DOI: https://doi.org/10.70656/ijcse.v2i01.499

Segmentation of Problematic Loan Customers Using the K-Means Clustering Algorithm to Support Strategic Decision-Making (Case Study: Bank Mega Finance Bengkulu)

Willi Novrian
Universitas Bengkulu
Bengkulu, Indonesia
willinovrian@unib.ac.id

Annisa Afriani
Universitas Bengkulu
Bengkulu Indonesia
Annisaafriani790@gmai.com

Julia Purnama Sari
Universitas Bengkulu
Bengkulu, Indonesia
Juliapurnamasari@unib.ac.id

Yusran Panca Putra
Universitas Bengkulu
Bengkulu Indonesia
Yusranpanca@unib.ac.id

***Abstract*—This study analyzes 5,305 records of non-performing loan customers from Bank Mega Finance Bengkulu using the K-Means Clustering algorithm within the CRISP-DM framework. Based on variables such as tenure, outstanding balance, installment amount, and payment delay duration, the analysis identified three customer risk clusters (high, medium, and low) with a Davies-Bouldin Index (DBI) of 0.201, indicating good clustering quality. The segmentation results can help determine collection priorities, loan restructuring, and risk mitigation strategies, demonstrating the effectiveness of data mining in supporting stra**

Keywords- Non-Performing Loan, K-Means Clustering, CRISP-DM, Data Mining, Davies-Bouldin Index (DBI), Customer Segmentation.

I. INTRODUCTION

The rapid growth of digital technology, particularly in Financial Technology (FinTech), has transformed financial services in Indonesia by enhancing transaction efficiency, expanding access, and reaching unbanked communities. In the banking sector, effective customer data management is crucial for understanding customer behavior and assessing credit risk. Customer segmentation based on payment characteristics plays a key role in minimizing non-performing loans (NPLs) that threaten financial stability.

This study applies the K-Means Clustering algorithm within a Data Mining framework to segment problematic customers of Bank Mega Finance Bengkulu based on payment behavior, including outstanding balance, delay duration, and payment history. Using the Davies-Bouldin Index (DBI) as an evaluation metric, cluster quality can be measured objectively. The results are expected to assist the bank in making strategic decisions for debt collection, loan restructuring, and credit risk mitigation more effectively and efficiently.

II. METHODOLOGY

A. Problematic Loan Customer

Problematic loans among customers can be caused by various factors, both internal and external. Internal factors may include weaknesses in the customer's business management, inability to make payments, or even errors by the bank in establishing credit policies and supervision. Meanwhile, external factors may involve macroeconomic conditions, natural disasters, or other unforeseen events beyond the customer's control, which ultimately hinder their ability to fulfill their obligations [7].

The resolution of problematic loans typically involves several restructuring methods, such as rescheduling, reconditioning, restructuring, and liquidation of collateral. The choice of method is determined by considering the customer's good faith and the sustainability prospects of their business [8].

B. Data Mining

Data Mining is a process that involves the use of statistical, mathematical, artificial intelligence, and machine learning techniques to identify and extract useful information and knowledge from large databases. This process aims to discover relationships, patterns, and trends relevant to the data by analyzing large stored datasets using pattern recognition techniques such as statistical and mathematical methods [4]..

C. Clustering

Clustering is one of the main techniques in data mining and machine learning used to group a set of data into several groups (clusters) based on the similarity of their characteristics. In the clustering process, data with high similarity are grouped into the same cluster, while different data are placed in other clusters [9]. This method is a type of unsupervised learning, meaning that the data used do not have predefined labels or classes. The main goal of clustering is to increase homogeneity within each cluster (intra-cluster similarity) and maximize the differences between clusters

(inter-cluster dissimilarity), so that hidden patterns in the data can be naturally identified without human intervention [10].

D. Algoritma K-Means

K-Means is a non-hierarchical data clustering method that aims to divide data into one or more groups (clusters). In this process, data with similar characteristics are grouped together in the same cluster, while data with different characteristics are placed in other clusters. This method is based on the distance between data points and works by partitioning the data into a predetermined number of clusters. It can only be applied to attributes that are numerical in nature [11].

The steps in applying the K-Means clustering algorithm include the following:[12].

1. Prepare the dataset.
2. Determine the number of clusters (K).
3. Initialize the centroids by randomly selecting K points as the initial centroids.
4. Calculate the distance of each data point to the existing centroids and group the points based on the shortest distance, using the Euclidean distance formula:

$$d(x, y) = \sqrt{\sum n}\ (xi - yi)2 \quad (1)$$

Where (x, y) represents the distance between data point x and centroid y; x = data point, y = centroid.

5. After all points are grouped into clusters, recalculate the new centroid positions by taking the average of all points within each cluster:

$$\mu_k = \frac{1}{N_k} \sum_{i=0}^{N_k} X^i \quad (2)$$

6. Assign each object to a cluster based on the minimum distance between the object and the centroid. The membership value of data in the distance matrix is represented by 0 or 1, where a value of 1 indicates that the data point belongs to a specific cluster, while 0 indicates that it is allocated to another cluster.
7. Repeat the process starting from step two until the centroids no longer change and the cluster members remain fixed in their respective groups.

E. CRISP-DM

This study applies the Cross Industry Standard Process for Data Mining (CRISP-DM) model, which includes the stages of Business Understanding, Data Understanding, Data Preparation, Modeling, and Evaluation. The main objective is to cluster problematic loan customers of Bank Mega Finance Bengkulu using the K-Means Clustering algorithm in RapidMiner. The data were cleaned from duplicates, missing values, and outliers, then normalized to ensure optimal modeling performance. The clustering results were evaluated using the Davies-Bouldin Index (DBI) to assess the quality of segmentation. This study does not include the Deployment stage, as the focus is on historical data analysis and strategic recommendations rather than operational system implementation. The stages of this research are illustrated in Figure 1.

Figure 1. Research Stages

F. Rapid Miner

RapidMiner is an advanced software platform for data science and machine learning. The platform provides a wide range of tools for data preparation, modeling, evaluation, and deployment. RapidMiner is designed for ease of use, allowing users to build and test various models without requiring programming experience [13].

With its drag-and-drop interface, RapidMiner enables users to create workflows that process and analyze data. The platform supports various data sources, including flat files, databases, and big data platforms such as Hadoop and Spark. In addition, the software is equipped with numerous built-in operators that serve as the fundamental components of workflows, covering all stages of the Data Mining process, such as data cleaning, feature selection, and modeling [14].

III. RESEARCH AND DISCUSSION

A. Type of Research

This study uses a quantitative approach with a dataset obtained from primary data (internal data of Bank Mega Finance Bengkulu). The data is processed to cluster problematic loan customers using the K-Means Clustering algorithm within the CRISP-DM framework. This approach aims to provide insights into customer risk profiles, support decision-making in managing non-performing loans, and offer strategic recommendations based on customer segmentation.

B. Research Subjects

The research subjects consist of 5,305 primary data records obtained from the credit records of Bank Mega Finance Bengkulu during the period 2008–2023.

C. Research Stages

This study follows the CRISP-DM (Cross Industry Standard Process for Data Mining) framework, which consists of the following stages:

1. Business Understanding The focus is on understanding the business objective: clustering delinquent loan customers to support decision-making at Bank Mega Finance Bengkulu. Project planning, resource assessment, and success criteria are established to ensure the analysis provides practical benefits.
2. Data Understanding This stage involves collecting and exploring customer data, including credit information, age, gender, loan amount, and delinquency status. Initial analysis identifies data quality issues, patterns, and inconsistencies to ensure reliable clustering results.
3. Data Preparation Data cleaning and transformation are performed by removing missing values, duplicates, and outliers. Relevant attributes are selected, and the dataset is normalized to optimize the K-Means algorithm performance.
4. Modeling
   The K-Means Clustering algorithm is applied in RapidMiner to group customers. The number of clusters (K) is determined through experimentation or domain knowledge. Iterative processing ensures meaningful and representative clusters, helping the bank identify distinct delinquent customer groups.
5. Evaluation
   Cluster quality is assessed using metrics such as the Davies-Bouldin Index (DBI) to ensure the segmentation aligns with business objectives and can support decision-making.
6. Deployment
   This study does not include deployment, as the focus is on analyzing historical data and providing strategic recommendations rather than operational system implementation. Deployment would require complex infrastructure and extensive testing, which is beyond the scope of this research.

IV. Analysis And Discussion

A. *Business Understanding*

This study focuses on the main issue faced by Bank Mega Finance Bengkulu**:** the high number of customers with delinquent loans, which increases credit risk and may affect the portfolio's health. Using internal data, customers are segmented based on loan amount, delinquency duration, and payment history. This approach aims to provide management with clearer insights for decision-making, enabling more targeted collection strategies, loan restructuring, and risk mitigation policies.

B. *Data understending*

This study uses primary data from Bank Mega Finance Bengkulu's credit records for 2008–2023, containing customers with delinquent loans. The dataset includes eight variables: Name, Start Period, Tenor, Initial Balance, Ending Balance, Installment, Days Aging, and Months Aging. Data were obtained in Excel format and processed through cleaning and variable selection to prepare it for analysis. The dataset is shown in Table 1.

Tabel 1. Tabel Dataset

ID Nasa	Period Awal	Tenor	OP Awal	OP Akhir	Angs	Hari Aging	Bln Aging
NS-1	26	35	17.290.000	3.831.953	494.000	5321	175
NS-2	1	35	21.140.000	12.638.340	604.000	6067	200
NS-3	7	29	9.976.000	5.451.648	344.000	5855	193
NS-4	30	35	25.200.000	3.258.924	720.000	5064	167
NS-5	12	35	16.695.000	7.675.189	477.000	5577	183
NS-6	6	23	14.398.000	8.123.235	626.000	5637	185
NS-7	34	35	18.235.000	508.243	521.000	4557	150
NS-8	34	35	15.050.000	419.481	430.000	4557	150
NS-9	31	35	17.255.000	1.850.906	493.000	4606	152
NS-10	34	35	17.605.000	490.327	503.000	4493	148
NS-11	34	35	17.220.000	479.198	492.000	4488	148
NS-12	29	35	17.850.000	2.809.223	510.000	4600	151
NS-13	33	35	18.200.000	1.002.424	520.000	4470	147
NS-14	34	35	17.675.000	491.171	505.000	4427	146
NS-15	34	35	19.530.000	543.453	558.000	4424	146
NS-16	4	35	19.600.000	12.654.706	560.000	5340	176
NS-17	27	35	19.670.000	4.094.154	562.000	4641	153
NS-18	33	35	17.675.000	970.547	505.000	4442	146
NS-19	33	35	17.850.000	979.357	510.000	4439	146
NS-20	28	35	19.075.000	3.454.623	545.000	4584	151
NS-21	29	35	17.675.000	2.771.469	505.000	4542	149
NS-22	34	35	18.900.000	526.051	540.000	4388	144
NS-23	34	35	17.500.000	486.504	500.000	4380	144
NS-24	31	35	17.500.000	1.877.736	500.000	4470	147
NS-25	24	29	14.790.000	2.364.715	510.000	4684	154
NS-26	33	35	16.975.000	936.567	485.000	4410	145
NS-27	31	35	18.550.000	1.971.744	530.000	4469	147
NS-28	32	35	22.400.000	1.825.830	640.000	4441	146

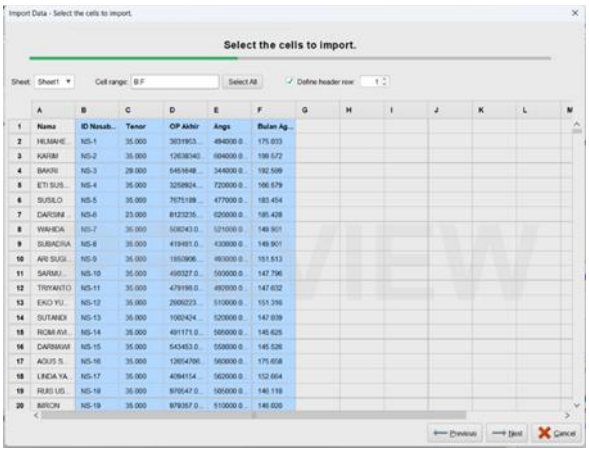
N S-29	32	3 5	18.2 00.000	1.47 9.223	52 0.000	4 437	1 46
N S-30	34	3 5	17.8 50.000	494. 699	51 0.000	4 369	1 44
.....							
N S-5305	33	3 5	21.0 00.000	1.14 6.017	60 0.000	2 137	7 0

C. *Data Preparation*

Next, the data processing in RapidMiner is carried out by following three predetermined stages, namely:

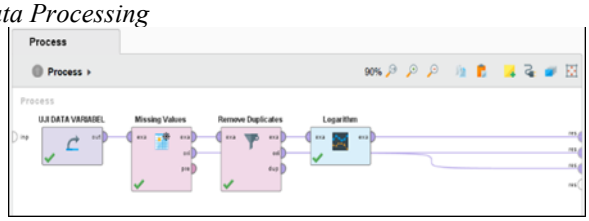
1. Data Selection

The data selection stage was conducted to choose relevant variables from the initial eight, resulting in five main variables: Customer ID, Tenor, Ending Balance (Outstanding Loan), Installment, and Months Aging (Delinquency Duration), as shown in Figure 2.



Picture 2. Proses Data *Selection*

2. *Data Processing*



Picture 2. Proses Data proccesing

The figure above illustrates the preprocessing workflow in RapidMiner. The process begins with data input, followed by handling missing values and removing duplicates, and concludes with a logarithmic transformation to reduce the impact of outliers, making the data more suitable for clustering analysis.

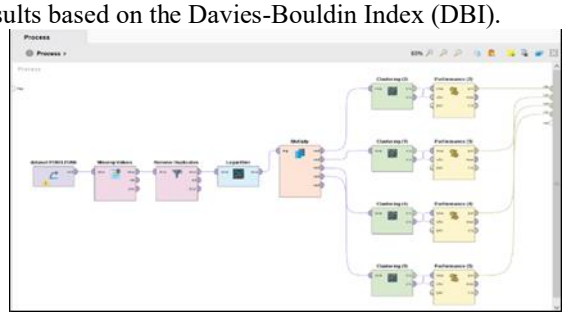
3. Data Transformation

In this study, data transformation was not required because all the data used were already numerical and did not show significant scale differences that could affect the analysis results. Transformation is typically applied to non-numerical data, such as text or categorical variables, or when there are large scale differences between numerical features that may impact the algorithm.

D. *Modeling*

At the modeling stage, the K-Means algorithm is applied using RapidMiner. The steps for determining the number of clusters are carried out through several stages.

1. Import the dataset into the analysis process.
2. Add the Logarithm operator to reduce the range of extreme values and improve the data distribution.
3. Use the Multiply operator so that the data can be copied and processed simultaneously across multiple branches.
4. Apply the K-Means Clustering operator with varying numbers of clusters (2, 3, 4, and 5)
5. Add the Performance operator to evaluate the clustering results based on the Davies-Bouldin Index (DBI).



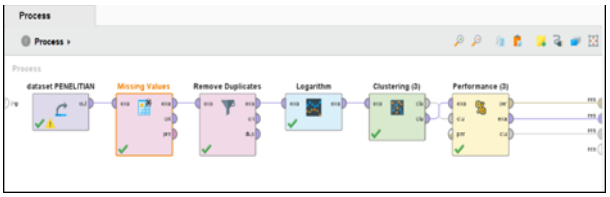
Picture 3. Cluster Number Determination Process

In the next stage, the obtained DBI values are used as the evaluation criterion in this study, as shown in Table 2. The DBI values correspond to the different numbers of clusters modeled.

Table 2. DBI Values for Each Cluster

Klaster	Nilai DBI	Average *Centroid Distance*
2	0.227	0.183
3	0.201	0.145
4	0.242	0.110
5	0.293	0.098

According to the Davies-Bouldin Index (DBI) rule, a smaller DBI value or one closer to zero indicates better clustering quality in data mining (Sigit, 2024). Based on Table 4.2**,** Cluster 3 has the best DBI value of 0.201, which is closer to zero compared to the other clusters. Therefore, the data is grouped using three clusters**.**



Picture 4. Proses Modeling Dataset 3 Klaster

The clustering process was carried out using three clusters**.** Figure 5 below shows the number of members in each cluster. Figure 5 shows the number of members in each cluster. Out of a total of 5,262 data items, the items are grouped into three

clusters**:** Cluster 0 with 3,154 data points, Cluster 1 with 1,911 data points, and Cluster 2 with 197 data points.

Cluster Model

```
Cluster 0: 3154 items
Cluster 1: 1911 items
Cluster 2: 197 items
Total number of items: 5262
```

Picture 5. Cluster Model

Figure 5 shows the number of members in each cluster. Out of a total of 5,262 data items, the data are grouped into three clusters**:** Cluster 0 with 3,154 items, Cluster 1 with 1,911 items, and Cluster 2 with 197 items.

Attribute	cluster_0	cluster_1	cluster_2
Tenor	3.443	3.172	3.045
OP Akhir	16.010	14.352	15.553
Angs	13.411	13.141	13.516
Bulan Aging	4.585	4.524	2.454

Picture 6. Centroid Values Table

To determine the number of clusters, three clusters were tested: C0, C1, and C2, using the variables Tenor, Ending Balance (OP Akhir), Installment (Angs), and Months Aging**.**

a. Tenor relates to the loan duration. The average Tenor values for each cluster are: Cluster 0 (3.443), Cluster 1 (3.172), and Cluster 2 (3.045)**.**

b. Ending Balance (OP Akhir) represents the remaining outstanding loan used in the analysis. The average values for each cluster are: Cluster 0 (16,010), Cluster 1 (14,352), and Cluster 2 (15,553)**.**

c. Installment (Angs) indicates the average installment amount. The average Installment for each cluster is: Cluster 0 (13,411), Cluster 1 (13,141), and Cluster 2 (13,516)**.**

d**.** Months Aging shows the duration the entity has been in a particular status. The values for this attribute are: Cluster 0 (4.585), Cluster 1 (4.524), and Cluster 2 (2.454)**.**

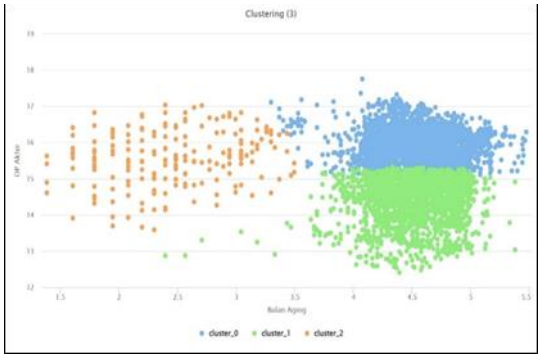


Figure 4.11. Dataset Visualization per Cluster

The three clusters can also be categorized into three risk levels based on their centroid values:

Cluster 0 (blue): has centroid values indicating relatively high Ending Balance (OP Akhir) and high Months Aging, so it is categorized as “High Risk.”

Cluster 1 (green): although lower than Cluster 0, its values are still higher than Cluster 2, and it is categorized as “Medium Risk.”

Cluster 2 (orange): shows the lowest centroid values for Ending Balance and Months Aging, and is therefore categorized as “Low Risk.”

E. Evaluation

The evaluation was conducted using the Davies-Bouldin Index (DBI) to assess the three clusters from a total of 5,262 recorded data points.

PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: 0.145
Avg. within centroid distance_cluster_0: 0.106
Avg. within centroid distance_cluster_1: 0.196
Avg. within centroid distance_cluster_2: 0.280
Davies Bouldin: 0.201
```

Figure 4.12. DBI Values

Figure 4.12 presents the evaluation results of the clustering model using centroid distance and the Davies-Bouldin Index (DBI). The average distance of data points to their centroids across all clusters is 0.145, indicating the overall spread of data within the clusters. Cluster 0 has a lower value of 0.106, showing that its data points are closer to the centroid. In contrast, Cluster 1 and Cluster 2 have higher values, 0.196 and 0.280, respectively, indicating that data points in these clusters are more dispersed. The Davies-Bouldin Index of 0.201 indicates good cluster separation, as lower values reflect clearly distinct clusters (Fauzi, Resmi, and Hermanto, 2023). The clustering analysis shows that problematic loan customers are divided into three clusters with different risk levels: Cluster 0 has 3,154 customers in the high-risk category, Cluster 1 has 1,911 customers in the medium-risk category, and Cluster 2 includes 197 customers in the low-risk category. These varying risk levels require the implementation of specific handling strategies for each cluster.

V. CONCLUSION

A. Conclusion

This study successfully implemented the K-Means Clustering algorithm to group problematic loan customers at Bank Mega Finance Bengkulu using the CRISP-DM methodology. The segmentation results identified three optimal clusters**:**

1. Cluster 0 (High Risk): Customers with significant outstanding loans and long delinquency periods, posing a high potential loss to the company.
2. Cluster 1 (Medium Risk): Customers with moderate delinquency levels requiring closer attention.

3. Cluster 2 (Low Risk): Customers with small outstanding loans and relatively short delinquency periods.

B. Recommendations

Based on the research results and interviews with Mr. Erlan Saputra, Remedial Supervisor**,** several strategic recommendations are proposed for managing problematic loan customers to optimize decision-making in credit risk management at Bank Mega Finance Bengkulu**:**

1. Cluster 0

management should focus on intensive collection and preventing larger losses. Measures may include loan restructuring for debtors who are still capable of paying, direct collection, and collateral seizure if necessary. Additionally, implementing an early warning system three times a week is important to monitor potential defaults from an early stage.

2. Cluster 1**,**

the strategy should aim to prevent risk escalation through regular monitoring**,** financial management education, and providing **incentives** to encourage timely payments. Historical data analysis can be used to identify potential declines in credit quality, allowing for quicker intervention.

3. Cluster 2

focuses on maintaining credit quality by fostering good relationships with customers, providing rewards, and leveraging opportunities for cross-selling additional products**.** This cluster-based approach is expected to optimize resource utilization, reduce non-performing loans, and improve the overall quality of the customer portfolio.

REFERENCES

[1] I. Rahadiyan, "Perkembangan Financial Technology Di Indonesia Dan Tantangan Pengaturan Yang Dihadapi," Mimb. Huk., vol. 34, no. 1, pp. 210–236, 2022, doi: 10.22146/mh.v34i1.3451.

[2] A. Aswirah, A. Arfah, and S. Alam, "Perkembangan Dan Dampak Financial Technology Terhadap Inklusi Keuangan Di Indonesia: Studi Literatur," J. Bisnis dan Kewirausahaan, vol. 13, no. 2, pp. 180–186, 2024, doi: 10.37476/jbk.v13i2.4642.

[3] Habibi, Wahyuni, and I. Afryanti, "CLUSTERING NASABAH MENGGUNAKAN ALGORITMAK-MEANSCLUSTERING PADA BANKALTIMTARA," 2024.

[4] D. Irawan, G. Wijaya, and T. T. Warisaji, "Penerapan Algoritma K-MeansClusteringuntuk Segmentasi Nasabah Bank," J. Teknol. Inf. dan Rekayasa Komput., vol. 6, pp. 47–53, 2025, [Online]. Available: https://www.bios.sinergis.org/bios/article/view/162

[5] S. Nur Illah, N. Suarna, I. Ali, and D. Solihudin, "Journal of Artificial Intelligence and Engineering Applications K-Means Clustering Method to Make Credit Payment Groupinhg Efficient," vol. 4, no. 2, pp. 2808–4519, 2025, [Online]. Available: https://ioinformatic.org/

[6] H. Rizkyanto and F. L. Gaol, "Customer Segmentation of Personal Credit using Recency, Frequency, Monetary (RFM) and K-means on Financial Industry," Int. J. Adv. Comput. Sci. Appl., vol. 14, no. 4, pp. 152–162, 2023, doi: 10.14569/IJACSA.2023.0140417.

[7] P. N. Sari and A. F. Mustoffa, "Analisis Strategi Dalam Penyelesaian Kredit Bermasalah pada PT.BPR Aswaja Ponorogo," Japp J. Akuntansi, Perpajak. Dan Portofolio, vol. 3, no. 1, pp. 1–7, 2023, doi: 10.24269/japp.v3i1.5019.

[8] A. Hilman, "Kewajiban Debitur Kredit Usaha Rakyat Atas Tunggakan Pembayaran Angsuran Kredit," J. Huk. Resp., vol. 23, no. 2, pp. 12–21, 2024, [Online]. Available: https://journal.unilak.ac.id/index.php/Respublica

[9] D. F. Surianto and D. F. Surianto, "Enhancing K-Means Clustering for Journal Articles using TF-IDF and LDA Feature Extraction," Brill. Res. Artif. Intell., vol. 4, no. 2, pp. 964–972, 2025, doi: 10.47709/brilliance.v4i2.5547.

[10] I. F. Fauzi, M. G. Resmi, and T. I. Hermanto, "Penentuan Jumlah Cluster Optimal Menggunakan Davies Bouldin Index pada Algoritma K-Means untuk Menentukan Kelompok Penyakit," JUMANJI (Jurnal Masy. Inform. Unjani), vol. 7, no. 2, pp. 1–15, 2023, [Online]. Available: https://jumanji.unjani.ac.id/index.php/jumanji/article/view/321

[11] R. N. Rahmadiana and D. Lestarini, "Implementasi Algoritma K-Means untuk Pengelompokan Customer Churn dalam Menentukan Strategi Pemasaran Bank," Sist. J. Sist. Inf., vol. 14, no. 4, pp. 1909–1919, 2025, [Online]. Available: http://sistemasi.ftik.unisi.ac.id

[12] N. Febrian and N. Noviandi, "Perbandingan Manhattan dan Euclidean Distance Untuk Pengelompokan Penyakit Jantung Menggunakan Algoritma K-Means," ICIT J., vol. 10, no. 1, pp. 61–70, 2024, doi: 10.33050/icit.v10i1.2860.

[13] J. F. Andry, H. Hartono, Honni, A. Chakir, and Rafael, "Data Set Analysis Using Rapid Miner to Predict Cost Insurance Forecast with Data Mining Methods," J. Hunan Univ. Nat. Sci., vol. 49, no. 6, pp. 167–175, 2022, doi: 10.55463/issn.1674-2974.49.6.17.

[14] M. Rafi Nahjan, Nono Heryana, and Apriade Voutama, "Implementasi Rapidminer Dengan Metode Clustering K-Means Untuk Analisa Penjualan Pada Toko Oj Cell," JATI (Jurnal Mhs. Tek. Inform., vol. 7, no. 1, pp. 101–104, 2023, doi: 10.36040/jati.v7i1.6094.

[15] S. Butsianto and N. T. Mayangwulan, "Penerapan Data Mining Untuk Prediksi Penjualan Mobil Menggunakan Metode K-Means Clustering," J. Nas. Komputasi dan Teknol. Inf., vol. 3, no. 3, pp. 187–201, 2020, doi: 10.32672/jnkti.v3i3.2428