

APPLICATION OF K-MEANS CLUSTERING FOR GROUPING STUDENT ACADEMIC PERFORMANCE IN A WEB-BASED DATA MINING SYSTEM

Eka Sri Rahayu^{1*}, Octaviana Anugrah Ade Purnama²

Universitas Pamulang / Tangerang Selatan

E-mail: dosen02839@unpam.ac.id^{1*},

dosen02780@unpam.ac.id²

Received: 13/07/2026 | Revised: 20/07/2026 | Accepted: 19/08/2026 | Published: 09/09/2026

Abstract

Improving the quality of education is one of the main priorities in managing academic institutions in the modern era. This effort is not only achieved through the improvement of facilities and infrastructure but also through the utilization of information technology to analyze academic data more comprehensively. One approach that can be applied to understand patterns of student achievement is data mining, which is the process of extracting meaningful information from large datasets to identify relevant patterns and relationships. This study aims to implement the K-Means Clustering method in the data mining process to classify students based on their levels of academic achievement. The K-Means Clustering method was selected because of its ability to efficiently group data based on the similarity of individual students' academic scores. Through this method, student scores from various subjects are analyzed and grouped into several categories, such as high, medium, and low academic achievement. The research stages include data collection, data cleaning and normalization, implementation of the K-Means algorithm, and evaluation of the clustering results. The results of this study are expected to provide a clearer overview of the distribution of students' academic achievement within the school environment. The information obtained can be utilized as a basis for strategic decision-making, such as designing more effective learning programs, providing specialized guidance for students with low achievement, and increasing motivation among high-achieving students. Therefore, this study is expected to make a meaningful contribution to the application of data mining technology in supporting data-driven educational quality improvement and promoting the optimal development of students' potential.

Keywords: Data Mining, K-Means Clustering, Student Achievement, Academic Data, Education.

INTRODUCTION

Education is a fundamental aspect of developing high-quality human resources. The success of the educational process can be assessed through various indicators, one of which is student academic achievement. Academic achievement reflects the extent to which students have mastered the learning materials provided. However, students may demonstrate different levels of achievement due to various factors, including cognitive ability, learning motivation, social environment, teaching methods, and family support. Therefore, schools need to understand the distribution of student achievement to support more appropriate and targeted educational strategies. One of the challenges faced by educational institutions is the suboptimal utilization of academic data for evaluating student achievement. Student grades are often used primarily for administrative purposes and are analyzed using conventional approaches, such as calculating average scores or conducting observations based on teacher assessments. Although these approaches provide general information, they may not reveal hidden patterns or similarities among students. Consequently, a systematic and data-driven approach is needed to process academic data into meaningful information that can support educational decision-making.

Data mining can be applied as an approach to identify meaningful patterns and information from large datasets. In the educational context, data mining can be used to analyze student academic data and support the evaluation of learning outcomes. Several data mining techniques, such as Decision Tree, Naive Bayes, and Neural Network, are commonly used for prediction and classification. However, these approaches generally

require predefined classes or labels. In contrast, this study focuses on identifying natural groupings within student academic data without predefined categories. Therefore, an unsupervised learning approach, particularly clustering, is considered more appropriate. One clustering method that can be applied to this problem is K-Means Clustering. K-Means is an unsupervised learning algorithm that groups data into a predetermined number of clusters based on the similarity between data points and their respective cluster centroids. The method is suitable for numerical academic data because student scores can be represented as numerical attributes. By applying K-Means, students can be grouped into categories representing different levels of academic achievement, such as high, medium, and low achievement, based on the characteristics of the analyzed data. This study aims to implement the K-Means Clustering method to group students according to their academic achievement. The research process consists of several stages, including data collection, data preprocessing, data cleaning, normalization, determination of the number of clusters, implementation of the K-Means algorithm, and evaluation of the clustering results. The quality of the resulting clusters can be evaluated using appropriate clustering evaluation measures, such as Sum of Squared Errors (SSE) or the Silhouette Coefficient. The clustering results can subsequently be interpreted to provide an overview of the distribution and characteristics of student academic achievement. The results of this study are expected to contribute to the application of data mining in the educational field, particularly in analyzing student academic achievement using a data-driven approach. The resulting student clusters can provide useful information for schools and teachers in designing more targeted learning programs, providing appropriate guidance to students with lower achievement, and developing strategies to support high-achieving students. Thus, the implementation of K-Means Clustering can help transform academic data into meaningful information that supports objective, measurable, and evidence-based educational decision-making.

LITERATURE REVIEW

Data mining is a computational approach used to discover meaningful patterns, relationships, and knowledge from large datasets. It combines statistical, mathematical, and machine learning techniques to support data analysis and decision-making. In education, Educational Data Mining (EDM) is applied to analyze academic data, identify learning patterns, and support educational evaluation. Clustering is a data mining technique that groups data into several clusters based on the similarity of their characteristics. Unlike classification, clustering does not require predefined class labels. In education, clustering can be used to identify groups of students with similar academic characteristics based on their scores across several subjects. K-Means Clustering is an algorithm used to group numerical data into a predetermined number of K clusters based on the distance between data points and centroids. Each data point is assigned to the nearest centroid, followed by centroid recalculation. This process is repeated until the clusters become stable or a predefined stopping condition is reached. The distance between data points and centroids is commonly calculated using Euclidean Distance:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

where x and y represent the data points being compared, and n represents the number of attributes. In this study, the attributes consist of student scores from several subjects. Student academic achievement represents the level of students' success in the learning process and can be measured through academic scores. Differences in student achievement can provide useful information for teachers and schools in developing appropriate learning strategies. However, analyzing scores individually may make it difficult to identify overall achievement patterns. Therefore, clustering can be used to analyze multiple academic attributes simultaneously. K-Means can provide a systematic approach to grouping students based on their academic characteristics. In this study, student scores from several subjects are processed through data preprocessing, including data cleaning and normalization, before clustering is performed. Normalization helps ensure that differences in attribute scales do not disproportionately affect distance calculations. The resulting clusters can represent groups of students with relatively high, medium, or low academic achievement. Although clustering has been widely applied to educational data, academic datasets may differ in terms of subjects, number of students, and academic periods. Therefore, K-Means needs to be adapted to the

characteristics of the dataset being analyzed, and the resulting clusters should be evaluated to ensure their meaningfulness. This study focuses on implementing K-Means Clustering to analyze patterns of student academic achievement based on scores from several subjects. The research aims to transform academic score data into meaningful student groups that can support academic evaluation and the development of more targeted educational programs.

METHOD

This study uses a descriptive quantitative approach to implement the K-Means Clustering algorithm for grouping students based on academic achievement. The research focuses on processing and analyzing numerical academic data to identify meaningful patterns and student groups.

1. Research Stages

The research consists of several main stages as follows:

a. Data Collection

The dataset consists of students' academic scores from several subjects. The data are organized in tabular form, with attributes such as student names, average scores, and other relevant academic achievement indicators. The dataset serves as the primary input for the clustering process.

b. Data Preprocessing

Data preprocessing is performed to ensure data quality before clustering. This stage includes data cleaning to remove invalid, duplicate, or missing data; data transformation to convert the data into numerical form and normalize the attributes; and data selection to select only relevant attributes for the clustering process.

c. Determining the Number of Clusters (K)

Before applying K-Means, the optimal number of clusters is determined using the Elbow Method. This method analyzes the change in the Sum of Squared Errors (SSE) for different numbers of clusters. The elbow point indicates the appropriate value of K for the clustering process.

d. K-Means Clustering

After determining the value of K, the data are grouped using the K-Means algorithm. The process consists of initializing the centroids, calculating the distance between each data point and the centroids using Euclidean Distance, assigning each data point to the nearest cluster, recalculating the cluster centroids, and repeating the process until the centroids become stable or the maximum number of iterations is reached.

e. Clustering Evaluation

The clustering results are evaluated to measure the quality of the generated groups. Evaluation can be performed using the Silhouette Coefficient, which measures how well each data point fits within its assigned cluster, and the Davies-Bouldin Index (DBI), which evaluates the similarity and separation between clusters. These measures are used to determine whether the resulting clusters are meaningful and well separated.

f. Interpretation and Visualization

The resulting clusters are analyzed and presented using visualizations such as scatter plots or bar charts. Each cluster is interpreted based on the characteristics of students' academic scores. The clusters may represent groups with relatively high, medium, or low academic achievement. The results can then be used to provide recommendations for educational evaluation and planning.

2. Web-Based Implementation

To support practical implementation, the K-Means algorithm is implemented through a web-based application using PHP and MySQL. The system allows users to upload academic data in CSV or Excel format, which is then processed to generate clustering results. The web interface is designed to be user-friendly by presenting the clustering results in the form of tables, charts, and reports. The system also provides an export feature so that the clustering results can be used for further analysis.

3. Research Results Analysis

The implementation of K-Means is expected to produce several student clusters representing different levels of academic achievement. For example, the clusters may consist of students with high and consistent achievement, moderate and stable achievement, and lower achievement requiring further improvement. The clustering results can provide educational institutions with information to support more targeted interventions. For example, students in the lower-achievement cluster may receive additional academic guidance, while high-achieving students may be provided with more challenging learning activities. Thus, the results of the clustering process can support more objective and data-driven educational decision-making.

RESULTS AND DISCUSSION

This section presents the results of implementing the K-Means Clustering algorithm to group students based on their academic achievement. The results include the clustering process, determination of the optimal number of clusters, evaluation of clustering quality, and interpretation of the resulting student groups. Tables and graphical visualizations are used to present the results clearly. The findings are discussed based on the characteristics of each cluster and their potential implications for academic evaluation and educational decision-making.

1. Determination of the Optimal Number of Clusters

The optimal number of clusters is determined using the Elbow Method by comparing the Sum of Squared Errors (SSE) values for different values of K. The SSE generally decreases as the number of clusters increases. The optimal K is selected at the point where the decrease in SSE begins to slow significantly.

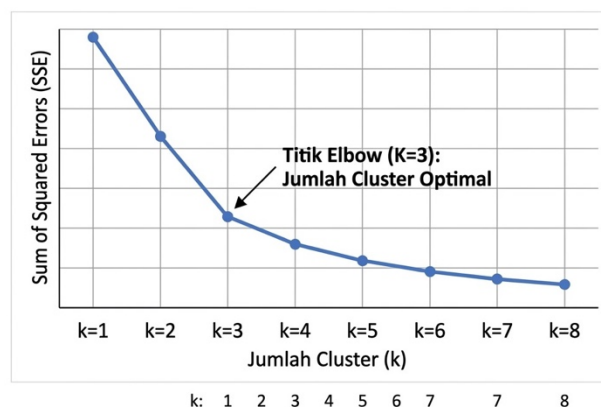


Figure 2. Elbow Method for Determining the Optimal Number of Clusters

The Elbow Method provides an objective basis for determining the number of clusters before applying K-Means. The selected value of K is then used in the subsequent clustering process. In this study, the selected clusters are interpreted according to the academic characteristics of the students in each group.

2. K-Means Clustering Results

After the optimal number of clusters is determined, the K-Means algorithm is applied to the preprocessed academic data. Each student is assigned to the cluster with the nearest centroid based on Euclidean Distance. The centroids are recalculated iteratively until the clustering process reaches a stable condition.

Table 1. Student Clustering Results

Cluster	Academic Achievement Characteristics	Number of Students
Cluster 1	High academic achievement	[n]
Cluster 2	Medium academic achievement	[n]
Cluster 3	Low academic achievement	[n]

The clustering results show that students can be grouped according to similarities in their academic scores. The interpretation of each cluster is based on the average score and characteristics of the students within the cluster. A cluster with a higher average score represents relatively higher academic achievement, while a cluster with a lower average score indicates students who may require additional academic support.

3. Clustering Evaluation

The quality of the resulting clusters is evaluated using the Silhouette Coefficient and Davies-Bouldin Index (DBI). The Silhouette Coefficient measures the similarity of data points within their assigned clusters and their separation from other clusters. A value closer to 1 indicates better-defined clusters. The Davies-Bouldin Index measures the similarity between clusters based on their compactness and separation. A lower DBI value indicates better clustering quality. The evaluation results can be presented as follows.

Table 2. Clustering Evaluation Results

Evaluation Metric	Result	Interpretation
Silhouette Coefficient	[value]	[Good/Moderate/Poor]
Davies-Bouldin Index	[value]	[Good/Moderate/Poor]

The evaluation results indicate the quality of the clusters generated by the K-Means algorithm. These evaluation measures provide additional evidence that the resulting groups are sufficiently distinct and can be interpreted based on their academic characteristics.

4. Visualization and Interpretation of Clusters

The clustering results can be visualized using scatter plots or bar charts to illustrate the distribution of students across the generated clusters.

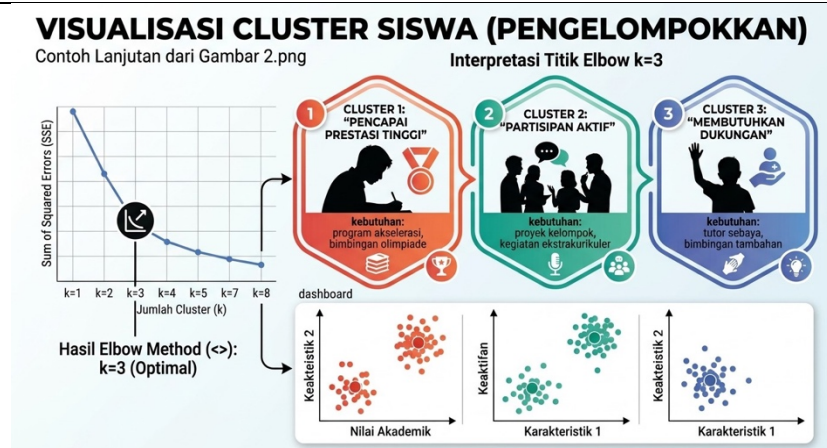


Figure 3. Visualization of Student Clusters

The visualization provides a clearer representation of the differences between student groups. Students belonging to the same cluster have relatively similar academic characteristics, while students in different clusters show greater differences in their academic scores. Based on the clustering characteristics, the groups can be interpreted as high-, medium-, and low-achievement groups. Students in the high-achievement group can be supported through enrichment activities, while students in the medium-achievement group can receive learning support to improve their performance. Students in the lower-achievement group may require more targeted academic guidance and additional learning assistance.

5. Web-Based Implementation

The K-Means clustering process is implemented through a web-based application using PHP and MySQL. The application allows users to upload academic data, perform data preprocessing, execute the clustering process, and display the resulting clusters.

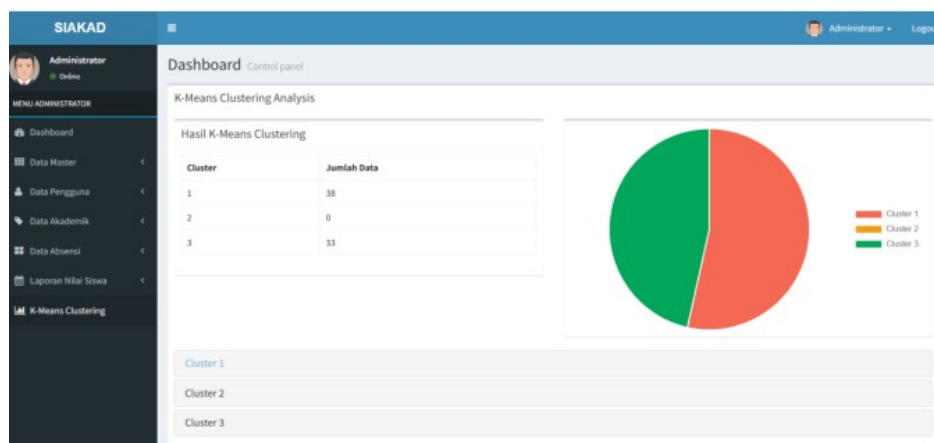


Figure 4. Web-Based K-Means Clustering Interface

The web-based implementation makes the clustering process more practical and accessible for educational users. The results are presented through tables and graphical visualizations, allowing users to identify student groups more easily. The export feature also allows the clustering results to be stored and used for further analysis or academic documentation.

6. Discussion

The implementation of K-Means Clustering demonstrates that academic score data can be transformed into meaningful student groups based on similarities in their academic characteristics. Unlike conventional analysis based only on average scores, clustering provides a broader view of the distribution of student achievement across several academic attributes. The resulting groups can support data-driven academic evaluation. By identifying groups with different achievement characteristics, educational institutions can develop more targeted strategies rather than applying the same intervention to all students. Therefore, K-Means Clustering has potential as a data mining approach for supporting academic evaluation and the development of appropriate learning strategies.

CONCLUSION

This study concludes that the K-Means Clustering algorithm can be implemented to group students based on similarities in their academic achievement. The research process consists of data collection, data preprocessing, attribute selection, normalization, determination of the optimal number of clusters using the Elbow Method, K-Means clustering, cluster evaluation, and interpretation of the resulting groups. The clustering results provide a clearer representation of student academic achievement by grouping students into categories with relatively high, medium, and low achievement characteristics. The evaluation using the Silhouette Coefficient and Davies-Bouldin Index can be used to assess the quality and separation of the resulting clusters.

The findings indicate that academic data mining can provide useful information for supporting data-driven educational decision-making. The resulting student groups can be used as a basis for developing more targeted learning programs, providing additional guidance to students who require improvement, and designing enrichment activities for high-achieving students. For future research, the study can be developed by using larger datasets, additional academic attributes, alternative clustering algorithms, or comparative evaluation with other clustering methods. The web-based implementation can also be further developed with additional analytical and reporting features to support academic evaluation more comprehensively.

REFERENCES

- Ariyanto, A., & Prasetyo, D. (2020). Implementation of data mining for student score clustering using the K-Means Clustering algorithm. *Jurnal Teknologi Informasi dan Komputer*, 6(2), 45–53.
- Badrul, M., & Ramadhan, F. (2022). Performance analysis of the K-Means and DBSCAN algorithms for academic data clustering. *Jurnal Ilmu Komputer dan Sistem Informasi*, 10(1), 14–22.
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3), 37–54.
- Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann Publishers.
- Hidayat, R., & Nugroho, D. (2021). Implementation of K-Means Clustering for web-based student learning achievement analysis. *Jurnal Teknologi dan Sistem Informasi*, 9(3), 120–128.
- Jain, A. K. (2010). Data clustering: 50 years beyond K-Means. *Pattern Recognition Letters*, 31(8), 651–666.
- Kumar, V., & Singh, S. (2021). Educational data mining: A review of the application of data mining techniques in education. *International Journal of Advanced Computer Science and Applications*, 12(4), 89–97.
- Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, 28(2), 129–137.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (Vol. 1, No. 14, pp. 281–297). University of California, Berkeley.
- Putra, M. A., Suryani, R., & Handoko, B. (2023). Application of K-Means Clustering in student academic achievement analysis to support educational decision-making. *Jurnal Teknologi Informasi dan Komunikasi*, 11(1), 34–42.
- Saputra, D., & Nataliani, Y. (2021). Application of data mining using the K-Means algorithm for student academic data clustering. *Jurnal Ilmiah Teknologi Informasi*, 7(2), 101–110.
- Setiawan, A. (2021). *Data mining: Concepts, methods, and applications in education*. Deepublish

APPLICATION OF K-MEANS CLUSTERING FOR GROUPING STUDENT ACADEMIC PERFORMANCE IN A WEB-BASED DATA MINING SYSTEM

Eka Sri Rahayu **et al**

Publisher.

Suryani, A., & Wibowo, T. (2020). Analysis of student score clustering using the K-Means Algorithm with Z-Score normalization. *Jurnal Ilmiah Informatika*, 8(4), 210–219.

Tan, P.-N., Steinbach, M., & Kumar, V. (2019). *Introduction to data mining* (2nd ed.). Pearson Education. Witten,

I. H., Frank, E., & Hall, M. A. (2017). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.