JERIN: Jurnal Riset Elektro dan Informatika
Vol. 1, No. 1, Mei 2025, pp. 10-16

Analisis Pengelompokan Faktor Risiko Diabetes Berdasarkan Indikator Kesehatan Menggunakan K-Means Clustering

Mildayanti Mildayanti[1], Nova Febryna. A[2*], Sahra Zhafirah[3]
[1,2,3] Program Studi Teknik Informatika, Universitas Muslim Indonesia, Kota Makassar

***Email Korespondensi:** 13020230040@student.umi.ac.id

Riwayat Artikel: Diterima: 11/02/2025; Direvisi: 18/04/2025; Disetujui: 02/05/2025

ABSTRAK

Penelitian ini bertujuan untuk menganalisis pengelompokan faktor risiko diabetes melitus berdasarkan indikator kesehatan menggunakan metode K-Means Clustering. Data yang digunakan merupakan data sekunder dari *platform* Kaggle yang terdiri dari 768 data pasien dengan beberapa indikator klinis, antara lain kadar glukosa darah, tekanan darah, insulin, indeks massa tubuh, riwayat keluarga diabetes, dan usia. Tahapan penelitian meliputi preprocessing data yang mencakup penanganan nilai tidak valid, imputasi data hilang menggunakan *median*, penanganan *outlier* dengan metode *Interquartile Range*, serta normalisasi data menggunakan StandardScaler. Proses *clustering* dilakukan dengan algoritma K-Means dengan jumlah *cluster* optimal sebanyak tiga yang ditentukan menggunakan *Elbow Method*, dan kualitas pengelompokan dievaluasi menggunakan Silhouette Score. Hasil penelitian menunjukkan bahwa data pasien dapat dikelompokkan ke dalam tiga *cluster* dengan karakteristik risiko diabetes yang berbeda, yaitu risiko rendah, risiko sedang, serta risiko sedang hingga tinggi. Meskipun nilai Silhouette Score menunjukkan adanya tumpang tindih antar *cluster*, hasil pengelompokan tetap memberikan informasi yang bermakna dalam mengidentifikasi profil risiko diabetes. Penelitian ini menunjukkan bahwa metode K-Means Clustering efektif digunakan untuk memetakan faktor risiko diabetes berdasarkan indikator kesehatan dan dapat menjadi dasar dalam analisis serta pengambilan keputusan di bidang kesehatan.

Kata Kunci: clustering data medis; diabetes melitus; indikator kesehatan; K-Means Clustering; pengelompokan diabetes

1. PENDAHULUAN

Diabetes melitus (DM) adalah salah satu penyakit tidak menular yang paling umum dan semakin meningkat di seluruh dunia, termasuk di Indonesia [1]. Meningkatnya kadar glukosa darah yang disebabkan oleh gangguan dalam produksi atau kerja insulin [2] adalah tanda kondisi ini. Menurut penelitian Irbah dan kawan kawan [7], gaya hidup yang tidak sehat, termasuk konsumsi makanan tinggi gula, kurangnya aktivitas fisik, dan faktor usia dan obesitas, berkontribusi pada peningkatan jumlah kasus diabetes di wilayah Aceh. Ini juga sejalan dengan temuan Sari [8] yang menunjukkan bahwa pola hidup dan kebiasaan makan memiliki hubungan signifikan dengan kemungkinan terkena diabetes tipe 2.

Analisis faktor risiko diabetes sangat penting untuk membantu mencegah dan mendeteksi diabetes sejak dini. Kadar glukosa darah, tekanan darah, indeks massa tubuh (IMT), kolesterol, dan usia adalah beberapa indikator kesehatan yang dapat digunakan untuk menentukan tingkat risiko seseorang terhadap diabetes [3]. Ramadhani dkk. [4] mengelompokkan pasien diabetes berdasarkan tingkat glukosa dan tekanan darah mereka. Hasilnya menunjukkan bahwa ada perbedaan signifikan dalam fitur kesehatan antara masing-masing kelompok pasien.

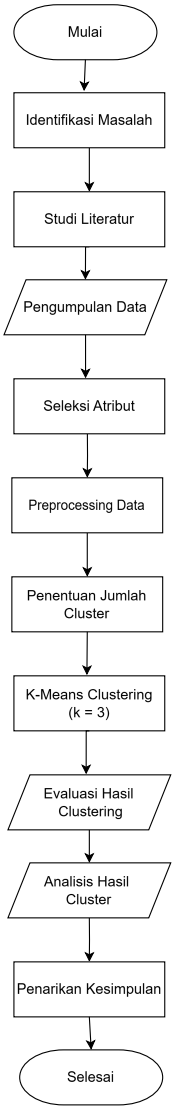
Pola tersembunyi dalam data kesehatan yang besar dan kompleks dapat ditemukan melalui penggunaan teknik pengelompokan, juga dikenal sebagai clustering [4]. Karena kemampuan untuk membagi data ke dalam kelompok berdasarkan kemiripan karakteristik, K-Means Clustering adalah salah satu algoritma yang paling banyak digunakan [2], [5]. Menurut penelitian [6], metode K-Means berhasil menentukan tingkat risiko diabetes berdasarkan IMT dan kadar gula darah. Di sisi lain, penelitian oleh Hussain [3] dan Kumar dan Kawan Kawan [7] menunjukkan bahwa menggunakan K-Means pada data klinis pasien dapat menghasilkan pola distribusi risiko yang bermanfaat untuk analisis prediktif. Penelitian oleh Simanjuntak dan Hasibuan [8] juga menunjukkan bahwa kombinasi analisis faktor risiko dan metode K-Means mampu mengelompokkan pasien diabetes ke dalam tingkat risiko ringan, sedang, dan berat secara efektif berdasarkan data klinis.

Tiga klaster risiko diciptakan oleh penelitian sebelumnya oleh Gestavito dan Kawan Kawan [2] yang menganalisis kadar glukosa dan tekanan darah sebagai faktor risiko penyakit diabetes melitus. Selain itu, Basyir [4] mengkonfirmasi bahwa algoritma K-Means memiliki kemampuan untuk meningkatkan visualisasi dan pemahaman pola risiko penyakit. Berdasarkan temuan berbagai penelitian, metode K-Means terbukti relevan untuk analisis pengelompokan data kesehatan dan dapat membantu tenaga medis dalam menentukan strategi pencegahan.

Meninjau uraian sebelumnya, penelitian ini bertujuan untuk menganalisis pengelompokan faktor risiko diabetes berdasarkan indikator kesehatan menggunakan metode K-Means Clustering. Diharapkan melalui pengelompokan ini, informasi mendalam tentang fitur kesehatan masing-masing kelompok risiko akan dikumpulkan, yang akan menjadi dasar untuk pencegahan dan pengelolaan penyakit diabetes melitus di masyarakat.

2. METODOLOGI PENELITIAN

Bagian Metodologi Penelitian menjelaskan secara sistematis rancangan penelitian, tahapan pelaksanaan, serta pendekatan yang digunakan untuk mencapai tujuan penelitian. Uraian metodologi disusun secara jelas dan logis agar proses penelitian dapat dipahami serta direplikasi oleh peneliti lain. Pada bagian ini dijelaskan secara runtut mulai dari desain penelitian, sumber dan variabel data yang digunakan, tahapan pengolahan dan perancangan sistem, hingga metode analisis dan evaluasi hasil yang diterapkan dalam penelitian.



Gambar 1. Diagram Alur Metodologi Penelitian

2.1. Akuisisi Data

Dataset yang digunakan dalam penelitian ini merupakan data sekunder yang diperoleh dari platform Kaggle, yaitu Diabetes *Dataset* yang dipublikasikan oleh Akshay Dattatray Khare. *Dataset* ini dapat diakses melalui tautan berikut:
https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset/data

Dataset diperoleh dengan cara mengunduh langsung (download) dari platform Kaggle dalam format *Comma-Separated Values* (CSV). *Dataset* mencakup 768 data pasien dengan berbagai karakteristik klinis yang berkaitan dengan faktor risiko diabetes melitus.
Atribut yang digunakan dalam penelitian ini meliputi:

a. Glukosa darah
b. Tekanan darah
c. Insulin
d. Indeks Massa Tubuh (*Body Mass Index* / BMI)
e. Riwayat keluarga diabetes
f. Usia

Atribut Pregnancies tidak digunakan karena bersifat spesifik gender dan tidak relevan secara universal terhadap seluruh populasi pasien. *Atribut Outcome* tidak digunakan dalam proses pembentukan *cluster*, tetapi dimanfaatkan pada pada tahap analisis hasil, atribut ini digunakan untuk membantu memahami kondisi diabetes pada masing-masing *cluster*.

2.2. Preprocessing Data

Tahap *preprocessing data* dilakukan untuk meningkatkan kualitas data sebelum diterapkan algoritma clustering. Tujuan dari proses ini adalah untuk mengurangi *noise*, menangani data yang tidak valid, dan menyamakan skala antar atribut sehingga hasil pengelompokan menjadi lebih baik.
Tahapan *preprocessing* yang dilakukan adalah sebagai berikut:

a. Penanganan Nilai Tidak Valid
   Nilai nol pada atribut klinis seperti glukosa darah, tekanan darah, BMI dan insulin diperlakukan sebagai missing value karena secara medis tidak mungkin bernilai nol. Metode ini biasanya digunakan untuk mengolah data Kesehatan [2].
b. Imputasi Data Hilang
   Nilai missing digantikan menggunakan nilai median dari masing-masing atribut. Penggunaan Median lebih tahan terhadap nilai ekstrem dari pada mean, terutama dalam data klinis [8].
c. Penanganan *Outlier*
   Outlier diidentifikasi dan ditangani menggunakan metode *Interquartile Range* (IQR) digunakan untuk mengurangi pengaruh nilai ekstrem terhadap proses clustering, mengingat algoritma K-Means sensitif terhadap *outlier* [1].
d. Normalisasi Data
   Normalisasi dilakukan menggunakan metode *StandardScaler* untuk melakukan normalisasi data sehingga setiap atribut memiliki skala yang seimbang dalam perhitungan jarak geometris [4].

2.3. Model K-Means Clustering

Algoritma yang digunakan dalam penelitian ini adalah K-Means Clustering, yaitu metode *unsupervised learning* bertujuan untuk mengelompokkan data ke dalam berbagai cluster berdasarkan tingkat kemiripan data.
Tahapan algoritma K-Means Clustering adalah sebagai berikut:

a. Menentukan jumlah *cluster* (*k*).
b. Menginisialisasi pusat *cluster* (*centroid*) secara acak.
c. Menghitung jarak setiap data terhadap *centroid* menggunakan jarak *Euclidean*.
d. Mengelompokkan data ke dalam cluster dengan jarak terdekat.
e. Memperbarui posisi *centroid* berdasarkan rata-rata data dalam *cluster*.
f. Mengulangi langkah 3–5 hingga *centroid* tidak mengalami perubahan signifikan.

Euclidean distance dirumuskan sebagai berikut:

$$d(x_i, c_j) = \sqrt{\sum_{k=1}^{n} (x_{ik} - c_{jk})^2} \quad (1)$$

di mana:

x_i: data ke-i
c_j: adalah *centroid cluster* ke-j
n: adalah jumlah atribut

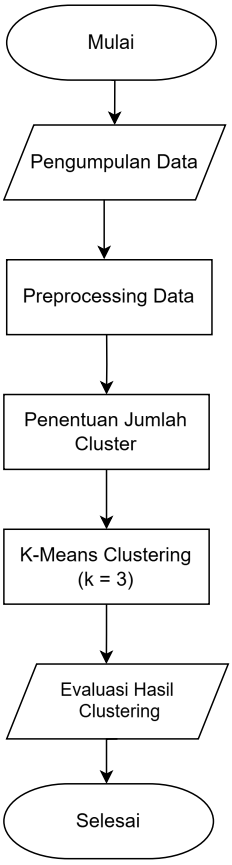
Tujuan algoritma K-Means adalah meminimalkan fungsi objektif:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_i - c_j \|^2 \quad (2)$$

Parameter yang digunakan dalam penelitian ini adalah:
Jumlah *cluster*: k = 3
Metode jarak: *Euclidean distance*

2.4. Alur Proses Clustering

Alur proses *clustering* dalam penelitian ini ditunjukkan pada Gambar 2. *Flowchart* tersebut menggambarkan tahapan penelitian secara keseluruhan mulai dari pengumpulan dataset hingga evaluasi hasil *clustering*.



Gambar 2. Alur Proses Clustering

2.5. Evaluasi Kinerja Algoritma

Evaluasi kinerja algoritma *clustering* dilakukan menggunakan dua metode, yaitu Elbow Method dan Silhouette Score.

a. Elbow Method

Elbow Method digunakan untuk menentukan jumlah *cluster* optimal dengan mengamati perubahan nilai *inertian* (*Sum of Squares*) terhadap jumlah *cluster*, di mana titik siku menunjukkan jumlah *cluster* yang

paling optimal. Hal ini telah banyak diterapkan dalam berbagai studi *clustering* seperti dalam analisis *crime datasets* dan K-Means clustering generik [9].

b. Silhouette Score

Silhouette Score digunakan untuk mengukur kualitas pemisahan antar cluster dengan rentang nilai antara -1 hingga 1. Nilai yang mendekati 1 menunjukkan struktur *cluster* yang baik dan metode ini sering digunakan bersama Elbow Method untuk mengevaluasi hasil *cluster* yang optimal [10].

Rumus Silhouette Score dirumuskan sebagai berikut:

$$s(i) = \frac{b(i) - a(i)}{\max\big(a(i), b(i)\big)} \tag{3}$$

di mana:
$a(i)$: jarak rata-rata data ke-i dengan data lain dalam *cluster* yang sama
$b(i)$: jarak rata-rata data ke-i dengan data pada *cluster* terdekat lainnya

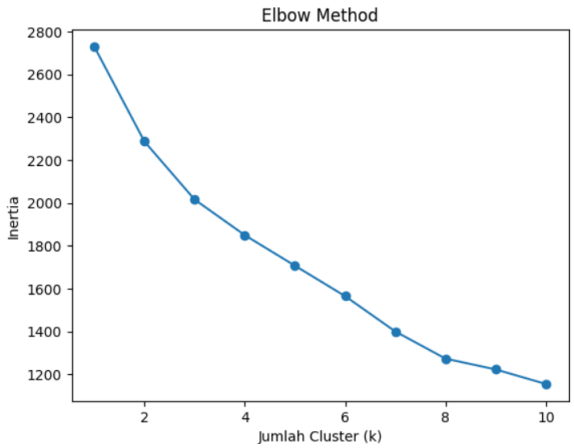
3. HASIL DAN PEMBAHASAN

3.1 Hasil Penelitian

Bagian ini menyajikan temuan penelitian yang diperoleh dari pengolahan data diabetes menggunakan metode K-Means Clustering. Hasil yang disajikan meliputi tahapan akuisisi data, pra-pemrosesan, pembentukan klaster, dan evaluasi kualitas klaster berdasarkan indikator kesehatan yang digunakan. Meskipun data mencatat adanya tren kenaikan penderita diabetes melitus secara nasional yang terus berlanjut [11]. Namun, untuk menghasilkan hasil yang lebih spesifik, penelitian ini berkonsentrasi pada pembagian risiko berdasarkan kondisi kesehatan terbaru [12].

3.1.1. Penentuan Jumlah Cluster Optimal Menggunakan Elbow Method

Metode Elbow digunakan untuk menentukan ukuran *cluster* optimal. Jumlah *cluster* yang digunakan dalam penelitian ini adalah tiga karena *Graphic* Elbow menunjukkan titik siku pada cluster k = 3.



Gambar 3. Grafik Elbow *Method* untuk Menentukan Jumlah *Cluster* Optimal

3.1.2. Evaluasi Kualitas Clustering Menggunakan Silhouette Score

Evaluasi kualitas hasil *clustering* juga dilakukan menggunakan metode Silhouette Score untuk mengukur tingkat kohesi dan separasi antar *cluster* yang terbentuk. Berdasarkan hasil perhitungan dengan jumlah *cluster* k = 3, diperoleh nilai Silhouette Score sebesar 0,178. Nilai ini menunjukkan bahwa terdapat tumpang tindih antar *cluster*, namun struktur pengelompokan masih dapat diinterpretasikan dan sesuai dengan karakteristik data medis yang memiliki indikator kesehatan yang saling beririsan. Meskipun skor pemisahan *cluster* tidak maksimal, namun algoritma ini terbukti andal dalam mengklasifikasikan tingkat keparahan pasien diabetes pada implementasi sistem yang adaptif [13]. Selain itu, prosedur ini bukan sekadar mengolah data, melainkan bertujuan mengoptimalkan peran atribut klinis dalam proses klastering [14]. Dengan demikian, K-Means tetap

menjadi solusi relevan untuk memetakan risiko kesehatan melalui pendekatan penilaian nutrisi yang inovatif [15].

Tabel 1. Ringkasan Karakteristik dan Kategori Risiko Setiap *Cluster*

Cluster	Glucose	BloodPressure	Insulin	BMI	Age	Outcome	Kategori Risiko
0	137,23	74,55	200,56	37,81	32,82	0,54	Risiko Sedang–Tinggi
1	102,14	66,42	110,66	28,87	26,45	0,13	Risiko Rendah
2	136,87	78,59	126,91	31,65	44,76	0,50	Risiko Sedang

3.2. Pembahasan

Berdasarkan dari penemuan hasil clustering ini adalah data pasien dapat dibagi menjadi tiga kelompok dengan ciri-ciri ketiga variabel *feature* yang berbeda satu sama lain menggunakan algoritma K-Means. Kelompok 1 menunjukan persentase *Outcome* yang relatif rendah, yaitu 0.13, dengan nilai rerata Glucose, Insulin, dan BMI yang lebih rendah dibandingkan dengan kelompok yang lainnya.

Cluster 0 memiliki proporsi *Outcome* terbesar dengan nilai 0.54 yang memiliki nilai rata-rata Glucose tinggi dan BMI. Atribut ini mengisyaratkan bahwa orang yang tergolong dalam cluster tersebut memiliki karakteristik lebih tinggi untuk mengalami penyakit diabetes dibandingkan pada *cluster* yang lain, yang membuat orang tersebut masuk ke dalam kategori Risiko Sedang hingga tinggi.

Sementara itu, cluster 2 memiliki proporsi Outcome sebesar 0,50 dengan nilai Glucose yang relatif tinggi serta rata-rata usia tertinggi di antara seluruh *cluster*. Ini berarti bahwa factor usia menjadi salah satu factor yang berkontribusi pada risiko diabetes bagi rombongan tersebut karena dengan demikian *cluster* 2 dapat dikategorikan ke dalam golongan yang berisiko sedang.

Nilai Silhouette Score sebesar 0,178 mengindikasikan bahwa kualitas pemisahan antar *cluster* belum terbentuk secara optimal. Kondisi ini di sebabkan oleh sifat variabel risiko diabetes yang multifaset, dengan instrumen seperti konsentrasi glukosa, indeks massa tubuh, insulin, serta usia pemakaian batas nilai yang saling menggantung. Meskipun demikian, hasil clustering masih menginformasikan data yang berguna dalam mengidentifikasi kategori risiko diabetes kategori rendah, sedang, serta sedang hingga tinggi.

Secara keseluruhan, hasil penelitian ini menunjukkan bahwa pendekatan *unsupervised learning* menggunakan algoritma K-Means dapat digunakan untuk mengidentifikasi kelompok risiko diabetes berdasarkan karakteristik klinis pasien. Hasil clustering ini diharapkan dapat menjadi dasar untuk analisis lebih lanjut atau sebagai pendukung dalam pengambilan keputusan di bidang kesehatan.

4. KESIMPULAN

Penelitian ini bertujuan untuk menganalisis pengelompokan faktor risiko diabetes berdasarkan indikator kesehatan menggunakan metode K-Means Clustering. Berdasarkan hasil pengolahan data yang telah melalui tahapan akuisisi, *preprocessing*, penanganan *outlier*, dan normalisasi, metode K-Means berhasil mengelompokkan data pasien ke dalam tiga *cluster* dengan karakteristik risiko yang berbeda.

Hasil *clustering* menunjukkan bahwa *cluster* yang terbentuk dapat diinterpretasikan sebagai kelompok risiko rendah, risiko sedang, dan risiko sedang hingga tinggi, berdasarkan perbedaan nilai rata-rata indikator klinis serta proporsi variabel *Outcome*. Temuan ini menunjukkan bahwa pendekatan *unsupervised learning* mampu mengidentifikasi pola pengelompokan data yang bermakna secara klinis tanpa menggunakan label kelas pada tahap pembentukan model.

Meskipun evaluasi menggunakan Silhouette Score menunjukkan bahwa kualitas pemisahan antar *cluster* belum terbentuk secara optimal, hasil pengelompokan yang diperoleh tetap memberikan informasi yang relevan dalam memetakan profil risiko diabetes berdasarkan indikator kesehatan. Dengan demikian, metode yang digunakan dinilai mampu mencapai tujuan penelitian. Sebagai arah penelitian selanjutnya, pengembangan dapat dilakukan dengan penerapan metode *clustering* lain atau perluasan dataset guna memperoleh hasil pengelompokan yang lebih optimal.

DAFTAR PUSTAKA

[1] H. Irbah, N. Zara, and R. Ikhsan, "Analisis Faktor Risiko Pasien Diabetes Mellitus di Puskesmas Dewantara Kecamatan Dewantara Kabupaten Aceh Utara," *Galen. J. Kedokt. dan Kesehat. Mhs. Malikussaleh*, vol. 1, no. 1, p. 1, 2022, doi: 10.29103/jkkmm.v1i1.8030.

[2] R. Gestavito, A. Id Hadiana, F. Rakhmat Umbara, and U. Jenderal Achmad Yani Jl Terusan Jenderal Sudirman, "Pengelompokan Tingkat Risiko Penyakit Diabetes Melitus Menggunakan Algoritma K-Means Clustering," *J. Masy. Inform. Unjani*, vol. 8, no. 1, pp. 16–35, 2024.

[3] B. Usia, K. Glukosa, and T. Darah, "Analisis Klaster Pasien Diabetes Menggunakan Algoritma K-Means," vol. 4, no. 2, pp. 374–378, 2025.

[4] M. K. Basyir, "Klastering Penyakit Diabetes dengan Metode K-Means," *BIIKMA Bul. Ilm. Ilmu Komput. dan Multimed.*, vol. 2, no. 5, pp. 904–909, 2025.
[5] D. D. Onthoni *et al.*, "Clustering-based risk strati fi cation of prediabetes populations : Insights from the Taiwan and UK Biobanks," vol. 16, no. 1, pp. 25–35, 2025, doi: 10.1111/jdi.14328.
[6] E. Jurnal, F. Feronika, N. A. Ramdhan, R. Muhammad, and H. Bhakti, "Penggunaan Algoritma K-Means dalam Pengelompokan Pasien Diabetes Mellitus Berdasarkan Parameter Klinis di Puskesmas Brebes," vol. 18, no. 1, pp. 356–366, 2025.
[7] R. M. Carrillo- and M. Castillo-, "Clusters of people with type 2 diabetes in the general population : unsupervised machine learning approach using national surveys in Latin America and the Caribbean," pp. 1–8, 2021, doi: 10.1136/bmjdrc-2020-001889.
[8] A. Simanjuntak and M. S. Hasibuan, "Application of PCA and K-Means Clustering Methods to Identify Diabetes Mellitus Patient Groups Based on Risk Factors," *Prism. Sains J. Pengkaj. Ilmu dan Pembelajaran Mat. dan IPA IKIP Mataram*, vol. 11, no. 4, p. 1002, 2023, doi: 10.33394/j-ps.v11i4.9263.
[9] N. T. M. Sagala and A. A. S. Gunawan, "Discovering the Optimal Number of Crime Cluster Using Elbow, Silhouette, Gap Statistics, and NbClust Methods," *ComTech Comput. Math. Eng. Appl.*, vol. 13, no. 1, pp. 1–10, 2022, doi: 10.21512/comtech.v13i1.7270.
[10] P. Khairani Ritonga and M. S. Hasibuan, "Analisis Perbandingan Silhouette dengan Elbow pada Algoritma K-Means dan DBSCAN," *Metik J.*, vol. 9, p. 2025, 2025, doi: 10.47002/metik.v9i1.1027.
[11] Z. Nurman, Ismanilda, D. Yuska, and I. Y. Puri, "Uji Kadar Protein, Serat dan Organoleptik Cake Tempe sebagaiAlternatif Snack Pasien Diabetes Melitus (DM)," *J. Ilm. Kesehat.*, vol. 6, no. 3, pp. 352–364, 2024.
[12] R. A. Sigit, U. Rio, L. Efrizoni, and E. Ali, "Penerapan K-Means Clustering untuk Mengelompokkan Risiko Diabetes Berdasarkan Gaya Hidup dan Kesehatan," vol. 12, no. 4, pp. 8–18, 2025.
[13] S. Maharani, Y. Wendra, and R. Rahim, "TOFEDU : The Future of Education Journal The Implementation of the K-Means Clustering Algorithm Based on the Severity Level of Diabetes in Patients Using a Website Platform," vol. 4, no. 7, pp. 3749–3761, 2025.
[14] R. Ishak, "Optimalisasi Seleksi Atribut K-Means Menggunakan Correlation Matrix pada Clustering Penyakit Pasien," *Jambura J. Electr. Electron. Eng.*, vol. 7, no. 2, pp. 2–9, 2025.
[15] I. Darmayanti, D. Mustofa, N. Hidayati, and I. Saputri, "K-Means and Fuzzy C-Means Cluster Food Nutrients for Innovative Diabetes Risk Assessment," *Sistemasi*, vol. 13, no. 5, p. 2175, 2024, doi: 10.32520/stmsi.v13i5.4552.