ILKOM Jurnal Ilmiah Vol. 15, No. 2, August 2023, pp.390-397
Accredited 2nd by RISTEKBRIN No. 200/M/KPT/2020; E-ISSN 2548-7779 | P-ISSN 2087-1716

The Support Vector Regression Method Performance Analysis in Predicting National Staple Commodity Prices

Huzain Azis [a,1,*]; Purnawansyah [a,2]; Nirwana [a,3]; Felix Andika Dwiyanto [b,4]
[a] *Universitas Muslim Indonesia, Jln. Urip Sumiharjo No.km.5, Panaikang, Makassar, 90231, Indonesia*
[b] *AGH University of Science and Technology, al. Adama Mickiewicza 30, 30-059 Kraków, Polandia*
[1] *huzain.azis@umi.ac.id;* [2] *purnanwansyah@gmail.com;* [3] *nirwana.labfik@umi.ac.id;* [4] *dwiyanto@agh.edu.pl*
[*] *Corresponding Author*

Article history: Received April 17, 2023; Revised May 30, 2023; Accepted June 09, 2023; Available online August 16, 2023

Abstract

Support Vector Regression (SVR) is a supervised learning algorithm to predict continuous variable values. The basic goal of the SVR algorithm is to find the most suitable decision line. SVR has been successfully applied to several issues in time series prediction. In this research, SVR is used to predict the price of staple commodity, which are constantly changing in price at any time due to several factors making it difficult for the public to get groceries that are easy to reach. National staple commodity data consisting of 17 commodities, including shallots, honan garlic, kating garlic, medium rice, premium rice, red cayenne peppers, curly red chilies, red chili peppers, meat of broiler chicken, beef hamstrings, granulated sugar, imported soybeans, bulk cooking oil, premium packaged cooking oil, simple packaged cooking oil, broiler chicken eggs, and wheat flour. With a data set for the last 3 years, including from January 1, 2020, to December 31, 2022. There are 3 variables in the data set, namely commodity, date, and price. This research divides the entire dataset into 80% training and 20% testing data. The results of this research show that SVR using the RBF kernel produces good forecasting accuracy for all datasets with an average MSE training data of 6,005 while data testing is 6,062, MAD of training data is 6,730 while data testing is 6.6831, MAPE training data is 0.0148 while data testing is 0.0147, and RMSE training data is 7.772 while data testing is 7.746.

Keywords: Food Prices; Machine Learning; Predictions; Support Vector Regression.

Introduction

The stability of staple commodity prices indicates economic progress in Indonesia [1]. Maintaining a stable level of staple commodity prices in the community in the current and future periods is essential to pay attention to [2]. According to the Decree of the Minister of Industry and Trade Number 115/MPP/Kep/2/1998 dated 27 February 1998, the nine staple commodities are the basic needs of the community. The nine staple commodities are rice, sugar, cooking oil and butter, beef and chicken meat, chicken eggs, milk, shallots, garlic, LPG gas and kerosene, and salt [3]. The need for staple commodity is needed by the community [4]. The community's need for groceries will increase yearly, directly proportional to the increase in Indonesia's population [5].

The high demand for staple commodity can result in an increase in the prices of groceries in the market [6]. However, aside from the increase in population, the price of staple commodity can also be affected by climate and weather factors that are erratic. Thus, causing farmers to fail to harvest, as a result the distribution to the community is reduced [4]. Then infrastructure factors such as damaged roads can hamper distribution channels. Apart from that, it is undeniable that there are still irresponsible parties who want to make more profits, these naughty parties usually buy goods/products from farmers in large quantities, then stockpile and sell them when market supplies are running low at high prices [7].

The Support Vector Regression (SVR) algorithm is a tool that can predict staple commodity prices using hyperplanes in a limited spatial dimension, which can be used for regression to time series data [8]. SVR is an algorithm that can overcome overfitting, resulting in excellent performance [9]. Based on research using SVR, namely forecasting bread sales, RMSE values for sweet bread, cake bread, and white bread are 0.00176, 0.00019, and 0.00010 [10] and in coal forecasting, a MAPE value of 9.64% is obtained [11]. Based on this, the Support Vector Regression Algorithm is planned to predict the prices of staple commodity materials, which are expected to provide the best results.

Method

The staple commodity material data used in this study is data from the Ministry of Trade's Market Monitoring System and Basic Needs (https://ews.kemendag.go.id) which is managed by the Ministry of Trade of the Republic of Indonesia to monitor national staple commodity prices. In the process of predicting the price of basic necessities, data is collected in monthly periodic form starting from 1 January 2020 to 31 December 2022 as many as 18,368 datasets, consisting of 17 types of commodities including shallots, honan garlic, kating garlic, medium rice, premium rice, red cayenne peppers, curly red chilies, red chili peppers, meat of broiler chicken, beef hamstrings, granulated sugar, imported soybeans, bulk cooking oil, premium packaged cooking oil, simple packaged cooking oil, broiler chicken eggs, and wheat flour [12].

A. Exploratory Data Analysis

Exploratory Data Analysis (EDA) is an initial investigative test process that aims to identify patterns, find anomalies, test hypotheses and check assumptions. EDA is used in this research to identify the condition of the data set collected in statistical form and find outliers in 17 data frames or each commodity data type using price and date variables.

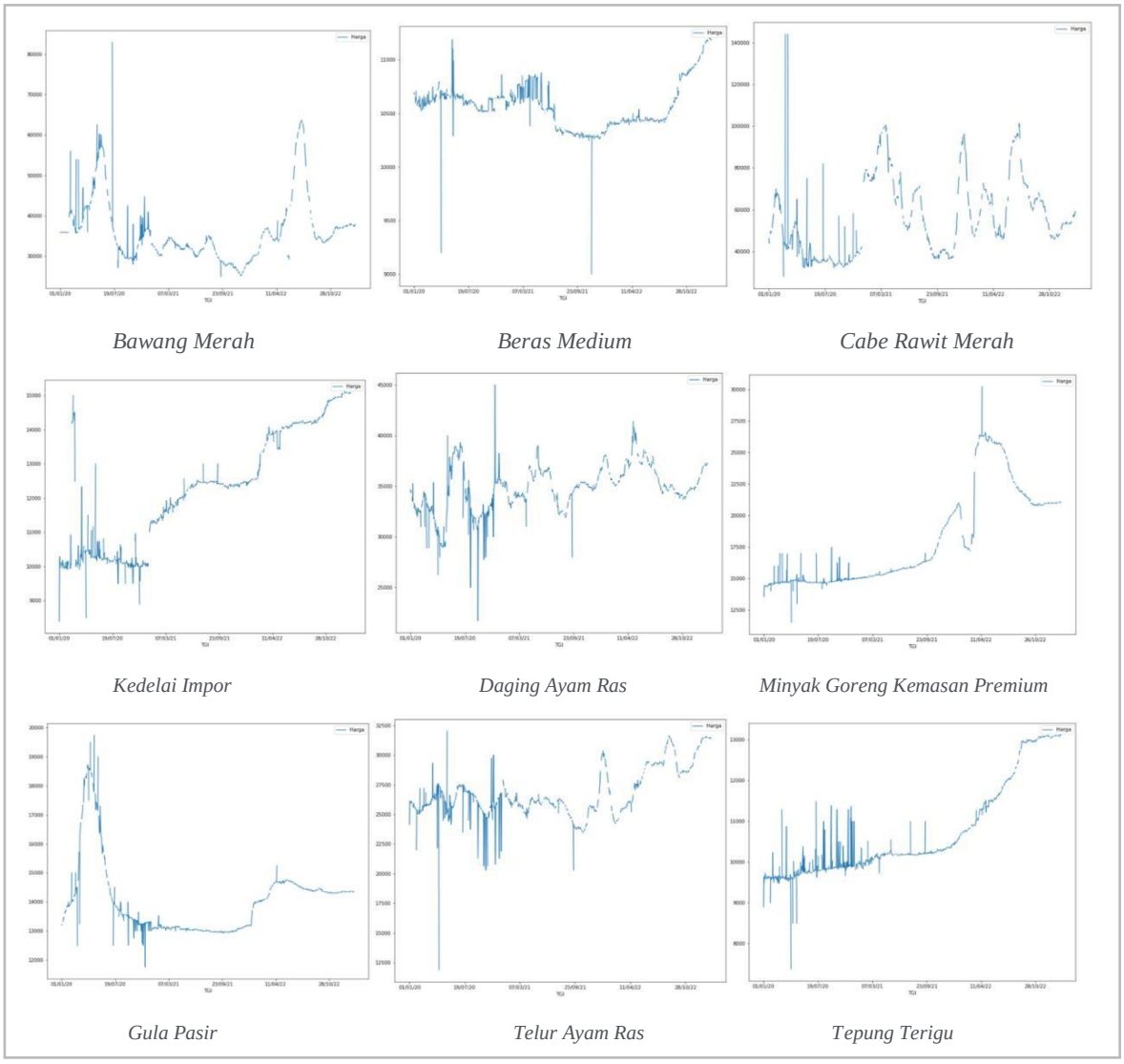

Figure 1. Exploratory Data Analysis Results

Figure 1 shows that some time series data have null or incomplete values. So it produces a graph by showing dotted or unconnected lines. This indicates that the processed data has a missing value on a certain date.

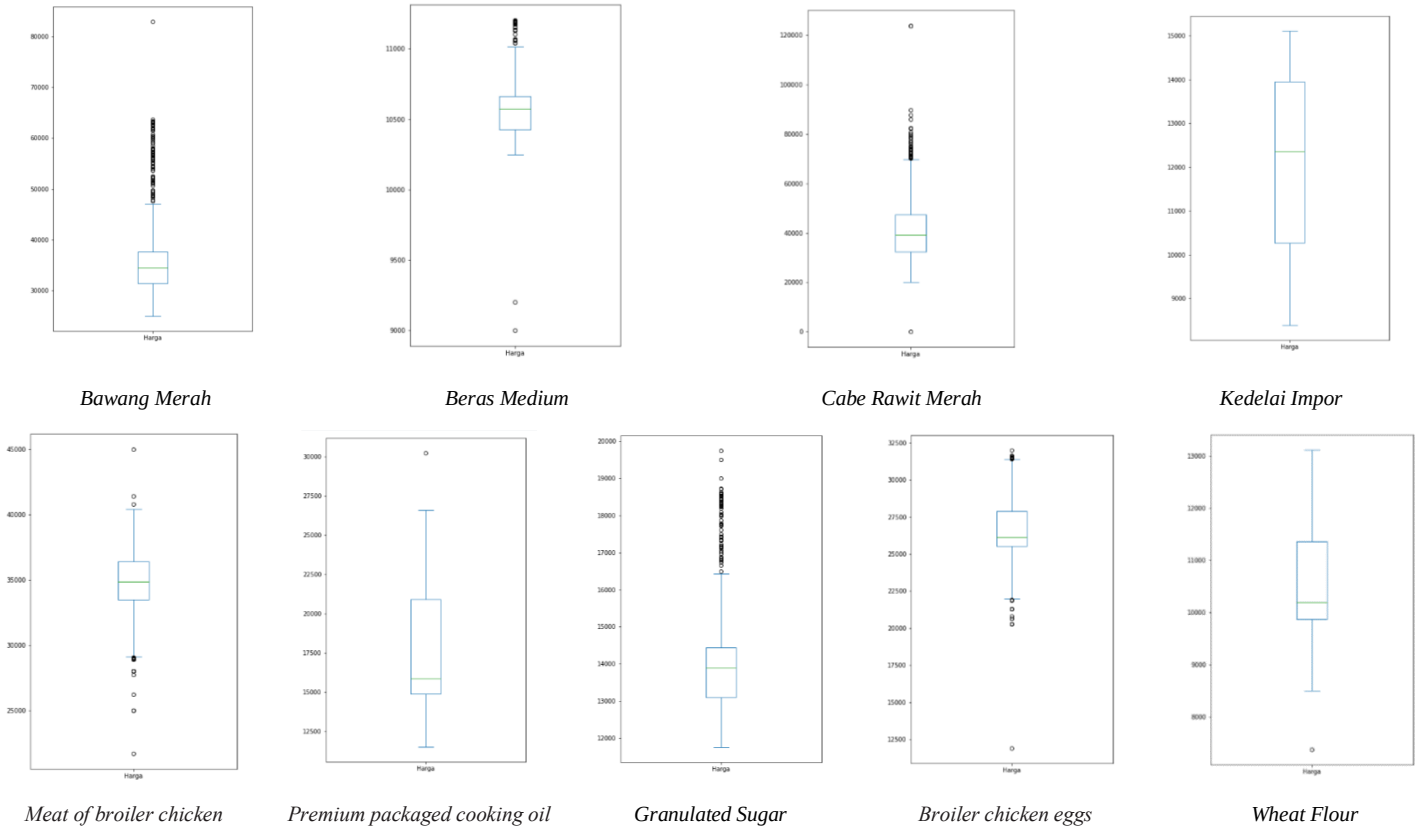

Figure 2. Box-Plot Graph Results

In **Figure 2**, the results of the box-plot graph indicate that the 9 sample data frames that were tested based on data on staple commodity prices experienced outliers or values that were significantly distant from the data value limit. This demonstrates that price data have changed significantly from the standard price data limits for each commodity.

B. Preprocessing Data

At the data preprocessing stage, there are 2 activities, namely Data Cleaning and Data Splitting. Data Cleaning is cleaning up missing values in data by completing empty data. There are 3 ways used to fill in the blank data, namely using, Interpolation, Forward fill, and backfill. Interpolation is a technique of filling in missing values by predicting the available data around it, Forward fill is a technique of filling in missing values using the final values available previously and backfill is a technique of filling in missing values using the leading values available afterward. Meanwhile, Data splitting is the process of dividing the dataset into two subsets, namely testing data and training data, with a ratio of 20% and 80% for the overall food commodity data [13].

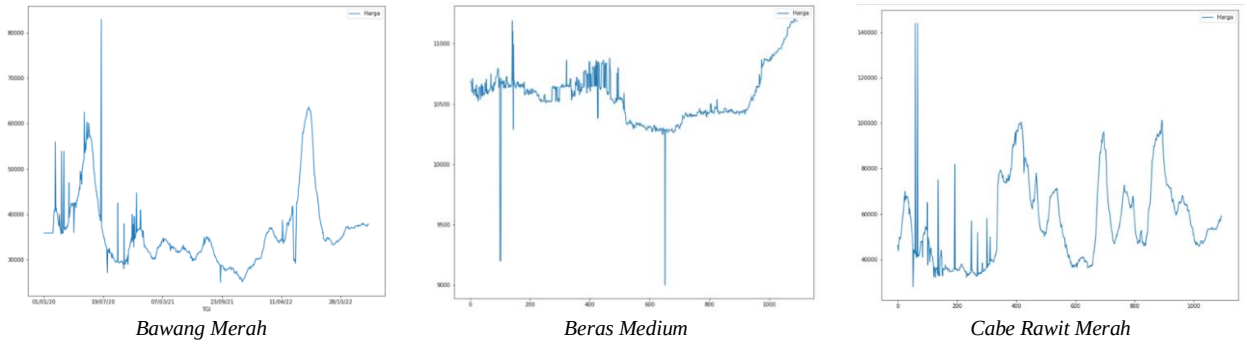

Figure 3. (a) Graph of the area after the data cleaning process

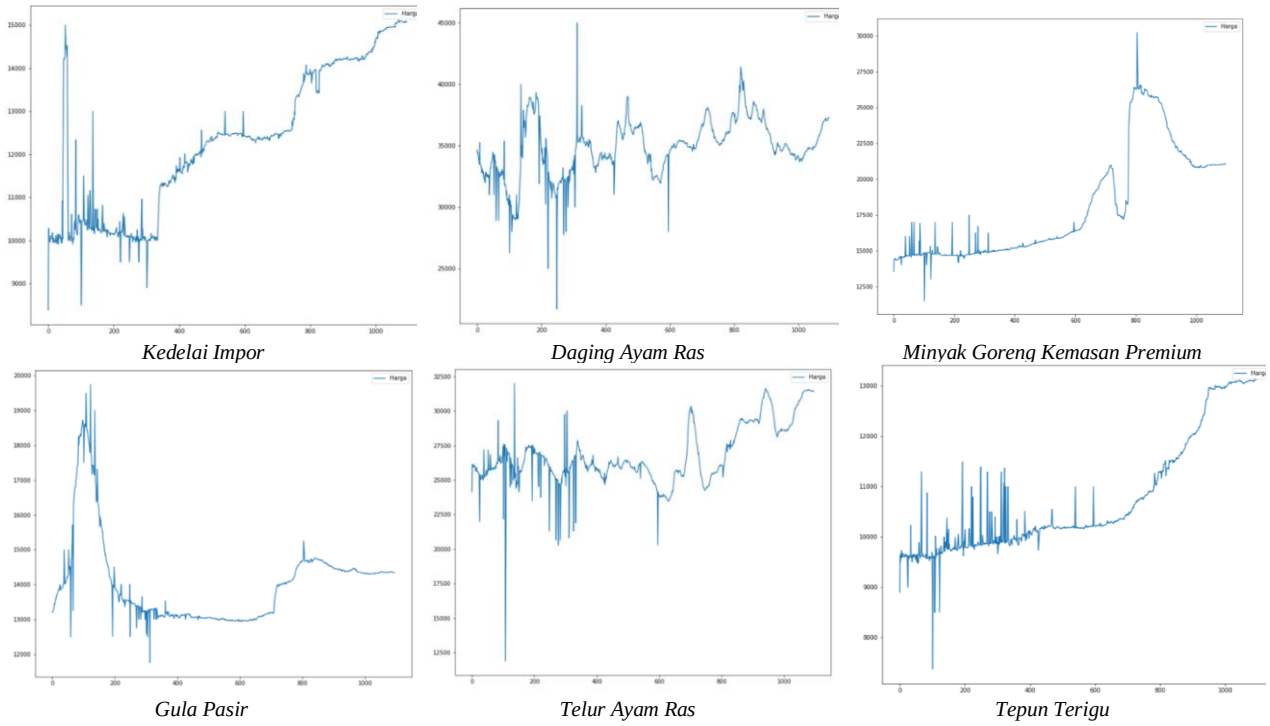

Figure 3. (b) Graph of the area after the data cleaning process

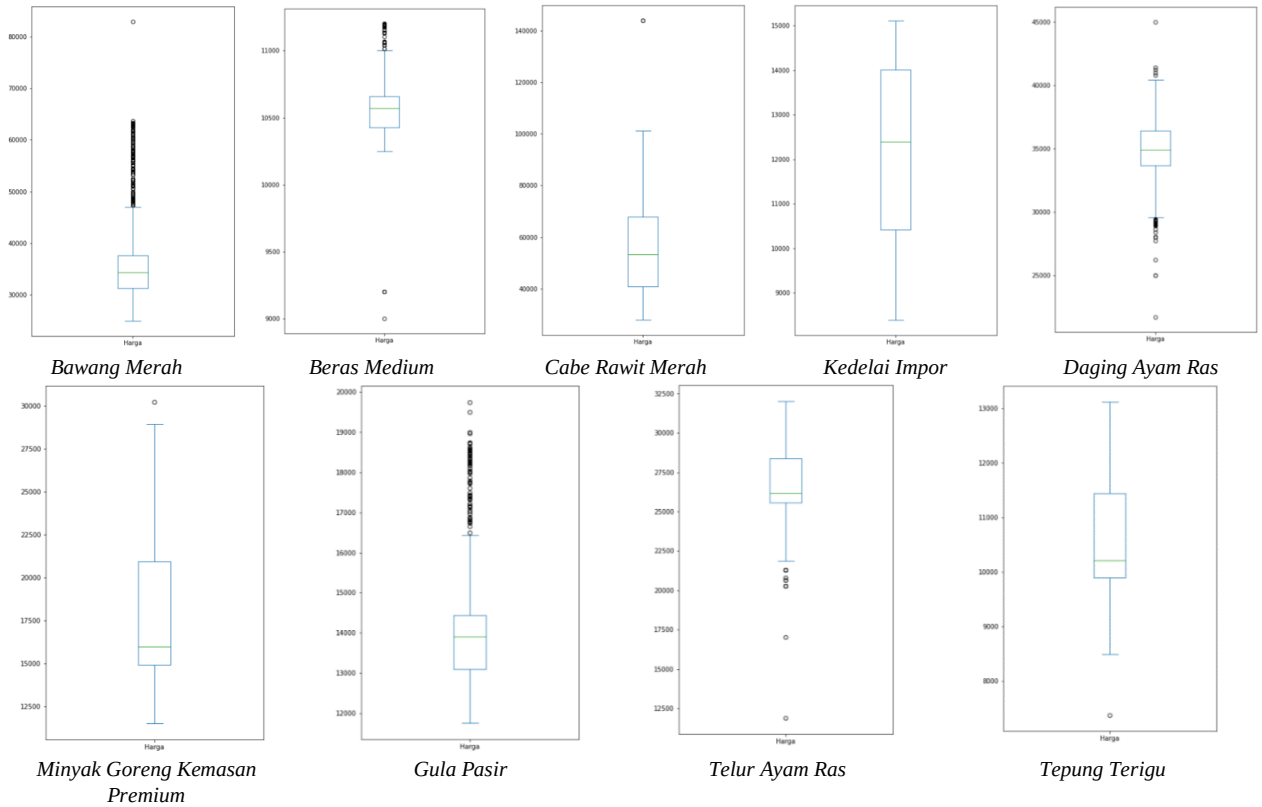

Figure 4. Box-plot graph after the data cleaning process

C. Support Vector Regression

SVR is a supervised learning algorithm to predict continuous variable values [14]. SVR uses the same principles as SVM. SVR has two objectives to be achieved at the same time. The first objective is to minimize prediction error by finding a solution with the most deviations ε (epsilon) from the observed reference value for each observation. The second goal of SVR is to minimize the value of w to prevent overfitting [15]. When working on data with high

dimensions, a small w value is a major concern because each dimension adds flexibility to the linear function, so high dimensions can increase the risk of overfitting [16].

With SVR, we want to find the function that has the largest deviation from the actual target, for all training data [8]. Support Vector is used in training data that lies within and outside the boundary f of the decision function, therefore the number of Support Vector decreases with increasing ε [14]. If the value is equal to 0 then a perfect regression equation is obtained [15]. The basic principle of SVR is linear regression which is stated in Equation 1 [6]. In Equation 1, x is the independent variable for a single sample, w is the regression vector, b is the offset, and e is the prediction error [17].

$$f(x) = w^T x + b + e \quad (1)$$

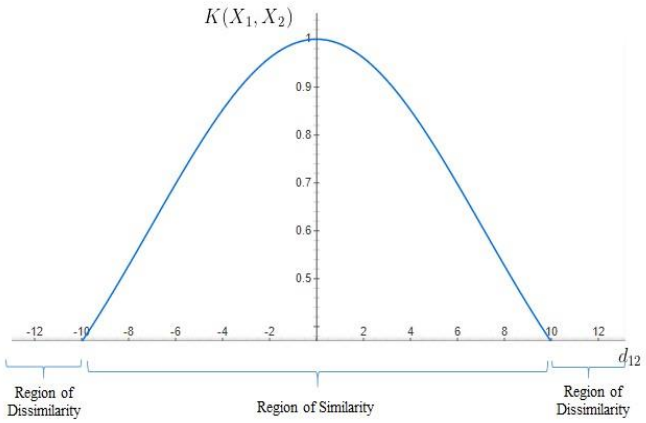

Figure 5. Illustration of SVR

In **Figure 5** is an SVR illustration where the hyperplane line is the dividing line between the two data classes in a higher dimension than the actual dimension and the boundary lines are two lines drawn around the hyperplane at a distance (epsilon). It is used to create margins between data points.

C. Radial Basis Function

The Radial Basis Function (RBF) kernel is the most frequently used kernel due to the selection of the correct σ value [16]. The idea of using the kernel function as an inner product in the feature space was introduced to machine learning techniques in 1964 with the potential function method [12]. RBF kernel can be used for various types of data [18]. The RBF kernel is defined into Equation 2.

$$\mathrm{K}(\mathrm{x}, \mathrm{x}') = \exp\left(-\frac{||X - X''||^2}{2\sigma^2}\right)^2 \quad (2)$$

In Equation 2, ||x-x"||2 is the Euclidean distance of the data in two different feature spaces, σ is an independent parameter in the RBF kernel that determines the kernel weight. The default value of σ is 1 [2].

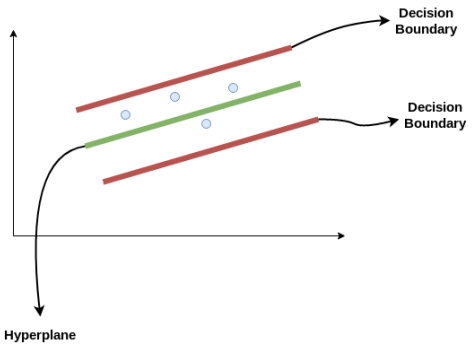

Figure 6. RBF Kernel Illustration

Figure 6 illustrates the RBF kernel, which has a value of σ = 10, which forms a large curve width, where points are considered similar for distances of up to 10 units, and beyond 10 units, they are different. This explains that using the RBF kernel provides an overview of time series data for proper regression modeling.

D. Model Evaluation

Model evaluation is carried out to calculate the resulting error value in the regression model that has been formed. In this study, four evaluation indices were used, namely Mean Square Error (MSE), Root Mean Square Error (RMSE), Mean Absolute Deviation (MAD), and Mean Absolute Percentage Error (MAPE) [6].

MSE is the average squared error between the actual value and the forecast value. The Mean Squared Error method is generally used to check the estimate of the error value in forecasting [19]. In Equation 3, A_t is the actual demand value, F_t is the forecasted value and *n* is the amount of data [18].

$$\text{MSE} = \frac{\sum_t^n = 1\,(A_t - F_t)^2}{n} \quad (3)$$

RMSE is the magnitude of the error rate of the prediction results, where the smaller (closer to 0) the RMSE value, the more accurate the prediction results will be [19]. In Equation 4, where n is the total amount of data, e is the error rate, ŷ is the output value or output (prediction) and y is the actual or actual value [7].

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{y}_i - y_i)^2}{n}} \quad (4)$$

MAD is the calculation used to calculate the average absolute (absolute) error [19]. In Equation 5, where *y* is the *actual* result value, $\hat{y}$ is the predicted result value and *n* is the amount of data.

$$\text{MAD} = \frac{\sum |\widehat{y}_i - y_i|}{n} \quad (5)$$

MAPE is the absolute (absolute) average percentage error. MAPE is a statistical measurement of estimates (predictions) accuracy in forecasting methods [19]. In Equation 6, where, n is the amount of data, y is the actual result value and $\hat{y}$ m is the predicted result value [20].

$$\text{MAPE} = \sum_{t=1}^{n} \left| \frac{\widehat{y}_i - y_i}{\widehat{y}_i} \right| \times 100\% \quad (6)$$

Results and Discussions

Following are the results of testing data on staple commodity from 2020 to 2022 by implementing the SVR method and producing the following predictive results:

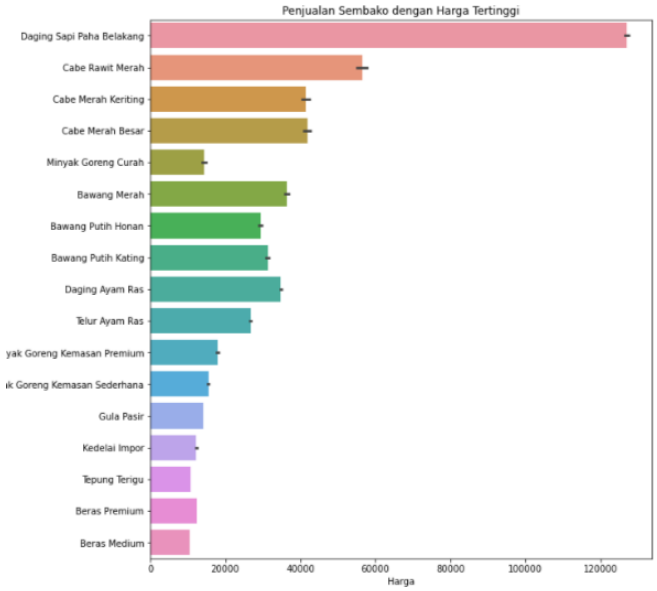

Figure 7. Price Levels of Staple commodities

Figure 7 indicates that the hamstring beef commodity is the highest price yearly among other commodities. This provides information to the public that the price of the highest commodity is the hamstring beef commodity based on the results of the analysis carried out.

In order to determine SVR's performance in predicting the price of staple commodities, the RBF kernel was used to analyze 17 staple commodities. Then it calculated the average MSE, RMSE, MAPE, and MAD values.

Table 1. Test Results

Commodity	MSE	RMSE	MAD	MAPE
Shallots (*Bawang Merah*)	7.51E+47	2.74E+32	2.35E+31	0.014418712223520768
Honan Garlic (*Bawang Putih Honan*)	8.01E+46	2.83E+32	2.44E+31	0.015009651822431304
Kating Garlic (*Bawang Putih Kating*)	7.43E+47	2.73E+32	2.36E+32	0.014495324421074093
Medium Rice (*Beras Medium*)	7.88E+47	2.81E+31	2.36E+32	0.014495324421074093
Premium Rice (*Beras Premium*)	8.31E+46	2.88E+32	2.52E+32	0.01553820949478279
Red Cayenne Peppers (*Cabai Merah Besar*)	7.66E+47	2.77E+32	2.36E+31	0.014613409988810824
Curly Red Chilies (*Cabe Merah Keriting*)	8.01E+47	2.83E+31	2.46E+32	0.015157741308579926
Red Chili Peppers (*Cabe Rawit Merah*)	7.56E+47	2.75E+31	2.36E+32	0.014573507316569815
Meat of Broiler Chicken (*Daging Ayam Ras*)	7.40E+46	2.72E+32	2.35E+32	0.014429023823871456
Beef Hamstrings (*Daging Sapi Paha Belakang*)	7.49E+47	2.74E+31	2.35E+32	0.014457197719871634
Granulated Sugar (*Gula Pasir*)	7.90E+47	2.81E+31	2.42E+31	0.014914231494744384
Imported Soybeans (*Kedelai Import*)	8.20E+47	2.86E+32	2.50E+31	0.015396046224686771
Bulk Cooking Oil (*Minyak Goreng Curah*)	8.16E+47	2.86E+32	2.49E+32	0.015303407275598235
Premium Packaged Cooking Oil (*Minyak Goreng Kemasan Premium*)	6.93E+47	2.63E+31	2.24E+31	0.013803715090595855
Simple Packaged Cooking Oil (*Minyak Goreng Kemasan Sederhana*)	7.02E+47	2.65E+31	2.26E+32	0.013857904145152049
Broiler Chicken Eggs (*Telur Ayam Ras*)	7.76E+47	2.79E+32	2.38E+31	0.014664970920887617
Wheat Flour (*Tepung Terigu*)	8.49E+47	2.91E+32	2.54E+30	0.015643840130898333

The results of testing the dataset of 17 staple foods in Table 1 show that the use of the SVR algorithm used gives an overall average value of 17 staple food commodities, namely the MSE value indicates the average prediction squared error is 6.49, which is also relatively low, the RMSE value indicates that the average the prediction error is 1.77 which is also relatively low, the MAD value indicates the average absolute prediction error is 1.37 which is also relatively low. According to MAPE, the average prediction percentage error is 0.0121, which is relatively low. This demonstrates that the SVR algorithm gives a value of relatively low error number, indicating good accuracy.

Conclusion

Based on the results of testing the time series data on the prices of staple commodity, it is found that there are 18,368 data with 17 types of commodities available. With a data set for the last 3 years from January 1, 2020, to December 31, 2022. There are 3 variables in the data set, namely commodity, date, and price. This research divides the entire dataset into 80% training and 20% testing data. The results of this research show that SVR using the RBF kernel produces very good predicting accuracy for all datasets with an average MSE training data of 6,005 while data testing is 6,062, MAD of training data is 6.730 while data testing is 6.6831, MAPE training data is 0.0148 while data testing is 0.0147, and RMSE training data is 7.772 while data testing is 7.746.

References

[1] A. E. Wardoyo and A. Nilogiri, "Prediksi Harga Sembako Menggunakan Algoritma Memetika Dan Scatter Search Studi Kasus Di Kabupaten Jember," *J. Apl. Sist. Inf. Dan Elektron.*, vol. 2, no. 1, pp. 1–6, 2020.

[2] R. E. Cahyono, J. P. Sugiono, and S. Tjandra, "Analisis Kinerja Metode Support Vector Regression (SVR) dalam Memprediksi Indeks Harga Konsumen (Performance Analysis of Support Vector Regression (SVR) Methods in Predicting the Consumer Price Index)," *J. Teknol. Inf. dan Multimed.*, vol. 1, no. 2, pp. 106–116, 2019.

[3] A. Nastuti, "Teknik Data Mining Untuk Penentuan Paket Hemat Sembako Dan Kebutuhan Harian Dengan Menggunakan Algoritma Fp-Growth (Studi Kasus Di Ulfamart Lubuk Alung)," *J. Ilm. Fak. Sains dan Teknol.*, vol. 7, no. 3, pp. 111–119, 2019.

[4] S. Pangesti, C. Suhery, and T. Rismawan, "Aplikasi Prediksi Harga Sembako Menggunakan Metode Box-Jenkins Berbasis Website," *J. Coding, Sist. Komput. Untan*, vol. 06, no. 03, pp. 139–149, 2018.

[5] J. T. Informatika, U. Jenderal, and A. Yani, "Prediksi Harga Sembako di DKI Jakarta Menggunakan Artificial Neural Network," *JUMANJI*, vol. 3, no. 2, pp. 34–41, 2019.

[6] M. Astiningrum, I. Putri, and V. Wijayaningrum, "Peramalan Harga Bahan Pokok Menggunakan Support Vector Regression," *Sentia 2020*, vol. 12, no. October, pp. 77–82, 2020.

[7] J. E. Islam, "Analisis Pandemic Covid-19 Terhadap Harga Sembako," *Al-Sharf J. Ekon. Islam*, vol. 1, no. 1, pp. 43–51, 2020, doi: 10.56114/al-sharf.v1i1.34.

[8] S. Saadah, F. Z. Z, and H. H. Z, "Support Vector Regression (SVR) Dalam Memprediksi Harga Minyak Kelapa Sawit di Indonesia dan Nilai Tukar Mata Uang EUR/USD," *J. Comput. Sci. Informatics Eng.*, vol. 5, no. 1, pp. 85–92, 2021, doi: 10.29303/jcosine.v5i1.403.

[9] L. Y. Kurniawati, H. Tjandrasa, and I. Arieshanti, "Prediksi pergerakan harga saham menggunakan support vector regression," *J. SimanteC*, vol. VIII, no. 1, pp. 11–21, 2013.

[10] N. D. Maulana, B. D. Setiawan, and C. Dewi, "Implementasi Metode Support Vector Regression (SVR) Dalam Peramalan Penjualan Roti (Studi Kasus : Harum Bakery)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 3, pp. 2986–2995, 2019.

[11] R. K. D. Olivia Bonita, Lailil Muflikhah and Program, "Prediksi Harga Batu Bara Menggunakan Support Vector Regression (SVR)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 12, pp. 6603–6609, 2018.

[12] K. Puteri and A. Silvanie, "Machine Learning Untuk Model Prediksi Harga Sembako Dengan Metode Regresi Linier Berganda," *J. Nas. Inform.*, vol. 1, no. 2, pp. 82–94, 2020.

[13] H. Sumarno, "Analisis Support Vector Regression Dengan Algoritma Grid Search Untuk Memprediksi Harga," *Math. Subj. Classif.*, vol. 18, no. 1, pp. 41–60, 2022.

[14] K. Dewi, P. P. Adikara, and S. Adinugroho, "Prediksi Indeks Harga Konsumen (IHK) Kelompok Perumahan , Air , Listrik , Gas Dan Bahan Bakar Menggunakan Metode Support Vector Regression," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 2, no. 10, pp. 3856–3862, 2018.

[15] N. P. N. Hendayanti and M. Nurhidayati, "Perbandingan Metode Seasonal Autoregressive Integrated Moving Average (SARIMA) dengan Support Vector Regression (SVR) dalam Memprediksi Jumlah Kunjungan Wisatawan Mancanegara ke Bali," *J. Varian*, vol. 3, no. 2, pp. 149–162, 2020, doi: 10.30812/varian.v3i2.668.

[16] R. Tange, M. A. Rasmussen, E. Tairac, and R. Bro, "Application of support vector regression for simultaneous modelling of near infrared spectra from multiple process steps," *J. Near Infrared Spectrosc.*, vol. 23, no. 2, pp. 75–84, 2015, doi: 10.1255/jnirs.1149.

[17] S. Lutfia Septiningrum1, Hasbi Yasin2, "Prediksi Indeks Harga Saham Gabungan Menggunakan Support Vector Regression (Svr) Dengan Algoritma Grid Search," *Gaussian*, vol. 4, no. 2, pp. 315–321, 2015.

[18] R. E. Caraka, H. Yasin, and A. W. Basyiruddin, "Peramalan Crude Palm Oil (CPO) Menggunakan Support Vector Regression Kernel Radial Basis," *J. Mat.*, vol. 7, no. 1, pp. 43–57, 2017, doi: 10.24843/jmat.2017.v07.i01.p81.

[19] S. S. M. Evy Sulistianingsih, "Prediksi Nilai Tukar Dolar Amerika Serikat Terhadap Rupiah Dengan Metode Support Vector Regression (Svr)," *Bimaster Bul. Ilm. Mat. Stat. dan Ter.*, vol. 8, no. 1, pp. 1–10, 2018, doi: 10.26418/bbimst.v8i1.30503.

[20] D. Sepri, A. Fauzi, R. Wandira, O. S. Riza, Y. F. Wahyuni, and H. Hutagaol, "Prediksi Harga Cabai Merah Menggunakan Support Vector Regression," *Comput. Based Inf. Syst. J.*, vol. 08, no. 02, pp. 1–5, 2020.