JISKA (Jurnal Informatika Sunan Kalijaga)
Vol. 9, No. 3, SEPTEMBER, 2024, Pp. 205 – 216
ISSN: 2527 – 5836 (print) | 2528 – 0074 (online)

Implementasi K-Means *Clustering* pada Pengelompokan Pasien Penyakit Jantung

Jihan Wala [(1)*], Herman [(2)], Rusydi Umar [(3)]
Magister Informatika, Fakultas Teknologi Industri, Universitas Ahmad Dahlan, Yogyakarta, Indonesia
e-mail : 2307048013@webmail.uad.ac.id, {hermankaha,rusydi}@mti.uad.ac.id.
* Penulis korespondensi.
Artikel ini diajukan 15 April 2024, direvisi 25 Juni 2024, diterima 26 Juni 2024, dan dipublikasikan 25 September 2024.

Abstract

Heart disease is a prominent global health concern, necessitating early identification and patient grouping for effective management. This study employs the K-Means clustering algorithm with a medical dataset of 303 patients, encompassing various attributes. These include Age, Gender, Chest Pain Type, Blood Pressure, Serum Cholesterol Level, Fasting Blood Sugar, Resting Electrocardiographic Results, Maximum Heart Rate, Angina, ST Depression, and Slope of the ST Segment. The goal is to categorize patients into four clusters based on chest pain types, a crucial symptom indicating disease severity. The computation concludes after the sixth iteration, revealing Cluster 1 (27 patients), Cluster 2 (135 patients), Cluster 3 (15 patients), and Cluster 4 (126 patients). Collaborative analysis with medical experts highlights that Cluster 1, mainly comprising older males, exhibits high-risk indicators. While this grouping aids in personalized treatment strategy development, further clinical validation involving more experts and datasets is imperative for enhanced reliability.

Keywords: Implementation, K-Means, Clustering, Grouping, Heart Disease

Abstrak

Penyakit jantung menjadi permasalahan kesehatan serius diseluruh dunia. Pendeteksian dini dan pengelompokan pasien berdasarkan ciri-ciri khusus dapat mendukung manajemen penanganan penyakit jantung. Penelitian ini mengusulkan algoritma K-Means *clustering* untuk mengelompokkan pasien penyakit jantung dengan *dataset* medis sebanyak 303 pasien. *Dataset* mencakup atribut Umur, Jenis Kelamin, Jenis Nyeri Dada, Tekanan Darah, Kadar Serum Kolesterol, Gula Darah, Hasil Elektrokardiografi, Denyut Jantung Maksimum, Angina, Depresi ST, dan Kemiringan Segmen ST. Tujuan penelitian ini adalah mengelompokkan pasien penyakit jantung berdasarkan tingkat keparahan atau kegawatdaruratan pasien menggunakan algoritma K-Means *clustering*. Wawancara bersama ahli medis untuk pembagian kelompok menjadi empat *cluster* berdasarkan jenis nyeri dada yang merupakan gejala utama tingkat keparahan penyakit jantung. Interpretasi menghasilkan 5 *cluster* dengan *cluster k*1 berjumlah 27 pasien, *k*2 berjumlah 135 pasien, *k*3 berjumlah 15 pasien, dan *k*4 berjumlah 126 pasien. Analisis data menunjukan, *cluster* 1 (*k*1), cenderung terdiri dari pasien yang lebih tua, mayoritas laki-laki, menunjukan risiko tinggi dengan gejala nyeri dada parah, tekanan darah, dan kadar kolesterol tinggi. Sementara itu, *cluster k*2, *k*3, dan *k*4 menunjukkan risiko lebih rendah, dengan variasi respons terhadap aktivitas fisik. Pengelompokan ini memberikan dukungan kepada dokter dan peneliti dalam memahami pola penyakit jantung serta merancang strategi pengobatan yang lebih spesifik dan personal.

Kata Kunci: Implementasi, K-Means, *Clustering*, Pengelompokkan, Penyakit Jantung

1. PENDAHULUAN

Penyakit jantung merupakan penyebab utama kematian di seluruh dunia. Menurut World Health Organization tahun 2020 terdapat 17,9 juta kematian dan 80% disebabkan oleh penyakit arteri koroner dan *stroke* serebral (Ali et al., 2021). Jumlah kematian yang besar ini umum terjadi di negara-negara berpenghasilan rendah dan menengah (Shah et al., 2020). Penyakit jantung dapat disebabkan oleh berbagai faktor yang berkaitan dengan kebiasaan hidup, seperti merokok, penggunaan alkohol dan kafein secara berlebihan, stres, aktifitas fisik yang kurang. Sebab lain

adalah faktor-faktor fisiologis (obesitas, hipertensi, kolesterol, darah tinggi, dan kondisi jantung). Identifikasi pasien penyakit jantung dilakukan dengan melihat atribut terkait data pasien yang memiliki signifikansi besar untuk membantu dokter memberikan perawatan yang lebih terfokus (Singh & Kumar, 2020). Salah satu teknik yang dapat membantu dokter dalam mengidentifikasi dan mengelompokan pasien penyakit jantung adalah teknik *data mining*.

Clustering adalah teknik yang populer digunakan pada *data mining.* Teknik ini merupakan proses pengelompokan data menjadi beberapa *cluster* data berdasarkan kemiripan atribut-atribut yang dimiliki data (Haris Kurniawan et al., 2020). Pada penelitian ini, digunakan algoritma K-Means *clustering*. K-Means adalah teknik pengelompokan data di mana atribut data dikelompokkan ke dalam partisi set data, kemudian ditetapkan ke dalam kelompok yang berbeda (Ikotun et al., 2023). Penelitian ini relavan dengan beberapa penelitian sebelumnya dalam penerapan K-Means *clustering* sebagai sumber referensi dan perbandingan hasil penelitian.

Penelitian Ariefandi et al., menggunakan *k-medoids clustering* untuk klasterisasi wilayah terinfeksi kasus *covid*-19 di DKI Jakarta. Penelitian ini menghasilkan 3 *cluster* dengan kasus yang paling tertinggi yakni *cluster* 0 terdiri dari 31 kelurahan sedangkan *cluster* paling rendah diketahui *cluster* 2 terdiri 66 kelurahan (Arifandi et al., 2021). Kemudian penelitian Purba et al., menggunakan K-Means *clustering* untuk mengelompokkan penyebab penyakit ISPA. Menghasilkan 2 *cluster*, di mana *cluster* 1 memberikan rekomendasi tinggi berjumlah 10 Kabupaten, *cluster* 2 memberikan rekomendasi rendah berjumlah 2 Kabupaten (Purba et al., 2021). Solechati & Jananto mengelompokan *profile* pasien. Dengan interpretasi menghasilkan 5 *cluster* dengan *cluster* 1 memiliki 228 *record*, pada *cluster* 2 memiliki 248 *record*, *cluster* 3 memiliki 1551 *record*, *cluster* 4 memiliki 2592 *record*, dan *cluster* 5 memiliki 362 *record* (Solechati & Jananto, 2023). Mashita et al., melakukan klasifikasi pada pasien penyakit jantung. penelitian ini menghasilkan akurasi yang diperoleh adalah *k*=7 dan *k*=9, yang merupakan hasil paling optimal karena memiliki akurasi tertinggi dibandingkan dengan nilai *k* lainnya, dengan akurasi sebesar 88% (Masitha et al., 2023). Novidianto el al., menggunakan Metode *k-prototypes cluster mix algorithm* untuk mengidentifikasi faktor kematian pada pasien gagal jantung. Hasil klasterisasi membentuk 2 *cluster* yang dianggap optimal berdasarkan nilai koefisien *silhouette* tertinggi sebesar 0,5777. Analisis hasil menunjukan bahwa *cluster* 1 adalah *cluster* pasien yang memiliki risiko rendah terhadap kemungkinan kematian akibat gagal jantung dan *cluster* 2 adalah *cluster* pasien dengan risiko tinggi terhadap kematian akibat gagal jantung (Novidianto et al., 2021)

Berdasarkan kajian literatur penelitian terdahulu di atas, maka penelitian ini dilakukan dengan tujuan untuk mengelompokkan pasien penyakit jantung berdasarkann keparahan atau kegawatdaruratan pasien menggunakan pendekatan K-Means *clustering*. Harapan dilakukan penelitian ini dapat memberikan wawasan alaternatif dalam pengelompokan pasien penyakit jantung dan berpotensi menjadi landasan untuk pengembangan strategi pengobatan yang lebih efektif di masa depan. Kontribusi penelitian ini terletak pada pengembangan strategi pengobatan yang lebih spesifik dan personal.

2. METODE PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini melibatkan serangkaian langkah yang penting untuk mempersiapkan dan merencanakan studi secara menyeluruh. Lima langkah tahapan yang dilakukan dalam penelitian ini adalah studi pustaka, pengumpulan data, implementasi K-Means *clustering*, dan analisis hasil *cluster*. Adapun tahapan penelitian secara lengkap dapat dilihat pada Gambar 1.

Gambar 1 Tahapan Penelitian

2.1.1 Studi Pustaka

Langkah awal penting dalam melakukan penelitian pendahuluan adalah melakukan studi pustaka yang komprehens. Ini melibatkan peninjauan yang luas terhadap literatur yang berkaitan dengan topik penelitian yang akan dilakukan (Muslimah, 2024). Pada tahap ini dilakukan pengumpulan, peninjauan, dan analisis berbagai sumber informasi seperti jurnal ilmiah, buku, artikel, dan publikasi terkait lainnya. Studi pustaka bertujuan untuk menghimpun, menelaah, dan menganalisis literatur terkait yang relevan dengan topik penelitian yang sedang diselidiki. Melalui proses ini, dapat memperoleh pemahaman yang mendalam tentang status terkini dari penelitian yang sudah ada, mengidentifikasi pengetahuan yang telah dikembangkan, serta menemukan area-area di mana pengetahuan masih terbatas dan memerlukan penelitian lebih lanjut.

2.1.2 Pengumpulan Data

Tabel 1 Deskripsi *Dataset Heart Disease* (Penyakit Jantung)

Atribut	Keterangan	Penjelasan
Id	Id Pasien	Kode pasien dari 1-303
Age	umur pasien (tahun)	Minimal = 29, Maksimal = 77
Sex	Jenis kelamin pasien	1 = laki-laki, 0 = perempuan
Cp	Jenis nyeri dada	*Cp* (*Chest pain*) yaitu tipe nyeri dada yang diderita pasien. Atribut ini memiliki 4 nilai yaitu: Nilai 1: tidak nyeri dada (*no chest pain*) Nilai 2: nyeri dada ringan (*mild chest pain*) Nilai 3: nyeri dada sedang (*moderate chest pain*) Nilai 4: nyeri dada parah (*severe chest pain*)
Trestbps	Tekanan darah istirahat (mm Hg)	*Trestbps* (*Resting blood pressure*) yaitu tekanan darah pasien ketika dalam keadaan istirahat. Rendah < 120, normal = 120, tinggi > 120
Chol	Serum kolesterol (mg/dl)	*Chol* (*Cholesterol*) yaitu kadar kolesterol dalam darah pasien. Rendah < 140, normal = 140, tinggi > 140
Fbs	Gula darah puasa > 120 mg/dl	*Fbs (Fasting blood sugar)* yaitu kadar gula darah pasien, atribut fbs ini memiliki 2 nilai yaitu 1 jika kadar gula darah pasien melebihi 120 mg/dl, dan 0 jika tidak melebihi atau sama dengan 120 mg/dl.
Restecg	Hasil elektrokardiografi istirahat	*Resting electrocardiographic* memiliki 3 nilai yaitu nilai 0 = normal, nilai 1 = ST-T *wave abnormality* nilai 2 = ventricular kiri mengalami hipertrop
Thalach	Denyut jantung maksimum	Tingkat detak jantung maksimum yang dicapai. Jika nilai “thalac” semakin tinggi dapat dianggap sebagai tanda risiko yang lebih tinggi untuk penyakit jantung
Exang	Angina yang dipicu oleh Latihan	*Exang* (*Exercise-induced angina*) keadaan dimana pasien akan mengalami nyeri dada apabila berolah raga, 0 = tidak nyeri, dan 1 = menyebabkan nyeri.
Oldpeak	Depresi ST	Depresi ST yang diinduksi oleh latihan relatif terhadap istirahat. Penurunan ST akibat olahraga. Nilai "*Oldpeak*" yang tinggi dapat dianggap sebagai tanda risiko yang lebih tinggi untuk penyakit jantung
Slope	Kemiringan segmen ST latihan puncak.	*Slope* dari puncak ST setelah berolah raga. Atribut ini memiliki 3 nilai yaitu 0 untuk *downsloping*, 1 untuk flat, dan 2 untuk *upsloping*

Data yang digunakan merupakan data sekunder yang di ambil dari internet situs Kaggle, *dataset* Penyakit Jantung (*heart disease*) oleh Awan (2020) sebagai objek penelitian. Penjelasan lebih

spesifik *dataset* penyakit jantung dilakukan bersama ahli medis (Dokter Spesialis Jantung dan Pembuluh Darah) dan sumber referensi dari beberapa artikel jurnal. Berikut pada Tabel 1 penjelasan setiap atribut *dataset* (Ali et al., 2021; Shah et al., 2020; Singh & Kumar, 2020; V. Ramalingam et al., 2018).

2.1.3 *Preprocessing*

Setelah melakukan pengumpulan data, tahap selanjutnya adalah *pre-processing* data. Tahap ini meliputi proses pengecekan *missing value* dan normalisasi data. *Missing value* mengindikasikan ketiadaan informasi untuk suatu variabel pada observasi tertentu. Pentingnya pengecekan *missing value* dalam analisis data karena hal tersebut dapat membantu mencegah adanya bias dalam penarikan kesimpulan (Han & Kang, 2023). Normalisasi bertujuan untuk membuat skala variabel dalam *dataset* menjadi seragam, sehingga setiap variabel memiliki kontribusi yang seimbang dalam analisis (Mishra et al., 2020). Metode normalisasi yang akan dilakukan pada penelitian ini yaitu *feature scaling*. *Feature scaling* dilakukan dengan tujuan untuk membandingkan atau mengintegrasikan data dari berbagai sumber atau variabel yang memiliki rentang nilai yang berbeda-beda. *Feature scaling* mengubah nilai-nilai yang diperkirakan ke dalam rentang yang lebih kecil atau seragam memiliki skala, Di mana nilai-nilai diperkirakan dikonversi ke rentang antara 0 sampai 1. Normalisasi *feature scaling* dapat dilakukan menggunakan Pers. (1) (Sun & Yu, 2021). Dalam rumus *feature scaling*, X_{baru} adalah nilai atribut baru setelah normalisasi, X_{awal} adalah nilai atribut asli yang akan dinormalisasi, dan X_{max} adalah nilai maksimum dari semua data pada atribut yang sama.

$$X_{baru} = \frac{X_{awal}}{X_{max}} \quad (1)$$

2.1.4 Penerapan K-Means *Clustering*

Proses selanjutnya adalah penerapan algoritma K-Means *clustering*. K-Means merupakan salah satu teknik pengelompokan yang paling *prominent* dalam ilmu dan teknologi (Das et al., 2023). Tujuan utama dari K-Means *clustering* adalah untuk membagi *dataset* menjadi kelompok-kelompok yang homogen, di mana setiap kelompok memiliki kesamaan internal yang tinggi dan perbedaan yang signifikan antar kelompok (Qi et al., 2023). *Flowchart* K-Means *clustering* dapat dilihat pada Gambar 2 (Rizki et al., 2020).

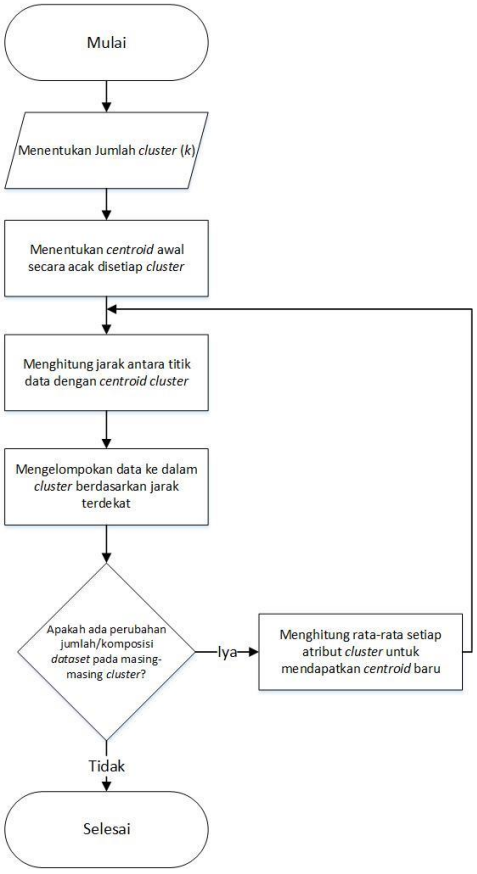
Berikut penjelasan tahapan *flowchart* K-Means *clustering* pada Gambar 2:

1) Menentukan jumlah *cluster* (*k*) merupakan tahap pertama dalam penentuan jumlah *cluster* yang optimal untuk data yang akan dikelompokan.
2) Menentukan *centroid* awal secara acak di setiap *cluster*. Pemilihan awal *centroid* dilakukan dengan cara mengambil secara acak titik data yang ada dalam *dataset*. Titik data yang akan menjadi *centroid* awal dipilih tanpa mempertimbangkan distribusi atau karakteristik khusus dari data. Secara praktis, setiap titik dalam *dataset* memiliki kesempatan yang sama untuk dipilih sebagai *centroid* awal.
3) Menghitung jarak antara titik data dengan setiap *centroid cluster* yaitu proses dalam K-Means *clustering* di mana jarak antara setiap titik data dengan setiap *centroid cluster* dihitung, menghitung jarak digunakan rumus Euclidean *distance*. Untuk menghitung jarak antara data *x* baris ke-*i* (*i*=1,2,3,…*n*), data *c* baris ke-*h* (*h*=1,2,3,…*k*) yang disimbolkan $d\ (x_i, c_h)$, dengan *n* merupakan jumlah total baris data, *m* adalah jumlah atribut, dan *k* adalah jumlah *cluster*. Rumus jarak $d\ (x_i, c_h)$ ditampilkan pada Pers. (2). Di mana x_{ij} adalah atribut ke *j* dari data ke *i* dan c_{hj} adalah atribut ke *j* dari *cluster h*. Jarak terkecil dari data ke *i* ke *cluster h* menunjukan bahwa data ke *i* masuk dalam *cluster h.* Jika jarak dari data ke 5 paling kecil adalah dengan *cluster* 3, maka data 5 dikelompokkan dalam *cluster* 3.

$$d\ (x_i, c_h) = \sqrt{\sum_{j=1}^{m}\sum_{h=1}^{k}(x_{ij} - c_{hj})^2} \quad (2)$$

$$d\ (x_i, c_h) = \sqrt{(x_{i1} - c_{h1})^2 + (x_{i2} - c_{h2})^2 + (x_{i3} - c_{h3})^2 + \ldots + (x_{im} - c_{hm})^2}$$

4) Kemudian memeriksa apakah terjadi perubahan dari *centroid* baru terhadap *centroid* sebelumnya setelah pengelompokan data ke dalam *cluster*. Jika terjadi perubahan pada nilai *centroid*, maka menunjukan bahwa proses masih berjalan dan pengelompokan data harus terus dilakukan pada iterasi berikutnya.
5) Tahap terakhir yaitu jika terjadi perubahan pada nilai *centroid*, maka lanjut ke tahap selanjutnya yaitu menghitung nilai rata-ratanya untuk menghasilkan *centroid* baru pada *cluster* tersebut. Kemudian ulangi langkah 3, dan 4 pada iterasi berikutnya sampai tidak ada perubahan lagi pada *centroid* setiap *cluster*. Jika tidak terjadi perubahan pada *centroid* maka proses *clustering* dinyatakan selesai.



Gambar 2 *Flowchart* K-Means *Clustering*

2.1.5 Analisis Hasil *Cluster*

Pada tahap analisis *cluster*, dilakukan pemilihan kelompok yang diprioritaskan untuk penanganan dalam pengobatan penyakit jantung. Pemilihan ini didasarkan pada hasil wawancara dengan ahli medis, yang memberikan wawasan terkait kelompok pasien dengan tingkat risiko tertinggi. Dengan menggunakan teknik *cluster*ing, data pasien dibagi ke dalam beberapa kelompok berdasarkan karakteristik medis mereka, seperti usia, riwayat penyakit, dan faktor risiko lainnya. Kelompok yang diidentifikasi sebagai prioritas merupakan fokus utama dalam pemberian

intervensi medis untuk meningkatkan efektivitas pengobatan dan mengurangi risiko komplikasi penyakit jantung.

3. HASIL DAN PEMBAHASAN

3.1 *Dataset* Penyakit Jantung

Dataset penelitian ini, yaitu data pasien berpenyakit jantung (*heart disease)* yang diambil dari data repositori Kaggle sebanyak 303 titik data *(data point*). Masing-masing titik data memiliki 12 atribut. Sebelum diproses *dataset* ini disortir terlebih dahulu berdasarkan *Cp* (*Chest pain*) karena hasil wawancara bersama ahli medis bahwa *Cp* merupakan gejala utama resiko penyakit jantung. Berikut pada Tabel 2 disajikan *row dataset*.

Tabel 2 *Row Dataset* Sebelum Normalisasi

Id	*Age*	*Sex*	*Cp*	*Trestbps*	*Chol*	*Fbs*	*Restecg*	*Thalach*	*Exang*	*Oldpeak*	*Slope*
1	63	1	1	145	233	1	2	150	0	2.3	3
21	64	1	1	110	211	0	2	144	1	1.8	2
22	58	0	1	150	283	1	2	162	0	1	1
28	66	0	1	150	226	0	0	114	0	2.6	3
31	69	0	1	140	239	0	0	151	0	1.8	1
42	40	1	1	140	199	0	0	178	1	1.4	1
60	51	1	1	125	213	0	2	125	1	1.4	1
102	34	1	1	118	182	0	2	174	0	0	1
113	52	1	1	118	186	0	2	190	0	0	2
125	65	1	1	138	282	1	2	174	0	1.4	2
142	59	1	1	170	288	0	2	159	0	0.2	2
151	52	1	1	152	298	1	0	178	0	1.2	2
…	…	…	…	…	…	…	…	…	…	…	…
298	57	0	4	140	241	0	0	123	1	0.2	2
300	68	1	4	144	193	1	0	141	0	3.4	2
301	57	1	4	130	131	0	0	115	1	1.2	2

3.2 *Preprocessing*

Pada tahap *pre-processing* dalam penelitian ini, dilakukan pengecekan *missing value* dan normalisasi pada *dataset*. Hasil pengecekan terhadap *missing value*, tidak terdeteksi adanya *missing value* dalam *dataset* sehingga jumlah data yang diproses tetap 303 *record*. Selanjutnya dilakukan proses normalisasi data, sehingga nilai-nilai dalam *dataset* tersebut berada dalam rentang skala 0 sampai 1. Pada Tabel 3 disajikan row *dataset* setelah proses normalisasi.

Tabel 3 *Row Dataset* Setelah Normalisasi

Age	*Sex*	*Cp*	*Trestbps*	*Chol*	*Fbs*	*Restecg*	*Thalach*	*Exang*	*Oldpeak*	*Slope*
0.818	1	0.25	0.725	0.413	1	1	0.743	0	0.371	1.000
0.831	1	0.25	0.550	0.374	0	1	0.713	1	0.290	0.667
0.753	0	0.25	0.750	0.502	1	1	0.802	0	0.161	0.333
0.857	0	0.25	0.750	0.401	0	0	0.564	0	0.419	1.000
0.896	0	0.25	0.700	0.424	0	0	0.748	0	0.290	0.333
0.519	1	0.25	0.700	0.353	0	0	0.881	1	0.226	0.333
0.662	1	0.25	0.625	0.378	0	1	0.619	1	0.226	0.333
0.442	1	0.25	0.590	0.323	0	1	0.861	0	0.000	0.333
0.675	1	0.25	0.590	0.330	0	1	0.941	0	0.000	0.667
0.844	1	0.25	0.690	0.500	1	1	0.861	0	0.226	0.667
0.766	1	0.25	0.850	0.511	0	1	0.787	0	0.032	0.667
0.675	1	0.25	0.760	0.528	1	0	0.881	0	0.194	0.667
…	…	…	…	…	…	…	…	…	…	…
0.740	0	1	0.700	0.427	0	0	0.609	1	0.032	0.667
0.883	1	1	0.720	0.342	1	0	0.698	0	0.548	0.667
0.740	1	1	0.650	0.232	0	0	0.569	1	0.194	0.667

3.3 Implementasi K-Means *Clustering*

3.3.1 Menentukan jumlah *cluster* (*k*)

Berdasarkan hasil wawancara dengan ahli medis, penentuan jumlah *cluster* (*k*) dalam analisis K-Means *cluster*ing dilakukan dengan mempertimbangkan atribut *Cp* (*Chest Pain*), yang merupakan faktor risiko utama penyakit jantung. Untuk tahap awal, diputuskan untuk menggunakan 4 *cluster*, sesuai dengan empat tingkat nyeri dada yang terdefinisi pada atribut *Cp* (*k*=4). Sebelum implementasi algoritma K-Means, *dataset* disusun terlebih dahulu dengan cara menyortir data berdasarkan urutan nilai atribut *Cp*. Proses penyortiran ini ditampilkan pada Tabel 4 dan dilakukan untuk memastikan bahwa pengelompokan data lebih terarah dan relevan dengan risiko penyakit jantung yang dikaitkan dengan nyeri dada.

3.3.2 Menentukan titik pusat *cluster* awal

Tahap ini merupakan tahap iterasi 1 dengan penentuan titik pusat *cluster* atau *centroid* awal. Penentuan *centroid* awal dilakukan secara acak pada *dataset* penyakit jantung yang berjumlah 303 pasien. *Pada centroid* (*c*1) merupakan titik pusat pada *cluster* (*k*1), *centroid* (*c*2) merupakan titik pusat pada *cluster* (*k*2), *centroid* (*c*2) merupakan titik pusat pada *cluster* (*k*2), *centroid* (*c*3) merupakan titik pusat pada *cluster* (*k*3), *centroid* (*c*4) merupakan titik pusat pada *cluster* (*k*4). *Centroid* awal setiap *cluster* disajikan pada Tabel 4.

Tabel 4 *Centroid* Awal

Centroid	*Age*	*Sex*	*Cp*	*Trestbps*	*Chol*	*Fbs*	*Restecg*	*Thalach*	*Exang*	*Oldpeak*	*Slope*
c1	0.844	1	0.25	0.69	0.500	1	1	0.861	0	0.226	0.667
c2	0.818	0	0.5	0.7	0.346	0	0	0.886	0	0	0.333
c3	0.623	1	0.75	0.62	0.452	1	0	0.866	0	0	0.333
c4	0.844	1	1	0.55	0.440	0	1	0.782	0	0.097	0.333

3.3.3 Menghitung jarak data ke *centroid* setiap *cluster*

Perhitungan jarak data pertama dengan titik *centroid* awal (Tabel 4) menggunakan Pers. (2). Pada tahap ini, jarak data pertama dihitung terhadap masing-masing *centroid* dari setiap *cluster*. Pertama, dihitung jarak data pertama dengan *centroid* pertama (*cluster* 1) yang dinyatakan sebagai $d(1,1)$. Selanjutnya, jarak data pertama dengan *centroid* kedua (*cluster* 2) dihitung sebagai $d(1,2)$, diikuti dengan jarak ke *centroid* ketiga (*cluster* 3) yang ditunjukkan oleh $d(1,3)$. Terakhir, jarak data pertama dengan *centroid* keempat (*cluster* 4) dihitung sebagai $d(1,4)$. Proses ini memastikan bahwa setiap jarak antara data dan *centroid* dapat dianalisis untuk menentukan *cluster* yang paling relevan bagi data tersebut.

$$d\,(x_i, c_h) = \sqrt{(x_{i1} - c_{h1})^2 + (x_{i2} - c_{h2})^2 + (x_{i3} - c_{h3})^2 + \ldots + (x_{im} - c_{hm})^2}$$

$$d(1,1) = \sqrt{(0.818 - 0.844)^2 + (1-1)^2 + (0.25 - 0.25)^2 + (0.725 - 0.69)^2 + (0.413 - 0.500)^2 + (1-1)^2 + (1-1)^2 + (0.743 - 0.861)^2 + (0-0)^2 + (0.371 - 0.226)^2 + (1 - 0.667)^2}$$

$$d(1,1) = 0.66$$

$$d(1,2) = \sqrt{(0.818 - 0.818)^2 + (1-0)^2 + (0.25 - 0.5)^2 + (0.725 - 0.7)^2 + (0.413 - 0.346)^2 + (1-0)^2 + (1-0)^2 + (0.743 - 0.886)^2 + (0-0)^2 + (0.371 - 0)^2 + (1 - 0.333)^2}$$

$$d(1,2) = 1.75$$

$$d(1,3) = \sqrt{(0.818 - 0.623)^2 + (1-1)^2 + (0.25 - 0.75)^2 + (0.725 - 0.62)^2 + (0.413 - 0.452)^2 + (1-1)^2 + (1-0)^2 + (0.743 - 0.886)^2 + (0-0)^2 + (0.371 - 0.0)^2 + (1 - 0.333)^2}$$

$$d(1,3) = 1.18$$

$$d(1,4) = \sqrt{\begin{matrix}(0.818-0.844)^2+\ (1-1)^2+\ (0.25-1)^2+\ (0.725-0.55)^2+(0.413-0.440)^2+(1-0)^2+(1-1)^2+\\ (0.743-0.782)^2+(0-0)^2+\ (0.371-0.097)^2+(1-0.333)^2\end{matrix}}$$

$$d(1,4) = 1.05$$

Perhitungan jarak (d) dari data pertama terhadap *centroid* awal (*c1*, *c2*, *c3*, dan *c4*) yang terdiri dari 11 atribut menunjukkan bahwa data pertama dikelompokkan ke dalam *cluster* 1. Hal ini disebabkan oleh jarak antara data pertama dengan *centroid* pertama *cluster* 1, yaitu $d(1,1)$, yang menghasilkan nilai terdekat sebesar 0,66. Sementara itu, jarak data pertama dengan *centroid* kedua *cluster* 2 $d(1,2)$ bernilai 1,75, jarak dengan *centroid* ketiga *cluster* 3 $d(1,3)$ bernilai 1,18, dan jarak dengan *centroid* keempat *cluster* $d(1,4)$ bernilai 1,05. Proses perhitungan jarak untuk data kedua hingga data terakhir dilakukan dengan menggunakan metode yang sama seperti pada perhitungan $d(1,1)$ hingga $d(1,4)$. Hasil dari *cluster* pada iterasi pertama berdasarkan perhitungan jarak ditampilkan pada Tabel 5.

Tabel 5 Hasil *Cluster* Iterasi 1

Id	*Age*	*Sex*	*Cp*	*Trestbps*	*Chol*	*Fbs*	*Restecg*	*Thalach*	*Exang*	*Oldpeak*	*Slope*	*Cluster*
117	0.753	1	0.75	0.7	0.374	1	1	0.817	0	0	0.333	1
125	0.844	1	0.25	0.25	0.5	1	1	0.861	0	0.226	0.667	1
140	0.662	1	0.75	0.625	0.434	1	1	0.822	0	0.387	0.667	1
…	…	…	…	…	…	…	…	…	…	…	…	…
262	0.753	0	0.5	0.68	0.566	1	1	0.752	0	0	0.333	1
214	0.857	0	1	0.89	0.404	1	0	0.817	1	0.161	0.667	2
6	0.727	1	0.5	0.6	0.418	0	0	0.881	0	0.129	0.333	2
14	0.571	1	0.5	0.6	0.466	0	0	0.856	0	0	0.333	2
…	…	…	…	…	…	…	…	…	…	…	…	…
295	0.818	0	1	0.62	0.349	0	0	0.673	1	0	0.667	2
32	0.779	1	1	0.585	0.408	1	0	0.792	1	0.226	0.333	3
40	0.792	1	0.75	0.75	0.431	1	0	0.678	1	0.161	0.667	3
267	0.675	1	1	0.64	0.362	1	0	0.722	1	0.161	0.667	3
…	…	…	…	…	…	…	…	…	…	…	…	…
300	0.883	1	1	0.72	0.342	1	0	0.698	0	0.548	0.667	3
2	0.870	1	1	0.8	0.507	0	1	0.535	1	0.242	0.667	4
3	0.870	1	1	0.6	0.406	0	1	0.639	1	0.419	0.667	4
9	0.818	1	1	0.65	0.450	0	1	0.728	0	0.226	0.667	4
…	…	…	…	…	…	…	…	…	…	…	…	…
254	0.662	0	0.75	0.6	0.523	0	1	0.777	0	0.097	0.333	4

Berdasarkan hasil *cluster* pada iterasi pertama, dilakukan perhitungan rata-rata untuk setiap *cluster* pada 11 atribut, yang disajikan dalam Tabel 6. Proses ini menghasilkan *centroid* baru untuk masing-masing *cluster*. Hasilnya menunjukkan bahwa *cluster* 1 (*k*1) terdiri dari 27 pasien, *cluster* 2 (*k*2) berjumlah 96 pasien, *cluster* 3 (*k*3) memiliki 15 pasien, dan *cluster* 4 (*k*4) mencakup 165 pasien. Penentuan jumlah pasien dalam setiap *cluster* ini penting untuk memahami distribusi data dan karakteristik masing-masing kelompok dalam analisis penyakit jantung.

Tabel 6 Hasil *Centroid* Iterasi 1

k	*Age*	*Sex*	*Cp*	*Trestbps*	*Chol*	*Fbs*	*Restecg*	*Thalach*	*Exang*	*Oldpeak*	*Slope*	*Jumlah Data*
1	0.748	0.704	0.731	0.717	0.467	0.963	1	0.716	0.444	0.210	0.617	27
2	0.688	0.354	0.667	0.652	0.424	0.042	0.120	0.775	0.188	0.099	0.490	96
3	0.716	1	0.767	0.665	0.391	1	0	0.798	0.2	0.146	0.511	15
4	0.711	0.836	0.873	0.652	0.445	0	0.676	0.720	0.4	0.203	0.547	165

Tabel 6 menunjukan data *centroid* setiap *cluster* (*k*) beserta jumlah data. Perhitungan iterasi kedua menggunakan proses yang sama seperti pada tahap pertama dengan perhitungan jarak setiap data dengan *centroid* baru pada Tabel 6. Perhitungan K-Means *clustering* berakhir pada iterasi keenam karena *centroid* baru pada iterasi ini tidak berubah dari *centroid* sebelumnya seperti yang terlihat pada tabel 8. Hasil *cluster* ditunjukkan pada Tabel 7.

Perhitungan rata-rata setiap atribut pada iterasi keenam menghasilkan *centroid* baru dengan menggunakan rumus AVERAGE di Excel, yang berfungsi untuk menghitung rata-rata. Hasil perhitungan tersebut ditampilkan dalam Tabel 8, yang menunjukkan data *centroid* untuk setiap *cluster* (*k*) beserta jumlah pasien dalam masing-masing *cluster*. Dalam hasil ini, *cluster* 1 (*k*1) tediri 27 pasien, *cluster* 2 (*k*2) berjumlah 135 pasien, *cluster* 3 (*k*3) memiliki 15 pasien, *cluster* 4

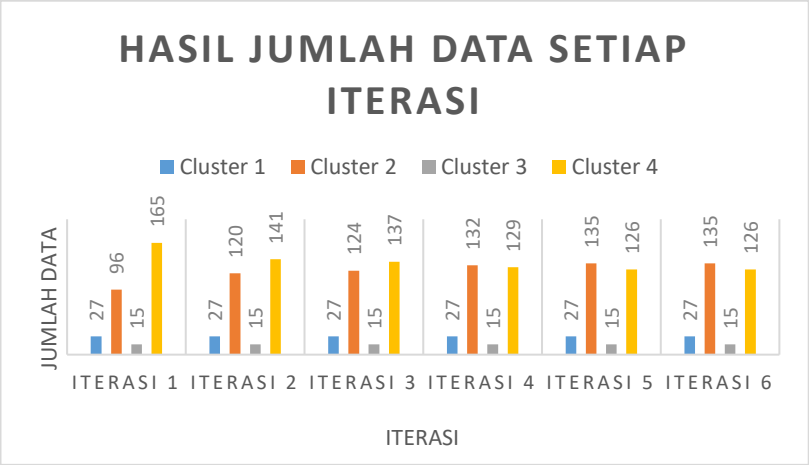
(*k*4) mencakup 126 pasien. Hasil *centroid* dan jumlah data pada *cluster* iterasi 5 sama dengan hasil iterasi 6 sehingga perhitungan dihentikan dan dinyatakan telah selesai. Jumlah data pada setiap iterasi ditampilkan dalam diagram batang yang disajikan pada Gambar 3.

Tabel 7 Hasil *Cluster* Iterasi 6

Id	*Age*	*Sex*	*Cp*	*Trestbps*	*Chol*	*Fbs*	*Restecg*	*Thalach*	*Exang*	*Oldpeak*	*Slope*	*Cluster*
117	0.753	1	0.75	0.7	0.374	1	1	0.817	0	0	0.333	1
125	0.844	1	0.25	0.25	0.5	1	1	0.861	0	0.226	0.667	1
140	0.662	1	0.75	0.625	0.434	1	1	0.822	0	0.387	0.667	1
…	…	…	…	…	…	…	…	…	…	…	…	…
262	0.753	0	0.5	0.68	0.566	1	1	0.752	0	0	0.333	1
4	0.481	1	0.75	0.65	0.443	0	0	0.926	0	0.565	1	2
6	0.727	1	0.5	0.6	0.418	0	0	0.881	0	0.129	0.333	2
11	0.740	1	1	0.7	0.340	0	0	0.733	0	0.065	0.667	2
…	…	…	…	…	…	…	…	…	…	…	…	…
122	0.818	0	1	0.75	0.722	0	1	0.762	0	0.645	0.667	2
32	0.779	1	1	0.585	0.408	1	0	0.792	1	0.226	0.333	3
40	0.792	1	0.75	0.75	0.431	1	0	0.678	1	0.161	0.667	3
267	0.675	1	1	0.64	0.362	1	0	0.722	1	0.161	0.667	3
…	…	…	…	…	…	…	…	…	…	…	…	…
300	0.883	1	1	0.72	0.342	1	0	0.698	0	0.548	0.667	3
283	0.714	0	1	0.64	0.363	0	0.5	0.644	1	0.323	0.667	4
2	0.870	1	1	0.8	0.507	0	1	0.535	1	0.242	0.667	4
3	0.870	1	1	0.6	0.406	0	1	0.639	1	0.419	0.667	4
…	…	…	…	…	…	…	…	…	…	…	…	…
62	0.597	0	0.75	0.71	0.314	0	1	0.792	1	0.226	1	4

Tabel 8 Hasil *Centroid* Iterasi 6

k	*Age*	*Sex*	*Cp*	*Trestbps*	*Chol*	*Fbs*	*Restecg*	*Thalach*	*Exang*	*Oldpeak*	*Slope*	*Jumlah Data*
1	0.751	0.667	0.759	0.717	0.464	1	0.963	0.719	0.481	0.191	0.605	27
2	0.694	0.415	0.757	0.645	0.437	0.022	0.226	0.763	0.059	0.118	0.489	135
3	0.716	1	0.767	0.665	0.391	1	0	0.798	0.200	0.146	0.511	15
4	0.711	0.929	0.833	0.659	0.438	0	0.742	0.714	0.595	0.219	0.569	126



Gambar 3 Hasil Jumlah Data Setiap Iterasi

Pada Gambar 3, terlihat bahwa proses iterasi dari 1 hingga 6 menunjukkan jumlah data yang konsisten pada *cluster* 1, yaitu sebanyak 27 pasien, dan pada *cluster* 3, yang terdiri dari 15 pasien. Namun, terdapat perubahan jumlah pasien pada *cluster* 2 dan *cluster* 4. Pada akhir iterasi, jumlah pasien di *cluster* 2 mencapai 135 pasien, sedangkan *cluster* 4 berjumlah 126 pasien. Perubahan ini mencerminkan dinamika pengelompokan data selama proses iterasi, di mana *cluster* 2 dan *cluster* 4 mengalami penyesuaian jumlah pasien sesuai dengan perhitungan *centroid* yang dilakukan.

3.4 Analisis Hasil *Cluster*

Pada hasil iterasi keenam, telah diperoleh *cluster* dan *centroid* berdasarkan data yang telah dinormalisasi. Namun, untuk melakukan analisis yang lebih komprehensif, penting untuk menerjemahkan hasil *cluster* kembali ke dalam bentuk data awal atau data sebelum dinormalisasi. Tabel 9 menyajikan terjemahan dari data hasil *cluster*, sementara Tabel 10 menampilkan terjemahan *centroid* pada iterasi keenam.

Tabel 9 Terjemahan Data Hasil *Cluster* Iterasi 6

Id	*Age*	*Sex*	*Cp*	*Trestbps*	*Chol*	*Fbs*	*Restecg*	*Thalach*	*Exang*	*Oldpeak*	*Slope*	*Cluster*
117	58	1	3	140	211	1	2	165	0	0	1	1
125	65	1	1	138	282	1	2	174	0	1.4	1.4	1
140	51	1	3	125	245	1	2	166	0	2.4	2.4	1
…	…	…	…	…	…	…	…	…	…	…	…	…
262	58	0	2	136	319	1	2	152	0	0	0	1
4	37	1	3	130	250	0	0	187	0	3.5	3.5	2
6	56	1	2	120	236	0	0	178	0	0.8	0.8	2
11	57	1	4	140	192	0	0	148	0	0.4	2	2
…	…	…	…	…	…	…	…	…	…	…	…	…
122	63	0	4	150	407	0	2	154	0	4	2	2
32	60	1	4	117	230	1	0	160	1	1.4	1	3
40	61	1	4	150	243	1	0	137	1	1	2	3
267	52	1	3	128	204	1	0	156	1	1	2	3
…	…	…	…	…	…	…	…	…	…	…	…	…
300	68	1	4	144	193	1	0	141	0	3.4	2	3
283	55	0	4	128	205	0	1	130	1	2	2	4
2	67	1	4	160	286	0	2	108	1	1.5	2	4
3	67	1	4	120	229	0	2	129	1	2.6	2	4
…	…	…	…	…	…	…	…	…	…	…	…	…
62	46	0	3	142	177	0	2	160	1	1.4	13	4

Tabel 10 Terjemahan Data Hasil *Centroid* Iterasi 6

k	*Age*	*Sex*	*Cp*	*Trestbps*	*Chol*	*Fbs*	*Restecg*	*Thalach*	*Exang*	*Oldpeak*	*Slope*	*Jumlah Data*
1	58	1	3	143	262	1	2	145	0	1.2	2	27
2	53	0	3	129	246	0	0	154	0	0.7	1	135
3	55	1	3	133	220	1	0	161	0	0.9	2	15
4	55	1	3	132	247	0	1	144	1	1.4	2	126

Interpertasi data hasil *cluster* pada Tabel 9 dan Tabel 10 yang dilakukan bersama ahli medis dihasilkan bahwa *cluster* 1 (*k*1) memiliki rata-rata usia 58 tahun, mayoritas laki-laki, dengan tipe nyeri dada (*Cp*) yang menunjukkan gejala nyeri dada parah. Pasien dalam kelompok ini cenderung memiliki tekanan darah istirahat (*Trestbps*) dan kadar *cholesterol* (*Chol*) yang tinggi, serta *fasting blood sugar* (*Fbs*) yang lebih dari 120 mg/dl. *Resting electrocardiographic* (*Restecg*) dan *exercise-induced angina* (*Exang*) menunjukkan tingkat abnormalitas yang cukup signifikan. Tingkat detak jantung maksimum (*Thalach*) cenderung rendah, dan nilai *Oldpeak* yang tinggi mengindikasikan depresi ST yang mungkin menjadi tanda risiko lebih tinggi untuk penyakit jantung. *Slope* puncak ST setelah berolah raga (*Slope*) cenderung meningkat seiring dengan peningkatan tingkat nyeri dada.

Cluster 2 (*k*2), dengan rata-rata usia 53 tahun, mayoritas perempuan, menunjukkan tipe nyeri dada yang cenderung parah. Pasien dalam kelompok ini memiliki tekanan darah (*Trestbps*) dan kadar *cholesterol* (*Chol*) yang relatif normal, *fasting blood sugar* (*Fbs*) rendah, serta *resting electrocardiographic* (*Restecg*) dan *exercise-induced angina* (*Exang*) yang cenderung normal. Tingkat detak jantung maksimum (*Thalach*) dan nilai *Oldpeak* yang rendah menunjukkan adanya respon yang lebih baik terhadap aktivitas fisik. *Slope* puncak ST setelah berolah raga (*Slope*) cenderung datar.

Cluster 3 (*k*3), dengan rata-rata usia 55 tahun, mayoritas laki-laki, menunjukkan tipe nyeri dada yang cenderung parah. Pasien dalam kelompok ini memiliki tekanan darah (*Trestbps*) dan kadar *cholesterol* (*Chol*) yang relatif normal, *fasting blood sugar* (*Fbs*) tinggi, serta *resting electrocardiographic* (*Restecg*) yang normal. *Exercise-induced angina* (*Exang*) cenderung rendah. Tingkat detak jantung maksimum (*Thalach*) dan nilai *Oldpeak* menunjukkan respon yang baik terhadap aktivitas fisik. *Slope* puncak ST setelah berolah raga (*Slope*) cenderung meningkat.

Cluster 4 (*k*4), dengan rata-rata usia 55 tahun, mayoritas laki-laki, menunjukkan tipe nyeri dada yang cenderung parah. Pasien dalam kelompok ini memiliki tekanan darah (*Trestbps*) dan kadar *cholesterol* (*Chol*) yang relatif normal, *fasting blood sugar* (*Fbs*) rendah, serta *resting electrocardiographic* (*Restecg*) dan *exercise-induced angina* (*Exang*) yang cenderung tinggi. Tingkat detak jantung maksimum (*Thalach*) dan nilai *Oldpeak* menunjukkan respon yang beragam, sementara *Slope* puncak ST setelah berolah raga (*Slope*) cenderung rendah.

Berdasarkan hasil tersebut, dapat disarankan bahwa *cluster k*1 menunjukan tingkat risiko penyakit jantung yang lebih tinggi karena pasien dalam kelompok ini memiliki resiko sangat tinggi terhadap penyakit jantung, pasien menunjukan gejala serius seperti nyeri dada (*Cp*) tinggi, tekanan darah (*Trestbps*) tinggi, *cholesterol* (*Chol*) tinggi, *fasting blood sugar* (*Fbs*) tinggi, *electrocardiographic* (*Restecg*) tinggi dan faktor resiko lainnya, sementara *k*2, *k*3, dan *k*4 menunjukkan risiko yang lebih rendah, dengan variasi respon terhadap aktivitas fisik. Pengelompokan ini dapat memberikan informasi awal untuk merancang strategi pengelolaan dan perawatan yang lebih spesifik sesuai dengan karakteristik dari setiap kelompok. Tetapi, perlu diingat bahwa validasi klinis lebih lanjut dan pertimbangan medis lebih mendalam tetap diperlukan untuk penanganan pasien secara lebih tepat dan efektif.

4. KESIMPULAN

Kesimpulan dari penelitian ini menunjukkan bahwa *cluster k*1 memiliki profil risiko yang paling tinggi. Pasien dalam kelompok *k*1 sebagian besar adalah laki-laki dengan rata-rata usia lebih tua yang menunjukkan gejala nyeri dada parah, tekanan darah tinggi, kadar kolesterol tinggi. Sementara itu, *cluster k*2, *k*3, dan *k*4 menunjukkan profil risiko yang lebih rendah dengan parameter kesehatan yang lebih normal dan respon yang baik terhadap aktivitas fisik, meskipun nyeri dada tetap menjadi gejala yang dominan.
Dalam sintesis hasil penelitian, dapat disimpulkan bahwa pengelompokan pasien berdasarkan karakteristik klinis dan demografis dapat memberikan wawasan penting untuk strategi pengelolaan dan perawatan yang lebih terarah. *Cluster k*1 memerlukan intervensi medis yang lebih intensif dan pemantauan ketat untuk mengelola faktor risiko yang tinggi, sedangkan *cluster k*2, *k*3, dan *k*4 memerlukan pendekatan yang lebih disesuaikan dengan profil risiko masing-masing. Pendekatan yang berbeda ini memungkinkan penyedia layanan kesehatan untuk memberikan perawatan yang lebih efektif dan efisien, serta meningkatkan kualitas hidup pasien melalui pengelolaan penyakit yang lebih personal dan tepat sasaran. Validasi klinis lebih lanjut diperlukan untuk memastikan bahwa pendekatan ini dapat diterapkan secara luas dan memberikan manfaat yang maksimal bagi pasien.

DAFTAR PUSTAKA

Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, *136*, 104672. https://doi.org/10.1016/j.compbiomed.2021.104672

Arifandi, M., Hermawan, A., Hermawan, A., Avianto, D., & Avianto, D. (2021). Implementasi Algoritma K-Medoids untuk Clustering Wilayah Terinfeksi Kasus Covid-19 di DKI Jakarta. *JTT (Jurnal Teknologi Terapan)*, *7*(2), 120–128. https://doi.org/10.31884/jtt.v7i2.353

Awan, A. A. (2020). *Heart Disease patients*. Kaggle. https://www.kaggle.com/datasets/kingabzpro/heart-disease-patients

Das, D., Kayal, P., & Maiti, M. (2023). A K-means clustering model for analyzing the Bitcoin extreme value returns. *Decision Analytics Journal*, *6*(2022), 100152. https://doi.org/10.1016/j.dajour.2022.100152

Han, J., & Kang, S. (2023). Optimization of missing value imputation for neural networks. *Information Sciences*, *649*, 119668. https://doi.org/10.1016/j.ins.2023.119668

Haris Kurniawan, Sarjon Defit, & Sumijan. (2020). Data Mining Menggunakan Metode K-Means Clustering Untuk Menentukan Besaran Uang Kuliah Tunggal. *Journal of Applied Computer Science and Technology*, *1*(2), 80–89. https://doi.org/10.52158/jacost.v1i2.102

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

Masitha, A., Biddinika, M. K., & Herman, H. (2023). K Value Effect on Accuracy Using the K-NN for Heart Failure Dataset. *MATRIK : Jurnal Manajemen, Teknik Informatika Dan Rekayasa Komputer*, *22*(3), 593–604. https://doi.org/10.30812/matrik.v22i3.2984

Mishra, P., Biancolillo, A., Roger, J. M., Marini, F., & Rutledge, D. N. (2020). New data preprocessing trends based on ensemble of multiple preprocessing techniques. *TrAC Trends in Analytical Chemistry*, *132*, 116045. https://doi.org/10.1016/j.trac.2020.116045

Muslimah, V. (2024). Implementing Bayes' Theorem Method in Expert System to Determine Infant Disease. *Khazanah Informatika : Jurnal Ilmu Komputer Dan Informatika*, *10*(1), 1–14. https://doi.org/10.23917/KHIF.V10I1.4837

Novidianto, R., Wibowo, H., & Chandranegara, D. R. (2021). ClusterMix K-Prototypes Algorithm to Capture Variable Characteristics of Patient Mortality With Heart Failure. *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, *6*(2), 109–116. https://doi.org/10.22219/kinetik.v6i2.1209

Purba, N., Poningsih, P., & Tambunan, H. S. (2021). Penerapan Algoritma K-Means Clustering Pada Penyebaran Penyakit Infeksi Saluran Pernapasan Akut (ISPA) di Provinsi Riau. *Journal of Information System Research (JOSH)*, *2*(3), 220–226. https://ejurnal.seminar-id.com/index.php/josh/article/view/736

Qi, K.-T., Zhang, H.-S., Zheng, Y.-G., Zhang, Y., & Ding, L.-Y. (2023). Stripe segmentation of oceanic internal waves in SAR images based on Gabor transform and K-means clustering. *Oceanologia*, *65*(4), 548–555. https://doi.org/10.1016/j.oceano.2023.06.006

Rizki, B., Ginasta, N. G., Tamrin, M. A., & Rahman, A. (2020). Customer Loyality Segmentation on Point of Sale System Using Recency-Frequency-Monetary (RFM) and K-Means. *Jurnal Online Informatika*, *5*(2), 130–136. https://doi.org/10.15575/join.v5i2.511

Shah, D., Patel, S., & Bharti, S. K. (2020). Heart Disease Prediction using Machine Learning Techniques. *SN Computer Science*, *1*(6), 345. https://doi.org/10.1007/s42979-020-00365-y

Singh, A., & Kumar, R. (2020). Heart Disease Prediction Using Machine Learning Algorithms. *2020 International Conference on Electrical and Electronics Engineering (ICE3)*, 452–457. https://doi.org/10.1109/ICE348803.2020.9122958

Solechati, R. G., & Jananto, A. (2023). Penerapan Algoritma K-Means Clustering pada Data Brain Stroke untuk Pengelompokan Profile Pasien. *SemanTIK*, *9*(1), 39–46. https://doi.org/10.55679/semantik.v9i1.29446

Sun, F., & Yu, J. (2021). Improved energy performance evaluating and ranking approach for office buildings using Simple-normalization, Entropy-based TOPSIS and K-means method. *Energy Reports*, *7*, 1560–1570. https://doi.org/10.1016/j.egyr.2021.03.007

V. Ramalingam, V., Dandapath, A., & Karthik Raja, M. (2018). Heart disease prediction using machine learning techniques : a survey. *International Journal of Engineering & Technology*, *7*(2.8), 684–687. https://doi.org/10.14419/ijet.v7i2.8.10557