Bulletin of Computer Science and Electrical Engineering
Vol. 3, No. 2, December 2022, pp. 106~116
ISSN: 2722-7324, DOI: 10.25008/bcsee.v3i2.1163

Application of K-Means Clustering on School Identification in the Distribution of Assistance Funds for DPRD Members (Case Study in North Padang Lawas DPRD)

Eka Hayana Hasibuan[1], Aripin Rambe[2], Dinur Syahputra[3]
[1]Informatics Study Program, Faculty of Technology, Battuta University
[2]Information Technology Study Program, Faculty of Technology, Battuta University
[3]Informatics Study Program, Faculty of Technology, Battuta University

Article Info

Article history:

Received Dec 03, 2022
Revised Dec 15, 2022
Accepted Dec 24, 2022

Keywords:

Distribution DPRD
School.
K-Means Clustering
Software RapidMiner

ABSTRACT

In this study, the k-means algorithm was used to group schools and categorize DPRD grants into very feasible, feasible, and impractical categories for better focus. Based on the results of computational analysis using the K-Means clustering algorithm using the Euclidean distance equation for the distribution of DPRD suction subsidies from 52 schools, 28 schools are in the very decent category and 11 schools are decent. In that category, 13 schools were found with fewer categories. executable category. RapidMiner Studio v.7.6 software can group schools based on the distribution needs of DPRD suction tools for more effective and efficient results.

Corresponding Author:

Eka Hayana Hasibuan,
Informatics Program & Technology Information Program,
Battuta University,
Adress. Sekip, Adress. Simpang Sikambing Medan City
Email: ekahayana@man1medan.sch.id

1. INTRODUCTION

Education is the foundation and most important factor in human character and personality formation. According to current norms and regulations, education has an important task and role in shaping human character and the good and bad of personality. Faced with these problems, the government has endeavored to improve the education system as much as possible by implementing various curriculums and educational programs. For the country. Therefore, in education, in addition to the role of teachers as part of the educational process, the importance of building infrastructure and containers as an educational institution is also at the forefront. Educational institutions, including educational facilities and infrastructure, are a place for knowledge exchange and daily learning activities. Some of the government's obligations and functions related to education are closely related. When studying the above issue that the importance of aid allocation is the focus of government, especially North Padang Lawas DPRD, the K-means clustering algorithm is a good way to classify and infer expected aid allocation. Knowledge discovery in Database (KDD) is a way to get knowledge from an existing database.

The database has tables that are related to each other. The results of the knowledge gained can be used as a knowledge base for decision-making purposes (Mardi, 2016). In another sense, Database

Knowledge Discovery (KDD) is the application of scientific methods to data mining. In this context, data mining is considered a step in the KDD process (Defit, 2013). Data mining is a new technology with tremendous potential to allow businesses, marketing, etc. to focus on the most important information in master data. Data mining is the process of extracting data from a large database to reveal hidden information. Eko Prasetyo (2012) proposes in his study that there are four working groups related to data mining. The first is predictive modeling, the second is group analysis (cluster analysis), and the third is anomaly detection (anomaly detection). And related analysis (related analysis). Cluster analysis is a decision support technique used to collect objects into multivariate groups based on the properties of the objects (Mulyati, 2015). K-means clustering [x] is one of the well-known clustering techniques aimed at finding a specific number of clusters (Ngamsuriyaroj and Thepsutum, 2017). K-Means is an unsupervised machine learning algorithm that forms a series of clusters.

As input (Shetty and Kallimani, 2017). Widely used in various fields such as text mining, machine learning, image analysis, image processing, web cluster engine, bioinformatics, and meteorological analysis reports (Garg and Rani, 2017). K-Means clustering is a method of classifying or grouping groups of objects into several k_{groups} (number of positive integers) according to the same attributes or characteristics (Penangsang, Putra, Kurniawan, 2017). Clusters are defined by the average mass of the clusters (Ngamsuriyaroj and Thepsutum, 2017). In data mining, a common cluster analysis is K-means. This is a vector quantization method (Shetty and Kallimani, 2017). This is due to the problem I found in this area: the efforts of members of the DPRD state of Padang Lawas Utara, where data mining techniques used the K-Means clustering algorithm to collect aid funds and group them into several groups. Consistent with the distribution of aid funds. Category, that is, very feasible, feasible, and low value.

The Regional Revenue and Expenditure Budget (APBD) is a financial plan for implementing finalized and ratified allocation of funds (Permendagri No. 13 of 2006). The plenary session held at the time of the establishment of APBD was attended by the local government and the Local People's Congress (DPRD), and it is expected that each village will receive a part of APBD and be ratified from the results of the APBD debate. rice field. The guidelines helped support progress and local needs in various aspects of both infrastructure and welfare. APBD is a tool used as a tool aimed at improving service to both the community and the general public in the community. The ratified APBD is expected to explain the needs and capabilities of each region, depending on the income, uniqueness and potential of each region. Even in this APBD, schools distributed in all regions, districts and cities are central to the distribution of APBD funds as a place of education. Since one of the missions of Committee V working for welfare is education, Committee V plays an important role in distributing aid funds to each school. To help identify schools that will distribute funds to members of the Padang Lawas Utara DPRD, we need a way to rank and compare schools that are eligible for assistance. Several studies have been done so far to analyze data, images, visuals, and more. In this study, we proposed a data processing method for grouping data from school data using the K-Means algorithm and allocating aid funds. The findings are compared to identification using the K-Means clustering algorithm, so it is hoped that the support received from the school will be accurate.

2. RESEARCH METHOD

This survey method is systematically performed to explain the process of implementing the survey framework. Several stages are described in the form of a picture frame, which can be traced from needs analysis to the results of this study. The research phases: data needs analysis, data collection process, data analysis using the K-Means algorithm, data processing using RapidMiner, and the results of this research. Based on the framework, we will walk you through the steps of this study. The following are the steps that will be carried out in this research method:

a) Identifying the problem
b) Analyzing data needs
c) Studying literature
d) Collecting data
e) Analyzing problems

f) Testing using K-Means cluster clusters

As for the concepts and data processing stages in this K-means cluster as outlined in the figure 1 below.

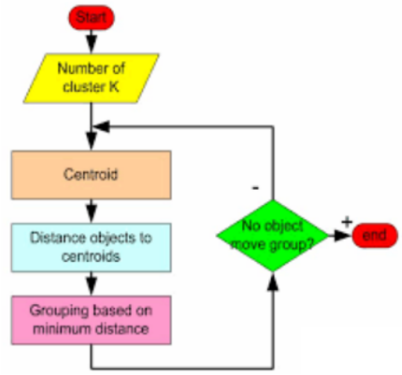


Figure 1. Flowchart K-Means Clustering

Let's say we have 4 objects as training data points and each object has 2 attributes.
Each attribute represents the coordinates of the object:
Object Attribute 1 (X): index weight
Object Attribute 2 (Y): pH

3. RESULTS AND DISCUSSION

3.1. Analysis and Design Phase

Table 1. Data on the Distribution of DPRD Aspiration Funds

No	School Name	Total Students	Total Study Group
1	SMAN 1 Batang Onang	861	27
2	SMAN 1 Dolok	804	24
3	SMAN 1 Halongonan	675	13
4	SMAN 1 Padang Bolak	420	9
5	SMAN 1 Padang Bolak Julu	490	14
6	SMAN 1 Portibi	246	7
7	SMAN 1 Simangambat	118	5
8	SMAN 2 Padang Bolak	63	3
9	SMAN 1 Dolok Sigoppulon	198	6
10	SMAN Nagasaribu	118	4
11	MAS Abu Bakar Sidik	420	9
12	MAS Al-Imron	356	8
13	MAS Al-Bahriyyah Purba Tua	114	4
14	MAS AL-Mukhtariyyah S. Dua	196	8
15	MAS Al-Yunusiyah	118	5
16	MAS Al-Yususiyah Sionggotan	377	14
17	MAS Baiturrahman	350	12
18	MAS Darul Ulum Nabundong	178	6
19	MAS Darul Ulum Sipaho	795	22
20	MAS Darul Ulum Kampung Banjir	722	19
21	MAS Darussalam Parmeraan	528	14
22	MAS Darussalam Siunggamjae	488	12
23	MAS Nurul Hidayah	388	8
24	MAS Nurul Iman	482	12
25	MAS Pubaganal Sosopan	812	20

3.2. Determining the Number of Clusters

In this study, 3 types of clusters were formed from data on the DPRD aspiration assistance funds, the clusters were which schools were very worthy, worthy, and less worthy to get DPRD aspiration

assistance funds. The recipients of the aspiration fund assistance are determined based on the number of students and the number of study rooms (rombel) from each school.

3.3. Determining the Cluster Center Point Randomly

The distance from each object to each centroid uses the correlation formula between three objects, namely Euclidean Distance (D). To determine An, it is taken from the number of students (X) and the number of study rooms (Y).

- All data will be grouped into three clusters
- Center points of the three clusters that are determined randomly are:
  Cluster center 0 (C0) : taken from number 23 (388, 8)
  Center of cluster 1 (C1) : taken from number 1 (861, 27)
  Center of cluster 2 (C2) : taken from number 7 (63, 3)

In this study, the factors that influence the recipients of DPRD aspiration funds are:

- Number of students
- Number of Study Rooms

To prove that the recipients of the DPRD aspiration funds are determined from what has been described previously, a tool is needed to prove it, namely the number of students and the number of study rooms. The experiment will be carried out using the following parameters:

Number of data : 25
Number of clusters : 3 (very feasible, decent, not feasible)
Number of attributes : 4 (School Name, Code, Number of students, and number of rooms study)

The sample data on the number of students and the number of study rooms are shown in Table 2 below:

Table 2. Random Centroid Value

Centroid	Total Students	Number of Study Rooms
C0	388	8
C1	861	27
C2	63	3

3.4. Calculating the Nearest Centroid

Calculate the distance of each data to each centroid using the correlation formula between three objects (Euclidean Distance).

$$D_{(i,j)} = \sqrt{(X_{1i} - X_{1j})^2 + (X_{2i} - X_{2j})^2 + \cdots + (X_{ki} - X_{kj})^2}$$

Where:
$D_{(i,j)=}$ distance of data to to center of cluster j
$X_{ki=}$ Data to i on attribute data to k
$X_{k=}$ The jth center of the kth attribute

3.5. Test Data

The system implementation phase is one of the phases of the systems development life cycle, the phase where the information system is designed to work. In this phase, several activities are carried out sequentially, starting from the implementation of the implementation plan, the implementation of implementation activities, and the implementation of follow-up. In order for your implementation to work and function as expected, you must first create an implementation plan. This plan is designed to manage the costs and time spent in the implementation phase. The implementation process and data analysis were carried out using the RapidMiner application program to test all data taken by the DPRD Secretariat of North Padang Lawas Utara from January to April 2020/2021. When using Microsoft Excel, it is converted to Excel application program document format.

3.6. RapidMiner Application Testing

There are several file extensions that you can import into RapidMiner. B. CSV files, Excel Sheen files, Access database table files, database tables, binary files. In this case, the data file to use is the data saved in Microsoft Excel with the extension.

1) Step 1 - Open the RapidMiner Application Double click the RapidMiner icon, as shown in Figure 1.

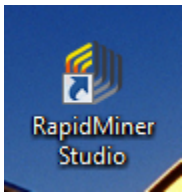


Figure 1. Icon RapidMiner

Then the Welcome Perspectives window will appear as shown in Figure 2.

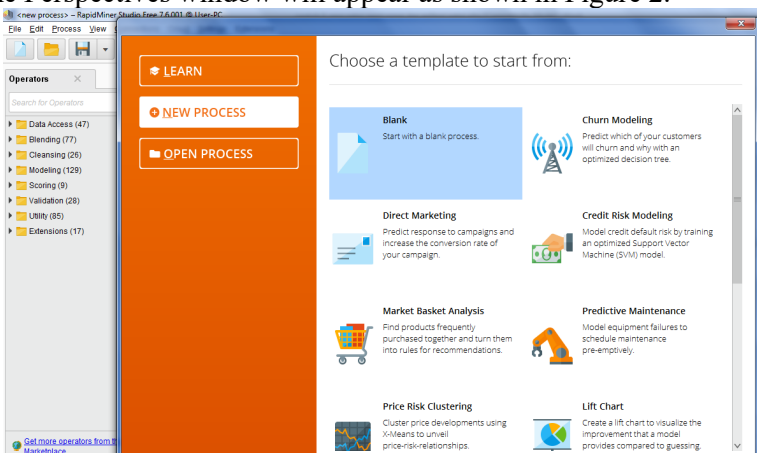


Figure 2. Window Welcome Perspectives

2) Step 2 -Creating a New Process
From the Welcome Perspectives window, click New Process or click Blank, as shown in Figure 3. Then the Design Perspectives Window will appear which is the RapidMiner work area, as shown in Figure 3.

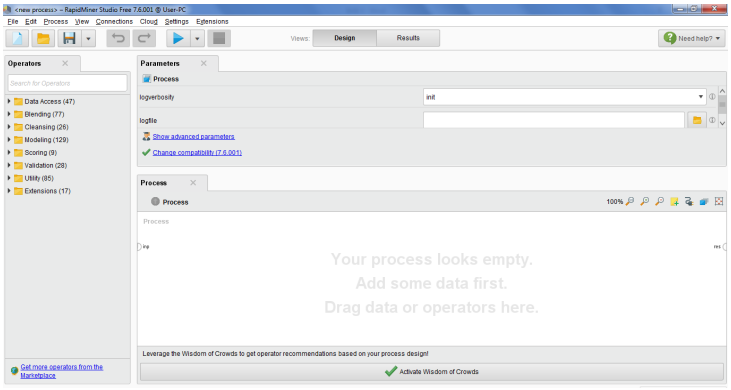


Figure 3. Window Design Perspectives

3) Step 3 - Import Data
Unlike most other Data Mining Tools, RapidMiner has its own advantages, namely that it can directly import files with the extension .xls or .xlsx, namely files from Ms. Excel. You do this by hovering over Operator and Drag and drop read excel to Main Process then clicking Add Data > My Computer as shown in Figure 4.

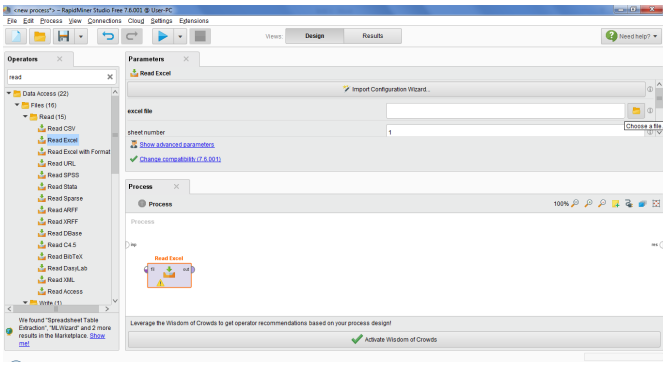


Figure 4. Steps to Import Data

Then hover over the excel file menu and click it, a window like Figure 5. will appear, continue by selecting the data to be processed.

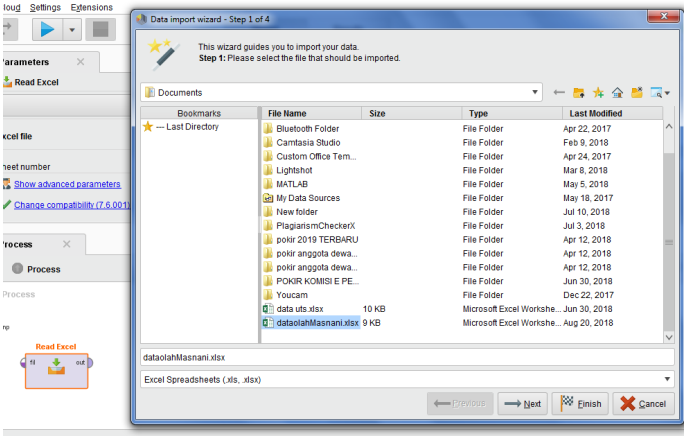


Figure 5. Step 1 of 3 Import Data Excel

Next in step 2 is, select the sheet to be inserted then click the Next button, as shown in Figure 5.6.

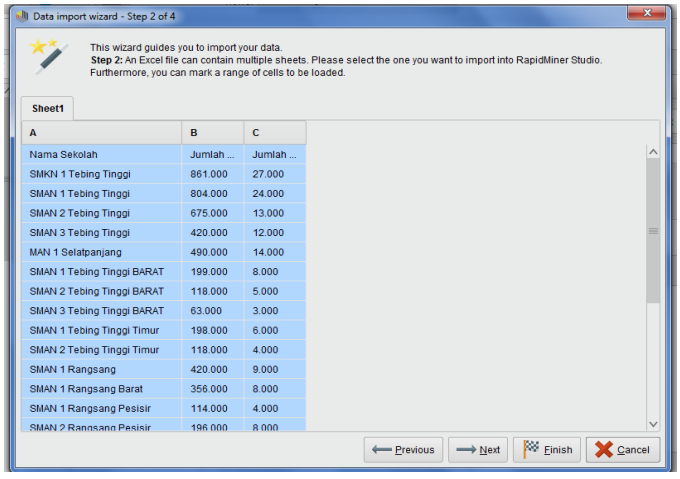
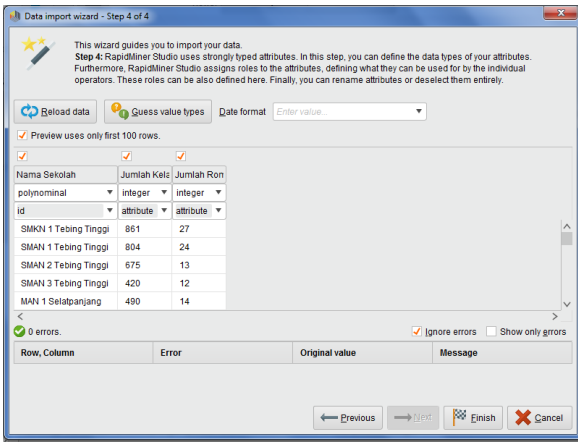


Figure 6. Step 2 of 3 Data Import Wizard

Step 3 is to give the data type to the table. RapidMiner will provide the right data type automatically. However, if you feel that the data type provided by RapidMiner is not suitable, the user can change it. Click the Finish button, as shown in Figure 7.



Figuren 7. Step 3 of 3 Data Import

Next do this step; in the views operator, select the Modeling > Segmentation > K-Means folder. Drag and drop K-Means to the Main Process, specify the number of clusters. K-means is a command to process data with the k-means clustering algorithm. After entering the k-means command, the next step is to assign a value to column k, which means cluster or category. Where in this research data for the category is 3, as in Figure 8.

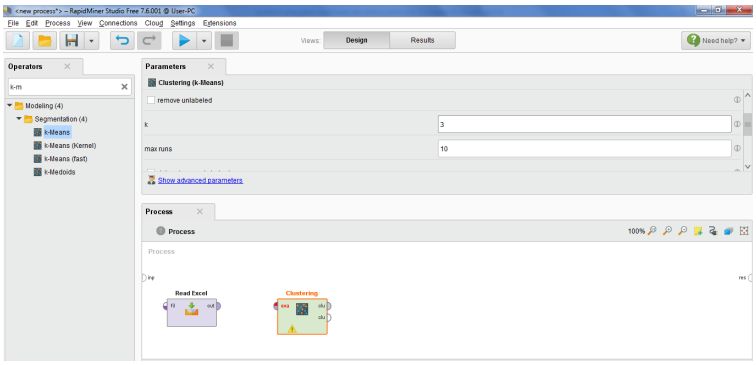


Figure 8. Step 2 of 3 Setting K-Means

Next on the views operator, select the Validation > Predictive > Segmentation > Cluster Distance Performance folder. Drag and drop Cluster Distance Performance to Main Process, connect output in repository to clustering direction then clustering to result. Finally click the Play button, as shown in Figure 9.

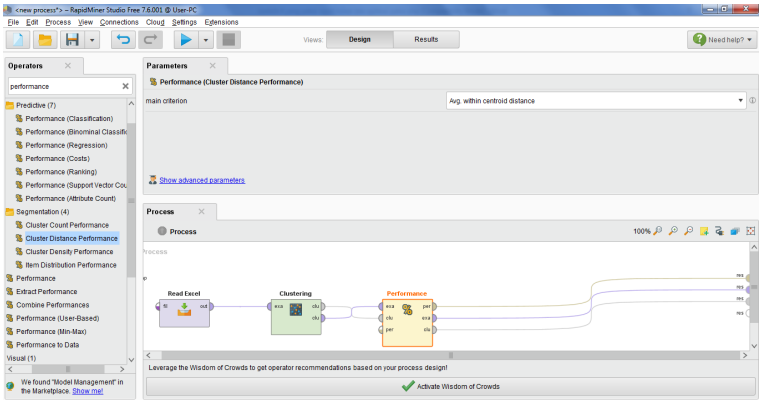


Figure 9. Step 3 of 3 Setting K-Means

After the process is running, the Results Perspectives window will appear, as shown in Figure 10.

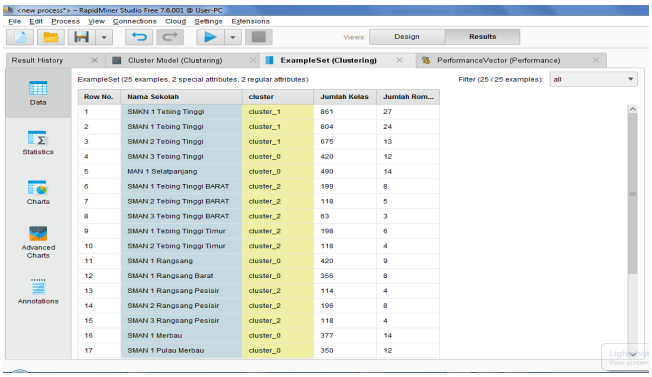


Figure 10. Window Results Perspectives

The Result Perspectives window provides several forms of cluster results such as; Folder View, Graph, Centroid Table and Plot.

3.7. Transaction Data Execution Results

The results of the execution using RapidMiner are:

a. Text View

The results of the global transaction data cluster from March 2020 to December 2021 can be seen in Figure 11.

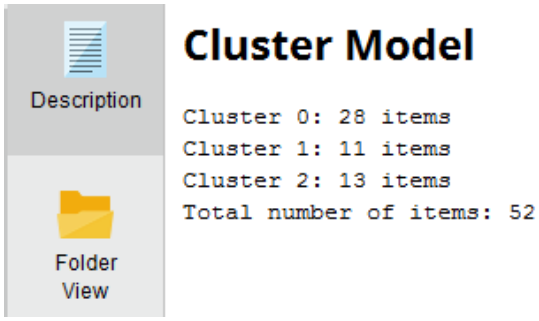


Figure 11. Cluster Model

In the form of a text view, it can be seen globally that the members of cluster 0 are 28 items, cluster 1 is 11 items, and cluster 2 is 13 items.

Figure 12 represents some of the members of cluster 0, where cluster 0 represents very worthy members in receiving the DPRD aspiration assistance funds.

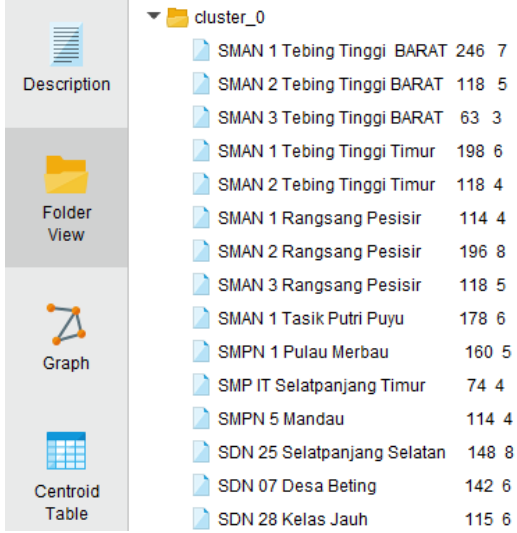


Figure 12. Folder View - Cluster 0

Figure 13. represents some of the members of cluster 1, where cluster 1 represents eligible members in receiving the DPRD aspiration assistance funds.

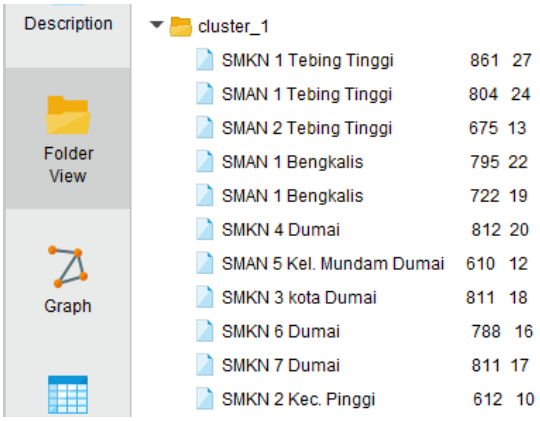


Figure 13. Folder View - Cluster 1

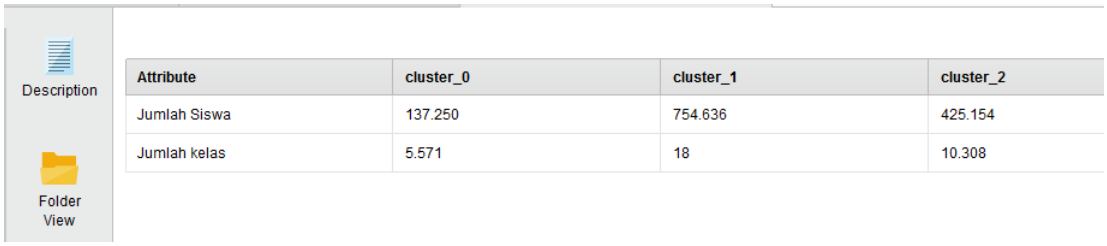
Figure 14. represents some of the members of cluster 2, where cluster 2 represents members who are less worthy in receiving the DPRD aspiration assistance funds.

Figure 14. Folder View - Cluster 2

In Folder View displays members who are in cluster groups c0, c1, and c2. This makes it easier for researchers to see the results of membership in the cluster.

b. Centroid Table

At the Centroid Table process stage, the results of the cluster category calculations will be displayed, which can be seen in Figure 5.16 for the number of students category with a distance value of cluster 0 is 137,250 and cluster 1 is 754,636, and cluster 2 is 425,154. while for the number of classes, the distance for cluster 0 is 5,571, cluster 1 is 18, and cluster 2 is 10,308. Centroid Table can be seen as Figure 15.



Attribute	cluster_0	cluster_1	cluster_2
Jumlah Siswa	137.250	754.636	425.154
Jumlah kelas	5.571	18	10.308

Figure15. Centroid Table

c. Charts Histogram

In order for the Histogram to give an accurate picture of the condition of the results of the distribution of aid, it is necessary to do accurate data processing first, starting with data collection, not less than 50 samples, namely the number that is considered to be able to meet the population to be observed. Data processing on histograms is very important, especially in determining the size of the mean (standard) and how many data classes will describe the spread of the data created. To view the histogram charts of schools receiving DPRD assistance funds, see Figure 16.

Figure 16. Charts

Seen in the chart above, the red line is for cluster 0 with the number of students below 250, the blue line is for cluster 1 with the number of students between 350 to 500, and for the green line is cluster 2 with the number of students above 600. Based on the results of Data Mining testing using the K-means Clustering method, testing was carried out with the RapidMiner 7.6 application for data processing of samples of schools receiving DPRD aspiration funds which resulted in three clusters. Where cluster 0 has 28 members, cluster 1 has 11 members, and cluster 2 has 13 members. This is no different from the manual processing of sample data from schools receiving DPRD aspiration funds, which resulted in 28 cluster members, 11 members for cluster 1, and 13 members for cluster 2.

4. CONCLUSION

Based on the research that has been done, the following conclusions can be drawn: The k-means algorithm can be used to group schools to get the DPRD's aspiration assistance funds into very decent, decent, and less feasible categories so that they are more targeted. Based on the results of the calculation analysis using the K-Means clustering algorithm using the Euclidean Distance formula for the distribution of DPRD aspiration assistance funds from 52 schools, 28 schools were found in the very decent category, 11 schools in the decent category, and 13 schools in the less feasible category. RapidMiner Studio v.7.6 software can group schools based on requirements for distribution of DPRD aspiration assistance funds with more effective and efficient results.

ACKNOWLEDGEMENTS

We thank those who recommended us and the employees who assisted us in managing school data that received assistance from the DPRD. The purpose of this research was made so that decision making in providing assistance to schools was in accordance with what was expected.

REFERENCES

[1] P. S. Bradley and U. M. Fayyad, "Refining initial points for k-means clustering.," in *ICML*, 1998, vol. 98, pp. 91–99.
[2] K. P. Sinaga and M.-S. Yang, "Unsupervised K-means clustering algorithm," *IEEE access*, vol. 8, pp. 80716–80727, 2020.
[3] S. Na, L. Xumin, and G. Yong, "Research on k-means clustering algorithm: An improved k-means clustering algorithm," in *2010 Third International Symposium on intelligent information technology and security informatics*, 2010, pp. 63–67.
[4] J. Xu and K. Lange, "Power k-means clustering," in *International conference on machine learning*, 2019, pp. 6921–6931.
[5] Y.-M. Cheung, "k∗-Means: A new generalized k-means clustering algorithm," *Pattern Recognit. Lett.*, vol. 24, no. 15, pp. 2883–2893, 2003.
[6] H. Zha, X. He, C. Ding, M. Gu, and H. Simon, "Spectral relaxation for k-means clustering," *Adv. Neural Inf. Process. Syst.*, vol. 14, 2001.
[7] I. Bin Mohamad and D. Usman, "Standardization and its effects on K-means clustering algorithm," *Res. J. Appl. Sci. Eng. Technol.*, vol. 6, no. 17, pp. 3299–3303, 2013.
[8] A. Likas, N. Vlassis, and J. J. Verbeek, "The global k-means clustering algorithm," *Pattern Recognit.*, vol. 36, no. 2, pp. 451–461, 2003.
[9] K. Alsabti, S. Ranka, and V. Singh, "An efficient k-means clustering algorithm," 1997.
[10] D. T. Pham, S. S. Dimov, and C. D. Nguyen, "Selection of K in K-means clustering," *Proc. Inst. Mech. Eng. Part C J. Mech. Eng. Sci.*, vol. 219, no. 1, pp. 103–119, 2005.
[11] P. S. Bradley, K. P. Bennett, and A. Demiriz, "Constrained k-means clustering," *Microsoft Res. Redmond*, vol. 20, no. 0, p. 0, 2000.
[12] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *J. R. Stat. Soc. Ser. c (applied Stat.*, vol. 28, no. 1, pp. 100–108, 1979.
[13] J. Burkardt, "K-means clustering," *Virginia Tech, Adv. Res. Comput. Interdiscip. Cent. Appl. Math.*, 2009.
[14] K. A. A. Nazeer and M. P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm," in *Proceedings of the world congress on engineering*, 2009, vol. 1, pp. 1–3.
[15] L. Morissette and S. Chartier, "The k-means clustering technique: General considerations and implementation in Mathematica," *Tutor. Quant. Methods Psychol.*, vol. 9, no. 1, pp. 15–24, 2013.
[16] D. Steinley, "K-means clustering: a half-century synthesis," *Br. J. Math. Stat. Psychol.*, vol. 59, no. 1, pp. 1–34, 2006.
[17] T. M. Kodinariya and P. R. Makwana, "Review on determining number of Cluster in K-Means Clustering," *Int. J.*, vol. 1, no. 6, pp. 90–95, 2013.
[18] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, 2002.
[19] K. Teknomo, "K-means clustering tutorial," *Medicine (Baltimore).*, vol. 100, no. 4, p. 3, 2006.
[20] K. Wagstaff, C. Cardie, S. Rogers, S. Schrödl, and others, "Constrained k-means clustering with background knowledge," in *Icml*, 2001, vol. 1, pp. 577–584.
[21] B. Harahap, "Penerapan Algoritma K-Means Untuk Menentukan Bahan Bangunan Laris (Studi Kasus Pada UD. Toko Bangunan YD Indarung)," *Ready Star*, vol. 2, no. 1, pp. 394–403, 2019.
[22] B. Harahap and A. Rambe, "Implementasi K-Means Clustering Terhadap Mahasiswa yang Menerima Beasiswa Yayasan Pendidikan Battuta di Universitas Battuta Tahun 2020/2021 Studi Kasus Prodi Informatika," *Informatika*, vol. 9, no. 3, pp. 90–97, 2021, doi: 10.36987/informatika.v9i3.2185.
[23] "K-Means Analysis in Grouping Abilities of Battuta University Informatics Study Program Students".