School Zoning System for Student Admission using Constrained K-Means Algorithms

Andi Alviadi Nur Risal
Postgraduate Student of Electrical Engineering Department
Hasanuddin University
Makassar, Indonesia
risalaan19d@student.unhas.ac.id

Zahir Zainuddin
Department of Informatics
Hasanuddin University
Makassar, Indonesia
zahir@unhas.ac.id

Muhammad Niswar
Department of Informatics
Hasanuddin University
Makassar, Indonesia
niswar@unhas.ac.id

***Abstract*—The issuance of the Regulation Minister of Education and Culture number 51 of 2018 regulates new student admission by implementing a zoning system to achieve equal distribution of education quality in every school, especially at the high school level in Makassar city. This study aims to cluster the school zoning area based on the closest distance between the student's domicile and the school location. The dataset used is 22 school locations and 2248 student location data. In this paper, the method used is constrained k-means to cluster the prospective new students to each school. The constrained k-means method works based on the value of K as the closest cluster center to the value of N (cluster members) with a linear programming algorithm (LPA) approach so that each cluster has a balanced N member. The results of this study can overcome the unbalanced data distribution problem with an average cluster member value of 103 and the absence of empty clusters in each school/centroid. Thus, the system can be implemented in the new student admissions process as a reference in determining the optimal and accurate school zoning area based on the cluster center.**

Keywords—Student Admission, School Zoning System, Clustering, Constrained K-Means

I. Introduction

Registration for new students is an activity program at school educational institutions every year to attract prospective new students. Regarding the acceptance of new students, the government, through the Ministry of Education and Culture, has issued a policy rule, namely the Regulation of the Minister of Education and Culture number 51 of 2018, concerning the admission of new students [1]. One of the government's strategies is allocating equal distribution of education through the implementation of school zoning policies. By paying attention to the multicultural situation of the Indonesian people, it is very appropriate to have a zoning system policy for admitting new students. The zoning system is a rule that requires prospective new students to determine the school with the closest radius to their domicile [1].

The government's zoning system is a policy through the Ministry of Education and Culture to provide equal access to education services and equal distribution of national education quality. Muhadjir Effendy stated that the zoning system is also one of the integrated strategies of the government's hope of accelerating equity in education. The purpose of zoning is to eliminate gaps in the quality of education, especially in the school system, to make the educational environment closer to the community, to increase students' creativity in independent learning, and eliminate school discrimination. The problem that has occurred so far is the disparity in society between schools that are assumed to be schools that are considered more advanced or called superior (favorite) and schools that are considered to be in the category of schools that are not liked (non-favorite) [1].

Government regulations regarding the equal distribution of the quality of education levels in Indonesia as contained in Regulation Minister of Education and Culture number 51 of 2018 in its implementation of disturbances as stated by the Indonesian Teachers' Federation, namely the problem of school capacity with an unbalanced number of students. As a result, many students were not accommodated on the registration by the school. Meanwhile, the distance from the prospective student's house to the destination school is not far from each other. Another problem with this zoning system is the sudden modification of the domicile. A new prospective student puts his name on the identity card of his relative's family so that he can be accommodated by registering at the school in the transfer area. It can be rigged from the government policy in Regulation Minister of Education and Culture number 51 of 2018 regarding the student admission system with this zoning route [2].

Previous research on the school zoning system, such as in the study of Syaripudin et al., prepared an engineering system for determining the closest school using the haversine formula to calculate the distance between the school location and students' homes. This study uses location-based services to indicate the initial position using Global Positioning System (GPS). The results of this study are excellent compared to the results of the default google distance prediction from google. However, in predicting the location of the nearest school based on the zoning system, the accuracy level is not always optimal because this engineering system refers to the accuracy level of the GPS on the device used [3].

Furthermore, related research regarding zoning-based student admissions uses the generate and test method to recommend the best and most accurate shortest-distance recommendations. This research does not use other floor plan-based applications. The study's results were tested by calculating the haversine formula and predicting distance measurements using google maps [4].

Another research to identify the zoning system for student admissions is to use the k-means algorithm. This study calculates the distance between the student's domicile and the school location so that a non-circular zoning-based student admissions area is established. The results obtained in this study stated that the k-means algorithm's performance on the distribution of student data was relatively good. However, the results of this study are not optimal because each school has an unbalanced cluster member, which is the distribution of students in each school or cluster center [5].

The k-nearest neighbor and k-means algorithms use supervised and unsupervised learning methods. However, they are not performing optimally in carrying out their duties on large datasets. In addition, they assume the distribution of data classes is balanced in reality. However, many large data sets are significantly unbalanced. Most data are in one class, while the minority data are in another [6].

Based on previous research regarding the school zoning system for student admissions, most of them only focus on the dimensions of the distance between the student's residence and the distance from the school without paying attention to the capacity volume between the school. So that it causes problems in student admission through the zoning system. However, this paper offers an integrated method for determining the student admission zoning system against unbalanced data. The proposed approach uses a constrained k-means Algorithm. Data distribution must be optimal and not cause overlaps in certain data conditions, such as that schools are close to each other. So that schools that are a short distance apart are not grouped as the same cluster, likewise with the cluster members.

This paper's systematic is as follows: Part I explains the introduction, related and relevant previous research, and research objectives. Part II contains a discussion of the proposed method and the dataset used. Part III presents the experimental results. Part IV summarizes the conclusions.

II. PROPOSED METHOD

This research is located in Makassar City, the capital city of South Sulawesi Province, Indonesia. Makassar City is one of the largest municipalities in the central region of Indonesia. Makassar City has an area of 177.3 km^2. Geographically, Makassar City is located at coordinates between 119°, 18' 27.97" to 119° 32' 31.03" east longitude and 5°30'18" to 5°14' 49 "south latitude [7].

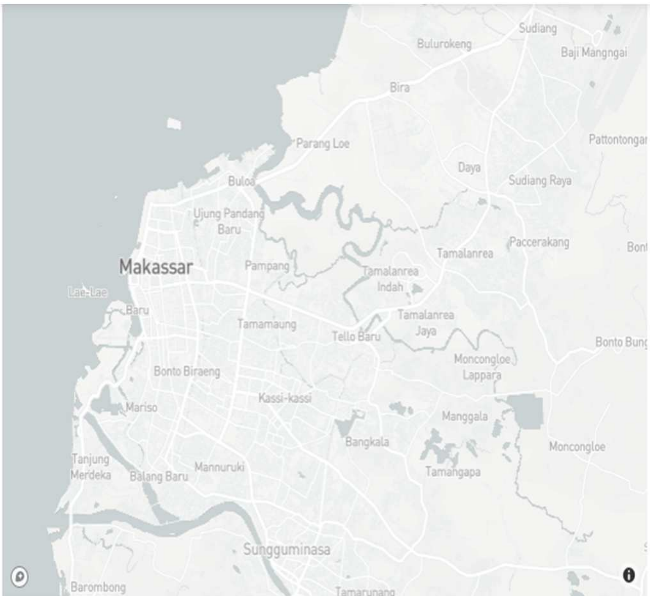

Fig. 1. Research sites

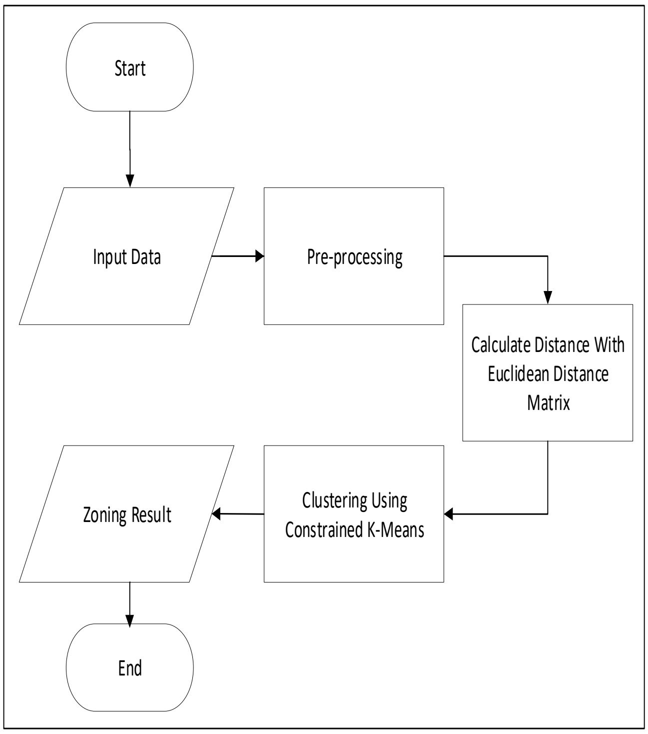

Fig. 2. System design flow

This study consists of several proposals to cluster the system for determining the zoning area of student admissions based on distance sorting with the constrained k-means method. The stages are shown in Fig. 2.

A. Input data

In this study, the data used were 2248 student domiciles, including their location coordinates and 22 school locations in .csv format. The attributes used are the latitude and longitude of each student's domicile and the location of the school site. All of the data is based on Makassar City.

B. Pre-processing

This stage aims to clean noise and transform data to get efficient and more accurate results [5], [8]. Generally, the pre-processing process is carried out before processing the input data. The data used has several irrelevant attributes, such as school names and student addresses, so the attribute selection is carried out for further processing. In addition, the data selection process is also carried out in order to avoid duplicate data.

C. Calculate distance using euclidean distance

This euclidean distance formula is centered on studies that relate angles and distances and is very closely related to the Pythagorean theorem used in 1, 2, and 3 dimensions [9]. The Euclidean distance matrix is a calculation of the distance measured from one coordinate point to another coordinate point. It can generally search for multiple conveyors, transport, and distribution models from Euclidean distances [10].

The Euclidean distance calculation is carried out at this stage to find the distance between 2 coordinate points. In this study, the distance calculation is carried out using latitude and longitude [11]. The following is the euclidean distance formula defined:

$$d = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$$

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \quad (1)$$

Information :

d = distance between coordinates.

x_1 = value of the first x coordinate.

x_2 = second x coordinate value.

y_1 = value of first y coordinate.

y_2 = Second y coordinate value.

D. Clustering using constrained k-means

Clustering in data mining is a data segmentation process that aims to group data so that similar data types become part of certain groups that are similar (or related) to each other and are referred to as clusters [12] [13].

K-means is the most popular partitioning algorithm for data clustering because it is assumed to be efficient, but in some cases, the problem is found in the cluster size constraint, which requires each cluster to be in the same range. In such cases, k-means is often combined with one or more empty clusters. Thus k-means summarizes a small amount of data [14]. In addition, the k-means clustering algorithm is prone to noise in the data clustering results, resulting in unbalanced data clusters that are not even optimal. Meanwhile, the computational nature of euclidean metrics defines that the k-means algorithm in the clustering computing process can only be implemented on a data set in a circular pattern with the exact dimensions and density scale of the dataset. In this case, the k-means algorithm's limitations are shown in the Fig. 3 [15]. At the clustering stage, the system for determining the zoning area for new student admissions is the core of this research.

Therefore, this study proposes a constrained k-means formula to overcome the problems that occur in the k-means algorithm.

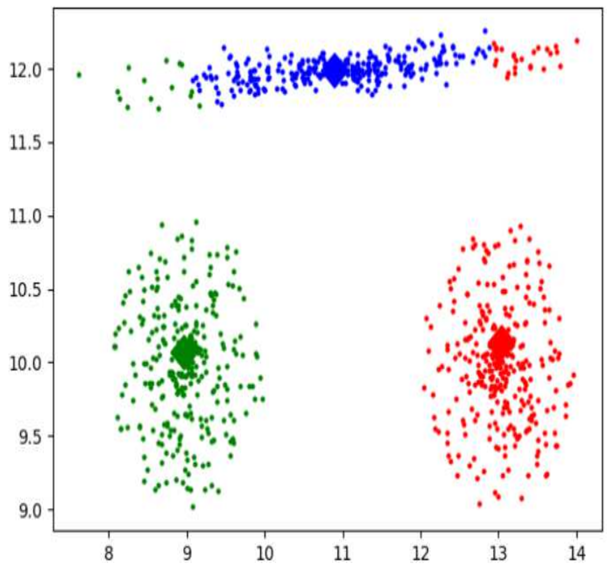

Fig. 3. Illustration of the limitations of the k-means algorithm

Constrained k-means is a modified algorithm from the traditional k-means formula. In carrying out its function, it approaches the linear programming algorithm (LPA) by requiring each cluster to have a subject [16] [17].

Briefly, the stages of constrained k-means are described as follows:

- Extract the input data in the form of the coordinates of the student's domicile and the coordinate of the school site. Use this coordinate attribute to the cluster.
- Choose the coordinates of the school site location as the center point of the cluster.
- Using a linear programming procedure to find optimal clustering with constraints that require each cluster to have a subject and not be empty by adding an indicator variable based on equation (2).

$$\min_{C,T} \sum_{i=1}^{m} \sum_{h=1}^{k} T_{i,h} \left(\|x_i - c_h\|^2 \right)$$

Subject to:

$$\sum_{h=1}^{m} T_{i,h} \geq 2 \,; h = 1, \ldots k$$

$$\sum_{h=1}^{k} T_{i,h} = 1; i = 1, \ldots m$$

$$T_{i,h} \geq 0, i = 1, \ldots m; h = 1, \ldots k \quad (2)$$

- Allocate each object's data to the nearest centroid (cluster center) based on comparing the distance between the object and the centroid (cluster center) of each existing cluster from the results of the linear programming procedure.

$$C_{h,t+1} = \frac{\sum_{i=1}^{m} T_{i,h}^{t} X_i}{\sum_{i=1}^{m} T_{i,h}^{t}} \quad (3)$$

- If the cluster member changes, repeat steps 2 and 3
- Determine the results at each cluster center in the form of data and coordinates.

III. RESULT AND DISCUSSION

This paper proposes a system for determining high school admissions based on a balanced zoning system. The system was developed using the python programming language with the constrained k-means algorithm, which is the development of the traditional k-means formula. Firstly, determining the central point of the data or the number of k values. In this case, there is 22 school location, as shown in Fig 4. The school data attributes used in the form of latitude and longitude coordinates are shown in Table I. Then, the data clustering process is performed in the next stage using the constrained k-means algorithm. According to the formed cluster center, the data were mapped in 22 clusters. This formula uses the euclidean distance function to calculate the distance between 2 coordinates. The proposed approach uses the constrained k-means algorithm to optimize the acceptance parameters of the number of students in each school. This method works by optimizing clusters that have empty members for the process of determining the closest cluster center. The number of student data in this study was 2248. The attributes used are latitude and longitude coordinates, as shown in Table II.

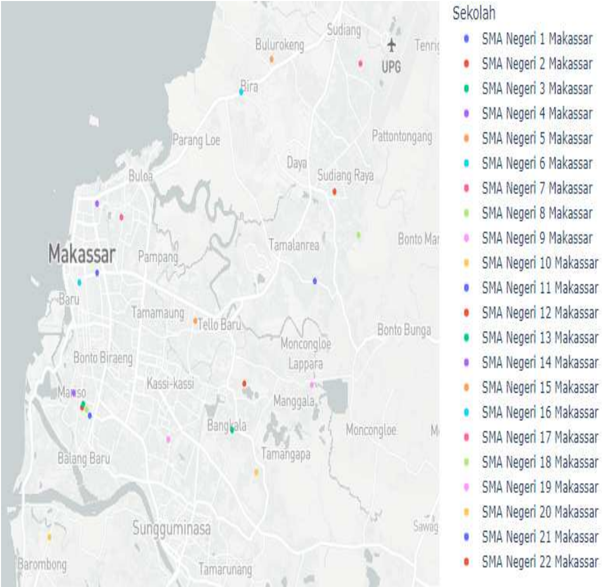

Fig. 4. School location map

TABLE I. SCHOOL DATA DETAILS

No	School	Latitude	Longitude
1	SMA Negeri 1 Makassar	-5,1349823	119,4191
2	SMA Negeri 2 Makassar	-5,1695047	119,4126
3	SMA Negeri 3 Makassar	-5,1686272	119,4131
4	SMA Negeri 4 Makassar	-5,11728	119,4191
5	SMA Negeri 5 Makassar	-5,1473156	119,4618
6	SMA Negeri 6 Makassar	-5,0886288	119,4818
7	SMA Negeri 7 Makassar	-5,0813625	119,5337
8	SMA Negeri 8 Makassar	-5,1700418	119,4145
9	SMA Negeri 9 Makassar	-5,1777373	119,4501
10	SMA Negeri 10 Makassar	-5,1860887	119,4884
11	SMA Negeri 11 Makassar	-5,1715849	119,4159
12	SMA Negeri 12 Makassar	-5,1633799	119,4832
13	SMA Negeri 13 Makassar	-5,1752924	119,4779
14	SMA Negeri 14 Makassar	-5,1658281	119,4089
15	SMA Negeri 15 Makassar	-5,0802534	119,495
16	SMA Negeri 16 Makassar	-5,1374913	119,4114
17	SMA Negeri 17 Makassar	-5,120742	119,4297
18	SMA Negeri 18 Makassar	-5,1253379	119,5328
19	SMA Negeri 19 Makassar	-5,1637625	119,5125
20	SMA Negeri 20 Makassar	-5,2027611	119,3984
21	SMA Negeri 21 Makassar	-5,1371232	119,5139
22	SMA Negeri 22 Makassar	-5,1141946	119,5224

TABLE II. STUDENT DATA DETAILS

No	Student Address	Latitude	Longitude
1	Jl. Baji Rupa, Balang Baru, Kec. Tamalate, Kota Makassar, 90224	-5,18047	119,4095
2	Jl. Kedamaian Selatan II 40-29, Tamalanrea, Kec. Tamalanrea, Kota Makassar, 90245	-5,13938	119,5095
3	Jl. Baji Gau 34-30, Bongaya, Kec. Tamalate, Kota Makassar, 90134	-5,17096	119,4136
4	Jl. Barukang III, Pattingaloang, Kec. Ujung Tanah, Kota Makassar, 90162	-5,11464	119,4212
5	Jl. Moha Lrg. 3 25-1, Antang, Kec. Manggala, Kota Makassar, 90234	-5,16423	119,4801
6	Jl. Andi Mangerangi 38, Bongaya, Kec. Tamalate, Kota Makassar, 90131	-5,16734	119,4151
7	Jl. Tamangapa Raya 5 No.6, Rt.2, Tamangapa, Kec. Manggala, Kota Makassar, 90235	-5,18534	119,487
8	Jl. Karunrung Raya I No.13, Mappala, Kec. Rappocini, Kota Makassar, 90222	-5,17629	119,451
9	Jl. Rahmatullah 5, Tamangapa, Kec. Manggala, Kota Makassar, 90235	-5,18321	119,4927
10	Griya Atirah, Jl. Perintis Kemerdekaan, Pai, Biringkanaya, Kota Makassar, 90242	-5,09857	119,5137

Based on the results of the data mining process with the reliability of the constrained k-means algorithm function based on the number of clusters = 22 as the center point of the school centroid to the distribution of student domicile distribution data as much as 2248. The clustering results are visualized in a map, as shown in Fig 5.

The clustering results are also detailed in Table III. 21 schools received 103 students or cluster members, while one school only received 85 cluster members. It can be seen that the constrained k-means algorithm assigns students to each of the closest schools evenly, even though there is one school that only has different cluster members from the others. This also means that there are no schools that do not have cluster members. In addition, the clustering results also show that there is no overlap in cluster members or cluster centers.

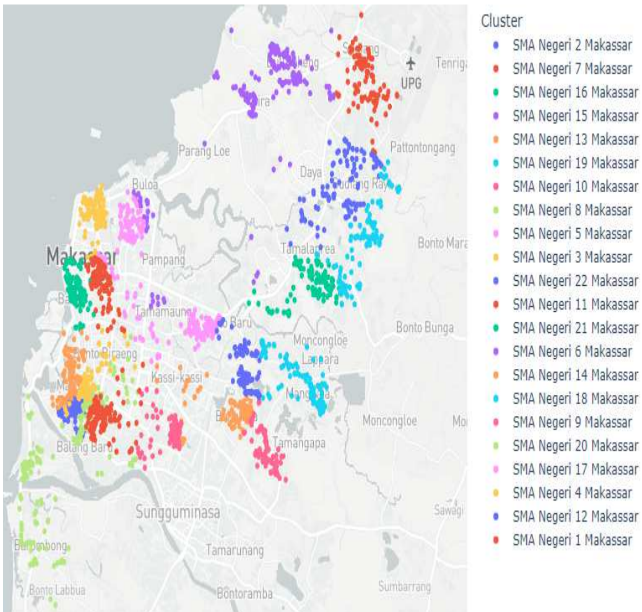

Fig. 5. Clustering results map

TABLE III. CLUSTERING RESULTS DETAIL

No	School	Latitude	Longitude	Cluster	Cluster Member
1	SMA Negeri 1 Makassar	-5,13498	119,4191	10	103
2	SMA Negeri 10 Makassar	-5,18609	119,4884	0	103
3	SMA Negeri 11 Makassar	-5,17158	119,4159	20	103
4	SMA Negeri 12 Makassar	-5,16338	119,4832	5	103
5	SMA Negeri 13 Makassar	-5,17529	119,4779	16	103
6	SMA Negeri 14 Makassar	-5,16583	119,4089	1	85
7	SMA Negeri 15 Makassar	-5,08025	119,495	13	103
8	SMA Negeri 16 Makassar	-5,13749	119,4114	17	103
9	SMA Negeri 17 Makassar	-5,12074	119,4297	3	103
10	SMA Negeri 18 Makassar	-5,12534	119,5328	14	103
11	SMA Negeri 19 Makassar	-5,16376	119,5125	11	103
12	SMA Negeri 2 Makassar	-5,1695	119,4126	19	103
13	SMA Negeri 20 Makassar	-5,20276	119,3984	6	103
14	SMA Negeri 21 Makassar	-5,13712	119,5139	12	103
15	SMA Negeri 22 Makassar	-5,11419	119,5224	2	103
16	SMA Negeri 3 Makassar	-5,16863	119,4131	21	103
17	SMA Negeri 4 Makassar	-5,11728	119,4191	18	103
18	SMA Negeri 5 Makassar	-5,14732	119,4618	8	103
19	SMA Negeri 6 Makassar	-5,08863	119,4818	4	103
20	SMA Negeri 7 Makassar	-5,08136	119,5337	7	103
21	SMA Negeri 8 Makassar	-5,17004	119,4145	15	103
22	SMA Negeri 9 Makassar	-5,17774	119,4501	9	103

IV. CONCLUSION

This paper proposes determining the admission system for new high school students in Makassar City using a zoning approach with the constrained k-means algorithm to solve unbalanced data problems. The system developed can map the number of new students in each school based on the distance from where the student lives to the school location based on the zoning area. The clustering results show that the distribution of students is more effective with a balanced number of cluster members. With 22 schools and 2248 new prospective students, it produces cluster members in each school, with 103 students in 21 schools and 85 students in one school. So that there is no overlap in cluster members, and there is no cluster center that does not have cluster members. This system can be used as a reference by the government in making an efficient and accurate school zoning system.

REFERENCES

[1] T. Widayati and A. Sudrajat, "Conflict and Overlapping Authorities in the Newly Implemented School Zoning Policy in Indonesia the Case in the Urban–Rural Regency of Magelang," *2nd Int. Conf. Soc. Sci. Character Educ. (ICoSSCE 2019)*, vol. 398, pp. 277–282, doi: 10.2991/assehr.k.200130.056.

[2] R. Apinino, "Empat Kelemahan Sistem Zonasi dalam PPDB 2018," *Tirto.id*, Jul. 10, 2018.

[3] U. Syaripudin, N. Fauzi, W. Uriawan, W. Z, and A. Rahman, "Haversine Formula Implementation to Determine Bandung City School Zoning Using Android Based Location Based Service," *1st Int. Conf. Islam. Sci. Technol. ICONISTECH 2019*, doi: http://dx.doi.org/10.4108/eai.11-7-2019.2303558.

[4] N. Ratnasar *et al.*, "Implementation of Generate and Test Algorithm for Junior High School Zoning System in Malang," *4th Int. Conf. Vocat. Educ. Training, ICOVET 2020*, pp. 167–170, doi: 10.1109/ICOVET50258.2020.9230207.

[5] M. D. Febriana, Z. Zainuddin, and I. Nurtanio, "School Zoning System Using K-Means Algorithm for High School Students in Makassar City," *2nd Int. Semin. Res. Inf. Technol. Intell. Syst. ISRITI 2019*, pp. 368–372, doi: 10.1109/ISRITI48646.2019.9034601.

[6] M. Ghanavati, R. K. Wong, F. Chen, Y. Wang, and C. S. Perng, "An Effective Integrated Method for Learning Big Imbalanced Data," *Proc. - IEEE Int. Congr. Big Data, BigData Congr. 2014*, pp. 691–698, 2014, doi: 10.1109/BigData.Congress.2014.102.

[7] F. Ismiati, A. Damayanti, and Muhammad Dimyati, "Determining location of tsunami disaster temporary evacuation shelter (TES) utilizes network analysis in City of Makassar , South Sulawesi Province," *IOP Conf. Ser. Earth Environ. Sci. Fifth Int. Conf. Indones. Soc. Remote Sens.*, 2020, doi: 10.1088/1755-1315/500/1/012062.

[8] P. H. P. Rosa and R. Gunawan, "The clustering of high schools based on national and school examinations: A case study at Daerah Istimewa Yogyakarta Province," *Int. Conf. Data Softw. Eng.*, pp. 231–236, 2015, doi: 10.1109/ICODSE.2015.7437003.

[9] R. Purbaningtyas and A. Arizal, "Nearest Excellent Potential Location Using Distance Algorithm," *J. Phys. Conf. Ser.*, vol. 1413, no. 1, 2019, doi: 10.1088/1742-6596/1413/1/012032.

[10] A. C. Arrumdany, P. P. Sari, P. Rahmadani, and A. I. Lubis, "Web-Based Geographic Information System (GIS) in Determining Shortest Path of MSME Medan City Using Bellman-Ford Algorithm," *J. Phys. Conf. Ser. Int. Conf. Comput. Sci. Appl. Math.*, vol. 1255, no. 1, 2019, doi: 10.1088/1742-6596/1255/1/012075.

[11] D. Ramdania, R. Andrian, M. Irfan, and R. Abidin, "On Designing Application of Finding Nearby Islamic Boarding Schools in West Java using Haversine Formula and Euclidean Distance Algorithms," *1st Int. Conf. Islam. Sci. Technol. ICONISTECH 2019*, doi: 10.4108/eai.11-7-2019.2297517.

[12] M. Faisal, E. M. Zamzami, and Sutarman, "Comparative Analysis of Inter-Centroid K-Means Performance using Euclidean Distance , Canberra Distance and Manhattan Distance," *J. Phys. Conf. Ser.*, 2019, doi: 10.1088/1742-6596/1566/1/012112.

[13] D. Saini and M. Singh, "Achieving Balance in Clusters-A Survey," *Int. Res. J. Eng. Technol.*, vol. 02, no. 09, pp. 2611–2614, 2015, [Online]. Available: www.irjet.net.

[14] K. Lei, S. Wang, A. Weiwei Song, and Qilin Li, "Size-Constrained Clustering Using an Initial Points Selection Method," *Int. Conf. Knowl. Sci. Eng. Manag.*, vol. 8041, pp. 373–383, 2013, doi: 10.1007/978-3-642-39787-5.

[15] T. Wang and J. Gao, "An Improved K-Means Algorithm Based on Kurtosis Test," *J. Phys. Conf. Ser.*, vol. 1267, no. 1, 2019, doi: 10.1088/1742-6596/1267/1/012027.

[16] J. Zhao, "Optimal Clustering: Genetic Constrained K-Means and Linear Programming Algorithms," Virginia Commonwealth University, 2006.

[17] P. Bradley, K. Bennett, and A. Demiriz, "Constrained K-Means Clustering," *Microsoft Res.*, p. 9, 2000, [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.33.3257.