Volume 7 Nomor 2 Juli 2025

JJEEE

e-ISSN : 2715-0887
p-ISSN : 2654-7813

Jambura Journal of Electrical and Electronics Engineering

Optimalisasi Seleksi Atribut *K-Means* Menggunakan *Correlation Matrix* pada *Clustering* Penyakit Pasien

Optimization of K-Means Attribute Selection Using Correlation Matrix in Patient Disease Clustering

Amiruddin
Program Studi Teknik Informatika
Universitas Ichsan Gorontalo
Gorontalo, Indonesia
amier.76 @gmail.com

Rezqiwati Ishak
Program Studi Teknik Informatika
Universitas Ichsan Gorontalo
Gorontalo, Indonesia
rezqi.uig @gmail.com

Diterima : Oktober 2024
Disetujui : Mei 2025
Dipublikasi : Juli 2025

***Abstrak*— Kesehatan pasien merupakan elemen penting dalam sistem kesehatan masyarakat, di mana pengelompokan data penyakit dapat memfasilitasi identifikasi risiko dan perencanaan perawatan yang lebih efisien. Namun metode *clustering* konvensional seperti *K-Means* sering mengalami kesulitan dalam memisahkan *cluster* secara optimal, terutama ketika atribut yang digunakan tidak relevan atau berlebihan. Penelitian ini bertujuan untuk mengoptimalkan proses *clustering* data kesehatan pasien dengan menerapkan seleksi atribut menggunakan *Correlation Matrix* dan *Heatmap* dalam algoritma *K-Means*. Metode yang digunakan melibatkan normalisasi data dengan *StandardScaler* dan penentuan jumlah *cluster* optimal melalui *Elbow Method*, yang menghasilkan tiga *cluster* optimal. Seleksi atribut dilakukan untuk mengurangi redundansi, menyisakan fitur-fitur penting seperti umur, tinggi badan, dan indeks massa tubuh (IMT). Hasil analisis menunjukkan bahwa seleksi atribut secara signifikan meningkatkan *performa clustering*, dengan *Silhouette Score* meningkat dari 0,20 menjadi 0,54 dan *Davies-Bouldin Index (DBI)* menurun dari 1,60 menjadi 0,63. Visualisasi hasil *clustering* menggunakan *Principal Component Analysis (PCA)* menunjukkan pemisahan yang lebih jelas antar *cluster*, mencerminkan karakteristik pasien yang berbeda. Temuan ini menegaskan pentingnya seleksi atribut dalam proses *clustering* untuk mencapai hasil yang lebih optimal yang dapat membantu dalam memahami pola kesehatan pasien dan merancang intervensi yang lebih tepat.**

***Kata Kunci*— *Heatmap; Silhouette Score; K-Means; Principal Component Analysis (PCA); Elbow*.**

Abstract* — *Patient health is a critical element in public health systems, where grouping disease data can facilitate risk identification and more efficient treatment planning. However, conventional clustering methods such as K-Means often have difficulty in separating clusters optimally, especially when the attributes used are irrelevant or redundant. This study aims to optimize the clustering process of patient health data by applying attribute selection using Correlation Matrix and Heatmap in the K-Means algorithm. The method used involves normalizing the data with a StandardScaler and determining the optimal number of clusters through the Elbow Method, which results in three optimal clusters. Attribute selection is carried out to reduce redundancy, leaving important features such as age, height, and body mass index (BMI). The results of the analysis showed that attribute selection significantly improved clustering performance, with the Silhouette Score increasing from 0.20 to 0.54 and the Davies-Bouldin Index (DBI) decreasing from 1.60 to 0.63. Visualization of clustering results using Principal Component Analysis (PCA) shows a clearer separation between clusters, reflecting different patient characteristics. These findings confirm the importance of attribute selection in the clustering process to achieve more optimal results that can help in understanding patient health patterns and designing more appropriate interventions.

***Keywords*— *Heatmap; Silhouette Score; K-Means; Principal Component Analysis (PCA); Elbow*.**

I. PENDAHULUAN

Kesehatan masyarakat adalah salah satu aspek penting yang memengaruhi kualitas hidup suatu populasi. Dalam beberapa dekade terakhir, pengelolaan data kesehatan pasien menjadi semakin krusial, terutama dengan meningkatnya jumlah kasus penyakit kronis dan kompleksitas data yang dihasilkan oleh sistem perawatan kesehatan modern. Analisis data dalam bentuk pengelompokan (*clustering*) merupakan salah satu pendekatan yang dapat digunakan untuk memahami pola-pola penyakit, sehingga dapat membantu dalam proses diagnosis, prognosis, serta pengambilan keputusan klinis. Obyek penelitian ini adalah data kesehatan pasien yang terdiri dari variabel-variabel penting seperti jenis kelamin, umur, tinggi badan, berat badan, lingkar perut, Indeks Massa Tubuh (IMT), tekanan darah sistole, diastole, frekuensi nafas, detak nadi, dan diagnosa penyakit [1]. Dengan menganalisis data tersebut, kita dapat mengelompokkan pasien berdasarkan pola kesehatan dan penyakit mereka melalui teknik *clustering*. Algoritma *K-*

Means sering digunakan untuk tujuan ini karena efisiensi dan kemudahan penggunaannya dalam *clustering* data besar [2].

Beberapa metode seleksi atribut yang digunakan untuk meningkatkan performa algoritma *K-Means* antara lain *Principal Component Analysis* (PCA) [3][4], *Random Forest Feature Importance*[5][6], *Chi-Square Test*[7][8] dan *XGBoost*[9][10]. PCA secara efektif mereduksi dimensi data dengan memproyeksikan atribut ke dalam komponen utama, tetapi interpretasi hasilnya seringkali sulit karena transformasi linier yang dilakukan tidak mudah dikaitkan dengan atribut asli[3][4]. *Random Forest Feature Importance* dapat memberikan skor untuk setiap atribut berdasarkan kontribusinya terhadap klasifikasi, namun metode ini lebih cocok untuk tugas supervisi dibandingkan *unsupervised clustering* seperti *K-Means*[5][6]. *Chi-Square Test* sangat bermanfaat untuk data kategorikal, tetapi kurang efektif untuk data numerik yang umum dalam dataset medis[7][8]. *XGBoost* mampu menangani data besar dan memiliki fitur regularisasi yang meningkatkan akurasi model tetapi memerlukan pemahaman yang lebih mendalam tentang parameter tuning, dan lebih kompleks dibandingkan penggunaan matriks korelasi[9][10].

Salah satu masalah utama dalam penerapan *K-Means* adalah sensitivitas algoritma terhadap pemilihan atribut dan skala data. Tanpa pemilihan atribut yang baik, *K-Means* dapat menghasilkan *cluster* yang kurang akurat dan interpretasi yang salah karena pengaruh dari atribut yang tidak relevan atau redundan [11]. Kegagalan dalam mengidentifikasi atribut penting dapat menyebabkan hasil *clustering* yang tidak sesuai dengan kebutuhan klinis, yang pada akhirnya mempengaruhi pengambilan keputusan medis.

Solusi untuk mengatasi masalah ini adalah dengan menerapkan *Correlation Matrix with Heatmap* sebagai metode seleksi atribut. Metode ini memungkinkan visualisasi korelasi antar atribut, sehingga atribut yang tidak relevan atau yang sangat berkorelasi dapat diidentifikasi dan dihilangkan. Dengan *Heatmap*, peneliti dapat dengan cepat mengidentifikasi atribut yang paling berpengaruh terhadap hasil clustering, mengoptimalkan kinerja *K-Means*, dan meningkatkan akurasi *clustering*. Penggunaan metode ini telah terbukti efektif dalam mengurangi kompleksitas komputasi sekaligus menjaga kualitas hasil *clustering* [2][12].

Beberapa penelitian terkait telah mengkaji penggunaan berbagai metode seleksi atribut untuk algoritma *K-Means*. Misalnya, penelitian oleh [11] membahas bagaimana seleksi atribut dapat meningkatkan hasil *clustering* dalam konteks data besar. Sementara itu [13] dan [14] menunjukkan bagaimana metode-metode seleksi atribut dapat mengoptimalkan stabilitas dan efisiensi algoritma *K-Means,* terutama dalam pengelompokan dataset yang besar dan kompleks.

Penelitian ini memiliki nilai kebaruan penelitian terletak pada kombinasi teknik seleksi atribut yang baru (*Correlation Matrix with Heatmap),* pengoptimalan kinerja *K-Means* dalam data medis, serta fokus pada atribut spesifik pasien yang berkontribusi langsung pada praktik klinis. Ini memberi pendekatan baru yang lebih akurat dan relevan dalam *clustering* data pasien, yang dapat membantu membuat keputusan klinis yang lebih baik dan personal.

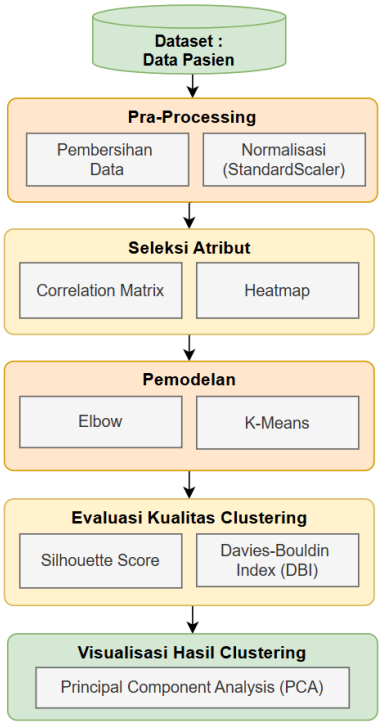
Penelitian ini akan menggunakan *StandardScaler* untuk normalisasi data, memastikan semua fitur memiliki skala yang sebanding. Setelah itu, algoritma *K-Means* akan digunakan untuk melakukan *clustering*, dengan teknik *Elbow* untuk menentukan jumlah *cluster* yang optimum. Penilaian kualitas hasil *clustering* akan dilakukan menggunakan *Silhouette Score* dan *Davies-Bouldin Index (DBI).* Selanjutnya, visualisasi hasil *clustering* akan dilakukan menggunakan *Principal Component Analysis (PCA)* untuk memahami distribusi dan separasi antar *cluster*.

Tujuan penelitian ini adalah untuk mengoptimalkan *clustering* penyakit pasien dengan menerapkan *Correlation Matrix with Heatmap* untuk seleksi atribut, dan menggunakan *K-Means* sebagai algoritma *clustering* utama. Dengan pendekatan ini, diharapkan dapat dihasilkan *cluster* yang lebih akurat dan relevan dalam konteks diagnosis dan manajemen pasien. Penelitian ini juga berfokus pada evaluasi hasil *clustering* menggunakan *Silhouette Score* dan *Davies-Bouldin Index* untuk memastikan kualitas pembagian *cluster* yang dihasilkan.

II. METODE

A. *Desain Penelitian*

Desain penelitian yang digunakan dalam penelitian ini terdiri dari pengumpulan dataset, *pra-processing*, seleski atribut, pemodelan, evaluasi kualitas *clustering* dan visualisasi hasil *clustering*, seperti pada Gambar 1.



Gambar 1. Desain Penelitian

1. Sumber Dataset: Dataset yang digunakan dalam penelitian ini diperoleh dari Dinas Kesehatan Kabupaten Bone Bolango sebanyak 488[1]. Data ini mencakup sejumlah atribut, seperti umur, jenis kelamin, tinggi badan, berat badan, lingkar perut, indeks massa tubuh (IMT), tekanan darah sistole dan diastole, frekuensi pernapasan, dan detak nadi.
2. *Pra-Processing*: Sebelum dilakukan analisis, data mengalami proses pra-pemrosesan sebagai berikut:

- Pembersihan Data: Menghapus kolom yang tidak relevan dan menghapus satuan nilai pada atribut.
- Normalisasi: Data dinormalisasi menggunakan *StandardScaler* dari pustaka *scikit-learn* untuk memastikan semua fitur memiliki skala yang sebanding, sehingga tidak ada fitur yang mendominasi hasil *clustering*.

3. Seleksi Atribut: *Correlation Matrix* dihitung untuk mengidentifikasi hubungan antar atribut dalam dataset. Visualisasi menggunakan *Heatmap* dilakukan untuk membantu dalam seleksi atribut. Atribut dengan korelasi tinggi satu sama lain dievaluasi dan fitur yang *redundant* dihilangkan untuk meningkatkan efisiensi algoritma *clustering*.
4. Pemodelan: Setelah proses seleksi atribut, algoritma *K-Means* diterapkan pada data yang telah dinormalisasi. Proses *clustering* dilakukan dengan langkah-langkah berikut:
   - Penentuan Jumlah *Cluster*: Menggunakan *Elbow Method* untuk menentukan jumlah *cluster* optimal k. Grafik inersia terhadap jumlah *cluster* k dianalisis untuk menemukan titik *elbow*.
   - Pelatihan Model: Model *K-Means* dilatih dengan jumlah *cluster* yang telah ditentukan. Data pasien dikelompokkan ke dalam *cluster* berdasarkan kemiripan fitur.
5. Evaluasi Hasil *Clustering*: Kualitas hasil *clustering* dievaluasi menggunakan dua metrik, yaitu *Silhouette Score* dan *Davies-Bouldin Index (DBI)*. *Silhouette Score* diukur untuk menilai seberapa baik data dalam *cluster* terpisah dari *cluster* lain, sedangkan *Davies-Bouldin Index* (DBI) digunakan untuk menilai rasio jarak antar *cluster* dengan ukuran *cluster*. Metrik ini memberikan gambaran tentang kejelasan dan efektivitas pengelompokan yang dilakukan.
6. Visualisasi Hasil *Clustering*: Untuk memahami distribusi dan separasi antar *cluster*, visualisasi hasil *clustering* dilakukan dengan menggunakan *Principal Component Analysis (PCA)*. PCA digunakan untuk mereduksi dimensi data ke dalam dua komponen utama yang dapat divisualisasikan secara grafis.

B. StandardScaler

StandardScaler merupakan teknik normalisasi yang digunakan untuk memastikan bahwa data memiliki distribusi yang seimbang sebelum diterapkan pada algoritma pembelajaran mesin[15]. Normalisasi sangat penting, terutama ketika data yang digunakan memiliki skala yang berbeda antar fitur, seperti tinggi badan dalam cm, berat badan dalam kg, atau tekanan darah dalam mmHg. *StandardScaler* adalah metode normalisasi yang berfungsi dengan menstandarisasi fitur agar memiliki *mean* nol dan *varian* satu.

C. *Correlation Matrix with Heatmap*

Correlation Matrix adalah representasi tabel dari hubungan antar variabel dalam dataset, biasanya diukur dengan koefisien korelasi Pearson atau metode lain seperti *Spearman* atau *Kendall*. Koefisien korelasi Pearson adalah ukuran linier dari kekuatan dan arah hubungan antara dua variabel. Nilainya berkisar antara -1 hingga 1, di mana: 1 menunjukkan korelasi positif sempurna, 0 berarti tidak ada korelasi, -1 menunjukkan korelasi negatif sempurna.
Heatmap adalah representasi grafis dari data yang menunjukkan nilai melalui warna. Dalam konteks *Correlation Matrix*, *Heatmap* membantu memvisualisasikan korelasi antar variabel, membuatnya lebih mudah untuk mendeteksi pola hubungan antar atribut. *Heatmap* memberikan warna untuk mewakili kekuatan korelasi, di mana warna lebih terang atau lebih gelap menunjukkan korelasi yang lebih kuat atau lemah [12].

D. Metode Elbow

Metode *Elbow* bekerja dengan menghitung nilai within-*cluster sum of squares (WCSS)* atau dikenal sebagai *inertia*, yaitu total jarak kuadrat dari setiap titik data ke *centroid cluster* terdekat. Secara umum, semakin besar jumlah *cluster* (k), semakin kecil nilai *WCSS*, karena titik data akan lebih dekat ke *centroid* yang lebih banyak [16][17][18].
Berikut tahapan dalam metode *Elbow*:

1. Lakukan *clustering* pada dataset dengan berbagai jumlah *cluster*.
2. Hitung variabilitas dari setiap hasil *clustering*.
3. Buat grafik yang membandingkan jumlah *cluster* dengan variabilitas.
4. Cari titik siku pada grafik, yang menunjukkan di mana penurunan variabilitas mulai melambat secara signifikan seiring dengan peningkatan jumlah *cluster*.

E. K-Means Clustering

Algoritma *K-Means* adalah salah satu teknik *unsupervised learning* yang paling populer untuk *clustering*. Tujuan utamanya adalah mengelompokkan dataset ke dalam beberapa *cluster* berdasarkan kesamaan antar data. Algoritma ini bekerja dengan meminimalkan variansi di dalam *cluster* sehingga data dalam satu *cluster* lebih dekat satu sama lain dibandingkan dengan data di cluster lainnya. Salah satu algoritma *clustering* yang paling umum digunakan adalah *K-Means* [19][20].
Secara umum, langkah-langkah dalam algoritma *K-Means* dapat dijelaskan sebagai berikut:

Langkah 1: Inisialisasi

a. Tentukan jumlah *cluster* yang akan dibentuk.
b. Pilih titik awal sebagai *centroid* untuk setiap *cluster*, bisa dilakukan secara acak atau dengan metode inisialisasi yang dipilih.tertentu.

Langkah 2: *Assignment*

a. Hitung jarak antara setiap titik data dan *centroid cluster* menggunakan metrik jarak tertentu, seperti jarak *Euclidean* atau sesuai dengan rumus pada persamaan (1).

$$d_{ik} = \sqrt{\sum_{i=1}^{n} \left(C_{ij} - C_{kj}\right)^2} \quad (1)$$

di mana:

C_{ij} = pusat *cluster*
C_{kj} = data

b. Tentukan *cluster* terdekat untuk setiap titik data dengan memilih *centroid* yang memiliki jarak terkecil.
c. Tetapkan setiap titik ke *cluster* yang sesuai.

Langkah 3: *Update Centroid*

a. Hitung rata-rata posisi dari semua titik dalam setiap *cluster* yang terbentuk.
b. Gunakan nilai rata-rata tersebut sebagai *centroid* baru untuk setiap *cluster*.

Langkah 4: Iterasi, ulangi langkah 2 dan 3 sampai tercapai konvergensi, yaitu ketika tidak ada perubahan dalam penetapan *cluster* atau jumlah iterasi maksimum tercapai.

F. Davies-Bouldin Index

Davies-Bouldin Index (DBI) adalah metode evaluasi untuk menilai kualitas hasil *clustering* dengan membandingkan jarak antar *cluster* dengan ukuran dalam-*cluster*. *DBI* adalah ukuran rata-rata *similarity* antar *cluster*, di mana nilai lebih rendah menunjukkan *clustering* yang lebih baik. Nilai *DBI* memperhitungkan baik jarak antar *cluster* maupun ukuran *cluster* [16][21][22].

Langkah-langkah penghitungan *Davies-Bouldin Index (DBI)* adalah sebagai berikut:

1. Lakukan proses *clustering* pada dataset menggunakan algoritma *K-Means*.
2. Hitung *centroid* atau pusat dari setiap *cluster* yang terbentuk.
3. Hitung jarak antar *centroid* untuk setiap pasangan *cluster*, kemudian simpan hasilnya dalam bentuk matriks jarak antar-*cluster*.
4. Untuk setiap *cluster*, hitung nilai R_{ij} , yaitu rata-rata jarak antara *centroid cluster* tersebut dengan semua titik yang berada di dalam *cluster* menggunakan persamaan (2).

$$R_{ij} = \frac{S_i + S_j}{d_{ij}} \quad (2)$$

di mana:
- R_{ij} adalah nilai rata-rata jarak untuk *cluster* i dan j.
- d_{ij} adalah jarak antara *centroid cluster* i dan j.
- S_i dan S_j adalah jarak rata-rata antara setiap titik cluster I dan *centroid* dari *cluster* j

5. Hitung nilai DB, yang merupakan rata-rata dari semua nilai DB untuk setiap *cluster*, sesuai dengan persamaan (3)

$$DB = \frac{1}{k}\sum_{i=1}^{k} max R_{ij, i \neq j} \quad (3)$$

G. Silhouette Score

Silhouette Score adalah metode untuk mengukur kualitas hasil *clustering*, memberikan penilaian seberapa baik data terkelompok. Ini digunakan untuk menentukan seberapa dekat suatu titik data dengan *cluster* yang menjadi anggotanya dibandingkan dengan *cluster* lain. Skor ini berkisar antara -1 hingga 1 [23]:

- Nilai 1 menunjukkan bahwa data sangat cocok dengan cluster-nya dan jauh dari *cluster* lainnya.
- Nilai 0 menunjukkan bahwa data berada di batas antara dua *cluster*.
- Nilai -1 menunjukkan bahwa data lebih dekat ke *cluster* lain daripada *cluster*-nya sendiri, menunjukkan pengelompokan yang buruk.

H. Principal Component Analysis (PCA)

Principal Component Analysis (PCA) adalah metode statistik yang digunakan untuk mereduksi dimensi data multivariat tanpa kehilangan informasi yang signifikan. PCA bertujuan untuk menemukan komponen utama *(principal components)* dari sebuah dataset, yang merupakan kombinasi linier dari variabel asli dengan tujuan memaksimalkan variasi dalam data. Dalam konteks *clustering*, PCA sering digunakan untuk memproyeksikan data berdimensi tinggi ke dalam dua atau tiga dimensi agar dapat divisualisasikan, sehingga hasil *clustering* menjadi lebih mudah diinterpretasikan[24].

III. HASIL DAN PEMBAHASAN

Proses eksperimen pada penelitian ini menggunakan bahasa pemrograman *Python* dengan *tools* editor *Google Colaboratory*.

A. *Pengumpulan Dataset*

Dataset yang digunakan merupakan *dataset private* yaitu data data pasien pada Dinas Kesehatan Kabupaten Bone Bolango seperti terlihat pada Tabel 1[1].

TABEL 1. SAMPEL DATASET

No	JK	Umur Tahun	Tinggi Badan	Berat Badan	Lingkar Perut	IMT	Sistole	Diastole	Nafas	Detak Nadi
1	0	22	155	70	80	29,14	130	100	20	80
2	0	69	157	49	76	19,88	120	80	20	90
3	0	48	155	69	65	28,72	130	90	22	80
4	1	49	155	63	90	26,22	110	79	22	92
5	0	1	72	8,8	47	16,98	90	60	32	110
6	1	26	143	51	76	24,94	116	70	20	80
7	1	1	65	8	41	18,93	100	90	30	80
8	1	20	157	43,2	78	17,53	91	77	28	80
9	1	68	153	45	82	19,22	132	74	20	88
10	0	68	151	51,8	84	22,72	103	55	22	88

B. Pra-Processing

Dataset yang digunakan dalam penelitian ini terdiri dari sejumlah variabel kesehatan pasien seperti umur, jenis kelamin, tinggi badan, berat badan, lingkar perut, indeks massa tubuh (IMT), tekanan darah sistole dan diastole, frekuensi pernapasan, dan detak nadi. Setelah dilakukan pembersihan data, kolom-kolom yang tidak relevan seperti nama pasien dan diagnosa penyakit dihapus. Selain itu, data yang bersifat numerik seperti umur, tinggi badan, berat badan, dan variabel lainnya dikonversi ke dalam bentuk numerik yang dapat dianalisis lebih lanjut.

Untuk memastikan semua fitur memiliki skala yang sebanding, proses normalisasi dilakukan menggunakan *StandardScaler*. Langkah ini penting agar tidak ada fitur yang mendominasi *clustering* karena perbedaan skala yang besar. Setelah proses normalisasi selesai, data diubah ke dalam bentuk standar dengan nilai rata-rata nol dan standar deviasi satu.. Hasil normalisasi data hasilnya seperti pada Tabel 2.

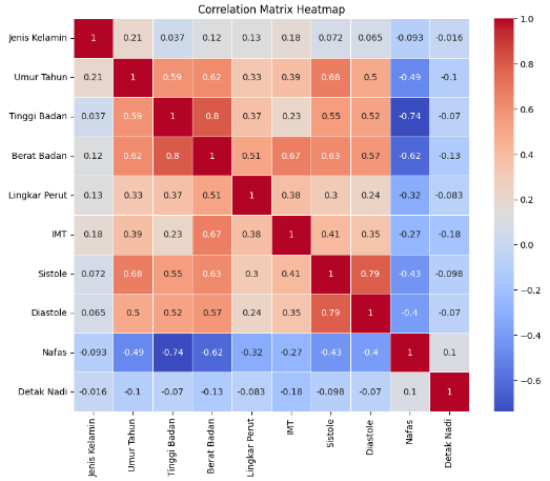
TABEL 2. SAMPEL HASIL NORMALISASI DATA

JK	Umur Tahun	Tinggi Badan	Berat Badan	Lingkar Perut	IMT	Sistole	Diastole	Nafas	Detak Nadi
-1,37	-0,45	0,43	0,90	0,10	0,76	0,88	1,94	-0,50	-0,43
-1,37	1,90	0,50	-0,07	-0,02	-0,41	0,40	0,39	-0,50	0,53

JK	Umur Tahun	Tinggi Badan	Berat Badan	Lingkar Perut	IMT	Sistole	Diastole	Nafas	Detak Nadi
-1,37	0,85	0,43	0,85	-0,34	0,71	0,88	1,17	-0,18	-0,43
0,73	0,90	0,43	0,57	0,39	0,39	-0,08	0,32	-0,18	0,72
-1,37	-1,51	-2,68	-1,93	-0,87	-0,77	-1,03	-1,15	1,41	2,45
0,73	-0,25	-0,02	0,02	-0,02	0,23	0,21	-0,38	-0,50	-0,43
0,73	-1,51	-2,94	-1,97	-1,05	-0,53	-0,55	1,17	1,09	-0,43
0,73	-0,55	0,50	-0,34	0,04	-0,70	-0,98	0,16	0,78	-0,43
0,73	1,85	0,35	-0,26	0,16	-0,49	0,97	-0,07	-0,50	0,34
-1,37	1,85	0,28	0,06	0,22	-0,05	-0,41	-1,54	-0,18	0,34

C. *Correlation Matrix dan Seleksi Atribut*

Langkah pertama dalam analisis ini adalah menghitung *Correlation Matrix* untuk memahami hubungan antar fitur dalam dataset. Hasil *Correlation Matrix* divisualisasikan menggunakan *Heatmap* seperti yang ditunjukkan pada Gambar 3.

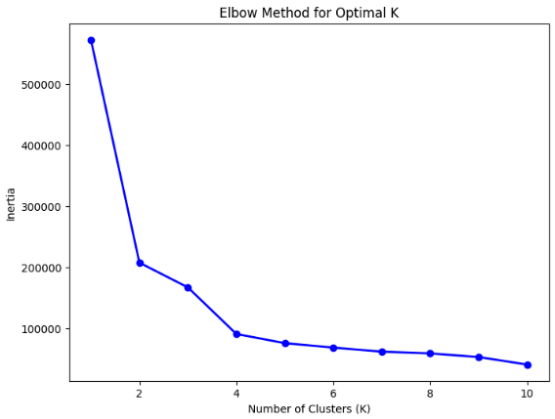


Gambar 3. *Correlation Matrix Heatmap*

Berdasarkan dari Gambar 3, terlihat bahwa terdapat beberapa fitur yang memiliki korelasi tinggi, seperti berat badan dan lingkar perut dengan indeks massa tubuh (IMT). Karena korelasi yang tinggi, fitur seperti berat badan atau IMT dapat dipertimbangkan untuk dihilangkan guna mengurangi redundansi informasi. Pada penelitian ini, kami mempertahankan IMT sebagai representasi dari hubungan berat badan dan tinggi badan, serta menghapus berat badan dari analisis *clustering*.

D. *Penentuan Jumlah Cluster dengan Elbow Method*

Penentuan jumlah *cluster* yang optimal, digunakan *Elbow Method*. Metode ini mengevaluasi nilai inersia *(within-cluster sum of squares)* untuk berbagai jumlah *cluster* dan mencari titik "siku" di mana penurunan inersia mulai melambat. Hasil dari *Elbow Method* ditunjukkan pada Gambar 4.



Gambar 4. *Elbow Method*

Berdasarkan Gambar 4, terlihat bahwa jumlah *cluster* optimal berada pada tiga *cluster*. Hal ini dapat diidentifikasi dari grafik, di mana penurunan *inersia* mulai tidak signifikan setelah titik ketiga. Oleh karena itu, nilai k = 3 ditetapkan sebagai jumlah *cluster* optimal untuk pemodelan *K-Means* dalam penelitian ini.

E. *Hasil Clustering Tanpa Seleksi Atribut*

Pertama dilakukan *clustering* tanpa seleksi atribut, menggunakan semua fitur yang ada. Hasil *clustering* ditampilkan pada Tabel 2.

TABEL 2. HASIL *CLUSTERING* TANPA SELEKSI ATRIBUT

Atribut	*Cluster* 1	*Cluster* 2	*Cluster* 3
Umur Tahun	30,11	46,74	3,30
Tinggi Badan	153,89	155,77	94,82
Berat Badan	52,77	67,29	13,11
IMT	22,29	27,99	15,75
Sistole	105,52	132,14	86,95
Diastole	72,12	85,80	60,66
Nafas	21,05	20,98	32,33
Detak Nadi	83,60	84,19	87,21
Lingkar Perut	76,47	91,83	48,04
Jenis Kelamin	0,72	0,67	0,46

Hasil menunjukkan bahwa meskipun *cluster* terdefinisi, beberapa fitur tidak memberikan kontribusi signifikan dalam membedakan *cluster* secara jelas, dan terdapat tumpang tindih antara *cluster* yang divisualisasikan hasilnya pada Gambar 7.

F. *Hasil Clustering dengan Seleksi Atribut*

Selanjutnya *clustering* diterapkan menggunakan fitur yang telah diseleksi. Hasil *clustering* ditampilkan pada Tabel 3.

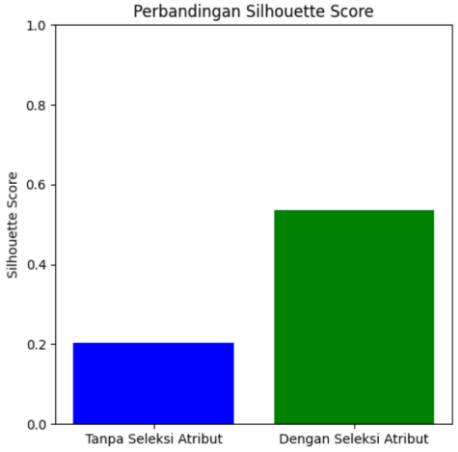
TABEL 3. HASIL *CLUSTERING* DENGAN SELEKSI ATRIBUT

Atribut	*Cluster* 1	*Cluster* 2	*Cluster* 3
Umur Tahun	38,01	2,61	6,91
Tinggi Badan	155,41	75,57	120,01
IMT	24,76	19,94	14,02

Dibandingkan dengan hasil sebelumnya, *cluster* yang terbentuk setelah seleksi atribut menunjukkan perbedaan yang lebih jelas dalam karakteristik. Misalnya, nilai IMT dan umur menunjukkan distribusi yang lebih konsisten dalam setiap *cluster* sebagaimana hasil visualissinya pada Gambar 8.

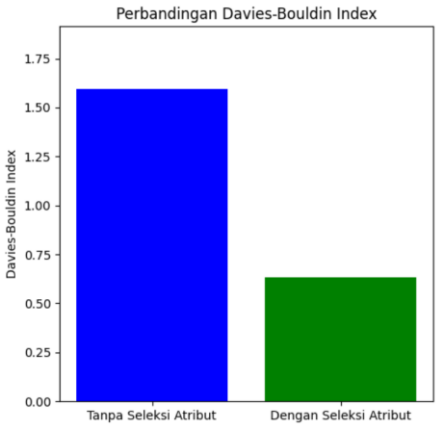
G. *Evaluasi Hasil Clustering*

Kualitas *clustering* dievaluasi menggunakan *Silhouette Score* dan *Davies-Bouldin Index* untuk kedua pendekatan. Hasil evaluasi hasil *clustering* tanpa seleksi atribut dan dengan seleksi atribut dapat dilihat pada Gambar 5.



Gambar 5. Perbandingan *Silhouette Score*

Berdasarkan *Silhouette Score*, pengelompokan dengan seleksi atribut menghasilkan skor yang lebih tinggi, yang berarti *cluster* yang terbentuk lebih kompak dan terpisah dengan jelas. Ini menunjukkan bahwa fitur-fitur yang dipilih (umur, tinggi badan, IMT) adalah atribut yang cukup kuat untuk membedakan pasien ke dalam kelompok-kelompok yang berbeda, tanpa membutuhkan fitur tambahan lainnya.



Gambar 6. Perbandingan *Davies-Bouldin Index* (DBI)

Berdasarkan *Davies-Bouldin Index (DBI)* untuk *clustering* dengan seleksi atribut lebih rendah dibandingkan tanpa seleksi. Ini menunjukkan bahwa setelah seleksi atribut, *cluster* yang terbentuk lebih kompak dan terpisah dengan lebih baik, yang berarti pemisahan antar kelompok pasien menjadi lebih jelas. Kedua hasil metrik evaluasi di atas lebih detail diuraikan pada Tabel 4.

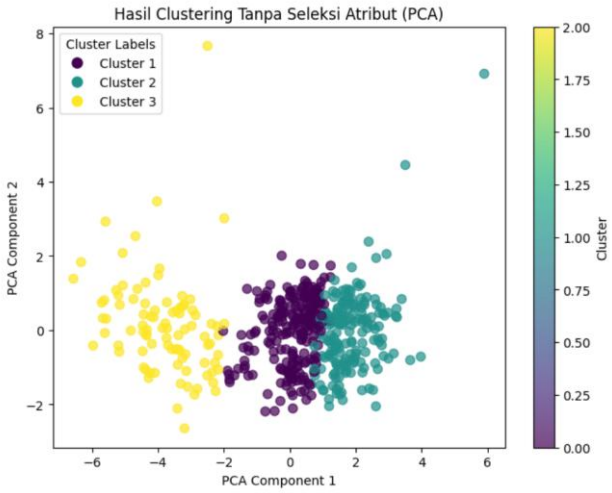
TABEL 4. METRIKS EVALUASI

Metriks Evaluasi	Tanpa Seleksi Atribut (10 Atribut)	Dengan Seleksi Atribut (3 Atribut)	Selisih
Silhouette Score	0,20	0,54	0,34
Davies-Bouldin Index (DBI)	1,60	0,63	-0,97

Berdasarkan Tabel 4. terlihat bahwa *Silhouette Score* meningkat dari 0,20 (tanpa seleksi atribut) menjadi 0,54 (dengan seleksi atribut), menunjukkan pemisahan *cluster* yang lebih baik dan *Davies-Bouldin Index (DBI)* menurun dari 1,60 menjadi 0,63, mengindikasikan bahwa *cluster* lebih terpisah dan memiliki karakteristik yang berbeda. Selisih antara kedua kondisi menunjukkan peningkatan performa *clustering*, dengan *Silhouette Score* meningkat sebesar 0,34 dan DBI menurun sebesar 0,97. Temuan ini menegaskan pentingnya seleksi atribut dalam analisis data kesehatan pasien untuk menghasilkan pemisahan *cluster* yang lebih jelas dan relevan.

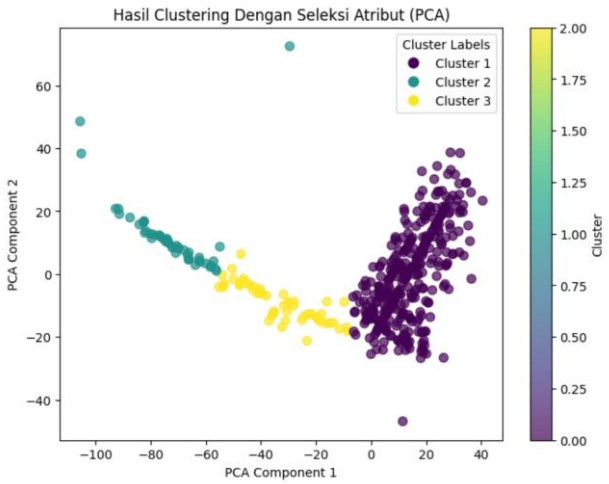
H. *Visualisasi dan Analisis Hasil Clustering*

Hasil *clustering* juga divisualisasikan menggunakan PCA untuk kedua pendekatan. Visualisasi tanpa seleksi atribut menunjukkan tumpang tindih yang signifikan antara *cluster*, sedangkan visualisasi setelah seleksi atribut menunjukkan pemisahan yang lebih jelas, seperti pada Gambar 7 dan Gambar 8.



Gambar 7. Visualisasi Hasil *Clustering* tanpa Seleksi Atribut

Visualisasi hasil *clustering* tanpa seleksi atribut pada Gambar 7, menunjukkan tumpang tindih antara *cluster*, terutama antara *Cluster* 1 (Ungu) dan *Cluster* 2 (Hijau). *Cluster* yang terbentuk juga terlihat kurang kompak dan lebih menyebar, terutama pada *Cluster* 2, yang menunjukkan bahwa ada beberapa atribut yang mungkin tidak relevan atau terlalu beragam sehingga menyebabkan pemisahan *cluster* yang kurang jelas



Gambar 8. Visualisasi Hasil *Clustering* dengan Seleksi Atribut

Setelah seleksi atribut dilakukan, maka hasil *clustering* menunjukkan pemisahan yang jauh lebih jelas antara *cluster*. *Cluster* 1 (Ungu) dan *Cluster* 3 (Kuning) menjadi lebih kompak dan terisolasi dari *cluster* lain. Ini menunjukkan bahwa atribut yang diseleksi (misalnya, umur, tinggi badan, dan IMT) lebih efektif dalam membedakan karakteristik pasien sehingga menghasilkan pemisahan yang lebih baik antar *cluster.*

Perbandingan hasil *clustering* menunjukkan bahwa seleksi atribut berkontribusi positif terhadap pemisahan *cluster*. Seleksi atribut yang tepat tidak hanya menghilangkan redundansi, tetapi juga meningkatkan kualitas hasil *clustering.*

IV. KESIMPULAN

Penelitian ini berhasil menunjukkan bahwa penerapan seleksi atribut menggunakan *Correlation Matrix* dan *Heatmap* secara signifikan meningkatkan kualitas *clustering* dalam data kesehatan pasien. Proses seleksi atribut tidak hanya mengurangi redundansi data, tetapi juga meningkatkan pemisahan antar *cluster*. Hasil evaluasi menunjukkan peningkatan *Silhouette Score* dari 0,20 menjadi 0,54, serta penurunan *Davies-Bouldin Index (DBI)* dari 1,60 menjadi 0,63 setelah seleksi atribut, yang mengindikasikan bahwa *cluster* yang terbentuk lebih terdefinisi dengan baik. Selain itu, visualisasi menggunakan *Principal Component Analysis (PCA)* mengonfirmasi adanya pemisahan yang lebih jelas antar *cluster* setelah seleksi atribut. Tiga *cluster* yang relevan yang diidentifikasi mencakup umur, tinggi badan, dan indeks massa tubuh (IMT), yang mencerminkan pola kesehatan yang berbeda. Temuan ini memberikan wawasan berharga untuk pengelompokan pasien berdasarkan karakteristik kesehatan mereka. Dengan demikian, seleksi atribut terbukti berperan penting dalam mengoptimalkan algoritma *K-Means* untuk *clustering* data Kesehatan pasien, yang dapat membantu meningkatkan kualitas analisis dalam konteks medis.

REFFERENSI

[1] Dinkes, "Laporan Harian Pelayanan Pasien," Gorontalo, 2024.

[2] J. Zhao, Y. Bao, D. Li, and X. Guan, "An Improved K-Means Algorithm Based on Contour Similarity," *Mathematics*, vol. 12, no. 14, p. 2211, Jul. 2024, doi: 10.3390/math12142211.

[3] W. Lv, W. Tang, H. Huang, and T. Chen, "Research and Application of Intersection Clustering Algorithm Based on PCA Feature Extraction and K-Means," *J. Phys. Conf. Ser.*, vol. 1861, no. 1, p. 012001, Mar. 2021, doi: 10.1088/1742-6596/1861/1/012001.

[4] D. Andra and A. B. Baizal, "E-commerce Recommender System Using PCA and K-Means Clustering," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 6, no. 1, pp. 57–63, Feb. 2022, doi: 10.29207/resti.v6i1.3782.

[5] H. A. Rosyid, U. Pujianto, and M. R. Yudhistira, "Classification of Lexile Level Reading Load Using the K-Means Clustering and Random Forest Method," *Kinet. Game Technol. Inf. Syst. Comput. Network, Comput. Electron. Control*, pp. 139–146, May 2020, doi: 10.22219/kinetik.v5i2.897.

[6] X. Zhao *et al.*, "ECFS-DEA: an ensemble classifier-based feature selection for differential expression analysis on expression profiles," *BMC Bioinformatics*, vol. 21, no. 1, p. 43, Dec. 2020, doi: 10.1186/s12859-020-3388-y.

[7] N. Yusliani, S. A. Q. Aruda, M. D. Marieska, D. M. Saputra, and A. Abdiansah, "The effect of Chi-Square Feature Selection on Question Classification using Multinomial Naïve Bayes," *Sinkron*, vol. 7, no. 4, pp. 2430–2436, Oct. 2022, doi: 10.33395/sinkron.v7i4.11788.

[8] M. R. Mahmood, "Two Feature Selection Methods Comparison Chi-square and Relief-F for Facial Expression Recognition," *J. Phys. Conf. Ser.*, vol. 1804, no. 1, p. 012056, Feb. 2021, doi: 10.1088/1742-6596/1804/1/012056.

[9] A. Bengnga and R. Ishak, "Penerapan XGBoost untuk Seleksi Atribut pada K-Means dalam Clustering Penerima KIP Kuliah," *Jambura J. Electr. Electron. Eng.*, vol. 5, no. 2, pp. 192–196, 2023, doi: 10.37905/jjeee.v5i2.20253.

[10] J. Henriques, F. Caldeira, T. Cruz, and P. Simões, "Combining K-Means and XGBoost Models for Anomaly Detection Using Log Datasets," *Electronics*, vol. 9, no. 7, p. 1164, Jul. 2020, doi: 10.3390/electronics9071164.

[11] C. Jie, Z. Jiyue, W. Junhui, W. Yusheng, S. Huiping, and L. Kaiyan, "Review on the Research of K-means Clustering Algorithm in Big Data," in *2020 IEEE 3rd International Conference on Electronics and Communication Engineering (ICECE)*, Dec. 2020, pp. 107–111, doi: 10.1109/ICECE51594.2020.9353036.

[12] A. Bengnga and R. Ishak, "Implementasi Seleksi Fitur Klasifikasi Waktu Kelulusan Mahasiswa Menggunakan Correlation Matrix with Heatmap," *Jambura J. Electr. Electron. Eng.*, vol. 4, no. 2, pp. 169–174, Jul. 2022, doi: 10.37905/jjeee.v4i2.14403.

[13] A. C. Pellicelli, "Application of an K-means Improved Clustering Analysis Algorithm in the Design of Resource Management Information System," 2022.

[14] M. Lv, "Application of an K-means Improved Clustering Analysis Algorithm in the Design of Resource Management Information System," in *2022 World Automation Congress (WAC)*, Oct. 2022, pp.

158–162, doi: 10.23919/WAC55640.2022.9934387.
[15] S. Rajesh, P. Praveen, and D. N, "Performance Analysis of Machine Learning Algorithms on Parkinson's Disease Data," in *2024 IEEE International Conference on Contemporary Computing and Communications (InC4)*, Mar. 2024, pp. 1–10, doi: 10.1109/InC460750.2024.10649372.
[16] "Clustering Performance Evaluation," *scikit-learn developers*. https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation (accessed May 20, 2023).
[17] A. B. H. Kiat, Y. Azhar, and V. Rahmayanti, "Penerapan Metode K-Means Dengan Metode Elbow Untuk Segmentasi Pelanggan Menggunakan Model RFM (Recency, Frequency & Monetary)," *Repositor*, vol. 2, no. 7, pp. 945–952, 2020.
[18] R. Ishak and Amiruddin, "Clustering Tingkat Pemahaman Dasar Mahasiswa Pada Pra-Perkuliahan Probabilitas Statistika Dengan Metode K-Means," *Jambura J. Electr. Electron. Eng.*, vol. 4, pp. 65–69, 2022, doi: 10.37905/jjeee.v4i1.11997.
[19] R. Primartha, *Algoritma Machine Learning*. Bandung: Informatika, 2021.
[20] "Clustering," *scikit-learn developers*. https://scikit-learn.org/stable/modules/clustering.html# (accessed May 20, 2023).
[21] A. K. Singh, S. Mittal, P. Malhotra, and Y. V. Srivastava, "Clustering Evaluation by Davies-Bouldin Index(DBI) in Cereal data using K-Means," in *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, Mar. 2020, pp. 306–310, doi: 10.1109/ICCMC48092.2020.ICCMC-00057.
[22] E. Muningsih, I. Maryani, and V. R. Handayani, "Penerapan Metode K-Means dan Optimasi Jumlah Cluster dengan Index Davies Bouldin untuk Clustering Propinsi Berdasarkan Potensi Desa," *J. Sains dan Manaj.*, vol. 9, no. 1, p. 96, 2021, doi: 10.31294/evolusi.v9i1.10428.
[23] Suyanto, *Data Mining untuk Klasifikasi dan Klasterisasi Data*. Bandung: Informatika, 2019.
[24] I. Turbay, P. Ortiz, and R. Ortiz, "Statistical analysis of principal components (PCA) in the study of the vulnerability of Heritage Churches," *Procedia Struct. Integr.*, vol. 55, pp. 168–176, 2024, doi: 10.1016/j.prostr.2024.02.022.