JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 7, No. 2, MEI, 2022, Pp. 91 – 99
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Comparative Study of *K-Means* Clustering Algorithm and *K-Medoids* Clustering in Student Data Clustering

Qomariyah [(1)], Maria Ulfah Siregar [(2)*]
[1,2] Teknik Informatika, Fakultas Sains dan Teknologi, UIN Sunan Kalijaga, Yogyakarta
[2] Magister Informatika, Fakultas Sains dan Teknologi, UIN Sunan Kalijaga, Yogyakarta
e-mail : qomariyah.app@gmail.com, maria.siregar@uin-suka.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 15 September 2021, direvisi 10 Desember 2021, diterima 11 Desember 2021, dan dipublikasikan 25 Mei 2022.

Abstract

Universities as educational institutions have very large amounts of academic data which may not be used properly. The data needs to be analyzed to produce information that can map the distribution of students. Student academic data processing utilizes data mining processes using clustering techniques, K-Means and K-Medoids. This study aims to implement and analyze the comparison of which algorithm is more optimal based on the cluster validation test with the Davies Bouldin Index. The data used are academic data of UIN Sunan Kalijaga students in the 2013-2015 batch. In the K-Means process, the best number of clusters is 5 with a DBI value of 0.781. In the K-Medoids process, the best number of clusters is 3 with a DBI value of 0.929. Based on the value of the DBI validation test, the K-Means algorithm is more optimal than the K-Medoids. So that the cluster of students with the highest average GPA of 3,325 is 401 students.

Keywords: Data Mining, Data Pre-Processing, Validation Test, Davies Bouldin Index, Optimal

Abstrak

Perguruan tinggi sebagai institusi pendidikan memiliki data akademik dalam jumlah yang sangat besar, yang mungkin saja data tersebut belum dimanfaatkan dengan baik. Pada data tersebut perlu dilakukan analisis untuk dihasilkannya informasi yang bisa memetakan persebaran mahasiswa. Pengolahan data akademik mahasiswa memanfaatkan proses *data mining* dengan menggunakan teknik *clustering* yaitu *K-Means* dan *K-Medoids*. Penelitian ini bertujuan mengimplementasikan dan menganalisis perbandingan algoritma tersebut mana yang lebih optimal berdasarkan uji validasi klaster dengan *Davies Bouldin Index*. Adapun data yang digunakan adalah data akademik mahasiswa UIN Sunan Kalijaga tahun angkatan 2013–2015. Pada proses *K-Means* mendapatkan jumlah klaster terbaik yaitu 5 dengan nilai DBI sebesar 0,781. Pada proses *K-Medoids* mendapatkan jumlah klaster terbaik yaitu 3 dengan nilai DBI sebesar 0,929. Berdasarkan nilai uji validasi DBI bahwa algoritma *K-Means* lebih optimal dari *K-Medoids*. Sehingga klaster mahasiswa dengan rata-rata IPK tertinggi sebesar 3,325 sejumlah 401 mahasiswa.

Kata Kunci: *Data Mining*, Pra-Pemrosesan Data, Uji Validasi, Davies Bouldin Index, Optimal

1. INTRODUCTION

The abundance of data sets is the accumulation of transaction data recorded for years. So it is interesting to process the data into useful and useful information and knowledge. Universities as educational institutions have very large amounts of student academic and administrative data, which may not be used properly (especially in the preparation of evaluations). On data student academics that accumulate from year to year need to do analysis to be able to open up opportunities to generate information in the manufacture of decisions.

Data mining is a term used to describe the discovery of knowledge in databases. Data mining is a process that uses statistics, mathematics, artificial intelligence, and machine learning to extract and identify useful information and knowledge that is assembled from various large databases (Kusrini & Taufiq, 2019). Study Comparison is a study conducted to compare variables (objects

of research), between different subjects or at different timestamps to find cause-and-effect relationships (Sudijono, 2010). The study comparison is similar to the simulation method. One of the research performed simulation methods is the research of Nurhayati, et al which analyses K-Means and K-Medoids' performance (Nurhayati et al., 2018).

The *K-Means* method is a clustering method that is quite simple and common in its use (Santosa, 2007). *K-Means* is often used in clustering problems (Kharisma & Yazid, 2018) because it has the ability to group data in large enough quantities and with relatively fast and efficient computation time. *K-Means* uses objects in a collection of objects that represent a cluster with an average value, while objects that represent a cluster in the k-Medoid method are medoid. So *K-Medoids* Clustering or Partitioning Around Method is a clustering method that is a variant of the *K-Means* Clustering method (Sindi et al., 2020). In this study, the author will implement and analyze the comparison of which algorithm is more optimal based on the cluster validation test with the Davies Bouldin Index (DBI). DBI is a metric to evaluate the results of the clustering algorithm (Davies & Bouldin, 1979). By using DBI, a cluster will be considered to have an optimal clustering scheme that has a minimum DBI (Farissa et al., 2021).

According to Ruaika (2019), the *K-Means* method is a clustering method that is quite simple and common in use. In every research that has been done on the clustering method, the problem which has not been discussed is the preprocessing stage of outlier removal because an object with a large value may automatically substantially deviate from the distribution of the data. Therefore, deletion outliers are contained in the dataset with the k-Nearest Neighbors algorithm (kNN).

The *K-Medoids* algorithm emerged as a solution to the algorithm's weaknesses in outlier-sensitive *K-Means* due to an object with a large value that may substantially deviate from the data distribution (Farissa et al., 2021). The definition of an outlier is observational data that has extreme values in univariate and multivariate (Alhamdani et al., 2021). Mark extreme in the observation data is a value that is different from some other values in a group.

The existence of outliers in the dataset can cause low accuracy results in the classification process. Outliers in the dataset can be removed at the stage of classification algorithm pre-processing (Sugriyono & Siregar, 2020). Furthermore, to eliminate datasets containing outliers, pre-processing is carried out with outlier removal with the k-Nearest Neighbors (kNN) algorithm. In every research that has been done on the clustering method, the problem that has not been discussed is the outlier removal pre-processing stage because an object with a large value may automatically substantially deviate from the distribution of the data. Therefore, deletion outliers are contained in the dataset with kNN.

In this study, the authors used academic data for students of the Faculty of Science and Technology of UIN Sunan Kalijaga class of 2013-2015. The academic data attributes used are school origin, place of residence, the final cumulative student index (GPA), and the student's study period. Data will be processed using the *K-Means* Clustering algorithm and *K-Medoids* Clustering. With this research, it is also hoped that it can help related parties who need information analysis of student distribution maps.

2. METHODS

The method used is *K-Means* clustering and *K-Medoids* clustering to generate information that can map the distribution of students. Data used is the student data of the Faculty of Science and Technology of UIN Sunan Kalijaga Yogyakarta batch 2013-2015. The data that can be collected is 1557 data. Then the pre-processing stage is changing the data format without changing the data content so that it is easy to process. The data that has gone through the pre-processing stage is then processed using the *K-Means* clustering and *K-Medoids* clustering algorithms. The implementation of the algorithm will be processed using the Python programming language and RapidMiner tool. Then the analysis stage shows the value of the Davies Bouldin Index (DBI) as the best reference for grouping clusters (Supriyadi et al., 2021). The test is carried out by

determining the number of clusters 2, 3, 4, 5, and 6. The method aims to maximize the inter-cluster distance and minimize the intra-cluster distance. Clusters that are considered to have an optimal clustering scheme are those that have a minimum DBI value. On the other hand, a confusion matrix is usually used to evaluate the performance of machine learning algorithms (Fitriyadi, 2021). The variable used is a student area, school origin, cumulative grade point (GPA), and student study period. Examples of raw data can be seen in Table 1.

Table 1 Training Data

Province	School	Period (*semester*)	GPA
D.I Yogyakarta	SMA	13	3,62
Jawa Tengah	MA	10	3,00
D.I Yogyakarta	SMK	8	2,98
Jawa Timur	MA	8	3,56
D.I Yogyakarta	SMA	9	3,33
Jawa Barat	SMA	11	3,34
Jawa Barat	MA	8	2,50
Jawa Tengah	SMA	8	3,25
Jawa Tengah	SMA		3,25
Jawa Timur		10	3,33

The flow of the *K-Means* and *K-Medoids* algorithm is shown in Figure 1 and 2.

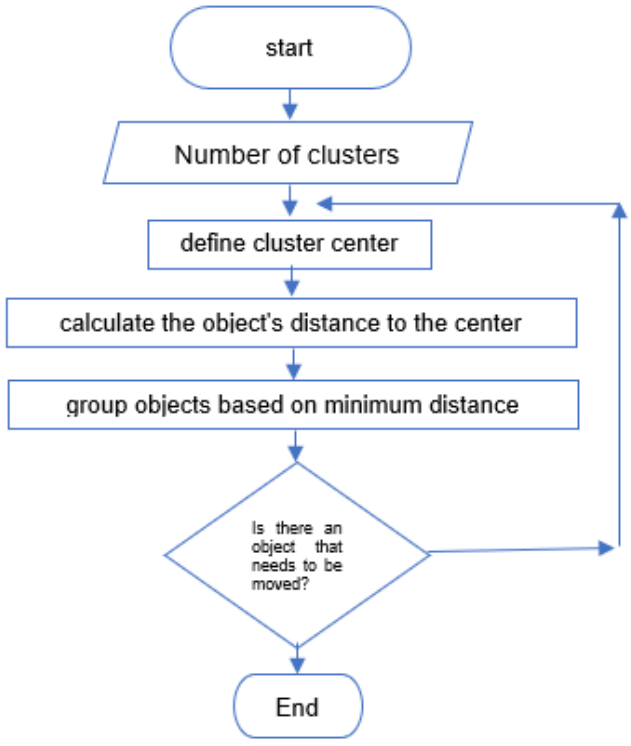

Figure 1 The Flow of the *K-Means* Algorithm

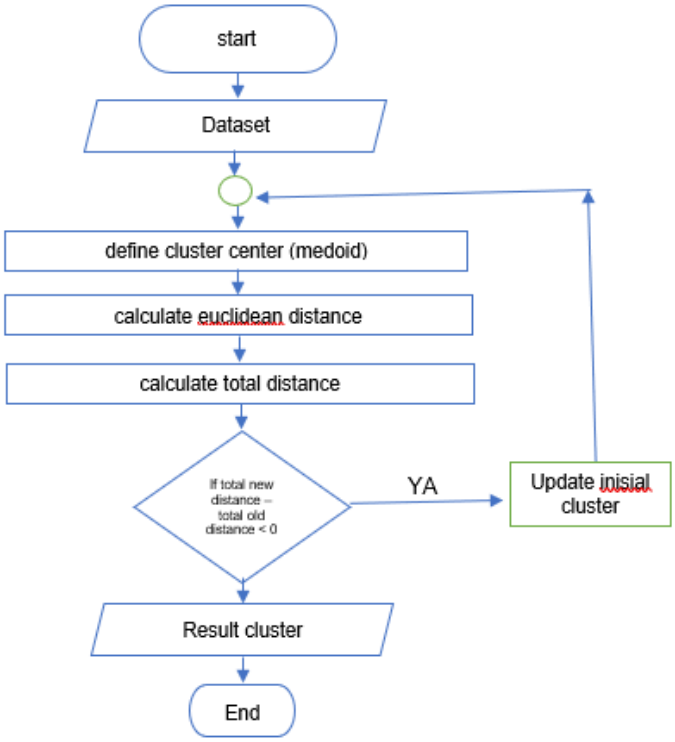

Figure 2 The Flow of the *K-Medoids* Algorithm

3. RESULTS AND DISCUSSION

Several stages that are carried out in this research are as follows:

3.1 Data Collection

The data used are student data of UIN Sunan Kalijaga class of 2013–2015. The data was obtained from the Student Affairs section of UIN Sunan Kalijaga. The data contains 1557 student data consisting of several data attributes, including NIM, address, sub-district, district, province, study program & architecture, school origin, number of study periods, and GPA.

3.2 Data Preprocessing

Data preprocessing (preprocessing) is a step carried out to prepare and process the raw data that has been obtained so that the data can be processed at the next stage optimally. Data preprocessing could transform unstructured data into structured data (Oktarina et al., 2020). Following are the stages of data pre-processing (Susanto, 2013).

3.2.1 Data Reduction

Data reduction is applied to reduce the size of the data by combining or eliminating data, dimensions, and attributes that are not needed. After the raw data were analyzed, there were still attribute data that were not needed in this study. Therefore, researchers reduce some attributes that are not used. So that the final result of the reduction process is the attribute data of the student's NIM, the student's school origin, the student's province of origin, the number of semesters taken by the student, and the student GPA with a total of 1557 students.

3.2.2 Data Integration

Data integration is merging data from various databases into one new database or making data into a single source file. This integration process must be carried out if the data source to be processed has several files. In collecting data in this study, researchers obtained one data source in excel format. The data has been integrated into one data source and is in accordance with the needs of this research, making it easier for the next process.

3.2.3 Data Cleaning

Data cleaning is a process of removing noise, correcting inconsistent data, and filling in missing value data. Such as blank data, duplication of data, data typos, missing letters, or excess letters, and so on. In this data, there are still 181 missing value data on the school origin attribute. Then replace the missing value data with the mode of the school origin attribute data. Many factors lead to inconsistent data. In addition to missing value data, there are data typos in writing the attribute data from schools and provinces. This causes the data processing process to be less than optimal, so researchers must find and change the data so that it can be processed further optimally.

3.2.4 Data Transformation

Table 2 The Label of Province

No.	Student Province Origin	Frequency	Label
1	Bali	5	0
2	Banten	16	1
3	Bengkulu	7	2
4	D.I Aceh	4	3
5	D.I Yogyakarta	424	4
6	D.K.I Jakarta	11	5
7	Jambi	14	6
8	Jawa Barat	112	7
9	Jawa Tengah	583	8
10	Jawa Timur	191	9
11	Kalimantan Barat	5	10
12	Kalimantan Selatan	4	11
13	Kalimantan Tengah	8	12
14	Kalimantan Timur	6	13
15	Kalimantan Utara	3	14
16	Kep.Bangka Belitung	16	15
17	Kepulauan Riau	12	16
18	Lampung	33	17
19	Luar Negeri	3	18
20	Maluku	1	19
21	Nusa Tenggara Barat	20	20
22	Nusa Tenggara Timur	3	21
23	Papua Barat	1	22
24	Riau	22	23
25	Sulawesi Barat	1	24
26	Sulawesi Selatan	8	25
27	Sulawesi Tengah	2	26
28	Sulawesi Tenggara	3	27
29	Sumatera Barat	7	28
30	Sumatera Selatan	16	29
31	Sumatera Utara	16	30

Table 3 The Label of School Origin

No.	School Origin	Frequency	Label
1	MA	532	0
2	SMA	906	1
3	SMK	119	2

Data transformation functions to change data into a suitable format for processing in data mining. The clustering method can only accept input data in the form of numeric (numbers). The attribute data format of the student's school origin and the student's province of origin is in the form of string data, so researchers need to convert the data into numeric data so that it can be processed in clustering. The following data format can be seen in Table 2 and Table 3.

The next stage is the process of removing dataset outliers. At this stage, the author uses a library in Python. The key parameter in kNN is *n_neighbors*, which specifies the number of neighbors to use to calculate the distance from the measurement point. The researcher chose the closest neighbors with as many as four data points. The threshold for outlier detection is 0.01. This threshold value is set after doing the experiment and 0.01 is the best. The resulting output is in the form of data that has been cleaned of outlier data with a total of 1120 data.

Normalization is a transformation process to change data values. Min-Max Normalization is a normalization technique by performing linear transformations on the original data attributes to produce the same range of values. In this case, the researcher equates the data attribute scale to a specific range that is smaller from 0 to 1 (Ningsih et al., 2019). Min-Max Normalization maps a value *v* from attribute *A* to *v'* into the range [new_minA, new_MaxA] with the Equation (1) as follows (Pramesti et al., 2017).

$$v' = \frac{v - min_A}{max_A - min_A}(new\ max_A - new\ min_A) + new\ min_A \qquad (1)$$

3.3 Data Modelling

In this clustering process, experiments were carried out by determining the number of *k* (clusters) namely 2, 3, 4, 5, and 6.

3.4 The Validity of Clustering

Davies Bouldin index is one of the internal evaluation methods measuring cluster evaluation on a grouping method based on the value of cohesion and separation (Muhammad, 2015). Clusters that are considered to have an optimal clustering scheme are those that have a minimum DBI value. The Davies Bouldin Index is based on the similarity of the cluster size based on cluster spread and cluster size inequalities. This approach is to maximize the inter-cluster distance and minimize the intra-cluster distance. The ratio value obtained is used to find the DBI value from Equation (2) below (Iskandar et al., 2018).

$$\frac{1}{K}\sum_{i=1}^{k} max_{i \neq j}\ (R_{i,j}) \qquad (2)$$

3.5 Analysis

Based on the data, we calculate the DBI of each cluster, and the time required for each algorithm is obtained as follows in Table 4.

The Davies Bouldin Index is based on the similarity of the cluster size based on cluster spread and cluster size inequalities. A cluster will have an optimal clustering scheme if it has minimal DBI (Farissa et al., 2021). Based on experiments on this dataset, *K-Means* has a Davies Bouldin Index (DBI) value that is smaller than that owned by *K-Medoids*. Starting from k amounted to 2 to 6. This shows that *K-Means* is a more optimal algorithm.

JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 7, No. 2, MEI, 2022: 91 – 99

Table 4 The DBI for Several Cluster

Tools	Cluster	*K-Means*	Duration (second)	*K-Medoids*	Duration (second)
RapidMiner Python	k = 2	1.086	4	1.186	5
	k = 3	0.925	4	0.929	5
	k = 4	0.879	4	1.403	5
	k = 5	0.781	4	1.094	5
	k = 6	0.873	4	1.126	5
	k = 2	1.021	1	0.832	38
	k = 3	0.975	1	1.234	38
	k = 4	0.861	1	1.029	38
	k = 5	0.869	1	1.022	38
	k = 6	0.882	1	0.974	38

The time needed in processing data in this study, *K-Means* only takes an average of 1-4 seconds while processing data with *K-Medoids* takes an average of 5-33 seconds. That matter indicates that the *K-Means* algorithm is faster in processing processes clustering.

It is stated in Table 4 in the *K-Means* column that the Davies Bouldin value the smallest index (DBI) both processed using Python and RapidMiner is owned by a cluster of 5. Where in cluster 1 there are students with the most residence and school origins than other clusters. The highest number of students from the province Yogyakarta and most graduates from high school (SMA).

Based on the research that has been done, the smallest value of the DBI is owned by cluster 5. The DBI value for cluster 5 is 0.781. Thus, a cluster grouping pattern of 401 students was obtained. Where the cluster has an average GPA value of 3,325. The cluster grouping can be presented in Table 5 below.

Table 5 The DBI for Several Cluster

The pattern of student grouping comes from:	
Bali: 2	SMA: 401
Banten: 12	
Bengkulu: 4	
D.I. Aceh: 2	
D.I. Yogyakarta: 119	
D.K.I. Jakarta: 6	
Jambi: 4	
Jawa Barat: 49	
Jawa Tengah: 120	
Jawa Timur: 54	
Kalimantan Barat: 2	
Kalimantan Selatan: 2	
Kalimantan Tengah: 4	
Kalimantan Timur: 4	
Kalimantan Utara: 2	
Kep.Bangka Belitung: 13	
Kepulauan Riau: 2	

Based on research in clusters of 5, students who have the highest average GPA of 3,325 are high school graduates from Bali, Banten, Bengkulu, D.I. Aceh, D.I. Jakarta, D.I. Yogyakarta, Jambi, West Java, Central Java, East Java, Kalimantan Island, Bangka Belitung, and Riau Islands.

4. CONCLUSIONS

Based on research that has been done on the clustering process, the number of clusters is 2,3,4,5, and 6. In the *K-Means* process, the best number of clusters is 5 with a DBI value of 0.781. In the *K-Medoids* process, the best number of clusters is 3 with a DBI value of 0.929. Based on the value of the DBI validation test, the *K-Means* algorithm is more optimal than the *K-Medoids*. This is because *K-Means* have a low computational performance compared to *K-Medoids*. So the cluster of students with the highest average GPA of 3.325 is 401 students.

ACKNOWLEDGEMENT

Thank you very much to the student body of UIN Sunan Kalijaga and all parties for all input, moral and material support, corrections, and assistance in this research activity so that this research can be completed.

REFERENCES

Alhamdani, F. D. S., Dianti, A. A., & Azhar, Y. (2021). Segmentasi Pelanggan Berdasarkan Perilaku Penggunaan Kartu Kredit Menggunakan Metode K-Means Clustering. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *6*(2), 70–77. https://doi.org/10.14421/jiska.2021.6.2.70-77

Davies, D. L., & Bouldin, D. W. (1979). A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PAMI-1*(2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909

Farissa, R. A., Mayasari, R., & Umaidah, Y. (2021). Perbandingan Algoritma K-Means dan K-Medoids Untuk Pengelompokkan Data Obat dengan Silhouette Coefficient di Puskesmas Karangsambung. *Journal of Applied Informatics and Computing*, *5*(2), 109–116. https://doi.org/10.30871/jaic.v5i1.3237

Fitriyadi, A. U. (2021). Algoritma K-Means dan K-Medoids Analisis Algoritma K-Means dan K-Medoids Untuk Clustering Data Kinerja Karyawan Pada Perusahaan Perumahan Nasional. *KILAT*, *10*(1), 157–168. https://doi.org/10.33322/kilat.v10i1.1174

Iskandar, I. D., Pertiwi, M. W., Kusmira, M., & Amirulloh, I. (2018). Komparasi Algoritma Clustering Data Media Online. *IKRA-ITH INFORMATIKA : Jurnal Komputer Dan Informatika*, *2*(3), 1–8.

Kharisma, R. B., & Yazid, A. S. (2018). The Mapping of Access Point Workloads at UIN Sunan Kalijaga Based on Log Analysis using K-Means Algorithm. *IJID (International Journal on Informatics for Development)*, *6*(1), 17. https://doi.org/10.14421/ijid.2017.06105

Kusrini, E. T. L., & Taufiq, E. (2019). *Algoritma Data Mining*. Penerbit Andi.

Muhammad, A. F. (2015). *Klasterisasi Proses Seleksi Pemain Menggunakan Algoritma K-Means (Study Kasus : Tim Hockey Kabupaten Kendal)*. Universitas Dian Nuswantoro.

Ningsih, W. A., Indriani, F., & Farmadi, A. (2019). Klasifikasi Detak Jantung Janin Dengan Learning Vector Quantization (LVQ). *Seminar Nasional Ilmu Komputer (SOLITER)*, *2*, 130–135.

Nurhayati, Sinatrya, N. S., Wardhani, L. K., & Busman. (2018). Analysis of K-Means and K-Medoids's Performance Using Big Data Technology. *2018 6th International Conference on Cyber and IT Service Management (CITSM)*, 1–5. https://doi.org/10.1109/CITSM.2018.8674251

Oktarina, C., Notodiputro, K. A., & Indahwati, I. (2020). Comparison of K-Means Clustering Method and K-Medoids on Twitter Data. *Indonesian Journal of Statistics and Its Applications*, *4*(1), 189–202. https://doi.org/10.29244/ijsa.v4i1.599

Pramesti, D. F., Furqon, M. T., & Dewi, C. (2017). Implementasi Metode K-Medoids Clustering Untuk Pengelompokan Data. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer*, *1*(9), 723–732.

Santosa, B. (2007). *Teknik Pemanfaatan Data untuk Keperluan Bisnis* (1st ed.). Graha Ilmu.

Sindi, S., Ningse, W. R. O., Sihombing, I. A., R.H.Zer, F. I., & Hartama, D. (2020). Analisis Algoritma K-Medoids Clustering dalam Pengelompokan Penyebaran Covid-19 di Indonesia. *Jurnal Teknologi Informasi*, *4*(1), 166–173. https://doi.org/10.36294/jurti.v4i1.1296

Sudijono, A. (2010). *Pengantar Statistik Pendidikan*. RajaGrafindo Persada.

Sugriyono, S., & Siregar, M. U. (2020). Preprocessing kNN algorithm classification using K-means and distance matrix with students' academic performance dataset. *Jurnal Teknologi Dan Sistem Komputer*, *8*(4). https://doi.org/10.14710/jtsiskom.2020.13874

Supriyadi, A., Triayudi, A., & Sholihati, I. D. (2021). Perbandingan Algoritma K-Means dengan K-Medoids pada Pengelompokan Armada Kendaraan Truk Berdasarkan Produktivitas. *JIPI (Jurnal Ilmiah Penelitian Dan Pembelajaran Informatika)*, *6*(2), 229–240. https://doi.org/10.29100/jipi.v6i2.2008

Susanto, B. (2013). *Data Preprocessing*.