Matrik: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer
Vol. 25, No. 1, November 2025, pp. 217∼230
ISSN: 2476-9843, accredited by Kemenristekdikti, Decree No: 10/C/C3/DT.05.00/2025
DOI: 10.30812/matrik.v25i1.5843

Comparing K-Means, GMM, and BIRCH for Student Academic Performance Data: Evaluation on Two Public Datasets

Ricky Aurelius Nurtanto Diaz[1], Ni Luh Gede Pivin Suwirmayanti[1], Emy Setyaningsih[2], I Wayan Budi Sentana[3]
[1] Institut Teknologi dan Bisnis STIKOM Bali, Bali, Indonesia
[2] Universitas AKPRIND, Yogyakarta, Indonesia
[3] Politeknik Negeri Bali, Bali,Indonesia

Article Info

Article history:

Received October 19, 2025
Revised November 26, 2025
Accepted November 29, 2025

Keywords:

Birch;
Clustering;
Comparison;
GMM;
K-Means;
Student academic.

ABSTRACT

Academic data contains complex patterns that require appropriate clustering approaches to support informed educational decision-making. However, comparative studies that regularly evaluate various clustering methods for student academic performance, using diverse public data sets and consistent evaluation criteria, are limited. This study aims to identify the most effective clustering algorithm for modeling student academic performance by comparing three techniques: K-Means, GMM, and BIRCH, on two publicly available datasets: the Student Performance Metrics (SPM) Dataset with 16 features and 493 instances, and the Higher Education Students Performance Evaluation (HESPE) dataset with 32 features and 145 instances. Algorithm evaluation was performed using Sum of Squared Errors (SSE), Davies–Bouldin Index (DBI), Silhouette Score, and computational time. The results show that K-Means consistently delivers superior clustering quality across both datasets, outperforming the other algorithms on four evaluation criteria. At the same time, BIRCH demonstrates superiority in two metrics and achieves the shortest computational time. These findings highlight that clustering effectiveness is strongly influenced by algorithm characteristics and data structure, with K-Means being more suitable for accuracy-oriented clustering and BIRCH for time-critical applications. Overall, this study contributes to educational data mining by providing comparative evidence on algorithm performance and demonstrating how methodological choices influence the interpretation of student performance patterns. In practice, institutions can choose clustering methods that best suit their needs, such as K-Means for precise academic profiling or BIRCH for rapid, large-scale analysis, to help students graduate successfully.

Corresponding Author:

Ricky Aurelius Nurtanto Diaz, +6285858653508
Department of Computer Systems, Faculty of Informatics and Computer,
Institut Teknologi dan Bisnis STIKOM Bali, Bali, Indonesia,
Email: ricky@stikom-bali.ac.id

How to Cite:
R. A. N. Diaz, N. L. G. P. Suwirmayanti, E. Setyaningsih, and I. W. B. . Sentana, ”Comparing K-Means, GMM, and BIRCH for Student Academic Performance Data: Evaluation on Two Public Datasets”, *MATRIK: Jurnal Manajemen, Teknik Informatika, dan Rekayasa Komputer*, Vol. 25, No. 1, pp. 217-230, November, 2025.

1. INTRODUCTION

Student academic performance is one of the main indicators in assessing the quality of education [1, 2]. To enhance the quality of education, it is essential to gain a deeper understanding of the factors that influence student's academic performance [3]. Student academic data, which includes various variables such as attendance, study habits, and test results, can provide valuable insights for teaching and learning improvement [2]. However, the analysis of academic data often requires an appropriate approach to uncover the hidden patterns in the data. In the Big Data Era, data mining approaches are increasingly being used to address these challenges [4]. The application of data mining techniques in education allows the identification of hidden patterns in data that cannot be found with traditional approaches [5].

One of the methods widely applied in academic data analysis is clustering, which is used to group students based on the similarities of their characteristics [6]. Previous research leverages an Educational Data Mining (EDM) approach to extract insights from large-scale institutional data encompassing student demographics, academic records, and activity logs. By applying clustering techniques, the study identifies groups of student performance differentiated by their characteristics and school achievements [7]. In addition, research using data from 117,069 undergraduates across 20 public HEIs and three clustering algorithms (k-means, BIRCH, DBSCAN): the optimized k-means model (KMoB) performed best, yielding five performance clusters primarily shaped by CGPA, activity count, employment status, and dropout status, enabling early identification and targeted interventions for B40 students [8]. This is important in the context of academic clustering, which requires an accurate evaluation of techniques that can effectively group students.

Three cluster methods used in this study were deliberately chosen as clustering techniques that represent the three main families of clustering, namely K-Means as a centroid-based partition method [9], Gaussian Mixture Models(GMM) as a distribution modeling (probabilistic) method that models data as a Gaussian mixture with soft assignments [10, 11] and BIRCH as a hierarchical-incremental approach based on cluster feature (CF) trees that perform a gradual hierarchical construction for very large data through a summary of local statistics (CF) before the global clustering stage [12–15]. In the realm of retail customer segmentation, existing research shows that GMM tends to excel when the distribution of existing clusters overlaps or is known as elliptical, while K-Means remains an efficient method, and the BIRCH method is effective when data size and incremental processing needs are the main factors of need [16, 17]. A study using GMM, K-Means Algorithm, and hierarchical methods to group students based on quizzes, forums, and CGPA data, showed results that both K-Means and GMM produced similar Silhouette scores, which shows the effectiveness of these two algorithms in reflecting consistent learning patterns [18]. K-Means, with its simplicity and speed of computing, is ideal for identifying groups [2], GMM provides more flexibility through soft assignments, useful when the boundaries between clusters are not sharp, or when there is overlap in performance. Hierarchical approaches such as BIRCH, currently not widely used in education data clustering, still offer efficiency in large data or incremental processes without losing cluster quality.

Previous research using student performance analysis datasets has focused on supervised classification, particularly ensemble-based algorithms. However, the exploration of unsupervised clustering approaches for the dataset is still very limited, conducted with only one algorithm (e.g., the centroid-based algorithm), and has not shown how clustering results compare with other clustering techniques. Similarly, the Student Performance Metrics Dataset, which is relatively new, is still more widely used in initial studies and experimental projects, without a systematic study of the effectiveness of clustering algorithms in uncovering latent patterns in student performance data.

To address this gap, this study presents a comparative analysis of three clustering techniques, namely K-Means (centroid-based), GMM (probabilistic-based), and BIRCH (hierarchical incremental-based), applied to the Student Performance Metrics Dataset and the Higher Education Students Performance Evaluation. The contribution of this study is to provide a unified comparative evaluation of K-Means, GMM, and BIRCH on two public academic datasets using consistent clustering metrics (SSE, DBI, Silhouette Score, and runtime).

2. RESEARCH METHOD

In this study, we used two public datasets, namely the Student Performance Metric Dataset and the Higher Education Students Performance Evaluation. Furthermore, these two public datasets will be processed using three different clustering methods, namely K-Means, Gaussian Mixture Model, and BIRCH. Each method will be used to process the two public datasets, and the clustering results of each method will be assessed using three clustering test techniques, namely SSE, DBI, and Silhouette Score. Figure 1 shows the research method we used in this study.

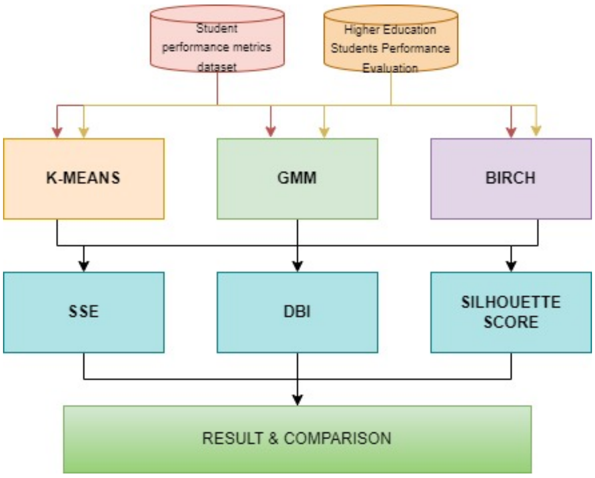

Student performance metrics dataset
Higher Education Students Performance Evaluation
K-MEANS
GMM
BIRCH
SSE
DBI
SILHOUETTE SCORE
RESULT & COMPARISON

Figure 1. Research method

Figure 1 shows, student performance in higher education is evaluated using a clustering technique to group students based on their performance metrics. The datasets used in this study included various student performance metrics, which included factors such as test scores, attendance, and time spent studying. The purpose of this study is to determine which clustering algorithm is most effective in grouping students with similar performance characteristics. There are three clustering algorithms, namely K-Means, Gaussian Mixture Model (GMM), and BIRCH, used to process the two existing datasets. Each of these algorithms was evaluated using three different performance metrics, namely Sum of Squared Errors (SSE), Davies-Bouldin Index (DBI), and Silhouette Score. SSE measures cluster density by calculating the square distance between the data point and its cluster center, where a lower value indicates better clustering. The DBI evaluates the separation between the cluster and the internal cohesion of the cluster, with lower values indicating better cluster separation. The Silhouette Score measures the extent to which each data point fits into its cluster compared to other clusters, with higher scores indicating a clearer, more separate cluster. The results of this evaluation are then compared to determine which clustering algorithm provides the best clustering by looking at the computational time required.

2.1. Experimental Environment

System resources may influence algorithm performance and reproducibility. The experiments in this study were executed within a controlled local computing environment to ensure consistency and reliability throughout the clustering process. All computations were performed on a machine equipped with the hardware and software configurations listed in Table 1. These specifications include details on the processor, memory capacity, storage type, operating system, and the software libraries utilized during the analysis. The following table describes the resources used in this research

Table 1. Hardware and Software Specifications

Component	Specification
CPU	Intel Core i5 @ 2.40GHz
RAM	8 GB DDR4
Storage	512 GB SSD
Operating System	Windows 11 Home 64-bit
Software	Python 3.9.13, scikit-learn 1.6.1, NumPy 2.0.2, Pandas 2.2.3

2.2. Data Preprocessing and Parameter Setting

To process the dataset, three clustering techniques for numerical data will be applied. Therefore, the preprocessing process will utilize the Standard Scaler technique to ensure each feature has a comparable scale, allowing the clustering algorithm to perform optimally and avoid bias toward variables with a wider range of values. For each clustering algorithm, several parameters will be used, as can be seen in the following table.

Table 2. Parameters Setting

Algorithm	Parameters
K-Means	n_clusters=3, n_init=10, random_state=42.
GMM	n_components=3, covariance_type='full', random_state=42.
BIRCH	n_clusters=3, threshold=0.5, branching_factor=50.

Table 2 shows the parameter configurations used in this study. K-Means was configured with the default values of 10 initializations (n_init = 10) and a fixed random seed (random_state = 42). The Gaussian Mixture Model (GMM) also used default parameter settings with a full covariance structure, allowing for flexible cluster shapes. For BIRCH, we used a default threshold of 0.5, which sets the maximum radius for data insertion into feature cluster nodes. In BIRCH, a lower threshold value maintains tighter cluster boundaries, whereas a larger value results in looser, less compact subclusters, potentially reducing clustering precision. In this study, the number of clusters was set at three (k = 3) to reflect the most common and pedagogically meaningful grouping structure found in recent educational data mining research. Some previous studies have used three performance groups, typically representing high, medium, and low achieving students because these categories align with academic evaluation practices and allow institutions to implement effectively targeted interventions [19].

2.3. Clustering Method

K-Means: Because of its ease of use and effectiveness, K-Means is among the most widely used non-hierarchical clustering algorithms. K-Means is one of the most widely used non-hierarchical clustering algorithms due to its simplicity, speed, and effectiveness in dividing data into clusters. This algorithm works by minimizing the squared distance between each data point and the cluster centroid, resulting in clusters with the lowest possible internal variability. The iterative process involves assigning data to the nearest centroid and updating the centroid positions until convergence is achieved [20]. Gaussian Mixture Model (GMM): A probabilistic clustering technique called the Gaussian Mixture Model (GMM) represents data distributions as mixtures of Gaussian distributions. The Gaussian Mixture Model is a probabilistic clustering method that represents data distribution as a combination of several Gaussian distributions. Unlike K-Means, which performs hard clustering, GMM allows for soft clustering, where each data point has a probability of belonging to more than one cluster. GMM parameter estimation, including the mean, covariance, and weight of each Gaussian component, is generally performed using the Expectation-Maximization (EM) algorithm, which optimizes the likelihood value of the data. GMM's flexibility makes it superior for data with elliptical cluster shapes and varying variances, although this algorithm is also susceptible to overfitting if the number of components is not carefully determined [21]. BIRCH is an efficient hierarchical clustering algorithm for large datasets. This algorithm uses a special tree structure called the Clustering Feature (CF) Tree, which stores summary statistics such as the number of elements, the number of vectors, and the sum of squares of vectors at each node, allowing for efficient data compression without explicitly storing all data points. As new data points arrive, the algorithm gradually assigns them to the most appropriate leaf nodes; if the capacity limit is exceeded, the node is split to maintain the balance of the tree structure. Once the CF-Tree is formed, further clustering is performed using a partitioning method to produce more structured final clusters [22].

2.4. Evaluation Method

Sum of Squared Errors (SSE): The Sum of Squared Errors (SSE) is used as a fitness function or cost function that is minimized by an evolutionary algorithm (in this case, IWO). This allows the system to evaluate the quality of the solution (centroid/cluster position) in each individual generation, selecting the best solution based on the lowest SSE value, achieving convergence towards the optimal clustering solution. The Sum of Squared Errors (SSE) metric is used as the main evaluation function to assess the quality of data clustering results. SSE is used because it directly reflects the level of compactness of a cluster, which is one of the main objectives of clustering, namely grouping similar data into one cluster and separating dissimilar data into other clusters. [20, 23]. Davies-Bouldin Index (DBI): The Davies-Bouldin Index (DBI) is an internal evaluation metric used to assess the quality of clustering results by considering the compactness of a cluster and the distance between clusters. The DBI value is calculated by comparing the average dispersion within a cluster with the distance between centroids, so a smaller DBI value indicates better cluster separation and a more consistent cluster structure. This metric is popular because it is simple, based on cluster geometry calculations, and can be used in various algorithms without requiring data labels [24, 25]. Silhouette Score: The Silhouette Score is an internal evaluation index that measures clustering quality by combining two important aspects: cohesion (the average proximity of a point to its own cluster members) and separation (the average distance of the point to the nearest other cluster members). The Silhouette Score ranges from -1 to +1; values close to +1 indicate excellent cluster separation, values close to 0 indicate that a point is on the boundary between clusters, and negative values indicate that the point is likely misplaced. The advantage of the Silhouette Score is its ability to provide intuitive interpretations both at the cluster level and across the entire dataset [26].

3. RESULT AND ANALYSIS

In this section, we present the results and analysis obtained from the experiments conducted in this study. The discussion begins with a description of the datasets used, including their sources and features. This is followed by a detailed presentation of the clustering outcomes for each method, evaluated using three validation metrics: Sum of Squared Errors (SSE), Davies–Bouldin Index (DBI), and Silhouette Score. Each metric is reported separately to highlight the performance of the clustering techniques across the the two datasets. Finally, an integrated analysis is provided to interpret the comparative results across all evaluation metrics and an ANOVA test result.

3.1. Dataset

In this study, we used two public datasets, namely the Student Performance Metrics (SPM) Dataset and the Higher Education Students Performance Evaluation (HESPE) which were obtained from two public dataset sources, namely UCI and Mendeley dataset. The selection of these two public datasets is seen from the form of the existing dataset, which is based on the comparison of the number of features, as well as the form of data used, such as categorical and numeric forms. Table 3 shows the source of the dataset that is shown through the source information of the dataset, the number of features and the amount of data for each dataset.

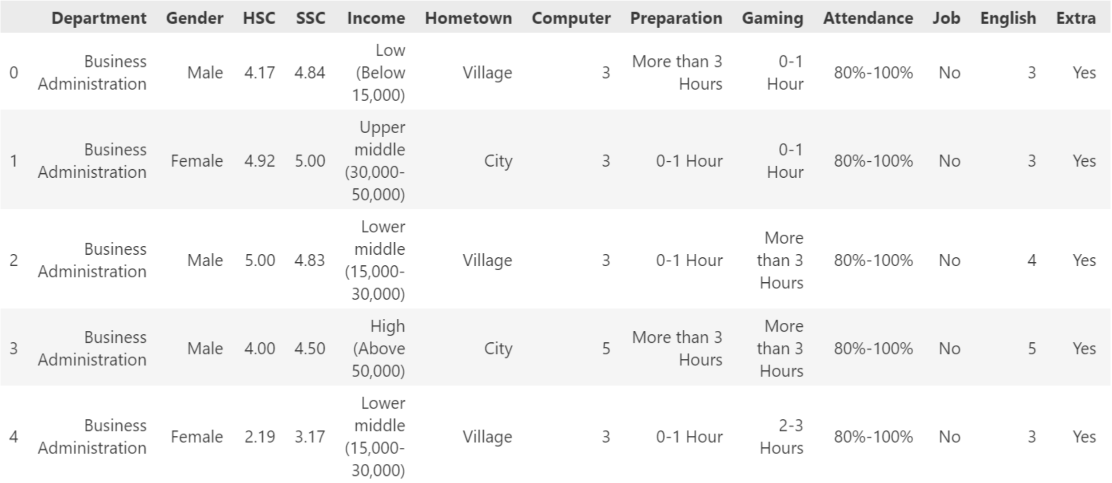
Table 3. Dataset Information

Dataset Name	Source	Number of instances	Number of features
Student Performance Metrics Dataset	https://data.mendeley.com/datasets/5b82ytz489/1	493	16
Higher Education Students Performance Evaluation	https://archive.ics.uci.edu/dataset/856/higher+ https://education+students+performance+evaluation	145	32

The student metric dataset has 16 features, consisting of 10 features in the form of categorical data and 6 features in the form of numerical data. The following in Table 4 is the detailed information from the Student metric dataset, as well as Figure 2, which shows the appearance of the dataset before it is processed using the clustering method.

Table 4. Student Metrics Dataset

No	Feature	Data Type	Description
1	Academic Department	Categorical	Refers to the student's enrolled major, for instance: Business Administration, Computer Engineering, Economics, Electrical and Electronics, English Literature, Journalism and Media Communication, Law and Human Rights, Political Science, or Public Health.
2	Gender	Categorical	A student's gender identity is categorized as Male or Female.
3	High School Certificate (HSC)	Numeric	Numerical value representing the student's Higher Secondary exam performance.
4	Secondary School Certificate (SSC)	Numeric	A numeric indicator of student achievement at the secondary education level.
5	Family Income Level	Categorical	Economic class grouping: 1 = Low income (below 15,000), 2 = Lower middle (15,000-30,000), 3 = Middle (30,000-50,000), 4 = Upper middle (50,000 and above).
6	Hometown Type	Categorical	Indicates where the student resides-either in a rural/village area or an urban/city location.
7	Computer Skills	Numeric	Quantitative score showing the student's computer proficiency.
8	Study Preparation Time	Categorical	Average daily preparation time before classes:1=Less than 1 hour;2=Between 1–3 hours;3=More than 3 hours.
9	Gaming Frequency	Categorical	Represents how often the student plays digital games: 1=Less than 1 hour;2=Between 1–3 hours;3=More than 3 hours.
10	Class Attendance	Categorical	Attendance percentage group: 1 = 80-100%, 2 = 60-79%, 3 = Below 60%.
11	Part-Time Job	Categorical	Employment status of the student: 1 = Employed, 2 = Not employed.
12	English Proficiency	Numeric	Score or numerical measure indicating English language skill.
13	Extra Activities	Categorical	Indicates participation in additional activities such as clubs, art, or sports -1 = Yes, 2 = No.
14	Current Semester	Categorical	Denotes which semester the student is currently attending (e.g., 2nd, 3rd, 4th, . . . up to 12th).
15	Previous Semester GPA	Numeric	Numerical record of grade point average (GPA) from the previous term.
16	Cumulative GPA	Numeric	Overall GPA calculated across all completed semesters.



	Department	Gender	HSC	SSC	Income	Hometown	Computer	Preparation	Gaming	Attendance	Job	English	Extra
0	Business Administration	Male	4.17	4.84	Low (Below 15,000)	Village	3	More than 3 Hours	0-1 Hour	80%-100%	No	3	Yes
1	Business Administration	Female	4.92	5.00	Upper middle (30,000-50,000)	City	3	0-1 Hour	0-1 Hour	80%-100%	No	3	Yes
2	Business Administration	Male	5.00	4.83	Lower middle (15,000-30,000)	Village	3	0-1 Hour	More than 3 Hours	80%-100%	No	4	Yes
3	Business Administration	Male	4.00	4.50	High (Above 50,000)	City	5	More than 3 Hours	More than 3 Hours	80%-100%	No	5	Yes
4	Business Administration	Female	2.19	3.17	Lower middle (15,000-30,000)	Village	3	0-1 Hour	2-3 Hours	80%-100%	No	3	Yes

Figure 2. Student Performance Metrics Dataset

Furthermore, the second public dataset used is the Higher Education Students Performance Evaluation dataset which has 32 features consisting of 25 features in the form of categorical data and 7 features in the form of numerical data. The following in Table 5 is the details of information from the Higher Education Students Performance Evaluation dataset, as well as Figure 3 which shows the appearance of the dataset before it is processed using the clustering method.

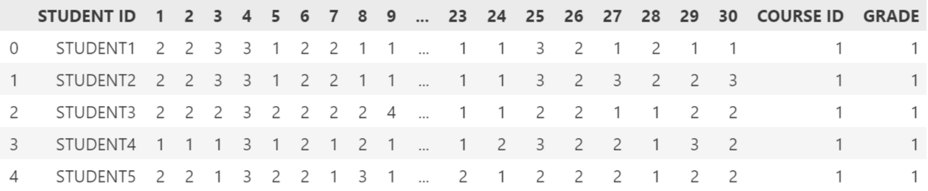
Table 5. Higher Education Students Performance Evaluation

No	Feature	Data Type	Description
1	Student Age	Categorical	1=18-2 years;2=22-25 years;3=Above 26 years
2	Gender	Binary	Encoded as 1 for female and 2 for male students.
3	Type of High School Graduated	Categorical	1=Private school;2=Public school; 3=other types of institution.
4	Scholarship Coverage	Categorical	Scholarship level:1=None; 2=25% discount; 3=50% discount; 4=75% discount; 5=Full scholarship.
5	Has Additional Work	Binary	Employment status coded as 1=Working; 2=Not working.
6	Engaged in Art or Sports Activities	Binary	Participation status: 1=Yes; 2=No.
7	Relationship Status	Binary	1=In a relationship or married; 2=Single or not in a relationship.
8	Monthly Income (if available)	Categorical	1=135-200; 2 = 201-270; 3 = 271-340; 4 = 341-410; 5 = Above 410 (all in USD)
9	Transportation Mode	Categorical	Encoded as 1 for Public bus; 2 for Private car or taxi; 3 for Bicycle and 4 for Other.
10	Accommodation Type	Categorical	Encoded as 1 for Rental housing; 2 for Dormitory or campus housing; 3 for Living with family and 4 for Other arrangements.
11	Mother’s Education Level	Categorical	Encoded as 1 for Primary; 2 for Secondary; 3 for High School; 4 for University; 5 for Master’s and 6 for Doctorate
12	Father’s Education Level	Categorical	Encoded as 1 for Primary; 2 for Secondary; 3 for High School; 4 for University; 5 for Master’s and 6 for Doctorate
13	Number of Siblings	Integer	1=One; 2=Two; 3=Three; 4=Four; 5=Five or More
14	Parental Marital Status	Categorical	1=Married; 2=Divorced; 3=Deceased (one or both)
15	Mother’s Occupation	Categorical	Encoded as 1 for Retired; Encoded as 2 for Homemaker; Encoded as 3 for Government Employee; Encoded as 4 for Private Sector; Encoded as 5 for Self-employed and encoded as 6 for Other
16	Father’s Occupation	Categorical	Encoded as 1 for Retired; Encoded as 2 for Homemaker; Encoded as 3 for Government Employee; Encoded as 4 for Private Sector; Encoded as 5 for Self-employed and encoded as 6 for Other
17	Weekly Study Hours	Categorical	Encoded as 1 for None; 2 for ¡5hours; 3 for 6–10hours; 4 for 11–20 hours; and 5 for = >20hours
18	Reading Frequency (Non-Academic)	Categorical	1=None;2=Occasionally;3=Frequently
19	Reading Frequency (Academic)	Categorical	1=None;2=Occasionally;3=Frequently
20	Participation in Department Activities	Binary	Encoded as 1 for yes and 2 for no
21	Impact of Projects or Activities on Achievement	Categorical	Encoded as 1 for positive; 2 for negative and 3 for neutral.
22	Class Attendance Regularity	Categorical	Encoded as 1 for always; 2 for sometimes and 3 for never.

(dilanjutkan di halaman berikutnya)

Tabel 5 (lanjutan)

No	Feature	Data Type	Description
23	Midterm Preparation Method	Categorical	1= Individual;2=Group;3=Not Applicable
24	Midterm Study Timing	Categorical	1=Near Exam Period;2=Consistently During Semester; 3=Never
25	Listening Attention in Class	Categorical	Encoded as 1 for never; 2 for sometimes and 3 for always.
26	Effect of Discussion on Interest and Performance	Categorical	Encoded as 1 for never; 2 for sometimes and 3 for always.
27	Experience with Flipped Learning	Categorical	Encoded as 1 for never; 2 for sometimes and 3 for always.
28	GPA in Previous Semester (/4.00)	Categorical	Encoded as 1 for not useful; 2 for useful and 3 for not aplicable.
29	Expected GPA at Graduation (/4.00)	Categorical	Encoded 1 for GPA <2.00;2 for GPA 2.00-2.49;3 for GPA 2.50-2.99; 4 for GPA 3.00-3.49; and 5 for GPA >3.49
30	Course Identifier	Categorical	1= <2.00;2=2.00-2.49; 3=2.50-2.99; 4=3.00-3.49; 5 = >3.49
31	Final Grade (Output Variable)	Integer	0= Fail; 1=DD; 2=DC; 3=CC; 4=CB; 5 =BB; 6=BA; 7=AA
32	Student Identifier	Categorical	Unique numeric ID assigned to each student



	STUDENT ID	1	2	3	4	5	6	7	8	9	...	23	24	25	26	27	28	29	30	COURSE ID	GRADE
0	STUDENT1	2	2	3	3	1	2	2	1	1	...	1	1	3	2	1	2	1	1	1	1
1	STUDENT2	2	2	3	3	1	2	2	1	1	...	1	1	3	2	3	2	2	3	1	1
2	STUDENT3	2	2	2	3	2	2	2	2	4	...	1	1	2	2	1	1	2	2	1	1
3	STUDENT4	1	1	1	3	1	2	1	2	1	...	1	2	3	2	2	1	3	2	1	1
4	STUDENT5	2	2	1	3	2	2	1	3	1	...	2	1	2	2	2	1	2	2	1	1

Figure 3. Higher education students performance evaluation

Figure 3 shows a snippet of the dataset to be processed using the clustering algorithm. From the image, we can see that the features of this dataset have been adjusted to the descriptions in Table 5 above: all categorical data has been converted to integer values. We did the same process on the first dataset, namely the Student Performance Metrics Dataset, where all categorical data was converted to numeric values, with the conversion carried out manually for both datasets.

3.2. SSE Results

The clustering evaluation using the Sum of Squared Errors (SSE) was performed on two publicly available datasets, namely DS1 (SPM) and DS2 (HESPE), employing three clustering techniques: K-Means, Gaussian Mixture Model (GMM), and BIRCH. SSE serves as an internal evaluation metric that assesses the compactness of clusters, where lower values indicate tighter and more cohesive groupings. This metric is particularly useful for comparing clustering algorithms, as it provides an objective measure of how well each method minimizes within-cluster variance. Table 6 presents a detailed comparison of the SSE values produced by the three clustering techniques across both datasets. Figure 4 and Figure 5 show a comparison of the SSE values for the two datasets used along with the required computation time.

Table 6. SSE Result for Each Dataset

Clustering Algorithm	DS1	DS2
K-Means	1.921.704.838	4.072.625.273
GMM	2.421.476.568	4.103.112.057
BIRCH	1.980.492.211	4.165.186.848

The clustering evaluation using Sum of Squared Errors (SSE) was conducted on two public datasets, namely DS1 (SPM) and DS2 (HESPE), using three clustering techniques: K-Means, Gaussian Mixture Model (GMM), and BIRCH. SSE is an internal evaluation metric that measures the compactness of clusters; lower SSE values indicate better clustering performance, as they indicate tighter, more cohesive clusters.

For the DS1 (SPM) dataset, the K-Means algorithm produced the lowest SSE value of 1921.70, outperforming GMM (2421.48) and BIRCH (1980.49). This result suggests that K-Means produces more compact clusters for DS1 than the other two methods. The relatively higher SSE values from GMM and BIRCH indicate that the clusters formed by these algorithms are less cohesive or contain higher intra-cluster variance. A similar pattern is observed in the DS2 (HESPE) dataset, where K-Means again achieved the lowest SSE value of 4072.63, followed closely by GMM (4103.11) and BIRCH (4165.19). Although the differences among the three techniques in DS2 are smaller compared to DS1, K-Means still demonstrates superior compactness in cluster formation.

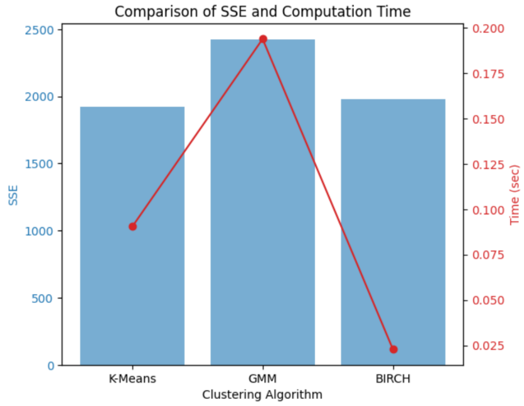


Figure 4. Comparison of sse and computation time with student performance metrics dataset

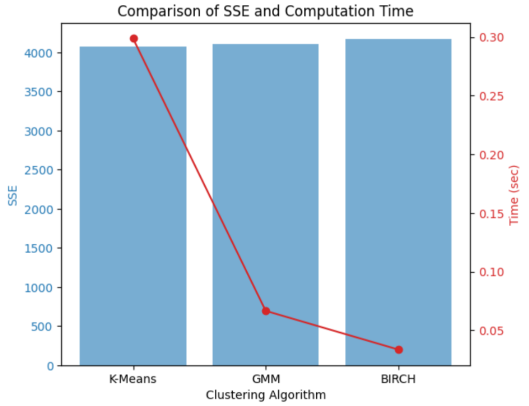


Figure 5. Comparison of sse and computation time with higher education students performance evaluation

Figures 4 and 5 visualize the comparison between the SSE values and computation times for the three clustering algorithms using a dual-axis representation. The left axis shows SSE values (blue bars) indicating the quality of the clustering results; lower values indicate more compact, well-separated clusters. The right axis shows computation time in seconds as a red line plot, reflecting each algorithm's efficiency. Figure 4 shows that in the first dataset, K-Means produced the lowest SSE value, followed by BIRCH and finally GMM. BIRCH had the lowest computation time in this dataset, while GMM had the highest. In Figure 5, we see that K-Means consistently produces the lowest SSE, followed by GMM and BIRCH. In this dataset, K-Means has the highest computation time, while BIRCH consistently has the lowest. Overall, based on the SSE comparison summarized in Table 6, K-Means consistently outperforms GMM and BIRCH across both datasets. This finding indicates that K-Means is more effective in forming compact clusters for the given datasets, making it the most suitable method among the three evaluated techniques for this particular clustering task. The results also imply that the underlying data distribution in both DS1 and DS2 is likely aligned with the assumptions of K-Means, enabling it to partition the data more efficiently.

3.3. DBI Results

The next step is to test the clustering results by examining the DBI values for three different clustering methods. Table 7 compares DBI values for the three clustering techniques. Figure 6 and Figure 7 compare the DBI values for the two datasets, along with the required computation time. The second evaluation of clustering performance was conducted using the Davies–Bouldin Index (DBI), an internal validation metric that assesses clustering quality by comparing the compactness and separation of clusters. In DBI, lower values indicate better cluster quality, as they reflect compact, well-separated clusters.

Table 7. DBI Result for Each Dataset

Clustering Algorithm	**DS1**	**DS2**
K-Means	1.455.965.449	3.241.410.082
GMM	1.002.104.355	331.262.754
BIRCH	1.484.417.027	3.065.569.194

For the DS1 (SPM) dataset, the K-Means algorithm achieved a DBI of 1.4559, slightly better than BIRCH (1.4844) and significantly better than GMM, which produced a much higher DBI of 10.0210. The large DBI value from GMM indicates poor separation between clusters and higher intra-cluster dispersion, suggesting that GMM struggled to form distinct cluster boundaries on DS1. Meanwhile, the close DBI values for K-Means and BIRCH indicate that both algorithms can generate relatively compact, well-separated clusters, with K-Means performing marginally better. For the DS2 (HESPE) dataset, the clustering performance shows a slightly different pattern. BIRCH produced the lowest DBI value of 3.0656, followed by GMM (3.3126) and K-Means (3.2414). Although the differences among the three are smaller than in DS1, BIRCH demonstrates better overall cluster separation and compactness for DS2. The results imply that the data distribution in DS2 may be better captured by the hierarchical, incremental nature of BIRCH than by the centroid-based approach of K-Means or the probabilistic modeling of GMM.

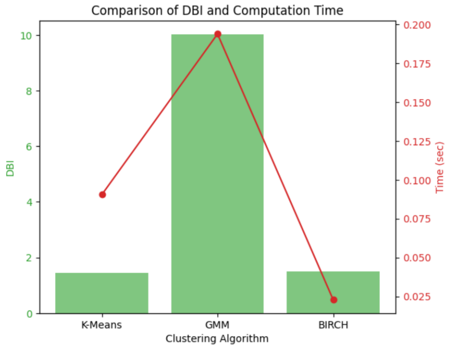


Figure 6. Comparison of dbi and computation time with student performance metrics dataset

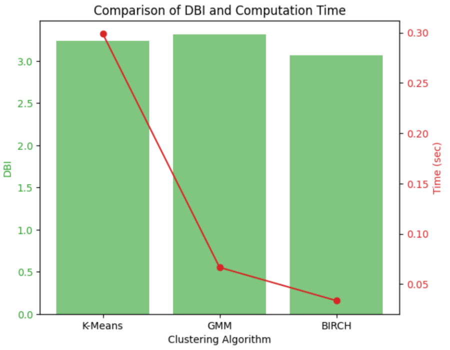


Figure 7. Comparison of dbi and computation time with higher education students performance evaluation

Figures 6 and 7 visualize the comparison between the Davies–Bouldin Index (DBI) and computation time for the three clustering algorithms using a dual-axis representation. The left axis shows DBI values (green bars) indicating the quality of the clustering results; lower values indicate more compact, well-separated clusters. The right axis shows computation time in seconds as a red line plot, reflecting each algorithm's efficiency. The figures above show that K-Means and BIRCH consistently have lower DBI values than GMM. In terms of computation time, Figures 6 and 7 also show that K-Means requires high computation time when processing dataset 2, whereas BIRCH is consistently the lowest on both datasets.

Overall, the DBI evaluation reveals that no single clustering algorithm consistently outperforms the others across both datasets. These results suggest that each dataset's characteristics influence clustering quality, and different algorithms may be more suitable depending on the underlying data structure.

3.4. Sihouette Score Results

The last part is the clustering results by looking at the Silhouette Score value by comparing three different clustering. The Silhouette Score evaluation was conducted to further assess clustering quality by measuring both cohesion and separation. Cohesion means how close data points are to their assigned cluster and separation means how far they are from other clusters. The Silhouette Score ranges from –1 to 1, where values closer to 1 indicate well-formed clusters, while values near 0 suggest overlapping or poorly separated clusters. Table 8 shows the comparison of the Silhouette Score values for the three clustering techniques used.

Table 8. Silhouette Score for Each Dataset

Clustering Algorithm	DS1	DS2
K-Means	0.240673565	0.053837111
GMM	0.050364075	0.050364075
BIRCH	0.250677039	0.05328957

The Silhouette Score evaluation was conducted to further assess clustering quality by measuring both cohesion (how close data points are to their assigned cluster) and separation (how far they are from other clusters). The Silhouette Score ranges from -1 to 1, where values closer to 1 indicate well-formed clusters, while values near 0 suggest overlapping or poorly separated clusters.

For the DS1 (SPM) dataset, the BIRCH algorithm achieved the highest Silhouette Score of 0.2507, indicating comparatively better cluster cohesion and separation than the other methods. K-Means produced a Silhouette Score of 0.2407, only slightly below BIRCH, suggesting similarly reasonable but not strongly separated clusters. Meanwhile, GMM obtained the lowest score (0.0503), reflecting weak cohesion and substantial overlap among clusters. This result indicates that GMM struggles to form distinct clusters for DS1, consistent with its poorer performance in the DBI evaluation. For the DS2 (HESPE) dataset, all algorithms yielded relatively low Silhouette Scores around 0.05, indicating minimal separation between clusters. BIRCH again produced the highest score (0.0533), followed by K-Means (0.0538) and GMM (0.0503). Although the differences are minor, the overall low scores suggest that DS2 has less distinct cluster structures, making it more difficult for all algorithms to form well-separated groups.

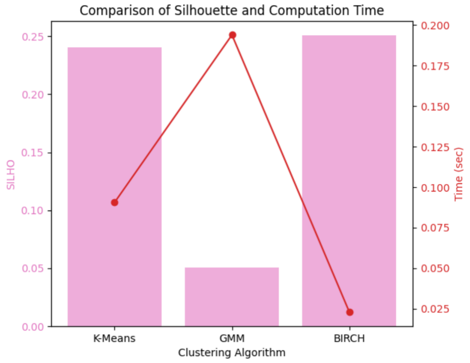


Figure 8. Comparison of silhouette score and computation time with the student performance metrics dataset

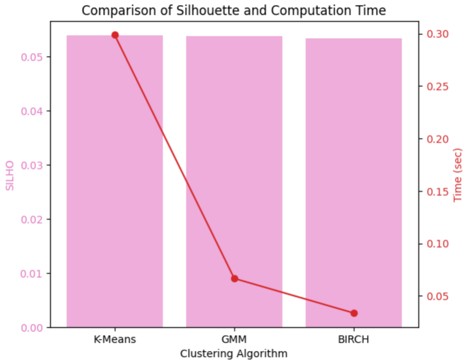


Figure 9. Comparison of silhouette score and computation time with the performance evaluation of higher education students

Figures 8 and 9 present a comparison of Silhouette Scores and computation time for clustering algorithms in a dual-axis plot. The left axis displays the Silhouette Score in pink bars, representing the cohesion and separation of the resulting clusters; higher Silhouette Scores indicate more well-defined cluster boundaries. The right axis illustrates computation time in seconds using a red line plot. As in the previous test pattern, in the test with silhouette score, BIRCH consistently produces the lowest computation time for each dataset. In contrast, K-Means in the first dataset takes more time than BIRCH but less than GMM, whereas in the second dataset, it consistently requires the highest computation time. In summary, the Silhouette Score results reinforce earlier findings from SSE and DBI, showing that BIRCH and K-Means generally produce more cohesive clusters. In contrast, GMM tends to form less distinct clustering structures across both datasets.

3.5. Analysis

Based on the test results, for the Student performance metrics (SPM) dataset (DS1), the best SSE value was obtained with the K-Means algorithm (1921.704838), followed by the BIRCH algorithm (1980.492211), and finally GMM (2421.476568). For the HESPE dataset (DS2), the best SSE score was obtained when using K-Means with an SSE value of 4072.625273, followed by GMM with an SSE value of 4103.112057, and finally BIRCH with an SSE value of 4165.186848. In the evaluation with DBI, based on the test results for the SPM dataset (DS1), the K-Means algorithm yielded the best DBI of 1.455965449. For the HESPE dataset (DS2), the best DBI was achieved with BIRCH at 3.065569194. Finally, for the SPM dataset (DS1), the BIRCH algorithm achieved the best Silhouette Score of 0.250677039. In contrast, for the HESPE dataset (DS2), the best Silhouette Score was achieved with K-Means (0.053837111).

The results of the SSE analysis indicate that K-Means consistently outperformed the other methods on both datasets, achieving the lowest SSE values and demonstrating superior compactness. This suggests that centroid-based K-Means optimization is particularly effective at producing cohesive clusters for the data characteristics observed in DS1 and DS2. In contrast, the evaluation using DBI shows a more nuanced performance pattern. For DS1, K-Means yielded the lowest DBI value, followed closely by BIRCH, indicating well-separated and compact clusters. However, for DS2, BIRCH obtained the best DBI score, showing stronger cluster separation than both K-Means and GMM. GMM, on the other hand, recorded substantially higher DBI values, especially for DS1, suggesting weaker separation and less distinct clustering structures. The Silhouette Score results reinforce these findings, where BIRCH achieved the highest score on DS1 and marginally outperformed the other methods on DS2. Nevertheless, the generally low Silhouette Scores on DS2 indicate that this dataset contains more overlapping cluster structures, making it inherently more challenging for all algorithms.

By integrating the findings across all evaluation metrics, it can be concluded that K-Means and BIRCH demonstrate the most reliable performance in this study, albeit with different strengths. K-Means excels in forming compact clusters, as consistently shown through SSE, while BIRCH exhibits superior separation quality based on DBI and Silhouette Score. Conversely, GMM shows comparatively weaker performance across the datasets. Overall, these results highlight that the underlying characteristics of each dataset strongly influence clustering effectiveness and that selecting an appropriate clustering method requires considering multiple validation metrics rather than relying on a single criterion. The test results also show that the k-means algorithm produces good clusters while maintaining low computational cost [9, 27]. In terms of computational time, BIRCH is much superior to the other two algorithms, with the computational time required by BIRCH for each test technique in both datasets consistently lower than that required by GMM and K-Means, which are superior in test value to BIRCH.To assess whether the performance differences among the three clustering algorithms were statistically significant, a one-way ANOVA test was performed across 30 repeated trials with varying random seeds. The test yielded an F-statistic of 1904.16 and a p-value of $1.51 \times 10{-}72$, indicating that the differences in SSE produced by K-Means, GMM, and BIRCH were statistically significant. These ANOVA test results indicate that the SSE value obtained by K-Means in the previous clustering test was not due to random variability but rather reflected the algorithm's superior performance. This research provides comprehensive comparative evidence on the performance of several widely used clustering algorithms and demonstrates how the choice of methodology significantly affects the identification and interpretation of student performance patterns. In processing the dataset, the algorithms used in this study operate on a distance-based or probabilistic basis, which is more suited to numeric and continuous data, making them less effective for purely categorical datasets without adequate preprocessing. Previous literature also indicates that distance measures, such as the Euclidean distance, which underlie algorithms like K-Means and BIRCH, lack semantic meaning for categorical attributes or high-dimensional data [28]. In contrast, probabilistic models like GMM assume a continuous Gaussian distribution, which is not suited to discrete data. Furthermore, existing data, particularly private student data, is largely categorical. Therefore, using and testing various techniques for processing categorical data is essential for comparing with existing research results.

4. CONCLUSION

This research was conducted to identify clustering techniques suitable for students' academic performance. The research used two public datasets and three clustering algorithms: K-Means, Gaussian Mixture Model, and BIRCH. Furthermore, the clustering results for each technique were evaluated using three clustering metrics: SSE, DBI, and Silhouette Score. In general, K-Means is superior to GMM and BIRCH, excelling in 4 clustering test results compared to BIRCH, which excels in only 2. In terms of computational time, BIRCH outperforms K-Means and GMM for clustering the two public datasets. Based on these results, if the computing resources are sufficient, we can use the K-Means clustering technique to achieve good clustering. On the other hand, if the available computing resources are minimal, we can consider the BIRCH algorithm for its fast computation and clustering results that are not too far from the K-Means results. Also, from the result, we can find that clustering effectiveness is strongly influenced by algorithm characteristics and dataset structure, with K-Means being more suitable for accuracy-oriented grouping and BIRCH for time-critical applications.

Overall, this research contributes to the field of educational data mining by providing comparative evidence of algorithm dominance and demonstrating how methodological choices in clustering can substantially affect the interpretation of student performance patterns. In practice, the results offer actionable guidance for educational institutions, enabling them to select clustering techniques that align with their operational needs, whether for precise academic profiling with K-Means or for rapid, large-scale analysis with BIRCH, thereby supporting more effective student monitoring, early decision-making, and data-driven policy formulation. This study has limitations: the use of two public datasets, testing with only three clusters, and the use of default parameter settings for each algorithm. Future research could include more diverse datasets, private datasets, hyperparameter optimization, deep learning techniques, capabilities for automatically processing categorical data, and interpretable clustering results.

5. ACKNOWLEDGEMENTS

The Acknowledgments section is optional. Research sources can be included in this section.

6. DECLARATIONS

AI USAGE STATEMENT

The authors used ChatGPT and Grammarly to improve the language and clarity of the manuscript. The authors still reviewed and edited the content and took full responsibility for it.

AUTHOR CONTIBUTION

Ricky Aurelius Nurtanto Diaz, as the first author, is in charge of modeling and coding, Ni Luh Gede Pivin Suwirmayanti, as the second author, is in charge for data collection and processing, Emy Setyaningsih, as the third author, is in charge of model evaluation, and I Wayan Budi Sentana as fourth author is in charge of model evaluation and proofreading.

FUNDING STATEMENT

This study used independent funding from the authors.

COMPETING INTEREST

The authors confirm that there are no conflicts of interest, financial or non-financial, that could influence the research results or the interpretation of the data in this article.

REFERENCES

[1] M. Alvarez-Garcia, M. Arenas-Parra, and R. Ibar-Alonso, "Uncovering student profiles. An explainable cluster analysis approach to PISA 2022," *Computers & Education*, vol. 223, p. 105166, Dec. 2024, https://doi.org/10.1016/j.compedu.2024.105166.

[2] Z. Zhang, X. Zeng, H. Bao, and B. Li, "Intelligent Student Performance Clustering and Personalized Teaching Suggestions Based on K-Means," in *Proceedings of the 2024 International Conference on Digital Society and Artificial Intelligence*. Qingdao China: ACM, May 2024, pp. 38–42, https://doi.org/10.1145/3677892.3677898.

[3] W. Chen, Z. Wu, S. Zeng, H. Guo, and J. Li, "Diverse behavior clustering of students on campus with macroscopic attention," *Scientific Reports*, vol. 15, no. 1, p. 29800, Aug. 2025, https://doi.org/10.1038/s41598-025-15103-8.

[4] M. Gul and M. A. Rehman, “Big data: An optimized approach for cluster initialization,” *Journal of Big Data*, vol. 10, no. 1, p. 120, Jul. 2023, https://doi.org/10.1186/s40537-023-00798-1.

[5] K. Ouassif, B. Ziani, J. Herrera-Tapia, and C. A. Kerrache, “Empowering Education: Leveraging Clustering and Recommendations for Enhanced Student Insights,” *Education Sciences*, vol. 15, no. 7, p. 819, Jun. 2025, https://doi.org/10.3390/educsci15070819.

[6] E. Kalita, S. S. Oyelere, S. Gaftandzhieva, K. N. V. P. S. Rajesh, S. K. Jagatheesaperumal, A. Mohamed, Y. M. Elbarawy, A. S. Desuky, S. Hussain, M. A. Cifci, P. Theodorou, S. Hilčenko, J. Hazarika, and T. Ali, “Educational data mining: A 10-year review,” *Discover Computing*, vol. 28, no. 1, p. 81, May 2025, https://doi.org/10.1007/s10791-025-09589-z.

[7] S. J. Sultan Alalawi, I. N. Mohd Shaharanee, and J. Mohd Jamil, “Clustering Student Performance Data Using k-Means Algorithms,” *Journal of Computational Innovation and Analytics (JCIA)*, vol. 2, no. 1, pp. 41–55, Jan. 2023, https://doi.org/10.32890/jcia2023.2.1.3.

[8] A. F. Mohamed Nafuri, N. S. Sani, N. F. A. Zainudin, A. H. A. Rahman, and M. Aliff, “Clustering Analysis for Classifying Student Academic Performance in Higher Education,” *Applied Sciences*, vol. 12, no. 19, p. 9467, Sep. 2022, https://doi.org/10.3390/app12199467.

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

[10] P. Economou, “A clustering algorithm for overlapping Gaussian mixtures,” *Research in Statistics*, vol. 1, no. 1, p. 2242337, Oct. 2023, https://doi.org/10.1080/27684520.2023.2242337.

[11] B. Chassagnol, A. Bichat, C. Boudjeniba, P.-H. Wuillemin, M. Guedj, D. Gohel, G. Nuel, and E. Becht, “Gaussian Mixture Models in R,” *The R Journal*, vol. 15, no. 2, pp. 56–76, Nov. 2023, https://doi.org/10.32614/RJ-2023-043.

[12] A. Lang and E. Schubert, “BETULA: Fast clustering of large data with improved BIRCH CF-Trees,” *Information Systems*, vol. 108, p. 101918, Sep. 2022, https://doi.org/10.1016/j.is.2021.101918.

[13] R. Wang and J. Li, “Fast sparse representative tree splitting via local density for large-scale clustering,” *Scientific Reports*, vol. 15, no. 1, p. 29398, Aug. 2025, https://doi.org/10.1038/s41598-025-13848-w.

[14] A. A. Wani, “Comprehensive analysis of clustering algorithms: Exploring limitations and innovative solutions,” *PeerJ Computer Science*, vol. 10, p. e2286, Aug. 2024, https://doi.org/10.7717/peerj-cs.2286.

[15] S. Pitafi, T. Anwar, and Z. Sharif, “A Taxonomy of Machine Learning Clustering Algorithms, Challenges, and Future Realms,” *Applied Sciences*, vol. 13, no. 6, p. 3529, Mar. 2023, https://doi.org/10.3390/app13063529.

[16] P. Artioli, A. Maci, and A. Magrì, “A comprehensive investigation of clustering algorithms for User and Entity Behavior Analytics,” *Frontiers in Big Data*, vol. 7, p. 1375818, May 2024, https://doi.org/10.3389/fdata.2024.1375818.

[17] J. M. John, O. Shobayo, and B. Ogunleye, “An Exploration of Clustering Algorithms for Customer Segmentation in the UK Retail Market,” *Analytics*, vol. 2, no. 4, pp. 809–823, Oct. 2023, https://doi.org/10.3390/analytics2040042.

[18] S. Bhurre, S. Prajapat, and S. Raikwar, “Performance Pattern Mining for Higher Education Students in Blended Learning Using Clustering Algorithms,” in *Contributions Presented at The International Conference on Computing, Communication, Cybersecurity and AI, July 3–4, 2024, London, UK*, N. Naik, P. Jenkins, S. Prajapat, and P. Grace, Eds. Cham: Springer Nature Switzerland, 2024, vol. 884, pp. 362–386, https://doi.org/10.1007/978-3-031-74443-3_22.

[19] J. Dong, R. Sun, Z. Yan, M. Shi, and X. Bi, “Research on learning achievement classification based on machine learning,” *PLOS One*, vol. 20, no. 6, p. e0325713, Jun. 2025, https://doi.org/10.1371/journal.pone.0325713.

[20] I. K. Khan, H. B. Daud, N. B. Zainuddin, R. Sokkalingam, M. Farooq, M. E. Baig, G. Ayub, and M. Zafar, “Determining the optimal number of clusters by Enhanced Gap Statistic in K-mean algorithm,” *Egyptian Informatics Journal*, vol. 27, p. 100504, Sep. 2024, https://doi.org/10.1016/j.eij.2024.100504.

[21] T.-C. Liu, P. N. Kalugin, J. L. Wilding, and W. F. Bodmer, "GMMchi: Gene expression clustering using Gaussian mixture modeling," *BMC Bioinformatics*, vol. 23, no. 1, p. 457, Nov. 2022, https://doi.org/10.1186/s12859-022-05006-0.

[22] R. Ulug, "Implementation of the BIRCH algorithm to construct a data-adaptive network design for regional gravity field modeling via SRBF," *Earth Science Informatics*, vol. 18, no. 2, p. 202, Jun. 2025, https://doi.org/10.1007/s12145-025-01712-4.

[23] A. Y. Raya-Tapia, F. J. López-Flores, C. Ramírez-Márquez, and J. M. Ponce-Ortega, *Fundamentals of Clustering: Methods, Metrics, and Optimization*. Cham: Springer Nature Switzerland, 2025, vol. 1233, pp. 13–50, https://doi.org/10.1007/978-3-032-03876-0_2.

[24] B. Sadeghi, "Clustering in geo-data science: Navigating uncertainty to select the most reliable method," *Ore Geology Reviews*, vol. 181, p. 106591, Jun. 2025, https://doi.org/10.1016/j.oregeorev.2025.106591.

[25] L. E. Ekemeyong Awong and T. Zielinska, "Comparative Analysis of the Clustering Quality in Self-Organizing Maps for Human Posture Classification," *Sensors*, vol. 23, no. 18, p. 7925, Sep. 2023, https://doi.org/10.3390/s23187925.

[26] K. Amrulloh, T. H. Pudjiantoro, P. N. Sabrina, and A. I. Hadiana, "Comparison Between Davies-Bouldin Index and Silhouette Coefficient Evaluation Methods in Retail Store Sales Transaction Data Clusterization Using K-Medoids Algorithm," in *Proceedings of the International Conference on Industrial Engineering and Operations Management*. Asuncion, Paraguay: IEOM Society International, Jul. 2022, pp. 1952–1961, https://doi.org/10.46254/SA03.20220384.

[27] N. L. G. P. Suwirmayanti, E. Setyaningsih, R. A. N. Diaz, and K. Budiarta, "Optimization of the K-Means Method for Clustering Banking Data Using the Hybrid Model of Invasive Weed Optimization and K-Means (IWOKM)," 2024, https://doi.org/10.24507/icicel.18.04.413.

[28] N. L. G. P. Suwirmayanti, I. K. G. D. Putra, M. Sudarma, I. M. Sukarsa, E. Setyaningsih, and R. A. N. Diaz, "Invasive Weed Optimization K-Means Performance Robust Operations (IWOKM PRO) in High-Dimensional Datasets," *Engineering, Technology & Applied Science Research*, vol. 15, no. 4, pp. 24 390–24 395, Aug. 2025, https://doi.org/10.48084/etasr.11112.