Penerapan K-Means Clustering untuk Mengelompokkan Risiko Diabetes Berdasarkan Gaya Hidup dan Kesehatan

Rapel Aprilius Sigit*[1], Unang Rio[2], Lusiana Efrizoni[3], Edwar Ali[4]
[1,2,3,4]Universitas Sains dan Teknologi Indonesia; Jl. Purwodadi Indah, Kel. Sialangmunggu
[1,2,3,4]Jurusan Teknik Informatika, Riau
e-mail: *[1]rafelas04@gmail.com, [2]unangrio@usti.ac.id, [3]lusiana@usti.ac.id, [4]edwarali@usti.ac.id

Abstrak

Diabetes melitus merupakan penyakit kronis yang prevalensinya terus meningkat secara global seiring dengan perubahan pola hidup masyarakat modern. Deteksi dini terhadap risiko diabetes menjadi krusial dalam upaya pencegahan dan pengurangan komplikasi jangka panjang. Penelitian ini bertujuan untuk mengelompokkan individu berdasarkan tingkat risiko diabetes menggunakan algoritma K-Means Clustering dengan mempertimbangkan atribut gaya hidup dan kondisi kesehatan. Data yang digunakan bersumber dari platform Kaggle, terdiri atas 5.452 entri dan 22 atribut. Proses pre-processing mencakup data cleaning, normalization, serta feature selection secara manual. Penentuan jumlah Cluster optimal menggunakan metode Elbow menunjukkan hasil terbaik pada k = 3. Evaluasi kualitas Cluster dilakukan dengan Davies-Bouldin Index (DBI), yang menghasilkan nilai sebesar 0,7678, menandakan kualitas pengelompokan yang cukup baik. Hasil akhir membentuk tiga Cluster risiko, yaitu rendah, sedang, dan tinggi, dengan distribusi masing-masing 424, 819, dan 615 data. Segmentasi ini dapat menjadi dasar bagi instansi kesehatan dalam merancang intervensi preventif yang lebih terarah dan berbasis data.

Kata kunci— Data Mining, Davies-Bouldin Index, Diabetes, Elbow Method, Gaya Hidup, K-Means Clustering

Abstract

Diabetes mellitus is a chronic disease with a globally increasing prevalence, driven by modern lifestyle changes. Early detection of diabetes risk is crucial in preventing and mitigating long-term complications. This study aims to cluster individuals based on their diabetes risk levels using the K-Means Clustering algorithm by considering lifestyle and health condition attributes. The dataset used was obtained from the Kaggle platform, consisting of 5,452 entries and 22 attributes. The pre-processing stage involved data cleaning, normalization, and manual feature selection. The optimal number of clusters was determined using the Elbow Method, which indicated the best result at k = 3. Cluster quality evaluation was performed using the Davies-Bouldin Index (DBI), which yielded a score of 0.7678, indicating a reasonably good level of cluster compactness and separation. The final output formed three risk clusters: low, medium, and high, with distributions of 424, 819, and 615 records, respectively. This segmentation is expected to serve as a basis for healthcare institutions in designing more targeted and data-driven preventive interventions.

Keywords— Data Mining, Davies-Bouldin Index, Diabetes, Elbow Method, Lifestyle, K-Means Clustering

1. PENDAHULUAN

Diabetes melitus adalah penyakit kronis yang jumlah penderitanya terus meningkat secara signifikan, termasuk di Indonesia. Hal ini banyak dipengaruhi oleh perubahan pola hidup masyarakat modern seperti kurangnya aktivitas fisik, konsumsi makanan tidak sehat, serta tingginya tingkat stres. Kondisi tersebut menjadikan diabetes sebagai ancaman serius karena berpotensi menimbulkan berbagai komplikasi, seperti gangguan jantung, kerusakan ginjal, hingga gangguan penglihatan[1].

Deteksi risiko diabetes sejak dini menjadi langkah penting untuk menekan dampak jangka panjang dan menurunkan beban biaya pengobatan. Salah satu pendekatan yang dapat dimanfaatkan adalah teknik pengelompokan data berdasarkan tingkat risiko menggunakan metode analisis data. Dalam hal ini, data mining sangat efektif dalam mengidentifikasi pola tersembunyi dari kumpulan data besar, terutama dalam bidang kesehatan[2].

K-Means Clustering merupakan salah satu metode pembelajaran tanpa pengawasan (unsupervised learning) yang berguna dalam mengelompokkan data berdasarkan kesamaan karakteristik. Dalam konteks risiko diabetes, algoritma ini memungkinkan pemetaan kelompok individu berdasarkan atribut gaya hidup dan kondisi kesehatan mereka sehingga distribusi risiko dapat dipahami dengan lebih sistematis[3].

Penelitian ini bertujuan untuk melakukan segmentasi risiko diabetes menggunakan algoritma K-Means Clustering berdasarkan atribut gaya hidup dan kesehatan. Atribut yang dianalisis mencakup BMI, aktivitas fisik, kesehatan mental dan fisik, serta variabel pendukung lainnya. Selain itu, penelitian ini juga mengevaluasi kualitas pengelompokan menggunakan Davies-Bouldin Index (DBI) sebagai metrik untuk menilai struktur cluster yang terbentuk[4].

Peningkatan jumlah penderita diabetes juga memperbesar tantangan dalam keterbatasan layanan medis untuk melakukan skrining massal. Oleh karena itu, pendekatan berbasis teknologi informasi menjadi penting untuk mengidentifikasi kelompok dengan risiko tinggi. Strategi ini mampu memberikan gambaran awal terhadap distribusi risiko tanpa perlu prosedur klinis yang mahal dan memakan waktu[5].

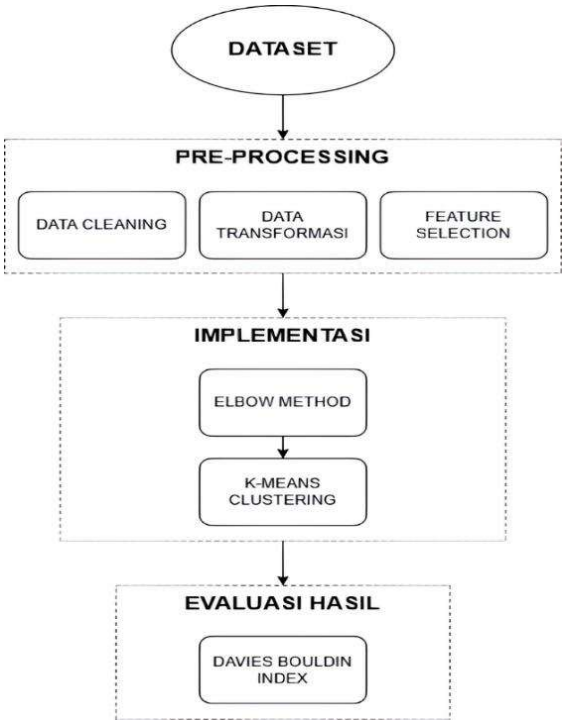
Teknik data mining merupakan bagian dari kecerdasan buatan yang dapat mengekstraksi pola tersembunyi dari data berukuran besar. Dalam konteks kesehatan, metode ini sangat berguna untuk menemukan keterkaitan antara gayahidup dan risiko penyakit kronis seperti diabetes. Algoritma seperti K-Means Clustering mampu membagi individu ke dalam kelompok berdasarkan kesamaan fitur tanpa memerlukan label sebelumnya, sehingga cocok diterapkan pada data yang belum teranotasi[6].

Dengan menggabungkan metode clustering dan atribut gaya hidup, pendekatan ini memberikan potensi untuk menganalisis penyebaran risiko diabetes secara lebih prediktif. Hasil segmentasi yang dihasilkan dapat dimanfaatkan sebagai dasar dalam penyusunan program intervensi kesehatan yang lebih tepat sasaran. Penelitian ini juga memberikan kontribusi baru dengan menggunakan kumpulan atribut yang cukup komprehensif dan evaluasi kuantitatif terhadap kualitas cluster yang terbentuk, sehingga dapat memperkaya literatur teknologi kesehatan serta menjadi pijakan awal dalam pengembangan sistem pendukung keputusan berbasis data[7].

Selain itu, penelitian ini juga menyoroti potensi penerapan metode ini dalam sistem informasi kesehatan digital yang dapat terintegrasi dengan layanan kesehatan masyarakat. Melalui sistem ini, instansi terkait dapat lebih mudah memantau kelompok populasi berisiko dan menyusun strategi pencegahan yang lebih efektif. Integrasi antara teknik data mining dan kebijakan kesehatan publik tidak hanya meningkatkan efisiensi, tetapi juga memungkinkan pelayanan kesehatan yang lebih personal dan sesuai kebutuhan individu[8].

Penerapan algoritma K-Means Clustering dalam studi ini juga menambah kontribusi terhadap pengembangan metode analisis risiko berbasis data yang bersifat non-invasif. Dengan menggunakan data sekunder yang tersedia secara publik, pendekatan ini menjadi lebih ekonomis dan memungkinkan untuk direplikasi di berbagai wilayah dan konteks. Hal ini membuka peluang penerapan lebih luas, baik di lembaga pendidikan, fasilitas layanan kesehatan, maupun oleh peneliti independen, sebagai bagian dari upaya deteksi dan pencegahan penyakit kronis[9].

2. METODE PENELITIAN



Gambar 1 Tahapan Penelitian

2.1 Dataset

Penelitian ini menggunakan data yang diperoleh dari situs Kaggle dalam format CSV. Dataset mencakup 5.452 baris data, masing-masing mewakili satu individu, dengan 22 atribut yang mencerminkan aspek gaya hidup dan kondisi kesehatan. Atribut-atribut tersebut meliputi indikator seperti status diabetes, tekanan darah, kadar kolesterol, indeks massa tubuh (BMI), kebiasaan merokok, aktivitas fisik, konsumsi alkohol, serta variabel demografis seperti usia, pendidikan, dan pendapatan. Untuk keperluan analisis, atribut diklasifikasikan menjadi dua kelompok utama: atribut gaya hidup dan atribut Kesehatan[10].

2.2 Pre-Processing

Tahapan pre-processing bertujuan untuk mempersiapkan data mentah agar siap dianalisis. Proses ini meliputi tiga tahap utama.

2.2.1 Data Cleaning

Langkah awal dalam tahap pre-processing adalah proses pembersihan data (data cleaning), yang bertujuan untuk meningkatkan kualitas data dan menghilangkan potensi gangguan pada analisis. Proses ini mencakup penghapusan data yang bersifat duplikat, pengisian atau penghapusan nilai kosong (missing values), serta identifikasi dan eliminasi nilai pencilan (outlier). Metode yang digunakan untuk mendeteksi outlier adalah pendekatan Interquartile Range (IQR), di mana data dianggap sebagai outlier apabila berada di luar rentang Q1 − 1,5×IQR atau Q3 + 1,5×IQR. Penghapusan nilai-nilai ekstrem ini penting agar distribusi data menjadi lebih representatif dan model clustering tidak terganggu oleh keberadaan data yang menyimpang secara drastis[11].

2.2.2 Data Transformation

Tahap berikutnya adalah normalisasi atau transformasi data numerik. Teknik yang digunakan adalah *StandardScaler*, yaitu metode normalisasi yang mengubah setiap fitur menjadi distribusi dengan nilai rata-rata (mean) nol dan deviasi standar satu. Rumus transformasi dengan metode *StandardScaler* ditunjukkan pada **Persamaan (1)**[12].

$$X^1 = \frac{X-\mu}{\sigma} \qquad (1)$$

Dimana:

X^1 : Nilai data setelah dinormalisasi

X : Nilai asli dari data

μ : Nilai rata-rata (mean)

σ : Standar deviasi

2.2.3 Feature Selection

Langkah terakhir dalam tahap pre-processing adalah pemilihan fitur atau feature selection. Dari total 22 atribut yang tersedia dalam dataset, tidak semuanya digunakan dalam proses clustering. Pemilihan atribut dilakukan secara manual berdasarkan kajian literatur medis dan analisis korelasi awal terhadap variabel risiko diabetes. Lima atribut yang dipilih sebagai representasi utama dalam proses clustering adalah: indeks massa tubuh (BMI), tingkat aktivitas fisik, usia, kondisi kesehatan fisik, dan kondisi kesehatan mental. Kelima atribut ini dinilai memiliki pengaruh paling signifikan terhadap risiko seseorang terkena diabetes, sehingga layak dijadikan dasar dalam segmentasi[10].

2.3 Implementasi

Setelah seluruh tahapan pre-processing selesai dilaksanakan dan data berada dalam kondisi optimal untuk dianalisis.

2.3.1 Elbow Method

Untuk keperluan ini digunakan pendekatan Elbow Method, yang merupakan teknik populer dalam penentuan jumlah cluster optimal dalam analisis K-Means. Metode ini dilakukan dengan menghitung nilai Within-Cluster Sum of Squares (WCSS) untuk sejumlah nilai k yang berbeda, dalam penelitian ini dari k = 1 hingga k = 6. Hasil perhitungan tersebut kemudian divisualisasikan dalam bentuk grafik, di mana titik "siku" atau elbow point—yaitu titik ketika penurunan WCSS mulai melambat—diidentifikasi sebagai indikator jumlah cluster yang paling efisien. Rumus *Elbow Method* ditunjukkan pada **Persamaan (2)**[13].

$$WCSS = \sum_{i=1}^{k} \sum_{x \in Ci} \| x - ui \| 2 \qquad (2)$$

Dimana:

k : Jumlah *Cluster*
x : Data dalam *Cluster* C_i
u_i : *Centroid* dari *Cluster* C_i
$\| x - ui \| 2$: Jarak Euclidean kuadrat antara data dan centroid

2.3.2 K-Means Clustering

Setelah jumlah cluster ditentukan, algoritma K-Means mulai diterapkan pada dataset yang telah dinormalisasi. Prosedur K-Means melibatkan beberapa langkah utama, yakni: inisialisasi awal centroid secara acak, pengelompokan data berdasarkan jarak Euclidean terdekat dari masing-masing centroid, lalu pembaruan posisi centroid berdasarkan rata-rata dari anggota dalam masing-masing cluster. Rumus *K-Means Clustering* ditunjukkan pada **Persamaan (3)**[14].

$$d(X_i, C_k) = \sqrt{\sum_{j=1}^{n}(X_{ij} - C_{kj})}\, 2 \qquad (3)$$

Dimana:
$d(X_i, C_k)$: Jarak antara data X_i dan *centroid* C_k
X_{ij} : Nilai fitur ke-j dari data i
C_{kj} : Nilai fitur ke-j dari *centroid* k
n : Jumlah fitur dalam Dataset

2.4 Evaluasi Hasil

Untuk menilai kualitas hasil pengelompokan yang dihasilkan oleh algoritma K-Means, digunakan metode evaluasi yang bernama *Davies-Bouldin Index* (DBI). Metrik ini bertujuan untuk mengevaluasi sejauh mana cluster yang terbentuk memiliki pemisahan yang jelas dan struktur internal yang rapat. Secara teknis, DBI menghitung rata-rata rasio antara jarak antar centroid dengan tingkat dispersi (variansi) di dalam masing-masing cluster. Semakin kecil nilai DBI yang diperoleh, semakin baik performa clusterisasi yang dihasilkan, karena menunjukkan bahwa anggota dalam satu cluster memiliki kesamaan yang tinggi dan berbeda secara signifikan dengan anggota dari cluster lain. Rumus *Davies-Bouldin Index* ditunjukkan pada **Persamaan (4)**[15].

$$D\ B\ I = \frac{1}{k} \sum_{i=1}^{k} \underset{j \neq i}{max} \left(\frac{s_i + s_j}{d_{i,j}}\right) \qquad (4)$$

Dimana:
s_i : Rata-rata jarak data dalam *Cluster* i ke *centroid Cluster* i.
$d_{i,j}$: Jarak antara *centroid Cluster* i dan *centroid Cluster* j.

3. HASIL DAN PEMBAHASAN

3.1 Dataset

Dataset yang digunakan dalam tahap pengujian dan analisis penelitian ini diambil dari platform berbasis data publik, yaitu Kaggle. Dataset tersebut terdiri atas 5.452 entri yang masing-masing mewakili satu individu, serta mencakup 22 atribut yang memberikan informasi menyeluruh tentang status kesehatan dan kebiasaan gaya hidup para responden.

Beberapa atribut utama yang digunakan antara lain mencakup status diagnosis diabetes, tekanan darah tinggi, kadar kolesterol, indeks massa tubuh (Body Mass Index), kebiasaan merokok, riwayat penyakit jantung dan stroke, tingkat aktivitas fisik, serta konsumsi makanan sehat seperti buah dan sayur. Kombinasi atribut-atribut ini memungkinkan dilakukannya pemetaan profil risiko kesehatan dengan pendekatan segmentasi menggunakan algoritma K-Means dan hasil ini dapat diamati pada **Gambar 2.**

	Diabetes_Biner	Tekanan_Darah_Tinggi	Kolesterol_Tinggi	Cek_Kolesterol	BMI	Perokok	Stroke	Serangan_Jantung_Penyakit_Jantung	Aktivitas_Fisik	Bua
0	0.0	1.0	1.0	1.0	40.0	1.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	25.0	1.0	0.0	0.0	1.0	
2	0.0	1.0	1.0	1.0	28.0	0.0	0.0	0.0	0.0	
3	0.0	1.0	0.0	1.0	27.0	0.0	0.0	0.0	1.0	
4	0.0	1.0	1.0	1.0	24.0	0.0	0.0	0.0	1.0	
...	...	...	...	...	...	...	...	...	...	
5447	0.0	0.0	0.0	1.0	28.0	0.0	0.0	0.0	1.0	
5448	0.0	0.0	0.0	1.0	40.0	1.0	0.0	0.0	0.0	
5449	0.0	0.0	0.0	1.0	27.0	1.0	0.0	0.0	1.0	
5450	0.0	1.0	1.0	1.0	34.0	1.0	1.0	0.0	1.0	
5451	0.0	0.0	0.0	1.0	39.0	1.0	0.0	0.0	1.0	

5452 rows × 22 columns

Gambar 2 Pengolahan Data

Gambar 2 menampilkan cuplikan dari dataset sebelum dilakukan proses pre-processing. Terlihat bahwa data bersifat numerik biner (0 dan 1) untuk sebagian besar atribut kategorikal, dan numerik kontinu untuk atribut seperti BMI.

3.2 Pre-Processing

Tahapan pre-processing dilakukan untuk memastikan bahwa data yang digunakan dalam proses clustering memiliki kualitas yang baik dan siap dianalisis menggunakan algoritma K-Means. Langkah pertama adalah data cleaning, yaitu dengan mengidentifikasi dan menghapus entri yang memiliki nilai kosong (missing value) serta mendeteksi dan menghilangkan nilai pencilan (outlier) menggunakan metode Interquartile Range (IQR). IQR digunakan untuk menentukan batas bawah dan batas atas data, di mana nilai-nilai yang berada di luar rentang (Q1 - 1,5 × IQR) dan (Q3 + 1,5 × IQR) dianggap sebagai outlier dan dihapus dari dataset. Dari total 5.452 entri awal, sebanyak 327 entri dihapus karena mengandung nilai kosong atau outlier, sehingga menyisakan 5.125 entri yang layak untuk dianalisis lebih lanjut.

Setelah data dibersihkan, langkah berikutnya adalah melakukan normalisasi terhadap atribut numerik menggunakan metode StandardScaler. Metode ini mengubah skala data menjadi distribusi dengan nilai rata-rata 0 dan standar deviasi 1, sehingga semua fitur berada dalam skala yang sebanding. Normalisasi ini sangat penting karena algoritma K-Means menghitung jarak antar titik data menggunakan rumus Euclidean, yang sangat sensitif terhadap perbedaan skala antar fitur.

Langkah terakhir dalam pre-processing adalah feature selection, yaitu pemilihan atribut yang paling relevan untuk dianalisis. Seleksi dilakukan secara manual berdasarkan tinjauan literatur, analisis korelasi, dan pemahaman terhadap domain kesehatan. Dari 22 atribut awal yang tersedia, dipilih lima atribut utama yang dinilai paling mewakili faktor risiko diabetes, yaitu indeks massa tubuh (BMI), aktivitas fisik, usia, kesehatan mental, dan kesehatan fisik. Atribut-atribut ini dianggap memiliki kontribusi signifikan dalam mengidentifikasi kelompok risiko diabetes berdasarkan karakteristik gaya hidup dan kondisi kesehatan individu.

Dari total 5.452 entri awal, sebanyak 327 entri dihapus karena mengandung nilai kosong atau outlier, sehingga menyisakan 5.125 entri yang layak untuk dianalisis lebih lanjut. Namun, setelah dilakukan proses *feature selection* terhadap lima atribut utama (BMI, aktivitas fisik, usia, kesehatan fisik, dan kesehatan mental), serta penyaringan data yang hanya memiliki nilai lengkap pada kelima atribut tersebut, jumlah entri berkurang menjadi **1.858**. Dataset inilah yang kemudian digunakan dalam tahap clustering hasil ini dapat diamati pada **Gambar 3.**

Data setelah pre-processing:

	BMI	Aktivitas_Fisik	Usia	Kesehatan_Mental	Kesehatan_Fisik
0	-0.113542	0.455277	1.198175	-0.393685	-0.396581
1	-0.724435	0.455277	1.198175	1.845499	-0.396581
2	-0.520804	0.455277	0.852060	-0.393685	0.708967
3	-0.724435	-2.196463	0.159831	-0.393685	-0.396581
4	-1.335327	0.455277	0.852060	-0.393685	-0.396581
...	...	...	...	...	...
1853	-0.724435	0.455277	-1.570742	-0.393685	0.708967
1854	0.700981	0.455277	-0.186283	-0.393685	-0.396581
1855	-0.520804	0.455277	-0.532398	-0.393685	0.708967
1856	-0.724435	0.455277	0.159831	-0.393685	-0.396581
1857	2.330027	0.455277	-1.570742	1.099104	-0.396581

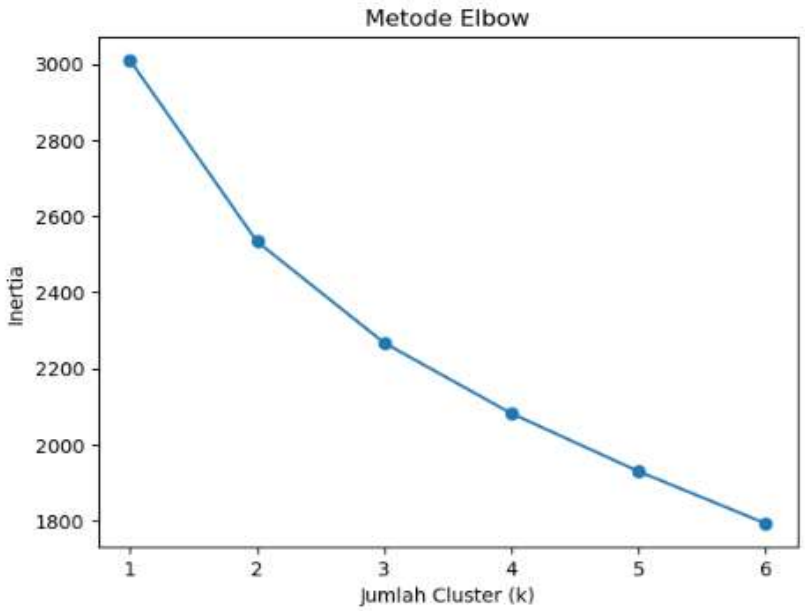
[1858 rows x 5 columns]

Gambar 3 Hasil Pre-Processing Data

Dari **Gambar 3** dapat diamati bahwa hasil akhir dari tahapan pre-processing menghasilkan sebuah dataset baru yang terdiri dari 1.858 entri dengan lima atribut utama, yaitu Body Mass Index (BMI), aktivitas fisik, usia, kesehatan mental, dan kesehatan fisik. Seluruh atribut dalam dataset ini telah melalui proses normalisasi menggunakan metode StandardScaler, yang bertujuan untuk menyeragamkan skala setiap fitur agar memiliki rata-rata nol dan deviasi standar satu. Hasil normalisasi tersebut menghasilkan nilai-nilai seperti −0,72, 0,45, dan 1,19 yang menggambarkan posisi relatif masing-masing data terhadap distribusi keseluruhan populasi.

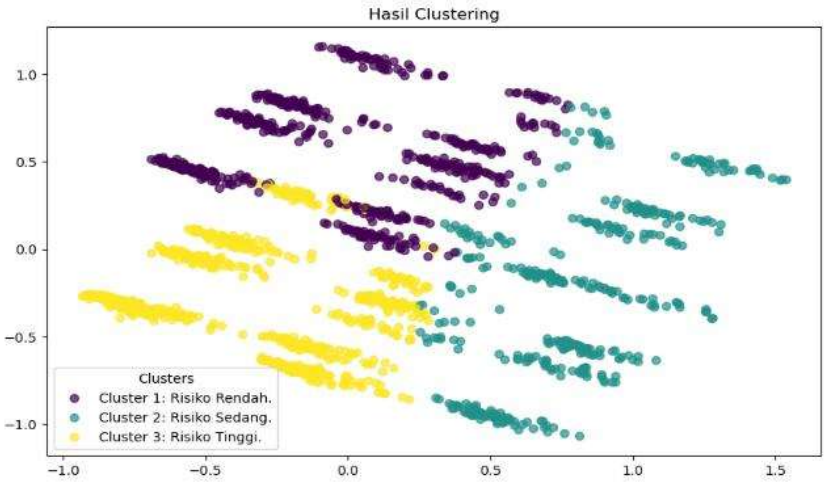
3.3 Implementasi

Dalam rangka menentukan jumlah cluster yang paling sesuai dalam analisis clustering, penelitian ini menerapkan pendekatan Elbow Method sebagai strategi penentu parameter k. Langkah ini dilakukan dengan menghitung nilai Within-Cluster Sum of Squares (WCSS) untuk rentang nilai k dari 1 hingga 6. Hasil dari proses ini divisualisasikan ke dalam grafik untuk mengamati tren penurunan nilai WCSS. Metode Elbow divisualisasikan pada **Gambar 4.**



Gambar 4 Jumlah Cluster Optimal

Gambar 4 menunjukkan hasil analisis menggunakan metode Elbow, di mana terlihat adanya penurunan signifikan pada nilai Within-Cluster Sum of Squares (WCSS) hingga titik k = 3. Setelah titik tersebut, penurunan nilai WCSS mulai melandai, yang mengindikasikan bahwa tambahan jumlah cluster setelah k = 3 tidak memberikan peningkatan signifikan dalam kualitas pemisahan data. Pola ini menunjukkan bahwa jumlah cluster sebesar tiga (k = 3) merupakan pilihan yang optimal karena mampu memberikan keseimbangan antara kompleksitas model dan efektivitas pengelompokan data. Berdasarkan hasil tersebut, algoritma K-Means kemudian diterapkan dengan k = 3 sebagai jumlah cluster. Proses clustering yang dilakukan menghasilkan tiga kelompok utama yang merepresentasikan variasi tingkat risiko diabetes pada individu dalam dataset dan hasil ini dapat diamati pada **Gambar 5.**



Gambar 5 Hasil Cluster Data

Setelah jumlah cluster optimal ditentukan, algoritma K-Means diterapkan pada data yang telah diproses. **Gambar 5** menunjukkan hasil visualisasi proses clustering menggunakan

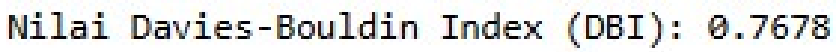
nilai k = 3, di mana data berhasil dikelompokkan ke dalam tiga kategori tingkat risiko diabetes. Warna ungu mewakili Cluster 1 (Risiko Rendah) dengan total 424 individu, warna hijau kebiruan menunjukkan Cluster 2 (Risiko Sedang) dengan 819 individu, dan warna kuning menggambarkan Cluster 3 (Risiko Tinggi) yang terdiri atas 615 individu. Dari ketiga kelompok tersebut, Cluster 2 merupakan kelompok dengan jumlah anggota terbanyak.

Individu yang termasuk dalam Cluster 2 (Risiko Sedang) menunjukkan karakteristik kombinasi antara gaya hidup yang kurang aktif dengan kondisi kesehatan yang mulai menurun. Mereka memiliki nilai BMI yang mendekati batas atas kategori normal, serta kondisi kesehatan fisik dan mental yang tidak optimal. Cluster ini merepresentasikan kelompok yang berada dalam fase transisi antara kondisi sehat dan risiko tinggi, sehingga perlu menjadi fokus utama dalam upaya intervensi preventif untuk mencegah peningkatan risiko diabetes di masa mendatang.

3.4 Evaluasi Hasil

Evaluasi terhadap hasil clustering dalam penelitian ini menggunakan metrik Davies-Bouldin Index (DBI), yaitu salah satu metode validasi internal yang berfungsi untuk mengukur seberapa baik kualitas pengelompokan data yang telah dilakukan. DBI menghitung rasio antara rata-rata jarak dalam cluster (intra-cluster dispersion) dan jarak antar cluster (inter-cluster distance).

Metrik ini memberikan informasi tentang seberapa kompak data dalam satu cluster dan seberapa jauh jarak antar cluster yang terbentuk. Semakin kecil nilai DBI, semakin baik kualitas clustering-nya, karena hal tersebut menandakan bahwa anggota dalam setiap cluster saling berdekatan dan antar cluster memiliki pemisahan yang jelas. Nilai ideal DBI adalah mendekati nol (0), yang menunjukkan clustering optimal-di mana tiap kelompok terbentuk secara kompak dan tidak tumpang tindih serta hasil ini dapat diamati pada **Gambar 6.**



Gambar 6 Hasil Output Davies-Boudin Index

Berdasarkan hasil yang ditampilkan pada **Gambar 6,** diperoleh nilai DBI sebesar 0,7678. Nilai ini mengindikasikan bahwa proses clustering menggunakan algoritma K-Means dalam penelitian ini mampu menghasilkan struktur kelompok yang baik. Angka DBI yang berada di bawah 1 merupakan indikasi bahwa pembentukan cluster telah memenuhi kriteria efektivitas dalam pengelompokan data. Hal ini berarti bahwa masing-masing cluster memiliki kesamaan karakteristik internal (kompak), serta antar cluster memiliki perbedaan yang signifikan (terpisah dengan baik).

4. KESIMPULAN

Berdasarkan hasil penelitian yang meliputi perumusan masalah, pengumpulan data, pre-processing, implementasi algoritma K-Means Clustering, dan evaluasi menggunakan Davies-Bouldin Index (DBI), dapat disimpulkan bahwa algoritma K-Means berhasil mengelompokkan data menjadi tiga cluster yang mewakili tingkat risiko diabetes yang berbeda. Penentuan jumlah cluster optimal dilakukan menggunakan Elbow Method dengan hasil terbaik pada k = 3. Dari ketiga cluster yang terbentuk, Cluster 2 merupakan kelompok dengan jumlah data terbanyak, yaitu sebanyak 819 individu, yang berdasarkan karakteristik gaya hidup dan kesehatannya dapat dikategorikan sebagai kelompok dengan risiko diabetes sedang. Evaluasi dengan DBI menghasilkan nilai sebesar 0.7678, yang menunjukkan kualitas pengelompokan yang baik, dengan pemisahan antar cluster yang jelas dan kekompakan data dalam masing-masing cluster.

Penelitian ini membuktikan bahwa K-Means dapat digunakan secara efektif untuk segmentasi risiko diabetes, dan dapat menjadi dasar dalam pengembangan sistem pendukung keputusan untuk pencegahan dan pengelolaan penyakit diabetes.

5. SARAN

Berdasarkan hasil penelitian ini, disarankan agar penelitian selanjutnya dapat memperluas cakupan atribut dengan menambahkan faktor genetika, riwayat keluarga, maupun pola konsumsi makanan yang lebih detail, serta mencoba metode clustering lain seperti *Fuzzy C-Means* atau *Hierarchical Clustering* untuk memperoleh hasil yang lebih komprehensif. Hasil pengelompokan risiko yang diperoleh juga diharapkan dapat dimanfaatkan oleh instansi kesehatan dalam merancang program pencegahan yang lebih tepat sasaran, khususnya bagi kelompok dengan risiko sedang dan tinggi, dengan dukungan teknologi analitik berbasis *data mining* yang mampu meningkatkan efisiensi proses skrining kesehatan masyarakat. Bagi masyarakat, temuan ini dapat menjadi dorongan untuk lebih menjaga gaya hidup sehat melalui pengaturan pola makan, peningkatan aktivitas fisik, serta pemeriksaan kesehatan secara berkala guna mendeteksi dini risiko diabetes dan mencegah dampak yang lebih serius di masa mendatang.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Universitas Sains dan Teknologi Indonesia, khususnya Program Studi Teknik Informatika, atas segala dukungan dan fasilitas yang diberikan selama proses penelitian ini. Ucapan terima kasih juga disampaikan kepada Bapak Unang Rio, M.Kom., selaku dosen pembimbing, atas bimbingan dan masukan berharga. Tak lupa, penulis juga menghargai semua pihak yang telah membantu baik secara langsung maupun tidak langsung hingga artikel ini dapat diselesaikan dengan baik.

DAFTAR PUSTAKA

[1] J. J. Khanam and S. Y. Foo, “A comparison of machine learning algorithms for diabetes prediction,” *ICT Express*, vol. 7, no. 4, pp. 432–439, Dec. 2021, doi: 10.1016/j.icte.2021.02.004.

[2] X. Lu *et al.*, “Type 2 diabetes mellitus in adults: pathogenesis, prevention and therapy,” Dec. 01, 2024. doi: 10.1038/s41392-024-01951-9.

[3] M. Melladia, D. E. Putra, and L. Muhelni, “Penerapan Data Mining Pemasaran Produk Menggunakan Metode Clustering,” *Jurnal Teknik Informasi dan Komputer (Tekinkom)*, vol. 5, no. 1, p. 160, Jun. 2022, doi: 10.37600/tekinkom.v5i1.458.

[4] N. Sepriyanti, R. Sani Nahampun, M. H. Zikri, I. Ambarani, and A. Rahmadeyan, “Penerapan K-Means Clustering Untuk Mengelompokkan Tingkat Kemiskinan di Provinsi Riau,” Aug. 2022. [Online]. Available: https://journal.irpi.or.id/index.php/sentimas

[5] P. Chandrasekaran and R. Weiskirchen, "The Role of Obesity in Type 2 Diabetes Mellitus—An Overview," Feb. 01, 2024, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/ijms25031882.

[6] J. Pebralia, "Analisis Curah Hujan Menggunakan Machine Learning Metode Regresi Linier Berganda Berbasis Python dan Jupyter Notebook," *JIFP (Jurnal Ilmu Fisika dan Pembelajarannya)*, vol. 6, no. 2, pp. 23–30, 2022, [Online]. Available: http://jurnal.radenfatah.ac.id/index.php/jifp/

[7] T. Santoso, A. Darmawan, N. Sari, M. A. F. Syadza, E. C. B. Himawan, and W. A. Rahman, "Clusterization of Agroforestry Farmers using K-Means Cluster Algorithm and Elbow Method," *Jurnal Sylva Lestari*, vol. 11, no. 1, pp. 107–122, Jan. 2023, doi: 10.23960/jsl.v11i1.646.

[8] R. R. Asyrofi and R. Asyrofi, "Implementasi Aplikasi Jupyter Notebook Sebagai Analisis Kreteria Plagiasi Dengan Teknik Simantik," *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, vol. 8, no. 2, pp. 627–637, May 2023, doi: 10.29100/jipi.v8i2.3699.

[9] A. Praja, C. Lubis, and D. E. Herdiwindiati, "Deteksi Penyakit Diabetes Dengan Metode Fuzzy C-Means Clustering Dan K-Means Clustering," 2017.

[10] R. Fitriyanto and Mohamad Ardi, "Feature Selection Comparative Performance For Unsupervised Learning On Categorical Dataset," *Jurnal Techno Nusa Mandiri*, vol. 22, no. 1, pp. 61–69, Mar. 2025, doi: 10.33480/techno.v22i1.6512.

[11] N. Septiani and S. Wahyuni, "Implementasi Data Mining Dalam Mengelompokkan Tingkat Kepuasan Pemakaian Jasa Cleaning Service Dengan Menggunakan Algoritma K-Means Clustering," vol. 5, no. 4, pp. 340–354, 2024, doi: 10.47065/bit.v5i2.1729.

[12] S. A. Hendrawan, Afdhal Chatra, Nurul Iman, Soemarno Hidayatullah, and Degdo Suprayitno, "Digital Transformation in MSMEs: Challenges and Opportunities in Technology Management," *Jurnal Informasi dan Teknologi*, pp. 141–149, Jun. 2024, doi: 10.60083/jidt.v6i2.551.

[13] V. A. Permadi, S. P. Tahalea, and R. P. Agusdin, "K-Means And Elbow Method For Cluster Analysis Of Elementary School Data," *PROGRES PENDIDIKAN*, vol. 4, no. 1, pp. 50–57, Jan. 2023, doi: 10.29303/prospek.v4i1.328.

[14] T. Tendean and W. Purba, "Analisis Cluster Provinsi Indonesia Berdasarkan Produksi Bahan Pangan Menggunakan Algoritma K-Means," *Jurnal Sains dan Teknologi*, vol. 1, no. 2, pp. 5–11, Mar. 2020.

[15] C. Debora Mait, J. Armando Watuseke, P. David Gibrael Saerang, S. Reynaldo Joshua, and U. Sam Ratulangi, "Sistem Pendukung Keputusan Menggunakan Fuzzy Logic Tahani Untuk Penentuan Golongan Obat Sesuai Dengan," *Jurnal Media Infotama*, vol. 18, no. 2, p. 344, 2022.