IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 4, August 2026, pp. 3476~3491
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i4.pp3476-3491

Comparison algorithms with optimization for clustering multi-criteria

Ida Mulyadi[1], Titik Khawa Abdul Rahman[2], Muhammad Faisal[1]
[1]Department of Informatics, Universitas Muhammadiyah Makassar, Makassar, Indonesia
[2]Department of School of Science and Technology, Asia E University, Kuala Lumpur, Malaysia

Article Info

Article history:

Received Sep 3, 2025
Revised Jun 5, 2026
Accepted Jul 9, 2026

Keywords:

Elbow
Fuzzy C-means
K-means
Optimization
Silhouette

ABSTRACT

In higher education, forecasting student graduation is crucial for early intervention development, policy formulation, and class planning. This categorizes the factors that affect student graduation according to both academic and non-academic traits. To compare various clustering techniques, such as K-means, fuzzy C-means (FCM), K-medoids, density-based spatial clustering of applications with noise (DBSCAN), spectral clustering, Gaussian mixture model (GMM), and deep learning (DL), as well as particle swarm optimization (PSO) and genetic algorithm (GA) optimization for K-means and FCM, this article applies the hybrid elbow + silhouette optimization prior to the clustering process. Two clusters were found in the pre-cluster optimization research results. K-means, FCM, GMM, and DL clustering all demonstrated more distinct centroid separation; K-means and GMM were the most visually stable and comprehensible. Inter-cluster separation is the main goal of post-cluster optimization, and the K-means + PSO method is the best option. The cross-dataset validation findings indicate that the clustering model exhibits moderate consistency on the new dataset, with an adjusted Rand index (ARI) of 0.478 and a normalized mutual information (NMI) of 0.433.

Corresponding Author:

Ida Mulyadi
Department of Informatics, Universitas Muhammadiyah Makassar
Sultan Alauddin No.259, Gn. Sari, Rappocini, Makassar City, South Sulawesi, Indonesia
Email: idamulyadi@unismuh.ac.id

1. INTRODUCTION

Understanding the factors associated with student graduation is an important issue in higher education because it supports academic planning, policy development, and early intervention strategies. Educational datasets are typically multidimensional and contain complex relationships among variables. Clustering techniques are machine learning methods used to categorize related data points without the need for predetermined class labels. K-means and fuzzy C-means (FCM) are frequently used in educational analytics for their versatility and effectiveness across diverse, heterogeneous datasets [1]–[5].

K-means is often used for its computational efficiency and straightforward implementation. Nonetheless, its efficacy depends on the choice of initial centroids. FCM is more appropriate for overlapping datasets, permitting data to be incorporated into numerous clusters. Determining the number of clusters before the clustering procedure yields highly accurate results [6]. The elbow technique serves as a preliminary reference; however, the elbow point may not always be distinctly apparent. The silhouette approach enhances the elbow method's efficacy by analyzing cluster quality through intra-cluster cohesiveness and inter-cluster separation [7]. The task of grouping graduation-related criteria shares similarities with segmentation problems in healthcare, business, and social analytics. Across these domains,

researchers must address heterogeneous data characteristics, boundary cases, and the need to generate clusters that are both meaningful and actionable for decision makers.

Previous studies reveal research gaps, particularly the employment of restricted algorithms, as evidenced by [8]–[10], which compares K-means, K-medoids, and FCM algorithms for clustering jobs without including further optimizations to enhance the accuracy of these algorithms in clustering multi-criteria data. This research enhances the accuracy and efficiency of clustering, particularly in managing multi-criteria data, by incorporating hitherto unutilized metaheuristic optimization strategies [11]. Determine the ideal number of clusters by using the elbow technique or gap statistic, as outlined by [12]. This research addresses the ambiguity encountered in determining the elbow point by applying silhouette optimization approaches to enhance the selection of the number of clusters. Prior studies also address multi-criteria clustering as shown in [13], concentrated solely on clustering data using a limited number of criteria. This research examines multi-criteria data clustering, incorporating multiple factors to assess student graduation and academic performance, so offering a more holistic and pertinent approach to educational realities.

This research is novel in incorporating elbow–silhouette optimization to ascertain the cluster layout prior to clustering. A comparison was conducted among several clustering techniques, including K-means, FCM, K-medoids, density-based spatial clustering of applications with noise (DBSCAN), spectral clustering, Gaussian mixture models (GMM), and deep learning (DL) [14]–[16]. The clustering outcomes from K-means and FCM are improved using particle swarm optimization (PSO) and genetic algorithm (GA) to enhance cluster quality [17]. The primary donations of this research are:

- A comparison analysis of K-means and FCM focusing on factors influencing student graduation.
- Evaluation of each algorithm's performance in terms of cluster separation, overlap handling, computational efficiency, and interpretability within an educational context.
- Elbow method for determining cluster quantity.
- Integration of metaheuristic optimization techniques to enhance the accuracy of clustering.

2. METHOD

The clustering process, which begins with academic and non-academic datasets, is depicted in the research design in Figure 1. Elbow optimization utilizing silhouette analysis to determine the ideal number of clusters. The optimization outcomes are applied in K-means, FCM, K-medoids, DBSCAN, spectral clustering, GMM, and DL [18]. To make sure the clustering results are consistent and dependable, algorithms then optimized using PSO and GA, assessed using silhouette score, purity, normalized mutual information (NMI), and adjusted Rand index (ARI), and validated using bootstrapping and cross-validation.

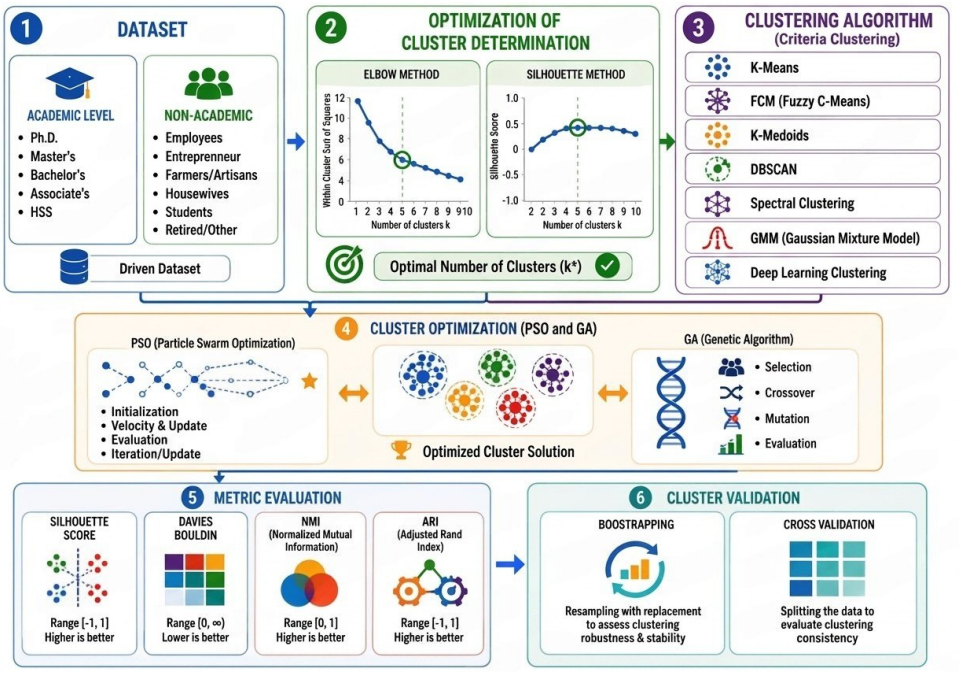


Figure 1. Research methodology

2.1. Data collection

Samples from academic datasets at several universities are used in this study and are supported by references to relevant research. Direct observation on campus and a review of the literature were used to gather data. The criteria were established by referencing previous research in the domain of educational data mining, which demonstrated that these factors have a significant impact on student graduation rates [19].

The proposed dataset includes 33 criteria that represent contextual academic and non-academic performance variables associated with student achievement [20], [21]. The cumulative grade point average (GPA) [22], course completion rates, the number of repeated courses, semester credit load, and general trends in academic achievement [23]–[26] are examples of academic factors. Conversely, involvement in extracurricular activities, school attendance, socioeconomic circumstances, and demographic traits such as age and gender [27] are examples of non-academic influences. These characteristics were selected because a number of interrelated factors affect student performance and graduation prospects.

After that, all data values are standardized to the same range, 0 to 1, using min-max normalization. For distance-based clustering techniques like K-means and FCM, this step is crucial. 16 specialists examined 33 factors to determine the final dataset. The K-means and FCM clustering techniques were then used to process the data in order to find groupings of criteria with comparable academic and non-academic traits that might have an impact on student graduation. By converting input data into fuzzy membership degrees, assessing these values using a set of fuzzy rules, and applying defuzzification to obtain the final clustering result, fuzzy inference facilitates flexible decision-making. Membership values in the ambiguous range of 0.5023 to 0.5903 indicate that this method effectively controls uncertainty for students whose data placements are near the cluster boundary.

2.2. Cluster optimization using elbow method

Establishing the appropriate cluster configuration is essential in clustering research, as it directly influences cluster interpretability and model reliability [28]. A model that overfits the data or has poor cluster separation can result from selecting an incorrect value of k, thereby lowering the accuracy and interpretability of the study's findings. To determine the best cluster structure, this study used the elbow method, with the sum of squared errors (SSE) as the benchmark.

SSE computes the aggregate squared distance between each data point and the weighted centroid in FCM or the cluster centroid assigned to the data point in K-means. Stated differently, SSE quantifies the degree to which the data points cluster around the center of each individual cluster. SSE can be stated mathematically as (1). Where x is the data point, C_i represents the i^{th} cluster, and μ_i of the centroid of cluster i.

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

For K-means, SSE only shows how close together groups are around clear centers [7], [29]–[32]. Fuzzy SSE is more sensitive in identifying overlapping data groups in FCM since it takes into account the membership weights of each data point. In the meantime, the elbow method is employed to ascertain the proper value of k so that the resulting clusters are both easily interpretable and statistically supported in both approaches.

The degree of membership is taken into account while using the FCM approach. Because each data point may partially belong to more than one cluster, the SSE can be measured in a fuzzy form. The elbow method is then used by computing the SSE values for range of potential cluster counts, from k = 2 to k = 10. The SSE value will often drop as the number of clusters rises. But the decline becomes less noticeable after a certain point. This point is then used to balance error reduction and model complexity to find the ideal number of clusters.

The silhouette coefficient (SC) is used to verify the quality of the clustering results after the elbow method determines the ideal number of clusters. Strong internal similarity and distinct separation from other clusters are guaranteed by this validation. Therefore, more dependable and accurate clustering findings on datasets linked to graduation requirements can be obtained using a two-stage optimization procedure that includes elbow-SSE to select the candidate number of clusters and silhouette analysis for cluster validation.

2.3. Validating clusters with the silhouette

Silhouette analysis is employed to assess cluster validity by evaluating cohesion and the distinctions generated by the proposed method. This method allows for a more objective assessment of cluster formation accuracy by comparing the extent of resemblance among data within the identical cluster and the degree of difference between clusters [33]–[36]. This formula is defined in (2) [37].

$$s(i) = \frac{b(i) - a(i)}{\max \{ a(i), b(i) \}} \quad (2)$$

Where $b(i)$ denotes the lowest average distance between point i and objects in the closest neighboring cluster; and $a(i)$ represents the average distance between point i and all other objects in the same cluster. The range of values for the SC is -1 to +1. A value around 0 indicates that the data point is close to the boundary among two clusters; a value near +1 shows that the point has been correctly clustered inside its cluster, and a value near -1 indicates that the point is probably in the wrong cluster.

The overall SC is the mean of all the data points. A cluster with higher numbers is closer together and less spread out. After K-means and FCM clustering on multiple factors (GPA, course completion, credit load, attendance, extracurricular activities, socioeconomic background, age, and gender), SC was used in this study. For K-means, SC is used to check whether the student grouping is correct, while in FCM, SC is used to see fuzzy membership. When used with elbow–SSE, SC guarantees accurate and simple grouping for studying factors that affect graduation.

2.4. Clustering algorithms

The study examined data sets from previous research on multi-criteria, which contain both academic and non-academic elements that influence a student's capacity to graduate. Using different types of analysis in different studies supports the idea that using both academic and non-academic factors together makes graduation expectations more accurate than using only one type of factor. Strategies to raise the graduation rate of students don't just focus on improving academic performance; they also look at things like students' drive, the support they get in the classroom, and their social situations to make a more complete and long-lasting plan [38].

2.4.1. Clustering using K-means for multiple criteria

Graduation-related criteria are grouped into academic and non-academic categories using the K-means clustering approach. This technique partitions the data into a fixed number of clusters, each with a centroid. To limit variance within the cluster until a convergence condition is met, the cluster membership is updated by first computing the Euclidean distance from each data point to the centroid [39]–[45]. This approach partitions the data into a predetermined quantity. of clusters, k. A central point, known as the centroid, is then used to represent each cluster. The K-means algorithm iteratively assigns observations to the closest centroid and updates the centroid positions until a stable partition is attained. The calculation typically uses Euclidean distance as shown in (3).

$$d(x,c) = \sqrt{(x_1 - c_1)^2 + (x_2 - c_2)^2 + \cdots + (x_n - c_n)^2} \quad (3)$$

Where $x = (x_1, x_2, \ldots, x_n)$ is a data point and $c = (c_1, c_2, \ldots, c_n)$ is the centroid, update centroid in (4).

$$c_i = \frac{1}{|C_i|}\sum_{x_j \epsilon C_i} x_j \quad (4)$$

Where C_i is a collection of data in cluster I, $|C_i|$ is the number of data in the cluster.

- Convergence: until the maximum iteration limit is reached or the centroid changes only slightly, this process is repeated.
- Results: final centroid: the centroid's ultimate location for every cluster. The cluster label is based on how close it is to the nearest cluster; each data point is labeled accordingly.

2.4.2. Clustering using fuzzy C-means for multiple criteria

The method known as FCM was used in this study to group students together based on academic and non-academic factors that affect their chances of graduating [46]. By using the soft clustering approach, FCM assigns a membership level to one or more clusters in addition to placing each criterion absolutely in one cluster [47]–[49]. How the algorithm continuously modifies centroids and membership values:

i) Set up the membership matrix (u) = [u_{ij}] randomly, ensuring $\sum_{j=1}^{k} u_{ij} = 1$

ii) Compute cluster centroids in (5).

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{i=1}^{N} u_{ij}^{m}} \quad (5)$$

iii) Update membership values are updated according to (6).

$$u_{ij} = \frac{1}{\sum_{i=1}^{k} (\frac{||x_i - c_j||}{||x_i - c_j||})^{\frac{2}{m-1}}} \quad (6)$$

iv) The objective function to be minimized is given in (7).

$$J_m^{(t)} = \sum_{i=1}^{c}\sum_{k=1}^{N} u_{i_k}^{(t)^m} . d^2(x_k, v_i^{(t)}) \quad (7)$$

v) Repeat steps ii–iii again and again until convergence, which means that the change in U is less than the key value.

GPA course completion rate, number of repeated courses, credit load, attendance rate, extracurricular involvement, socioeconomic background, age, and gender are some of the factors that were used in this study. U all values are adjusted to fall between 0 and 1 using the min-max scale.

2.5. Optimization using particle swarm optimization and genetic algorithm

Clustering optimization can be carried out using GA and PSO to improve the quality of clustering findings. While PSO modifies centroid placements based on the best particle's experience and the best global solution, GA creates the optimal centroid configuration through selection, crossover, and mutation operations. The objective function to be minimized is either Jm in FCM or inertia in K-means. By using both strategies, this method can prevent local minima, the PSO and GA improve cluster cohesion and dissociation, and result in more effective clustering procedure on complicated datasets. The formulas are as in (8) and (9).

$$PSO\ Particle\ Position\ v_i(t+1) = w.v_i(t) + c_1 r_1\big(pbest_i - x_i(t)\big) + c_2 r_2\big(gbest - x_i(t)\big)$$
$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (8)$$

$$Fitness\ Function\ GA = \frac{1}{J_m(FCM) or\ Inertia\ (K-Means)} \quad (9)$$

3. RESULTS AND DISCUSSION

3.1. Dataset

Table 1 shows that the data used in this study consists of 33 criteria and 16 experts who provided assessments for each criterion. Weighting is given based on the level of factors influencing student graduation, such as weights 0.43796 (very influential), 0.21898 (influential), and 0.14599 (less influential). These differences in what experts agree on are important in multi-criteria decision-making because they add value to the review and help us understand why people have different ideas. By uniting the views of the majority and minority, the decision will be more complete, stronger, and open to everyone. Table 2 displays the outcomes of applying high-dimensional handling with an autoencoder prior to clustering, utilizing two methods: encoder (compressed features) and decoder (reconstructing the original data), which reduces the number of criteria from 33, 16 features to 2 features.

Table 1. Dataset of assessments for each criterion

Expert	Crit1	Crit2	Crit3	Crit4	Crit5	Crit6	Crit7	Crit8..	Crit33
1	0.43796	0.21898	0.43796	0.43796	0.43796	0.21898	0.43796	0.43796	0.43796
2	0.43796	0.21898	0.43796	0.43796	0.43796	0.21898	0.43796	0.43796	0.43796
3	0.43796	0.43796	0.21898	0.21898	0.43796	0.43796	0.43796	0.43796	0.43796
4	0.43796	0.43796	0.43796	0.43796	0.43796	0.43796	0.21898	0.43796	0.43796
5..	0.43796	0.43796	0.43796	0.43796	0.21898	0.43796	0.21898	0.21898	0.43796
16	0.21898	0.43796	0.43796	0.21898	0.43796	0.21898	0.21898	0.21898	0.21898

Table 2. Feature compression results with autoencoder

Number	0	1	2	3..	32
Latent_Feature_1	0.112381	0.112381	0.112381	0.112381	-2.827449
Latent_Feature_2	0.816452	0.816452	0.816452	0.816452	-2.721458
Cluster	0	0	0	0	1

3.2. Elbow with silhouette

A more objective foundation for figuring out the ideal number of clusters can be obtained by combining the elbow technique with the silhouette score. This combination evaluates the quality of cluster creation by considering the degree of separation between clusters and the potential for overlap. The elbow technique uses the pattern of SSE reduction to ascertain the optimal quantity of clusters.

3.2.1. Elbow

Table 2 displays the outcomes of the SSE computation using (1). Every data point (xi) is determined, grouped according to the centroid (μi), and the squared distance between each data point and the associated centroid $(||x - \mu i||2)$ is then calculated. The SSE is then determined by summing the squared distances of each point within each cluster. One of the computation outputs, for instance, produced an SSE value of 371.911898. In the following round, this value dropped dramatically until it reached an SSE of 323.771577 at k =3. This decline suggests that the best clustering outcome is three clusters [50].

$$x = [0.43796,0.43796,0.43796,0.43796,0.43796,0.43796,0.43796,0.43796,0.43796,0.43796,0.43796,\dots]$$

$$\mu i = [0.3384,0.3008,0.3296,0.3119,0.3097,0.3362,0.2964,0.3008,0.2853,0.2853,0.2920,0.2676,0.3163,..]$$

$$||x - \mu|| = [(0.43796 - 0.3384)^2 + (43796 - 0.3008)^2 + (0.43796 - 0.3296)^2 + (0.43796 - 0.3119)^2 + \cdots]$$

$$SSE = 371.911898$$

3.2.2. Silhouette

Silhouette is then used to process the elbow's k =3 value. For every data point i, two values are computed: $b(i)$, which is the average distance to the closest other cluster used to measure separation, and $a(i)$ [51]–[53]. In (2) is used to calculate the silhouette value, which ranges from −1 to 1. The cluster's quality is evaluated using the silhouette score, where a number near 1 denotes a good cluster, a value near 0 indicates that the data is close to the cluster boundary, and a negative value suggests a potential clustering error. The total silhouette value is calculated by averaging each data score. Finding the average distance between members of each cluster is the first stage in this computation. For instance, cluster 0 has ten members, while cluster 1 has twenty-three. Computing the distances $d(a)$, $d(b)$, and $b(i)$.

$$d(a) = \sqrt{(0.43796 - 0.21898)^2 + \cdots + (0.43796 - 0.21898)^2} = 3.574779$$
$$d(b) = \sqrt{(0.43796 - 0.21898)^2 + \cdots + (0.43796 - 0.21898)^2} = 6.803666$$
$$Silhouette = \frac{6.803666 - 3.574779}{6.803666} = 0.266789$$

Table 3 shows the SSE value dropped from 323.9119 at k = 3 to 173.15291 at k = 10, suggesting that as the number of clusters rises, the distance between data points and centroids decreases. But the SSE value is not the only factor that determines the optimal number of clusters. The silhouette score reached its maximum value of 0.266789 at k = 2. From k = 3 to k = 7, the score dropped; it then increased slightly at k = 8 and k = 9, but remained below the peak value at k = 2. This demonstrates that the optimal number of clusters for this dataset is k = 2. With the elbow point apparent at k = 3, the elbow technique demonstrates that the SSE value falls as the number of clusters rises. Since the silhouette score reaches its greatest value at k = 2, k = 2 is the ideal number of clusters for this dataset.

Table 3. Result of elbow and silhouette average

k	Inertia (SSE)	Silhouette score
2	371.9119	**0.266789**
3	**323.77158**	0.196341
4	290.65492	0.189455
5..	265.10648	0.183155
10	173.15291	0.152549

3.3. Cluster algorithms

This study employs an unlabeled data clustering technique to cluster a multi-criteria dataset based on similarity patterns. It also compares several clustering techniques, including K-means, FCM, K-medoids, DBSCAN, and spectral clustering [54]–[56]. The comparison aims to identify the best way to create precise and significant clusters.

3.3.1. K-means

The value of k is determined by selecting the optimal cluster using the hybrid elbow-silhouette method applied to K-Means, as illustrated in Figure 2. The K-means algorithm resolution process consists of five stages. Table 4 shows the initial stage of randomly determining the centroid [57]–[59], where the data is taken from Table 1. Table 5. shown Euclidean distance to initial centroid, two initial centroids

(cluster 0, cluster 1) are shown in rows (0, 1). One of the data's features is column (0...15). As a result, every centroid possesses multidimensional coordinates based on Euclidean distance in (1).

In comparison to C1, the distance between all data (criteria 1–33) and centroid C0 is the shortest. This indicates that all of the data will probably be placed in cluster C0 during the first clustering cycle. The clustering results, which show two clusters—cluster 0 with 10 members and cluster 1 with 23 members—are shown in Table 6. The member requirements are displayed for each cluster.

$$d,(x_0C_1) = \sqrt{(0.474039-(-0.2061))^2+(0.776439-(-0.33758))^2+(0.353666-(-0.15377))^2+\cdots}$$
$$d(x_0,C_1) = 2.083304$$

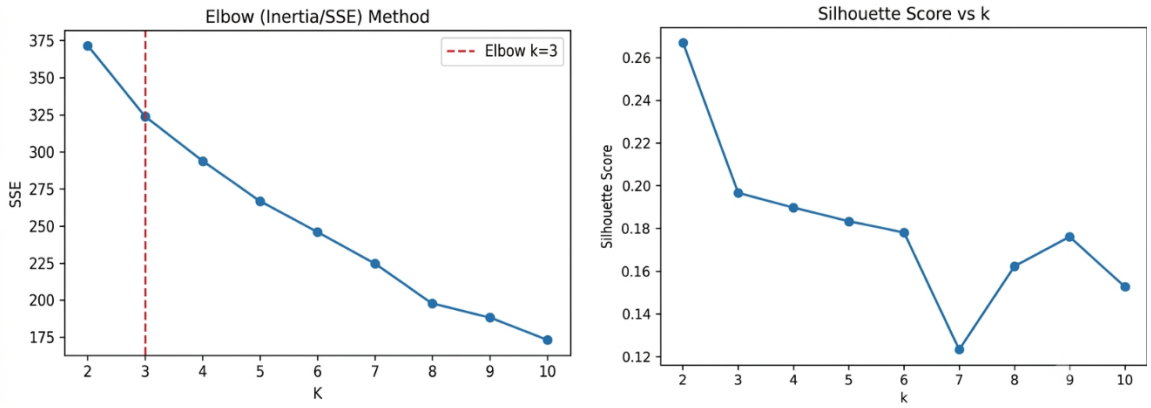


Figure 2. Elbow and silhouette visualization

Table 4. Initial centroid

Cluster	0	1	2	3	4	5	6	7	8..	15
Cluster 0	0.474039	0.776439	0.353666	0.834446	0.341325	0.481077	1.216459	1.140741	1.389483	-0.04153
Cluster 1	-0.2061	-0.33758	-0.15377	-0.3628	-0.1484	-0.20916	-0.5289	-0.49597	-0.60412	0.018057

Table 5. Euclidean distance to initial centroid

Criteria	Centroid 0	Centroid 1
Crit1	2.083304	5.7933
Crit2	3.112823	5.666317
Crit3	2.819333	5.623008
Crit 4 ..	2.091753	5.686948
Crit 33	6.520207	3.55264

Table 6. Cluster membership of K-means

Cluster	Total member (n)	Criteria
C0	10	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
C1	23	10,11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 27, 28, 30, 31, 32

3.3.2. Fuzzy C-means

The clustering procedure in FCM generates multiple primary outputs that depict the generated cluster structure. Each data point's membership matrix (u), cluster centroid values, objective function values, and cluster labels are among these outputs [60]–[65].

i) Membership matrix (u): the membership matrix (u) for 33 data points in the FCM algorithm, indicating the membership levels toward two clusters, is displayed in Table 7. Cluster_0 has a membership of 0.778303 for data point 0, while cluster_1 has a membership of 0.221697, suggesting a greater correlation with cluster_0. For data point 32, cluster_0 is 0.323105, and cluster_1 is 0.676895, suggesting a stronger correlation with cluster_1 [66]. Each column sums to 1, and defuzzification is used to determine the final cluster label: label 0 if cluster_0 membership is higher, and label 1 if cluster_1 membership is higher.

Table 7. Membership (U) and label

Cluster	0	1	2	3	4	5	32
Cluster_0	0.778303	0.714785	0.735585	0.776149	0.698873	0.737631	0.323105
Cluster_1	0.221697	0.285215	0.264415	0.223851	0.301127	0.262369	0.676895
Label	0	0	0	0	0	0	1

ii) Cluster centroids: the weighted sum of F1 values, based on the square of the membership degrees, is divided by the total square of the memberships. To determine the centroid value of cluster_0 for feature F1 in FCM (m = 2). The fuzzy weighted average of F1 in cluster_0 is represented by the resulting centroid value 0.275272.

$$v_{0.1} = \frac{((0.9030)^2 \times 0.43796) + ((0.0970)^2 \times 0.21898) + \cdots + ((0.903)^2 \times 0.43796)}{(0.9030)^2 + (0.903)^2 + \cdots + (0.097)^2} = 0.275272$$

Table 8 presents the centroid values of the two clusters derived by fuzzy C-means. Each number from F1 to F16 denotes the fuzzy weighted average of each feature in a classification. Cluster_0 exhibits centroid values that are predominantly positive, including F1 = 0.275272 and F4 = 0.454959. This signifies that the data inside this cluster are comparatively above average, assuming the data have been normalized. Conversely, cluster_1 exhibits a predominant negative centroid value, such as F1 = -0.21055 and F4 = -0.35017. This signifies that the data within this cluster is subpar.

Table 8. Centroid cluster

Cluster	F1	F2	F3	F4	F5 …	F16
Cluster_0	0.275272	0.437355	0.199145	0.454959	0.147754	0.016061
Cluster_1	-0.21055	-0.32511	-0.15805	-0.35017	-0.09593	-0.07279

iii) The objective function value: Table 9 shows that the fuzzy C-means' objective function value dropped from 309.672222 in the first iteration to 264.242724 in the second, demonstrating notable improvements in the membership matrix (u) and centroid adjustments. The value decline slowed and stabilized after the third iteration. Its value reached 259.352419 at iteration 33, suggesting that the optimization process has stabilized and the algorithm is almost convergent.

Table 9. Iteration and objective function

Iteration	Objective function (Jm)
1	309.672222
2	264.242724
3	263.348574
4	262.827033
5…	262.291954
33	259.352419

The objective function (Jm) is determined by summing the Euclidean distance (d) for each row and the membership matrix (u) according to (5). This procedure is executed for all data and all clusters. The cumulative findings yield a total objective function value of 309.672222 in that iteration.

$$Jm = (0.778303)2\ x\ (0.43796 - 0.275272)\ 2 + (0.221697)2\ x\ (0.21898 - (-0.21055))2 + ..) = 309.672222$$

iv) Label cluster: Table 10 presents the conclusive outcomes of data clustering employing FCM subsequent to the defuzzification procedure. Cluster C0 comprises 10 data points (indices 0–9), representing the data with the highest affiliation to Cluster_0. Cluster C1 comprises 23 data points (indices 10–24, 27, 28, 30, 31, and 32), exhibiting a greater affinity for Cluster_1.

Table 10. Cluster membership of fuzzy C-means

Cluster	Total member (n)	Criteria
C0	11	0, 1, 2, 3, 4, 5, 6, 7, 8, 9,15
C1	22	10,11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 27, 28, 30, 31, 32

With 11 data points in cluster C0 and 22 in cluster C1, Table 11 demonstrates that both K-means and FCM generate the same cluster structure. The primary distinction lies in the assignment mechanism: FCM uses soft assignment, which allows data points to have membership values in multiple clusters simultaneously, making it more appropriate for data with ambiguous cluster boundaries, whereas K-means utilizes hard assignment.

Table 11. Analysis comparison hybrid elbow+ silhouette, K-means, and FCM

Comparison	Cluster	Number of members	Criteria
Hybrid elbow + silhouette + K-means	C0	10	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
	C1	23	10,11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 27, 28, 30, 31, 32
Hybrid elbow + silhouette + fuzzy C-means	C0	11	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 15
	C1	22	10,11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 27, 28, 30, 31, 32

This study clarifies the factors that affect student graduation in terms of ethics and justice. Two clusters are produced by this study: cluster 1 (influential non-academic criteria) and cluster 0 (very influential academic criteria). Demographic and socioeconomic factors are positioned as supplementary settings rather than as the primary foundation for clustering or determining a criteria.

3.4. Depth of comparison using K-medoids, DBSCAN, spectral cluster, GMM, and DL

Table 12 shows the results of the comparison analysis of K-means, FCM, K-medoids, DBSCAN, and Spectral with the elbow+silhouette hybrid. K-means, FCM, GMM, and DL clustering methods yield the same results: a small cluster C0 and a large cluster C1. For centroid-based techniques, the dataset structure is quite obvious. A more equal distribution of K-medoids shifts some data from the big cluster to C0, strengthening its resistance to outliers. Spectral clustering with inverted clusters emphasizes data similarity and global connectedness [67]–[71]. DBSCAN generates a large number of minor clusters that are not directly comparable to two major clusters but are useful for detecting outliers and local density structures [72]–[76].

Using the K-means, FCM, K-medoids, DBSCAN, spectral clustering, GMM, and DL algorithms based on the distance of data to the centroid and the distance between centroids, Figure 3 illustrates the differences between the hybrid elbow and silhouette. K-means, fuzzy C-means, GMM, and DL clustering are the most straightforward methods. The centroids are quite clearly separated in all four techniques. Nonetheless, GMM and K-means appear to be the most practical options if we must select the most visually stable, as they yield two comparatively distinct and fairly explicable groups. Following a comparison of each algorithm's clustering performance in Table 12, a criterion selection technique was applied to identify the most influential factors. The Shapley additive explanations (SHAP) method was used to evaluate the significance of these factors in shaping clustering outcomes.

Table 12. Comparison analysis of K-means, FCM, K-medoids, DBSCAN, spectral, GMM, and DL with elbow + silhouette

Comparison	Cluster	Number of members	Criteria
Hybrid elbow + silhouette + K-means	C0	10	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
	C1	23	10,11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 32
Hybrid elbow + silhouette + Fuzzy C-means	C0	11	0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 15
	C1	22	10,11, 12, 13, 14, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 32
Hybrid elbow + silhouette + K-medoids	C0	22	0,1,2,3,5,8,10,11,13,15,16,19,20,22,23,24,25,26,28,29,31,32
	C1	11	4,6,7,9,12,14,17,18,21,27,30
Hybrid elbow + silhouette + spectral clustering	C0	32	10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32
	C1	10	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Hybrid elbow + silhouette + DBSCAN	32 clusters are formed, with cluster 6 having two criteria—6 and 7.		
Hybrid elbow + silhouette + GMM	C0	10	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
	C1	23	10,11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 32
Hybrid elbow + silhouette + DL	C0	10	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
	C1	23	10,11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30, 31, 32

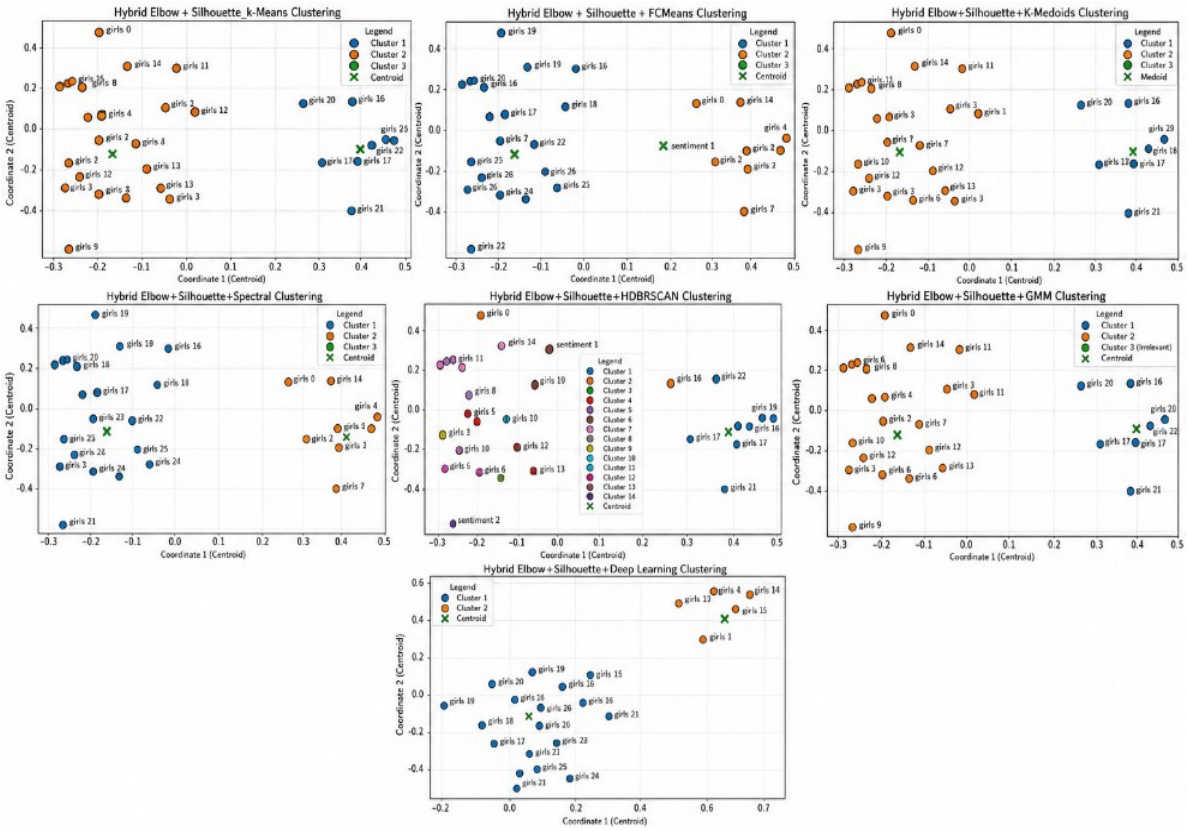


Figure 3. Comparison of hybrid elbow+silhouette using K-means, FCM, K-medoids, spectral clustering, DBSCAN, GMM, and DL clustering

The range of influence from strong to low is illustrated in Figure 4. K-means is quick, memory-efficient, and still generates two clusters, Table 13 demonstrates that it is the most balanced method. Although FCMean is faster than K-means, it uses a little more memory. Because DBSCAN can only create a single cluster with the given parameters, it is not advised. While K-medoids is still useful but needs further cluster quality assessment, spectral clustering is less effective for large data.

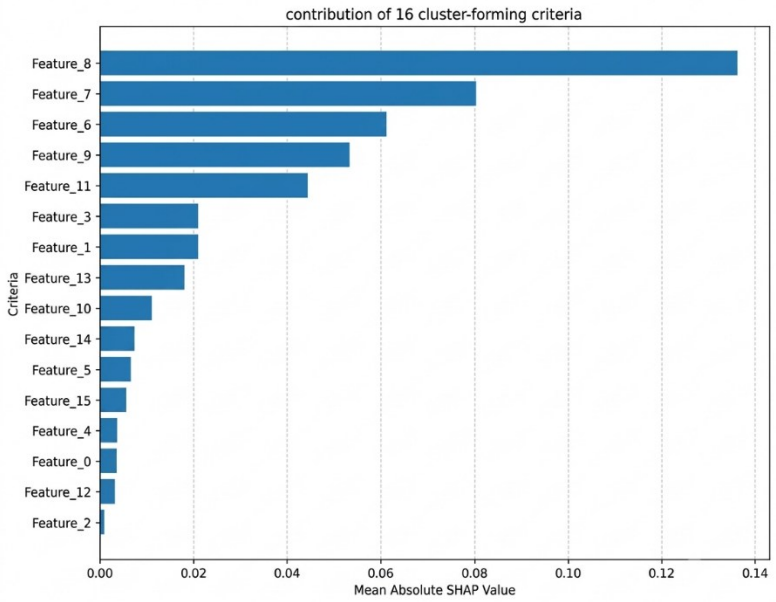


Figure 4. Feature importance ranking SHAP

Table 13. Efficiency benchmark

No	Algorithm	Scalability	n_F	Runtime_mean (sec)	Memory_mean (MB)	C	Noise_ Points	Labels_Last_Run
0	K-means	500	16	0.031039	0.188286	2	0	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, ...
1	FC-mean	500	16	0.012761	0.363861	2	0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0, ...
2	DBSCAN	500	16	0.035176	4.065701	1	0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
3	Spectral clustering	500	16	0.102301	7.647919	2	0	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, ...
4	K-medoids	500	16	0.013960	3.426704	2	0	[0, 0, 0, 1, 1, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, ...
5	K-means	700	16	0.054815	0.238579	2	0	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, ...
6	FCMean	700	16	0.032087	0.482880	2	0	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, ...
7	DBSCAN	700	16	0.079708	7.959934	1	0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, ...
8	Spectral clustering	700	16	0.281860	14.975927	2	0	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, ...
9	K-medoids	700	16	0.018340	6.539619	2	0	[1, 1, 0, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 1, 1, ...

3.5. Optimization using particle swarm optimization and genetic algorithm

Table 14 shows that if inter-cluster separation is the primary indicator, K-means + PSO is the most logical approach to select as the main outcome. The optimal fuzzy comparison is FCM + PSO if the study requires interpreting membership degrees. Because the distance between centroids is too small, FCM + GA approach produces fewer distinguishable clusters and should not be employed as the primary method.

Table 14. Comparative result: baseline, PSO, and GA

Method	C0	C1	Objective value	Distance between centroids	Data to centroid C0	Data to centroid C1	Interpretation
Elbow + silhouette + K-means	10	23		0.550765	0.584326	0.490721	K-means shows two clusters with a clear separation.
Elbow + silhouette + fuzzy C-means	11	22		0.332845	0.504167	0.485086	FCM is able to show the degree of membership, but the separation of centroids is lower.
Elbow + silhouette + K-means + PSO	10	23	5.601771	0.577250	0.609683	0.490822	The separation of the centroid is the strongest.
Elbow + silhouette + K-means + GA	23	10	5.391484	0.543394	0.492977	0.578223	Lowest SSE so that it is the most compact internally.
Elbow + Silhouette + FCM + PSO	11	22	3.701355	0.267310	0.493568	0.481103	Best fuzzy objective; more feasible than GA-FCM.
Elbow + Silhouette + FCM + GA	11	22	3.740730	0.103449	0.476656	0.474827	The centroids are too close to one another, resulting in a very poor separation.

3.6. Cluster evaluation

Table 15 shows that K-means + PSO is the best clustering assessment method since, together with K-means + GA, it has the highest silhouette score but a lower objective value. In comparison to the first silhouette scores, the last silhouette scores values remain lower. Thus, altogether, it is not yet possible to say that the optimization findings have successfully enhanced clustering performance. To improve the results, more testing is required, such as experimenting with different values of k, increasing the number of PSO/GA iterations, expanding the population/particle size, carrying out feature selection or normalization, and incorporating the original labels if the values of purity, NMI, and ARI are to be computed.

Table 15. Cluster evaluation

No	Method	Objective value	Last_silhouette score	Purity	NMI	ARI	Cluster label	First_ silhouette
1	K-means + PSO	5.782470	0.254177	NaN	NaN	NaN	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, ...]	0.266789
2	K-means + GA	5.925785	0.254177	NaN	NaN	NaN	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, ...]	0.266789
3	FCM + PSO	3.721401	0.223247	NaN	NaN	NaN	[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, ...]	0.266789
4	FCM + GA	3.959529	0.212373	NaN	NaN	NaN	[1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 0, 0, ...]	0.266789

3.7. Validation

Using a new dataset and the cross-dataset validation metrics method, Table 16 displays the clustering findings. Although there are still many mistakes and overlaps between clusters, the ARI of 0.478 indicates moderate similarity. Although there is still considerable uncertainty, the NMI value of 0.433 indicates moderate shared knowledge, with the clusters capturing some of the original label patterns. Table 17 compares clustering

between AI and experts. Members of clusters 0 and 1 correspond to the categories highly influential and influential, as defined by AI and experts. This indicates that the clustering algorithm can capture patterns in academic criteria that align with expert assessments. The input values in Table 1 of the dataset are those provided by the expert. Figure 5 shows the clustering K-means and FCM are fairly robust at differentiating the two primary groups, particularly along the X-axis, as indicated by the error bars and standard deviations. The consistency of results within the cluster must be preserved throughout the experimental assessment process, though, as there is still uncertainty or internal variance on the Y-axis.

Table 16. New dataset (cross-dataset prediction)

No.	Cr1	Cr2	Cr3	Cr4..	Cr33	Actual	Cluster
1	0.2304	0.2304	0.2304	0.2304	0.1152	On time	1
2..	0.2304	0.2304	0.2304	0.2304	0.1152	On time	1
15	0.2304	0.2304	0.2304	0.2304	0.1152	Not on time	1
16	0.2304	0.2304	0.2304	0.2304	0.1152	On time	1

Table 17. Comparison of AI and expert grouping

AI clustering results	Factors	Criteria	Expert	Interpretation
Cluster 0	Very influential	GPA, course values, amount of credits taken, and number of credits passed/not passed.	0.43796	According to the expert
Cluster 1	Influential	Family income financial aid, attendance and participation, extracurricular activities, tuition fees, salary status	0.21898	According to the expert
Cluster 2	Quite influential	Extracurricular activities, and college status	0.14599	According to the expert

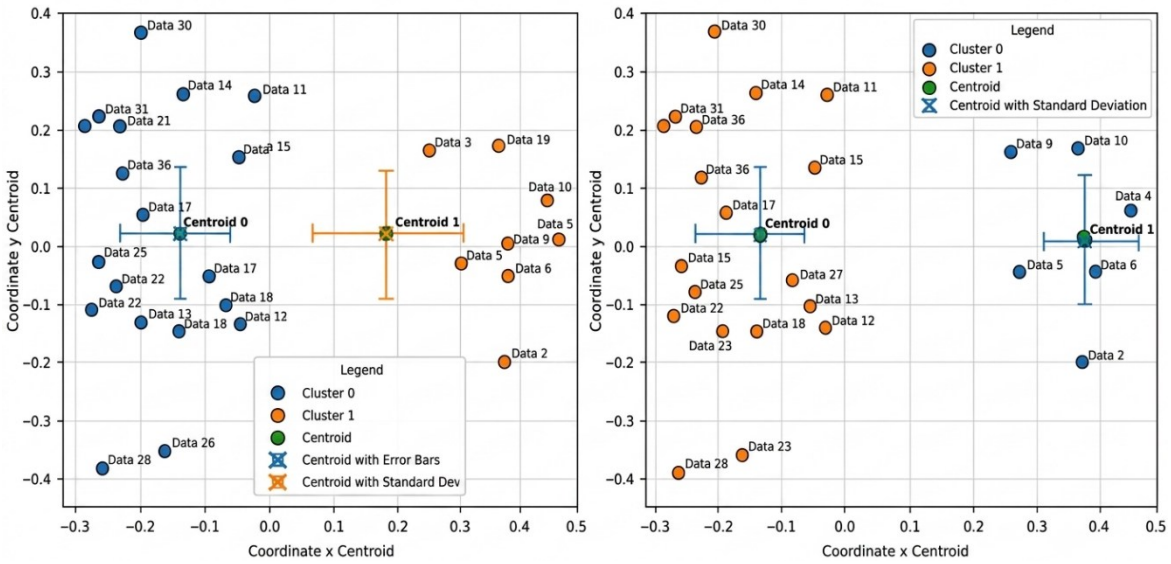


Figure 5. Error bars of K-means and FCM

3. CONCLUSION

This analysis determines that two clusters represent the best quantity according to the integrated elbow and silhouette methodologies. K-means and GMM exhibited the most distinct and persistent cluster separation among the evaluated techniques. Post-optimization, K-means + PSO emerged as the preferred method for maximizing inter-cluster separation, whilst FCM + PSO was more appropriate for interpreting unclear or overlapping membership degrees. The FCM + GA method is inadvisable, as the resulting clusters are difficult to discern due to the minimal distances between centroids. Furthermore, the optimization outcomes have not significantly improved clustering performance, as the final silhouette value remains inferior to the original. Additional testing is required to verify the purity, NMI, and ARI values. This encompasses testing with various k values, increasing the number of PSO/GA iterations, increasing the particle or population size, performing feature selection or normalization, and incorporating original labels. The cross-dataset validation results, with an ARI of 0.478 and an NMI of 0.433, demonstrate that the

clustering model exhibits considerable consistency on the new dataset. The findings suggest that the model can discern some patterns in the reference labels, despite the presence of overlaps and categorization inaccuracies. The clustering outcomes align with expert evaluations, indicating that academic aspects are highly influential, whereas socio-economic and institutional elements are also impactful in student graduation. This paradigm facilitates more precise institutional decision-making by associating each cluster with actions such as achievement development, remedial programs, counseling, monitoring, or academic coaching. This method, while akin to segmentation techniques in other domains, encounters obstacles like data heterogeneity, instances situated at the peripheries of clusters, and practical implementation issues. Consequently, the model ought to serve as a decision-support instrument rather than a replacement for institutional discernment. Subsequent development must integrate a continuous learning mechanism to enable the model to adapt to changes in criteria that influence student graduation.

FUNDING INFORMATION

Support for this study comes from the University of Muhammadiyah Makassar in providing publication costs for this journal.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ida Mulyadi	✓	✓	✓		✓	✓			✓	✓	✓			
Titik Khawa Abdul Rahman		✓			✓					✓				
Muhammad Faisal					✓	✓			✓		✓			

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

All of the participants in this study have given their individual informed consent.

ETHICAL APPROVAL

Human use research has been authorized by the institutional review board or comparable committee of the writing institution and has adhered with all applicable national rules and institutional procedures in line with the Declaration of Helsinki's principles.

DATA AVAILABILITY

Since no new data were generated or examined for this investigation, data availability is not relevant to this paper.

REFERENCES

[1] B. A. D. S. R. Rocha and A. T. Júnior, "Predictive factors of graduation delay in a medical program: a retrospective cohort study in Brazil, 2010-2016," *Revista Brasileira de Educação Médica*, vol. 44, no. 1, 2020, doi: 10.1590/1981-5271v44.1-20190205.ing.

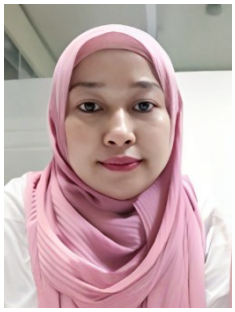
[2] P. Atorkey and C. Owiredua, "Clustering of multiple health risk behaviours and association with socio-demographic characteristics and psychological distress among adolescents in Ghana: a latent class analysis," *SSM - Population Health*, vol. 13, 2021, doi: 10.1016/j.ssmph.2020.100707.

[3] Y. Hasnataeni, M. R. Nurhambali, R. Ardhani, S. Hafsah, and A. M. Soleh, "Comparison of clustering analysis of K-means, K-medoids, and fuzzy C-means methods: case study of school accreditation in West Java," *Journal of Soft Computing Exploration*, vol. 6, no. 2, pp. 79–88, 2025, doi: 10.52465/joscex.v6i2.575.
[4] E. Sanjari, F. M. Dehkordi, and H. R. Shahraki, "Clustering undergraduate students based on academic burnout and satisfaction from the field using partitioning around medoid," *Computational and Mathematical Methods in Medicine*, vol. 2023, 2023, doi: 10.1155/2023/8898939.
[5] E. M. S. Rochman, Miswanto, and H. Suprajitno, "Comparison of clustering in tuberculosis using fuzzy C-means and K-means methods," *Communications in Mathematical Biology and Neuroscience*, vol. 2022, 2022, doi: 10.28919/cmbn/7335.
[6] Z. Salman and A. Alomary, "Performance of the K-means and fuzzy C-means algorithms in big data analytics," *International Journal of Information Technology*, vol. 16, no. 1, pp. 465–470, 2024, doi: 10.1007/s41870-023-01436-y.
[7] W. Junthopas and C. Wongoutong, "Pre-determining the optimal number of clusters for K-means clustering using the parameters package in R and distance metrics," *Applied Sciences*, vol. 15, no. 21, 2025, doi: 10.3390/app152111372.
[8] A. N. Shabrina, M. Afdal, and S. Monalisa, "Comparison of K-means, K-medoids, and fuzzy C-means algorithms for clustering drug user's addiction levels," *Journal of Intelligent Systems*, vol. 6, no. 2, pp. 113–122, 2023, doi: 10.37396/jsc.v6i2.313.
[9] O. N. Kenger, Z. D. Kenger, E. Ozceylan, and B. Mrugalska, "Clustering of cities based on their smart performances: a comparative approach of fuzzy C-means, K-means, and K-medoids," *IEEE Access*, vol. 11, pp. 134446–134459, 2023, doi: 10.1109/ACCESS.2023.3333753.
[10] B. M. Angadi and M. S. Kakkasageri, "K-means and fuzzy-based hybrid clustering algorithm for WSN," *International Journal of Electronics and Telecommunications*, vol. 69, no. 4, pp. 793–801, 2023, doi: 10.24425/ijet.2023.147703.
[11] A. Fadlurohman, N. A. N. Roosyidah, and N. A. Annisa, "Social vulnerability analysis in Central Java with K-medoids algorithm," *Parameter: Journal of Statistics*, vol. 4, no. 2, pp. 83–92, 2024, doi: 10.22487/27765660.2024.v4.i2.17131.
[12] C. Shi, B. Wei, S. Wei, W. Wang, H. Liu, and J. Liu, "A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm," *Eurasip Journal on Wireless Communications and Networking*, vol. 2021, no. 1, 2021, doi: 10.1186/s13638-021-01910-w.
[13] S. Wang, W. Ma, and J. Zhan, "A clustering multi-criteria decision-making method for large-scale discrete and continuous uncertain evaluation," *Entropy*, vol. 24, no. 11, 2022, doi: 10.3390/e24111621.
[14] S. M. Miraftabzadeh, C. G. Colombo, M. Longo, and F. Foiadelli, "K-means and alternative clustering methods in modern power systems," *IEEE Access*, vol. 11, pp. 119596–119633, 2023, doi: 10.1109/ACCESS.2023.3327640.
[15] A. Aljohani, "Optimizing patient stratification in healthcare: a comparative analysis of clustering algorithms for EHR data," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, 2024, doi: 10.1007/s44196-024-00568-8.
[16] H. Li and J. Wang, "From soft clustering to hard clustering: a collaborative annealing fuzzy C-means algorithm," *IEEE Transactions on Fuzzy Systems*, vol. 32, no. 3, pp. 1181–1194, 2024, doi: 10.1109/TFUZZ.2023.3319663.
[17] X. Wu, X. Shen, C. Wei, X. Xie, and J. Li, "A hybrid particle swarm optimization-genetic algorithm for multiobjective reservoir ecological dispatching," *Water Resources Management*, vol. 38, no. 6, pp. 2229–2249, 2024, doi: 10.1007/s11269-024-03755-6.
[18] S. Cordero, "A multi-algorithm clustering framework to optimize plant-knowledge pattern detection in ethnobotanical research," *Journal of Ethnobiology and Ethnomedicine*, vol. 22, no. 1, 2026, doi: 10.1186/s13002-026-00849-w.
[19] M. Yağcı, "Educational data mining: prediction of students' academic performance using machine learning algorithms," *Smart Learning Environments*, vol. 9, no. 1, 2022, doi: 10.1186/s40561-022-00192-z.
[20] H. Elkadry, M. Shamsuzzaman, S. Piya, S. Haridy, H. Bashir, and M. Khadem, "A fuzzy Delphi-AHP framework for identifying and prioritizing factors affecting students' satisfaction in public high schools: Insights from the United Arab Emirates," *Journal of Engineering Research*, vol. 13, no. 2, pp. 596–610, 2025, doi: 10.1016/j.jer.2023.12.008.
[21] C. A. Silvera, J. C.-Nino, J. D. Manotas, N. S. Quintero, and Z. H.-Fontalvo, "Model based in Crowdfunding and institutional accompaniment for the timely graduation in undergraduate students," *Procedia Computer Science*, vol. 220, pp. 991–997, 2023, doi: 10.1016/j.procs.2023.03.137.
[22] D. Sekiwu, "Investigating the relationship between school attendance and academic performance in universal primary education: the case of Uganda," *African Educational Research Journal*, vol. 8, no. 2, pp. 152–160, 2020, doi: 10.30918/aerj.82.20.017.
[23] A. I. Adekitan and O. Salau, "The impact of engineering students' performance in the first three years on their graduation result using educational data mining," *Heliyon*, vol. 5, no. 2, 2019, doi: 10.1016/j.heliyon.2019.e01250.
[24] H. Mastour, T. Dehghani, E. Moradi, and S. Eslami, "Early prediction of medical students' performance in high-stakes examinations using machine learning approaches," *Heliyon*, vol. 9, no. 7, 2023, doi: 10.1016/j.heliyon.2023.e18248.
[25] I. Y. Panessai *et al.*, "Predicting students' academic performance: a review for the attribute used," *International Journal of Academic Research in Business and Social Sciences*, vol. 11, no. 4, 2020, doi: 10.6007/ijarbss/v11-i4/9706.
[26] M. Jüttler, "Predicting economics student retention in higher education: the effects of students' economic competencies at the end of upper secondary school on their intention to leave their studies in economics," *PLoS ONE*, vol. 15, no. 2, 2020, doi: 10.1371/journal.pone.0228505.
[27] N. Lounsbury, A. C. Perry, and L. Arric, "The effect of socioeconomic disadvantage for on-time graduation in an accelerated doctor of pharmacy program," *Currents in Pharmacy Teaching and Learning*, vol. 15, no. 10, pp. 868–873, 2023, doi: 10.1016/j.cptl.2023.07.023.
[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 Computer Science*, vol. 234, pp. 96–105, 2024, doi: 10.1016/j.procs.2024.02.156.
[29] M. Gul and M. A. Rehman, "Big data: an optimized approach for cluster initialization," *Journal of Big Data*, vol. 10, no. 1, 2023, doi: 10.1186/s40537-023-00798-1.
[30] P. Raval *et al.*, "Custom ulna megaprosthesis use in revision total elbow replacement," *Journal of Shoulder and Elbow Surgery*, vol. 35, no. 7, pp. e739–e746, 2026, doi: 10.1016/j.jse.2025.11.022.
[31] X. Cheow, S. Schramm, A. Voss, and S. Greiner, "Injury chronicity does not affect outcomes following ligament repair with suture tape augmentation for post-traumatic elbow instability," *Journal of Shoulder and Elbow Surgery*, vol. 35, no. 6, pp. 1529–1538, 2026, doi: 10.1016/j.jse.2025.11.011.
[32] S. C. Clark *et al.*, "Operative management of posterolateral rotatory instability in females: favorable 6-year outcomes in primary and revision elbows," *Journal of Shoulder and Elbow Surgery*, vol. 35, no. 5, pp. 1291–1295, 2026, doi: 10.1016/j.jse.2025.08.025.
[33] A. Davison, S. Barron, A. Blood, and R. Taylor, "Fractal complexity in visual nature: perceptual preferences of leaf silhouettes and implications for biophilic design," *Trees, Forests and People*, vol. 22, 2025, doi: 10.1016/j.tfp.2025.100996.
[34] A. M. Ikotun, F. Habyarimana, and A. E. Ezugwu, "Cluster validity indices for automatic clustering: a comprehensive review," *Heliyon*, vol. 11, no. 2, 2025, doi: 10.1016/j.heliyon.2025.e41953.
[35] M. Jamalian and H. V.-Nejad, "Effective clustering of big text data: empirical evaluation and comparative analysis," *Egyptian Informatics Journal*, vol. 32, 2025, doi: 10.1016/j.eij.2025.100812.

[36] H. Y. Chuang and W. Chou, "SilhouetteScoreinR: beyond traditional network layouts by leveraging local cohesion and nearest neighbor separation," *MethodsX*, vol. 15, 2025, doi: 10.1016/j.mex.2025.103622.
[37] S. Alrabie and A. Barnawi, "Enhancing heart sound classification with iterative clustering and silhouette analysis: an effective preprocessing selective method to diagnose rare and difficult cardiovascular cases," *CMES - Computer Modeling in Engineering and Sciences*, vol. 144, no. 2, pp. 2481–2519, 2025, doi: 10.32604/cmes.2025.067977.
[38] K. E. Setiawan, A. Kurniawan, A. Chowanda, and D. Suhartono, "Clustering models for hospitals in Jakarta using fuzzy C-means and K-means," *Procedia Computer Science*, vol. 216, pp. 356–363, 2022, doi: 10.1016/j.procs.2022.12.146.
[39] A. Jahwar, "Meta-heuristic algorithms for K-means clustering: a review," *PalArch's Journal of Archaeology of Egypt/Egyptology*, vol. 17, no. 7, pp. 7–9, 2021.
[40] I. K. Khan *et al.*, "Addressing limitations of the K-means clustering algorithm: outliers, non-spherical data, and optimal cluster selection," *AIMS Mathematics*, vol. 9, no. 9, pp. 25070–25097, 2024, doi: 10.3934/math.20241222.
[41] E. Hassan *et al.*, "A hybrid K-means++ and particle swarm optimization approach for enhanced document clustering," *IEEE Access*, vol. 13, pp. 48818–48840, 2025, doi: 10.1109/ACCESS.2025.3535226.
[42] M. Barzegar, S. M. Cherati, D. J. Pasadas, C. Pernechele, A. L. Ribeiro, and H. G. Ramos, "Baseline-free damage imaging of CFRP lap joints using K-means clustering of guided wave signals," *Mechanical Systems and Signal Processing*, vol. 229, 2025, doi: 10.1016/j.ymssp.2025.112562.
[43] F. J. Kanedi, A. Utamima, E. N. S. Mufida, and S. M. Kimani, "IT-enabled route planning: k-means clustering and GA-SA metaheuristics for high-school brochure delivery," *Procedia Computer Science*, vol. 269, pp. 1191–1200, 2025, doi: 10.1016/j.procs.2025.09.060.
[44] L. Liu, "Application of K-means supported by clustered systems in big data association rule mining," *Systems and Soft Computing*, vol. 7, 2025, doi: 10.1016/j.sasc.2025.200211.
[45] L. Zhang, "Research on K-means clustering algorithm based on MapReduce distributed programming framework," *Procedia Computer Science*, vol. 228, pp. 262–270, 2023, doi: 10.1016/j.procs.2023.11.030.
[46] H. Han, "Fuzzy clustering algorithm for university students' psychological fitness and performance detection," *Heliyon*, vol. 9, no. 8, 2023, doi: 10.1016/j.heliyon.2023.e18550.
[47] G. S. M. Khamis, N. S. Alqahtani, S. M. Alanazi, M. M. Alruwaili, M. S. Alenazi, and M. A. Alrawaili, "Using fuzzy C-means clustering and PCA in public health: a machine learning approach to combat CVD and obesity," *Informatics in Medicine Unlocked*, vol. 57, 2025, doi: 10.1016/j.imu.2025.101666.
[48] A. A. Subuh, S. H. A. Kaboli, M. Waqar, and F. Vallée, "Hybrid model for cleaning abnormal data of wind turbine power curve based on machine learning approaches," *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 13, 2025, doi: 10.1016/j.prime.2025.101043.
[49] S. E. Hashemi, F. G.-Jouybari, and M. H.-Keshteli, "A fuzzy C-means algorithm for optimizing data clustering," *Expert Systems with Applications*, vol. 227, 2023, doi: 10.1016/j.eswa.2023.120377.
[50] R. E. Yuliana, D. M. Ulya, and M. Jamhuri, "Mall customer segmentation using k-means clustering optimized by the elbow method," *Jurnal Riset Mahasiswa Matematika*, vol. 4, no. 5, pp. 273–286, 2025, doi: 10.18860/jrmm.v4i5.33389.
[51] H. Lai, T. Huang, B. L. Lu, S. Zhang, and R. Xiaog, "Silhouette coefficient-based weighting K-means algorithm," *Neural Computing and Applications*, vol. 37, no. 5, pp. 3061–3075, 2025, doi: 10.1007/s00521-024-10706-0.
[52] P. M. Hasugian, D. Mathelinea, S. Simamora, and P. B. N. Simangunsong, "Comparative evaluation of data clustering accuracy through integration of dimensionality reduction and distance metric," *MATRIK : Jurnal Manajemen, Teknik Informatika dan Rekayasa Komputer*, vol. 24, no. 3, pp. 577–588, 2025, doi: 10.30812/matrik.v24i3.5057.
[53] K. Patidar, D. P. Tiwari, M. Saxena, I. Khan, V. Khatarkar, and K. N. Pandagre, "Performance prediction of students data by using K-means and KNN algorithms," in *AIP Conference Proceedings*, 2025, vol. 3175, no. 1. doi: 10.1063/5.0256108.
[54] C. Wu *et al.*, "Efficient boundary-seeking clustering based on reverse k-nearest neighbor and average direction centrality metric," *Expert Systems with Applications*, vol. 303, 2026, doi: 10.1016/j.eswa.2025.130589.
[55] B. Rolf *et al.*, "A review on unsupervised learning algorithms and applications in supply chain management," *International Journal of Production Research*, vol. 63, no. 5, pp. 1933–1983, 2025, doi: 10.1080/00207543.2024.2390968.
[56] Q. Huang and Y. Huang, "A forest firefighting task area division method based on the SW-DBSCAN algorithm," *Scientific Reports*, vol. 16, no. 1, 2026, doi: 10.1038/s41598-026-42407-0.
[57] M. Li, S. Guo, and J. Chen, "Construction of aesthetic education evaluation model based on K-means clustering analysis in the context of artificial intelligence," *Journal of Computational Methods in Sciences and Engineering*, vol. 25, no. 1, pp. 728–743, 2025, doi: 10.1177/14727978251321946.
[58] Y. Cao, P. Leung, and A. Monod, "K-means clustering for persistent homology," *Advances in Data Analysis and Classification*, vol. 19, no. 1, pp. 95–119, 2025, doi: 10.1007/s11634-023-00578-y.
[59] G. Qiu, Y. Zhao, and X. Gui, "Efficient privacy-preserving outsourced K-means clustering on distributed data," *Information Sciences*, vol. 674, 2024, doi: 10.1016/j.ins.2024.120687.
[60] S. Askari, "Fuzzy C-means clustering algorithm for data with unequal cluster sizes and contaminated with noise and outliers: Review and development," *Expert Systems with Applications*, vol. 165, 2021, doi: 10.1016/j.eswa.2020.113856.
[61] R. Saltos, I. Carvajal, F. Crespo, and R. Weber, "Novel validity indices for dynamic clustering and an improved dynamic fuzzy C-means," *Egyptian Informatics Journal*, vol. 29, 2025, doi: 10.1016/j.eij.2025.100613.
[62] A. A. Wani, "Comprehensive analysis of clustering algorithms: exploring limitations and innovative solutions," *PeerJ Computer Science*, vol. 10, pp. 1–45, 2024, doi: 10.7717/PEERJ-CS.2286.
[63] C. Wu and J. Hou, "Asymmetric deviation entropy regularization for semi-supervised fuzzy C-means clustering and its fast Algorithm," *Big Data Research*, vol. 43, 2026, doi: 10.1016/j.bdr.2025.100586.
[64] W. Li, Z. Yu, and J. Qiao, "Recognition of imbalanced working conditions in wastewater treatment process with a robust possibilistic fuzzy C-means algorithm," *Process Safety and Environmental Protection*, vol. 210, 2026, doi: 10.1016/j.psep.2026.108666.
[65] X. Yang, H. Xing, X. Su, and W. Pedrycz, "Dynamic thunderstorm capture and path imaging driven by information granulation and fuzzy C-means clustering," *Knowledge-Based Systems*, vol. 336, 2026, doi: 10.1016/j.knosys.2026.115312.
[66] I. Mikrou and N. S. Sapidis, "A systematic evaluation of clustering algorithms against expert-derived clustering," *Operations Research Forum*, vol. 6, no. 2, 2025, doi: 10.1007/s43069-025-00453-w.
[67] Y. Zhu and D. Huang, "A communication efficient boosting method for distributed spectral clustering," *Pattern Recognition*, vol. 176, 2026, doi: 10.1016/j.patcog.2026.113168.
[68] B. Yang, X. Zhang, Y. Luo, F. Nie, F. Wang, and B. Chen, "Efficient superpixel-guided global-local spectral clustering for large-scale HSI," *Neurocomputing*, vol. 678, 2026, doi: 10.1016/j.neucom.2026.133190.

[69] H. Jiang *et al.*, "Spectral clustering dimensionality reduction in wheat quality detection based on hyperspectral data," *Infrared Physics and Technology*, vol. 154, 2026, doi: 10.1016/j.infrared.2026.106400.
[70] G. Jiang, W. Chen, Y. Liu, L. Shi, J. Peng, and H. Wang, "Hybrid anchor graph learning and tensorized spectral embedding fusion for multi-view clustering," *Neurocomputing*, vol. 673, 2026, doi: 10.1016/j.neucom.2026.132768.
[71] R. Feng, C. Zhong, J. Qian, T. Pan, and X. Yue, "A fair spectral clustering with weighted fairness constraints," *Pattern Recognition*, vol. 173, 2026, doi: 10.1016/j.patcog.2025.112821.
[72] R. Luo, T. Li, R. Pu, J. Yang, D. Tang, and L. Yang, "NaGB-DBSCAN: an improved DBSCAN clustering algorithm by natural neighbor and granular-ball," *Information Sciences*, vol. 719, 2025, doi: 10.1016/j.ins.2025.122445.
[73] H. Wan, C. Lu, and Y. Cui, "Improved WOA-DBSCAN online clustering algorithm for radar signal data streams," *Sensors*, vol. 25, no. 16, 2025, doi: 10.3390/s25165184.
[74] K. Song, H. Wang, S. Yang, S. Liu, and L. Yan, "Research on parallel DBSCAN algorithm based on ternary optical computer," *Optics Communications*, vol. 608, 2026, doi: 10.1016/j.optcom.2026.133079.
[75] A. Ailane *et al.*, "DADM2D-SFL: decentralised aggregation framework based on DBSCAN malicious model detection for secure federated learning," *Computer Networks*, vol. 276, 2026, doi: 10.1016/j.comnet.2025.111987.
[76] A. Garah, N. Mbarek, and S. Kirgizov, "Enhancing IoT object integrity and energy efficiency through DBSCAN and fuzzy logic-based self-management framework," *Journal of Systems Architecture*, vol. 175, 2026, doi: 10.1016/j.sysarc.2026.103751.

BIOGRAPHIES OF AUTHORS



Ida Mulyadi received her S.Kom. degree from the Department of Information Systems, Universitas Diponegoro, Makassar, Indonesia, at 2012, and an M.T. degree from the Masters in Electrical Engineering, Hasanuddin University, Makassar, Indonesia in 2016. She is currently a lecturer and researcher at the Department of Informatics, Universitas Muhammadiyah Makassar, Indonesia. Her current research is on big data and data mining. She can be contacted at email: idamulyadi@unismuh.ac.id.

Titik Khawa Abdul Rahman received B.Sc. in Electrical and Electronics Engineers from Loughborough University of Technology, United Kingdom. She holds a Ph.D. in Power Systems from the Universiti Malaya. She is the current deputy vice chancellor of department of School of Science and Technology, Asia E University, Malaysia. Her expertise includes artificial intelligence, biologically inspired optimization and machine learning, power system analysis and operation, and smart grid. She can be contacted at email: titik.khawa@aeu.edu.my.

Muhammad Faisal holds a S.SI. degree from the Information Systems Study Program, STMIK Profesional, Makassar, Indonesia, in 2011, and an M.T. degree from the Master of Electrical Technology Study Program, Graduate School, Hasanuddin University, Makassar, Indonesia, in 2014. Furthermore, he Ph.D. in ICT from Asia E University, Selangor, Malaysia, in 2024. He is a lecturer and researcher at the Department of Informatics, Universitas Muhammadiyah Makassar, Indonesia. His current research concentrations include mobile programming, data mining, decision support systems, expert systems, web applications, and image processing. He can be contacted at email: muh.faisal.art@gmail.com.