JUMANJI
ISSN (p): 2598-8050
ISSN (e): 2598-8069
| Vol. 8 | No. 1 | Halaman 16 - 35
April 2024

PENGELOMPOKAN TINGKAT RISIKO PENYAKIT DIABETES MELITUS MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING

Rio Gestavito[1], Asep Id Hadiana[2], Fajri Rakhmat Umbara[3]

[1,2,3] Jurusan Teknik Informatika, Universitas Jenderal Achmad Yani, Indonesia
Universitas Jenderal Achmad Yani
Jl. Terusan Jenderal Sudirman, Cimahi
rio.gestavito@unjani.ac.id

ABSTRAK

Penelitian ini fokus pada diabetes melitus (DM), kondisi metabolik kronis dengan tingkat gula darah tinggi karena kurangnya insulin. Faktor penyebab DM bervariasi, termasuk kurangnya produksi insulin oleh sel beta Langerhans di pankreas dan ketidakresponsifan tubuh terhadap insulin. Penyakit ini prevalen di negara berkembang dan diperkirakan terus meningkat. Studi ini menggunakan algoritma K-Means Clustering untuk mengelompokkan risiko DM. Evaluasi pada k = 2 menunjukkan data dalam klaster cenderung bercampur, dengan nilai Silhouette Coefficient 0.5716 dan Davies Bouldin Index 0.672. Visualisasi scatter menunjukkan penyebaran data yang seragam dalam klaster, memberikan pemahaman mendalam tentang pola data. Hasilnya dapat mendukung pemahaman dan penanganan lebih lanjut terhadap DM.

***Kata kunci**: Diabetes Melitus, K-Means, Clustering.*

ABSTRACT

This study focuses on diabetes mellitus (DM), a chronic metabolic condition characterized by high blood sugar levels due to insufficient insulin. The causes of DM vary, including a lack of insulin production by beta cells in the pancreas and the body's resistance to insulin. The disease is prevalent in developing countries and is expected to continue increasing. The study utilizes the K-Means Clustering algorithm to categorize DM risk levels. Evaluation at k = 2 indicates that data in the clusters tend to mix, with a Silhouette Coefficient value of 0.5716 and Davies Bouldin Index of 0.672. Scatter plot visualization shows uniform data distribution within clusters, providing a deeper understanding of data patterns. The results can support further understanding and management of DM.

***Keywords**: Diabetes Melitus, K-Means, Clustering.*

1. PENDAHULUAN

Pengklasteran, atau klasterisasi, merupakan metode dalam data mining untuk mengelompokkan objek berdasarkan variasi dan kesamaan karakteristik data (Syakur et al., 2018). Algoritma klasterisasi banyak digunakan dalam pemrosesan gambar, eksplorasi data, dan bioinformatika (Debatty et al., 2014). Klasterisasi membagi objek data menjadi klaster berdasarkan kesamaan, di mana klaster adalah kumpulan data yang berbeda atau mirip (Larose and Larose, 2014),(Chen et al., 2020).

Algoritma-algoritma klasterisasi memanfaatkan struktur data dan menetapkan aturan untuk mengelompokkan data serupa (Jain et al., 1999). Proses klasterisasi menghasilkan partisi dari kumpulan data tanpa pengetahuan awal tentang komposisinya. Idealnya, setiap klaster terdiri dari data serupa yang berbeda secara signifikan dari klaster lainnya. Pengukuran ketidakserupaan bergantung pada data asal dan tujuan algoritma. Klasterisasi merupakan inti dari aplikasi berbasis data dan dianggap tugas penting dalam pembelajaran mesin (Ahmed et al., 2020).

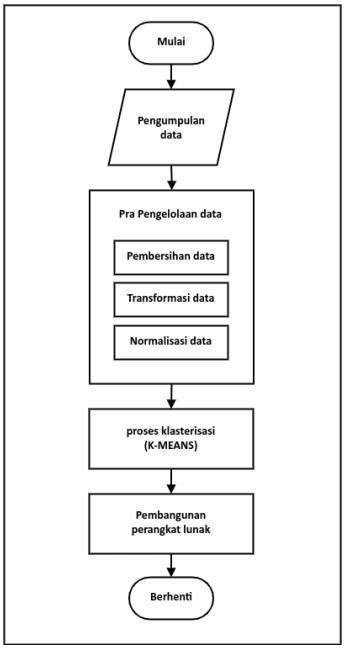
Meskipun ada berbagai metode klasterisasi yang tersedia, algoritma K-means tetap menjadi salah satu yang paling populer (Berkhin, 2006),(Jain, 2010). Bahkan, algoritma ini diidentifikasi sebagai salah satu dari 10 algoritma teratas dalam penambangan data (Wu et al., 2008, p. 10). Penelitian ini akan menggunakan K-Means untuk menentukan jumlah cluster risiko penyakit. K-Means dipilih karena efektivitas, efisiensi, dan kesederhanaannya, menjadi dasar pemahaman konsep dasar dalam pengelompokan data (Na et al., 2010).

Penilaian hasil klaster melibatkan pertimbangan jarak antar objek data dan menentukan jumlah pusat klaster (centroid) (Vora, 2013). K-Means menunjukkan akurasi signifikan terhadap variasi ukuran objek, efisien dalam memproses kumpulan data besar (Ediyanto, 2013). Studi perbandingan algoritma untuk siswa Tunagrahita menunjukkan keefektifan K-Means pada dataset kecil (Harahap, 2021). Namun, algoritma ini memiliki keterbatasan dalam menentukan jumlah klaster optimal (Bosnic et al., 2021).

Metode elbow digunakan untuk mencari nilai optimal k, dengan pendekatan sederhana dan evaluasi WCSS pada setiap nilai klaster (Merliana et al., 2015). Hasil analisis ini menjadi dasar untuk klasterisasi dengan K-Means pada dataset penyakit diabetes melitus, diharapkan menghasilkan kelompok optimal dan informasi berharga (Bholowalia, n.d.).

Hasil dari analisis ini akan menentukan jumlah klaster k yang paling optimal, yang selanjutnya akan menjadi landasan untuk menjalankan proses klasterisasi dengan algoritma K-Means. Proses ini diimplementasikan dalam konteks studi kasus Dataset Penyakit Diabetes Melitus. Harapannya, hasil klasterisasi menggunakan metode K-Means akan memberikan kelompok yang optimal dan mampu menghasilkan informasi berharga. Ini membantu mengungkap pola-pola yang berbeda di antara kelompok klaster yang berbeda.

2. METODE



Gambar 1. Metode penelitian

2.1. Pengumpulan Data

Tahap ini dimaksudkan untuk menentukan data yang akan diproses. Dalam langkah ini, semua informasi yang dibutuhkan akan dikumpulkan menjadi satu set data, termasuk atribut yang diperlukan untuk proses tersebut. Caranya adalah dengan mencari informasi yang sudah tersedia dan juga mencari informasi tambahan yang dibutuhkan.

2.2. Prapengelolaan Data

Pada tahap ini dilakukan pengolahan data agar menjadi data yang siap untuk digunakan. Pemrosesan data yang dilakukan meliputi pembersihan data, transformasi data, dan normalisasi data.

2.2.1. Pembersihan Data

Dalam tahap ini dilakukan pembersihan data. Dengan menghilangkan nilai 0 dan null pada tiap atribut ID, No_Pation, Gender, AGE, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, BMI, CLASS.

2.2.2. Transformasi Data

Dalam tahap ini melakukan transformasi nilai pada atribut Gender yaitu dengan mengubah variabel data menjadi bentuk data seperti numerik dan pengurangan. Kedua langkah tersebut dilakukan untuk mendapatkan data yang representatif serta untuk mengekstrak informasi berharga dari data yang diperoleh.

2.2.3. Normalisasi Data

Normalisasi adalah proses untuk mengubah data ke dalam skala yang sama atau rentang yang seragam. Tujuan dari normalisasi data adalah untuk menghilangkan perbedaan skala antara fitur-fitur atau variabel dalam dataset.

2.3. Proses Klasterisasi

Pada tahap ini data yang telah melewati tahap pre-processing akan masuk pada Metode Elbow untuk mencari nilai K optimal yang akan digunakan pada nilai K dalam K-Means Clustering. Implementasi klasterisasi dengan K-Means ini menggunakan library Scikitlearn, Jupiter dan bahasa pemrograman Python.

2.4. Pembangunan Perangkat Lunak

Pada tahap ini, dibuat sebuah program komputer untuk mempermudah prediksi dan visualisasi data yang digunakan dengan melakukan perancangan dan pembangunan. Program tersebut menggunakan bahasa pemrograman Python dan *Framework* Flask.

2.5. Pengujian dan Evaluasi

Pada pengujian eksperimen, dilakukan analisis klasterisasi menggunakan algoritma K-Means dan pengujian akurasi dengan menggunakan Silhouette Scores pada setiap nilai K yang digunakan untuk optimalisasi jumlah klaster. Serta Davies Bouldin Index (DBI) suatu ukuran untuk mengevaluasi kinerja pengelompokan.

3. HASIL DAN PEMBAHASAN

3.1. Deskripsi Atribut Data

Sebelum memasuki tahap proses data mining, dataset ini perlu dijelaskan tiap atribut yang ada pada data tersebut, agar dapat menentukan atribut yang digunakan dari data pasien ini. Deskripsi atribut dapat dilihat pada Tabel 1.

Tabel 1. Deskripsi atribut data

No.	Atribut Data	Deskripsi
1	ID.	ID unik dari dataset
2	No_Pation	Nomer pasien
3	Gender	Jenis kelamin dari pasien M = Male, F = Female
4	Age	Umur dari pasien
5	Urea	Kadar ureum atau blood urea nitrogen (BUN). zat sisa dari pemecahan protein dan asam amino di dalam hati
6	Cr	albumin to creatinine ratio atau rasio albumin-kreatinin. Rasio albumin-kreatinin urine (ACR) atau ACR sewaktu digunakan untuk mengidentifikasi penyakit ginjal yang dapat terjadi akibat komplikasi seperti diabetes.
7	HbA1c	Pemeriksaan dengan mengkur kada atau prosentase glukosa terkait dengan hemogoblin
8	Chol	Cholesterol atau kolesterol, adalah lemak yang diproduksi oleh tubuh, dan juga berasal dari makanan hewani.
9	TG	Triglycerides atau Trigliserida adalah salah satu jenis lemak yang mengalir di dalam darah

No.	Atribut Data	Deskripsi
10	HDL	HDL (high-density lipoprotein) atau kolesterol baik. HDL adalah kolesterol yang berfungsi untuk membersihkan kelebihan kolesterol yang berbahaya di dalam darah dan membawanya kembali ke hati untuk dikeluarkan dari tubuh
11	LDL	Low-density lipoproteins termasuk golongan kolesterol jahat. Mengandung lebih banyak kolesterol
12	VLDL	very low-density lipoproteins termasuk golongan kolesterol jahat. Mengandung lebih banyak trigliserida
13	BMI	Masa indeks tubuh, pengukuran yang digunakan untuk menentukan golongan berat badan sehat dan tidak sehat
14	CLASS	N = tidak diabetes Y = diabetes P = mungkin diabetes

3.2. Dataset Diabetes

Berikut adalah sampel dari data diabates melitus dengan jumlah atribut 14 yang dapat dilihat pada Tabel 2.

Tabel 2. Sampel Dataset Diabetes Melitus

ID.	No_Pation	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	CLASS
502	17975	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24	N
735	34221	M	26	4.5	62	4.9	3.7	1.4	1.1	2.1	0.6	23	N
420	47975	F	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24	N
...	...	...	...	...	...	...	...	...	...	...	...	...	...
99	24004	M	38	5.8	59	6.7	5.3	2	1.6	2.9	14	40.5	Y
248	24054	M	54	5	67	6.9	3.8	1.7	1.1	3	0.7	33	Y

3.3. Prapengelolaan Data

Tahap prapengolaan data adalah tahapan untuk mengubah data menjadi data yang siap diuji. Pada tahap prapengolaan data, data diproses melalui tahap pembersihan data, dan transformasi data.

3.4. Pembersihan Data

Pada tahap pembersihan data, identifikasi merupakan langkah yang penting untuk memastikan kualitas data yang digunakan dalam analisis. Oleh karena itu, dalam penelitian ini, proses pembersihan data dilakukan dengan mengidentifikasi nilai kosong pada atribut ID, No_Pation, Gender, AGE, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, BMI. CLASS. Untuk mengatasi nilai kosong tersebut, dilakukan penghapusan baris data untuk yang mengandung nilai 0 atau missing value. Melakukan pemilihan atribut yang relevan yakni menghilangkan atribut ID, dan No_Pation.

3.5. Transformasi data

Tahap ini melibatkan transformasi data setelah proses pembersihan data selesai. Dalam tahapan ini, variabel data diubah menjadi bentuk numerik atau format data yang sesuai. Dataset yang digunakan dalam penelitian ini terdiri dari 11 atribut yang mencakup gender, age, urea, cr, hba1c, chol, TG, HDL, LDL, VLDL, dan BMI. Dari dataset tersebut, 10 atribut memiliki jenis data numerik sedangkan 1 atribut memiliki jenis data nominal. Untuk mengatasi hal ini dilakukan transformasi data, transformasi ini dilakukan untuk atribut gender, yang mana gender merupakan bentuk data nominal, menjadi data numerik dengan menerapkan teknik one-hot encode. Teknik ini mewakili bentuk data kategori menjadi numerik. Pada penelitian ini, untuk atribut Gender dengan nilai M menjadi 2, dan F menjadi 1. hasil transformasi data dapat dilihat seperti pada Tabel 3.

Tabel 3. Setelah tahapan transformasi data

Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
1	50	4.7	46	4.9	4.2	0.9	2.4	1.4	0.5	24
2	26	4.5	62	4.9	3.7	1.4	1.1	2.1	0.6	23
…	…	…	…	…	…	…	…	…	…	…
2	38	5.8	59	6.7	5.3	2	1.6	2.9	14	40.5
2	54	5	67	6.9	3.8	1.7	1.1	3	0.7	33

3.6. Normalisasi Data

$$X_{new} = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

Untuk metode normalisasi yang digunakan Min-Max Scaling. Cara kerja dari metode ini setiap nilai pada sebuah fitur dikurangi dengan nilai minimum fitur tersebut, kemudian dibagi dengan rentang nilai atau nilai maksimum dikurangi nilai minimum dari fitur, seperti pada persamaan (1).

Atribut Gender :

0 = (1-1)/(2-1) Data Ke-1

1 = (2 – 1) / (2 - 1) Data Ke-2

Dan seterusnya sampai dengan baris terakhir atribut, begitu juga pada atribut lainnya sehingga di dapat hasil pada Tabel 4.

Tabel 4. Setelah tahapan normalisasi

Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
0.00	0.51	0.11	0.05	0.26	0.41	0.04	0.23	0.11	0.01	0.17
1.00	0.10	0.10	0.07	0.26	0.36	0.08	0.09	0.19	0.01	0.14
…	…	…	…	…	…	…	…	…	…	…
1.00	0.31	0.14	0.07	0.38	0.51	0.13	0.14	0.27	0.40	0.75

Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
1.00	0.58	0.12	0.08	0.40	0.37	0.10	0.09	0.28	0.02	0.49

3.7. Algoritma Elbow

$$WCSS(k) = \sum d\,(P_i C_i) \tag{2}$$

Metode Elbow digunakan untuk menentukan jumlah klaster yang optimal melalui analisis perbandingan antara jumlah klaster yang dibentuk dan persentase penurunan nilai yang signifikan pada suatu titik tertentu. Seperti persamaan (2) berikut adalah perhitungan mencari nilai WCSS pada elbow menentukan titik siku dalam grafik.

WCSS = (0.302 + 0.142 + 142 + 0.162 + 0.162) + (0.152 + 0.302 + 0.262 + 0.262 + 0.162 + 0.162 + 0.222 + 0.222 + 0.442)

WCSS = 196.75

Pada penelitian ini didapatkan nilai k terbaik adalah 2. Untuk mempercepat waktu disini peneliti menampilkan satu kali perhitungan untuk k 2, juga dihitung pada data sampel sebanyak 14 data. Sebaiknya dilakukan juga pada nilai k lainnya untuk mendapatkan nilai berbandingan.

3.8. Algoritma K-Means Clustering

Menentukan nilai K sebagai klaster yang ingin dibentuk, dalam penelitian ini dibentuk kelompok atau klaster sebanyak 2 klaster. Sampel data yang digunakan dapat dilihat pada Tabel 5.

Tabel 5. Sampel data

Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
0.00	0.51	0.11	0.05	0.26	0.41	0.04	0.23	0.11	0.01	0.17
1.00	0.10	0.10	0.07	0.26	0.36	0.08	0.09	0.19	0.01	0.14
0.00	0.51	0.11	0.05	0.26	0.41	0.04	0.23	0.11	0.01	0.17
0.00	0.51	0.11	0.05	0.26	0.41	0.04	0.23	0.11	0.01	0.17
1.00	0.22	0.17	0.05	0.26	0.48	0.05	0.06	0.18	0.01	0.07
0.00	0.42	0.05	0.02	0.21	0.28	0.05	0.08	0.13	0.01	0.07
0.00	0.51	0.04	0.06	0.21	0.35	0.07	0.07	0.19	0.01	0.17
1.00	0.47	0.11	0.05	0.21	0.28	0.04	0.07	0.14	0.01	0.17
1.00	0.39	0.05	0.08	0.21	0.37	0.04	0.23	0.35	0.03	0.07
0.00	0.20	0.08	0.03	0.21	0.37	0.13	0.23	0.36	0.03	0.17
0.00	0.19	0.10	0.06	0.22	0.35	0.03	0.15	0.14	0.01	0.14
0.00	0.22	0.07	0.06	0.21	0.39	0.06	0.07	0.25	0.03	0.07
0.00	0.17	0.07	0.05	0.21	0.48	0.07	0.10	0.30	0.01	0.10
0.00	0.42	0.11	0.06	0.28	0.41	0.10	0.10	0.20	0.02	0.14

Inisialisasi K pusat klaster (centroid) awal dengan cara mengambil data dari sumber secara random acak. Adapun centroid awal sebagai berikut, dapat dilihat pada Tabel 6.

Tabel 6. Centroid awal

Centroid awal / iterasi 1	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
C1	1.00	0.19	0.07	0.05	0.74	0.36	0.14	0.05	0.20	0.44	0.63
C2	1.00	0.17	0.12	0.08	0.35	0.40	0.11	0.10	0.22	0.36	0.54

Hitung jarak setiap data ke pusat klaster (centroid) menggunakan rumus perhitungan jarak Euclidean Distance pada persamaan (3) sebagai berikut:

$$d(x,y) = \sum (x_i - y_i)^2 + (x_i - y_i)^2 + \ldots + (x_n - y_n)^2 \quad (3)$$

Lakukan perhitungan jarak sampai Data ke-N terhadap awal klaster, sehingga hasil didapatkan seperti Tabel 7 berikut:

Tabel 7. Hasil perhitungan tiap data iterasi ke - 1

Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI	Jarak C1	Jarak C2
1.00	0.19	0.07	0.05	0.74	0.36	0.14	0.05	0.20	0.44	0.63	0.00	0.42
1.00	0.17	0.17	0.08	0.34	0.36	0.06	0.10	0.20	0.23	0.29	0.58	0.30
1.00	0.42	0.09	0.06	0.59	0.43	0.07	0.07	0.30	0.27	0.53	0.37	0.37
1.00	0.42	0.09	0.06	0.59	0.43	0.07	0.07	0.30	0.27	0.53	0.37	0.37
1.00	0.42	0.13	0.07	0.66	0.34	0.09	0.11	0.16	0.30	0.37	0.41	0.45
1.00	0.19	0.08	0.04	0.32	0.44	0.10	0.08	0.29	0.02	0.17	0.76	0.51
1.00	0.42	0.13	0.07	0.66	0.34	0.09	0.11	0.16	0.30	0.37	0.41	0.45
1.00	0.17	0.12	0.08	0.35	0.40	0.11	0.10	0.22	0.36	0.54	0.42	0.00
1.00	0.17	0.12	0.08	0.35	0.40	0.11	0.10	0.22	0.36	0.54	0.42	0.00
1.00	0.25	0.11	0.06	0.41	0.32	0.05	0.10	0.16	0.20	0.29	0.55	0.34

Lakukan pengelompokan data klaster dengan jarak yang terpendek. Tetapkan data pada masing-masing klaster dengan jarak terpendek. Adapun hasil penempatan klaster setiap data pada iterasi ke-1 dapat dilihat pada Tabel 8 berikut.

Tabel 8. Hasil pengelompokkan iterasi ke-1

No.	Gender	AGE	...	VLDL	BMI	Jarak C1	Jarak C2	Hasil i - 1
1	1.00	0.19	...	0.44	0.63	0.00	0.42	C1
2	1.00	0.17	...	0.23	0.29	0.58	0.30	C2
3	1.00	0.42	...	0.27	0.53	0.37	0.37	C1
4	1.00	0.42	...	0.27	0.53	0.37	0.37	C1
5	1.00	0.42	...	0.30	0.37	0.41	0.45	C1
6	1.00	0.19	...	0.02	0.17	0.76	0.51	C2

No.	Gender	AGE	...	VLDL	BMI	Jarak C1	Jarak C2	Hasil i - 1
7	1.00	0.42	...	0.30	0.37	0.41	0.45	C1
8	1.00	0.17	...	0.36	0.54	0.42	0.00	C2
9	1.00	0.17	...	0.36	0.54	0.42	0.00	C2
10	1.00	0.25	...	0.20	0.29	0.55	0.34	C2

Setelah didapatkan anggota dari setiap klaster kemudian hitung pusat klaster atau centroid yang baru.

$$ClusterCenter = \frac{(x_1 + x_2 + \ldots + x_n)}{n} \quad (4)$$

Pada centroid C1 dan C2 hitung tiap atribut Gender sampai dengan BMI dengan menggunakan rumus persamaan (4).

Sehingga nilai centroid baru C1 dan C2 yang didapat seperti pada Tabel 9.

Tabel 9. Centroid baru C1 & C2

Centroid baru / iterasi 2	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
C1	1.00	0.38	0.10	0.06	0.64	0.38	0.09	0.08	0.22	0.31	0.49
C2	1.00	0.27	0.11	0.07	0.35	0.41	0.11	0.09	0.24	0.24	0.34

Lakukan persamaan (1) pada iterasi ke-2 menggunakan nilai centroid baru sehingga didapatkan hasil seperti pada Tabel 10 berikut.

Tabel 10. Hasil jarak dari centroid baru pada iterasi ke 2

No.	Gender	AGE	...	VLDL	BMI	Jarak C1	Jarak C2	Hasil i - 2
1	1.00	0.19	...	0.44	0.63	0.30	0.54	C1
2	1.00	0.17	...	0.23	0.29	0.44	0.15	C2
3	1.00	0.42	...	0.27	0.53	0.14	0.35	C1
4	1.00	0.42	...	0.27	0.53	0.14	0.35	C1
5	1.00	0.42	...	0.3	0.37	0.16	0.37	C1
6	1.00	0.19	...	0.02	0.17	0.58	0.30	C2
7	1.00	0.42	...	0.3	0.37	0.16	0.37	C1
8	1.00	0.17	...	0.36	0.54	0.37	0.26	C2
9	1.00	0.17	...	0.36	0.54	0.37	0.26	C2
10	1.00	0.25	...	0.2	0.29	0.36	0.16	C2

Pada perhitungan iterasi ke-2 tidak ada data yang berpindah dari satu cluster ke cluster yang lain atau kovergen, maka proses iterasi pun berhenti.

3.9. Algoritma Silhouette Coefficient

$$ClusterCenter = \frac{(x_1 + x_2 + \ldots + x_n)}{n} \quad (5)$$

Metode silhouette salah satu metode yang umum digunakan untuk mengukur seberapa baik sebuah objek berada dalam klaster telah dibentuk. Metode ini memberikan skor untuk setiap klasternya. Tafsiran hasil jika mendekati 1 maka pengelompokan menunjukkan baik.

Untuk mencari nilai Silhouette Coefficient terdapat beberapa tahapan. Tahapan pertama mencari nilai *a(i)*. *a(i)* adalah rata-rata jarak titik data *i* dengan semua titik data dalam cluster yang sama dengan *i* (intra-cluster distance). Persamaan (5)

a(i) = $\sqrt{(1\text{-}1)^2 + (0.42\text{-}0.19)^2 + (0.09\text{-}0.07)^2 + (0.06\text{+}0.05)^2 + (0.59\text{+}0.74)^2 + (0.43\text{-}0.36)^2 + (0.07\text{-}0.14)^2 + (0.07\text{-}0.05)^2 + (0.30\text{-}0.20)^2 + (0.27\text{-}0.44)^2 + (0.53\text{-}0.63)^2 + \ldots + \sqrt{(i - j)^2}}$

a(i) = 0.3902

Tahap kedua mencari nilai b(i), b(i) adalah rata-rata jarak antara titik data i dengan semua titik data dalam cluster terdekat yang berbeda dengan i (inter-cluster distance).

b(i) = $\sqrt{(1\text{-}1)^2 + (0.42\text{-}0.19)^2 + (0.09\text{-}0.07)^2 + (0.06\text{+}0.05)^2 + (0.59\text{+}0.74)^2 + (0.43\text{-}0.36)^2 + (0.07\text{-}0.14)^2 + (0.07\text{-}0.05)^2 + (0.30\text{-}0.20)^2 + (0.27\text{-}0.44)^2 + (0.53\text{-}0.63)^2 + \ldots + \sqrt{(i - j)^2}}$

b(i) = 0.5774

Langkah yang sama dilakukan pada data lainnya. Sehingga setelah semua hasil *a(i)* dan *b(i)* didapatkan.

Tahap ketiga menghitung nilai *s(i)* atau *silhouette*.

s(i) $\frac{0.3902 - 0.5774}{max(0.3902; 0.5774)}$

Hitungan berlanjut sampai kepada sampel data terakhir yaitu *s* (14), dan hasil keseluruhan akan tampak seperti pada Tabel 11 berikut ini.

Tabel 11. Hasil perhitungan nilai silhouette coefficient

No.	Gender	AGE	...	VLDL	BMI	a(i)	b(i)	s(i)
1	1	0.19	...	0.44	0.63	0.3900	0.5774	0.3243
2	1	0.17	...	0.23	0.29	0.2969	0.4707	0.3693
3	1	0.42	...	0.27	0.53	0.2198	0.4154	0.4708
4	1	0.42	...	0.27	0.53	0.2198	0.4154	0.4708
5	1	0.42	...	0.30	0.37	0.2294	0.4324	0.4695
6	1	0.19	...	0.02	0.17	0.4091	0.6021	0.3206
7	1	0.42	...	0.30	0.37	0.2294	0.4324	0.4695
8	1	0.17	...	0.36	0.54	0.3419	0.4123	0.1709
9	1	0.17	...	0.36	0.54	0.3419	0.4123	0.1709

No.	Gender	AGE	...	VLDL	BMI	a(i)	b(i)	s(i)
10	1	0.25	...	0.20	0.29	0.2866	0.4004	0.2843
11	1	0.25	...	0.20	0.29	0.2866	0.4004	0.2843
12	1	0.42	...	0.36	0.42	0.3231	0.3400	0.0499
13	1	0.42	...	0.36	0.42	0.3231	0.3400	0.0499
14	1	0.42	...	0.05	0.10	0.5357	0.7132	0.2489

Setelah semua nilai *s(i)* diketahui hitung rata rata nilai seluruh *s(i)*, sehingga didapatkan nilai *silhouette coefficient* yaitu 0.2967.

3.10. Algoritma Davies Bouldin Index

$$SSW_i = 1\frac{1}{m_i}\sum_{j=i}^{m_i} d(x_j, c_i) \quad (6)$$

$$SSB_{i,j} = d(c_j, c_i) \quad (7)$$

$$R_{i,j} = \frac{SSW_i + SSW_j}{SSB_{i,j}} \quad (8)$$

$$DBI = \frac{1}{k}\sum_{j=1}^{k} max(R_{i,j}) \quad (9)$$

Diketahui sebuah *centroid akhir* yang didapatkan pada Tabel 12.

Tabel 12. Centroid akhir

centroid	Gender	AGE	Urea	Cr	HbA1c	Chol	TG	HDL	LDL	VLDL	BMI
1	1.00	0.38	0.10	0.06	0.64	0.38	0.09	0.08	0.22	0.31	0.49
2	1.00	0.27	0.11	0.07	0.35	0.41	0.11	0.09	0.24	0.24	0.34

Pertama, mencari nilai SSW berdasarkan persamaan (6)

Jarak data – klaster

Data ke-1 = $\sqrt{(((1.0\text{-}1.0)^2) + ((0.19\text{-}0.38)^2)+((0.07\text{-}0.10)^2)+((0.05\text{-}0.06)^2)+((0.74\text{-}0.64)^2)+((0.36\text{-}0.38)^2)+((0.14\text{-}0.09)^2)+((0.05\text{-}0.08)^2)+((0.20\text{-}0.22)^2)+((0.44\text{-}0.31)^2)+((0.63\text{-}0.49)^2)))}$

= 0.0878

Data ke-2 = $\sqrt{(((1.0\text{-}1.0)^2) + ((0.17\text{-}0.27)^2)+((0.17\text{-}0.11)^2)+((0.08\text{-}0.07)^2)+((0.34\text{-}0.35)^2)+((0.36\text{-}0.41)^2)+((0.06\text{-}0.11)^2)+((0.10\text{-}0.09)^2)+((0.20\text{-}0.24)^2)+((0.23\text{-}0.24)^2)+((0.29\text{-}0.34)^2)))}$

= 0.0231

Lakukan sampai data ke-*n*. Lalu cari rata-rata nilai pada klaster 1 dan juga klaster 2. Sehingga didapatkan keseluruhan hasil, pada tiap klasternya.

Klaster 1	0.15
Klaster 2	0.24

Kemudian mencari nilai SSB berdasarkan persamaan (7).

$SSB_{1,1}$ = $\sqrt{(((1-1)^2)+((0.38-0.38)^2)+((0.10-0.10)^2)+((0.06-0.06)^2)+((0.64-0.64)^2)+((0.38-0.38)^2)+((0.09-0.09)^2)+((0.08-0.08)^2)+((0.22-0.22)^2)+((0.31-0.31)^2)+((0.49-0.49)^2))}$

$SSB_{1,1}$ = 0.00

$SSB_{1,2}$ = $\sqrt{(((1-1)^2)+((0.38-0.27)^2)+((0.10-0.11)^2)+((0.06-0.07)^2)+((0.64-0.35)^2)+((0.38-0.41)^2)+((0.09-0.11)^2)+((0.08-0.09)^2)+((0.22-0.24)^2)+((0.31-0.24)^2)+((0.49-0.34)^2))}$

$SSB_{1,2}$ = 0.35

Untuk $SSB_{2,1}$ mempunyai nilai sama dengan $SSB_{1,2}$ yaitu 0.35 begitu juga pada $SSB_{2,2}$ yaitu 0.00. Sehingga akan tampak seperti Tabel 13.

Tabel 13. Nilai SBB

	Centroid	
SBB	1	2
1	0.00	0.35
2	0.35	0.00

Kemudian menghitung rasio berdasarkan persamaan (8)

$R_{i,j}$ = 0.15/0.24 x 0.35

$R_{i,j}$ = 1.10

kemudian menghitung DBI berdasarkan persamaan (9)

DBI = 1/2 x 1.10

DBI = 0.55

3.11. Eksperimen dan Pengujian

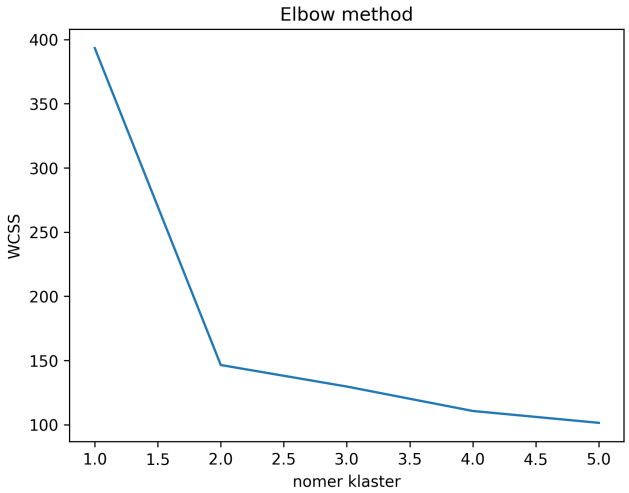
Eksperimen yang dilakukan pada penelitian ini adalah dengan melakukan pengujian nilai k sebanyak k = 1 sampai dengan k = 5, hasil yang dapatkan seperti pada Tabel 14.

Tabel 14. Pengujian dan nilai elbow

Nilai K	WCSS
1	393.27
2	146.48
3	129.7
4	110.68

Nilai K	WCSS
5	101.44

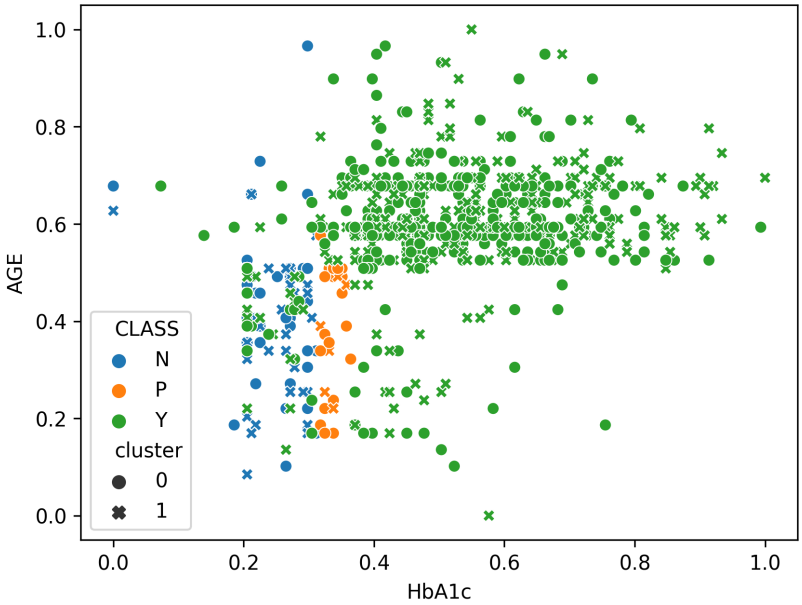
Sehingga di dapat grafik seperti pada Gambar 2.



Gambar 2. Titik elbow

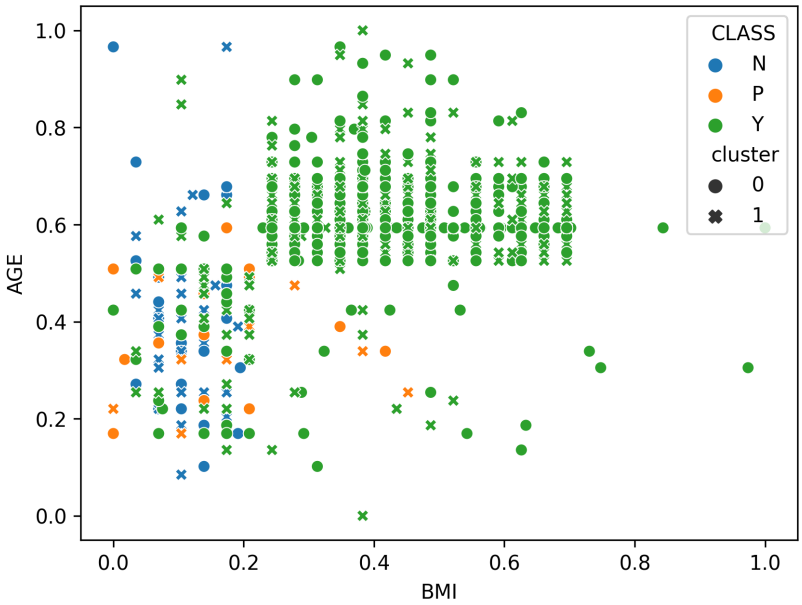
Gambar 2 menunjukkan bahwa nilai K yang baik berada pada nilai K = 2. Ini di gambarkan karena titik siku berada pada nilai 2 yaitu dengan WCSS 146.48 yang adalah nilai kekompakan atau kedekatan titik data dengan pusat cluster.

Eksperimen juga dilakukan untuk mengetahui visual beberapa tiap atribut pada Gambar 3. Hasil visualisasi menggunakan scatter pada atribut AGE dan HbA1c untuk Memeriksa apakah ada pola antara usia dan kontrol gula darah jangka panjang, di dapatkan seperti pada Gambar 3.



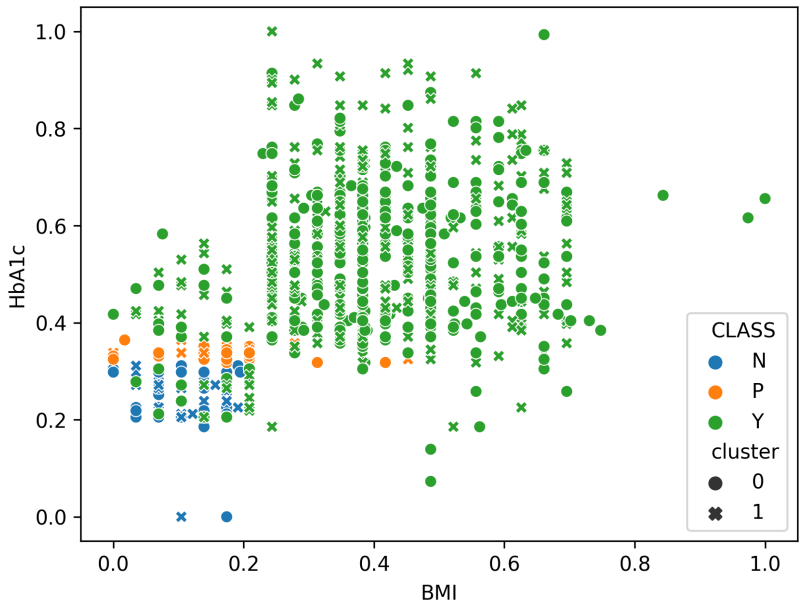
Gambar 3 Visualisasi titik atribut AGE dan HbA1c

Eksperimen juga dilakukan untuk mengetahui visual beberapa tiap atribut. Hasil visualisasi menggunakan scatter pada atribut AGE dan BMI untuk Melihat apakah ada hubungan antara usia dan indeks massa tubuh, didapatkan seperti pada Gambar 4.



Gambar 4 Visualisasi titik atribut AGE dan BMI

Eksperimen juga dilakukan untuk mengetahui visual beberapa tiap atribut. Hasil visualisasi menggunakan scatter pada atribut HbA1c dan BMI untuk Melihat apakah ada hubungan antara usia dan indeks massa tubuh, didapatkan seperti pada Gambar 5.



Gambar 5 Visualisasi titik atribut HbA1c dan BMI

pada hasil visualisasi Gambar 3, Gambar 4, Gambar 5 menunjukkan bahwa custer 0 dan 1 tumpang tindih antar klaster. Untuk cluster 0 dan 1 masing masing cluster tersebar dalam class pasien yang terjangkit diabetes, mungkin terjangkit diabetes, dan tidak terjangkit diabetes. Tumpang tindih antara klaster 0 dan 1 dalam kategori pasien yang terjangkit

diabetes, mungkin terjangkit diabetes, dan tidak terjangkit diabetes bisa terjadi karena kemiripan fitur yang digunakan dalam klastering, keterbatasan atribut yang tidak cukup membedakan risiko diabetes, variasi yang kompleks dalam data pasien diabetes, serta kemungkinan kesalahan dalam proses klastering atau pengumpulan data yang mengakibatkan pemisahan yang tidak optimal antara kelompok pasien.

Evaluasi dilakukan menggunakan 2 teknik yaitu Silhouette Coefficient (SC) dan Davies Bouldin Index (DBI). Hasil dari evaluasi pada tiap tiap SC dan DBI. Penggunaan kedua metrik ini secara bersama-sama memberikan gambaran yang lebih komprehensif tentang kualitas klaster yang dihasilkan. SC memberikan informasi tentang seberapa baik setiap titik data dikelompokkan dalam klaster, sementara DBI memberikan informasi tentang seberapa baik klaster itu sendiri dibentuk. Dengan demikian, kombinasi SC dan DBI dapat membantu dalam memilih jumlah klaster yang optimal dan mengevaluasi kekompakan serta pemisahan klaster yang dihasilkan oleh algoritma klaster yang digunakan.

Evaluasi yang dilakukan pada penelitian ini, menggunakan Silhouette Coefficient dan Davies Bouldin Index dengan nilai minimal k = 2 sampai k = 5, dan 5 kali percobaan pembelajaran. Ditunjukkan pada Tabel 15 berikut.

Tabel 15. Evaluasi silhouette coefficient

Nilai K	Kali Percobaan	Silhouette Coefficient	Davies Bouldin Index
2	1	0.5716	0.672
	2	0.5716	0.672
	3	0.5716	0.672
	4	0.5716	0.672
	5	0.5716	0.672
3	1	0.4243	1.2525
	2	0.319	1.5474
	3	0.4243	1.2525
	4	0.3201	0.7299
	5	0.3929	1.8344
4	1	0.1492	2.1856
	2	0.1475	2.2817
	3	0.3939	2.1996
	4	0.1653	1.757
	5	0.2507	2.1872
5	1	0.1549	2.0451
	2	0.1515	2.005
	3	0.1505	2.0019
	4	0.1468	2.0617
	5	0.1574	2.001

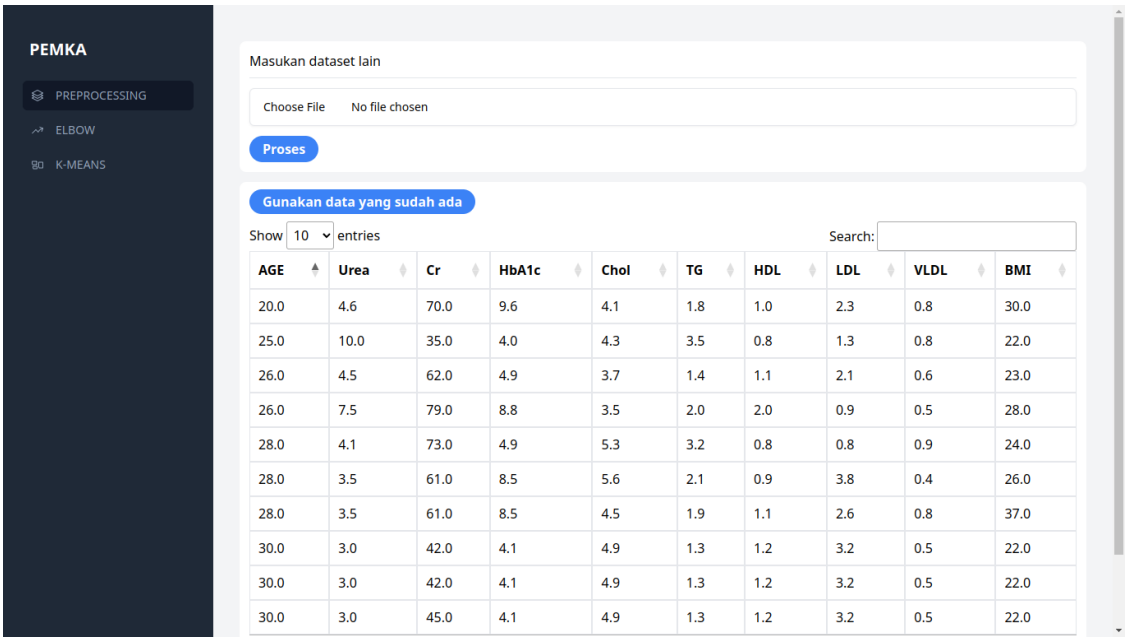
Pada hasil eksperimen dan pengujian diatas menunjukkan bahwa k = 2 menunjukkan hasil yang konsisten dengan nilai 0.5716 untuk Silhouette Score dan 0.672 untuk Davies Bouldin Index pada setiap kali percobaannya.

3.12. Implementasi Perangkat Lunak

Implementasi perangkat lunak memperlihatkan hasil dari perangkat lunak atau produk yang telat di buat.

1. Halaman Preprocessing

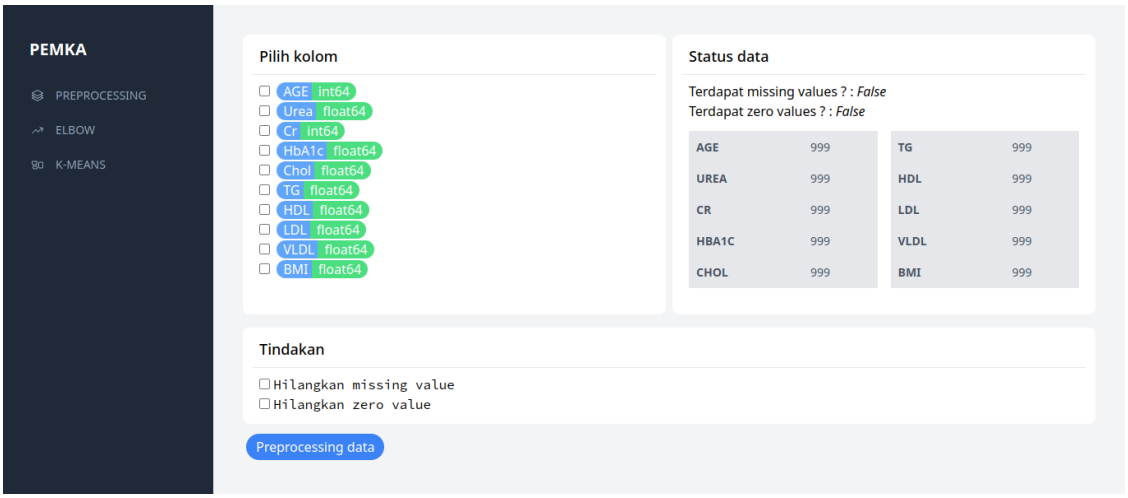
Halaman Antarmuka Preprocessing adalah halaman yang pertama kali dijalankan. Memungkinkan pengguna untuk memasukkan data baru atau menggunakan data yang telah ada untuk kemudian di lakukan preprocessing. Dapat dilihat pada Gambar 6.



Gambar 6 Halaman preprocessing

2. Halaman Preprocessing data

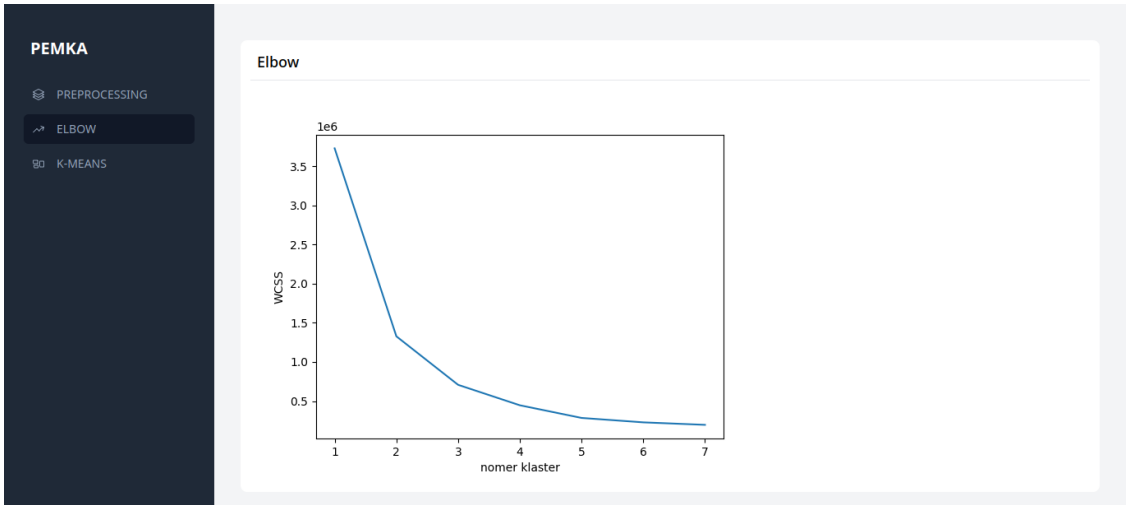
Pada halaman ini pengguna memungkinkan untuk memilih tindakan preprocessing yang akan di pilih. Pada halaman ini menampilkan kolom pada data, status data, dan tindakan preprocessing. Dapat dilihat pada Gambar 7.



Gambar 7 Halaman preprocessing data

3. Halaman Elbow

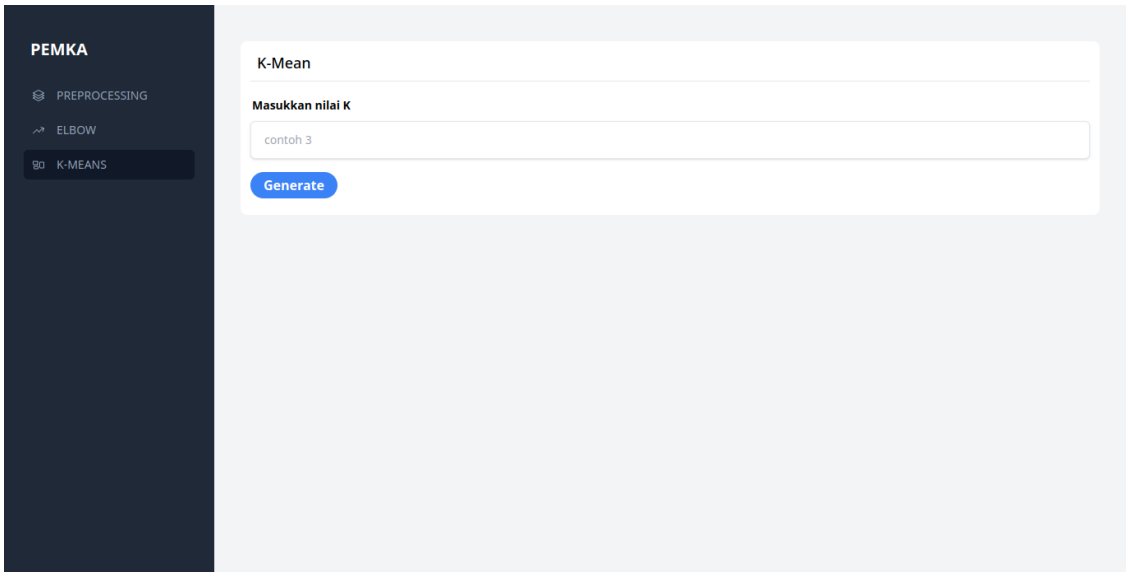
Pada saat halaman ini dibuka sistem akan memproses metode Elbow dan menampilkan visualisasinya. Dapat dilihat pada Gambar 8.



Gambar 8 Halaman elbow

4. Halaman K-Means

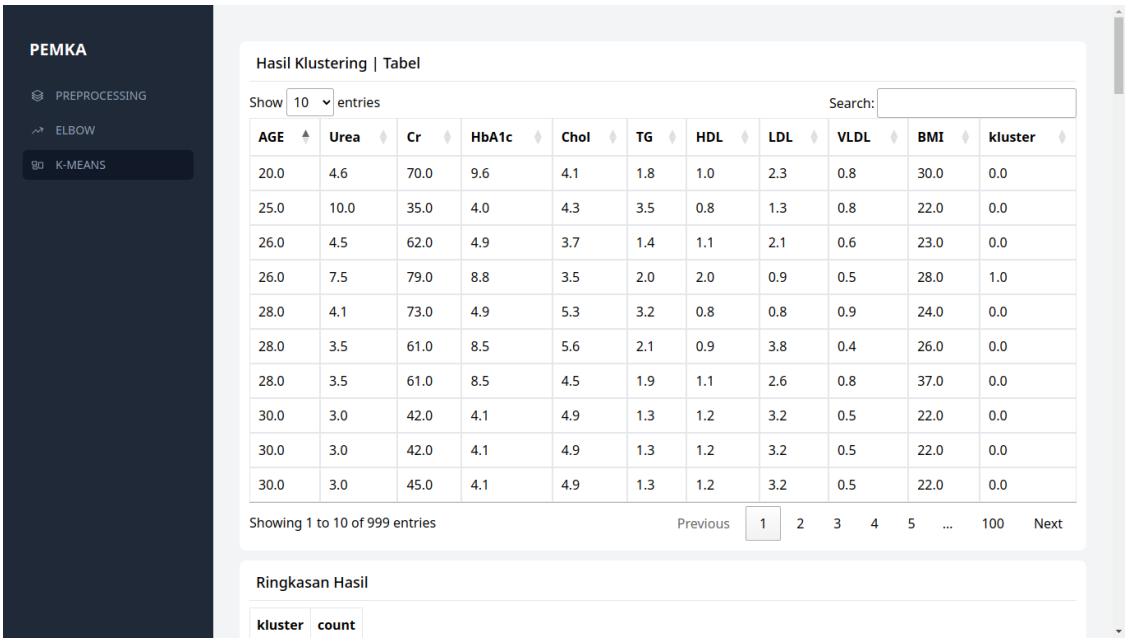
Pada halaman ini memungkinkan pengguna untuk memasukkan nilai K atau berapa banyak kelompok data yang akan di klasterisasi. Dapat dilihat pada Gambar 9.



Gambar 9 Halaman k-means

5. Halaman Hasil Klasterisasi

Pada halaman ini menampilkan hasil dari klasterisasi dengan nilai K yang dimasukkan sebelumnya. Pada halaman ini informasi yang diberikan adalah hasil klasterisasi, visualisasi diagram, dan evaluasi. Dapat dilihat pada Gambar 10.



Gambar 10 Halaman hasil klasterisasi

4. KESIMPULAN

Berdasarkan hasil eksperimen sudut siku terdapat pada nilai K = 2 sebagai nilai K optimal dengan nilai WCSS = 146.48. Pada implementasi di dapatkan 2 cluster dengan banyak cluster 0 = 490 data dan cluster 1 = 354 data, dan berikut akurasi atau evaluasi Silhouette Coefficient dan Davies Bouldin Index menunjukkan bahwa data tiap klaster cenderung becampur dengan tiap klaster. Dimana menggunakan Silhouette Coefficient menghasilkan nilai 0.5716, Nilai Silhouette Coefficient yang mendekati 1 menandakan bahwa pengelompokan klaster relatif baik. Begitu juga dengan evaluasi Davies Bouldin Index menghasilkan nilai 0.672. Yang mana menunjukkan tingkat kompaknya klaster yang lebih rendah. Dari hasil eksperimen visualisasi scatter menunjukkan bahwa penyebaran data pada setiap klaster memiliki sebaran yang relatif seragam, hal ini mengindikasikan bahwa klaster-klaster tersebut memiliki karakterisik yang mendekati seragam didalam masing masing kelompok. Pendekatan menggunakan elbow memberikan panduan visual yang kuat dalam menentukan jumlah klaster yang optimal, selain itu penggunaan elbow menghindari keputusan jumlah klaster yang sepenuhnya subyektif. Pendekatan tanpa menggunakan elbow lebih bergantung pada penilaian domain dan kebutuhan bisnis, beberapa dataset yang kompleks dan tidak teratur sulit diidentifikasi dengan titik elbow, dan pada kasus ini pendekatan tanpa elbow dapat lebih bermanfaat.

DAFTAR RUJUKAN

Ahmed, M., Seraj, R., Islam, S.M.S., 2020. The k-means Algorithm: A Comprehensive Survey and Performance Evaluation. Electronics 9, 1295. https://doi.org/10.3390/electronics9081295

Beginner's Guide To K-Means Clustering [WWW Document], n.d. URL https://analyticsindiamag.com/beginners-guide-to-k-means-clustering/ (accessed 8.19.23).

Berkhin, P., 2006. A Survey of Clustering Data Mining Techniques, in: Kogan, J., Nicholas, C., Teboulle, M. (Eds.), Grouping Multidimensional Data: Recent Advances in Clustering. Springer, Berlin, Heidelberg, pp. 25–71. https://doi.org/10.1007/3-540-28349-8_2

Bholowalia, P., n.d. EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. International Journal of Computer Applications 105.

Bosnic, Z., Yildirim, P., Babič, F., Šahinović, I., Wittlinger, T., Martinović, I., Majnaric, L.T., 2021. Clustering Inflammatory Markers with Sociodemographic and Clinical Characteristics of Patients with Diabetes Type 2 Can Support Family Physicians' Clinical Reasoning by Reducing Patients' Complexity. Healthcare (Basel) 9, 1687. https://doi.org/10.3390/healthcare9121687

Chen, J., Qi, X., Chen, L., Chen, F., Cheng, G., 2020. Quantum-inspired ant lion optimized hybrid k-means for cluster analysis and intrusion detection. Knowledge-Based Systems 203, 106167. https://doi.org/10.1016/j.knosys.2020.106167

Debatty, T., Michiardi, P., Mees, W., Thonnard, O., 2014. Determining the k in k-means with MapReduce, in: EDBT/ICDT 2014 Joint Conference, Proceedings of the Workshops of the EDBT/ICDT 2014 Joint Conference. Athènes, Greece.

Ediyanto, M.N.M., 2013. PENGKLASIFIKASIAN KARAKTERISTIK DENGAN METODE K-MEANS CLUSTER ANALYSIS. Bimaster : Buletin Ilmiah Matematika, Statistika dan Terapannya 2. https://doi.org/10.26418/bbimst.v2i02.3033

Harahap, F., 2021. Perbandingan Algoritma K Means dan K Medoids Untuk Clustering Kelas Siswa Tunagrahita. TIN: Terapan Informatika Nusantara 2, 191–197.

Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, Award winning papers from the 19th International Conference on Pattern Recognition (ICPR) 31, 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Jain, A.K., Murty, M.N., Flynn, P.J., 1999. Data clustering: a review. ACM Comput. Surv. 31, 264–323. https://doi.org/10.1145/331499.331504

Larose, D.T., Larose, C.D., 2014. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons.

Merliana, N.P.E., Ernawati, Santoso, A.J., 2015. ANALISA PENENTUAN JUMLAH CLUSTER TERBAIK PADA METODE K-MEANS CLUSTERING. Proceeding SENDI_U.

Na, S., Xumin, L., Yong, G., 2010. Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm, in: 2010 Third International Symposium on Intelligent Information Technology and Security Informatics. Presented at the 2010 Third International Symposium on Intelligent Information Technology and Security Informatics, pp. 63–67. https://doi.org/10.1109/IITSI.2010.74

Syakur, M.A., Khotimah, B.K., Rochman, E.M.S., Satoto, B.D., 2018. Integration K-Means Clustering Method and Elbow Method For Identification of The Best Customer Profile Cluster. IOP Conf. Ser.: Mater. Sci. Eng. 336, 012017. https://doi.org/10.1088/1757-899X/336/1/012017

Tan, P.-N., Steinbach, M., Kumar, V., 2006. Introduction to Data Mining. Pearson Addison Wesley.

Vora, P., 2013. A Survey on K-mean Clustering and Particle Swarm Optimization 1.

Wu, X., Kumar, V., Ross Quinlan, J., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G.J., Ng, A., Liu, B., Yu, P.S., Zhou, Z.-H., Steinbach, M., Hand, D.J., Steinberg, D., 2008. Top 10 algorithms in data mining. Knowl Inf Syst 14, 1–37. https://doi.org/10.1007/s10115-007-0114-2