VOL. 11. NO. 3 FEBRUARY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i3.7101

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)

HYBRID PSO K-MEANS AND ROBUST SPARSE K-MEANS FOR EMPLOYEE STUDY DECISIONS

Luh Dwi Ari Sudawati[1*]; Roy Rudolf Huizen[2]; Dandy Pramana Hostiadi[2]

Magister Program, Department of Magister Information Systems[1]
Department of Magister Information Systems[2]
Institut Teknologi dan Bisnis STIKOM Bali, Indonesia[1,2]
https://www.stikom-bali.ac.id[1,2]
232011025@stikom-bali.ac.id[*], roy@stikom-bali.ac.id, dandy@stikom-bali.ac.id

(*) Corresponding Author
(Responsible for the Quality of Paper Content)

Abstract— *Human Resources (HR) are a strategic asset in institutional advancement, so employee performance evaluation must be conducted objectively and based on data. This study aims to cluster employee performance data at XYZ University for determining further studies, using the K-Means, PSO K-Means, and Robust Sparse K-Means methods, as well as three types of distance measurements: Euclidean, Manhattan, and Mahalanobis Distance. The dataset consists of 17 attributes. The evaluation was conducted using the Silhouette Score, Davies-Bouldin Index, and visualization using PCA. The results indicate that the combination of PSO K-Means with Euclidean Distance provides the best balance between quantitative performance (Silhouette Score 0.1253 and DBI 2.0521) and a more visually representative distribution of cluster members. The interpretation of the clustering results yielded three clusters: Cluster 0 (no further study) consisting of 8 employees, Cluster 1 (further study) consisting of 97 employees, and Cluster 2 (awaiting study decision) consisting of 58 employees. These findings can be utilized by institutions to design more targeted and data-driven human resource development strategies.*

Keywords: *Clustering, Data Mining, Distance Metric, K-Means, Optimization*

Intisari— *Sumber Daya Manusia (SDM) merupakan aset strategis dalam kemajuan institusi, sehingga evaluasi kinerja karyawan perlu dilakukan secara objektif dan berbasis data. Penelitian ini bertujuan mengklasterisasi data kinerja karyawan di Perguruan Tinggi XYZ dalam penentuan studi lanjut, dengan menggunakan metode K-Means, PSO K-Means, dan Robust Sparse K-Means, serta tiga jenis pengukuran jarak: Euclidean, Manhattan, dan Mahalanobis Distance. Dataset penelitian terdiri dari 17 atribut. Evaluasi dilakukan melalui Silhouette Score, Davies-Bouldin Index, dan visualisasi menggunakan PCA. Hasil menunjukkan bahwa kombinasi PSO K-Means dengan Euclidean Distance memberikan keseimbangan terbaik antara performa kuantitatif (Silhouette Score 0,1253 dan DBI 2,0521) dan distribusi anggota klaster yang lebih representatif secara visual. Interpretasi hasil klasterisasi menghasilkan tiga klaster: Klaster 0 (tidak lanjut studi) sebanyak 8 karyawan, Klaster 1 (lanjut studi) sebanyak 97 karyawan , dan Klaster 2 (menunggu keputusan studi) sebanyak 58 karyawan. Temuan ini dapat dimanfaatkan oleh institusi untuk merancang strategi pengembangan Sumber Daya Manusia yang lebih tepat sasaran dan berbasis data.*

Kata Kunci: *Pengelompokan, Penambangan Data, Metrik Jarak, K-Means, Optimasi.*

INTRODUCTION

Human Resources (HR) represent a vital asset within an organization for achieving its goals and enhancing employee performance. One of the more challenging processes in HR management is performance evaluation, which aims to provide employees with feedback on their performance and to support decisions regarding career advancement and future compensation[1]. One possible approach

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7101

to this is analyzing employee performance data to identify patterns and characteristics that can serve as a foundation for further studies. At XYZ University, employees come from diverse backgrounds and exhibit varying levels of performance.

Currently, the only requirement for determining an employee's eligibility for further studies is the individual's own consent. However, field reviews indicate that this approach is not ideal for improving the organization's quality, productivity, and overall performance. Therefore, a more systematic performance analysis is needed. Clustering analysis is one of the methods that can be employed for this purpose. K-Means Clustering algorithm was utilized in this study because of the ability to uncover patterns that are hiden in multi variate complex data which is difficult to bee intepreted manually [2].

K-Means Clustering is considered to be one of the most used partition based clustering methods in data mining field. It is unsupervised data anaalysisi method that partition objects into classes or *k* clusters. [3]. Previous studies by Apriliani Nur et al. (2023) [4], Goran W et.al (2023) [5] and Safitri Juanita (2024) [6] applied the Elbow Method to calculate the optimal sum of clusters in analyses using the K-Means Clustering algorithm.

These studies demonstrated that the Elbow Method yielded more optimal clustering results compared to approaches that did not incorporate the method. This study adopts an approach consistent with the Elbow Method, that utilized to identify the best number of clusters as an initial step prior to the implementation of the clustering process. This study also explores a comparison of various K-Means Clustering optimization methods, similar to the approaches taken in studies by Ci Fan (2021) [7],

Hamida Amdouni (2024) [8], I Made Satria Bimantara et.al (2023) [9], Mingchao Qi et.al (2024) [10] and Yully Sofyah Waode (2024) [11]. These studies compared different optimization strategies with the aim of enhancing the quality of clustering results. The results obtained indicate performance variations depending on the optimization method used, highlighting the value of comparative analysis in supporting more informed decision-making.

This demonstrates the advantages of enhanced clustering approaches, confirming the usage of K-Means method to cluster employee performance data is sufficiently accurate to support further study recommendations. This study also analyzes the use of combined distance measurement methods, in line with the studies conducted by Farid Akhmatshin et al. (2024) [12], Poonam et al. (2024) [13], Regita Putri Permata et.al (2025) [14] and Relita Buatond et al. (2024) [15]. These studies aimed to conduct a comparative analysis to identify the most effective distance calculation techniques in influencing clustering outcomes.

Although the results varied, the combination of distance metrics provided a more comprehensive evaluation of each metric's performance in the context of clustering. Other studies that also serve as references in this research include the use of hybrid methods to achieve higher accuracy and stability of clustering results compared to traditional K-Means. Among these are studies conducted by Aina Latifa Riyana Putri dkk (2025) [16], Hongwei Yue dkk (2025) [17], Manish Mahajan dkk (2021) [18] Their research successfully demonstrated that the use of hybrid methods led to a significant improvement in the quality of the resulting clusters.

While previous studies have mainly focused on improving the technical accuracy and optimization of clustering algorithms, few have applied these methods in the context of educational institutions for strategic human resource development. This study advances the field by integrating Particle Swarm Optimization (PSO), Robust Sparse with K-Means and combining multiple distance metrics to enhance cluster stability and interpretability. Moreover, it contributes a novel application of data mining. The resulting classification is expected to serve as a foundation for strategic policy formulation, supporting the Human Resources Directorate in identifying employees eligible for further study based on available performance data.

MATERIALS AND METHODS

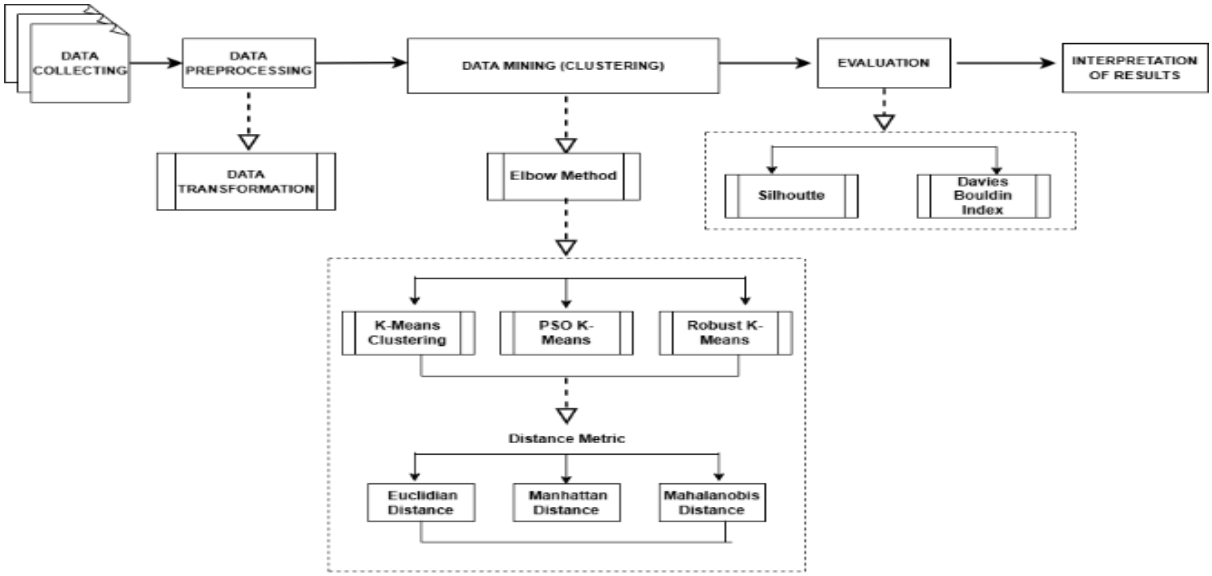
This study utilized a quantitative approach using explorative-descriptive method. The objective is to cluster employee performance data at XYZ University using the K-Means algorithm, which is also used to analyze the impact of different distance metric combinations on clustering outcomes and compare clustering optimization methods to interpret the characteristics of each resulting group.

Research Flow Framework

The conceptual framework of the research flow is a visual representation used to outline the key ideas to be investigated in a study in order to achieve its research objectives. The stages undertaken by the author are as follow.

VOL. 11. NO. 3 FEBRUARY 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480 /jitk.v11i3.7101

JITK (JURNAL ILMU PENGETAHUAN DAN TEKNOLOGI KOMPUTER)



Source : (Research Result, 2025)

Figure 1 Research Flow Framework

A. Data Collection

The questionnaire method was employed as the primary technique for data collection. The respondents involved in this study consisted of permanent lecturers and staff at XYZ University, with the criteria of having at least a Master's degree and not currently enrolled in further studies. Based on calculations using the Slovin formula , with a population size of 250 and a margin of error of 5%, the minimum required sample size is 154 respondents. This study successfully collected data from 163 respondents, thus meeting the statistically recommended minimum sample size. However, the researcher acknowledges that this number remains relatively limited when compared to a broader population beyond the study context [19]. Therefore, the findings of this research should be interpreted within the scope of the studied institution. The slovin formula is :

$$n = \frac{N}{1+N\,(e)^2} \quad (1)$$

The data atribute which is utilized in this study are :

Table 1 Performance Data Attributes

No	Atributes	Data Type
1	Gender	Categorical
2	Age	numerical
3	Marital status	Categorical
4	Number of chidren	numerical
5	Finansial status	Categorical
6	Highest level of education	Categorical
7	Length of employment	Categorical
8	Desire to further study	Categorical
9	Suvervisor recommendation	Categorical
10	Employment status	Categorical
11	Position held	Categorical
12	Impact of study on the institution	Categorical
13	Professional certification	Categorical
14	Study as self-actualization	Categorical
15	Study as self-development	Categorical
16	Reason for further study	Categorical
17	Location of study	Categorical

Source : (Research Results, 2025)

B. Data Preprocessing

After the questionnaire data were collected, a preprocessing stage was conducted to prepare the dataset for the K-Means Clustering algorithm, which only accepts numerical input and is sensitive to feature scale differences. This process was carried out manually by converting categorical data into numerical form using appropriate encoding techniques. Each categorical response option was assigned a numerical label or binary value depending on the nature of the attribute, ensuring consistency across all data entries. The outcome of this transformation process is a fully numerical dataset that is ready for normalization and further analysis.

The details of the transformation applied to each feature are presented in the following table:

Tabel 2 Data Transform

No	Atributes	Data Type	Tranformation Method
1	Gender	Categorical	One-Hot-Encoding
2	Age	numerical	-
3	Marital status	Categorical	One-Hot-Encoding
4	Number of chidren	numerical	-
5	Finansial status	Categorical	One-Hot-Encoding
6	Highest level of education	Categorical	Label Encoding

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7101

No	Atributes	Data Type	Tranformation Method
7	Length of employment	Categorical	Label Encoding
8	Desire to further study	Categorical	One-Hot-Encoding
9	Suvervisor recommendation	Categorical	One-Hot-Encoding
10	Employment status	Categorical	One-Hot-Encoding
11	Position held	Categorical	One-Hot-Encoding
12	Impact of study on the institution	Categorical	One-Hot-Encoding
13	Professional certification	Categorical	Label Encoding
14	Study as self-actualization	Categorical	One-Hot-Encoding
15	Study as self-development	Categorical	One-Hot-Encoding
16	Reason for further study	Categorical	One-Hot-Encoding
17	Location of study	Categorical	One-Hot-Encoding

Source : (Research Results, 2025)

After the encoding process, all numerical features were normalized using the Min-Max normalization method to ensure that each attribute contributed equally to the Euclidean distance calculation used in the K-Means algorithm. The normalization was applied using the following equation:

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

Where X′ represents the normalized value, X is the original value, and Xmin and Xmax denote the minimum and maximum values of each attribute.

All numerical attributes were then rescaled using Min–Max Normalization to the range [0,1]. For example, an "Age" value of 40 with a minimum of 23 and a maximum of 56 was normalized to 0.515, ensuring comparable scale among all features in distance computation.

In addition, several of the 17 attributes used in this study may contain subjective factors, such as motivation and supervisor recommendations, which could introduce bias in the clustering process. However, a previous study revealed that motivation can be considered a measurable and quantitatively analyzable factor, and therefore can be categorized as an objective and quantifiable variable rather than a subjective one. This is supported by the explanation that motivation stems from identifiable and measurable factors such as achievement, recognition, challenge, development, engagement, and opportunity [20]. Hence, this approach aligns with the authors' perspective in including motivation and supervisor recommendations as attributes that can be appropriately utilized in the analysis.

Furthermore, statistical validation was performed to strengthen the reliability of the clustering process. A one-way ANOVA test was applied to selected numerical features across clusters to examine whether the differences in mean values were statistically significant. This validation ensured that the clusters represented meaningful structural distinctions rather than random variations within the dataset.

C. Data Mining

At this stage, the study focuses on clustering techniques, in which the method employed is the Elbow Method in determining the optimal number of clusters [6]. This is continued by the clustering phase, which involves comparing the standard K-Means Clustering's performance with its optimized variants, namely PSO K-Means and Robust Sparse K-Means. These three methods are combined with different distance measurements: Euclidean distance, Manhattan distance, and Mahalanobis distance. This approach provides insights into how these combinations influence the clustering outcomes.

1. Elbow Method

The Elbow Method is a technique used to determine the optimal number of clusters. This method is presented in the form of a graph with a characteristic "elbow" shape [5]. The graph displays a decrease in inertia as the number of clusters increases. However, after a certain point, the rate of decrease becomes insignificant. This point is known as the "elbow point", which indicates the maximum value of k [2][21].

2. K-Means Clustering Method

K-Means Clustering is an algorithm publicised by MacQueen and is considered one of the famous and well-known techniques of clustering [22]. K-Means is an algorithm for clustering which groups certain data into some clusters depending on the shared characteristics [16]. This method enables datasets to ne organized into distinct groups, minimizing overlaps between clusters [23]. The K-Means algorithm's basic calculation include:

a. Determine the sums of clusters (*k*) to form.
b. Determining the first centroid for each of the cluster
c. Calculate how far is the object to the centroids
d. The formula of Euclidean Distance used is as follows.

$$d_{euclid}\,(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)^2} \quad (3)$$

Where *i* is the index of the attribute, *n* is the total amount of the data points, which is X_i, refer to the value's attribute of data point from *i-th*, and Y_i refers to the value's attribute of the *i-th* cluster center [24][13].

a. The clusters are made by grouping the objects based on the objects' closest distance to the centroid
b. New New centroid is determined and calculated by utilizing the average of members in the cluster:

$$C_k = \frac{1}{n_k}\sum d_i \quad (4)$$

Where C_k is the new centroid for the cluster of *k-th*, n_k is the amount of data points in the *k-th* cluster, and di is the *i-th* data point within cluster k [24].

The general workflow of the K-Means clustering process is summarized in Algorithm 1 :
Algorithm K-Means Clustering:
Initialize centroids
Randomly select k data points from the dataset as initial centroids.

Repeat until convergence
Assignment step:
For each data point in the dataset:
a) Calculate the distance between the data point and each centroid.
b) Assign the data point to the nearest centroid.
Update step:
For each centroid:
Calculate the new centroid by taking the mean of all data points assigned to it.

Convergence criteria
a) Check if the centroids have stopped moving (i.e., the changes in centroid positions are below a certain threshold).
b) If centroids have converged, terminate the algorithm.
c) If not, repeat steps 2a and 2b.
d) End Algorithm

The choice of distance metric (Euclidean, Manhattan, or Mahalanobis) used in Step 2(a) will affect how the distances between data points and centroids are computed. Each metric is discussed in the subsequent sections.

3. Particle Swarm Optimization (PSO) K-Means Method

The PSO K-Means Algorithm is a hybrid approach which unifies Particle Swarm Optimization (PSO) with K-Means Clustering, in which PSO is employed to enhance the selection of the first centroids [9]. This method takes advantage of PSO's ability to search for optimal solutions with the purpose to enhance the cluster's quality results thus generating better clusters compared with conventional K-Means algorithm [5]. The process begins with determining the number of clusters, followed by the random initialization of centroids. Data points are then grouped based on the nearest centroid using distance calculations. Next, the algorithm computes the fitness value of each particle, updates the personal best (Pbest) and global best (Gbest) positions, and adjusts the momentum and position of the centroids accordingly. These steps are repeated until a predefined optimal number of iterations is achieved. The optimal centroids obtained from the PSO process are subsequently utilized as the first input for the clustering procedure in the K-Means algorithm [11].

The general workflow of the K-Means clustering process is summarized in Algorithm 2.
Procedure:
1) Initialization
a. Generate *P* particles, where each particle represents a potential set of *k* centroids.
b. Randomly initialize each particle's position X_p and velocity V_p.
c. Evaluate each particle's fitness using the K-Means objective function (e.g., Sum of Squared Errors — SSE).
d. Set each particle's best-known position $pbest_p = X_p$ and identify the global best position gbestgbestgbest among all particles.

2) Repeat for each interation t = 1 to T_{max}:
a. Update particle velocity

$$V_p(t+1) = w\,.\,V_p(t) + c1\,.r1\,.\left(pbest_p - X_p(t)\right) + c2\,.r2\,.(gbest - X_p(t)) \quad (5)$$

Where $r_1, r_2 \in [0,1]$ are random numbers.
b. Update particle position

$$X_p(t+1) = X_p(t) + V_p(t+1) \quad (6)$$

c. Evaluate fitness
For each particle, assign data points to the nearest centroid (as in K-Means) and compute fitness using:

$$f(X_p) = \sum_{i=1}^{n} min\,||x_i - c\,||^2 \quad (7)$$

d. Update personal and global best
If $f(X_p) < f(pbest_b)$, then set $pbest_p = X_p$.
If $f(pbest_b) < f(gbest)$, then set $gbest = pbest_p$.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7101

3) Post-Optimization Refinement (Optional)
Apply the standard K-Means procedure using gbestgbestgbest as the initial centroids to fine-tune the clustering results.

4) Return
The optimized cluster centroids *C* = *gbestC* and their corresponding cluster assignments.

5) End Algorithm
The optimized centroids obtained from PSO are then used as the initial positions for the K-Means clustering process, resulting in more stable and globally optimal clusters.

4. Robust Sparse K-Means Method

The Robust Sparse K-Means Clustering (RSKC) algorithm is made to be defiance to the outliers by temporarily excluding outlier data during the clustering process [25]. It assigns non-outlier data to clusters first, and subsequently reassigns the outliers to their nearest clusters. This approach enhances clustering accuracy, particularly when dealing with datasets that contain noise or outlier elements [11] [26].

The general workflow of the K-Means clustering process is summarized in Algorithm 3.
Procedure:

1) Initialization
   a. Initialize feature weights $w_j = 1$ for all features $j = 1, 2, \ldots, m$.
   b. Randomly select k data points as the initial centroids $C = \{c_1, \ldots, c_k\}$.
   c. Normalize the dataset to reduce the influence of outliers.
2) Repeat until convergence or maximum iteration T_{max}:
   a. Weight Distance Computation:
   For each data point x_i and centroid c_j, compute the weighted disatance:

$$D_w(x_i c_j) = \sum_{p=1}^{m} w_p (x_{ip} - c_{jp})^2 \quad (8)$$

   Asign each x_i to the cluster with the smallest weighted distance.
   b. Update Centroids:
   For each cluster *j*:

$$C_j = \frac{1}{n_j} \sum_{x_i \in C_j} x_i \quad (9)$$

   Where n_j is the number of data points in cluster *j*.
   c. Update Feature Weights:
   Update feature weights based on within-cluster variance and apply sparsity constraint:

$$w_p = \frac{max(0, s_p - \lambda)}{\sqrt{\sum_{q=1}^{m} [max(0, s_p - \lambda)]^2}} \quad (10)$$

   Where s_p represents the between-cluster sum of squares for feature *p*, and λ is a tuning parameter controlling sparsity.
   d. Outlier Adjustment (Robustness):
   Down-weight or exclude data points with large distances from all centroids to reduce the effect of outliers on cluster formation.
3) Convergence check:
   Stop if centroids and feature weights no longer change significantly, or if the maximum iteration limit is reached.
4) Return
   Optimized centroids *C*, feature weights *w*, and final cluster assignments.
5) End Algorithm
   The integration of sparsity and robustness mechanisms in the clustering process allows the Robust Sparse K-Means to handle noisy data and high-dimensional features more effectively than the standard K-Means approach.

5. Manhattan Distance

Manhattan Distance is a distance measurement method that calculates the total absolute differences between the coordinates of two points, similar to the path taken along a city street grid. This distance is often referred to as the "city block distance" and is commonly used to measure the dissimilarity between objects along a grid-like path [15][27]. The Manhattan The formula of Manhattan Distance used is as follows :

$$d\,(\mathrm{x,y}) = \sum_{i=1}^{n} | Xi - Yi | \quad (11)$$

Where d(x, y) is the Manhattan Distance separating point x and point y, n is the amount of the attributes, X_i is the value of the *i-th* attribute of point x, Y_i is the amount value of the *i-th* attribute of point y, and [X_i – Y_i] is the magnitude of the difference amongst the two attribute values [27].

6. Mahalanobis Distance

Mahalanobis Distance is a statistical measure that calculates the distance between two points in a multidimensional space by taking into account the variance and covariance among variables. Introduced by Prasanta Chandra Mahalanobis in 1936, this method is particularly useful for

accurately measuring the separation between clusters in datasets where variables are interrelated. This metric is highly effective when dealing with high-dimensional data or datasets with correlated features [13]. The Mahalanobis Distance formula is used :

$$d_{mahalanobis}(x,y)= \sqrt{(x-y)^t \sum^{-1}(x-y)} \quad (12)$$

Where $d_{mahalanobis}(x, y)$ is the Mahalanobis distance between two vectors x and y, $(x - y)^t$ is the transpose of the difference vector between x and y, and $\sum^{-1}$ is the inverse covariance matrix of the data (typically calculated from the entire dataset) [13].

However, it should be noted that the Mahalanobis distance requires a well-conditioned covariance matrix to produce stable and meaningful results. When applied to relatively small datasets, such as the 163 samples used in this study, the covariance matrix may become unstable or nearly singular, which can lead to unreliable or fluctuating clustering outcomes. Despite this inherent risk, the Mahalanobis distance metric ultimately yielded the highest performance metrics in our comparative analysis Therefore, results obtained using this metric were interpreted with caution and compared against those produced by Euclidean and Manhattan distances for validation.

D. Evaluation

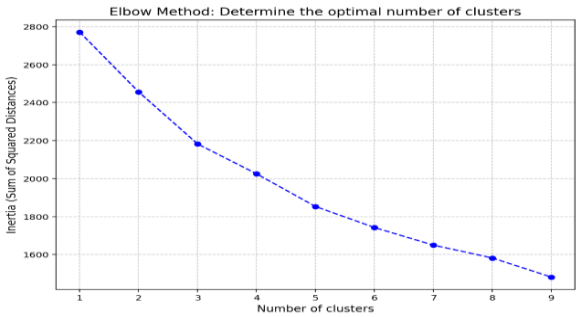
At this stage, two evaluation methods are used. The first is the Silhouette Score, which calculates how well-separated are the clusters. It extends from -1 to 1, whereas values closest to 1 means better-defined and more distinct clusters [5][28]. The second evaluation method is the Davies-Eauldin Index (DBI), which assesses cluster quality determined by the similarity in intra-cluster and the differences between inter-cluster. In this metric, lower DBI values indicate better clustering performance [15].

RESULTS AND DISCUSSION

This study utilized an employee performance dataset from XYZ University, collected through a questionnaire survey. A total of 163 responses were successfully obtained. The collected data were then transformed into numerical form to facilitate the clustering analysis process.

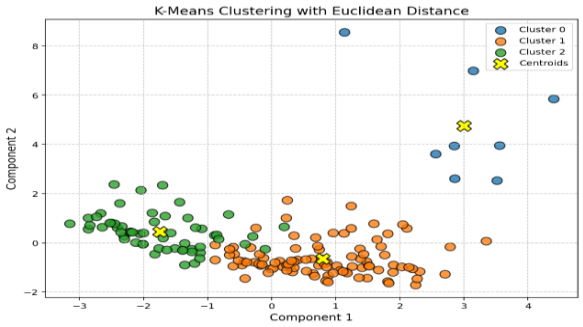
A. Elbow Method Results

This graph presents the visualization result of the Elbow Method. A significant decrease in inertia is observed up to the third cluster, after which the rate of change becomes more gradual. Thus, the optimal number of clusters is decided to be there (3).



Source : (Research Results , 2025)
Figure 2 Elbow Method Result

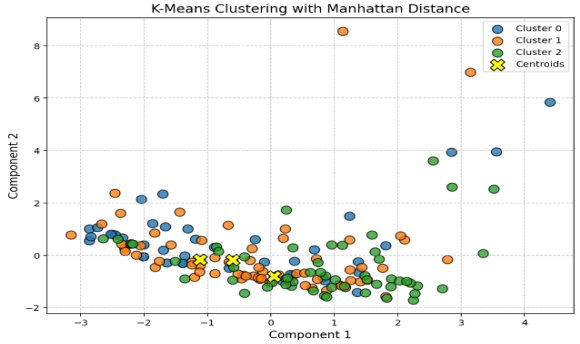
B. Results of K- Means Clustering with Euclidean Distance



Source : (Research Result, 2025)
Figure 3 K-Means with Euclidean Distance

The visualization of the results from the clustering is determined from the combination between K-Means Clustering and Euclidean Distance, using PCA for dimensionality reduction, shows that the data points from each cluster are well-separated. This indicates that the segmentation process successfully formed distinct groups with differing characteristics.

C. Results of K- Means Clustering with Manhattan Distance



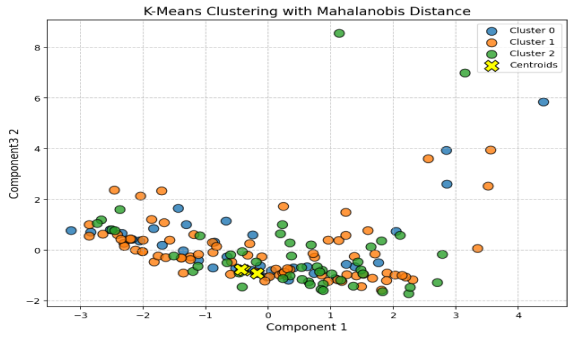
Source : (Research Result, 2025)
Figure 4 K-Means with Manhattan Distance

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7101

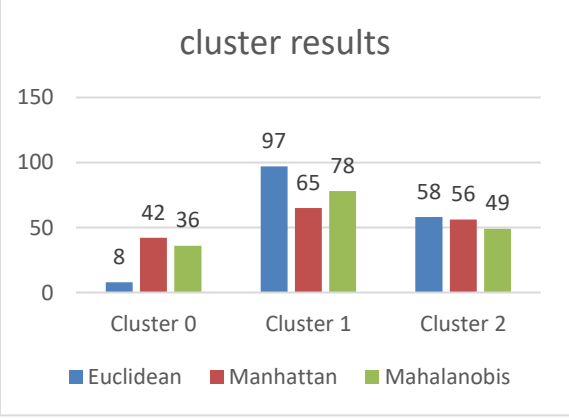
The visualization of the results from the clustering is determined from the combination between K-Means Clustering and Manhattan Distance, utilizing PCA to reduce dimensionality, shows that the data distribution among clusters is not clearly separated, particularly between Cluster 1 and Cluster 2. Data points in Cluster 0 appear to be more widely dispersed and positioned farther from the other clusters, with some overlapping points observed near the cluster boundaries.

D. Results of K-Means with Mahalanobis Distance



Source : (Research Result, 2025)
Figure 5 K-Means with Mahalanobis Distance

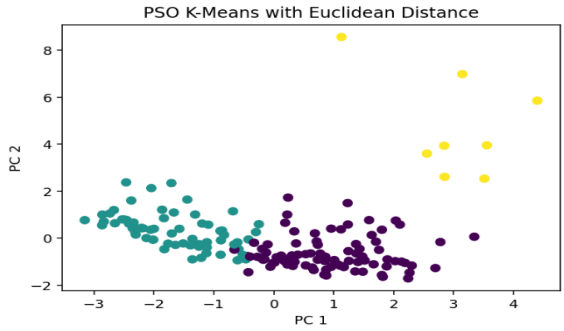
The visualization of the results from the clustering is determined from the combination between K-Means Clustering and Mahalanobis Distance, utilizing PCA to reduce dimensionality, shows that Cluster 1 and Cluster 2 overlap significantly, with data points scattered within the same area, particularly around the central coordinate (0,0). The proximity of the centroids of these two clusters indicates a similarity in feature characteristics among the objects, resulting in less distinct separation between the clusters.



Source : (Research Result, 2025)
Figure 6 Distribution of Data Quantity in K-Means Clustering with Three Distance Matrices

Figure 6 shows the number of participants in three clusters based on three types of metrics: Euclidean, Manhattan, and Mahalanobis. It is evident that the Euclidean metric results in a dominant cluster in Cluster 1 (97), whereas the Manhattan and Mahalanobis metrics indicate a more evenly distributed clustering.

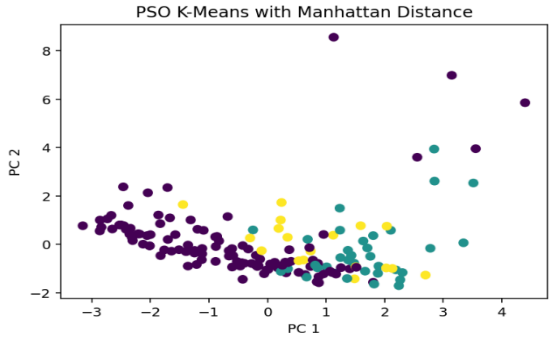
E. Results of PSO K-Means with Euclidean Distance



Source : (Research Result, 2025)
Figure 7 PSO K-Means with Euclidean Distance

The visualization of the clustering results from the combination of PSO K-Means with Euclidean Distance, using PCA, demonstrates clearer and more well-defined cluster separation, with greater distances between cluster centroids and more compact point distributions. This indicates that PSO successfully optimized the centroid positions, thereby enhancing the quality of cluster separation.

F. Results of PSO K-Means with Manhattan Distance

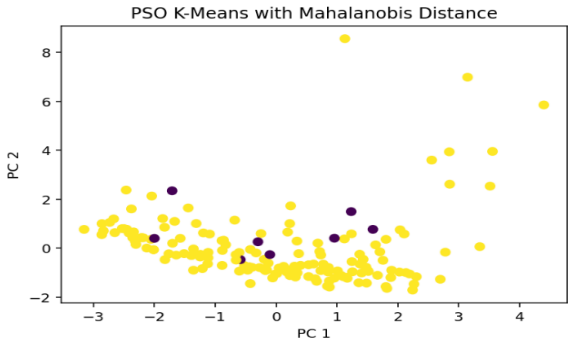


Source : (Research Result, 2025)
Figure 8 PSO K-Means with Manhattan Distance

The visualization of the clustering results from the combination of PSO K-Means with Manhattan Distance, using PCA, shows a reasonably clear cluster separation. However, some cluster

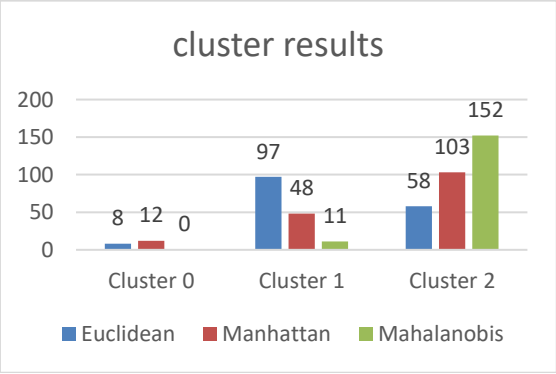
overlap is observed, indicating that a high degree of similarity in characteristics still exists among certain data points, making them difficult to separate distinctly.

G. Results of PSO K-Means with Mahalanobis Distance



Source : (Reserach Result, 2025)
Figure 9 PSO K-Means with Mahalanobis Distance

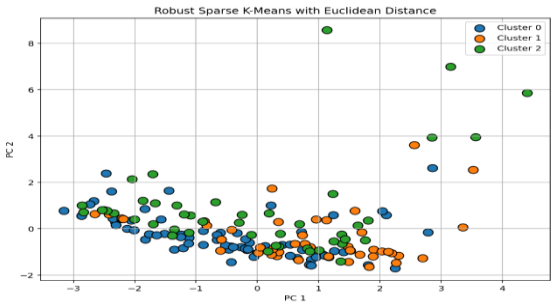
The visualization of the clustering results from the combination of PSO K-Means with Mahalanobis Distance, using PCA, reveals the formation of two main clusters. The data distribution shows that the majority of the data points are concentrated within one cluster (represented in yellow), while the other cluster (represented in purple) contains fewer data points that are more widely dispersed.



Source : (Research Result, 2025)
Figure 10 Distribution of Data Quantity in PSO K-Means Clustering with Three Distance Matrices

Figure 10 shows the number of parts in three categories based on the Manhattan, Euclidean, and Mahalanobis metrics. The results show that while Euclidean is firmly established in Cluster 1 (97), Manhattan is positioned with the highest concentration in Cluster 2 (103), while Mahalanobis dominates Cluster 2 (152).

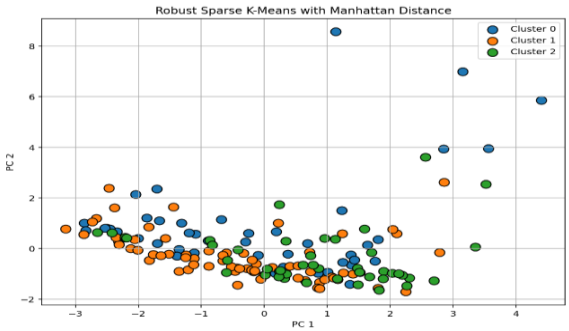
H. Result of Robust Sparse K-Means with Euclidean Distance Result



Source : (Research Result, 2025)
Figure 11 Robust Sparse K-Means with Euclidean Distance

The visualization of the clustering results from the combination of Sparse Robust K-Means with Euclidean Distance, using PCA, shows the formation of three well-defined clusters. Cluster 0 (blue) appears compact and dominant on the middle-left of the plot, Cluster 1 (orange) is spread across the center to lower-right area but remains clearly grouped, while Cluster 2 (green) is more dispersed across the upper and partially right regions of the plot, possibly representing data points with extreme or distinct characteristics.

I. Results of Robust Sparse K-Means with Manhattan Distance



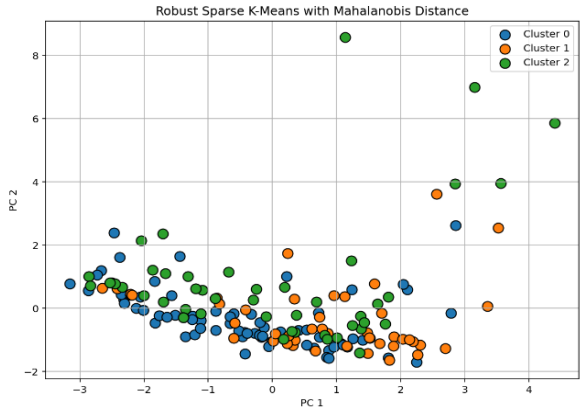
Source : (Research Result, 2025)
Figure 12 Robust Sparse K-Means with Manhattan Distance

The visualization of the clustering results from the combination of Sparse Robust K-Means with Manhattan Distance, using PCA, reveals three main groups that are fairly evenly distributed along the principal dimensions. Although there is a tendency toward separation along certain sides of the main dimensions, many data points exhibit closely related characteristics, resulting in clusters that are not entirely distinct within the two-dimensional space.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7101

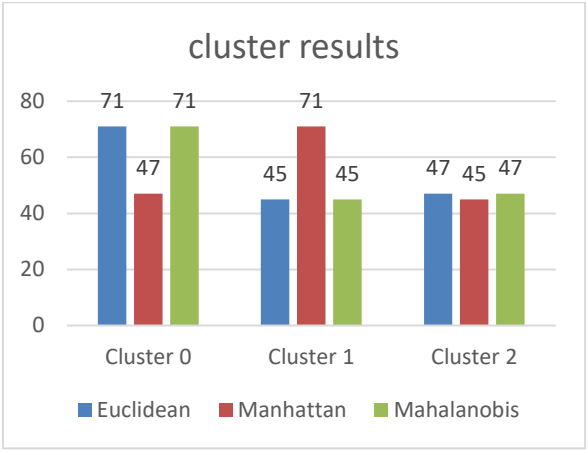
J. Result of Robust Sparse K-Means with Mahalanobis Distance



Source : (Research Result, 2025)

Figure 13 Robust Sparse K-Means with Mahalanobis Distance

The visualization of the clustering results from the combination of Sparse Robust K-Means with Mahalanobis Distance, using PCA, shows the formation of three clusters (Cluster 0, Cluster 1, and Cluster 2). Although overlapping areas are present in the center of the plot, this indicates similarities in characteristics among objects from different clusters.



Source : (Research Result, 2025)

Figure 14 Distribution of Data Quantity in Sparse Robust K-Means Clustering with Three Distance Matrices

Figure 14 shows the distribution of participants across three clusters based on Euclidean, Manhattan, and Mahalanobis distance metrics. All three metrics result in relatively balanced distributions, with minor variations between clusters, indicating consistent clustering performance.

K. Silhoutte Score

Table 3 Silhouette Score

No.	Combination of Methods	Silhoutte Score
1	K-Means with Euclidian Distance	0.1253
2	K-Means with Manhattan Distance	0.0457
3	K-Means with Mahalanobis Distance	-0.0037
4	PSO K-Means with Euclidian Distance	0.1253
5	PSO K-Means with Manhattan Distance	0.0349
6	PSO K-Means with Mahalanobis Distance	0.3263
7	Robust Sparse K-Means with Euclidian Distance	0.0728
8	Robust Sparse K-Means with Manhattan Distance	0.0728
9	Robust Sparse K-Means with Mahalanobis Distance	0.0728

Source : (Research Results, 2025)

Based on the evaluation using the Silhouette Score metric across nine clustering method variations, PSO K-Means with Mahalanobis Distance demonstrated the best performance, achieving a score of 0.3263. This value indicates a good clustering quality, characterized by high intra-cluster similarity and clear separation between clusters. In contrast, the standard K-Means method showed lower performance, where its combination with Euclidean and Manhattan Distance yielded scores of 0.1253 and 0.0457, respectively. The use of Mahalanobis Distance with standard K-Means even resulted in a negative score (-0.0037), indicating a poor clustering outcome.

L. Davies Bouldin Index (DBI)

Table 4 Davies Bouldin Index

No.	Combination of Methods	DBI
1	K-Means with Euclidian Distance	2.0521
2	K-Means with Manhattan Distance	3.5771
3	K-Means with Mahalanobis Distance	7.6727
4	PSO K-Means with Euclidian Distance	2.0521
5	PSO K-Means with Manhattan Distance	2.9190
6	PSO K-Means with Mahalanobis Distance	2.2233
7	Robust Sparse K-Means with Euclidian Distance	3.2974
8	Robust Sparse K-Means with Manhattan Distance	3.2974
9	*Robust Sparse* K-Means with *Mahalanobis Distance*	3.2974

Source : (Research Results, 2025)

Based on the evaluation using the Davies-Bouldin Index (DBI) across nine clustering method variations, the lowest DBI value of 2.0521 was obtained by two methods: K-Means with Euclidean Distance and PSO K-Means with Euclidean Distance. However, to comprehensively assess clustering quality, two internal evaluation metrics were considered: the Silhouette Score and the Davies-Bouldin Index. From this perspective, the PSO K-

Means with Mahalanobis Distance combination delivered the best overall performance, achieving the highest Silhouette Score of 0.3263 and a relatively low DBI value of 2.2233. In contrast, the K-Means method with Mahalanobis Distance showed the poorest performance, indicated by a negative Silhouette Score (-0.0037) and the highest DBI value (7.6727), suggesting that the clustering result was highly suboptimal. The PSO K-Means with Euclidean Distance and K-Means with Euclidean Distance methods yielded the lowest DBI value of 2.0521, indicating that the resulting clusters were fairly compact and well-separated. However, the Silhouette Score for both methods was only 0.1253, suggesting that the overall cohesion and separation of the clusters were relatively weak despite the low DBI.

The Robust Sparse K-Means (RSK-Means) method demonstrated stable performance across all three types of distance metrics, with a Silhouette Score of 0.0728 and a DBI value of 3.2974. This indicates that, although the results were consistent, the clustering quality was not superior compared to that achieved by the PSO K-Means method. Although PSO K-Means with Mahalanobis Distance recorded the best internal evaluation scores quantitatively, the clustering visualization revealed an uneven distribution of cluster members, with one cluster (Cluster 0) containing no data points. Therefore, the researcher selected PSO K-Means with Euclidean Distance as the most optimal method, as it offers a better balance between quantitative evaluation performance and visually representative cluster member distribution.

M. Intepretation of Result

At this stage, the researcher conducted an in-depth analysis of the unlabeled data structure through a series of clustering method combinations. After comparing various combinations, PSO K-Means with Euclidean Distance was identified as the most optimal method, based on a balanced consideration of quantitative performance and visual cluster distribution. The clustering results produced three main groups: Cluster 0 with 8 members, Cluster 1 with 97 members, and Cluster 2 with 58 members. To interpret the clustering results, the researcher focused on four key attributes considered most relevant in determining the tendency for pursuing further studies, namely:

1) Age – represents the respondent's life or career stage.
2) Financial status – Indicates financial readiness for further study.
3) Supervisor Recommendation – reflects institutional or supervisory support.
4) Desire for further study – reflects personal interest in pursuing further education.

The average values of the four key attributes were analyzed across each cluster, with the results as follows:

1) Cluster 0: The average age is 35.25 years, with a financial readiness score of 0.38, supervisor's recommendation score of 1.75, and desire for further study score of 0.88. This cluster is interpreted as a group unlikely to pursue further education due to low levels of readiness and motivation.
2) Cluster 1: The average age is 40.14 years, with a financial readiness score of 0.66, supervisor's recommendation score of 1.75, and desire for further study score of 1.05. This group is categorized as the cluster likely to pursue further education, due to adequate readiness and support.
3) Cluster 2: The average age is 30.28 years, with a financial readiness score of 0.24, supervisor's recommendation score of 1.55, and desire for further study score of 1.16. This cluster represents a group that is waiting for the right time to pursue further study, characterized by high motivation but still in the planning stage due to limited support and low financial readiness.

It is important to note that these cluster labels were inferred based on dominant attribute patterns rather than predefined categories. Since K-Means is an unsupervised learning method without ground truth data, the assigned labels ("Unlikely to Study Further," "Likely to Study Further," and "Planning Stage") should be interpreted as descriptive tendencies rather than definitive classifications. There remains a potential risk of misclassification or overgeneralization, especially in the absence of external validation data.

This interpretation provides a conceptual understanding of employee behavior and readiness in considering further education, and serves as a foundation for formulating more targeted policy recommendations for XYZ University. However, to ensure that the PSO K-Means with Euclidean Distance model accurately reflects real-world conditions, further validation in practical settings is required. This aspect is acknowledged as one of the limitations of the present study.

CONCLUSION

Based on the analysis of questionnaire data from 163 lecturers and permanent staff at XYZ University, this study successfully classified the respondents into three clusters using PSO K-Means

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri

VOL. 11. NO. 3 FEBRUARI 2026
P-ISSN: 2685-8223 | E-ISSN: 2527-4864
DOI: 10.33480/jitk.v11i3.7101

with Euclidean Distance as the most effective method. The selection of this method was based on a combination of quantitative evaluation metrics, namely the Silhouette Score and Davies-Bouldin Index, as well as considerations of visual clarity and proportional data distribution. The three resulting clusters represent a segmentation of employees' readiness and inclination to pursue further education. Cluster 0 reflects a group with low financial readiness and moderate interest in further studies, despite having received supervisor recommendations, suggesting that they are unlikely to continue their studies in the near future. Cluster 1 represents employees of more mature age, with good financial readiness, strong institutional support, and high motivation to pursue further education—indicating a group that is well-prepared and likely to proceed with further studies soon. In contrast, Cluster 2 consists of younger respondents who show strong interest in further education but face financial limitations and lack of supervisor support, indicating that they are still in the planning stage or awaiting the right opportunity.

The results of this clustering analysis provide a clear overview of the readiness of human resources at XYZ University to pursue formal education. These findings can serve as a foundation for institutional management in formulating strategic policies, such as the provision of scholarships, mentoring programs, and other forms of structural support, to encourage increased participation in further education among lecturers and staff. To strengthen future studies, the dataset could be expanded to include a larger and more diverse group of respondents, enabling broader generalization of findings. Additionally, validating and refining the selected features through expert evaluation or advanced statistical techniques could further enhance the accuracy and interpretability of the clustering model, supporting more precise recommendations for policy development.

REFERENCE

[1] Z. M. Fadhil, "Hybrid of K-means clustering and naive Bayes classifier for predicting performance of an employee," *Period. Eng. Nat. Sci.*, vol. 9, no. 2, pp. 799–807, 2021, doi: 10.21533/pen.v9i2.1898.

[2] Y. C. Fadilah, A. Sani, A. Andrianingsih, and U. Nasional, "APPLYING K-MEANS CLUSTERING FOR GROUPING PAPUA ' S DISTRICTS," vol. 10, no. 3, pp. 543–553, 2025, doi: 10.33480/jitk.v10i3.5865.APPLYING.

[3] M. I. C. Rachmatullah, "Proposed Modification of K-Means Clustering Algorithm with Distance Calculation Based on Correlation," *J. Ilm. Tek. Elektro Komput. dan Inform.*, vol. 8, no. 1, p. 136, 2022, doi: 10.26555/jiteki.v8i1.23696.

[4] A. N. Afifah, V. Nurcahyawati, and V. R. Hananto, "Analisis Clustering dan Pemetaan Sebaran Pelanggan Perusahaan Properti di Sidoarjo," *J. Edukasi dan Penelit. Inform.*, vol. 9, no. 3, p. 502, 2023, doi: 10.26418/jp.v9i3.67935.

[5] G. W. Hamaali, K. A. Abduljabbar, and D. R. Sulaiman, "K-means Clustering and PSO Algorithm for Wireless Sensor Networks Optimization," *Univ. Thi-Qar J. Eng. Sci.*, vol. 13, no. 1, pp. 40–50, 2023, doi: 10.31663/tqujes13.1.457(2023).

[6] S. Juanita and R. D. Cahyono, "K-Means Clustering With Comparison of Elbow and Silhouette Methods for Medicines Clustering Based on User Reviews," *J. Tek. Inform.*, vol. 5, no. 1, pp. 283–289, 2024, [Online]. Available: https://doi.org/10.52436/1.jutif.2024.5.1.1349

[7] C. Fan, "Evaluating Employee Performance with an Improved Clustering Algorithm," *Inform.*, vol. 46, no. 5, pp. 123–128, 2022, doi: 10.31449/inf.v46i5.4079.

[8] H. Amdouni, G. Manita, D. Oliva, E. H. Houssein, O. Korbaa, and S. Zapotecas-Martínez, "Dynamic Social Particle Swarm Optimization For Automatic Clustering," *Procedia Comput. Sci.*, vol. 246, no. C, pp. 1409–1418, 2024, doi: 10.1016/j.procs.2024.09.583.

[9] I. M. S. Bimantara and I. M. Widiartha, "Optimization of K-Means Clustering Using Particle Swarm Optimization Algorithm for Grouping Traveler Reviews Data on Tripadvisor Sites," *J. Ilm. Kursor*, vol. 12, no. 1, pp. 1–10, 2023, doi: 10.21107/kursor.v12i01.269.

[10] M. Qi, J. Q. Zhao, and Y. Feng, "An optimized public opinion communication system in social media networks based on K-means cluster analysis," *Heliyon*, vol. 10, no. 24, p. e40033, 2024, doi: 10.1016/j.heliyon.2024.e40033.

[11] Y. S. Waode, A. Kurnia, and Y. Angraini, "K-Means Optimization Algorithm to Improve Cluster Quality on Sparse Data," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 23, no. 3, pp. 641–652, 2024, doi: 10.30812/matrik.v23i3.3936.

[12] F. Akhmatshin, P. Egarmin, M. Gerasimova, I.

Petrova, and S. Mikitchak, "Clustering of k-means based on Euclidean distance metric and Mahalanobis metric," *E3S Web Conf.*, vol. 531, pp. 0–5, 2024, doi: 10.1051/e3sconf/202453103002.

[13] P. Kumari and S. Gupta, "Comparative analysis between Euclidean distance metric and Mahalanobis Distance Metric," vol. 12, no. 2, 2024.

[14] R. P. Permata, "Optimizing K-Means Clustering through Distance Metric Simulation for Strategic Enrollment Segmentation in Private Universities," vol. 10, no. 2, pp. 616–629, 2025.

[15] R. Buaton *et al.*, "COMPARATIVE EVALUATING NUMERICAL MEASURE VARIATIONS IN K- MEDOIDS CLUSTERING FOR EFFECTIVE DATA GROUPING," vol. 10, no. 2, pp. 394–403, 2024, doi: 10.33480/jitk.v10i2.5545.INTRODUCTION.

[16] A. M. P. Swarm and O. Approach, "A Multi-Objective Particle Swarm Optimization Approach," vol. 9, no. 3, pp. 542–550, 2025.

[17] H. Yue, H. Zhang, and Y. Dai, "Application of PSO-integrated K-means algorithm in resident digital portrait classification," pp. 1–15, 2025, doi: 10.1371/journal.pone.0329123.

[18] M. Mahajan, S. Kumar, and B. Pant, "Prediction of Environmental Pollution Using Hybrid PSO-K-Means Approach," vol. 12, no. 2, pp. 65–76, 2021, doi: 10.4018/IJEHMC.2021030104.

[19] A. Penerapan, F. Slovin, D. A. N. Kesalahan, and D. Perspektif, "Jurnal multidisiplin sosisal humaniora," vol. 1, pp. 53–63, 2024.

[20] T. Third, I. Conference, L. Ganasih, R. Rama, and R. Aswin, "The Influence Of Motivation , Perceptions And Attitudes On Student Decisions In Choosing A Major (Study Of Management Students , Faculty Of Economics And Business , University Of Riau)," 2023.

[21] N. A. Maori and E. Evanita, "Metode Elbow dalam Optimasi Jumlah Cluster pada K-Means Clustering," *Simetris J. Tek. Mesin, Elektro dan Ilmu Komput.*, vol. 14, no. 2, pp. 277–288, 2023, doi: 10.24176/simet.v14i2.9630.

[22] Baiq Nikum Yulisasih, H. Herman, and S. Sunardi, "K-Means Clustering Method For Customer Segmentation Based On Potential Purchases," *J. ELTIKOM*, vol. 8, no. 1, pp. 83–90, 2024, doi: 10.31961/eltikom.v8i1.1137.

[23] N. L. Ratniasih and R. A. N. Diaz, "Comparison of Clustering Algorithm in Employee Training Management Recommendations," *3rd Int. Conf. Cybern. Intell. Syst. ICORIS 2021*, pp. 5–8, 2021, doi: 10.1109/ICORIS52787.2021.9649503.

[24] N. Rahmadani, E. Rahayu, and A. Lestari, "K-Means Clustering Areas Prone To Traffic Accidents in Asahan Regency," *JITK (Jurnal Ilmu Pengetah. dan Teknol. Komputer)*, vol. 6, no. 2, pp. 181–186, 2021, doi: 10.33480/jitk.v6i2.1519.K-MEANS.

[25] A. B. Madjoukeng and B. Edith, "Robust Sparse k-means Clustering Against STOF Observations Robust Sparse k -means Clustering Against STOF Observations," no. March, 2025, doi: 10.36227/techrxiv.174114537.78354861/v1.

[26] M. Muhajir and D. Rosadi, "Robust clustering using sparse K means on text analitycs of PPKM policy in Indonesia," *AIP Conf. Proc.*, vol. 2720, no. December, 2023, doi: 10.1063/5.0137220.

[27] T. M. Ghazal *et al.*, "Performances of k-means clustering algorithm with different distance metrics," *Intell. Autom. Soft Comput.*, vol. 30, no. 2, pp. 735–742, 2021, doi: 10.32604/iasc.2021.019067.

[28] W. A. Prastyabudi, A. N. Alifah, and A. Nurdin, "Segmenting the Higher Education Market: An Analysis of Admissions Data Using K-Means Clustering," *Procedia Comput. Sci.*, vol. 234, no. 2023, pp. 96–105, 2024, doi: 10.1016/j.procs.2024.02.156.

Accredited Rank 2 (Sinta 2) based on the Decree of the Dirjen Penguatan RisBang Kemenristekdikti No.225/E/KPT/2022, December 07, 2022. Published by LPPM Universitas Nusa Mandiri