Ilkogretim Online - Elementary Education Online, 2021; 20 (1): pp. 691-698
http://ilkogretim-online.org
doi: 10.17051/ilkonline.2021.01.59

Educational Data Mining to Improve Decision Support on the Ratio of students and Study Groups in Elementary Schools in Indonesia using K-Means Method

Etika Kartikadarma, *Universitas Dian Nuswantoro, Semarang, Indonesia*
Sri Jumini, *Universitas Sains Al-Qur'an, Wonosobo, Indonesia*
Nurulisma Binti Hj. Ismail, *School of Computer and Communication Engineering, Universiti Malaysia Perlis, Malaysia*
Barany Fachri, *Universitas Pembangunan Panca Budi, Medan, Indonesia*
Dadang Sudrajat, *STMIK IKMI CIREBON, Cirebon, Indonesia*
***Robbi Rahim,** *Sekolah Tinggi Ilmu Manajemen Sukma, Medan, Indonesia*, usurobbi85@zoho.com
*Corresponding Author

Abstract. In the world of education, the process of processing space in learning activities is very important. This is in order to prevent unwanted behavior and can direct student activities. The purpose of the study was to analyze whether the mapping of regions (provinces) on the ratio of students to study groups (abbreviated as *rombel*) could be done by utilizing artificial intelligence techniques. The data source was obtained from the Ministry of Education and Culture which was processed by the Central Statistics Agency (abbreviated as BPS) for the academic year 2018/2019 which consisted of 34 records. The research is aimed at student to class ratios at the primary school level. The technique used is the k-means method which is part of data mining. The analysis process was carried out with the help of Rapid Miner software. Two clusters of mapping labels were used, namely the largest student-to-class ratio cluster (K1) and the smallest student-to-class ratio cluster (K2). The analysis results show that 14 provinces are in the largest cluster (K2) with an average per class = 24.2 and 20 provinces are in the smallest cluster (K2) with an average per class = 19.1. The validity test was carried out by testing the cluster results (k = 2) with Davies Bouldin was 0.570. The cluster results are optimal. The validity test was also conducted on the results of the cluster with Performance (Classification) with the results of classification error: 0.00%. The results of the analysis can be used as input for the government in making policies so that the quality of human resources is increasingly competitive and can compete regionally and internationally.

Keywords: Data Mining, k-means method, clustering, student ratio, Indonesia.

Received: 03.11.2020 Accepted: 14.12.2020 Published: 05.01.2021

INTRODUCTION

Human Resources are a very important factor in supporting economic growth [1]. In Indonesia, increasing human resources is always carried out to produce quality education graduates who have high quality competence and productivity [2,3]. One of the efforts to improve the quality of graduates is through an effective learning atmosphere for student participants. This has been regulated in the Government's policy in creating an effective learning atmosphere by issuing *Permendikbud* Number 22 of 2016 concerning Basic and Secondary Education Process Standards limiting the number of study groups (abbreviated as "rombel") in each educational unit and the number of students per study group[4] Development of media kocerin (Smart box interactive) to learning mathematics in Junior High School,. In the research conducted by Finn and Achilles in [2] states that small classes improve learning behavior and result in fewer class disruptions and disciplinary problems. Based on data sources from the Central Bureau of Statistics (BPS) in the BPS Catalog: 4301008 on the Portrait of Indonesian Education, mapping of regions in Indonesia regarding the ratio of students per class for primary school level to create an effective teaching and learning atmosphere and can improve the quality of graduates. There are many techniques that can be used for mapping [5,6]. One of them is Data Mining [7]. Several techniques that can be used for mapping are classification (C4.5, ID3) [8] and clustering (k-means) [9].

The purpose of this research is to map the regions in Indonesia regarding the ratio of the number of students per study group to create an effective teaching and learning atmosphere and to improve the quality of graduates. K-means is a clustering method that minimizes variations between existing data in a cluster and maximizes variations with existing data in other clusters [10]. Several studies that use the advantages of the k-means method are [9] on K-Means Clustering With Incomplete Data. The researcher proposes a new K-means-based clustering method that unites grouping and imputation into a single objective function. The experimental results have clearly demonstrated the effectiveness of the method that outperforms some commonly used methods for incomplete data. Next [11], In this paper, a new method is proposed to find a better initial centroid and to provide an efficient way of assigning data points to suitable clusters with reduced time complexity. The experimental results of the proposed method have much more accuracy with less computation time. Based on these advantages, it is hoped that the k-means method can map the number of students per study group in Indonesia so that the results of the analysis can be input to the government in making policies so that the quality of human resources is more competitive and can compete regionally and internationally.

METHODOLOGY

Data Mining

The related explanation of data mining is the discovery of hidden patterns and modeling using several techniques that can thoroughly explore complex relations in large sets. Data sets can be tabulated and can be implemented into other representations such as multimedia, text, images and spatial data [10]. Several data mining methods have been used by previous researchers such as K-Means, Improved K-Means, K-Medoids (PAM), Fuzzy C-Means, DBSCAN, CLARANS and Fuzzy Substractive [12].

K-Means Method

The k-means method is the simplest and most widely used method for separating datasets into "k" groups [9]. The goal is to separate objects into several groups of different characteristics from one group to another [10]. The k-means method is also called the unsupervised modeling process. The following is an example of cluster results using the k-means method:

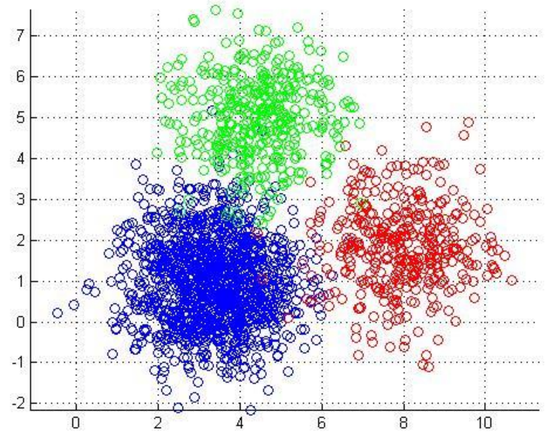


Figure 1. *Clustering using the k-means method*

The procedure used in optimizing using k-means is as follows [13]:
Step 1. Determine the number of clusters.
Step 2. Allocate data into clusters randomly.
Step 3. Calculate the centroid / average of the data in each cluster.
Step 4. Allocate each data to the nearest centroid / average
Step 5. Return to Step 3, if there is still data moving clusters or if the centroid value changes.

Data

The data used in the study is secondary data sourced from the Ministry of Education and Culture which is processed by the Central Statistics Agency (abbreviated as BPS) in the BPS Catalog: 4301008 on the Portrait of Indonesian Education. The research leads to student and class ratios at the primary school level for the 2018/2019 academic year which consists of 34 records. The following is the source of the research data used in mapping the area to the ratio of the number of students per study group.

Table 1. *Research data*

Province	Student to Study Group Ratio
Aceh	20
North Sumatra	23
West Sumatra	21
Riau	23
Jambi	21
South Sumatra	23
Bengkulu	20
Lampung	22
Kep. Bangka Belitung	25
Kep. Riau	24
DKI Jakarta	28
West Java	27
Central Java	22
DIYogyakarta	21
East Java	21
Banten	28
Bali	23
West Nusa Tenggara	23
East Nusa Tenggara	20
West Kalimantan	19
Central Kalimantan	16
South Borneo	18
East Kalimantan	23
North Kalimantan	20
North Sulawesi	16
Central Sulawesi	17
South Sulawesi	20
Southeast Sulawesi	19
Gorontalo	19
West Sulawesi	18
Maluku	19
North Maluku	18
West Papua	19
Papua	25

source: BPS

Here's how research works using the k-means method in mapping the area to the ratio of the number of students per study group as described in Figure 2 below:

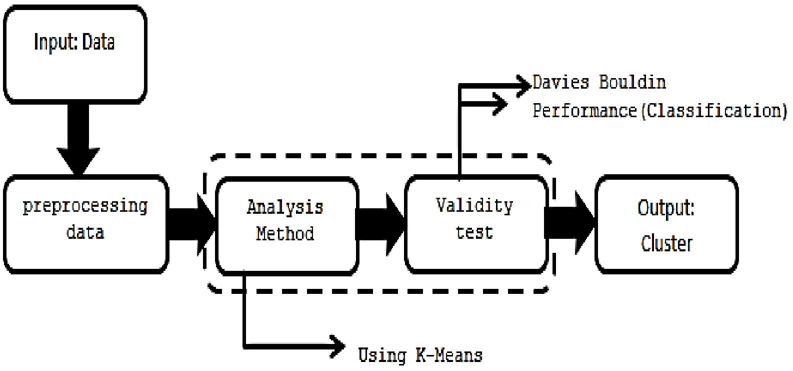
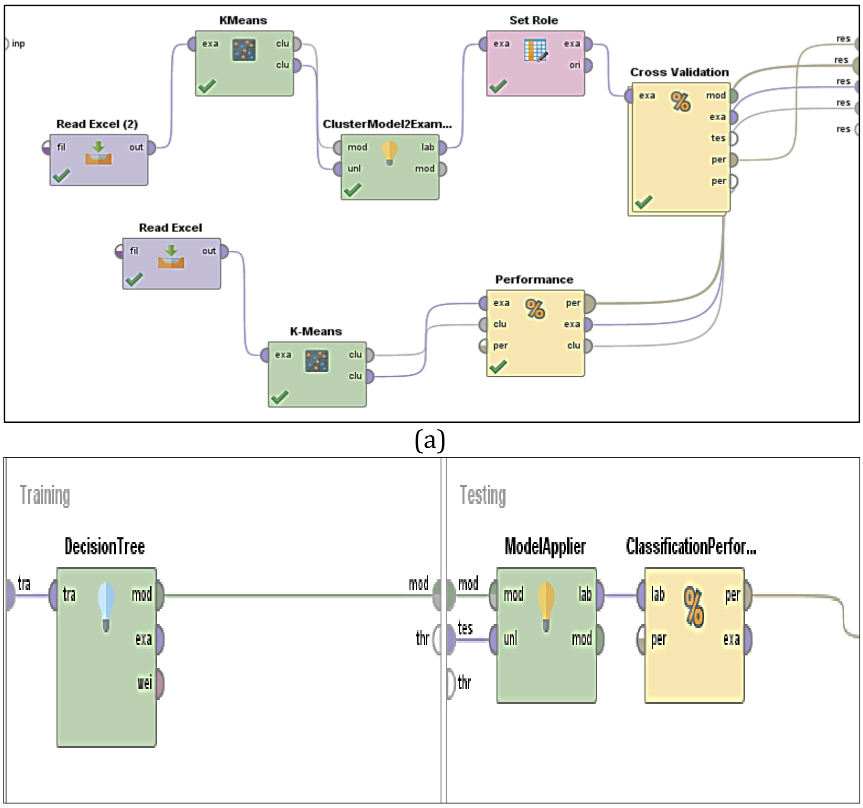


Figure 2. *Research workflow*

In the figure 2, it is explained that the data source was obtained from the Ministry of Education and Culture which was processed by the Central Statistics Agency (abbreviated as BPS) in excel. The data preprocessing process is carried out to see incomplete data so that when processing it gets maximum results. The analysis process uses the help of Rapid Miner software. The analysis process using the k-means method is carried out. The results of the analysis were tested for validity using several parameters to see the results of the cluster formed (K = 1,2,3 .. n) and the performance of the cluster. If feasible, the cluster formed is the best cluster in this study.

RESULTS AND DISCUSSION

At this stage, the analysis process is carried out using the k-means method. Two clusters of mapping labels were used, namely the largest student-class ratio cluster (K1) and the smallest student-class ratio cluster (K2). The attributes used are the province and the cluster student ratio per class in each province. The following is the design of the k-means method using the RapidMiner software as shown in the following figure:

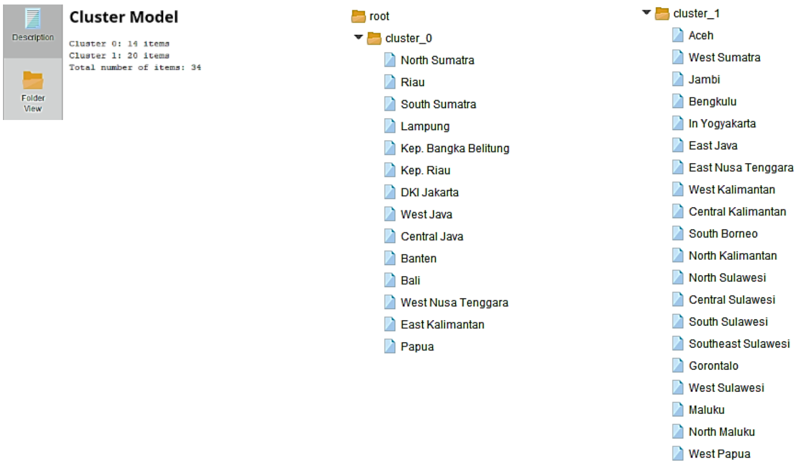


(a)

(b)

Figure 3. *The k-means model in the RapidMiner design*

In Figure 3 (a) the data is processed based on the input entered (.xls). By using the k-means method, a cluster mapping process is generated. In Figure 3 (b), the cluster results are tested with the Davies Bouldin Index (DBI) parameter. DBI is a reference for the cluster formed. The smaller the DBI value, the more optimal the cluster formed. Then the cluster results are re-tested with performance (classification) parameters to see the results of the accuracy (%) of the cluster formed. The following are cluster results using the k-means method on the ratio of the number of students per study group based on province as shown in the following figure:



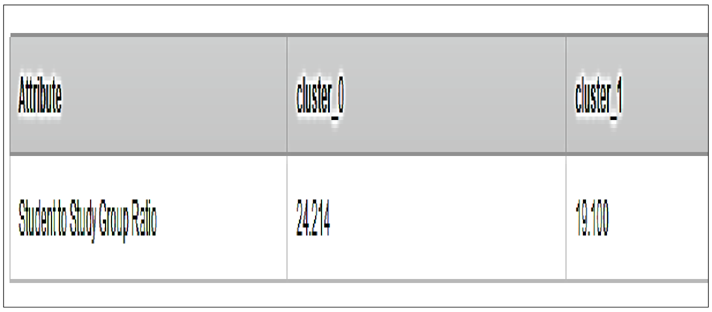
Cluster Model

```
Cluster 0: 14 items
Cluster 1: 20 items
Total number of items: 34
```

```
root
  cluster_0
    North Sumatra
    Riau
    South Sumatra
    Lampung
    Kep. Bangka Belitung
    Kep. Riau
    DKI Jakarta
    West Java
    Central Java
    Banten
    Bali
    West Nusa Tenggara
    East Kalimantan
    Papua
  cluster_1
    Aceh
    West Sumatra
    Jambi
    Bengkulu
    In Yogyakarta
    East Java
    East Nusa Tenggara
    West Kalimantan
    Central Kalimantan
    South Borneo
    North Kalimantan
    North Sulawesi
    Central Sulawesi
    South Sulawesi
    Southeast Sulawesi
    Gorontalo
    West Sulawesi
    Maluku
    North Maluku
    West Papua
```

Figure 4. *Results of clustering with k-means*

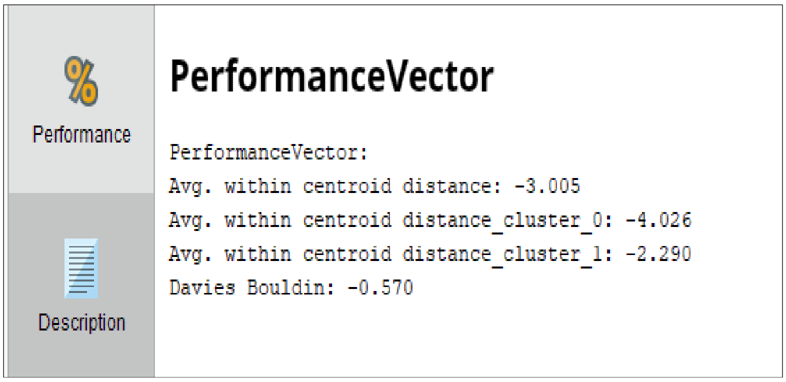
In Figure 4, it can be explained that the results of the cluster show that 14 provinces are in the largest cluster student ratio per class (K1 = cluster_0) and 20 provinces are in the smallest student ratio cluster per class (K2 = cluster_1). Clusters on KI are North Sumatra, Riau, South Sumatra, Lampung, Kep. Bangka Belitung, Kep. Riau, DKI Jakarta, West Java, Central Java, Banten, Bali, West Nusa Tenggara, East Kalimantan, Papua. Clusters in K2 are Aceh, West Sumatra, Jambi, Bengkulu, DI Yogyakarta, East Java, East Nusa Tenggara, West Kalimantan, Central Kalimantan, South Borneo, North Kalimantan, North Sulawesi, Central Sulawesi, South Sulawesi, Southeast Sulawesi, Gorontalo, West Sulawesi, Maluku, North Maluku, West Papua. The results of cluster determination can be seen from the final centroid which is formed as shown in the following figure:



Attribute	cluster_0	cluster_1
Student to Study Group Ratio	24.214	19.100

Figure 5. *The final centroid results*

In Figure 5 the final result of the centroid is 24,214 for the largest cluster (cluster_0) and 19.1 for the smallest cluster (cluster_1).

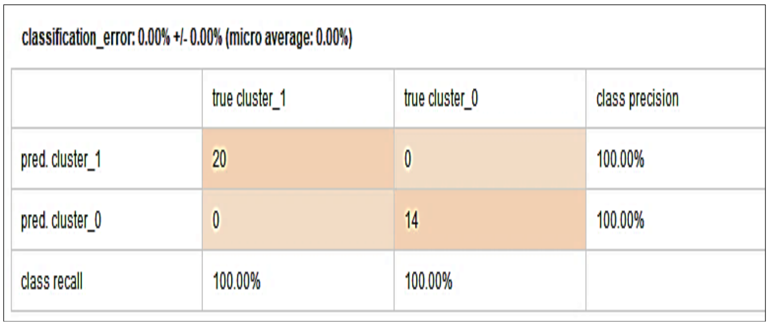


PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: -3.005
Avg. within centroid distance_cluster_0: -4.026
Avg. within centroid distance_cluster_1: -2.290
Davies Bouldin: -0.570
```

Figure 6. *Performance of the Davies Bouldin Index*

In figure 6, the test results using the Davies Bouldin Index parameter is 0.570 with a value of k = 2. The results of the cluster formed are optimal. Because the smaller the generated DBI value, the better the cluster results will be formed.



classification_error: 0.00% +/- 0.00% (micro average: 0.00%)

	true cluster_1	true cluster_0	class precision
pred. cluster_1	20	0	100.00%
pred. cluster_0	0	14	100.00%
class recall	100.00%	100.00%	

Figure 7. *Performance of the performance (classification)*

In figure 7, the test results use performance (classification) where the classification error = 0%, which means that all data can be predicted correctly. The following is the complete cluster results on mapping the ratio of the number of students per class at the elementary school level using the k-means method as shown in the following table:

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

Ratio	Province	cluster
20	Aceh	cluster_1
23	North Sumatra	cluster_0
21	West Sumatra	cluster_1
23	Riau	cluster_0
21	Jambi	cluster_1
23	South Sumatra	cluster_0
20	Bengkulu	cluster_1
22	Lampung	cluster_0
25	Kep. Bangka Belitung	cluster_0
24	Kep. Riau	cluster_0
28	DKI Jakarta	cluster_0
27	West Java	cluster_0
22	Central Java	cluster_0
21	DI Yogyakarta	cluster_1
21	East Java	cluster_1
28	Banten	cluster_0
23	Bali	cluster_0
23	West Nusa Tenggara	cluster_0
20	East Nusa Tenggara	cluster_1
19	West Kalimantan	cluster_1
16	Central Kalimantan	cluster_1
18	South Borneo	cluster_1
23	East Kalimantan	cluster_0
20	North Kalimantan	cluster_1
16	North Sulawesi	cluster_1
17	Central Sulawesi	cluster_1
20	South Sulawesi	cluster_1
19	Southeast Sulawesi	cluster_1
19	Gorontalo	cluster_1
18	West Sulawesi	cluster_1
19	Maluku	cluster_1
18	North Maluku	cluster_1
19	West Papua	cluster_1
25	Papua	cluster_0

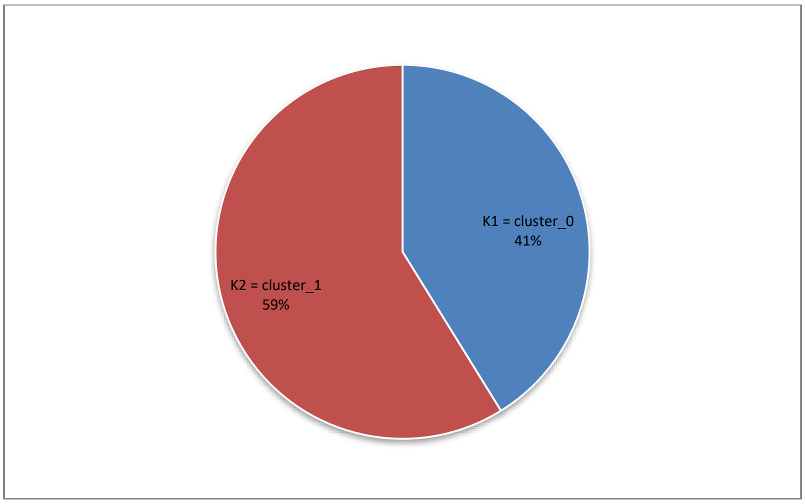
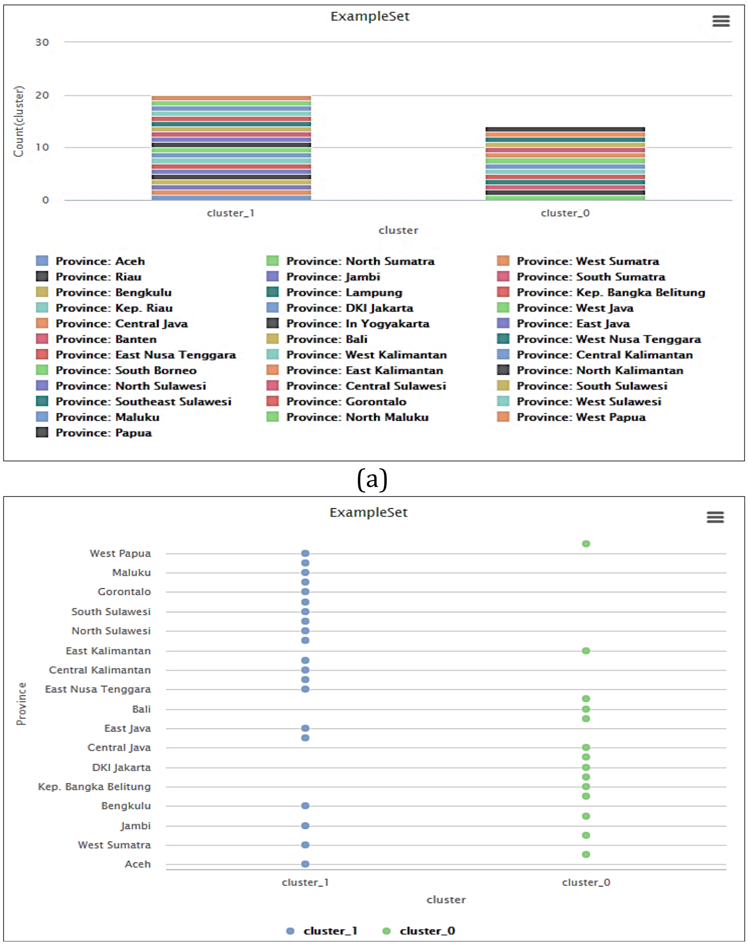


Figure 8. *Percentage of cluster results*

In Figure 8, it is explained that the percentage cluster ratio of students per class is based on province. The analysis results show that 14 provinces are in the largest cluster (K2) with an average per class = 24.2 and 20 provinces are in the smallest cluster (K2) with an average per class = 19.1. The following is a mapping graph with bars and scatter as shown in the following image:



(a)

(b)

Figure 9. *Visualization of clustering results with bat chat and scatter plotter (a)(b)*

CONCLUSION

The results of the study concluded that the k-means method can be applied to mapping the number of students per study group at the primary school level by province. The results of the analysis show that 14 provinces are in the largest cluster (K2) with an average per class = 24.2 and 20 provinces are in the smallest cluster (K2) with an average per class = 19.1. The validity test was done by testing the cluster results (k = 2) with Davies Bouldin was 0.570. The cluster results are optimal. The validity test was also carried out on the results of the cluster with Performance (Classification) with the results of classification error: 0.00%.

REFERENCES

N.S. Perdana, Analisis Capaian Rombongan Belajar Di Provinsi Lampung Tahun 2018 Dalam Upaya Implementasi Permendikbud Nomor 17 Tahun 2017, Dewantara. V (2018) 1–16.

N.S. Perdana, Pengelolaan Ukuran Rombongan Belajar Dan Siswa Per- Rombel dalam Upaya Peningkatan Kualitas Lulusan Menyongsong Society 5.0, in: Pros. SEMDIKJAR (Seminar Nas. Pendidik. Dan Pembelajaran), 2019: pp. 570–580.

Sunandar, A. Buchori, N.D. Rahmawati, W. Kusdaryani, Mobilemath (mobile learning math) media design with seamless learning model on analytical geometry course, Int. J. Appl. Eng. Res. 12 (2017) 8076–8081. https://www2.scopus.com/inward/record.uri?eid=2-s2.0-85040255903&partnerID=40&md5=e3d36bcb06c17d82ebd20ff6c483f1ff.

Sunandar, A. Buchori, N.D. Rahmawati, Development of media kocerin (Smart box interactive) to learning mathematics in Junior High School, Glob. J. Pure Appl. Math. (2016).

B. Supriyadi, A.P. Windarto, T. Soemartono, Mungad, Classification of natural disaster prone areas in Indonesia using K-means, Int. J. Grid Distrib. Comput. 11 (2018) 87–98.

S. Sriyanto, A. Buchori, A. Handayani, P.T. Nguyen, H. Usman, Implementation multi factor evaluation process (MFEP) decision support system for choosing the best elementary school teacher, Int. J. Control Autom. (2020).

R.A. Haraty, M. Dimishkieh, M. Masud, An enhanced k-means clustering algorithm for pattern discovery in healthcare data, Int. J. Distrib. Sens. Networks. 2015 (2015). https://doi.org/10.1155/2015/615740.

A. Waluyo, H. Jatnika, M.R.S. Permatasari, T. Tuslaela, I. Purnamasari, A.P. Windarto, Data Mining Optimization uses C4.5 Classification and Particle Swarm Optimization (PSO) in the location selection of Student Boardinghouses, IOP Conf. Ser. Mater. Sci. Eng. 874 (2020) 1–9. https://doi.org/10.1088/1757-899X/874/1/012024.

S. Wang, M. Li, N. Hu, E. Zhu, J. Hu, X. Liu, J. Yin, K-Means Clustering With Incomplete Data, IEEE Access. 7 (2019) 69162–69171. https://doi.org/10.1109/ACCESS.2019.2910287.

H.L. Sari, D. Suranti, L.N. Zulita, Implementation of k-means clustering method for electronic learning model, J. Phys. Conf. Ser. 930 (2017). https://doi.org/10.1088/1742-6596/930/1/012021.

C. Zhang, S. Xia, K-means clustering algorithm with improved initial center, Proc. - 2009 2nd Int. Work. Knowl. Discov. Data Mining, WKKD 2009. 1 (2009) 790–792. https://doi.org/10.1109/WKDD.2009.210.

D. Marlina, N. Lina, A. Fernando, A. Ramadhan, Implementasi Algoritma K-Medoids dan K-Means untuk Pengelompokkan Wilayah Sebaran Cacat pada Anak, J. CoreIT J. Has. Penelit. Ilmu Komput. Dan Teknol. Inf. 4 (2018) 64. https://doi.org/10.24014/coreit.v4i2.4498.

P. Arora, Deepali, S. Varshney, Analysis of K-Means and K-Medoids Algorithm for Big Data, Phys. Procedia. 78 (2016) 507–512. https://doi.org/10.1016/j.procs.2016.02.095.