# Clustering Analysis of School Grant Recipients Based on Grant Allocation Characteristics Using the K-Means Algorithm in North Sumatra Province

**Hengki Gunawan, Muhammad Syahputra Novelan**

**Abstract**

The distribution of educational grants requires systematic analysis to understand funding allocation patterns among educational institutions. This study aims to analyze the characteristics of school grant recipients in North Sumatra Province using the K-Means clustering algorithm. The study uses administrative grant data covering the 2023–2026 period, consisting of recipient institution names and addresses and nominal grant amounts. The research process includes data preprocessing, numerical feature selection, data transformation, K-Means clustering, and cluster evaluation. The grant amounts are transformed using the natural logarithm of one plus the nominal value to reduce the influence of extreme values in the clustering process. The algorithm is configured to generate three clusters representing lower, moderate, and higher grant allocation levels. The clustering results are evaluated using the Silhouette Coefficient to provide an initial assessment of cluster cohesion and separation. The findings are expected to provide a structured overview of grant allocation characteristics and support the descriptive evaluation of educational funding distribution. However, the resulting clusters represent differences in nominal grant amounts rather than direct measurements of institutional needs or funding priorities. Further research incorporating student enrollment, infrastructure conditions, socioeconomic characteristics, and educational facility availability is necessary to support a more comprehensive, evidence-based approach to grant allocation.

***Keywords**: K-Means, clustering, educational grants, grant allocation, North Sumatra Province.*

[1]Hengki Gunawan
Information Technology, Universitas Pembangunan Panca Budi, Indonesia
e-mail: hengkigunawan1591@gmail.com

[2]Muhammad Syahputra Novelan
e-mail: putranovelan@dosen.pancabudi.ac.id
2nd International Cofference on the Epicentrum of Economic Global Framework (ICEEGLOF)
Theme: Navigating The Future: Business and Social Paradigms in a Transformative Era.
https://proceeding.pancabudi.ac.id/index.php/ICEEGLOF

## Introduction

The distribution of educational grants plays an important role in supporting the development of educational institutions, improving educational facilities, and expanding access to quality education. In North Sumatra Province, grant allocation involves educational institutions with different characteristics, locations, and funding requirements. Differences in institutional conditions and the amount of financial assistance received indicate the need for a systematic approach to understanding grant distribution patterns. An appropriate classification process can help identify groups of grant recipients with similar allocation characteristics and provide a clearer overview of how educational grants are distributed across the region. Therefore, data-driven approaches are increasingly relevant for supporting more structured, transparent, and evidence-based grant management [1] [2].

The increasing availability of grant recipient data provides an opportunity to apply data mining techniques to the analysis of educational funding distribution. One of the main challenges in managing these data is identifying patterns among recipients based on their grant allocation characteristics. Conventional administrative records generally present recipients individually, making it difficult to identify similarities and differences across institutions when the dataset becomes large. Clustering techniques can address this challenge by grouping data objects according to similarities in their selected attributes. Through this approach, grant recipient records can be organized into distinct groups that represent different allocation characteristics, thereby facilitating the interpretation of funding distribution patterns and supporting further evaluation of grant management practices [3] [4] .

This study uses grant recipient data from North Sumatra Province covering the period from 2023 to 2026. The dataset is organized into four annual worksheets, consisting of 1,804 records for 2023, 1,586 records for 2024, 469 records for 2025, and 373 records for 2026, resulting in a total of 4,232 records. Each worksheet contains three main fields: a record number, the name and address of the recipient institution, and the nominal amount of the grant received. The institutional information includes various educational entities, such as early childhood education institutions, madrasahs, schools, Islamic boarding schools, foundations, and other educational organizations. The nominal grant values provide quantitative information that can be examined to identify differences in funding allocation among recipients. These data serve as the basis for analyzing grant distribution characteristics across the four-year observation period [5] [6].

The K-Means algorithm is employed in this study because it is a widely used unsupervised learning technique for grouping data based on similarities in their numerical characteristics. The algorithm partitions observations into a predetermined number of clusters by minimizing the distance between individual observations and their respective cluster centroids. In the context of this research, K-Means can be used to group grant recipient records according to the nominal amount of funding received, provided that the data are appropriately prepared and validated before clustering. The resulting clusters can subsequently be interpreted to describe different grant allocation profiles, such as groups characterized by relatively lower, intermediate, or higher funding amounts. However, these groups represent differences in allocation characteristics rather than direct measurements of institutional need, educational quality, or the fairness of grant distribution [7] [8] .

Based on this background, this study aims to analyze the characteristics of educational grant recipients in North Sumatra Province using the K-Means clustering algorithm and grant allocation data from 2023 to 2026. The analysis is intended to identify groups of recipients with similar funding characteristics and provide a structured overview of grant distribution patterns across the observed period. The contribution of this study lies in applying a data mining approach to administrative grant records to transform individual recipient data into more interpretable clusters. The findings are expected to provide descriptive information that can support further evaluation of educational grant allocation and inform the development of more

comprehensive data-driven decision-support processes. Nevertheless, the interpretation of clustering results must consider the available attributes, data quality, and differences in grant allocation across years. Additional indicators, such as institutional infrastructure, student enrollment, socioeconomic conditions, and unmet educational needs, would be required to establish a more comprehensive assessment of funding priorities[9] .

## Literature Review

### 2.1 Educational Grant Allocation

Educational grant allocation plays an important role in supporting the development of educational institutions, improving learning facilities, and strengthening the quality of educational services. The distribution of financial assistance involves institutions with different characteristics, geographical locations, and funding requirements. Therefore, analyzing grant allocation data is necessary to understand funding distribution patterns and identify similarities among recipients. Administrative records containing institutional names, addresses, and grant amounts provide valuable information for examining the characteristics of funding distribution across different periods. In this study, educational grant data from North Sumatra Province covering the period from 2023 to 2026 are used to examine differences in the nominal amounts received by grant recipient institutions. This analysis provides a basis for identifying groups of recipients with similar grant allocation characteristics [10] [11].

### 2.2 Data Mining and Clustering

Data mining is a process of extracting meaningful patterns and useful information from datasets using computational and statistical techniques. It is widely applied in various fields to transform raw data into structured information that supports analysis and decision-making. One of the important techniques in data mining is clustering, an unsupervised learning method that groups observations according to similarities in their attributes without requiring predefined class labels. Clustering enables researchers to explore data structures, identify groups with comparable characteristics, and understand variations within a dataset. In educational grant management, clustering can be used to group recipient institutions according to the nominal amount of funding received. The resulting groups can help describe grant distribution patterns and provide an overview of differences in funding allocation. However, clustering based only on grant amounts does not directly establish the level of institutional need or the adequacy of the funding received [12] [13] .

### 2.3 K-Means Clustering Algorithm

K-Means is an unsupervised learning algorithm that partitions a dataset into a predetermined number of clusters based on similarities among observations. The algorithm assigns each observation to the cluster with the nearest centroid, commonly calculated using Euclidean distance. The centroid represents the mean value of the observations within a cluster and is updated iteratively until the assignments stabilize or a stopping criterion is reached. The primary objective of K-Means is to minimize the within-cluster sum of squared distances between observations and their corresponding centroids. Due to its relatively simple implementation, computational efficiency, and ease of interpretation, K-Means is frequently used for numerical data analysis. In this study, K-Means is applied to the nominal grant amounts received by educational institutions in North Sumatra Province during 2023–2026. The resulting clusters can be interpreted as groups with relatively low, moderate, or high grant allocations, depending on the observed distribution and the selected number of clusters. These results are expected to provide a structured description of grant allocation characteristics and support further evaluation of educational funding distribution. [14].

## Research Methodology

This study employs a quantitative approach using data mining techniques to analyze the characteristics of educational grant recipients in North Sumatra Province. The research focuses on applying the K-Means clustering algorithm to identify groups of recipients based on the nominal amounts of grants received during the 2023–2026 period. The dataset consists of administrative records containing recipient identification information, institutional names and addresses, and grant allocation amounts. The research process is organized into several systematic stages, beginning with data collection and continuing through data preprocessing, feature selection, data transformation, K-Means clustering, and evaluation and interpretation of the resulting clusters. This structured approach is intended to ensure that the analysis is conducted consistently and that the resulting groups can be interpreted in relation to the characteristics of grant allocation.

The research methodology is designed to transform administrative grant records into meaningful information about funding distribution patterns. Before clustering is performed, the dataset is examined to identify missing values, duplicate records, inconsistent formats, and potential anomalies in the grant amounts. The relevant numerical attributes are then prepared for analysis, while institutional names and addresses are retained as supporting information for identifying recipients and interpreting the results. The K-Means algorithm is subsequently applied to group observations according to similarities in grant allocation amounts. Finally, the clustering results are evaluated and interpreted to describe the characteristics of each group. Since the primary quantitative attribute is the nominal grant amount, the findings represent grant allocation patterns rather than direct measurements of institutional need or funding priority.

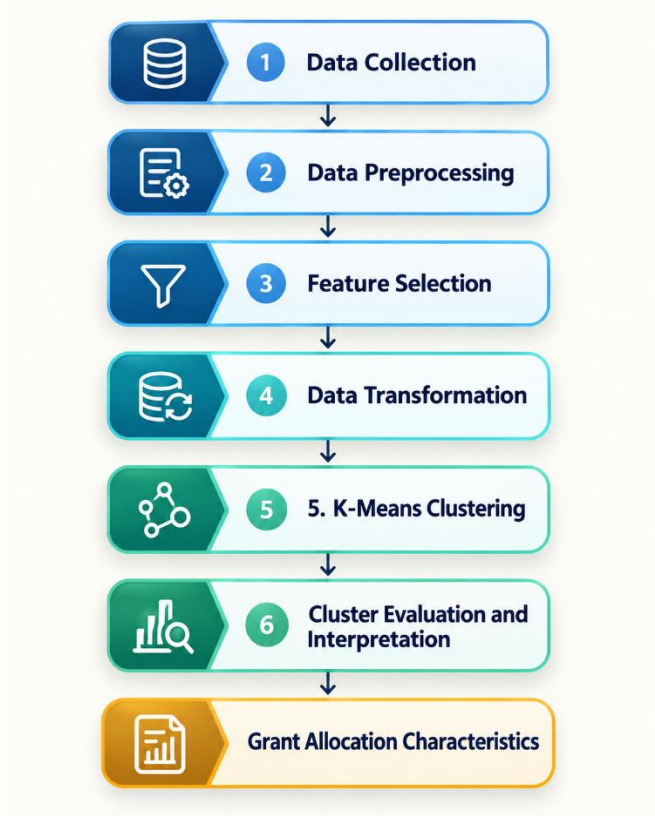



**Figure 1.** Research Methodology

1. Data Collection
   The first stage involves collecting educational grant recipient data from North Sumatra Province for the 2023–2026 period. The dataset contains administrative information, including recipient record numbers, institutional names and addresses, and nominal grant

amounts. The data are organized into annual worksheets, enabling the researcher to examine funding allocation characteristics across different years. These records provide the primary data source for the clustering process. At this stage, the structure and completeness of the available attributes are also examined to determine their suitability for subsequent analysis.

2. Data Preprocessing
   Data preprocessing is conducted to improve data quality before applying the clustering algorithm. This stage includes checking missing values, identifying duplicate records, standardizing numerical formats, and examining inconsistent entries in the grant amount field. Potentially unusual values are also investigated to determine whether they represent valid grant allocations or data entry errors. Institutional names and addresses are reviewed to maintain consistent recipient identification. These procedures are important because inaccurate or inconsistent data may affect the calculation of cluster centroids and the assignment of observations to clusters.
3. Feature Selection
   Feature selection determines the attributes used as input variables for the clustering process. In this study, the nominal grant amount is selected as the primary numerical feature because it directly represents the funding allocation received by each institution. Recipient names and addresses are retained as identification attributes to facilitate the interpretation of clustering results but are not directly included as numerical inputs in the K-Means algorithm. This separation ensures that the clustering process is based on meaningful quantitative information rather than arbitrary numerical representations of institutional identities.
4. Data Transformation
   Data transformation prepares the selected numerical attributes for analysis. The grant amount field is converted into a consistent numerical format to ensure compatibility with the K-Means algorithm. Descriptive statistics and the distribution of grant amounts are examined to identify differences in scale and potential extreme values. If necessary, an appropriate transformation or scaling technique can be considered, provided that its effect on the interpretation of grant amounts is clearly evaluated. Because the analysis primarily uses one numerical feature, the transformation process focuses on maintaining data consistency and ensuring that the resulting clusters remain interpretable in terms of funding allocation.
5. K-Means Clustering
   At this stage, the K-Means algorithm is applied to group grant recipient records according to similarities in their nominal grant amounts. The algorithm begins by determining the number of clusters and initializing their centroids. Each observation is assigned to the nearest centroid based on Euclidean distance, after which the centroids are recalculated using the mean values of the observations assigned to each group. These steps are repeated until the cluster assignments stabilize or the stopping criterion is satisfied. The resulting groups may represent relatively low, moderate, and high grant allocation levels, depending on the observed data distribution and the selected number of clusters. The number of clusters should be determined through an appropriate evaluation procedure rather than assumed in advance.
6. Cluster Evaluation and Interpretation
   The final stage evaluates the clustering results and interprets the characteristics of each group. Evaluation can be performed using the Elbow Method to examine the relationship between the number of clusters and the within-cluster sum of squares, together with the Silhouette Coefficient to assess cluster cohesion and separation. The number of observations, mean grant amount, median grant amount, and range of funding values can then be summarized for each cluster. These results provide a descriptive overview of how grant recipients are distributed across different allocation levels. The findings can support further examination of grant distribution patterns in North Sumatra Province, although identifying

schools with the greatest funding needs would require additional indicators, such as student enrollment, infrastructure conditions, and socioeconomic characteristics.

## Results

This study analyzed 4,228 individual grant recipient records from the annual worksheets covering 2023–2026. The records contain institutional names and addresses and nominal grant amounts, which serve as the principal quantitative attribute in the clustering process. Annual summary rows that did not identify individual recipients were excluded so that the analysis reflects recipient-level observations rather than duplicated totals.

The results are presented through annual descriptive statistics, cluster profiles, and visual summaries. Annual recipient counts and total grant values provide complementary perspectives on the available records and the nominal distribution of funding. The number of records in a year should be interpreted in light of data coverage, while mean and median grant amounts help describe the distribution and the potential influence of unusually large allocations.

The K-Means algorithm was applied to the natural-log-transformed grant amount, ln(1 + grant amount), and configured to produce three clusters. The clusters were ordered by their median nominal grant values and described as lower, moderate, and higher allocation profiles. The silhouette coefficient for this one-feature clustering was 0.609. The resulting groups describe funding allocation characteristics only and do not directly measure institutional need, educational quality, or eligibility for future grants.

### 4.1 Annual Recipient Records and Grant Allocation

The annual record counts show the volume of recipient-level data available for each year. Total grant values summarize the nominal funding recorded in the individual entries, while mean and median grant amounts describe the distribution from different perspectives. A difference between the mean and median may indicate that a small number of large allocations influence the average. These figures provide a descriptive overview; year-to-year differences should not be interpreted as changes in educational need unless data completeness and annual recording criteria are comparable.

Table 1 summarizes annual record counts and the nominal grant allocation statistics.

| Year | Recipient records | Total grant | Mean grant | Median grant |
|---|---|---|---|---|
| 2023 | 1,803 | Rp 283.037.000.000 | Rp 156.981.143 | Rp 100.000.000 |
| 2024 | 1,585 | Rp 268.828.070.000 | Rp 169.607.615 | Rp 100.000.000 |
| 2025 | 468 | Rp 115.492.000.000 | Rp 246.777.778 | Rp 105.000.000 |
| 2026 | 372 | Rp 64.035.000.000 | Rp 172.137.097 | Rp 150.000.000 |

### 4.2 K-Means Cluster Characteristics

The three groups distinguish lower, moderate, and higher allocation profiles based on their median grant values. The number and share of records show the prevalence of each profile in the dataset. Median values are useful alongside means because unusually large grants can raise the average. The cluster labels are descriptive and should not be treated as rankings of institutional need, since the clustering input is the nominal grant amount rather than direct indicators of educational conditions.

**Table 2 presents the three K-Means clusters, ordered by increasing median grant amount.**

| Cluster | Allocation profile | Records | Share | Median grant | Mean grant |
|---|---|---|---|---|---|
| Cluster 1 | Lower allocation | 1,120 | 26.5% | Rp 50.000.000 | Rp 55.110.911 |
| Cluster 2 | Moderate allocation | 1,829 | 43.3% | Rp 100.000.000 | Rp 116.376.982 |
| Cluster 3 | Higher allocation | 1,279 | 30.3% | Rp 250.000.000 | Rp 357.165.246 |

### 4.3 Cluster Distribution by Year

This table makes it possible to compare the prevalence of each allocation profile within each annual worksheet. Absolute cluster counts should be considered alongside each year's total records, because years with fewer observations will naturally tend to have fewer entries in each cluster. The table supports descriptive comparison of allocation patterns, but it does not establish why these patterns occurred or whether one group should receive future grants first.

**Table 3 provides a year-by-year breakdown of the records assigned to each cluster.**

| Year | Cluster 1 Lower | Cluster 2 Moderate | Cluster 3 Higher | Total |
|---|---|---|---|---|
| 2023 | 534 | 797 | 472 | 1,803 |
| 2024 | 364 | 706 | 515 | 1,585 |
| 2025 | 154 | 178 | 136 | 468 |
| 2026 | 68 | 148 | 156 | 372 |

### 4.4 Visual Analysis

Figure 2 makes annual data coverage easy to compare. The height of each bar represents the number of individual recipient entries available in that year. Lower counts in later years should be understood as a feature of the supplied records, not automatically as evidence of declining demand or reduced support.

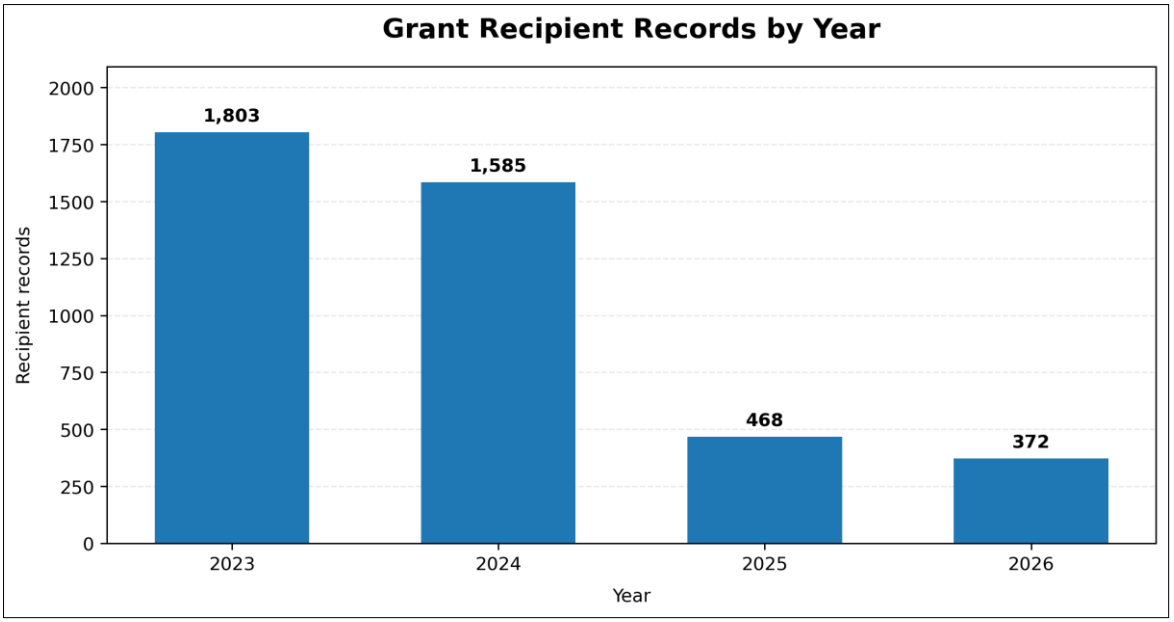



**Figure 2. Number of grant recipient records by year**.

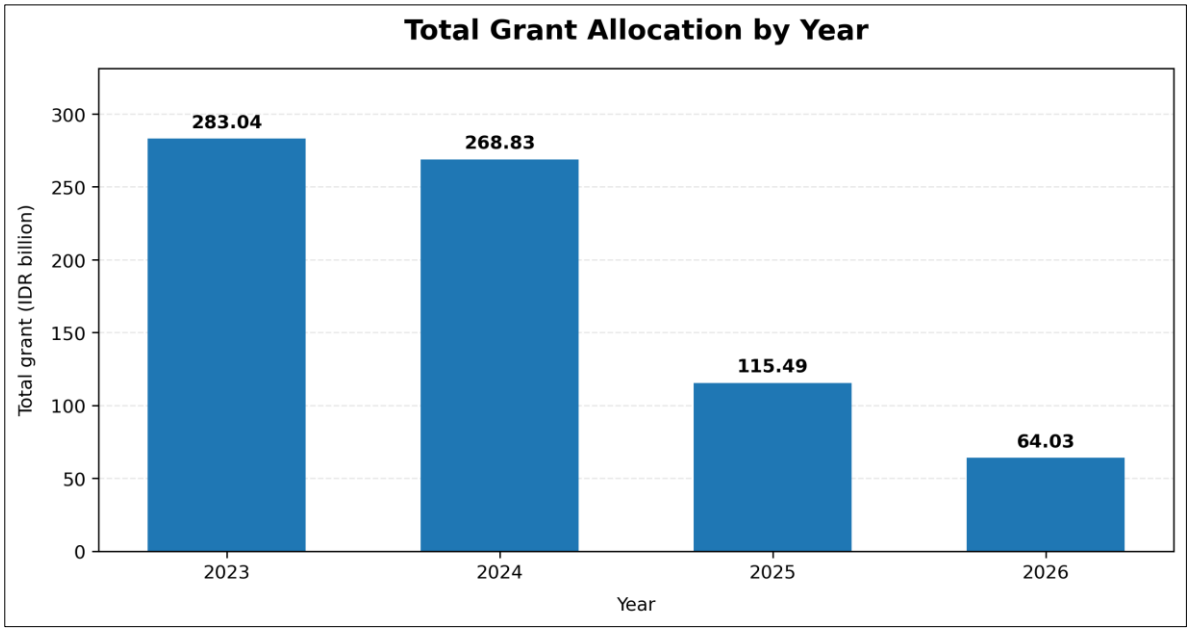



**Figure 3. Total nominal grant allocation by year**.

Figure 3 compares the total recorded grant amount for each year, expressed in billions of Indonesian rupiah. This view complements recipient counts: a year with fewer records may still have a large total if the amounts recorded for recipients are higher. Both the volume of records and the grant values are therefore important when describing annual allocation.

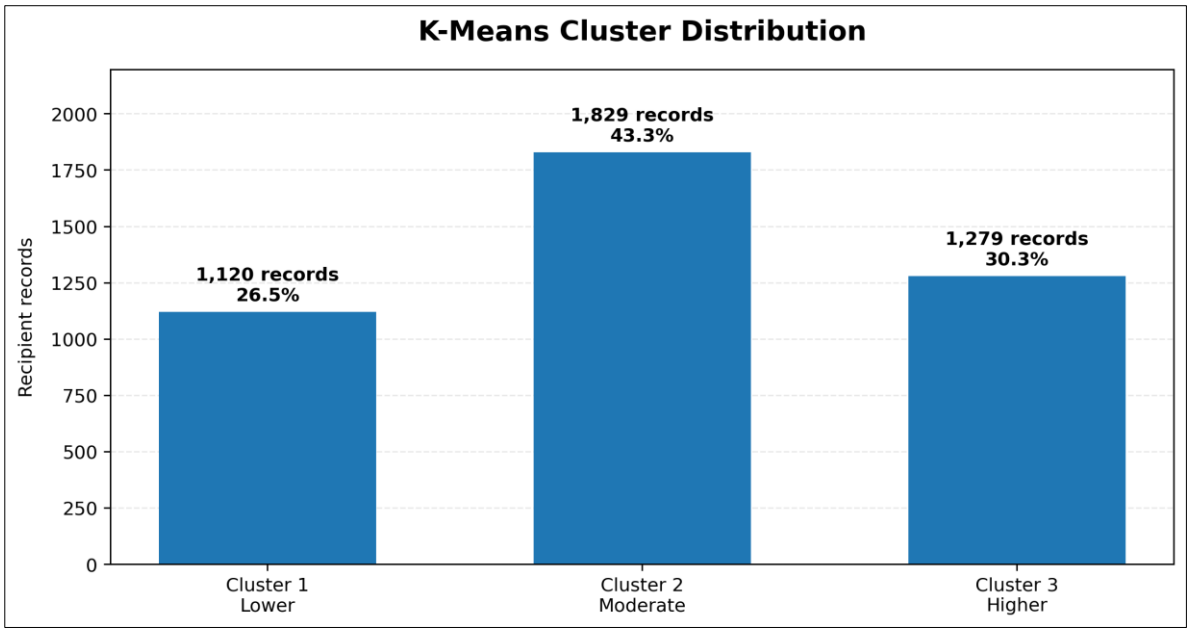



**Figure 4. Distribution of recipient records across K-Means clusters.**

Figure 4 shows the size and percentage of the lower, moderate, and higher allocation clusters. The labels above the bars help readers compare the frequency of each profile without needing to calculate percentages from the table. The chart describes grant-size patterns only and should not be interpreted as a direct measure of need or priority.

.

## Discussion

The results demonstrate that the K-Means algorithm can be used to organize educational grant recipient records into three distinct allocation clusters based on the nominal amounts of grants received during the 2023–2026 period. The resulting clusters represent lower, moderate, and higher grant allocation profiles, providing a structured overview of the distribution of financial assistance among recipient institutions. This grouping helps transform administrative records into interpretable information that can facilitate the examination of funding distribution patterns. Furthermore, the combination of annual descriptive statistics and cluster profiles provides complementary perspectives on the available data, allowing differences in recipient counts and grant amounts to be examined more systematically. However, the interpretation of these patterns must consider variations in annual record coverage and the possibility that a small number of large allocations may influence the overall distribution. Consequently, the cluster results should be understood as descriptive representations of grant allocation characteristics rather than definitive evidence of funding adequacy or institutional priority.

The findings also highlight the potential and limitations of applying unsupervised learning to educational grant management. K-Means provides a practical approach for identifying groups with similar numerical characteristics without requiring predefined recipient categories. Nevertheless, because the clustering process in this study primarily uses nominal grant amounts, the resulting groups cannot independently determine which institutions have the greatest educational needs or should receive future funding first. A comprehensive needs-based assessment would require additional variables, including student enrollment, classroom and infrastructure conditions, availability of educational facilities, socioeconomic characteristics, and existing institutional resources. Moreover, the differences in record counts across the four annual worksheets should be considered when interpreting changes in allocation patterns over time. Future research could incorporate these additional indicators, compare alternative clustering configurations, and evaluate cluster stability to develop a more comprehensive decision-support framework for educational grant allocation in North Sumatra Province.

## Conclusion

This study applied the K-Means clustering algorithm to analyze the characteristics of educational grant recipients in North Sumatra Province using administrative data from 2023 to 2026. The analysis grouped recipient records into three clusters representing lower, moderate, and higher grant allocation levels based on the nominal amounts received. The results provide a structured overview of grant distribution patterns and demonstrate the potential of K-Means to organize administrative funding data into more interpretable groups. The combination of annual descriptive statistics and cluster analysis also facilitates the examination of differences in recipient records and grant allocation characteristics across the observed period.

The findings indicate that K-Means can serve as an exploratory analytical tool to support the evaluation of educational grant distribution. However, the resulting clusters describe funding allocation characteristics and do not directly measure institutional needs or determine which schools should receive priority in future funding programs. Further research should incorporate additional indicators, such as student enrollment, infrastructure conditions, socioeconomic characteristics, and the availability of educational facilities, to develop a more comprehensive needs-based assessment. The integration of these indicators with clustering techniques could provide stronger analytical support for transparent, evidence-based, and more equitable educational grant allocation in North Sumatra Province.

## References


[1] H. W. Dhany, F. Izhari, and S. Komputer, "Jurnal ICT : Information and Communication Technologies, 15 (2) (2024) 48-54 Prediction analysis condition animal use algorithm (SVM+KNN)," 2024. [Online]. Available: www.ejournal.marqchainstitute.or.id/index.php/JICT

[2] Z. Sitorus, D. Prayogi, M. A. Rizko, A. G. Suteja, and M. R. Harahap, "Implementation of the Insertion Sort Algorithm to Sort Positive Integers in Ascending Order Using Flowgorithm," *Journal of Information Technology, computer science and Electrical Engineering (JITCSE)*, vol. 1, no. 3, pp. 323–328, 2024, doi: 10.30596/jitcse.

[3] K. Khairul, A. H. Nasyuha, A. Ikhwan, M. H. Aly, and A. Ahyanuardi, "Implementation of Multiple Linear Regression to Estimate Profit on Sales of Screen Printing Equipment," *JURNAL INFOTEL*, vol. 15, no. 2, pp. 55–61, Jun. 2023, doi: 10.20895/infotel.v15i2.934.

[4] M. S. Novelan and S. Aryza, "OPTIMIZATION CVRP WITH MACHINE LEARNING FOR IMPROVED CLASSIFICATION OF IMBALANCED DATA FOOD DISTRIBUTION," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 10, no. 4, pp. 917–925, Jun. 2025, doi: 10.33480/jitk.v10i4.6467.

[5] R. Farta Wijaya, F. Kurniawan, R. R. Putrai, A. Alvin, F. Sains, and D. Teknologi, "OPTIMIZING THE MANAGEMENT OF VILLAGE ACTIVITIES THROUGH INFORMATION SYSTEMS FOR TRANSPARENCY AND ACCOUNTABILITY IN PERTUMBUKAN VILLAGE, WAMPU DISTRICT," 2024.

[6] M. M. Maulana, A. I. Suroso, Y. Nurhadryani, and K. B. Seminar, "The Smart Governance Framework and Enterprise System's Capability for Improving Bio-Business Licensing Services," *Informatics*, vol. 10, no. 2, Jun. 2023, doi: 10.3390/informatics10020053.

[7] A. Putera Utama Siahaan, N. Azizah Harahap, R. Yuni Simanullang, and P. Wanny, "Analysis of Inpatient Data Using Cluster Analysis on Simulation Dataset," *Bulletin of Information Technology (BIT)*, vol. 6, no. 1, pp. 33–39, 2025, doi: 10.47065/bit.v5i2.1830.

[8] M. S. Novelan, S. Efendi, P. Sihombing, and H. Mawengkang, "VEHICLE ROUTING PROBLEM OPTIMIZATION WITH MACHINE LEARNING IN IMBALANCED CLASSIFICATION VEHICLE ROUTE DATA," *Eastern-European Journal of Enterprise Technologies*, vol. 5, no. 3(125), pp. 49–56, 2023, doi: 10.15587/1729-4061.2023.288280.

[9] M. Iqbal and S. Efendi, "Data-Driven Approach for Credit Risk Analysis Using C4.5 Algorithm," *ComTech: Computer, Mathematics and Engineering Applications*, vol. 14, no. 1, pp. 11–20, May 2023, doi: 10.21512/comtech.v14i1.8243.

[10] J. Asharf, N. Moustafa, H. Khurshid, E. Debie, W. Haider, and A. Wahab, "A review of intrusion detection systems using machine and deep learning in internet of things: Challenges, solutions and future directions," Jul. 01, 2020, *MDPI AG*. doi: 10.3390/electronics9071177.

[11] W. Cetera, W. Gogołek, A. Żołnierski, and D. Jaruga, "Potential for the use of large unstructured data resources by public innovation support institutions," *J. Big Data*, vol. 9, no. 1, Dec. 2022, doi: 10.1186/s40537-022-00610-6.

[12] G. Rajeshkumar *et al.*, "Smart office automation via faster R-CNN based face recognition and internet of things," *Measurement: Sensors*, vol. 27, Jun. 2023, doi: 10.1016/j.measen.2023.100719.

[13] A. Rejeb *et al.*, "The Internet of Things (IoT) in healthcare: Taking stock and moving forward," Jul. 01, 2023, *Elsevier B.V.* doi: 10.1016/j.iot.2023.100721.

[14] M. Overton, S. Larson, L. J. Carlson, and S. Kleinschmit, "Public data primacy: the changing landscape of public service delivery as big data gets bigger," *Global Public Policy and Governance*, vol. 2, no. 4, pp. 381–399, Dec. 2022, doi: 10.1007/s43508-022-00052-z.

[15] J. Ilmiah, E. & Erudisi, A. Zahara, M. I. Kabullah, and R. E. Putera, "JIEE: Effectiveness of the OSSRBA (Online Single Submission Risk Based Approach) System in Business Licensing Services in Payakumbuh City DPMPTSP," vol. 3, pp. 22–40, 2023.

[16] W. Jin, "Challenges and Innovative Countermeasures Faced by Public Administration in the Context of Big Data and Internet of Things," *Math. Probl. Eng.*, vol. 2022, 2022, doi: 10.1155/2022/8949365.