PROGRES PENDIDIKAN
Vol. 4, No. 1, January 2023, pp. 50∼57
p-ISSN: 2721-3374, e-ISSN: 2721-9348, DOI: 10.29303/prospek.v4i1.328

K-MEANS AND ELBOW METHOD FOR CLUSTER ANALYSIS OF ELEMENTARY SCHOOL DATA

Vynska Amalia Permadi[1], Sylvert Prian Tahalea[2], Riza Prapascatama Agusdin[3]
[1,2]Informatics, Universitas Pembangunan Nasional Veteran Yogyakarta, Indonesia
[3]Information System, Universitas Pembangunan Nasional Veteran Yogyakarta, Indonesia

Article Info

Article history:

Received: 07-01-2023
Revised: 13-01-2023
Published: 30-01-2023

Keywords:

clustering
k-means
elbow methods
education

ABSTRACT

This research was conducted to find the groups of elementary schools in the Special Capital Region of Jakarta, also known as DKI Jakarta. Elementary school data were selected because it is the first stage of formal education in Indonesia. This research used K-means clustering with the elbow method to determine optimal cluster numbers. The optimal cluster number is three $(k = 3)$ with Cluster 2 having the most members, followed by Cluster 1 and Cluster 0. The data distribution of Cluster 2 shows that the second-most student body and public schools located in East and West Jakarta have an adequate student-to-teacher ratio based on Article 17 of Government Regulation 74, 2008.

Corresponding Author:

Riza Prapascatama Agusdin,
Information System, Faculty of Industrial Engineering,
Universitas Pembangunan Nasional Veteran Yogyakarta,
Jl. Tambak Bayan No. 2, Yogyakarta, Indonesia
Email: rizapra@upnyk.ac.id

1. INTRODUCTION

Education is widely known as one of the programs that can help to prepare and engineer the path that human growth will take in the foreseeable future (Riski, Rusdinal, & Gistituti, 2021). Education is the primary emphasis of the government of the Republic of Indonesia since it serves as the cornerstone upon which the nation's progress is built. The Constitution of the Republic of Indonesia in 1945 states that one of the aims of the Government of the Republic of Indonesia is to educate the people of the nation. Nowadays, education is meeting the challenges of the modern world in order to satisfy one's most fundamental needs (Priatmoko, 2018). It is undeniable that education plays a significant part in forming future generations, and the process remains relevant over time (Sudarsana, 2015). Education must be carried out properly to be able to produce qualified and responsible future generations (Atnawi, 2019).

Based on Indonesia Law number 20 of 2003, formal education has four levels: early childhood education, basic education, secondary education, and higher education. One official educational institution that serves as a platform for information transfer and acquisition is the school. The four formal education stages mentioned earlier are each carried out in four-level schools respectively. Therefore, we might think of schools as a place to cultivate or educate the next generation (Suwartini, 2017). In order to fulfill its responsibility to educate its citizen, the government must administer the facilities that make up the educational system. The government is obligated to work toward ensuring that all areas have access to similarly high-quality educational administration. In a system with equitable management, every school would have an equal opportunity

to acquire appropriate facilities and qualified teaching staff. How educational institutions, such as schools, are managed will affect the overall level of education provided across the country (Julaeha, 2019).

The first stage of primary education in Indonesia is completed at the elementary school level. Elementary schools are prominent places for character-building for their students. Children in elementary school range in age from six to twelve years old, which is the age range in which they are able to think by themselves and replicate what they see. Therefore, elementary schools must pay attention to developing students' personalities. In elementary schools, teachers play a crucial part in developing students' personalities and values. It is essential to remember that a teacher paying attention to their class will be more remarkable if there are fewer pupils than there are teachers, as this will contribute to the overall efficiency and efficacy of the teaching and learning process.

The DKI Jakarta Province is a metropolitan province containing Indonesia's most populous, Jakarta. This province is divided into forty-four districts and six regencies or cities. In the province of DKI Jakarta, the number of elementary schools is expected to reach 2,845 in 2021, with 36,978 teachers and 780,959 students. The Provincial Government of DKI Jakarta needs to consider whether or not it should make an effort to equalize education by increasing the number of teachers in each region. Regional grouping in DKI Jakarta Province needs to be done through the DKI Jakarta Education Office so that the Provincial Government can focus on regional clusters that require education improvement, particularly concerning the teacher-student ratio. This will allow the Provincial Government to pay attention to regional clusters in DKI Jakarta Province.

In this research, the K-Means Algorithm is utilized to categorize data on teachers, students, regencies, and districts at the elementary school level in DKI Jakarta Province during 2021. The data that was used in this research were considered secondary data. Data was acquired from reading the Open Data Jakarta publication. It is anticipated that the findings of this research will be utilized as a contribution by the DKI Jakarta Education Office in the process of formulating policies for the education sector in the following year. These policies will take into consideration the ratio of teachers to students, as well as the distribution of schools according to regencies and districts.

2. RESEARCH METHOD

This research was conducted through data collection, data preprocessing, clustering, and data visualization as shown in Figure 1.

Figure 1. Research flow

2.1. Data Collection

Data was collected from Jakarta Open Data titled "Data Jumlah Siswa dan Guru SD Provinsi DKI Jakarta Tahun 2021", which contains information regarding the numbers of students and teachers in elementary schools in the Special Capital Region of Jakarta, Indonesia. There are 2,312 data were utilized, and the sample is shown in Table 1.

Table 1. Sample of top five data from the data set

	School Names	NPSN	Level	School Status	Students	Teachers	District	Region
0	SDN Nurul Iman	20109091	Elementary	Private	496	44	Duren Sawit	East Jakarta
1	SD Negeri Cakung Barat 15 Pagi	20104283	Elementary	Public	850	38	Cakung	East Jakarta
2	SD Negeri Penjagalan 01 PG	20104891	Elementary	Public	885	38	Penjarungan	North Jakarta
3	SDS Muhammadiyah 24	20109150	Elementary	Private	685	34	Pulo Gadung	East Jakarta
4	SDN Lubang Buaya 06 Pagi	20104293	Elementary	Public	761	32	Cipayung	East Jakarta

2.2. Data Preprocessing

Data preprocessing is conducted after data collection is finished. Data preprocessing was used in this research, such as parameter selection and data transformation. Parameters used in this research are school status, students, teachers, districts, and regions. Meanwhile, the data transformation is used to encode the status, district, and region attributes from non-numerical into numerical ones using the Python programming function Ordinal Encoder from the Sklearn package.

Clustering is a method to group data by their similarities in attributes (Adhitama, Burhanuddin, & Ananda, 2020; Irhamni, Damayanti, & Khusnul K, 2014). This method has no specific target output since the data will be put into different groups depending on their attributes. This research was conducted using the K-means algorithm and elbow method.

The k-means algorithm is an algorithm that groups data based on the closest similarity to a centre point, also known as a centroid (Nainggolan, Perangin-Angin, Simarmata, & Tarigan, 2019). The similarity was obtained by calculating the distance between the data and the centroid, this research in particular, using the euclidean distance formulated as shown in equation (1).

$$D\left(i,j\right)=\sqrt{\left(x_1i-x_1j\right)^2+\left(x_2i-x_2j\right)^2+\cdots+\left(x_ki-x_kj\right)^2} \tag{1}$$

Where, $D(i,j)$ is the distance from data $[i]$ to centroid $[j]$, x_ki is data $[i]$ from k attribute, and x_k j is centroid $[j]$ from k attribute. The K-means algorithm is one of the most used algorithms because of its simplicity and efficiency (Nainggolan et al., 2019).

2.3. Data Visualization

Data visualization is a method that is often conducted to effectively analyze data (Runkler, 2020). There are many techniques to visualize data effectively, but this research only uses histogram plots.

3. RESULT AND DISCUSSION

This study utilized the Jakarta Open Data titled "Data Jumlah Siswa dan Guru SD Provinsi DKI Jakarta Tahun 2021". This dataset contains information regarding the number of students and teachers in SD DKI Jakarta Province, including: school name, NPSN number, school education level, school status (public/private), amount of students, the quantity of teachers, school area districts, and school area administrative cities/towns. A total of 2,312 data were utilized, with a preview in Table 2.

Table 2. Sample of top five data from the data set

	School Names	NPSN	Level	School Status	Students	Teachers	District	Region
0	SDN Nurul Iman	20109091	Elementary	Private	496	44	Duren Sawit	East Jakarta
1	SD Negeri Cakung Barat 15 Pagi	20104283	Elementary	Public	850	38	Cakung	East Jakarta
2	SD Negeri Penjagalan 01 PG	20104891	Elementary	Public	885	38	Penjarungan	North Jakarta
3	SDS Muhammadiyah 24	20109150	Elementary	Private	685	34	Pulo Gadung	East Jakarta
4	SDN Lubang Buaya 06 Pagi	20104293	Elementary	Public	761	32	Cipayung	East Jakarta

In the first stage, data preparation is carried out by selecting attributes/parameters to construct the cluster. Only the last five data items shown in Table 2 will be used: school status, students, teachers, district, and region. Figure 2 depicts the distribution of values for each of these characteristics. The next step in data preparation is to encode the data to adhere to the K-Means model, which can only be applied to numerical or categorical data. The status, district, and region attributes which were initially composed of non-numeric characters converted into numeric-categorical data starting from zero. The last encoding value is the maximum amount of unique values for each variable. The dataset, for example, includes 44 sub-districts in DKI Jakarta, resulting in an encoding value range of 0 to 43. The dataset for this study is encoded using the Python programming function OrdinalEncoder from the Sklearn package. Table 3 illustrates the results of data encoding.

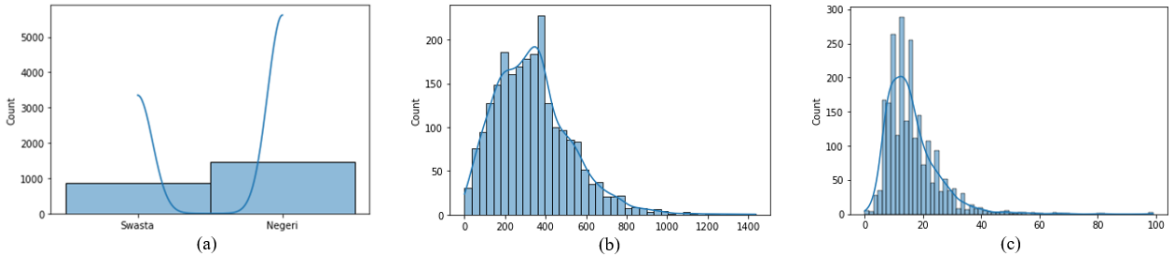

Figure 2. Distribution of attribute values: status (a), amount of students (b), and quantity of teachers (c)

Table 3. Sample of top five data from the data set

	School Status	Students	Teachers	District	Region
0	01.00	496	44	07.00	04.00
1	00.00	850	38	00.00	04.00
2	00.00	885	38	33.00.00	05.00
3	01.00	685	34	35.00.00	04.00
4	00.00	761	32	05.00	04.00

Finding insightful patterns or information in selected data using specific techniques or methods, the so-called data mining process, is the following step. Data mining techniques, procedures, and algorithms vary greatly, but in this study, the K-Means algorithm and the Elbow method will be used to group data into clusters. K-means clustering is used to generate k clusters from given data based on their similarity in characteristics. In contrast, the elbow is cast-off to determine the best quantity of clusters so that the K-means data grouping is more optimal by performing a quadratic difference calculation on several different test k values (1 to 10). The greater the value of k, the smaller the average degree of distortion becomes. At the optimal value of k, the highest level of distortion increases, and an elbow forms. As the simplest term, the best k can be determined by visualizing the data; the optimal number of k clusters occurs when each cluster forms a data group that is not stacked on top of each other, that results in a grouping error. Figure 3(a) depicts the results of implementing the elbow method to determine the optimal k value, with a black broken line indicating the formation of an elbow at $k = 3$.

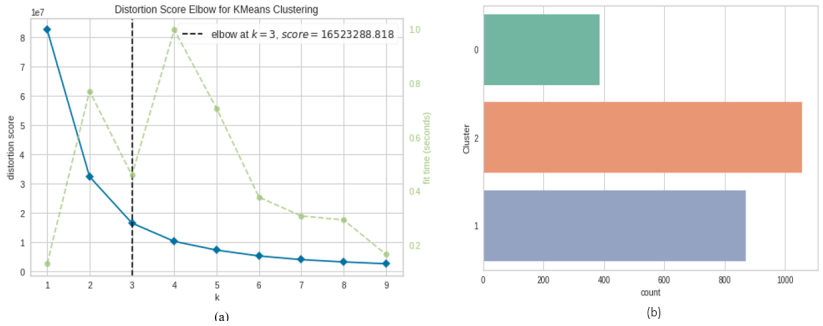

Figure 3. K-Means clustering results: Optimal k determination by elbow method (a) and cluster distribution (b)

After determining the optimal number of k, which is 3, the K-Means algorithm will be used to recognize the pattern in each row of data and divide it into three clusters. Figure 3(b) depicts the cluster distribution results. Cluster 2 has the most members, followed by Clusters 1 and 0. The clustering results will then be saved and added to the dataset for future analysis. Table 4 shows a dataset preview following the K-Means algorithm's implementation. The interpretation of cluster results formed by the K-Means algorithm is the final step in this study after obtaining the cluster categorization. To interpret the data, we first plot or visualize the data distribution of each cluster, as shown in Figures 4 to 6. Following that, the characteristics of each cluster are investigated based on the data characteristics for each attribute.

Table 4. Top five data of clustering result

	School Status	Students	Teachers	District	Region	Cluster
0	01.00	496	44	07.00	04.00	2
1	00.00	850	38	00.00	04.00	0
2	00.00	885	38	33.00.00	05.00	0
3	01.00	685	34	35.00.00	04.00	0
4	00.00	761	32	05.00	04.00	0

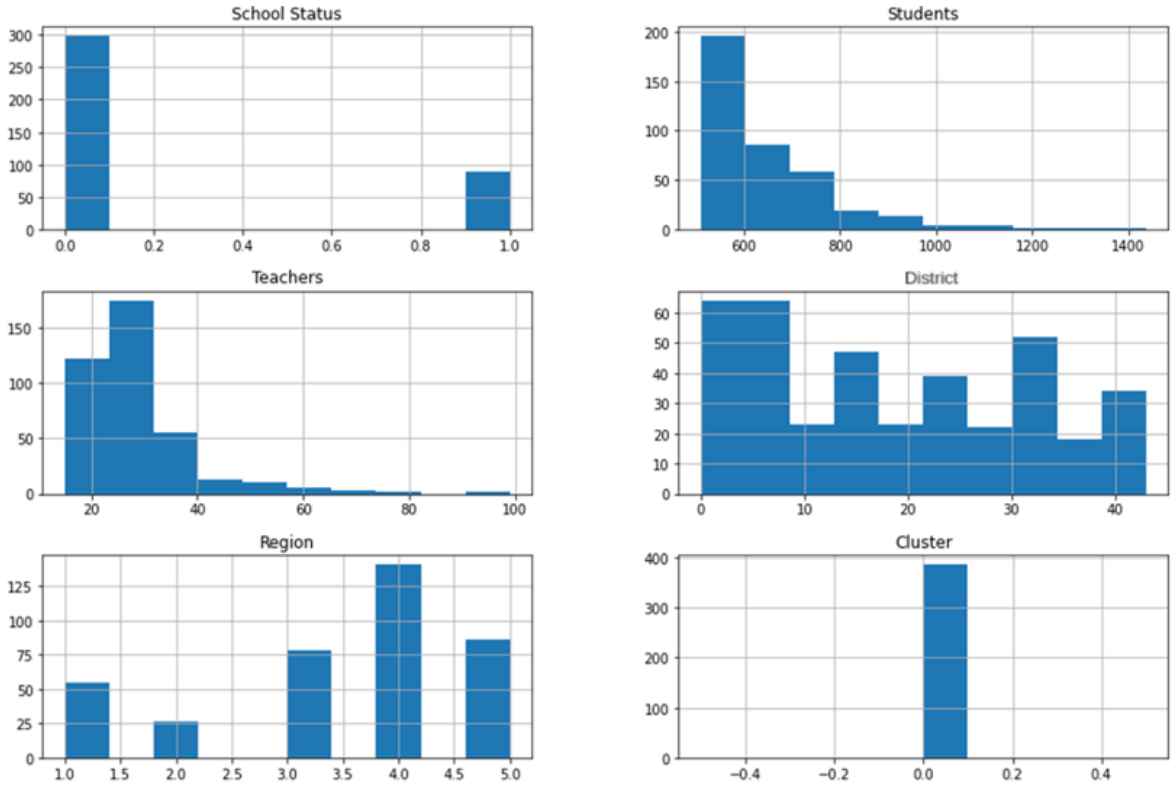

Figure 4. Information summaries for each attribute of Cluster 0

The data distribution characteristics of each attribute for Cluster 0 are depicted in Figure 4. This cluster includes 386 schools, the majority of which are state schools with 511 to 1436 students and a minimum of 15 to 99 teachers. East Jakarta (region code 4.0) has the most schools (141), followed by North Jakarta (area code 5.0), which has 86 schools, and South Jakarta (region code 3.0) with 78 schools. Despite having the fewest members, this cluster has a solid character: a disproportionate number of students, which is over twice the size of the schools in Cluster 1. With a significant number of students, this cluster has the most teaching personnel compared to other clusters. The fact that several schools in this group have examined the teacher-student ratio, though it may not be perfect, is a plus. This cluster has 10 schools with over 1000 students, which are distributed throughout eight districts in three regions: Central Jakarta, North Jakarta, and East Jakarta.

Furthermore, compared to the population of DKI Jakarta, which is expected to be 10.6 million in 2021, according to data from Statistics Indonesia (BPS), East Jakarta has the most residents, 3.05 million people, or 28.81% of the total population of DKI Jakarta. As a result, it is not surprising that 141 schools in East Jakarta are included in this cluster. With such a significant population, there must be adequate schools, particularly at the primary level. However, more research is needed to determine whether the number of residents aged 6-7 (who should be starting primary school) surpasses the set quota.

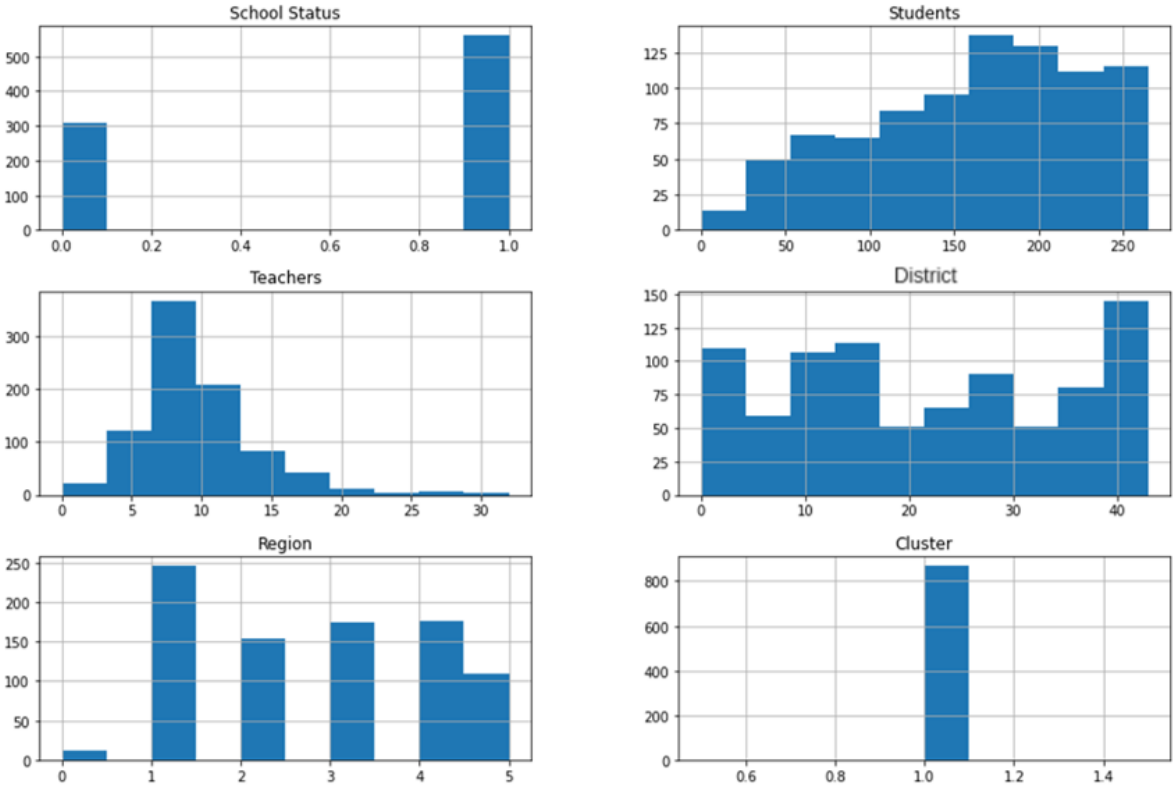

Figure 5. Information summaries for each attribute of Cluster 1

Figure 5 displays the frequency analysis of each attribute value for the 864 schools (out of 2311) that constitute Cluster 1. This cluster contains 65% of the private schools in DKI Jakarta and 30% located in the West Jakarta neighborhood (area code 1.0). The proportion of students is quite exciting, having only 263 students at most and the number of teachers ranging from 3 to 49 individuals. However, despite having a small student body, 2% of schools in this cluster have a student-to-teacher ratio greater than 1:20. As the number of schools in Cluster 0 and Custer 2 can reach up to twice as high than this cluster, it should be anticipated that the number of schools in this cluster will be reduced in the future in order to achieve a more equitable distribution of students among all schools in DKI Jakarta.

Compared to the other two clusters, Cluster 2 data distribution characteristics (given in Figure 6) consists of schools having the second-most student body and generally are public schools located in East Jakarta (region code 4.0) and West Jakarta (region code 1.0). A further significant and intriguing aspect of this cluster is its relatively small average teaching staff size of 16 people. As stated in Article 17 of Government Regulation 74 of 2008 Concerning Teachers, the recommended student-to-teacher ratio for primary schools is 20:1; hence this cluster has an adequate student-to-teacher ratio. Moreover, one district appears to have hundreds more schools than the others. Still, because the members of this cluster are the largest compared to those of the other two clusters, we observe that the difference is still within tolerable limits.

Afterward, we compare the facts about these schools to the primary school accreditation rating from the National Accreditation Board for Schools/Madrasah (BAN-S/M). According to the comparison result, several educational institutions in this cluster are ranked among the ten top elementary schools in each region. Similarly, Cluster 0 schools are underrepresented in the top 10 in most regions. Based on these findings, we can conclude that having a large number of students in clusters 0 and 2 does not necessarily have a negative impact as long as the school maintains a student-to-teacher ratio lower than the established standard. We can also imply that these schools have a high level of interest, which also influences the ranking and accreditation of affiliated institutions. Schools in Clusters 0 and 2 may benefit from having more students and educators than schools in Cluster 1, as their children will have further access to more comprehensive educational experiences; hence the proportion of exceptional students will likely increase too.

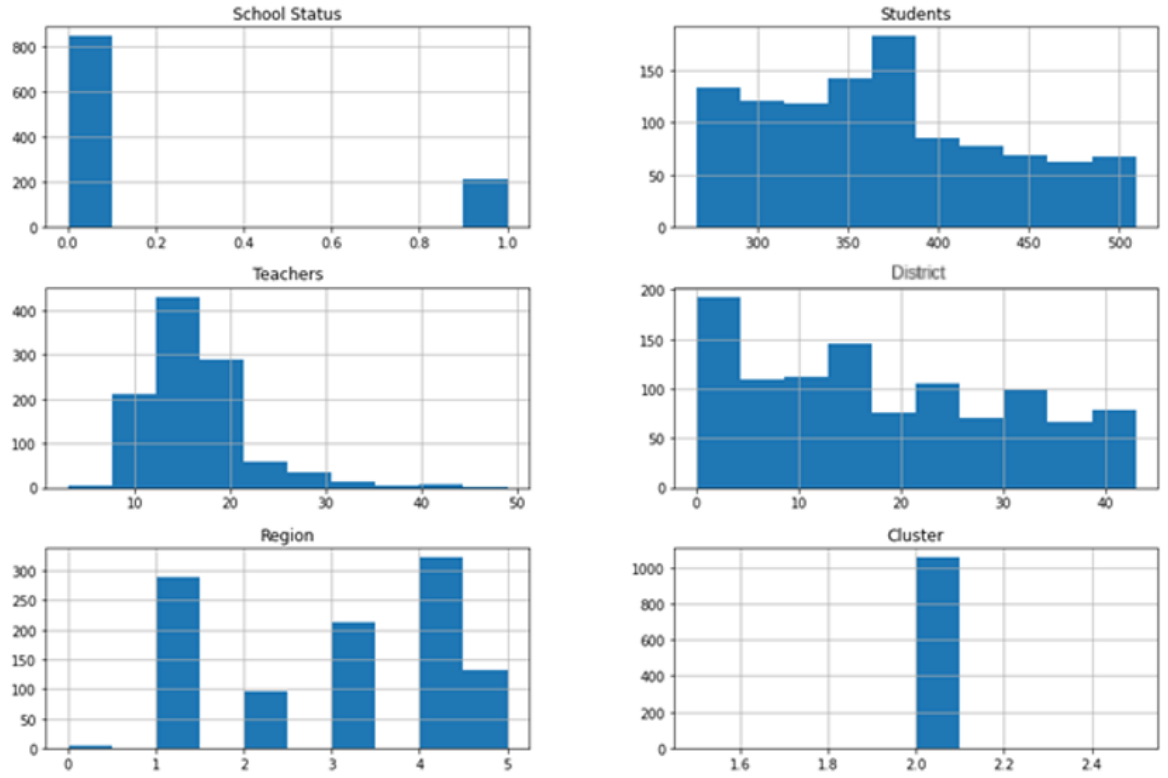

Figure 6. Information summaries for each attribute of Cluster 2

4. CONCLUSION

This research was conducted using K-means clustering with the elbow method to determine an optimal number of clusters. The result of the elbow method shows that the optimal value is three $(k = 3)$. The biggest cluster is Cluster 2, followed by Cluster 1 and 0. Cluster 2 has a relatively small number of teachers, approximately 16 teachers; also its member primarily located in East and West Jakarta. Cluster 0 has 386 members, mostly in East Jakarta, with 86 schools. Schools in this cluster have the most teachers compared to another cluster despite having the fewest members. Meanwhile, Cluster 1 has more private school as members, stated in West Jakarta, and 2% of its member has a student-to-teacher ratio greater than $1 : 20$. Members in cluster 0 benefit with the student body, and member in cluster 2 has the most teacher. Still, schools in cluster 1 have more comprehensive educational experiences because they have a better student-to-teacher ratio. Hence, the schools in cluster 1 will likely produce more exceptional students.

REFERENCES

Adhitama, R., Burhanuddin, A., & Ananda, R. (2020). Penentuan Jumlah Cluster Ideal SMK di Jawa Tengah Dengan Metode X-Means Clustering Dan K-Means Clusterin. *JIKO (Jurnal Informatika dan Komputer)*, *3*(1), 1–5. doi:10.33387/jiko.v3i1.1635

Atnawi, A. (2019). Pengaruh Kedisiplinan Terhadap Tingkat Prestasi Belajar Siswa Di SDN Murtajih Pamekasan. *Al-Ulum Jurnal Pemikiran Dan Penelitian Ke Islaman*, *6*(2), 1–10.

Irhamni, F., Damayanti, F., & Khusnul K, B. (2014). Optimalisasi Pengelompokan Kecamatan Berdasarkan Indikator Pendidikan Menggunakan Metode Clustering dan Davies Bouldin Index. In *Seminar nasional dan teknologi umj* (pp. 1–5).

Julaeha, S. (2019). Problematika kurikulum dan pembelajaran pendidikan karakter. *Jurnal Penelitian Pendidikan Islam*, *7*(2), 157.

Nainggolan, R., Perangin-Angin, R., Simarmata, E., & Tarigan, A. F. (2019). Improved the Performance of the K-Means Cluster Using the Sum of Squared Error (SSE) optimized by using the Elbow Method. *Journal of Physics: Conference Series*, *1361*(1). doi:10.1088/1742-6596/1361/1/012015

Priatmoko, S. (2018). Memperkuat Eksistensi pendidikan Islam di era 4.0. *TA'LIM: Jurnal Studi Pendidikan Islam*, *1*(2), 221–239.

Riski, H., Rusdinal, R., & Gistituti, N. (2021). Kepemimpinan Kepala Sekolah di Sekolah Menengah Pertama. *EDUKATIF: JURNAL ILMU PENDIDIKAN*, *3*(6), 3531–3537.

Runkler, T. A. (2020). Data Visualization. In *Data analytics* (pp. 37–59). doi:10.1007/978-3-658-29779-4_4

Sudarsana, I. K. (2015). Peningkatan mutu pendidikan luar sekolah dalam upayapembangunan sumber daya manusia. *Jurnal Penjaminan Mutu*, *1*(1), 1–14.

Suwartini, S. (2017). Pendidikan karakter dan pembangunan sumber daya manusia keberlanjutan. *Trihayu: Jurnal Pendidikan Ke-SD-An*, *4*(1).