=Journal of Intelligent Decision Support System (IDSS) 9 (2) (2026), pp. 114-123

Published by: Institute of Computer Science (IOCS)

Journal of Intelligent Decision Support System (IDSS)

Journal homepage: www.idss.iocspublisher.org

Improving K-Means clustering performance on non-linear data using variance-weighted distance metrics

Elsya Sabrina Asmita Simorangkir[1*], Efori Bu'ulolo[2]

[1] *Teknologi Informasi, Universitas Senior Medan, Medan, Indonesia*
[2] *Teknik Komputer, Politeknik Negeri Medan, Medan, Indonesia*

Article Info

Article history:

Received May 25, 2026
Revised Jun 20, 2026
Accepted Jun 28, 2026

Keywords:

Attribute Weighting
K-Means
Non-Linear Data
Variance-Weighted Distance Metrics

ABSTRACT

K-Means is one of the most widely used clustering algorithms because of its simplicity and computational efficiency. However, its performance often decreases when handling non-linear data due to the assumption that all attributes contribute equally to the distance calculation process. This study proposes a Variance-Weighted Distance Metrics K-Means (VWDM-KMeans) method that assigns attribute weights based on variance values to improve clustering quality. The proposed approach consists of Min-Max Normalization, variance calculation, weight generation, and integration of variance-based weights into the distance metric used by K-Means. Experiments were conducted on a non-linear dataset containing 103 records and 3 attributes (x, y, and z) with K = 3 clusters. The generated attribute weights were 0.3207, 0.3342, and 0.3451 for attributes x, y, and z, respectively. The performance of VWDM-KMeans was compared with conventional K-Means and K-Medoids using the number of iterations, Sum of Squared Errors (SSE), and Silhouette Score (SS). The results showed that VWDM-KMeans converged in 5 iterations, compared to 6 iterations for K-Means and 3 iterations for K-Medoids. In terms of cluster compactness, VWDM-KMeans achieved the lowest SSE value of 2.7932, outperforming K-Means (8.2429) and K-Medoids (8.9602). Furthermore, VWDM-KMeans obtained a Silhouette Score of 0.4854, equal to K-Means and higher than K-Medoids (0.4696). These findings demonstrate that incorporating variance-based attribute weighting into the distance calculation process improves cluster compactness while maintaining cluster separation quality and stability. Therefore, VWDM-KMeans can serve as an effective and computationally efficient alternative for clustering non-linear data.

Corresponding Author:

Elsya Sabrina Asmita Simorangkir,
Teknologi Informasi,
Universitas Senior Medan,
Djamin Ginting Km. 8,5 No 13, Mangga, Kec. Medan Tuntungan, Kota Medan, Sumatera Utara 20141, Indonesia
elsyasabrinaas@gmail.com

Introduction

Advances in digital technology, the Internet of Things (IoT), cloud computing and modern information systems have generated vast amounts of data with increasingly complex characteristics(Bu'ulolo et al., 2026). This situation has driven the need for data analysis methods capable of effectively identifying hidden patterns without requiring class labels. One widely used technique is clustering, which is the process of grouping objects based on the degree of similarity in their characteristics(Simorangkir et al.,

2024). Clustering is a key component in various fields such as customer analysis, bioinformatics, cybersecurity, pattern recognition, and recommendation systems(Ikotun et al., 2023). Among the various clustering algorithms available, K-Means is the most popular method due to the simplicity of the algorithm, computational efficiency, and ease of implementation across various types of data K-means clustering algorithms. A comprehensive review, variants analysis, and advances in the era of big data (Bu'ulolo;Efori, Mesran, Hasibuan;Nelly Astuti, Utomo;Aripin;Soeb, Putro Utomo, 2023). Nevertheless, the increasing volume and complexity of data today necessitate the development of clustering methods that are more adaptable to diverse data structures(Thrun, 2021). Various studies indicate that the performance of K-Means is still significantly influenced by data characteristics and the distance measurement methods employed. Consequently, research into improving clustering quality remains a relevant and important topic for further development(Ikotun et al., 2023).

The K-Means algorithm works by dividing data into a number of clusters based on the shortest distance between data points and cluster centroids. In its implementation, K-Means typically uses Euclidean distance as a measure of similarity between objects(Bu'ulolo et al., 2025). This approach is effective for data with a linear distribution and relatively spherical cluster shapes. However, for data with non-linear patterns, irregular distributions, or varying densities, Euclidean distance often fails to represent the true relationships between data points (Zhang & Liu, 2023), (L. S. et al., 2023). Consequently, the clustering process becomes sub-optimal and yields poor-quality clusters. Furthermore, all attributes in the dataset are assumed to contribute equally to the clustering process, meaning that less informative attributes can significantly influence the clustering results (Du, 2023), (H. Irwandi O. S. Sitompul & Sutarman, 2023). Several studies report that the assumption of equal attribute contribution is one of the main causes of the decline in K-Means performance on complex data. Consequently, an approach is required that can account for the relative importance of each attribute in the distance measurement process(Golzari Oskouei, Balafar, et al., 2021).

This issue becomes increasingly significant when K-Means is applied to non-linear data. Non-linear data is commonly found in various real-world applications, such as internet user behaviour data, medical data, digital image data, and network security data(Ikotun et al., 2021). Such data structures often form circular, curved, hierarchical, or overlapping patterns, making them difficult to separate using linear cluster boundaries (Ikotun et al., 2023), (Pramudya et al., 2026). To address these limitations, various approaches have been developed, including Kernel K-Means, Spectral Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), and graph-based methods(E. B. et al., 2022). Although capable of improving the quality of clustering on non-linear data, most of these methods have higher computational complexity compared to conventional K-Means(Galis & Onchis, 2025). Furthermore, some methods require data transformation processes or the determination of additional parameters, which is not straightforward. Therefore, there remains a need to develop an approach that retains the simplicity of K-Means whilst enhancing its ability to handle non-linear data.

One potential approach to improving the performance of K-Means is to assign weights to attributes based on the statistical characteristics of the data. Several previous studies have applied feature weighting techniques using entropy, correlation, variance, or optimisation methods to improve cluster quality (Golzari Oskouei, Hashemzadeh, et al., 2021), (Wu et al., 2022). Among these various statistical measures, variance has the ability to describe the degree of data dispersion and can therefore be used as an indicator of the level of information contained within an attribute (Syahputra et al., 2022). Attributes with higher variance values generally have a greater ability to distinguish objects between clusters compared to attributes with low variance. However, most previous studies have utilised attribute weighting for general data and have not specifically evaluated the effectiveness of variance-based weighting on non-linear data using K-Means. Furthermore, the use of variance as a primary component in the formulation of distance metrics remains relatively limited compared to kernel-based or optimisation approaches (Irwandi et al., 2022). This situation presents an opportunity for research to develop distance measurement methods that are more adaptive to the characteristics of data distributions.

Various feature weighting techniques have been proposed to improve clustering performance, including entropy weighting, mutual information weighting, feature relevance analysis, optimization-based weighting, and variance-based weighting. Entropy and mutual information methods focus on

measuring information content and feature dependency, while optimization-based approaches iteratively learn feature weights to maximize clustering quality. Although effective, these methods often require additional computational costs, parameter tuning, or complex optimization procedures. In contrast, variance-based weighting offers a simple and computationally efficient mechanism by assigning greater importance to attributes with higher data dispersion, which are generally more capable of distinguishing objects within a dataset (Khan et al., 2024), (Peng et al., 2021). Nevertheless, high variance does not always indicate high discriminative power, as large variance may also result from noise or outliers rather than meaningful cluster structure (Romanuke, 2023). Despite this limitation, the direct integration of variance-derived weights into the K-Means distance metric remains relatively underexplored, particularly for non-linear datasets. Therefore, this study proposes the Variance-Weighted Distance Metrics K-Means (VWDM-KMeans) method, which incorporates variance-based attribute weighting into the clustering process to improve cluster quality on non-linear data while preserving the simplicity and computational efficiency of conventional K-Means.

Based on the above, this study proposes the Variance-Weighted Distance Metric (VWDM) as a modification to the distance calculation in the K-Means algorithm to improve the quality of clustering in non-linear data. Unlike previous studies, which generally use standard Euclidean distance or weighting techniques that do not directly account for the variance distribution of attributes, the proposed method assigns weights to each attribute based on its proportion of variance within the dataset. These weights are used in the distance calculation so that attributes containing more information make a more dominant contribution to the cluster formation process. The novelty of this study lies in the integration of the Variance-Weighted Distance Metrics mechanism into the K-Means clustering process, specifically to improve cluster separation capabilities on non-linear data without significantly increasing the algorithm's complexity. Furthermore, this research evaluates the effectiveness of the proposed method using various non-linear datasets and comprehensive cluster validation metrics. The results of this research are expected to provide a simple, efficient, and easily implementable alternative solution to improve the performance of K-Means on non-linear data, as well as serve as a foundation for the development of further clustering methods.

Method

This study proposes the Variance-Weighted Distance Metrics K-Means (VWDM-KMeans) method to improve the quality of clustering in non-linear data. The main idea behind the proposed method is to assign different weights to each attribute based on its variance, so that attributes containing more information have a more dominant contribution to the distance measurement process.

In the conventional K-Means algorithm, all attributes are considered to have the same level of importance. Consequently, less informative attributes still contribute equally to cluster formation. In non-linear data, this often results in the centroids failing to represent the data structure optimally. To address this issue, this study calculates attribute weights using variance information and integrates them into the calculation of the distance between the data and the centroid.

The Variance-Weighted Distance Metrics (VWDM) K-Means method is illustrated in Figure 1:

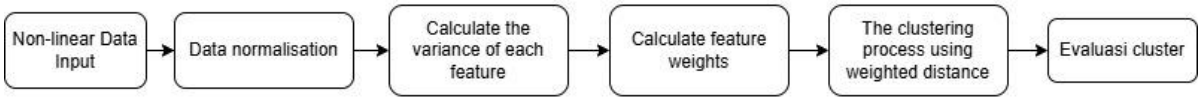

Figure 1. The VWDM-KMeans algorithm

2.1 Data Preprocessing

The first step is preprocessing to ensure that all attributes fall within a comparable range of values. As variance is heavily influenced by the scale of the data, normalisation is performed using Mini-Max Normalisation(Khan et al., 2024).

$$x'_{ij} = \frac{x_{ij} - x_i^{min}}{x_i^{max} - x_i^{min}} \quad (1)$$

Where x_{ij} is the original value of the i-th attribute in the j-th object, x'_{ij} is the normalised value, x_i^{min} is the minimum value of the i-th attribute, and x_i^{max} is the maximum value of the i-th attribute.

Normalisation is performed to prevent attributes with a wider range of values from dominating the data.

2.2 Variance Calculation

Once the data has been normalised, the next step is to calculate the variance of each attribute. Variance is used to measure the degree of dispersion of the data relative to its mean (Peng et al., 2021). For the i-th attribute:

$$\boldsymbol{Var_i} = \frac{\mathbf{1}}{\boldsymbol{n}}\sum_{j}^{n}(\boldsymbol{x_{ij}} - \bar{\boldsymbol{x}}_{\boldsymbol{ij}})^{\mathbf{2}} \quad (2)$$

Where, $\boldsymbol{Var_i}$ is variance of the i-th attribute, n number of data objects, and $\bar{\boldsymbol{x}}_{\boldsymbol{i}}$ is the mean of the i-th attribute. A high variance value indicates that the attribute is better at distinguishing between data objects. Conversely, attributes with low variance tend to provide less discriminatory information.

2.3 Variance-Based Weight Generation

Once all the variances have been obtained, the attribute weights are calculated. The weights are calculated based on the proportion of the variance relative to the total variance of all attributes.

$$\boldsymbol{w_i} = \frac{\boldsymbol{Var_i}}{\sum_{k=1}^{m}\boldsymbol{Var_k}} \quad (3)$$

Where, $\boldsymbol{w_i}$is weight of the i-th attribute, and m is number of attributes. Sifat bobot Properties of weights: o≤$\boldsymbol{w_i}$≤1 and $\sum_{i=m}^{m}\boldsymbol{w_i} = \mathbf{1}$. Through this mechanism, attributes with a wider distribution of information are assigned a higher weight than those with low variation.

2.4 Variance-Weighted Distance Metrics

In standard K-Means, the distance between data points and centroids is calculated using Euclidean distance(Romanuke, 2023; Sinaga et al., 2021).

$$\boldsymbol{d(x,c)} = \sqrt{\sum_{i=1}^{m}(\boldsymbol{x_i} - \boldsymbol{c_i})^{\mathbf{2}}} \quad (4)$$

The main drawback of this equation is that all attributes are given equal weight. This study proposes the following modifications.

$$\boldsymbol{dvw\ DM} = \sqrt{\sum_{i=1}^{m}\boldsymbol{w_i}(\boldsymbol{x_i} - \boldsymbol{c_i})^{\mathbf{2}}} \quad (5)$$

Where, $\boldsymbol{dv\ DM}$ is Variance-Weighted Distance and $\boldsymbol{w_i}$is variance-based attribute weights

2.5 Clustering Process

The clustering process is carried out as follows.

a. Initialization
   Determine the number of clusters K, which is done at random.
   $$\boldsymbol{C} = \{\boldsymbol{c_1}, \boldsymbol{c_2}, \dots, \boldsymbol{c_K}\} \quad (6)$$
b. Distance Computation
   Calculate the distance of each object from the centroid using VWDM
   $$\boldsymbol{dvw\ DM}(\boldsymbol{x_j}, \boldsymbol{c_K}) \quad (7)$$
c. Cluster Assignment
   Data objects are placed in clusters with the minimum distance.
   $$\boldsymbol{Cluster}\ (\boldsymbol{x_j}) = \mathbf{arg}\ \boldsymbol{min_k d_{vw\ DM}}(\boldsymbol{x_j}, \boldsymbol{c_K}) \quad (8)$$
d. Centroid Update
   The new centroids are calculated using the average of the cluster members.
   $$\boldsymbol{C_K} = \frac{\mathbf{1}}{\boldsymbol{N_K}}\sum_{\boldsymbol{x_j}\boldsymbol{\epsilon}\boldsymbol{C_K}}\boldsymbol{x_j} \quad (9)$$
   Where, $\boldsymbol{N_K}$ is jumlah anggota cluster ke-k
e. Convergence Check
   The iteration is terminated if there are no changes to the cluster members, or if the change in the

centroid is less than a certain threshold, or if the maximum number of iterations is reached.

2.6 **Algorithm of VWDM-KMeans**

Algorithm Variance-Weighted Distance Metrics K-Means
Input (Database X, Jumlah Cluster K)
Output (Cluster Label, Final Centroid)

```
Begin
Normalize dataset
For each attribute Ai
   Compute variance Vari
End For
Compute weight:
   wi = Vari / ΣVari
Initialize K centroids
Repeat

   For each data point xj
      For each centroid ck
         Compute VWDM distance
      End For
      Assign xj to nearest cluster
   End For
   Update centroids
Until convergence
Return clusters and centroids
End
```

2.7 Evaluation of VWDM-KMeans

To evaluate the effectiveness of the proposed VWDM-KMeans method, experiments were conducted on a non-linear dataset containing 103 records and three attributes (x, y, and z). All attributes were normalised using Min-Max Normalization before clustering. The performance of VWDM-KMeans was compared with K-Means and K-Medoids using K=3. The evaluation was based on the number of iterations, Sum of Squared Errors (SSE), Silhouette Score (SS), and cluster stability to measure clustering quality and convergence performance.

Results and Discussions

To evaluate the effectiveness of the proposed method, this study utilises a synthetic non-linear dataset comprising 103 records and 3 attributes, namely x, y, and z. The dataset is designed to have non-linear relationships between variables, where the y variable is generated using a sine function of x, whilst the z variable is generated using a quadratic function and additional non-linear components. These characteristics result in data patterns that cannot be optimally represented by simple linear relationships. Data visualisation in the form of a three-dimensional graph shows that the data distribution forms a curved surface, indicating the presence of non-linear relationships between attributes. This dataset was selected to test the ability of the Variance-Weighted Distance Metrics K-Means (VWDM-KMeans) method to handle data structures that are more complex than linear data. Furthermore, the clustering results obtained were compared with conventional K-Means using several evaluation metrics to measure the quality of the clusters produced.

Tabel 1. Data Non-Linear

No	X	y	z
1	-3	-0,1627	9,9695
2	-2,9412	-0,2158	9,6975
…	…	…	…
102	2,9412	0,2517	9,5699
103	3	0,0375	10,0055

Table 1 shows the characteristics of the dataset used in this study. The dataset consists of 103 records and three numerical attributes, namely x, y, and z. The x variable acts as the primary attribute

used to establish non-linear relationships with the other attributes. The y variable is generated using a sine function that produces a wave pattern, whilst the z variable is formed using a quadratic function combined with an additional non-linear component. The combination of these three attributes produces a data distribution that does not follow a simple linear pattern. Therefore, this dataset is suitable for evaluating the ability of clustering algorithms to identify complex and non-linear data structures.

To provide a clearer picture of the data structure used, the dataset was visualised in the form of a three-dimensional scatter plot, as shown in Figure 1.

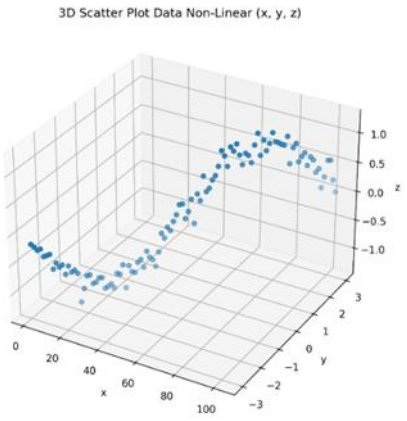

Figure 1. Three-Dimensional Visualization of the Non-Linear Dataset

Figure 1 shows a three-dimensional visualisation of the non-linear dataset used in the study. The dataset consists of 103 data points with three numerical attributes: x, y, and z. Based on the distribution of the data points, it is evident that the relationship between the attributes does not form a flat plane (linear plane), but rather follows a curved pattern that changes gradually throughout three-dimensional space. The y variable exhibits a fluctuating pattern in response to changes in the value of x, whilst the z variable undergoes changes that are not proportional to the other two attributes. This pattern indicates the presence of non-linear relationships between the attributes, making the data structure more complex than that of linear data.

Before the clustering process is carried out, all attributes in the dataset are normalised using the Min-Max Normalisation method. This step aims to standardise the range of values for each attribute so that no single attribute dominates the distance calculation process. Normalisation is important because the K-Means algorithm and the proposed method use distance measures that are sensitive to differences in data scale. If attributes have significantly different ranges of values, those with larger values will exert a more dominant influence on the cluster formation process. The results of the normalisation of the dataset are presented in Table 2.

Table 2. Data After Min-Max Normalization

No	X	y	z
1	0	0.4489	0.9961
2	0.0098	0.4275	0.9664
...			
102	0.9902	0.6159	0.9525
103	1	0.5296	1

As shown in Table 2, all attributes have been successfully transformed into a uniform range of values between 0 and 1. The value of attribute x in the first record has a normalised value of 0 as it is the minimum value of that attribute, whilst the 103rd record has a normalised value of 1 as it is the maximum value. A similar pattern also occurs for attributes y and z, although their value distributions do not change linearly due to the inherently non-linear nature of the data. This normalisation process

aims to eliminate the influence of scale differences between attributes so that each attribute makes a balanced contribution to the distance calculation process. The normalised data is subsequently used as input in the stage of calculating attribute variance and forming the Variance-Weighted Distance Metrics (VWDM) proposed in this study.

Based on the dataset normalised using Min-Max Normalisation, the variance and attribute weights (Variance-Weighted Distance Metrics) shown in Table 3 were obtained.

Table 3. Variance and Attribute Weights

Attribute	Variance ((Var_i))	Weight ((w_i))
x	0.084967	0.320685
y	0.088560	0.334243
z	0.091429	0.345072
Total	0.264956	1.000000

Next, the clustering process was carried out; the K-Means algorithm requires a number of initial centroids to serve as cluster centres. In this study, the number of clusters was set to K=3, thus requiring three initial centroids. The centroid initialisation process was carried out randomly by selecting three data points from the normalised dataset as the initial centroids. The initial random centroids are shown in Table 4.

Table 4. Initial Random Centroids (K=3)

Centroid	Record	x	y	z
C1	18	0.1667	0.2134	0.6125
C2	56	0.5392	0.4857	0.0841
C3	91	0.8824	0.7816	0.6338

Once the initial centroids have been randomly determined, the next step is to perform the cluster assignment process in the first iteration, followed by the second, third, and subsequent iterations until convergence is reached. At this stage, the distance of each data object to all centroids is calculated using Variance-Weighted Distance Metrics (VWDM). Unlike conventional K-Means, which uses Euclidean Distance, the proposed method takes into account attribute weights derived from variance values, so that more informative attributes contribute more significantly to the distance measurement process. The data objects are then placed in the cluster with the smallest distance value relative to the centroid. This process aims to form initial groups that will be used in updating the centroids in the next iteration. The cluster results of the iteration can be seen in Table 5.

Table 5. Cluster Assignment Results for Each Iteration

Iteration	Cluster	Number of Members	Member IDs
I	C1	39	1–38, 42
I	C2	38	39–41, 43–77
I	C3	26	78–103
II	C1	41	1–38, 40–42
II	C2	37	39, 43–78
II	C3	25	79–103
III	C1	43	1–43
III	C2	36	44–79
III	C3	24	80–103
IV	C1	43	1–43
IV	C2	38	44–81
IV	C3	22	82–103
V	C1	43	1–43
V	C2	38	44–81
V	C3	22	82–103

In Iteration I, the initial clustering process was carried out based on randomly selected centroids. The clustering results showed that Cluster 1 contained 39 data points, Cluster 2 contained 38 data points, and Cluster 3 contained 26 data points. As the centroids used were still the initial centroids, the distribution of cluster members was not yet fully stable. In Iteration II, following an update of the centroids based on the average of the cluster members from Iteration I, some data points moved between clusters. The number of members in Cluster 1 increased to 41 data points, whilst Cluster 2 and Cluster 3 saw only minor changes. This indicates that the cluster optimisation process is still ongoing.

In Iteration III, changes in cluster membership still occur but are becoming fewer. Cluster 1 increased to 43 members, Cluster 2 contained 36 members, and Cluster 3 contained 24 members. The reduction in member movement indicates that the centroid positions are beginning to approach an optimal state. In Iteration IV, changes to the clusters only occurred for a few data points on the cluster boundaries. The clustering results yielded 43 members in Cluster 1, 38 members in Cluster 2, and 22 members in Cluster 3. Compared to the previous iteration, the changes that occurred were relatively minor.

In Iteration V, there were no changes to the cluster members compared to Iteration IV. Consequently, the VWDM-KMeans algorithm is deemed to have reached a state of convergence in Iteration V. A visualisation of the cluster results from the final iteration can be seen in Figure 2.

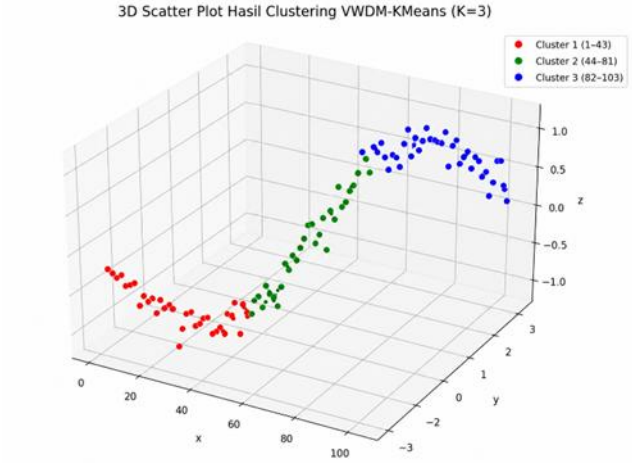

Figure 2. 3D Visualization of VWDM-KMeans Clustering Result

Figure 2 shows that the use of the Variance-Weighted Distance Metric (VWDM) is capable of producing a stable clustering process over five iterations and forming distinct clusters in the non-linear dataset used.

To obtain a more comprehensive picture of the effectiveness of the proposed method, a comparison of clustering results was carried out between K-Means, K-Medoids, and VWDM-KMeans on the same non-linear dataset. These three methods employ different cluster formation mechanisms. K-Means uses centroids as cluster centres and calculates the proximity of objects using Euclidean distance. K-Medoids uses actual data points (medoids) as cluster centres, making it more robust against extreme data and outliers. Meanwhile, VWDM-KMeans is an extension of K-Means that integrates Variance-Weighted Distance Metrics (VWDM) to assign different weights to each attribute based on its level of variance. Consequently, attributes containing more information will make a more dominant contribution to the distance measurement process. The comparison was conducted using several evaluation indicators, namely the number of iterations required to achieve convergence, the Sum of Squared Errors (SSE) value as a measure of cluster compactness, and the Silhouette Score as a measure of cluster separation quality. The use of these evaluation indicators aims to provide a more objective assessment of each method's ability to handle the non-linear data structures used in this study.

Table 6. Summary of Clustering Results

Parameter	VWDM-KMeans	K-Means	K-Medoids
Number of Records	103	103	103
Number of Attributes	3	3	3
Number of Clusters (K)	3	3	3
Distance Metric	Variance-Weighted Distance	Euclidean Distance	Euclidean Distance
Number of Iterations	5	6	3
SSE	2.7932	8.2429	8.9602
Silhouette Score	0.4854	0.4854	0.4696

As shown in Table 5, all three methods successfully clustered the non-linear dataset into three stable clusters. K-Medoids achieved convergence the fastest, in just 3 iterations, whilst VWDM-KMeans and K-Means required more iterations. In terms of cluster quality, VWDM-KMeans produced the lowest SSE value, indicating that the resulting clusters were more compact than those produced by the other methods. Furthermore, all methods achieved 100% cluster stability at the final iteration, indicating no changes in cluster membership once convergence was reached. These results demonstrate that the application of Variance-Weighted Distance Metrics enhances the quality of clustering on non-linear data by assigning greater weight to attributes containing more significant information based on their variance values.

Conclusions

This study proposed a Variance-Weighted Distance Metrics K-Means (VWDM-KMeans) method to improve clustering performance on non-linear data by incorporating attribute weights derived from variance values into the distance calculation process. Experimental results on a non-linear dataset consisting of 103 records and three attributes demonstrated that the proposed method was able to produce more compact clusters, indicated by a lower SSE value while maintaining stable clustering results. The main contribution of this research lies in the integration of variance-based attribute weighting into the K-Means algorithm, enabling more informative attributes to have a greater influence on cluster formation without significantly increasing computational complexity. These findings suggest that VWDM-KMeans can serve as an effective alternative for clustering non-linear data. However, this study was limited to a single dataset and used static attribute weights calculated only once during preprocessing. Therefore, future research should evaluate the proposed method on larger and more diverse benchmark datasets and investigate adaptive weighting strategies that update attribute weights dynamically during the clustering process to further improve clustering quality.

References

Bu'ulolo;Efori, Mesran, Hasibuan;Nelly Astuti, Utomo;Aripin;Soeb, Putro Utomo, S. (2023). *Big Data Analysis dengan Phyton untuk Perguruan Tinggi* (I).

Bu'ulolo, E., Sihombing, P., Sutarman, & Budiman, M. A. (2025). Variance-Weighted Centroid: A Centroid Estimation Approach for High-Dimensional Data Clustering. *2025 Tenth International Conference on Informatics and Computing (ICIC)*, 1–7. https://doi.org/10.1109/ICIC68054.2025.11309407

Bu'ulolo, E., Sihombing, P., Sutarman, S., & Budiman, M. (2026). K-Cube Consensus Clustering with Centroid Improvement and Variance-Based Metrics on High-Dimensional Data. *Journal of Applied Data Sciences*, *7*(2), 1440–1454. https://doi.org/10.47738/jads.v7i2.1209

Du, X. (2023). A Robust and High-Dimensional Clustering Algorithm Based on Feature Weight and Entropy. *Entropy*, *25*(3). https://doi.org/10.3390/e25030510

et al., E. B. (2022). A Review of Clustering Algorithms: Comparison of DBSCAN and K-Mean with Oversampling and t-SNE. *Recent Patents on Engineering*, *16*(2). https://doi.org/10.2174/1872212115666210208222231

et al., L. S. (2023). Reproducible Clustering with Non-Euclidean Distances: A Simulation and Case Study. *International Journal of Data Science and Analytics*.

Galis, F., & Onchis, D. (2025). *Refining Filter Global Feature Weighting for Fully-Unsupervised Clustering*.

Golzari Oskouei, A., Balafar, M. A., & Motamed, C. (2021). FKMAWCW: Categorical fuzzy k-modes clustering with

automated attribute-weight and cluster-weight learning. *Chaos, Solitons & Fractals*, *153*, 111494. https://doi.org/https://doi.org/10.1016/j.chaos.2021.111494

Golzari Oskouei, A., Hashemzadeh, M., Asheghi, B., & Balafar, M. A. (2021). CGFFCM: Cluster-weight and Group-local Feature-weight learning in Fuzzy C-Means clustering algorithm for color image segmentation. *Applied Soft Computing*, *113*, 108005. https://doi.org/https://doi.org/10.1016/j.asoc.2021.108005

H. Irwandi O. S. Sitompul, & Sutarman. (2023). K-Means Performance Optimization Using Rank Order Centroid (ROC) and Braycurtis Distance. *Sinkron*, *7*(2). https://doi.org/10.33395/sinkron.v7i2.11371

Ikotun, A. M., Almutari, M. S., & Ezugwu, A. E. (2021). K-Means-Based Nature-Inspired Metaheuristic Algorithms for Automatic Data Clustering Problems: Recent Advances and Future Directions. *Applied Sciences*, *11*(23). https://doi.org/10.3390/app112311246

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Irwandi, H., Sitompul, O. S., & Sutarman, S. (2022). K-Means Performance Optimization Using Rank Order Centroid (ROC) And Braycurtis Distance. *Sinkron : Jurnal Dan Penelitian Teknik Informatika*, *6*(2), 472–478. https://doi.org/10.33395/sinkron.v7i2.11371

Khan, A. A., Bashir, M. S., Batool, A., Raza, M. S., & Bashir, M. A. (2024). K-Means Centroids Initialization Based on Differentiation Between Instances Attributes. *International Journal of Intelligent Systems*, *2024*(1). https://doi.org/10.1155/2024/7086878

Peng, J., Wang, D., & Wang, S. (2021). Feature-weighted distance metric learning for clustering. *Pattern Recognition*, *114*, 107867. https://doi.org/10.1016/j.patcog.2021.107867

Pramudya, R. I., Kurniawan, T. A., Candra, M. H., Onn, C. W., & Dissanayake, K. K. (2026). Advancing unsupervised clustering: A systematic review of hybrid K-means and metaheuristic optimization algorithms in data mining. *Computer Science Review*, *61*, 100972. https://doi.org/https://doi.org/10.1016/j.cosrev.2026.100972

Romanuke, V. V. (2023). Random Centroid Initialization for Improving Centroid-Based Clustering. *Decision Making: Applications in Management and Engineering*, *6*(2), 734–746. https://doi.org/10.31181/dmame622023742

Simorangkir, E. S. A., Siahaan, A. P. U., Marlina, L., Nasution, D., & Sitorus, Z. (2024). Deteksi Outlier Hasil Clustering Algoritma K-Medoids Menggunakan Metode Boxplot Pada Data KIP Kuliah. *Journal of Computer System and Informatics (JoSYC)*, *5*(4), 893–902. https://doi.org/10.47065/josyc.v5i4.5479

Sinaga, K. P., Hussain, I., & Yang, M. S. (2021). Entropy K-Means Clustering with Feature Reduction under Unknown Number of Clusters. *IEEE Access*, *9*, 67736–67751. https://doi.org/10.1109/ACCESS.2021.3077622

Syahputra, N., Zarlis, M., & Efendi, S. (2022). Seleksi Fitur Menggunakan Eigen Vector Untuk Peningkatan Kinerja K-Means Clustering Dalam Pengelompokan Data. *Building of Informatics, Technology and Science (BITS)*, *4*(2 SE-Articles). https://doi.org/10.47065/bits.v4i2.2022

Thrun, M. C. (2021). Distance-based clustering challenges for unbiased benchmarking studies. *Scientific Reports*, *11*(1), 18988. https://doi.org/10.1038/s41598-021-98126-1

Wu, Z., Wang, B., & Li, C. (2022). A new robust fuzzy clustering framework considering different data weights in different clusters. *Expert Systems with Applications*, *206*, 117728. https://doi.org/https://doi.org/10.1016/j.eswa.2022.117728

Zhang, R.-L., & Liu, X.-H. (2023). A Novel Hybrid High-Dimensional PSO Clustering Algorithm Based on the Cloud Model and Entropy. *Applied Sciences*, *13*(3), 1246.