Computer Engineering and Applications Vol. 13, No. 3, October 2024

Cluster Analysis Of Obesity Risk Levels Using K-Means And Dbscan Methods

Dite Geovani[1*], Zainal Umari[2], Suci Ramadini[1]

[1]*Master of Computer Science, Faculty of Computer Science, Universitas Sriwijaya*
[2]*Department of Reliability, PT. Pusri Palembang*
**09012682327011@student.unsri.ac.id*

ABSTRACT

Obesity is defined as excessive fat accumulation and abnormal accumulation of adipose tissue in the human body that poses health risks. The causes of obesity are multifactorial and include environmental and individual factors. Several factors that cause obesity include genetic, behavioral and environmental factors. Obesity causes various problems in various fields, including health, employment, demographics, economics and family. The problem of obesity has a significant impact on public health. Therefore, understanding and predicting the level of obesity risk is important in efforts to prevent and treat obesity risk. Data on eating habits, physical activity, and other factors associated with obesity levels in certain populations can provide an important basis for understanding obesity risk. This research clusters the risk of obesity to find hidden patterns in the data. The stages in this research consist of pre-processing, clustering, and analysis. The clustering methods used are K-means and DBSCAN. In clustering using the K-means method with a parameter value of k = 2, results are obtained with the same pattern as clustering using the DBSCAN method with a parameter value of epsilon = 1.4 and a minimum sample = 5. In clustering using the K-means method with a parameter value of k = 4, Four clusters were formed which had different patterns. The clustering results obtained in this research can be used as an effort to prevent and treat the risk of obesity.
.

Keywords: Obesity, Clustering, K-means, DBCAN, risk.

1. INTRODUCTION

Obesity is a disease that causes uncontrolled body weight gain due to low energy expenditure and high calorie intake [1]. According to the World Health Organization (WHO), obesity is defined as excessive fat deposition and abnormal accumulation of adipose tissue in the human body that poses health risks [2], [3]. Individuals who have a Body Mass Index (BMI) greater than 30 are considered obese, while individuals who have a BMI between 25 and 30 are considered overweight [2]. The causes of obesity are multifactorial and include both environmental and individual factors [4], [5]. Several factors that cause obesity include genetic, behavioral, and environmental factors [5], [6]. Genetic factors play a role in the pathogenesis of obesity in approximately 40-70%. It was found that the

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

risk of obesity is 4-5 times higher if one parent is obese [5]. In addition, children of mothers who are obese have an almost 13 times higher incidence of obesity than children of mothers with normal weight [7]. Behavioral factors that cause obesity include poor eating patterns, not paying attention to food nutrition, low physical activity, and lack of sleep [5], [8]. Based on environmental factors, obesity is more common in urban areas than in rural areas. This is because urban areas have more fast food and easier mobility access than villages [5].

Obesity causes various problems in various fields, including health, labor, demographics, economics, and family [2], [3], [9]. In the health sector, obesity increases the risk of chronic disease, cardiovascular disease, various types of cancer, musculoskeletal disorders, metabolic syndrome, diabetes mellitus, and kidney disease. It also increases the inflammatory process and results in adverse vascular changes such as arterial stiffness [1], [2], [3], [4]. The problem of obesity has had a significant impact on public health. Therefore, understanding and predicting the level of risk of obesity is important in efforts to prevent and treat the risk of obesity. Data on eating habits, physical activity, and other factors related to obesity levels in certain populations can be an important basis for understanding the risk of obesity. In an effort to understand and predict the level of risk of obesity, this study carried out a clustering of the level of risk of obesity.

Clustering is one of the data mining techniques that can be used to group data [10]. Clustering is a process of grouping data based on similarity into different clusters [10], [11], [12]. A cluster is a collection of objects or data that are similar or have similarities to each other in the same cluster and are different from data in other clusters [10], [12]. Clustering is an unsupervised method, which means that clustering is an unsupervised method that works on data sets that do not have target or outcome variables, which are often called unlabeled data [13], [14]. Clustering is often used in data analysis to find hidden structures or patterns in large data sets [15], [16].

One of the clustering methods that is widely used is the K-means method [11], [17], [18]. The K-means method has great potential to handle very large data sets [18]. The K-means method divides data into groups, where each group has its own centroid value. Then calculate the Euclidean distance of the data to the centroid of each group. Data that has the closest distance to the centroid will be included in the same group. This process is carried out until all data is divided into k-clusters [17], [18]. Several previous studies have used the K-means method, including clustering of Covid-19 disease [19], clustering of nutritional status [20], and clustering of obesity disease [21].

Apart from the K-means method, a method that can be used for clustering is the Density-Based Spatial Clustering Algorithm with Noise (DBSCAN). In contrast to the K-means method, the DBSCAN method performs grouping based on the minimum points and epsilon values used [22], [23]. In the DBSCAN method, clusters are formed based on data density. The DBSCAN method has the ability to detect clusters with changing shapes, is efficient in detecting noise, and automatically detects the number of clusters that can be formed without determining the number of clusters first [24], [25]. Several previous studies have used the DBSCAN method, including clustering of diabetes [26], clustering of heart disease [27], and clustering of Covid-19 [28].

In this study, a cluster analysis will be carried out on the risk of obesity using the K-means and DBSCAN methods. In the K-means and DBSCAN methods, the parameters to be used have several different values so that several clusters are

formed. For each parameter value in both the K-means and DBSCAN methods, an analysis of the clusters formed will be carried out. This is done to find hidden patterns in clusters formed from data used in an effort to prevent and treat the risk of obesity.

2. MATERIAL AND METHODS

The dataset used in this paper is a collection of estimates of obesity levels from Mexico, Peru, and Colombia based on eating habits and physical condition [29], obtained from survey results. Data contains 17 attributes and 2111 records from ages 14-61 years old with features related to obesity levels, which can be seen in Table 1 [30].

TABLE 1.
Features and Description of Obesity Levels

Features Name	Description	Values
Gender	Gender	Female or Male
Age	Contains data of 14-61 years	Numeric
Height	Contains data of 1.45 – 1.98 meters	Numeric in meters
Weight	Contains data of 39 – 173 Kg	Numeric in kilograms
SWO	Family history of overweight	Yes or No
FAVC	Frequent consumption of high-caloric food	Yes or No
FCVC	Frequency of consumption of vegetables	‘Never’, ‘Sometimes’, ‘Always’
NCP	Number of main meals	‘Between 1-2’, ‘Three’, ‘More than three’
CAEC	Consumption of food between meals	‘No’, ‘Sometimes’, ‘Frequently’, ‘Always’
SMOKE	Have a smoke	Yes or No
CH2O	Consumption of water daily	‘Less than 1 liter’, ‘Between 1 – 2 liters’, ‘More than 2 liters’
SCC	Calorie consumption monitoring	Yes or No
FAF	Physical activity frequency	‘I do not have’, ‘1 or 2 days’, ‘2 or 4 days’, ‘4 or 5 days’
TUE	Time using technological	‘0-2 hours’, ‘3-5 hours’, ‘More than 5 hours’
CALC	Consumption of alcohol	‘No’, ‘Sometimes’, ‘Frequently’, ‘Always’
MTRANS	Transportation used	‘Automobile’, ‘Motorbike’, ‘Bike’, ‘Public transportation’ ‘Walking’
NObeyesdad	Estimate of overweight level	‘Underweight’, ‘Normal weight’, ‘Overweight level I’, ‘Overweight level II’, ‘Obesity type I’, ‘Obesity type II’, ‘Obesity type III’

From Table 1, it can be seen that features related to eating habits are FAVC, FCVC, NPC, CAEC, CH2O, and CALC. Meanwhile, features related to physical condition are SCC, FAF, TUE, and MTRANS. After understanding the data structures, the next step taken in this research is the data processing to produce information using data mining techniques. Data mining is a method used to find relationships and find patterns in large datasets, resulting in information and involving statistical analysis of the data [31],[32].

Data mining itself is closely related to Knowledge Discovery in Database (KDD), which is the overall process of discovering knowledge in data [33],[34]. The stages used in this research can be seen in Figure 1.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

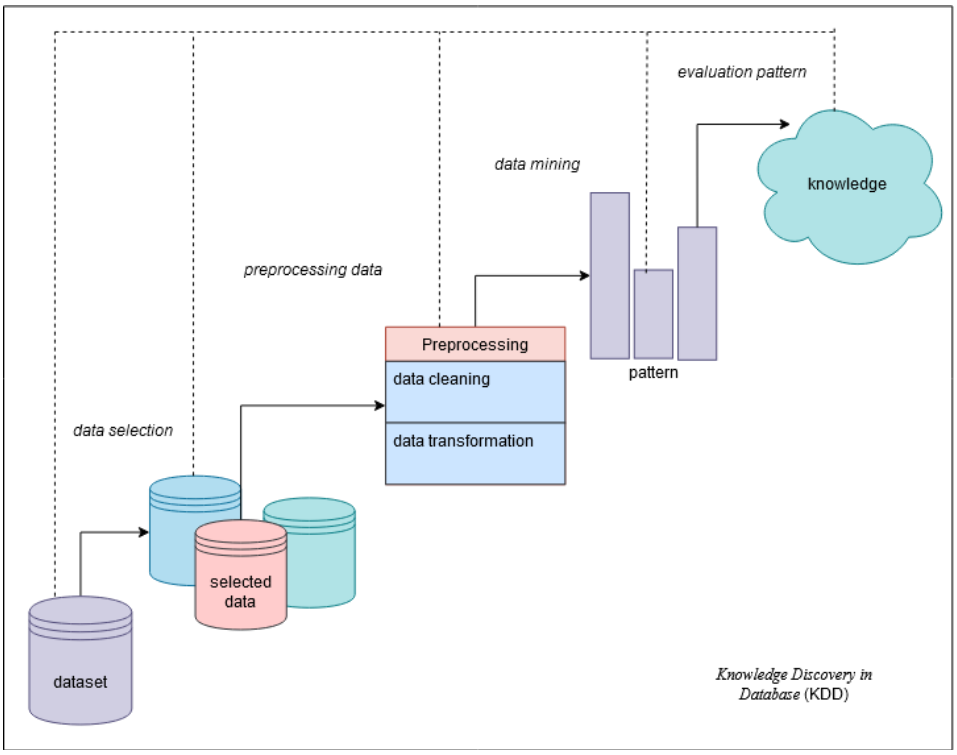

FIGURE 1. Knowledge Discovery in Database (KDD)

There are 5 stages in the KDD process, namely data selection, data cleaning, data transformation, data mining, and the last stage is pattern evaluation. In the application of KDD, each process can be repeated [35], so that the results obtained can be improved and obtain more accurate final results.

In the first stage of KDD, namely data selection, the data that will be used for research is selected [36]. The dataset taken from the UCI repository titled Estimation of Obesity Levels Based on Eating Habits and Physical Condition was prepared, and then its characteristics, data structure, and existing patterns were understood. Then, a portion of the data to be used is selected, ignoring irrelevant data so that the data used is easier to analyze and produces more accurate results.

The second stage, namely data cleaning, involves cleaning the data with the aim of eliminating noise and inconsistent data, such as checking and removing duplicate data if any. Data cleaning is also carried out to handle missing values in the dataset.[35].

The third stage, namely data transformation is performed to change data into the appropriate format, for example changing the format from string to integer, converting nominal data into numeric data, performing normalization, merging data, and transforming data into an encoded form [37].

Then the fourth stage is data mining, the step to extract or discover patterns to produce the required information. At this stage, selecting the appropriate method is based on the objectives to be achieved in the research, such as choosing a model for characterization, classification, clustering, association, or regression [38]. Then, select the algorithm that suits the chosen method. In this research, the data mining method used is clustering with the k-means and DBSCAN methods.

The final stage is pattern evaluation, which is the process of evaluating the patterns produced in the previous stages in the form of visualization to make them

easier to understand. The obtained patterns are then validated to determine their accuracy or to measure the performance of the created model [38].

2. 1 PRE-PROCESSING

The pre-processing stage is a phase conducted to ensure that the data used in the research can be adjusted before being processed, thereby obtaining accurate results. Data pre-processing is a component of the second and third stages of the KDD (Knowledge Discovery in Database), which include data cleaning and data transformation [39].

This study's data cleaning procedures include looking for duplicate data, eliminating outlier data from the age characteristics, and throwing away any data that is not needed, in this case, the height and weight features.

The data transformation process involves normalization using the min-max method to scale the data between 0 and 1 [40]. This includes combining unbalanced categorical data where categories with more than two values are merged into two categories specifically in the features CAEC, CALC, and MTRANS. Data transformation is also performed to convert categorical values into binary and numerical values. One-hot encoding is used to convert the gender category into binary values of 0 and 1 [41]. Label encoding is used to convert categories into numerical values [42], and the features transformed using label encoding are family history with overweight, FAVC, CALC, CAEC, MTRANS, and obesity levels.

2. 2 K-MEANS METHOD

The k-means clustering method involves putting data into groups according to shared criteria, even though each group has unique traits [32]. Based solely on numerical properties, the group in k-means is divided into one or more clusters using distance computation [43]. The following are the steps for using the k-means algorithm [44],[45],[46]:

1. Choose the value of k for the number of clusters to be formed
2. Randomly initialize k values to be the centroids in the initial clusters, using the Euclidean distance formula Equation 1 [47], to calculate the distance of each centroid.
3. Sort each data point according to its closest centroid.
4. Then, update the centroid values in each cluster by recalculating the centroid using the average value of all the data points in the cluster.
5. Until convergence, repeat steps 3 and 4 together.

K-means is affected by the selection of initial centroid values and the suitable number of clusters, choosing the right initial centroids and number of clusters to be grouped is essential to obtaining the correct cluster result.

$$D_e = \sqrt{(x_i - s_i)^2 + (y_i - t_i)^2} \quad (1)$$

where, D_e is the Euclidean distance, i is the number of a certain object, (x, y) is the coordinate of the object, and (s, t) is the coordinate of the centroid.

2.3 METODE DBSCAN

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

The Density-Based Spatial Clustering of Applications with Noise (DBSCAN) method is a clustering method based on density or connectedness, forming clusters where areas of high density are considered clusters, while areas of low density not included in a cluster are considered noise [48],[49]. The smallest number of data points inside an epsilon radius, the greatest distance between two points in a cluster, is used to compute the density in DBSCAN. The minimum quantity of data points, also known as minimum sample, is the bare minimum of points required to establish a cluster. Epsilon and minimum sample are two main parameters in the DBSCAN [50],[51]. The following are the steps for using the DBSCAN algorithm [52],[49]:

1. Randomly determine the initial point (p).
2. Determine or initialize the epsilon and minimum sample parameters.
3. Calculate epsilon or all density/connectedness distance to p using Euclidean distance formula Equation 1.
4. Take all points within the epsilon value, if it is met and if the value is greater than the predefined minimum sample. Then, p is a core point, and the cluster is fulfilled.
5. Then, repeat steps 3 and 4 until all existing points are processed. If p is a border point, meaning it is within a cluster but has fewer neighboring points than minimum sample, then the process continues to the other point.

3. RESULTS AND DISCUSSION

3.1 PRE-PROCESSING

In the dataset, there is categorical data with unbalanced categories, some of which are very extreme. To reduce this imbalance, similar categories will be merged, as shown in Figure 2.

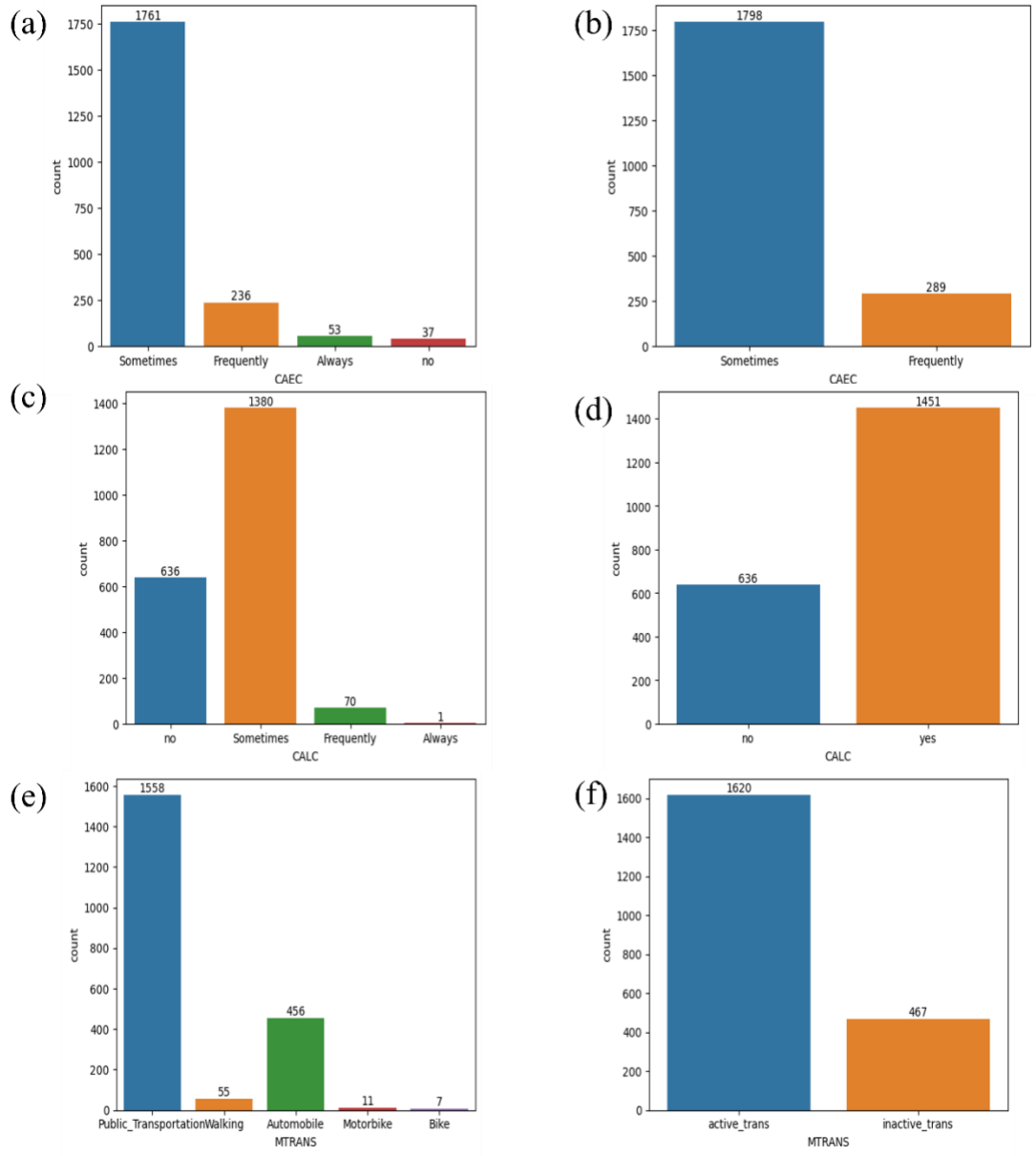

FIGURE 2. Merging Similar Categories
(a) CAEC Before (b) CAEC After (c) CALC Before (d) CALC After (e) MTRANS Before (f) MTRANS After

Based on Figure 1, it can be seen that category merging was carried out on the CAEC, CALC, and MTRANS features. In the CAEC feature, the categories 'frequently', 'always', and 'no' are combined into one category, namely 'frequently', so that in the CAEC feature there are only two categories, namely 'sometimes' and 'frequently'. In the CALC feature, the categories 'sometimes', 'frequently', and 'always' are combined into one category, namely 'yes', so that in the CALC feature there are only two categories, namely 'no' and 'yes'. In the MTRANS feature, the categories 'public_transportation' and 'walking' are combined into one category, namely 'active_trans', and the categories 'automobile', 'motorbike', and 'bike' are combined into one category, namely 'inactive_trans', so that in the MTRANS feature there are only two categories namely 'active_trans' and 'inactive_trans'. For numerical

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

data, outliers will be detected using the Interquartile Range (IQR) method. The results of outlier detection in the Age feature can be seen in Figure 3.

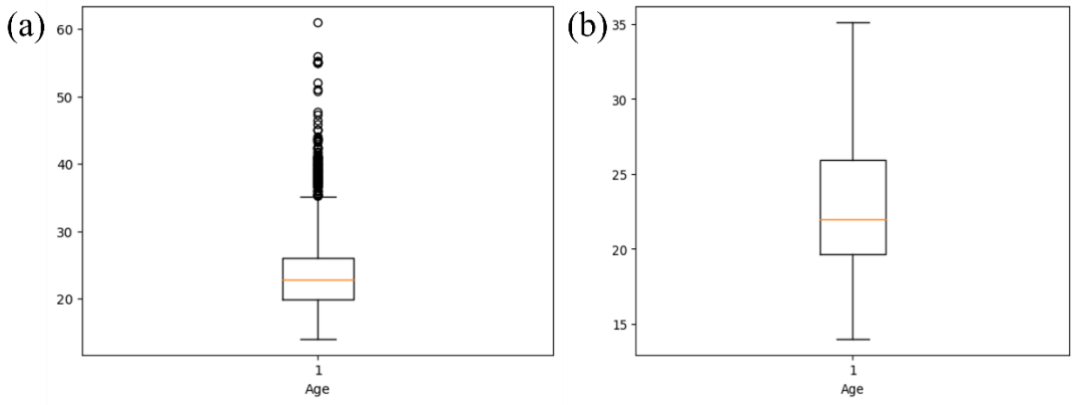

FIGURE 3. Remove Outliers from The Age Data (a) Before (b) After

2.2 CLUSTERING AND ANALYSIS

In the clustering stage, the K-means and DBSCAN methods are used. In the K-means method, the parameters used are the values k = 2 and 4. In the DBSCAN method, the parameters used are the value epsilon = 1.4. In the clustering results using the DBSCAN method with the value epsilon = 1.4 and a minimum sample = 5, two clusters are formed like the results of clustering using the method K-means with a value of k = 2. The cluster results formed using the K-means method with a value of k = 2 and the DBSCAN method with an epsilon value = 1.4 and a minimum sample = 5 have the same pattern.

2.2.1 ANALYSIS OF CLUSTERING RESULTS USING K-MEANS (K = 2) AND DBSCAN (EPSILON = 1.4 AND MINIMUM SAMPLE = 5) METHODS

The first experiment used the K-means (k = 2) algorithm and the DBSCAN algorithm. Coincidentally, the results of the K-means and DBSCAN clusters were exactly the same. The results can be seen in Figure 4.

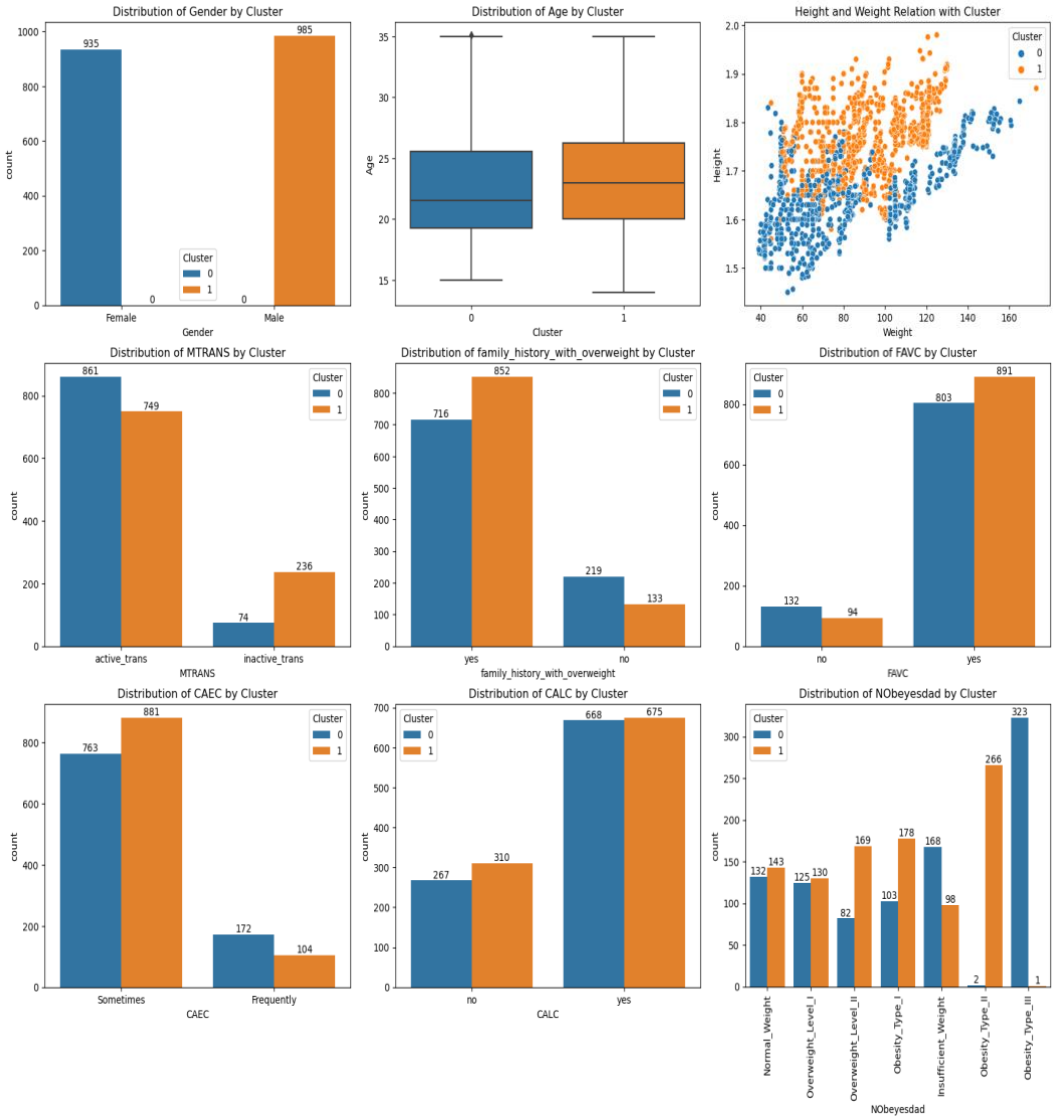

FIGURE 4. Distribution of Two Clusters of Data

In this experiment, two clusters were obtained, which clearly divided the female group with a total of 935 patients in the first cluster, and the male group with a total of 985 patients in the second cluster. More specifically, the first cluster is a group of women who, compared to the second cluster, are slightly younger, shorter, use active transportation more often, have less of a family history of being overweight, consume fewer high-calorie foods, snack less, drink less alcohol, and tend to be underweight or have level 3 obesity. The second cluster is a group of men who, compared to the first cluster, are slightly older, taller, use active transportation less often, more frequently have a family history of being overweight, consume high-calorie foods more often, snack less frequently, drink alcohol more often, and tend to be overweight to the point of level 2 obesity.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

2.2.2 ANALYSIS OF CLUSTERING RESULTS USING THE K-MEANS METHOD (K = 4)

In clustering using the K-means method with parameter value (k = 4), four clusters are formed which have different patterns. The results of clustering using the K-means method with parameter value (k = 4) can be seen in Figure 5.

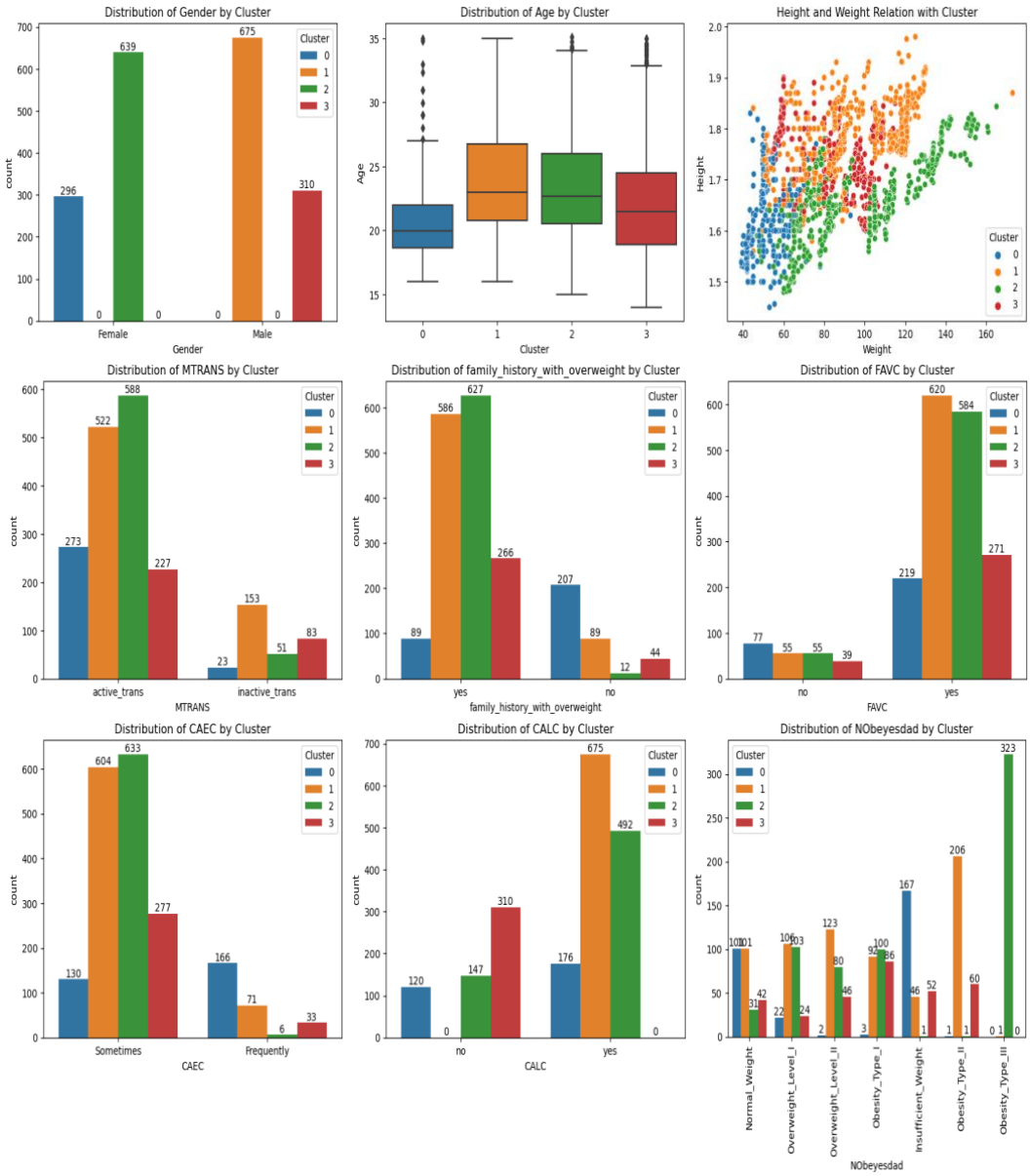

FIGURE 5. Distribution of Four Clusters of Data

In this second experiment, there were 2 male clusters and 2 female clusters. The first cluster is a group of underweight and normal weight women with a total of 296 patients. The second cluster is a group of men who consume alcohol with a total of 679 patients. The third cluster is a group of overweight and obese women with a total of 675 patients. While the fourth cluster is a group of men who do not consume alcohol with a total of 310 patients.

More specifically, the first cluster is a group of women who, compared to the other clusters, have a younger average age, have a lower body weight, often use

active transportation, have less of a family history of being overweight, consume little high-calorie food, snack most often, rarely consume alcohol, and are underweight to normal. The second cluster is a group of men who, compared to the other clusters, have an older age range, have the highest height, tend to use inactive transportation, have a family history of being overweight, most often consume high-calorie foods, rarely snack, consume alcohol, and tend to be overweight to the point of level 2 obesity. The third cluster is a group of women who, compared to the other clusters, tend to be older, have the heaviest body weight, often use active transportation, have the most family history of being overweight, often consume high-calorie foods, snack the least frequently, often consume alcohol, and many have level 3 obesity. The fourth cluster is a group of men who, compared to the other clusters, are quite young, have a balanced body weight and height, often use inactive transportation, tend to have a family history of being overweight, consume few high-calorie foods, tend to snack, do not consume alcohol, and tend to be overweight.

4. CONCLUSION

Based on the research results obtained, the K-means and DBSCAN methods were able to cluster the risk of obesity from the data used. In the clustering results using the K-means method with parameter values k = 2 and k = 4, two and four clusters were formed, respectively. From each cluster formed, hidden patterns in the data were discovered. In the clustering results using the DBSCAN method with parameter value epsilon = 1.4 and minimum sample = 5, two clusters were formed. The two clusters formed as a result of clustering using the DBSCAN method with a parameter value of epsilon = 1.4 and a minimum sample = 5 have the same pattern or data distribution as the results of clustering using the K-means method with a parameter value of k = 2. This shows that the K-means method means and DBSCAN are able to cluster the risk of obesity well. The obesity risk clustering carried out in this study aims to be an effort to prevent and overcome the high risk of obesity.

REFERENCES

[1] F. H. Yagin *et al.*, "Estimation of Obesity Levels with a Trained Neural Network Approach optimized by the Bayesian Technique," *Appl. Sci.*, vol. 13, no. 6, pp. 1–14, 2023, doi: 10.3390/app13063875.

[2] H. G. G. Bag *et al.*, "Estimation of Obesity Levels through the Proposed Predictive Approach Based on Physical Activity and Nutritional Habits," *Diagnostics*, vol. 13, no. 18, pp. 1–16, 2023, doi: 10.3390/diagnostics13182949.

[3] E. Rodríguez, E. Rodríguez, L. Nascimento, A. da Silva, and F. Marins, "Machine Learning Techniques to Predict Overweight or Obesity," in *CEUR Workshop Proceedings*, 2021, pp. 190–204.

[4] M. Manafe, P. K. Chelule, and S. Madiba, "The Perception of Overweight and Obesity among South African Adults: Implications for Intervention Strategies," *Int. J. Environ. Res. Public Health*, vol. 19, no. 19, 2022, doi: 10.3390/ijerph191912335.

[5] A. D. Maślak, M. Kusz, P. Pawluczuk, A. Alzubedi, and P. Polski, "Causes of Overweight and Obesity in Children and Adolescents," *J. Educ. Heal. Sport*,

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

vol. 10, no. 5, pp. 11–18, 2020, doi: 10.12775/jehs.2020.10.05.001.

[6] A. R. Kansra, S. Lakkunarajah, and M. S. Jay, "Childhood and Adolescent Obesity: A Review," *Front. Pediatr.*, vol. 8, no. January, pp. 1–16, 2021, doi: 10.3389/fped.2020.581461.

[7] G. Sobek *et al.*, "Preferences for Sweet and Fatty Taste in Children and Their Mothers in Association with Weight Status," *Int. J. Environ. Res. Public Health*, vol. 17, no. 2, 2020, doi: 10.3390/ijerph17020538.

[8] N. T. Aramburu and M. A. Izaga, "Risk Factors of Overweight/Obesity-Related Lifestyles in University Students: Results from the EHU12/24 Study," *Br. J. Nutr.*, vol. 127, no. 6, pp. 914–926, 2022, doi: 10.1017/S0007114521001483.

[9] D. D. Solomon *et al.*, "Hybrid Majority Voting: Prediction and Classification Model for Obesity," *Diagnostics*, vol. 13, no. 15, pp. 1–15, 2023, doi: 10.3390/diagnostics13152610.

[10] S. P. Tamba, M. D. Batubara, W. Purba, M. Sihombing, V. M. Mulia Siregar, and J. Banjarnahor, "Book data grouping in libraries using the k-means clustering method," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Sep. 2019. doi: 10.1088/1742-6596/1230/1/012074.

[11] T. M. Ghazal *et al.*, "Performances of K-Means Clustering Algorithm with Different Distance Metrics," *Intell. Autom. Soft Comput.*, vol. 30, no. 2, pp. 735–742, 2021, doi: 10.32604/iasc.2021.019067.

[12] A. Alam, M. Muqeem, and S. Ahmad, "Comprehensive Review on Clustering Techniques and Its Application on High Dimensional Data," *Int. J. Comput. Sci. Netw. Secur.*, vol. 21, no. 6, pp. 237–244, 2021, doi: 10.22937/IJCSNS.2021.21.6.31.

[13] H. Alashwal, M. El Halaby, J. J. Crouse, A. Abdalla, and A. A. Moustafa, "The Application of Unsupervised Clustering Methods to Alzheimer's Disease," *Front. Comput. Neurosci.*, vol. 13, no. May, pp. 1–9, 2019, doi: 10.3389/fncom.2019.00031.

[14] H. Iwasawa, T. Ueno, T. Masui, and S. Tajima, "Unsupervised Clustering for Identifying Spatial Inhomogeneity on Local Electronic Structures," *npj Quantum Mater.*, vol. 7, no. 1, 2022, doi: 10.1038/s41535-021-00407-5.

[15] G. Kansal and M. Rawat, "Similarity Measure Based Cluster Generation of Text Documents," *NeuroQuantology*, vol. 20, no. 8, pp. 1008–1016, 2022, doi: 10.14704/nq.2022.20.8.NQ44109.

[16] M. Cendana and R. Kuo, "Categorical Data Clustering : A Bibliometric Analysis and Taxonomy," *Mach. Learn. Knowl. Extr.*, vol. 6, no. 2, pp. 1009–1054, 2024.

[17] N. Sujatha, L. N. Valli, A. Prema, S. Rathiha, and V. Raja, "Initial Centroid Selection for K- Means Clustering Algorithm using The Statistical Method," *Int. J. Sci. Res. Arch.*, vol. 7, no. 2, pp. 474–478, 2022, doi: 10.30574/ijsra.2022.7.2.0309.

[18] I. Ali, A. U. Rehman, D. M. Khan, Z. Khan, M. Shafiq, and J. G. Choi, "Model Selection Using K-Means Clustering Algorithm for the Symmetrical Segmentation of Remote Sensing Datasets," *Symmetry (Basel).*, vol. 14, no. 6, pp. 1–19, 2022, doi: 10.3390/sym14061149.

[19] J. Hutagalung, N. L. W. S. R. Ginantra, G. W. Bhawika, W. G. S. Parwita, A. Wanto, and P. D. Panjaitan, "COVID-19 Cases and Deaths in Southeast Asia Clustering using K-Means Algorithm," *J. Phys. Conf. Ser.*, vol. 1783, no. 1, pp. 0–6, 2021, doi: 10.1088/1742-6596/1783/1/012027.

[20] S. S. Nagari and L. Inayati, "Implementation of Clustering Using K-Means Method To Determine Nutritional Status," *J. Biometrika dan Kependud.*, vol. 9, no. 1, pp. 62–68, 2020, doi: 10.20473/jbk.v9i1.2020.62-68.
[21] E. Setiawati, U. D. Fernanda, and S. Agesti, "Implementation of K-Means , K-Medoid and DBSCAN Algorithms In Obesity Data Clustering," *Indones. J. Appl. Technol. Innov. Sci.*, vol. 1, no. February, pp. 23–29, 2024.
[22] H. S. Ramadan, H. A. Maghawry, M. El-Eleamy, and K. El-Bahnasy, "A Heuristic Novel Approach for Determination of Optimal Epsilon for Dbscan Clustering Algorithm," *J. Theor. Appl. Inf. Technol.*, vol. 100, no. 7, pp. 2243–2253, 2022.
[23] H. T. Nguyen, E. H. Lee, C. H. Bae, and S. Lee, "Multiple Object Detection based on Clustering and Deep Learning ethods," *Sensors (Switzerland)*, vol. 20, no. 16, pp. 1–14, 2020, doi: 10.3390/s20164424.
[24] N. Hanafi and H. Saadatfar, "A fast DBSCAN algorithm for big data based on efficient density calculation," *Expert Syst. Appl.*, vol. 203, p. 117501, Oct. 2022, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417422008302
[25] I. de M. Ventorim, D. Luchi, A. L. Rodrigues, and F. M. Varejão, "BIRCHSCAN: A Sampling Method for Applying DBSCAN to Large Datasets," *Expert Syst. Appl.*, vol. 184, no. July, 2021, doi: 10.1016/j.eswa.2021.115518.
[26] S. Krivtsov, I. Meniailov, and K. Korobchynskyi, "Detection of Patients with Diabetes Mellitus using Density-Based Spatial Clustering of Applications with Noize," in *CEUR Workshop Proceedings*, 2022, pp. 97–103.
[27] P. Dileep, K. N. Rao, P. Bodapati, S. Gokuruboyina, and R. Peddi, "Impact of K-Means and DBSCAN Clustering on Supervised Learning for Heart Disease Prediction," *Turkish J. Comput. Math. Educ.*, vol. 12, no. 9, pp. 1173–1183, 2021.
[28] I. Fitri, R. Asmar, and A. Rubhasy, "Data Cluster Mapping Of Global Covid-19 Pandemic Based on Geo-Location," *J. Mantik*, vol. 4, no. 1, pp. 511–520, 2020.
[29] "Estimation of Obesity Levels Based On Eating Habits and Physical Condition ." 2019.
[30] F. M. Palechor and A. de la H. Manotas, "Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico," *Data Br.*, vol. 25, Aug. 2019.
[31] C. Zai and T. Komputer, "Implementasi Data Mining Sebagai Pengolahan Data," *Portaldata.org*, vol. Vol.2, no. 3, 2022.
[32] G. Abdurrahman, "Clustering Data Ujian Tengah Semester (UTS) Data Mining Menggunakan Algoritma K-Means," vol. Vol 1, No, 2016.
[33] J. Han, M. Kamber, and J. Pei, *Data Mining. Concepts and Techniques, 3rd Edition (The Morgan Kaufmann Series in Data Management Systems)*, Third Edit. Morgan Kaufmann, 2012.
[34] Haris Kurniawan, Sarjon Defit, and Sumijan, "Data Mining Menggunakan Metode K-Means Clustering Untuk Menentukan Besaran Uang Kuliah Tunggal," *J. Appl. Comput. Sci. Technol.*, vol. 1, no. 2, pp. 80–89, Dec. 2020.
[35] J. Amalia, N. Yosevin Nababan, K. G. Tambunan, and I. S. Sinaga, "Decision Tree Dengan Binary Bat Algoruthm Optimization Pada Heart Catheterization

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)

Prediction," *Hexag. J. Tek. dan Sains*, vol. 3, no. 2, pp. 46–51, 2022.
[36] W. Prasetyo Aji, "Analisa Pengelompokkan Data Nilai Rapot Siswa Menggunakan Pendekatan Metode K-Means Di SMK Ponpes Manba'ul Ullum Cirebon," *Kopertip J. Ilm. Manaj. Inform. dan Komput.*, vol. 8, no. 1, pp. 14–18, 2024.
[37] A. Nugraha, O. Nurdiawan, and G. Dwilestari, "Penerapan Data Mining Metode K-Means Clustering Untuk Analisa Penjualan Pada Toko Yana Sport," *JATI (Jurnal Mhs. Tek. Inform.*, vol. 6, no. 2, pp. 849–855, 2022.
[38] A. Agung, A. Daniswara, I. Kadek, and D. Nuryana, "Data Preprocessing Pola Pada Penilaian Mahasiswa Program Profesi Guru," *J. Informatics Comput. Sci.*, vol. 05, pp. 97–100, 2023.
[39] Z. Yamani, S. Nurmaini, and D. P. Rini, "Author Matching Classification with Anomaly Detection Approach for Bibliomethric Repository Data," *Comput. Eng. Appl. J.*, vol. 9, no. 2, pp. 79–92, 2020.
[40] D. A. Nasution, H. H. Khotimah, and N. Chamidah, "Perbandingan Normalisasi Data untuk Klasifikasi Wine Menggunakan Algoritma K-NN," *Comput. Eng. Sci. Syst. J.*, vol. 4, no. 1, p. 78, 2019.
[41] T. Al-shehari and R. A. Alsowail, "An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques," *Entropy*, vol. 23, no. 10, 2021.
[42] M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," *IEEE Access*, vol. 9, pp. 114381–114391, 2021.
[43] M. Benri, H. Metisen, and S. Latipa, "Analisis Clustering Menggunakan Metode K-Means dalam Pengelompokkan Penjualan Produk pada Swalayan Fadhila," *J. Media Infotama*, vol. 11, no. 2, 2015.
[44] A. A. Abdulnassar and L. R. Nair, "Performance analysis of Kmeans with modified initial centroid selection algorithms and developed Kmeans9+ model," *Meas. Sensors*, vol. 25, Feb. 2023.
[45] F. Nurdiyansyah, I. Akbar, A. History, and C. Author, "Jurnal Teknologi dan Manajemen Informatika Implementasi Algoritma K-Means untuk Menentukan Persediaan Barang pada Poultry Shop Article Info ABSTRACT," vol. 7, no. 2, pp. 86–94, 2021.
[46] "Perbandingan Kinerja Algoritma Kmeans dengan Kmeans Median pada Deteksi Kanker Payudara," *J. Inf. dan Teknol.*, vol.5, no. 2, pp. 88–91, 2023.
[47] A. Febiyati Ayutrisula and A. Fanani, "Customer Profiling dengan Menggunakan Metode K-Means Euclidean Distance di BPJS Ketenagakerjaan Tanjung Perak," *J. Mhs. Mat. Algebr.*, vol. 1, no. 1, pp. 157–168, 2020.
[48] N. Arsih, N. Hajarisman, S. Darwis, P. Statistika, and F. Matematika dan Ilmu Pengetahuan Alam, "Metode Pengclusteran Berbasis Densitas Menggunakan Algoritma DBSCAN Methods of Density-Based Clustering Algorithm using," *Pros. Stat.*, vol.2, no. 2, 2016.
[49] D. Pri Indini, S. R. Siburian, and D. Putro Utomo, "Implementasi Algoritma Dbscan untuk Clustering Seleksi Penentuan Mahasiswa yang Berhak Menerima Beasiswa Yayasan," *Pros. Semin. Nas. Sos. Humaniora, dan Teknol.*, no. Senashtek, 2022.
[50] A. Putri, W. Hadi, H. Pratiwi, and I. Slamet, "Pengelompokan Data Gempa Bumi di Indonesia dengan Algoritma K-Means dan DBSCAN," *Semin. Nas. Pendidik. Mat. Ahmad Dahlan*, vol.7, 2023.

[51] M. Mieczynska and I. Czarnowski, "DBSCAN algorithm for AIS data reconstruction," in *Procedia Computer Science*, Elsevier B.V., 2021, pp. 2512–2521.

[52] N. Arsih, N. Hajarisman, S. Darwis, P. Statistika, and F. Matematika dan Ilmu Pengetahuan Alam, "Metode Pengclusteran Berbasis Densitas Menggunakan Algoritma DBSCAN Methods of Density-Based Clustering Algorithm using," *Pros. Stat.*, vol.2, no. 2, 2016.

ISSN: 2252-4274 (Print)
ISSN: 2252-5459 (Online)