Journal of Artificial Intelligence and Engineering Applications

Website: https://ioinformatic.org/

15th February 2025. Vol. 4. No. 2; e-ISSN: 2808-4519

Sales Data Analysis Using Linear Regression Algorithm on Raw Water Sales

Eti Rohayati[1]*, Martanto[2], Arif Rinaldi Dikananda[3], Dede Rohman[4]

[1,4]Teknik Informatika, STMIK IKMI Cirebon
[2]Manajemen Informatika, STMIK IKMI Cirebon
[3]Rekayasa Perangkat Lunak, STMIK IKMI Cirebon
rohayatieti408@gmail.com[1], martantomusijo@gmail.com[2], rinaldi21crb@gmail.com[3], deden.ikmi@gmail.com[4]*

Abstract

This study aims to assess the effectiveness of linear regression algorithm in predicting raw water demand by considering customer transaction data, raw water volume, and seasonal variables. The method used is Knowledge Discovery in Databases (KDD), including data selection, preprocessing, transformation, data mining, and result evaluation. The dataset is divided 80% for training and 20% for testing. The analysis results show that the linear regression model has a coefficient of determination (R^2) of 0.77, which means that the model can explain 77% of the data variability. The prediction error value is low, with Mean Absolute Error (MAE) 0.06, Mean Squared Error (MSE) 0.01, and Root Mean Squared Error (RMSE) 0.08, indicating good accuracy. In the comparison between actual and predicted values, for actual data of 7,000 liters, the model predicts 7,984.70 liters. The variable number of customer transactions has the greatest influence on raw water demand, with a coefficient of 16,940.46, while seasonal factors have less influence. Based on these findings, it can be concluded that the linear regression algorithm is effective in predicting raw water demand, however further development is required to improve accuracy at extreme values, by adding variables or using more complex algorithms.

Keywords: *linear regression, sales prediction, raw water, business strategy, data analysis*

1. Introduction

The rapid development of informatics technology has brought significant impacts in various sectors of life, including in the business world. This technology enables more in-depth data analysis, which supports strategic decision-making, such as in sales forecasting and production planning. The use of linear regression algorithms has been widely applied to analyze the relationship between variables and produce accurate predictions, in analyzing business sales data [1], and for cash flow and sales projections [2].
Today's digital era faces major challenges in data management and data-driven decision-making, particularly with regard to prediction accuracy in sales projections, inventory management, and marketing. These obstacles often arise due to low data quality and inappropriate variable selection. It has been shown that without proper preprocessing or selection of relevant variables, linear regression can produce biased and inconsistent predictions, signaling the need for a more systematic approach in its application [3].

Demonstrating the effectiveness of linear regression in predicting property sales and seasonal production [4]. These algorithms are able to identify clear linear patterns, providing easy-to-understand insights for decision makers. However, most of the existing research is still focused on specific case studies, with little exploration of linear regression optimization to handle data with high variation or non-linear factors. This highlights the importance of further research for the development and optimization of linear regression algorithms.
This research aims to develop a more optimal application of linear regression algorithms to improve the accuracy and efficiency of predictive analysis, especially for sales data that has high variation. Emphasize the importance of data clustering to improve prediction accuracy [5]. Demonstrates the use of linear regression in predicting the number of Umrah registrants. This research is expected to provide practical insights in the application of linear regression algorithms to complex and varied data [6].

The quantitative approach used in this study focuses on the application of linear regression to predict sales, with an emphasis on preprocessing techniques to improve model accuracy. This research adopts a similar approach to that which uses multiple linear regression to predict new student enrollment [7]. Using variable clustering techniques to improve prediction accuracy. This research aims to produce a more reliable model, especially in handling dynamic sales data [8].
The results obtained from this study are expected to contribute to the development of linear regression algorithms, especially in predicting complex sales data, such as seasonal or highly fluctuating data. Similar findings in the context of trend-based logistics service prediction. This research can provide practical guidance for the industry in adopting linear regression algorithms to improve prediction accuracy, as well as filling the knowledge gap that exists in the literature related to the application of data preprocessing in linear regression [9].

2. Research Methods

The following are the KDD (Knowledge Discovery in Databases) stages used in the research and are organized in Figure 1.

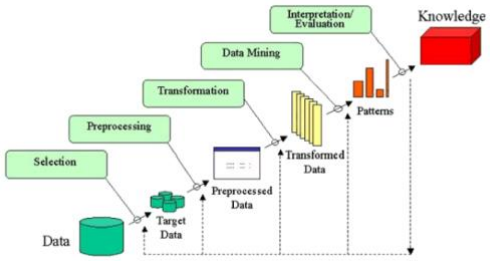

Fig. 1: Knowledge Discovery in Database (KDD).

2.1. Data Selection

Raw water sales data, which was originally a daily format from January 2021 to October 2024, was converted into a monthly digital format using Microsoft Excel. Rainfall data obtained from BMKG was added to provide seasonal context. The main variables analyzed included number of transactions, water volume, and season, with a total of 1610 records.

2.2. Preprocessing

The data was cleaned by removing blank values, converting categorical variables into numerical format, and removing outliers to ensure data quality and consistency.

2.3. Data Transformation

The normalization process was performed on the dataset using the Min-Max Scaling technique to transform the values into the range [0, 1]. Next, the normalized numerical data was combined with the coded categorical data, forming a dataset ready for analysis.

2.4. Data Mining

Model evaluation was carried out using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), comparison between actual and predicted values used to assess the accuracy of the model in projecting raw water demand and coefficient of determination (R^2). Analysis of the regression coefficient and intercept helps in understanding the relationship between variables.

2.5. Interpretation/ Evaluation

Model evaluation was carried out using metrics such as Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), comparison between actual and predicted values used to assess the accuracy of the model in projecting raw water demand and coefficient of determination (R^2). Analysis of the regression coefficient and intercept helps in understanding the relationship between variables.

2.6. Knowledge

The knowledge stage involves analyzing the evaluation results to identify significant variables, such as the number of transactions that have the most influence. These findings are then used to formulate recommendations for model development, including the addition of relevant variables or the application of more complex algorithms to improve accuracy, especially in the face of extreme values. This process is instrumental in transforming data into insights that are useful for more effective decision-making and business strategy planning.

3. Result and Discussion

The following are the results and discussion of the research:

The dataset was divided into two subsets, 20% for test data and 80% for training data, with the aim of ensuring the model obtained sufficient data for training and could be tested using separate data. The independent variables used in this model are “Transaction Amount” and “Season_encoded”, while the predicted dependent variable is “Water Amount (liters)”. This division of data is important to objectively evaluate the accuracy and predictive ability of the model.

Various evaluation metrics are used to measure the extent to which the model is able to explain the variability of the data and its accuracy in making predictions.

Table 1: Model Evaluation

Model Evaluation	Value
Mean Absolute Error	0.06

Model Evaluation	Value
Mean Squared Error	0.01
Root Mean Squared Error	0.08
R-squared	0.77

Table 1 displays the acceptable prediction error rate metrics, where the Mean Absolute Error (MAE) value is 0.06, the Mean Squared Error (MSE) is 0.01, and the Root Mean Squared Error (RMSE) is recorded at 0.08. These figures show that the errors are quite small, which indicates satisfactory model performance in this study. Such a decrease in error value indicates good model quality in terms of prediction [13]. In this study, the coefficient of determination (R^2) value was recorded at 0.77, which indicates that the model can explain 77% of the data variation. This indicates that the model has adequate performance in the context of prediction. This result is in line with studies that reveal that a high R^2 value indicates the model's adequate ability to describe the data [14].

Table 2: Actual and Predicted Values

Actual Value	Predicted Value
6500	7412.966633
7000	7984.698346
…	…
9500	8027.946887
11000	10314.873735

Table 2 presents a comparison between the actual values and the predicted results generated by the model. In general, the model performs well in predicting values with a high degree of accuracy for most of the data, although there are some more significant deviations in data with extreme values. For example, when the actual value was 7,000 liters, the model predicted 7,984.70 liters, but on extreme data such as 13,500 liters, the prediction was 12,558.55 liters. This shows that although the model is quite effective for most data, there is still potential for improvement on extreme data [15].

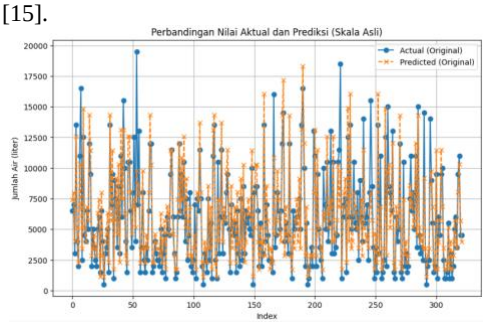

Fig. 2: Comparison of Actual and Predicted Values

Figure 2 shows a visualization of the comparison between actual values (in blue) and predicted values (in orange). This visualization shows that the model works effectively for data below 10,000 liters, but starts to show significant deviations for data exceeding 15,000 liters. This finding is in line with studies highlighting that linear regression models can capture general data patterns, but have difficulty predicting data with extreme values [16].

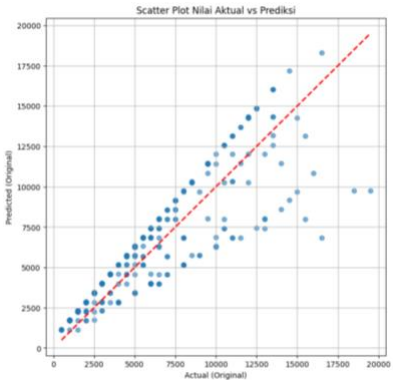

Fig. 3: Scatter Plot of Actual vs Predicted Values

Figure 3 displays a scatter plot illustrating the relationship between the actual and predicted values. Overall, most of the data points are close to the perfect prediction line, indicating good model accuracy. However, at the extremes of the data, there are larger deviations, indicating a potential for improvement in these areas. This finding is in line with research in the field of linear regression-based prediction which also shows similar challenges in predicting extreme values [17].

Table 3: Intercept and Regression Coefficient

Intersept	Transaction Amount	Season
1172.43	19640.46	-23.92

Table 3 presents the intercept result of 1172.43 indicating the initial value that occurs when all independent variables are zero. The coefficient for the variable “Number of Transactions” is 16,940.46, which indicates a significant positive relationship, where every one unit increase in the number of transactions is predicted to increase the amount of water by 16,940.46 liters. This coefficient indicates a

strong relationship between the number of transactions and water demand, which is in line with other studies that emphasize the importance of the relationship between the independent variables and the prediction results (Pratama et al., 2024). The variable "Season_encoded" shows a negative coefficient of -23.92, indicating that a change from the "Dry" (0) to "Rainy" (1) season is predicted to reduce water demand by 23.92 liters. This decrease can be understood as a consequence of increased natural water availability in the rainy season, which has implications for decreasing raw water demand. This finding is in line with research discussing the effect of seasonality on consumption patterns [18].

Table 4: Correlation Matrix

Correlation Matrix	Transaction Amount	Water Amount (liters)	Season_encoded
Transaction Amount	1.000000	0.851130	-0.063484
Water Amount (liters)	0.851130	1.000000	-0.056889
Season_encoded	-0.063484	-0.056889	1.000000

Table 4 shows the results of the correlation matrix analysis where there is a very strong positive relationship between the variables "Number of Transactions" and "Amount of Water (liters)" with a correlation value of 0.851, which indicates a significant linear relationship. This finding is in line with previous studies which confirmed that an increase in transaction activity is often associated with an increase in resource consumption [19]. On the other hand, the very low relationship between "Number of Transactions" and "Season_encoded" (-0.063) as well as between "Amount of Water (liters)" and "Season_encoded" (-0.057) indicates that seasonal factors do not have a significant impact on transaction patterns or water consumption. This supports the assumption that factors more related to human activities have a greater influence in determining water demand than external conditions such as seasonality [20].

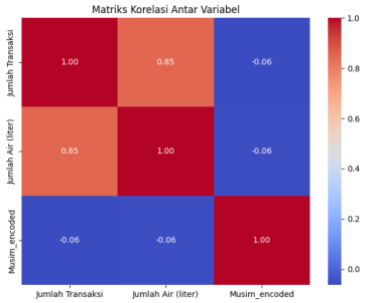

Fig. 4: Correlation Matrix between Variables

Figure 4 displays a visualization in the form of a correlation heatmap that corroborates the results of the analysis by displaying a very strong relationship between "Total Transactions" and "Total Water (liters)" with a correlation value of 0.85, which is consistent with previous research that reveals that transaction data is often used as a leading indicator in projecting resource demand [21]. This visualization also highlights the weakness of the relationship with the seasonality variable, which supports the conclusion that seasonality does not significantly affect consumption patterns.

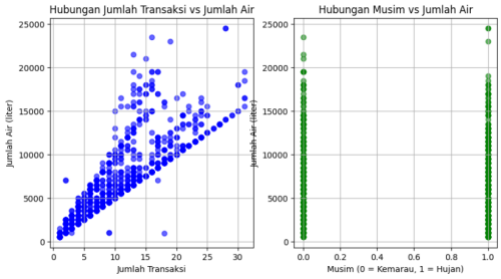

Fig. 5: Visualization of the Relationship between X and Y Variables

The presentation of the data in the form of a scatter plot between "Number of Transactions" and "Amount of Water (liters)" shows a consistent positive relationship pattern, although there are slight variations in the distribution of the data. This is consistent with the literature which states that transactions have a linear relationship with consumption levels in various economic and social contexts (Kristianto & Rudianto, 2020). In contrast, the scatter plot showing the relationship between "Season_encoded" and "Amount of Water (liters)" shows a relatively even distribution of data in both seasons, indicating that water demand tends to be stable despite external factors such as seasonality, consistent with findings in a previous study [22].

Table 5: Prediction Results

Name	Month_encoded	Year	Water amount predicted
Haris	1	2025	8047.445906
A Emung	1	2025	4082.395764
…	…	…	…
Yeyet	1	2025	3075.788607
Yoyo	1	2025	11865.187685

Table 5 shows the prediction table that presents the estimated water demand for January 2025. The results of this prediction are in line with previous research that utilizes a data-driven approach in estimating resource requirements [23].

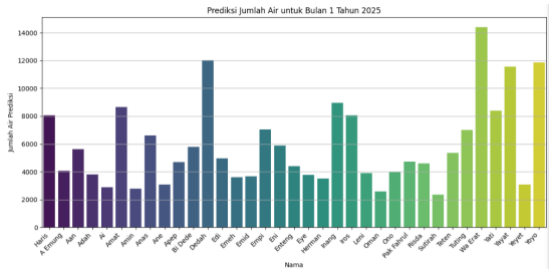

Fig. 6: Visualization of Prediction Results

Figure 6 displays the prediction that customers such as “Wa Erat” and “Dedah” fall into the category with the highest water demand, while customers such as “Sutirah” and “Oman” are predicted to have lower demand. This information can support more efficient water resource distribution planning. Visualization of the prediction results displayed in the form of bar charts shows the distribution of water demand among customers. This visualization not only strengthens the quantitative analysis, but also provides additional insights that can be useful for operational water resources management [24].

4. Conclusion

This study utilizes a linear regression algorithm to estimate raw water demand by considering two main variables: number of transactions and seasonality. The dataset used is divided into two parts, 80% for training data and 20% for testing data. The evaluation results show that the model performs adequately, with Mean Absolute Error (MAE) of 0.06, Mean Squared Error (MSE) of 0.01, Root Mean Squared Error (RMSE) of 0.08, and coefficient of determination (R^2) of 0.77, indicating that the model is able to explain 77% of the variation in the data. As an illustration, for an actual water demand of 7,000 liters, the model predicts 7,984.70 liters. Therefore, further development is required, for example through the addition of more relevant variables or the application of more complex data processing methods. Further analysis revealed that the number of transactions variable had the most significant influence on the prediction of raw water demand, with a coefficient of 16,940.46, meaning that each increase of one transaction unit contributed to an increase in water demand of 16,940.46 liters. On the other hand, the season variable has a relatively small influence, with a coefficient of -23.92, indicating that a change in season from dry to rainy only decreases the predicted water demand by 23.92 liters. The model intercept of 1,172.43 indicates the initial estimate of water demand when all independent variables are zero. The results of the correlation analysis support this finding, with a very strong positive relationship (r = 0.85) between the number of transactions and water quantity, while the relationship between seasonality and water quantity is very weak (r = -0.06). The scatter plot visualization shows a linear relationship between the number of transactions and water demand, while the seasonality variable does not show a significant pattern. This linear regression model can also be used to project future raw water demand based on historical data patterns. For example, a customer with a high number of transactions such as “Wa Erat” is predicted to require 14,393.98 liters of water in January 2025. Such predictions can provide strategic insights for companies in planning for more efficient distribution and management of water resources, so the linear regression algorithm can be considered as an effective predictive tool in supporting strategic decision-making regarding water demand.

References

[1] A. Adri, N. D. Rumlaklak, and D. R. Sina, “Implementasi Algoritma Apriori Untuk Analisa Data Penjualan (Studi Kasus: Toko Ud. Suryani),” *J. Komput. dan Inform.*, vol. 9, no. 2, pp. 182–188, 2021, doi: 10.35508/jicon.v9i2.5132.

[2] S. Alfarisi, R. Astuti, and F. M Basysyar, “Implementasi Data Minning Menggunakan Algoritma Regresi Linear Untuk Prediksi Penjualan Dan Cashflow Di Ayam Geprek Cap Cangkir,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 3, pp. 3392–3395, 2024, doi: 10.36040/jati.v8i3.9717.

[3] O. J. Ababil, S. A. Wibowo, and H. Zulfia Zahro’, “Penerapan Metode Regresi Linier Dalam Prediksi Penjualan Liquid Vape Di Toko Vapor Pandaan Berbasis Website,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 1, pp. 186–195, 2022, doi: 10.36040/jati.v6i1.4537.

[4] G. N. Ayuni and D. Fitrianah, “Penerapan Metode Regresi Linear Untuk Prediksi Penjualan Properti pada PT XYZ,” *J. Telemat.*, vol. 14, no. 2, pp. 79–86, 2020, doi: 10.61769/telematika.v14i2.321.

[5] Y. Diana *et al.*, “Analisa Penjualan Menggunakan Algoritma K-Medoids Untuk Mengoptimalkan Penjualan Barang,” *JOISIE J. Inf. Syst. Informatics Eng.*, vol. 7, no. 1, pp. 97–103, 2023.

[6] H. M. Fikri, I. Permana, R. M. Mundzir, and Megawati, “Prediksi Jumlah Pendaftar Jemaah Umrah Menggunakan Backpropagation dan Regresi Linear pada PT. Hajar Aswad Mubaroq,” vol. 4, no. October, pp. 1209–1217, 2024.

[7] Hendra Di Kesuma, D. Apriadi, H. Juliansa, and E. Etriyanti, “Implementasi Data Mining Prediksi Mahasiswa Baru Menggunakan Algoritma Regresi Linear Berganda,” *J. Ilm. Bin. STMIK Bina Nusant. Jaya Lubuklinggau*, vol. 4, no. 2, pp. 62–66, 2022, doi: 10.52303/jb.v4i2.74.

[8] S. Herdyansyah, E. H. Hermaliani, L. Kurniawati, and S. R. Sri Rahayu, “Analisa Metode Association Rule Menggunakan Algoritma Fp-Growth Terhadap Data Penjualan (Study Kasus Toko Berkah),” *J. Khatulistiwa Inform.*, vol. 8, no. 2, pp. 127–133, 2020, doi: 10.31294/jki.v8i2.9277.

[9] N. Hidayat and T. Wahyudi, “Prediksi Jasa Pengiriman Barang Top Trend Logistik Menggunakan Algoritma Regresi Linear pada PT. XNH Nurhikmah,” vol. 4, no. October, pp. 1448–1455, 2024.

[10] A. N. Afiati *et al.*, “Implementasi Algoritma Regresi Linear dalam Prediksi Persediaan Voucher di Raffa Cell Sukabumi,” vol. 8, no. 5, pp. 10801–10808, 2024.

[11] A. Sugiyarta, Sumiati, and H. Maulana, “Implementasi Data Mining Pola Penjualan Dengan Pendekatan Regresi Linear,” *JSiI (Jurnal Sist. Informasi)*, vol. 11, no. 1, pp. 54–61, 2024, doi: 10.30656/jsii.v11i1.8411.

[12] F. A. Bilawa, H. Hikmayanti, S. T. Informatika, F. I. Komputer, and U. B. Perjuangan, “Prediksi Harga Beras Medium Di Indonesia Dengan Membandingkan Metode Regresi Linear Dan Regresi Polinomial,” *Ris. Sist. Inf. Dan Tek. Inform.*, vol. 9, pp. 774–787, 2024.

[13] A. Pradita and Rasiban, “Implementasi Data Mining dengan Metode Regresi Linear untuk Prediksi Hasil Penjualan di PT Awitama Cyndo,” vol. 5, no. 3, pp. 2709–2723, 2024.

[14] A. Kurniadi Hermawan, A. Nugroho, and Edora, “Analisa Data Mining Untuk Prediksi Penyakit Ginjal Kronik Dengan Algoritma Regresi Linier,” *Bull. Inf. Technol.*, vol. 4, no. 1, pp. 37–48, 2023, doi: 10.47065/bit.v4i1.475.

[15] M. J. C. Sianturi, D. M. Sinaga, B. S. N. Sembiring, and E. Ginting, “Metode Regresi Linear Berganda dalam Prediksi Penjualan Produk Berbasis Web,” vol. 16, no. 1, pp. 236–242, 2024.

[16] Miftahuljannah, A. S. Sunge, and A. T. Zy, “Analisis Prediksi Penjualan Dengan Metode Regresi Linear Di Pt. Eagle Industry Indonesia,” *J. Inform.*

Teknol. dan Sains, vol. 5, no. 3, pp. 398–403, 2023, doi: 10.51401/jinteks.v5i3.3325.
[17] T. K. Tarumingkeng, A. Jacobus, W. Patty, and M. Sciences, "Prediksi Hasil Tangkapan Ikan Cakalang dengan Metode Regresi Linear dan Recurrent Neural Network," vol. 19, no. 03, pp. 229–238, 2024.
[18] H. A. Wibowo, K. Faisal, and Y. Devianto, "Analisa Dan Visualisasi Data Penjualan Menggunakan Exploratory Data Analysis Pada PT. Telkominfra," *JATISI (Jurnal Tek. Inform. dan Sist. Informasi)*, vol. 9, no. 3, pp. 2292–2304, 2022, doi: 10.35957/jatisi.v9i3.2737.
[19] E. Triyanto, H. Sismoro, and A. D. Laksito, "Implementasi Algoritma Regresi Linear Berganda Untuk Memprediksi Produksi Padi Di Kabupaten Bantul," *Rabit J. Teknol. dan Sist. Inf. Univrab*, vol. 4, no. 2, pp. 66–75, 2019, doi: 10.36341/rabit.v4i2.666.
[20] A. G. Purwaningsih and Nurhadi, "Pengaruh Promosi Penjualan Dan Gender Terhadap Perilaku Impulse Buying Pada E-Commerce Shopee," *J. Ilm. STIE MDP*, vol. 10, no. 2, pp. 159–167, 2021.
[21] Q. Yumansyah Qori, A. Turmudi Zy, and M. Fatchan, "Prediksi Jumlah Kasus Klaim Indemnity Dengan Menggunakan Algoritma Regresi Linear Pada Asuransi Mandiri Inhealth," *Bull. Inf. Technol.*, vol. 4, no. 3, pp. 299–305, 2023, doi: 10.47065/bit.v4i3.733.
[22] D. S. O. Panggabean, E. Buulolo, and N. Silalahi, "Penerapan Data Mining Untuk Memprediksi Pemesanan Bibit Pohon Dengan Regresi Linear Berganda," *JURIKOM (Jurnal Ris. Komputer)*, vol. 7, no. 1, p. 56, 2020, doi: 10.30865/jurikom.v7i1.1947.
[23] G. Khalda Rifdan, N. Rahaningsih, A. Bahtiar, I. Ali, and N. Dienwati Nuris, "Ramalan Penjualan Rumah Menggunakan Algoritma Linear Regresi Di Tebet Jakarta Selatan," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 8, no. 2, pp. 1847–1851, 2024, doi: 10.36040/jati.v8i2.9022.
[24] M. Sholeh, E. K. Nurnawati, and U. Lestari, "Penerapan Data Mining dengan Metode Regresi Linear untuk Memprediksi Data Nilai Hasil Ujian Menggunakan RapidMiner," *JISKA (Jurnal Inform. Sunan Kalijaga)*, vol. 8, no. 1, pp. 10–21, 2023, doi: 10.14421/jiska.2023.8.1.10-21.