Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2827-2840
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5727

Regional Segmentation of School Dropouts Based on Economic and Accessibility Factors Using K-Means Clustering

Juna Eska*[1], Dinda Djesmedi[2], Yuhandri[3]

[1,2,3]Information System, Putra Indonesia University, Indonesia

Email: [1]dosen.junaeska@gmail.com

Received : Feb 7, 2026; Revised : Mar 15, 2026; Accepted : Mar 16, 2026; Published : Jun 15, 2026

Abstract

The high dropout rate in Asahan Regency has become a serious problem affecting the quality of human resources and equitable access to education across various regions. This study aims to identify patterns and characteristics of dropout-prone areas using the K-Means clustering technique. The research method involves collecting dropout data from the Asahan Regency Education Office for the period 2022–2025, followed by data pre-processing for cleaning and normalization, and then clustering analysis to generate three regional clusters based on dropout vulnerability levels. The results indicate that clusters with high dropout rates are largely influenced by economic factors, followed by limited access to education and social conditions in the community. The resulting regional segmentation provides a spatial overview of dropout vulnerability levels in Asahan Regency. These findings offer data-driven insights that can support the formulation of more targeted education policies and programs to encourage inclusive education development in the region. Scientifically, this study contributes to strengthening the validity and effectiveness of the K-Means algorithm as a quantitative approach in mapping and identifying complex patterns in socio-educational data, thereby expanding its application in data-driven analytical studies in the field of education.

Keywords *: Data mining, Dropout, K-Means Clustering, Regional Segmentation, Spatial Analysis.*

1. INTRODUCTION

Education is widely acknowledged as a fundamental pillar of socio-economic development, human capital formation, and long-term national competitiveness[1]. Empirical research consistently demonstrates that educational attainment is strongly associated with poverty reduction, labor productivity, and social mobility[2]. However, school dropout remains a persistent challenge in many developing countries, including Indonesia[3]. Although compulsory basic education policies have been implemented, regional disparities in dropout rates continue to exist[4]. In North Sumatra Province—particularly in Asahan Regency—dropout patterns vary considerably across sub-districts, reflecting differences in socio-economic status, school accessibility, and family awareness regarding the importance of education[5]. These disparities indicate that aggregated statistical reports are insufficient to fully capture the structural patterns of dropout vulnerability at the regional level[6].

Previous studies on school dropout in Indonesia and comparable contexts have primarily focused on identifying determinant factors using conventional statistical approaches such as linear regression, logistic regression, and descriptive correlation analysis[7]. These studies typically emphasize economic hardship, parental education, geographic isolation, or cultural factors as dominant predictors of dropout[8]. While such findings are valuable for understanding causal relationships, they often treat variables independently and rely on predefined dependent outcomes [9]. Consequently, they provide

Vol. 7, No. 3, June 2026, Page. 2827-2840
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5727

limited insight into how multiple vulnerability factors interact simultaneously across heterogeneous regions[10]. Moreover, most studies concentrate on individual student-level analysis rather than examining regional segmentation patterns that are directly relevant for public policy planning[11].

Recent advances in *Educational Data Mining* (EDM) and machine learning have introduced alternative analytical frameworks capable of uncovering hidden patterns in complex datasets [12]. *Supervised learning* models such as *Random Forests, Gradient Boosting, and Neural Networks* have shown strong predictive performance in dropout classification tasks [13]. However, these models require labeled data and are primarily designed to maximize prediction accuracy rather than reveal latent structural groupings[14]. In contrast, unsupervised learning methods—particularly clustering algorithms such as K-Means—enable the discovery of natural group structures within data without prior assumptions about dependent variables [15]. Clustering has been widely applied to student performance grouping and academic risk profiling, yet its application to regional dropout segmentation remains limited, especially in developing country contexts [2], [16].

Several recent studies have employed K-Means clustering to classify dropout risk among university or secondary school students based on academic performance indicators [17]. Others have combined clustering with decision tree methods to enhance classification interpretability [18]. Nevertheless, these studies largely focus on academic attributes and rarely integrate multidimensional socioeconomic and accessibility variables [19]. Furthermore, prior clustering research often stops at pattern identification without translating findings into a policy-oriented segmentation framework [20]. Thus, the existing literature reveals three major gaps: (1) limited regional-level dropout segmentation using unsupervised learning; (2) insufficient integration of economic, geographic, and family awareness factors within a single clustering model; and (3) lack of policy-driven interpretation that supports targeted educational intervention.

To address these gaps, this study applies the K-Means clustering algorithm to multidimensional dropout data from Asahan Regency, integrating socio-economic indicators, school accessibility measures, and parental awareness variables. Unlike previous regression-based or student-level clustering studies, this research focuses on regional segmentation to identify *high-, moderate-, and low-risk* dropout clusters across sub-districts[21]. By uncovering latent vulnerability structures, the proposed approach provides a more comprehensive and policy-relevant analytical framework for educational authorities[22]. This study therefore contributes to the state of the art in educational data mining by extending clustering applications from individual academic profiling to regional education policy analytics, offering a replicable model for other districts facing similar dropout challenges [23].

2. METHOD

This chapter discusses the research methods used to cluster data on school dropouts at the Asahan Regency Education Office using the K-Means algorithm. This study employed a quantitative descriptive approach with data mining techniques. The data consisted of elementary school dropouts and grade repetitions in Asahan between 2022 and 2025, obtained from the Asahan Regency Education Office. Evaluation was conducted using Euclidean Distance, two criteria, the Davies-Bouldin Index (DBI), and three cluster groupings.

This study integrated data collection, interview instruments, data preprocessing, K determination, centroid initialization, distance calculation, clustering, centroid updating, and convergence checking.

After pre-processing, the next step is clustering using the K-Means Clustering algorithm. This step identifies natural patterns in student anxiety. The data is grouped into three clusters: “C1” = Low, “C2” = Medium, and “C3” = “High”. The number of clusters is determined based on conceptual considerations and policy analysis needs. After the number of clusters is determined, the next step is to randomly determine the initial centroids from the available data[24]. The centroid is the center point of

Vol. 7, No. 3, June 2026, Page. 2827-2840
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5727

each cluster, which serves as a reference in calculating the distance of each data object. Each data item is calculated for its distance from each centroid using the *Euclidean Distance* formula. After all distances are calculated, each object will be assigned to the cluster with the closest distance (minimum value). After all data is clustered, the centroids are updated by calculating the average of all members in each cluster. This new centroid will be reused in the next iteration process. The final step is to check whether there is a change between the old and new centroids. However, if there is still a significant difference, the process returns to the distance calculation stage until convergence is achieved[25].

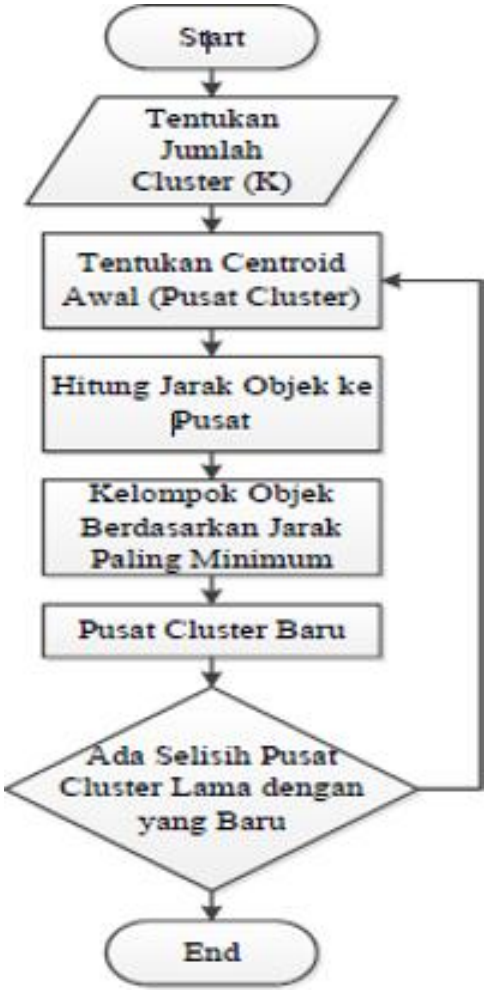

Figure 1. illustrates the research flow diagram

2.1. Research Materials

The data in this study were obtained from the Asahan Regency Education Office for the period 2022–2025 which can be seen in Table 1. The data used include the number of students dropping out of school, the main causal factors, and geographical locations as shown in Table 2 below.

Table 1. Summary of Basic Education Data

Year	Status	
	Dropouts	Repeating
2022/2023	48	428
2023/2024	42	430
2024/2025	45	453

Vol. 7, No. 3, June 2026, Page. 2827-2840
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5727

Table 2. Summary of Sub-district Name Data

Code	Alternative Names	Drop-Out Students	Repeating Students
J01	Aek Kuasan	0	1
J02	Aek Ledong	0	7
J03	Aek Songsongan	5	25
J04	Air Batu	0	17
J05	Air Joman	0	7
J06	Bandar Pasir Mandoge	3	35
J07	Bandar Pulau	7	37
J08	Buntu Pane	0	4
J09	Kisaran Barat	0	6
J10	Kisaran Timur	3	22
J11	Meranti	0	7
J12	Pulau Rakyat	3	28
J13	Pulo Bandring	1	12
J14	Rahuning	3	18
J15	Rawang Panca Arga	0	3
J16	Sei dadap	1	10
J17	Sei Kepayang	1	15
J18	Sei Kepayang Barat	1	5
J19	Sei Kepayang Timur	4	20
J20	Setia Janji	1	19
J21	Silau Laut	0	0
J22	Simpang Empat	7	38
J23	Tanjung Balai	3	33
J24	Teluk Dalam	3	19
J25	Tinggi Raja	1	25

2.2. Data Collection Techniques

In this study, data were collected through various information gathering techniques, including interviews, direct observation, and literature review. Interviews were conducted with relevant parties, such as representatives from the Education Office and educators, to obtain information on factors influencing school dropout rates. Direct observations were conducted by visiting the research locations to understand the social and economic conditions in areas experiencing high dropout rates. Additionally, literature review was conducted by reviewing various academic references, official reports, and statistical documents to support data analysis[26].

2.3. K-Means Clustering

Data processing is conducted to ensure the dataset is suitable for clustering using the K-Means algorithm. This stage includes data transformation, initial centroid determination, distance calculation, cluster formation, and convergence evaluation.

1. Conversion of Nominal Data into Numeric Values
   Some dataset attributes are categorical or nominal and therefore cannot be directly used in mathematical calculations. Consequently, data are converted into numeric form using label encoding. This conversion enables all variables to be processed mathematically within the K-Means algorithm.

2. Determination of Number of Clusters (K)
   The number of clusters is defined based on the research objective, which is to classify regions according to dropout vulnerability levels:
   - Cluster 1: High vulnerability
   - Cluster 2: Moderate vulnerability
   - Cluster 3: Low vulnerability

   Thus, the number of clusters used is: ***K=3***
3. Initialization of Initial Centroids
   Initial centroids are randomly selected from the available dataset. Suppose the dataset consists of data points:

$$x = (x_1, x_2, \dots, x_n) \quad (1)$$

   Three data points are randomly selected as initial centroids:
   C1,C2,C3
   Each centroid contains values for all attributes used in clustering.

4. Distance Calculation Using Euclidean Distance
   The distance between each data point and each centroid is calculated using the ***Euclidean Distance*** formula:

$$d(\mathbf{x}, \mathbf{c}) = \sqrt{\sum_{i=1}^{m}(x_i - c_i)^2} \quad (2)$$

   Where:
   xi = value of the data point on attribute i
   ci = value of the centroid on attribute i
   m = number of attributes
   d = distance between data point and centroid
   Each data point is assigned to the cluster with the smallest distance.

5. After assigning all data points to clusters, centroids are recalculated as the average of all members in each cluster:

$$c_j = \frac{1}{N_j}\sum_{i=1}^{N_j} x_i \quad (3)$$

   Where:
   cj = updated centroid of cluster j
   Nj = number of data points in cluster j
   xi = data points in cluster j
   This step generates new centroid positions.

6. Iterative Clustering Process
   The following steps are repeated:
   1. Compute distances between data points and new centroids.
   2. Reassign data points to clusters.
   3. Recalculate centroids.

   This process continues until stability is achieved.

Vol. 7, No. 3, June 2026, Page. 2827-2840
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5727

7. Stopping Criteria Based on Centroid Change Ratio

$$R = \frac{|C_{new} - C_{old}|}{|C_{old}|} < \varepsilon \quad (4)$$

2.4. Model Validation

Once the clustering process for each k is complete, the Davies-Bouldin Index (DBI) can be used to determine the optimal number of clusters. This measurement approach aims to maximize the distance between clusters while simultaneously minimizing the distance between objects within a cluster. The clustering with the optimal number of clusters is the one with the minimum DBI value.

The optimal number of clusters was tested by calculating the average Silhouette Score value for several candidate number of clusters, namely K = 2, 3, 4, and 5. The Silhouette Score graph against the number of clusters shows that the highest value was obtained at K = 3 with a score of 0.57.

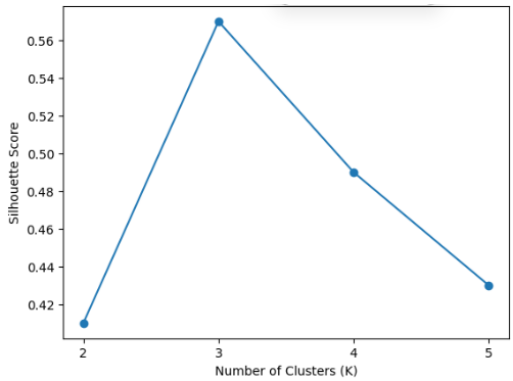

Figure 2. Silhouette Score Curve for Determining the Optimal Number of Clusters

The significant increase from K = 2 to K = 3 indicates that the formation of three clusters resulted in better separation between groups and greater cohesiveness within each cluster. However, after K > 3, the Silhouette Score decreased, indicating that increasing the number of clusters did not improve segmentation quality and, in fact, tended to cause data fragmentation. Methodologically, the Silhouette Score of 0.57 qualifies as a reasonable cluster structure, thus concluding that K = 3 is the optimal number of clusters for segmenting dropout-prone areas in this study.

3. RESULT

3.1. Research Data Description

The Dataset Used In This Study Consists Of Two Main Variables, Namely The Number Of Students Dropping Out Of School And The Number Of Students Repeating A Grade Across 25 Districts. These Data Were Used To Identify Patterns In The Distribution Of Educational Problems Using The K-Means Clustering Method.

In Aggregate, The Educational Data Show That The Number Of Students Repeating A Grade Is Relatively Higher Than The Number Of Students Dropping Out Of School. In The 2022/2023 Academic Year, There Were 48 Students Who Dropped Out And 428 Students Who Repeated A Grade. In The 2023/2024 Academic Year, There Were 42 Students Who Dropped Out And 430 Students Who Repeated A Grade, While In The 2024/2025 Academic Year There Were 45 Students Who Dropped Out And 453 Students Who Repeated A Grade. This Condition Indicates That The Phenomenon Of Grade Repetition Remains An Important Indicator Of Educational Problems. The Dataset Was Then

Vol. 7, No. 3, June 2026, Page. 2827-2840
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5727

Analyzed Using The K-Means Method To Group Districts Based On The Level Of Educational Problems.

3.2. Determination of the Number of Clusters Using the Elbow Method

Before performing the clustering process, the optimal number of clusters was determined using the Elbow Method. This method evaluates the Sum of Squared Error (SSE) value for different numbers of clusters.

Table 3. SSE Value Table

K	SSE
1	3370.88
2	1038.74
3	381.67
4	273.23
5	145.61
6	137.83

The Elbow Method graph is shown in the following figure.

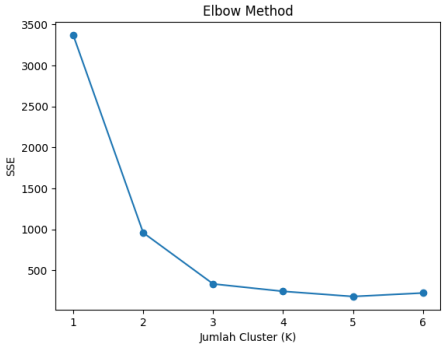

Figure 3. Elbow Method Graph

The Elbow graph shows that the most significant decrease in SSE occurs up to K = 3. After this point, the decrease in SSE becomes relatively small, forming an elbow pattern. Therefore, the optimal number of clusters used in this study is three clusters.

3.3. K-Means Iteration Process

After determining the number of clusters, the clustering process was carried out using the K-Means algorithm. This process was performed iteratively until the centroid values no longer changed significantly.

Table 4. SSE Value for Each Iteration

Iteration	SSE
1	2764
2	1009.03
3	723.76
4	419.44
5	346.61
6	335.48

Vol. 7, No. 3, June 2026, Page. 2827-2840
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5727

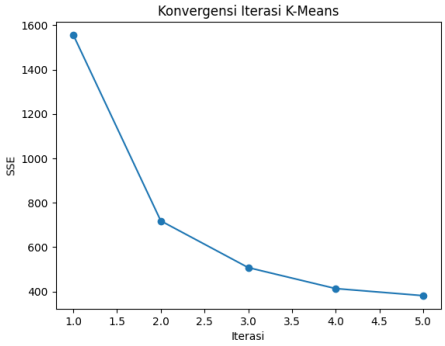

Figure 4. K-Means Iteration Convergence Graph

The convergence graph shows that the SSE value decreases significantly in the early iterations and begins to stabilize after the 4th to 5th iteration. This indicates that the K-Means algorithm has reached a convergent condition, so the resulting centroid values can be considered stable.

3.4. District Clustering Results

Based on the clustering process using K-Means with K = 3, the districts were grouped as follows.

Table 5. Cluster Result Table

District	Dropout	Repeating	Cluster
Aek Kuasan	0	1	C1
Aek Ledong	0	7	C1
Aek Songsongan	5	25	C3
Air Batu	0	17	C2
Air Joman	0	7	C1
Bandar Pasir Mandoge	3	35	C3
Bandar Pulau	7	37	C3
Buntu Pane	0	4	C1
Kisaran Barat	0	6	C1
Kisaran Timur	3	22	C2
Meranti	0	7	C1
Pulau Rakyat	3	28	C3
Pulo Bandring	1	12	C2
Rahuning	3	18	C2
Rawang Panca Arga	0	3	C1
Sei Dadap	1	10	C1
Sei Kepayang	1	15	C2
Sei Kepayang Barat	1	5	C1
Sei Kepayang Timur	4	20	C2
Setia Janji	1	19	C2
Silau Laut	0	0	C1
Simpang Empat	7	38	C3
Tanjung Balai	3	33	C3
Teluk Dalam	3	19	C2
Tinggi Raja	1	25	C3

Jurnal Teknik Informatika (JUTIF)
P-ISSN: 2723-3863
E-ISSN: 2723-3871
Vol. 7, No. 3, June 2026, Page. 2827-2840
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5727

3.5. Cluster Centroids

The final centroid values for each cluster are shown below.

Table 6. Cluster Centroid

Cluster	Dropout	Repeating
C1	0.20	5.00
C2	2.00	17.75
C3	4.14	31.57

The centroid interpretation indicates that:

- Cluster 1 represents areas with a low level of educational problems.
- Cluster 2 represents areas with a moderate level of educational problems.
- Cluster 3 represents areas with a high level of educational problems.

3.6. Cluster Visualization

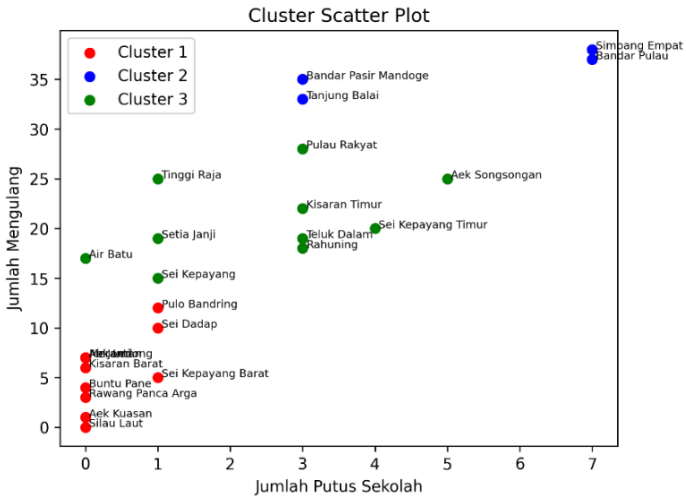

Figure 5. Scatter Plot of Clustering Results

To facilitate interpretation of the clustering results, visualization was performed using a scatter plot with two variables: the number of student dropouts and the number of students repeating a grade.

The scatter plot shows the distribution of data based on the number of student dropouts and the number of students repeating a grade in each district. The graph indicates that several districts, such as Simpang Empat and Bandar Pulau, have relatively higher repetition rates compared to other districts.

3.7. Clustering Evaluation Using Silhouette Score

After the clustering process using the K-Means algorithm was completed, the next step was to evaluate the quality of the resulting clusters. One of the commonly used evaluation methods in clustering analysis is the Silhouette Score.

The Silhouette Score is calculated using the following equation:

$$S(i) = \frac{b(i)-a(i)}{\max\{a(i), b(i)\}} \qquad (5)$$

where:

- a(i) = the average distance between data point *i* and all other data points in the same cluster
- b(i) = the average distance between data point *i* and all data points in the nearest neighboring cluster

Vol. 7, No. 3, June 2026, Page. 2827-2840
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5727

- S(i) = silhouette value of data point *i*

The Silhouette value ranges between:

$$-1 \leq S(i) \leq 1$$

To evaluate the clustering quality, the Silhouette Score was calculated, resulting in:

Silhouette Score=0.569

This value indicates that the cluster structure has reasonable clustering quality, meaning that the K-Means method is able to separate the data relatively well.

3.8. Result Analysis

Based on the clustering results, most districts fall into clusters with low to moderate levels of educational problems. However, several districts fall into the cluster with a high level of educational problems, such as Bandar Pulau, Simpang Empat, Bandar Pasir Mandoge, and Tanjung Balai.

These areas have relatively high numbers of students repeating grades and therefore can be prioritized in educational policy planning. Intervention programs such as improving learning quality, monitoring student progress, and implementing dropout prevention programs can be focused on these areas.

The results of this study can be seen in table 7 below, which shows that the K-Means Clustering method is able to identify distribution patterns of educational problems based on region effectively, so that it can help policy makers in determining more targeted handling strategies.

Table 7. Cluster Member Grouping

Cluster	Sub-district	Description
C1	Aek Kuasan, Buntu Pane, Kisaran Barat, Rawang Panca Arga, Sei Kepayang Barat, Silau Laut	Rarely Occurs
C2	Aek Ledong, Air Batu, Air Joman, Meranti, Pulo Bandring, Rahuning, Sei Dadap, Sei Kepayang	Fairly Frequently Occurs
C3	Aek Songsongan, Bandar Pasir Mandoge, Bandar Pulau, Kisaran Timur, Pulau Rakyat, Sei Kepayang Timur, Setia Janji, Simpang Empat, Tanjung Balai, Teluk Dalam, dan Tinggi Raja.	Very Frequently Occurs

4. DISCUSSION

4.1. INTERPRETATION OF RESEARCH FINDINGS

The results of this study indicate that the implementation of the *K-Means* algorithm successfully grouped the school dropout data into three main clusters representing low, medium, and high risk levels. This segmentation pattern suggests that the variables *dropouts* and *repeating* play an important role in shaping educational regions with different characteristics.

These findings demonstrate that clustering approaches are capable of revealing hidden patterns within educational datasets that are often difficult to detect using conventional statistical analysis. In the context of educational data analysis, clustering techniques enable researchers to identify groups of students or regions that share similar educational characteristics, which can serve as a basis for data-driven educational policy formulation.

Vol. 7, No. 3, June 2026, Page. 2827-2840
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5727

4.2. Comparison with Previous Studies

The findings of this study are consistent with several recent studies in the field of *Educational Data Mining,* which demonstrate that clustering techniques are effective methods for analyzing educational data.

Previous studies have shown that the K-Means algorithm can cluster student academic performance data into several groups based on learning characteristics. For example, a study analyzing student academic achievement data using K-Means demonstrated that the algorithm can form clusters of students based on similarities in academic performance, making it easier to identify groups with high and low academic outcomes[27].

Furthermore, other studies applying the *Knowledge Discovery in Databases (KDD)* framework in educational datasets have shown that K-Means can produce optimal segmentation results when evaluated using cluster validation indices such as the *Davies–Bouldin Index*. These studies indicate that the resulting clusters are capable of representing significant differences in student characteristics based on academic indicators and attendance data.

In the context of school dropout analysis, previous research also found that the K-Means algorithm effectively identifies regions with high and low dropout rates. In that study, the clustering results achieved a *Silhouette Score of 0.722*, indicating a relatively strong separation between clusters.

Additionally, recent systematic studies in educational data mining highlight that clustering methods play an important role in identifying students who are at risk of dropping out or experiencing academic difficulties. Clustering enables educational institutions to group students based on learning patterns, academic performance, and socioeconomic factors, allowing educational interventions to be implemented more effectively[28].

Therefore, the results of this study support previous findings indicating that clustering methods, particularly K-Means, represent an effective approach for analyzing educational datasets and mapping educational challenges in a structured manner.

4.3. Methodological Argument: K-Means vs Other Clustering Methods

From a methodological perspective, the use of the K-Means algorithm in this study is based on several advantages compared to other clustering methods.

First, K-Means is an *unsupervised learning* algorithm with high computational efficiency and the ability to process relatively large datasets quickly. This advantage makes K-Means one of the most widely used algorithms in data mining and educational data analysis.

Second, K-Means employs a simple yet effective iterative mechanism to minimize the distance between data objects within clusters using centroid-based grouping. This approach enables the algorithm to group objects according to similarity efficiently, producing clusters that are relatively easy to interpret.

Several comparative studies have also evaluated K-Means against other clustering methods such as *Agglomerative Hierarchical Clustering (AHC).* These studies report that K-Means often produces higher *Silhouette Score* values compared to hierarchical clustering in certain educational datasets, indicating better cluster separation[29].

In addition, comparisons with algorithms such as *DBSCAN* have shown that K-Means performs better in separating clusters when the dataset has a relatively homogeneous distribution.

However, some recent studies suggest that more advanced clustering methods, *including deep learning–based clustering approaches*, may produce higher predictive performance in dropout risk analysis. Nevertheless, these methods typically require more complex computational resources and larger datasets[30].

Vol. 7, No. 3, June 2026, Page. 2827-2840
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5727

Therefore, considering the characteristics of the dataset used in this study, which contains a limited number of variables, the K-Means algorithm is considered an appropriate method because it can generate clear data segmentation with relatively low computational complexity.

4.4. Scientific Implications for Educational Data Mining

This study provides an important contribution to the development of analytical approaches within the domain of *Educational Data Mining,* particularly in the application of clustering techniques for regional educational data analysis.

Conceptually, this research demonstrates that data mining approaches can be used to identify patterns in educational problems more systematically. By clustering regions based on dropout levels, this study highlights patterns of educational inequality that may not be easily observed using conventional statistical analysis.

Furthermore, this research strengthens the role of *unsupervised learning techniques* in educational policy analysis. The segmentation of educational regions into clusters allows policymakers and educational institutions to design more targeted interventions by focusing on regions with the highest dropout risk[31]

Thus, this study contributes to expanding the application of data mining methods in supporting evidence-based decision-making within the education sector.

4.5. Research Limitations

Despite demonstrating the effectiveness of the K-Means algorithm in clustering school dropout data, several limitations should be acknowledged.

First, this study used a limited number of variables, focusing only on two educational indicators: *dropouts* and *repeating*. In practice, school dropout phenomena are influenced by multiple factors, including socioeconomic conditions, accessibility of educational facilities, transportation, and family background.

Second, the K-Means algorithm assumes that clusters have relatively spherical shapes and homogeneous distributions. If the dataset contains more complex cluster structures or significant outliers, the clustering performance of K-Means may be reduced.

Third, this study applied only one clustering method without conducting a broader comparative experiment with other clustering algorithms.

4.6. Future Research

Future research may extend this study by incorporating additional educational variables such as poverty levels, school accessibility, student–teacher ratios, and other socioeconomic indicators.

In addition, future studies could conduct comparative analyses between multiple clustering algorithms, including *DBSCAN, Hierarchical Clustering, and Fuzzy C-Means*, in order to identify the most suitable approach for analyzing educational datasets.

Further research may also integrate clustering techniques with classification or predictive models such as *Random Forest, Gradient Boosting, or Neural Networks,* enabling not only the identification of cluster patterns but also the prediction of potential dropout risks with greater accuracy

5. CONCLUSION

This study applied the *K-Means* algorithm to analyze regional patterns of school dropout data. The results show that the algorithm successfully grouped the dataset into three distinct clusters representing low, moderate, and high dropout risk levels. The clustering process converged efficiently within a limited number of iterations, indicating that the dataset contains identifiable patterns that can be effectively segmented using centroid-based clustering methods. The resulting clusters reveal clear

Vol. 7, No. 3, June 2026, Page. 2827-2840
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5727

differences in the characteristics of dropout and repetition indicators, demonstrating that clustering techniques can provide meaningful insights into the distribution of educational challenges across regions.

From a scientific perspective, this research contributes to the development of analytical approaches in the field of *Educational Data Mining* by demonstrating the effectiveness of unsupervised learning techniques for analyzing educational datasets. The findings highlight the capability of clustering algorithms to uncover hidden patterns in educational data and provide structured segmentation that supports deeper interpretation of educational issues. By applying K-Means to regional dropout data, this study expands the application of data mining techniques in educational research and demonstrates how data-driven approaches can enhance the understanding of complex educational phenomena.

In terms of practical implications, the clustering results provide valuable insights for educational policymakers and institutions. By identifying regions that belong to clusters with higher dropout risk, policymakers can design targeted intervention programs and allocate educational resources more effectively. Such data-driven policy strategies can support early intervention efforts, improve educational planning, and ultimately contribute to reducing school dropout rates. Therefore, integrating data mining techniques into educational policy analysis can play an important role in supporting evidence-based decision-making in the education sector.

REFERENCES

[1] U. Arfan and Y. Pekei, "An Evaluation of Teacher – Student Distribution and School Availability in Supporting Educational Development Using the K-Means Clustering Algorithm Evaluasi Distribusi Guru-Siswa dan Ketersediaan Sekolah untuk Mendukung Pembangunan Pendidikan Menggunakan K-Means Clustering," vol. 5, no. October, pp. 1508–1516, 2025.

[2] P. R. Alvarez-p, M. Olga, and P. A. Toledo-delgado, "Acta Psychologica Dropping out of higher education : Analysis of variables that characterise students who interrupt their studies," vol. 252, no. December 2024, pp. 1–7, 2025, doi: 10.1016/j.actpsy.2024.104669.

[3] S. Sukriadi and E. W. Mawarni, "Analisis Faktor Penyebab Anak Putus Sekolah : Studi Kualitatif di Dusun Lestari Setia , Kecamatan Sokan Kabupaten Melawi," vol. 5, no. 2, pp. 1143–1152, 2025.

[4] E. Nurkhofifah, D. Athina, A. R. Tarida, and F. A. Pratiwi, "Clustering of Junior High School Education in West Java Based on Density and Dropout Ratios Using Quartile and K- Means Methods," pp. 483–511.

[5] M. Vaarma and H. Li, "Technology in Society Predicting student dropouts with machine learning : An empirical study in Finnish higher education," *Technol. Soc.*, vol. 76, no. December 2023, p. 102474, 2024, doi: 10.1016/j.techsoc.2024.102474.

[6] Y. Syahra *et al.*, "Customer Segmentation Using RFM and K- Means Clustering to Support CRM in Retail Industry," vol. 9, no. 3, pp. 1120–1131, 2025.

[7] D. Sugiarto and R. Fitriana, "BUSINESS INTELLIGENCE IN VEGETABLE ONLINE RETAILING," vol. 35, no. August, pp. 118–135, 2025.

[8] A. Selina *et al.*, "Permasalahan dalam Implementasi Pendidikan Inklusi di Sumatera Utara," vol. 9, pp. 8876–8883, 2025.

[9] J. T. Edward Revaldo Danuwinata1, "INFORMASI (Jurnal Informatika dan Sistem Informasi) Volume 17 No.2 / Nov / 2025," vol. 17, no. 2, pp. 245–264, 2025.

[10] H. Tan, "Dropout in Online Education : A Longitudinal Multilevel Analysis of Elementary Students ' Extracurricular English Course Engagement in China," vol. 12, 2025.

[11] M. Psyridou, F. Prezja, M. Torppa, M. Lerkkanen, and K. Vasalampi, "Machine Learning Predicts Upper Secondary Education Dropout as Early as the End of Primary," pp. 1–14.

[12] D. A. Imanuel, G. Alfian, S. Vokasi, U. G. Mada, P. Korespondensi, and C. Index, "VISUALISASI SEGMENTASI PELANGGAN BERDASARKAN ATRIBUT RFM MENGGUNAKAN ALGORITMA K-MEANS UNTUK MEMAHAMI KARAKTERISTIK

Vol. 7, No. 3, June 2026, Page. 2827-2840
https://jutif.if.unsoed.ac.id
DOI: https://doi.org/10.52436/1.jutif.2026.7.3.5727

PELANGGAN PADA TOKO RETAIL ONLINE VISUALIZATION OF CUSTOMER SEGMENTATION BASED ON RFM ATTRIBUTES USING K-MEANS ALGORITHM TO COMPREHEND CUSTOMER CHARACTERISTICS WITHIN AN ONLINE RETAIL STORE," vol. 12, no. 2, pp. 283–292, 2025, doi: 10.25126/jtiik.2025128619.

[13] M. Hammam and T. Utomo, "CLUSTERING DATA SISWA PUTUS SEKOLAH DENGAN ALGORITMA K-MEANS DAN DBSCAN," vol. 18, no. 2, pp. 150–159, 2023.

[14] H. Safitri, S. Putri, L. Geni, F. Merry, and M. Wati, "Penerapan K-Means Clustering untuk Segmentasi Konsumen E-Commerce Berdasarkan Pola Pembelian," vol. 7, pp. 89–99, 2025.

[15] M. R. Hardyanto, A. O. Prameswari, and M. Rizky, "Analisis Segmentasi Pelanggan Dalam Preferensi Pembelian Produk Menggunakan Metode K-Means : Studi Kasus Industri Sepatu," vol. 03, no. 11, pp. 1780–1792, 2024.

[16] N. H. Ahsina, F. Fatimah, and F. Rachmawati, "BERDASARKAN PENGAMBILAN KREDIT DENGAN MENGGUNAKAN METODE K-MEANS CLUSTERING," vol. 8, no. 3, 2022.

[17] T. C. Saputra, S. M. Fadhilah, and S. U. Mangkuto, "Segmentation , targeting and positioning analysis using k- means clustering model : A case study of the laptop market in Indonesia," vol. 12, no. 2, pp. 195–203, 2024.

[18] C. H. Ardana, A. Aldita, A. Aisyah, A. Khoyum, and M. Faisal, "Segmentasi Pelanggan Penjualan Online Menggunakan Metode K- means Clustering," vol. 9, no. 1, pp. 1–9, 2024.

[19] S. Anwar, "Application of K-Means and C4 . 5 Algorithms for dropout risk prediction in vocational high schools," 2024.

[20] K. Tauhid *et al.*, "KAJIAN UNTUK SEGMENTASI CUSTOMER BANK DENGAN ALGORITMA K-MEANS," vol. 3, pp. 3899–3906, 2024.

[21] D. Methods, R. Wati, B. Sembiring, F. A. Mohammed, and K. Chairuang, "Customer Segmentation Based on RFM Model Using," vol. 11, no. 1, pp. 32–43, 2020.

[22] N. T. Phudjashakty *et al.*, "Segmentasi Pelanggan Grosir Menggunakan K-Means : Analisis Outlier dan Ketidakseimbangan Data," vol. 4, no. 3, pp. 15993–16002, 2026.

[23] E. Omol, D. Onyangor, L. Mburu, and P. Abuonji, "Application Of K-Means Clustering For Customer Segmentation In Grocery Stores In Kenya," pp. 192–200.

[24] J. Chitra and J. Heikal, "Customer segmentation using the K-Means Clustering algorithm in Foreign Banks in Indonesia," vol. 11, no. 4, pp. 230–241, 2024.

[25] E. Nurmiati and A. A. Nurindah, "Segmentasi Customer Pada Industri Ritel Menggunakan Teknik Clustering," vol. 5, no. 1, pp. 344–353, 2025.

[26] E. Gari and R. Sari, "Analisis Segmentasi Wilayah Penjualan menggunakan Algoritma K-Means Clustering," vol. 7, no. 2, pp. 145–151, 2025.

[27] A. Akram, N. Risal, D. D. Andayani, M. I. Suherman, and A. B. Kaswar, "Utilizing the K-Means Clustering Algorithm for Analyzing Student Achievement Assessment at SMK Negeri 1 Gowa," vol. 05, no. March, pp. 60–67, 2024.

[28] Y. Lu, S. Yeom, J. Maktoubian, and M. M. Rahman, "Improve Student Risk Prediction with Clustering Techniques : A Systematic Review in Education Data Mining," vol. 1, pp. 1–38, 2025.

[29] S. Lestari, "Assessment Clusterization Teacher Performance with K-Means Algorithm Clustering and Agglomerative Hierarchical Clustering (AHC)," vol. 9, no. 1, pp. 357–365, 2025.

[30] P. G. Almeida, G. A. L. Silva, V. Santos, G. Moreira, P. Silva, and E. Luz, "P REDICTING A T -R ISK S TUDENTS," no. Dl, 2024.

[31] M. F. Fadhillah, A. Lovely, A. Suyoso, and I. Puspitasari, "Customer Segmentation with Clustering Algorithm Based on Recency , Frequency , and Monetary (RFM) Attributes Segmentasi Pelanggan dengan Algoritma Clustering Berdasarkan Atribut Recency , Frequency dan Monetary (RFM)," vol. 5, no. January, pp. 48–56, 2025.