Journal of Information Systems and Informatics
Vol. 4, No. 4, December 2022 e-ISSN: 2656-4882 p-ISSN: 2656-5935
http://journal-isi.org/index.php/isi Published By DRPM-UBD

Analysis of School Grouping Against Educational Teachers in NTB Using K-Prototypes Method

Gloria Milenda Jesika Prima[1], Charitas Fibriani[2]

[1,2]Program Studi Sistem Informasi, Universitas Kristen Satya Wacana, Salatiga, Idnonesia
Email: [1]682018073@student.uksw.edu, [2]charitas.fibriani@uksw.edu

Abstract

The shortage of teaching staff is still a problem in education in Indonesia, resulting in the number of teachers and students being unbalanced. It is very significant based on data from the Analysis of the Distribution of Teachers in Region 3T, which shows that the national ratio of teachers per school is 18.41, and many still have a ratio of teachers per school lower than the national. Therefore, it is necessary to group the data in one of the 3T regions, especially in the Province of West Nusa Tenggara (NTB). The data used for this study were 4997 schools from 8 districts and two cities. The results of this study were conducted to determine whether the grouping ratio of the number of teachers and students is ideal or not ideal. The data used are categorical and numeric data types, so the clustering analysis method used is K-Prototypes. Each cluster produces a different range of teachers and students. Cluster 1, which obtained the High category almost followed the ideal, while Cluster 2 and the Cluster 3 were still far from ideal. This needs to be considered to increase the number of teaching staff in the Education Office in each district.

Keywords: Clustering K-Prototypes, Educational Teachers, School Grouping

1. INTRODUCTION

Schools are educational facilities where teachers and students learn and teach. The teaching staff is professionally responsible for developing students' skills, teaching, and devoting themselves to education throughout Indonesia. Quality schools are not run alone and are not born because of complete facilities [1]. Thus, quality schools must be formed and appropriately implemented. One of the expectations for the quality of teaching and learning education to educate all students is the teacher or teaching staff.

Based on Law no. 20 of 2003 regarding the National Education System paragraph 6, Teachers, or in other words Educators, are educational staff who are qualified as teachers, lecturers, tutors, instructors, tutors, facilitators, or others following their fields and participate in providing education [2]. The teacher or teacher plays a role in the learning process to guide, teach, and educate students at the primary

to secondary education level. Meanwhile, the analysis applied to plan the need for teaching staff at elementary to high school

levels is the comparison of the number of teachers and students for education units based on the level of the education unit. The ratio of teaching staff that is applied based on Government Regulation no. 74 of 2008 explains about teachers' article 17 for elementary, junior high, high school, vocational education levels as follows: elementary education level is 20:1, junior high school education is 20:1, high school education is 20:1, and vocational education is 15:1 [3].

Based on this, the Government has provided services and facilities to ensure quality education for every citizen. In 2020-2024, the Government already has Presidential Regulation Number 63 of 2020 concerning the Determination of Disadvantaged Regions in every 3T region. The 3T regions are the Frontier, Outermost, and Disadvantaged Regions [4]. The province of NTB is included in the 3T area, namely in the district of North Lombok. However, based on an teaching staff in the Province of West Nusa Tenggara (NTB) and one from the volunteer education community in NTB, information was obtained that the teaching staff has almost been fulfilled for the capital area.

In contrast, outside the capital, there are still many shortages of teaching staff, especially in remote areas. In the village, only three teachers teach, and even one teacher teaches more than one subject. The shortage of teachers is one of the problems in Indonesia's education field. Some factors for the lack of teaching staff are the first lack of access to education services in 3T areas and more attention in urban areas than other areas. Second, communication difficulties in teaching and learning aids. The third is the unbalanced distribution of teaching staff [5].

West Nusa Tenggara (NTB) is one of the Disadvantaged Regions, which is part of the 3T Region. From 122 districts, the number of 3T regional schools is 35,478. NTB has 4,997 elementary, junior high, high school, and vocational school levels. NTB has eight regencies consisting of Bima, Dompu, West Lombok, East Lombok, North Lombok, Sumbawa, and West Sumbawa, and two cities consisting of Bima City and Mataram City. Based on data on the ratio of teachers per school in underdeveloped areas, the national average is 18.41 [6]. Then the learning process improves because the number of teachers for each subject is more than the number of subjects. The province of NTB, which has exceeded the national average, is Kab. Bima and Kab. Dompu, the rest is still below average. However, the analysis results do not indicate which regional schools need attention. The number of teachers is still tiny. Therefore, it is necessary to classify schools with many students, but the number of teachers is still lacking. So, it is necessary to know what the school teachers need so that all schools in West Nusa Tenggara are evenly distributed.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

School grouping analysis is done by implementing clustering. Clustering is a mining technique to group data based on similarities without knowing the previous group or cluster. One of the clustering methods in the form of a combination of numeric and categorical data is K-Prototypes because the data obtained from the Main Data of the Ministry of Education and Culture is in the form of Category and Numerical Data Types.

Based on previous related research, research has been carried out in the Tegal City area with the Chi-Square-based K-Means method supported by a decision system to identify disparities in teacher needs in 2018. The resulting discussion displays the accuracy of labeling in the resulting clusters as decision support in identifying teacher availability in Tegal City. The Chi-Square test results are in the form of nine clusters, and there are three clusters of teacher availability with the categories of High, Fair, and Less [7].

Then the Regional Grouping based on facilities and COVID-19 events in the Semarang City area using the Fuzzy K-Prototypes Clustering algorithm, a previous study, was carried out in 2021. The results obtained were 2 clusters of the number of events and COVID-19 patients directly proportional to the number of available health facilities. Comparison of Cluster 2 is 2.39 times than Cluster 1[8].

Furthermore, the Design and Development of a Geographic Information System for Educational Achievements as a supporter of the Education and Culture policy of the province of West Nusa Tenggara with the proposed results based on a digital map information system of high school education indicator data in Mataram City and West Lombok Regency based on per sub-district bounded by polygons and per year. [9].

The K-Means and Fuzzy K-Prototypes methods have been carried out based on previous research. The number of clusters obtained has its respective advantages. Overall, the visualization results of how many schools and the number of health facilities have not been transparent. Therefore, this study aims to group school data based on comparing the number of students and the number of teachers and see the number of schools that still lack teachers using data visualization.

This study took data from the Ministry of Education and Culture's Basic Education Data in the form of Variable Names of Schools, Education Boards, Number of Teachers, Number of Students, and others. The focus of the results of this study is to find out which schools need teachers as soon as possible. The result is a grouping of school data based on the number of students and teachers. The results provide input to the Regional Education Office to pay attention to allocating the number of teaching staff in each sub-district.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

2. METHODS

Figure 1 There are six stages of research conducted in the School Grouping of Education Teaching Personnel in NTB can be seen as follows.

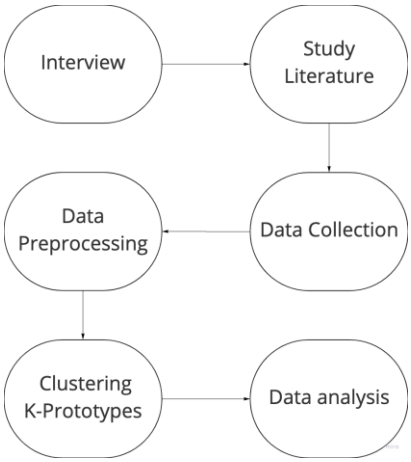

Figure 1. Stages of Research

2.1. Interview

The results of the first stage of research understand the identification of problems that exist in the problems of teaching staff in the Province of West Nusa Tenggara (NTB), so it is necessary to conduct interviews with schools in Central Lombok Regency and from the education office in NTB, sharpen information about the obstacles faced by the shortage of teaching staff in NTB.

2.2. Study Literature

The second stage, the literature study determines the purpose of this research to study the literature related to clustering, education problems in the province of NTB and the 3T regional education model. The literature sources were obtained from various journals, e-books, and articles that support the research.

2.3. Data Collection

In the third stage, the data collection in this study was taken from the Ministry of Education and Culture's primary data, which anyone can access. The data used is student and teaching staff data for the 2021/2022 academic year in West Nusa Tenggara[10]. The dataset consists of 10,495 data, all names of schools in NTB, and there are 14 variables in it.

2.4. Data Preprocessing

The fourth stage is data preprocessing which aims to prepare before clustering. Initial data processing will be divided from several variables used in this study, namely Name_School, Status, Regency, BP, PD, and Teacher. Furthermore, the data taken is only the level of education bodies from Elementary School (SD), Junior High School (SMP), Senior High School (SMA), and Vocational High School (SMK), so the data that has been filtered is 4997 data. At the Data Preprocessing stage, the data used in the process results are not always in ideal conditions for processing and sometimes there is data that fits the system [11]. Therefore, it is necessary to organize the data in a specific order. Selection of variables used to group data. Variables consist of categorical and numerical data types as follows.

Tabel 1. Data Type

No	Attribute Name	Data Type	Description
1	School_name	Categorical	All names of schools in NTB
2	Regency	Categorical	Consists of 10 districts located in NTB
3	Education egency (Badan Pendidikan)	Categorical	School Education Level from Early Childhood to High School/Vocational High School
4	Student (Peserta Didik)	Numerical	Number of Students or Students in each school
5	Teachers	Numerical	Number of teachers or teaching staff in each school

After collecting and knowing the categorical and numerical data types in Table 2, it groups the schools using Clustering.

2.5. Clustering K-Prototypes

The definition of *clustering research* is the process of grouping data into several groups so that the data in one cluster has a maximum and minimum level of similarity [12]. Clustering is one part of the Unsupervised Learning Algorithm. Unsupervised Learning is a Machine Learning technique without dependent variables and applies the Learning in the associated data set. The benefit of this cluster is that it is helpful as an object of research regarding data grouping, so that close distances between clusters and other clusters can be found.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

K-Prototypes is an integrated process of k-means and k-modes for data grouping methods with categorical and numeric data types [13]. K-prototypes measure the dissimilarity or dissimilarity measure used differently from k-modes. So the inequality measure combines the Euclidean distance equation with the dissimilarity measure contained in k-modes [14]. The following is the inequality between two mixed objects of type X and Y variables, with explanations by Huang A1r, A2r, …, Azr, Az+1c, …:

$$b_2(X, Y) = \sum_{j=1}^{p} \left(x_j - y_j\right)^2 + \gamma \sum_{j=p+1}^{m} \delta(\chi_j - y_j) \quad (1)$$

The first term measures the squared Euclidean distance on a numeric attribute and the second term is a simple matched dissimilarity measure on a categorical attribute. Weights are used to avoid selecting one type of attribute. The influence on the clustering process was discussed in a previous study [13]. According to research by Huang, there are several stages of explaining the stages using K-Prototypes, namely the first stage of selecting the initial k-prototype from the X data set for each cluster. The second stage allocates each object in X to the cluster whose prototype is closest to the equation. After that, the cluster prototype is updated after each allocation. In the third stage, after all, objects have been assigned to a cluster, retest the similarity of objects to the initial prototype if an object is obtained so that the result of the closest cluster distance becomes the object being measured. In the last stage, after allocating all the observation data in the dataset to the new prototype. Repeat if there are still clusters that change until no objects change clusters after being tested. Some of the testing two processes for K-prototypes for this research are as follows. Cluster results see that the values obtained have diversity. If a cluster of elements is the same and between clusters, the details are more different. In the sense that the variety within the cluster is narrower and the diversity between the clusters is wider [15]. The First is Elbow Method to determine the number of clusters in the K-Means algorithm. The Elbow method has the benefit of finding the optimal cluster to produce the best Clustering[16]. The Second is Cluster Centroids. After determining the cluster value, the next step is calculating the distance between the similarities from each data, namely the centroid result. The centroid will change Back to Steps to calculate the distance with the data, while if the centroid remains, it will stop, and the clustering results are obtained [17].

3. RESULTS AND DISCUSSION

3.1 Result of Clustering

The results and discussion of this study discuss the results of grouping school data based on the ratio of teachers to students in NTB. The first part discusses the

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

results of grouping school data from each cluster with the implementation of the results using Tableau visualization. Furthermore, the results of the analysis focus on the parameters that will be observed in each cluster, namely the number of teachers and the number of students, the ratio of students to teachers, the number of schools based on education level, and finally, the grouping of ratios by the district.

3.1.1 Elbow Method

The results of the distance value of the Elbow method are displayed with an SSE (Sum Square Error) score. The explanation is below.

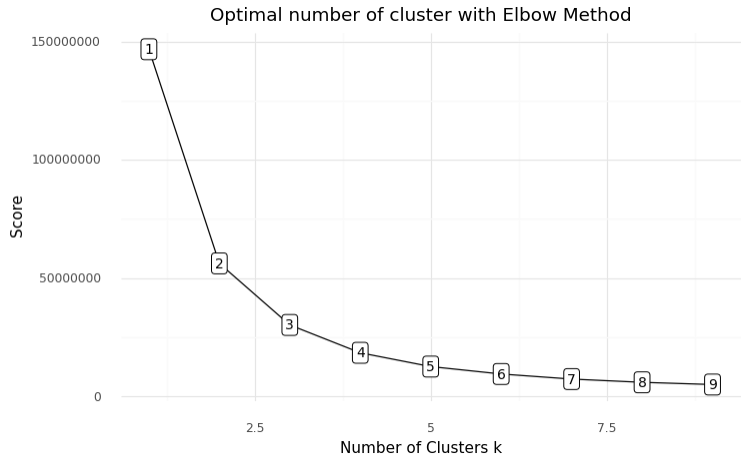

Figure 2. Elbow Method

Table 2. SSE Score

No	SSE Score
1	146685746.625
2	56237943.400
3	30233886.496
4	18475554.027
5	12653141.754

Figure 2 and Table 2 visualize the Elbow method and the Elbow Value using python. The optimal cluster value is k=3 because it shows an angled curve, and the SSE score shows the distance from Cluster 1 to Cluster 2. The distance value has decreased drastically. However, because the diagram does not show the point of a right triangle, the best result from the Elbow Method is k= 3. After knowing the best cluster, the next step is to use the K-Prototypes Clustering method.

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

3.1.2 Cluster Centroids

```
# Cluster centorid
kprototype.cluster_centroids_

array([['852.6424581005587', '57.100558659217874', 'SD IT ANAK SHOLEH',
        'Lombok_Timur', 'SMA'],
       ['283.5886699507389', '19.713464696223316', 'SD NEGERI 1 BELEKA',
        'Lombok_Timur', 'SD'],
       ['101.39277777777778', '10.643611111111111',
        'SMKS ISLAM DARUSSALAM', 'Lombok_Timur', 'SD']], dtype='<U32')
```

Figure 3. Centroid Results

The results of Centroid Figure 3 above are the frequency of data that often appears or the mode of each cluster using the Python programming language. In cluster 1, the result of the centroid of the number of students is 852.6424, and the centroid of the teacher is 57.1005, while clusters 2 and 3 are lower than the centroid of cluster 1. All Cluster 1 to Cluster 3 has the same centroid value, namely East Lombok Regency. This differs from the k-mean and k-modes centroids because k-prototypes have two data types.

3.2 Discussion Analysis

Previously, we already knew the best number of clusters through the K-prototypes method. Next, analyze the results of the clustering effectiveness of school grouping based on the number of teachers and students.

Figure 4. Exploration Data Analyst (EDA)

Figure 4 Exploration Data Analysis (EDA) using Tableau based on School has three groups cluster 1 in blue, cluster 2 in orange, and cluster 3 in red. The distance

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

between cluster 1 and cluster 2 is very close, while the distance between cluster 2 and cluster 3 is quite far. The explanation of each cluster is as follows:
Cluster 1: is the number of students in the school with a range of 498 to 1800 students and a range of teachers from 14 to 146.
Cluster 2: is the number of students in the school with a range of 0 to 139 students and a range of teachers from 0 to 64.
Cluster 3: is the number of students in the school with a range of 138 to 495 students and a range of teachers from 1 to 88.

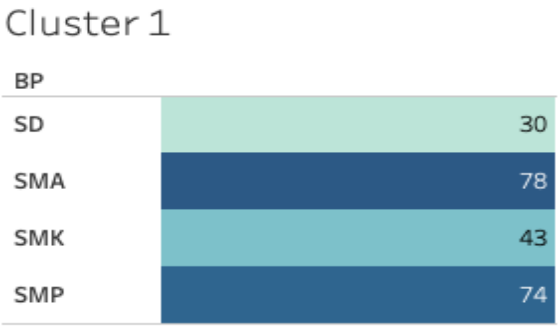

Figure 5. Cluster 3 Results of Grouping School

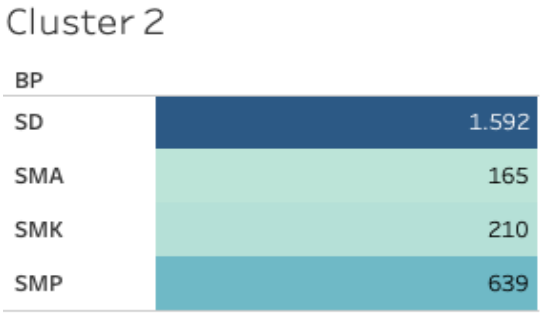

Figure 6. Cluster 2 Results of Grouping School

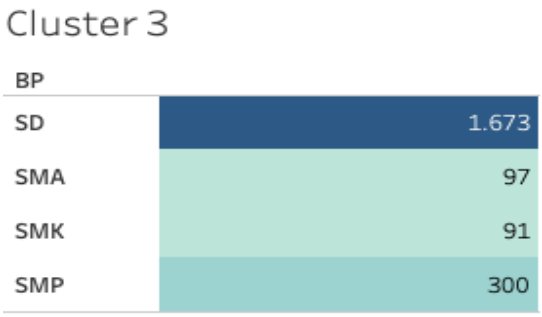

Figure 7. Cluster 3 Results of Grouping School

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

Figure 5 to 7 shows the results of grouping school education level data in NTB. In the number of schools in each cluster, it can be seen that Cluster 1 has the highest number of high school education levels, namely 78 schools. In contrast, Cluster 2 and Cluster 3 have the highest levels of elementary school education. The total number of schools in each cluster is Cluster 1, as many as 225 schools. Cluster 2 has as many as 2607 schools, and Cluster 3 has as many as 2161 schools.

Table 3. Results of Grouping Ratio

	Cluster 1	**Cluster 2**	**Cluster 3**
Total Student	176961	203191	482462
Total Teachers	11912	26373	34225
Avarage Teachers	52,94	10,08	15,83
SD, SMP, SMA Ratio	1,34	2.57	1,38
SMK Ratio	1,02	2,13	1,53

Table 3 shows the results of grouping the ratio of total teachers to total students in NTB. Following the Law, the ratio of elementary, middle, and high school education levels is 20:1 and for vocational schools is 15:1. In cluster 1, which is a HIGH category because the ratio results are almost ideal. Furthermore, cluster 2 is a LOW category because the results of the ratio are far if it is said to be ideal. Cluster 3 is in the MEDIUM category because the results of the ratio at the elementary to high school education levels are not much different from the Low category. A balanced ratio is 1, so the higher the ratio, the more than 1, and the more students there are than the teacher. Otherwise, if the ratio is less than one, the number of teachers is more than the number of students.

Judging from the elementary to high school levels in clusters 2 and 3, the ratio of teachers and students in NTB is still far from the ideal ratio, while in cluster 1, the average is almost ideal. However, there is a significant gap between the high to low category clusters, which means that the distribution of teachers is not yet fully distributed throughout the NTB region. If visualized, the grouping of ideal and not ideal schools from each district is as follows.

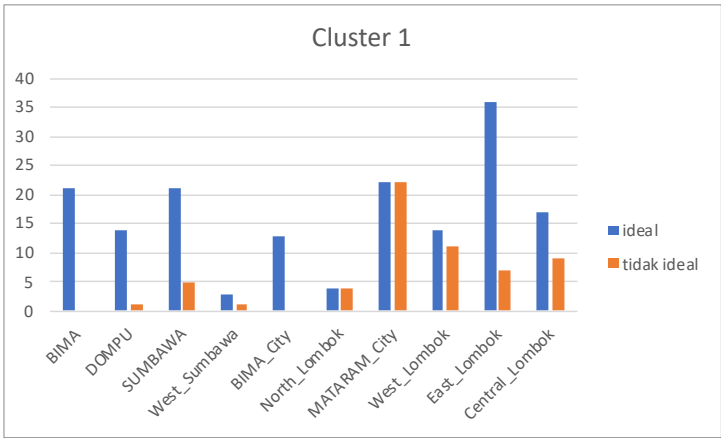

Figure 8. District Cluster 1 Bar Chart

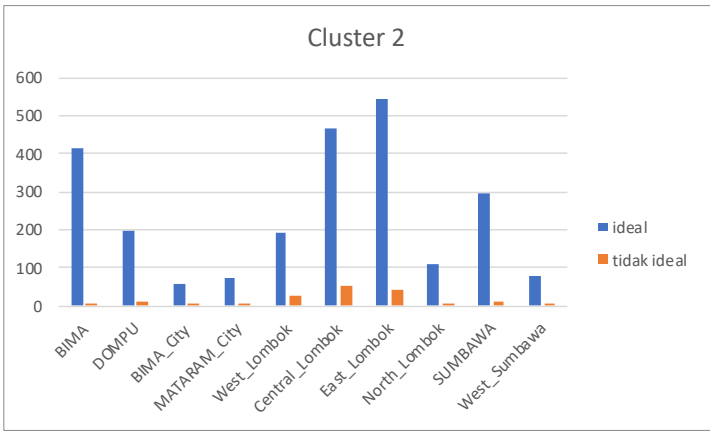

Figure 9. District Cluster 2 Bar Chart

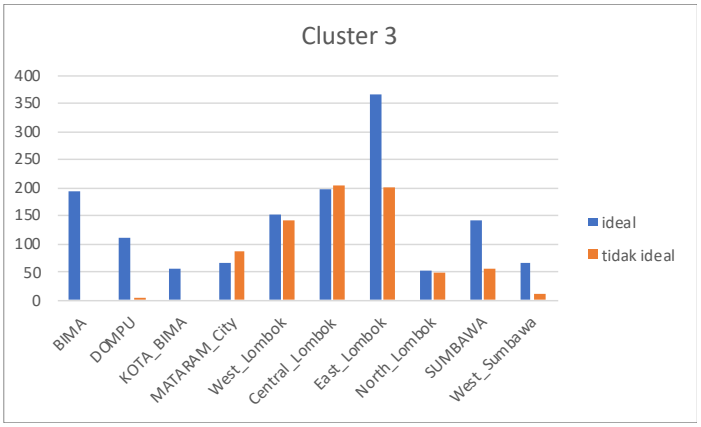

Figure 10. District Cluster 3 Bar Chart

Figure 8 to 10 shows the results of grouping the ratio of the total number of ideal and non-ideal schools from each district. Cluster 1 is the number of schools in each district with an ideal total of 165 schools and not ideal 60 schools. North Lombok and Sumbawa districts are said to be ideal because they have been fulfilled. Cluster 2 is the most significant number of schools than the other clusters, with an ideal total of 2436 schools and not ideal 175 schools. Cluster 3 has an ideal total of 1410 schools and not ideal 762 schools. The results of the entire cluster show that in East Lombok Regency, the number of schools that fit the ideal is 945 schools, while the number of schools that are not ideal is Central Lombok Regency with 266 schools. This is what needs to be considered at this time to require teaching staff, namely Cluster 1 because it has the most significant number of students compared to other clusters. Cluster 3 also needs to be considered in the future because it has a small number of teachers and students. This is due to the lack of adequate school facilities.

4. CONCLUSION

Based on the study's results, it can be concluded that clustering with the K-Prototypes method can be designed as a grouping of school data still lacking for teaching staff in NTB. Each cluster has different characteristics. Cluster 1 has a ratio difference of 0.34, which is almost ideal because the number of schools is small and the number of students is the most, so it is necessary to need teachers in a short time. Furthermore, Cluster 2 has a far ratio that is said to be ideal, and because it has the most significant number of schools and the fewest number of students and teachers, it is necessary to require teachers for a long time. Cluster 3 has an almost ideal ratio like Cluster 1 because it has an average number of schools and a variety of students and teachers. Some have the highest number of students or vice versa, so teachers are needed immediately after running Cluster 1. In the school grouping comparison of teachers and students, the grouping cannot be a reference because the number of teachers, both in quality and quantity, is due to the number of students and teachers in each school is not evenly distributed. Because the distribution of teachers in NTB still experiences a gap between districts and cities.

With this, improving future research requires the distribution of the latest teachers, especially in the 3T area. In addition, it can be developed through Geospatial data to see schools in each province, but the distribution of teachers is still lacking.

REFERENCES

[1] M. Minsih, R. Rusnilawati, and I. Mujahid, "Kepemimpinan Kepala Sekolah Dalam Membangun Sekolah Berkualitas Di Sekolah Dasar," *Profesi*

Pendidikan Dasar, vol. 1, no. 1, pp. 29–40, Jul. 2019, doi: 10.23917/ppd.v1i1.8467.

[2] Presiden Republik Indonesia, "Undang-Undang Republik Indonesia Nomor 20 Tahun 2003 Tentang Sistem Pendidikan Nasional."

[3] P. Data dan Statistik Pendidikan dan Kebudayaan, "Sistem Verifikasi dan Validasi Proses Pembelajaran," *https://vervalsp.data.kemdikbud.go.id/vervalpp/formula.php#:~:text=%22Guru%20tetap%20pemegang%20sertifikat%20pendidik,yang%20sederajat%2015%3A1%22*.

[4] P. Pembangunan, D. Tertinggal, P. Menetapkan, P. Presiden, and P. Daerah, "Dengan Rahmat Tuhan Yang Maha Esa Presiden Republik Indonesia, bahwa untuk melaksanakan ketentuan Pasal 6 ayat (3) Peraturan Pemerintah Nomor 78 Tahun 2OL4 tentang."

[5] Tim Penyusun Direktorat Sekolah Dasar, *Direktorat Sekolah Dasar Pendidikan Bagi Anak di Daerah 3T SERI 4*, vol. Cetakan 1, 2021. 2021.

[6] K. Pendidikan dan Kebudayaan and P. Data dan Statistik Pendidikan dan Kebudayaan, "Analisis Sebaran Guru Dikdasmen Di Wilayah 3 T," 2016.

[7] M. Nishom and D. S. Wibowo, "Implementasi Metode K-Means berbasis Chi-Square pada Sistem Pendukung Keputusan untuk Identifikasi Disparitas Kebutuhan Guru," *Jurnal Sistem Informasi Bisnis*, vol. 8, no. 2, p. 187, Nov. 2018, doi: 10.21456/vol8iss2pp187-194.

[8] H. Prasetyo, B. Pusat, S. Provinsi, and J. Tengah, "Pengelompokan Wilayah Menurut Potensi Fasilitas Kesehatan Dan Kejadian Covid-19 Menggunakan Algoritma Fuzzy K-Prototypes," 2021.

[9] P. P. Putra, I. Bagus, K. Widiartha, and A. Zubaidi, "Rancang Bangun Sistem Informasi Geografis Capaian Pendidikan Formal Sebagai Alat Pendukung Kebijakan Dinas Pendidikan dan Kebudayaan Provinsi NTB."

[10] P. D. dan P. M. D. Jenderal Pendidikan Anak Usia Dini, "Data Pokok Pendidikan," *https://dapo.kemdikbud.go.id/sp*, May 2022.

[11] B. G. Sudarsono, M. I. Leo, A. Santoso, and F. Hendrawan, "Analisis Data Mining Data Netflix Menggunakan Aplikasi Rapid Miner," *JBASE - Journal of Business and Audit Information Systems*, vol. 4, no. 1, Apr. 2021, doi: 10.30813/jbase.v4i1.2729.

[12] P.-N. Tan, M. Steinbach, and V. Kumar, "Introduction to Data Mining Instructor's Solution Manual."

[13] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," 1998.

[14] A. S. Sulthoni, R. Andreswari, and F. Hamami, "segmentasi pelanggan pt. Telekomunikasi seluler indonesia menggunakan clustering algoritma k-prototypes dan metode elbow sebagai perumusan strategi marketing customer segmentation pt. Telekomunikasi seluler indonesia uses clustering k-prototypes algorithm and elbow method for formulating marketing strategy."

p-ISSN: **2656-5935** http://journal-isi.org/index.php/isi e-ISSN: **2656-4882**

[15] S. Sulastri, L. Usman, and U. D. Syafitri, "K-prototypes Algorithm for Clustering Schools Based on The Student Admission Data in IPB University," *Indonesian Journal of Statistics and Its Applications*, vol. 5, no. 2, pp. 228–242, Jun. 2021, doi: 10.29244/ijsa.v5i2p228-242.

[16] N. Putu, E. Merliana, and A. J. Santoso, *Analisa Penentuan Jumlah Cluster Terbaik Pada Metode K-Means Clustering*.

[17] Y. Aprilia, P. Kartikasari, Y. A. Pranoto, and D. Rudhistiar, "Penerapan Metode K-Modes Untuk Proses Penentuan Penerima Bantuan Langsung Tunai (BLT)," 2021.