

PENERAPAN METODE K-MEDOIDS UNTUK PENGELOMPOKAN PASIEN GAGAL JANTUNG MENGGUNAKAN DATASET BMC MEDICAL INFORMATICS AND DECISION MAKING

Zidan Indra Nugraha¹, Arnita², Alsya Adelia Putri³, Rendy Derian⁴, Risna Tutiarna Simorangkir⁵

zidan.84748@gmail.com¹, arnita@unimed.ac.id², alsyaadelia00@gmail.com³,
rendiderian@gmail.com⁴, rismasmrgkr@gmail.com⁵

Universitas Negeri Medan

Abstrak

Gagal jantung adalah kondisi medis kritis yang memengaruhi jutaan orang di seluruh dunia, termasuk Indonesia, dengan tingkat kematian yang tinggi. Penelitian ini bertujuan untuk meningkatkan pengobatan pasien gagal jantung menggunakan algoritma K-Medoids guna mengelompokkan pasien berdasarkan data klinis. Dataset yang digunakan mencakup 300 pasien gagal jantung dengan 12 atribut klinis, seperti usia, anemia, fraksi ejeksi, dan tekanan darah. Algoritma K-Medoids dipilih karena ketahanannya terhadap outlier, yang memberikan hasil klustering lebih akurat dibandingkan K-Means. Metode Elbow dan Skor Silhouette digunakan untuk menentukan jumlah kluster optimal. Hasilnya, tiga kluster pasien berhasil diidentifikasi. Salah satu kluster berisi pasien yang lebih tua, dengan hipertensi dan fraksi ejeksi rendah, yang menunjukkan risiko kematian lebih tinggi. Penelitian ini menunjukkan bahwa K-Medoids efektif dalam mengelompokkan pasien gagal jantung, memberikan wawasan tentang profil pasien dan memungkinkan intervensi medis yang lebih tepat sasaran. Keunggulan K-Medoids dalam menangani outlier membuka peluang penelitian lebih lanjut dalam stratifikasi risiko di bidang kesehatan.

Kata Kunci: K-Medoids, Data Klinis, Clustering, Data Mining.

Abstract

Heart failure represents a significant medical concern, affecting millions of individuals globally, including in Indonesia, with a considerable mortality rate. The objective of this study is to enhance the treatment of patients with heart failure by employing the K-Medoids algorithm to categorize patients based on their clinical data. The dataset employed comprises 300 cases of heart failure, with 12 clinical attributes, including age, anemia, ejection fraction, and blood pressure. The K-Medoids algorithm was selected due to its resilience against outliers, which yields more precise clustering outcomes than K-Means. The Elbow Method and Silhouette Score were employed to ascertain the optimal number of clusters. Consequently, three distinct patient clusters were discerned. One cluster was comprised of older patients with hypertension and a low ejection fraction, suggestive of an elevated mortality risk. This study demonstrates the efficacy of the K-Medoids algorithm in clustering heart failure patients, offering insights into patient profiles and facilitating more precise medical interventions. The superiority of K-Medoids in handling outliers presents avenues for further research on risk stratification in healthcare.

Keywords : K-Medoids, Clinical Data, Clustering, Data Mining.

1. PENDAHULUAN

Gagal jantung adalah masalah yang sangat klinis yang sering kali dialami oleh Masyarakat dengan jumlah kasus yang

mencapai jumlah lebih dari 23 juta di seluruh dunia [1]. Gagal jantung sendiri adalah suatu kondisi di mana terjadi kegagalan pada saat jantung tidak mampu memompa darah yang

sesuai untuk memenuhi kebutuhan jaringan pada tubuh. Hal inilah yang menjadi salah satu dari penyebab paling utama kasus kematian di seluruh dunia termasuk di Indonesia, dan terutama pada kalangan yang sudah lanjut usia.

Gagal jantung merupakan suatu kondisi yang diakibatkan karena penyakit kardiovaskular. Dalam penelitian ini dataset yang akan digunakan dan dianalisis pada penelitian kali ini memiliki 13 fitur yang akan digunakan untuk memprediksi mortalitas akibat kondisi gagal jantung [2]. Dataset ini berisi informasi tentang pasien dengan gagal jantung dan mencakup berbagai fitur klinis yang penting untuk memprediksi risiko mortalitas akibat kondisi tersebut. Salah satu fitur utamanya adalah usia pasien, yang berperan signifikan dalam menentukan risiko gagal jantung, karena usia yang lebih tua sering kali dikaitkan dengan risiko yang lebih tinggi. Jenis kelamin juga dicatat, mengingat perbedaan risiko dan gejala gagal jantung antara pria dan wanita. Status anemia adalah fitur penting lainnya, karena anemia dapat memperburuk kondisi gagal jantung dan mempengaruhi prognosis pasien. Gagal jantung sendiri dapat terjadi dan terbagi menjadi empat kategori, kategori yang pertama yaitu disebabkan karena faktor risiko penyakit yang lama diketahui, seperti cedera pada iskemik, hipertensi, dan juga sindrom metabolic. Hal ini menjadi faktor penyebab pasien mengalami gagal jantung. Yang kedua biasanya disebabkan karena faktor genetik, yang biasa disebabkan karena mutasi autosom yang dominan atau kelompok keluarga dengan frekuensi alel yang jarang. Ketiga biasanya disebabkan karena faktor mekanik, hal ini biasanya diakibatkan karena tidak berfungsi nya katup atau tekanan yang berlebihan pada ventrikel kiri pada tubuh orang tua yaitu stenosis aorta. Dan yang keempat biasanya disebabkan karena imunitas, yang mencakup autoimun dan infeksi, baik dari bakteri ataupun dari virus.

Dengan menganalisis fitur-fitur ini menggunakan model machine learning, kita dapat membuat prediksi tentang kemungkinan mortalitas pasien akibat gagal jantung. Model ini dapat membantu dalam mengidentifikasi pasien yang mungkin memerlukan perawatan lebih intensif atau intervensi pencegahan yang lebih efektif. Sebagai contoh, pasien dengan usia lanjut, hipertensi, dan fraksi ejeksi rendah mungkin memiliki risiko lebih tinggi dan memerlukan perhatian lebih lanjut. Data ini memungkinkan kita untuk melakukan intervensi yang lebih tepat waktu dan personal, yang dapat meningkatkan hasil kesehatan pasien dan mengurangi tingkat kematian akibat gagal jantung.

Dalam menangani penyakit gagal jantung dan kondisi klinis lainnya yang hampir sama Pengelompokan pasien berdasarkan data klinis untuk mengembangkan strategi perawatan yang lebih personal dan lebih efektif juga lebih terarah [3]. Salah satu metode yang akan digunakan dalam penelitian ini yaitu data mining dengan metode clustering seperti algoritma K-Medoids dimainkan peran krusialnya.

Clustering memainkan peran penting dalam banyak bidang, seperti pengenalan pola, pengolahan citra serta analisis data. K-medoids adalah algoritma yang populer dan sering digunakan dalam Clustering. K-medoids ini merupakan algoritma metode pengelompokan yang hampir mirip dengan algoritma K-means, tetapi memiliki perbedaan penting dalam cara memilih pusat kelompok (centroid). Biasanya pada K-means, pusat kelompok dihitung sebagai rata-rata dari anggota dalam satu kelompok, sementara pada K-medoids, pusat kelompoknya disebut dengan "medoid", yaitu salah satu titik data aktual yang memiliki jarak terkecil terhadap anggota lain dari kelompok tersebut [2].

K-medoids dianggap lebih efisien untuk digunakan dibandingkan dengan K-

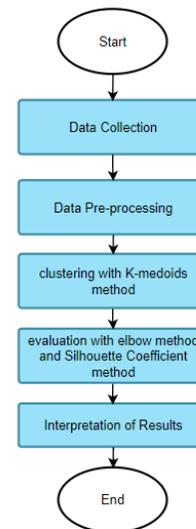
means, karena tidak terlalu sensitif terhadap outlier. Ini disebabkan karena penggunaan medoid yang merupakan data yang nyata, bukan nilai rata-rata, sehingga lebih stabil apabila terdapat data yang sedikit menyimpang. Algoritma ini sering digunakan dalam situasi di mana data mengandung noise ataupun outlier yang signifikan [5].

Walaupun algoritma K-medoids ini memberikan kelebihan dalam menghadapi outlier, tetapi algoritma ini cenderung lebih lambat jika dibandingkan dengan algoritma K-means [6]. Proses pemilihan medoid melibatkan perhitungan jarak antar data secara keseluruhan, untuk meningkatkan kompleksitas komputasi, terutama pada dataset yang besar. Oleh karena itu, berbagai pendekatan dan varian K-medoids telah dikembangkan untuk memperbaiki efisiensi komputasinya. Pada akhirnya, meskipun algoritma K-medoids ini menawarkan beberapa keuntungan penting dibandingkan dengan algoritma K-means, tantangan dalam meningkatkan efisiensi dan mengatasi kompleksitas komputasi menjadi fokus dalam penelitian lebih lanjut. Pendekatan seperti PAM (Partitioning Around Medoids) dan CLARA (Clustering Large Applications) telah dikembangkan untuk memperbaiki kinerja K-medoids pada dataset besar, memungkinkan algoritma ini tetap relevan dan efisien dalam aplikasi modern.

Dalam penelitian ini, akan membahas konsep lebih lanjut mengenai prinsip kerja dari algoritma K-medoids, dan membandingkannya dengan algoritma pengelompokan yang lain, dan juga mengeksplorasi penerapannya pada berbagai bidang, khususnya dalam analisis data [8]. Algoritma ini menjadi relevan di era big data di mana keakuratan dan ketangguhan metode pengelompokan sangat penting untuk mengambil Keputusan berbasis data, terutama di dalam konteks penanganan kondisi Kesehatan seperti gagal jantung.

2. METODE

Penelitian ini dibuat guna menerapkan algoritma K-Medoids dalam mengelompokkan Pasien Gagal Jantung berdasarkan data. Penelitian ini dimulai dengan mencari dataset, analisis eksplorasi data (EDA), pre-processing data, clustering dengan metode K-Medoids dan analisis hasil yang bertujuan untuk mengevaluasi efektivitas algoritma K-Medoids untuk mengelompokkan data pasien gagal jantung berdasarkan data. Langkah – langkah yang akan dilakukan adalah sebagai gambar 1 berikut:



Gambar 1. Alur Penelitian

Pengumpulan Data

Penggunaan dataset dalam penelitian ini diambil dari Kaggle. Kaggle ini menyediakan beragam dataset yang dapat digunakan dalam penelitian yang sangat berkaitan dengan data mining. Dataset yang kami ambil adalah file Heart Failure Prediction. Atribut yang digunakan dalam menentukan pasien gagal jantung mencakup 13 fitur dan jumlah data yang tersedia sebanyak 299.

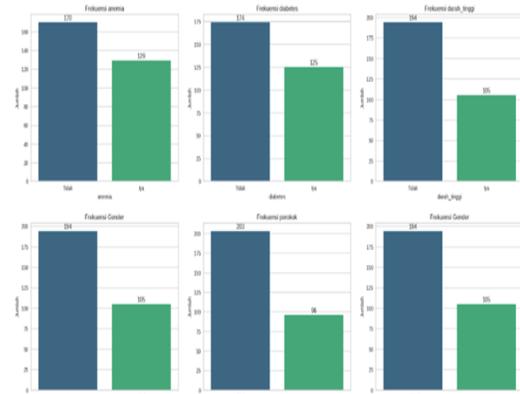
- a. Age : berisi data usia pasien
- b. Anemia : pasien mengalami anemia atau tidak (0 = tidak, 1 = ya)

- c. Creatinine phosphokinase : tingkat enzim CPK dalam darah (mcg/L)
- d. Diabetes : jika pasien mengalami diabetes (0 = tidak, 1 = ya)
- e. Ejection fraction : persentase darah yang keluar dari jantung pada setiap kontraksi (persentase)
- f. High blood pressure : jika pasien menderita hipertensi (0 = tidak, 1 = ya)
- g. Platelets : trombosit dalam darah (kilo platelets/mL)
- h. Serum creatinine : Tingkat kreatinin serum dalam darah (mg/dL)
- i. Serum sodium : kadar natrium serum dalam darah (mEq/L)
- j. Sex : jenis kelamin pasien (0 = Perempuan, 1 = laki-laki)
- k. Smoking : jika pasien perokok (0 = tidak, 1 = ya)
- l. Time : periode tindak lanjut (hari)
- m. DEATH EVENT: Apakah pasien meninggal (0 = tidak, 1 = ya)

Pre-processing Data

Tahap selanjutnya adalah pre-processing terhadap data yang akan digunakan sebelum melakukan clustering. Preprocessing adalah langkah awal dalam pemrosesan data sebelum digunakan dalam analisis atau model machine learning [9]. Tujuannya adalah untuk membersihkan dan mempersiapkan data agar lebih mudah dianalisis. Adapun tahapan-tahapan dalam pre-processing meliputi cleaning data, standarisasi data, reduksi data.

Nama Variabel	Jumlah Data Hilang
Umur	0
anemia	0
kadar_kreatinfosfokinase	0
diabetes	0
Ejectio_Fraction	0
darah_tinggi	0
trombosit	0
kadar_kreatinin	0
kadar_sodium	0
Gender	0
perokok	0
time	0
DEATH_EVENT	0



Gambar 2. Bar Plot Frekuensi Pada Fitur

Tahap ini digunakan untuk memisahkan kolom numerik dan biner dari dataset, kemudian menghitung frekuensi dari setiap variabel biner seperti anemia, diabetes, tekanan darah tinggi, dan status merokok. Setelah itu, frekuensi setiap variabel biner ditampilkan dalam bentuk plot batang. Setiap batang diberi anotasi dengan nilai jumlah di atasnya untuk memudahkan pembacaan. Khusus untuk variabel gender, dilakukan penyesuaian label agar lebih mudah dipahami, seperti mengubah nilai biner menjadi "Laki-laki" dan "Perempuan". Visualisasi ini membantu memahami distribusi data biner. Secara keseluruhan, ini digunakan untuk menganalisis dan memvisualisasikan frekuensi dari variabel biner dalam dataset.

Clustering

Clustering merupakan salah satu teknik dalam algoritma machine learning yang disebut Unsupervised learning. Clustering adalah teknik data mining yang tujuannya untuk mengelompokkan sekumpulan objek data menjadi beberapa cluster, dalam proses ini dibuat objek pada suatu cluster, apakah cluster ini memiliki persamaan atau kemiripan yang tinggi, akan tetapi sangat berbeda dengan objek dalam cluster yang lain[10]. Adapun perbedaan clustering dengan klasifikasi yaitu tidak adanya variabel target dalam melakukan suatu pengelompokan pada proses clustering[8]. Clustering membantu dalam memahami struktur tersembunyi dan pola dalam data yang tidak terlihat secara langsung [9]. Ada beberapa metode clustering yang dipilih sesuai dengan tujuan dan karakteristik data, seperti K-Means, Hierarchical Clustering, DBSCAN dan dalam penelitian ini algoritma yang dipakai yaitu algoritma K-Medoids.

Algoritma K-Medoids

K-medoids merupakan salah satu Teknik yang sangat populer untuk analisis cluster, yang di mana K-medoids menggunakan titik data yang sebenarnya sebagai pusat cluster (medoid). K-Medoids atau yang disebut dengan Partitioning Around Medoids (PAM) ataupun K-Medians merupakan algoritma clustering dengan varian dari algoritma K-Means[10]. Algoritma K-Means dan K-Medoid ini bekerja dengan cara meminimalkan jarak dari titik data yang diberikan label untuk berada di dalam satu klaster dan titik data yang dapat ditetapkan sebagai pusat klaster tersebut[9]. Keuntungan utama dari K-medoids adalah ketahanannya terhadap outlier, karena pusat cluster yang sebenarnya cenderung lebih representatif daripada centroid yang bisa terdistorsi oleh data ekstrem.

Menurut Hardiyanti, Tambunan, & Saragih (2019), adapun langkah-langkah dalam perhitungan K-Medoids diantaranya :

- 1) Inialisasikan pusat cluster sebanyak jumlah cluster (k).
- 2) Alokasikan setiap data (objek) ke cluster terdekat menggunakan persamaan ukuran jarak Euclidian Distance menggunakan persamaan 1 sebagai berikut:

$$d(x_{ij}, c_{kj}) = \sqrt{\sum_{j=i}^p \sum_{i=1}^n (x_{ij}, c_{kj})^2}$$

Dimana:

$d(x_{ij}, c_{kj})$ = Jarak Euclidean Distance antara pengamatan ke-i variabel ke-j ke pusat cluster ke-k pada variabel ke-j,

x_{ij} = objek pada pengamatan ke-i pada variabel ke-j,

c_{kj} = pusat kelompok ke-k pada variabel ke-j,

p = banyaknya variabel yang diamati,

n = banyaknya pengamatan yang diamati

- 3) Pilih secara acak objek pada masing-masing cluster sebagai kandidat medoid baru.
- 4) Hitung jarak setiap objek yang berada pada masing-masing cluster dengan kandidat medoids baru.
- 5) Hitung total simpangan (S) dengan menghitung nilai total distance baru – total distance lama. Jika $S < 0$, maka tukar objek dengan data cluster untuk membentuk sekumpulan k objek baru sebagai medoid.
- 6) Ulangi langkah 3 sampai 5 hingga tidak terjadi perubahan medoids, sehingga didapatkan cluster beserta anggota masing-masing cluster.

Silhouette Coefficient

Silhouette Coefficient adalah ukuran yang digunakan untuk mengevaluasi kualitas pengelompokan dalam sebuah dataset. Salah satu indikator evaluasi adalah koefisien siluet (SC), yang menilai kualitas klaster berdasarkan pengukuran jarak rata-rata antara satu titik data dengan titik data lainnya dalam klaster yang sama (kekompakan), serta jarak rata-rata antara klaster yang berbeda (pemisahan) [12]. Menurut Handrea, dkk,

Silhouette Coefficient mengevaluasi bagaimana sebuah titik data berada dalam clusternya sendiri (cohesion) dibandingkan dengan cluster lain (separation). Nilai SC berkisar antara -1 hingga 1. Nilai mendekati 1 yang menunjukkan bahwa data dikelompokkan dengan baik (cohesion tinggi dan separation jelas), sementara nilai tersebut mendekati -1 yang menunjukkan pengelompokan yang dapat disebut buruk (titik data lebih dekat dengan cluster lain daripada cluster-nya sendiri). Ukuran ini mengukur seberapa baik setiap titik data cocok dengan cluster yang ditugaskan dibandingkan dengan cluster lainnya.

Adapun beberapa tahapan perhitungan Silhouette Coefficient menurut (Wira, Budianto, & Wiguna, 2019) diantaranya :

- 1) Hitung jarak rata-rata dari suatu objek ke-*i* dengan semua objek lain yang berada dalam satu cluster, untuk menghitung jarak rata-rata tersebut dapat dilihat pada persamaan 3 berikut:

$$a(i) = \frac{1}{[A] - 1} \sum_{j \in A, j \neq i} d(i, j)$$

- 2) Hitung rata-rata dari objek ke-*i* tersebut dengan semua objek pada cluster yang lainnya, dan ambil nilai terkecilnya.

$$d(i) = \frac{1}{[A]} \sum_{j \in C} d(i, j)$$

- 3) Nilai Silhouette Coefficient nya adalah: Jumlah *s(i)* diperoleh dengan menggabungkan *a(i)* dan *b(i)*

$$s(i) = \begin{cases} 1 - \frac{a(i)}{b(i)}, & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & \text{if } a(i) > b(i) \end{cases}$$

Sehingga dapat dirumuskan seperti persamaan 5 berikut :

$$s(i) = \frac{b(i) - a(i)}{(a(i), b(i))}$$

Dimana:

s(i)=Nilai Silhouette.

a(i)=Rata – rata jarak antara *i* dengan semua objek pada cluster-nya.

b(i)=Rata – rata jarak antara data *i* pada semua objek pada cluster lain.

d(i,j)= Jarak antara objek *i* dengan *j*.

Tabel 2. Kriteria Silhoutte Coefficient

Silhouette Coefficient	Kriteria Penilaian
$0.7 < SC \leq 1.0$	Strong Structure
$0.5 < SC \leq 0.7$	Medium Structure
$0.25 < SC \leq 0.5$	Wake Structure
$SC \leq 0.25$	No Structure

Selain Elbow Method, metode lain seperti metode Silhouette Coefficient juga digunakan untuk menentukan *K* optimal. Dengan metode ini ditunjukkan bahwa 6 cluster mungkin merupakan jumlah yang optimal berdasarkan data yang digunakan.

Tabel 3. Silhoutte Coefficient Score Dataset

K	Silhouette Coefficient
2	0.366593
3	0.363395
4	0.315541
5	0.295963
6	0.342448
7	0.315293
8	0.312683
9	0.300091

Interpretasi Hasil

Analisis clustering K-Medoids menghasilkan beberapa kelompok data dengan karakteristik unik. Setiap klaster dianalisis untuk memahami ciri-ciri yang membedakan satu dengan lainnya, seperti usia, tekanan darah, kadar natrium, dan ejection fraction. Misalnya, klaster tertentu mungkin menunjukkan pasien yang lebih tua dengan ejection fraction rendah, mengindikasikan kondisi gagal jantung yang parah, sedangkan klaster lain berisi pasien lebih muda dengan kadar natrium tinggi namun lebih mudah dikendalikan secara medis. Visualisasi hasil clustering, seperti scatterplot dan PCA, membantu melihat seberapa baik klaster terpisah, mengidentifikasi outlier, dan memberi insight untuk pengambilan keputusan klinis. Contohnya, klaster dengan pasien lanjut usia dan ejection fraction rendah dapat menjadi

fokus intervensi dini, sementara klaster lain mungkin memerlukan pendekatan berbeda.

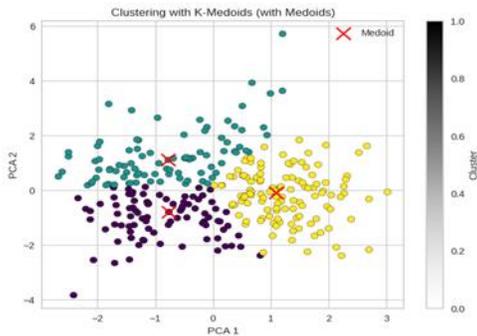
3. HASIL DAN PEMBAHASAN

Penelitian ini bertujuan untuk mengelompokkan data klinis dari pasien gagal jantung. Proses implementasi yang dilakukan pertama yaitu mengimpor dataset yang akan digunakan. Nilai dataset yang digunakan dapat dilihat pada Gambar 3.

	umur	ameta_kadar_ireaktif	forfokinae	diabetes	ejection_fraction	darah_tinggi	trombosit	kadar_ireaktif	kadar_sodium	Gender	perokok	time	DEATH_EVENT	cluster
0	75.0	0	582	0	20	1	265000.00	1.9	130	1	0	4	1	1
1	55.0	0	7881	0	38	0	263358.03	1.1	136	1	0	6	1	2
2	65.0	0	146	0	20	0	162000.00	1.3	129	1	1	7	1	2
3	50.0	1	111	0	20	0	210000.00	1.9	137	1	0	7	1	2
4	65.0	1	180	1	20	0	327000.00	2.7	116	0	0	8	1	1
...
294	62.0	0	61	1	38	1	155000.00	1.1	143	1	1	270	0	2
295	55.0	0	1820	0	38	0	270000.00	1.2	139	0	0	271	0	0
296	45.0	0	2060	1	60	0	742000.00	0.8	138	0	0	278	0	0
297	45.0	0	2413	0	38	0	140000.00	1.4	140	1	1	280	0	2
298	50.0	0	196	0	45	0	385000.00	1.6	136	1	1	285	0	2

Gambar 3. Tabel Dataset

Dalam pengolahan clustering pada data klinis pasien gagal jantung menggunakan Python, langkah pertama setelah menyiapkan dataset adalah mengimpor data. Gambar 4 memperlihatkan hasil visualisasi proses clustering dengan algoritma K-Medoids. Teknik reduksi dimensi PCA (Principal Component Analysis) digunakan untuk memproyeksikan data berdimensi tinggi ke dalam ruang dua dimensi (PCA 1 dan PCA 2) agar lebih mudah divisualisasikan.



Gambar 4. Visualisasi PCA

a) Tiga Cluster Distink: Data terbagi menjadi tiga kelompok yang jelas, ditandai dengan warna yang berbeda. Hal ini mengindikasikan bahwa algoritma

berhasil mengidentifikasi tiga cluster utama dalam data.

- b) Medoid sebagai Pusat Cluster: Tanda silang merah mewakili medoid dari setiap cluster. Medoid adalah titik data dalam cluster yang meminimalkan jumlah total disimilaritas ke semua titik data lainnya dalam cluster tersebut. Posisi medoid ini mengindikasikan pusat atau representasi dari masing-masing cluster.
- c) Distribusi Data: Data dalam setiap cluster terlihat cukup terdistribusi dengan baik di sekitar medoidnya. Ini menunjukkan bahwa pembagian data menjadi tiga cluster cukup masuk akal dan mewakili struktur yang sebenarnya dalam data.
- d) Variasi Densitas: Densitas data dalam setiap cluster terlihat berbeda. Ada cluster yang lebih padat dan ada yang lebih renggang. Ini mengindikasikan bahwa karakteristik data dalam setiap cluster mungkin berbeda.

Pada Tabel 4 distribusi jumlah data dalam masing-masing klaster serta jumlah data yang terkait dengan beberapa fitur numerik spesifik untuk setiap klaster. Terdapat tiga klaster yang dihasilkan dari proses clustering, yaitu klaster 0, klaster 1, dan klaster 2. Klaster 0 dan klaster 1 masing-masing memiliki 89 data, sedangkan klaster 2 memiliki 121 data. Hal ini mencerminkan bahwa jumlah data terbanyak terdapat di klaster 2, sementara klaster 0 dan 1 memiliki jumlah data yang sama. [3]

Tabel 4. Jumlah Data di Setiap Klaster

Klaster	Jumlah Data
1	89
2	89
3	121

Pada Gambar 5, terlihat bahwa data berhasil dikelompokkan menjadi tiga klaster: 0, 1, dan 2. Setiap baris dalam tabel mewakili satu klaster, dan perbandingan nilai statistik antara klaster membantu mengidentifikasi perbedaan karakteristik. Misalnya, jika rata-rata usia pada klaster 1 lebih tinggi, maka

kluster tersebut cenderung terdiri dari individu yang lebih tua. [11]

The image contains three small data tables. The first table shows statistics for 'umur' (age) across three clusters. The second table shows statistics for 'diabetes', 'Ejection_fraction', and 'kadar_kreatinin'. The third table shows statistics for 'kadar_sodium', 'Gender', and 'perokok'. Each table lists cluster 0, 1, and 2 with their respective mean and sum values.

Gambar 5. Mengelompokkan Data Di Cluster Berbeda

Dapat kita lihat bahwa, Kluster 0: Memiliki rata-rata kadar kreatinin yang lebih rendah dan rata-rata usia yang lebih muda dibandingkan dengan kluster lainnya. Kluster 1: Memiliki rata-rata kadar kreatinin yang lebih tinggi dan persentase pasien dengan penyakit jantung (Ejection Fraction rendah) yang lebih tinggi. Kluster 2: Memiliki rata-rata kadar sodium yang lebih tinggi dan persentase pasien perokok yang lebih tinggi.

Dari Keterangan diatas, dapat disimpulkan bahwa:

- a) Kluster 0: Menunjukkan profil pasien yang umumnya lebih muda dengan kadar kreatinin yang lebih rendah, mengindikasikan fungsi ginjal yang lebih baik dibandingkan dengan kluster lainnya. Hal ini mungkin mencerminkan kelompok pasien yang relatif lebih sehat dengan risiko gagal jantung yang lebih rendah.
- b) Kluster 1: mewakili kelompok pasien dengan risiko penyakit ginjal kronis yang lebih tinggi, yang juga berhubungan erat dengan peningkatan risiko gagal jantung. Kelompok ini memiliki kadar kreatinin yang tinggi dan fraksi ejeksi yang rendah, menunjukkan bahwa mereka mungkin memerlukan intervensi medis yang lebih intensif untuk mengelola kedua kondisi ini secara bersamaan.
- c) Kluster 2: Terdiri dari pasien dengan risiko tinggi untuk penyakit

kardiovaskular serius, ditandai dengan kadar sodium yang tinggi dan prevalensi merokok yang lebih besar. Kondisi ini dapat meningkatkan risiko tekanan darah tinggi dan komplikasi kardiovaskular, yang memerlukan pendekatan perawatan yang khusus untuk mengurangi risiko komplikasi lebih lanjut.

Hasil analisis yang ditunjukkan pada tabel 5 adanya perbedaan yang signifikan dalam distribusi karakteristik antara ketiga kluster. Hal ini mengindikasikan bahwa proses clustering telah berhasil mengidentifikasi sub kelompok individu dengan profil yang berbeda.

Kluster 0 memiliki jumlah individu dengan kondisi medis tertentu seperti diabetes dan darah tinggi yang relatif tinggi dibandingkan dengan kluster lainnya. Hal ini mengindikasikan bahwa kluster 0 mungkin mewakili kelompok individu dengan risiko kesehatan yang lebih tinggi[12]. Sebaliknya, kluster 2 memiliki jumlah perokok yang jauh lebih tinggi dibandingkan dengan kluster lainnya, yang menunjukkan bahwa kluster ini mungkin mewakili kelompok individu dengan gaya hidup yang berbeda. Kluster 1 memiliki proporsi individu dengan anemia yang lebih tinggi dibandingkan Kluster 2, namun lebih rendah dibandingkan Kluster 0. Ini menunjukkan bahwa kondisi anemia mungkin menjadi faktor risiko tambahan pada Kluster 1 dan memiliki jumlah kematian yang cukup tinggi, sebanding dengan Kluster 0. Hal ini mengindikasikan bahwa kedua kluster ini mungkin memiliki profil risiko kematian yang serupa[13].

Tabel 5. Jumlah Data Bernilai 1 Pada Beberapa

Kluster	Gender	Anemia	Diabetes	Darah_tinggi	Perokok	Death_EVENT
0	36	30	58	23	2	9
1	38	65	37	53	9	50
2	120	34	30	29	85	37

Selanjutnya gambar 6 memberikan gambaran yang lebih lengkap tentang distribusi data pada setiap kluster. Dengan

adanya informasi jumlah data dengan nilai 0, kita dapat membandingkan proporsi individu dengan dan tanpa kondisi tertentu pada setiap klaster. Ini memberikan gambaran yang lebih komprehensif tentang karakteristik unik setiap klaster [14]. Dengan membandingkan proporsi antara klaster, kita dapat mengidentifikasi pola tertentu. Misalnya, jika suatu kondisi medis lebih sering ditemukan pada satu klaster dibandingkan dengan klaster lainnya, ini dapat mengindikasikan adanya faktor risiko yang unik pada klaster tersebut.

cluster	Gender	anemia	diabetes	darah_tinggi	perokok	DEATH_EVENT	total \
0	36	30	58	23	2	9	89
1	38	65	37	53	9	50	89
2	120	34	30	29	85	37	121

cluster	Gender_0	anemia_0	diabetes_0	darah_tinggi_0	perokok_0 \
0	53	59	31	66	87
1	51	24	52	36	80
2	1	87	91	92	36

cluster	DEATH_EVENT_0
0	80
1	39
2	84

Gambar 6. Menghitung Jumlah Data Total di Setiap Klaster

Visualisasi pada gambar 6 juga secara jelas menunjukkan adanya perbedaan yang signifikan dalam distribusi fitur antara ketiga klaster. Setiap klaster memiliki profil yang unik berdasarkan distribusi fitur-fiturnya. Klaster 0, misalnya, memiliki proporsi penderita diabetes dan darah tinggi yang relatif tinggi dibandingkan dengan klaster lainnya. Hal ini mengindikasikan bahwa klaster 0 mungkin mewakili kelompok individu dengan risiko penyakit kronis yang lebih tinggi[15]. Sebaliknya, Klaster 2 memiliki proporsi perokok yang jauh lebih tinggi dibandingkan dengan klaster lainnya, yang menunjukkan bahwa klaster ini mungkin mewakili kelompok individu dengan gaya hidup yang berbeda.

Analisis distribusi fitur pada setiap klaster mengungkapkan adanya heterogenitas yang signifikan dalam populasi yang diteliti. Hasil visualisasi menunjukkan bahwa clustering telah berhasil mengidentifikasi sub-kelompok individu dengan karakteristik yang berbeda.

Interpretasi Setiap Fitur :

- Gender: Klaster 2 memiliki proporsi individu laki-laki yang jauh lebih tinggi dibandingkan dengan klaster lainnya.
- Anemia: Klaster 1 memiliki proporsi individu dengan anemia yang lebih tinggi dibandingkan dengan klaster lainnya.
- Diabetes: Klaster 0 memiliki proporsi penderita diabetes yang paling tinggi.
- Darah Tinggi: Klaster 2 memiliki proporsi individu dengan darah tinggi yang relatif rendah.
- Perokok: Klaster 2 memiliki proporsi perokok yang jauh lebih tinggi dibandingkan dengan klaster lainnya.
- DEATH_EVENT: Klaster 1 memiliki proporsi kematian yang lebih tinggi dibandingkan dengan klaster lainnya.

Analisis visualisasi menunjukkan bahwa Klaster 2 memiliki profil risiko yang unik, ditandai dengan proporsi perokok yang tinggi dan kejadian kematian yang lebih sering. Hal ini mengindikasikan bahwa gaya hidup mungkin menjadi faktor risiko utama pada kelompok ini. Sebaliknya, Klaster 0 memiliki proporsi penderita diabetes dan darah tinggi yang tinggi, yang menunjukkan adanya komorbiditas yang signifikan pada kelompok ini.

4. KESIMPULAN

Penelitian ini berhasil menerapkan algoritma K-Medoids untuk mengelompokkan pasien gagal jantung berdasarkan data klinis mereka, yang mengungkapkan kelompok pasien dengan faktor risiko yang berbeda-beda. Hasil penelitian menunjukkan bahwa K-Medoids efektif dalam mengidentifikasi pasien dengan risiko kematian yang lebih tinggi, terutama mereka yang berusia lanjut, memiliki hipertensi, dan fraksi ejeksi yang rendah. Ketahanan algoritma K-Medoids terhadap outlier menjadikannya cocok untuk menganalisis dataset klinis yang sering mengandung data yang tidak teratur. Temuan ini menunjukkan potensi penggunaan

pendekatan berbasis data seperti clustering untuk mendukung strategi perawatan yang lebih personal bagi pasien gagal jantung. Dengan mengelompokkan pasien secara akurat berdasarkan profil klinis mereka, tenaga medis dapat memberikan perawatan yang lebih terarah, yang berpotensi meningkatkan hasil kesehatan pasien. Penelitian lebih lanjut perlu difokuskan pada peningkatan efisiensi komputasi K-Medoids untuk dataset yang lebih besar dan eksplorasi penerapannya pada kondisi medis lainnya.

5. DAFTAR PUSTAKA

- A. P. Lumi, V. F. F. Joseph, dan N. C. I. Polii, "Rehabilitasi Jantung pada Pasien Gagal Jantung Kronik," *Jurnal Biomedik*, vol. 13, no. 3, hlm. 309–316, 2021, doi: 10.35790/jbm.13.3.2021.33448.
- S. Bahri, D. Marisa Midyanti, dan P. Korespondensi, "PENERAPAN METODE K-MEDOIDS UNTUK PENGELOMPOKAN MAHASISWA BERPOTENSI DROP OUT APPLICATION OF K-MEDOIDS METHOD FOR DROPOUT POTENTIAL STUDENT GROUPING," vol. 10, no. 1, hlm. 165–172, 2023, doi: 10.25126/jtiik.2023106643.
- M. Arifandi, A. Hermawan, dan D. Avianto, "IMPLEMENTASI ALGORITMA K-MEDOIDS UNTUK CLUSTERING WILAYAH TERINFEKSI KASUS COVID19 DI DKI JAKARTA," *Jurnal Teknologi Terapan* |, vol. 7, no. 2, 2021.
- N. Gusmantoni, "Application of Data Mining Clustering the Development of Covid-19 Using K-Medoids Method," *Journal of Computer Science and Information Technology*, hlm. 22–26, Jan 2022, doi: 10.35134/jcsitech.v8i1.29.
- M. Hafiih, M. W. Prasetyo, dan D. Dwi Prasetya, "ANALISIS PENGGUNAAN LISTRIK PLN DI JAWA TIMUR MENGGUNAKAN ALGORITMA K-MEDOIDS," *Jurnal Riset dan Aplikasi Mahasiswa Informatika (JRAMI)*, vol. 05, 2024.
- L. Penelitian dan Penerbitan Hasil Penelitian Ensiklopedia dan J. Oliando, "PENGELOMPOKAN BERITA KESEHATAN PADA SOSIAL MEDIA TWITTER DENGAN METODE K-MEANS CLUSTERING," *Ensiklopedia of Journal*, vol. 4, 2022, [Daring]. Tersedia pada: <http://jurnal.ensiklopediaku.org>
- N. P. Dharshinni dan C. Fandi, "Penerapan Metode K-Medoids Clustering Untuk Mengelompokkan Ketahanan Pangan," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 6, no. 4, hlm. 2301, Okt 2022, doi: 10.30865/mib.v6i4.4939.
- Sukma Sindi, W. R. (2020). *ANALISIS ALGORITMA K-MEDOIDS CLUSTERING DALAM PENGELOMPOKAN PENYEBARAN COVID-19 DI INDONESIA*. *Jurnal Teknologi Informasi*, 166-173.
- B. Karthikeyan, D. J. George, G. Manikandan, and T. Thomas, "A Comparative Study on K-Means Clustering and Agglomerative Hierarchical Clustering," *International Journal of Emerging Trends in Engineering Research*, vol. 8, no. 5, 2020, doi: 10.30534/ijeter/2020/20852020.
- K. Pola et al., "Clustering Sales Patterns of Best Selling and Less Selling Products at El Jhon Bengkulu Stores Using the K-Medoid Method," *JURNAL KOMITEK*, vol. 2, no. 2, pp. 637–642, 2022, doi: 10.53697/jkomitek.v2i2.
- Siti Nurlaela, A. . *ALGORITMA K-MEDOIDS UNTUK CLUSTERING PENYAKIT MAAG DI KABUPATEN KARAWANG*. *Jurnal Informatika, Manajemen dan Komputer*, vol 12, no. 2, hal 56-62. 2020.
- Y. H. Susanti and E. Widodo, "Perbandingan K-Means dan K-Medoids Clustering terhadap Kelayakan Puskesmas di DIY Tahun 2015, Prosiding SI MaNIs (Seminar Nasional Integrasi Matematika dan Nilai Islami), vol. 1, no. 1, pp. 116122, 2017.
- Nurafni Syahfitri, E. B. Pengelompokan Produk Berdasarkan Data Persediaan Barang Menggunakan Metode Elbow dan K-Medoid . *KLIK: Kajian Ilmiah Informatika dan Komputer*, 1668-1675.2023
- N. Adawiyah, N. Sulistiyowati, and M. Jajuli, "Klasterisasi Kasus Kekerasan Terhadap Anak dan Perempuan Berdasarkan

Algoritma K-Means,” Generation Journal, vol. 5, no. 2, 2021.

- D. Ramdhan, G. Dwilestari, R. Danar Dana, and A. Ajiz, “Clustering Data Persediaan Barang dengan Menggunakan Metode K-Means. Clustering Data Persediaan Barang dengan Menggunakan Metode K-Means,” MEANS (Media Informasi Analisa dan Sistem, vol. 7, no. 1, 2022, [Online]. Available:
http://ejournal.ust.ac.id/index.php/Jurnal_Means/