Published online on the journal's webpage: **http://journal.ittelkom-pwt.ac.id/index.php/dinda**

Journal of Dinda

Data Science, Information Technology, and Data Analytics

Vol. 4 No. 1 (2024) 14 - 22 E-ISSN: 2809-8064

Prediction of Obesity Classification Using K-Means Clustering

Aditya Wildan[1*], Helmy Akmal Burhansyah[2], Choki Ferdiansyah[3]

[1,2,3*]Software Engineering, Faculty of Informatics, Institut Teknologi Telkom Purwokerto

[1*]21104077@ittelkom-pwt.ac.id, [2]21104079@ittelkom-pwt.ac.id, [3]21104017@ittelkom-pwt.ac.id

Abstract

This paper aims to determine the difference between someone who is obese and who is not and classify the level of obesity by utilizing the K-Means clustering algorithm to group them. The move was taken as part of obesity prevention efforts, with the hope that a deeper understanding of the distribution of obesity within specific categories could help design more specific and effective interventions. Using this approach, it is hoped that this study can contribute to our understanding of the complexities of obesity and encourage more precise and targeted preventive measures. In this study we used datasets from Kaggle. It is used to classify the difference between underweight and overweight people. In this study, data was processed using Data Mining techniques with the K-Means method. Based on the classification, four clusters were categorized. Cluster 0 in this cluster only has women, with an age range ranging from 45 to 60 years. Relatively thin to normal weight. Cluster 1 only has men, with an age range of more than 40 years and 55 to 60 years. People in this cluster are overweight or obese. Cluster 2 women aged 15-70 years make up the majority in this group, with women aged 55-60 years as the highest proportion. In general, they have a normal weight. Many underweight individuals aged 10-45 years, with the highest proportion at the age of 20-25 years. The classification results show that men have a higher likelihood of suffering from obesity than women. Therefore, obesity prevention needs to be done, one of which is by applying a healthy lifestyle.

Keywords: BMI, Prediction, Obesity, K-Means

© 2024 Journal of DINDA

1. Introduction

Obesity is caused by excessive accumulation of fat tissue in the body. This happens because of the imbalance between incoming energy and outgoing energy [1]. Obesity itself is an increasingly serious health problem around the world. Even the World Health Organization (WHO) has declared that obesity is a global epidemic. Therefore, it is a health problem that must be addressed immediately [2]. The Ministry of Health of the Republic of Indonesia (Kemkes RI) announced that the number of obese people worldwide has more than doubled since 1980. More than 1.9 billion adults over the age of 18 are obese [3].

A commonly used method to measure obesity rates is body mass index (BMI), which is calculated by dividing body weight (kg) by height (meters) [4]. A person with a BMI of 30 or higher is generally considered obese. People with a BMI of 25 or higher are considered overweight [5]. Obesity is one of the causative factors of various degenerative diseases such as heart disease and stroke. This disease is the leading cause of death of the world's population, especially in the elderly group. In addition, obesity increases the risk of bone and joint damage [6]. In addition, the increasing prevalence of global obesity will also increase mortality and morbidity at a relatively young age. The consequences of this obesity epidemic will have an impact on the degree of Public Health [7].

From the prevalence and consequences of the obesity epidemic, immediate treatment is needed. One of them is to take a computational approach to make predictions between someone who is obese and who is not [8]. Where someone can check the level of obesity independently. If the results of independent examination are found symptoms of diseases such as obesity, then the community can continue the examination to the doctor [9].

One of them uses grouping techniques with the clustering method. Clustering is a method of finding and grouping data with similar characteristics among each data set [10]. The group or cluster obtained becomes useful information for the user's decision-making process [11].

Received: 08-01-2024 | Revised: 15-02-2024 | Published: 15-02-2024

K-Means is one of the clustering algorithms. K-Means cluster analysis is a non-hierarchical cluster analysis method that seeks to divide existing objects into one or more clusters or groups of objects based on their characteristics, so that objects that have the same characteristics are grouped in the same cluster and similar objects have different characteristics will be grouped into other clusters [12]. A few previous studies have consistently demonstrated the use of K-Means clustering methods in various contexts. Research by [13], using K-Means clustering to group the spread of diarrhea. Meanwhile research by [14], illustrates the success of this method in the classification of nutritional value of toddlers. Such positive findings demonstrate the adaptive ability of K-Means clustering in handling different types of data.

Among them are Li Li, Qifa Song, and Xi Yang (2020) conducted two studies related to obesity. The first study used K-means clustering and quantile transformations in metabolic data to classify overweight and obese individuals. The results show hope for accurate handling. The second study, also by Li Li, Qifa Song, and Xi Yang, focused on classifying obese patients based on insulin resistance and insulin release by K-means clustering method using HOMA-IR and IGI 30 minutes [15]-[16].

Furthermore, Vervoort, T. Naets, L. Goossens, S. Verbeken, L. Claes, A. Tanghe, and C. Braet (2022) examined the relationship of emotional eating and obesity in children. The results showed that emotional eating was associated with higher body weight and poor weight loss outcomes [17]. I Made Satria Bimantara and I Wayan Supriana (2022) in their research used Euclidean distance and K-means indexing methods in machine learning and data mining to estimate obesity rates with a Case Based Reasoning (CBR) approach [3]. Other research conducted by M. G. Ahamad, M. F. Ahmed, and M. Y. Uddin (2016) utilizing Weka in exploration for automatic classification, regression, and feature selection in bioinformatics. Their research deals with the analysis of risk factors for diabetes, hypertension, and obesity [18].

Research by Pebriani, Frethernety, and Trinovita (2022) found a positive relationship between junk food consumption and obesity through a systematic review of 15 journals. Mahdiah [19], Hadi, and Susetyowati (2004) highlight the change in diet to junk food because of technological developments, supporting the need for prevention of adolescent obesity in cities and villages [20]. Ni Putu Lia Juliantini and I Gusti Lanang Sidiartha (2014) confirmed the relationship between parental obesity and child obesity [21]. Sesilia Effendy, Gunawan, Argoputra, Anggraeni, Abraham, and Fenty (2018) found the complexity of factors affecting obesity, without a significant relationship between physical activity and obesity at the study site [22]. Mulyani, Ngo, and Yudia (2021) highlight the risk of complications in obese pregnant women, emphasizing the need for obesity monitoring and management during pregnancy [23].

This paper aims to determine the difference between someone who is obese and who is not and classify the level of obesity by utilizing the K-Means clustering algorithm to group them. The move was taken as part of obesity prevention efforts, with the hope that a deeper understanding of the distribution of obesity within specific categories could help design more specific and effective interventions. Using this approach, it is hoped that this study can contribute to our understanding of the complexities of obesity and encourage more precise and targeted preventive measures.

2. Research Methods

In this study we used datasets from kaggle. It is used to classify the difference between underweight and overweight people. In this study, data was processed using Data Mining techniques with the K-Means method.

2.1. *Dataset*

The dataset used in this study was sourced from kaggle. This dataset consists of 110 data containing information ranging from id, age, gender, height, weight, BMI (Body Mass Index), and obesity classification labels which can be normal weight, underweight, or overweight.

Table 1. Dataset Obesity

ID	Age	Gender	Height	Weight	BMI	Label
1	25	Male	175	80	25.3	Normal Weight
2	30	Female	160	60	22.5	Normal Weight
3	35	Male	180	90	27.3	Overweigh t
4	40	Female	150	50	20	Underweig ht
5	45	Male	190	100	31.2	Obese
6	50	Female	140	40	16.7	Underweig ht
7	55	Male	200	110	34.2	Obese
8	60	Female	130	30	13.3	Underweig ht

9	65	Male	210	120	37.2	Obese
10	70	Female	120	20	10	Underweight
11	18	Male	175	70	23.4	Normal Weight
12	23	Female	160	50	20	Underweight
13	28	Male	180	80	25.3	Normal Weight
14	33	Female	150	60	22.5	Normal Weight
15	38	Male	190	90	27.3	Overweight
16	43	Female	140	50	20	Underweight
17	48	Male	200	100	31.2	Obese
19	53	Female	130	40	16.7	Underweight
20	58	Male	210	110	34.2	Obese
21	63	Female	120	30	13.3	Underweight
22	68	Male	175	80	25.3	Normal Weight
23	73	Female	160	60	22.5	Normal Weight
24	78	Male	180	90	27.3	Overweight
25	83	Female	150	50	20	Underweight
26	88	Male	190	100	31.2	Obese
27	93	Female	140	40	16.7	Underweight
28	98	Male	200	110	34.2	Obese
29	103	Female	130	30	13.3	Underweight
30	108	Male	210	120	37.2	Obese
31	19	Male	175	75	24.2	Normal Weight
32	24	Female	160	55	21.2	Normal Weight
33	29	Male	180	85	26.1	Overweight
35	39	Male	190	95	27	Overweight
36	44	Female	140	75	25	Overweight
37	49	Male	200	105	28.9	Obese
38	54	Female	130	85	27.5	Overweight
39	59	Male	210	115	30.8	Obese
40	64	Female	120	95	29.1	Overweight
41	17	Male	175	65	22.7	Normal Weight
42	22	Female	160	45	18.7	Underweight
43	27	Male	180	75	24.2	Normal Weight
44	32	Female	150	55	21.2	Normal Weight
45	37	Male	190	85	26.1	Overweight
46	42	Female	140	65	22.7	Normal Weight
47	47	Male	200	95	27	Overweight
48	52	Female	130	75	25	Overweight
49	57	Male	210	105	28.9	Obese
50	62	Female	120	85	27.5	Overweight

51	67	Male	175	65	22.7	Normal Weight
52	72	Female	160	45	18.7	Underweight
53	77	Male	180	75	24.2	Normal Weight
54	82	Female	150	55	21.2	Normal Weight
55	87	Male	190	85	26.1	Overweight
56	92	Female	140	65	22.7	Normal Weight
57	97	Male	200	95	27	Overweight
58	102	Female	130	75	25	Overweight
59	107	Male	210	105	28.9	Obese
60	112	Female	120	85	27.5	Overweight
61	16	Male	175	55	21.2	Normal Weight
62	21	Female	160	35	16.7	Underweight
63	26	Male	180	65	22.7	Normal Weight
64	31	Female	150	45	18.7	Underweight
65	36	Male	190	75	24.2	Normal Weight
66	41	Female	140	55	21.2	Normal Weight
67	46	Male	200	85	26.1	Overweight
68	51	Female	130	65	22.7	Normal Weight

69	56	Male	210	95	27	Overweight
70	61	Female	120	75	25	Overweight
71	15	Male	175	45	18.7	Underweight
72	20	Female	160	30	13.3	Underweight
73	25	Male	180	55	21.2	Normal Weight
74	30	Female	150	40	16.7	Underweight
75	35	Male	190	65	22.7	Normal Weight
76	40	Female	140	50	20	Underweight
77	45	Male	200	75	24.2	Normal Weight
78	50	Female	130	60	22.5	Normal Weight
79	55	Male	210	85	26.1	Overweight
80	60	Female	120	70	23.4	Normal Weight
81	14	Male	175	35	16.7	Underweight
82	19	Female	160	25	10	Underweight
83	24	Male	180	45	18.7	Underweight
84	29	Female	150	30	13.3	Underweight
85	34	Male	190	55	21.2	Normal Weight
86	13	Male	175	25	10	Underweight

87	18	Female	160	20	8.3	Underweight
88	23	Male	180	30	13.3	Underweight
89	28	Female	150	25	10	Underweight
90	33	Male	190	40	16.7	Underweight
91	38	Female	140	35	16.7	Underweight
92	43	Male	200	50	20	Underweight
93	48	Female	130	40	16.7	Underweight
94	53	Male	210	55	21.2	Normal Weight
95	58	Female	120	35	16.7	Underweight
96	12	Male	175	15	5.6	Underweight
97	17	Female	160	15	5.6	Underweight
98	22	Male	180	20	8.3	Underweight
99	27	Female	150	20	8.3	Underweight
100	32	Male	190	25	10	Underweight
101	37	Female	140	25	10	Underweight
102	42	Male	200	30	13.3	Underweight
103	47	Female	130	30	13.3	Underweight
104	52	Male	210	35	16.7	Underweight
105	57	Female	120	25	10	Underweight
106	11	Male	175	10	3.9	Underweight
107	16	Female	160	10	3.9	Underweight
108	21	Male	180	15	5.6	Underweight
109	26	Female	150	15	5.6	Underweight
110	31	Male	190	20	8.3	Underweight

2.2. *Clustering*

Cluster is a collection of data objects that have similarities between one another in the same group and different objects with other group data [14]. While clustering is basically a method to find and group data that has similar characteristics (similarity) between one data with another data. Clustering is one of the data mining methods that is unsupervised, meaning that this method is applied without training (taining) and without a teacher (teacher) and does not require an output target [10].

2.3. *K-Means Clustering*

K-Means is a method of analyzing data in Data Mining where the modeling process is without supervision and is one method that groups data in partitions. This method minimizes differences between data in one cluster and maximizes differences with other clusters.

There are several characteristics in the K-Means method, including:

1) K-Means is a simple grouping method that can be used easily.
2) In certain types of data sets, K-Means cannot segment data properly where the segmentation results cannot determine group patterns that represent the characteristics of the natural form of the data.
3) K-Means can run into problems when grouping data that contains outliers.

In general, the K-Means method uses the following algorithm:

1) Specify k as the number of clusters in the form. The determination of the number of k clusters is carried out by several factors such as theoretical and conceptual considerations proposed to determine how many clusters.
2) Randomly generate the initial Centroid k (cluster center point). To determine the initial centroid is done randomly from several objects available as many as k clusters, to calculate the next i-th cluster centroid, using the following formula:

$$v = \frac{\sum_{i=1}^{n} Xi}{n} \quad ; i = 1, 2, 3, \dots n \qquad (1)$$

Where:

v = centroid on cluster
Xi = i-th object
N = The number of objects or the number of objects that are members of the cluster

Calculate the distance of each object to each centroid of each cluster. Then calculate the distance between objects with centroids, in this study using Euclidean Distance.

$$d(x,y) = ||x - y|| = \sqrt{\sum_{i=1}^{n} (Xi - Yi)^2} \qquad (2)$$

Where:

Xi = i-th x object
Yi = Data i-th
N = The number of objects

3) Allocate each object into the closest centroid. Iterate, then determine the position of the new centroid using the equation.
4) Repeat step 3 if the new centroid positions are not the same.

The point merging process is done by comparing the task set matrix in the previous iteration with the task set matrix in the current iteration. If the results are the same then the k-means cluster analysis algorithm is already convergent, but if it is different than it has not converged so it needs to be done the next iteration.

3. Results and Discussion

We used 110 obesity data from kaggle. This data contains information ranging from id, age, sex, height, weight, BMI (Body Mass Index), and obesity classification labels which can be normal weight, underweight, or overweight. The purpose of this study was to find differences and classify into clusters from those who were overweight to those who were not.

From the data in table 1, several analyses of the data will be carried out.

Figure 1. Distribution of Age

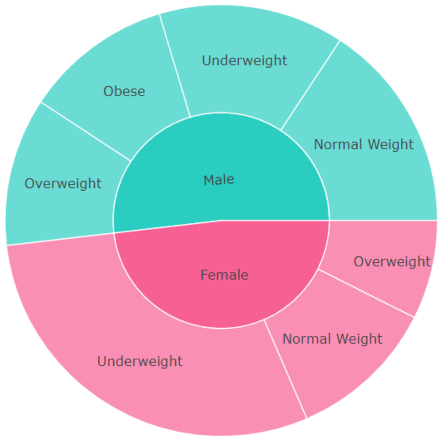

Figure 2. Distribution of Sex

The age distribution of the study in figure 1 shows that the age range ranges from 10 - 60 years. Where the ratio between men and women is almost equal or almost half. Figure 2 also showed that more than half of the women were underweight and none were obese.

The application of the K-Means algorithm successfully divided the study participants into four clusters based on certain characteristics, such as age, gender, and BMI values. This division provides deeper insight into patterns that might influence prediction of obesity classifications, helping to be more specific in clustering approaches.

Table 1. Clustering

n_cluster	Correctly labeled	accuracy
3	31	0,29
4	56	0,52
5	16	0,15

Based on table 2, the number of cluster 3 is 31 correctly labeled data where the accuracy is 0.29. While in the number of clusters equal to 5, there are only 16 data with an accuracy level of 0.15. Then the number of cluster 4 is the highest level of accuracy with labeled data is 56 data.

Therefore, in this method 4 clusters are used, as shown in figure 3.

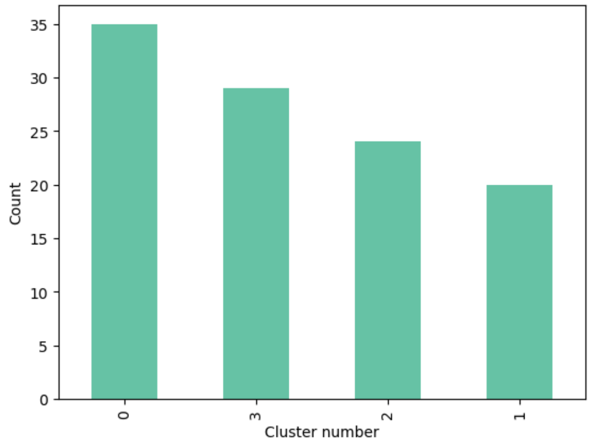

Figure 3. N-Cluster

Furthermore, classification was carried out based on the 4 clusters with data in the form of gender, age, and obesity level labels. Based on figures 4, 5, and 6, it is found that the four clusters are categorized, as follows:

1) Cluster 0
In this cluster, there are only women, with an age range ranging from 45 to 60 years. Relatively thin to normal weight.
2) Cluster 1
In this cluster, there are only men, with an age range of more than 40 years and 55 to 60 years. People in this cluster are overweight or obese.
3) Cluster 2
Women aged 15-70 years make up the majority in this group, with women aged 55-60 years as the highest proportion. In general, they have a normal weight.
4) Cluster 3
Most underweight individuals in the study were aged 10-45 years, with the highest proportion at the age of 20-25 years.

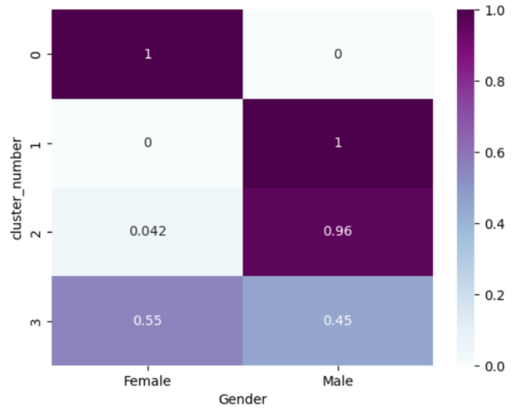

Figure 4. Classification Based on Gender

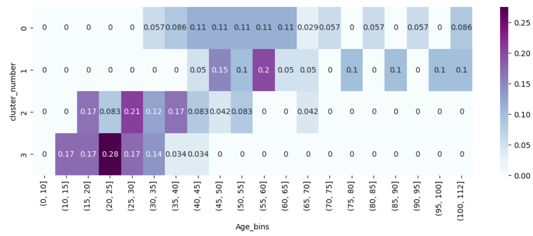

Figure 5. Classification Based on Age

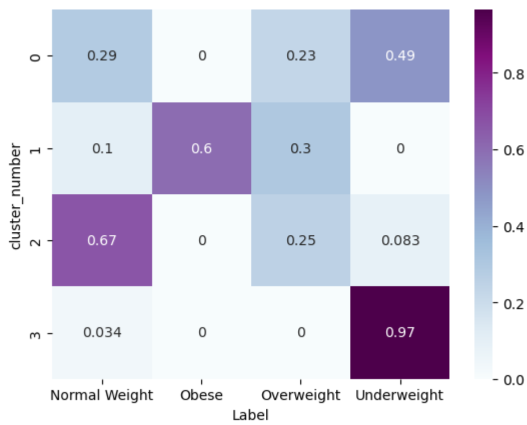

Figure 6. Classification Based on Label

4. Conclusion

The results of this study suggest that this approach may contribute to our understanding of the complexities of obesity and encourage more precise and targeted preventive measures. Using the K-Means clustering algorithm, the study was successful in grouping obesity data into several similar groups, which could aid in medical decision-making and obesity prevention.

Where this study found that obesity levels can be classified into four clusters based on three parameters, namely gender, age, and obesity level labels. The classification results show that men have a higher likelihood of suffering from obesity than women. Therefore, obesity prevention needs to be done, one of which is by applying a healthy lifestyle.

References

[1] S. K. Saraswati *et al.*, "Literature Review : Faktor Risiko Penyebab Obesitas," *Media Kesehat. Masy. Indones.*, vol. 20, no. 1, pp. 70–74, 2021, doi: 10.14710/mkmi.20.1.70-74.

[2] M. Mauliza, "Obesitas Dan Pengaruhnya Terhadap Kardiovaskular," *AVERROUS J. Kedokt. dan Kesehat. Malikussaleh*, vol. 4, no. 2, p. 89, 2018, doi: 10.29103/averrous.v4i2.1040.

[3] I. M. S. Bimantara and I. W. Supriana, "Case Based Reasoning (Cbr) for Obesity Level Estimation Using K-Means Indexing Method," *J. Ilm. Kursor*, vol. 11, no. 4, pp. 145–156, 2023, [Online]. Available: http://www.kursorjournal.org/index.php/kursor/article/view/268%0Ahttps://www.kursorjournal.org/index.php/kursor/article/download/268/140

[4] H. Harahap, Y. Widodo, and S. Mulyati, "Penggunaan Berbagai Cut-Off Indeks Massa Tubuh Sebagai Indikator Obesitas Terkait Penyakit Degeneratif Di Indonesia," *Gizi Indones.*, vol. 28, no. 2, 2014, doi: 10.36457/gizindo.v28i2.20.

[5] H. Ariyanti and D. I. Angraini, "Penatalaksanaan Holistik Obesitas di Puskesmas Rawat Inap Kemiling," *Majority*, vol. 7, no. 3, pp. 191–196, 2018, [Online]. Available: http://juke.kedokteran.unila.ac.id/index.php/majority/article/view/2075/2043

[6] E. Supartini, P. F. Rahmah, F. F. Rahmadanti, M. Antikasari, and R. S. Pontoh, "ANALYSIS OF OBESITY RATES ON CALORIE CONSUMPTION OF SOME FOODS IN 40 ASIAN COUNTRIES Obesity occurs when the body becomes obese caused by a buildup of adipose (adipocytes , i . e ., particular fat tissue stored by the body), which is excessive and cause," pp. 1–21, 2022.

[7] M. Masrul, "Epidemi obesitas dan dampaknya terhadap status kesehatan masyarakat serta sosial ekonomi bangsa," *Maj. Kedokt. Andalas*, vol. 41, no. 3, p. 152, 2018, doi: 10.25077/mka.v41.i3.p152-162.2018.

[8] R. C. Cervantes and U. M. Palacio, "Estimation of obesity levels based on computational intelligence," *Informatics Med. Unlocked*, vol. 21, no. November, 2020, doi: 10.1016/j.imu.2020.100472.

[9] S. Y. Sibi, A. R. Widiarti, and U. S. Dharma, "Klasifikasi Tingkat Obesitas mempergunakan Algoritma KNN," *Semin. Nas. Corisindo*, vol. 7, no. 2, pp. 370–375, 2022.

[10] J. O. Ong, "Implementasi Algotritma K-means clustering untuk menentukan strategi marketing president university," *J. Ilm. Tek. Ind.*, vol. vol.12, no, no. juni, pp. 10–20, 2013.

[11] A. Nur Khormarudin, "Teknik Data Mining: Algoritma K-Means Clustering," *J. Ilmu Komput.*, pp. 1–12, 2016, [Online]. Available: https://ilmukomputer.org/category/datamining/

[12] Ediyanto, M. N. Mara, and N. Satyahadewi, "Pengklasifikasian karakteristik dengan metode k-means cluster analysis," *Bul. Ilm. Mat. Stat. dan Ter.*, vol. 02, no. 2, pp. 133–136, 2013.

[13] F. Nasari and C. J. M. Sianturi, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokkan Penyebaran Diare Di Kabupaten Langkat," *CogITo Smart J.*, vol. 2, no. 2, pp. 108–119, 2016, doi: 10.31154/cogito.v2i2.19.108-119.

[14] Ramadhan, Nur Ghaniaviyanto. "Indonesian online news topics classification using Word2Vec and K-nearest neighbor." *Jurnal RESTI (Rekayasa Sistem Dan Teknologi Informasi)* 5.6 (2021): 1083-1089.

[15] L. Li, Q. Song, and X. Yang, "K-means clustering of overweight and obese population using quantile-transformed metabolic data," *Diabetes, Metab. Syndr. Obes.*, vol. 12, pp. 1573–1582, 2019, doi: 10.2147/DMSO.S206640.

[16] L. Li, Q. Song, and X. Yang, "Categorization of β-cell capacity in patients with obesity via ogtt using k-means clustering," *Endocrine Connections*, vol. 9, no. 2. pp. 135–143, 2020. doi: 10.1530/EC-19-0476.

[17] L. Vervoort *et al.*, "Subtyping youngsters with obesity: A theory-based cluster analysis," *Appetite*, vol. 168, no. August 2020, p. 105723, 2022, doi: 10.1016/j.appet.2021.105723.

[18] M. G. Ahamad, M. F. Ahmed, and M. Y. Uddin,

"Clustering as Data Mining Technique in Risk Factors Analysis of Diabetes, Hypertension and Obesity.," *Eur. J. Eng. Technol. Res.*, vol. 1, no. 6, pp. 88–93, 2018, doi: 10.24018/ejeng.2016.1.6.202.

[19] L. Pebriani, A. Frethernety, and E. Trinovita, "Studi Literatur: Pengaruh Konsumsi Junk Food terhadap Obesitas," *J. Surya Med.*, vol. 8, no. 2, pp. 270–280, 2022, doi: 10.33084/jsm.v8i2.3103.

[20] Mahdiah, H. Hadi, and Susetyowati, "Prevalensi obesitas dan hubungan konsumsi fast food dengan kejadian obesitas pada remaja SLTP kota dan desa di Daerah Istimewa Yogyakarta," *JURNAL GIZI KLINIK INDONESIA*, vol. 1. pp. 69–77, 2004. [Online]. Available: https://jurnal.ugm.ac.id/jgki/article/view/17397/11323

[21] I. G. L. Sidiartha and N. P. L. Juliantini, "Hubungan Riwayat Obesitas Pada Orangtua Dengan Kejadian Obesitas Pada Anak Sekolah Dasar," *E-Jurnal Med. Udayana*, vol. 3, no. 12, pp. 1–13, 2018.

[22] S. Effendy, M. F. Gunawan, D. Lintang, A. Argoputra, P. D. Anggraeni, and Y. B. Abraham, "the Relationship Between Physical Activity and Obesity Based on Body Fat Percentage in Banjaroyo Village," *Int. Phys. Act. J. Farm. Sains dan Komunitas*, vol. 15, no. 1, pp. 29–36, 2018, [Online]. Available: http://dx.doi.org/10.24071/jpsc.151963.

[23] L. Mulyani, N. F. Ngo, and R. C. P. Yudia, "Hubungan Obesitas dengan Komplikasi Maternal dan Luaran Perinatal," *J. Sains dan Kesehat.*, vol. 3, no. 2, pp. 343–350, 2021, doi: 10.25026/jsk.v3i2.483.

[24] W. Van Der Aalst, *Process Mining Data Science In Action*. Springer Heidelberg New York Dordrecht London, 2016.