Unsupervised Data Mining with K-Medoids Method in Mapping Areas of Student and Teacher Ratio in Indonesia

Adya Hermawati [1], Sri Jumini [2], Mardiah Astuti [3], Fajri Ismail [3], Robbi Rahim [4]

[1]*Universitas Widyagama Malang, Kota Malang, Indonesia*
[2]*Universitas Sains Al-Qur'an, Wonosobo, Indonesia*
[3]*Universitas Islam Negeri Raden Fatah Palembang, Palembang, Indonesia*
[4]*Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia*

Abstract **– The purpose of this study was to analyze the k-medoids method in conducting cluster mapping in the ratio of the number of students and teachers in Indonesia by region, especially at the elementary school level. The data source is secondary obtained from the Ministry of Education and Culture which is processed by the Central Statistics Agency (abbreviated as BPS) in the BPS Catalog: 4301008 concerning the Portrait of Indonesian Education. The analysis process uses the help of Rapid Miner software by using parameters of the Davies Bouldin Index (DBI) and Performance (Classification). By using three cluster labels, namely the high cluster (K1), normal cluster (K2) and poor cluster (K3), it was found that 3 provinces were in the high cluster, 9 provinces were in the normal cluster and 22 provinces were in the fewer clusters. By testing the cluster results (k = 3) through the DBI parameter the value = 0.587 was obtained. This shows that the results of the cluster formed are optimal (the smaller the better). The test results with the parameter Performance (Classification) show the results of classification error = 2.50%. The results of the research can be used as information to determine the ratio of students and teachers because the higher the value of this ratio means that the level of teacher supervision and attention to students is reduced so that the quality of teaching tends to be lower.**

DOI: 10.18421/TEM94-37
https://doi.org/10.18421/TEM94-37

Corresponding author: Robbi Rahim,
Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia.
Email: usurobbi85@zoho.com

Received: 20 August 2020.
Revised: 09 October 2020.
Accepted: 15 October 2020.
Published: 27 November 2020.

Keywords **– Data Mining, k-medoids method, Davies Bouldin Index, student and teacher, ratio.**

1. Introduction

In Indonesia, the ratio of teachers to students is said to be ideal compared to developed countries such as South Korea, Japan, and Malaysia. For the ratio of teachers to students in Indonesia 1:14, while South Korea 1:30, Malaysia 1:25, and Japan 1:20. But the problem is the unequal distribution because there are many teachers in urban schools, while in rural areas there is still a shortage of teachers because the ratio of teachers to students is a picture of the workload of teachers in teaching and seeing the quality of teaching in class. The higher the value of this ratio, the lower the level of teacher supervision and attention to students so that the quality of teaching tends to be lower. The purpose of this research is to analyze regions in Indonesia in the form of a mapping in the form of clusters to the ratio of teachers to students at the primary school level. In this case the ratio of teachers and students has been regulated in paragraph 1 article 17 of Law Number 74 in 2008 concerning Teachers and Lecturers, which states that for the Elementary School level or equivalent is 1:20 [1].

Many techniques in computer science work for mapping. One of them is data mining [2]. Data mining is one of the Unsupervised Learning techniques where the expected results cannot be known by anyone [3]. Data processed by data mining techniques creates new information from old data, and data generation results can be used to decide a decision in the future [4]. In other words, data mining is a learning method that is suitable for finding patterns of many objects in the form of clusters that are not entirely the same [5]. Data mining methods are quite popular in business, academia, and industry, one of which is k-medoids [6]. The following is an illustration of Unsupervised Learning techniques with clustering techniques as shown in the following figure:

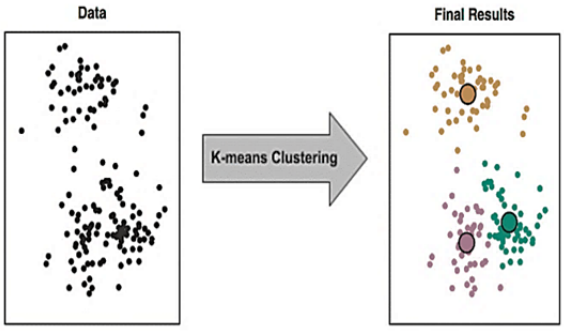


Figure 1. Unsupervised Learning techniques in data mining clustering

Based on Aishwaya's research on [7], it was concluded that the k-means method is efficient for small and large datasets and K-Medoids has better performance than k-means for small and large datasets. In addition, the k-medoids method is used to solve the drawback of the technique of k-means, which is very susceptible to outliers since these items are far away. [8].

The advantages of the k-medoids method were also applied to the research conducted [6]. This paper proposes two of the most popular clustering algorithms k-means and k-medoids to be evaluated on the KEEL 10k transaction dataset. The comparison results show that k-medoids are much better than K-Means both in terms of execution time, and they are insensitive to outliers and reduce noise. Next research is also conducted [2]. This paper proposes a k-medoid clustering algorithm by utilizing a measure of similarity in the form of vectors compared to traditional clustering algorithms. The results show that the k-medoids algorithm with the latest measure of similarity outperforms the clustering of k-means for the mixed data set. In addition, analysis was conducted [9]. This paper proposes a k-medoids method to measure performance in the case of applicants for Student Learning Assistance (BBM) scholarships. The results show that the K-Medoids algorithm is more suitable for use in datasets with the entire encoded attribute format. Based on these advantages, it is hoped that the research results can provide information in the form of cluster mapping on the ratio of students to teachers at the elementary school level because the higher the value of this ratio means the less level of supervision and attention of teachers towards students so that the quality of teaching tends to be lower.

2. Methodology

2.1. Data Mining

Data determines a decision in the future, the results of data processed using data mining techniques can be done by generating new knowledge that comes from old data [10] Data mining processing consists of predictive classification, modelling, classification, and association [11]. Clustering is often done as the first step in the data mining process. There are many clustering algorithms that have been used by previous researchers such as K-Means, Improved K-Means, K-Medoids (PAM), Fuzzy C-Means, DBSCAN, CLARANS and Fuzzy Subtractive [7].

2.2. K - Medoids Method

The k-medoids method uses the object as a representative cluster center (medoid) for each cluster and is suitable for grouping data compared to the k-means method [12].

2.3. Data

Research on the mapping of areas from teacher to student ratios used secondary data sources obtained from the Ministry of Education and Culture processed by the Central Statistics Agency (abbreviated as BPS) in the BPS Catalog: 4301008 on Portraits of Indonesian Education. The analysis process uses the help of Rapid Miner software by using parameters of the Davies Bouldin Index (DBI) and Performance (Classification). The cluster uses three cluster labels, namely the high cluster (K1), the normal cluster (K2) and the fewer clusters (K3). Before determining the number of clusters formed, they are first tested with DBI to see the value of the cluster formed. The cluster value formed is the key to mapping. The results of the mapping in the form of clusters will be tested with Performance (Classification) to see the results of the error (%). The smaller the error value, the better. The following is the research design as shown in the following figure:

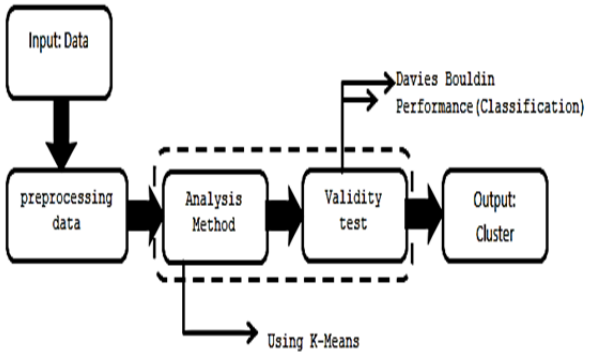


Figure 2. Research workflow

3. Results and Discussion

The research data used is the ratio of teachers and students at the primary school level for the 2018/2019 academic year as shown in the following Table:

Table 1. Research data

Province	Student and Teacher Ratio
Aceh	10
North Sumatra	16
West Sumatra	14
Riau	16
Jambi	14
South Sumatra	16
Bengkulu	14
Lampung	15
Kep. Bangka Belitung	18
Kep. Riau	17
DKI Jakarta	20
West Java	21
Central Java	16
DI Yogyakarta	14
East Java	14
Banten	21
Bali	15
West Nusa Tenggara	13
East Nusa Tenggara	14
West Kalimantan	15
Central Kalimantan	11
South Borneo	13
East Kalimantan	16
North Kalimantan	13
North Sulawesi	12
Central Sulawesi	12
South Sulawesi	13
Southeast Sulawesi	13
Gorontalo	14
West Sulawesi	12
Maluku	13
North Maluku	14
West Papua	17
Papua	24

source: BPS processed data

Before conducting cluster mapping, determining the number of clusters was carried out using the Davies Bouldin Index (DBI) parameter. The DBI results become a reference for the number of clusters used. The following is the result of the number of clusters with DBI parameters

Table 2. the results of the comparison of k values

Cluster	DBI value
K=2	0.705
K=3	0.587

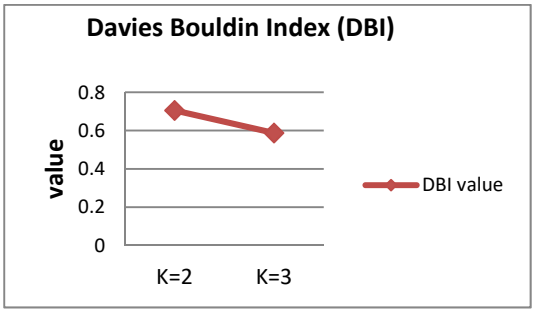


Figure 3. Comparison Chart

In Figure 3 the graph showing the comparison of the value (k) with the DBI parameter becomes a reference if the value is getting smaller. If the DBI value gets smaller, the resulting cluster will be better and optimal. For the value of k = 3, it is much better than the value of k = 2. Then the cluster used is k = 3 with a value of 0.587. The following is a research design using Rapidminer software to map the ratio of teachers to elementary school students in Indonesia.

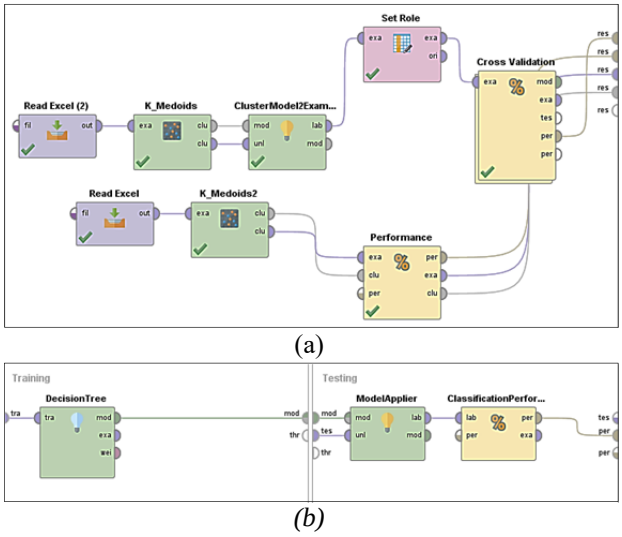


Figure 4. The k-medoids model in the RapidMiner design (a)(b)

In Figure 4 (a) the data obtained from BPS is processed using Microsoft Excel. At this stage the data has passed from preprocessing data. The input data used is (.xls). By using the k-medoids method with the number of clusters (k = 3), a cluster mapping process will be generated (after being tested with the Davies Bouldin Index parameter). In Figure 4 (b), the results of the clusters formed are tested with performance parameters (classification) to see the error results (%) of the formed clusters. The following are the results of the k-medoids method cluster on the ratio of teachers to students by province as shown in the following figure:

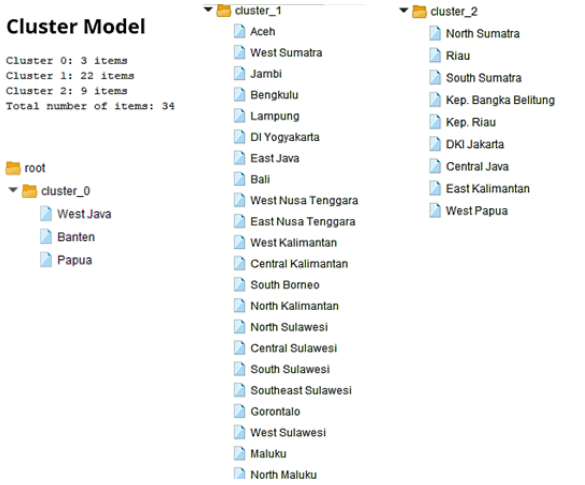


Figure 5. Results of k-medoids

In table 5, it can be explained that the results of the mapping are in the form of clusters, namely 3 provinces are in the high cluster (Group_0), 9 provinces are in the normal cluster (Group_2) and 22 provinces are in the less cluster (Group_1). List of clusters can be seen in Figure 5 above. The determination of high, normal and under-clusters is done by looking at the final result of the centroid as shown in the following Figure:

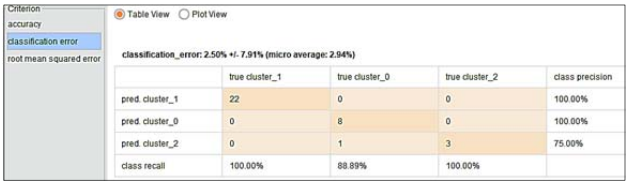
Figure 6. The final centroid results

Following are the complete results of the clusters that have been exported from Rapid Miner to Excel, namely high cluster (Group_0), normal cluster (Group_2) and fewer clusters (Group_1).

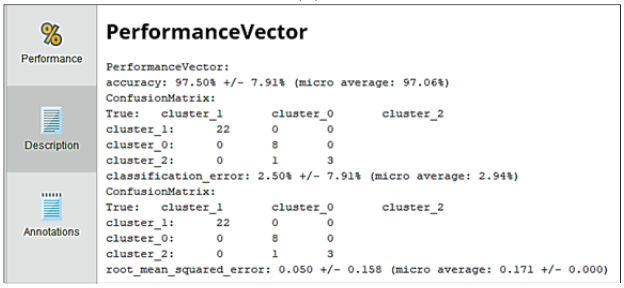
Table 2. The results of the RapidMiner export file to Excel

Ratio	Province	Group
10	Aceh	Group_1
16	North Sumatra	Group_2
14	West Sumatra	Group_1
16	Riau	Group_2
14	Jambi	Group_1
16	South Sumatra	Group_2
14	Bengkulu	Group_1
15	Lampung	Group_1
18	Kep. Bangka Belitung	Group_2
17	Kep. Riau	Group_2
20	DKI Jakarta	Group_2
21	West Java	Group_0
16	Central Java	Group_2
14	DI Yogyakarta	Group_1
14	East Java	Group_1
21	Banten	Group_0
15	Bali	Group_1
13	West Nusa Tenggara	Group_1
14	East Nusa Tenggara	Group_1
15	West Kalimantan	Group_1
11	Central Kalimantan	Group_1
13	South Borneo	Group_1
16	East Kalimantan	Group_2
13	North Kalimantan	Group_1
12	North Sulawesi	Group_1
12	Central Sulawesi	Group_1
13	South Sulawesi	Group_1
13	Southeast Sulawesi	Group_1
14	Gorontalo	Group_1
12	West Sulawesi	Group_1
13	Maluku	Group_1
14	North Maluku	Group_1
17	West Papua	Group_2
24	Papua	Group_0

The results of the cluster formed will be tested with the Performance (Classification) parameter, showing the results of the classification error (%) as it is shown below:



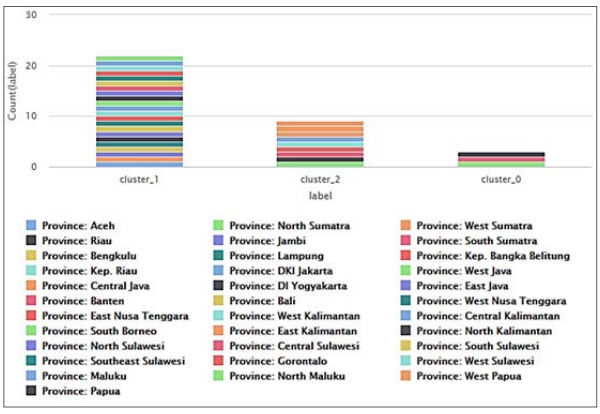
(a)



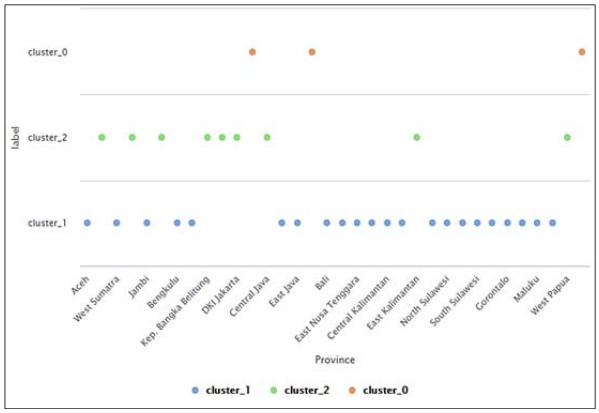
(b)

Figure 7. Result of Performance (Classification)

In Figure 7 it can be explained that classification error = 2.5% where Group_0 incorrectly predicts 9 data into 8 data so that the predication result for Group_0 is 88.89%. Overall, the accuracy value for the three clusters is 97.5%. The following is a cluster graph using chat bars and a scatter plot on the ratio of teachers to elementary school students.



(a)



(b)

Figure 8. Visualization of clustering results with scatter bar chat and plotter

4. Conclusion

Based on the results of the study, it can be explained that the application of the k-medoids method can be applied to the cluster mapping of teacher and student ratios at the elementary school level by maximizing the parameters of the Davies Bouldin Index (DBI) and Performance (Classification). The results show that 65% of provinces in Indonesia have a poor teacher to student ratio. This means that the government has to rethink how to distribute the ratio of teachers to students in both urban and rural areas.

References

[1]. Perdana, N. S. (2018). Analisis Capaian Rombongan Belajar di Provinsi Lampung Tahun 2018 Dalam Upaya Implementasi Permendikbud Nomor 17 Tahun 2017. *Jurnal Dewantara*, *5*(01), 1-16.

[2]. Harikumar, S., & Surya, P. V. (2015). K-medoid clustering for heterogeneous datasets. *Procedia Computer Science*, *70*, 226-237. https://doi.org/10.1016/j.procs.2015.10.077.

[3]. Dutt, A., Ismail, M. A., & Herawan, T. (2017). A systematic review on educational data mining. *Ieee Access*, *5*, 15991-16005. https://doi.org/10.1109/ACCESS.2017.2654247.

[4]. Zhang, C., & Xia, S. (2009, January). K-means clustering algorithm with improved initial center. In *2009 Second International Workshop on Knowledge Discovery and Data Mining* (pp. 790-792). IEEE. https://doi.org/10.1109/WKDD.2009.210.

[5]. Sun, D., Fei, H., & Li, Q. (2018). A Bisecting K-Medoids clustering Algorithm Based on Cloud Model. *IFAC-PapersOnLine*, *51*(11), 308-315. https://doi.org/10.1016/j.ifacol.2018.08.301.

[6]. Arora, P., & Varshney, S. (2016). Analysis of k-means and k-medoids algorithm for big data. *Procedia Computer Science*, *78*, 507-512. https://doi.org/10.1016/j.procs.2016.02.095.

[7]. Marlina, D., Fernando, A., & Ramadhan, A. (2018). Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak. *Jurnal CoreIT: Jurnal Hasil Penelitian Ilmu Komputer dan Teknologi Informasi*, *4*(2), 64-71. https://doi.org/10.24014/coreit.v4i2.4498.

[8]. Wira, B., Budianto, A. E., & Wiguna, A. S. (2019). Implementasi Metode K-Medoids Clustering Untuk Mengetahui Pola Pemilihan Program Studi Mahasiswa Baru Tahun 2018 Di Universitas Kanjuruhan Malang. *Rainstek: Jurnal Terapan Sains & Teknologi*, *1*(3), 53-68.

[9]. Defiyanti, S., Jajuli, M., & Rohmawati, N. (2017). K-medoid algorithm in clustering student scholarship applicants. *Scientific Journal of Informatics*, *4*(1), 27-33. https://doi.org/10.15294/sji.v4i1.8212.

[10]. Atmaja, E. H. S. (2019). Implementation of k-Medoids Clustering Algorithm to Cluster Crime Patterns in Yogyakarta. *International Journal of Applied Sciences and Smart Technologies*, *1*(1), 33-44. https://doi.org/10.24071/ijasst.v1i1.1859.

[11]. Waluyo, A., Jatnika, H., Permatasari, M. R. S., Tuslaela, T., Purnamasari, I., & Windarto, A. P. (2020, June). Data Mining Optimization uses C4. 5 Classification and Particle Swarm Optimization (PSO) in the location selection of Student Boardinghouses. In *IOP Conference Series: Materials Science and Engineering* (Vol. 874, No. 1, p. 012024). IOP Publishing.

[12]. Senduk, F. R., Indwiarti, I., & Nhita, F. (2019). Clustering of earthquake prone areas in indonesia using k-medoids algorithm. *Indonesia Journal on Computing (Indo-JC)*, *4*(3), 65-76. https://doi.org/10.21108/indojc.2019.4.3.359.