International Journal of Information Technology and Computer Science Applications (IJITCSA)
p-ISSN: 2964-3139 e-ISSN: 2985-5330
Vol. 04, No. 02, page 115 - 124
Submitted 14/7/2025; Accepted 8/6/2026; Published 23/6/2026

Clustering and Sales Prediction Using K-Means and Simple Linear Regression

[1]Tia Aulia [2]Wowon Priatna [3]Muhammad Yasir
[1]Informatics, Universitas Bhayangkara Jakarta Raya, Bekasi, INDONESIA
[2]Information Department, Universitas Bhayangkara Jakarta Raya, Bekasi, INDONESIA
e-mail : [1]202110715185@mhs.ubharajaya.ac.id, [2]wowon.priatna@dsn.ubharajaya.ac.id, [2]muhammad.yasir@dsn.ubharajaya.ac.id

Corresponding Autor: Wowon Priatna

Abstract
CV. Cipta Usaha Selaras faces challenges in identifying customer purchasing patterns and accurately projecting sales values. The importance of this research lies in the company's need for data-driven marketing strategies and efficient operational planning. This study employs the K-Means algorithm to cluster customers based on purchase frequency and total transaction value, as well as Simple Linear Regression to predict total purchases based on transaction frequency. The data analyzed consists of 358 sales transaction entries from the year 2024. The clustering results reveal three customer segments with distinct characteristics, with a Silhouette Score of 0.7913, indicating good segmentation quality. The regression model produced an equation with a coefficient of determination (R^2) of 0.6910, a MAE of IDR 213 million, and a MSE of IDR 206 trillion. These results indicate that the applied approach provides a reasonably representative overview of customer purchasing behavior. This research offers a significant contribution to data-driven decision-making within the company, particularly in the development of marketing strategies and estimation of potential revenue.

Keywords—Clustering, K-Means, Sales Prediction, Simple Linear Regression, Customer Segmentation

1 Introduction
CV. Cipta Usaha Selaras is a company engaged in the production of jumbo bags for industrial needs such as chemicals, silica sand, coal, and wood pellets. In its business operations, the company faces challenges including sales fluctuations and difficulties in identifying customer purchasing patterns. Irregular transactions, both from regular and non-regular customers, hinder the effectiveness of production planning and marketing strategy development.

In its business practices, the company struggles with inconsistent purchase behavior and challenges in recognizing transactional patterns. The inconsistency of purchases regardless of whether they come from loyal or occasional customers makes it difficult to formulate precise production and marketing strategies. Consequently, the company's decision-making process becomes inefficient, as it is reactive to incoming demand rather than based on measurable and predictive patterns.

One potential solution to address this issue is the implementation of data-driven analytical approaches (data mining), which can help the company better understand customer behavior and predict future purchasing values [1]. The K-Means method is used to cluster customers based on purchasing data, while the Simple Linear Regression method is used to estimate total purchases based on transaction frequency [2].

Previous studies have supported the application of these approaches. For instance, [3] successfully differentiated between high- and low-purchasing customers through sales data clustering. Study [4] applied linear regression to forecast purchasing trends at Toko 99 and achieved good accuracy results based on MAPE values. Meanwhile, [5] combined K-Means and regression methods in a network analysis, demonstrating that the integration of both methods leads to more accurate analytical outcomes. K-Means is considered advantageous because it can cluster customers

without requiring strict assumptions about data distribution or specific purchasing patterns, making it suitable for highly variable data contexts [6].

Based on these considerations, this study aims to develop a clustering and purchasing value prediction model for CV. Cipta Usaha Selaras using 2024 customer transaction data as the basis for analysis. The model employs the K-Means algorithm and Simple Linear Regression. It is expected to assist the company in making more accurate and data-driven decisions, particularly in formulating production and marketing strategies within the manufacturing sector.

2 Research methods

This study adopts the CRISP-DM (Cross Industry Standard Process for Data Mining) approach as the primary methodology. CRISP-DM is an industry-standard framework for data processing that is iterative and flexible, consisting of six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. This approach was chosen because it ensures that the analytical process is conducted systematically, enabling the development of accurate models that can support data-driven decision-making [7][8].

2.1 Business Understanding

This phase focuses on understanding customer needs and business objectives. It involves defining what the business aims to achieve, assessing available resources, planning data collection methods, and developing a comprehensive project plan.

2.2 *Data Understanding*

This phase involves the process of collecting and conducting an initial analysis of the required data. Activities in this stage include gathering raw data, describing its characteristics, performing deeper data exploration, and evaluating the quality of the available data.

2.3 *Data Preparation*

This phase focuses on preparing the data for the modeling process. The data is cleaned and adjusted to meet the quality standards required for analytical processing. The activities carried out include:

- Data cleaning: Removing duplicate entries and records with missing values to ensure optimal data quality.
- Normalization**:** Applying the Min-Max Scaling technique to the attributes *Qty PO* and *Total Price* to bring them to the same scale, thereby preventing certain variables from dominating the clustering process.
- Feature engineering: Creating a new attribute representing the purchase frequency per customer, calculated based on the number of transactions, to be used as a predictor variable in the regression model..
- Attribute selection: Selecting relevant attributes Qty PO, Total Price, and Purchase Frequency—for use in both the clustering and prediction processes.

2.4 *Modeling*

This study builds two main models: K-Means Clustering and Simple Linear Regression. K-Means is an unsupervised learning algorithm designed to group data into a specified number of clusters based on the similarity between data points. The algorithm works by minimizing the Within-Cluster Sum of Squares (WCSS), which is the sum of squared distances between each data point and its corresponding cluster centroid. The process begins by defining the number of clusters (k), after which the initial centroids are randomly selected. Each data point is then assigned to the nearest centroid using the Euclidean Distance formula, as follows [9]:

$$d = \sqrt{(x_2 - \mu_x)^2 + (y_2 - \mu_y)^2} \tag{1}$$

The values of x_2 and y_2 represent customer data attributes (such as order quantity and total price), while μ_x and μ_y are the coordinates of the cluster centroid for each attribute. The result of this calculation yields d, the distance between a customer data point and the cluster center. The smaller the value of d, the closer the data point is to the centroid, and thus it will be assigned to the corresponding cluster. This model is used to segment customers into three groups high, medium, and low based on their purchasing patterns.

The Simple Linear Regression model is used to represent a linear relationship between one independent variable and one dependent variable[10]. In this study, it is applied to predict the Total Price or customer purchase value based on the frequency of transactions. By using this approach, the company can estimate the potential purchase value in the future, enabling more proactive and data-driven planning for production and marketing strategies The general form of the linear regression equation used is:

$$Y = a + bX \tag{2}$$

In this equation, the variable Y represents the Total Price as the predicted purchase value, while X denotes the Purchase Frequency or the number of transactions made by the customer. The value a is the constant (intercept), which indicates the predicted purchase value when the purchase frequency is zero. The coefficient b is the regression slope, representing how much the purchase value is expected to change with each additional transaction. This method is chosen for its simplicity in interpretation and its ability to directly explain the linear relationship between variables. A positive relationship between X and Y indicates that the more frequently a customer makes transactions, the greater their potential purchase value.

2.5 *Evaluation*

Model evaluation was conducted on the clustering and prediction results using a quantitative metric-based approach. For the K-Means model, the quality of segmentation was assessed using the Silhouette Score, a metric that measures how well each data point fits within its assigned cluster. This score ranges from –1 to 1, where values closer to 1 indicate that a data point is well matched to its own cluster and poorly matched to neighboring clusters. The Silhouette Score is calculated using the following formula [12] :

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))} \tag{3}$$

Where$a(i)$ is the average distance between data point *i* and all other points within the same cluster, and $b(i)$ is the average distance between data point i and all points in the nearest neighboring cluster. A higher Silhouette Score indicates better clustering quality and clearer separation between clusters.

Meanwhile, the Simple Linear Regression model is evaluated using three primary metrics: R-Squared (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE). R^2 measures the proportion of variance in the target variable (Total Price) that can be explained by the predictor variable (Purchase Frequency)[13]. The formula for R^2 is:

$$R^2 = 1 - \frac{\sum(y_i + \hat{y}_i)^2}{\sum(y_i + \bar{y}_i)^2} \tag{4}$$

Where y_i is the actual value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ adalah rata-rata nilai aktual. The closer the R^2 value is to 1, the better the model explains the variability of the data. MAE (Mean Absolute Error) is used to measure the average absolute error between actual and predicted values, and is calculated using the following formula:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i + \hat{y}_i| \tag{5}$$

Meanwhile, MSE (Mean Squared Error) calculates the average of the squared differences between the actual and predicted values. It is defined by the following formula:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i + \hat{y}_i)^2 \tag{6}$$

MSE is more sensitive to outliers than MAE because the error values are squared. These three metrics provide an overview of how accurately the model predicts customer purchase values and help identify deviations or prediction errors that may occur.

2.6. Deployment

At this stage, the clustering and prediction results are presented in the form of visualizations to facilitate interpretation. These visualizations help the company understand purchasing patterns and estimate the potential value of customer transactions. The developed model has not yet been implemented directly into the operational system; rather, it is used as a preliminary analytical tool in the form of visual reports that support decision-making processes.

3 Results and Discussion

The presentation of results in this study follows the CRISP-DM framework, starting from business understanding to model deployment. The discussion is structured systematically according to the analytical process flow.

3.1. Business Understanding

CV. Cipta Usaha Selaras is a manufacturing company that produces jumbo bags for various industrial needs, such as chemicals, silica sand, coal, and wood pellets. In its operations, the company faces challenges related to irregular purchasing patterns from both regular and non-regular customers. These fluctuations make it difficult for the company to develop effective production and marketing strategies. In addition, the company does not yet have a predictive system capable of estimating customer purchasing potential based on existing historical data.

Based on these problems, this study aims to develop a data-driven analytical model to support the decision-making process. The objective is to generate customer segmentation using the K-Means algorithm and to predict purchase value using the Simple Linear Regression method. With this model, the company is expected to gain a more structured understanding of customer behavior and formulate more targeted business strategies.

3.2. Data Understanding

This stage is carried out to thoroughly understand the structure and characteristics of the data used in the study. The data analyzed consists of sales transaction records from CV. Cipta Usaha Selaras for the year 2024, comprising 358 entries. Each entry includes key information such as customer name, transaction date, order quantity (Qty PO), and total transaction value (Total Price). The initial analysis involved identifying the data types of each attribute, checking for completeness, and detecting any extreme values (outliers) that could affect the accuracy of the analysis.

Additionally, a new attribute was created to represent the purchase frequency per customer, calculated based on the number of transactions during the data period. This variable was then used as one of the main inputs for predicting purchase value in the modeling stage.

3.3. Data Preparation

This stage is carried out to ensure that the data used for modeling is clean, consistent, and ready for analysis. The first step is data cleaning, which involves removing duplicate entries and missing values to maintain data quality and prevent distortion in the model results. Next, normalization is performed using the Min-Max Scaling technique on the Qty PO and Total Price attributes. The goal is to bring both variables to the same scale, preventing one from dominating the other during the clustering process using the K-Means algorithm.

In addition, feature engineering is conducted by creating a new attribute representing purchase frequency per customer, calculated based on the number of transactions in the dataset. This feature is used as the independent variable in the purchase value prediction process using linear regression. Finally, attribute selection is performed by choosing the three most relevant variables: Qty PO, Total Price, and Purchase Frequency, which are then used in the modeling stage.

3.4. Modeling

3.4.1. K-Means Clustering

At this stage, the K-Means algorithm is used to group customers based on two main attributes: Total Purchase and Purchase Frequency. The objective of this clustering process is to categorize customers according to their transaction patterns, allowing the company to apply differentiated strategies for each segment. The clustering process

resulted in three clusters, labeled C1 (high), C2 (medium), and C3 (low). Table 1 below presents a portion of the customer segmentation results based on the output of the K-Means model:

Table 1 Final Results of K-Means Clustering

No	Customer	Total Purchase	Purchase Frequency	Cluster	Segment
1	Bpk. Erza Raharjo	2.300.000	1	C3	Low
2	CV. Mitra Bersama Anugrah	135.000.000	2	C3	Low
3	CV. Mitra Sarana Anugrah	83.700.000	1	C3	Low
4	CV. Sumber Rejeki	40.500.000	1	C3	Low
5	Graha Bintang Metalindo	5.500.000	1	C3	Low
6	MG Industrial Supplier	5.040.000	1	C3	Low
7	Pak Yogi	7.800.000	1	C3	Low
8	PT. Anoa Bintang Metalindo	2.230.000	1	C3	Low
9	PT. Arisa Pratama Abadi	2.730.000	1	C3	Low
10	PT. ARMB Energi Indonesia	47.450.000	1	C3	Low
…	…	…	…	…	…
48	PT. Sumber Wahana Sejati	1.054.000.000	15	C2	Medium
49	PT. Tsuchiyoshi Hoasana Indonesia	74.500.000	7	C3	Low
50	PT. Visi Energi Sukses	74.100.000	5	C3	Low

The clustering results using the K-Means algorithm produced three customer segments: High, Medium, and Low, based on total purchases and transaction frequency. The majority of customers fall into the Low segment, indicating low purchase frequency and small transaction values. The Medium segment includes customers with relatively stable purchasing activity, while the High segment consists of customers with very high transaction values and frequencies. This segmentation provides a clear picture of customer profiles and can serve as a foundation for formulating more targeted marketing and production strategies.

3.4.2. Purchase Prediction Using Simple Linear Regression

After customer segmentation is completed, the next step is to develop a predictive model using the Simple Linear Regression method. This model aims to estimate the Total Purchase value of each customer based on the Purchase Frequency variable. This approach is intended to help the company project the potential transaction value of individual customers in the future, thereby serving as a basis for more structured production and marketing strategy planning [4]. From the model training process, a simple linear regression equation was obtained, as shown in Equation (2) below:

$$Y = -2.688.402.807{,}94 + 372.258.401{,}52 \times Frekuensi$$

In the equation, Y represents the predicted Total Purchase, while the Frequency variable indicates the number of transactions made by the customer. The negative intercept implies that when the transaction frequency approaches zero, the predicted total purchase is also very low potentially to the point of having no economic value. The regression coefficient, valued at Rp 372,258,401.52, indicates that each additional transaction is estimated to increase the total purchase by that amount.

The results of this model are compared with the actual data in Table 2 to evaluate how closely the predictions align with the actual purchase values. Model evaluation is conducted using metrics such as R-Squared and MAE to assess the accuracy and effectiveness of the predictions.

Table 2 Predicted Total Purchase Based on Transaction Frequency

No	Purchase Frequency	Total Purchase (Actual)	Predicted Total Purchase
1	1	2.300.000	58.415.377
2	2	135.000.000	110.085.800
3	1	83.700.000	58.415.377
4	1	40.500.000	58.415.377
5	1	5.500.000	58.415.377
6	1	5.040.000	58.415.377
7	1	7.800.000	58.415.377
8	1	2.230.000	58.415.377
9	1	2.730.000	58.415.377
10	1	47.450.000	58.415.377
…	…	…	…
48	15	1.054.000.000	781.801.298
49	7	74.500.000	368.437.915
50	5	74.100.000	265.097.069

The prediction results show that the total purchase value increases as customer purchase frequency rises. Although there are differences between the actual and predicted values, the model provides a rough estimate that can be used as a basis for grouping customers according to their revenue-generating potential.

3.5. Evaluation

3.5.1. Evaluation of the K-Means Model

The evaluation of the K-Means model was conducted using the Silhouette Score, a metric that measures how similar a data point is to its own cluster compared to other clusters. The Silhouette Score ranges from –1 to 1, with higher values indicating better separation between clusters [12]. The evaluation results show that the average Silhouette Score for the entire clustering model is 0.7913, indicating a well-structured segmentation with clearly separated clusters. This suggests that the K-Means model has successfully grouped customers based on Total Purchase and Purchase Frequency with meaningful and solid cluster structures. Table 3 presents the detailed evaluation results for each individual cluster.

Table 3 Average Silhouette Score for Each Cluster

Cluster	Segment	Number of Customers	Average Silhouette Score	Cluster Quality	Interpretation
C3	Low	4	0.3038	Fair	Overlap exists between clusters
C2	Medium	44	0.8843	Good	Clusters are dense and well-separated
C1	High	4	0.4004	Very Good	Clusters are reasonably well-defined

Table 3 shows that Cluster C2 has the highest quality, with the highest Silhouette Score value of 0.8843, indicating a clear segmentation. Clusters C1 and C3 have lower values, suggesting a slight overlap between clusters. Overall, the average score of 0.7913 indicates that the clustering model is effective and can be used as a reliable basis for customer segmentation.

3.5.2. Evaluasi Model Regresi Linear

To evaluate the performance of the Simple Linear Regression model, three evaluation metrics are used: R-Squared (R^2), Mean Absolute Error (MAE), and Mean Squared Error (MSE). The evaluation results are presented in Table 4 below:

Table 4 Evaluation Results of the Simple Linear Regression Model

Evaluation Metric	Value
R-Squared (R^2)	0,6910
MAE	Rp213.183.958
MSE	Rp206.204.672.226.881.088

The R^2 value of 0.6910 indicates that the model is able to explain approximately 69.10% of the variation in total purchase value based on transaction frequency. The MAE of Rp213 million is still acceptable within the industrial context, although the high MSE suggests the presence of extreme values (outliers) influencing the model. Overall, the model is reasonably effective as an initial estimation tool, and its performance can be further improved by incorporating additional variables or applying more advanced prediction methods.

3.6. Deployment

3.6.1. K-Means Model Visualization

The K-Means algorithm was applied to the normalized Qty PO and Total Price data. The value of K = 3 was selected based on the need to segment customers into three distinct groups. The clustering results are presented in Table 1 and visualized in Figure 1 and Figure 2.

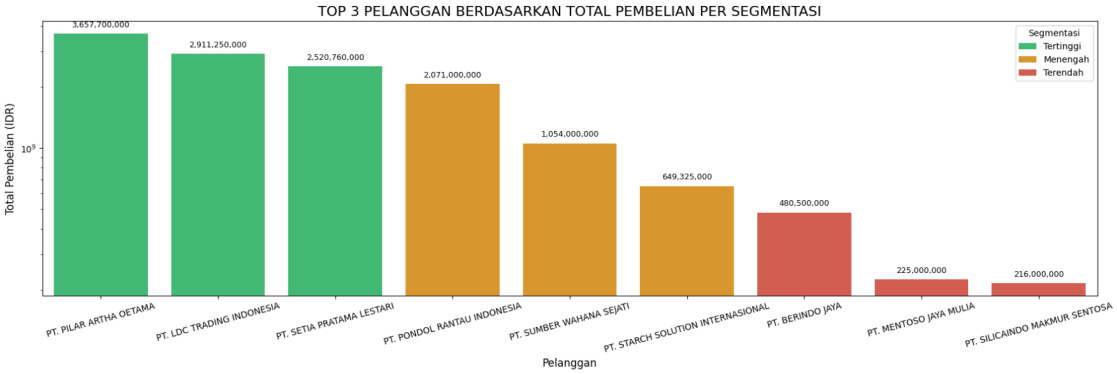

Figure 1 Top 3 Customers Based on Total Purchase per Segment

Figure 1 displays the top ten customers with the highest total purchases, categorized into three segments: High, Medium, and Low. In the High segment are PT. Pilar Artha Oetama (Rp3,587,700,000), PT. LDC Trading Indonesia (Rp2,911,250,000), and PT. Setia Pratama Lestari (Rp2,520,760,000). These three customers contribute the most significantly to the company's total revenue and should be considered top priorities in marketing and service strategies. The Medium segment includes PT. Pondok Rantau Indonesia (Rp2,071,000,000) and PT. Sumber Wahana Sejati (Rp1,054,000,000). Meanwhile, the Low segment consists of customers with purchase values below Rp1 billion.

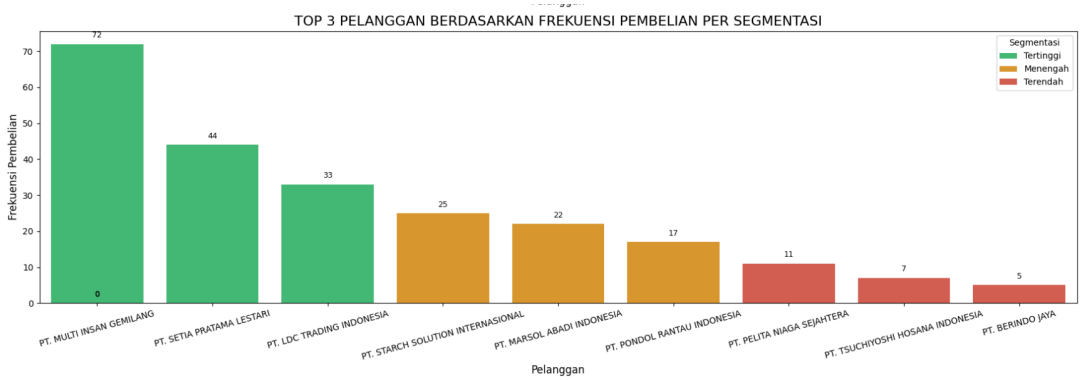

Figure 2 Top 3 Customers Based on Purchase Frequency per Segment

Figure 2 complements the previous analysis by presenting the purchase frequency data of the top ten customers. The customer with the highest number of transactions is PT. Multi Insan Gemilang (72 transactions), followed by PT. Setia Pratama Lestari (44 transactions) and PT. LDC Trading Indonesia (33 transactions). Interestingly, PT. Pilar Artha Oetama, despite recording the highest total purchase value, does not rank among the top three in terms of purchase frequency. This indicates that PT. Pilar Artha Oetama tends to make large but infrequent transactions. In

contrast, PT. Multi Insan Gemilang exhibits very high transaction frequency, albeit with relatively smaller purchase values per transaction.

Based on the analysis of total purchase and transaction frequency, it can be concluded that the company has two primary customer segments: customers with high purchase value but low transaction frequency, and customers with high transaction frequency but relatively low purchase value per transaction. This segmentation provides deeper insight into customer behavior characteristics, enabling the company to design more targeted marketing strategies, both to improve customer retention and to maximize revenue contribution from each segment.

3.6.2. Visualization of the Simple Linear Regression Model

Following the segmentation process, a prediction model for Total Purchase Value was developed using Simple Linear Regression, based on purchase frequency. The analysis results indicate a positive relationship between transaction frequency and purchase value. The derived regression equation is: $Y = -2.688.402.807{,}94 + 372.258.401{,}52\ X$. Where: Y: Total Purchase Value (in IDR) X: Purchase Frequency.

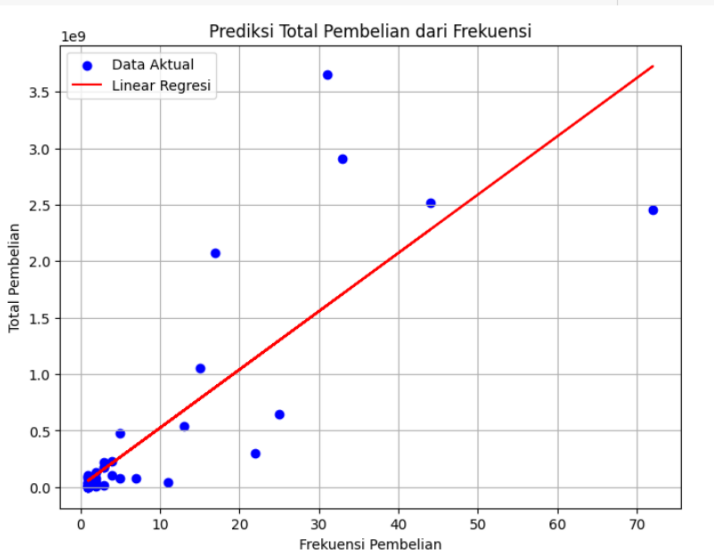

Figure 3 Graph of Total Purchase Prediction Based on Transaction Frequency

Figure 3 illustrates the relationship between purchase frequency and total purchase value based on the results of the simple linear regression model. The blue dots on the graph represent the actual data, while the red line depicts the model's predicted values. A positive relationship can be observed, indicating that the more frequently a customer makes transactions, the higher their total purchase value tends to be. Most of the data points are concentrated in the low-frequency area, which aligns with the clustering results that show the majority of customers belong to the low-activity segment.

However, there are a few outliers that deviate significantly from the regression line, suggesting that some customers' total purchases either fall below or exceed the model's general trend. Overall, this model provides a reasonably good initial estimate of purchase value based on transaction frequency, though there is still room for improvement either through the use of more complex prediction methods or by incorporating additional relevant variables.

3.7. Discussion

The results of this study indicate that the K-Means method effectively grouped customers into three distinct segments. This segmentation benefits the company by enabling more targeted marketing strategies—for example, offering special promotions to high-value customers or implementing loyalty programs for mid-tier customers.

Meanwhile, the Simple Linear Regression model provides an initial overview of customer purchase potential based on transaction intensity. Despite the presence of some outliers, the model still delivers reasonably accurate estimates, particularly for customers with low to medium transaction frequencies.

4 Conclusion

This study successfully developed two analytical models for CV. Cipta Usaha Selaras: the K-Means algorithm for customer segmentation and Simple Linear Regression for predicting purchase value. The K-Means model effectively grouped customers into three segments based on total purchase and transaction frequency, with a

Silhouette Score of 0.7913, indicating good segmentation quality. Meanwhile, the Simple Linear Regression model achieved an R^2 value of 0.6910, which is sufficiently effective in explaining the relationship between transaction intensity and total purchase value.

These findings are consistent with previous studies. Research by [3] demonstrated that clustering methods are effective in distinguishing customers based on their purchase levels. Similarly, the study in [4] confirmed that linear regression can predict purchasing trends with a reasonable degree of accuracy. The key distinction of this research lies in the combined application of both methods within a single study, as well as the direct implementation on actual data from a manufacturing company, thereby providing practical value for data-driven decision-making processes.

Nevertheless, this study has several limitations. The prediction model is still influenced by outliers, resulting in a relatively high MSE. The use of a single predictor variable (purchase frequency) limits estimation accuracy. The clustering segmentation is not yet adaptive to future changes in customer behavior. The dataset is limited to transactions from the year 2024, and the models have not yet been deployed into the company's operational systems.

5 Suggestion

This research still has room for further development. It is recommended to incorporate additional predictor variables such as product type, customer category, or transaction timing to improve the predictive model's accuracy. Moreover, the application of more advanced modeling algorithms, such as Random Forest, can be explored as a benchmark to evaluate model performance more comprehensively. Integrating the model into the company's real-time information system is also advised, in order to support more responsive, data-driven, and measurable decision-making processes based on actual transactional data.

6 Acknowledgments

The author would like to express sincere gratitude to CV. Cipta Usaha Selaras for granting permission to use the sales transaction data in this study. Special thanks are also extended to Mr. Wakhid, Production Coordinator, for his assistance and the valuable information provided during the data collection process. Appreciation is also given to the academic advisor for their guidance and direction, as well as to all parties who have provided moral and technical support, enabling this research to be successfully completed.

.

BIBLIOGRAPHY

[1] A. W. Zunan Setiawan, Muhammad Fajar , Arif Mudi Priyatno, Anggi Yhurinda Perdana Putri, Mediana Aryuni , Siti Yuliyanti, Harya Widiputra , Budanis Dwi Meilani, Rohmat Nur Ibrahim, Rezania Agramanisti Azdy, Satrio Junaidi, *BUKU AJAR DATA MINING*. PT. Sonpedia Publishing Indonesia, 2023. [Online]. Available: https://www.google.co.id/books/edition/BUKU_AJAR_DATA_MINING/1nLVEAAAQBAJ?hl=id&gbpv=1

[2] I. Safira, R. Salkiawati, and W. Priatna, "Penerapan Algoritma K-Means untuk Mengetahui Pola Persediaan Barang pada Toko Raja Bekasi," *Journal of Informatic and Information Security*, vol. 3, no. 1, pp. 99–110, 2022, doi: 10.31599/jiforty.v3i1.1253.

[3] A. Nugraha, O. Nurdiawan, and G. Dwilestari, "Penerapan Data Mining Metode K-Means Clustering Untuk Analisa Penjualan Pada Toko Yana Sport," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 6, no. 2, pp. 849–855, 2022, doi: 10.36040/jati.v6i2.5755.

[4] P. A. Duran, A. V. Vitianingsih, M. S. Riza, A. L. Maukar, and S. F. A. Wati, "Data Mining Untuk Prediksi Penjualan Menggunakan Metode Simple Linear Regression," *Teknika*, vol. 13, no. 1, pp. 27–34, 2024, doi: 10.34148/teknika.v13i1.712.

[5] M. Yasir, F. Sinlae, and C. Author, "Penerapan Algoritma K-Means dan Linear Reggression Sederhana Dalam Klasterisasi Grafik Bandwidth," vol. 1, no. 4, pp. 150–158, 2023, [Online]. Available: https://creativecommons.org/licenses/by/4.0/

[6] U. Arfan and N. Paraga, "Perbandingan Algoritma K-Means, Naïve Bayes dan Decision Tree Dalam Memprediksi Penjualan Bahan Bakar Minyak," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 4, pp. 1379–1389, 2024, doi: 10.57152/malcom.v4i4.1566.

[7] B. Sutara, F. Vulture, and R. Novianti, "Application of K-Means algorithm with CRISP-DM method in student data analysis as a support for promotion strategy," *Side: Scientific Development …*, vol. 1, no. 1, pp. 1–7, 2024, [Online]. Available: https://ojs.arbain.co.id/index.php/side/article/view/6%0Ahttps://ojs.arbain.co.id/index.php/side/article/down

load/6/6
[8] E. Muningsih, I. Maryani, and V. R. Handayani, "Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa," *Jurnal Sains dan Manajemen*, vol. 9, no. 1, p. 96, 2021, [Online]. Available: www.bps.go.id
[9] R. Primartha, *Algoritma Machine Learning*. Bandung: Informatika Bandung, 2021.
[10] G. N. Ayuni and D. Fitrianah, "Penerapan Metode Regresi Linear Untuk Prediksi Penjualan Properti pada PT XYZ," *Jurnal Telematika*, vol. 14, no. 2, pp. 79–86, 2020, doi: 10.61769/telematika.v14i2.321.
[11] F. Ramdhani and K. Setiawan, "Penerapan Data Mining untuk Prediksi Pelanggan di PT. XYZ Menggunakan Algoritma Linear Regression," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, pp. 490–497, 2024, doi: 10.57152/malcom.v4i2.1217.
[12] M. Piao Tan and C. A. Floudas, "Determining the Optimal Number of Clusters," *Encyclopedia of Optimization*, vol. 1, pp. 687–694, 2023, doi: 10.1007/978-0-387-74759-0_123.
[13] A. Novalas *et al.*, "Analisis prediksi penjualan iklan media massa dan elektronik menggunakan metode linear regression," vol. 7, pp. 203–209, 2024.