Customer Segmentation of Potential Customers Based on Rejection Loans Using K-Means Clustering

1st Narita Ayu Prahastiwi
Telkom University
Bandung, Jakarta
naritaayupp@student.telkomuniversity.ac.id

2nd Muharman Lubis
Telkom University
Bandung, Jakarta
muharmanlubis@telkomuniversity.ac.id

3rd Hanif Fakhrurroja
Telkom University
Bandung, Jakarta
haniffakhrurroja@telkomuniversity.ac.id

***Abstract*— This study explores customer segmentation for analyzing rejected loan applicants at Bank XY using the K-Means clustering algorithm. With the rise of digital banking and Fintech lending in Indonesia, understanding customer behavior and optimizing services are critical to staying competitive. Data from 94,573 rejected loan records were processed to identify potential borrowers overlooked by the credit scoring system. The research methodology involved data preprocessing, transformation, and feature selection using PCA to reduce 306 attributes to 99. Using the Elbow method, four clusters were identified, representing customer segments categorized as *Most Potential*, *Potential*, *Medium Potential*, and *Low Potential*. Key insights include correlations between loan attributes like credit score, installment ratio, and income. Clusters 2 and 3 were identified as priorities for re-engagement due to their high likelihood of successful loan repayment. This segmentation model can help the bank enhance profitability by targeting viable customers, optimizing marketing strategies, and minimizing the risk of non-performing loans.**

Keywords—Customer Segmentation, Clustering, K-Means, Data Mining, Loans.

I. INTRODUCTION (*HEADING 1*)

The rapid development of technology and information in the current digital era significantly impacts business growth in various sectors in Indonesia, one of which is the financial sector, particularly conventional banks, which are beginning to transform into digital banks. According to Law Number 10 of 1998 concerning Banking, credit or loans are one of the activities or business sources of banking institutions [1]. The impact of technological advancements and modern developments has intensified business competition, urging companies to maximize their capabilities to stay competitive. Survey results indicate that credit or loan users in Indonesia have experienced fluctuating growth. However, the demand and interest of the public in using credit remain high due to the convenience it provides in financial assistance, including working capital loans, investment loans, and consumer loans [2].

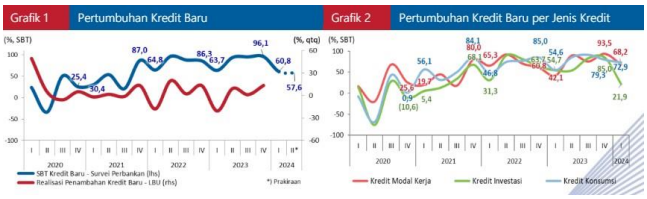

Fig. 1. Credit Growth in Indonesia [2]

Fig 1 indicates that, on a quarterly basis (Q1 2024), new credit growth showed positive trends. This is reflected in the SBT value (Survey of Bank Tendency) of new credit disbursement in Q1 2024, which reached 60.8%. Although lower than 96.1% in the previous quarter, the data illustrates that the need for credit disbursement in Indonesia remains substantial. Figure 2, which categorizes loans by usage type, also demonstrates active growth across all loan categories, although not as high as Q4 2023, following historical patterns. This underscores the ongoing necessity for credit in Indonesia and presents an opportunity to enhance corporate profitability while also posing challenges due to the increasing competition among lending platforms.

The application of machine learning in the form of credit scoring facilitates the process of determining whether applicants are eligible for loans. This technology is implemented in one of Bank XY's mobile app-based digital loan products to assess the creditworthiness of borrowers and prevent Non-Performing Loans (NPL) [3]. However, loan applications that have been rejected by the current system are not yet given special attention, either through telesales or alternative program offers that could help customers receive desired loan offerings. Furthermore, data rejected by the system may lead to false positives, predicting eligibility incorrectly due to potential inaccuracies in the model. These inaccuracies could result in misclassifying borrowers who are capable of repayment or those who might default [4]. To address this, customer segmentation analysis is required to evaluate customer value, enabling companies to identify high-value customers and less-profitable ones [5],[6]. This can be achieved through data mining, a process of uncovering patterns or insights within large datasets, which can also be applied to customer segmentation [7]. Clustering is an unsupervised data mining technique that can be used for customer segmentation [8]. It has proven to be efficient in uncovering patterns or relationships hidden within repositories of unlabeled datasets by partitioning a set of data elements into groups [9]. The number of clusters formed can be determined using methods such as the elbow method, gap statistics, or dendrograms with average linkage.

In this case, clustering is applied to loan-rejected data or data rejected by the system using the K-Means approach. Referring to the study by [9] titled *"Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services,"* it was found that customer segmentation results using the K-Means approach on retail business training patterns were identified very effectively. Based on previous research, the K-Means algorithm is one of

the clustering methods that can perform customer segmentation with high accuracy for both numerical and categorical data [10]. Therefore, we see the need to analyze and segment customers whose loan applications have been rejected by conducting potential customer segmentation based on rejected loans using K-Means clustering in the case study of Pinang Flexi. This effort indirectly assists customers while preventing them from turning to competitors. It enables the company to target specific consumer groups, leading to more efficient marketing resource allocation and greater potential for cross-selling and up-selling.

II. Literature Review

The data within the banking sector is highly diverse, one of which is customer loan transaction data, which will be discussed in this paper concerning Bank XY. This data can be utilized to enhance the company's profitability by segmenting customers appropriately to align with the company's business targets while considering the design and distribution of products [11]. Customer Segmentation is one approach that can be employed to analyze data by dividing the business customer base into homogeneous groups consisting of customers with similar market characteristics [12]. This segmentation is based on factors that directly or indirectly influence the market or business, such as customer demographic data in this paper [13]. The importance of customer segmentation includes, among others [14] :

1. The ability of the company to adjust market strategies tailored to each customer group.
2. A tool for business decision-making to address risks, such as in establishing credit relationships with customers.
3. Identifying suitable products for each segment and managing the dynamics between supply and demand.

One of the efforts in conducting customer segmentation is data analysis through data mining. Data mining is the process of collecting, processing, and analyzing data to uncover hidden patterns that can provide new insights. These data sources may come from databases, data warehouses, or dynamically streamed data [15]. Data mining is also a step in the data processing cycle. Through data mining, patterns for clustering or classification can be identified and applied [8]. This technique is often used in business to group customers based on certain variables, such as age, gender, or account balance, to improve revenue and customer satisfaction [16]. Grouping unlabeled data based on similarities among customers is a clustering technique used for customer segmentation [17]. One commonly used method in data mining for clustering is K-Means, which divides data into several relevant groups. This method partitions or separates objects into distinct regions, grouping data with similar characteristics into one cluster and data with different characteristics into another cluster[18].

In short, the literature review emphasizes the importance of understanding customer segments by clustering using the K-Means method on rejected customer loan data at Bank XY and then utilizing the segmentation results to increase profitability and optimize the company's business needs.

III. Metodology and Workflow

The research method employed in this study is illustrated in Fig. 2, which outlines seven steps for clustering rejected loan application data at Bank XY. The process begins with data collection and understanding, followed by data preprocessing, data transformation, determining the number of clusters to be formed, applying the K-Means method, and concludes with deriving clustering insights.

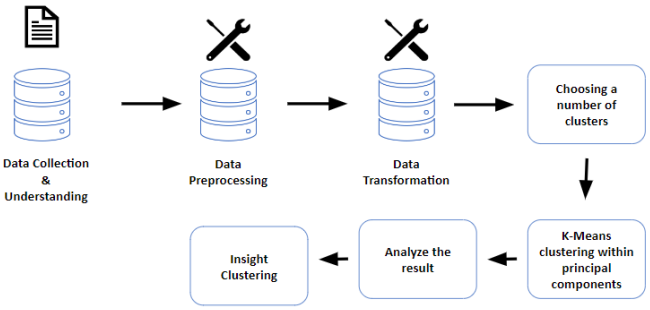

Fig. 2. Research Method

A. Data Collection & Understanding

Data collection and understanding are conducted to comprehend the contents of the data, examine correlations among attributes in the dataset, and gain detailed insights into these attributes [19].

B. Data Preprocessing

Data preprocessing involves preparing raw data into a form ready for analysis. This includes **data cleaning**, which addresses missing, inconsistent, or erroneous data that could impede the data mining process. This step is also part of the Knowledge Discovery in Database (KDD) process, which aims to extract useful, hidden, and relevant knowledge from large datasets [19].

C. Data Transformation

Data transformation refers to the process of modifying the format or structure of data to make it more suitable for analysis. This step is performed after data cleaning.

D. Choosing several cluster

The determination of how many clusters to form is achieved using the K-Means method. K-Means is a partitioning technique that divides objects into separate regions, grouping data with similar characteristics into one cluster and data with different characteristics into another cluster [20]. The first step in the K-Means method is to determine the number of clusters (K), which is done using the elbow method. This process identifies the optimal value for K by calculating the Sum of Squares Error (SSE), where a significant decrease in SSE helps pinpoint the best number of clusters [21].

$$SSE = \sum_{K=1}^{K} \sum_{x_i \in S_K} ||X_i - C_K||_2^2 \quad (1)$$

Where K = number of clusters, C = midpoint, X = different data in each cluster [21]. The second step involves selecting the centroid or the central point of the cluster, which can be

chosen randomly [21]. The third step is calculating the distance between the centroid and each object using the Euclidean distance, a widely used distance measurement method [12], [21]. This technique groups new data based on its proximity to several nearest data points [22].

$$dist(i,j) = \sqrt{(x_{i1} - x_{j1})^2 + \cdots + (x_{in} - x_{jn})^2} \quad (2)$$

Where, x = objek, i=(xi,$xi2$,…,xin), j= ($xj1$,$xj2$,…,xjn) represent two-dimensional data objects[12], [21]. The next step is assigning data points to clusters based on the minimum distance between the centroid and the data points, repeating the process until the centroids stabilize [18], [21].

E. PCA *(Principal Component Analysis)*

Principal Component Analysis (PCA) is a multivariate analysis technique used to evaluate data tables where observations are described by numerous interrelated variables [23]. The goal of PCA is to reduce the dimensionality of the data, making interpretation easier. PCA extracts important information from statistical data by representing it as a set of new orthogonal variables, known as principal components. This technique also enables visualization of similarity patterns among observations and variables in the form of scatter plots [24].

IV. Result and Discusion

In this section, the case study involves performing customer segmentation analysis using the K-Means clustering algorithm based on rejected loan data (CR_SCORE_REJ) or data from customers whose loan applications were rejected by the system. The study also examines whether the rejected loan applications were due to model errors, which could allow for a re-approach to offer other products, or if the rejection was justified because the customer did not meet the business or credit engine criteria necessary for a loan.

A. Dataset

The dataset used comprises records of rejected loan applications over two years, from 2021 to 2023, in Comma-Separated Values (CSV) format. It contains 94,573 rejected records, consisting of 84,018 entries labeled as CR_SCORE_REJ and 10,555 entries labeled as USR_REJECT. Description of these statuses is provided in the table below.

TABLE I. Dataset

Dataset	
CR_REJ	Users do not qualify for the credit scoring engine
USR_REJ	Users pass the credit scoring model but fail in borrowing

B. Data Understanding

Data understanding is done to understand the contents of the data we have, see the correlation of the attributes in the dataset and find out the details of the attributes [25]. The clustering carried out in customer segmentation was obtained from the rejected loan dataset with 306 features or attributes and 94,573 rows of data. Fig 3 shows that the volume of users who do not pass the credit scoring engine (CR_SCORE_REJ) is high, making it possible to carry out further analysis to gain insights that can increase profitability from rejected loan data.

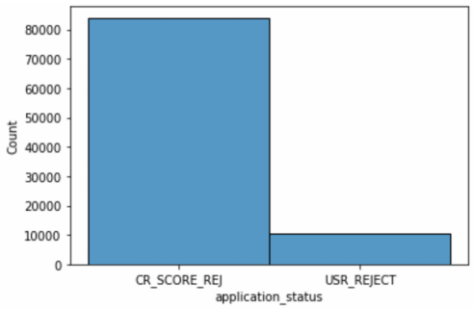

Fig. 3. Data Understanding Application Status

Then Fig 4 explains how many columns there are, knowing the start to end columns, the type of data in the dataset and also the memory used in the dataset.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 37935 entries, 0 to 94572
Columns: 306 entries, application_id to ratio_med_abal_base_dependents_6m
dtypes: float64(285), int64(11), object(10)
memory usage: 88.9+ MB
```

Fig. 4. Information Data Type

After that, try doing Exploratory Data Analysis (EDA) on the dataset to find out basic information that can be used as a reference or consideration for the next process [26]. Based on Fig 5 below, it states that most customers have grade_cs 6 to 8.

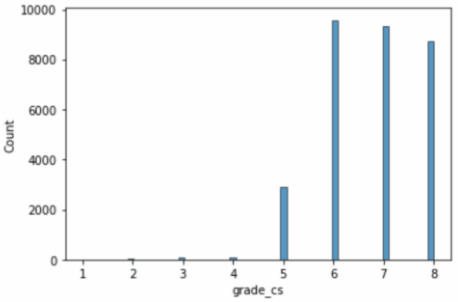

Fig. 5. EDA grade_cs

Then, based on Fig 6, it states that the majority of customers are asking for a ceiling in the range of five hundred thousand to one million rupiah and a tenor of 12 months.

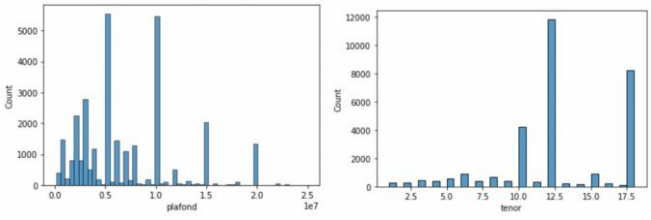

Fig. 6. EDA plafond & tenor

C. Data Cleansing

Data cleansing or data cleaning where this dataset removes outliers and garbage data such as removing missing values or unnecessary data as in the image below.

```
df = df[(df['grade_cs']>0)]
df = df[df['credit_score']>0]
```

Fig. 7. Data Cleansing

Fig 7 explains that in this study only selected data from the grade_cs and credit_score columns which were above 0. Because for grade_cs in the dataset there are still -1 and 0 where this data is not needed and also credit_score which has data 0 is also owned by grade_cs which is -1 and 0.

.

D. Data Preprocessing

Data preprocessing is the process of preparing raw data into data that is ready to be processed. At this stage, filter or only select the application_status attribute with CS_SCORE_REJ data according to Fig 8. This process is important because raw data often does not have a regular format. In this section, we will adjust the data type as shown in the image below, adjusting the date data type from previously object to datetime64 as shown in Fig 9.

application_status
CR_SCORE_REJ
CR_SCORE_REJ
CR_SCORE_REJ
CR_SCORE_REJ
CR_SCORE_REJ
...

Fig. 8. Filter Application Status

```
application_id                          object
cre_time                          datetime64[ns]
cre_time_month                    datetime64[ns]
grade_cs                                 int64
rekening_debet                           int64
                                           ...
ratio_med_abal_base_dependents_2m      float64
ratio_med_abal_base_dependents_3m      float64
ratio_med_abal_base_dependents_4m      float64
ratio_med_abal_base_dependents_5m      float64
ratio_med_abal_base_dependents_6m      float64
Length: 306, dtype: object
```

Fig. 9. Preprocessing datetime

E. Data Transformation

Data transformation is used to change data in an appropriate form in the data mining process and also change the scale of data into another form so that the data has the expected distribution.

```
encoder = OneHotEncoder()
```

Fig. 10. Data Transformation

In Fig 10, we use the label encoder package to help convert categorical data into data. The data that is converted to non-numeric is the marital_status, gender, last_education and home_ownership_status attributes, because these four attributes are still categorical (non-numerical) data. Then Fig 11, because too many attributes make the correlation plot difficult to analyze, in this paper PCA is carried out for feature selection so that dimensions can be reduced and to select suitable main components to be included in the model. Following are the results of PCA from 306 attributes to 99 attributes.

```
new feature dimension (28496, 99)
```

Fig. 11. PCA Reduction Result

F. Clustering Process

Clustering was carried out using K-Means using PCA result data consisting of 99 attributes which were saved into X.transformed to determine the best cluster using the elbow method. Based on Figure 12, it can be seen that the elbows produced form 4 clusters, where the customer segmentation results will be based on these 4 clusters.

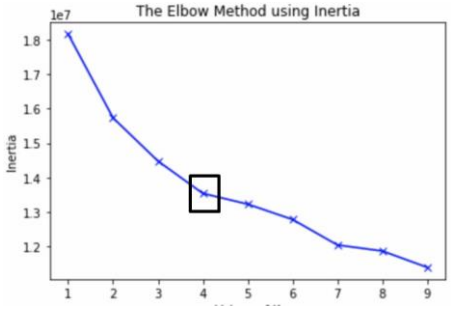

Fig. 12. Elbow

Then do a Heatmap to see the correlation of each attribute by giving a score and color. Figure 13 is the heatap result where the color is closer to bright, meaning the score is closer to value 1 and the attribute will be selected to be included in the customer segmentation cluster analysis.

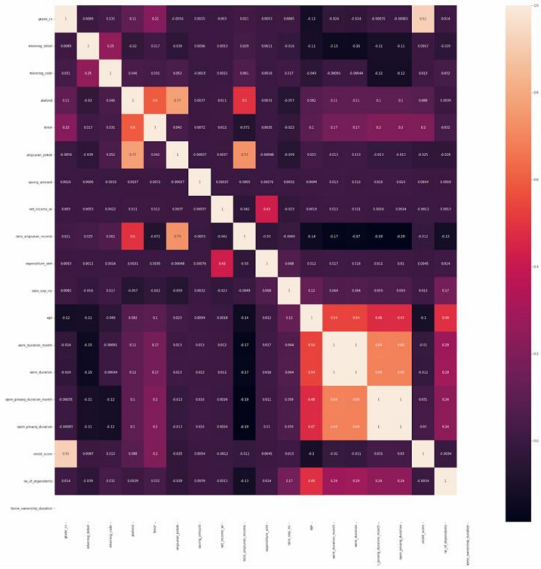

Fig. 13. Heatmap fiture correlation

The following attributes are selected based on the highest score according to the table below.

TABLE II. SCORE HEATMAP

No	Attribute	Score
1	credit_score	0.92
2	saving_amount	0.25
3	ratio_angsuran_income	0.74
4	work_duration	0.73
5	tenor	0.6
6	plafond	0.77
7	angsuran_pokok	0.83
8	net_income_wl	0.43

EDA was carried out on the attributes entered into K-Means processing, the author made several EDAs obtained from ratio_angsuran_income, grade_cs, plafond and work_duration. Below you can see the insights obtained, one of which is the installment ratio as in Fig 14.

	cluster	ratio_angsuran_income
0	0	0.07
1	1	0.10
2	2	0.03
3	3	0.08

Fig. 14. EDA Ratio Angsuran

Based on Fig 14, the highest income installment ratio is owned by cluster 1 with a value of 0.10 and the smallest is owned by cluster 2 with a value of 0.3. Analysis of other attributes has been carried out and cluster results have been obtained which can be categorized into Most Potential, Potential, Medium Potential and Low Potential. The cluster results obtained are as follows in Fig 15.

	cluster	credit_score	saving_amount	ratio_angsuran_income	work_duration	tenor	plafond	angsuran_pokok	net_income_wl	priority
0	0	620.0	126675.73	0.07	8.0	15.0	7500000.0	555555.56	6535400.0	Medium Potential
1	1	592.0	70256.63	0.10	5.0	12.0	5000000.0	500000.00	4500000.0	Low Potential
2	2	640.0	652380.00	0.03	3.0	12.0	2700000.0	168750.00	5500000.0	Most Potential
3	3	623.0	706250.00	0.08	7.0	18.0	10000000.0	555555.56	6535000.0	Potential

Fig. 15. Clustering Result

V. CONCLUSION

The conclusions drawn from this study on customer segmentation to identify customers with the potential to reapply for loans in the rejected loan dataset using the K-Means algorithm indicate the following: Cluster 2 is identified as the priority group "Most Potential," while Cluster 3 is categorized as the priority group "Potential."

Based on these findings, it is recommended that the business team re-engage with customers in Cluster 2 and Cluster 3. Out of the 30,709 processed data points, 3,901 customers (categorized as "Most Potential," "Potential," and "Medium Potential") have the potential to apply for loans but were rejected by the system. Among these, 426 customers are identified as having strong potential and minimal risk, based on the "Most Potential" and "Potential" priorities.

REFERENCES

[1] Badan Pembinaan Hukum Nasional, "UNDANG-UNDANG REPUBLIK INDONESIA NOMOR 10 TAHUN 1998," pp. 2–3, 1998, [Online]. Available: www.bphn.go.id

[2] Bank Indonesia, "Survei Perbankan Triwulan I 2024: Penyaluran Kredit Baru Tumbuh Positif," https://www.bi.go.id/id/publikasi/ruang-media/news-release/Pages/sp_268224.aspx.

[3] H. E. Riwayati, A. Aviliani, and F. Y. Prastika, "The Implementation of Credit Scoring in Order to Analyze the Importance of Non Performing Loans on Peer To Peer Lending towards Credit Distribution for Micro, Small and Medium Enterprises," *International Business and Accounting Research Journal*, vol. 6, no. 2, pp. 137–147, 2022, [Online]. Available: http://journal.stebilampung.ac.id/index.php/ibarj

[4] D. Riana *et al.*, "Identifikasi Citra Pap Smear RepoMedUNM dengan Menggunakan K-Means Clustering dan GLCM," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 1, pp. 1–8, Jan. 2022, doi: 10.29207/resti.v6i1.3495.

[5] S. Y. Kim, T. S. Jung, E. H. Suh, and H. S. Hwang, "Customer segmentation and strategy development based on customer lifetime value: A case study," *Expert Syst Appl*, vol. 31, no. 1, pp. 101–107, Jul. 2006, doi: 10.1016/j.eswa.2005.09.004.

[6] S. Monalisa and N. Nurahmah, "Segmentasi Pelanggan B2B Berdasarkan Perilaku Pembelian dan Firmografi pada PT. Sukses Riau Permata (SRP)," *Jurnal Teknologi Informasi dan Ilmu Komputer*, vol. 11, no. 1, pp. 11–18, Feb. 2024, doi: 10.25126/jtiik.20241116487.

[7] M. K. Gupta and P. Chandra, "A comprehensive survey of data mining," *International Journal of Information Technology (Singapore)*, vol. 12, no. 4, pp. 1243–1257, Dec. 2020, doi: 10.1007/s41870-020-00427-7.

[8] D. A. Maharani, H. Fakhrurroja, Riyanto, and C. Machbub, "Hand gesture recognition using K-means clustering and Support Vector Machine," in *ISCAIE 2018 - 2018 IEEE Symposium on Computer Applications and Industrial Electronics*, Institute of Electrical and Electronics Engineers Inc., Jul. 2018, pp. 1–6. doi: 10.1109/ISCAIE.2018.8405435.

[9] C. P. Ezenkwu, S. Ozuomba, and C. Kalu, "Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services," 2015. [Online]. Available: www.ijarai.thesai.org

[10] A. Subhan, A. Faqih, and B. Irawan, "CLUSTERING ITEM FAST MOVING DAN SLOW MOVING PADA PRODUK UNILEVER MENGGUNAKAN ALGORITMA K-PROTOTYPE (Studi Kasus: YOGYA PURWAKARTA)," 2022.

[11] S. Kulkarni, V. Lokhande, N. Bhalerao, Y. Tajane, and J. Kharat, "Advance Customer Segmentation," 2022. [Online]. Available: www.ijcrt.org

[12] S. F. Djun, I. G. A. Gunadi, and S. Sariyasa, "Analisis Segmentasi Pelanggan pada Bisnis dengan Menggunakan Metode K-Means Clustering pada Model Data RFM," *JTIM : Jurnal Teknologi Informasi dan Multimedia*, vol. 5, no. 4, pp. 354–364, Feb. 2024, doi: 10.35746/jtim.v5i4.434.

[13] C. P. Ezenkwu, S. Ozuomba, and C. Kalu, "Application of K-Means Algorithm for Efficient Customer Segmentation: A Strategy for Targeted Customer Services," 2015. [Online]. Available: www.ijarai.thesai.org

[14] A. O. Siagian and Y. Cahyono, "Strategi Pemulihan Pemasaran UMKM di Masa Pandemi Covid-19 Pada Sektor Ekonomi Kreatif," *Jurnal Teknologi Dan Sistem Informasi Bisnis*, vol. 3, no. 1, pp. 206–217, Feb. 2021, doi: 10.47233/jiteksis.v3i1.212.

[15] J. Han, M. Kamber, and J. Pei, "Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)," 2012.

[16] R. Maart Adi Waskita, I. Ferli, R. Rizqullah Fahrizal, and J. Heikal, "CUSTOMER SEGMENTATION BASED ON AGE, GENDER, PRODUCT AND TOTAL CUSTOMER BALANCE AT BANK XYZ USING THE K-MEANS CLUSTERING MODEL," *Jurnal Ekonomi, Manajemen dan Akuntansi*, 2024, [Online]. Available: http://jurnal.kolibi.org/index.php/neraca

[17] R. W. Sembiring Brahmana, F. A. Mohammed, and K. Chairuang, "Customer Segmentation Based on RFM Model Using K-Means, K-Medoids, and DBSCAN Methods," *Lontar Komputer : Jurnal Ilmiah Teknologi Informasi*, vol. 11, no. 1, p. 32, Apr. 2020, doi: 10.24843/lkjiti.2020.v11.i01.p04.

[18] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman, and B. D. Satoto, "Integration K-Means Clustering Method and Elbow Method for Identification of the Best Customer Profile Cluster," in *IOP Conference Series: Materials Science and Engineering*, Institute of Physics Publishing, Apr. 2018. doi: 10.1088/1757-899X/336/1/012017.

[19] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases)," 1996. [Online]. Available: www.ffly.com/

[20] E. C. H. Chen, L. F. Pau, P. S. P. Wang, and R. C. Dubes, "Handbook of Pattern Recognition and Computer Vision," 1998. [Online]. Available: www.worldscientific.com

[21] V. Purwayoga, "Optimasi Jumlah Cluster pada Algoritme K-Means untuk Evaluasi Kinerja Dosen," *Jurnal Informatika Universitas Pamulang*, vol. 6, no. 1, p. 118, Mar. 2021, doi: 10.32493/informatika.v6i1.9522.

[22] A. R. Lubis, M. Lubis, and Al-Khowarizmi, "Optimization of distance formula in k-nearest neighbor method," *Bulletin of Electrical Engineering and Informatics*, vol. 9, no. 1, pp. 326–338, Feb. 2020, doi: 10.11591/eei.v9i1.1464.

[23] D. L. Anne-Leen *et al.*, "Principal component analysis of texture features derived from FDG PET images of melanoma lesions," *EJNMMI Phys*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40658-022-00491-x.

[24] S. Mishra *et al.*, "Principal Component Analysis," *International Journal of Livestock Research*, p. 1, 2017, doi: 10.5455/ijlr.20170415115235.

[25] N. A. Prahastiwi, R. Andreswari, and R. Fauzi, "STUDENTS GRADUATION PREDICTION BASED ON ACADEMIC DATA RECORD USING THE DECISION TREE ALGORITHM C4.5 METHOD," *JURTEKSI (Jurnal Teknologi dan Sistem Informasi)*, vol. 8, no. 3, pp. 295–304, Aug. 2022, doi: 10.33330/jurteksi.v8i3.1680.

[26] P. A. Riyantoko, K. M. Hindrayani, T. M. Fahrudin, and M. Idhom, "Exploratory Data Analysis and Machine Learning Algorithms to Classifying Stroke Disease," 2021.