K-Means Clustering Optimization Using the Elbow Method and Early Centroid Determination Based on Mean and Median Formula

Edy Umargono[1,*], Jatmiko Endro Suseno[2], Vincensius Gunawan S.K[2]

[1]*Departement of Information System, Post Graduated School, Diponegoro University*
[2]*Departement of Physics, Faculty of Science and Mathematics, Diponegoro University*
[*] *Corresponding author:*edyumargono@ymail.com

ABSTRACT

The most widely used algorithm in the cluster partitioning method is the K-Means algorithm, K-Means is an iteration algorithm with the user determining the number of clusters that need to be grouped and determining the centroid for each cluster so that the level of similarity between members in one group is high while the level of similarity with members in other groups is very low. Historically K-Means is still the best grouping algorithm among other grouping algorithms with the ability to group a number of data with relatively fast and efficient computing time. The K-Means algorithm is widely implemented in various fields in industrial and scientific applications and is very suitable for processing quantitative data with numeric attributes, but there are still weaknesses in this algorithm. Weaknesses of the K-Means algorithm include determining the number of clusters based on assumptions and relying heavily on the initial selection of centroids to overcome this weakness, in this study, we propose the use of the elbow method to determine the best number of clusters and initials. Centroid determination based on average and median data. The results of this study indicate that using initial cluster center determination based on average data makes the number of iterations needed to achieve uniformity in clusters 22.58% less than initial random cluster determination and determining the best number of clusters using the elbow method makes the required iteration 25% less than using the number of other clusters.

Keywords: *clustering, optimization, elbow method, mean data.*

1. INTRODUCTION

Clustering is the process of grouping a set of data objects by dividing data into groups or clusters based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity so that objects in one cluster have high similarity but are very different from objects in other clusters [1]. Dissimilarities and similarities are assessed based on the attribute values describing the objects and often involve distance measures [2]. Clustering methods can generally be classified as partition method with mean or medoid to represent the cluster center with one level of grouping, a hierarchical method with several levels of grouping, density-based method, grid-based Method to handle spatial data. Judging from the characteristics, ease of implementation and computational performance for grouping small and medium data partition method with the average (mean) is the most effective and efficient [3]. The data to be grouped in this study are relatively small and of medium-size, so the partition-based grouping method using the average (mean) is the most suitable method used in this study.

The most widely used algorithm in the partition method is the K-Means algorithm, K-Means is an iteration algorithm with the user determining the number of clusters that need

to be grouped in the dataset and determining the centroid for each cluster [4] so that the level the similarity between members in one cluster is high whereas the level of similarity with members in other clusters is very low [5]. Historically K-Means is still the best clustering algorithm among other clustering algorithms with the ability to group a number of data with relatively fast and efficient computing time [6], [7], [8] and became one of the most important algorithms in the field of data mining [9]. K-Means algorithm is widely implemented in various fields in industrial and scientific applications [5]and is very suitable for processing quantitative data with numerical attributes, however, there are still weaknesses in this algorithm.

The weaknesses of the K-Means algorithm include determining the number of clusters based on assumptions and relying heavily on the initial selection of centroids to overcome these weaknesses. It is necessary to optimize, one of the popular cluster optimization methods is the Elbow method [6], [7], [8], [10]. This method is a visual method to test the consistency of the best number of clusters by comparing the difference of the sum of square error (SSE) of each cluster, the most extreme difference forming the angle of the elbow shows the best cluster number. In some of these studies, the focus is still on optimizing the determination of the best number of clusters

by the elbow method while the initial selection of cluster centers is still random. This allows the number of iterations to place objects in the cluster based on the center of the new cluster to be more numerous so that the achievement of similarity of patterns formed becomes longer.

The K-means clustering algorithm in this research will be implemented in the case study of mapping data of teaching staff in public schools in districts, cities in province of Central Java, with this grouping it can be seen which schools have excess teaching staff or lack of teaching staff so that they can be used as a basis for distribution teaching staff as an effort to equalize teaching staff placement in public schools in Central Java so that there are no problems with excess or lack of teaching staff, excessive concentration of teaching staff in certain areas, and aging teaching staff population in placement in major city centers because of the distribution of teaching staff which is uneven [11]. In this study, it is assumed that three main groups are formed that represent deficiency, sufficient and excess conditions, to improve the performance of the K-means algorithm in this study using the elbow method to evaluate the determination of the best number of clusters and combined with determining the initial cluster center by comparing the minimum value, the median value, the average and maximum object values of the results of the

comparison of this experiment are used as a determinant of the initial cluster center, so it is expected to reduce the number of iterations to achieve similarity in the formed cluster rather than using the initial random center cluster determination.

2. MATERIALS AND METHODS

2.1. Materials

The data used in this study is a sample recapitulation of state high school data in Central Java Province covering 16 of State Senior High School data in Semarang City, 11 of State Senior High School data Semarang Regency and 3 of State Senior High School data Salatiga City and its attributes including the name of the school, number of students, number of teachers, number of study groups, number of subjects, data obtained from http://sekolah.data.kemdikbud.go.id and the Office of Education and Culture of the Central Java Provincial Government, the data is downloaded in the form of CSV file. Statistics of the data used can be seen in **Figure 1.** and **Figure 2.**
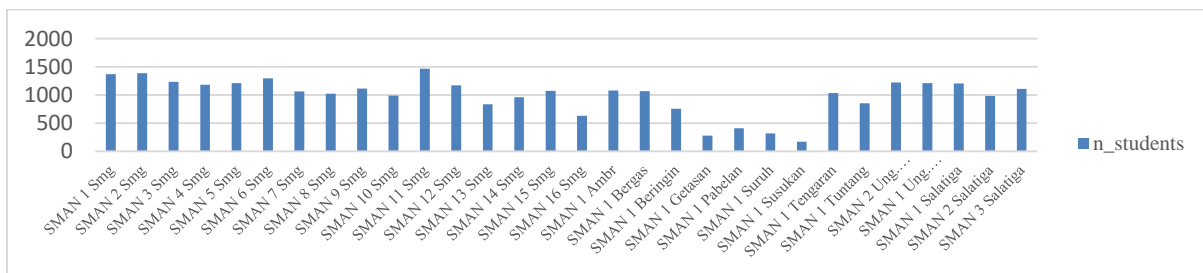

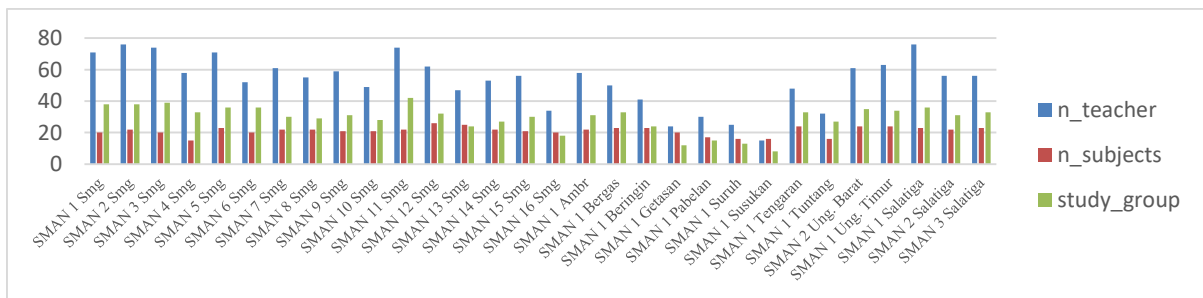**Fig. 1.** Data on the number of students at school

Figure 2. Data on the number of teachers, subjects and study group at school

The tools used in this study are a set of computers with AMD Dual Core A9-9420 3.6 GHz CPU hardware specifications, 4GB RAM with Windows 10 operating system, Excel applications, Orange Data Mining,

Python programming and Visual Studio Code editors as supporting software.

2.2. Methods

In this study conducted using the following stages, the first stage is problem analysis at this stage a problem analysis is

carried out in the case study of equal distribution of teaching staff, especially in public schools in districts,

cities in Central Java, which results in the formulation of the problem needed by mapping the teaching staff in public schools.

The second stage is the literature review, at this stage a literature study method is conducted which will be used to clustering data on existing case studies, the result of this stage is the use of the K-Means clustering method for grouping data that enables improved performance by determining the best number of clusters and determining the initial centroid.

The third stage is data collection. The data used are secondary data obtained from the website of the Ministry of Education and Culture of the Republic of Indonesia at the address http: //sekolah.data.kemdikbud.go.id and other data from the Office of Education and Culture of Central Java Province. Data attribute relationships are determined with the Pearson Correlation Coefficient are shown in **Table 1**[12].

Table 1. Pearson Correlation Coefficient P. Schober and L. A. Schwarte, "Correlation coefficients: *Anesth. Analg.*, vol. 126, no. 5, pp. 1763–1768, 2018.

Correlation Coefficient	Interpretation
0.00–0.10	Negligible correlation
0.10–0.39	Weak correlation
0.40–0.69	Moderate correlation
0.70–0.89	Strong correlation
0.90–1.00	Very strong correlation

The fourth stage is data preparation; at this stage, data cleaning is done from inconsistent data, incomplete data and then integrating data from different sources. To eliminate attribute dominance, normalization using the min-max method, which places data in the range of 0 as a minimum value up to 1 as a maximum value. With the following formula

$$X^* = \frac{x - minValue}{maxValue - minValue}$$

(1)

With x is the value to be normalized.

The fifth stage is modeling, at this stage using the K-Means algorithm to group data into a number of groups that have been determined. K-Means Clustering is a prototype based clustering method where the dataset is divided into (k) - clusters, in this method the user determines the number of clusters (k) that need to be grouped in the dataset. The purpose of K-Means Clustering is to find prototype data points for each cluster; all data points are then assigned to the nearest prototype, which then forms a cluster. The prototype is called centroid, the center of the cluster. Cluster center can be the average of all data objects in a cluster, such as in K-Means, or the most represented data object. K-Means clustering creates partitions (k) in n-dimensional space, where n is the number of attributes in the given dataset. To partition data, proximity measurements must be determined. The most commonly used measure for numerical attributes is the Euclidean distance. Following are the stages of the K-Means Clustering process [4]

1. The number of clusters k must be specified by the user

2. Centroid initialization

The first step in the k-means algorithm after determining the number of clusters is to start randomly determining centroids. In this study, it is proposed to use the mean formula as a description of the average value of the cluster members shown in formula

$$\mu_i = \frac{x_1 + x_2 + x_3 + \cdots + x_j}{j}$$

(2)

with x_j is object value in cluster and j number of objects in cluster, and the formula for the single data median as follows, if the data is odd and n is the amount of data then the medians formula becomes

$$ME = X_{(n+1)/2}$$

(3)

If the amount of data is even then the median formula becomes

$$ME = (X_{n/2} + X_{(n/2)})/2$$

(4)

Median data by dividing the smallest and largest addition values by formula

$$ME = \frac{minimum\ value + maximum\ value}{2}$$

(5)

2.3. Data Allocation

Every data or object will be located to the closest cluster. The proximity of two objects is determined based on the distance between the two objects. The closest network between one data and one particular cluster will determine

most common proximity measure, although other measurements such as the Manhattan size and Jaccard coefficient can be used. Euclidean distance (d) between two data points x (x1, x2, …xn) and c (c1, c2, ..., cn) with n attributes. Here is the Euclidean Distance formula

$$d = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \cdots + (x_n - c_n)^2}$$
(6)

with d is the distance of points x and c, x_n is the criterion data and c_n is the centroid of the nth cluster.

4. Calculate the new centroid

For each cluster, after the first iteration data allocation stage is completed then a new centroid is calculated for the next iteration data allocation. This new centroid is the most representative data point of all data points in the cluster obtained from the cluster average. Mathematically, this step can be expressed as minimizing the sum squared errors (SSE) from all data points in a cluster to the center of mass of the cluster. The overall goal of this step is to minimize
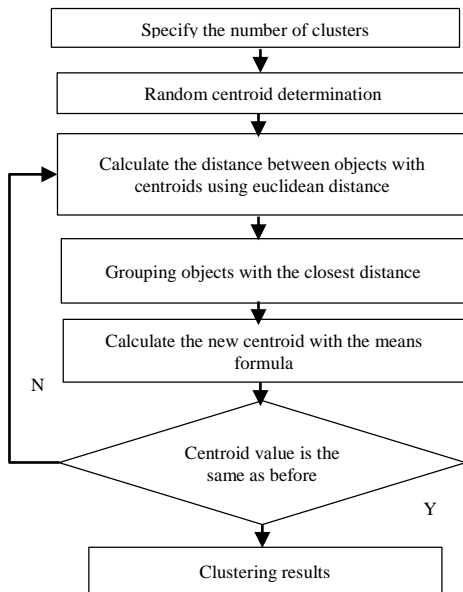
a data entry into the cluster which in this context the "closest" is calculated by the proximity measure. The Euclidean distance measurement is the

the SSE of each group. Here is the new centroid determination formula

$$\mu_i = \frac{1}{j_i} \sum_{x \in C_i} X$$
(7)

with μ_i is center point (centroid), X is object in the cluster, j_i the number of objects in the cluster.

5. Termination

The step of calculating a new centroid, and the step to assign data points to the new centroid are repeated until no more changes in the assignment of data points occur. In other words, no significant changes in centroids were noted. The final centroid is declared as a prototype cluster and is used to describe the whole grouping model. The stages of the original K-Means algorithm are illustrated in **Figure. 3.** and the proposed K-Means algorithm by determining the initial center of the cluster using the median formula and the mean formula illustrated in **Figure 4.**

Figure. 3. Original K-Means
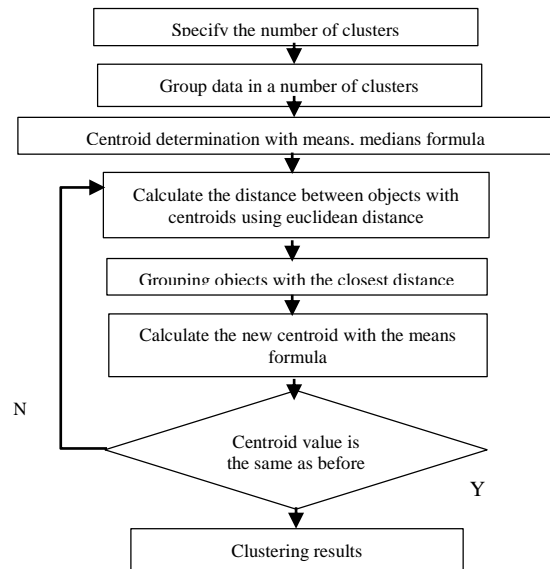
Figure. 4. Proposed K-Means algorithm optimization

The sixth stage is Evaluation, evaluate the results of the cluster using the Elbow method, this method is a visual method to test the consistency of the best number of clusters. The idea is to determine the number of clusters then add clusters, calculate the sum squared error (SSE) per cluster until the maximum number of clusters that have

been determined, then by comparing the difference SSE of each cluster, the most extreme difference forming the angle of the elbow shows the best cluster number.[7], [13], here is the SSE formula

$$SSE = \sum_{i=1}^{k} \sum_{x_j \in c_i} \left\| x_j - \mu_i \right\|^2$$

(8)

with x_j is object in the cluster c_i and centroid of the cluster

The Elbow method algorithm in determining the value of k in KMeans is as follows: a. Initialization $ k = 1 b. Starting c. $ k = $ k + 1 d. Calculate SSE e. SSE value has decreased dramatically f. The value of $ k at the time of a drastic decrease is the optimal k value g. completed. The complete research method is illustrated in the **Figure. 5.**
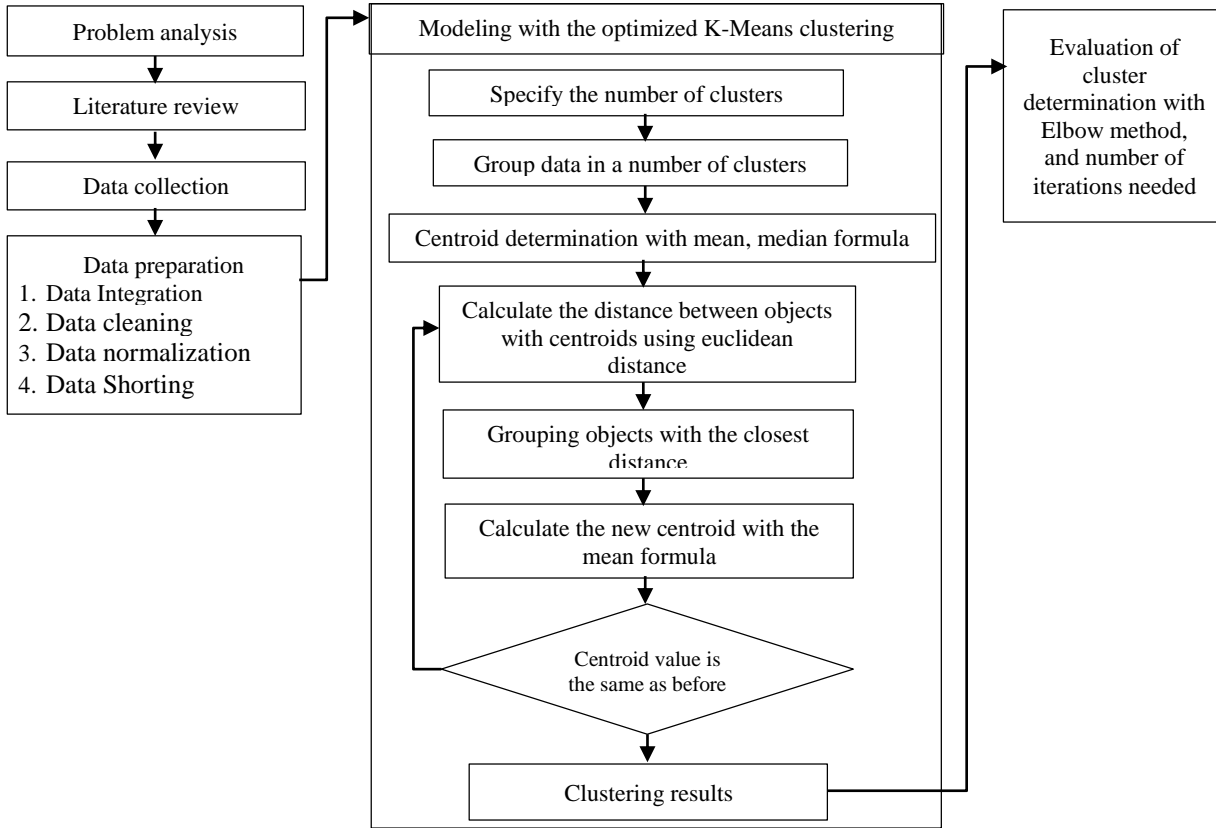

Figure 5. Research methods

3. RESULTS AND DISCUSSION

3.1. Result of Qualitative Analysis

In this study, the data preparation stage is performed feature correlation analysis using Pearson Correlation on data mining tools with the final result using four features with interpretation using the Pearson Correlation coefficient table [12], a comparison of correlations between features is presented in **Table 2.**

Table 2. Pearson Correlation Coefficient result.

Correlation Coefficient	Feature	Interpretation
0.986	n_student, n_studygroup	Very strong correlation
0.932	n_student, n_teacher	Very strong correlation
0.927	n_teacher, n_studygroup	Very strong correlation
0.523	n_subjects, n_teacher	Moderate correlation
0.486	n_subjects, n_student	Moderate correlation
0.476	n_subjects, n_studygroup	Moderate correlation

Base on **Table. 2**. it can be seen that features of the number of students with the number of study groups, the number of students to the number of teachers, the number of teachers with the number of study groups has a very strong correlation and the others has a moderate correlation. Base

on **Figure 1**. students features are too dominant over other feature so the data is normalized using the min-max normalization method with the following results in **Table 3.**

Table 3. Sample of data normalization results

school name	n_students	n_teacher	n_subjects	n_studygroup
SMAN 1 Susukan	0.000000	0.000000	0.090909	0.000000
SMAN 1 Getasan	0.084235	0.147541	0.454545	0.117647
SMAN 1 Suruh	0.115147	0.163934	0.090909	0.147059
SMAN 1 Pabelan	0.183153	0.245902	0.181818	0.205882
SMAN 16 Semarang	0.357032	0.311475	0.454545	0.294118
……………………	………..	………..	……….	………..
SMAN 11 Semarang	1	0,967213	0,636363	1

Before the clustering process is carried out with the K-Means algorithm the data is divided into a number of (k) clusters and the data is sorted from the smallest to the largest value, this is done so that the median and average group data can be calculated. The results of the K-means clustering algorithm are very dependent on the initial determination of the number of clusters to be formed and the selection of the initial centroids on grouping objects in the cluster, related to these problems in this study conducted experiments using the mean statistical formula, and the median formula to determine initial centroids compared to random initial centroid determination to test the effect of initial determination of centroids in the K-Means algorithm on the number of iterations needed to reach the data uniformity in the cluster formed. In the K-Means algorithm, the initial determination of the centroid is done by randomly selecting an object assuming each data object has a level of similarity to be chosen as the initial centroid. In this research, the results of three clusters that represent the excess teaching staff, the adequacy of the teaching staff and the lack of teaching staff are needed then a random value is used that refers to the initial, middle and end values to represent the distribution of existing data objects. The median formula is used in this experiment as a representation of the mean value of the data in the group assuming that the middle value is a counterbalance to the minimum and maximum values that exist in the data group, also conducted an experiment with the average formula (mean) with the assumption of representing all the values of objects in the group because the mean formula divides all the values of data objects in the group with the amount of data in the group. The initial centroid value is taken from the number of students as centroid 1, the number of teachers as centroid 2, the number of subjects as centroid 3 and the number of study groups as centroid 4, here is the centroid table of all clusters

Table 4. Early centroid base on a median value.

Cluster	Centroid 1	Centroid 2	Centroid 3	Centroid 4
1	0,405719	0,295082	0,454545	0,382353
2	0,696291	0,672131	0,636364	0,676471
3	0,817620	0,918033	0,636364	0,823529

Table 5. Early Centroid base on median 2 value.

Cluster	Centroid 1	Centroid 2	Centroid 3	Centroid 4
1	0,314915	0,336066	0,363636	0,338235
2	0,703246	0,663934	0,772727	0,647059
3	0,891036	0,836066	0,318182	0,867647

Table 6. Early centroid base on mean value.

Cluster	Centroid 1	Centroid 2	Centroid 3	Centroid 4
1	0,347295	0,339344	0,427273	0,350000
2	0,696676	0,662295	0,681818	0,676471
3	0,855564	0,862295	0,572727	0,844118

Table 7. Early centroid base on minimum value.

Cluster	Centroid 1	Centroid 2	Centroid 3	Centroid 4
1	0	0	0,090909	0
2	0,6329211	0,540984	0,545455	0,588235
3	0,7820710	0,606557	0	0,735294

Table 8. Early centroid base on maximum value.

Cluster	Centroid 1	Centroid 2	Centroid 3	Centroid 4
1	0,629830	0,672131	0,909091	0,676471
2	0,773570	0,770491	1	0,735294
3	1	1	0,818182	1

Table 9. Early centroid base on random1 value.

Cluster	Centroid 1	Centroid 2	Centroid 3	Centroid 4
1	0,5139103	0,52459	0,909091	0,470588
2	0,6885625	0,754098	0,636364	0,647059
3	0,9381761	1	0,636364	0,882353

Table 10. Early centroid base on random2 value.

Cluster	Centroid 1	Centroid 2	Centroid 3	Centroid 4
1	0,454405	0,426229	0,727273	0,470588
2	0,668470	0,540983	0,818182	0,735294
3	0,925039	0,918033	0,454545	0,882353

Determination of the number of clusters that are formed greatly affects the results of grouping data using K-Means clustering because this is the earliest stage of this algorithm if incorrectly determining the number of clusters that will be formed causes the number of iterations needed more and results in a bad cluster, in this case, it allows there empty cluster has no members.

3.2. Quantitative Analysis

In this study, the evaluation of determining the best number of clusters to be used in clustering uses the Elbow method which in principle observes the most drastic decrease in SSE that forms the angle of the elbow, the evaluation results of determining the best cluster number in this study are presented in **Figure 5.**
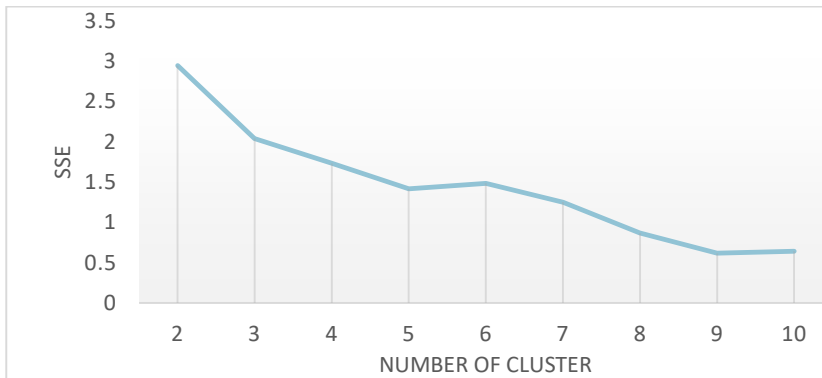

Figure 5. Determination of the best cluster using the Elbow method.

Graph in **Figure 5.** is arranged based on the summary results of the SSE total calculation experiment from each number of clusters in the range of experiments starting from 2 to 10 clusters which are presented in **Table 8.**

Table 8. Comparison table of the number of clusters with SSE and the number of iterations needed.

Number of clusters	SSE	Difference SSE	Number of iterations
2	2,945106	2,945106	4
3	2,001710	0,943396	3
4	1,733812	0,307676	5

5	1,416148	0,317664	4
6	1,484771	-0,06862	6
7	1,249919	0,234852	5
8	0,866852	0,383067	6
9	0,617018	0,249834	5
10	0,642253	-0,02524	5

Based on **Figure 5**. it can be observed that the most drastic decrease in SSE refers to the number of clusters 3 with SSE difference at 2 data points 0.943396 then the SSE gradually decreases without a drastic surge towards the lowest point forming an angular angle between clusters 2, 3 and 5 so that from observation in the graph forming a right angle, the number of clusters 3 is the best number of clusters to use. This is reinforced by the number of iterations needed to achieve the convergence of cluster members formed referring to the number of 3 clusters with 3 iterations needed or 25% less iteration than the number of iterations needed if using another cluster.

Statistics on the results of the K-Means clustering process with various experiments conducted are presented in the form of bar charts in **Figure 6.**
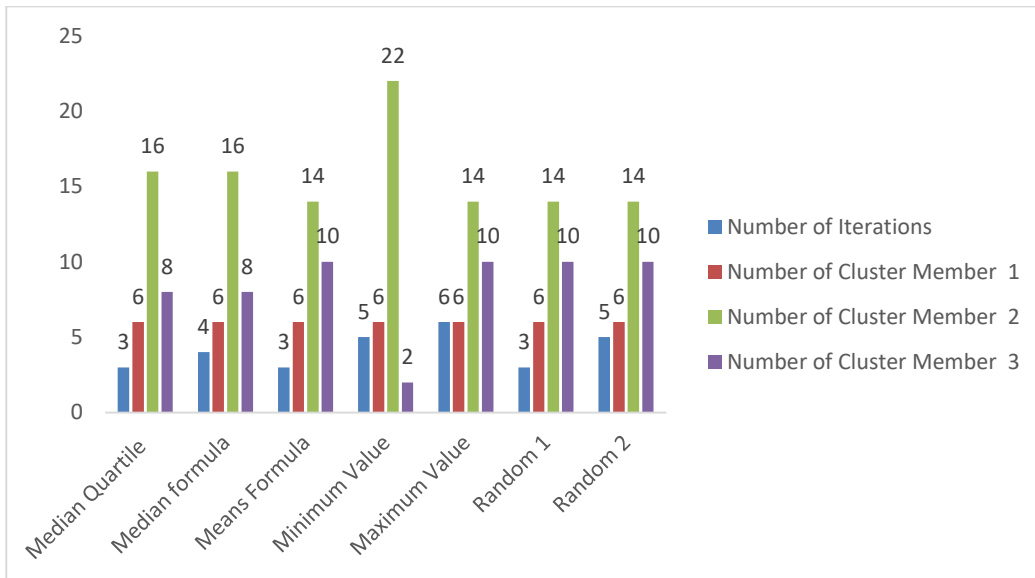

Figure 6. Results of K-Means clustering with various centroid value experiments.

Based on **Figure 6.** The results of clustering using the K-Means algorithm with the initial determination of centroid values using various formulas can be explained as follows, clustering K-Means using initial centroid values based on the median formula (5) requires 3 iterations with 6 members of cluster 1, 16 members of cluster 2 and 8 cluster members 3 while the median-based centroid value based on formula (6) requires the same iteration as the initial centroid based on formula average (3) but has a difference in the number of cluster members in clusters 2 and 3. The cluster member results use a minimum based value centroid is very much different in the number of cluster members 2 and 3 compared to the results of the cluster using other centroid values. Clustering uses centroid values based on maximum values, random 1 value, random 2 value produces the same number of members, namely 6 members in cluster 1, 14 members in cluster 2, 10 members in cluster 3, but requires a different number of iterations, centroid values based on maximum values is the one that requires the most iteration, as many as 6 iterations. Based on the results of the determination of experimental centroid clusters using the average value (average) requires at least iteration with 3 iterations or 22.58% less than the average iteration using random values and the number of cluster members produced is more stable compared to median and random values.

4. CONCLUSION

Based on the results of the clustering data experiment using the K-Means algorithm with optimization of the

determination of the best number of clusters using the Elbow method in terms of the iteration needed to achieve the convergence of the members of the cluster results 25% less efficient than using the number of other clusters and the determination of the initial cluster center using a flat formula - the average is more consistent in producing convergence of the number of members in the cluster as well as the most efficient number of iterations with 22.58% less than the iteration needed if using a random initial centroid determination.

REFERENCES

[1] Han, J., Kamber, M., Pei, J. (2012). Data Mining Concepts and Techniques (Third). Waltham. 1-38.

[2] Li, Y., Wu, H. (2012). 2012 International Conference on Solid State Devices and Materials Science A Clustering Method Based on K-Means Algorithm. Phys. Procedia. 25. 1104–1109.

[3] Wilks, D. S. (2011). Cluster Analysis. 100.

[4] Simovici, D. A., Djeraba, C. (2104). Clustering. 9781447164067.

[5] Shakeel, P. M., Baskar, S., Dhulipala, V. S., Jaber, M. M. (2018). Cloud Based Framework for Diagnosis of Diabetes Mellitus using K-means Clustering. Health Information Science and Systems. 6(1): 16.

[6] López-Rubio, E., Palomo, E. J., Ortega-Zamorano, F. (2018). Unsupervised Learning by Cluster Quality Optimization. Information Sciences. 436: 31-55.

[7] Bholowalia, P., Kumar, A. (2014). EBK-Means: A Clustering Technique Based on Elbow Method and K-Means in WSN. International Journal of Computer Applications. 105(9): 17-24.

[8] Kodinariya, T., Makwana, P. (2013). Review on Determining Number of Cluster in K-Means Clustering. Int. J. Adv. Res. Comput. Sci. Manag. Stud. 1(6): 90–95.

[9] Wu, X. (2008). Top 10 Algorithms in Data Mining. 14(1).

[10] Liu, L., Peng, Z., Wu, H., Jiao, H., Yu, Y., Zhao, J. (2018). Fast Identification Of Urban Sprawl Based On K-Means Clustering With Population Density And Local Spatial Entropy. Sustainability. 10(8): 2683.

[11] Szelągowska-Rudzka, K. (2018). Human Resources Management in Higher Education Institutions in Poland. Management. 22(1): 208-225.

[12] Schober, P., Boer, C., Schwarte, L. A. (2018). Correlation Coefficients: Appropriate Use and Interpretation. Anesthesia & Analgesia. 126(5): 1763-1768.

[13] Madhulatha, T. S. (2012). An Overview on Clustering Methods. 2(4): 719–725