Institut Riset dan Publikasi Indonesia (IRPI)
IJATIS: Indonesian Journal of Applied Technology and Innovation Science
Journal Homepage: https://journal.irpi.or.id/index.php/ijatis
Vol. 1 Iss. 1 February 2024, pp: 23-29
ISSN(P): 3032-7466 | ISSN(E): 3032-7474

Implementation of K-Means, K-Medoid and DBSCAN Algorithms In Obesity Data Clustering

Elsa Setiawati[1*], Ustara Dwi Fernanda[2], Suci Agesti[3],
Muhammad Iqbal[4], Muhammad Okten Adetama Herjho[5]

[1,2,3]Department of Information System, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasm, Indonesia
[4]Department of Arabic Language and Literature, Arabic Language Al-Azhar University, Egypt
[5]Department of Syariah Islamiyyah, Syariah Islamiyyah Al-Azhar University, Egypt

E-Mail: [1]12150325178@students.uin-suska.ac.id, [2]12150314341@students.uin-suska.ac.id,
[3]1212150323229@students.uin-suska.ac.id, [4]Azhar@azhar.eun.eg, [5]panglimaongga@gmail.com.

Received Dec 02nd 2023; Revised Dec 22th 2023; Accepted Jan 05th 2024
Corresponding Author: Elsa Setiawati

Abstract

Obesity is an excessive accumulation of body fat and can be harmful to health. This study aims to understand the patterns and relationships between obesity data that have been obtained, so a data clustering step will be carried out using the K-Means, K-Medoid and DBSCAN algorithms. This study utilizes the Davies Bouldin Index (DBI) to determine the best cluster value comparison and validated. So the results of the best cluster value in processing obesity data are using the K-Means K2 algorithm with a value of 0.604. The K-Medoid algorithm obtained the best cluster k2, with a DBI value of around 0.614. and the DBSCAN algorithm clustering trial K3, with a value of 1.040. Thus in this study the comparison results of the application of 3 clustering algorithms, the results obtained are the K-Means algorithm shows the value of the resulting cluster is the best of other algorithms in clustering obesity data with a value of 0.604.

Keyword: Clustering, DBSCAN, DBI, K-Means, K-Medoid, Obesity

1. INTRODUCTION

In line with the World Health Organization (WHO) definition, obesity is characterized as an abnormal or excessive buildup of fat that indicates an elevated health risk [1]. In general, healthy people have a balanced body weight according to their age. This means that as you age, you also need to control your weight and avoid excessive weight gain.But if this balance is disrupted and you gain weight, you will become obese.

Obesity is an imbalance in energy homeostasis caused by the difference between energy intake and energy use. Obesity can elevate the likelihood of encountering various health conditions, including but not limited to high blood pressure, cardiovascular disease, diabetes, cancer, osteoporosis, and several other ailments [2]. These health issues may adversely affect an individual's level of productivity and life expectancy. Obesity involves the buildup of surplus body fat and may present a risk to one's health.Obesity is closely related to many chronic diseases, and chronic diseases have long-term negative impacts on those affected. Obesity is caused by an imbalance between energy intake and energy production in the body [3]. Obesity not only affects physical health, but also psychological, social, and even spiritual health. Obesity also contributes to increased health care costs due to social impacts and decreased quality of life due to the many diseases and problems faced by people with obesity.

Obesity, or overweight, has become an alarming global epidemic and one of the most serious health problems worldwide. This phenomenon occurs when fat in the body accumulates excessively or abnormally and poses a serious negative impact on human health [5]. Also, if your weight and height are balanced, you are considered to be in good health. Obesity impacts individuals across various age groups and is prevalent among both adolescents and adults in numerous countries. In 2001, the prevalence was reported to be 30% in the United States and 22% in the United Kingdom[6].

In the contemporary age of information technology, a novel method has surfaced to tackle the issue of obesity by employing data mining methodologies. Data mining, characterized as the automated exploration for valuable insights within extensive datasets, presents an avenue for scrutinizing and unveiling previously undiscovered correlations in obesity data. In the analysis of this data, clustering techniques, including the

DOI: https://doi.org/10.57152/IJATIS.v1i1.1109

application of K-Means, K-Medoids, and DBSCAN algorithms, are employed to streamline the intricacy of the information and enhance comprehension of the associations between variables.

The clustering process involves the creation of data groups according to similarities and serves as a crucial foundation for data analysis. The K-Means algorithm is used as a way to index the major cases within the case base. The optimal number of k groups to be formed was determined using the DBI method [24]. The K-Means method is a method included in the distance-based clustering algorithm that partitions data into a set of clusters. This algorithm only works for numeric attributes[25]. The K-Means algorithm, a clustering technique, organizes data by its closeness to a center point within a group, referred to as a centroid [7]. K-means belongs to the category of the simplest unsupervised learning algorithms to handle clustering problems. It relies on a simple iterative approach to find the optimal local solution [8]. On the other hand, K-Medoids is utilized to identify the medoid within a cluster, serving as the representative center of the cluster [9]. K-Medoids have an advantage over K-Means in that the number of differences between data objects can be reduced by selecting representative objects. K-medoid is a classical partitioning technique of clustering that clusters the data set of n objects into k number of clusters. This k: the number of clusters required is to be given by user. This algorithm works on the principle of minimizing the sum of dissimilarities between each object and its corresponding reference point [26][27].

From the previous explanation of the application of clustering algorithms that researchers will do in this study with the title "Application of K-Means, K-Medoid and DBSCAN Algorithms in Obesity Data Grouping", namely to determine the optimal number of groups in the context of obesity data, this is to understand the complex variations in the population of obese patients and detail the clinical characteristics that can be used to further cluster.

2. MATERIAL AND METHOD

In this study, the first stage is data collection, the dataset applied is obesity data taken from keggle. The second stage is data pre-processing, then the data clustering stage is carried out using 3 algorithms, namely, K-Means, K-Medoid, and DBSCAN. The research method can be presented visually in Figure 1.

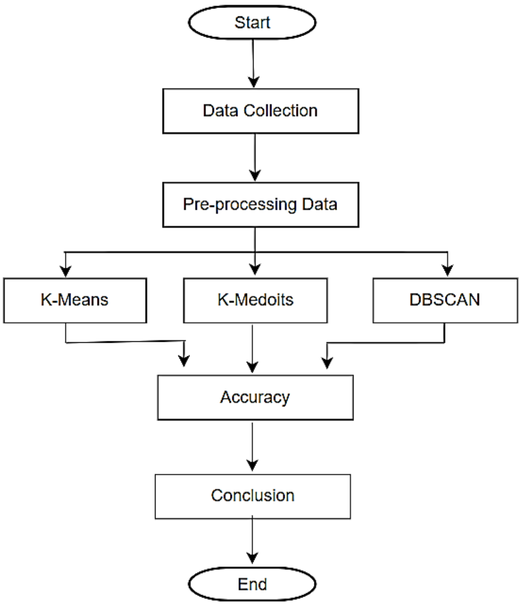

Figure 1. Research Methodology

2.1. Data Collecting

Data collection is a major challenge in the development of machine learning and is the focus of research conducted by various communities[12]. The information utilized in this research was gathered from a dataset acquired from Kaggle.com, consisting of data related to obesity.

2.2. Data Mining

Data mining involves the examination of data from diverse viewpoints, with the aim of generating valuable information that can increase profits, reduce costs, or even achieve both goals. In its technical aspect, data mining can be explained as an attempt to find correlations or patterns from hundreds or thousands of data fields in an extensive relational database [13].

2.3. Clustering

Clustering is a method of categorizing data into two or more groups, where the data points within a group exhibit greater similarity to each other than to those in other groups. This categorization relies on the information contained within the data points [12]. Clustering is the act of classifying data into clusters or groups based on their degree of similarity. In the process of clustering, akin data is assembled in a single group, while dissimilar data is segregated into distinct groups [14][15]. In this study, the primary approach employed for data management is clustering, wherein medical record data is collected and processed using a programming language through the application of clustering methodology.

2.4. K-Means

K-Means Clustering is an unsupervised modeling technique or data mining method for data analysis.It is one of the approaches used to cluster data with a partitioning approach. The K-Means method seeks to group data into clusters, where data in one cluster have similar characteristics, while different characteristics are grouped in other clusters [16].The primary objective of K-Means is to enhance similarity within a cluster while diminishing similarity between distinct clusters. The pivotal parameter utilized to boost data similarity within a cluster is the distance function, where data similarity is determined by the nearest distance from the cluster center [17].

The procedures of the k-means algorithm include:

1. Choose a value of k as the desired total number of clusters.
2. Initializing the cluster center k can be done using various methods, but the most common is by random sampling from the available data.
3. Measuring the distance to each centroid of all input data is done using the Euclidean distance formula until the closest distance to the centroid of all data is found. The following is the Euclidean distance equation 1:

$$De = \sqrt{(x_i - s_i)^2 (y_i - t_i)^2} \tag{1}$$

4. Classify all data based on proximity to centroid (shortest distance).
5. Update the center of gravity value. The new centroid value is determined from the average of the associated clusters using equation 2.

$$v_{ij} = \frac{1}{N} \sum_{k=0}^{n_i} X_{ij} \tag{2}$$

6. Repeat steps 2 through 10 until no members of each cluster change.

2.5. K-Medoids

In addition, the K-Medoids algorithm is used to find the medoid in the group that serves as the center of the group. K-Medoids has many advantages over K-Means. This is because K-Medoids finds k objects that represent a group by minimizing the sum of differences between data objects, while K-Means uses the sum of squared Euclidean distances. data objects. This approach helps reduce noise and values that are far from the average. But the K-Means algorithm also has drawbacks [18]: finding the value of K is a difficult task. It does not give effective results when used on global clusters. Your cluster results may change if different boot partitions are chosen. Differences in cluster size and density are not taken into account by the algorithm. To overcome the shortcomings of the K-Means mathematical algorithm, we use the K-Medoids algorithm, which is based on object representation technology [19]. A medoid is the most central cluster data object. A medoid is randomly selected from Ky data objects to form the Ky cluster, and the remaining data objects are placed near the medoid in the cluster. It then processes all data objects in the cluster to iteratively find a new medoid that better represents the new cluster. After finding the new medoid, merge all data objects into the cluster. The position of the medoid changes at each iteration. Therefore, a cluster representing n data objects is formed [20].

$$deu(xij\ , ckj) = \sqrt{}\sum\ \sum\ (xij\ -\ ckj)\ 2\ n\ i{=}1\ p\ j{=}1 \tag{3}$$

2.6. DBSCAN

DBSCAN clustering does not have all these drawbacks, and most importantly, it can handle noisy data and outliers very efficiently. [21] DBSCAN is a partition clustering technology.[22] DBSCAN is the first density-based clustering algorithm. Reported by Ester et al. he suggested. Developed in 1996, it is used to cluster data of various formats in the presence of noise in high-dimensional spatial and non-spatial databases. The main idea of DBSCAN is that for each neighborhood cluster object, the specified radius (Eps) must contain at least the minimum number of objects (MinPts). That is, the cardinality of the neighborhood must be above a certain threshold. - Define the neighborhood of any point p as .

$$\text{Formula of DBSCAN 1. } N_{Eps} = \{q \text{ € } D \,/\, dist\ (p, q) < Eps\} \quad (4)$$

Here D is a database object. A point P is called a key point if there are at least a certain number of points around it. A key point is defined as:

$$\text{Formula of DBSCAN 2. } N_{Eps}(P) > MinPts \quad (5)$$

Here, Eps and MinPts are user-defined parameters, which mean the neighborhood radius of the core point and the minimum number of neighborhood points, respectively. If these conditions are not met, then the point is fulfilled as a non-core point [23].

2.7. Google Collaboratory

Google Collaboratory (commonly known as Google Collab) is an open-source service provided by Google to anyone with a Gmail account. It gives you the flexibility to use any type of code and and any size of data set you want. Once your Google Drive is connected to Google Collab, you can use it as [24].

3. RESULTS AND ANALYSIS

3.1. K-Means

By employing clustering methods on obesity data, specifically utilizing the K-Means algorithm along with the Davies Bouldin Index (DBI) technique, the outcomes of the clustering process can be visualized through a cluster diagram. This diagram provides a clear representation of how the data is distributed within each group as figure 2.

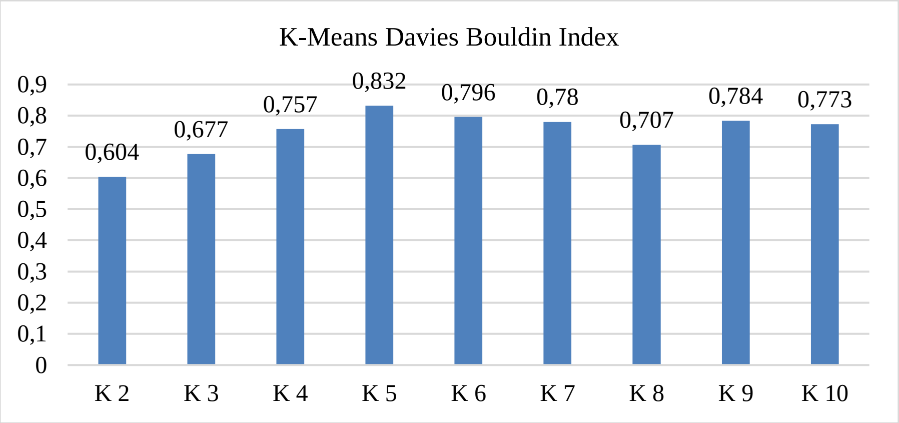

Figure 2. K-Means Davies Bouldin Index

In the framework of this research, a series of 9 cluster tests were conducted, starting from k=2 to k=10, with the aim of identifying the number of clusters that provide optimal results. The success of the cluster is measured by the degree to which the value approaches 0 and is non-negative. From the illustration above, The optimal clustering results in cluster K2 with a value of 0.604 are evident in the performance of the K-Means Algorithm.

3.2. K-Medoids

For the subsequent clustering phase, the utilization of the K-Medoids Algorithm aimed to identify the most favorable cluster by employing the Davies Bouldin Index (DBI) technique. An evaluation of the quality of the data clustering process with the K-Medoids Algorithm can be found in Figure 3, which shows the DBI value as figure 3.

Based on the DBI value outcomes, it can be deduced that the optimal outcome of the experiment employing the K-Medoids Algorithm is observed at k2, with a DBI value approximately equal to 0.614. In this configuration, the dataset is divided into two clusters, where cluster 0 has 1243 data, and cluster 1 has 757 data. This evaluation provides a deeper understanding of the quality and distribution pattern of the clusters generated by the K-Medoids Algorithm in a particular experiment.

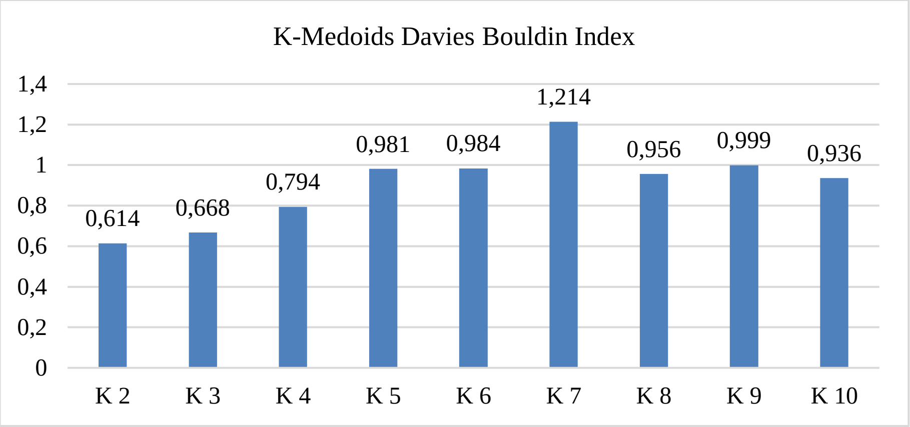

Figure 3. K-Medoids Davies Bouldin Index

3.3. DBSCAN

In testing clusterization using the DBSCAN method, 9 tests were carried out starting from K = 2 to K = 10 using the Davies Bouldin index (DBI). This aims to determine the number of clusters that provide optimal performance, as shown in Figure 4.

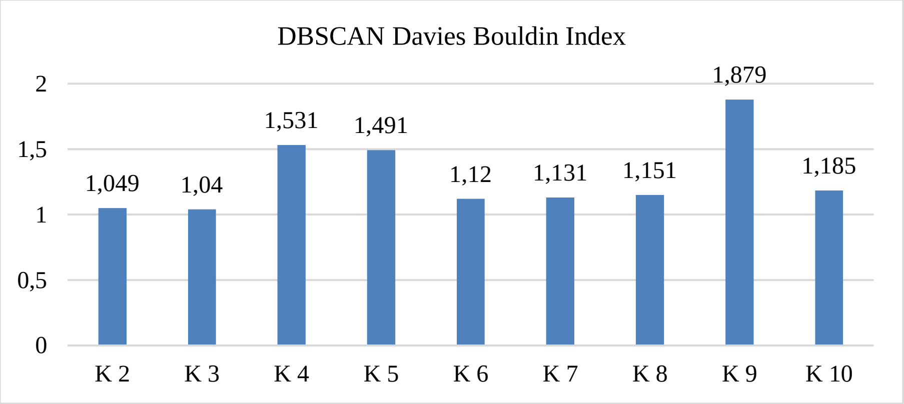

Figure 4. DBSCAN Davies Bouldin Index

Based on the figure above, it can be concluded that the DBSCAN algorithm using the Davies Bouldin Index (DBI) technique achieved the best performance in experiment K3, with a value of 1.040. This is because this number is closest to the value of 0, indicating that the number of clusters 3 provides optimal results in this test.

3.4. Comparison of K-Means, K-Medoids, and DBSCAN Algoritma

A comparison between data mining algorithms in the context of Clustering methods, namely K-means, K-Medoids, and DBSCAN, can be seen in Figure 5.

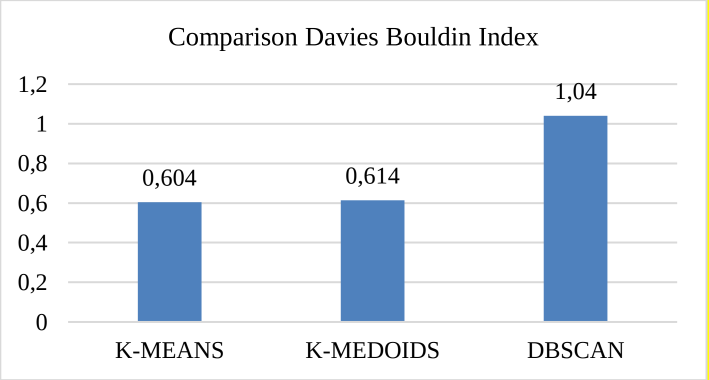

Figure 5. Comparison Davies-Bouldin Index

Figure 4 illustrates that among the three clustering algorithms, the most effective cluster is evident in the K-Means algorithm, achieving a score of 0.604. Consequently, this research asserts that the optimal cluster is achieved using the K-Means algorithm in the K2 experiment, where the validity test outcome attains 0.604.

4. CONCLUSION

This research focuses on determining the optimal number of clusters in the context of obesity data, this is to understand the complex variation in the obese patient population and detail the clinical characteristics that can be used to perform further clustering using K-Means, K-Medoid and DBSCAN algorithms. From this research, it can be concluded that the K-Means algorithm is the best choice for clustering obesity data. This algorithm shows optimal performance on the K2 cluster, with a Davies Bouldin Index (DBI) value of about 0.604. Although K-Medoids also gave good results on the K2 test (DBI of about 0.614), and DBSCAN showed the best performance on K3 (DBI of 1.040), the comparison confirmed that K-Means on K2 was the optimal choice with a validity test of about 0.604.

REFERENCES

[1] Muscogiuri, G., Verde, L., Sulu, C., Katsiki, N., Hassapidou, M., Frias-Toral, E., ... & Barrea, L. (2022). Mediterranean diet and obesity-related disorders: what is the evidence?. Current Obesity Reports, 11(4), 287-304..

[2] Thamrin, S. A., Arsyad, D. S., Kuswanto, H., Lawi, A., & Nasir, S. (2021). Predicting Obesity in Adults Using Machine Learning Techniques: An Analysis of Indonesian Basic Health Research 2018. *Frontiers in Nutrition*, *8*. https://doi.org/10.3389/fnut.2021.669155

[3] Septiyanti, S., & Seniwati, S. (2020). Obesity and Central Obesity in Indonesian Urban Communities. *Jurnal Ilmiah Kesehatan (JIKA)*, *2*(3), 118–127. https://doi.org/10.36590/jika.v2i3.74

[4] Susi Muktiharti, Purwanto, Imam Purnomo, Rosmiati Saleh. Fakultas Ilmu Kesehatan, Program Studi Kesehatan Masyarakat, Universitas Pekalongan. Faktor Risiko Kejadian Obesitas pada Remaja SMA Negeri 2 dan SMA Negeri 3 di Kota Pekalongan Tahun 2010. Diunduh dari:http://www.download.portalgaru da.org/ipi21062.pdf. Akses: 17 September 2015.

[5] Susi Muktiharti, Purwanto, Imam Purnomo, Rosmiati Saleh. Fakultas Ilmu Kesehatan, Program Studi Kesehatan Masyarakat, Universitas Pekalongan. Faktor Risiko Kejadian Obesitas pada Remaja SMA Negeri 2 dan SMA Negeri 3 di Kota Pekalongan Tahun 2010.

[6] Hadi, H. (2004). Gizi lebih sebagai tantangan baru dan implikasinya terhadap kebijakan pembangunan kesehatan nasional. *Jurnal Gizi Klinik Indonesia*, *1*(2), 47-53.2

[7] Brunner, E. J., Chandola, T., & Marmot, M. G. (2007). Prospective effect of job strain on general and central obesity in the Whitehall II Study. *American Journal of Epidemiology*, *165*(7), 828–837. https://doi.org/10.1093/aje/kwk058

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

[9] Chen, J., Qi, X., Chen, L., Chen, F., & Cheng, G. (2020). Quantum-inspired ant lion optimized hybrid k-means for cluster analysis and intrusion detection. *Knowledge-Based Systems*, *203*. https://doi.org/10.1016/j.knosys.2020.106167

[10] Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm for Big Data. *Physics Procedia*, *78*, 507–512. https://doi.org/10.1016/j.procs.2016.02.095

[11] Ahmed, K. N., & Razak, T. A. (2016). IJARCCE An Overview of Various Improvements of DBSCAN Algorithm in Clustering Spatial Databases An Overview of Various Improvements of DBSCAN Algorithm in Clustering Spatial Databases. *International Journal of Advanced Research in Computer and Communication Engineering*, *5*(2). https://doi.org/10.17148/IJARCCE.2016.5277

[12] Govindasamy, K., & Velmurugan, T. (2018). Analysis of student academic performance using clustering techniques. *International Journal of Pure and Applied Mathematics*, *119*(15), 309-323.

[13] Aggarwal, D., & Sharma, D. (2019). Application of clustering for student result analysis. In *International Journal of Recent Technology and Engineering*. https://www.researchgate.net/publication/333115249

[14] Ananda, L. R. (2018). Clustering Untuk Menentukan Calon Mahasiswa Berprestasi. Jiti, 1(2), 16–19.

[15] Alfina, Santosa, Barkbah. “Analisa Perbandingan Metode Hierarchical Clustering, KMeans dan Gabungan Kedua dalam Cluster Data”, Jurnal Teknik ITS, Vol. 1, No. 1, 2012.

[16] Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *J*, *2*(2), 226–235. https://doi.org/10.3390/j2020016

[17] Murni, D., Efendi, B., Rahmadani, N., Informasi, S., & Tinggi Manajemen Informatika dan Komputer Royal Kisaran, S. (2022). IMPLEMENTATION OF EMPLOYEE DISCIPLINE CLUSTERING AT GOTTING SIDODADI VILLAGE OFFICE BANDAR PASIR MANDOGE USING K-MEANS

ALGORITHM. *Jurnal Teknik Informatika (JUTIF)*, *3*(2), 295–304. https://doi.org/10.20884/1.jutif.2022.3.2.236

[18] Saurabh Shah & Manmohan Singh “Comparison of A Time Efficient Modified K-Mean Algorithm with K-Mean and K-Medoids algorithm”, International Conference on Communication Systems and Network Technologies, 2012I.

[19] Jiawei Han, Han Kamber “Data Mining Concept and Techniques" ,2nd Edition

[20] Shalini S Singh & N C Chauhan ,“K-Means v/s KMedoidss: A Comparative Study”, National Conference on Recent Trends in Engineering & Technology, 2011.

[21] Saputra, R., Mustakim, M., Okfalisa, O., & Ridwan, M. Menentukan Popularitas Calon Presiden dan Tren pada Pilpres 2019 menggunakan Algoritma DBSCAN. In *Seminar Nasional Teknologi Informasi Komunikasi dan Industri* (pp. 123-130).

[22] Chakraborty NKNagwani Lopamudra Dey, S. (2011). Performance Comparison of Incremental K-means and Incremental DBSCAN Algorithms. In *International Journal of Computer Applications* (Vol. 27, Issue 11).

[23] Khan, Kamran, et al. "DBSCAN: Past, present and future." *The fifth international conference on the applications of digital information and web technologies (ICADIWT 2014)*. IEEE, 2014.

[24] Made, I., Bimantara, S., & Supriana, W. (2022). *CASE BASED REASONING (CBR) FOR OBESITY LEVEL ESTIMATION USING K-MEANS INDEXING METHOD*. *11*(4). https://archive.ics.

[25] Mega, W. (2015). *CLUSTERING MENGGUNAKAN METODE K-MEANS UNTUK MENENTUKAN STATUS GIZI BALITA* (Vol. 15, Issue 2).

[26] Radhika Kyadagiri ,Prof. D. Jamuna ,Masthan Mohammed, “An Efficient Density based Improved K-Medoids Clustering algorithm” ,International Journal of Computers and Distributed Systems Vol. No.2, Issue 1, December 2012

[27] R. Pratap, K. Suvarna, J. Rama, and D. . Nageswara, “An Efficient Dens Improved K-Medoids Clustering algorithm, “*Int. J. Adv. Comput. Sci. Appl., vol. 2, no. 6, 2011*