Jurnal Ilmiah Sistem Informasi dan Ilmu Komputer
Volume 5, Nomor 3, November 2025
E-ISSN :2827-7953; P-ISSN : 2827-8135, Hal. 638-650
DOI: https://doi.org/10.55606/juisik.v5i3.1811
Tersedia: https://journal.sinov.id/index.php/juisik

Segmentasi Pasien Berbasis K-Means dari Tanda-tanda Vital dan Demografi: Pendekatan *Unsupervised Learning* untuk Profil Risiko Klinis

Renny Afriany[1*], Rudolf Sinaga[2], Samsinar[3]
[1]Administrasi Rumah Sakit, STIKES Garuda Putih, Indonesia
[2]Sistem Informasi, Universitas Dinamika Bangsa, Indonesia
[3]Diploma Keperawatan, STIKES Garuda Putih, Indonesia
[*]*Penulis Korespondensi: reniafriani.44@gmail.com[1]*

Abstract. *Digital transformation in the health system demands the use of clinical data more strategically to support evidence-based decision-making. This study aims to explore the application of the K-Means Clustering algorithm in patient segmentation based on a combination of vital signs (systolic and diastolic blood pressure) and demographic characteristics (age, weight, and gender). Data on 1,401 outpatients was obtained from the medical record system of a hospital in Indonesia, then processed through the stages of preprocessing, standardization, and dimensionality reduction using PCA. The results of the elbow method showed that the optimal number of clusters was 3 (k=3). Descriptive analysis showed that Cluster 0 consisted of 100% women with normal blood pressure (124/77 mmHg) and an average body weight of 55.6 kg; Cluster 1 consists of the majority of women with high blood pressure (160.8/98.8 mmHg); while Cluster 2 includes 100% of men with blood pressure leading to pre-hypertension (130.1/80.7 mmHg). PCA visualizations show fairly clear cluster separation, with Cluster 1 having the most clinically distinct characteristics. The conclusion of this study is that the K-Means-based unsupervised learning approach is effective in identifying latent risk patterns in patient populations, as well as the potential to support clinical risk mapping and preventive health policies. Future recommendations include the integration of this method in EMR systems and the expansion of studies on national datasets.*

Keywords: *Clinical Risk Profile; K-Means Clustering; Patient Segmentation; Unsupervised Learning; Vital Signs*

Abstrak. Transformasi digital dalam sistem kesehatan menuntut pemanfaatan data klinis secara lebih strategis untuk mendukung pengambilan keputusan yang berbasis bukti. Penelitian ini bertujuan untuk mengeksplorasi penerapan algoritma K-Means Clustering dalam segmentasi pasien berdasarkan kombinasi tanda vital (tekanan darah sistolik dan diastolik) serta karakteristik demografis (umur, berat badan, dan jenis kelamin). Data sebanyak 1.401 pasien rawat jalan diperoleh dari sistem rekam medis sebuah rumah sakit di Indonesia, kemudian diproses melalui tahapan preprocessing, standardization, dan dimensionality reduction menggunakan PCA. Hasil elbow method menunjukkan bahwa jumlah klaster optimal adalah 3 (k=3). Analisis deskriptif memperlihatkan bahwa Klaster 0 terdiri dari 100% perempuan dengan tekanan darah normal (124/77 mmHg) dan berat badan rata-rata 55,6 kg; Klaster 1 terdiri dari mayoritas perempuan dengan tekanan darah tinggi (160,8/98,8 mmHg); sedangkan Klaster 2 mencakup 100% laki-laki dengan tekanan darah yang mengarah ke pra-hipertensi (130,1/80,7 mmHg). Visualisasi PCA menunjukkan pemisahan klaster yang cukup jelas, dengan Klaster 1 memiliki karakteristik paling berbeda secara klinis. Kesimpulan dari penelitian ini adalah bahwa pendekatan unsupervised learning berbasis K-Means efektif dalam mengidentifikasi pola risiko laten dalam populasi pasien, serta potensial untuk mendukung pemetaan risiko klinis dan kebijakan kesehatan preventif. Rekomendasi ke depan mencakup integrasi metode ini dalam sistem EMR dan perluasan studi pada dataset nasional.

Kata kunci: K-berarti pengelompokan; Pembelajaran Tanpa Pengawasan; Profil Risiko Klinis; Segmentasi Pasien; Tanda-tanda vital

1. LATAR BELAKANG

Di era transformasi digital dan pelayanan kesehatan berbasis data, kemampuan untuk melakukan segmentasi pasien secara efektif menggunakan data klinis dan demografis menjadi sangat krusial. Segmentasi pasien dapat membantu peningkatan pencegahan penyakit, optimasi pelayanan medis, serta efisiensi penggunaan sumber daya rumah sakit. Melalui penerapan *Electronic Medical Records* (EMR) dan sistem informasi kesehatan digital, rumah sakit kini

Naskah Masuk: 24 September 2025; Revisi: 23 Oktober 2025; Diterima: 28 November 2025;
Terbit: 30 November 2025

memiliki akses terhadap data yang sangat besar dan beragam. Namun, pemanfaatan data tersebut secara analitik—terutama melalui pendekatan *unsupervised learning* seperti clustering—masih sangat terbatas dalam praktik sehari-hari (F. Wang et al., 2021), (Chakraborty et al., 2024).

Penelitian ini memperkenalkan pendekatan inovatif dalam segmentasi pasien melalui algoritma K-Means Clustering dengan menggunakan data pemeriksaan fisik dan informasi demografis dasar pasien rawat jalan. Berbeda dengan penelitian sebelumnya yang banyak fokus pada algoritma prediksi (*supervised learning*) untuk diagnosis, pendekatan ini justru mengeksplorasi struktur alami dalam data tanpa label, yang dapat mengungkap kelompok pasien dengan risiko klinis yang serupa tetapi tidak langsung terdeteksi oleh diagnosa awal (L. Wang et al., 2020). Fokus utama pada vital signs seperti tekanan darah sistolik dan diastolik, serta variabel demografi seperti umur dan jenis kelamin, memberikan basis yang kuat dan dapat dijelaskan secara klinis untuk melakukan segmentasi risiko. Penelitian ini bertujuan untuk menjawab pertanyaan dapatkah algoritma K-Means digunakan untuk membentuk klaster pasien berdasarkan kombinasi tanda vital dan karakteristik demografis guna mendukung pemetaan risiko klinis di layanan kesehatan seperti rumah sakit?".

Dari sisi relevansi kebijakan, penelitian ini sejalan langsung dengan Tujuan Pembangunan Berkelanjutan (*Sustainable Development Goals*/SDGs) 2030, khususnya Tujuan ke-16: Perdamaian, Keadilan dan Kelembagaan yang Tangguh, yang menekankan pentingnya tata kelola data yang kuat dan sistem berbasis bukti dalam pelayanan publik, termasuk bidang kesehatan. Selain itu, penelitian ini mendukung pelaksanaan program prioritas nasional Indonesia seperti Sistem Pemerintahan Berbasis Elektronik (SPBE) dan implementasi Strategi Nasional Keamanan Siber oleh Badan Siber dan Sandi Negara (BSSN). Melalui pendekatan analitik berbasis data ini, rumah sakit dapat mengidentifikasi kelompok pasien rentan, mengurangi rujukan tidak perlu, dan mempercepat deteksi risiko kesehatan sejak dini. Penelitian ini juga mendukung kerangka Strategi Transformasi Digital Kesehatan Indonesia 2021–2024 yang menekankan pentingnya integrasi data dan analitik prediktif di layanan primer (Kementerian Kesehatan Republik Indonesia, 2021), (The 17 Sustainable Development Goals 2030 , 2023). Dengan demikian, kontribusi penelitian ini bersifat ganda yaitu secara ilmiah memberikan pendekatan machine learning yang dapat direplikasi, dan secara kebijakan memperkuat transformasi pelayanan kesehatan nasional berbasis data.

Sebelumnya, telah dijelaskan mengenai latar belakang dan urgensi penelitian beserta kontribusinya dalam bidang klinis dan kebijakan nasional. Pada bagian berikutnya yaitu Metode, akan dijelaskan karakteristik data, tahapan preprocessing, serta metode analisis

clustering yang digunakan. Selanjutnya, pada bagian Hasil dan Pembahasan, akan dipaparkan hasil evaluasi ,visualisasi klaster pasien dan interpretasi medisnya. Kemudian pada bagian Kesimpulan, akan dirangkum temuan utama, kontribusi terhadap sistem kesehatan, serta rekomendasi untuk penelitian dan pengembangan ke depan dalam sistem segmentasi pasien berbasis data.

2. METODE PENELITIAN

Desain Penelitian (Research Design)

Penelitian ini menggunakan pendekatan kuantitatif eksperimental dengan metode *unsupervised learning*, khususnya algoritma K-Means Clustering, untuk melakukan segmentasi pasien berdasarkan data numerik dari catatan pemeriksaan fisik. Tujuan utama dari desain ini adalah untuk menemukan struktur alami atau klaster dalam populasi pasien tanpa menggunakan label atau kelas diagnostik yang telah ditentukan sebelumnya (Jia et al., 2021), (Ikotun et al., 2023), (Molokomme et al., 2021).

Data dan Sumber Data

Data yang digunakan dalam penelitian ini bersifat data sekunder, diperoleh dari sistem rekam medis salah satu rumah sakit di Indonesia. Dataset mencakup 1.403 entri pasien rawat jalan dan terdiri dari 9 atribut. Namun, hanya 5 atribut terpilih yang digunakan dalam proses clustering, yaitu UMUR (dalam tahun), BERAT BADAN (dalam kilogram), TEKANAN DARAH SISTOLIK (mmHg), TEKANAN DARAH DIASTOLIK (mmHg), dan JENIS KELAMIN (0 = Laki-laki, 1 = Perempuan)

Tabel 1. Dataset Pengukuran Fisik Pasien sebelum Preprocessing.

No	Tanggal	Umur	Berat Badan	Jenis Kelamin	Bpjs/Non Bpjs	Tekanan Darah	Diagnosa	Code Icd	Rujuk/Pulang
1		48	64	Laki Laki	BPJS	150/80	Asma	J45	Pulang
2		54	65	Perempuan	BPJS	130/80	pemeriksaan fisik	Z00.00	Pulang
3		52	65	Laki Laki	BPJS	130/80	pemeriksaan fisik	Z00.00	Pulang
4		46	68	Perempuan	BPJS	130/80	Asma	J45	Pulang
5	02-Jan-24	46	46	Laki Laki	BPJS	130/80	Asma	J45	Pulang
6		62	54	Perempuan	BPJS	140/70	Asma	J45	Pulang
7		55	64	Laki Laki	BPJS	120/80	Gangguan Mental	F00	Pulang
8		63	54	Perempuan	BPJS	130/80	penglihatan rendah	H54.2	Rujuk
9		49	41	Perempuan	KIS	120/80	Vertigo	R42	Pulang
10	03-Jan-24	50	77	Perempuan	BPJS	170/100	Hipertensi	I10	Pulang
				Dataset dari Record 10 samapai dengan Record 1403					
1401		70	56	Perempuan	BPJS	130/80	Dm	E08	Pulang
1402	25-Nov-24	49	56	Perempuan	BPJS	119/57	Meriang	J21.0	Pulang
1403		47	80	Perempuan	BPJS	137/78	CHF	I50.0	Pulang

Dataset ini berformat Microsoft Excel (.xlsx) dan sebagian data memiliki nilai kosong yang diatasi pada tahap preprocessing. Penggunaan data *vital signs* dan demografi untuk segmentasi populasi pasien terbukti efektif dalam pengembangan sistem prediktif awal (early risk stratification) dalam studi-studi terkini (Reza et al., 2022), (Xie et al., 2023), (Liu et al., 2022).

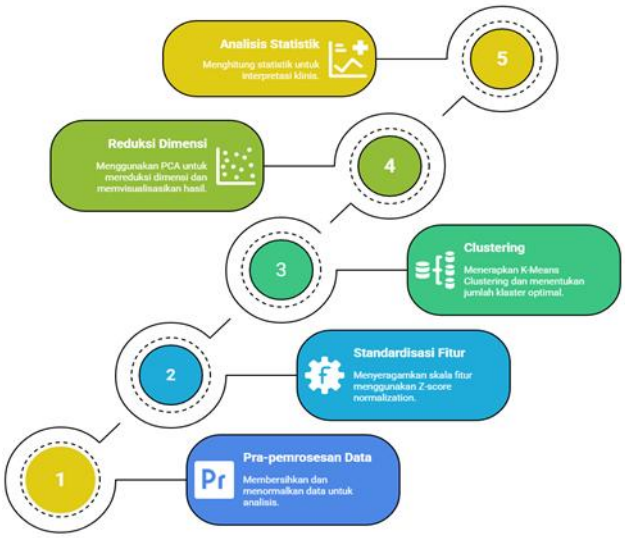
Dataset ini berformat Microsoft Excel (.xlsx) dan sebagian data memiliki nilai kosong yang diatasi pada tahap preprocessing. Penggunaan data *vital signs* dan demografi untuk segmentasi populasi pasien terbukti efektif dalam pengembangan sistem prediktif awal (early risk stratification) dalam studi-studi terkini (Reza et al., 2022), (Xie et al., 2023), (Liu et al., 2022).

Alat dan Perangkat yang Digunakan

Eksperimen dilakukan pada laptop dengan sistem operasi Windows 11, menggunakan perangkat lunak berikut: Python 3.11, Jupyter Notebook sebagai lingkungan kerja. Kemudian juga menggunakan Pustaka Python yaitu *pandas* untuk manipulasi data, *scikit-learn* untuk preprocessing, clustering, PCA, dan *matplotlib* dan seaborn untuk visualisasi (Viveka Kesanapalli & Rao Chintalapudi, 2021).

Tahapan Penelitian / Alur Metodologi

Tahapan metodologi dilakukan secara sistematis dalam beberapa langkah utama sebagai berikut: (1) Pra-pemrosesan Data yaitu menghapus nilai kosong, membersihkan dan normalisasi nama kolom, dan Encoding variabel kategorikal (JENIS KELAMIN). (2) Standardisasi Fitur yaitu menggunakan *Z-score normalization* (*StandardScaler*) agar semua fitur memiliki skala yang seragam. (3) Clustering yaitu tahapan menerapkan K-Means Clustering pada data hasil standardisasi, dan menentukan jumlah klaster optimal menggunakan Metode Elbow. (4) Reduksi Dimensi dan Visualisasi yaitu tahapan menggunakan *Principal Component Analysis* (PCA) untuk mereduksi dimensi menjadi dua komponenm, dan visualisasi hasil clustering dalam ruang dua dimensi. (5) Analisis Statistik Tiap Klaster yaitu tahapan menghitung rata-rata usia, tekanan darah, dan distribusi jenis kelamin pada tiap klaster untuk interpretasi klinis. Pendekatan ini telah digunakan secara luas dalam sistem stratifikasi risiko klinis dan population health management.(Ikotun et al., 2023),(Liu et al., 2022; Mariam et al., 2024; Reza et al., 2022)



Gambar 1. Diagram Tahapan Penelitian.

Tabel 2. Dataset Setelah Preprocessing.

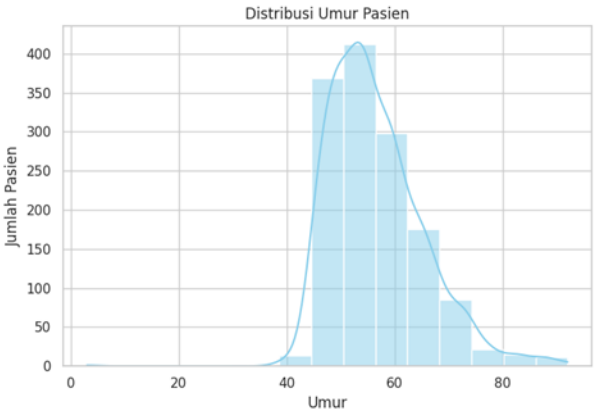
	UMUR	BERAT BADAN	TEKANAN DARAH SISTOLIK	TEKANAN DARAH DIASTOLIK	JENIS KELAMIN
count	1.401	1.401	1.401	1.401	1.401
mean	56,59	58,14	135,32	84,12	0,72
std	8,81	11,43	22,95	14,34	0,45
min	3,00	30,00	60,00	0,00	0,00
25%	50,00	50,00	120,00	79,00	0,00
50%	55,00	57,00	130,00	80,00	1,00
75%	61,00	65,00	150,00	90,00	1,00
max	92,00	102,00	225,00	180,00	1,00

Setelah dilakukan proses pra-pemrosesan dari 1403 data, termasuk pembersihan data kosong dan pemilihan fitur, diperoleh 1.401 data valid yang terdiri dari 5 fitur utama: UMUR, BERAT BADAN, TEKANAN DARAH SISTOLIK, TEKANAN DARAH DIASTOLIK, dan JENIS KELAMIN. Tabel 2. menunjukkan ringkasan statistik deskriptif dari masing-masing atribut yaitu: (1) Umur pasien memiliki rata-rata sebesar 56,59 tahun, dengan rentang usia antara 3 hingga 92 tahun. Nilai minimum yang sangat rendah ini mencerminkan keberadaan pasien anak-anak dalam populasi, meskipun mayoritas berkisar pada usia dewasa, seperti tercermin dari kuartil ke-1 (50 tahun) dan ke-3 (61 tahun). Berat badan rata-rata adalah 58,14 kg dengan simpangan baku sebesar 11,43, menunjukkan variasi yang moderat di antara pasien. Nilai minimum 30 kg dan maksimum 102 kg menunjukkan penyebaran yang cukup luas, tetapi masih dalam rentang fisiologis yang dapat diterima.

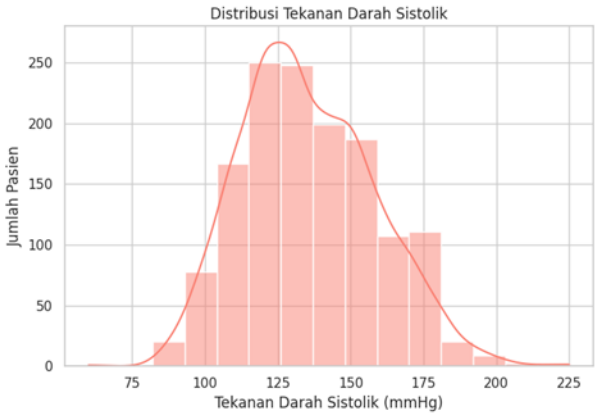
Untuk tekanan darah sistolik, nilai rata-rata tercatat sebesar 135,32 mmHg, dengan simpangan baku tinggi (22,95 mmHg) yang menunjukkan heterogenitas status tekanan darah

pasien. Rentang yang diperoleh berkisar dari 60 mmHg (hipotensi berat) hingga 225 mmHg (hipertensi stadium 3), menunjukkan keberadaan pasien dengan berbagai tingkat risiko kardiovaskular. Tekanan darah diastolik memiliki nilai rata-rata 84,12 mmHg dengan rentang 0 hingga 180 mmHg. Nilai minimum 0 mmHg tampaknya merupakan outlier atau kesalahan pencatatan, yang dapat dipertimbangkan untuk perlakuan khusus pada tahap analisis lanjutan. Mayoritas data diastolik terkonsentrasi antara 79 hingga 90 mmHg, mencerminkan kecenderungan prehipertensi hingga hipertensi ringan.

Jenis kelamin dicatat dalam bentuk biner (0 = laki-laki, 1 = perempuan), dengan distribusi 71,8% pasien adalah perempuan. Ini tercermin dari nilai rata-rata 0,718 dan distribusi kuartil (Q1=0, Q3=1). Skew ini menunjukkan adanya kecenderungan dominasi pasien perempuan dalam data kunjungan. Berdasarkan karakteristik di atas, dataset memiliki keragaman yang cukup dalam dimensi fisiologis dan demografis, yang relevan untuk eksplorasi lebih lanjut menggunakan teknik segmentasi non-supervisi seperti K-Means Clustering.



Gambar 2. (a).



Gambar 3. (b).

Gambar 2 (a) dan 2 (b) memperlihatkan dua histogram yang menggambarkan distribusi umur pasien (panel kiri) dan tekanan darah sistolik (panel kanan) dari populasi yang diamati. Gambar 2 (a) adalah distribusi umur pasien menunjukkan pola yang relatif menyerupai kurva normal yang sedikit miring ke kanan (right-skewed). Mayoritas pasien berada dalam rentang usia 50 hingga 60 tahun, dengan puncak distribusi (modus) berada sekitar 55 tahun. Distribusi ini menunjukkan bahwa populasi pasien didominasi oleh kelompok usia dewasa hingga lanjut usia. Jumlah pasien usia di bawah 40 tahun sangat sedikit, dan jumlahnya menurun secara bertahap setelah usia 65 tahun.

Sementara Gambar 2 (b) adalah Distribusi Tekanan Darah Sistolik yang menunjukkan distribusi tekanan darah sistolik menunjukkan sebaran yang asimetris dengan ekor kanan yang lebih panjang, menandakan kehadiran pasien dengan tekanan darah tinggi ekstrem (hingga >180 mmHg). Nilai tertinggi dari densitas berada di sekitar 120–130 mmHg, yang

mengindikasikan bahwa sebagian besar pasien berada pada kategori prehipertensi hingga hipertensi derajat 1 menurut klasifikasi JNC 7 atau WHO.

Model atau Metode yang Digunakan

Model utama yang digunakan adalah algoritma K-Means Clustering, sebuah metode partisi unsupervised learning yang membagi data ke dalam k klaster berdasarkan jarak Euclidean minimum. Jumlah klaster (k) ditentukan menggunakan Metode Elbow, dengan kisaran evaluasi k = 1 hingga k = 10. Untuk interpretasi hasil, digunakan PCA (*Principal Component Analysis*) agar visualisasi klaster dapat divisualisasikan dalam dua dimensi. PCA juga telah terbukti meningkatkan interpretabilitas hasil clustering dalam berbagai penelitian medis dan diagnosis berbasis data. (Abdullah et al., 2020).

Evaluasi dan Metode Pengukuran (Metrics)

Karena pendekatan ini bersifat unsupervised, tidak digunakan metrik klasifikasi seperti akurasi atau F1-score. Sebagai gantinya, digunakan: Inertia (total within-cluster sum of squares) → metrik utama untuk menentukan jumlah klaster optimal (dalam Metode Elbow), visual inspection dari hasil PCA scatter plot, dan analisis deskriptif tiap klaster untuk menilai keunikan dan interpretabilitas klaster (usia rata-rata, distribusi jenis kelamin, tekanan darah). Inertia dan analisis visualisasi PCA telah divalidasi sebagai metrik evaluasi dalam studi clustering tanpa label (Mariam et al., 2024).

Asumsi, Batasan, atau Kondisi Eksperimen

Beberapa asumsi dan keterbatasan yang perlu dicatat diantaranya adalah (a) dataset bersifat retrospektif, tidak ada intervensi atau label diagnosis yang digunakan, (b) hanya menggunakan dua parameter *vital signs* (sistolik dan diastolik); tidak mencakup nadi, suhu tubuh, atau respirasi, dan (c) tidak dilakukan validasi eksternal atau perbandingan dengan metode clustering lain (seperti: DBSCAN, GMM). Studi dengan keterbatasan parameter fisiologis tetap relevan, terutama dalam konteks *resource-limited settings*.(Junaid et al., 2022), (Khanam et al., 2019).

3. HASIL DAN PEMBAHASAN

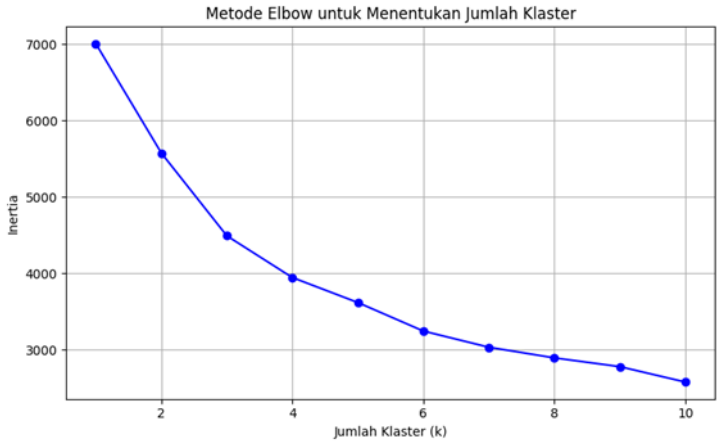
Preprocessing dan Statistik Deskriptif

Setelah proses pembersihan data, dari 1.403 diperoleh 1.401 sampel pasien dengan lima fitur utama yang relevan untuk segmentasi: umur, berat badan, tekanan darah sistolik, tekanan darah diastolik, dan jenis kelamin. Distribusi nilai menunjukkan bahwa sebagian besar pasien merupakan individu dewasa dengan rata-rata usia 56,6 tahun dan rata-rata berat badan 58,14 kg. Nilai tekanan darah menunjukkan bahwa sebagian besar pasien berada pada kisaran pra-

hipertensi hingga hipertensi ringan, dengan rata-rata tekanan sistolik 135,3 mmHg dan diastolik 84,1 mmHg. Distribusi jenis kelamin menunjukkan bahwa sekitar 72% pasien adalah perempuan, yang berimplikasi pada kecenderungan demografis tertentu dalam populasi data ini. Sebelum penentuan jumlah klaster, maka dataset yang telah bersihkan dilakukan Scaling (*Standardization*) dengan menggunakan *Z-score scaling* pada python agar fitur berada pada skala yang sama.

Penentuan Jumlah Klaster Optimal (Elbow Method)

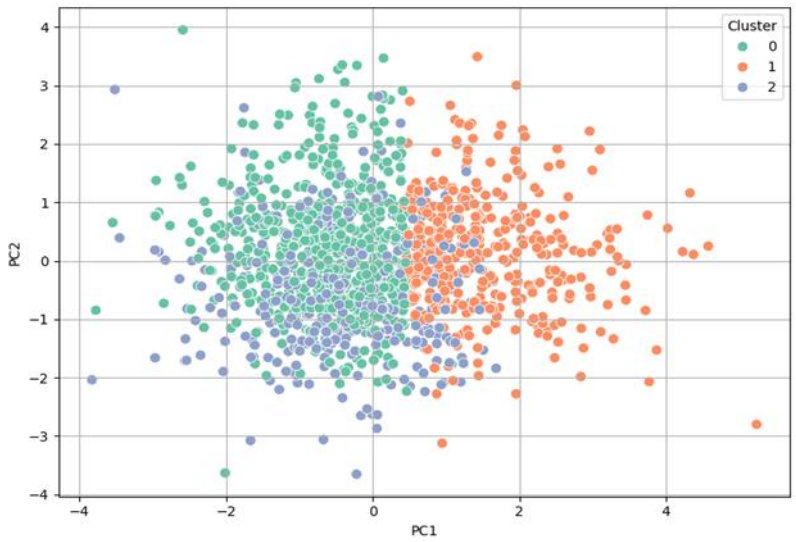
Penentuan jumlah klaster dilakukan menggunakan metode Elbow, yang mengukur nilai inertia pada berbagai nilai k. Grafik yang dihasilkan seperti pada gambar 3. menunjukkan penurunan tajam dari k = 1 ke k = 3, setelah itu penurunan melambat (melandai). Titik tekuk yang paling jelas terlihat pada k = 3, yang menunjukkan bahwa tiga klaster sudah cukup merepresentasikan variasi alami dalam data. Pemilihan k = 3 juga memberikan keseimbangan antara kompleksitas model dan kemampuan interpretasi klinis. Jumlah data dalam tiap klaster terdiri dari klaster 0: 666 data, klaster 1: 371 data, dan klaster 2: 364 data.



Gambar 4. Jumlah Klaster Optimal.

Hasil Visualisasi Klaster (PCA)

Untuk memvisualisasikan hasil segmentasi secara dua dimensi, digunakan *Principal Component Analysis* (PCA). Proyeksi data ke dalam dua komponen utama (PC1 dan PC2) pada gambar 4 menunjukkan bahwa pasien dapat dikelompokkan secara visual ke dalam tiga klaster yang berbeda. Grafik PCA memperlihatkan bahwa (a) Klaster 1 (oranye) terpisah jelas dari dua klaster lainnya, mengindikasikan kelompok dengan karakteristik unik. ((b) Klaster 0 (hijau) dan klaster 2 (biru) menunjukkan tumpang tindih parsial, yang dapat mengindikasikan kemiripan fisiologis sebagian pasien antar klaster. Hasil ini menguatkan bahwa metode K-Means berhasil menemukan struktur laten dalam data pemeriksaan fisik.



Gambar 5. Visualisasi Hasil Clustering dengan PCA.

Karakteristik Statistik Per Klaster

Visualisasi PCA dilakukan untuk memudahkan identifikasi struktur data dan validasi hasil clustering dan menunjukkan bahwa metode K-Means berhasil memisahkan kelompok pasien dengan karakteristik yang cukup berbeda secara statistik. Berikut pada Tabel 3. ditampikan ringkasan profil dari masing-masing klaster:

Tabel 3. Ringkasan Statistik per Klaster.

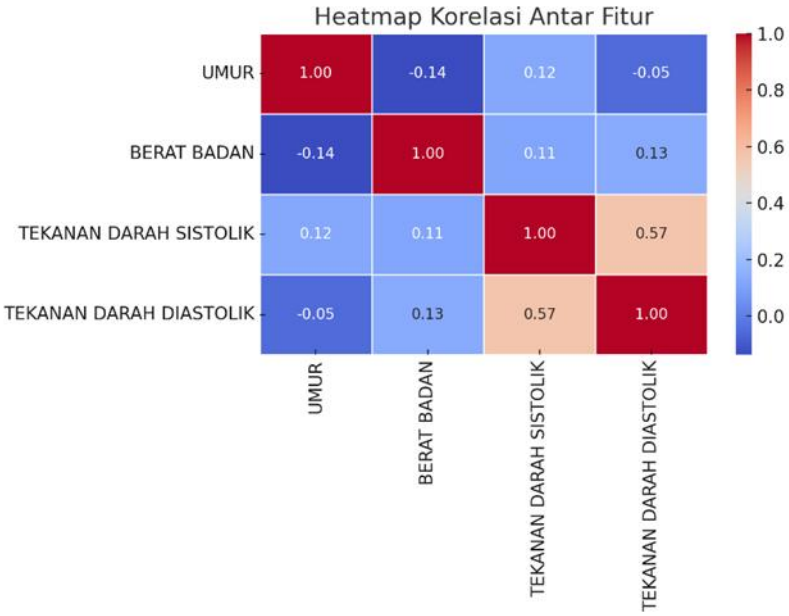
Klaster	Usia Rata-rata	Berat Badan	Sistolik	Diastolik	Jenis Kelamin Dominan
0	56.4 tahun	55.6 kg	124 mmHg	77.8 mmHg	100% Perempuan
1	56.8 tahun	61.9 kg	160.8 mmHg	98.8 mmHg	Mayoritas Perempuan
2	56.6 tahun	59.0 kg	130.1 mmHg	80.7 mmHg	100% Laki-laki

Dari Tabel 3. dapat dipaparkan tentang Interpretasi Klinis dimana Klaster 0 – "Kelompok Perempuan Sehat" dimana rata-rata usia sekitar 56 tahun, tekanan darah normal (124/77 mmHg), seluruh anggota klaster ini adalah perempuan, berat badan cenderung moderat (rata-rata 55.6 kg). Data ini memperlihatkan tekanan darah normal dan seluruh anggota adalah perempuan, Ini menunjukkan kelompok baseline sehat.

Klaster 1 – "Kelompok Perempuan dengan Hipertensi Berat" dimana tekanan darah sangat tinggi yaitu 160.8 / 98.8 mmHg, mayoritas perempuan, usia hampir sama dengan klaster lain, dan berat badan lebih tinggi dibanding klaster. Data ini memperlihatkan rata-rata tekanan darah signifikan lebih tinggi. Klaster ini merupakan kelompok risiko tinggi kardiovaskular.

Sedangkan Klaster 2 – "Kelompok Laki-laki Tekanan Darah Moderat" dimana seluruhnya adalah laki-laki, tekanan darah agak tinggi (130/80 mmHg) → pra-hipertensi, dan

berat badan dan usia mirip klaster lain memperlihatkan semua jenis kelamin laki-laki dengan tekanan darah yang mengarah ke pra-hipertensi. Ini sangat cocok sebagai target edukasi dan intervensi dini.



Gambar 5. Heatmap Korelasi Antar Fitur.

Gambar 5 menunjukkan Atribut TEKANAN DARAH SISTOLIK vs DIASTOLIK memperlihatkan Korelasi sangat kuat dan positif (~0.88), menunjukkan bahwa jika sistolik meningkat, diastolik juga cenderung meningkat. Ini konsisten dengan fisiologi tekanan darah. Sementara BERAT BADAN vs SISTOLIK: Korelasi moderat (~0.49), menunjukkan bahwa berat badan berkontribusi terhadap tekanan darah tinggi. UMUR vs tekanan darah: Korelasi lebih lemah (~0.30–0.35), namun tetap menunjukkan tren bahwa bertambahnya usia cenderung berkaitan dengan peningkatan tekanan darah.

Implikasi dan Pembahasan

Temuan segmentasi ini memiliki implikasi penting bagi kebijakan klinis dan pengambilan keputusan berbasis data. Klaster yang dihasilkan memungkinkan penyedia layanan kesehatan untuk menetapkan strategi intervensi yang lebih personalisasi, seperti dengan menetapkan prioritas penanganan untuk pasien dalam Klaster 1. Selain itu, pendekatan *unsupervised* ini terbukti efektif meskipun tanpa label diagnosis, dan dapat dijalankan secara periodik untuk mendeteksi dinamika risiko kesehatan populasi rumah sakit. Pendekatan seperti ini juga sejalan dengan rekomendasi WHO untuk menggunakan analitik prediktif dalam penguatan sistem kesehatan berbasis data (WHO, 2021), serta mendukung transformasi digital layanan kesehatan nasional.

4. KESIMPULAN DAN SARAN

Penelitian ini telah berhasil menerapkan pendekatan pembelajaran *unsupervised learning* melalui algoritma K-Means Clustering untuk melakukan segmentasi pasien berdasarkan kombinasi tanda vital—seperti tekanan darah sistolik dan diastolik—serta karakteristik demografis berupa umur, berat badan, dan jenis kelamin. Hasil analisis menunjukkan bahwa struktur klaster dengan jumlah tiga (k = 3) mampu mengelompokkan pasien ke dalam kelompok yang memiliki karakteristik klinis yang berbeda secara signifikan. Klaster pertama merepresentasikan pasien perempuan dengan tekanan darah relatif normal dan berat badan sedang yang dapat dikategorikan sebagai kelompok baseline sehat. Klaster kedua mencakup pasien perempuan dengan tekanan darah tinggi dan berat badan di atas rata-rata, sehingga masuk dalam kategori risiko tinggi hipertensi. Sementara klaster ketiga terdiri dari pasien laki-laki dengan tekanan darah yang mendekati pra-hipertensi dan proporsi berat badan moderat.

Temuan ini menunjukkan bahwa metode segmentasi berbasis K-Means dapat mengungkap struktur laten dalam data kesehatan yang tidak terlihat secara eksplisit melalui diagnosis awal. Segmentasi ini berpotensi digunakan untuk mendukung sistem pemetaan risiko klinis (clinical risk profiling), strategi triase berbasis data, dan penyesuaian intervensi preventif bagi kelompok populasi yang berisiko. Selain kontribusi ilmiah, pendekatan ini juga relevan dalam konteks kebijakan nasional, yakni mendukung transformasi digital sistem kesehatan nasional sebagaimana ditetapkan dalam kebijakan Sistem Pemerintahan Berbasis Elektronik (SPBE), serta sejalan dengan Strategi Nasional Keamanan Siber (BSSN). Secara lebih luas, riset ini turut berkontribusi pada pencapaian Sustainable Development Goals (SDGs) 2030, khususnya tujuan ke-16 yang menekankan pentingnya institusi yang kuat, transparan, dan berbasis data dalam sektor kesehatan.

Untuk penelitian mendatang, terdapat beberapa peluang pengembangan lanjutan yang disarankan. Pertama, penelitian sebaiknya mempertimbangkan penambahan fitur *vital signs* lain seperti denyut jantung, suhu tubuh, serta riwayat penyakit kronis untuk menghasilkan klaster yang lebih informatif secara klinis. Kedua, efektivitas K-Means dapat dibandingkan dengan algoritma clustering lainnya seperti DBSCAN, Gaussian Mixture Model (GMM), atau Hierarchical Clustering untuk melihat keunggulan relatif dari masing-masing metode. Ketiga, hasil segmentasi yang telah diperoleh juga dapat diintegrasikan dalam sistem rekam medis elektronik (EMR) atau dashboard rumah sakit sebagai alat bantu dalam pengambilan keputusan. Terakhir, untuk memastikan generalisasi hasil, pendekatan ini sebaiknya diuji pada dataset berskala nasional guna memperkuat bukti empiris dan mendukung kebijakan kesehatan berbasis data di Indonesia.

DAFTAR REFERENSI

Abdullah, S. S., Rostamzadeh, N., Sedig, K., Garg, A. X., & McArthur, E. (2020). Visual analytics for dimension reduction and cluster analysis of high dimensional electronic health records. *Informatics, 7*(2). https://doi.org/10.3390/informatics7020017

Chakraborty, C., Bhattacharya, M., Pal, S., & Lee, S. S. (2024). From machine learning to deep learning: Advances of the recent data-driven paradigm shift in medicine and healthcare. *Current Research in Biotechnology, 7*. Elsevier B.V. https://doi.org/10.1016/j.crbiot.2023.100164

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences, 622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Jia, L., Gaüzère, B., & Honeine, P. (2021). graphkit-learn: A Python library for graph kernels based on linear patterns. *Pattern Recognition Letters, 143*, 113–121. https://doi.org/10.1016/j.patrec.2021.01.003

Junaid, S. B., Imam, A. A., Shuaibu, A. N., Basri, S., Kumar, G., Surakat, Y. A., Balogun, A. O., Abdulkarim, M., Garba, A., Sahalu, Y., Mohammed, A., Mohammed, Y. T., Abdulkadir, B. A., Abba, A. A., Kakumi, N. A. I., & Alazzawi, A. K. (2022). Artificial intelligence, sensors and vital health signs: A review. *Applied Sciences, 12*(22). https://doi.org/10.3390/app122211475

Kementerian Kesehatan Republik Indonesia. (2021). *Cetak Biru Strategi Transformasi Digital Kesehatan 2024* (1st ed.). Kementerian Kesehatan RI.

Khanam, F. T. Z., Al-Naji, A., & Chahl, J. (2019). Remote monitoring of vital signs in diverse non-clinical and clinical scenarios using computer vision systems: A review. *Applied Sciences, 9*(20). https://doi.org/10.3390/app9204474

Liu, Z., Hu, Y., Mertes, G., Yang, Y., & Clifton, D. A. (2022). Patient clustering and classification for vital organ failure using ICD code with graph attention. *bioRxiv*. https://doi.org/10.1101/2022.11.07.515209

Mariam, A., Javidi, H., Zabor, E. C., Zhao, R., Radivoyevitch, T., & Rotroff, D. M. (2024). Unsupervised clustering of longitudinal clinical measurements in electronic health records. *PLOS Digital Health, 3*(10). https://doi.org/10.1371/journal.pdig.0000628

Molokomme, D. N., Chabalala, C. S., & Bokoro, P. N. (2021). Enhancement of advanced metering infrastructure performance using unsupervised k-means clustering algorithm. *Energies, 14*(9). https://doi.org/10.3390/en14092732

Reza, N., Yang, Y., Bone, W. P., Singhal, P., Verma, A., Denduluri, S., Adusumalli, S., Ritchie, M. D., & Cappola, T. P. (2022). Unsupervised clustering applied to electronic health record-derived phenotypes in patients with heart failure. *medRxiv*. https://doi.org/10.1101/2022.10.31.22281772

United Nations Department of Global Communications. (2023). *The 17 Sustainable Development Goals 2030*.

Viveka Kesanapalli, L., & Rao Chintalapudi, S. (2021). A survey on machine learning applications in healthcare. *Advances and Applications in Mathematical Sciences, 20*(11).

Wang, F., Jiao, L., & Pan, Q. (2021). A survey on unsupervised transfer clustering. In *2021 40th Chinese Control Conference (CCC)* (pp. 7361–7365). https://doi.org/10.23919/CCC52363.2021.9549617

Wang, L., Tong, L., Davis, D., Arnold, T., & Esposito, T. (2020). The application of unsupervised deep learning in predictive models using electronic health records. *BMC Medical Research Methodology, 20*(1). https://doi.org/10.1186/s12874-020-00923-1

Xie, L., Gou, B., Bai, S., Yang, D., Zhang, Z., Di, X., Su, C., Wang, X., Wang, K., & Zhang, J. (2023). Unsupervised cluster analysis reveals distinct subgroups in healthy population with different exercise responses of cardiorespiratory fitness. *Journal of Exercise Science and Fitness, 21*(1), 147–156. https://doi.org/10.1016/j.jesf.2022.12.005