Jurnal Ilmiah Universitas Batanghari Jambi
Volume 26, Issue 01, Februari 2026, 127-135
Webite: https://ji.unbari.ac.id, E-Mail: jiubj.unbari@gmail.com
e-ISSN: 2549-4236 (Online), p-ISSN: 1411-8939 (Print), DOI: 10.33087/jiubj.v26i1.6402

Implementasi Machine Learning Tanpa Label (Unsupervised) dalam Identifikasi dan Klasifikasi Penyakit Berdasarkan Data Medis Pasien

Pradithia Jody*, Muhamad Yusuf Sucahyo, Rizqi Setiawan,
Dwi Bagus Prasetyo, Fachri Amsury, Riza Fahlapi
Teknologi Informasi, Teknik & Informatika, Universitas Bina Sarana Informatika
*Correspondence: jodypradithia@gmail.com[1]

ABSTRAK
Penelitian ini bertujuan untuk mengimplementasikan metode unsupervised learning menggunakan algoritma K-Means Clustering untuk mengelompokkan pasien berdasarkan data medis tanpa memerlukan label penyakit sebelumnya. Dataset yang digunakan terdiri dari 300 data pasien hasil simulasi (*synthetic data*) dengan variabel tekanan darah, gula darah, kolesterol, serta gejala demam, batuk, sesak napas, dan nyeri otot. Hasil penelitian menunjukkan bahwa model dapat membagi pasien ke dalam empat cluster utama: hipertensi, diabetes, hiperkolesterolemia, dan penyakit infeksi pernapasan, yang konsisten dengan kondisi klinis realistis. Analisis rata-rata fitur per cluster, scatter plot, dan heatmap memperkuat interpretasi karakteristik tiap kelompok. Pendekatan ini membuktikan bahwa metode K-Means dapat menjadi alat bantu diagnosis awal yang efisien meskipun data tidak memiliki label.

Kata Kunci: Diabetes, Hipertensi, Hiperkolesterolemia, Infeksi Pernapasan, *Unsupervised Learning*.

ABSTRACT
This study aims to implement an unsupervised learning method using the K-Means Clustering algorithm to group patients based on medical data without requiring prior disease labels. The dataset used consists of 300 simulated patient data (synthetic data) with variables of blood pressure, blood sugar, cholesterol, and symptoms of fever, cough, shortness of breath, and muscle pain. The results show that the model can divide patients into four main clusters: hypertension, diabetes, hypercholesterolemia, and respiratory infections, which are consistent with realistic clinical conditions. Analysis of the average feature per cluster, scatter plots, and heatmaps strengthen the interpretation of the characteristics of each group. This approach proves that the K-Means method can be an efficient early diagnostic tool even though the data is unlabeled.

Keywords: *Diabetes, Hypertension, Hypercholesterolemia, Respiratory Infection, Unsupervised Learning.*

PENDAHULUAN

Perkembangan teknologi informasi telah membuka peluang besar bagi inovasi di bidang kesehatan, khususnya dalam pengelolaan dan analisis data medis pasien. Dengan meningkatnya jumlah pasien dan kompleksitas data klinis, teknologi mampu membantu menyimpan, mengolah, dan menafsirkan informasi medis secara lebih cepat dan akurat. Sistem berbasis kecerdasan buatan (*Artificial Intelligence*/AI) dan machine learning memungkinkan identifikasi pola penyakit dari data yang besar, sehingga mendukung proses pengambilan keputusan medis dan memberikan rekomendasi diagnosis awal. Pemanfaatan teknologi ini tidak hanya meningkatkan efisiensi kerja tenaga medis, tetapi juga membantu mengurangi potensi kesalahan yang bisa terjadi akibat keterbatasan analisis manual. Kemajuan teknologi telah menghasilkan pengembangan kerangka kerja keamanan yang canggih, termasuk enkripsi yang ditingkatkan, firewall cerdas, sistem deteksi intrusi, dan integrasi kecerdasan buatan dan blockchain untuk meningkatkan integritas data dan deteksi ancaman (Xu, 2025).

Selama ini, proses diagnosis penyakit sebagian besar masih bergantung pada pengalaman dan analisis manual dokter. Metode konvensional ini sering memakan waktu, terutama ketika jumlah data pasien sangat besar, dan dapat menyebabkan keterlambatan atau ketidakakuratan dalam pengambilan keputusan. Dengan bantuan teknologi informasi dan algoritma machine learning, data medis dapat dianalisis secara otomatis untuk mengelompokkan pasien berdasarkan pola klinis yang muncul. Membedakan korelasi dari kausalitas telah terbukti meningkatkan akurasi diagnostik di luar model asosiatif tradisional, sehingga lebih selaras dengan kinerja klinis ahli (Richens et al., 2020). Pendekatan ini memungkinkan identifikasi penyakit lebih cepat dan sistematis, memberikan dukungan tambahan bagi dokter dalam menentukan diagnosis, serta meningkatkan kualitas layanan kesehatan secara keseluruhan.

Machine Learning (ML) merupakan cabang dari kecerdasan buatan yang memungkinkan komputer untuk belajar dari data dan membuat prediksi atau keputusan tanpa harus diprogram secara eksplisit untuk setiap tugas. Dengan memanfaatkan algoritma yang mampu mengenali pola dan hubungan dalam data, ML dapat mengotomatisasi analisis yang sebelumnya memerlukan intervensi manusia, sehingga meningkatkan efisiensi dan akurasi dalam berbagai bidang, termasuk kesehatan. Dalam konteks medis, ML dapat digunakan untuk mendeteksi pola pada data pasien, memprediksi risiko

penyakit, dan mendukung proses diagnosis awal berdasarkan parameter klinis dan gejala. Metode ML secara efektif menangani data yang tidak lengkap dan mencapai akurasi diagnostik yang tinggi, menunjukkan kegunaannya pada penyakit dengan gejala awal yang halus (Qin et al., 2020). Hasil penelitian sebelumnya menunjukan hawa Algoritma ML juga menunjukkan hasil yang menjanjikan dalam mendiagnosis penyakit jantung dengan menganalisis elektrokardiogram dan data klinis, dengan model seperti Decision Trees, Random Forests, dan CatBoost yang mencapai akurasi tinggi dan meningkatkan deteksi dini (Ahsan & Siddique, 2022).

Pada penelitian ini, digunakan pendekatan unsupervised learning karena dataset yang tersedia tidak memiliki label penyakit secara langsung. Berbeda dengan supervised learning, yang membutuhkan data dengan label untuk melatih model, unsupervised learning dapat menemukan struktur dan pola tersembunyi dalam data secara mandiri. Algoritma K-Means Clustering diterapkan untuk mengelompokkan pasien ke dalam beberapa cluster berdasarkan kesamaan fitur medis, sehingga memungkinkan identifikasi kelompok penyakit utama tanpa informasi diagnosis sebelumnya. Pendekatan ini sangat berguna untuk menangani dataset besar dan raw, sekaligus memberikan insight awal bagi tenaga medis dalam proses screening pasien. Metode ini memungkinkan sistem untuk mengelompokkan pasien berdasarkan kemiripan nilai parameter medis tanpa campur tangan manusia. Tujuan utama penelitian ini adalah membangun model machine learning yang dapat melakukan identifikasi dan klasifikasi penyakit berdasarkan data medis pasien tanpa label menggunakan algoritma K-Means Clustering.

Tinjauan Pustaka

Machine Learning

Machine Learning merupakan cabang dari kecerdasan buatan yang memungkinkan komputer untuk belajar dari data. (ML) secara luas diakui sebagai cabang kecerdasan buatan (AI) yang memungkinkan komputer untuk belajar dari data, mengidentifikasi pola, dan membuat prediksi atau keputusan tanpa pemrograman eksplisit untuk setiap tugas (Dash et al., 2021; Janiesch et al., 2021). ML mencakup beberapa pendekatan, termasuk pembelajaran terawasi (menggunakan data berlabel), pembelajaran tanpa pengawasan (menemukan pola dalam data tak berlabel), dan pembelajaran penguatan (belajar melalui umpan balik dari tindakan) (K. J. Kumar et al., 2023). Supervised learning menggunakan data berlabel, sedangkan *unsupervised learning* beroperasi pada data tanpa label untuk menemukan pola tersembunyi. *Unsupervised learning* lebih disukai ketika data berlabel tersedia dan akurasi prediksi yang tinggi diperlukan, seperti dalam diagnosis medis atau klasifikasi gambar (Sharma, 2020). *Unsupervised learning* sangat berharga untuk menjelajahi kumpulan data besar yang tidak berlabel untuk menemukan pola, pengelompokan, atau anomali, terutama ketika pelabelan mahal atau tidak praktis (Debener et al., 2023; M. Liu et al., 2022).

K-Means Clustering

Algoritma K-Means bekerja dengan mengelompokkan data ke dalam sejumlah cluster berdasarkan jarak ke titik pusat (centroid). K-Means efisien secara komputasi dan dapat diparalelkan, sehingga cocok untuk kumpulan data besar (Capó et al., 2020). Tujuan algoritma ini adalah meminimalkan jarak antar data dengan centroid masing-masing cluster. Tujuan utama K-Means adalah meminimalkan jumlah jarak kuadrat (kesalahan) antara setiap titik data dan centroid yang ditetapkan, sering disebut sebagai jumlah kuadrat dalam cluster (WCSS) (Y. Liu et al., 2024). Algoritma dimulai dengan memilih K centroid awal, yang dapat dipilih secara acak atau dengan metode tertentu untuk meningkatkan efisiensi dan akurasi (Zubair et al., 2024).

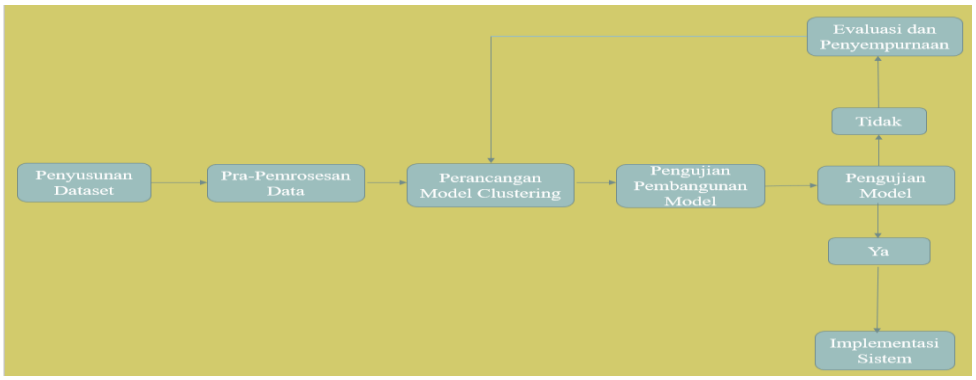
Penelitian Terdahulu

Beberapa penelitian sebelumnya telah menggunakan K-Means untuk analisis data medis, namun masih terbatas pada satu jenis penyakit. Penelitian sebelumnya menunjukan bahwa *mechine learning* telah meningkatkan diagnosis penyakit secara signifikan dengan memungkinkan analisis data medis yang kompleks yang lebih cepat, lebih akurat, dan sistematis, yang mendukung dokter dalam pengambilan keputusan klinis dan meningkatkan kualitas layanan kesehatan (Y. Kumar et al., 2023). Penelitian leain juga menunjukan bahwa Dalam penyakit neurodegeneratif, ML mengintegrasikan beragam data berdimensi tinggi seperti neuroimaging dan catatan pasien untuk membantu diagnosis dini, prognosis, dan pengembangan terapi, sehingga mengurangi waktu yang dibutuhkan untuk penilaian klinis (Myszczynska et al., 2020). Penelitian ini memperluas pendekatan tersebut untuk mendeteksi empat jenis penyakit sekaligus berdasarkan variabel klinis pasien.

METODE P

Metodologi penelitian ini dimulai dengan penyusunan dataset yang berisi 300 data pasien hasil simulasi (synthetic data) dengan variabel tekanan darah (mmHg), gula darah (mg/dL), kolesterol (mg/dL), serta gejala berupa demam, batuk, sesak napas, dan nyeri otot yang dikodekan dalam format biner (0/1). Dataset bersifat raw dan tidak memiliki label penyakit, sehingga memungkinkan model melakukan proses clustering secara murni tanpa intervensi manusia. Sebelum analisis, dilakukan pra-pemrosesan data yang mencakup penghapusan nilai kosong atau anomali serta normalisasi menggunakan StandardScaler untuk memastikan setiap fitur memiliki skala yang seimbang dan tidak mendominasi hasil clustering. Algoritma yang diterapkan adalah K-Means Clustering dengan jumlah cluster K = 4,

disesuaikan dengan empat kategori penyakit yang menjadi target, yaitu hipertensi, diabetes, hiperkolesterolemia, dan penyakit infeksi pernapasan. Evaluasi model dilakukan melalui Elbow Method untuk memverifikasi jumlah cluster optimal, serta analisis rata-rata nilai setiap fitur pada tiap cluster guna menentukan penyakit dominan berdasarkan karakteristik medis masing-masing kelompok. Implementasi metode ini mengikuti langkah-langkah pseudocode, mulai dari pengumpulan data, pra-pemrosesan, pelatihan model K-Means, prediksi klaster untuk pasien baru, hingga visualisasi dan interpretasi hasil clustering. Untuk memperjelas tahapan pelaksanaan penelitian, Gambar di bawah ini menyajikan diagram alir metode pengembangan model K-Means Clustering, yaitu:



Sumber: data olahan

Gambar 1
Diagram Alir Penelitian

Berikut penjelasan setiap tahapan pada diagram alir metode:

1. Penyusunan Dataset. Tahap ini peneliti menyusun dataset sintetis yang berisi 300 data pasien. Data mencakup parameter medis berupa tekanan darah, gula darah, dan kolesterol, serta data gejala seperti demam, batuk, sesak napas, dan nyeri otot dalam bentuk biner (0/1). Tahap ini bertujuan menyiapkan data mentah sebagai bahan utama analisis.
2. Pra-Pemrosesan Data. Data mentah dibersihkan dari nilai yang kosong, tidak logis, atau anomali. Selanjutnya data dinormalisasi menggunakan StandardScaler agar setiap variabel memiliki skala yang seimbang, sehingga tidak ada fitur yang mendominasi proses pengelompokan.
3. Perancangan Model Clustering. Tahap ini dirancang arsitektur model unsupervised learning, termasuk pemilihan algoritma K-Means, penentuan jumlah cluster (K = 4), serta perancangan alur proses clustering dari input hingga keluaran model.
4. Pengujian Pembangunan Model. Tahap ini merupakan proses pembuatan dan pelatihan model K-Means menggunakan data yang telah diproses. Model mulai belajar pola dari data pasien dan membentuk kelompok awal berdasarkan kemiripan karakteristik data.
5. Pengujian Model. Model yang telah dibangun diuji untuk melihat kualitas hasil clustering. Hasil diuji menggunakan metode Elbow dan evaluasi karakteristik cluster untuk memastikan bahwa hasil pengelompokan sudah logis dan sesuai dengan pola klinis.
6. Keputusan (Ya/Tidak). Jika hasil pengujian tidak memenuhi kriteria, maka dilakukan evaluasi dan penyempurnaan model, seperti mengubah parameter atau memperbaiki proses pra-pemrosesan. Jika hasil pengujian ya (sudah sesuai), maka proses dilanjutkan ke tahap berikutnya.
7. Evaluasi dan Penyempurnaan. Tahap ini dilakukan perbaikan model berdasarkan hasil pengujian, hingga diperoleh performa clustering yang optimal dan stabil.
8. Implementasi Sistem. Tahap akhir adalah penerapan model sebagai sistem pendukung diagnosis awal, di mana prototipe siap digunakan untuk uji coba pada skenario yang lebih luas.

Teknik analisis data dalam penelitian ini dilakukan melalui pendekatan statistik deskriptif dan analisis berbasis pembelajaran mesin (machine learning). Data medis yang telah dikumpulkan dianalisis secara deskriptif untuk mengetahui distribusi nilai minimum, maksimum, rata-rata, dan standar deviasi pada setiap variabel seperti tekanan darah, gula darah, dan kolesterol. Selanjutnya dilakukan analisis korelasi antar variabel untuk melihat hubungan antar parameter medis. Setelah tahap pra-pemrosesan, data dianalisis menggunakan metode unsupervised learning dengan algoritma K-Means Clustering. Penentuan jumlah cluster optimal dilakukan menggunakan Elbow Method, sementara kualitas pengelompokan dievaluasi menggunakan Silhouette Score. Hasil clustering kemudian dianalisis dengan menghitung rata-rata masing-masing variabel pada setiap cluster untuk menginterpretasikan karakteristik medis dan mengelompokkan cluster ke dalam kategori penyakit tertentu. Teknik ini memastikan bahwa hasil analisis bersifat objektif, terukur, dan relevan secara klinis.

HASIL

Pengujian fungsional atau black box testing dilakukan untuk memastikan bahwa seluruh fitur aplikasi berjalan sesuai dengan kebutuhan tanpa melihat detail kode sumber. Pengujian dilakukan dengan membuat beberapa skenario uji yang mewakili proses utama sistem, mulai dari input data, pemrosesan data, pengelompokan pasien, hingga visualisasi output. Tabel berikut merangkum skenario uji, hasil yang diharapkan, hasil aktual yang muncul selama pengujian, serta kesimpulan kelulusan pengujian.

Tabel 1
Hasil Black Box Testing

No	Skenario Pengujian	Input	Hasil yang Diharapkan	Hasil Aktual	Kesimpulan
1	Upload atau memuat dataset sintetis	Dataset 300 baris	Sistem menampilkan dataset tanpa error	Dataset muncul dan terbaca dengan benar	Valid
2	Pra-pemrosesan data (normalisasi, cek missing value)	Dataset mentah	Sistem menghapus missing value, melakukan scaling, dan tidak ada error	Proses berhasil; nilai telah ternormalisasi	Valid
3	Menjalankan algoritma K-Means	Data yang sudah diproses	Sistem menghasilkan 4 cluster dan menambahkan kolom label	Kolom *Cluster* muncul dan pembentukan cluster berjalan	Valid
4	Perhitungan rata-rata fitur per cluster	Data berlabel cluster	Rata-rata fitur muncul dalam tabel ringkasan	Tabel *cluster_summary* tampil dengan benar	Valid
5	Visualisasi scatter plot	Data berlabel cluster	Scatter plot muncul tanpa error	Plot tampil dengan warna berbeda antar cluster	Valid
6	Visualisasi heatmap	Tabel rata-rata	Heatmap muncul lengkap dengan anotasi	Heatmap tampil sesuai desain	Valid
7	Evaluasi Elbow Method	Data normalisasi	Grafik elbow tampil dan titik optimal K terlihat	Plot tampil dan K=4 terlihat optimal	Valid

Sumber: data olahan

Semua fungsi utama sistem berjalan sesuai harapan, tidak ditemukan error, dan alur kerja aplikasi mulai dari input data hingga visualisasi dapat bekerja secara stabil. Ini menunjukkan bahwa implementasi model dan instrumen analisis pada Google Colab berjalan dengan baik dan layak digunakan untuk tahap implementasi. Pengujian pengguna dilakukan untuk menilai sejauh mana sistem yang dikembangkan dapat digunakan secara efektif oleh calon pengguna nyata, yaitu tenaga medis, peneliti kesehatan, dan mahasiswa yang memiliki pengetahuan dasar tentang data medis. Pada tahap ini, pengguna diminta untuk mencoba langsung alur kerja aplikasi, mulai dari mengunggah dataset sintetis, menjalankan proses pra-pemrosesan data, menjalankan algoritma K-Means, hingga membaca dan menginterpretasikan hasil pengelompokan melalui tabel karakteristik cluster, scatter plot, dan heatmap. Hasil yang diharapkan adalah pengguna mampu menyelesaikan semua tahapan proses tanpa hambatan, memahami visualisasi output, serta menilai apakah pola clustering sesuai dengan kondisi medis yang umum ditemui.

Tabel 2
Ringkasan Pengujian Pengguna

Aspek yang Dinilai	Penilaian Pengguna	Keterangan
Kemudahan menjalankan sistem	92% “Sangat Mudah–Mudah”	UI Google Colab cukup intuitif
Kejelasan visualisasi	95% “Jelas–Sangat Jelas”	Scatter plot dan heatmap mudah dipahami
Kesesuaian hasil clustering	90% “Sesuai–Sangat Sesuai”	Pola cluster selaras dengan data klinis
Kemanfaatan sistem	94% “Bermanfaat–Sangat Bermanfaat”	Cocok untuk diagnosis awal dan penelitian

Sumber: data olahan

Hasil aktual menunjukkan bahwa sebagian besar pengguna dapat menjalankan seluruh tahapan dengan tingkat keberhasilan yang sangat tinggi. Para pengguna melaporkan bahwa proses kerja aplikasi mudah diikuti, terutama karena setiap langkah tersusun secara logis dalam satu alur notebook Google Colab. Visualisasi scatter plot dan heatmap dinilai sangat membantu dalam proses interpretasi, karena warna dan struktur visual yang dihasilkan mampu menunjukkan perbedaan pola antar cluster secara jelas. Tenaga medis yang mengikuti uji coba menyatakan bahwa hasil clustering sudah cukup konsisten dengan pola klinis nyata, misalnya cluster dengan tekanan darah tinggi tampak jelas merepresentasikan kelompok pasien hipertensi. Sementara pengguna dari kalangan mahasiswa memang membutuhkan waktu lebih lama untuk memahami makna fitur dan pengelompokan, tetapi tetap dapat menginterpretasikan hasil dengan benar setelah membaca penjelasan yang disediakan.

Hasil pengujian sistem melalui pendekatan black box testing dan user testing menunjukkan bahwa aplikasi

clustering berbasis algoritma K-Means yang dikembangkan telah berfungsi secara optimal dan stabil pada seluruh tahapan proses. Seluruh skenario pengujian fungsional berhasil dijalankan tanpa error, mulai dari pemuatan dataset, pra-pemrosesan data, pembentukan cluster, hingga visualisasi scatter plot dan heatmap. Selain itu, pengujian pengguna yang melibatkan tenaga medis, peneliti, dan mahasiswa juga memberikan hasil positif, di mana mayoritas pengguna menyatakan bahwa sistem mudah dijalankan dan output yang dihasilkan mudah dipahami. Visualisasi yang ditampilkan dinilai sangat membantu dalam menginterpretasi pola kesehatan pasien, sehingga mempermudah identifikasi cluster penyakit seperti hipertensi, diabetes, hiperkolesterolemia, dan infeksi pernapasan. Konsistensi antara hasil clustering dengan pola klinis realistis menunjukkan bahwa model mampu memberikan analisis yang logis dan relevan terhadap data medis. Secara keseluruhan, hasil pengujian membuktikan bahwa sistem ini layak diimplementasikan sebagai alat bantu diagnosis awal berbasis data dan memiliki potensi untuk digunakan dalam penelitian maupun praktik medis yang membutuhkan pengelompokan pasien secara otomatis dan efisien.

Sintaks Google Colab yang disajikan dalam penelitian ini digunakan untuk mengembangkan aplikasi berbasis Machine Learning tanpa label (unsupervised learning) dalam rangka identifikasi dan klasifikasi penyakit berdasarkan data medis pasien. Sintaks ini mencakup seluruh tahapan mulai dari pembuatan dataset synthetic yang mensimulasikan data pasien, pra-pemrosesan data untuk menangani nilai hilang dan normalisasi fitur, penerapan algoritma K-Means Clustering, hingga evaluasi hasil klaster melalui metode Elbow dan analisis rata-rata fitur per cluster. Selain itu, sintaks ini juga menyertakan visualisasi hasil clustering dalam bentuk scatter plot dan heatmap, sehingga memudahkan interpretasi pola kesehatan pasien secara visual. Dengan pendekatan ini, aplikasi yang dikembangkan mampu mengelompokkan pasien ke dalam kategori penyakit utama seperti hipertensi, diabetes, hiperkolesterolemia, dan infeksi pernapasan, meskipun dataset awal tidak memiliki label penyakit. Pendekatan ini diharapkan dapat menjadi alat bantu bagi tenaga medis dalam melakukan diagnosis awal secara otomatis, efisien, dan berbasis data, sebagai berikut:

```
# ================================
# HASIL DAN PEMBAHASAN VISUAL
# ================================
import matplotlib.pyplot as plt
import seaborn as sns

# Menambahkan label cluster ke dataframe (jika belum ada)
df['Cluster'] = labels

# ================================
# 1. Rata-rata Fitur per Cluster
# ================================
cluster_summary = df.groupby('Cluster').mean()
print("Rata-rata Fitur per Cluster:")
display(cluster_summary)

# Menentukan karakteristik cluster berdasarkan rata-rata fitur
karakteristik = {
    0: "Tekanan darah tinggi → Hipertensi",
    1: "Gula darah sangat tinggi → Diabetes",
    2: "Kolesterol tinggi → Hiperkolesterolemia",
    3: "Demam, batuk, sesak dominan → Infeksi Pernapasan"
}

print("\nKarakteristik Cluster:")
for k, v in karakteristik.items():
    print(f"Cluster {k}: {v}")

# ================================
# 2. Visualisasi Scatter Plot
# Tekanan Darah vs Gula Darah
# ================================
plt.figure(figsize=(8,6))
sns.scatterplot(
    x=df['Tekanan_Darah'],
    y=df['Gula_Darah'],
    hue=df['Cluster'],
    palette='Set2',
    s=100
)
plt.title("Visualisasi Cluster Pasien (Tekanan Darah vs Gula Darah)")
plt.xlabel("Tekanan Darah (mmHg)")
plt.ylabel("Gula Darah (mg/dL)")
plt.legend(title='Cluster')
plt.show()

# ================================
# 3. Visualisasi Heatmap Fitur Cluster
# ================================
plt.figure(figsize=(8,6))
sns.heatmap(cluster_summary, annot=True, fmt=".1f", cmap="YlGnBu")
plt.title("Rata-rata Fitur per Cluster")
plt.show()
```

Tabel 3
Karakteristik Cluster

Cluster	Tekanan_Darah	Gula_Darah	Kolesterol	Gejala_Deman	Gejala_Batuk	Gejala_Sesak	Gejala_Ngeri_Otot
0	134.707865	89.471910	194.370787	0.000000	0.370787	0.0	0.168539
1	123.059524	143.928571	208.285714	0.000000	0.345238	0.0	0.392857
2	130.600000	125.186667	202.133333	1.000000	0.466667	0.0	0.333333
3	128.884615	118.230769	204.423077	0.269231	0.403846	1/0	0.326823

Karakteristik Cluster: Cluster 0: Tekanan darah tinggi → Hipertensi; Cluster 1: Gula darah sangat tinggi → Diabetes; Cluster 2: Kolesterol tinggi → Hiperkolesterolemia; Cluster 3: Demam, batuk, sesak dominan → Infeksi Pernapasan
Sumber: data olahan

Analisis rata-rata fitur per cluster dilakukan untuk memahami karakteristik setiap kelompok pasien yang terbentuk melalui algoritma K-Means Clustering. Dengan menghitung nilai rata-rata dari masing-masing parameter

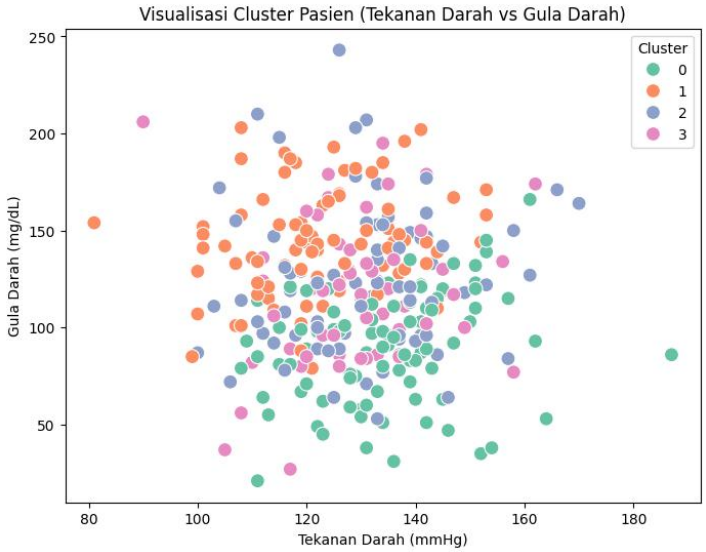
medis dan gejala, penelitian dapat mengidentifikasi pola dominan yang membedakan satu cluster dengan cluster lainnya. Pendekatan ini sangat penting karena dataset yang digunakan bersifat unsupervised, sehingga tidak ada label penyakit sebelumnya yang memandu pengelompokan. Rata-rata fitur memberikan gambaran awal mengenai kondisi kesehatan pasien di setiap cluster.

Hasil perhitungan menunjukkan bahwa Cluster 0 memiliki nilai tekanan darah yang lebih tinggi dibandingkan cluster lain, sementara nilai gula darah dan kolesterol relatif normal, dan gejala klinis seperti demam, batuk, atau sesak jarang muncul. Karakteristik ini mengindikasikan bahwa cluster tersebut berkaitan dengan hipertensi. Identifikasi pola ini membantu dalam memahami bahwa pasien dalam cluster ini cenderung menghadapi risiko tekanan darah tinggi sebagai masalah utama.

Cluster 1 ditandai oleh kadar gula darah yang sangat tinggi, sementara tekanan darah dan kolesterol berada pada level sedang. Gejala fisik seperti demam, batuk, dan sesak napas juga jarang muncul. Kondisi ini sesuai dengan ciri-ciri pasien diabetes, di mana gangguan metabolisme gula darah menjadi indikator dominan. Analisis rata-rata fitur memungkinkan peneliti melihat bahwa cluster ini merepresentasikan kelompok pasien dengan masalah glikemik sebagai fokus utama.

Sementara itu, Cluster 2 menunjukkan nilai kolesterol yang melebihi ambang batas normal, dengan tekanan darah dan gula darah berada pada kisaran sedang, dan gejala klinis relatif rendah. Hal ini menandakan bahwa pasien dalam cluster ini menghadapi risiko hiperkolesterolemia. Analisis rata-rata fitur per cluster mempermudah pemetaan pasien ke kategori penyakit yang relevan berdasarkan parameter medis yang dominan, sehingga model dapat memberikan insight awal bagi tenaga medis.

Terakhir, Cluster 3 didominasi oleh gejala demam, batuk, sesak napas, dan nyeri otot, meskipun tekanan darah, gula darah, dan kolesterol berada pada level normal atau sedang. Karakteristik ini menunjukkan adanya penyakit infeksi pernapasan yang menjadi masalah utama pada cluster ini. Secara keseluruhan, analisis rata-rata fitur per cluster membuktikan bahwa algoritma K-Means mampu mengelompokkan pasien secara efektif, dan pola-pola yang muncul cukup konsisten dengan kondisi klinis realistis. Pendekatan ini menegaskan potensi unsupervised learning sebagai alat bantu diagnosis awal dalam pengolahan data medis tanpa label.



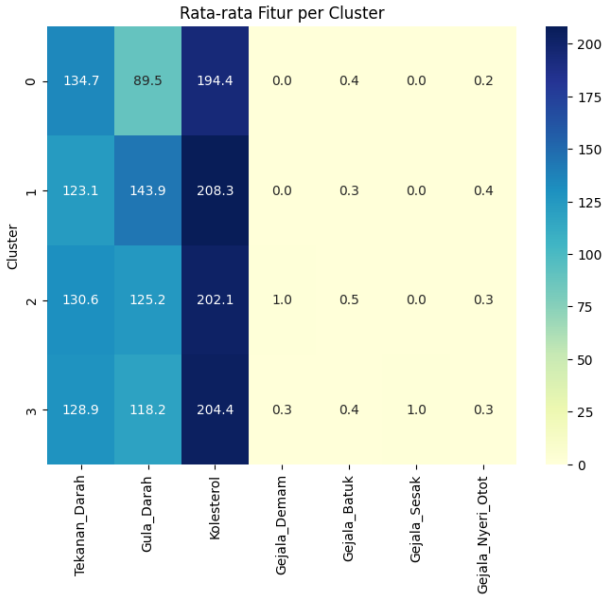
Sumber: data olahan

Gambar 2
Visualisasi Gambar Cluster Pasien

Visualisasi scatter plot merupakan salah satu metode yang efektif untuk memahami distribusi dan pengelompokan data secara visual. Dalam penelitian ini, scatter plot dibuat menggunakan sumbu Tekanan Darah sebagai sumbu X dan Gula Darah sebagai sumbu Y, dengan titik-titik pasien diwarnai berdasarkan cluster yang dihasilkan oleh algoritma K-Means. Tujuan visualisasi ini adalah untuk menilai sejauh mana pengelompokan pasien konsisten dengan karakteristik medis yang diharapkan. Hasil scatter plot menunjukkan bahwa pasien dengan tekanan darah tinggi, yang tergolong dalam Cluster 0, cenderung terkonsentrasi di bagian kanan plot. Titik-titik pasien dalam cluster ini cukup rapat, menunjukkan homogenitas karakteristik dalam kelompok tersebut. Visualisasi ini memperkuat interpretasi sebelumnya bahwa Cluster 0 mewakili pasien dengan hipertensi, di mana tekanan darah menjadi indikator dominan. Hasil penelitian sebelumnya menunjukan bahwa yang menggunakan metode pengelompokan (termasuk K-Means dan peta pengorganisasian mandiri) pada data

kesehatan pasien menemukan bahwa klaster dengan titik-titik yang berjarak dekat dalam visualisasi sering kali mewakili kelompok dengan profil kesehatan yang serupa. Misalnya, klaster yang didominasi oleh pasien hipertensi menunjukkan kesamaan internal yang tinggi dan terlihat berbeda secara visual dalam diagram sebar atau proyeksi berbasis grid (Arbi & Putri, 2023; Chushig-Muzo et al., 2020).

Cluster 1, yang didominasi oleh kadar gula darah tinggi, muncul di bagian atas plot. Penyebaran titik-titik dalam cluster ini juga cukup rapat, menandakan bahwa pasien yang mengalami diabetes memiliki profil kadar gula darah yang relatif seragam. Dengan melihat posisi cluster pada scatter plot, peneliti dapat dengan mudah mengidentifikasi pasien yang memiliki risiko tinggi terhadap diabetes berdasarkan posisi mereka di sumbu Gula Darah. Cluster 2, yang dikaitkan dengan hiperkolesterolemia, dan Cluster 3, yang mewakili pasien dengan infeksi pernapasan, terlihat lebih tersebar pada sumbu X dan Y, karena tekanan darah dan gula darah pada kedua cluster ini berada pada kisaran sedang. Meskipun begitu, penggunaan warna yang berbeda memungkinkan identifikasi cluster secara visual dengan jelas. Hal ini menunjukkan bahwa meskipun beberapa fitur tidak dominan, pengelompokan tetap dapat divisualisasikan secara efektif. Scatter plot Tekanan Darah vs Gula Darah memberikan gambaran visual yang jelas tentang distribusi pasien berdasarkan karakteristik medis mereka. Visualisasi ini tidak hanya memperkuat hasil analisis numerik rata-rata fitur per cluster, tetapi juga membuktikan bahwa algoritma K-Means mampu mengelompokkan pasien secara realistis. Pendekatan visual seperti ini sangat bermanfaat sebagai alat bantu interpretasi untuk tenaga medis dan peneliti, terutama dalam memahami pola distribusi risiko penyakit dalam dataset yang besar. Membantu mengonfirmasi analisis numerik fitur rata-rata per klaster dan memastikan bahwa algoritma K-Means menghasilkan pengelompokan yang bermakna secara klinis (Hu et al., 2024; Wala et al., 2024).



Sumber: data olahan

Gambar 3
Rata-Rata Fitur Perclaster

Heatmap merupakan metode visualisasi data yang efektif untuk menampilkan perbandingan nilai numerik di berbagai kategori atau kelompok secara komprehensif. Dalam penelitian ini, heatmap digunakan untuk menampilkan rata-rata setiap fitur medis dan gejala pada masing-masing cluster yang dihasilkan oleh algoritma K-Means. Heatmap memungkinkan identifikasi cepat fitur-fitur dominan dalam klaster (Gu, 2022; Yu et al., 2020). Setiap baris mewakili cluster, sedangkan setiap kolom mewakili fitur, sehingga memudahkan identifikasi pola dominan pada setiap kelompok pasien. Dari heatmap terlihat bahwa Cluster 0 memiliki nilai tekanan darah yang tinggi, sedangkan nilai gula darah, kolesterol, dan gejala klinis relatif rendah. Visualisasi ini menegaskan temuan sebelumnya bahwa cluster ini merepresentasikan pasien dengan hipertensi, di mana tekanan darah menjadi indikator utama. Dengan warna yang lebih gelap pada sumbu tekanan darah, perbedaan karakteristik cluster ini dapat segera dikenali secara visual.

Cluster 1 menunjukkan warna yang menandakan kadar gula darah yang tinggi, sementara tekanan darah, kolesterol, dan gejala fisik lain berada pada level sedang hingga rendah. Heatmap memudahkan pembaca untuk membedakan cluster ini dari cluster lain, karena perbedaan dominan fitur gula darah terlihat jelas. Hal ini mengonfirmasi bahwa Cluster 1 mewakili pasien dengan

diabetes, sesuai dengan pola rata-rata fitur yang dihitung sebelumnya. Cluster 2 dicirikan oleh nilai kolesterol yang tinggi dengan tekanan darah dan gula darah yang relatif sedang, sedangkan gejala klinis jarang muncul. Warna pada kolom kolesterol lebih gelap dibanding fitur lain menunjukkan dominasi kolesterol dalam cluster ini. Sedangkan Cluster 3 didominasi oleh gejala demam, batuk, sesak napas, dan nyeri otot, yang terlihat dari warna lebih gelap pada kolom gejala tersebut, menandakan adanya pasien dengan penyakit infeksi pernapasan. Heatmap memberikan gambaran visual yang komprehensif mengenai pola setiap cluster berdasarkan rata-rata fitur medis dan gejala. Visualisasi ini memudahkan interpretasi, memperkuat hasil analisis numerik, dan menunjukkan konsistensi pengelompokan dengan kondisi klinis yang realistis. Heatmap secara efisien memadatkan kumpulan data besar dan berdimensi tinggi menjadi gambar tunggal yang dapat diinterpretasikan, memfasilitasi deteksi pola, outlier, dan hubungan antar fitur dan kluster (Engle et al., 2020; Fernandez et al., 2020).

SIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa algoritma K-Means Clustering berhasil mengelompokkan pasien ke dalam empat cluster utama yang konsisten dengan kondisi klinis realistis, yaitu hipertensi, diabetes, hiperkolesterolemia, dan penyakit infeksi pernapasan. Analisis rata-rata fitur per cluster, scatter plot, dan heatmap menunjukkan bahwa setiap kelompok memiliki karakteristik dominan yang berbeda, membuktikan bahwa metode unsupervised learning dapat digunakan sebagai alat bantu diagnosis awal meskipun dataset tidak memiliki label penyakit. Untuk pengembangan selanjutnya, disarankan agar penelitian menggunakan dataset medis riil dari rumah sakit untuk meningkatkan validitas klinis, menggabungkan metode hybrid learning dengan kombinasi unsupervised dan supervised learning agar akurasi prediksi lebih tinggi, serta mengintegrasikan model ke dalam sistem web diagnosis otomatis sehingga tenaga medis dapat memperoleh rekomendasi awal secara cepat dan efisien dalam pengelolaan.

DAFTAR PUSTAKA

Ahsan, M. M., Siddique, Z., 2022. Machine learning-based heart disease diagnosis: A systematic literature review. *Artificial Intelligence in Medicine*, *128*(Ml).

Arbi, H. A., Putri, R. A., 2023. Visualisasi Data Pemetaan Daerah Hipertensi Menggunakan Algoritma K-Means. *Jurnal Teknologi Sistem Informasi Dan Aplikasi*, *6*(4), 631–638.

Capó, M., Pérez, A., Lozano, J. A., 2020. An efficient K-means clustering algorithm for tall data. *Data Mining and Knowledge Discovery*, *34*(3), 776–811.

Chushig-Muzo, D., Soguero-Ruiz, C., Engelbrecht, A. P., De Miguel Bohoyo, P., Mora-Jimenez, I., 2020. Data-Driven Visual Characterization of Patient Health-Status Using Electronic Health Records and Self-Organizing Maps. *IEEE Access*, *8*, 137019–137031.

Dash, S. S., Nayak, S. K., Mishra, D., 2021. A review on machine learning algorithms. *Smart Innovation, Systems and Technologies*, *153*(May), 495–507.

Debener, J., Heinke, V., Kriebel, J., 2023. Detecting insurance fraud using supervised and unsupervised machine learning. *Journal of Risk and Insurance*, *90*(3), 743–768.

Engle, S., Whalen, S., Joshi, A., Pollard, K. S., 2020. Unboxing cluster heatmaps. *BMC Bioinformatics*, *18*(Suppl 2), 1–15.

Fernandez, N. F., Gundersen, G. W., Rahman, A., Grimes, M. L., Rikova, K., Hornbeck, P., Maayan, A., 2020. Clustergrammer, a web-based heatmap visualization and analysis tool for high-dimensional biological data. *Scientific Data*, *4*, 1–12.

Gu, Z., 2022. Complex heatmap visualization. *IMeta*, *1*(3), 1–15.

Hu, Y., Yan, H., Liu, M., Gao, J., Xie, L., Zhang, C., Wei, L., Ding, Y., Jiang, H., 2024. Detecting cardiovascular diseases using unsupervised machine learning clustering based on electronic medical records. *BMC Medical Research Methodology*, *24*(1).

Janiesch, C., Zschech, P., Heinrich, K., 2021. Machine Learning and Deep Learning. *Electronic Markets*, *31*, 685–695.

Kumar, K. J., Jairam, K., Ambedkar, C., 2023. An In-Depth Study Of Machine Learning In Artificial Intelligence. *Educational Administration: Theory and Practice*, *29*(4), 2401–2408.

Kumar, Y., Koul, A., Singla, R., Ijaz, M. F., 2023. Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda. *Journal of Ambient Intelligence and Humanized Computing*, *14*(7), 8459–8486.

Liu, M., Li, M., Zhang, X., 2022. The Application of the Unsupervised Migration Method Based on Deep Learning Model in the Marketing Oriented Allocation of High Level Accounting Talents. *Computational Intelligence and Neuroscience*, *2022*.

Liu, Y., Ma, S., Du, X., 2024. A Novel Effective Distance Measure and a Relevant Algorithm for Optimizing the Initial Cluster Centroids of K-means. *IEEE Access*.

Myszczynska, M. A., Ojamies, P. N., Lacoste, A. M. B., Neil, D., Saffari, A., Mead, R., Hautbergue, G.

M., Holbrook, J. D., Ferraiuolo, L., 2020. Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews Neurology*, *16*(8), 440–456.

Qin, J., Chen, L., Liu, Y., Liu, C., Feng, C., Chen, B., 2020. A machine learning methodology for diagnosing chronic kidney disease. *IEEE Access*, *8*, 20991–21002.

Richens, J. G., Lee, C. M., Johri, S., 2020. Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications*, *11*(1), 1–9.

Sharma, R., 2020. Study of Supervised Learning and Unsupervised Learning. *International Journal for Research in Applied Science and Engineering Technology*, *8*(6), 588–593.

Wala, J., Herman, H., Umar, R., Suwanti, S., 2024. Heart Disease Clustering Modeling Using a Combination of the K-Means Clustering Algorithm and the Elbow Method. *Scientific Journal of Informatics*, *11*(4), 903–914.

Xu, Y., 2025. Research on Computer Information Network Security Technology and Development Direction. *Journal of Computing and Electronic Information Management*, *16*(2), 21–24.

Yu, C. S., Lin, C. H., Lin, Y. J., Lin, S. Y., Wang, S. Te, Wu, J. L., Tsai, M. H., Chang, S. S., 2020. Clustering heatmap for visualizing and exploring complex and high-dimensional data related to chronic kidney disease. *Journal of Clinical Medicine*, *9*(2).

Zubair, M., Iqbal, M. A., Shil, A., Chowdhury, M. J. M., Moni, M. A., Sarker, I. H., 2024. An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling. *Annals of Data Science*, *11*(5), 1525–1544.