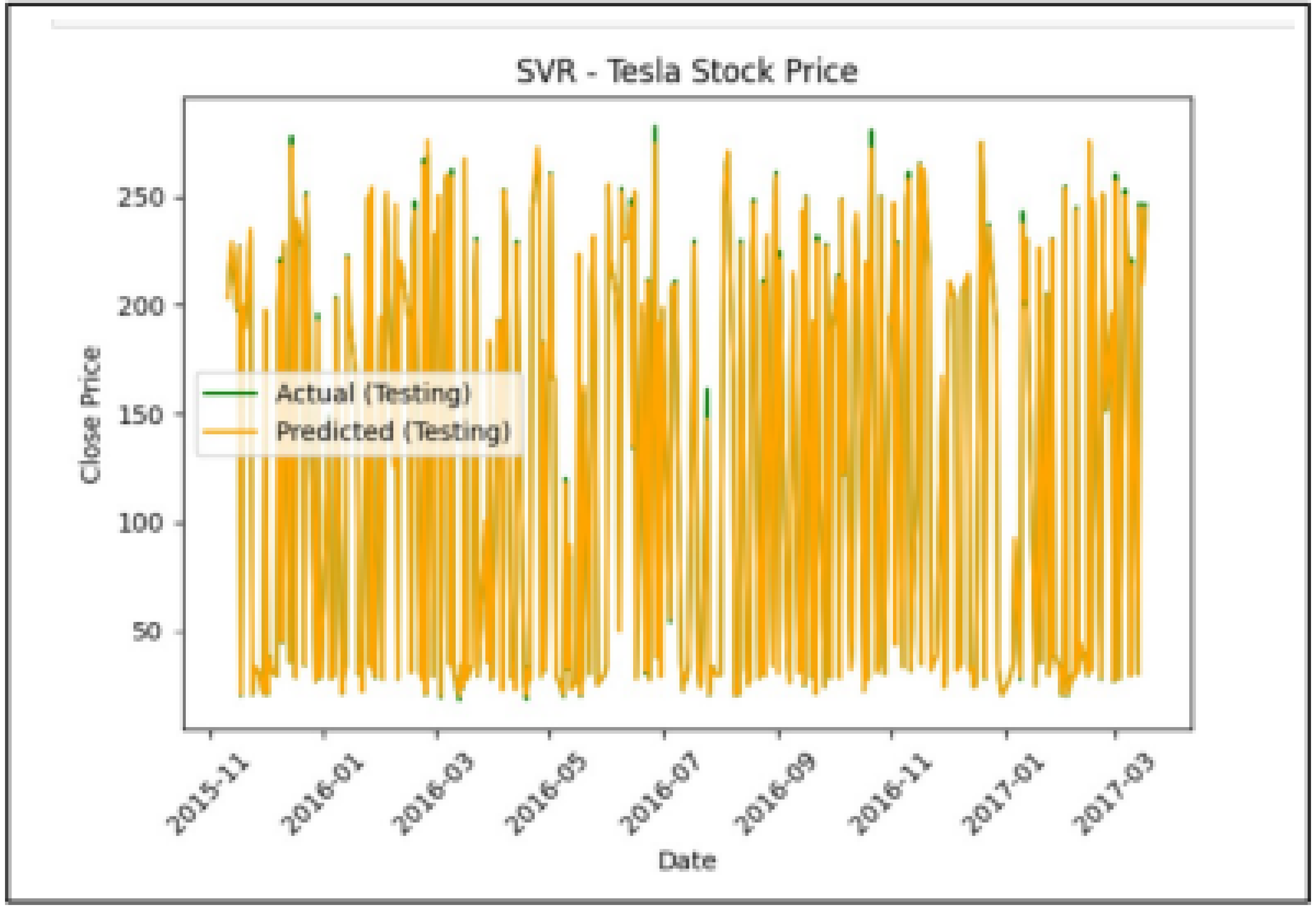

Fig. 2. Actual and Predicted Results Graph from Testing Data

CONCLUSION

Based on the analysis conducted, it can be concluded that the best SVR model uses the RBF kernel function with the parameters C = 10, gamma = 0.001, and epsilon = 0.001. This model achieved a high level of accuracy, with 99.94% for the training data and 99.96% for the testing data, making it reliable for predicting Tesla stock prices. Based on the conducted investigation, the following recommendations can be proposed for future research development and practical implications:

1) Explore alternative prediction methods beyond SVR, such as Neural Networks or Random Forests, and consider ensemble models for improved accuracy.

2) Incorporate relevant additional data, such as financial reports, industry news, and market sentiment, to enhance prediction quality.

CONCLUSION

3) Collaborate with practitioners and investors to gain valuable insights and practical implications for investment decision-making.
4) Expand the research scope to include other companies in the electric vehicle sector for comparative analysis. By implementing these recommendations, future research can contribute to the development of more accurate prediction methods, better understanding of stock price dynamics, and improved guidance for investment decisions.

REFERENCES

[1] P. Anoraga and P. Pakarti, Pengantar Pasar Modal Edisi Revisi, 5th ed. Rineka Cipta Jakarta, 2001.
[2] F. Pai, Ping and S. Lin, Chih, “A Hybrid ARIMA and Support Vector Machine Model in Stock Price Forecasting,” Omega, vol. 33, no. 6, pp. 497–505, 2005.
[3] S. R. Gunn, Support vector machines for classification and regression. University of Southampton, 1998.
[4] D. Bertsimas and R. Shioda, “Classification and Regression via Integer Optimization,” Operation Research, vol. 55, no. 2, 2007.
[5] C. Hong, Wei, “Rainfall Forecasting by Technological Learning Models,” International Journal of Applied Mathematics and Computation, vol. 200, pp. 41–57, 2008.

REFERENCES

[6] Y. Chen, Li, “Application of SVR with chaotic GASA algorithm to forecast Taiwanese 3G mobile phone demand,” Journal of Neurocomputing, vol. 127, pp. 206–213, 2014.
[7] K. Goyal, Manish, B. Bharti, J. Quilty, J. Adamowski, and A. Pandey, “Modeling of Daily Pan Evaporation in Sub Tropical Climates using ANN, LS-SVR, Fuzzy Logic, and ANFIS,” Journal international of Expert Sistem with Applications, vol. 41, pp. 5267–5276, 2014.
[8] K. Trapsilasiwi, Rani and Abdillah, “Peramalan Beban Listrik Jangka Pendek pada PT. PLN area JawaTimur – Bali Menggunakan Support Vector Machine,” Jurnal Statistika, vol. 3, no. 1, 2010.
[9] J. Kao, Ling, C. Chiu, Chih, J. Lu, Chi, and L. Yang, Jung, “Integration on Nonlinear Independent Component Analysis and Support Vector Regression for Stock Price Forecasting,” Neurocomputing, vol. 99, pp. 534–542, 2013.
[10] J. Wang, Ju, Z. Wang, Jian, G. Zhang, Zhe, and P. Guo, Shu, “Stock Index Forecasting Based on a Hybrid Model,” Omega, vol. 40, no. 6, pp. 758–766, 2012.
[11] A. Kazem, E. Sharifi, K. Hussain, Farookh, M. Saberi, and K. Hussain, Omar, “Support Vector Regression with Chaos Based Firefly Algorithm for Stock Market Price Forecasting,” Applied Soft Computing, vol. 13, no. 2, pp. 947–958, 2013.

Group Member Work Role

Dava Virgio Kertawijaya (00000056848) mengerjakan Bab 1 & 2 + abstrak + PPT
Nigel Andrian (00000055946) mengerjakan Bab 2 & 3 + daftar pustaka + PPT
Michael Owen Kohar (00000056755) mengerjakan Bab 3 & 4 + coding + PPT
Samuel Andrew (00000056975) mengerjakan Bab 4 & 5 + coding + PPT

Terima Kasih