**Jurnal Varian**
Vol. x, No. x, Oktober 202x, Hal. xx-xx
e-ISSN 2581-2017

# Clustering of Province Based on Education Indicators Using K-Medoids

**Annisa Zuhri Apridayanti[1], M.Fathurahman[2], Surya Prangga[3]**
[1,2,3]Mulawarman University, Indonesia

**Article Info**
**Article history:**
Received : mm-dd-yyyy
Revised : mm-dd-yyyy
Accepted : mm-dd-yyyy

**Keyword:**
Clusters;
Data Mining;
DBI;
Education;
K-Medoids.

**ABSTRACT**
Data mining is searching for interesting patterns or information by selecting data using specific techniques or methods. One method that can be used in data mining is K-Medoids. K-Medoids is a method used to group objects into a cluster. This research aimed to obtain the optimal number of clusters using the K-Medoids method based on Davies-Bouldin Index (DBI) validity on education indicators data by province in Indonesia in 2021. The results showed that the optimal number of clusters using the K-Medoids method based on DBI validity is 5 clusters. Cluster 1 consists of 1 province with a higher average dropout rate, average length of schooling, and well-owned classrooms compared to other clusters. Cluster 2 consists of 15 provinces with an average proportion of school libraries lower than Clusters 3 and 4 and higher than Clusters 1 and 5. Cluster 3 consists of 9 provinces with an average proportion of school libraries, proportions of school laboratories, net enrollment rates, and higher school enrollment rates than other clusters. Cluster 4 consists of 8 provinces with a higher average enrollment rate than the other clusters. Cluster 5 consists of 1 province with a higher average repetition rate and student-per-teacher ratio than other clusters.

Accredited by Kemenristekdikti, Decree No: 200/M/KPT/2020
DOI: https://doi.org/10.30812/varian.v4i1.xxxxx

***Corresponding Author:***

M. Fathurahman,
Department of Statistics, Mulawarman University, Indonesia.
Email: fathur@fmipa.unmul.ac.id



## A. INTRODUCTION

Data mining is discovering new information by looking for patterns or rules in vast amounts of data (Defiyanti, 2017). Data mining can also be used to cluster, aiming to discover universal patterns from existing data. According to Suyanto (2017), clustering is the process of grouping a data set into several groups or clusters such that the objects in a cluster have high similarity (homogeneous) but have high dissimilarity with objects in other clusters.

Many experts, including K-Means, K-Modes, Fuzzy C-Means, and K-Medoids have developed cluster analysis. Each has characteristics, advantages, and disadvantages, which are then grouped into four categories. In this study, the clustering method used is the partitioning method, where the algorithm included in the partitioning-based method is K-Medoids (Suyanto, 2017).

The application of the K-Medoids method was developed in 1987 by Leonard Kauffman and Peter J. Rousseww. The two researchers stated that K-Medoids can cluster data that has outliers. One of the methods used to see the optimal cluster in K-Medoids is the Davies-Bouldin Index (DBI). According to Badruttamam et al. (2019), the DBI method can measure how good a cluster is by maximizing the distance between clusters and, at the same time, minimizing the distance between objects in a cluster. The K-Medoids method can be applied in various fields, one of which is regional clustering based on the quality of education.

Improving the quality of education in Indonesia must also begin with improving the quality of education in the region. The quality of education in Indonesian regions needs to be maximally equitable, and the lack of equal distribution of education occurs in remote provinces; this is mainly due to the distribution of education subsidies that have not been comprehensive. Based on data collected in the Socio-Economic Survey in Education Statistics at the SMA/MA level, Indonesia still needs to meet the criteria for completing compulsory education because the participation rate in Indonesia has not reached 95%. In handling this problem, a process of clustering the quality of education in each province in Indonesia is needed to determine which regions have developed or still need to catch up in the quality of education (Gibran et al., 2018).

Several previous studies grouped indicators of education, namely research conducted by Praokta (2021), to find out which provincial groups need improvement in problems in education. Data on indicators of education for SMA/MA equivalent level in Indonesia in the 2019/2020 school year were obtained from the Socio-Economic Survey in March 2020 in Education Statistics 2020. This research found that the first cluster contained six provinces with higher education indicators, the second cluster contained 13 provinces with moderate education indicators, and the third cluster had 15 provinces with low education indicators. Zulfa (2019) conducted research on the grouping of provinces in Indonesia based on educational indicators using the K-Means and K-Medoids methods to know the results of grouping the best between the K-Means and K-Medoids methods, and the results show that the K-Medoids method is a better method than the K-Means method. Subsequent research was conducted by Dewi et al. (2021) regarding the grouping of districts/cities in Maluku Province based on educational indicators using the Ward method, which aims to find out results of district/city clustering in Maluku Province based on education indicators using the Ward method.

## B. RESEARCH METHOD

### 1. Data and Research Variable

This study's data for research variables is education indicators in Indonesia in 2021. This data is secondary data obtained from the Ministry of Education, Culture, Research and Technology of the Republic of Indonesia. The research variables used are ten, as shown in Table 1.

**Table 1.** Research Variable

| Variable | Description |
|---|---|
| $X_1$ | Percentage of school libraries |
| $X_2$ | Percentage of school laboratories |
| $X_3$ | School participation rate |
| $X_4$ | Pure participation rate |
| $X_5$ | Gross participation rate |
| $X_6$ | Number repeating |
| $X_7$ | Drop out rate |
| $X_8$ | The average length of school |
| $X_9$ | Student-to-teacher ratio |
| $X_{10}$ | Percentage of well-owned classrooms |

### 2. Analysis Step

The research flow is explained through the flowchart according to Figure 1 and explained in detail through the following steps:

1. Data standardization is to scale the data into another form so that the data has the expected distribution. Each data performed the same mathematical operations on the original data. Standardize data using the defined in Equation 1:

$$\hat{x}_{ij} = \frac{x_{ij} - \overline{x}_j}{S_j} \tag{1}$$

2. Detect multicollinearity. According to Mega et al. (2018), multicollinearity is a perfect linear relationship between one variable and another. One way to determine whether there is multicollinearity is by looking at the Variance Inflation Factor (VIF) value. The VIF value can be obtained using Equation 2 as follows:

$$VIF_k = \frac{1}{1 - R_k^2}, k = 1,2, \dots, q. \tag{2}$$

3. Clustering data using the K-Medoids algorithm for each $K = 2,3, \dots,10$ with the following stages (Rizky et al., 2020; Wahyu et al., 2022):
   a. Randomly initialize the cluster center with as many $K$ objects as representative objects $o_m$ (medoids).
   b. Calculating the Euclidean distance for each object against each medoid using the following equation:

$$d(x_{ij}, o_{mj}) = \sqrt{(x_{i1}, o_{m1})^2 + (x_{i2}, o_{m2})^2 + \cdots + (x_{iq}, o_{mq})^2} \tag{3}$$

   where $d(x_{ij}, o_{mj})$ is the distance between the data at $i$th observation and $j$th observation on the $m$th medoid for $i = 1,2, \dots, n$; $j = 1,2, \dots, q$ and $m = 1,2, \dots, k$.
   c. Assigns each object to the cluster closest to its medoids and calculates the objective function, which is the sum of the proximity of all objects to the nearest medoids based on the minimum distance between objects to each medoid.
   d. Selecting cluster centers randomly as many as $K$ objects as non-representative objects $o_h$ (non-medoids).

e. Calculating the Euclidean distance for each object against each of the non-medoids using the following equation:

$$d(x_{ij}, o_{hj}) = \sqrt{(x_{i1}, o_{h1})^2 + (x_{i2}, o_{h2})^2 + \cdots + (x_{iq}, o_{hq})^2} \quad (4)$$

where $d(x_{ij}, o_{mj})$ is the distance between the data at $i$th observation and $j$th observation on the $h$th medoid for $i = 1,2, \ldots, n$; $j = 1,2, \ldots, q$ and $h = 1,2, \ldots, k$.

f. Place each object into the cluster closest to its non-medoids and calculate the objective function, which is the sum of the proximity of all objects to the nearest non-medoids based on the minimum distance between objects to each non-medoids.

g. Compute the difference of the objective function by subtracting the non-medoid objective function from the medoid objective function.

h. Replace medoids with non-medoids if the objective function value is < 0.

i. Repeat step d to step g until there are no more medoids changes.

j. After clustering, there is no change in the representative object.

4. Compute the DBI value to find the optimal cluster from the clustering results with the following stages:

a. The Sum of Square Within-cluster (SSW) is an equation used to determine cohesion (homogeneity) in a $K$-cluster. Compute the SSW value for each $K$ cluster using Equation 5.

$$SSW_k = \sqrt{\frac{1}{n}\sum_{i=k}^{n_k} d(x_i, c_k)} \quad (5)$$

b. The Sum of Square Between-cluster (SSB) is an equation used to determine cluster separation (heterogeneity). Compute the SSB value using Equation 6.

$$SSB_{k,t} = d(c_k, c_t) \quad (6)$$

c. Ratio measurement is then used to determine the comparison value between the $k$th and $t$th clusters. A good cluster has the smallest possible cohesion value and the vast possible separation. The ratio value can be obtained by:

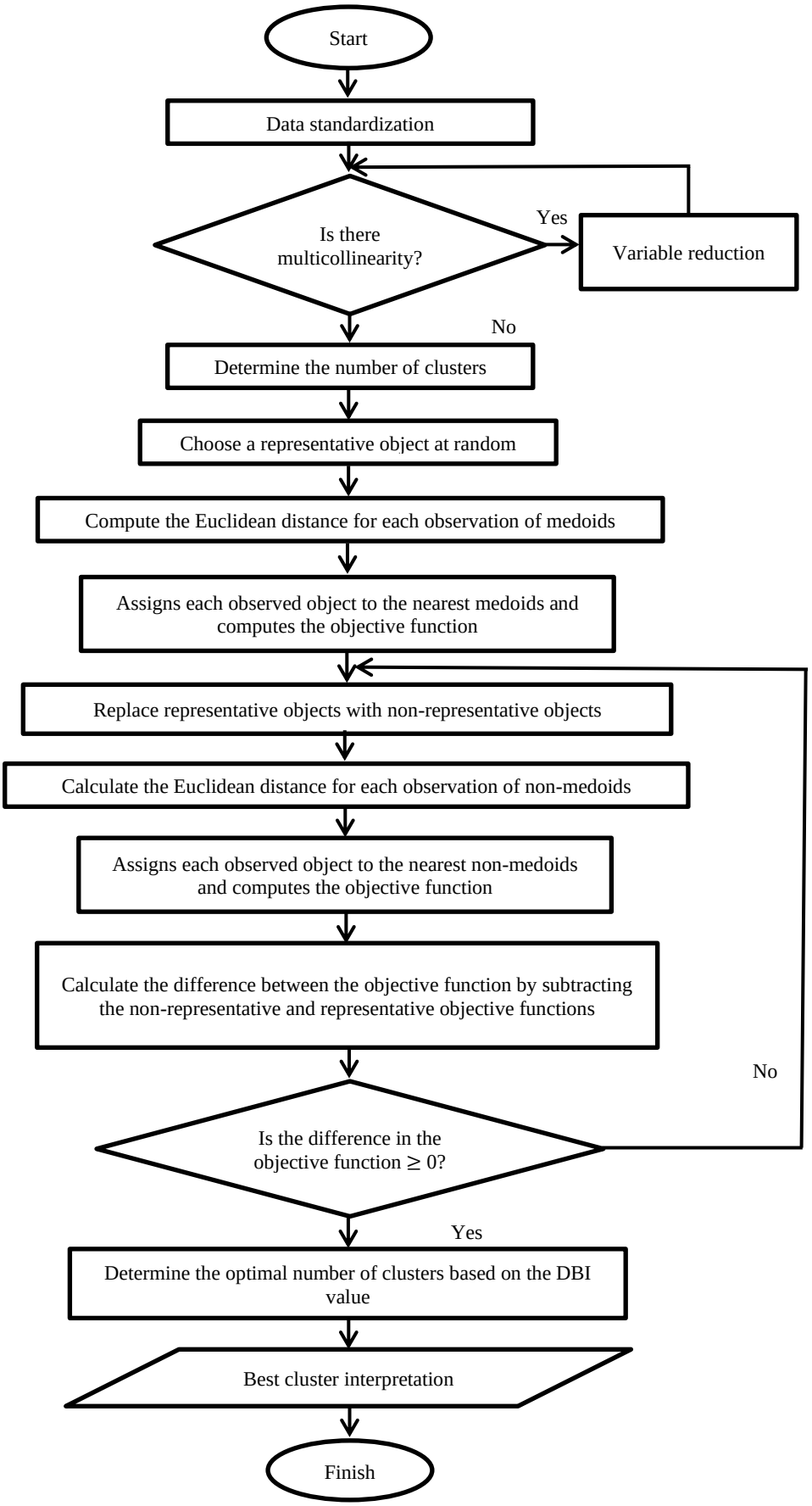
$$R_{k,t} = \frac{SSW_k + SSW_t}{SSB_{k,t}} \quad (7)$$

d. Compute the DBI. The ratio value obtained from Equation (8) is used to find the DBI value using the following formula:

$$DBI = \frac{1}{K} max_{k \neq t}(R_{k,t}) \quad (8)$$

The complete analysis stages of applying the K-Medoids method to clustering provinces in Indonesia are shown in Figure 1.


DOI: https://doi.org/10.30812/varian.v4i1.xxxxx


Start
Data standardization
Is there multicollinearity?
Yes
Variable reduction
No
Determine the number of clusters
Choose a representative object at random
Compute the Euclidean distance for each observation of medoids
Assigns each observed object to the nearest medoids and computes the objective function
Replace representative objects with non-representative objects
Calculate the Euclidean distance for each observation of non-medoids
Assigns each observed object to the nearest non-medoids and computes the objective function
Calculate the difference between the objective function by subtracting the non-representative and representative objective functions
No
Is the difference in the objective function ≥ 0?
Yes
Determine the optimal number of clusters based on the DBI value
Best cluster interpretation
Finish

## C. RESULTS AND DISCUSSION

### 1. Data Standardization

Data standardization is changing the original data values into Z-score form. Data standardization aims to create the same range of values for all variables. The results of data standardization using the Z-score can be seen in Table 2.

**Table 2.** Standardization Result Data

| Province | $X_1$ | $X_2$ | ... | $X_{10}$ |
|---|---|---|---|---|
| DKI Jakarta | -0,69 | 0,41 | ... | 2,37 |
| West Java | -0,39 | 0,48 | ... | 0,48 |
| Central Java | 0,65 | 1,80 | ... | 0,33 |
| DI Yogyakarta | 2,66 | 2,31 | ... | 0,81 |
| ⋮ | ⋮ | ⋮ | ... | ⋮ |
| North Kalimantan | -0,46 | -1,73 | ... | 0,17 |

### 2. Multicollinearity Detection

Multicollinearity detection can be done by looking at the VIF value. The following is the result of calculating the VIF value for each variable.

**Table 3.** VIF Value of Variable

| Variable | VIF | Variable | VIF |
|---|---|---|---|
| $X_1$ | 1,51 | $X_6$ | 2,13 |
| $X_2$ | 2,56 | $X_7$ | 1,41 |
| $X_3$ | 4,77 | $X_8$ | 3,02 |
| $X_4$ | 7,48 | $X_9$ | 2,46 |
| $X_5$ | 4,45 | $X_{10}$ | 2,75 |

Based on Table 3, it can be seen that all variables have a VIF value of less than 10, so there is no multicollinearity between variables. Therefore, all variables can be used for the process of clustering provinces in Indonesia based on indicators using the K-Medoids method.

### 3. K-Medoids Method

K-Medoids is a non-hierarchical grouping where it is necessary to determine the number of clusters at the beginning and also the center points (medoids) to group research objects. After the clustering process is carried out, then determine the optimal cluster using DBI.

### 4. Validation of Clustering Results with the DBI

In this study, the DBI value will be calculated to determine the quality of each clustering result. The results of Cluster Validation Based on value can be seen in Table 4.

**Table 4.** DBI Value

| $K$ | DBI |
|---|---|
| 2 | 1,78 |
| 3 | 1,91 |
| 4 | 1,50 |
| 5 | 1,31 |
| 6 | 1,57 |
| 7 | 1,36 |
| 8 | 1,32 |
| 9 | 1,56 |
| 10 | 1,77 |

Based on Table 4 it can be seen that the DBI value for validating data from the results of clustering Provinces in Indonesia based on education indicators using the K-Medoids method has different values. The smallest value is the clustering of 5 clusters, namely 1.31. Therefore, it can be decided that the most optimal clustering using the K-Medoids method is at 5 clusters. The result are displayed in Table 5.

**Table 5**. Cluster Member for 5 Cluster

| Cluster | Amount | Province |
|---|---|---|
| 1 | 1 | DKI Jakarta |
| 2 | 15 | West Java, Central Java, East Java, Jambi, South Sumatra, Lampung, West Kalimantan, Central Kalimantan, South Kalimantan, South Sulawesi, East Nusa Tenggara, Banten, Bangka Belitung Islands, Gorontalo and West Sulawesi |


DOI: https://doi.org/10.30812/varian.v4i1.xxxxx

| | | |
|---|---|---|
| 3 | 9 | DI Yogyakarta, Aceh, West Sumatra, Riau, East Kalimantan, Central Sulawesi, Bali, Bengkulu and the Riau Archipelago |
| 4 | 8 | North Sumatra, North Sulawesi, Southeast Sulawesi, Maluku, West Nusa Tenggara, North Maluku, West Papua and North Kalimantan |
| 5 | 1 | Papua |

Based on Table 5, it can be interpreted that only 1 province joined in cluster 1, then 15 provinces joined in cluster 2, after that 9 provinces joined in cluster 3, then 8 provinces joined in cluster 5 and only 1 province joined in cluster 5.

## D. CONCLUSION AND SUGGESTION

Based on the results of research and discussion, the following conclusions are obtained, the optimal number of clusters using the K-Medoids method based on DBI validity on education indicator data by the province in Indonesia in 2021 is 5 clusters. The characteristics of the optimal cluster formed using the K-Medoids method based on DBI validity on education indicator data by province in Indonesia in 2021 are Cluster 1 consists of 1 province, namely DKI Jakarta Province, which has a higher average dropout rate, average length of schooling, and well-owned classrooms compared to other clusters. Cluster 2 consists of 15 provinces, namely West Java, Central Java, East Java, Jambi, South Sumatra, Lampung, West Kalimantan, Central Kalimantan, South Kalimantan, South Sulawesi, East Nusa Tenggara, Banten, Bangka Belitung Islands, Gorontalo, and West Sulawesi. It has an average percentage of school libraries lower than clusters 3 and 4 and higher than clusters 1 and 5. Cluster 3 consists of 9 provinces, namely DI Yogyakarta, Aceh, West Sumatra, Riau, East Kalimantan, Central Sulawesi, Bali, Bengkulu, and the Riau Islands which have an average percentage of school libraries, school laboratories, net enrollment rates and school enrollment rates which are higher than the other clusters. Cluster 4 consists of 8 provinces, namely North Sumatra, North Sulawesi, Southeast Sulawesi, Maluku, West Nusa Tenggara, North Maluku, West Papua, and North Kalimantan, which have an average gross enrollment rate that is higher than the other clusters. Cluster 5 consists of 1 province, namely Papua, which has a higher average repetition rate and student-per-teacher ratio than other clusters.

Based on the results of this study, suggestions that can be given are for further research, to use other methods such as CLARA as an alternative method to obtain optimal clustering results and to add other educational indicators. Suggestions for related agencies or parties can maximize development equity programs in the education sector by prioritizing provinces that have low education indicators.

## REFERENCES


Astri, I., Almira, F., Intan, P., & Roghibah, S. (2021). Penerapan Algoritma K-Modes Clustering dengan Validasi Davies Bouldin Index pada Pengelompokan Tingkat Minat Belajar Online di Provinsi Daerah Istimewa Yogyakarta. *Jurnal Matematika dan Statistika beserta Aplikasinya*, 30.

Badan Pusat Statistik (2018). *Indikator Pendidikan.* Jakarta: Badan Pusat Statistik.

Badruttamam, A., Sudarno, & Maruddani, D. A. (2019). Penerapan Analisis Klaster K-Modes dengan Validasi Davies Bouldin Index dalam Menentukan Karakteristik Kanal Youtube di Indonesia. *Jurnal Gaussian*, 263-272.

Defiyanti. (2017). Optimalisasi *K-Medoid* dalam Pengklasteran Mahasiswa Pelamar Beasiswa dengan Cubic Clustering Criterion. *Jurnal TEKNOSI*, 03(01).

Dewi L.S, M. W. Talakua, Y., & Lessnusa, M. M. (2021). Analisis Klaster untuk Pengelompokkan Kabupaten/Kota di Provinsi Maluku Berdasarkan Indikator Pendidikan dengan Menggunakan Metode Ward. Jurnal Statistika dan Aplikasinya, 5(1), e-ISSN: 2620-8369

Gibran, S., Hairani, & Raden, F. (2018). Aplikasi Pemetaan Kualitas Pendidikan di Indonesia Menggunakan Metode K-Means. *Jurnal Matrik*, 17(2).

Kementerian Pendidikan dan Kebudayaan. (2020). *Profil Pendidikan Dasar, Menengah dan Atas Ikhtisar Data Pendidikan* . Jakarta.

Nur, R., I., Memi, N., H., & Fidia, D., T., A. (2020). Penerapan Algoritma K-Medoids pada Pengelompokan Wilayah Desa atau Kelurahan di Kabupaten Kutai Kartanegara. *Jurnal Eksponensial*, 11(2).

Praokta, A. D. (2021). Implementasi Metode K-Medoids Clustering untuk Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Pendidikan. *Journal of Mathematics Education and Applied*, 02(02).

Rofiqi, Y. A. (2017). Clustering Berita Olahraga Berbahasa Indonesia Menggunakan Metode K-Medoid Bersayarat. *Jurnal Simantec*, 6(1).

Santoso, S. (2015). *Pengolahan Data Statistik di Era Informasi.* Jakarta: PT. Alex Media Komputindo.

Sopyan, Y., Agrian, D., & Christiana. (2022). Analisis Algoritma *K-Means* dan *Davies Bouldin Index* dalam Mencari Cluster Terbaik Kasus Perceraian di Kabupaten Kuningan. *Building of Informatics, Technologhy and Science (BITS)*, 10.

Suyanto. (2017). *Data Mining untuk Klasifikasi dan Klasterisasi Data.* Bandung: Informatika.

Sriningsih, M., Djoni, H., & Jantje, D.P. (2018). Penanganan Multikolinearitas dengan Menggunakan Analisis Regresi Komponen Utama pada Kasus Impor Beras di Provinsi Sulut. *Jurnal Ilmiah Sains*, 18(01).

Wahyu, I. S., Fauzan, A. C., & Muhamat, M. (2022). Implementasi Algoritma *K-Medoids* dengan Evaluasi *Davies Bouldin Index* untuk Klasterisasi Harapan Hidup Pasca Operasi pada Pasien Penderita Kanker Paru-paru. *Jurnal Sistem Komputer dan Informatika*, 556-566.

Zulfa, F. (2019). Pengelompokan Provinsi di Indonesia Berdasarkan Indikator Pendidikan Menggunakan K-Means dan K-Medoids. *Jurnal Sains dan Seni Pomits*, 2(2).