

Balanced clustering for student admission school zoning by parameter tuning of constrained k-means

Zahir Zainuddin¹, Andi Alviadi Nur Risal²

¹Department of Informatics, Faculty of Engineering, Hasanuddin University, Makassar, Indonesia

²Department of Electrical Engineering, Faculty of Engineering, Hasanuddin University, Makassar, Indonesia

Article Info

Article history:

Received Aug 18, 2023

Revised Oct 24, 2023

Accepted Dec 14, 2023

Keywords:

Balanced
Clustering
Constrained k-means
K-means
School zoning

ABSTRACT

The Indonesian government issued a regulation through the Ministry of Education and Culture, number 51 of 2018, which contains zoning rules to improve the quality of education in school educational institutions. This research aims to compare the performance of the k-means algorithm with the constrained k-means algorithm to model the zoning of each school area based on the shortest distance parameter between the school location and the domicile of prospective students. The study used data from 2,248 prospective students and 22 public school locations. The results of testing the k-means algorithm in grouping showed the formation of non-circular patterns in the cluster membership with different numbers of centroid cluster members. In contrast, testing the constrained k-means algorithm showed balanced outcomes in cluster membership with a membership value of 103 for each school as the cluster center. The research findings state that the developed constrained k-means algorithm solves the problem of unbalanced data clustering and overlapping issues in the process of new student admissions. In other words, the constrained k-means algorithm can be a reference for the government in making decisions on new student admissions.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Zahir Zainuddin

Department of Informatics, Faculty of Engineering, Hasanuddin University

Makassar, Indonesia

Email: zahir@unhas.ac.id

1. INTRODUCTION

Education is central to improving human resource development in a country. The quality of its education system determines the sustainability of a nation's social life. Therefore, the influence of the role of education is significant in Indonesia [1].

Every new school year, every school education institution in Indonesia implements a new student admission program agenda. The Government of Indonesia regulates the process of accepting prospective new students in every educational institution at all levels, starting from Kindergarten, Elementary, and Middle School to High School and vocational High School, which is guided by the Regulation of the Minister of Education and Culture of 2018 no 51, which is the work program of the Indonesian government which sustainable to realize equity in the education sector by implementing a policy of accepting new prospective students based on system zoning. Considering that the Indonesian people live in a multicultural nation, the government's policies are considered very appropriate to apply deep zoning system implementation acceptance of prospective students in public and private school education institutions. The basic rule of this zoning system is to oblige prospective new students to identify schools within the radius of the closest area to where they [2], [3]. Government directives through the Ministry of Education and Culture Regulation number 51 of 2018 regulate several pathways in implementing the acceptance of prospective new students, namely the zoning system path, achievement pathway,

and parent or guardian isolation pathway. Each admission pathway has a different quota allocation, such as the zoning system, which has 90% of the total quota of prospective student admissions for the community whose domicile is closest to the school [4].

The purpose of the school zoning system is to guarantee the quality of equal distribution of access to education services as a government way for the community regarding the importance of education, in addition to bringing families closer to the school environment for students as a place for creativity in independent learning, and most importantly eliminating the community's paradigm of school discrimination. In improving aspects in the field of education, the government still assesses the existence of socio-economic disparities in society. In this case, there is a difference between middle and upper economic-class students and lower middle economic-class students, causing caste differences between schools, namely between excellent schools and regular schools in the community. There is a view of the social paradigm between schools in accessing educational services. This means that outstanding schools are filled by middle- to upper-class students, while the less fortunate are in regular schools [5], [6].

Education regulations in Indonesia, as stated in Permendikbud No. 51 of 2018, aim to facilitate access to education services, eliminate public views regarding exclusivity in school education institutions, and improve education services and quality. However, in the implementation of Permendikbud No. 51 of 2018, the Indonesian Teachers' Federation found problems regarding school capacity with an uneven number of prospective students. This resulted in some prospective students whose registration was not further processed by the nearest school, even though the student's domicile was within one radius of the nearest school.

Another thing that needs to be improved in this school zoning system policy is that prospective students manipulate and modify domicile data and identity documents. In one case, it was found that prospective students included their identity on the ownership identity of their close relative on the family card document so that the registration completeness could be processed. Furthermore, the school is in the area of moving relatives, in the sense that government regulations regarding the candidate acceptance system can be rigged, especially in the zoning acceptance pathway.

The application of the school zoning system based on government directives is to prioritize prospective new students on the primary determination based on the closest distance between the school and the domicile of prospective new students, which is one of the determining aspects in new student registration. Problems arising from the government application regarding the school zoning system, among others, cannot accommodate the enrollment of prospective new students, especially those who live in blind spot areas (not within the school zone radius). Another problem concerns the interpretation of the student's domicile distance to the destination school based on the address of the identity of the student's parent or guardian in the process of implementing the acceptance of prospective new students.

In principle, the K-Means algorithm uses the unsupervised learning method, but when clustering large datasets, the K-Means algorithm does not work well. Therefore, it is identified that the data classes are distributed and balanced. In contrast, those that occur in large data sets should be accounted for in a balanced manner in the class distribution of the data. That is, most of the information is in one category, while the minority information is in another category [7].

The current approach to determining school zoning systems, which relies solely on the shortest distance between a prospective new student's residence and the school, can result in challenges for schools located within the same radius, particularly in accommodating new students. This study addresses this issue by introducing each school's "capacity ratio" parameter and comparing the shortest distances between schools. Considering this parameter, the study aims to prevent problems such as overlapping student admissions within the zoning radius. In light of these challenges, the research sets out to evaluate and compare the performance of the K-Means algorithm with the K-Means Constrained clustering algorithm to find an optimized solution for school zoning and student allocation. This research aims to improve the effectiveness and fairness of the school zoning system, providing better solutions for both schools and students.

- Analyzing the K-Means clustering algorithm and the Constrained K-algorithm in the computational process of clustering data.
- The results of a comparative analysis of clustering data for the two algorithms can be used as government guidelines, especially for schools and institutions in the education sector, to determine optimal school zoning areas.

2. RESEARCH BACKGROUND

2.1. Data mining

Data mining is a systematic and iterative process focused on recognizing patterns, extracting valuable insights, and enhancing our understanding of data. This method leverages essential and relevant data to establish connections and complement the existing information. It is employed in various fields, including

statistical segmentation within Mathematics, artificial intelligence, and machine learning [8]. Furthermore, data mining is an integral component of knowledge discovery in databases (KDD), which plays a pivotal role in transforming raw data into valuable and practical information [9]. The data mining process encompasses multiple stages, beginning with pre-processing to prepare the data and culminating in post-processing for interpreting and applying the mining results, making it a vital tool in data-driven decision-making and problem-solving.

2.2. Clustering

Clustering is an analytical activity focused on understanding the transformation process and segmenting data comprehensively. This process involves partitioning data to form distinct groups known as clusters, where similar and interconnected data points are grouped together. The primary goal of clustering is to facilitate the organization and analysis of data, enabling deeper insights and more effective decision-making [10]–[12].

2.3. The k-means algorithm

The K-Means clustering algorithm is the predominant choice for data analysis, particularly in data clustering. This algorithm effectively groups data into distinct subclass partitions through an unsupervised learning modeling process [13]. The main goal is to minimize the overall variance within each cluster, which means improving the squared error function of the calculation [14]. The K-Means clustering algorithm follows a series of structured stages during data analysis outlined below [15], [16]. These stages guide the systematic process of partitioning data, making it a valuable tool in diverse applications such as data segmentation, pattern recognition, and customer profiling.

- Input data that will be analyzed clustering.
- Determine the number of "K" values as clusters that will be formed based on the distance matrix used.
- Determine "K" from all data on the number of clusters in a random way as the center of the cluster (centroid data).
- They were computing each data to be partitioned to the nearest cluster center with predetermined distance matrix parameters.
- Determine the cluster center "K" if there is still a change, then repeat steps 4-5 until there is no change in the position of the cluster center.

2.4. The constrained k-means clustering algorithm

The K-Means clustering algorithm is a method that is often used in a list of other clustering algorithms because it has an advantage in terms of speed in completing the clustering process so that it is efficient. However, the K-Means algorithm has weaknesses when operating. The fault found when the K-Means clustering algorithm performs computations lies in the scale of the cluster dimension boundaries formed. Each cluster formed must be on the same or balanced dimensional scale of the total data in each cluster formed. K-Means clustering is in a bad state in such situations because it encapsulates a small amount of data [17].

The K-Means clustering algorithm encounters another issue related to its computational output, leading to noise within the resulting cluster data. Consequently, this noise undermines the optimality and balance of the clustering process. In contrast, the algorithm's reliance on Euclidean matrix calculations limits its applicability to datasets with circular conditions and uniform density across the dataset scale. This constraint restricts the effectiveness of K-Means clustering when applied to datasets with varying shapes or uneven data distribution [18].

Processing data by positioning two data objects systematically in one cluster is an incorrect systematic operation, so a Constrained clustering approach is applied. Constrained clustering analyzes the boundaries of data points stated to exist or not exist in the same cluster or different clusters from the computational results of the clustering process so that the Constrained clustering approach does not only analyze the distance between each data object. Therefore, limited clustering modeling is called semi-supervised clustering [19].

The future of Constrained K-means is a development of the K-means clustering algorithm. The Constrained K-Means formula, when carrying out its computational tasks, has a reference stating that each cluster has a data object. Therefore, this Constrained K-Means formula utilizes the function of the linear programming algorithm (LPA) [20], [21]. The description of the stages of the Constrained K-Means modeling formula is as follows:

- Determine the initial cluster center randomly from k values or Constrained K-Means solutions.
- In the search for optimal clusters against constraints, provided that each cluster center has at least two subjects and uses the Linear Programming Algorithm (LPA) procedure.
- Cluster center updated from Linear Programming Algorithm (LPA) computational results.
- Repeat steps 2 and 3 if there are still changes in the cluster membership.

- Repeat steps 1-5 for many datasets on the initial object. The cluster with the objective minimum score is the final cluster solution.

2.5. Related work

Research that has relevance in the school zoning system, among others, proposes a generate and test algorithm to provide accommodation for the closest, best, and most accurate suitability with the design of a system performance model without involving the help of floor plan software. The test results from this study use the haversine formula calculations and utilize Google Maps to predict distance measurements [22]. Subsequent research as a reference is to design a new student admissions system that is guided by the parameter of the distance of the domicile of prospective new students to the nearest school by modeling using the Haversine distance algorithm to analyze the distance between school locations and the location of domicile of prospective new students. This research involves global positioning system (GPS) services as a determinant of the initial location point. System performance results are superior to Google's GPS distance predictions. However, the resulting level of accuracy could be more optimal because in predicting the closest distance to a school, this system is based on the GPS accuracy of the device used by the user [23].

In addition, research related to the school zoning system about new student admissions was carried out using the K-means algorithm approach. This case study research focuses on the high school level in Makassar City in accepting new prospective students by calculating the distance from the school location to the student's residence and forming a zoning-based student enrollment area with a non-circular pattern. The results obtained in this study stated that the K-Means algorithm in dividing student data was relatively good based on the reference parameter of the shortest distance between the school location and the student's domicile. However, based on the clustering results, this study produced unequal data between cluster centers and cluster members, resulting in overlapping of the membership of each cluster center and thus making the registration process not optimal [5].

Furthermore, research related to the zoning system was carried out in 2017. This study identified crime areas in India using the K-Means clustering algorithm. The results of this study can visualize crime-prone areas on a unique map to display crime information so that they can assist law enforcement officials, especially the police, in carrying out crime prevention strategies [24].

3. PROPOSED METHOD

The system proposed in this research consists of two main stages, namely data preparation and modeling, as shown in Figure 1. Testing is implemented using the Python programming language. Data preparation is an essential first step, where raw data is collected, cleaned, and transformed to suit the analysis needs. This process involves various data preprocessing techniques, such as integration, filtering, cleaning, and feature engineering. After the data is prepared, the modeling stage begins.

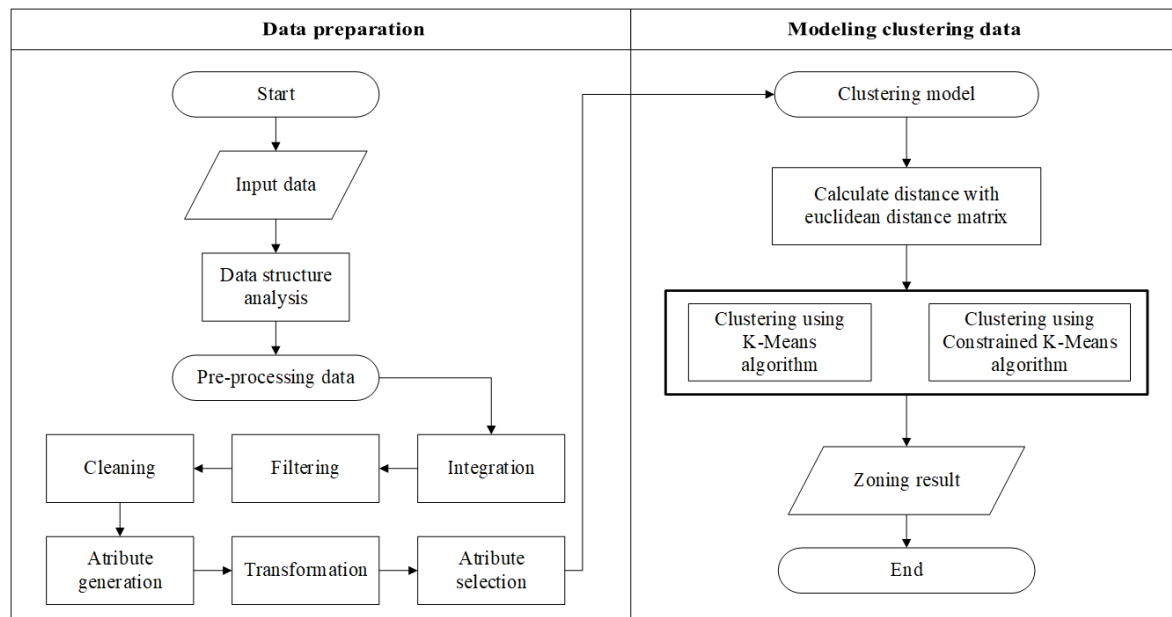


Figure 1. System flowchart

The processed data is used to build and evaluate machine learning or statistical models in the modeling stage. This step includes selecting an appropriate algorithm and training the model on the prepared data. The algorithms used in this research case study are K-Means and Constrained K-Means, and their performance using relevant metrics. Python is a popular choice for implementing these modeling tasks because of its extensive libraries and tools for performing data analysis and machine learning. The programming language provides various resources, making it versatile for data scientists and researchers in this field. As a result, using Python simplifies model development and testing, which is essential for the success of the proposed system.

Combining the essential stages of data preparation and modeling is pivotal for developing robust systems across diverse applications like data analysis, predictive modeling, and decision support. This integration ensures that data is properly processed and utilized for accurate modeling, leading to informed decision-making. Furthermore, leveraging Python as a programming language enhances the efficiency and effectiveness of system testing and deployment. Its versatility and extensive libraries make it an invaluable tool for researchers and practitioners in various fields, enabling them to create sophisticated data-driven solutions with ease.

3.1. Data preparation

3.1.1. Data input

Data acquisition is the initial stage of presentation that must be prepared before processing data on the system. In this study, data collection included input data of 2,248 residences of prospective new students, and their location coordinates shown in Table 1 and 22 school locations shown in Table 2 in the .csv file format, which consists of attributes of latitude and longitude coordinates of residence for each prospective student and geographic location at each school location. All the data is in Makassar, where new prospective students live and schools. The school data used refers to the level of public high school (SMA) education shown in Figure 2.

Table 1. Student distribution coordinate data

No	Id Studet	Student Address	Latitude	Longitude
1	N-2116	Jl. Gatot Subroto Iv No.17, Ujung Pandang Baru, Kec. Tallo, Kota Makassar, 90215	-5.1174224	119.4362051
2	N-2117	Jl. Mutiara Kirana Utama No.72, Bulurokeng, Kec. Biringkanaya, Kota Makassar, 90243	-5.076964	119.4905702
3	N-2118	Jl. Muh. Tahir No.135, Parang Tambung, Kec. Tamalate, Kota Makassar, 90221	-5.1806464	119.4175839
4	N-2119	Jl. Caddika No.26, Bulurokeng, Kec. Biringkanaya, Kota Makassar, 90242	-5.0751925	119.5046118
5	N-2120	Jl. Bulu Salaka No.12, Lariang Bangi, Kec. Makassar, Kota Makassar, 90000	-5.1394387	119.4213233
6	N-2121	Jl. Toa Daeng Iii No.49, Batua, Kec. Manggala, Kota Makassar, 90233	-5.1523296	119.4656025
7	N-2122	Jalan Biola 25 No.L/194, Manggala, Kec. Manggala, Kota Makassar, 90562	-5.1717966	119.5109946
8	N-2123	Jl. Cakalang Iii No.29, Totaka, Kec. Ujung Tanah, Kota Makassar, 90165	-5.1162767	119.4201751
9	N-2124	Jl. Andi Mappaodang Kompleks Perwira Lama No.H.47, Jongaya, Kec. Tamalate, Kota Makassar, 90223	-5.1732544	119.4128737
10	N-2125	Jl. Bdp Lrg. 5, Sudiang Raya, Kec. Biringkanaya, Kota Makassar, 90242	-5.103744	119.5207274

Table 2. School coordinate data

No	School	Latitude	Longitude
1	SMAN 1 Kota Makassar	-5.1349823	119.4191
2	SMAN 2 Kota Makassar	-5.1695047	119.4126
3	SMAN 3 Kota Makassar	-5.1686272	119.4131
4	SMAN 4 Kota Makassar	-5.11728	119.4191
5	SMAN 5 Kota Makassar	-5.1473156	119.4618
6	SMAN 6 Kota Makassar	-5.0886288	119.4818
7	SMAN 7 Kota Makassar	-5.0813625	119.5337
8	SMAN 8 Kota Makassar	-5.1700418	119.4145
9	SMAN 9 Kota Makassar	-5.1777373	119.4501
10	SMAN 10 Kota Makassar	-5.1860887	119.4884
11	SMAN 11 Kota Makassar	-5.1715849	119.4159
12	SMAN 12 Kota Makassar	-5.1633799	119.4832
13	SMAN 13 Kota Makassar	-5.1752924	119.4779
14	SMAN 14 Kota Makassar	-5.1658281	119.4089
15	SMAN 15 Kota Makassar	-5.0802534	119.495
16	SMAN 16 Kota Makassar	-5.1374913	119.4114
17	SMAN 17 Kota Makassar	-5.120742	119.4297
18	SMAN 18 Kota Makassar	-5.1253379	119.5328
19	SMAN 19 Kota Makassar	-5.1637625	119.5125
20	SMAN 20 Kota Makassar	-5.2027611	119.3984
21	SMAN 21 Kota Makassar	-5.1371232	119.5139
22	SMAN 22 Kota Makassar	-5.1141946	119.5224

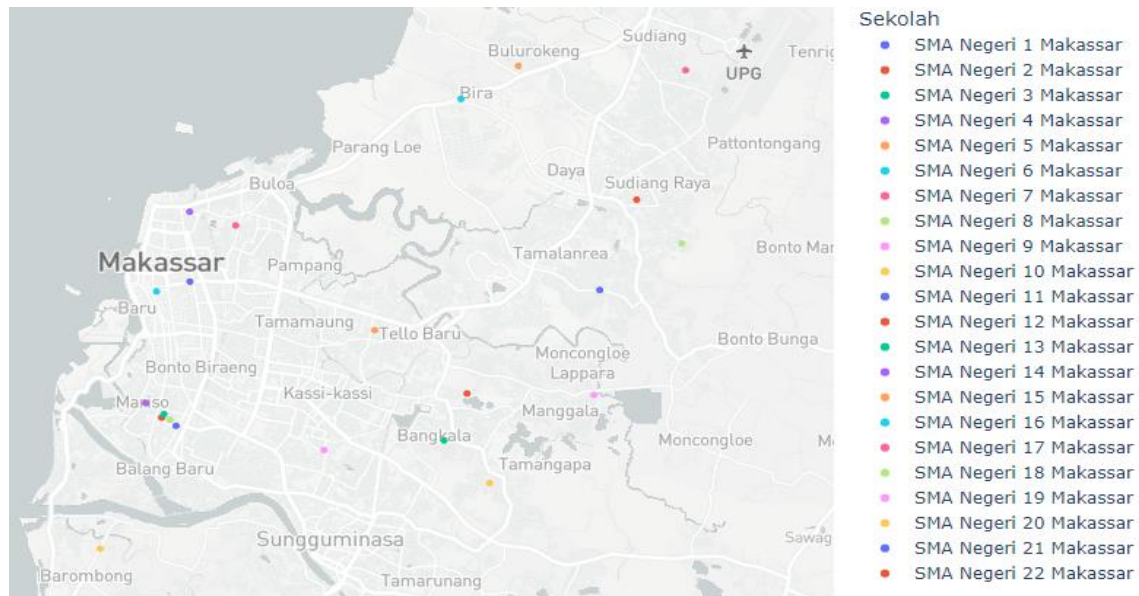


Figure 2. School location coordinates

3.1.2. Pre-processing

Before starting the clustering process stage, it is essential to perform data pre-processing, which is vital in selecting relevant attributes, eliminating data duplication, and ensuring data accuracy [3], [25]. Pre-processing actions help transform the data, reduce noise, and make it suitable for analysis with clustering algorithms and data mining tools. In Figure 1, six pre-processing techniques are used, including data integration, data filtering, data cleaning, attribute creation, data transformation, and attribute selection. Data integration combines data from various sources into a unified data set. Specifically in this research, data filtering includes selecting data related to prospective new students continuing their education at Makassar State secondary schools.

Following data filtering, the data cleaning stage removes irrelevant attribute categories, such as student names, addresses, and duplicate data. The subsequent data transformation phase adapts the data format for compatibility with the clustering algorithm. Finally, the attribute selection process is implemented to prevent data duplication and ensure that only pertinent attributes are considered in the clustering analysis. These preprocessing steps are essential for preparing the data for effective and accurate clustering analysis.

3.2. Modeling clustering data

3.2.1. Perform distance operations with euclidean distance

In this process stage, the study involves determining the distance between two coordinate points using the Euclidean distance calculation approach [26]. The longitude and latitude attributes were used to analyze the computational calculation of the coordinate distance data of school locations with the distribution of student domiciles in this study. Specifically, the attributes of longitude and latitude are employed to perform the computational calculations of the distance between school locations and the distribution of student domiciles. The Euclidean Distance equation, used to compute these distances, plays a fundamental role in assessing the spatial relationships between schools and students' residences, enabling the development of a more accurate school zoning system. This step is crucial for ensuring efficient student allocation to schools based on proximity, thereby enhancing the effectiveness of the school zoning process. The Euclidean Distance equation is as follows,

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

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

Information Where X and Y are longitudes and latitudes

3.2.2. Application of the k-means algorithm

The K-Means clustering algorithm is a clustering modeling method that groups data by dividing and dividing it into several classes so that data that have a relationship with each other are in the same cluster class. However, data with different characteristics are of various types. The K-Means algorithm is as follows:

- Choose student and school input data that states the coordinates of each variable for clustering
- Determine the number of cluster centers formed, namely 22 school coordinate locations.
- Alloc randomly allocates each student's domicile data to the cluster center (school).
- Calculating the average value of each k (V_{ij}) student domicile data for the centroid/cluster center with the following equation:

$$V_{ij} = \frac{\sum_{k=1}^{N_i} x_{kj}}{N_i} \quad (2)$$

Information:

N_i = lots of data from the center of the cluster

- Perform calculations to allocate student domicile data closest to the cluster's center (school). The process of allocating student domicile data is by comparing it with other student domicile data against the centroid of each existing cluster. Defined in the following equation:

$$a_{ik} = \begin{cases} 1, & d = \min \{D(x_k, v_i)\} \\ 0, & \text{other} \end{cases} \quad (3)$$

Information:

a_{ik} : Is membership of K-data to I-cluster

V_i : Is the cluster centroid value i

- If data changes still occur in the clusters that have been formed or the centroid values have changed, then return to step 4. To perform calculations based on the Euclidean distance formula on each student's domicile and centroid (school) data.
- The clustering results for determining school zoning refer to the coordinate points for the distribution of student data to the nearest cluster center point (school).

3.2.3. Implementation of the constrained k-means clustering algorithm

Clustering data processing using the Constrained K-means algorithm is a developed version of the K-means algorithm. Regarding performance during computation, the Constrained K-Means clustering algorithm can provide an additional limit on the number of points in each centroid for data clustering processes with the linear programming algorithm (LPA) function [3]. Limitations of the K-Means clustering algorithm, when computing determines the initial center (centroid) used in defining clusters by dividing the data set into classes (partitions) from the given data into k clusters to be formed, then used to initialize the distance on each data, point to k different centroids. The following is the pseudocode for the assignment of the constrained k-means algorithm in this study:

Input:

- Data: The data points to cluster
- K: The desired number of cluster
- Constraints: Given constraints

Procedure ConstrainedKMeans(Data, K, Constraints):

1. Initialize centers randomly or using a specific method like K-Means++
2. Compute distances between data points and center points
3. Create a constraint matrix based on the given constraints
4. Repeat until convergence:
 - a. Assign each data point to the nearest cluster center, considering the constraints
 - b. Evaluate constraint violations and penalties
 - c. Optimize the clustering by adjusting data point assignments or center positions to reduce constraint violations
5. Return the final clustering results

Procedure AssignDataToClusters(Data, Centers, Constraints):

1. Initialize an empty list of clusters for each center

2. For each data point in Data:
 - a. Initialize the minimum distance to a large value
 - b. Initialize the assigned cluster index to -1
 - c. For each cluster index from 0 to K-1:
 - i. Calculate the distance between the data point and the cluster center
 - ii. If the distance is smaller than the current minimum distance and satisfies the constraints, update the minimum distance and assigned cluster index
 - d. Assign the data point to the assigned cluster index
 - e. Add the data point to the list of clusters for the assigned cluster index
3. Return the list of clusters

Procedure EvaluateConstraintViolations(Clusters, Constraints):

1. Initialize the total violation count to 0
2. For each constraint in Constraints:
 - a. If the constraint is a hard constraint:
 - i. If the constraint is violated in any cluster, increment the total violation count
 - b. If the constraint is a soft constraint:
 - i. For each cluster, calculate the constraint violation count based on the specific constraint criteria and add it to the total violation count
3. Return the total violation count

Procedure OptimizeClusterAssignments(Data, Centers, Clusters, Constraints):

1. For each data point in Data:
 - a. For each cluster index from 0 to K-1:
 - i. Calculate the violation count if the data point is assigned to the cluster index
 - b. Find the cluster index that results in the minimum violation count and assign the data point to that cluster
 - c. Update the cluster assignment in the Clusters list
2. For each cluster index from 0 to K-1:
 - a. Update the center position based on the data points assigned to the cluster
3. Return the updated Clusters and Centers

Main:

Initialize Data, K, and Constraints

Centers = InitializeCenters(Data, K) // Random initialization or K-Means++

Clusters = AssignDataToClusters(Data, Centers, Constraints)

PreviousViolationCount = EvaluateConstraintViolations(Clusters, Constraints)

Repeat until convergence:

 Clusters = OptimizeClusterAssignments(Data, Centers, Clusters, Constraints)

 CurrentViolationCount = EvaluateConstraintViolations(Clusters, Constraints)

 If CurrentViolationCount >= PreviousViolationCount:

 Break // Convergence reached

 PreviousViolationCount = CurrentViolationCount

Display Clusters

Firstly, the algorithm is initialized with the desired number of clusters or centroids and the constraints that define the relationships between data points. Next, the algorithm iteratively assigns data points to the nearest centroids while ensuring it adheres to the specified constraints. Finally, the algorithm refines the cluster assignments and updates the centroids to optimize the clustering while maintaining the given constraints. This constrained k-means approach allows for incorporating domain-specific knowledge into the clustering process, making it a valuable tool in various applications, such as image and customer segmentation for marketing. The stages of the constrained k-means algorithm task in this study are as follows:

- Choose student and school input data that states the coordinates of each variable for clustering.
- Determine the cluster's center, in this case, the coordinates of each school.
- Perform calculations based on the linear procedure method programming to find optimal clustering results with the requirement that there is a constraint on each cluster center (school) with the indicator variable $T_{i,h} \geq 2$, the equation used is as follows:

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

Subjek to :

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

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

$$T_{i,h} \geq 0, i = 1, \dots, m ; h = 1, \dots, k \tag{4}$$

- Enter data for each student to the nearest centroid (school) by comparing the distance between each student data point and each existing centroid cluster (school) concerning the computational results of linear programming. The equation of this stage is as follows:

$$C_{h,t+1} = \frac{\sum_{i=1}^m T_{i,h}^t X_i}{\sum_{i=1}^m T_{i,h}^t} \tag{5}$$

Where $C_{h,t+1}$ is the membership of k-data to i-cluster

- If there are still changes in the cluster membership data, go back to steps 3 - 4.
- The results of the data clustering process are the coordinates of the cluster center (school) and cluster members (students).

3.2.4. Parameter tuning

In machine learning, parameter tuning is used to optimize the algorithm by experimenting with limiting the data at each cluster center. Tuning these parameters serves as a reference control in the learning process. In this study, two clustering algorithms are compared in admitting new students to the city of Makassar, consisting of the k-means clustering algorithm and the constrained k-means algorithm, to find excellent and balanced optimal results for the amount of data in each cluster center. Each clustering algorithm is reviewed with control tuning parameters to obtain optimal clustering results. The implementation is done by looking at the specification requirements of the dataset to be used. The datasets used are the school location coordinates, and student domicile coordinates distribution data. The parameter tuning values are based on min-max computations on each prospective new student data and the number of schools as cluster centers.

4. RESULTS AND DISCUSSION

This study compares the results of determining school zoning for determining new student admissions at the high school level on the zoning registration route in Makassar. The steps taken to process the data processing as a whole apply clustering data mining modeling designed using the Python programming language for the clustering model are the K-Means clustering and Constrained K-Means algorithms model based on the research methodology described in section III of the proposed method. In particular, the research describes the stages of the analysis and discussion process and the flowchart chart of this research, which can be seen in Figure 3.

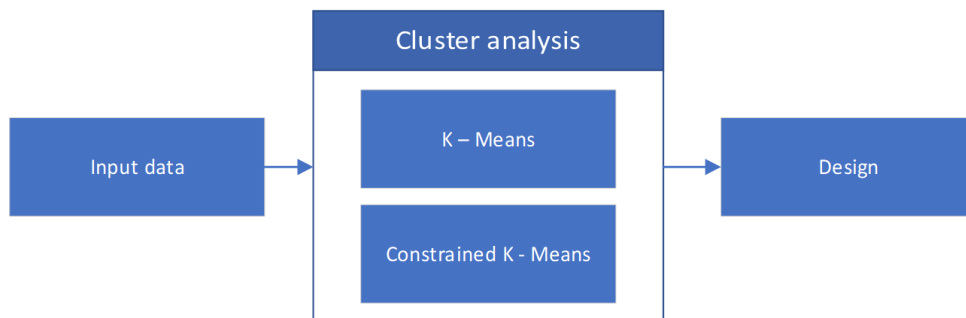


Figure 3. Flowchart of the stages of analysis and discussion

The process of clustering data using the K-Means algorithm involved utilizing school data as cluster center points with a set value of 22 and new student domicile data consisting of 2,248 data points. The outcomes of this clustering process were visually represented using Mapbox, as shown in Figure 4. Additionally, the results of testing the K-Means clustering algorithm were summarized in Table 3, where 22 cluster centers (schools) and the distribution of 2,248 student domiciles were used. The analysis revealed that the cluster membership was asymmetric, particularly among schools located within a specific radius, forming a non-circular pattern in the distribution of students. This information is valuable for understanding the spatial distribution of students about school locations and can have implications for optimizing school planning and resource allocation.

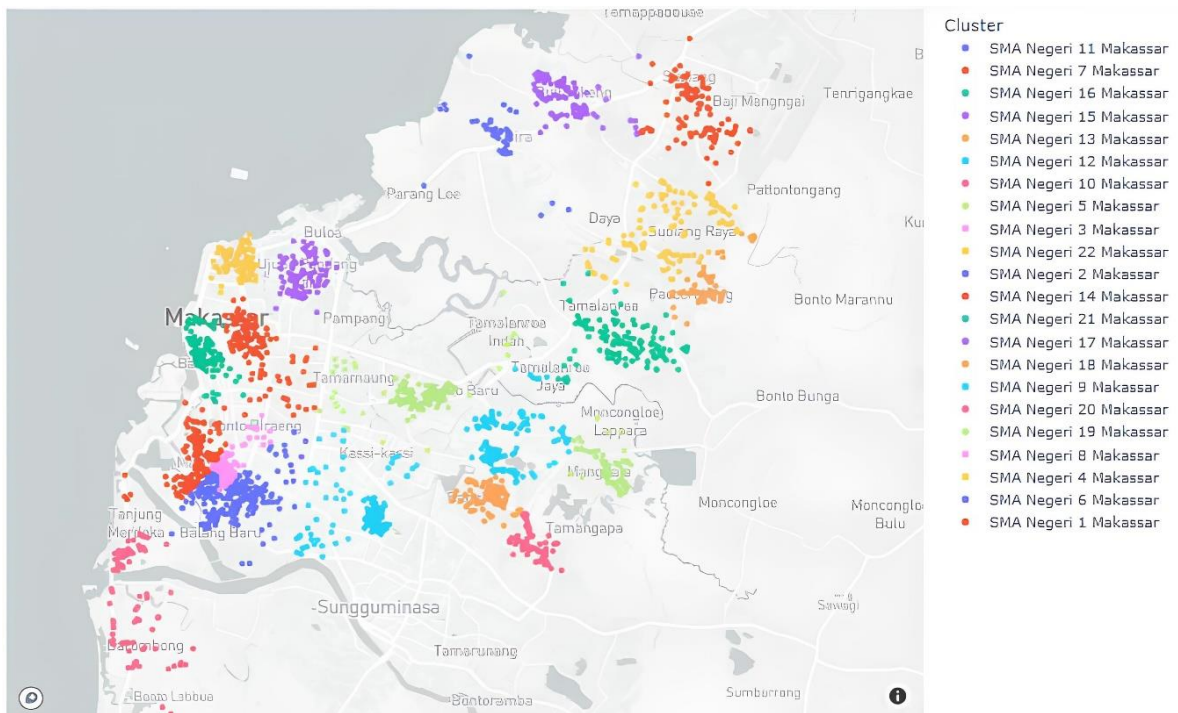


Figure 4. Mapbox visualization of algorithm k-means clustering results

Table 3. Data results of k-means clustering

No	School	Latitude	Longitude	Cluster	Result
1	SMAN 1 Kota Makassar	-5.13498	119.4191	18	136
2	SMAN 10 Kota Makassar	-5.18609	119.4884	14	81
3	SMAN 11 Kota Makassar	-5.17158	119.4159	2	162
4	SMAN 12 Kota Makassar	-5.16338	119.4832	16	139
5	SMAN 13 Kota Makassar	-5.17529	119.4779	6	95
6	SMAN 14 Kota Makassar	-5.16583	119.4089	19	148
7	SMAN 15 Kota Makassar	-5.08025	119.495	0	101
8	SMAN 16 Kota Makassar	-5.13749	119.4114	11	106
9	SMAN 17 Kota Makassar	-5.12074	119.4297	15	98
10	SMAN 18 Kota Makassar	-5.12534	119.5328	5	54
11	SMAN 19 Kota Makassar	-5.16376	119.5125	3	71
12	SMAN 2 Kota Makassar	-5.1695	119.4126	21	67
13	SMAN 20 Kota Makassar	-5.20276	119.3984	7	82
14	SMAN 21 Kota Makassar	-5.13712	119.5139	10	136
15	SMAN 22 Kota Makassar	-5.11419	119.5224	13	113
16	SMAN 3 Kota Makassar	-5.16863	119.4131	17	87
17	SMAN 4 Kota Makassar	-5.11728	119.4191	4	97
18	SMAN 5 Kota Makassar	-5.14732	119.4618	9	122
19	SMAN 6 Kota Makassar	-5.08863	119.4818	12	54
20	SMAN 7 Kota Makassar	-5.08136	119.5337	8	104
21	SMAN 8 Kota Makassar	-5.17004	119.4145	20	53
22	SMAN 9 Kota Makassar	-5.17774	119.4501	1	142

In the data mining testing utilizing the latest Constrained K-Means algorithm for clustering modeling, 22 school data points were used as cluster center points, and there were 2,248 new student domicile data points. The findings of this analysis were visually represented in Figure 5 using Mapbox, while a summary of the results can be found in Table 4. The results indicate that the Constrained K-Means algorithm effectively determined school zoning areas for students in the public high schools of Makassar City, allocating each student to the nearest school in a balanced manner. Notably, one school had a cluster membership of 85 students. In contrast, the remaining 21 schools had 103 students as cluster members, demonstrating that the Constrained K-Means algorithm approach ensured that cluster members (students) and cluster centers (schools) did not overlap. This information is crucial for optimizing school allocation and providing equitable access to educational resources.

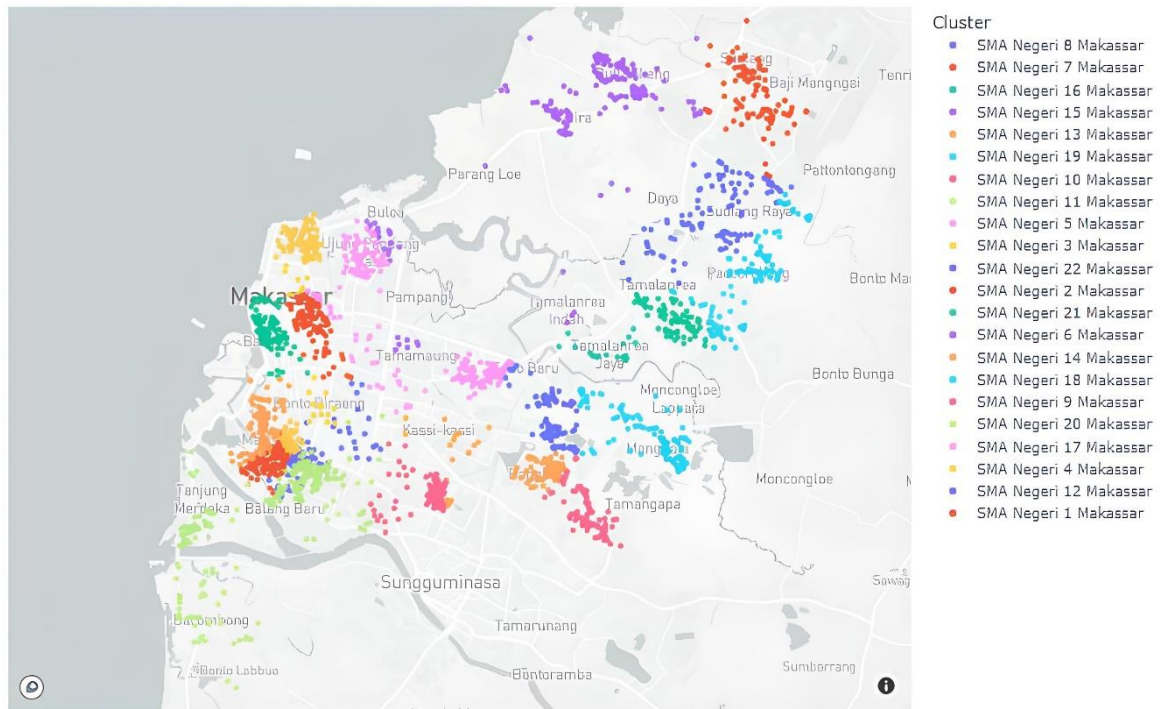


Figure 5. Mapbox visualization of the constrained k-means clustering results algorithm modeling

Table 4. Data results of constrained k-means clustering

No	School	Latitude	Longitude	Cluster	Results
1	SMAN 1 Kota Makassar	-5.13498	119.4191	4	103
2	SMAN 10 Kota Makassar	-5.18609	119.4884	14	103
3	SMAN 11 Kota Makassar	-5.17158	119.4159	20	103
4	SMAN 12 Kota Makassar	-5.16338	119.4832	16	103
5	SMAN 13 Kota Makassar	-5.17529	119.4779	5	103
6	SMAN 14 Kota Makassar	-5.16583	119.4089	1	103
7	SMAN 15 Kota Makassar	-5.08025	119.495	3	103
8	SMAN 16 Kota Makassar	-5.13749	119.4114	18	103
9	SMAN 17 Kota Makassar	-5.12074	119.4297	10	103
10	SMAN 18 Kota Makassar	-5.12534	119.5328	2	103
11	SMAN 19 Kota Makassar	-5.16376	119.5125	9	103
12	SMAN 2 Kota Makassar	-5.1695	119.4126	21	103
13	SMAN 20 Kota Makassar	-5.20276	119.3984	7	103
14	SMAN 21 Kota Makassar	-5.13712	119.5139	12	103
15	SMAN 22 Kota Makassar	-5.11419	119.5224	11	103
16	SMAN 3 Kota Makassar	-5.16863	119.4131	19	103
17	SMAN 4 Kota Makassar	-5.11728	119.4191	17	103
18	SMAN 5 Kota Makassar	-5.14732	119.4618	0	103
19	SMAN 6 Kota Makassar	-5.08863	119.4818	13	85
20	SMAN 7 Kota Makassar	-5.08136	119.5337	6	103
21	SMAN 8 Kota Makassar	-5.17004	119.4145	15	103
22	SMAN 9 Kota Makassar	-5.17774	119.4501	8	103

5. CONCLUSIONS

This study compares the performance of computational clustering on the K-Means algorithm with the Constrained K-Means clustering algorithm model in a case study to determine the regulation of accepting new students at the public high school level in Makassar City. An approach using models algorithm Constrained K-Means can map data on prospective new students at each school regarding the calculation parameters for the distance to the nearest school to the domicile of the prospective student. Based on the data clustering process results, the distribution of student zoning clustering in accepting new prospective students is optimal because each school's cluster membership data is balanced. In contrast, the data processing results of the old K-Means clustering algorithm show that the data distribution of students and cluster centers needs to be more balanced, especially in adjacent school areas. Therefore, a data clustering model with the Constrained K-Means algorithm can be used in government agencies engaged in education for efficient and accurate regulation of school zoning systems.

ACKNOWLEDGEMENTS

The author is delighted that he can express his heartfelt gratitude for the assistance he received from Hasanuddin University in Makassar, Indonesia.




REFERENCES

- [1] I. A. Rahmayanti, S. N. Apsariny, and A. Meganfi, "The effect of school zoning system to the quality of education in Senior High Schools (Case study of public Senior High Schools in Surabaya)," *International Journal of Academic and Applied Research*, vol. 5, no. 2, pp. 103–106, 2021.
- [2] 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," in *2nd International Conference on Social Science and Character Educations (ICoSSCE 2019)*, 2020, vol. 398, pp. 277–282, doi: 10.2991/assehr.k.200130.056.
- [3] A. A. N. Risal, Z. Zainuddin, and M. Niswar, "School zoning system for student admission using constrained k-means algorithms," in *2022 IEEE International Conference on Communication, Networks and Satellite (COMNETSAT)*, 2022, pp. 174–178, doi: 10.1109/COMNETSAT56033.2022.9994366.
- [4] M. S. R. Batita and P. H.-M. Tsai, "A study of student admission by school zoning system in Indonesia : Problem or Solution ?," National Chung Cheng University, 2020.
- [5] M. D. Febriana, Z. Zainuddin, and I. Nurtanio, "School zoning system using k-means algorithm for High School Students in Makassar City," in *2019 2nd International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2019*, 2019, pp. 368–372, doi: 10.1109/ISRITI48646.2019.9034601.
- [6] M. Hajaroh, R. Nurhayati, F. Sidiq, A. S. Raharjo, and E. Sholikhah, "School zoning policy and equalization of education access for poor students in Yogyakarta City," in *The 2nd International on Meaningful Education (2nd ICMEd)*, 2021, vol. 2021, pp. 245–255, doi: 10.18502/kss.v6i2.9992.
- [7] M. Ghanavati, R. K. Wong, F. Chen, Y. Wang, and C.-S. Perng, "An effective integrated method for learning big imbalanced data," in *Proceedings - 2014 IEEE International Congress on Big Data, BigData Congress 2014*, 2014, no. 2, pp. 691–698, doi: 10.1109/BigData.Congress.2014.102.
- [8] W. Utomo, "The comparison of k-means and k-medoids algorithms for clustering the spread of the Covid-19 outbreak in Indonesia," *ILKOM Jurnal Ilmiah*, vol. 13, no. 1, pp. 31–35, 2021, doi: 10.33096/ilkom.v13i1.763.31-35.
- [9] P.-N. Tan, M. Steinbach, A. Karpatne, and V. Kumar, *Introduction To Data Mining Second Edition*, Second Edi. New York: Pearson Education, 2019.
- [10] R. Panthong and T. Wongkanthiya, "Analysis of clustering and association using data mining technique for elderly health condition dataset," *IAES International Journal of Artificial Intelligence*, vol. 12, no. 4, pp. 1774–1783, 2023, doi: 10.11591/ijai.v12.i4.pp1774-1783.
- [11] M. Faisal, E. M. Zamzami, and Sutarman, "Comparative analysis of inter-centroid k-means performance using Euclidean Distance , Canberra Distance and Manhattan Distance," *Journal of Physics: Conference Series*, 2019, doi: 10.1088/1742-6596/1566/1/012112.
- [12] D. Saini and M. Singh, "Achieving balance in clusters-a survey," *International Research Journal of Engineering and Technology (IRJET)*, vol. 02, no. 09, pp. 2611–2614, 2015.
- [13] A. Azizah, A. A. Ilham, and Syafaruddin, "Spatial analysis of the spread of tuberculosis cases based on socio-economic factors using distance-based algorithm," in *2022 International Conference on Electrical Engineering and Informatics (ICELTICS)*, 2022, pp. 19–24, doi: 10.1109/iceltics56128.2022.9932043.
- [14] M. Pokharel, J. Bhatta, and N. Paudel, "Comparative analysis of k-means and enhanced k-means algorithms for clustering," *NUTA Journal*, vol. 8, no. 1–2, pp. 79–87, 2021, doi: 10.3126/nutaj.v8i1-2.44044.
- [15] Y. Duan, Q. Liu, and S. Xia, "An improved initialization center k-means clustering algorithm based on distance and density," 2018, doi: 10.1063/1.5033710.
- [16] T. Ma and W. Shen, "Research on a hybrid K-means clustering algorithm based on improved genetic algorithm," in *Proceedings - 2017 International Conference on Computer Technology, Electronics and Communication, ICCTEC 2017*, 2017, pp. 502–507, doi: 10.1109/ICCTEC.2017.00115.
- [17] K. Lei, S. Wang, W. Song, and Q. Li, "Size-constrained clustering using an initial points selection method," in *International Conference on Knowledge Science, Engineering and Management*, 2013, vol. 8041, pp. 373–383, doi: 10.1007/978-3-642-39787-5.
- [18] T. Wang and J. Gao, "An Improved K-Means algorithm based on kurtosis test," *Journal of Physics: Conference Series*, vol. 1267, no. 1, 2019, doi: 10.1088/1742-6596/1267/1/012027.
- [19] A. A. Zuenko, O. V. Fridman, O. N. Zuenko, and O. G. Zhuravleva, "An approach to solution of constrained clustering problems using the constraint programming paradigm and the multiset theory," *AIDTTS 2020. Journal of Physics: Conference Series*, vol. 1801, no. 1, 2021, doi: 10.1088/1742-6596/1801/1/012041.




- [20] J. Zhao, "Optimal clustering: genetic constrained k-means and linear programming algorithms," Virginia Commonwealth University, 2006.
- [21] P. S. Bradley, K. P. Bennett, and A. Demiriz, "Constrained K-Means clustering," *Microsoft Research*, p. 9, 2000.
- [22] N. Ratnasar *et al.*, "Implementation of Generate and Test Algorithm for Junior High School Zoning System in Malang," in *2020 The 4th International Conference on Vocational Education and Training*, 2020, pp. 167–170, doi: 10.1109/ICOVET50258.2020.9230207.
- [23] U. Syaripudin, N. A. Fauzi, W. Uriawan, W. B. Z, and A. Rahman, "Haversine formula implementation to determine Bandung City school zoning using android based location based service," 2019, doi: <http://dx.doi.org/10.4108/eai.11-7-2019.2303558>.
- [24] L. S. Thota, F. Fathima, S. B. Changalasetty, and M. Shiblee, "Cluster based zoning of crime info," in *2017 2nd International Conference on Anti-Cyber Crimes, ICACC 2017*, 2017, pp. 87–92, doi: 10.1109/Anti-Cybercrime.2017.7905269.
- [25] P. H. P. Rosa, R. Gunawan, and I. A. Dwiatmoko, "The clustering of high schools based on national and school examinations: a case study at Daerah Istimewa Yogyakarta Province," in *2015 International Conference on Data and Software Engineering*, 2015, pp. 231–236, doi: 10.1109/ICODSE.2015.7437003.
- [26] D. R. Ramdania, R. Andrian, M. Irfan, R. Z. Abidin, and F. M. Kaffah, "On designing application of finding nearby islamic boarding schools in West Java using haversine formula and euclidean distance algorithms," 2019, doi: 10.4108/eai.11-7-2019.2297517.

BIOGRAPHIES OF AUTHORS



Zahir Zainuddin    holds a Doctor of Computer Engineering from Bandung Institute of Technology, Indonesia, in 2004. He also received his B.Sc. in Electrical Engineering Department, Hasanuddin University, Indonesia, in 1988 and his M.Sc. (Computer Engineering) from Florida Institute of Technology USA in 1995. He is an associate professor at the Department of Informatics at Hasanuddin University in Indonesia. His research includes Computer Systems, intelligent systems, computer vision, and smart cities. He has published over 60 papers in international journals and conferences. In 1989, he was a JSPS research fellow at the Tokyo Institute of Technology. He can be contacted at email: zahir@unhas.ac.id.



Andi Alviadi Nur Risal    He earned his Bachelor of Education (S.Pd) in Informatics and Computer Engineering Education from Makassar State University, Indonesia, in 2017. He is pursuing his Master of Engineering (M.T) in the Department of Electrical Engineering with a concentration in Informatics Engineering at Hasanuddin University, Indonesia. He can be contacted via email: risalaan19d@student.unhas.ac.id