Jurnal Informatika dan Rekayasa Perangkat Lunak ISSN 2656-2855
Vol. 8, No. 1, March 2026, Page. 147-155 e-ISSN 2685-5518

Comparative Approaches to Clustering for Profiling Students in Educational Data Mining

Noor Azizah[1*], **Kusworo Adi**[2], **Catur Edi Widodo**[3]
[1] Doctoral Program of Information System, Postgraduate School, Universitas Diponegoro
[2,3] Departement of Physics, Faculty of Science and Mathematics, Universitas Diponegoro
[*]Email: azizah.dsi2024@gmail.com

Abstract

This study aims to compare the performance of five clustering algorithms, a K-Means, K-Medoids, Fuzzy C-Means (FCM), DBSCAN, and Gaussian Mixture Model (GMM) in profiling 239 students using quantitative data. The methodology includes data collection, refinement, transformation, application of clustering algorithms, and evaluation using the Silhouette Score, Davies–Bouldin Index, and execution time. The results indicate that K-Means provides the most balanced performance, achieving the highest Silhouette score with well-defined cluster separation. K-Medoids and GMM demonstrate competitive performance, while DBSCAN excels in detecting outliers but produces an excessive number of clusters, limiting its interpretability for profiling. FCM performs the weakest due to poor cluster separability. Overall, K-Means is recommended as the primary approach for student profiling, while other algorithms may complement specific analytical needs.

Keywords*: Clustering, Educational Data Mining, Student Profiling*

INTRODUCTION

Educational Data Mining (EDM) utilizes clustering for various purposes, such as mapping and grouping student thesis topics to provide a basis for academic guidance (Andre *et al.*, 2023). Clustering also tracks the evolution of students' learning behaviors over time, revealing migration patterns between groups and the dynamics of the learning process in digital environments (Priyambada *et al.*, 2021). The method is also used to generate personalized exercise recommendations in computational or data structure courses (Zhang *et al.*, 2023), and for analyzing student performance and enrollment selection criteria as a foundation for academic decision-making (Fahrudin *et al.*, 2023; Ghosh *et al.*, 2025). These practices follow a similar workflow, including feature engineering, data normalization, empirically determined cluster number selection, internal validation using metrics such as silhouette or Davies-Bouldin, and most importantly, interpreting student profiles to provide actionable insights within pedagogical (Ezugwu *et al.*, 2022; Ghosh *et al.*, 2025; Priyambada *et al.*, 2021; Xu and Tian, 2015; Zhang *et al.*, 2023).

In line with this, the development of clustering algorithms can be mapped into several complementary families: partitioning, hierarchical, density-based, model-based, and fuzzy. Comprehensive surveys position partitioning approaches as efficient and scalable for standardized quantitative data; hierarchical methods offer structure and interpretability via dendrograms; density-based methods effectively reveal arbitrarily shaped clusters and identify outliers; while model-based approaches (e.g., Gaussian mixture models/EM) and fuzzy (partial membership) provide flexibility for overlapping distributions and heterogeneous cluster variances (Ezugwu *et al.*, 2022). Modern practices have added better initialization, stringent validation criteria, more systematic cluster number determination procedures, and mini-batch variants for large-scale data, driving wider adoption in educational domains (Xu and Tian, 2015).

The context of EDM further drives the next evolution. For longitudinal digital learning, cluster evolution analysis maps shifts in student profiles, enabling institutions to not only know "who is in which cluster," but also "when and why migration occurs" insights crucial for adaptive interventions and academic support scheduling (Priyambada *et al.*, 2021). On the other hand, the increasing diversity of data sources (activity logs, grades, social interactions, and content) encourages the adoption of deep multimodal clustering to merge heterogeneous representations and uncover richer latent structures, all without losing the scalability

required for massive learning ecosystems (Raya *et al.*, 2024). At the operational level, hybrid approaches combining clustering with concept drift detection maintain the reliability of student profiles when learning behavior patterns shift such as due to curriculum changes or learning mode alterations ensuring models remain relevant to real-time classroom dynamics(Jain *et al.*, 2022).

Consequently, the selection of clustering algorithms in EDM becomes goal-specific. Partitioning clustering remains ideal for fast segmentation of clean, quantitative data; medoid-based variants are preferred when robustness to outliers and non-Euclidean distance matters; fuzzy clustering is prioritized when learning profiles overlap and partial membership is necessary; while density-based methods are useful for finding minority groups or identifying uncommon patterns that trigger early interventions (Ghosh *et al.*, 2025; Jaiswal, 2025). Practical evidence demonstrates the added value of these techniques: from academic topic mapping and thesis interests (Andre *et al.*, 2023), tracking migration of learning behaviors (Priyambada *et al.*, 2021), to adaptive exercise recommendations (Zhang *et al.*, 2023) all showing that the quality of pedagogical decisions improves when the algorithm selection aligns with data characteristics and intervention goals(Ghosh *et al.*, 2025).

Based on this review, a prominent research gap is the lack of a comprehensive and evidence-based decision framework that links educational data characteristics (e.g., outlier proportions, profile overlap, longitudinal dynamics) with the relative performance of various clustering algorithm families and their impact on pedagogical decisions. While some studies focus on internal metric performance, few systematically test robustness against dynamics (drift), explore ease of interpretation for educators, and assess the direct benefits of clustering on measurable academic interventions (Xu and Tian, 2015).

Therefore, the goal of this research is to develop and evaluate a comparative framework for student profiling based on clustering in real-world EDM contexts: a) modeling and comparing several families of algorithms in data scenarios representing educational conditions (static vs. longitudinal, clean vs. noisy, well-separated vs. overlapping), b) assessing cluster quality with a combination of internal metrics and pedagogical utility indicators (interpretability, stability against drift, and ease of action), and providing an actionable decision map to assist educators and program managers in selecting the appropriate clustering approach aligned with data characteristics and intervention targets in the EDM environment.

LITERATURE REVIEW

2.1 K-Means Clustering

K-Means clustering is a non-hierarchical data clustering technique that groups data into one or more clusters (Jain *et al.*, 2022). In this method, data points with similar characteristics are assigned to the same cluster, while those with differing characteristics are placed into separate clusters. As a result, data within the same cluster exhibit minimal internal variation, reflecting a high degree of homogeneity, whereas data across clusters demonstrate greater dissimilarity.

2.2 K-Medoids Clustering

The K-Medoids algorithm, also known as Partitioning Around the Medoids (PAM), is a variant of the K-Means algorithm. The key difference lies in how the cluster centers are determined. While K-Means identifies the cluster center based on the mean value (centroid), K-Medoids selects the center from actual data points (medoids) that best represent the cluster (Abbas *et al.*, 2020).A medoid can be defined as the object within a cluster that has the lowest average dissimilarity compared to all other objects in the same cluster. The K-Medoids process begins by selecting K initial medoids, after which each data point in the dataset is assigned to the nearest medoid using a chosen distance metric (e.g., Euclidean, Manhattan, or others).

2.3 Fuzzy C-Means

Fuzzy C-Means (FCM) is a clustering algorithm that assigns each data point to multiple clusters with varying degrees of membership. Introduced by Bezdek in 1981, FCM assigns a membership value between 0 and 1 to each data point, reflecting the degree to which a point belongs to a specific cluster. Unlike hard clustering, where data points strictly belong to one cluster, FCM allows partial membership, making it highly useful in scenarios where data points exhibit characteristics of multiple clusters

simultaneously, such as in educational data mining (Bezdek, 1981).

Unlike hard clustering methods like K-Means. It minimizes an objective function that incorporates both the distance between data points and cluster centroids, while considering partial membership for each data point. This flexibility allows FCM to handle data with overlapping characteristics, making it especially useful in Educational Data Mining (EDM), where students may exhibit multiple profiles, such as being academically successful and actively participating in extracurricular activities. The algorithm iteratively updates membership degrees and centroids until convergence, allowing for better cluster definitions in ambiguous datasets. FCM's key advantage lies in its ability to provide fuzzy memberships, which are valuable for personalized learning interventions and profiling students(Zhao *et al.*, 2022).

2.4 DBSCAN Clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that defines clusters based on the density of data points in a region (Ma et al., 2023). Unlike other clustering algorithms, such as K-Means, DBSCAN does not require the number of clusters to be predefined. It uses two parameters: ϵ(epsilon) and MinPts (minimum points), which together define the density of a cluster. The basic idea behind DBSCAN is to classify data points into three categories: 1) Core points: These are points that have at least $MinPts$ points within a radius of ϵ; 2) Border points: Points that are not core points but are within the ϵ-neighborhood of a core point. 3) Noise points: Points that are neither core points nor border points, essentially outliers.

DBSCAN starts with an arbitrary point and expands the cluster by recursively including all points that are density-reachable from the initial point. A point is considered density-reachable if it is within the neighborhood of a core point and can be reached by other points in the cluster. This method allows DBSCAN to discover clusters of arbitrary shapes and is particularly effective in identifying and handling noise or outliers within the dataset. These advancements reinforce DBSCAN's applicability across domains requiring robust cluster detection while mitigating traditional limitations (Huang et al., 2025).

2.5 Gaussian Mixture Model Clustering

Gaussian Mixture Model (GMM) Clustering is a probabilistic clustering algorithm that assumes data is generated from a mixture of Gaussian distributions. Unlike hard clustering methods like K-Means, where data points are strictly assigned to one cluster, GMM assigns probabilities for each data point to belong to multiple clusters, allowing for soft assignments. GMM uses the Expectation-Maximization (EM) algorithm, which alternates between two steps: the Expectation (E-step), where the probability of a data point belonging to each cluster is computed, and the Maximization (M-step), where the parameters (mean, covariance, and mixing coefficient) of the Gaussian components are updated to maximize the likelihood of the data (Kasa and Rajan, 2023).

GMM is ideal for data with overlapping clusters or complex shapes. It is widely used in areas such as image segmentation and student profiling in educational data mining. The algorithm's flexibility allows it to model clusters with different shapes and densities. However, choosing the correct number of clusters remains a challenge.

2.6 Silhouette Score

The Silhouette Score evaluates the quality of clustering results by simultaneously considering both intra-cluster cohesion and inter-cluster separation. For each point i, $a(i)$ is calculated as the average distance to other points within the same cluster (cohesion), and $b(i)$ as the average minimum distance to the nearest cluster (separation) (Yang *et al.*, 2024).

$$s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \quad (1)$$

The silhouette coefficient is defined as a value in the range [-1,1]. A score close to 1 indicates compact and well-separated clusters, around 0 suggests overlap between clusters, and negative values indicate poor assignment. The average silhouette score $s(i)$ is used to select the optimal number of clusters or to compare different algorithms. In practice, a score greater than 0.5 is often considered good, though it depends on the distance metric and feature scaling; thus,

preprocessing (such as standardization) and consistency of the metric are essential. The silhouette score is particularly effective for convex cluster structures but less stable for highly variable densities. Therefore, its interpretation should be complemented with visualizations (e.g., silhouette plots) and other metrics.

2.7 Davies Bouldin Score

Davies-Bouldin Index (DBI) is a metric used to evaluate the quality of clustering results by measuring both the compactness and separation of clusters (Hosen *et al.*, 2023). A lower DBI value indicates better clustering. It is calculated by evaluating the average similarity between each cluster and its most similar cluster, based on the ratio of intra-cluster distances to inter-cluster distances. The formula is:

$$DBI = \frac{1}{N} \sum_{i=1}^{N} \underset{i \neq j}{MAX} \left(\frac{s_i + s_j}{d(c_i, c_j)} \right) \quad (2)$$

Where:

- S_i is the average distance between each point in cluster iii and its centroid (intra-cluster scatter).
- $d(C_i, C_j)$ is the distance between the centroids of clusters iii and jjj (inter-cluster separation).
- N is the total number of clusters.

DBI is particularly useful for comparing clustering algorithms and determining the optimal number of clusters. A lower DBI indicates clusters that are more compact and well-separated. It is often used alongside other metrics, such as the Silhouette Score, to evaluate clustering quality.

METHOD

The research employs quantitative analysis to cluster and profile 239 students (referred to as "santri") based on their academic performance and extracurricular involvement. The methodology used in this study follows a systematic process, as defined in Figure 1.

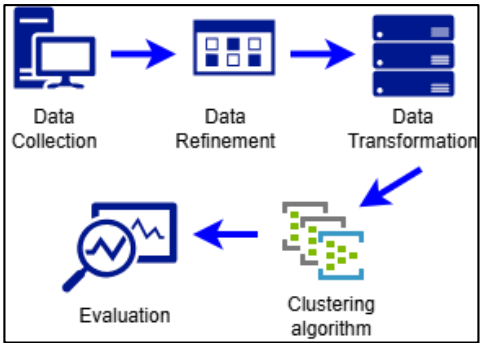

Figure 1. Research stages

The explanation of each stage of the research is as follows

1. Data Collection
   Data was gathered from the students, including academic records (such as exam scores), attendance, and participation in extracurricular activities. The dataset represents a broad spectrum of student behavior and performance indicators, which are critical for profiling. This collection process ensures that we have a rich, multidimensional view of student characteristics, enabling meaningful clustering. Data collection is the foundation of any EDM analysis, and its quality directly impacts the accuracy and validity of the findings.
2. Data Refinement
   Once the data is collected, data refinement is essential to ensure the dataset's integrity and reliability. This involves removing or correcting any missing, inconsistent, or outlier values to maintain accuracy and prevent bias in the clustering process. Outliers are carefully analyzed and removed when necessary, as they could distort the results. Additionally, numerical data is standardized to ensure that all features are on the same scale, which is crucial for algorithms like K-Means that rely on distance-based measures. This normalization ensures that no single feature dominates the clustering process. Furthermore, only the most relevant features, which directly contribute to profiling the students, are selected for clustering. This feature selection reduces noise and enhances the clustering algorithm's ability to identify meaningful patterns, ultimately improving the overall clustering accuracy.
3. Data Transformation

Following refinement, the data transformation stage ensures the data is in a format suitable for clustering algorithms. Categorical data, such as student groups or extracurricular activities, is converted into numerical values using encoding methods like one-hot encoding. In cases where the dataset contains many features, dimensionality reduction techniques, such as Principal Component Analysis (PCA), are applied to reduce the number of features and focus on the most relevant ones. This helps improve the clustering efficiency by eliminating redundant information and highlighting the key attributes necessary for meaningful segmentation.

4. Clustering Algorithm
   The core of the study is the application of various clustering algorithms to group the students based on their features. The algorithms selected for this study include:
   a) *K-Means Clustering*: This is a partitioning-based algorithm that divides the dataset into a predefined number of clusters (K). It minimizes the variance within clusters by assigning each data point to the nearest centroid.
   b) *K-Medoids Clustering:* Similar to K-Means, but instead of using the mean of the points, it uses actual data points (medoids) as the cluster center, making it more robust to outliers.
   c) *Fuzzy C-Means* (FCM): This algorithm allows for soft clustering, where each data point can belong to multiple clusters with varying degrees of membership. FCM is particularly useful when the boundaries between clusters are not well-defined, such as in student profiling, where students might fit into multiple categories (e.g., both academic achievers and active participants).
   d) DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*): This density-based algorithm groups points that are close to each other, while identifying outliers as noise. DBSCAN does not require the number of clusters to be predefined, which makes it useful for datasets with arbitrary cluster shapes.
5. Evaluation
   After clustering, the evaluation phase is crucial to assess the quality of the clusters formed by each algorithm. To evaluate the clustering performance, two key metrics are used: the Silhouette Score and the Davies-Bouldin Index (DBI). The Silhouette Score measures how similar each data point is to its own cluster compared to other clusters. A high Silhouette Score indicates that the clusters are well-separated and cohesive, with each data point closely matching its own cluster while being far from other clusters. On the other hand, the Davies-Bouldin Index (DBI) calculates the average similarity between each cluster and the most similar cluster. A lower DBI value signifies better clustering performance, where the clusters are compact (with low internal variance) and well-separated from each other. These metrics allow for a comprehensive assessment of the clustering results, providing insights into the effectiveness and quality of the clustering algorithms used.Data yang digunakan dalam penelitian ini adalah data santri sebanyak 239 siswa.

RESULT AND DISCUSSION

4.1 Clustering Result

The clustering process was carried out using five algorithms : K-Means, K-Medoids, Fuzzy C-Means (FCM), DBSCAN, and Gaussian Mixture Model (GMM), each producing distinct segmentation patterns in the student dataset. The results demonstrate how each algorithm responds to the structure, density, and variability of the data.

K-Means with K=3 generated three clusters with clear separation when visualized using two-dimensional PCA. The plot shows two dense, compact groups and one small, distant cluster indicating a minority group of students whose characteristics differ markedly from the majority. This visual structure aligns with the high Silhouette Score (0.339), suggesting strong cohesion within clusters and good separation between them. The compactness and clarity of these boundaries reflect K-Means' assumption of spherical clusters, which fits well with the underlying distribution of the data.

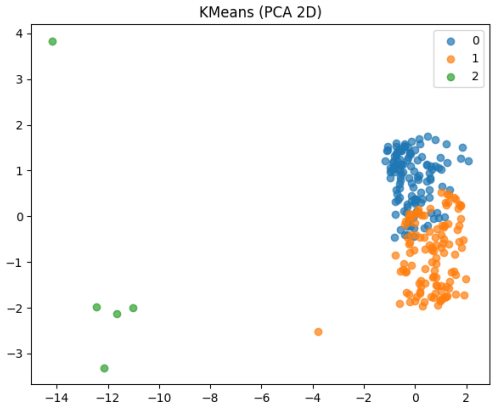

Figure 2. Visualization of K-Means scatter plot

K-Medoids (PAM) produced a clustering structure similar to K-Means but demonstrated tighter handling of outliers due to its medoid-based centroid selection. While the numerical results (Silhouette = 0.331; DBI = 0.915) indicate comparable performance to K-Means, K-Medoids formed slightly more stable clusters around data points with high local density. This reflects its robustness to extreme values, making it suitable when data variability is high.

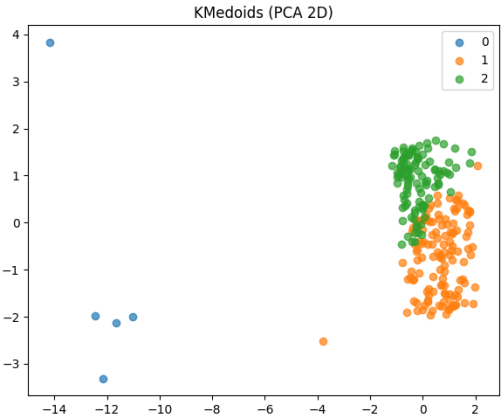

Figure 3. Visualization of K-Medoids scatter plot

FCM with c=3 resulted in clusters with considerable overlap, as reflected in its low Silhouette Score (0.201). The fuzziness parameter m permitted partial membership, but in this dataset, the resulting clusters lacked clear boundaries. Students were distributed with overlapping membership values across clusters, indicating that FCM struggled to differentiate meaningful student profiles due to the low separability of feature space.

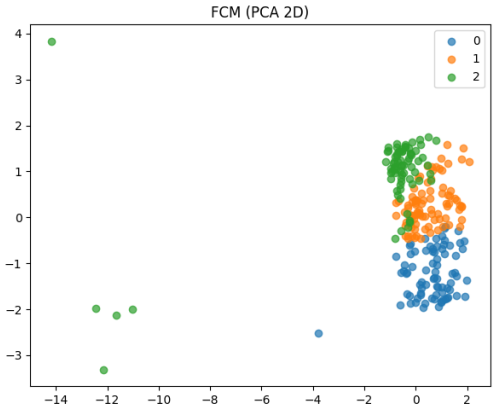

Figure 4. Visualization of FCM scatter plot

DBSCAN identified eight clusters, far more than the expected three, due to its sensitivity to local density variations. This caused several micro-clusters and noise points, consistent with its low Silhouette (0.311) but strong DBI (0.835). The algorithm successfully detected outliers but fractured the main data body into smaller subgroups, reducing interpretability for profiling purposes.

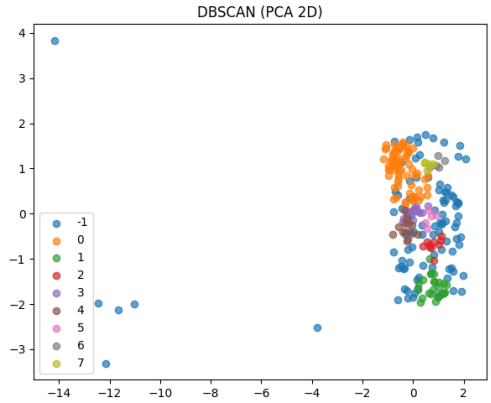

Figure 5. Visualization of DBSCAN scatter plot

GMM with three components generated Gaussian-shaped clusters that approximately matched the main structures observed in K-Means. With a Silhouette Score of 0.335, GMM captured the elliptical spread of clusters slightly better than centroid-only methods. However, its DBI (0.946) suggests looser separation among components, attributable to probabilistic overlaps inherent in mixture models.

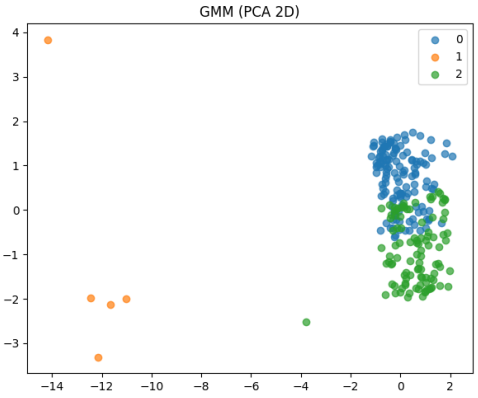

Figure 6. Visualization of GMM scatter plot

4.2 Performance Comparison and Evaluation

To provide a comprehensive comparison of the five clustering algorithms evaluated in this study, Table 1 summarizes their performance in terms of the number of clusters detected, Silhouette Score, Davies–Bouldin Index, and execution time.

Table 1. Performance comparasion of method

Algorithm	Number of Cluster	Silhoutte Score	DBI	Time
K-Means	3	0.339	0.923	0.017
K-Medoids	3	0.331	0.915	0.022
FCM	3	0.201	1.456	0.005
DBSCAN	8	0.311	0.835	0.003
GMM	3	0.335	0.945	0.011

K-Means offers the most balanced performance on this dataset, achieving the highest Silhouette (0.339264) with a competitive DBI (0.923717) at low runtime, yielding cohesive, well-separated clusters that are easy to interpret for profiling. K-Medoids is a close second: its slightly lower Silhouette (0.331206) is offset by the best DBI among the three-cluster methods (0.915448), consistent with stronger compactness from medoid centers, at a modest computational cost. GMM nearly matches K-Means on Silhouette (0.335414) with a somewhat higher DBI (0.945796) and efficient runtime, indicating flexible, anisotropic clusters that capture variance well but with looser separation.

DBSCAN attains the lowest DBI overall (0.835766) and the fastest runtime, yet its lower Silhouette (0.311005) and eight detected clusters suggest over-fragmentation driven by density sensitivity useful for anomaly discovery, less so for canonical profiling. FCM shows the weakest structure (Silhouette 0.201471; DBI 1.456483), reflecting substantial overlap. Practically, K-Means is the recommended default for actionable student segmentation; K-Medoids is preferred when outlier robustness matters; GMM suits scenarios needing probabilistic/elliptical clusters; DBSCAN is best reserved for irregular densities and outlier detection; FCM is less suitable here where crisp, interpretable clusters are required.

CONCLUSION

This study compared five clustering approaches for profiling 239 students and found that K-Means provides the best overall trade-off between cohesion, separation, and interpretability (highest Silhouette, competitive DBI, low runtime), with K-Medoids and GMM as close alternatives—respectively stronger to outliers and more flexible through probabilistic, elliptical components. DBSCAN achieved the lowest DBI and fastest runtime but fragmented the cohort into many small groups, suiting anomaly detection rather than canonical profiles, while FCM yielded the weakest separation for this dataset. Practically, K-Means is a sensible default for actionable segmentation; K-Medoids/GMM are situationally preferable, and DBSCAN is complementary for irregular densities and outliers. Future work will extend validation beyond internal indices by linking clusters to academic outcomes (retention, grades) and educator actions; test robustness under longitudinal drift; integrate multimodal features (logs, text) and deep/semi-supervised clustering; automate model selection and stability analysis; and incorporate explainability and fairness checks to support trustworthy, data-informed interventions.

ACKNOWLEDGMENT

The authors would like to express their gratitude to Beasiswa Pendidikan Indonesia (BPI) or Indonesia Education Scholarship, Pusat Pembiayaan dan Asesmen Pendidikan Tinggi (PPAPT) or Center for Higher Education Funding and Assessment, under The Ministry of

Higher Education, Science and Technology (Kemendikti Saintek) and Lembaga Pengelola Dana Pendidikan (LPDP) or Indonesian Endowment Fund for Education for funding their doctorate degree.

REFERENCES

Abbas, S.A., Aslam, A., Rehman, A.U., Abbasi, W.A., Arif, S. and Kazmi, S.Z.H. (2020), "K-Means and K-Medoids: Cluster Analysis on Birth Data Collected in City Muzaffarabad, Kashmir", *IEEE Access*, Vol. 8, pp. 151847–151855, doi: 10.1109/ACCESS.2020.3014021.

Andre, Suciati, N., Fabroyir, H. and Pardede, E. (2023), "Educational Data Mining Clustering Approach: Case Study of Undergraduate Student Thesis Topic", *IEEE Access*, IEEE, Vol. 11 No. September, pp. 130072–130088, doi: 10.1109/ACCESS.2023.3332818.

Bezdek, J.C. (1981), "Objective Function Clustering BT - Pattern Recognition with Fuzzy Objective Function Algorithms", in Bezdek, J.C. (Ed.), , Springer US, Boston, MA, pp. 43–93, doi: 10.1007/978-1-4757-0450-1_3.

Ezugwu, A.E., Ikotun, A.M., Oyelade, O.O., Abualigah, L., Agushaka, J.O., Eke, C.I. and Akinyelu, A.A. (2022), "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects", *Engineering Applications of Artificial Intelligence*, Elsevier Ltd, Vol. 110 No. January, p. 104743, doi: 10.1016/j.engappai.2022.104743.

Fahrudin, T., Asror, I. and Wibowo, Y.F.A. (2023), "Student Enrollment Performance of Telkom Schools in 23 / 24 schoolyear using k-Means Clustering", *International Conference on Informatics and Computing (ICIC)*, IEEE, pp. 1–5, doi: 10.1109/ICIC60109.2023.10381939.

Ghosh, A., Sengupta, P., Chaudhuri, A.K., Mukherjee, D., Das, A.K. and De, P. (2025), "Analyzing Student Achievement with Clustering Techniques in Educational Analytics", *Panamerican Mathematical Journal*, Vol. 35 No. 1, pp. 424–436, doi: https://doi.org/10.52783/pmj.v35.i4s.6014 .

Hosen, M.A., Moz, S.H., Kabir, S.S., Galib, S.M. and Adnan, M.N. (2023), "Enhancing Thyroid Patient Dietary Management with an Optimized Recommender System based on PSO and K-means", *Procedia Computer Science*, Elsevier B.V., Vol. 230 No. 2023, pp. 688–697, doi: 10.1016/j.procs.2023.12.124.

Huang, Z., Liang, Z., Zhou, S. and Zhang, S. (2025), "An Improved Density-Based Spatial Clustering of Applications with Noise Algorithm with an Adaptive Parameter Based on the Sparrow Search Algorithm", *Algorithms*.

Jain, M., Kaur, G. and Saxena, V. (2022), "A K-Means clustering and SVM based hybrid concept drift detection technique for network anomaly detection", *Expert Systems with Applications*, Elsevier Ltd, Vol. 193 No. June 2020, p. 116510, doi: 10.1016/j.eswa.2022.116510.

Jaiswal, S. (2025), "Clustering Students Based on Learning Styles : A Machine Learning Approach for Personalized Education", *2025 World Skills Conference on Universal Data Analytics and Sciences (WorldSUAS)*, India, doi: 10.1109/WorldSUAS66815.2025.11198965.

Kasa, S.R. and Rajan, V. (2023), "Avoiding inferior clusterings with misspecified Gaussian mixture models", *Scientific Reports*, Nature Publishing Group UK, pp. 1–13, doi: 10.1038/s41598-023-44608-3.

Ma, B., Yang, C., Li, A., Chi, Y. and Chen, L. (2023), "ScienceDirect ScienceDirect 10th International Conference on Information Technology and Quantitative Management A Faster DBSCAN Algorithm Based on Self-Adaptive A Faster DBSCAN Algorithm Based on Self-Adaptive Determination of Parameters Determination of Parameters", *Procedia Computer Science*, Elsevier B.V., Vol. 221, pp. 113–120, doi: 10.1016/j.procs.2023.07.017.

Priyambada, S.A., Er, M., Yahya, B.N. and Usagawa, T. (2021), "Profile-Based Cluster Evolution Analysis : Identification of Migration Patterns for Understanding Student Learning Behavior", *IEEE Access*, IEEE, Vol. 9, pp. 101718–101728, doi: 10.1109/ACCESS.2021.3095958.

Raya, S., Orabi, M., Afyouni, I. and Aghbari, Z. Al. (2024), "Neurocomputing Multi-modal data clustering using deep learning : A systematic review", *Neurocomputing*, Elsevier B.V., Vol. 607 No. September 2022, p. 128348, doi: 10.1016/j.neucom.2024.128348.

Xu, D. and Tian, Y. (2015), "A Comprehensive Survey of Clustering Algorithms", *Annals of Data Science*, Springer Berlin Heidelberg, Vol. 2 No. 2, pp. 165–193, doi: 10.1007/s40745-015-0040-1.

Yang, D., Wang, J., He, J. and Zhao, C. (2024), "A clustering mining method for sports behavior characteristics of athletes based on the ant colony optimization", *Heliyon*, Elsevier Ltd, Vol. 10 No. 12, p. e33297, doi: 10.1016/j.heliyon.2024.e33297.

Zhang, Z., Liu, H. and Wu, Z. (2023), "Student Profile Clustering Based Personalized Exercise Recommendation: Taking Data Structures Course as an Example", *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, IEEE, pp. 68–70, doi: 10.1109/ICALT58122.2023.00026.

Zhao, K., Dai, Y., Jia, Z. and Ji, Y. (2022), "General Fuzzy C-Means Clustering Strategy : Using Objective Function to Control Fuzziness of", *IEEE Transaction on Fuzzy System*, Vol. 30 No. 9, pp. 3601–3616, doi: 10.1109/TFUZZ.2021.3119240.