# Mapping and Clustering of Educational Facility Coverage across Regencies and Municipalities in North Sumatra Province Using K-Means in 2025

**Rahmat Hidayat, Eko Hariyanto, Solly Aryza**

**Abstract**

This study aims to identify the characteristics and cluster regencies and municipalities in North Sumatra Province based on educational facility coverage in 2025. The study uses secondary data from Statistics Indonesia (BPS), covering 33 regencies and municipalities and four variables: Elementary School (SD), Junior High School (SMP), Senior High School (SMA), and Vocational High School (SMK) facility coverage. The research applies the K-Means algorithm through several stages, including calculating educational facility coverage percentages, data standardization using Z-Score, determining the optimal number of clusters through the Elbow Method and Silhouette Score, and evaluating clustering performance using Inertia, Silhouette Score, Davies-Bouldin Index, and Calinski-Harabasz Score. The evaluation results indicate that K=2 is the optimal number of clusters, achieving a Silhouette Score of 0.4987. The clustering process produces Cluster 1, consisting of 22 regencies and municipalities (66.67%), and Cluster 2, consisting of 11 regions (33.33%). Cluster 2 demonstrates higher average educational facility coverage than Cluster 1 across all education levels. These findings indicate that K-Means can identify differences in educational facility coverage among regions. The results are expected to provide useful information for supporting educational facility distribution planning and promoting more equitable access to education in North Sumatra Province.

***Keywords**: Educational Facilities, K-Means, Clustering, North Sumatra.*

Rahmat Hidayat
Magister Teknologi Informasi, Universitas Pembangunan Panca Budi, Indonesia
e-mail: 92.rahmathidayat@gmail.com

Eko Hariyanto, Solly Aryza
e-mail: eko.hariyanto@dosen.pancabudi.ac.id, sollyaryzalubis@gmail.com
3rd International Conference Epicentrum of Economic Global Framework (ICEEGLOF)
Theme: Cross-Sector and Cross-Disciplinary Strategies for Achieving the SDGS
https://proceeding.pancabudi.ac.id/index.php/ICEEGLOF

## Introduction

Education is one of the key factors in human resource development and the improvement of societal well-being. The availability of educational facilities in a region is an important aspect to consider, as it is closely related to the equitable provision of educational services. However, the availability of educational facilities may vary across regencies and cities, particularly between primary and secondary education levels. Law No. 20 of 2003 concerning the National Education System emphasizes the importance of providing equitable and fair access to education [1]. In line with this, the Sustainable Development Goals (SDGs), particularly Goal 4, highlight the importance of ensuring inclusive, equitable, and quality education [2]. The Central Bureau of Statistics (BPS) of North Sumatra Province provides data on the number of villages and urban wards with school facilities by regency/city and education level [3]. The 2025 data can be utilized to identify the coverage of educational facilities across different regions [4], [5].

Various studies have examined the distribution and equity of educational facilities using statistical and computational approaches. Harahap et al. [6] investigated the spatial distribution of educational infrastructure development in North Sumatra. Nurfadhiilah et al. [7] analyzed the distribution of educational facilities across Indonesian provinces using the K-Means clustering method. Hadi and Mahi [8] analyzed the role of the Special Physical Allocation Fund for Affirmation (DAK Fisik Afirmasi) in educational development in underdeveloped regions. These studies demonstrate that educational equity can be examined through various indicators and analytical approaches. These studies demonstrate that educational equity can be examined through various indicators and analytical approaches. In the application of data mining, Wardani et al. [9] employed the K-Means algorithm to cluster provinces based on classroom conditions and primary school teaching staff. Salman et al. [10] compared the K-Means and K-Medoids algorithms for grouping schools based on the availability of facilities and infrastructure. Fadillah and Hasugian [11] also applied K-Means to cluster primary school facilities and infrastructure. Subaekti et al. [12] employed K-Means++ and the Davies–Bouldin Index to cluster educational resources across districts in Surabaya. These studies demonstrate the application of clustering algorithms to identify groups of regions or objects based on similarities in data characteristics. Unlike previous studies that focused on spatial distribution or the characteristics of individual schools, this study uses village-level coverage of educational facilities as the basis for grouping regencies and cities. The K-Means algorithm was selected because it can group data based on similarities in numerical characteristics. The study covers 33 regencies and cities in North Sumatra Province in 2025. Before clustering, the data are standardized to prevent differences in variable scales from dominating the distance calculations. The number of clusters is determined by considering the Elbow Method and Silhouette Score, while clustering quality is evaluated using the Silhouette Score and Davies–Bouldin Index. Diantika et al. [13] discussed the use of Silhouette, Elbow, and Gap Statistics methods to determine the optimal number of clusters in K-Means analysis.

The novelty of this study lies in the use of village-level educational facility coverage data to cluster regencies and cities in North Sumatra Province based on four levels of education during 2025. This study addresses three research questions: what are the characteristics of educational facility coverage in each regency and city; what clustering patterns emerge based on similarities in educational facility coverage; and how is the quality of the clusters generated by the K-Means algorithm? The objectives are to identify the characteristics of educational facility coverage, group regencies and cities according to their coverage levels, and evaluate the resulting clusters. The findings are expected to provide an overview of differences in educational facility coverage across regions and serve as a source of information for understanding the distribution of educational facilities in North Sumatra Province.

## Literature Review

Educational Facility Coverage, educational facility coverage represents the percentage of villages and urban wards that have school facilities at a particular education level. This indicator can be used to compare the availability of educational facilities across regions. In this study, coverage is calculated based on four education levels: primary schools (SD), junior high schools (SMP), senior high schools (SMA), and vocational high schools (SMK), using data from the Central Bureau of Statistics (BPS) of North Sumatra Province for 2025.

Data mining is a computational approach used to discover patterns and information from datasets. Yelansya and Hariyanto [14] applied the C4.5 algorithm to analyze tourist satisfaction using data collected from respondents, demonstrating the application of data mining techniques for extracting useful information and supporting decision-making, while Novelan and Aryza [15] outlined the fundamentals of machine learning in data processing. These studies provide a foundation for the use of computational methods in data analysis. This study applies an unsupervised clustering approach to group regions based on similarities in their characteristics. Hasibuan and Novelan [16] applied machine learning to analyze student sentiment for evaluating the quality of educational services, demonstrating its potential in supporting data-driven decision-making in education.

K-Means, Iqbal and Budianto [17] applied K-Means clustering to group training institutions based on several institutional characteristics, demonstrating the use of clustering to identify patterns and classify institutions based on similar attributes.

Clustering Evaluation, Clustering evaluation is performed to determine the number of clusters and assess the quality of the resulting groups. The Elbow Method is used by observing changes in the inertia values, while the Silhouette Score measures cluster cohesion and separation. The Davies–Bouldin Index is used as an additional measure to evaluate clustering quality. Apriliansyah et al. [18] discussed equitable access to education in Indonesia, including issues related to the distribution of educational facilities. This study reinforces the importance of analyzing differences in educational conditions across regions..

## Research Methodology

Research Location and Period

This study was conducted computationally using Google Colaboratory (Google Colab) and the Python programming language. Secondary data were obtained from the Central Bureau of Statistics (BPS) of North Sumatra Province for 2025, covering 33 regencies and cities across four education levels: primary schools (SD), junior high schools (SMP), senior high schools (SMA), and vocational high schools (SMK). The data processing stages included data cleaning, calculation of educational facility coverage percentages, data standardization, clustering using K-Means, as well as evaluation and analysis of the clustering results.

Research Stages

The research process was carried out through the following stages:

1. Problem Identification: Identifying differences in educational facility coverage across regencies and cities in North Sumatra Province.
2. Literature Review: Reviewing the concepts of clustering, K-Means, data standardization, and cluster evaluation.
3. Data Collection: Collecting secondary data from the Central Bureau of Statistics (BPS) of North Sumatra Province for 2025, covering 33 regencies and cities. An example of the data is presented in Table 1.

| Regency/City | SD | SMP | SMA | SMK |
|---|---|---|---|---|
| Nias | 131 | 53 | 13 | 18 |
| Mandailing Natal | 328 | 113 | 49 | 19 |
| Tapanuli Selatan | 206 | 79 | 42 | 13 |
| Tapanuli Tengah | 197 | 78 | 33 | 20 |
| Tapanuli Utara | 228 | 78 | 23 | 19 |

| | | | | |
|---|---|---|---|---|
| Toba Samosir | 179 | 46 | 17 | 12 |
| Labuhan Batu | 94 | 65 | 43 | 26 |
| Asahan | 196 | 117 | 63 | 31 |
| Simalungun | 393 | 147 | 52 | 31 |
| Dairi | 163 | 52 | 24 | 10 |
| Karo | 193 | 57 | 26 | 10 |
| Deli Serdang | 323 | 201 | 121 | 78 |
| Langkat | 270 | 178 | 85 | 58 |
| Nias Selatan | 319 | 137 | 58 | 47 |
| Humbang Hasundutan | 148 | 46 | 17 | 11 |
| Pakpak Bharat | 48 | 30 | 7 | 4 |
| Samosir | 123 | 33 | 14 | 5 |
| Serdang Bedagai | 201 | 96 | 51 | 29 |
| Batu Bara | 143 | 74 | 33 | 21 |
| Padang Lawas Utara | 207 | 80 | 46 | 10 |
| Padang Lawas | 151 | 72 | 40 | 15 |
| Labuhan Batu Selatan | 54 | 41 | 28 | 19 |
| Labuhan Batu Utara | 89 | 61 | 34 | 17 |
| Nias Utara | 100 | 55 | 17 | 22 |
| Nias Barat | 79 | 39 | 16 | 12 |
| Kota Sibolga | 14 | 9 | 8 | 4 |
| Kota Tanjung Balai | 31 | 19 | 12 | 6 |
| Kota Pematang Siantar | 52 | 29 | 23 | 20 |
| Kota Tebing Tinggi | 33 | 22 | 15 | 7 |
| Kota Medan | 147 | 140 | 110 | 69 |
| Kota Binjai | 37 | 32 | 19 | 13 |
| Kota Padangsidimpuan | 57 | 30 | 21 | 16 |
| Kota Gunungsitoli | 76 | 30 | 10 | 11 |

Table 1. Dataset Obtained from the Central Bureau of Statistics (BPS)

4. Data Preprocessing: Cleaning the data, removing rows that do not represent regency/city observations, and selecting the four research variables: primary schools (SD), junior high schools (SMP), senior high schools (SMA), and vocational high schools (SMK).
5. Data Transformation: Calculating the educational facility coverage percentage for each education level and performing standardization using the Z-score method (StandardScaler).
6. Determining the Number of Clusters: Testing the number of clusters from K = 2 to K = 8 using the Elbow Method and Silhouette Score.
7. K-Means Clustering: The K-Means algorithm is used to group 33 observations based on four educational facility variables. The grouping process is performed by calculating the distance between each data point and the nearest centroid using the Euclidean Distance formula:

$$d(x_i, \mu_j) = \sqrt{\sum_{k=1}^{p} (x_{ik} - \mu_{jk})^2}$$

8. Evaluation and Results Analysis: Evaluating the clustering results using the Elbow Method, Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Score. The results are analyzed based on the distribution and characteristics of each cluster.

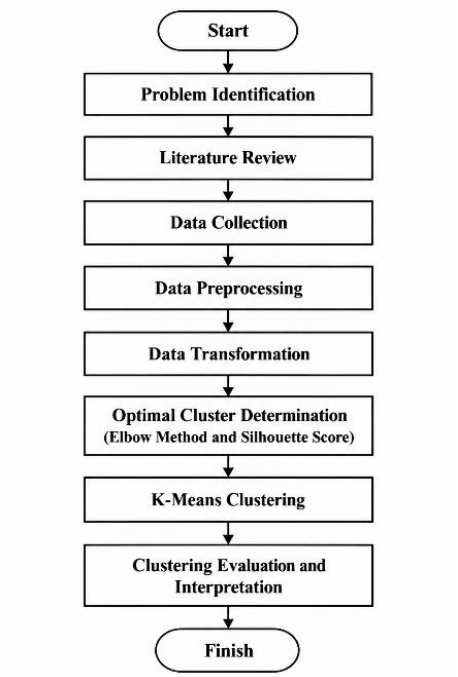



Figure 1. Analytical Methodology Flowchart

**Metode Analisis**

1. Data Collection and Preprocessing: The research data were obtained from the Central Bureau of Statistics (BPS) for 2025, covering the number of villages and urban wards with educational facilities across 33 regencies and cities in North Sumatra Province. The data included four education levels: primary schools (SD), junior high schools (SMP), senior high schools (SMA), and vocational high schools (SMK). Subsequently, the data were examined, cleaned, and reformatted.
2. Calculation of Educational Facility Coverage: Educational facility coverage was calculated by dividing the number of villages and urban wards with school facilities by the total number of villages and urban wards in each regency or city, and then multiplying the result by 100%.

$$\boldsymbol{C_{ij} = \frac{F_{ij}}{D_i} \times 100\backslash\%}$$

3. Data Standardization: The four educational facility coverage variables (SD, SMP, SMA, and SMK) were standardized using the Z-score method through StandardScaler to ensure that all variables had comparable scales.
4. Determining the Number of Clusters: The optimal number of clusters was determined by testing values of K ranging from 2 to 8 using the Elbow Method and Silhouette Score.
5. Implementation of K-Means Clustering: The K-Means algorithm was applied to the four standardized variables to group regencies and cities based on similarities in educational facility coverage.
6. Evaluation and Interpretation of Results: The clustering results were evaluated using Inertia, Silhouette Score, Davies–Bouldin Index, and Calinski–Harabasz Score. The resulting clusters were analyzed based on the distribution of cluster members and the average educational facility coverage, and were presented in tables and graphs.
7. Tools and Software: This study used Python in Google Colaboratory, supported by Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn for data processing, clustering, evaluation, and result visualization.

**Results**

1. **Determination of the Optimal Number of Clusters**

The optimal number of clusters was determined by testing K values ranging from 2 to 8. The evaluation was performed using the Elbow Method based on Inertia values and the Silhouette Score to assess the quality of cluster separation. The results are presented in Figures 2 and 3.

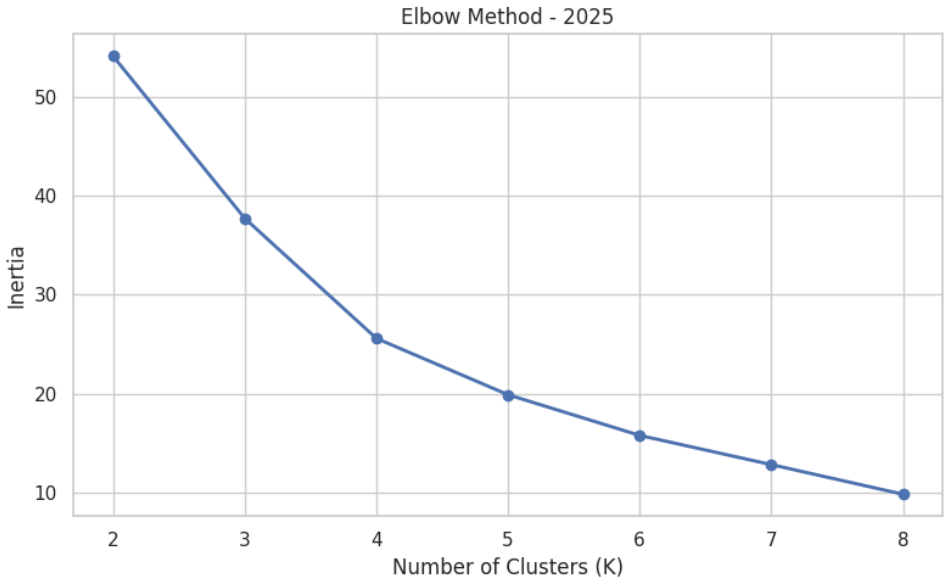



Figure 2. Elbow Method

Based on Figure 2, the Inertia value decreased from 54.1175 at K = 2 to 9.8632 at K = 8. A relatively substantial decrease was observed between K = 2 and K = 4, after which the decline in Inertia began to level off. This pattern indicates that gradually increasing the number of clusters results in progressively smaller reductions in within-cluster variation.

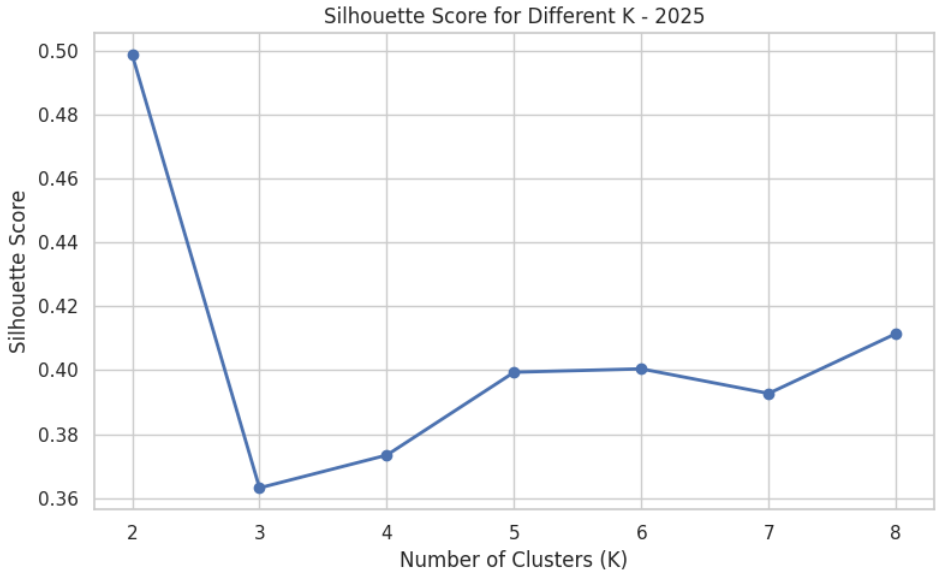



Figure 3. Silhouette Score

Based on Figure 3, the highest Silhouette Score was obtained at K = 2, with a value of 0.4987. At K = 3, the score decreased to 0.3631, then gradually increased to 0.4114 at K = 8. These results indicate that K = 2 provides the best cluster separation quality among the evaluated cluster configurations.

| | K | Inertia | Silhouette_Score | Davies_Bouldin_Index | Calinski_Harabasz_Score |
|---|---|---|---|---|---|
| 0 | 2 | 54.1175 | 0.4987 | 0.7072 | 44.6132 |
| 1 | 3 | 37.7384 | 0.3631 | 0.8948 | 37.4665 |
| 2 | 4 | 25.6087 | 0.3734 | 0.8447 | 40.1602 |
| 3 | 5 | 19.9137 | 0.3994 | 0.7914 | 39.4002 |
| 4 | 6 | 15.8051 | 0.4004 | 0.7863 | 39.6993 |
| 5 | 7 | 12.8527 | 0.3927 | 0.7224 | 40.1710 |
| 6 | 8 | 9.8632 | 0.4114 | 0.6851 | 44.2253 |

Figure 4. Cluster Evaluation Results

Based on Figure 4, the evaluation results show that K = 2 achieved the highest Silhouette Score of 0.4987 and the highest Calinski–Harabasz Score of 44.6132. Meanwhile, the lowest Davies–Bouldin Index was obtained at K = 8, with a value of 0.6851. Nevertheless, considering the results of all three evaluation metrics, particularly the Silhouette Score, this study selected K = 2 as the optimal number of clusters.

**2. K-Means Clustering Result**

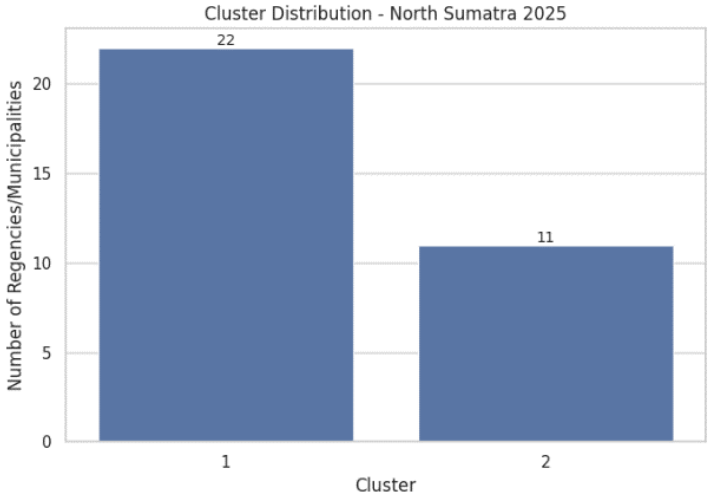



Figure 5. Cluster Distribution

Based on Figure 5, the clustering results show that Cluster 1 consists of 22 regencies/cities (66.67%), while Cluster 2 consists of 11 regencies/cities (33.33%), out of the 33 regions analyzed. These results indicate that the majority of regencies/cities belong to Cluster 1, whereas Cluster 2 has fewer members. The two clusters were subsequently analyzed to identify the characteristics of educational facility coverage at each education level.

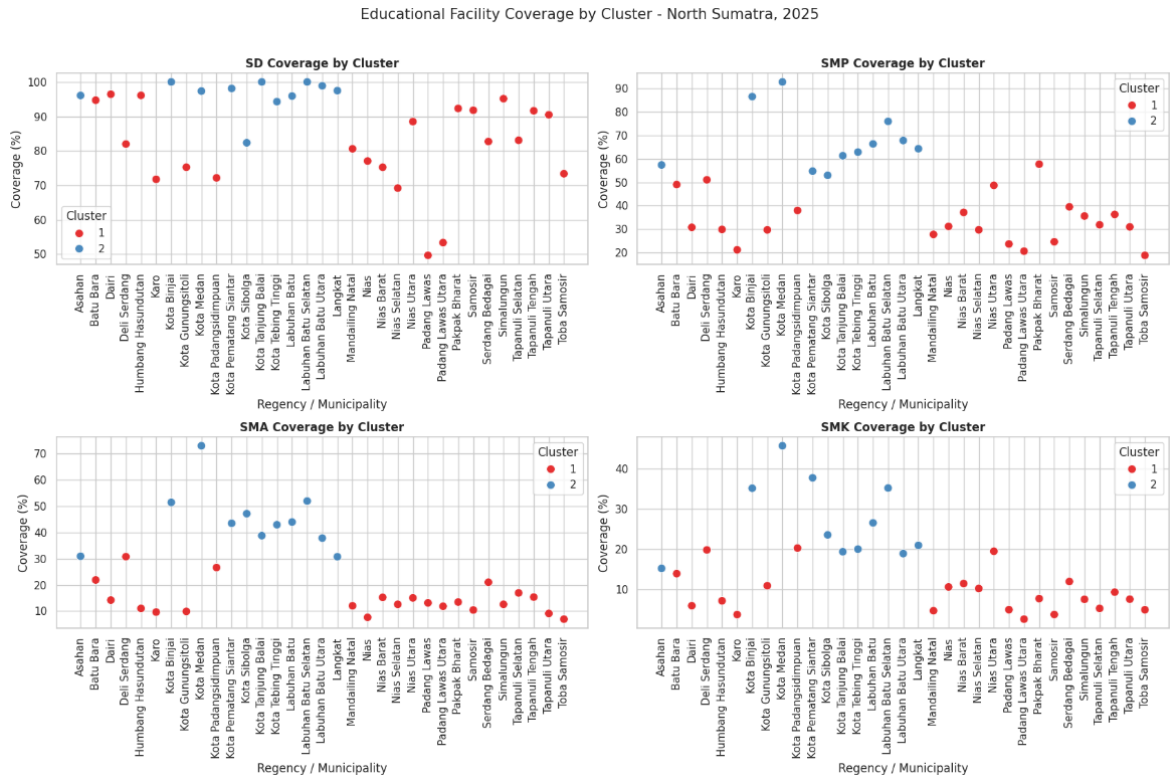



Figure 6. Distribution of Educational Facility Coverage by Cluster in North Sumatra Province in 2025

Based on Figure 6, educational facility coverage across the 33 regencies and cities in North Sumatra Province in 2025 exhibits different characteristics between Cluster 1 and Cluster 2. At the primary school level, both clusters show relatively high coverage percentages, whereas more noticeable differences are observed at the junior high school, senior high school, and vocational high school levels. Cluster 2 (blue) generally has higher educational facility coverage than Cluster 1 (red), particularly at the secondary and vocational education levels. These findings indicate that the K-Means algorithm can group regions based on similarities in educational facility coverage characteristics and illustrate differences in the distribution of educational facilities across North Sumatra Province.

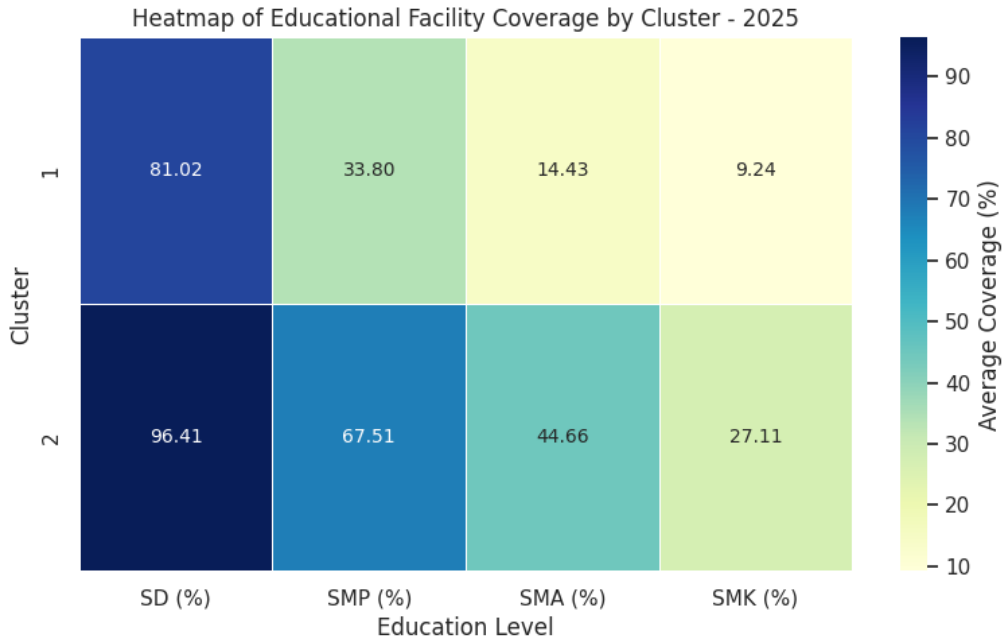



Figure 7. Heatmap of Average Educational Facility Coverage by Cluster in North Sumatra Province in 2025

Based on the heatmap in Figure 7, Cluster 2 has a higher average educational facility coverage than Cluster 1 across all education levels. Cluster 1 has coverage rates of 81.02% for primary schools (SD), 33.80% for junior high schools (SMP), 14.43% for senior high schools (SMA), and 9.24% for vocational high schools (SMK). Meanwhile, Cluster 2 achieves coverage rates of 96.41% for SD, 67.51% for SMP, 44.66% for SMA, and 27.11% for SMK. Both clusters exhibit a similar pattern, with the highest coverage at the primary school level and the lowest at the vocational high school level.

## Conclusion

This study successfully clustered 33 regencies and cities in North Sumatra Province based on educational facility coverage in 2025 using the K-Means algorithm with four variables: primary schools (SD), junior high schools (SMP), senior high schools (SMA), and vocational high schools (SMK). Based on the evaluation of cluster numbers ranging from K = 2 to K = 8, K = 2 was selected because it produced the highest Silhouette Score of 0.4987 and a Calinski–Harabasz Score of 44.6132, although the lowest Davies–Bouldin Index was obtained at K = 8. The clustering results produced Cluster 1, consisting of 22 regencies and cities (66.67%), and Cluster 2, consisting of 11 regencies and cities (33.33%). Cluster 1 exhibited relatively lower average educational facility coverage, with 81.02% for SD, 33.80% for SMP, 14.43% for SMA, and 9.24% for SMK. In contrast, Cluster 2 showed higher coverage rates of 96.41% for SD, 67.51% for SMP, 44.66% for SMA, and 27.11% for SMK. Both clusters indicate that educational facility coverage tends to decrease at higher education levels, particularly at the senior high school and vocational high school levels. Therefore, the K-Means algorithm can be used to identify differences in educational facility coverage characteristics across regencies and cities. The findings are expected to provide useful information for local governments in identifying regions that require greater attention to the equitable distribution of educational facilities. Future research could be expanded by incorporating additional indicators, such as population size, geographical conditions, and school availability, to provide a more comprehensive picture of educational equity.

## References


[1] Republik Indonesia, "Undang-Undang Republik Indonesia Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional," Lembaran Negara Republik Indonesia Tahun 2003 Nomor 78, 2003. [Online]. Available: https://peraturan.bpk.go.id/Details/43920/uu-no-20-tahun-2003.

[2] United Nations, "Goal 4: Ensure Inclusive and Equitable Quality Education and Promote Lifelong Learning Opportunities for All," United Nations Sustainable Development Goals. [Online]. Available: https://sdgs.un.org/goals/goal4.

[3] Badan Pusat Statistik Provinsi Sumatera Utara, "Provinsi Sumatera Utara dalam Angka 2025," BPS Provinsi Sumatera Utara, Medan, 2025. [Online]. Available: https://sumut.bps.go.id/.

[4] Badan Pusat Statistik Provinsi Sumatera Utara, "Jumlah Desa/Kelurahan yang Memiliki Fasilitas Sekolah Menurut Kabupaten/Kota dan Tingkat Pendidikan di Provinsi Sumatera Utara, 2024," 2025. [Online]. Available: https://sumut.bps.go.id/.

[5] Badan Pusat Statistik Provinsi Sumatera Utara, "Jumlah Desa/Kelurahan yang Memiliki Fasilitas Sekolah Menurut Kabupaten/Kota dan Tingkat Pendidikan di Provinsi Sumatera Utara, 2025," 2026. [Online]. Available: https://sumut.bps.go.id/.

[6] M. H. Harahap, H. Siregar, E. Rustiadi, and A. E. Pravitasari, "Analysis of the Spatial Distribution Patterns of Education Infrastructure Development (A Case Study of 33 Regencies/City in North Sumatra Province)," Journal of Infrastructure, Policy and Development, vol. 8, no. 11, Art. no. 8624, 2024.

[7] A. Nurfadhiilah, B. G. Martens, and A. Wibowo, "Analisis pengelompokkan distribusi fasilitas pendidikan di Indonesia dengan metode klasterisasi," Jurnal Education and Development, vol. 14, no. 1, pp. 708–715, 2026.

[8] S. Hadi and B. R. Mahi, "Peran Dana Alokasi Khusus Fisik Afirmasi terhadap Pembangunan Pendidikan Daerah Tertinggal di Indonesia," Indonesian Treasury Review, vol. 9, no. 1, 2024.

[9] M. T. Wardani, V. O. Ramadhani, N. Anggreani, Sumanto, and A. D. Kuswanto, "Pengelompokkan Provinsi Berdasarkan Kelayakan Ruang Kelas dan Tenaga Kependidikan Sekolah Dasar Menggunakan Algoritma K-Means: Analisis Data Periode 2023–2024," Riau Jurnal Teknik Informatika, vol. 4, no. 2, pp. 163–172, 2025.

[10] S. Simamora, M. Iqbal, A. P. U. Siahaan, Khairul, and Z. Sitorus, "Comparison of K-Means and Self Organizing Map Algorithms for Ground Acceleration Clustering," Sinkron: Jurnal dan Penelitian Teknik Informatika, vol. 8, no. 4, pp. 2345–2353, 2024.

[11] O. M. Z. Fadillah and A. H. Hasugian, "K-Means Clustering for Grouping Elementary School Facilities and Infrastructure in Indonesia," IDEALIS: InDonEsiA journaL Information System, vol. 9, no. 2, pp. 182–192, 2026.

[12] H. Subaekti, L. Hakim, H. Khaulasari, and D. Yuliati, "An integrated K-Means++ Davies–Bouldin Index approach for educational resource-based district clustering: A case study of districts in Surabaya," Jambura Journal of Mathematics, vol. 8, no. 1, 2026.

[13] M. W. Diantika, A. Rusgiyono, and B. A. Saputra, "Perbandingan metode optimasi Silhouette, Elbow, dan Gap Statistics dalam menentukan nilai K terbaik pada analisis K-Means clustering," Jurnal Gaussian, vol. 14, no. 2, pp. 335–344, 2025.

[14] Yelansya and E. Hariyanto, "Analisis Kepuasan Wisatawan Pada Daerah Kota Binjai Menggunakan Algoritma C4.5 Berbasis Mobile," Bulletin of Information Technology (BIT), vol. 4, no. 1, pp. 89–98, 2023.

[15] M. S. Novelan and S. Aryza, Belajar Mengenal Dasar Machine Learning. Medan: PT Serasi Media Teknologi, 2025.

[16] M. E. Hasibuan and M. S. Novelan, "Implementation of Machine Learning for Student Sentiment Analysis as a Basis for Evaluating the Quality of Educational Services at

Private Higher Education Institutions in Padang Lawas," JURIKOM (Jurnal Riset Komputer), vol. 15, no. 5, 2026.

[17] E. Budianto and M. Iqbal, "Model Predictive Analysis of Performance in Training and Course Institutions Using Naive Bayes and K-Means Clustering," Journal of Computer Science and Research (JoCoSiR), vol. 3, no. 1, pp. 10–16, 2025.

[18] F. Apriliansyah, T. T. D. Susanto, L. I. Mayasari, and K. Kamaludin, "Equitable Access to Education in Indonesia: A Systematic Review of the Literature with Bibliometric Analysis (2007–2025)," Journal of Educational Sciences, vol. 10, no. 7, pp. 436–446, 2026, doi: 10.31258/jes.10.7.p.436-446.