International Journal of Application on Sciences, Technology and Engineering (IJASTE)
Volume 1, Issue 4, 2023. ISSN:2987-249

CRYPTOCURRENCY PRICE PREDICTION USING SUPPORT VECTOR REGRESSION

Thomas Stephen

Information System Study Programs, Faculty of Infomation Technology, University Tarumanagara
Email : Thomas33stephen@gmail.com

Submitted: 27-09-2023, Revised: 27-10-2023, Accepted: 11-12-2023

ABSTRACT

The rise of cryptocurrencies in the wake of the Industrial Revolution 4.0 has changed the economic landscape, providing an innovative alternative to conventional currencies. These digital currencies based on blockchain technology offer unparalleled flexibility, transparency, speed and transaction costs. However, the volatile nature of cryptocurrency prices poses challenges, especially for novice investors. This research explores the application of Support Vector Regression (SVR) models, specifically Polynomial Kernel SVR, to predict cryptocurrency prices. Using real-time data from Yahoo Finance for popular cryptocurrencies such as Bitcoin, Ethereum, Binance Coin, Chainlink, XRP, Cardano, and Dogecoin, this study carefully evaluates various SVR scenarios. The results show that the Polynomial Kernel SVR method, with optimized parameters, achieves an average accuracy of 44.92% as measured by R2 Square and an average error of 11.3% as measured by RMSE (Root Mean Square Error).

1. INTRODUCTION

The Industrial Revolution 4.0 brings the development of the world of technology to a much more advanced direction and presents various innovations in almost all aspects of human life including in economic activities. One of the innovations that have emerged in the economic field is the presence of Cryptocurrency as an alternative to conventional currencies.

Cryptocurrency is a digital currency that is created from a series of codes or called blockchain. Cryptocurrency can be used as a means of payment where transactions are made virtually or via the internet. Cryptocurrencies are considered to have advantages over conventional currencies, including flexibility that can be used anywhere, transparency, speed, and low transaction costs. Unlike the widely known currencies, these currencies are intangible, and are not issued by a country or central bank so they are not under government control.

The growing interest in investing in cryptocurrencies makes prices more volatile and makes it more difficult for someone to do technical or fundamental analysis to predict the next price. This price movement makes it difficult for novice investors/traders to read the direction of price movements, because at any time the price can change considerably due to the psychology of investors/traders as a whole wanting to buy or sell the coin. Seeing the opportunity that cryptocurrencies will be used in general for transactions in the future and with market conditions like this, not a few of the investors suffered losses because they predicted the wrong movement. By using Machine Learning as one of the solutions in minimizing losses and maximizing profits, it is hoped that the machine will be able to predict coin prices within a certain period of time with good accuracy. Artificial Intelligence (AI) makes human problems easier to solve and can predict future problems. In recent years, there have been many models used to make predictions, one of which is the Support Vector Regression (SVR) model which is a model with a statistical approach. The Support Vector Regression (SVR) method is a forecasting method that can be used to predict time series data. SVR is a modification of the Support Vector Machine (SVM) used for regression problems. The advantage of SVR is the ability to overcome non-linear data problems with kernel tricks. In addition, this method is also able to find the function $f(x)$ as a hyperplane for all input data that has the largest deviation from the actual target training data and can make the error as thin as possible.

Support Vector Regression (SVR) polynomial possesses the ability to handle non-linear relationships between dependent and independent variables. Its strength lies in the flexibility to use a polynomial kernel function, enabling the model to adapt to complex patterns. Furthermore, SVR polynomial tends to be more robust against the influence of outliers due to its focus on support vectors less affected by outlier data. However, the use of SVR polynomial requires careful parameter selection, including the polynomial degree, and may involve computationally intensive processes, especially at higher polynomial degrees. The resulting model can also be complex and challenging to interpret. When compared to conventional polynomial regression, SVR polynomial offers advantages in seeking a globally optimal solution and provides greater flexibility in regulatory control through adjustable parameters.

2. METHOD

The data used in this study are open, highest, lowest, and close prices of Bitcoin coins in dollars with the BTC/USDT code, the data is obtained from the Yahoo Finance. The amount of data taken is as much as 1462 data in the form of each opening price data, highest, lowest, and closing price data per day. After the data collected will be divided into training data and testing data then model design for each method, and the last is to evaluate the method.

2. 1. Support Vector Regression Polynomial

Support Vector Regression (SVR) was first introduced by Vapnik in 1995. Support Vector Regression (SVR) is a development method of Support Vector Machine (SVM) used in regression problems. SVR is included in the supervised learning algorithm used to predict the value of continuous variables. Just like the SVM concept, the SVR method also finds the best hyperplane in the form of a regression function by making the error as small as possible by maximizing the margin. SVR aims to find a function f(x) as a hyperplane in the form of a regression function which fits all input data by making the smallest possible error .

The concept of SVM can be explained simply as an attempt to find the best hyperplane that functions as a separator of two classes in the input space. Figure 1 shows some data that is a member of two classes. Class -1 is symbolized in red while class +1 is yellow. On the left of figure 1 shows some alternatives to the dividing line.

The best hyperplane can be found by measuring the margin of the hyperplane and finding the maximum point of the margin. Margin is the distance between the hyperplane and the closest data from each class. The data closest to the margin is called a support vector. In Figure 1 on the right, a solid line showing the best hyperplane is located right in the middle of the two classes. Red and yellow dots that are in the black circle are support vectors.

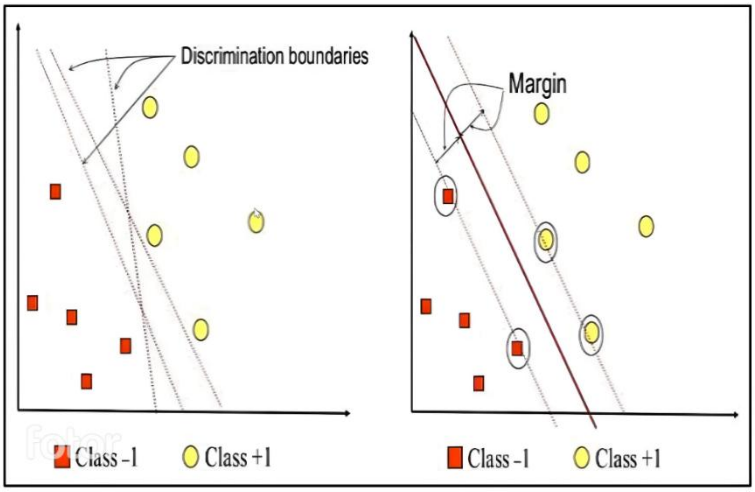

Figure 1. Hyperplane

Many data mining or machine learning techniques were developed with the assumption of linearity, so the resulting algorithm is limited for linear cases. With the kernel method, an x data in the input space is mapped to feature space with higher dimensions through.

$$\varphi: x \rightarrow \varphi(x) \tag{1}$$

where x are separate data points,φ is what would map our data onto higher dimensional space. Polynomial kernel is one of the kernel functions used in SVR, along with linear and radial kernels. The polynomial kernel function is used to transform the input data into a higher-dimensional space, making it possible to find a hyperplane that can separate the data points.The polynomial kernel does involve taking the inner product from a higher dimension space. The polynomial kernel can be expressed as:

$$\mathrm{K}(x, y\ ) = (ax^T y + C)^d \tag{2}$$

where x and y are input data points, c is a constant, and d is the degree of the polynomial. The degree of the polynomial determines the complexity of the decision boundary.

2. 2. Evaluation Metrics

Therefore, it is necessary to evaluate several types of evaluation metrics that exist to get a conclusion on which model is the best. In this paper, several evaluation metrics will be used, including the following:

- RMSE

  Root Mean Square Error (RMSE) is a commonly used metric to measure the accuracy of a regression model. It measures the difference between the predicted values and the actual values in the same unit. RMSE is calculated by taking the square root of the average of the squared differences between the predicted and actual values. A lower RMSE value indicates a better fit of the model to the data. In the study. The RMSE can be expressed as:

$$RMSE = \sqrt{\frac{\Sigma(y_t - y_{t+1})^2}{n}} \tag{3}$$

Information:
y_t = Actual value in period t
y_{t+1} = Average
n = Number of observation

- R2 Square

R2 evaluation, also known as the coefficient of determination, is a statistical measure that represents the proportion of the variance in the dependent variable that is predictable from the independent variable(s). It is a value between 0 and 1, where 0 indicates that the model does not explain any of the variability in the dependent variable, and 1 indicates that the model explains all of the variability in the dependent variable. The R2 Square can be expressed as:

$$R2 = 1 - \frac{SS\ Error}{SS\ Total} = 1 - \frac{\Sigma(y_i - y_i\prime)^2}{\Sigma(y_i - y\prime\prime)^2} \tag{4}$$

Information:

y_i = Observation of the i^{th} response
y'' = Average
y_i = Forecast of the i^{th} response
$SS\ Error$ = Variation values of the model residues
$SS\ Total$ = The value of total variation in data

2. 3. Research Data

In this study, the dataset is obtained from a web-based financial market platform used to provide real-time data, financial news, data and commentary including stock quotes, press releases, financial reports, as well as cryptocurrency data. The platform is called finance.yahoo.com. On this platform, the BTC to US Dollar dataset is taken from today's date to 4 years back. This dataset consists of 5 attributes, namely Open, Close, High, Low. The dataset has a total of 1462 data on each attribute. The flow of steps in retrieving the BTC to US Dollar exchange rate dataset is: The first step is to access the finance.yahoo.com platform. The second is to choose which dataset to take according to the research needs. The third is to get raw data that has not been processed. The file obtained is in CSV format.

2. 4. Proposed Method

This section describes the proposed method used in this paper and can be generally described in Figure 2. The proposed method is divided into 5 stages. The first stage is data collection. The second stage is data preprocessing, which is useful for preparing the data before it is applied to the algorithm. The third stage is to implement the data into the kernel model in SVR. The fourth stage is to evaluate the model using accuracy test and error test. The fifth stage is to display the performance results of the tested kernel.

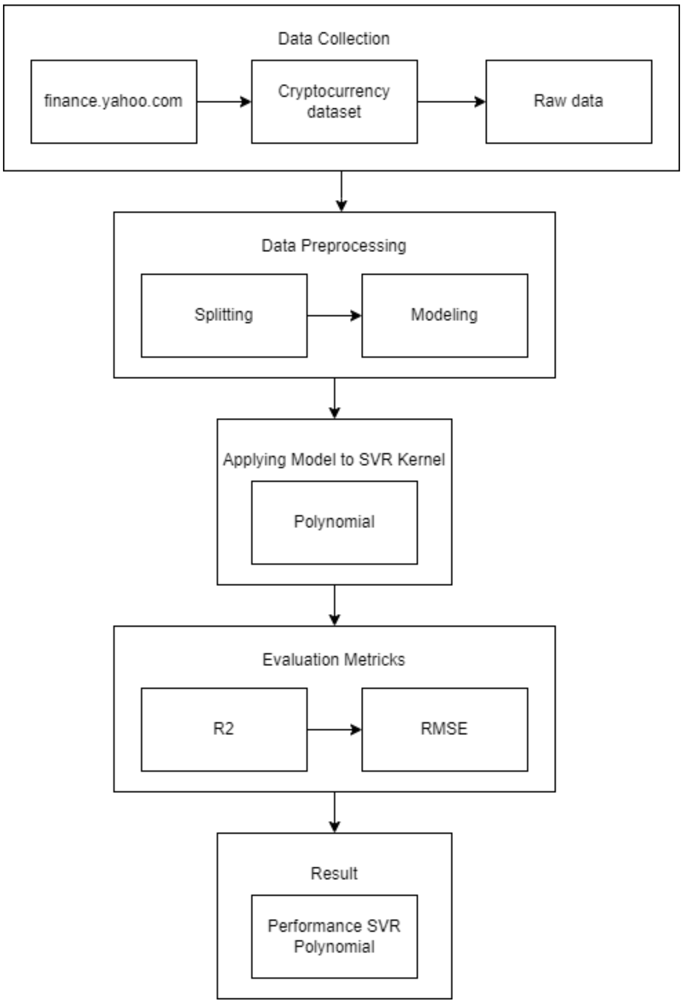

Figure 2. Research Stage.

Figure 2 shows the steps of this study which are divided into 5 main stages, namely:

1. The collected data on cryptocurrency against the US Dollar spans from the current real-time information up to four years back. For instance, considering Bitcoin data (Figure 3), the dataset is inherently dynamic, constantly refreshed as the system is utilized. This dynamism arises from the use of daily data attributes such as Open (the opening price of the day), High (the highest price of the day), Low (the lowest price of the day), and Close (the closing price on that day). This dynamic nature ensures that the information remains relevant and up-to-date, capturing the nuances of the market on a day-to-day basis. As the system operates, it systematically updates with the most recent actual data, incorporating it as a robust variable for predicting price movements. This immediate integration of new data postulates a stronger predictive capability, enhancing the system's adaptability to the ever-changing landscape of the cryptocurrency market.

Date	Open	High	Low	Close
2019-09-30	8104.226562	8314.231445	7830.758789	8293.868164
2019-10-01	8299.720703	8497.692383	8232.679688	8343.276367
2019-10-02	8344.212891	8393.041992	8227.695312	8393.041992
2019-10-03	8390.774414	8414.227539	8146.437012	8259.992188
2019-10-04	8259.494141	8260.055664	8151.236816	8205.939453
...	...	...	...	...
2023-09-26	26294.757812	26389.884766	26090.712891	26217.250000
2023-09-27	26209.498047	26817.841797	26111.464844	26352.716797
2023-09-28	26355.812500	27259.500000	26327.322266	27021.546875
2023-09-29	27024.841797	27225.937500	26721.763672	26911.720703
2023-09-30	26900.173828	26997.414062	26889.638672	26997.027344

[1462 rows x 4 columns]

Figure 3. BTC/USD price data.

2. Preprocessing data include:
   - Splitting Data. The dataset is divided into x and y data. Data x as the independent variable and data y as the dependent variable. This data splitting is done on each of the attributes such as Open, High, Low, and Close with a division of 80% training data and 20% test data. The allocation of 80% for training provides the model with sufficient data to learn patterns and trends in the cryptocurrency market. The remaining 20% for testing allows for a thorough evaluation of how well the model can predict on previously unseen data. Such a data split helps generate more consistent and valid estimates of the model's performance as it is tested on an independent dataset. Training and testing data be generally described in Figure 4 and Figure 5.

Training Data:

Date	Open	High	Low	Close
2019-10-01	8299.720703	8497.692383	8232.679688	8343.276367
2019-10-02	8344.212891	8393.041992	8227.695312	8393.041992
2019-10-03	8390.774414	8414.227539	8146.437012	8259.992188
2019-10-04	8259.494141	8260.055664	8151.236816	8205.939453
2019-10-05	8210.149414	8215.526367	8071.120605	8151.500488

Testing Data:

Date	Open	High	Low	Close
2022-12-13	17206.441406	17930.085938	17111.763672	17781.318359
2022-12-14	17782.066406	18318.531250	17739.513672	17815.650391
2022-12-15	17813.644531	17846.744141	17322.589844	17364.865234
2022-12-16	17364.546875	17505.525391	16584.701172	16647.484375
2022-12-17	16646.982422	16800.589844	16614.029297	16795.091797

Figure 4. Splitting training data and testing data.

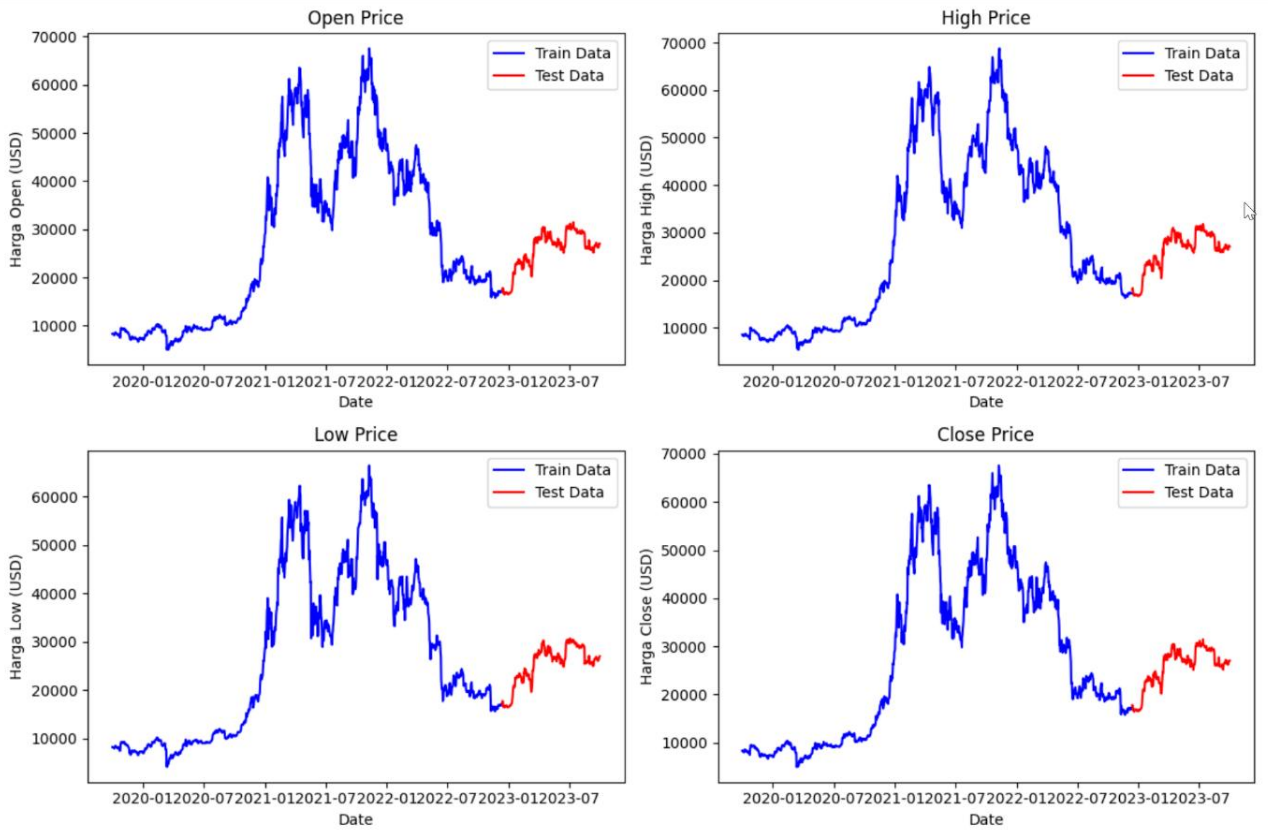

Figure 5. Splitting the training data and test data based on the oldest date.

- Data Modelling. Data for each attributes (Open, High, Low, Close) is prepared in the form of a numpy array. then train the SVR (Support Vector Regression) model with a polynomial kernel using the training data and test the model with the test data.

3. Implement the best model of each SVR kernel with test data.
4. Evaluate the metrics on the model by using R2 Square for accuracy values and RMSE for error values.
5. Showing the results of the performance, the accuracy and error values.

3. RESULTS AND DISCUSSION

In this section, the results and discussion regarding the application of the SVR method to the prediction process are explained. In addition, it also discusses the performance of the polynomial kernel based on the calculation of R2 Square for the accuracy value and RMSE (Root Mean Square Error) for the error value.

3. 1. SVR Polynomial Kernel

This study presents the computational results obtained from the Polynomial SVR model. The model is visualized using a line graph depicting the movement of the cryptocurrency value against the US Dollar over time. The horizontal axis represents time, while the vertical axis represents the exchange rate. The graph displays two lines: a red line representing the predicted data and a blue line representing the actual data. The closeness of these two lines indicates the accuracy of the prediction (Figure 6). The closer the lines are, the higher the accuracy. The model is derived from testing over 30 scenarios using SVR polynomial equations.

From the experiments of Polynomial kernel SVR, Trial and error was conducted with various scenarios to form the SVR equation for the Polynomial kernel. These scenarios amounted to more than 30 scenarios. The scenarios resulted in a model with Polynomial kernel parameters consisting of degree, epsilon and C. Degree is the degree of the Polynomial used to find the hyperplane to split the data. Epsilon determines the extent to which training data points can fall outside the prediction corridor without affecting the value of the loss function. The C parameter governs the trade-off between training error and model complexity by giving weight to larger training errors with higher values of C. In addition to measuring accuracy and error values, this study also measured the prediction time of each model. Prediction time is the length of time it takes for a model to predict a value. This prediction time is used to measure the performance of the model.

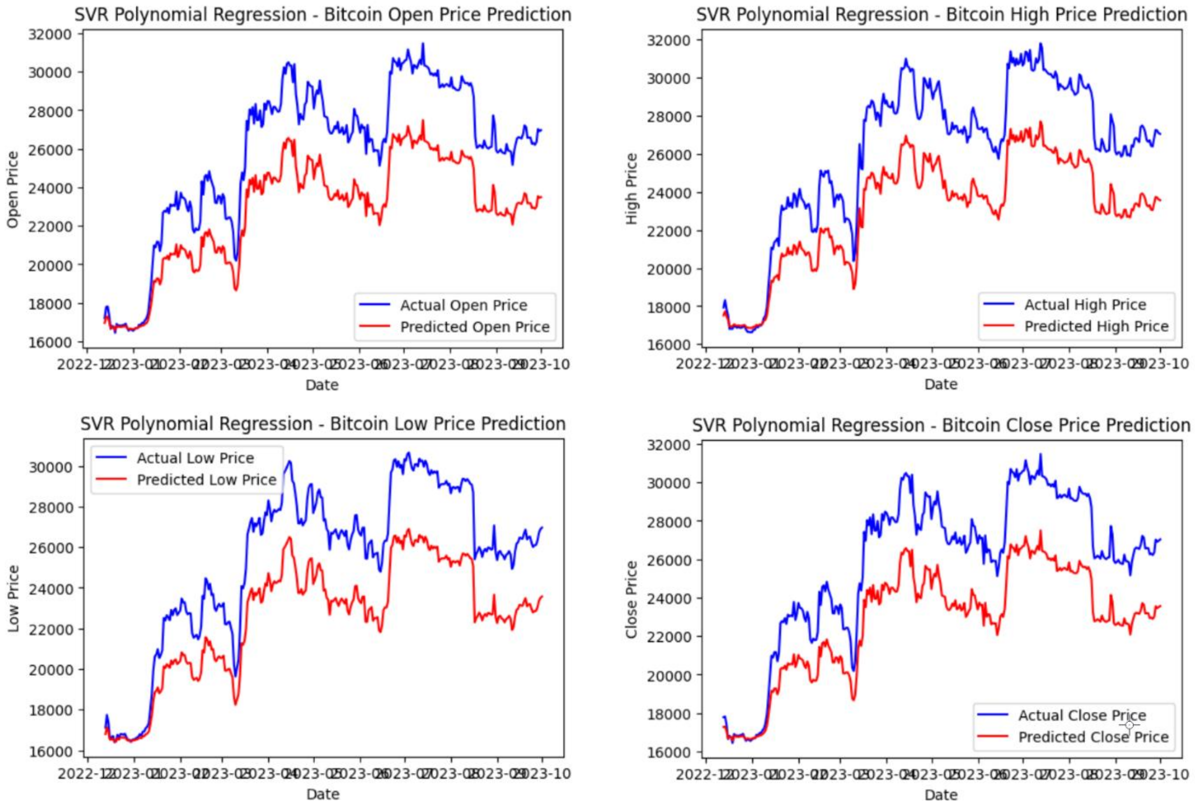

Figure 6. Actual data and predicted data in line chart (BTC/USD sample).

3. 1. Kernel Performance and Evaluation Metrics

This study also calculates each performance generated from polynomial SVR. This calculation is done from the 7 best scenarios. In this study, more than 30 scenarios were made when testing the value of accuracy, error, and processing time. This scenario consists of a combination of each Polynomial parameter including the Degree, C and Epsilon parameters. The evaluation is measured based on the accuracy value using R2 Square and error value using RMSE (Root Mean Square Error). Experimental results in the form of performance and evaluation can be seen in Table 1.

No	Coin Name	Attribute (price)	Degree	C	Epsilon	R2	RMSE	Prediction Time
1	BTC/USD	Open	2	100	1	35.74%	12.41%	0.66 sec
		Highest	2	100	1	34.39%	12.50%	
		Lowest	2	100	1	39.10%	12.04%	
		Close	2	100	1	35.19%	12.36%	
2	ETH/USD	Open	2	10	0.01	35.74%	11.61%	0.50 sec
		Highest	2	10	0.01	34.39%	11.70%	
		Lowest	2	10	0.01	39.10%	11.28%	
		Close	2	10	0.01	35.19%	11.57%	
3	BNB/USD	Open	2	1	0.01	57.26%	9.32%	0.79 sec
		Highest	2	1	0.01	52.20%	9.83%	
		Lowest	2	1	0.01	57.98%	9.25%	
		Close	2	1	0.01	56.84%	9.29%	
4	LINK/USD	Open	2	0.1	0.01	26.22%	8.86%	0.87 sec
		Highest	2	0.1	0.01	14.53%	9.77%	
		Lowest	2	0.1	0.01	40.85%	7.79%	
		Close	2	0.1	0.01	27.19%	8.85%	
5	XRP/USD	Open	2	0.1	0.01	58.32%	13.56%	0.79 sec
		Highest	2	0.1	0.01	61.04%	13.02%	
		Lowest	2	0.1	0.01	59.78%	13.46%	
		Close	2	0.1	0.01	58.19%	13.55%	
6	ADA/USD	Open	2	0.1	0.01	42.80%	12.70%	1.12 sec
		Highest	2	0.1	0.01	42.81%	12.88%	
		Lowest	2	0.1	0.01	43.36%	12.57%	
		Close	2	0.1	0.01	43.41%	12.67%	
7	DOGE/USD	Open	2	0.1	0.01	57.26%	9.32%	0.73 sec
		Highest	2	0.1	0.01	57.98%	9.83%	
		Lowest	2	0.1	0.01	56.84%	9.25%	
		Close	2	0.1	0.01	58.19%	9.29%	

Table 1. Sample experiments for implementation SVR polynomial.

4. CONCLUSION

This study demonstrated the potential of Support Vector Regression (SVR) models, particularly the Polynomial Kernel SVR, in predicting cryptocurrency prices. While showing promise, the use of SVR models requires careful parameter selection and may involve computationally intensive processes. The research also highlighted the challenges posed by the volatility of cryptocurrency prices, especially for novice investors.

Based on the experimental results of the combination of the best degree, epsilon, and C variables, the accuracy evaluation with R2 Square tested on approximately 30 popular cryptocurrencies resulted in an average of 44.92, which is quite accurate. However, there is an error with an average of 11.3% using RMSE. The polynomial kernel takes an average time of 0.64 seconds to process the entire computation. These findings underscore the need for further refinement and optimization to enhance prediction accuracy and mitigate the challenges posed by cryptocurrency price volatility.

REFERENCES

[1] Anisa, D., Anggraini, T., & Tambunan, K. (2023). “Analisis Cryptocurrency Sebagai Alat Alternatif Berinvestasi Di Indonesia”. Owner, 7(3), 2674–2682. doi: 10.33395/owner.v7i3.1698.

[2] Ben Fraj, M. (2018). In Depth: Parameter tuning for SVC. Retrieved from https://medium.com/all-things-ai/in-depth-parameter-tuning-for-svc-758215394769

[3] Cheng, Q., Liu, X., & Zhu, X. (2019). “Cryptocurrency momentum effect: DFA and MF-DFA analysis”. Phys. A Stat. Mech. its Appl., 526(80), 120847. doi: 10.1016/j.physa.2019.04.083.

[4] Chicco, D., Warrens, M. J., & Jurman, G. (2021). “The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation”. PeerJ Comput. Sci., 7, 1–24. doi: 10.7717/PEERJ-CS.623.

[5] Elsa. (2023). “PENERAPAN METODE SUPPORT VECTOR REGRESSION (SVR) MENGGUNAKAN KERNEL LINEAR, POLINOMIAL, DAN RADIAL DENGAN GRID SEARCH OPTIMIZATION”.no. Mi, pp. 5–24.

[6] Huda, N., & Hambali, R. (2020). “Risiko dan Tingkat Keuntungan Investasi Cryptocurrency. J. Manaj. dan Bisnis Performa”. 17(1), 72–84. doi: 10.29313/performa.v17i1.7236.

[7] Iskandar, D., Afriani, Pratiwi, & Effendi, E. (2021). “Analisis Teknik Penerjemahan pada Abstrak Jurnal IJAI”. J. Humanit. Soc. Sci., 3(1), 9–22. doi: 10.36079/lamintang.jhass-0301.187.

[8] Lopez-Martin, C., Azzeh, M., Bou-Nassif, A., & Banitaan, S. (2019). “Upsilon-SVR Polynomial Kernel for Predicting the Defect Density in New Software Projects”. Proc. - 17th IEEE Int. Conf. Mach. Learn. Appl. ICMLA 2018, pp. 1377–1382. doi: 10.1109/ICMLA.2018.00224.

[9] Panessai, I. Y., Iskandar, D., Afriani, Pratiwi, & Effendi, E. (2021). “Analisis Teknik Penerjemahan pada Abstrak Jurnal “. Humanit. Soc. Sci., 3(1), 9–22. doi: 10.36079/lamintang.jhass-0301.187.

[10] Purnama, D. I., & Setianingsih, S. (2020). “Support vector regression (SVR) model for forecasting the number of passengers on domestic flights at Sultan Hasanudin airport Makassar”. J. Mat. Stat. dan Komputasi, vol. 16, no. 3, p. 391. doi: 10.20956/jmsk.v16i3.9176.

[11] Sidik, A. P. (2020). Diagnosis of Types of Diseases in Cassava Plant by Bayes Method. J. Inform., vol. 4, no. 2, p. 69. doi: 10.15575/join.v4i2.379.

[12] Siregar, A. M., Faisal, S., Widiharto, B., Informatika, T., Perjuangan, U. B., & Ronggowaluyo, J. (2022). “Model Prediksi Penderita Covid 19 Di Indonesia Menggunakan Metode Support Vector”. Konf. Nas. Penelit. dan Pengabdi., pp. 79–90.