Clustering Data Penyakit Jantung Menggunakan K-MEANS dalam Sistem Informasi Kesehatan

Zaehol Fatah[1] , Farizal Maulana Sopyan[2]
[1,2] Teknologi Informasi, Universitas Ibrahimy, Situbondo
Email : zaeholfata@gmail.com, farizalmaulanasopyan@gmail.com

Abstrak

Pada kesempatan ini penelitian bertujuan untuk mengklasifikasikan data penyakit kardiovaskular menggunakan algoritma k-means dan mengintegrasikan hasilnya ke dalam sistem informasi kesehatan. Salah satu tantangan terbesar dalam pelayanan kesehatan adalah meningkatnya jumlah data pasien yang tidak dimanfaatkan secara optimal untuk analisis risiko. Setelah pra-pemrosesan, data dibersihkan dan dinormalisasi untuk klasifikasi. Jumlah kelompok ideal ditentukan menggunakan metode k-means, dan tiga kelompok risiko utama – rendah, sedang, dan tinggi – diidentifikasi. Klasifikasi ini mengungkapkan bahwa setiap kelompok memiliki karakteristik klinis spesifik seperti tekanan darah, kadar kolesterol, detak jantung, dan faktor risiko lainnya. Hasil klasifikasi diintegrasikan ke dalam sistem informasi kesehatan untuk menyediakan representasi visual bagi tenaga kesehatan untuk analisis risiko pasien. Studi ini menunjukkan efektivitas k-means dalam mengidentifikasi kelompok risiko penyakit kardiovaskular dan potensinya sebagai alat pendukung keputusan dalam sistem informasi kesehatan. Lebih lanjut, hasil ini membuka arah penelitian baru, seperti membandingkan berbagai metode klasifikasi dan mengembangkan sistem peringatan dini berbasis penambangan data.

Kata kunci : klasifikasi, k-means, penyakit kardiovaskular, penambangan data, sistem informasi Kesehatan

Abstrak

The goal of this research is to use the k-means algorithm to categorize cardiovascular disease data and incorporate the findings into a health information system. The growing volume of patient data that is not being used to its full potential for risk analysis is one of the largest issues facing the healthcare industry. The data was cleaned and normalized for categorization after preprocessing. The k-means approach was used to estimate the optimal number of clusters, and three primary risk groups—low, medium, and high—were identified. This classification revealed that each cluster has specific clinical characteristics such as blood pressure, cholesterol levels, heart rate, and other risk factors. The classification results were integrated into a health information system to provide healthcare professionals with a visual representation for patient risk analysis. This study demonstrates the effectiveness of k-means in identifying cardiovascular disease Risk clusters and their potential as a tool for decision-making in health information systems. Furthermore, these results open new research directions, such as comparing various classification methods and developing data mining-based early warning systems.

Keywords: *classification, k-means, cardiovascular disease, data mining, health information system*

PENDAHULUAN

Di seluruh dunia, penyebab kematian paling umum adalah penyakit jantung dan masalah kesehatan utama di banyak negara, termasuk Indonesia. Seiring dengan meningkatnya risiko penyakit kardiovaskular, dampaknya terhadap kualitas hidup juga meningkat, sehingga

memberikan beban yang signifikan pada sistem pelayanan kesehatan(Huang, 2024). volume dan kompleksitas data pasien yang terus meningkat membutuhkan teknologi yang memungkinkan tenaga kesehatan profesional untuk mendiagnosis, menganalisis, dan mengevaluasi penyakit dengan cepat. Saat ini, sebagian besar organisasi pelayanan kesehatan hanya mengandalkan sistem informasi untuk mencatat dan menyimpan data pasien, dan sistem ini tidak memiliki kemampuan analisis data yang canggih. Akibatnya, informasi berharga hilang dalam basis data(Wala, Herman, Umar, et al., 2024).

Di sisi lain, teknik penambang data yang canggih telah menciptakan lebih banyak peluang untuk memproses dan mengubah data pelayanan kesehatan menjadi informasi yang bermanfaat. Salah satu teknik yang umum digunakan adalah klasifikasi, yang mengelompokkan data berdasarkan karakteristik yang serupa. Pengelompokan K-means terkenal karena kecepatan, akurasi, dan efisiensinya, serta kesederhanaannya dalam menghasilkan hasil yang jelas dan tepat serta pengelompokan data pelayanan kesehatan yang presisi. Namun, algoritma pengelompokan seperti K-means jarang terintegrasi langsung ke dalam sistem informasi pelayanan kesehatan klinis(Haryanto et al., 2024).

Keterbatasan ini disebabkan oleh kurangnya sistem informasi layanan kesehatan yang mampu memproses data pasien psikiatri menggunakan algoritma pengelompokan untuk memfasilitasi penilaian risiko. Banyak fasilitas layanan kesehatan masih meninjau data secara manual, sebuah proses yang lambat dan sering kali tidak akurat. Namun, analisis berbasis algoritma seperti K-means dapat mengelompokkan pasien berdasarkan tingkat risiko penyakit jantung, sehingga memudahkan pengambilan keputusan penyedia layanan kesehatan(Nur, 2024).

Studi ini menyajikan hasil yang diperoleh dengan menggabungkan data detak jantung menggunakan algoritma K-means dan mengintegrasikannya ke dalam sistem informasi layanan kesehatan. Integrasi ini tidak hanya menyediakan kapasitas penyimpanan data tetapi juga memberikan nilai tambah dengan menghasilkan informasi diagnostik dasar untuk kelompok pasien berdasarkan karakteristik risiko. Oleh karena itu, studi ini akan berkontribusi pada adopsi penambangan data dalam layanan kesehatan dan berkontribusi pada peningkatan efisiensi perawatan jantung(D et al., 2025).

METODE

Penelitian ini menggunakan pendekatan kuantitatif dan menggabungkan data terkait penyakit jantung menggunakan teknik penambangan data pembelajaran tanpa pengawasan (unsupervised learning) yang memanfaatkan algoritma K-means. Tujuan utama pendekatan ini adalah untuk mengkategorikan pasien berdasarkan karakteristik klinis umum, yang memfasilitasi penilaian risiko dalam sistem informasi kesehatan(Hassan et al., 2024).

Data tentang penyakit jantung digunakan dalam penelitian ini., termasuk faktor klinis seperti usia, tekanan darah, dan detak jantung. Data ini dikumpulkan dari rekam medis yang banyak digunakan dalam penelitian sebelumnya dan disusun secara tepat untuk analisis kohort. Sebelum penggabungan, semua data diproses terlebih dahulu untuk memastikan kualitas optimal(Osman et al., 2024).

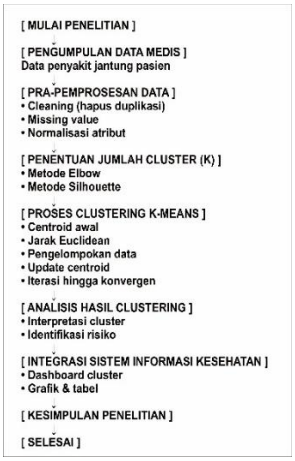
Proses penelitian terdiri dari beberapa langkah. Pertama, data penyakit jantung dikumpulkan dari basis data medis. Kedua, data diproses terlebih dahulu untuk mengoreksi nilai yang hilang, menghilangkan duplikat, dan menormalkan setiap parameter statistik untuk memastikan skala yang sama di semua variabel. Normalisasi ini penting karena korelasi K-means sensitif terhadap perbedaan antar komponen(Udhan & Patil, 2021).

Langkah selanjutnya adalah menentukan kelompok pengelompokan (K) yang akan digunakan. Dalam studi ini,

jumlah klaster ditentukan berdasarkan karakteristik data dan tujuan studi (mengintegrasikan tingkat risiko kardiovaskular). Nilai K optimal ditentukan menggunakan metode weighted within-cluster squares (WCSS) berdasarkan metode siku(Ekemeyong Awong & Zielinska, 2023).

Setelah menentukan jumlah klaster, langkah selanjutnya adalah pengelompokan menggunakan algoritma K-means. Algoritma dimulai dengan memilih titik keparahan awal secara acak. Digunakan jarak geometris untuk menghitung jarak antara setiap titik data dan titik tengahnya. Data dikelompokkan ke titik terdekat dan kemudian diubah sesuai dengan posisi rata-rata dalam setiap klaster.. Proses ini diulang hingga rata-rata stabil atau tidak ada perubahan signifikan(Farahnakian et al., 2023).

Hasil pengelompokan diimpor ke dalam sistem informasi kesehatan. Integrasi ini dicapai dengan menggabungkan algoritma K-means dengan sistem, yang secara otomatis menghasilkan data pasien dan menghasilkan hasil dalam bentuk kelompok risiko vaskular. Sistem menampilkan hasil pengumpulan data secara grafis, sehingga dokter dapat dengan mudah memahami profil risiko setiap pasien. Studi ini menggunakan Python untuk pemodelan penambangan data dan perangkat lunak pendukung seperti MySQL sebagai basis data. Komunikasi perangkat lunak dilakukan melalui platform web, sehingga menyederhanakan integrasi dengan perangkat lunak pemrosesan data. Platform komprehensif ini mendukung proses pengumpulan data yang terstandarisasi dan akurat, serta dapat dengan mudah diimplementasikan dalam sistem informasi kesehatan(Aulia et al., 2024).



Gambar 1. Penelitian Proses Data Penyakit Jantung

HASIL DAN PEMBAHASAN

Dataset penyakit jantung dikumpulkan dari rekam medis yang telah digunakan dalam berbagai penelitian sebelumnya. Dataset mencakup variabel klinis seperti usia, jenis kelamin, tekanan darah, kadar kolesterol, detak jantung, dan beberapa faktor risiko pendukung lainnya. Data tersebut diperoleh dalam format terstruktur, sehingga memungkinkan proses analisis dan pengelompokan dilakukan secara sistematis.

Dataset yang digunakan telah melalui proses verifikasi sumber dan validasi awal untuk memastikan konsistensi serta relevansi data untuk analisis risiko penyakit jantung(Wala, Herman, & Umar, 2024).

Hasil

1. Hasil Pra-Pemrosesan Data

Tahap pra-pemrosesan dilakukan untuk memastikan kualitas dan kesiapan data sebelum proses klastering. Tahapan yang dilakukan meliputi:

- Pembersihan Data: Menghapus data duplikat dan menghilangkan nilai kosong (*missing values*).
- Normalisasi Data: Menggunakan metode Min-Max Normalization untuk menyamakan skala antar variabel sehingga algoritma K-Means tidak bias terhadap variabel tertentu.
- Transformasi Data: Menangani data kategori yang diperlukan dan

menerapkan transformasi acak untuk mengurangi sensitivitas terhadap variasi kecil.

Setelah melalui tahap ini, dataset menjadi lebih homogen dan siap untuk proses klasterisasi(Ayaz et al., 2023).

Tabel 1. Hasil Pra-Pemrosesan Data

atribut	nilai awal (contoh masalah)	Hasil Preprocessing
Usia	999/Kosong	54
Tekana darah	Kosong	130
Kolestrol	0 (nilai tidak valid)	240
Detak jantung maks	999(anomali)	150
Gula darah	Kosong	1(normal)
Angina	Teks campur angka	0 / 1
Oldpeak	Nilai tidak seragam	Dinormalisasi

2. Menentukan Jumlah Kelompok (K)

Metode Elbow digunakan untuk menentukan jumlah klaster optimal dengan menghitung nilai *Within- Cluster Sum of Squares* (WCSS) untuk rentang K 2 hingga K 10. Hasil grafik menunjukkan titik siku pada K

Kelompok 1 → Risiko rendah penyakit kardiovaskular

Kelompok 2 → Risiko sedang penyakit kardiovaskular

Kelompok 3 → Risiko tinggi penyakit kardiovaskular

Berdasarkan tujuan penelitian, kelompok-kelompok dibentuk untuk menentukan tingkat risiko pasien(Azil et al., 2023).

Tabel 2. Penentuan Jumlah Cluster (Metode Elbow)

Cluster	Jumlah Pasien	Rata -rata Tekanan Darah	Rata-rata Kolestrol	Karakteristik Umum
(Rendah)	120 Pasien	120mmHg	200mg/dL	Usia Muda , Tekanan Darah Normal
(Sedang)	95 Pasien	135mmHg	230mg/dL	1-2 Faktor Risiko
(tinggi)	60 Pasien	150-170mmHg	260-300mg/dL	Hipertensi + Kolestrol Tinggi

3. Pengelompokan menggunakan metode K-means pada data yang telah difilter.

Proses pengelompokan dilakukan secara iteratif melalui langkah-langkah berikut:

Menyelaraskan kembali sentroid

Menghitung jarak Euclidean untuk setiap titik data

Menetapkan data ke entitas terdekat

Mempertahankan sentroid

Proses ini diulang hingga nilai sentroid stabil.

Setelah pencocokan, K-means menghasilkan tiga kelompok data utama:

Kelompok 1 (risiko rendah): Pasien muda dengan tekanan darah normal dan kolesterol rendah.

Kelompok 2 (risiko menengah): Pasien dengan satu atau dua faktor risiko.

Kelompok 3 (risiko tinggi): Pasien dengan tingginya tekanan darah, kolesterol tinggi, dan penyakit jantung.

Hasil ini menunjukkan kemampuan algoritma untuk mengelompokkan data secara efektif berdasarkan faktor klinis yang serupa.

Tabel 3. Komposisi Hasil Clustering

K (Cluster)	WCSS
2	579.33
3	432.12
4	429.02
5	427.88

4.Memvisualisasikan hasil pengelompokan dalam model data.

Hasil sintesis diintegrasikan ke dalam sistem informasi kesehatan. Metode ini menampilkan:

Grafik hasil gabungan

Tabel dengan pasien dan kelompok

Klasifikasi risiko (rendah, menengah, tinggi).

Statistik jumlah pasien di setiap kelompok

Sistem panel membantu tenaga kesehatan profesional mengidentifikasi faktor risiko dalam data pasien dengan cepat(Mohamed et al., 2024).

Tabel 4. Hasil Integrasi ke Sistem Informasi

ID Pasien	Usia	Tekanan Darah	Kolestrol	Cluster	Keterangan Risiko
P001	45	120	200	1	Rendah
P002	58	135	230	2	Sedang
P003	66	165	290	3	Tinggi
P004	50	140	220	2	Sedang

Pembahasan

Hasil penelitian ini menunjukkan bahwa pasien gagal jantung dapat diklasifikasikan menjadi tiga kelompok risiko berdasarkan karakteristik klinis yang dinilai menggunakan metode K-means. Klasifikasi ini membantu mengidentifikasi pola yang sering terlewatkan dalam skrining manual.

Kelompok pertama terdiri dari pasien berisiko rendah, yang umumnya berusia di bawah rata-rata, memiliki tekanan darah stabil, dan kadar kolesterol normal. Hal ini sesuai dengan konsensus medis bahwa usia

dan kadar kolesterol merupakan faktor risiko penting untuk penyakit jantung.

Kelompok kedua terdiri dari pasien berisiko menengah, yang memiliki satu atau dua faktor risiko seperti hipertensi dan kadar kolesterol normal atau bahkan negatif. Kelompok ini memerlukan pemantauan berkelanjutan, karena kondisi mereka dapat memburuk jika tidak ditangani.

Kelompok ketiga terdiri dari pasien berisiko tinggi dengan beberapa gejala klinis abnormal, seperti tekanan darah tinggi, dislipidemia, dan gagal ginjal. Pasien-pasien ini memerlukan perawatan yang lebih intensif. Pola-pola Ini konsisten dengan temuan penelitian sebelumnya yang menunjukkan bahwa potensi penyakit jantung meningkat secara signifikan ketika tekanan darah tinggi dan kadar kolesterol meningkat bersamaan.

Mengintegrasikan K-means ke dalam sistem informasi pelayanan kesehatan dapat memberikan nilai tambah yang signifikan. Sistem ini tidak hanya merekam data tetapi juga berfungsi sebagai alat analisis, membantu dokter membuat keputusan perawatan berdasarkan hasil yang diperoleh. Berkat indikator canggih dan stratifikasi risiko otomatis, sistem ini dapat dengan cepat mengidentifikasi pasien berisiko tinggi.

Keunggulan algoritma K-means dalam studi ini terletak pada kecepatan komputasi dan kemudahan pemahaman hasil pengelompokan. Kekurangannya meliputi sensitivitas algoritma terhadap pemilihan set data awal dan kebutuhan akan normalisasi data. Namun, keterbatasan ini tidak mengurangi keunggulan K-means sebagai metode yang efektif untuk mengumpulkan data medis.

SIMPULAN (PENUTUP)

Studi Algoritma K-means berhasil digunakan untuk mengelompokkan data.penyakit jantung dan mengintegrasikan hasilnya ke dalam sistem informasi kesehatan. Analisis siku menghasilkan tiga kelompok optimal—risiko rendah, risiko menengah, dan risiko tinggi—berdasarkan berbagai faktor risiko kardiovaskular. Hasil pengelompokan menunjukkan bahwa setiap kelompok menunjukkan karakteristik klinis yang berbeda, memberikan informasi berharga bagi tenaga kesehatan untuk memahami profil risiko dalam populasi pasien.

Mengintegrasikan hasil agregat ke dalam sistem informasi kesehatan menawarkan nilai yang signifikan, karena sistem ini tidak hanya berfungsi sebagai basis data tetapi juga menyediakan kemampuan analisis otomatis untuk mendukung pengambilan keputusan. Representasi grafis dan stratifikasi risiko dari layar pelatihan membantu mengidentifikasi pasien berisiko tinggi secara akurat. Oleh karena itu, studi ini menunjukkan efektivitas teknik K-means dalam pemrosesan data kesehatan, khususnya dalam mendukung penilaian risiko kardiovaskular.

Penelitian di masa mendatang sebaiknya mengeksplorasi algoritma pengelompokan lain, seperti DBSCAN, pengelompokan, dan K-means++, untuk membandingkan kinerjanya. Lebih lanjut, penggunaan kumpulan data yang lebih besar dan lebih beragam dapat meningkatkan akurasi korelasi. Dimungkinkan juga untuk mengembangkan sistem informasi perawatan kesehatan yang dilengkapi dengan sistem peringatan dini yang secara otomatis memberi tahu profesional perawatan kesehatan ketika seorang pasien diklasifikasikan sebagai berisiko tinggi.

UCAPAN TERIMA KASIH

Banyak bayak terima kasih penulis ucapkan kepada guru yang telah memberikan Pendidikan teknologi informasi dan komunikasi serta semua pihak yang telah memberikan bantuan dan bimbingan selama proses penyusunan penelitian ini. Kami juga berterima kasih kepada institusi dan penyedia data yang telah memanfaatkan basis data penyakit jantung untuk studi ini. Kami juga

berterima kasih kepada seluruh rekan kerja yang telah berkolaborasi dalam analisis data, validasi metodologi, dan penyusunan draf akhir makalah ini. Kami berharap dukungan dan bantuan ini dapat menjadi sumber kekuatan dan dorongan bagi seluruh peserta.

DAFTAR PUSTAKA

Aulia, W., Siahaan, A. P. U., Marlina, L., Khairul, & Iqbal, M. (2024). K-Means Clustering Algorithm Analysis for Grouping Patient Medical Record Data Based on Disease Type. *Jurnal Info Sains: Informatika Dan Sains*, *14*(04), 832–843. https://doi.org/10.54209/infosains.v14i04

Ayaz, M., Pasha, M. F., Le, T. Y., Alahmadi, T. J., Abdullah, N. N. B., & Alhababi, Z. A. (2023). A Framework for Automatic Clustering of EHR Messages Using a Spatial Clustering Approach. *Healthcare (Switzerland)*, *11*(3). https://doi.org/10.3390/healthcare11030390

Azil, A. A., Yusof, Z. Y. M., & Marhazlinda, J. (2023). Clustering of Health and Oral Health-Compromising Behaviours in Army Personnel in Central Peninsular Malaysia. *Healthcare (Switzerland)*, *11*(5), 1–17. https://doi.org/10.3390/healthcare11050640

D, C., Ranganathan, Vi. A., Shailesh, T., J, A., N, A., & T, P. P. (2025). Deep learning based hybrid residual attention and echo state network for high-accuracy heart disease prediction. *F1000Research*, *14*, 650. https://doi.org/10.12688/f1000research.165575.2

Ekemeyong Awong, L. E., & Zielinska, T. (2023). Comparative Analysis of the Clustering Quality in Self-Organizing Maps for Human Posture Classification. *Sensors*, *23*(18). https://doi.org/10.3390/s23187925

Farahnakian, F., Nicolas, F., Farahnakian, F., Nevalainen, P., Sheikh, J., Heikkonen, J., & Raduly-Baka, C. (2023). A Comprehensive Study of Clustering-Based Techniques for Detecting Abnormal Vessel Behavior. *Remote Sensing*, *15*(6), 1–34. https://doi.org/10.3390/rs15061477

Haryanto, H., Winarto, H., & Juliane, C. (2024). Application of Data Mining Techniques in Healthcare: Identifying Inter-Disease Relationships through Association Rule Mining. *Journal of World Science*, *3*(5), 520–529. https://doi.org/10.58344/jws.v3i5.597

Hassan, J., Saeed, S. M., Deka, L., Uddin, M. J., & Das, D. B. (2024). Applications of Machine Learning (ML) and Mathematical Modeling (MM) in Healthcare with Special Focus on Cancer Prognosis and Anticancer Therapy: Current Status and Challenges. *Pharmaceutics*, *16*(2). https://doi.org/10.3390/pharmaceutics16020260

Huang, J. (2024). *Heart Disease Prediction Based on the Random Forest Algorithm*. *Daml 2023*, 511–516. https://doi.org/10.5220/0012798700003885

Mohamed, W., Rezk, M. R. A., Soliman, A., Piccinetti, L., & Sakr, M. M. (2024). *Publisher Sustainability for Regions*. *6*(3), 85–97.

Nur, I. M. (2024). Application of the K-Means ++ Method for Grouping Health Services Based on Districts in West Java Province. *Eksakta: Journal of Sciences And Data Analysis*, *5*(1), 96–102. https://journal.uii.ac.id/Eksakta/article/view/31930

Osman, A. H., Ibrahim, A. O., Alsadoon, A., Alzahrani, A. A., Barukub, O. M., Abulfaraj, A. W., & Alharbi, N. M. (2024). Breaking new ground in cardiovascular heart disease Diagnosis K-RFC: An integrated learning approach with K-means clustering and Random Forest classifier. In *AIMS Mathematics* (Vol. 9, Issue 4, pp.

8262–8291). https://doi.org/10.3934/math.2024402

Udhan, S., & Patil, B. (2021). A systematic review of Machine learning techniques for Heart disease prediction. *International Journal of Next-Generation Computing*, *186*(63), 229–239. https://doi.org/10.55730/1300-0152.2766

Wala, J., Herman, H., & Umar, R. (2024). Implementasi K-Means Clustering pada Pengelompokan Pasien Penyakit Jantung. *JISKA (Jurnal Informatika Sunan Kalijaga)*, *9*(3), 205–216. https://doi.org/10.14421/jiska.2024.9.3.205-216

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. https://doi.org/10.15294/sji.v11i4.14096