JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 307–312
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.911

POTENTIAL CUSTOMER ANALYSIS USING K-MEANS WITH ELBOW METHOD

Fitri Marisa[1], **Arie Restu Wardhani**[2], **Wiwin Purnomowati**[3], **Anik Vega Vitianingsih**[4], **Anastasia L Maukar**[5], dan **Erri Wahyu Puspitarini**[6]
[1,2,3]Fakultas Teknik, Universitas Widyagama Malang, Indonesia
[4]Fakultas Ekonomi dan Bisnis, Universitas Widyagama Malang, Indonesia
[5] Fakultas Teknik, Universitas Presiden, Indonesia
[6]Fakultas Teknologi Informasi, Universitas Dr. Soetomo Surabaya, Indonesia
Email: fitrimasrisa@widyagama.ac.id[1], arierestu@widyagama.ac.id[2], wiwin@widyagama.ac.id[3] vega@unitomo.ac.id[4]
almaukar@president.ac.id[5], erri@stmik-yadika.ac.id[6]

Abstract

This study aims to obtain cluster data of potential customers using the K-Means clustering approach supported by the elbow method to determine the correct number of clusters. The data sample that was processed was 100 customer data from a minimarket containing three criteria (gender, age, and purchase retention). The number of initial clusters is determined as 5 and then processed by calculating K-Means. The calculation of the SSE value in the K-Means process produces the lowest SSE value, and the sharpest elbow angle graph visualization is in cluster 4. So, it can be stated that the best number of clusters in this K-Means calculation is four (4) which are used as material for further analysis. Then the analysis results of four (4) clusters state that potential customers are those with high purchase retention, consisting of female customers who dominate in the three (3) clusters. Most potential female customers are customers with an age range above 35 years. Meanwhile, customers with less potential are spread across each cluster with varied gender and age but are not dominant. Thus, this knowledge can be used as a consideration for the management in determining the right promotion strategy.
KeyWords: *Potential Customer, K-Means, Elbow Method.*

I. INTRODUCTION

Potential customer data is essential knowledge for every business owner, which is one of the considerations for the accuracy of promotional strategies [1], [2]. Potential customer knowledge can be extracted by extracting customer data by determining supporting criteria [3], [4]. One of the knowledge extraction methods that can be applied in this study is K-Means clustering which functions to group data that has close characteristics that produce several clusters [5]–[8]. These clusters can be analyzed by observing the results of the segmentation and distribution of datasets based on predetermined criteria [7], [8].

However, K-Means has a weakness in determining the number of clusters generated, where not all the resulting clusters follow the analysis's needs [5]. Thus, K-Means needs to be supported by the Elbow method or SSE value which serves to determine the accuracy of the clusters formed because not all clusters are according to the needs of analysis [5], [9], [10]. How the Elbow method works are to calculate each cluster's SSE value. The smallest value of SSE as a benchmark for a cluster is designated the best. This lowest SSE value forms an elbow angle that is on the X and Y axes, where the X axis represents the recommended number of clusters [5], [9]–[11].

Data analysis using K-Means has been done in many previous studies. K-Means is one of the popular algorithms in the clustering method that can present knowledge extraction in the form of clusters from a set of data-sets based on predetermined criteria [5], [9], [11], [12]. K-Means Clustering is used in analyzing data in various fields, one of which is in research [13] applying K-Means to identify the best customer profile. Research [14] applies clustering to obtain image segmentation, while research [15] applies K-Means in Customer Relationship Management (CRM). Research [16] applies K-Means to text analysis in the field of public opinion. In general, K-Means Clustering can meet and provide solutions for data analysis needs that result in the extraction of new knowledge related to the criteria and analysis needs of the issues raised.

Research on improvising K-Means with the Elbow method has been carried out in several studies. Including research [17] introducing the elbow method to improvise K-Means. Research [18] also applies the Elbow method, which supports determining the value of K. The same was done in research [19] by adding the firefly algorithm. Likewise, the study [20] used Sum Squared Error (SSE) to improve the performance of the Elbow method. Thus, the Elbow method is quite effectively applied to improvise K-Means to produce more distinct clusters.

Thus, this study uses the K-Means clustering approach, supported by the Elbow method, to provide recommendations and considerations for management in identifying potential customers to support management decisions in implementing new policy steps to develop the business. The research was conducted using a case study model involving a sample data-set from one of the mini-markets. At the same time, the calculation process produces clusters, segment graphs, Sum Squared Error (SSE) values, and Elbow graphs.

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 307–312
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.911

This research method emphasizes three (3) essential stages, namely grouping data with K-Means, updating clusters based on consideration of the highest SSE value/sharpness of the angle of the Elbow graph, and analysis of potential customers. The stages of the research method are depicted in Figure 1.

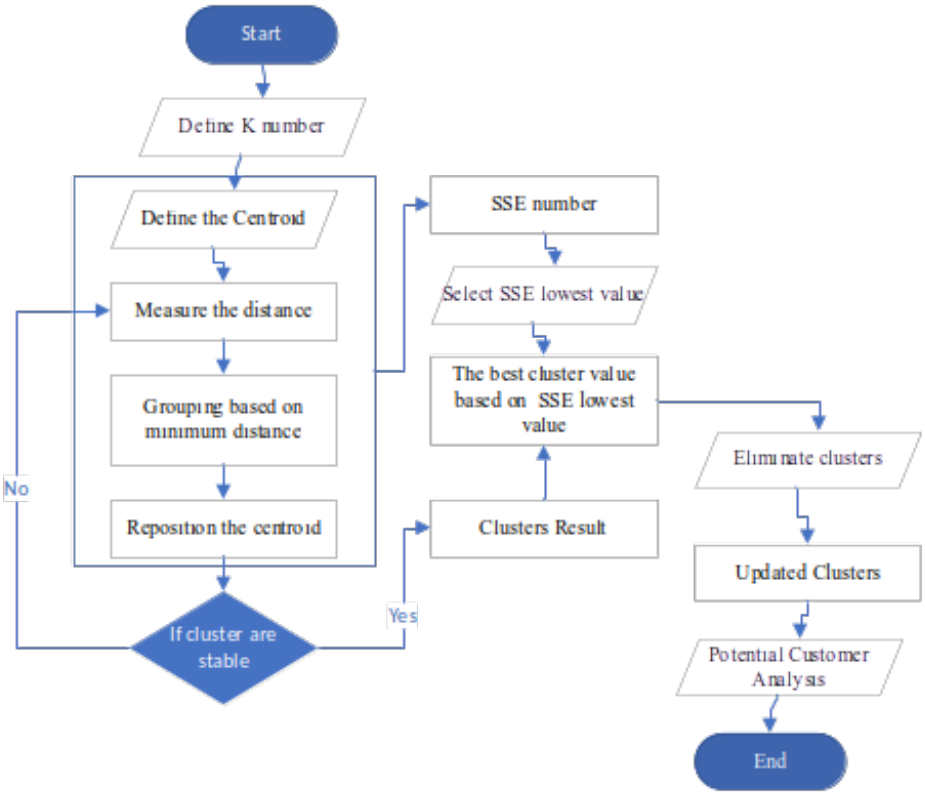

Figure. 1: Research Methods

A. K-Means Process

The K-Means process begins by determining the centroid ($K = n$). In the second step, determine the center of the centroid randomly. In the third step, measure the distance Euclidean distance with the formula (1) following:

$$\sqrt{(x_1 - x_2)^2 (y_1 - y_2)^2} \tag{1}$$

he measure of the distance or inequality between the i object and the j object, symbolized by d_{ij} and $k = 1, \ldots .., p$. The Value of d_{ij} is obtained by calculating the distance of the Euclidean square as formula (2):

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2} \tag{2}$$

Following:

d_{ij} = Euclidean Square Distance between object i and object j,
p = Number of cluster variables,
x_{ik} = Value or data from the i object in the k variable,
x_{jk} = Value or data from the j object in the k variable,

The fourth step is grouping data based on the minimum distance. The fifth step is repositioning the centroid, if the cluster is stable, then it stops, but if the cluster is unstable, then the distance calculation process is repeated.

B. Elbow Method / SSE Value

The K-Means process produces SSE values and an Elbow graph, where the lowest SSE value and the sharpest Elbow angle graph determine the best cluster value. The SSE formula can be described as Formula (3):

$$SSE = \sum (y_i - \hat{y}_i) \tag{3}$$

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 307–312
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.911

Where:
y_i = the observation value of the dependent variable for the $i$observations.
$\hat{y}_i$ = calculated Value of the dependent variable for the i observations.
The next step is eliminating clusters unrelated to the analysis and maintaining as many clusters as the number of clusters generated from the elbow method.

C. Data Analysis

The last step is to analyze the results of the cluster according to the grouped data. It is the extraction of knowledge resulting from the analysis of research data.

II. Result And Discussion

This section describes the results of calculating K-Means using the Elbow method from 100 data-sets with three (3) criteria, including Gender, Age, and Purchase retention, which have been normalized. The clusters formed amounted to 5, then reviewed the resulting SSE value. Based on the resulting SSE value (Table I), the lowest average value of SSE is in Cluster 4 at 65.4. Likewise, the resulting Elbow graph (Figure 2) produces the sharpest angle in cluster 4. Thus, the best cluster in this test is 4.

Table I: SSE Value

Cluster Number	SSE Value
1	238.2
2	186.7
3	134.9
4	65.4
5	101.9

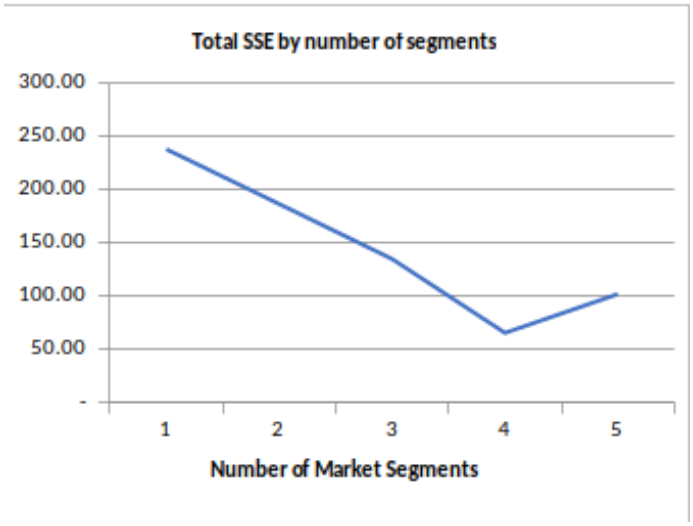

Figure. 2: Elbow Method Graph

After that, four (4) clusters were formed, which had good distribution and eliminated the highest SSE value. Then the four (4) clusters selected are cluster-2 which has two (2) segments; cluster-3, which has three (3) segments; cluster-4, which has four (4) segments; and cluster-5, which has five (5) segments. Each cluster is described in Figure 3, 4, 5, and 6.

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 307–312
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.911

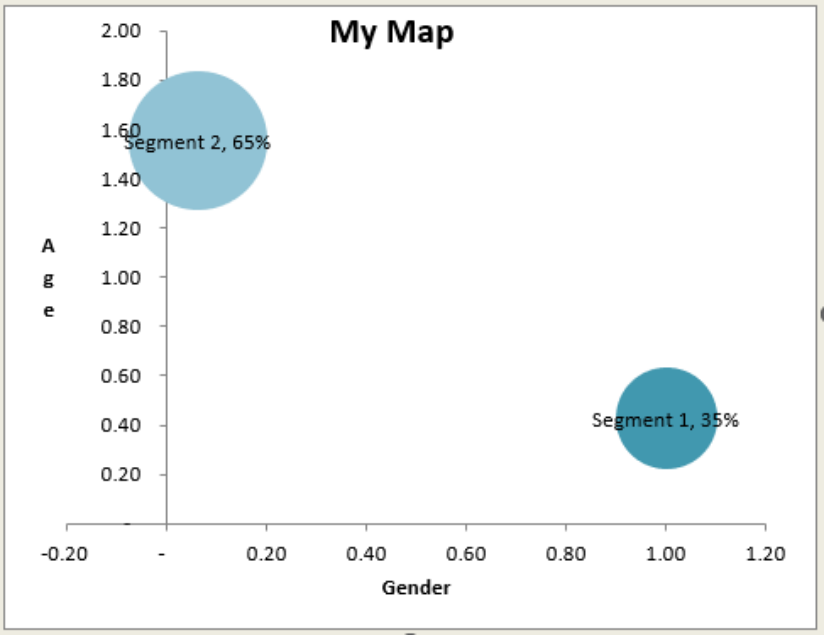

Figure. 3: Cluster-2 Segmentation Map

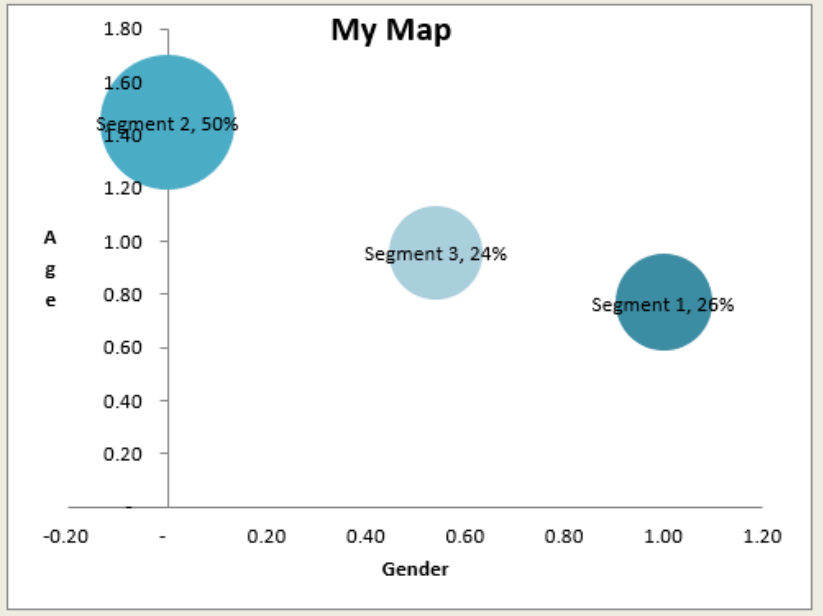

Figure. 4: Cluster-3 Segmentation Map

JIKO (Jurnal Informatika dan Komputer)
Vol. 7, no. 2, September 2023, hal. 307–312
e-ISSN : 2477-3964 — p-ISSN : 2477-4413
DOI: dx.doi.org/10.26798/jiko.v7i2.911

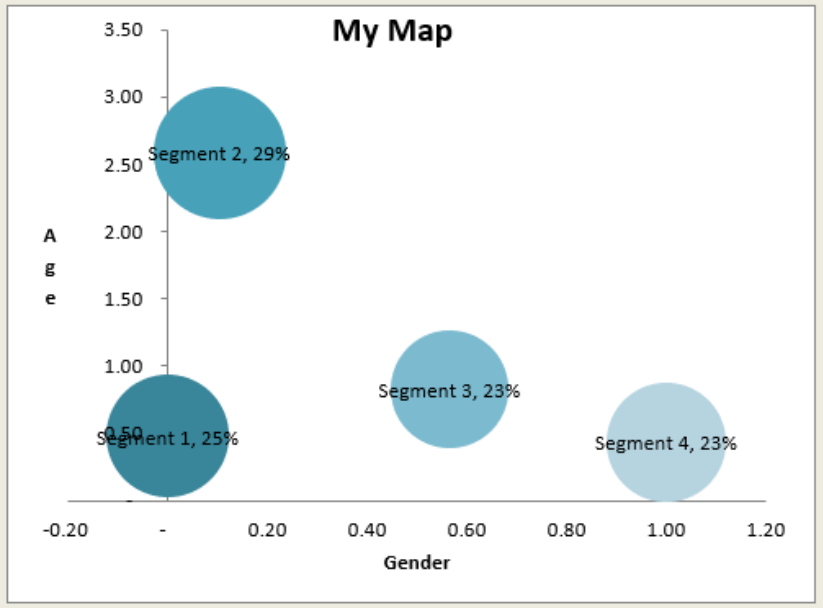

Figure. 5: Cluster-4 Segmentation Map

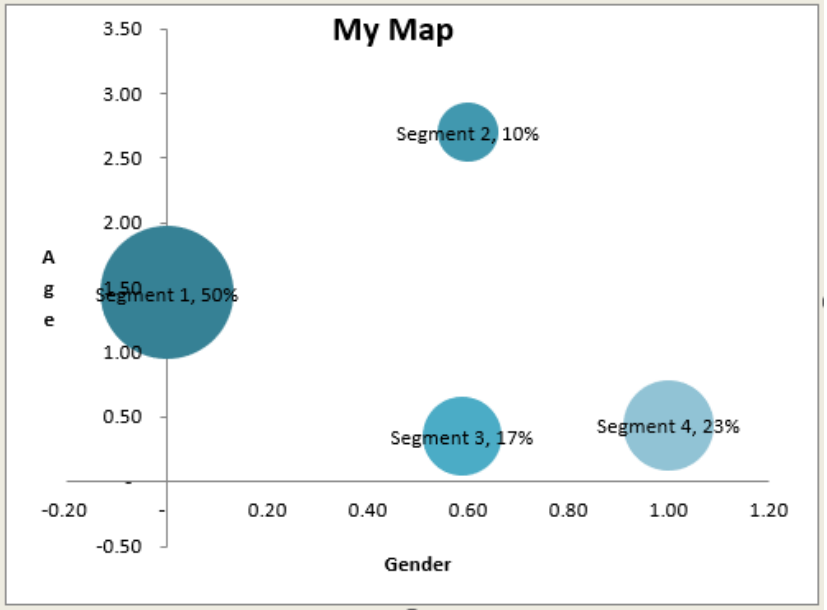

Figure. 6: Cluster-5 Segmentation Map

The next step is to continue analyzing cluster results by placing the data-sets into clusters formed to gain knowledge about potential customers. Table II describes the data groups in each segment in the cluster based on age, gender, and purchase retention criteria. By reviewing the cluster distribution, the criteria, and the number of item-set for each segment, several potential customer analyses have been produced as follows:

- Three clusters provide information that most customers are women of varying ages, including cluster-1 65%, cluster-2 50%, and cluster-4 50%.
- Cluster-3 provides information that most potential female customers are customers with an age range above 35 years. Meanwhile, potential male customers are with an age range of over 35 years.
- Four clusters provide information that non-potential customers do not dominate in each cluster but exist in each cluster.
- potential customer analysis

Table II: Potential Customer Analysis

Cluster	Segment	Cluster Data	%
2	1	Customer with male gender, variative age segmentation and variative purchase retention	35%
	2	Customer with female gender, variative age segmentation and variative purchase retention	65%
3	1	Customer with male gender, variative age segmentation and high purchase retention	26%
	2	Customer with female gender, variative age segmentation and high purchase retention	50%
	3	Customer with variative gender, variative age segmentation and low purchase retention	24%
4	1	Customer with female gender, age under 35 and high purchase retention	25%
	2	Customer with dominant female gender, age up to 35 and high purchase retention	29%
	3	Customer with variative gender, variative age and low purchase retention	23%
	4	Customer with male gender, variative age segmentation and high purchase retention	23%
5	1	Customer with female gender, variative age segmentation and high purchase retention	50%
	2	Customer with variative gender, age up to 35 and low purchase retention	10%
	3	Customer with variative gender, age under 35 and low purchase retention	17%
	4	Customer with male gender, variative age segmentation and high purchase retention	23%
	5	No member	0%

III. Conclusion

This research resulted in knowledge extraction for potential customers with the K-Means approach, which was enhanced by the Elbow method to get a more distinct cluster. This study resulted in four (4) clusters of potential customers with data distribution according to the proximity of three (3) criteria: age, gender, and purchase retention. The analysis of potential customer knowledge generated includes, among other things, that women of varying ages dominate the potential customers. However, the more dominant is the age range above 35 years. Potential male customers are those with an age range above 35 years. Based on the processed data, no potential customers dominate in the cluster, but the distribution is in all clusters. Thus, the results of this analysis can be used by management in determining promotional strategies for customers to make them more targeted. Future research can develop this analysis by adding criteria, and the data or methods can be compared or hybridized with other methods to develop potential customer analysis.

References

[1] R. Yazdani, M. J. Taghipourian, M. M. Pourpasha, and S. S. Hosseini, "Attracting Potential Customers in E-Commerce Environments: A Comparative Study of Metaheuristic Algorithms," Processes, vol. 10, no. 2, 2022, doi: 10.3390/pr10020369.
[2] R. Piehler, M. Schade, I. Hanisch, and C. Burmann, "Reacting to negative online customer reviews: Effects of accommodative management responses on potential customers," J. Serv. Theory Pract., vol. 29, no. 4, pp. 401–414, 2019, doi: 10.1108/JSTP-10-2018-0227.
[3] M. Zhan, H. Gao, H. Liu, Y. Peng, D. Lu, and H. Zhu, "Identifying market structure to monitor product competition using a consumer-behavior-based intelligence model," Asia Pacific J. Mark. Logist., vol. 33, no. 1, pp. 99–123, 2021, doi: 10.1108/APJML-08-2019-0497.
[4] T. Schulz, M. Böhm, H. Gewald, and H. Krcmar, "Smart mobility – an analysis of potential customers' preference structures," Electron. Mark., vol. 31, no. 1, pp. 105–124, 2021, doi: 10.1007/s12525-020-00446-z.
[5] B. Wang, G. Wang, Y. Wang, Z. Lou, S. Hu, and Y. Ye, "A K-means clustering method with feature learning for unbalanced vehicle fault diagnosis," Smart Resilient Transp., vol. 3, no. 2, pp. 162–176, 2021, doi: 10.1108/srt-01-2021-0003.
[6] D. Marutho, S. Hendra Handaka, E. Wijaya, and Muljono, "The Determination of Cluster Number at k-Mean Using Elbow Method and Purity Evaluation on Headline News," in Proceedings - 2018 International Seminar on Application for Technology of Information and Communication: Creative Technology for Human Life, iSemantic 2018, 2018, pp. 533–538, doi: 10.1109/ISEMANTIC.2018.8549751.
[7] K. Venkatachalam, V. P. Reddy, M. Amudhan, A. Raguraman, and E. Mohan, "An Implementation of K-Means Clustering for Efficient Image Segmentation," in 10th IEEE International Conference on Communication Systems and Network Technologies An, 2021, no. 2020, pp. 224–229, doi: 10.1109/csnt51715.2021.9509680.
[8] C. Baldassi, "Recombinator-k-means: An evolutionary algorithm that exploits k-means++ for recombination," in IEEE Transactions on Evolutionary Computation, 2022, no. c, pp. 1–13, doi: 10.1109/TEVC.2022.3144134.
[9] A. Aslam, U. Qamar, R. A. Khan, and P. Saqib, "Improving K-Mean Method by Finding Initial Centroid Points," in International Conference on Advanced Communication Technology, ICACT, 2020, vol. 2020, pp. 624–627, doi: 10.23919/ICACT48636.2020.9061522.
[10] C. Cai and L. Wang, "Application of improved k-means k-nearest neighbor algorithm in the movie recommendation system," in Proceedings - 2020 13th International Symposium on Computational Intelligence and Design, ISCID 2020, 2020, pp. 314–317, doi: 10.1109/ISCID51228.2020.00076.
[11] A.L Hananto, P. Assiroj, B. Priyanta, Nurhayati, A.Fauzi, A.Y. Rahman and S.S. Hilabi. "Analysis of Drug Data Mining with Clustering Technique Using K-Means Algorithm," in Journal of Physics: Conference Series, 2021, vol. 1908, no. 1, doi: 10.1088/1742-6596/1908/1/012024.
[12] A. M. A. Alan Fuad Jahwar, "Meta-Heuristic Algorithms For K-Means Clustering: A Review," Pjaee, vol. 17, no. 7, pp. 1–20, 2021.
[13] 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, 2018, vol. 336, no. 1, doi: 10.1088/1757-899X/336/1/012017.
[14] J. Zhou and D. Hu, "Applications of Improved Ant Colony Optimization Clustering Algorithm in Image Segmentation," TELKOMNIKA (Telecommunication Comput. Electron. Control., vol. 13, no. 3, p. 955, 2016, doi: 10.12928/telkomnika.v13i3.1803.
[15] A. Dwiastuti, A. Larasati, and E. Prahastuti, "The implementation of Customer Relationship Management (CRM) on textile supply chain using k-means clustering in data mining," MATEC Web Conf., vol. 204, 2018, doi: 10.1051/matecconf/201820404017.
[16] K. Yang and R. Miao, "Research on Improvement of Text Processing and Clustering Algorithms in Public Opinion Early Warning System," 2018 5th Int. Conf. Syst. Informatics, ICSAI 2018, no. Icsai, pp. 333–337, 2019, doi: 10.1109/ICSAI.2018.8599424.
[17] M. Cui, "Introduction to the K-Means Clustering Algorithm Based on the Elbow Method," Accounting, Audit. Financ., vol. 1, pp. 5–8, 2020, doi: 10.23977/accaf.2020.010102.
[18] D. M. SAPUTRA, D. SAPUTRA, and L. D. OSWARI, "Effect of Distance Metrics in Determining K-Value in K-Means Clustering Using Elbow and Silhouette Method," 2020, vol. 172, no. Siconian 2019, pp. 341–346, doi: 10.2991/aisr.k.200424.051.
[19] H. Xie, C.P. Lim, Y. Yu, C. Liu and H. Liu., "Improving K-means clustering with enhanced Firefly Algorithms," Appl. Soft Comput. J., vol. 84, 2019, doi: 10.1016/j.asoc.2019.105763.
[20] R. Nainggolan, R. Perangin-Angin, E. Simarmata, and A. F. Tarigan, "Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method," in Journal of Physics: Conference Series, 2019, vol. 1361, no. 1, doi: 10.1088/1742-6596/1361/1/012015.

