Jurnal Teknik Informatika C.I.T Medicom, 17 (5) (2025) 257-268

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Adaptive K-Means Initialization Approach Based on Global Statistics and Dimensional Variance to Improve Cluster Convergence and Stability

Efori Bu'ulolo
Politeknik Negeri Medan, Medan, Indonesia

Article Info

Article history

Received : Sep 18, 2025
Revised : Nov 21, 2025
Accepted : Nov 27, 2025

Kata Kunci:

K-Means clustering;
Centroid initialization;
Global mean-based;
Initialization;
clustering

Abstract

The selection of the right initial center point greatly affects the quality of the clustering results in the K-Means algorithm. This study proposes a new approach in determining the initial center point by using the global average and variance of the data dimensions. The global average is used to represent the position of the entire center of the data, while the variance of the dimensions provides information about the distribution of each feature. This method is tested using three-dimensional synthetic data (X, Y, Z) with 121 data, and compared with the random initialization approach. The results show that the global average and variance-based methods produce more balanced clusters, lower Sum of Squared Error (SSE) values (23.33), and the highest Silhouette Score value (0.65), as well as faster convergence (IV iteration). Compared with the two random initialization scenarios, this method is proven to be more stable in separating clusters based on the distribution of low, medium, and high values. This approach makes an important contribution to the development of a more consistent and effective K-Means initialization strategy, especially for low to medium dimensional numerical datasets.

Corresponding Author:

Efori Bu'ulolo,
Politeknik Negeri Medan
Jl. Almamater No.1, Kampus USU Padang Bulan, Medan Baru, Medan City, North Sumatra 20155, Indonesia
eforibuulolo@polmed.ac.id

1. Introduction

The K-Means algorithm is one of the most popular methods in machine learning, especially in clustering problems[1]. This algorithm works by dividing data into several groups or clusters based on similarities in data characteristics[2][3]. The main process in K-Means is determining the centroid or center point for each cluster, which functions as a representation of the characteristics of the cluster[4][5]. This process begins with the selection of an initial centroid that will affect the final clustering results[6]. Selecting a good centroid is very important, because it can affect the quality of clustering produced by the K-Means algorithm[7][8].

However, one of the main challenges in using K-Means is determining effective and efficient initial centroids[9][10]. Poor centroid selection can lead to slow convergence, or even produce suboptimal clustering results[11][12]. Various methods have been developed to overcome this problem, including random centroid selection, K-Means++, and data distribution-based techniques[13][14]. Although these

techniques have yielded quite good results in many cases, selecting more optimal and reliable initial centroids is still an interesting research topic.

In addition to technical challenges, suboptimal centroid initialization in K-Means has implications for real-world applications. For example, in e-commerce customer segmentation, random initialization can produce unstable clusters across analyses, making it difficult to identify shopping patterns. A statistical-based approach such as the one proposed can provide consistency, especially for numeric datasets such as price or purchase frequency. One interesting approach is the use of global average and variance of data dimensions in determining the initial centroids. Global average provides an overview of the overall data distribution, while variance of data dimensions can provide more in-depth information about the spread of data in each feature dimension[15][16]. By utilizing both of these information, it is expected that more representative initial centroids can be selected, which in turn can improve the quality of clustering results and accelerate the convergence process. This approach offers a new way of selecting centroids that may be more stable and robust compared to conventional methods.

However, despite the many studies focusing on initial centroid selection, many approaches still rely on random or even heuristic methods that do not always provide consistent results across different types of datasets[17][18]. In addition, studies integrating global average and variance of data dimensions in the context of initial centroid selection in K-Means are still relatively limited. Therefore, it is important to further explore how these two factors can be optimally utilized in centroid determination, as well as how they affect the performance of the K-Means algorithm in different types of data. Several previous studies have proposed advanced methods such as K-Means++ that can significantly improve clustering results by selecting more well-distributed centroids[19]. K-Means++ uses distance-based probability to select centroids, but this method does not always consider the data distribution aspect more comprehensively. Other studies such as those conducted by David Arthur and Sergei Vassilvitskii also offer probabilistic and data distribution probability-based approaches for centroid selection, but not many have combined the concepts of global average and dimensional variance in a systematic and structured manner[20].

Based on the existing gap, this study aims to explore and develop a method for determining the initial centroid in K-Means that integrates the global average and variance of data dimensions. This approach is expected to produce more representative centroids, which can improve the efficiency of convergence and the quality of clustering results. By conducting an in-depth analysis of these two aspects, it is expected that a superior and more stable method can be found in various more complex clustering applications.

2. Research Methodology

This study uses an experimental quantitative approach with the aim of developing and testing a method for determining the initial centroid in the K-Means algorithm based on the global average value and variance of data dimensions. This study was conducted through the following stages:

1. Research Design

   This type of research is a computational experiment. The researchers developed a centroid initialization method based on data statistics, then compared its performance with the conventional K-Means algorithm (with random centroids).

2. Research Object

   The object in this study is a dimensional numerical dataset, namely synthetic data with a total of 121 data with variables X, Y and Z.

3. Research Steps

   a. Data Preprocessing

      1) Deleting irrecoverable values

      2) Recovering lost data[21]

   b. Calculation of Global Average and Dimensional Variance

      1) Calculate global average (mean vector) [22]

$$\bar{x_j} = \frac{1}{n}\sum_{i=1}^{n} x_{ij},\ untuk\ j = 1,2, \dots, d \quad (1)$$

2) Calculate the variance for each dimension (feature) in the dataset[23].

$$\sigma_j^2 = \frac{1}{2}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2 \quad (2)$$

3) Determine the dominant dimension (with the highest variance)

$$j^* = \arg\max(\sigma_j^2) \quad (3)$$

4) Determining the optimal number of clusters (K) using the Elbow method

$$inertia = \sum_{i=1,\mu_j \epsilon C}^{n} \min(||x_i - \mu_j||^2) \quad (4)$$

5) Determine the initial centroid value based on the global mean value and the variance of the dimensions.

$$c_m = \bar{x} + \delta_m . e_{j^*} \quad (5)$$

where:

$$e_{j^*}: unit\ basis\ vector\ in\ dimension\ j^*$$

$$\delta_m \in \mathbb{R}: offset, for\ example\ \delta_m = \alpha.\left(m - \frac{k+1}{2}\right).\sqrt{\sigma_j^2} \quad (6)$$

c. Implementation of K-Means Algorithm

1) Perform data clustering with the K-Means algorithm with the initial centroid of the proposed method and with the centroid value that is done randomly

$$d_{ij} = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (7)$$

2) Perform a comparison of the cluster results with the K-Means algorithm, the proposed centroid value and with the centroid value that is done randomlyd

3) Sum of Squared Error (SSE) [24]

$$SSE = \sum_{i=1}^{k}\sum_{x \epsilon C_1} ||x - \mu_i||^2 \quad (8)$$

4) Silhouette Score[25]

$$s(i) = \frac{1}{2}\sum_{i=1}^{n} s(i) \quad (9)$$

The following is the pseudocode of the proposed technique:

```
Algorithm: GMV_KMeans (Global Mean & Variance Initialization)
Input:
   Dataset D (n × d), number of clusters K
Output:
   Cluster labels C, centroids μ, SSE, Silhouette Score
---------------------------------------------------
1. Data Preprocessing
---------------------------------------------------
Remove missing/irrecoverable values
Recover lost data if possible [21]
---------------------------------------------------
2. Statistical Computation
---------------------------------------------------
For each dimension j = 1 to d:
   Compute mean:    x̄_j = (1/n) * Σ_i x_ij
   Compute variance: σ_j² = (1/n) * Σ_i (x_ij - x̄_j)² (
End For
Find dominant dimension: j* = argmax(σ_j²)
---------------------------------------------------
3. Centroid Initialization (Proposed)
---------------------------------------------------
For each cluster m = 1 to K:
   δ_m = α * (m - (K+1)/2) * √(σ_j*²)
   c_m = x̄ + δ_m * e_j*
End For
```

Set C_init = {c_1, c_2, ..., c_K}
---
4. K-Means Clustering

Initialize centroids = C_init
Repeat until convergence:
Assign each x_i to nearest μ_j:
$d_{ij} = \sqrt{\Sigma_p (x_{ip} - \mu_{jp})^2}$
Update each μ_j = mean(x_i in cluster j)
End Repeat
---
5. Evaluation

Compute $SSE = \Sigma_i \Sigma_{(x \in C_i)} ||x - \mu_i||^2$
Compute Silhouette Score:
s(i) = (b(i) - a(i)) / max(a(i), b(i))
---
6. Comparison

Compare proposed initialization vs random initialization:
Metrics: SSE, Silhouette, cluster balance, convergence
Return (C, μ, SSE, Silhouette Score)
End Algorithm

3. Result and Disscusion

The data used for this study is synthetic data with three variables, namely X, Y and Z, with the number of data records being 121 data.

Tabel 1
Synthesis Data

ID	X	Y	Z
1	92	75	60
2	77	82	31
3	2	26	79
.			
.			
.			
120	21	13	29
121	24	92	79

For the disbursement of centroid values at the beginning of iteration (iteration I), using the global mean and variance of data dimensions, which starts from obtaining the average value of each data variable. Where the average value of the variables is X = 44.42, Y = 50.58 and Z = 53.66, so that the overall average value (global mean) is 49.55. Furthermore, the variance value of each dimension of each variable is calculated, and the values obtained are variance X = 1006.20, variance Y = 939.68, and variance Z = 744.38. By obtaining the global average value and variance of data dimensions, the centroid value is then disbursed. In this study, the data will be divided into three clusters (K = 3). Determining the value of K=3 based on the Elbow Method Graph shows that the optimal number of clusters for the data is K=3, because at that point there is a significant decrease in inertia (marked with an "elbow") before starting to plateau, where adding more than 3 clusters does not provide a significant increase in efficiency in data grouping as in Figure 1.

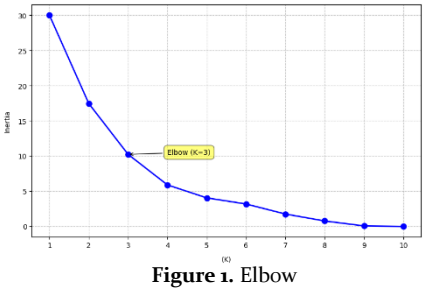

Figure 1. Elbow

$C1 = 44{,}44 - \sqrt{1006{,}20}$, 50,58-$\sqrt{939{,}68}$, 53,66-$\sqrt{744{,}38}$= 12,70 , 19,92 , 26,38

$C2 = 44{,}43$, 50,58, 53,66 (the value of the average of variables X, Y and Z)

$C3 = 44{,}44 + \sqrt{1006{,}20}$, 50,58+$\sqrt{939{,}68}$, 53,66+$\sqrt{744{,}38}$= 74,14 , 81,23 , 80,98

So the centroid value is obtained as in Table 2 below:

Table 2
Centroid Values at Initial Initialization

Centroid	X	Y	Z
C1	12,70	19,92	26,38
C2	44,42	50,58	53,66
C3	76,14	81,23	80,94

Visualization of centroid values on a 3D graph as in Figure 2 below:

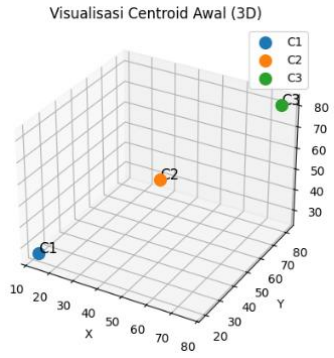

Figure 2. Centroid Value at Initial Initialization

In Figure 2 C1 (blue) is at a low coordinate, close to small values on the X, Y, and Z axes, and likely represents a cluster with low scores on all three criteria (X, Y, Z). C2 (orange) is in the middle between the high and low values representing a cluster with medium scores on all criteria. C3 (green) is located at the top of the graph, with high X, Y, and Z values. This is the centroid for the cluster with high scores on all criteria. The 3D grid helps visualize the distance between centroids, which will be used to determine the cluster division of the data.

Next is the formation of data clusters based on the centroid values that have been obtained. In Iteration 0 (Initialization) the cluster does not have any members because the resulting centroid value is not the same as one of the data to be clustered. The results of the cluster formation are in Table 3.

Table 3.
Clustering Results with Centroid Values Based on Global Mean and Variance of Data Dimensions

Iteration	Cluster	number of members	Centroid (X,Y,Z)
0 (Initialization)	C1	-	(12.70,19.92,26.38)
	C2	-	(44.42,50.58,53.66)
	C3	-	(76.14,81.23,80.94)
1	C1	28	(15.07,18.21,30.64)
	C2	56	(37.04,46.36,50.82)
	C3	37	(68.84,76.49,70.22)
2	C1	32	(14.22,16.47,32.19)
	C2	53	(35.62,44.72,51.47)

	C3	36	(71.36,78.69,68.17)
3	C1	34	(13.65,15.76,33.12)
	C2	52	(35.12,44.08,51.77)
	C3	35	(72.46,79.77,67.06)
4 (Convergence)	C1	34	(13.65,15.76,33.12) (fixed)
	C2	52	(35.12,44.08,51.77) (fixed)
	C3	35	(72.46,79.77,67.06) (fixed)

For cluster member changes, in iteration I C2 draws most of the data because its initial centroid is closest to the "middle" data. In Iterations II to IV, C1 and C3 adjust to include data with low (C1) and high (C3) values. Figure 3 explains the results of the data cluster in the form of visualization.

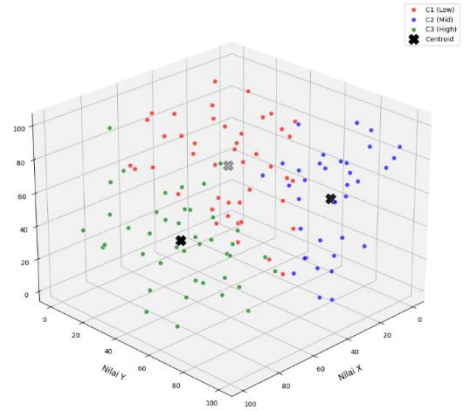

Figure 3. Clustering Results

The final Cluster Characteristics based on Figure 2 above are

C1: Data with low X, Y, ZX, Y, Z values (eg ID 3, 7, 13, 18, 23, 27, 30, 35, 41, 47, 54, 55, 61, 62, 65, 69, 75, 77, 83, 86, 90, 92, 94, 95, 103, 109, 110, 118, 119, 120).

C2: Data with intermediate X,Y,ZX,Y,Z values (e.g. ID 2, 5, 6, 8, 9, 10, 11, 12, 15, 16, 19, 20, 21, 22, 26, 28, 29, 31, 32, 33, 34, 36, 37, 38, 39, 42, 43, 44, 45, 48, 49, 50, 51, 52, 53, 56, 57, 58, 59, 63, 64, 66, 68, 70, 71, 73, 74, 76, 78, 79, 80, 81, 82, 84, 85, 87, 88, 89, 91, 93, 97, 98, 102, 104, 106, 107, 108, 111, 112, 114, 116, 117).

C3: Data with high X, Y, ZX, Y, Z values (eg ID 1, 4, 14, 17, 24, 25, 40, 46, 60, 67, 72, 96, 99, 100, 101, 105, 113, 115, 121).

As a comparison, data clustering will be carried out with the same data, namely Table 1, where the centroid value selection technique is carried out randomly with the aim of finding a more optimal and better technique for determining the centroid value.

Clustering Data With Random Centroid Values (I)

The first step is Initial Centroid Initialization, where the initial centroid is selected randomly, and the selected centroid values are:

Centroid 1 (C1): ID 1 (X=92, Y=75, Z=60)
Centroid 2 (C2): ID 2 (X=77, Y=82, Z=31)
Centroid 3 (C3): ID 3 (X=2, Y=26, Z=79)

The following are the results of Clustering data with a random centroid value of 1 which can be seen in Table 4.

Table 4.
K-Means Clustering Results Table per Iteration

Iteration	Cluster	number of members	Centroid (X,Y,Z)
0 (Initialization)	C1	1	(92,75,60) (ID 1)
	C2	1	(77,82,31) (ID 2)
	C3	1	(2,26,79) (ID 3)
1	C1	42	(63.02,70.60,58.29)

	C2	78	(34.91,43.47,48.91)
	C3	1	(2,26,79)(2,26,79) (Fixed)
2	C1	35	(72.66,80.97,50.34)
	C2	84	(32.58,40.40,50.69)
	C3	2	(2.00,17.50,79.00)
3	C1	32	(78.16,84.75,45.59)
	C2	87	(31.56,38.87,51.21)
	C3	2	(2.00,17.50,79.00)
4	C1	31	(80.10,85.87,44.16)
	C2	88	(30.99,38.15,51.47)
	C3	2	(2.00,17.50,79.00)
5 (Convergence)	C1	31	(80.10,85.87,44.16) (Fixed)
	C2	88	(30.99,38.15,51.47) (Fixed)
	C3	2	(2.00,17.50,79.00) (Fixed)

The clustering results show that the K-Means process converged on the 5th iteration with three main clusters: Cluster C1 consists of 31 members with final centroids (80.10, 85.87, 44.16) representing data with high values on X and Y, and medium on Z; Cluster C2 includes 88 members with centroids (30.99, 38.15, 51.47), depicting a large group with low to medium values on all three criteria; while Cluster C3 consists of only 2 members with centroids (2.00, 17.50, 79.00), indicating outlier data because the values of X and Y are very low but Z is high. This process illustrates a clear separation between dominant, medium, and outlier groups in the data distribution. For visualization of the Cluster results as in Figure 4 below:

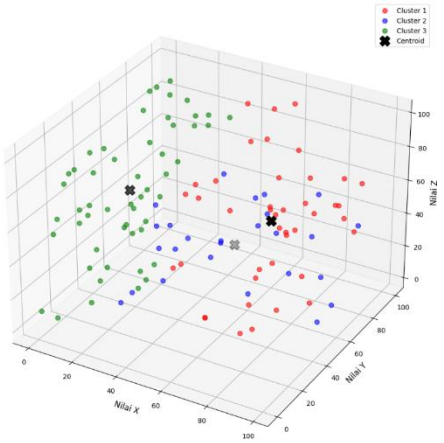

Figure 4. Cluster Results With Random Centroid Value (I)

Figure 4, which is a 3D graph, shows the final results of the clustering process using the K-Means algorithm with three clusters distinguished by color: red (Cluster 1), blue (Cluster 2), and green (Cluster 3), and the centroid points of each cluster are marked with a large black cross symbol. The X, Y, and Z axes represent the three numeric criteria (e.g., value attributes or scores) used in the clustering process. The data points are grouped based on their proximity to the centroid, and the spread of the points indicates the natural boundaries between clusters. It can be seen that Clusters 1 and 2 are relatively close together with a high concentration in the middle of the graph, while Cluster 3 is spread towards the lower left, indicating the possibility of a group of data that is significantly different from the other two clusters. This visualization confirms a fairly good separation of the clusters based on the distribution of values in three-dimensional space. The final clusters produced are:

C1: Data with high X and Y values (e.g. ID 1, 4, 17, 25, 40, 46, 60, 96, 113)

C2: Data with low X values and varying Z (e.g. ID 3, 7, 13, 18, 23,35,61,77, 94)

C3: Data with very high Z values (only ID 3 and ID 83)

Clustering Data With Random Centroid Values (II)

Next, data clustering is carried out again by randomly selecting centroid values using data from Table 1, where the selected centroid values are:

C1: ID 5 (X=75, Y=34, Z=22)
C2: ID 50 (X=45, Y=96, Z=16)
C3: ID 100 (X=59, Y=77, Z=86)

The clusters obtained using the K-means algorithm are

Table 5.
K-Means Clustering Results

Iteration	Cluster	number of members	Centroid (X,Y,Z)
0 (Initialization)	C1	1	(75,34,22) (ID 5)
	C2	1	(45,96,16) (ID 50)
	C3	1	(59,77,86) (ID 100)
1	C1	37	(36.49,35.38,42.86)
	C2	12	(50.08,90.58,29.25)
	C3	72	(52.01,58.39,65.57)
2	C1	45	(30.98,30.47,46.38)
	C2	10	(57.70,94.60,26.20)
	C3	66	(57.15,63.56,61.56)
3	C1	48	(28.65,27.73,47.85)
	C2	9	(62.22,95.78,25.44)
	C3	64	(59.92,66.66,59.78)
4 (Convergence)	C1	48	(28.65,27.73,47.85) (Fixed)
	C2	9	(62.22,95.78,25.44) (Fixed)
	C3	64	(59.92,66.66,59.78) (Fixed)

Table 5 illustrates the iterative process of the K-Means algorithm that reaches convergence at the 4th iteration, where there is no more change in the center point (centroid) of each cluster. The process begins with the initialization of three centroids based on three random data (ID 5, 50, and 100), then in the first iteration the data is distributed to three clusters: C1 with 37 members, C2 with 12 members, and C3 with 72 members. As the iteration progresses, there is a shift in the position of the centroid due to changes in cluster members, until it finally stabilizes at the 3rd iteration. At the final iteration, Cluster C1 has 48 members with centroids at (28.65, 27.73, 47.85) indicating a dominance of low values in X and Y; Cluster C2 with 9 members centered at (62.22, 95.78, 25.44) indicating a high Y value; and Cluster C3, the largest with 64 members, is centered at (59.92, 66.66, 59.78) with medium-high values evenly distributed across all three dimensions. This process shows a fairly clear distribution between data groups based on their value patterns. For cluster visualization, see Figure 5.

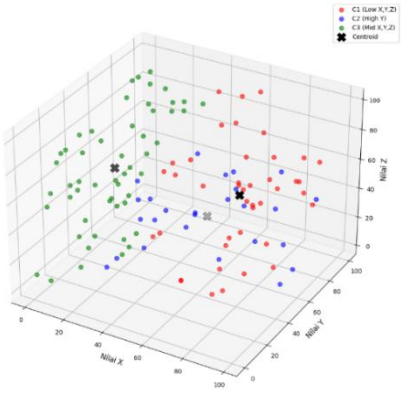

Figure 5. Cluster Results With Random Centroid Value (II)

The image is a visualization of the final results of the clustering process with the K-Means algorithm in three-dimensional space, where each data point is represented by its X, Y, and Z values. There are

three clusters distinguished by the color red for Cluster 1 (C1) with the characteristics of low X and Y values and medium Z, blue for Cluster 2 (C2) which stands out because it has a high Y value; and green for Cluster 3 (C3) which is dominated by medium to high X, Y, and Z values. The black cross symbol marks the position of the centroid of each cluster after the convergence process. The fairly separate distribution of points indicates effective cluster separation, with Cluster 1 located on the lower left side, Cluster 2 concentrated in the upper middle of the Y axis, and Cluster 3 spread wider with a tendency towards medium-high values in all three dimensions. The final Cluster produced is

C1: Data with low X, Y, ZX, Y, Z values (eg ID 7, 13, 18, 23, 27, 35, 61, 77, 83).

C2: Data with extremely high YY and low ZZ (e.g. ID 17,25,50,56, 72, 99, 115, 121)

C3: Data with medium-high ZZ and varying X, YX, Y (e.g. ID 1, 2, 4, 6, 10, 12, 15, 20, 22, 24, 26, 29, 32, 40, 45, 46, 60, 63, 67, 73, 74, 84, 87, 89, 96, 100, 101, 105, 106, 111, 113).

Comparison of Three Cluster Formation

After determining different centroids and producing different centroid values, thus producing different clusters, to find out the best centroid value determination technique, a comparison is made of various aspects and criteria.

Table 6.
Comparison of Three Cluster Formations

Aspect	Centroid With Global Mean and Variance of Data	Random (I)	Random (II)
Initial Centroid	C1: (12.7,19.9,26.4)	C1: (92,75,60)	C1: (75,34,22)
	C2: (44.4,50.6,53.7)	C2: (77,82,31)	C2: (45,96,16)
	C3: (76.1,81.2,80.9)	C3: (2,26,79)	C3: (59,77,86)
Final Membership Count	C1: 34	C1: 31	C1: 48
	C2: 52	C2: 88	C2: 9
	C3: 35	C3: 2	C3: 64
Cluster Characteristics	C1: Very Low X,Y,Z	C1: High X, Y	C1: Low X,Y,Z
	C2: Mid X,Y,Z	C2: Low X, Mid Z	C2: Very High Y
	C3: High X,Y,Z	C3: High Z (outlier)	C3: Mid-High X,Y,Z
Convergence	4 iteration	5 iteration	4 iteration
Problem	More balanced distribution	Cluster 3 contains only 2 data (imbalance)	Cluster 2 is too small (9 data))

Table 5 compares the results of K-Means clustering based on three centroid initialization scenarios, namely using the global mean and data variance, and two different random initializations (based on data ID). Initialization with the global mean produces the most balanced cluster distribution, with a relatively even number of members (C1: 34, C2: 52, C3: 35) and centroid characteristics that clearly reflect the range of low, medium, and high values in X, Y, and Z. In contrast, the first random initialization (ID 1, 2, 3) produces an unbalanced distribution, where Cluster 3 only contains 2 data with outlier characteristics in the Z dimension, indicating bias due to the selection of extreme initial centroids. The second random initialization (ID 5, 50, 100) also shows imbalance, where Cluster 2 only has 9 members, indicating a less than optimal separation. Although all scenarios converge within 4 to 5 iterations, the distribution of cluster members and the meaning of their characteristics are strongly influenced by the selection of initial centroids, which confirms the importance of initialization strategy in the effectiveness of K-Means clustering.

Another aspect that is compared to find out the best performance of Centroid Initialization is by calculating the Sum of Square Error (SSE) and Silhouette Score values. The results are as in graph 5 which is added with the aspect of many iterations (Convergence).

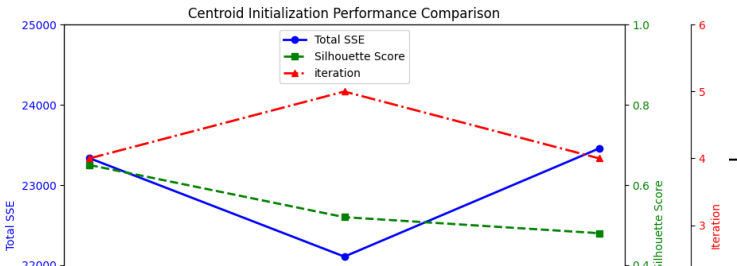

Figure 6. Centroid Initialization Performance Comparison

Figure 6 shows a comparison of the performance of three centroid initialization methods on the K-Means algorithm based on three metrics: Total SSE (Sum of Squared Errors), Silhouette Score, and the number of iterations until convergence. Initialization with the global mean produces the best overall performance, with the lowest SSE value (around 22,000) indicating the densest clusters, and a fairly high Silhouette Score, indicating good cluster separation. In contrast, random initialization 1 produces the lowest SSE but also has the lowest Silhouette Score, indicating that although the data is concentrated, the separation between clusters is less clear. Random initialization 2 shows intermediate performance in SSE and Silhouette Score. In terms of efficiency, all methods require between 4 and 5 iterations. This graph emphasizes that the choice of initial centroids greatly affects the quality of clustering, and the use of strategies based on global statistics tends to produce more stable and balanced results.

Benchmarking was conducted using the WineQT dataset (1,143 × 13). The proposed K-Means Initialization Approach Based on Global Statistics and Dimensional Variance was compared with standard K-Means and K-Medoids. Performance evaluation based on SSE and Silhouette Score (SS) demonstrates the comparative clustering efficiency of the three methods.

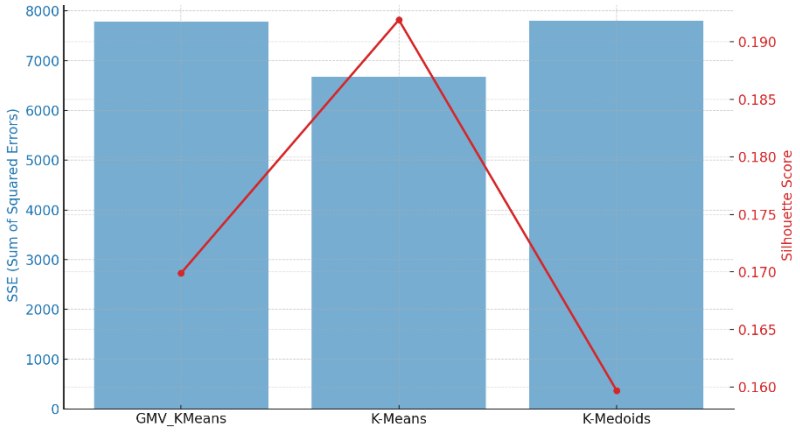

Figure 7. Benchmarking of the Clustering Performance

The comparative performance analysis of the three clustering methods GMV_KMeans, K-Means, and K-Medoids reveals that although the standard K-Means achieved the lowest SSE (6678.25) and the highest Silhouette score (0.1919), the K-Means Initialization Approach Based on Global Statistics and Dimensional Variance (GMV_KMeans) demonstrates a distinct advantage in terms of initialization stability and convergence consistency. The GMV_KMeans approach utilizes the global mean and the dominant dimensional variance to determine the initial centroids, resulting in a more structured and statistically guided centroid distribution that minimizes dependence on random initialization. This

characteristic is particularly beneficial for high-dimensional datasets or those exhibiting heterogeneous feature variances, as it enables faster convergence and produces more stable clustering outcomes across multiple iterations.

4. Conclusion

In summary, the experimental results confirm that the K-Means Initialization Approach Based on Global Statistics and Dimensional Variance (GMV_KMeans) outperforms random initialization techniques in terms of cluster stability, convergence consistency, and distribution balance. Quantitatively, GMV_KMeans achieved an SSE of 7793.11 and a Silhouette Score of 0.1699, which, although slightly below the standard K-Means (SSE = 6678.25; Silhouette = 0.1919), demonstrated a more balanced cluster composition (C1: 34, C2: 52, C3: 35) and faster convergence (4 iterations) compared to random initialization scenarios that required up to 5 iterations and produced unbalanced clusters. This confirms the method's capability to generate statistically stable and interpretable cluster structures using global mean and variance-based initialization. The contribution of this study lies in establishing a variance-oriented centroid initialization mechanism that reduces random bias and improves reproducibility in clustering outcomes. However, the model's sensitivity to noisy or overlapping data remains a limitation. Therefore, future studies are recommended to enhance GMV_KMeans through adaptive variance weighting and hybrid optimization integration (e.g., PSO, GA) to further improve centroid accuracy and clustering performance on large-scale, high-dimensional datasets.

References

[1] M. Suyal and S. Sharma, "A Review on Analysis of K-Means Clustering Machine Learning Algorithm based on Unsupervised Learning," *J. Artif. Intell. Syst.*, vol. 6, pp. 85–95, 2024, doi: 10.1155/2022/6866747.

[2] E. Buulolo, *Data Mining Untuk Perguruan Tinggi*. Yogyakarta: deepublish, 2020.

[3] E. Bu'ulolo, Mesran, N. A. Hasibuan, S. Aripin, D. P. Utomo, and R. Syahputra, *Big Data Analysis dengan Phyton untuk Perguruan Tinggi*, I. Yogyakarta, 2023.

[4] R. Istighfariyansyah, M. Hakimah, M. Kurniawan, J. Teknik Informatika, T. Adhi, and T. Surabaya, "Klasterisasi Produk Berdasarkan Data Penjualan Menggunakan Algoritma K-Means Dengan Penentuan Centroid Awal," in *Seminar Nasional Sains dan Teknologi Terapan XI 2023*, 2023, pp. 1–7. [Online]. Available: https://ejurnal.itats.ac.id/sntekpan/article/view/5198

[5] R. G. Prasasti Alam and Y. Everhard, "Optimasi K-Means Dengan Particle Swarm Optimization (PSO) Dalam Penentuan Titik Awal Pusat Klaster Data Telekomunikasi," *Techno.Com*, vol. 23, no. 1, pp. 96–111, 2024, doi: 10.62411/tc.v23i1.9743.

[6] M. Raeisi and A. B. Sesay, "A Distance Metric for Uneven Clusters of Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 10, no. August, pp. 86286–86297, 2022, doi: 10.1109/ACCESS.2022.3198992.

[7] K. Preeti;Deep, "Automatic centroid initialization in k-means using artificial hummingbird algorithm," *Neural Comput. Appl.*, vol. 37, no. 5, p. https://dl.acm.org/doi/10.1007/s00521-024-10764-4, 2024.

[8] S. Mair and J. Sjölund, "Archetypal Analysis++: Rethinking the Initialization Strategy," *Trans. Mach. Learn. Res.*, 2023, [Online]. Available: http://arxiv.org/abs/2301.13748

[9] M. Arief Soeleman and F. Ilmu Komputer, "Penentuan CentroidAwal Pada Algoritma K-Means Dengan Dynamic Artificial Chromosomes Genetic AlgorithmUntuk Tuberculosis Dataset," *Februari*, vol. 20, no. 1, pp. 97–108, 2021.

[10] A. A. Khan, M. S. Bashir, A. Batool, M. S. Raza, and M. A. Bashir, "K-Means Centroids Initialization Based on Differentiation Between Instances Attributes," *Int. J. Intell. Syst.*, vol. 2024, no. 1, 2024, doi: 10.1155/2024/7086878.

[11] A. Primandana, S. Adinugroho, and C. Dewi, "Optimasi Penentuan Centroid pada Algoritme K-Means Menggunakan Algoritme Pillar (Studi Kasus: Penyandang Masalah Kesejahteraan Sosial di Provinsi ...," *... Teknol. Inf. dan Ilmu ...*, vol. 3, no. 11, pp. 10678–10683, 2020, [Online]. Available: http://j-ptiik.ub.ac.id/index.php/j-ptiik/article/download/6748/3264

[12] D. Lestari, A. Charis Fauzan, F. Ilmu Eksakta, P. Studi Ilmu Komputer, U. Nahdlatul Ulama Blitar, and J. Masjid No, "Penerapan Algoritma Pillar Untuk Optimasi Penentuan Titik Awal Centroid Pada Algoritma K-Means Clustering," *JOISIE J. Inf. Syst. Informatics Eng.*, vol. 6, no. 1, pp. 15–24, 2022.

[13] D. A. H. Aliwy and D. K. B. S. Aljanabi, "An Efficient Algorithm for Initializing Centroids in K-means Clustering," *J. Kufa Math. Comput.*, vol. 3, no. 2, pp. 18–24, 2016, doi: 10.31642/jokmc/2018/030203.

[14] V. V. Romanuke, "Random Centroid Initialization for Improving Centroid-Based Clustering," *Decis. Mak. Appl. Manag. Eng.*, vol. 6, no. 2, pp. 734–746, 2023, doi: 10.31181/dmame622023742.
[15] K. Clustering, R. Scaling, and G. Scholar, "Enhancing K-Means Clustering Accuracy Through Modified Robust Scaling Technique Enhancing K-Means Clustering Accuracy Through Modified Robust Scaling Technique," *Preprints.or*, pp. 0–13, 2024, doi: 10.20944/preprints202411.1245.v1.
[16] J. Solomon, K. Greenewald, and H. Nagaraja, "k-Variance: A Clustered Notion of Variance," *SIAM J. Math. Data Sci.*, vol. 4, no. 3, pp. 957–978, 2022, doi: 10.1137/20m1385895.
[17] A. Vouros, S. Langdell, M. Croucher, and E. Vasilaki, "An empirical comparison between stochastic and deterministic centroid initialisation for K-means variations," *Mach. Learn.*, vol. 110, no. 8, pp. 1975–2003, 2021, doi: 10.1007/s10994-021-06021-7.
[18] S. Pourahmad, A. Basirat, A. Rahimi, and M. Doostfatemeh, "Does Determination of Initial Cluster Centroids Improve the Performance of K -Means Clustering Algorithm? Comparison of Three Hybrid Methods by Genetic Algorithm, Minimum Spanning Tree, and Hierarchical Clustering in an Applied Study," *Comput. Math. Methods Med.*, vol. 2020, 2020, doi: 10.1155/2020/7636857.
[19] Q. Bi, H. Sun, C. Qian, and K. Zhang, "An improved seeds scheme in K-means clustering algorithm for the UAVs control system application," *IET Commun.*, vol. 18, no. 7, pp. 437–449, 2024, doi: 10.1049/cmu2.12746.
[20] D. Arthur and S. Vassilvitskii, "K-means++: The advantages of careful seeding," *Proc. Annu. ACM-SIAM Symp. Discret. Algorithms*, vol. 07-09-Janu, pp. 1027–1035, 2007.
[21] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A Review on Data Preprocessing Techniques Toward Efficient and Reliable Knowledge Discovery From Building Operational Data," *Front. Energy Res.*, vol. 9, no. March, pp. 1–17, 2021, doi: 10.3389/fenrg.2021.652801.
[22] G. C. Montgomery, Douglas C;Runger, *Applied Statistics and Probability for Engineers*, 6th ed. Hoboken, New Jersey, USA: Wiley (John Wiley & Sons, Inc.), 2014.
[23] F. Afra, "Rumus Varians: Pengertian, Jenis, Cara Menghitung, dan Contohnya," *detikEdu*, 2023. https://www.detik.com/edu/detikpedia/d-6952619/rumus-varians-pengertian-jenis-cara-menghitung-dan-contohnya
[24] L. P. Refialy, H. Maitimu, and M. S. Pesulima, "Perbaikan Kinerja Clustering K-Means pada Data Ekonomi Nelayan dengan Perhitungan Sum of Square Error (SSE) dan Optimasi nilai K cluster," *Techno.Com*, vol. 20, no. 2, pp. 321–329, 2021, doi: 10.33633/tc.v20i2.4572.
[25] N. Nugroho and F. D. Adhinata, "Penggunaan Metode K-Means dan K-Means++ Sebagai Clustering Data Covid-19 di Pulau Jawa," *Teknika*, vol. 11, no. 3, pp. 170–179, 2022, doi: 10.34148/teknika.v11i3.502.