Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 863−872
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7462

Data Mining Dalam Clusterisasi Risiko Tinggi Obesitas Menggunakan Metode K-Means Clustering

Anzila Hasby*, Budianto Bangun, Masrizal Masrizal

Program Studi Sistem Informasi, Fakultas Sains dan Teknologi, Universitas Labuhanbatu, Rantauprapat, Indonesia
Email: [1,*]anzilahasby@gmail.com, [2]budiantobangun44@gmail.com, [3]masrizal120405@gmail.com
Email Penulis Korespondensi: anzilahasby@gmail.com
Submitted: **28/05/2025**; Accepted: **30/06/2025**; Published**: 30/06/2025**

Abstrak–Obesitas adalah kondisi kelebihan lemak tubuh akibat ketidakseimbangan antara asupan dan penggunaan kalori. Masalah ini telah menjadi epidemi global, termasuk di Indonesia, dengan dampak serius pada kesehatan fisik, mental, dan sosial. Perempuan lebih rentan mengalami obesitas karena faktor biologis dan gaya hidup, seperti terlihat dalam data sebuah puskesmas di mana 76,6% penderita obesitas sentral adalah perempuan. sehingga penelitian ini mengembangkan model segmentasi risiko obesitas pada perempuan menggunakan algoritma K-Means Clustering berbasis data sekunder dari Kaggle (n=898) dengan variabel usia, riwayat keluarga, pola konsumsi, aktivitas fisik, hingga mode transportasi yang digunakan. Hasil preprocessing dan normalisasi StandardScaler menunjukkan 2 cluster optimal (Silhouette Score: 0.267), di mana Cluster 1 (usia muda 24.53 tahun, riwayat keluarga obesitas 1.91, konsumsi fast food 1.84, aktivitas fisik rendah 2.71) berisiko lebih tinggi dibandingkan Cluster 0 (usia 41.41 tahun dengan pola hidup lebih sehat), mengungkap interaksi signifikan antara faktor genetik dan gaya hidup sebagai pemicu utama. Temuan ini menyediakan dasar ilmiah untuk intervensi berbasis kelompok, seperti program edukasi gizi terfokus bagi populasi usia muda, sekaligus mendemonstrasikan efektivitas pendekatan data mining dalam kesehatan masyarakat untuk klasifikasi risiko penyakit tidak menular.

Kata Kunci: Obesitas; K-Means Clustering; Analisis Cluster; Perempuan; Faktor Risiko

Abstract– Obesity is a condition of excess body fat due to an imbalance between calorie intake and expenditure. This problem has become a global epidemic, including in Indonesia, with serious impacts on physical, mental, and social health. Women are more susceptible to obesity due to biological factors and lifestyle choices, as evidenced by data from a community health centre where 76.6% of central obesity patients were women. This study developed an obesity risk segmentation model for women using the K-Means Clustering algorithm based on secondary data from Kaggle (n=898), incorporating variables such as age, family history, dietary patterns, physical activity levels, and mode of transportation used. The results of preprocessing and StandardScaler normalisation showed two optimal clusters (Silhouette Score: 0.267), where Cluster 1 (young age 24.53 years, family history of obesity 1.91, fast food consumption 1.84, low physical activity 2.71) has a higher risk compared to Cluster 0 (age 41.41 years with a healthier lifestyle), revealing a significant interaction between genetic factors and lifestyle as the main triggers. These findings provide a scientific basis for group-based interventions, such as targeted nutrition education programmes for the young population, while demonstrating the effectiveness of data mining approaches in public health for classifying the risk of non-communicable diseases.

Keywords: Obesity; K-Means Clustering; Cluster Analysis; Women; Risk Factors

1. PENDAHULUAN

Obesitas merupakan suatu kondisi kesehatan di mana terdapat penumpukan lemak berlebih di dalam tubuh. Hal ini terjadi ketika asupan kalori melebihi jumlah kalori yang digunakan oleh tubuh. Akibatnya, lemak tersebut tersimpan di berbagai area tubuh seperti perut, paha, dan lengan. Obesitas telah menjadi wabah global yang serius, tak terkecuali di Indonesia. Kondisi ini tidak hanya mengancam kesehatan fisik, tetapi juga memicu masalah psikologis dan sosial. Perempuan, khususnya, lebih rentan mengalami obesitas akibat kombinasi faktor biologis dan gaya hidup. Dari 156 orang yang memeriksakan dirinya ke sebuah puskesmas, sebanyak 71,20% yang mengalami obesitas sentral, dimana 76,60% adalah perempuan sedangkan laki-laki hanya 31,6%[1]. Perempuan lebih dominan terkena obesitas sehingga pada penelitian ini akan berfokus terhadap dataset perempuan yang terkena obesitas.

Berdasarkan permasalahan diatas, maka diperlukan identifikasi dini individu yang rentan terhadap obesitas untuk mencegah terjadinya komplikasi kesehatan yang lebih serius. Dengan mengelompokkan individu berdasarkan karakteristik risiko yang sama, kita dapat memberikan intervensi yang lebih efektif dan efisien. Untuk memahami obesitas secara lebih mendalam, kita dapat memanfaatkan teknik data mining. Metode ini memungkinkan kita untuk menganalisas data dengan mengidentifikasi tren, pola, dan faktor-faktor yang berkontribusi terhadap terjadinya obesitas dari kumpulan data yang luas.

Dalam analisis data, K-Means Clustering merupakan teknik yang sering digunakan untuk mengelompokkan data menjadi beberapa cluster berdasarkan kemiripan karakteristik[2]–[5]. Dalam konteks penelitian obesitas, metode ini dapat diaplikasikan untuk mengidentifikasi subkelompok wanita yang memiliki risiko tinggi mengalami obesitas dengan mempertimbangkan variabel-variabel seperti usia, indeks massa tubuh, pola konsumsi makanan, tingkat aktivitas fisik, dan faktor genetik.

K-Means menawarkan kemudahan dalam implementasi dan efisiensi komputasi yang tinggi, menjadikannya pilihan populer untuk analisis data[6]. Kemampuannya dalam menangani dataset besar dan kompleks, serta hasil yang umumnya baik, menjadikan K-Means sebagai alat yang sangat berguna dalam berbagai bidang, termasuk penelitian obesitas.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 863−872
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7462

Penelitian yang dilakukan pada tahun 2024 yaitu Cici Emilia Sukmawati, dkk[7]. Penelitian ini mengevaluasi kinerja algoritma AdaBoost dan XGBoost dalam klasifikasi penyakit obesitas menggunakan dataset yang diperoleh dari Kaggle. Setelah dilakukan preprocessing data, model dibangun dan dievaluasi menggunakan metrik akurasi, presisi, dan recall. Hasil penelitian menunjukkan bahwa XGBoost memiliki kinerja yang lebih baik dibandingkan AdaBoost, dengan tingkat akurasi, presisi, dan recall mencapai 92%.

Retno Wahyusari[8] melakukan penelitian pada tahun 2024, Obesitas menjadi masalah kesehatan yang semakin serius di seluruh dunia. Jumlah orang obesitas terus bertambah pesat. Penelitian ini mencoba mengelompokkan orang-orang yang obesitas menjadi beberapa kelompok menggunakan metode K-Medoids. Melalui evaluasi menggunakan metrik Davies-Bouldin Index sebesar 0,071, ditemukan bahwa pembagian data menjadi tiga kluster menghasilkan hasil pengelompokan yang paling optimal.

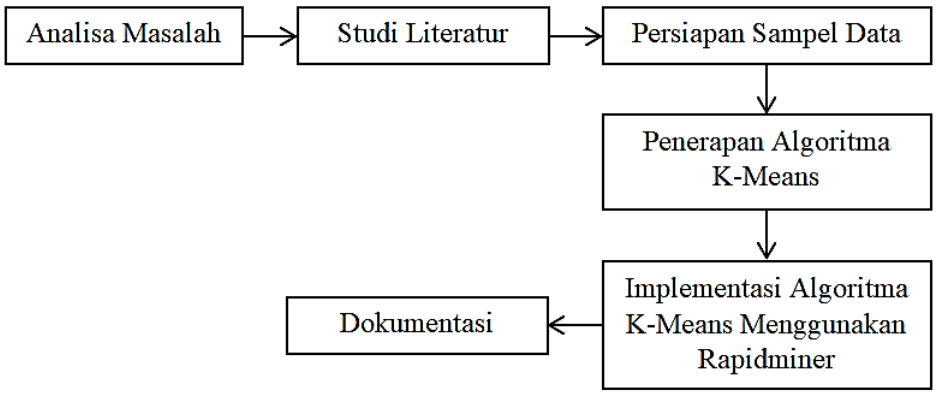
Faizal Shepyantoni, dkk[9] melakukan penelitian pada tahun 2024, dalam penelitian tersebut menjelaskan bahwa RSUD Kabupaten Kaur saat ini masih menghadapi tantangan dalam pengelolaan data pasien peserta BPJS Kesehatan yang bersifat manual. Penerapan metode K-Means Clustering pada data pasien rawat inap telah memberikan hasil yang signifikan dalam mengidentifikasi pola dan karakteristik pasien. Analisis cluster menunjukkan bahwa kelompok pasien dengan frekuensi kunjungan yang tinggi didominasi oleh perempuan berusia 40-59 tahun, kelas III, dengan diagnosis utama Diabetes Melitus Tipe 2.

Pada tahun 2023 Iwan Pii, dkk[10] melakukan penelitian, penelitian tersebut meneliti Dameyra Fashion, sebuah toko online pakaian di Cirebon yang beroperasi melalui Lazada, menghadapi tantangan dalam mengelola persediaan barang. Penelitian ini bertujuan untuk mengoptimalkan pengelolaan persediaan dengan mengidentifikasi produk laris dan tidak laris menggunakan metode k-means clustering. Analisis data penjualan melalui RapidMiner menunjukkan adanya 289 item produk laris seperti tunik dan kemeja, serta 9 item tidak laris seperti daster.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini mengikuti suatu proses yang terorganisir dengan baik, mulai dari tahap awal hingga akhir. Dengan menyusun tahapan penelitian secara sistematis, diharapkan penelitian dapat berjalan lebih terstruktur dan terfokus, sehingga memudahkan peneliti dalam mencapai tujuannya. Gambar 1 memberikan gambaran visual mengenai tahapan-tahapan yang akan dilalui dalam penelitian ini.



Gambar 1. Tahapan Penelitian

Berdasarkan gambar 1, berikut beberapa penjelasan dari tahapan penelitian yang akan dilakukan selama proses penyelesaian penelitian ini:

a. Analisis Masalah
   Pada tahap awal, kita akan mengidentifikasi secara spesifik masalah utama yang terkait dengan obesitas. Misalnya, kita akan mencari tahu faktor-faktor apa saja yang paling sering menyebabkan seseorang menjadi obesitas, terutama pada kelompok usia tertentu. Setelah mengetahui masalah utamanya, kita akan membatasi ruang lingkup penelitian agar lebih terarah. Pada penelitian ini akan memilih untuk fokus pada perempuan di suatu wilayah tertentu.
b. Studi Literatur
   Pada tahap ini akan mencari dan mengumpulkan berbagai informasi dari sumber-sumber yang terpercaya, seperti jurnal ilmiah atau laporan penelitian, yang berkaitan dengan obesitas. Informasi yang telah dikumpulkan kemudian akan kita analisis untuk mengetahui temuan-temuan penelitian sebelumnya, metode penelitian yang sudah digunakan, dan hal-hal apa saja yang belum diteliti oleh peneliti lain.
c. Persiapan Sampel Data
   Pada tahap ini akan menentukan kelompok orang yang akan menjadi objek penelitian, sebanyak 1610 dataset obesitas yang diambil dari kaggle dan dapat dilihat pada tabel 1 dibawah ini.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 863−872
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7462

Tabel 1. Dataset Obesitas

No.	Sex	Age	Height	Overweight_Obese_Family	...	...	Type_of_Transportation_Used
1	2	18	155	2	…	…	4
2	2	18	158	2	…	…	3
3	2	18	159	2	…	…	4
4	2	18	162	2	…	…	4
5	2	18	165	2	…	…	2
6	2	18	176	1	…	…	4
7	2	19	152	2	…	…	2
8	2	19	158	2	…	…	3
9	2	19	159	2	…	…	4
10	2	19	162	2	…	…	4
11	2	19	163	2	…	…	4
12	2	19	163	2	…	…	4
13	2	19	165	2	…	…	4
14	2	19	166	2	…	…	3
15	2	19	181	1	…	…	4
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
726	1	29	173	1	…	…	3
727	1	36	169	2	…	…	4
…	…	…	…	…	…	…	…
…	…	…	…	…	…	…	…
1609	2	53	168	2	…	…	1
1610	2	54	170	1	…	…	1

Berdasarkan tabel 1 diatas, penelitian ini akan berfokus pada data perempuan (sex = 2) yang ada dalam dataset tersebut, sehingga sampel data yang akan digunakan dalam penelitian berjumlah 898 data.

d. Penerapan Algoritma K-Means
Pada tahapan ini akan memilih beberapa faktor yang berpengaruh terhadap obesitas, seperti usia, riwayat keluarga, pola konsumsi, aktivitas fisik, hingga mode transportasi yang digunakan. Data akan dikelompokkan menggunakan algoritma K-Means menjadi beberapa kelompok berdasarkan kemiripan karakteristik. Misalnya, kita akan mengelompokkan peserta penelitian menjadi kelompok dengan risiko obesitas rendah, sedang, dan tinggi.

e. Implementasi Algoritma K-Means Menggunakan Python
Tahap selanjutnya akan membuat model menggunakan algoritma K-Means di dalam Python untuk mengelompokkan data. pada tahap ini juga akan memeriksa apakah model yang telah dibuat dapat mengelompokkan data dengan akurat dengan melakukan preprocessing terlebih dahulu. Preprocessing dilakukan dengan menormalisasikan skala data yang jauh berbeda antar variabel dengan menggunakan StandardScaler dari library scikit-learn.

f. Dokumentasi
Tahap akhir yaitu dokumentasi dengan cara menuliskan laporan dari penelitian yang telah dilakukan, hasil penelitian akan kita tulis dalam bentuk laporan yang lengkap. Laporan ini akan berisi latar belakang penelitian, metode yang digunakan, hasil yang diperoleh, pembahasan, dan kesimpulan.

2.2 KDD (*Knowledge Discovery in Databases*)

KDD adalah proses terstruktur untuk menggali pengetahuan berharga dari tumpukan data dalam database[11][12]. Proses ini tidak hanya sekadar mengumpulkan informasi, tetapi juga melibatkan serangkaian tahapan penting untuk memastikan data yang diolah akurat dan relevan. Tahapan-tahapan KDD yang dapat diterapkan antara lain[13]–[16]:

a. Memilih Data yang Tepat
Langkah awal adalah memilah data yang benar-benar dibutuhkan dari keseluruhan database. Tujuannya adalah agar analisis fokus pada informasi yang relevan dengan tujuan yang ingin dicapai.

b. Membersihkan dan Menata Data
Data yang sudah dipilih kemudian dibersihkan dari segala 'sampah' seperti data yang hilang, tidak konsisten, atau tidak relevan. Data juga diubah formatnya agar mudah diolah.

c. Mengubah Data
Data yang sudah bersih dan rapi selanjutnya diubah bentuknya agar sesuai dengan teknik analisis yang akan digunakan. Tujuannya agar pola-pola tersembunyi dalam data dapat lebih mudah diidentifikasi.

d. Mencari Pola dalam Data
Pada tahap ini, berbagai teknik 'penambangan data' diterapkan untuk menemukan pola-pola menarik yang mungkin tersembunyi dalam data. Contohnya, mencari kelompok pelanggan dengan karakteristik serupa atau menemukan hubungan antar produk yang sering dibeli bersamaan.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 863−872
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7462

e. Menilai Pola yang Ditemukan
Setelah pola-pola ditemukan, perlu dilakukan penilaian untuk menentukan pola mana yang paling berharga dan relevan. Tujuannya adalah untuk memastikan bahwa pengetahuan yang dihasilkan benar-benar berguna.
f. Menafsirkan Hasil
Tahap terakhir adalah menerjemahkan pola-pola yang telah dievaluasi menjadi pengetahuan yang mudah dipahami dan dapat digunakan. Pengetahuan ini dapat digunakan untuk membuat keputusan yang lebih baik, memprediksi tren, atau memahami data dengan lebih baik.

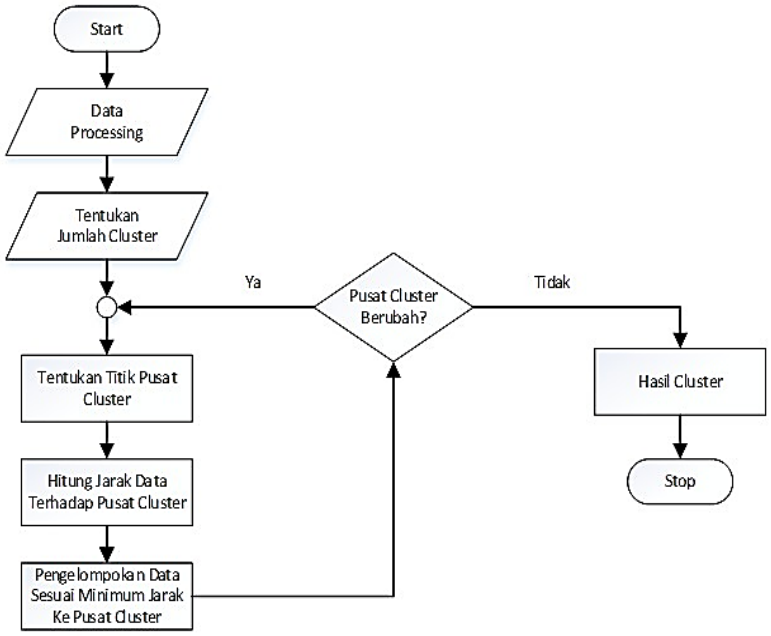
KDD adalah bidang yang terus berkembang seiring dengan pertumbuhan volume data. Dengan KDD, kita dapat mengubah data mentah menjadi pengetahuan berharga yang dapat mengubah cara kita hidup dan bekerja

2.3 Clustering

Clustering adalah proses mengelompokkan data menjadi beberapa kelompok atau cluster[17]–[19]. Data-data yang memiliki kesamaan atau ciri-ciri yang mirip akan dikelompokkan dalam satu cluster[18]. Misalnya, dalam data pelanggan suatu toko online, kita bisa mengelompokkan pelanggan berdasarkan kebiasaan belanja mereka. Ada yang sering membeli produk elektronik, ada yang lebih sering membeli pakaian, dan sebagainya. Bayangkan kamu sedang merapikan lemari pakaianmu. Kamu mengelompokkan baju-baju berdasarkan jenisnya, misalnya kaos, kemeja, celana. Baju-baju yang sejenis kamu taruh dalam satu laci atau rak. Nah, di dunia data, clustering itu seperti merapikan lemari pakaianmu, tapi dengan data-data.

2.4 Algoritma K-Means

K-Means merupakan metode dalam analisis data yang bertujuan mengelompokkan data menjadi beberapa cluster yang memiliki karakteristik serupa[20]–[23]. K-Means adalah teknik pengelompokan data yang melibatkan penentuan jumlah cluster yang diinginkan terlebih dahulu[24]. Selanjutnya, algoritma ini akan memilih secara acak beberapa titik data sebagai titik pusat awal dari setiap cluster. Kemudian, setiap data dalam dataset akan dihitung jaraknya terhadap semua titik pusat cluster. Setiap data akan dikelompokkan ke dalam cluster yang memiliki titik pusat terdekat. Setelah semua data dikelompokkan, titik pusat dari setiap cluster akan diperbarui dengan menghitung rata-rata dari semua data yang berada dalam cluster tersebut. Proses perhitungan jarak, pengelompokan, dan pembaruan titik pusat ini akan diulang secara berulang hingga tidak ada lagi perubahan signifikan dalam anggota setiap cluster[25][26].



Gambar 2. Flowchart K-Means Clustering

Berdasarkan Gambar 2 diatas, berikut rincian dari tahapan yang dilakukan dalam pengelompokkan menggunakan K-Means Clustering[27][28][29]:

a. Tentukan jumlah cluster (K) yang diinginkan setelah dilakukan preprocessing terhadap sampel data.
b. Tentukan titik pusat cluster (pilih secara acak K titik data sebagai pusat cluster awal (centroid)).
c. Hitung jarak setiap data ke semua pusat cluster (centroid) menggunakan rumus persamaan 1 dibawah ini.

$$d_{Euclidean}(X,Y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2} \quad (1)$$

Masukkan setiap data ke cluster yang centroid-nya paling dekat, hitung rata-rata dari semua data dalam setiap cluster dan update posisi centroid menjadi rata-rata yang baru dihitung.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 863−872
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7462

d. Bandingkan posisi centroid baru (pusat cluster) dengan posisi centroid sebelumnya. Jika ada perubahan signifikan, kembali ke langkah 2 dan gunakan rumus persamaan 2 dibawah ini untuk menentukan pusat centroid yang baru.

$$Ki = \frac{1}{M}\sum_{j=1}^{M} X_j \quad (2)$$

Jika tidak berubah, maka lanjut ke langkah terakhir (proses clustering selesai karena setiap data telah dikelompokkan ke dalam cluster yang sesuai)

Keterangan untuk rumus persamaan 1 dan 2 yaitu formula d(x,y) merupakan jarak data ke x ke pusat cluster y, Xi merupakan data ke-i pada atribut data ke n dan Yi merupakan data ke-j pada atribut data ke n.

3. HASIL DAN PEMBAHASAN

Penelitian ini memanfaatkan data sekunder yang bersumber dari platform Kaggle dengan fokus data yang digunakan adalah data perempuan. Variabel yang dianalisis mencakup usia, tinggi badan, riwayat keluarga dengan obesitas/kelebihan berat badan, kebiasaan konsumsi *fast food*, frekuensi makan sayur, jumlah makanan utama harian, asupan camilan, kebiasaan merokok, asupan cairan harian, perhitungan kalori, aktivitas fisik, durasi penggunaan gawai, serta moda transportasi. Sampel data penelitian dapat dilihat pada Tabel 2 di bawah.

Tabel 2. Sampel Data

No.	Age	Height	Overweight_Obese_Family	...	Type_of_Transportation_Used
1	18	155	2	...	4
2	18	158	2	...	3
3	18	159	2	...	4
4	18	162	2	...	4
5	18	165	2	...	2
6	18	176	1	...	4
7	19	152	2	...	2
8	19	158	2	...	3
9	19	159	2	...	4
10	19	162	2	...	4
…	…	…	…	...	…
…	…	…	…	...	…
897	53	168	2	...	1
898	54	170	1	...	1

Tabel 2 menyajikan sampel data penelitian yang terdiri dari 898 responden perempuan berdasarkan variabel penilaiannya. Data menunjukkan variasi usia responden antara 18 hingga 54 tahun, dengan tinggi badan berkisar 150–182 cm. Kolom Overweight_Obese_Family didominasi nilai 2 (menandakan adanya riwayat keluarga obesitas) pada sebagian besar responden, sementara untuk variabel terakhir seperti Type_of_Transportation_Used bervariasi dari nilai 1 hingga 4 (mewakili jenis transportasi seperti berjalan, motor, mobil, atau transportasi umum). Pola data awal ini mengindikasikan potensi pengaruh usia muda, faktor genetik, hingga kebiasaan mobilitas terhadap risiko obesitas, yang akan dianalisis lebih lanjut melalui clustering. Contohnya, responden usia 18–19 tahun cenderung menggunakan transportasi bernilai tinggi (3-4), sementara kelompok usia >50 tahun lebih banyak menggunakan transportasi bernilai rendah (1).

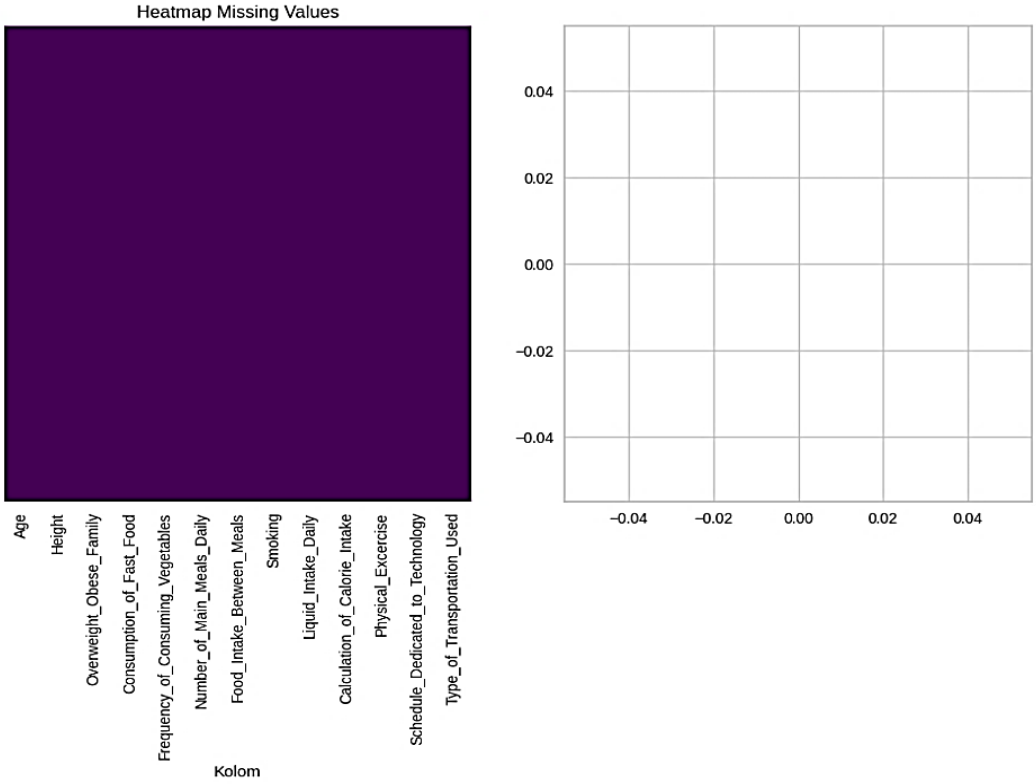
3.1 Hasil

Sebelum dilakukan analisis clustering, data terlebih dahulu melalui tahap preprocessing untuk memastikan kualitas dan konsistensi hasil. pada penelitian ini teknik preprocessing yang dilakukan adalah mencari data yang *missing values* dan melakukan normalisasi pada atribut usia dan tinggi badan untuk mempersiapkan data. Penelitian ini hanya melakukan normalisasi terhadap dua atribut dikarenakan atribut yang lain telah memiliki skala yang normal sehingga tidak perlu dilakukan normalisasi.

3.1.1 Preprocessing Data

Pada penelitian ini, dilakukan pengecekan kelengkapan dari setiap data yang akan dikelompokkan, berikut Gambar 3 adalah hasil pengecekaan terhadap missing values yang telah dilakukan.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 863−872
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7462



Gambar 3. Preprocessing Missing Value

Berdasarkan Gambar 3 diatas, terlihat bahwa tidak ada data yang mengalami missing value sehingga data dinyatakan lengkap. Tahap preprocessing selanjutnya adalah melakukan normalisasi terhadap taribut dengan skala berbeda seperti usia dalam tahun dan tinggi badan dalam cm yang dapat mempengaruhi performa algoritma K-Means karena sensitif terhadap jarak. Untuk mengatasi hal ini, dilakukan transformasi data menggunakan StandardScaler dari library scikit-learn. Proses ini mengubah nilai setiap fitur sehingga memiliki rata-rata (mean) 0 dan standar deviasi 1. Berdasarkan sampel data Tabel 4.1, rata-rata atribut usia 31,52115813 dan standar deviasi ≈10,01291471 sehingga proses perhitungan z-score sebagai berikut.

$$z = \frac{x-\mu}{\sigma} = \frac{18-31{,}52115813}{10{,}01291471} = -1{,}350371847$$

Tabel 3. Hasil Normalisasi

No.	Age	Height	Overweight_Obese_Family	...	Type_of_Transportation_Used
1	-1,350371847	-1,329794266	2	...	4
2	-1,350371847	-0,820848996	2	...	3
3	-1,350371847	-0,651200573	2	...	4
4	-1,350371847	-0,142255304	2	...	4
5	-1,350371847	0,366689966	2	...	2
6	-1,350371847	2,23282262	1	...	4
7	-1,250500827	-1,838739535	2	...	2
8	-1,250500827	-0,820848996	2	...	3
9	-1,250500827	-0,651200573	2	...	4
10	-1,250500827	-0,142255304	2	...	4
…	…	…	…	…	…
…	…	…	…	…	…
897	2,145113834	0,875635235	2	...	1
898	2,244984854	1,214932082	1	...	1

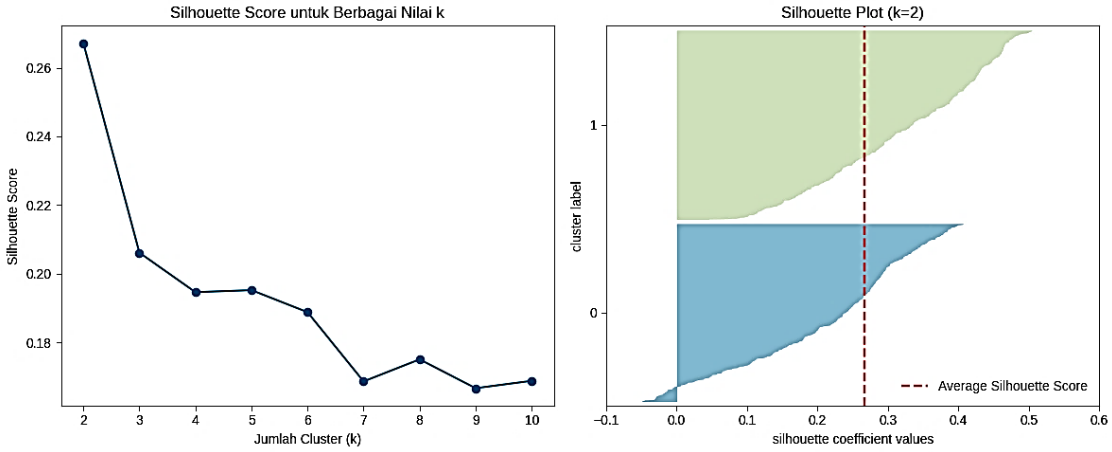
Berdasarkan Tabel 3, Setelah normalisasi dengan StandardScaler, nilai usia dan tinggi badan diubah menjadi skala z-score sehingga terlihat nilai negatif seperti --1,350371847 menunjukkan usia di bawah rata-rata, sedangkan nilai positif 2,244984854 menunjukkan usia di atas rata-rata.

3.1.2 Penentuan Jumlah Cluster Optimal

Penelitian ini menerapkan Silhouette Score sebagai indikator objektif untuk mengidentifikasi jumlah kelompok paling efektif dalam analisis K-Means Clustering. Metrik ini mengevaluasi kualitas pengelompokan dengan menganalisis

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 863−872
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7462

tingkat kedekatan data dalam satu cluster (kohesi) dibandingkan dengan jaraknya terhadap cluster lain (separasi), dengan skala penilaian antara -1 sampai 1. Skor mendekati 1 menandakan pengelompokan yang ideal, di mana objek memiliki karakteristik sangat mirip dengan anggota clusternya dan sangat berbeda dengan cluster lain, sementara skor negatif menunjukkan kesalahan pengelompokan. Proses pemilihan jumlah cluster dilakukan dengan membandingkan Silhouette Score pada berbagai kemungkinan jumlah cluster (misal k=2 sampai k=10), kemudian memilih nilai k dengan skor tertinggi sebagai solusi optimal. Teknik ini tidak hanya menjamin pembagian cluster yang kompak secara internal, tetapi juga menghasilkan pemisahan antar-cluster yang jelas, sehingga memungkinkan identifikasi kelompok risiko obesitas yang lebih akurat dan dapat ditindaklanjuti secara klinis. Berikut Gambar 3 yang memvisualisasikan Grafik Silhouette Score dalam pengelompokkan obesitas.



Gambar 3. Grafik Silhouette Score

Berdasarkan hasil analisis Silhouette Score untuk menentukan jumlah cluster optimal pada Gambar 3, terlihat bahwa nilai tertinggi dicapai ketika menggunakan 2 cluster (k=2) dengan skor 0.267, yang menunjukkan struktur pengelompokan paling baik dibandingkan opsi lainnya. Skor tersebut secara konsisten menurun seiring penambahan jumlah cluster, dengan nilai 0.206 untuk k=3 dan terus stabil di kisaran 0.195–0.169 untuk k=4 hingga k=10, mengindikasikan bahwa pembagian lebih dari 2 cluster justru mengurangi kualitas segmentasi data. Hasil ini membuktikan bahwa penggunaan 2 cluster merupakan pilihan paling optimal, karena tidak hanya mempertahankan kohesi internal yang kuat tetapi juga memaksimalkan separasi antar-cluster, sehingga memenuhi kriteria analisis clustering yang efektif untuk identifikasi risiko obesitas.

3.1.3 Penerapan K-Means

Penelitian ini menggunakan algoritma K-Means dengan jumlah cluster optimal sebanyak 2, yang telah ditentukan melalui analisis Silhouette Score sebelumnya. Setelah menentukan jumlah cluster, hasil pengelompokan K-Means menampilkan dua segmentasi risiko obesitas yang jelas seperti Cluster 1, Kelompok dengan karakteristik risiko obesitas lebih tinggi (misal: pola konsumsi tinggi kalori, aktivitas fisik rendah), dan cluster 2, Kelompok dengan risiko lebih rendah. Berikut Tabel 4 menampilkan hasil pengelompokkan menggunakan metode K-Means.

Tabel 4. Hasil Clustering dengan K-Means

No.	Age	Height	Overweight_Obese_Family	Type_of_Transportation_Used	Cluster	Risk_Score
1	-1,35037	-1,32979	2	…	1	2,25
2	-1,35037	-0,82085	2	…	1	2,25
3	-1,35037	-0,6512	2	…	1	2,5
4	-1,35037	-0,14226	2	…	1	2,5
5	-1,35037	0,36669	2	…	1	2
6	-1,35037	2,232823	1	…	0	1,75
7	-1,2505	-1,83874	2	…	1	2
8	-1,2505	-0,82085	2	…	1	3
9	-1,2505	-0,6512	2	…	1	2,25
10	-1,2505	-0,14226	2	…	1	2,25
…	…	…	…	…	…	…
897	2,145114	0,875635	2	…	0	2,5
898	2,244985	1,214932	1	…	0	1,75

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 863−872
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7462

Berdasarkan Tabel 4 hasil clustering K-Means, terlihat bahwa data terbagi menjadi dua cluster utama (Cluster 0 dan 1) dengan karakteristik yang berbeda. Cluster 1 didominasi oleh individu dengan nilai normalisasi Age negatif (usia lebih muda) dan Overweight_Obese_Family = 2 (riwayat keluarga obesitas), serta memiliki Risk_Score lebih tinggi (2-3), mengindikasikan kelompok berisiko obesitas. Sementara itu, Cluster 0 mencakup individu dengan variasi usia (termasuk nilai Age positif/lebih tua) dan Risk_Score relatif lebih rendah (1.75-2.5), meskipun beberapa kasus tetap menunjukkan risiko sedang. Pola ini menegaskan bahwa faktor usia muda dan riwayat keluarga berkontribusi signifikan terhadap peningkatan risiko obesitas, sementara penggunaan transportasi (kolom Type_of_Transportation_Used) dan tinggi badan (Height) juga memengaruhi, tetapi dengan pola yang lebih kompleks. Hasil ini selaras dengan analisis Silhouette Score sebelumnya yang mendukung k=2 sebagai jumlah cluster optimal, sekaligus memberikan dasar untuk intervensi berbasis karakteristik masing-masing kelompok. Rata-rata hasil clsutering secara keseluruhan dapat dilihat pada Tabel 5 berikut.

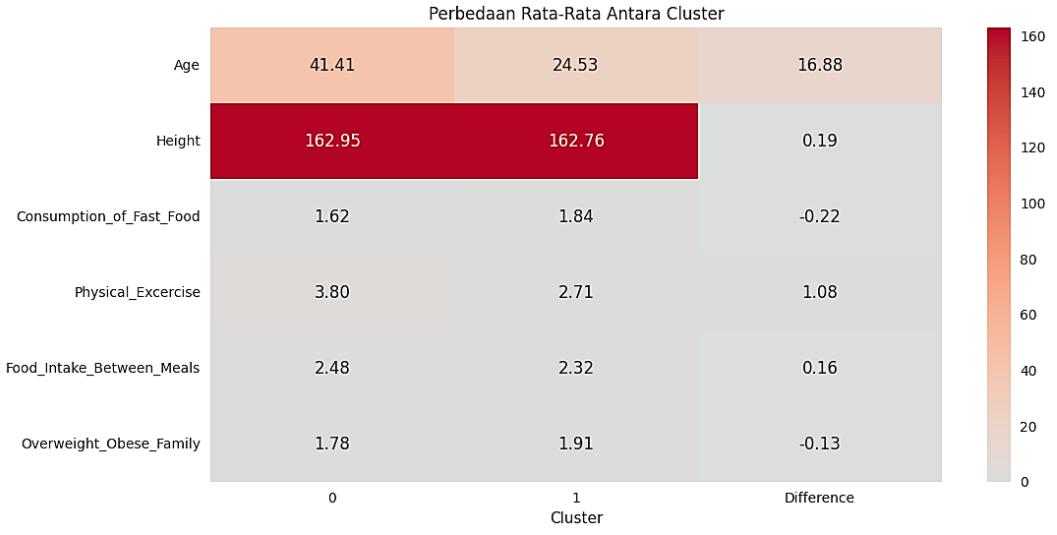
Tabel 5. Rata-Rata Hasil Clustering dengan K-Means

Cluster	Age	Height	Overweight_Obese_Family	Consumption_of_Fast_Food	...	Jumlah_Sample
0	0.581	0.027	1.775	1.586	...	435
1	-0.546	-0.025	1.935	1.909	...	463

Berdasarkan rata-rata hasil clustering Tabel 5, terlihat perbedaan karakteristik yang jelas antara kedua cluster. Cluster 1 menunjukkan nilai rata-rata Age negatif (-0.546) yang mengindikasikan dominasi usia lebih muda, dengan riwayat keluarga obesitas lebih tinggi (1.935) dan konsumsi fast food lebih sering (1.909), sehingga mengidentifikasi kelompok ini sebagai populasi berisiko tinggi obesitas. Sementara Cluster 0 memiliki rata-rata Age positif (0.581) yang mencerminkan usia relatif lebih tua, dengan riwayat keluarga obesitas (1.775) dan konsumsi fast food (1.586) yang sedikit lebih rendah, meskipun tetap perlu perhatian. Perbedaan signifikan pada variabel-variabel kunci ini memperkuat validitas pembagian cluster, di mana Cluster 1 memerlukan intervensi prioritas akibat kombinasi faktor usia muda, pola makan, dan riwayat keluarga, sedangkan Cluster 0 dapat menjadi fokus program pencegahan berbasis usia. Distribusi sampel yang seimbang (435 vs 463) juga menegaskan bahwa kedua kelompok memiliki representasi data yang memadai untuk analisis lebih lanjut.

3.2 Pembahasan

Berikut ini disajikan analisis perbedaan karakteristik antara dua cluster berdasarkan variabel-variabel kunci yang mempengaruhi risiko obesitas yang telah disesuaikan dengan data sebelum dilakukan normalisasi pada variabel usia dan tinggi badan berdasarkan Tabel 5. Gambar 4 akan membandingkan nilai rata-rata dari berbagai faktor seperti usia (Age) hingga riwayat keluarga obesitas (Overweight_Obese_Family) antara kedua cluster.



Gambar 4. Perbedaan Rata-rata antara Cluster

Berdasarkan hasil analisis perbedaan rata-rata antar cluster Gambar 4, terlihat bahwa Cluster 1 (kanan) didominasi oleh kelompok usia lebih muda (24.53 tahun) dibandingkan Cluster 0 (41.41 tahun) dengan selisih 16.88 tahun, sekaligus menunjukkan karakteristik risiko obesitas yang lebih tinggi seperti konsumsi fast food lebih sering (1.84 vs 1.62), riwayat keluarga obesitas lebih kuat (1.91 vs 1.78), dan aktivitas fisik lebih rendah (2.71 vs 3.80). Sementara itu, perbedaan tinggi badan (162.76 cm vs 162.95 cm) dan frekuensi makan camilan (2.32 vs 2.48) tidak signifikan, mengindikasikan bahwa faktor usia, pola makan, dan gaya hidup lebih berpengaruh dalam pembedaan cluster. Hasil ini mempertegas bahwa Cluster 1 merupakan kelompok prioritas untuk intervensi pencegahan obesitas, sementara Cluster 0 cenderung lebih sehat karena kombinasi usia lebih tua dan kebiasaan olahraga yang lebih rutin, meskipun tetap memerlukan monitoring pada aspek konsumsi fast food.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 863−872
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7462

Berdasarkan analisis clustering menggunakan algoritma K-Means dengan k=2 yang telah dioptimalkan melalui Silhouette Score, penelitian ini berhasil mengidentifikasi dua kelompok populasi dengan karakteristik risiko obesitas yang berbeda secara signifikan. Hasil clustering menunjukkan bahwa Cluster 1 didominasi oleh individu berusia lebih muda (rata-rata 24.53 tahun) dengan riwayat keluarga obesitas lebih tinggi (1.91), frekuensi konsumsi fast food lebih sering (1.84), dan tingkat aktivitas fisik lebih rendah (2.71). Sementara itu, Cluster 0 terdiri atas populasi berusia lebih tua (rata-rata 41.41 tahun) dengan pola hidup relatif lebih sehat, meskipun tetap memiliki faktor risiko seperti kebiasaan ngemil (2.48) dan riwayat keluarga obesitas (1.78). Perbedaan mencolok antara kedua cluster terlihat pada usia, gaya hidup dan Riwayat keluarga.

a. Usia: Selisih 16.88 tahun menunjukkan bahwa usia muda merupakan faktor kritis dalam kelompok berisiko tinggi.
b. Gaya Hidup: Cluster 1 memiliki konsumsi fast food lebih tinggi dan aktivitas fisik lebih rendah, yang konsisten dengan literatur terkait obesitas pada generasi muda.
c. Riwayat Keluarga: Nilai yang lebih tinggi pada Cluster 1 mengindikasikan pengaruh genetik yang kuat.

Temuan ini sejalan dengan studi sebelumnya yang menyebutkan bahwa interaksi antara faktor genetik dan gaya hidup tidak sehat memperburuk risiko obesitas[30]. Kelebihan analisis ini adalah penggunaan normalisasi StandardScaler yang memastikan komparasi antar variabel objektif, serta validasi Silhouette Score (0.267 untuk k=2) yang menjamin kualitas pengelompokan. Namun, keterbatasan terletak pada belum dimasukkannya variabel sosioekonomi yang mungkin memengaruhi pola konsumsi. Implikasi praktis dari penelitian ini adalah perlunya intervensi berbasis cluster:

a. Cluster 1: Program edukasi gizi dan aktivitas fisik intensif untuk kelompok muda.
b. Cluster 0: Skrining kesehatan berkala untuk memantau faktor risiko pada usia produktif.

Dengan demikian, pendekatan clustering ini tidak hanya memetakan risiko obesitas secara akurat tetapi juga menyediakan dasar untuk kebijakan kesehatan yang lebih terarah.

4. KESIMPULAN

Berdasarkan analisis K-Means Clustering terhadap 898 responden perempuan, penelitian ini berhasil mengidentifikasi dua kelompok risiko obesitas yang berbeda secara signifikan, di mana kelompok berisiko tinggi (Cluster 1) didominasi oleh perempuan usia muda (24.53 tahun) dengan riwayat keluarga obesitas, konsumsi fast food yang tinggi, dan aktivitas fisik rendah, sementara kelompok berisiko lebih rendah (Cluster 0) terdiri dari perempuan usia lebih tua (41.41 tahun) dengan pola hidup relatif lebih sehat. Temuan ini mengkonfirmasi bahwa interaksi antara faktor genetik dan gaya hidup tidak sehat, khususnya pada populasi usia muda, menjadi penyebab utama obesitas, sehingga diperlukan intervensi berbasis kelompok seperti program edukasi gizi dan promosi aktivitas fisik yang ditargetkan untuk populasi berisiko tinggi. Penelitian ini tidak hanya memperkuat penerapan data mining dalam bidang kesehatan masyarakat tetapi juga memberikan bukti empiris untuk pengembangan kebijakan pencegahan obesitas yang lebih efektif dan terarah di Indonesia.

REFERENCES

[1] D. Pratista, R. A. D. Sartika, dan P. N. Putri, “The Prevalence and Risk Factors of Central Obesity in Hypertensive Patients at Puskesmas Kemiri Muka, Depok City, West Jawa,” *J. Indones. Nutr. Assoc.*, vol. 47, no. 2, hal. 195–208, 2024, doi: 10.36457/gizindo.v47i2.1066.
[2] T. Hien, T. Nguyen, D. Tai, D. Songsak, dan S. Van Nam, “A method for k-means-like clustering of categorical data,” *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 11, hal. 15011–15021, 2023, doi: 10.1007/s12652-019-01445-5.
[3] S. M. Miraftabzadeh, C. G. Colombo, M. Longo, dan F. Foiadelli, “K-Means and Alternative Clustering Methods in Modern Power Systems,” *IEEE Access*, vol. 11, hal. 119596–119633, 2023, doi: 10.1109/ACCESS.2023.3327640.
[4] S. Suraya, M. Sholeh, dan U. Lestari, “Evaluation of Data Clustering Accuracy using K-Means Algorithm,” *Int. J. Multidiscip. Approach Res. Sci.*, vol. 2, no. 01, hal. 385–396, 2023, doi: https://doi.org/10.59653/ijmars.v2i01.504.
[5] B. Liu, C. Liu, Y. Zhou, D. Wang, dan Y. Dun, “An unsupervised chatter detection method based on AE and merging GMM and K-means,” *Mech. Syst. Signal Process.*, vol. 186, hal. 109861, 2023, doi: https://doi.org/10.1016/j.ymssp.2022.109861.
[6] H. Hu, J. Liu, X. Zhang, dan M. Fang, “An effective and adaptable K-means algorithm for big data cluster analysis,” *Pattern Recognit.*, vol. 139, hal. 109404, 2023, doi: https://doi.org/10.1016/j.patcog.2023.109404.
[7] C. E. Sukmawati, A. Fitri, N. Masruriyah, dan A. R. Juwita, “Efektivitas algoritma AdaBoost dan XGBoost pada dataset obesitas populasi dewasa,” *Jambura J. Informatics*, vol. 6, no. 2, hal. 101–111, 2024, doi: 10.37905/jji.
[8] R. Wahyusari, “Penerapan Algoritma K-Medoids Untuk Mengelompokkan Status Obesitas,” *Simetris*, vol. 18, no. 1, hal. 1–4, 2024, [Daring]. Tersedia pada: https://www.utrcepu.ac.id/index.php/simetris/article/download/405
[9] F. Shepyantoni, I. Kanedi, dan E. Suryana, “Penerapan Metode K-Means Clustering Dalam Pengelompokan Data Pasien Rawat Inap Peserta BPJS Di Rumah Sakit Umum Daerah Kabupaten Kaur,” *J. Media Infotama*, vol. 20, no. 2, hal. 493–500, 2024, doi: https://doi.org/10.37676/jmi.v20i2.6458.
[10] I. Pii, N. Suarna, dan N. Rahaningsih, “Penerapan Data Mining Pada Penjualan Produk Pakaian Dameyra Fashion Menggunakan Metode K-Means Clustering,” *JATI (Jurnal Mhs. Tek. Inform.*, vol. 7, no. 1, hal. 423–430, 2023, doi: https://doi.org/10.36040/jati.v7i1.6336.
[11] C. Llatas, B. Soust-Verdaguer, L. C. Torres, dan D. Cagigas, “Application of Knowledge Discovery in Databases (KDD) to environmental, economic, and social indicators used in BIM workflow to support sustainable design,” *J. Build. Eng.*, vol. 91, hal. 109546, 2024, doi: https://doi.org/10.1016/j.jobe.2024.109546.
[12] S. Głowania, J. Kozak, dan P. Juszczuk, “Knowledge discovery in databases for a football match result,” *Electronics*, vol.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 1, June 2025 Page: 863−872
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i1.7462

12, no. 12, hal. 2712, 2023, doi: https://doi.org/10.3390/electronics12122712.
[13] R. H. Sukarna dan Y. Ansori, "Implementasi Data Mining Menggunakan Metode Naive Bayes Dengan Feature Selection Untuk Prediksi Kelulusan Mahasiswa Tepat Waktu," *J. Ilm. Sains dan Teknol.*, vol. 6, no. 1, hal. 50–61, 2022, doi: 10.47080/saintek.v6i1.1467.
[14] F. O. Lusiana, I. Fatma, dan A. P. Windarto, "Estimasi Laju Pertumbuhan Penduduk Menggunakan Metode Regresi Linier Berganda Pada BPS Simalungun," *J. Informatics Manag. Inf. Technol.*, vol. 1, no. 2, hal. 79–84, 2021, [Daring]. Tersedia pada: https://hostjournals.com/
[15] Z. Nabila, A. Rahman Isnain, dan Z. Abidin, "Analisis Data Mining Untuk Clustering Kasus Covid-19 Di Provinsi Lampung Dengan Algoritma K-Means," *J. Teknol. dan Sist. Inf.*, vol. 2, no. 2, hal. 100, 2021, [Daring]. Tersedia pada: http://jim.teknokrat.ac.id/index.php/JTSI
[16] Y. L. Nainel, E. Buulolo, dan I. Lubis, "Penerapan Data Mining Untuk Estimasi Penjualan Obat Berdasarkan Pengaruh Brand Image Dengan Algoritma Expectation Maximization (Studi Kasus: PT. Pyridam Farma Tbk)," *JURIKOM (Jurnal Ris. Komputer)*, vol. 7, no. 2, hal. 214, 2020, doi: 10.30865/jurikom.v7i2.2097.
[17] G. J. Oyewole dan G. A. Thopil, "Data clustering: application and trends," *Artif. Intell. Rev.*, vol. 56, no. 7, hal. 6439–6475, 2023, doi: https://doi.org/10.1007/s10462-022-10325-y.
[18] S. E. Hashemi, F. Gholian-Jouybari, dan M. Hajiaghaei-Keshteli, "A fuzzy C-means algorithm for optimizing data clustering," *Expert Syst. Appl.*, vol. 227, hal. 120377, 2023, doi: https://doi.org/10.1016/j.eswa.2023.120377.
[19] S. Pitafi, T. Anwar, dan Z. Sharif, "A taxonomy of machine learning clustering algorithms, challenges, and future realms," *Appl. Sci.*, vol. 13, no. 6, hal. 3529, 2023, doi: https://doi.org/10.3390/app13063529.
[20] M. Annas dan S. N. Wahab, "Data mining methods: K-means clustering algorithms," *Int. J. Cyber IT Serv. Manag.*, vol. 3, no. 1, hal. 40–47, 2023, doi: https://doi.org/10.34306/ijcitsm.v3i1.122.
[21] T.-H. T. Nguyen, D.-T. Dinh, S. Sriboonchitta, dan V.-N. Huynh, "A method for k-means-like clustering of categorical data," *J. Ambient Intell. Humaniz. Comput.*, vol. 14, no. 11, hal. 15011–15021, 2023, doi: https://doi.org/10.1007/s12652-019-01445-5.
[22] P. Dubey dan A. Rajavat, "Effective K-means clustering algorithm for efficient data mining," in *2023 2nd International Conference on Vision Towards Emerging Trends in Communication and Networking Technologies (ViTECoN)*, 2023, hal. 1–6. doi: 10.1109/ViTECoN58111.2023.10157179.
[23] G. E. Okereke, M. C. Bali, C. N. Okwueze, E. C. Ukekwe, S. C. Echezona, dan C. I. Ugwu, "K-means clustering of electricity consumers using time-domain features from smart meter data," *J. Electr. Syst. Inf. Technol.*, vol. 10, no. 1, hal. 1–18, 2023, doi: https://doi.org/10.1186/s43067-023-00068-3.
[24] R. Zaib dan O. Ourabah, "Large scale data using K-means," *Mesopotamian J. Big Data*, vol. 2023, hal. 36–45, 2023, doi: https://doi.org/10.58496/MJBD/2023/006.
[25] S. N. Alaziz, B. Albayati, A. al-A. H. El-Bagoury, dan W. Shafik, "Clustering of COVID-19 multi-time series-based K-means and PCA with forecasting," *Int. J. Data Warehous. Min.*, vol. 19, no. 3, hal. 1–25, 2023, doi: 10.4018/IJDWM.317374.
[26] S. Kim, S. Cho, J. Y. Kim, dan D.-J. Kim, "Statistical assessment on student engagement in asynchronous online learning using the k-means clustering algorithm," *Sustainability*, vol. 15, no. 3, hal. 2049, 2023, doi: https://doi.org/10.3390/su15032049.
[27] I. F. Ashari, E. D. Nugroho, R. Baraku, I. N. Yanda, dan R. Liwardana, "Analysis of elbow, silhouette, Davies-Bouldin, Calinski-Harabasz, and rand-index evaluation on k-means algorithm for classifying flood-affected areas in Jakarta," *J. Appl. Informatics Comput.*, vol. 7, no. 1, hal. 95–103, 2023, doi: https://doi.org/10.30871/jaic.v7i1.4947.
[28] E. L. Cahapin, B. A. Malabag, C. S. Santiago Jr, J. L. Reyes, G. S. Legaspi, dan K. L. Adrales, "Clustering of students admission data using k-means, hierarchical, and DBSCAN algorithms," *Bull. Electr. Eng. Informatics*, vol. 12, no. 6, hal. 3647–3656, 2023, doi: https://doi.org/10.11591/eei.v12i6.4849.
[29] O. Khan *et al.*, "Exploring the performance of biodiesel-hydrogen blends with diverse nanoparticles in diesel engine: A hybrid machine learning K-means clustering approach with weighted performance metrics," *Int. J. Hydrogen Energy*, vol. 78, hal. 547–563, 2024, doi: https://doi.org/10.1016/j.ijhydene.2024.06.303.
[30] M. S. Kim *et al.*, "Association of genetic risk, lifestyle, and their interaction with obesity and obesity-related morbidities," *Cell Metab.*, vol. 36, no. 7, hal. 1494–1503, 2024, doi: 10.1016/j.cmet.2024.06.004.