KPJ 7(1) (2023)

KAPPA JOURNAL
Physics & Physics Education

https://e-journal.hamzanwadi.ac.id/index.php/kpj/index

Comparison of Algorithms K-Means and DBSCAN for Clustering Student Cognitive Learning Outcomes in Physics Subject

Kertanah[1], Wiwit Pura Nurmayanti[2*], Sri Rahmatun Aini[3], Lalu Muh. Amrullah[3], Muhammad Sya'roni[3]

[1] Kertanah, Department of Statistics, Universitas Hamzanwadi, Selong, Indonesia.
[2] Wiwit Pura Nurmayanti, Department of Statistics, Universitas Hamzanwadi, Selong, Indonesia.
[3] Sri Rahmatun Aini, SMAN 1 Sikur, Lombok Timur, Indonesia.
[4] Lalu. Muh. Amrullah, Department of Statistics, Universitas Hamzanwadi, Selong, Indonesia.
[5] Muhammad Sya'roni, Department of Statistics, Universitas Hamzanwadi, Selong, Indonesia.

Received: 16 June 2023
Revised: 16 August 2023
Accepted: 18 August 2023

Corresponding Author:
Author Name*: Wiwit Pura Nurmayanti
Email*: wiwit.adiwinata3@gmail.com

DOI:
https://doi.org/10.29408/kpj.v7i2.18428

Abstract: Clustering is an activity of grouping data into the same group based on similarity. The purpose of the study is to cluster and determine student cognitive learning outcomes characteristics. Cluster analysis was conducted on student cognitive learning outcomes using algorithms K-Means and DBSCAN. Both algorithms are appropriate to have been applied to the overlapping data such as student learning outcomes data. Also, their advantages are scaling large datasets and outliers. The data used in this study is student cognitive learning outcomes - final and mid-term exams grade X in physics subject. Applying the two proposed algorithms K-Means and DBSCAN, the best cluster algorithm to have been used for clustering analysis is K-Means which is based on the highest silhouette score of 0.43, while the silhouette score of DBSCAN is 0.39 respectively. Using the best cluster, the K-Means algorithm, found two types of clusters – cluster 1 consists of 132 students who have a high average score, and cluster 2 shows 183 students who have a low average score in both final and mid-term exams respectively. From the analysis results, most students still have low cognitive learning outcomes in physics subject.

Keywords: *K-Means Algorithms; DBSCAN Algorithms; Physics Subjects; Cognitive Learning Outcomes*

Introduction

The rapid development of science and technology has caused the development of methods that provide convenience and efficiency in accessing data in this big data era. Data mining is a popular method that is widely used to access data. Data mining is the process of looking for form or vital information in data by using a particular method (Mardi, 2016). For the convenience of filtering data and information for data grouping, we can use one of the data mining approaches, such as clustering algorithms. Clustering is defined as the process of grouping objects based on their similarity into different groups (Madhulatha, 2012). There are some kinds of clustering algorithms, but the K-Means algorithm and the Density-Based Spatial Clustering Algorithm with Noise (DBSCAN) are two popular algorithms that have more efficient clustering results. The K-Means algorithm is for grouping non-hierarchical data that divides data into two or more groups, so the data has similar characteristics when entered into the same group. DBSCAN is a clustering algorithm based on the density of data. The concept of density in DBSCAN produces three types of status from each data set: core, border, and noise (Budiman et al., 2016).

Clustering methods are widely applied in many sectors. However, applying those in educational data is still not

How to Cite:
Kertanah, K., Nurmayanti, W. P., Aini, S. R., Amrullah, L. M., & Sya'roni, M. (2023). Comparison of Algorithms K-Means and DBSCAN for Clustering Student Cognitive Learning Outcomes in Physics Subject. *Kappa Journal, 7*(1), 251-255. https://doi.org/10.29408/kpj.v7i2.18428

much conducted such as clustering student cognitive learning outcomes in physics subject. Senior High School (SMA) is the secondary level where all students in grade X have learned physics in the first semester. Physics is defined as one of the natural sciences that study theories related to natural phenomena and their relation to reality (Sapiruddin et al., 2021). The problem encountered is that student cognitive learning outcomes data are non-overlapping, so the methods needed for clustering its data are partitioning methods. K-Means and DBSCAN algorithms are partitioning methods for clustering.

Some previous researchers have studied some cases by applying K-Means and / or DBSCAN algorithms. Nisrina et al. (2022) have applied Algorithms Self Organizing Maps (SOM) and DBSCAN for clustering unmet needs of family planning (KB) in West Nusa Tenggara province. Therefore, Kertanah et al. (2022) have studied clustering analysis for the last twenty earthquake data using the K-Means algorithm in West Nusa Tenggara. Adha et al. (2021) have researched to compare DBSCAN and K-Means algorithms for clustering covid-19 cases in the world. Applying DBSCAN and K-Means has been conducted for clustering the basic data of laboratory competency (Qadrini, 2020), and the other clustering study has been done by applying K-Means in online sales data (Ashari et al., 2019).

According to the background of research delivered above, this study proposes models of algorithms K-Means and DBSCAN for clustering student cognitive learning outcomes and compares them to find out the better clustering algorithm by using Silhouette average method as the indicator. Therefore, the best algorithm has been applied for clustering analysis to interpret the overview and distribution of student cognitive learning outcomes in physics subject grade X.

Method

Datasets

The type of this study is quantitative research. Data utilized in this study is secondary data which recorded the grade X student cognitive learning outcomes in physics subject for one semester. This data is obtained from the Senior High School (SMA) number 1 of Sikur which provides 315 student cognitive learning outcomes consisting of mid-term exam and final exam marks for nine classes from grade X-4 to grade X-12 in the academic year 2022/2023. Both variables mid-term exam and final exam marks are utilized in this study.

This study applied data mining clustering methods using K-Means and DBSCAN algorithms. Both algorithms have different steps for the clustering process. The followings are steps of clustering analysis processes for the two different algorithms, respectively.

K-Means Algorithm

In K-Means clustering, data is partitioned into separate sets. In calculating the *i*-th distance to the *k*-th cluster center, the Euclidean equation is used as shown in Equation 1 (Adha et al., 2021).

$$d_{ik} = \sqrt{\sum_{j=1}^{m} \left(c_{ij} - x_{ik} \right)^2} \quad (1)$$

The followings are the steps of the clustering process using the K-Means algorithm.

Step 1: Choose an appropriate value of k, the number of clusters or centroids.
Step 2: Choose centroids at random for each cluster.
Step 3: Each data point is assigned to its nearest centroid.
Step 4: Step 4 involves adjusting the centroid for the newly formed cluster.
Step 5: Repeat steps 4 and 5 till all the data points are perfectly organized within a cluster space.

DBSCAN

The clustering process with DBSCAN can be shown as a tree starting with the point closest to the minimum number of points (MinPts) in the radius ε . The stages of the DBSCAN clustering process are as follows.

Step 1: Initialize minimum points (MinPts) and epsilon (ε).
Step 2: Determine starting point (p) randomly.
Step 3: Calculate the epsilon or all density reachable distances to p
Step 4: If the point that meets the epsilon is more than MinPts the point p is the core point and a cluster is formed. If not then it is considered as noise.
Step 5: Repeat steps 2 – 3 if all points have been processed.

Silhouette Method

Silhouette score is an indicator to choose the best number of clusters and also to look at the best algorithm which will be used for clustering. The purpose of the silhouette method is to explain the consistency within cluster data. The range of silhouette values will be between -1 and 1. A high silhouette score indicates that the objects are well-matched within clusters and weakly matched to neighboring clusters, respectively.

- A near +1 Silhouette score indicates that the sample is far away from its neighboring cluster.

- 0 Silhouette score suggests that the sample closes to the decision boundary separating two neighboring clusters.
- -1 Silhouette score indicates that the sample has been assigned to the wrong cluster.

The following equation is used to calculate the silhouette score (Swamynathan, 2017).

$$Silhouette\_index = \frac{(p-q)}{\max(p,q)} \quad (2)$$

where,

p: mean distance to the points in the nearest cluster.

q: mean intra-cluster distance to all the points

Data Processing Technique

The technique of data processing used in this study is clustering using K-Means and DBSCAN with RStudio software. The flow chart of data processing and analysis using RStudio is shown in Figure 1.

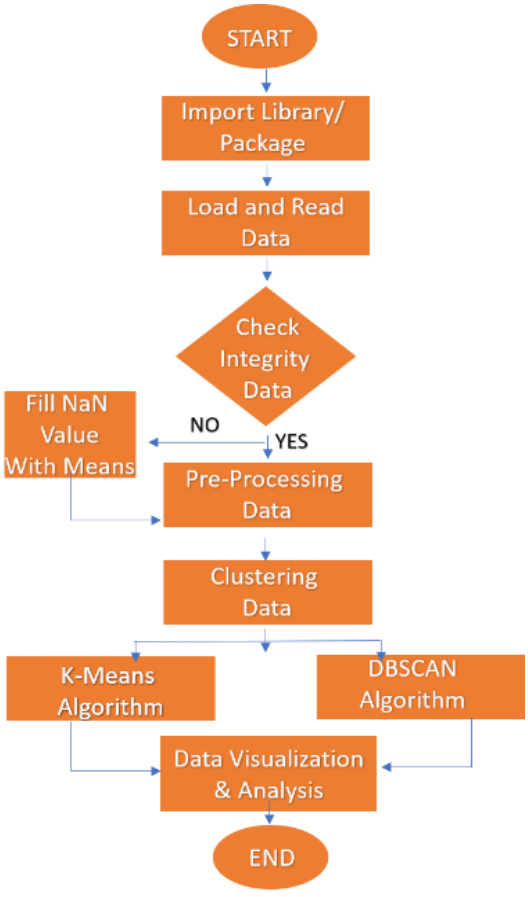

Figure 1. Flowchart of data processing

Result and Discussion

Clustering student cognitive learning outcomes was carried out using algorithms K-Means and DBSCAN, respectively. Of 315 students as a sample in this study, the followings are the clustering results using the K-Means algorithm with many clusters (k) equal to 2 as shown in Table 1.

Table 1. Clustering results using the K-Means algorithm

Clusters	Number of data
1	132
2	183

In addition, the clustering results of DBSCAN with an epsilon of 10 and minimum points of 4 are shown in Table 2.

Table 2. Clustering results using the DBSCAN algorithm

Clusters	Number of clusters
1	6
2	309

The clustering results have been conducted as shown in Table 1 and Table 2, respectively. Each cluster has the same number of clusters which equals 2 clusters. However, the best clustering algorithm needed to have been used in the next clustering analysis. According to the silhouette score using both algorithms K-Means and DBSCAN, the best silhouette score of which has been obtained as represented in Table 3 below.

Table 3. Silhouette scores of Algorithms K-Means and DBSCAN

Clustering Methods	Silhouette Score
K-Means	0.43
DBSCAN	0.39

Table 3 shows that algorithm K-Means has a higher silhouette score than DBSCAN's. It means that K-Means is a better algorithm than DBSCAN to have been used in clustering. Showing the distributions of clustering using K-Means can be looked at Figure 2.

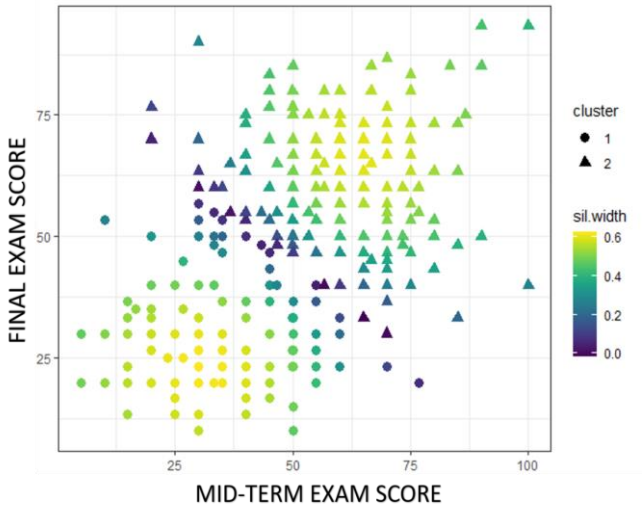

Figure 2. Clustering using the K-Means algorithm

Table 4. The results of clustering using K-Means

Clusters	Final Exam Average	Mid–Term Exam Average
1	60.89	59.74
2	30.60	36.42

Figure 2 depicts the number of student cognitive outcomes clusters presented in triangle-up and circle markers in the first cluster and the second cluster, respectively. Out of 315 student cognitive learning outcomes in the datasets, 132 data are in cluster 1 (triangle-up marker) and the remaining 183 data are in cluster 2 (circle marker). In addition, Table 4 shows cluster 1 that students have a high average score in both the final exams and midterms. On the other hand, cluster 2 shows students have a low average score of cognitive learning outcomes in both exams. From the results of the analysis above, more than 50% of students obtain low scores in both exams for physics subject.

Conclusion

Clustering analysis was carried out using Algorithms K-Means and DBSCAN in cognitive learning outcomes. The comparison of the two algorithms showed that K-Means is the better algorithm which has a high silhouette score of 0.43, and the K-Means algorithm has been applied to cluster the cognitive learning outcomes. The clustering analysis shows that students have high scores average in both exams in cluster 1, while cluster 2 shows students have low scores average in both exams, final and mid-term tests respectively. From the results, more than 50% of students have low scores in the both final and mid-term exams of the physics subject.

Acknowledgments

We are greatly to acknowledge to P3MP for the financial support of the grant for publication funding and also want to thank to SMAN 1 Sikur for permitting data used in this study.

References

Adha, R., Nurhaliza, N., Soleha, U., & Mustakim. (2021). Perbandingan Algoritma DBSCAN dan K-Means Clustering untuk Pengelompokan Kasus Covid-19 di Dunia. *Jurnal Sains Teknologi Dan Industri*, *18*(2), 206–211.

Ashari, S. bena, Otniel, S. C., & Rianto. (2019). PERBANDINGAN KINERJA K-MEANS DENGAN DBSCAN UNTUK METODE CLUSTERING DATA PENJUALAN ONLINE RETAIL. *Jurnal Siliwangi*, *5*(2), 64–67.

Budiman, S., Safitri, D., & Ispriyanti, D. (2016). Perbandingan Metode K-Means Dan Metode Dbscan Pada Pengelompokan Rumah Kost Mahasiswa Di Kelurahan Tembalang Semarang. *Jurnal Gaussian*, *5*(4), 757–762.

Kertanah, Rahadi, I., Novianti, B. A. ryan., Syahidi, K., Putra, H. M., Gazali, M., Hirzi, R. H., & Sabar. (2022). Applying K-Means Algorithm for Clustering Analysis Earthquakes Data in West Nusa Tenggara Province. *Indonesian Physical Review*, *5*(3), 197–207. https://doi.org/https://doi.org/10.29303/ipr.v5i3.148

Madhulatha, T. S. (2012). AN OVERVIEW ON CLUSTERING METHODS. *IOSR Journal of Engineering*, *2*(4), 719–725. https://doi.org/https://doi.org/10.48550/arXiv.1205.1117

Mardi, Y. (2016). Data Mining : Klasifikasi Menggunakan Algoritma C4 . 5. *Jurnal Edik Informatika*, *2*, 213–219. https://doi.org/https://doi.org/10.22202/ei.2016.v2i2.1465

Nisrina, S., Nurmayanti, W. P., Basirun, Kertanah, & Ghazali, M. (2022). Penerapan Metode Clustering SOM dan DBSCAN dalam Mengelompokkan Unmet Need Keluarga Berencana di Nusa Tenggara Barat Universitas Hamzanwadi. *J Statistika*, *15*(2), 237–244. https://doi.org/https//doi.org/10.36456/jstat.vol15.no2.a5549

Qadrini, L. (2020). Metode K-Means dan DBSCAN pada Pengelompokan Data Dasar Kompetensi Laboratorium ITS Tahun 2017. *J Statistika*, *13*(2), 5–11.

Sapiruddin, Novianti, B. A., & Kertanah. (2021). EDUKASI DAN PENDAMPINGAN PRAKTIKUM FISIKA PADA SISWA SEKOLAH MENENGAH ATAS NEGERI 1 SURALAGA KECAMATAN SURALAGA. *SELAPARANG: Jurnal Pengabdian Masyarakat Berkemajuan*, *5*, 738–742. https://doi.org/https://doi.org/10.31764/jpmb.v5i1.6286

Swamynathan, M. (2017). *Mastering Machine Learning with Python in Six Steps (p.201)*. Apress.