Indonesian Journal of Information Technology and Computing

Vol. 4, No.1 (2024), pp. 1-10 | e-ISSN: 2798-9216
https://journal.polhas.ac.id/index.php/imaging

Data Mining Processing Using K-Means and Support Vector Regression Methods on Indomie Products (Case Study: UIN Sunan Kalijaga Yogyakarta Student Cooperative)

Deddy Rahmadi[1] , Nida Huwaida Rizqya Zulfa[2], Ananda Muhammad Akbar[2], M. Rozzan Abdillah[2], Faiq Ziyaurrohman[2] Yusuf Al Humam [2]

[1] Mathematics Study Program, UIN Sunan Kalijaga Yogyakarta, Yogyakarta, Indonesia
[2] Industrial engineering Study Program, UIN Sunan Kalijaga Yogyakarta, Yogyakarta, Indonesia
E-mail: [1] deddy.rahmadi@uin-suka.ac.id

Article History: Received: June, 19 2024; Accepted: June 24, 2024; Published: June, 30 2024

ABSTRACT

This research was conducted to analyze the sales pattern of Indomie products at the UIN Sunan Kalijaga Yogyakarta Student Cooperative and predict future product prices. The data used is daily sales data from January to June 2023 with a total of 599 data into five clusters with the number of items cluster 0 consists of 32 items, cluster 1 consists of 409 items, cluster 2 consists of 102 items, cluster 3 consists of 48 items, and cluster 4 consists of 8 items. The methods used are K-Means for clustering and Support Vector Regression (SVR) for price prediction. The results of the K-Means analysis grouped the products into five clusters with different characteristics. In the Support Vector Regression (SVR) method, initially it has an accuracy rate of 70% with a fairly high Mean Squared Error (MSE) and Mean Absolute Error (MAE). After cleaning the data from outliers and tuning the hyperparameters, the model accuracy increased to 99%, showing a significant improvement in the model's predictive ability.

Keywords: *Clusterring, K-Means, SVR, Indomie*

INTRODUCTION

Data processing is an important thing that needs to be done to gain valuable insights from the collected information. One of the commonly used data processing techniques is data mining, which serves to identify patterns and relationships hidden in large data sets. The K-Means method is one of the clustering algorithms in data mining that is popular due to its simplicity and efficiency. This algorithm divides data into a number of clusters based on feature similarity, so that each data in one cluster has a high similarity with each other compared to data in other clusters. K-Means works by determining the cluster center (centroid) and grouping data into clusters closest to the centroid, through continuous iteration until convergence is achieved (Hutagalung & Sonata, 2021). One of the case studies used is in the student cooperative (KOPMA) UIN Sunan Kalijaga. This paper aims to determine the grouping of noodle sales with the indomie brand into several clusters based on the method. The initial stage in this case study is problem identification, followed by literature study, data collection and then data processed using the Cross-Industry Standard Process for Data Mining (CRISPDM) method and Exploratory Data Analysis (EDA) method. The use of the K-Means method allows analysts and researchers to simplify the complexity of data, making it easier to make more precise and effective data-based decisions. In the context of sales, this method can be used to find out the relationship between

data and help determine the profit obtained. Decision-making on which products sell best allows for increased profits.

METHODS

Methods thoose used in this paper are CRISP-DM for processing then continue by K-MEANS and SVR. The material explain below:

CRISP-DM

The data obtained is sales data per day of all products in KOPMA UIN Sunan Kalijaga in the period January - June 2023. The date provided from KOPMA only includes daily data so there are various Excel pages. The number of files was spread over 180 different excel pages (covering daily excel from January to June 2023) so they had to be sorted and combined into one. The amount of data in each excel is different, even the dates are not always in order. One Excel can input many items sold during a day. Using the help of Excel, the data can be processed into a table like the following. A date column is added to show the items sold. The following is an excel image that has been classified.

The total sales data of KOPMA during January - June 2023 is 58,512 data. Then the related attributes are code, product name, packaging, sold, selling price, net, gross profit, receipt, frequency, barcode, real stock, and additional date. After analysis using CRISPDM such as business understanding, data understanding, and data preparation, the data is ready to be processed like shown (Schröer et al., 2021).

The data processing method used in this research begins with the following steps.

1. Observation and direct data collection at UIN Sunan Kalijaga Yogyakarta Student Cooperative which was conducted on Thursday, March 14, 2024. Data collection is done directly through the process of looking at Kopma UIN Suka sales data and moving secondary data as the object of research.
2. Literature study is carried out through several relevant studies and other supporting references.Sales data processing in group 1 research uses the CRISP-DM (Cross Industry Standard Process for Data mining) model then further data processing is using K-Means and Support Vector Regression. The data used based on the results of data preparation processing are as follows:

a. Code, useful for unique identification of each product or service in the sales system.
b. Product Name, is the name or brand of a product.
c. Sold or "sell" indicates the number of products or services sold in one transaction or period of time.
d. Selling Price, the price charged to customers for each unit of product or service purchased.
e. Net, Indicates the original price that the cooperative purchased before selling to customers.
f. Gross Profit, Shows the difference between net minus selling price.
g. Real Stock, the actual amount of product available in inventory at a given point in time.
h. Date, Indicates the date the sales transaction was executed.

Clustering

Clustering is one of the techniques in data mining where algorithms are used to group data into specific groups (Sibuea & Sapta, 2017). This technique is used to group data based on similar characteristics. The clustering process involves determining or describing a quantitative value of the degree of similarity or difference between data (proximity measure), which is an important step in the process.

K-Means

K-Means is one of the algorithms in data mining that can be used for clustering data. There are many approaches to clustering, one of which is to create rules that dictate membership in the same group based on the degree of similarity among its members. Another approach is to measure a set of properties with those of the clustering as a function of some parameters of the clustering. Then, randomly select K starting points as cluster centers (centroids). Each data is then assigned to the cluster whose center is closest, based on the Euclidean distance. After all data has been assigned, the cluster center is updated by calculating the average of all points in the cluster. This process is repeated-reassigning the data to the closest cluster and updating the cluster center-until the cluster center does not change significantly or reaches a predetermined maximum number of iterations. The main goal of K-means is to minimize the total squared distance between the data and their cluster centers, thus achieving optimal clustering of the data (Nishom, 2019). This algorithm is efficient and easy to implement, but has some drawbacks such as dependence on the initial selection of cluster centers and difficulty in determining the optimal number of clusters. Unsupervised K-Means will give cluster that automatically found without given any boundaries (Sinaga & Yang, 2020).

Support Vector Regression

According to (Fu & Li, 2024) Support Vector Regression (SVR) has competitive overall prediction performance compared to other supervised learning algorithms in unbalanced regression problems. However, the performance of SVR in rare event prediction is slightly significantly different from other compared algorithms. Moreover, if the data set is highly imbalanced, the prediction performance of SVR will degrade significantly.

According to (Awad & Khanna, 2015) SVR is characterized by the use of kernels, sparse solutions, and VC control of the margin and number of support vectors. Although less popular than SVM, SVR has proven to be an effective tool in real value function estimation. As a supervised learning approach, SVR trains using a symmetric loss function, which penalizes high and low estimation errors equally.

Support Vector Regression (SVR) is a regression model developed from Support Vector Machine (SVM). This algorithm produces output in the form of real or continuous numbers (Amanda et al., 2014) SVR is able to reduce the risk of overfitting by minimizing the upper bound of generalization error. The main goal of this algorithm is to find the best dividing line, called a hyperplane (Ginting et al., 2021)The best hyperplane can be determined by measuring its margin, which is the distance between the hyperplane and the closest pattern. The pattern closest to this margin is called the Support Vector. The SVR algorithm is based on the concept of linear regression. As a result, SVR gives an viable instrument for taking care of high-dimensional information. In addition,SVR may be a machine learning strategy that learns a show to depict the factors that are critical in characterizing the relationship between input and and yields, not at all like conventional relapse strategies that depend on demonstrate suspicions that not exactly true (Zhang & O'Donnell, 2019).

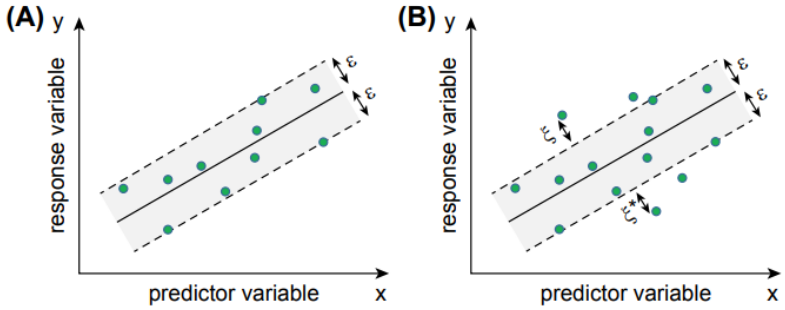

Figure 1. comparing represent data in SVR method
(Zhang & O'Donnell, 2019)

Figure 1 show that in A happen when the data is in boundaries and B show when the data is far from the boundaries. This part will measure in the percentage. Prediction will be more established when use optimization parameters (Quan et al., 2022). This also use to this paper use more optimization code. Reality show Predicting the results of sales of HT Motorola XiR C2660 products at CV.Alfacoms in the year 2023 using the SVR algorithm. Testing parameters that have passed the test stage, the best parameters are kernel with linear type and parameter C with a test value of 0.1. test value with MAPE error value of 11.23% (Wildwina & Kristianto, 2024).

Rapid Minner

RapidMiner is a tool that can be used to perform data mining, text mining, and predictive analysis. This tool uses a variety of descriptive and predictive approaches and techniques to build models that support decision making (Chisholm, 2013). In RapidMiner, model building does not require programming knowledge because all operations needed to build a model are already available in the form of operators that can be connected to each other (Sholeh et al., 2023). In RapidMiner K Means calculate by step find the cluster, then calculate the distance by using Euclidian distance, after that we can find the closest cluster (Tendean & Purba, 2020).

The research flow chart can be broken down into several practical steps that are carried out for problem solving. The steps can be seen in the flow chart as in the following figure.

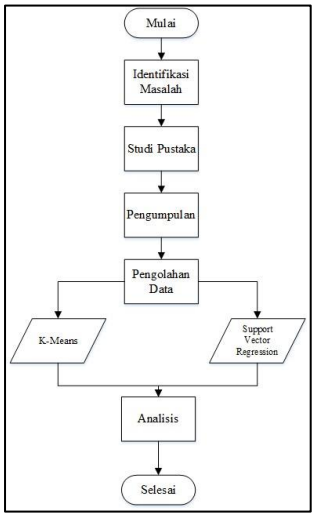

Figure 2. Flow Chart

The initial stage in this research is the process of identifying problems in the object of research and existing data, with the aim of formulating the purpose of this research. Furthermore,

scientific references are collected through literature studies of relevant research. After the literature study, the data collection process was carried out in the field, namely at the UIN Sunan Kalijaga Yogyakarta Student Cooperative. Data collection was carried out by observation and requesting data from Kopma. After the data was collected, data processing was carried out using EDA analysis at the beginning and the use of the CRISP-DM model to understand and approach the data. After the data is processed, the next stage is to read and analyze the results of the research to further conclude and evaluate this research. Furthermore, it will be processed using the K Means and Support Vector Regression methods. Modification used when writer use two methods instead one.

RESULTS AND DISCUSSION

CRISPDM (Data Preparation)

After data selection, namely eliminating minus data, 599 Indomie sales data at KOPMA are ready to be used. Analysis using EDA obtained that there is no missing data. Furthermore, the use of correlation to determine the variables used is based on the correlation matrix. In research, show that using this method can approach the creative problem solving process (Jaggia et al., 2020).

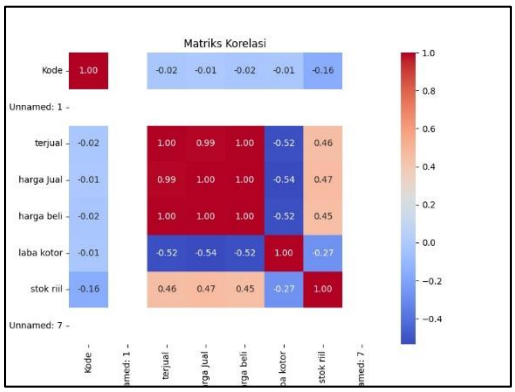

Figure 3. correlation precessing

The data that will be used to determine several interrelationships, including selling and selling price have an effect, selling and buying price have an effect, then gross profit has no effect on selling and real stock has no effect on selling. Based on the data, the variables selected are selling price, real stock, and gross profit. Based on previous processing, it was found that the best-selling Indomie brand is Indomie Goreng Special 85gr.

Modelling

K Means Method, have input

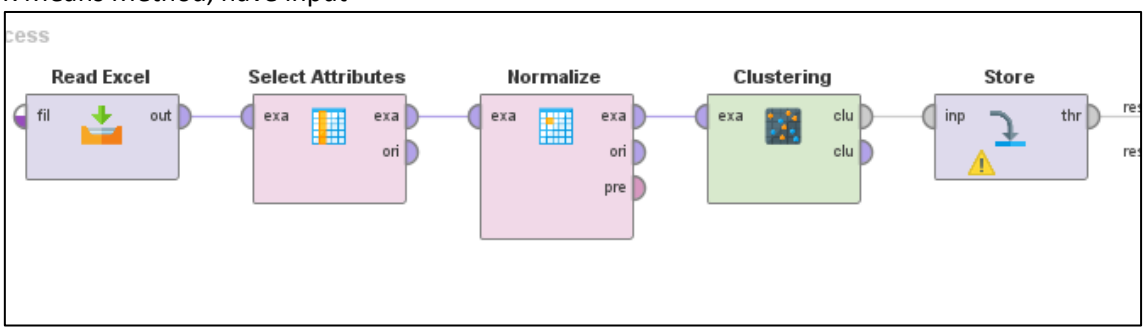

Figure 4. Input K Means

The figure above illustrates the stages of the data mining process starting from reading Excel files, selecting relevant attributes, normalizing data, clustering, and saving the results.

Cluster Model

```
Cluster 0: 32 items
Cluster 1: 409 items
Cluster 2: 102 items
Cluster 3: 48 items
Cluster 4: 8 items
Total number of items: 599
```

Figure 5. K-Means result

The figure above shows the clustering results that divide 599 items into five clusters with different numbers of items, namely Cluster 0 consists of 32 items, Cluster 1 consists of 409 items, Cluster 2 consists of 102 items, Cluster 3 consists of 48 items, and Cluster 4 consists of 8 items.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
Kode	-0.622	-0.038	0.669	-0.622	-0.346
Terjual	1.906	-0.458	0.824	0.007	5.249
Hrg Jual	2.056	-0.471	0.826	0.050	5.033
Netto	1.923	-0.460	0.828	0.004	5.264
Laba Kotor	2.548	-0.480	0.732	0.287	3.287
Stok Riil	2.204	-0.406	-0.300	2.299	1.967

Figure 6. Centroid Matrix

The figure above shows the centroid matrix of each cluster, which reflects the average value of the attributes within each cluster. For example, cluster 0 has high values on attributes such as 'Sold', 'Sold Hrg', 'Netto', 'Gross Profit', and 'Real Stock', indicating that items in this cluster tend to have high sales and stock and large gross profit. In contrast, cluster 4 shows the highest values on the same attributes, indicating that the items in this cluster are highly valuable and have strong financial performance.

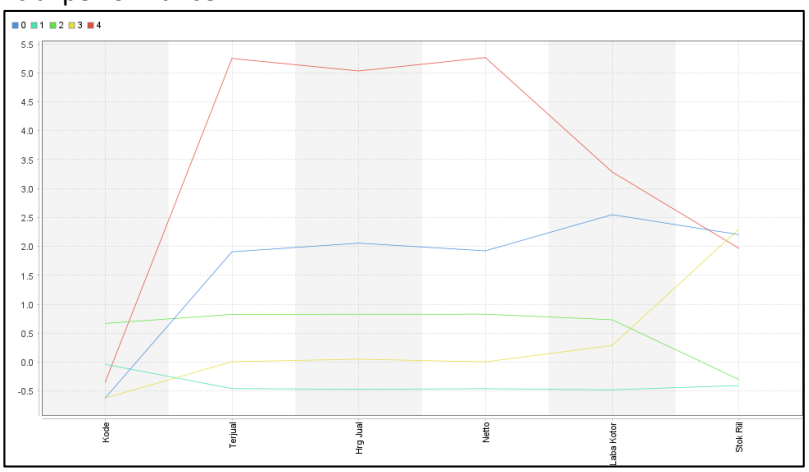

Figure 7. Value change graph

The graph shows the value changes of four variables (labeled 0, 1, 2, and 3) in six different categories, namely Code, Sold, Hrg Jual, Net, Discount, Gross Profit, and Rill Stock. From the

graph, it can be seen that the variable labeled 4 (marked with a red line) experiences a significant increase in the Sold and Hrg Jual categories, reaching its peak in Hrg Jual before finally decreasing in the next category. The variable labeled 0 (marked with a blue line) shows a more steady and gradual increase in all categories, while the variables labeled 2 and 3 (marked with green and yellow lines) remain relatively flat with little variation. In the last category, Starting Stock, all variables tend to converge or approach similar values. Overall, this graph indicates that variable 4 has the largest fluctuations, while the other variables show more stable changes.

Support Vector Regression Method

Using price data because you want to know the future price of the Indomie product. Based on the processing results, the following values are obtained:

```
Mounted at /content/drive
Best hyperparameters: {'svr__C': 1000, 'svr__epsilon': 10, 'svr__gamma': 'auto'}
Best score: -9063082.085213438
Mean Squared Error on test set: 14473982.449701456
Mean Absolute Error on test set: 611.1533479935298
R^2 Score on test set: 0.6920920462462458
```

Figure 8. Model results without cleaning

Based on the initial results, it is known that the initial value of hyperparamaters is C: 1000 (the margin is quite large) and the tolerance level is quite high the best value is -9063082.0852 with MSE: 14473982.449701456 and MAE: 611.153 which shows that most predictions are quite close to the actual value. And the data variability is 69%. The paper from Malik et al. (2020) show that we can know the model is better or no by see the RMS. In that paper also give good accurancy. From this study, the graphical results are as follows:

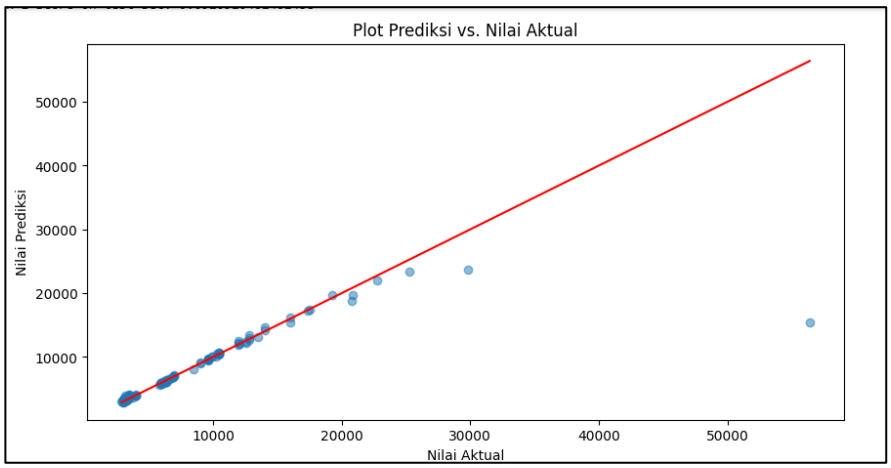

Figure 9. First graph image without data cleaning

First graph without data cleaningThe red line is the identity (ideal) line, where the prediction is equal to the actual value. Blue dot represents one observation with x axis as actual value and y as prediction. The blue dot indicates the prediction, if it is close then it is accurate, if it is far then it is less accurate.

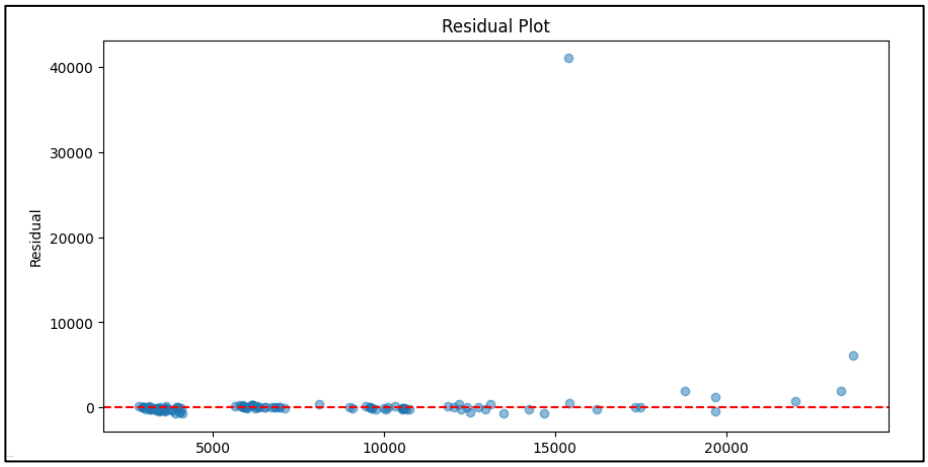

Figure 10. Residual image without data cleaning

The residual is the difference between the original value and the forecast. The red line is at zero point where the residual is zero meaning there is no difference. Then we will do a comparison using outlier cleaning and increased hyperparamater tuning then the second result becomes:

```
Best hyperparameters: {'svr__C': 1000, 'svr__epsilon': 10, 'svr__gamma': 'auto'}
Best score: -9721687.778275434
Mean Squared Error on test set: 13787977.231033009
Mean Absolute Error on test set: 484.85047779128183
R^2 Score on test set: 0.7066855739003404
```

Figure 11 The second result with increased hyperparameter

From the results it is known that the predicted value is 70%. Then it will be checked again with handling outliers and hyperparameters, the performance becomes

```
Best hyperparameters: {'svr__kernel': 'linear', 'svr__gamma': 'scale', 'svr__epsilon': 0.1, 'svr__C': 1000}
Best score: -772.2272512885403
Mean Squared Error on test set: 0.9451122679620059
Mean Absolute Error on test set: 0.3347688736734502
R^2 Score on test set: 0.9999999798944357
```

Figure 12. Third result image with outlier handling and hyperparameter

Based on the processing, it gives an increased value of 99% where the initial model only amounts to about 70%. And the graph is the best. From study, Caraka et al. (2020) we should put attention to MSE, gamma, and epsilon. Epsilon 0.1 show that margin error is low. Its different than epsilon use 10 that have 10 its show big margin error. It also be support by The RMSE and MAE were reduced by 85% and 61% respectively, and the R2value increased by by almost 10%. SVR model trained, validated and tested on data(Brkić & Larva, 2024). Here are the graphic of best moel.

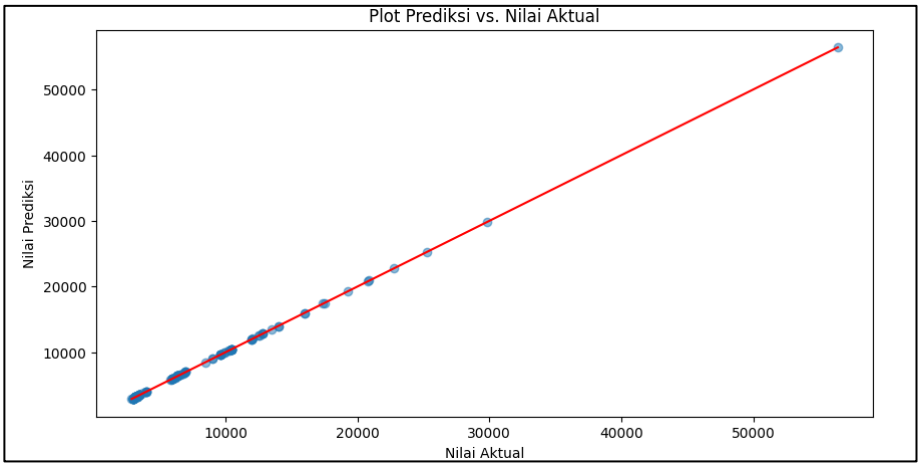

Figure 13. Image of the best graph

Support Vector Regression (SVR) does not provide an explicit model equation but SVR maps the data using a kernel function and then maximizes the margin and accuracy of the model. Based on the processing, the value of SVR is able to minimize non-linear data, high flexibility, and tuning handling. So to predict new prices can use the SVR model based on the most recent mse, mae, r2 values in coding. Lika research that tell that a model need to be more processed after one input, SVR model ilustrated less susceptibility to overfitting an underfitting (Ahmad et al., 2020). After so many long ways, it found most acurate model by remove the outlier that not shown in the process before.

CONCLUSION

The K-Means method successfully grouped Indomie product sales data at the UIN Sunan Kalijaga Yogyakarta Student Cooperative into five different clusters based on sales characteristics, such as code, sold, selling price, net, and real stock. Meanwhile, the Support Vector Regression (SVR) method is effective in predicting the future price of Indomie products. With data cleaning and hyperparameter tuning, the SVR model achieved high prediction accuracy, from 70% to 99%. These results show that SVR can be used to predict prices well.

For further research, it would be better to use other software as a comparison of results. Considering that unsupervised usually has high processing without any error comparison set.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest regarding the publication of this article. The research entitled "Data Mining Processing Using K-Means and Support Vector Regression Methods on Indomie Products (Case Study: UIN Sunan Kalijaga Yogyakarta Student Cooperative)" is published with full independence and academic integrity. There are no financial, commercial, or personal relationships that could influence the research process, results, or conclusions reached in this study.

REFERENCES

Ahmad, M. S., Adnan, S. M., Zaidi, S., & Bhargava, P. (2020). A novel support vector regression (SVR) model for the prediction of splice strength of the unconfined beam specimens. *Construction and Building Materials*, *248*. https://doi.org/10.1016/j.conbuildmat.2020.118475

Amanda, R., Yasin, H., & Prahutama, A. (2014). *ANALISIS SUPPORT VECTOR REGRESSION (SVR) DALAM MEMPREDIKSI KURS RUPIAH TERHADAP DOLLAR AMERIKA SERIKAT*. *3*(4), 849–857. http://ejournal-s1.undip.ac.id/index.php/gaussian

Awad, M., & Khanna, R. (2015). Support Vector Regression. In *Efficient Learning Machines* (pp. 67–80). Apress. https://doi.org/10.1007/978-1-4302-5990-9_4

Brkić, Ž., & Larva, O. (2024). Impact of climate change on the Vrana Lake surface water temperature in Croatia using support vector regression. *Journal of Hydrology: Regional Studies*, *54*. https://doi.org/10.1016/j.ejrh.2024.101858

Caraka, R. E., Chen, R. C., Bakar, S. A., Tahmid, M., Toharudin, T., Pardamean, B., & Huang, S. W. (2020). Employing best input SVR robust lost function with nature-inspired metaheuristics in wind speed energy forecasting. *IAENG Int. J. Comput. Sci*, *47*(3), 572-584.

Chisholm, A. (2013). *Exploring data with RapidMiner : explore, understand, and prepare real data using rapidminer's practical tips and tricks*.

Fu, G.-H., & Li, Z.-Z. (n.d.). *Rare event prediction in imbalanced regression with adaptive weighted support vector regression*. https://ssrn.com/abstract=4839324

Hutagalung, J., & Sonata, F. (2021). Penerapan Metode K-Means Untuk Menganalisis Minat Nasabah. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(3), 1187. https://doi.org/10.30865/mib.v5i3.3113

Jaggia, S., Kelly, A., Lertwachara, K., & Chen, L. (2020). Applying the CRISP-DM Framework for Teaching Business Analytics. In *Decision Sciences Journal of Innovative Education* (Vol. 18).

Ginting, L., MTSigiro, M., Delima Manurung, E., & Jasa Putra Sinurat, J. (2021). Perbandingan Metode Algoritma Support Vector Regression dan Multiple Linear Regression Untuk Memprediksi Stok Obat. In *Journal of Applied Technology and Informatics* (Vol. 1, Issue 2). http://journal-jati.del.ac.id/

Malik, A., Tikhamarine, Y., Souag-Gamane, D., Kisi, O., & Pham, Q. B. (2020). Support vector regression optimized by meta-heuristic algorithms for daily streamflow prediction. *Stochastic Environmental Research and Risk Assessment*, *34*(11), 1755–1773. https://doi.org/10.1007/s00477-020-01874-1

Nishom, M. (2019). Perbandingan Akurasi Euclidean Distance, Minkowski Distance, dan Manhattan Distance pada Algoritma K-Means Clustering berbasis Chi-Square. *Jurnal Informatika: Jurnal Pengembangan IT*, *4*(1), 20–24. https://doi.org/10.30591/jpit.v4i1.1253

Quan, Q., Hao, Z., Xifeng, H., & Jingchun, L. (2022). Research on water temperature prediction based on improved support vector regression. *Neural Computing and Applications*, *34*(11), 8501–8510. https://doi.org/10.1007/s00521-020-04836-4

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*(2019), 526–534. https://doi.org/10.1016/j.procs.2021.01.199

Sholeh, M., Kumalasari Nurnawati, E., & Lestari, U. (2023). Penerapan Data Mining dengan Metode Regresi Linear untuk Memprediksi Data Nilai Hasil Ujian Menggunakan RapidMiner. In *Jurnal Informatika Sunan Kalijaga)* (Vol. 8, Issue 1). https://archive.ics.uci.edu/ml/datasheets.php.

Sibuea, F. L., & Sapta, A. (2017). *PEMETAAN SISWA BERPRESTASI MENGGUNAKAN METODE K-MEANS CLUSTERING*. *1*, 85–92.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796

Tendean, T., & Purba, W. (2020). Analisis Cluster Provinsi Indonesia Berdasarkan Produksi Bahan Pangan Menggunakan Algoritma K-Means. *Jurnal Sains Dan Teknologi)*, *1*(2), 5–11.

Wildwina, & Ryan Putranda Kristianto. (2024). Prediksi Penjualan HT Motorola XiR C2660 Menggunakan Algoritma Support Vector Regression (Studi Kasus: CV. Alfacoms). *Jurnal Teknik Informatika Dan Komputer*, *3*(1), 11–16. https://doi.org/10.22236/jutikom.v3i1.13836

Zhang, F., & O'Donnell, L. J. (2019). Support vector regression. In *Machine Learning: Methods and Applications to Brain Disorders*. Elsevier Inc. https://doi.org/10.1016/B978-0-12-815739-8.00007-9