Bulletin of Information Technology (BIT)
Vol 6, No 4, Desember 2025, Hal. 361 - 366
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2293
https://journal.fkpt.org/index.php/BIT

Clustering Academic Data of Junior High School Student to Identify Learning Groups Using the DBSCAN Algorithm at SMP Muhammadiyah 5 Samarinda

Mini H[1*], Siti Lailiyah[2], Salmon[3,]

[1] Teknik Informatika, STMIK Widya Cipta Dharma, Samarinda, Indonesia
[2] Teknik Informatika, STMIK Widya Cipta Dharma, Samarinda, Indonesia
[3] Sistem Informasi, STMIK Widya Cipta Dharma, Samarinda, Indonesia
Email: [1*]2243074@wicida.ac.id , [2]lail.59a@gmail.com , [3]salmon@wicida.ac.id
(* : coressponding author: 2243074@wicida.ac.id)

Abstract- This study aims to identify learning groups based on student academic data at SMP Muhammadiyah 5 Samarinda. The data used includes exact and non-exact subject scores, exam results, assignment scores, attendance, and parents' educational backgrounds. The stages of the research include data collection, data preprocessing through cleaning, feature engineering, and transformation, data processing to determine the optimization values of the DBSCAN parameters, namely eps and minpts, and evaluation of the results using the Silhouette Score. The optimal parameters obtained were eps = 1.3 and min_samples = 3, resulting in three main clusters and some noise. The analysis results showed three main clusters, namely cluster 0 with 89 students (medium achievement), cluster 1 with 50 students (high achievement), and cluster 2 with 5 students (low achievement), as well as 14 students identified as noise. A Silhouette Score value of 0.217 indicates relatively weak cluster separation quality, but DBSCAN is able to detect noise that may not be detected by other algorithms. These findings indicate that even though the quality of the clusters is not yet optimal, the algorithm used is still useful for exploring student learning patterns and can serve as the basis for more targeted learning interventions.

Keywords: Clustering, DBSCAN, Academic Data, Study Group, Silhouette Score

1. INTRODUCTION

Lower secondary education is a strategic stage in forming the foundation of knowledge and skills, which plays an important role in determining students' readiness to continue to the next level of education. Students face various academic challenges, which require appropriate learning methods so that their learning potential can develop optimally. One method commonly applied in schools is the formation of study groups. This strategy is considered capable of improving students' understanding of the subject matter, motivating them to study harder, and fostering essential social skills in interactions between students[1]. In addition, group tutoring has also been proven to contribute to the development of good study habits, which are one of the main determinants of student academic success[2].

The formation of study groups in schools has not been implemented in a structured manner and is not supported by classification procedures based on students' academic abilities or learning styles. To overcome these obstacles, schools can take advantage of technology and data analysis, which are currently developing rapidly in the world of education. In educational data analysis, the application of machine learning has significant potential[[3]. One approach that shows great potential is the use of data mining to group students based on their similar characteristics [4],[5].

Data mining is the process of analyzing large amounts of data to discover hidden patterns and important information that can support systematic and objective decision-making. In the context of education, data mining enables educators to objectively identify student learning patterns, academic potential, and individual learning needs. One popular method in data mining is clustering, which aims to group data based on certain similarities between attributes. This technique is highly relevant in the world of education, as it can help form homogeneous learning groups, identify high-achieving students, and design personalized learning strategies[6].

The most commonly used clustering algorithm is K-Means, which is known for its performance in segmenting data quickly and efficiently. The effectiveness of K-Means has been proven in various studies, including grouping high school/MA students' national exam results[7], analyzing student academic data[8], forming study groups based on student performance at the junior high school level [9], and has been applied to religious-based educational institutions such as madrasas to analyze student learning achievements[10] . Other studies show that this method can help teachers determine learning strategies based on student grade groups [11],[12], and can be used in grouping student academic achievement indices [13]. Furthermore, at the university level, this algorithm has been implemented to support data-driven curriculum development[14].

In addition, the K-Means algorithm has limitations, especially in handling data that is not evenly distributed or contains noise. In the field of education, this condition is common in primary education, where student data is heterogeneous and has high variation in characteristics. To address this issue, the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is used as an alternative that can form clusters based on data density and effectively detect anomalies or outliers[15]. DBSCAN has the advantage of not requiring a predetermined number of clusters, which is often a weakness of K-Means. In practice, this algorithm has been used to analyze patterns of student visits to libraries [16]. Evaluating the quality of clustering results is also an important aspect that cannot be ignored. Several methods such

Bulletin of Information Technology (BIT)
Vol 6, No 4, Desember 2025, Hal. 361 - 366
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2293
https://journal.fkpt.org/index.php/BIT

as the Davies Bouldin Index, Elbow Method, and Silhouette Coefficient are commonly used to assess the effectiveness of clustering models in the context of student academic classification [17]. In addition, evaluation methods such as Silhouette Score and Davies-Bouldin Index have been proven effective for measuring the quality of clustering results in both K-Means and DBSCAN algorithms[18].

Based on the literature review, most previous studies still focus on the application of the K-Means algorithm as the main method. However, student data characteristics at the junior high school level tend to be more complex and heterogeneous. In addition, non-academic variables such as attendance rates and parental educational backgrounds are rarely considered as components of analysis. In fact, several studies explain that students' academic success is not only influenced by subject grades, but also by external factors. This study aims to apply the DBSCAN algorithm in clustering the academic data of students at SMP Muhammadiyah 5 Samarinda. By utilizing academic data, student attendance, and parental educational background, this study is expected to produce learning groups that can support the development of more effective learning strategies and provide practical contributions to teachers and schools in designing more targeted learning strategies.

2. RESEARCH METHODOLOGY

2.1 Research Stages

This research was systematically designed to produce academic data groupings that could represent learning patterns. In the initial stage, data was collected from the academic scores and attendance percentages of students at SMP Muhammadiyah 5 Samarinda, with a total of 158 entries. The collected data still contained possible duplicates and extreme values, so data cleaning was performed by removing duplicates and detecting outliers based on z-scores. After cleaning, 157 more representative data entries were obtained. The research flow is visualized in Figure 1.

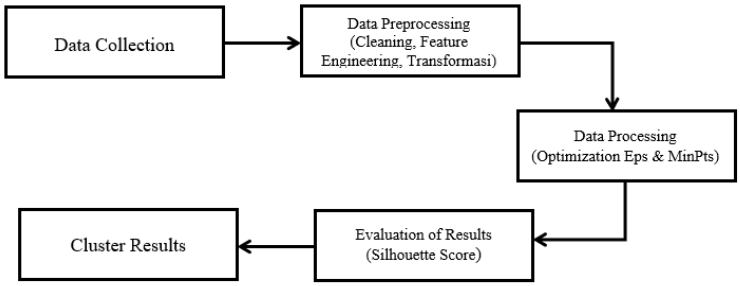

Figure 1. Research Stages

Research Stages:

a. Data Collection
   The dataset used contains the academic scores of students at Muhammadiyah 5 Samarinda Junior High School (exact, non-exact, exams, assignments, attendance, parental education). The exact sciences category includes mathematics and natural sciences, while the non-exact science category of Indonesia language, English language, sosial sciences, cultural arts, physical education, sports and health, and civics. Thus, this grouping of students provides a more structured picture of students' academic ability trends. The data consists of 158 data populations.
b. Data Preprocessing
   1. Cleaning: duplicates removed: extreme values (outliers) discarded using the z-score > 3 approach.
   2. Feature Engineering: created new features such as Academic_Score2 (combination of exact scores, non-exact scores, exams, assignments), Consistency_Score (difference between exams and assignments), Attendance_Score (normalized to 0–1), Parent Score (parents' education), and Achievement_Index as the final combined score. In addition, achievement categories (Low, Medium, High) are assigned based on quantiles.
   3. Transformation: all features are standardized using StandardScaler to ensure a uniform value range.
c. Data Processing
   1. The eps and min_samples parameters are determined using a grid search approach and k-distance plot to find the optimal values.
   2. In the experiment, the combination of eps=1.3 and min_samples=3 produced three clusters with minimal noise (14 students).
d. Evaluation of Results
   Evaluation using the Silhouette score to assess compactness and cluster separation. The silhouette value obtained was 0.217, indicating that the clusters were still in the weak-to-moderate category. Nevertheless, these results are still useful for exploring student learning patterns.
e. Cluster Results

Bulletin of Information Technology (BIT)
Vol 6, No 4, Desember 2025, Hal. 361 - 366
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2293
https://journal.fkpt.org/index.php/BIT

1. Cluster 0 (89 students): the majority group with moderate achievement indices
2. Cluster 1 (50 students): higher achievement group.
3. Cluster 2 (5 students): a small group with specific tendencies.
4. Noise (14 students): students who are not included in the cluster (outliers).

2.2 DBSCAN Algorithm

The DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm is used to cluster data based on point density. Two important parameters used are epsilon (ε) and Minpts. The value ε indicates the maximum distance between points to be considered neighbors, while Minpts indicates the minimum number of points required to form a cluster. The distance between two points p and q is calculated using Euclidean Distance using equation (1):

$$dist(p,q) = \sqrt{(\Sigma_{i=1}^{n}(p_i - q_i)^2)} \qquad (1)$$

Where :
$dist(p,q)$: Euclidean distance between point of data p and data point q
p_i : Value of the i-th feature at the data point p
q_i : Value of the i-th feature at the data point q
n : Number of feature or data dimensions

This equation calculates the degree of closeness between two points based on the difference in values of each feature. If the distance is smaller, the two points will be closer together in the data space. After the distance is calculated, the Epsilon-Neighborhood set of points within radius ε of the center point p is defined in equation (2):

$$N(p) = \{q \in D \mid dist(p,q) \leq \varepsilon\} \qquad (2)$$

Where :
$N(p)$: The set points that are neighbors of point p
$q \in D$: Point q, which is part of the entire dataset D
$dist(p,q)$: The distance between data point p dan data point q
ε : Maximum neighbors radius

This set is used to test the density of a point around point p. If the number of neighbors meets the minimum value criterion Minpts, then point p will be categorized as a core point in cluster formation. whereas, the number of neighbors who do not meet the minimum requirements is set as a border point or noise.

2.3 Application of Methods

This research method was applied in the following stages:

a. Implementation was carried out in the Python Jupyter Notebook environment, utilizing the pandas, scikit-learn, and matplotlib libraries..
b. Notebook Clustering_DBSCAN_Siswa.ipynb contains the entire pipeline from preprocessing to evaluation.
c. The modified dataset (data_final_mod_clustered_mid.csv) is used as input for DBSCAN after undergoing cleaning, feature engineering, and standardization..
d. The final results are visualized in the form of a 2D PCA scatter plot, which shows the distribution of clusters and student noise.

3. RESULTS AND DISCUSSION

This section presents the results of applying the DBSCAN algorithm to student academic data, including cluster distribution, grouping quality evaluation, and interpretation of each cluster's characteristics.

3.1 DBSCAN Clustering Results

The DBSCAN algorithm produced three main clusters and a number of noise data points. The distribution of members in each cluster is shown in Table 1.

Table 1. Distribution of DBSCAN Cluster Results

Cluster	Number of Members	General Description
0	89 students	Majority, average achievement index, stable attendance

Bulletin of Information Technology (BIT)
Vol 6, No 4, Desember 2025, Hal. 361 - 366
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2293
https://journal.fkpt.org/index.php/BIT

1	50 students	High achievement, consistent good grades
2	5 students	Small group, low/typical achievement
Noise	14 students	outlier

A total of 144 students (91.1%) were successfully grouped into three clusters, while 14 students (8.9%) were categorized as noise. This shows that DBSCAN is capable of identifying most patterns in the dataset, although there is still some data that does not meet the cluster density criteria.

3.2 Cluster Quality Evaluation

Cluster quality was measured using the Silhouette Score with a value of 0.217. This value indicates that the separation between clusters is still relatively weak, because some of the data is located at a similar distance between one cluster and another. Nevertheless, these values still indicate the existence of group structures, so that the clustering results can still be used as a basis for exploring student learning patterns.
The visualization of the clustering results is shown in Figure 2 using 2D PCA dimension reduction. This image shows the distribution of three main clusters and the position of data categorized as noise. Next, Figure 3 shows the K-distance plot used to determine the optimal eps parameter.

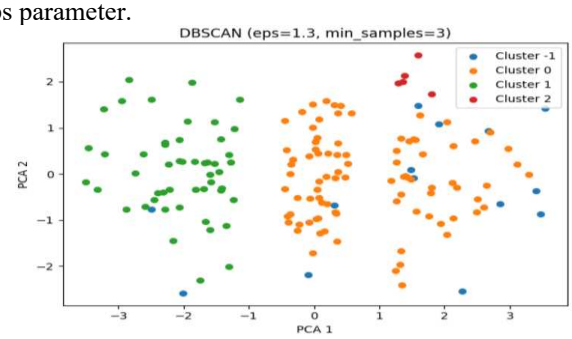

Figure 2. Visualization of DBSCAN Clustering Results with 2D PCA

The figure illustrates the results of data clustering using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm with an epsilon (eps) value of 1.3 and a min_samples parameter of 3. Prior to clustering, the data were reduced in dimensionality using Principal Component Analysis (PCA) to enable two-dimensional visualization. The results indicate that DBSCAN successfully identified three main clusters with varying density levels, as well as several data points classified as noise or outliers due to insufficient local density. These findings demonstrate that DBSCAN is effective for density-based clustering without requiring a predefined number of clusters and is well suited for datasets with irregular distributions and the presence of outliers.

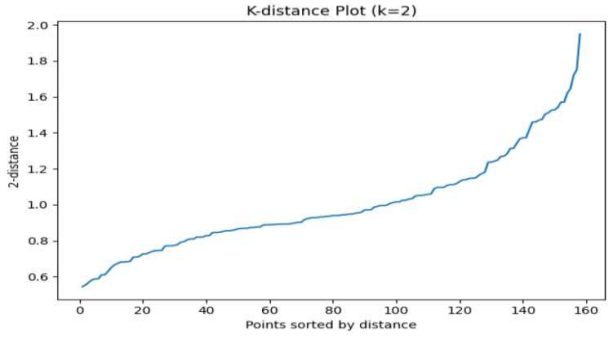

Figure 3. K-distance Plot

The figure presents a K-distance plot with k=2, which is commonly used to determine an appropriate epsilon (eps) value for the DBSCAN algorithm. The plot shows the distances to the second nearest neighbor for each data point, sorted in ascending order. A gradual increase in distance is observed for most points, followed by a sharp rise toward the end of the curve, forming an elbow-like shape. This point of rapid increase indicates a transition from dense regions to sparse

Bulletin of Information Technology (BIT)
Vol 6, No 4, Desember 2025, Hal. 361 - 366
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2293
https://journal.fkpt.org/index.php/BIT

regions and is typically selected as the optimal eps value. Therefore, the K-distance plot serves as an effective tool for identifying a suitable neighborhood radius that enables DBSCAN to distinguish clusters from noise accurately.

3.3 Interpretation of Clusters

The interpretation was done by analyzing the average features of each cluster, which included combined academic scores, achievement index, attendance, grade consistency, and parents' educational background. The interpretation results are shown in Table 2.

Table 2. Interpretation of Characteristics of Each Cluster

Cluster	Key Characteristics	Interpretation
0	Average academic performance stable attendance	Majority group, fairly good performance
1	High academic grades, consistent	High-achieving students, potential for acceleration
2	Nilai rendah/tidak stabil, kehadiran kurang	At-risk students, need additional guidance
Noise	Unique/extreme profile	Outlier, requires special investigation

Overall, the clustering results using the DBSCAN algorithm show three main groups and a number of students detected as noise. Although the Silhouette Score is still relatively low, these findings provide an initial overview of the variations in student learning patterns, which can be used as a basis for consideration in developing learning strategies.

4. CONCLUSION

This study applied the DBSCAN algorithm to cluster the academic data of students at Muhammadiyah 5 Junior High School in Samarinda.The clustering results show the formation of three main groups with relatively varying numbers of members, namely 89 students in cluster 0, 50 students in cluster 1, and 5 students in cluster 2, as well as 14 students categorized as noise. Evaluation with Silhouette Score produced a value of 0.217, indicating that the cluster separation quality was still weak. This may be due to the homogeneity of student academic data and the limitations of the variables used. Nevertheless, the results of the study still make an important contribution in the context of exploring student learning patterns. The majority cluster describes a group of students with average performance, the second cluster reflects high-achieving students who have the potential to be given enrichment programs, while the third cluster indicates students with low achievement risk who need more intensive guidance. These findings can serve as input for schools to implement more targeted learning interventions. For further research, it is recommended to add non-academic variables such as learning motivation, family support, and extracurricular activities to improve the quality of clustering and make the interpretation of results more comprehensive.

REFERENCES

[1] Nureki, Syamsuria, and E. Azis, "Pengaruh Kelompok Belajar Terhadap Peningkatan Hasil Belajar Siswa," *BEGIBUNG J. Penelit. Multidisiplin*, vol. 2, no. 1, pp. 293–301, 2024.

[2] M. Jannah, F. A. Alam, and Taufik, "Pengaruh Layanan Bimbingan Kelompok Dalam Meningkatkan Disiplin Belajar Siswa UPTD SMP Negeri 33 Barru," *J. Bimbing. dan Konseling*, vol. 1, pp. 27–38, 2023.

[3] H. Pratiwi, M. I. Sa'ad, and Salmon, "Strategi Manajemen Pendidikan Berbasis Machine Learning untuk Prediksi Prestasi Siswa," *Borneo Educ. Manag. Res. J.*, vol. 6, no. 1, 2025.

[4] R. K. Hapsari, T. Indriyani, and D. H. Sulaksono, *Buku Ajar Data Mining*. Yogyakarta: Deepublish, 2025.

[5] Mustika *et al.*, *Data Mining dan Aplikasinya*. Bandung: CV Widina Media Utama, 2021.

[6] Y. F. Sinurat, *Data Mining Pengelompokan Siswa Berprestasi Menggunakan Metode Clustering*. Jakarta: Penerbit NEM, 2024.

[7] I. Suputra, I. Candiasa, and I. Suryawan, "Klasterisasi Hasil Ujian Nasional SMA/MA dengan Algoritma K-Means," *Wahana Mat. dan Sains J. Mat. Sains, dan Pembelajarannya*, vol. 15, no. 1, 2021.

[8] Melizah, A. Anto Tri Susilo, N. Lestari, and Elmayati, "Implementasi Algoritma K-Means Clustering Untuk Analisis Data Nilai Akademik Mahasiswa," *J. Teknol. Inf. Mura*, vol. 16, no. 2, 2024.

[9] A. A. Mila, R. T. Abineno, and A. A. Pekuwali, "Pengelompokan Performa Siswa Dalam Pembelajaran Bahasa Indonesia

Bulletin of Information Technology (BIT)
Vol 6, No 4, Desember 2025, Hal. 361 - 366
ISSN 2722-0524 (media online)
DOI 10.47065/bit.v5i2.2293
https://journal.fkpt.org/index.php/BIT

Menggunakan Algoritma K-Means Clustering Di Smpn Satap Lambakara," *SATI Sustain. Agric. Technol. Innov.*, vol. 3, no. 1, pp. 593–603, 2024.

[10] Abdur Rohman Nurut Toyyibin and Zaehol Fatah, "Analisis Data Mining Menggunakan Metode Clustering Terhadap Prestasi Siswa I'Dadiyah Sukorejo," *J. Ilm. Multidisiplin Ilmu*, vol. 2, no. 1, pp. 96–105, 2025, doi: 10.69714/remqnx91.

[11] Aditia Yudhistira and R. Andika, "Pengelompokan Data Nilai Siswa Menggunakan Metode K-Means Clustering," *J. Ris. Sist. Inf.*, vol. 1, no. 1, pp. 20–28, 2025, doi: 10.69714/0v1pkz05.

[12] M. Syaefudulloh, A. Faqih, and F. M. Basysyar, "Clustering Kelompok Belajar Siswa Berdasarkan Hasil Ujian Sekolah Menggunakan Algoritma K-Means," *J. Sist. Inf. dan Manaj.*, vol. Volume 10, no. 1, pp. 195–199, 2022.

[13] S. Suraya, M. Sholeh, and D. Andayati, "Penerapan Metode Clustering Dengan Algoritma K-Means Pada Pengelompokan Indeks Prestasi Akademik Mahasiswa," *Skanika*, vol. 6, no. 1, pp. 51–60, 2023, doi: 10.36080/skanika.v6i1.2982.

[14] E. H. K. Ramang, S. Lailiyah, and M. I. Sa'ad, "Implementation of Data Clustering for Informatics Engineering Study Program Students at STMIK Widya Cipta Dharma Using the K-Means Method," *Sebatik*, vol. 29, no. 1, pp. 1–12, 2025, doi: 10.46984/sebatik.v29i1.0000.

[15] M. S. Hasibuan, A. H. Lubis, and M. N. Sari, "Perbandingan algoritma clustering dbscan dan k-means dalam pengelompokan siswa terbaik," *INFOTECH J. Inform. Teknol.*, vol. 5, no. 2, pp. 301–309, 2024, doi: 10.37373/infotech.v5i2.1457.

[16] A. Syafrianto and E. Riswanto, "Pengelompokkan Jumlah Kunjungan Mahasiswa ke Perpustakaan Kampus Menggunakan Algoritma DBSCAN," *G-Tech J. Teknol. Terap.*, vol. 8, no. 1, pp. 75–81, 2023.

[17] M. Sholeh and K. Aeni, "Perbandingan Evaluasi Metode Davies Bouldin, Elbow dan Silhouette pada Model Clustering dengan Menggunakan Algoritma K-Means," *STRING (Satuan Tulisan Ris. dan Inov. Teknol.*, vol. 8, no. 1, 2023, doi: 10.30998/string.v8i1.16388.

[18] Y. Hasan, "Pengukuran Silhouette Score Dan Davies-Bouldin Index Pada Hasil Cluster K-Means Dan Dbscan," *KAKIFIKOM*, vol. 06, no. 01, 2024, doi: 10.23960/jitet.v12i3s1.5001.