OPEN ACCESS
J. OF ICT, VOL. 3, NO. 2, PP.051-066, DES 2021
ISSN : 2686-1089 (ONLINE)

Journal of Informatics and Communications Technology (JICT)

Penerapan Algoritma K-Means *Clustering* Untuk Pengelompokan Tingkat Risiko Penyakit Jantung

Deny Haryadi [1], Dewi Marini Umi Atmaja [2]

Prodi S1 Teknologi Informasi, Prodi S1 Bisnis Digital
Institut Teknologi Telkom Jakarta, Universitas Medika Suherman
Jln. Daan Mogot KM 11, Cengkareng, Jakarta Barat Indonesia
Jln. Raya Industri Pasirgombong, Cikarang Utara-Bekasi, Jawa Barat Indonesia

[1] denyharyadi@ittelkom-jkt.ac.id, [2] dewi@medikasuherman.ac.id

Abstract

Penyakit jantung merupakan sebuah kondisi dimana jantung tidak dapat melaksanakan tugasnya dengan baik, penyakit ini terjadi bila darah ke otot jantung terhenti atau tersumbat sehingga mengakibatkan kerusakan berat pada jantung. Beberapa faktor yang menyebabkan penyakit jantung antara lain keturunan, usia, jenis kelamin, stres, kurang gerak, merokok, kolesterol tinggi, hipertensi, diabetes, dan obesitas. Pada dasarnya penyakit jantung dapat dicegah dengan berbagai faktor diantaranya pola hidup sehat, selain itu deteksi dini penyakit jantung juga diperlukan untuk mencegah terjadinya kematian pada penderitanya salah satu cara untuk melakukan deteksi dini ialah menggunakan *data mining.* Penggunaan algoritma *k-means* dapat dilakukan untuk melakukan klasterisasi pengelompokan penyakit jantung guna mengetahui seseorang terkena penyakit jantung maupun tidak. Metode klasterisasi dengan algoritma k-means pada penelitian ini menunjukkan sebuah wawasan baru yaitu pengelompokkan tingkat resiko penyakit jantung berdasarkan 3 *cluster*. *Cluster* 1 merupakan kategori usia dengan tingkat resiko penyakit jantung cukup rendah atau *Low* yaitu 355 dari 1025 kategori usia yang diuji, kemudian *cluster* 2 adalah kategori usia dengan tingkat resiko penyakit jantung sedang atau *Medium* yaitu 208 dari 1025 kategori usia yang diuji, dan terakhir adalah *cluster* 3 merupakan kategori usia dengan tingkat kategori usia cukup tinggi atau *High* yaitu 462 dari 1025 kategori usia yang diuji.

Keywords: Algoritma K-Means, *Cluster*, Data Mining, Penyakit Jantung.

I. INTRODUCTION

Jantung adalah sebuah organ tubuh manusia yang berongga serta berotot yang berperan dalam sistem peredaran darah manusia. Jantung memiliki empat ruang yang masing-masing memiliki fungsi tertentu. Organ ini terletak di dalam rongga dada tepatnya di bawah paru-paru sebelah kiri (pada umumnya), dan dilindungi oleh tulang dada (*sternum*) dan tulang rusuk (*costae*). Penyakit jantung merupakan sebuah kondisi dimana jantung tidak dapat melaksanakan tugasnya dengan baik, penyakit ini terjadi bila darah ke otot jantung terhenti atau tersumbat sehingga mengakibatkan kerusakan berat pada jantung [1]. Pada dasarnya penyakit jantung dapat dicegah dengan berbagai faktor, diantaranya pola hidup sehat selain itu deteksi dini penyakit jantung juga diperlukan untuk mencegah terjadinya kematian pada penderitanya salah satu cara untuk melakukan deteksi dini ialah menggunakan *data mining*. Gejala yang ditimbulkan penyakit jantung antara lain rasa tidak nyaman di dada, nyeri sampai ke lengan, sakit menjalar ke bagian rahang atau punggung dan detak jantung kerap tidak

teratur, gangguan pencernaan, pusing, mudah lelah, kerap berkeringat dingin, dan batuk. Dalam dunia kesehatan, penyakit jantung merupakan penyakit yang mendorong angka kematian yang cukup tinggi, sehingga banyak penelitian yang dilakukan sebelumnya untuk memprediksi penyakit jantung, diantaranya yaitu penelitian yang pernah dilakukan oleh Nur Aeni Widiastuti, Stefanus Santosa, Catur Supriyanto, dengan judul Algoritma Klasifikasi Data Mining *Naïve Bayes* Berbasis *Particle Swarm Optimization* untuk Deteksi Penyakit Jantung, menyebutkan bahwa algoritma klasifikasi data mining *naïve bayes* berbasis PSO untuk deteksi penyakit jantung menghasilkan akurasi yang lebih tinggi sebesar 92.86% dibandingkan algoritma klasifikasi data mining *naïve bayes* menghasilkan akurasi sebesar 82.14% [2]. Serta penelitian yang dilakukan oleh Dito Putra Utomo dan Mesran dengan Judul Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut pada *Dataset* Penyakit Jantung menyatakan bahwa algoritma *naïve bayes classifier* memiliki kinerja yang lebih baik dari pada algoritma C5.0 [3]. Berdasarkan penelitian yang telah dilakukan sebelumnya, perlu adanya pengelompokkan tingkat risiko penyakit jantung berdasarkan usia menggunakan salah satu teknik *data mining* yaitu metode *clustering*.

II. LITERATURE REVIEW

A. *Penyakit Jantung*

Penyakit jantung merupakan gangguan yang terjadi pada sistem pembuluh darah besar sehingga menyebabkan jantung dan peredaran darah tidak berfungsi sebagaimana mestinya. Penyakit- penyakit yang berhubungan dengan organ jantung dan pembuluh darah antara lain: gagal jantung, jantung koroner, dan jantung rematik. Penyakit jantung koroner (PJK) adalah penyakit jantung dan pembuluh darah yang disebabkan karena penyempitan arteri koroner. Penyempitan pembuluh darah terjadi karena proses aterosklerosis atau spasme atau kombinasi keduanya [1]. Aterosklerosis yang terjadi karena timbunan kolesterol dan jaringan ikat pada dinding pembuluh darah secara perlahan-lahan, hal ini sering ditandai dengan keluhan nyeri pada dada. Pada waktu jantung harus bekerja lebih keras terjadi ketidakseimbangan antara kebutuhan dan asupan oksigen, hal inilah yang menyebabkan nyeri dada. Kalau pembuluh darah tersumbat sama sekali, pemasokan darah ke jantung akan terhenti dan kejadian inilah yang disebut dengan serangan jantung. Penyakit jantung sering disebut "suddent death". Seseorang kemungkinan mengalami serangan jantung karena iskemia miokard atau kekurangan oksigen pada otot jantung atau sering disebut dengan nyeri dada. Beberapa faktor yang bisa menimbulkan penyakit jantung antara lain: bertambahnya usia, gaya hidup, stres, kurangnya waktu istirahat, kurangnya berolah raga, merokok, obesitas, dislipidemia, permasalahan dalam diagnosa klinis penyakit jantung.

B. *Algoritma K-Means Clustering*

Data mining adalah suatu proses ekstraksi atau penggalian data yang belum diketahui sebelumnya, namun dapat dipahami dan berguna dari *database* yang besar serta digunakan untuk membuat suatu keputusan bisnis yang sangat penting [4]. Berdasarkan fungsionalitas, *data mining* terbagi dalam 5 metode yaitu estimasi, prediksi, klasifikasi, *clustering*, dan asosiasi. Salah satu metode pada data mining, yaitu *Clustering* yang merupakan metode pengelompokan data. Tujuan dari pengelompokan *cluster* ini adalah untuk menemukan pengelompokan dari serangkaian pola, titik, objek maupun dokumen. Objek yang berada didalam pengelompokkan *cluster* yang sama memiliki kemiripan antar satu kelompok dan memiliki perbedaan dengan objek oleh kelompok *cluster* lain [5]. K-Means adalah sebuah metode *Clustering* yang termasuk dalam pendekatan *partitioning*. Algoritma K-Means merupakan model *Centroid*. *Mode Centroid* adalah model yang menggunakan *Centroid* untuk membuat *Cluster*. *Centroid* adalah titik tengah suatu *Cluster*. *Centroid* berupa *Centroid* nilai. *Centroid* digunakan untuk menghitung jarak suatu objek data terhadap *Centroid*. Suatu objek data termasuk dalam *Cluster* jika memiliki jarak terpendek terhadap *Centroid Cluster* tersebut. Algoritma K-Means dapat diartikan sebagai algoritma pembelajaran yang sederhana untuk memecahkan suatu permasalahan pengelompokkan yang bertujuan untuk meminimalkan kesalahan ganda [6]. Dalam menentukan nilai *Centroid* untuk awal interasi, nilai awal *Centroid* dilakukan secara acak. Sedangkan jika menentukan nilai *Centroid* yang merupakan tahap dari iterasi, maka digunakan rumus sebagai berikut:

$$Vij = \frac{1}{Ni}\sum_{k=0}^{Ni} Xkj$$

Keterangan:

*V*ij = *Centroid* rata-rata *cluster* ke-*i* untuk variabel ke-*i*

*N*i = jumlah *cluster* ke -i
i, *k* = indeks dari *cluster*
j = indeks dari variabel
Xkj = nilai data ke –k variabel ke –j dalam cluster

Menghitung jarak antara titik Centroid dengan titik tiap objek:

$$D = \sqrt{(xi - si)^2 + (yi - ti)^2}$$

Keterangan:
D = *Euclidean Distance*
i = banyaknya objek
(x,y) = koordinat objek
(s,t) = koordinat *centroid*

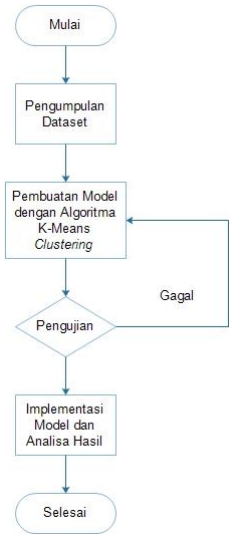
Metode *cross-validation* digunakan untuk menghindari *overlapping* pada data *testing*, adapun tahapan *cross-validation* adalah: a. Bagi data menjadi k *subset* yg berukuran sama. b. Gunakan setiap *subset* untuk data *testing* dan sisanya untuk data *training*. *Cross-validation* biasa disebut juga dengan *K-fold cross-validation*, seringkali subset dibuat *stratified* (bertingkat) sebelum *cross-validation* dilakukan, karena stratifikasi akan mengurangi variansi dari estimasi [7].

C. RapidMiner

RapidMiner merupakan perangkat lunak yang bersifat terbuka (*open source*). *RapidMiner* adalah sebuah solusi untuk melakukan analisis terhadap *data mining*, *text mining* dan analisis prediksi. *RapidMiner* ditulis dengan menggunakan bahasa java sehingga dapat bekerja di semua sistem operasi. *RapidMiner* sebelumnya bernama YALE (*Yet Another Learning Environment*), dimana versi awalnya mulai dikembangkan pada tahun 2001 oleh RalfKlinkenberg, Ingo Mierswa, dan Simon Fischer di *Artificial Intelligence Unit* dari *University of Dortmund*. *RapidMiner* didistribusikan di bawah lisensi AGPL (*GNU Affero General Public License*) versi 3. *RapidMiner* menyediakan GUI (*Graphic User Interface*) untuk merancang sebuah *pipeline* analitis. GUI ini akan menghasilkan file XML (*Extensible Markup Language*) yang mendefinisikan proses analitis keinginan pengguna untuk diterapkan ke data. *File* ini kemudian dibaca oleh *RapidMiner* untuk menjalankan analis secara otomatis [8].

III. RESEARCH METHOD

Dalam pembangunan sistem ini dimulai dari pengumpulan data penyakit jantung yang diambil dari situs Kaggle.com. Selanjutnya dilakukan pembuatan model dengan metode *clustering* menggunakan algoritma *k-means* di RapidMiner. Tujuan dilakukannya penelitian ini yaitu untuk mengelompokan tingkat risiko penyakit jantung berdasarkan usia.



Gambar 1 Metode Penelitian

Teknik pengumpulan data pada penelitian ini ialah dokumentasi dan studi pustaka. Data diolah dengan proses pemilihan atau seleksi data, proses pembersihan data (membuang *missing value*, duplikasi data, dan memeriksa inkonsistensi data dan memperbaiki kesalahan pada data), dan proses transformasi (mengubah format data awal menjadi sebuah format data standar untuk proses pembacaan data pada program maupun *tool* yang digunakan). Pemodelan pada penelitian ini dilakukan dengan metode *clustering* memakai algoritma k-means. Penelitian ini diimplementasikan dengan *tools RapidMiner*. Validasi model pada penelitian ini menggunakan *K-fold cross-validation*.

IV. RESULTS AND DISCUSSION

Penelitian ini menggunakan algoritma k-means untuk mengelompokkan mengelompokan tingkat risiko penyakit jantung berdasarkan usia. Sumber data sebagai objek pada penelitian ini adalah data yang diambil dari situs Kaggle.com. Data yang digunakan dalam penelitian ini terdiri dari atribut atau variabel seperti *age*, *trestbps*, dan *chol*.

Tabel I Dataset Penyakit Jantung

Age	**Trestbps**	**Chol**
52	125	212
53	140	203
70	145	174
61	148	203
62	138	294
58	100	248
58	114	318
55	160	289
46	120	249
54	122	286
….	….	….
59	140	221
60	125	258
47	110	275
50	110	254
54	120	188

Perhitungan algoritma K-Means ini menggunakan dataset yang akan diolah yaitu sebanyak 1025 usia yang akan dikelompokkan kedalam tiga *cluster* yaitu *low*, *medium*, dan *high*. Dengan pemodelan yang sudah ditetapkan sebelumnya, maka berikut adalah contoh untuk perhitungan *Euclidean Distance* pada record ke 1 dan 1025 dari proses iterasi pertama:

Data Ke- 1

$$D1(c1) = \sqrt{(125-130)^2+(212-242)^2} = 30$$

$$D1(c2) = \sqrt{(125-132)^2+(212-249)^2} = 38$$

$$D1(c3) = \sqrt{(125-132)^2+(212-247)^2} = 36$$

Data Ke- 1025

$$D1025(c1) = \sqrt{(120-130)^2+(188-242)^2} = 55$$

$$D1025(c2) = \sqrt{(120-132)^2+(188-249)^2} = 62$$

$$D1025(c3) = \sqrt{(120-132)^2+(188-247)^2} = 61$$

Tabel II Euclidean Distance dan Matriks Kelompok Data Iterasi 1

C1	C2	C3	Jarak Terpendek	C1	C2	C3
30	38	36	30	T		
40	47	45	40	T		
69	76	75	69	T		
42	49	47	42	T		
53	45	47	45		T	
31	32	32	31	T		
78	71	73	71		T	
56	49	50	49		T	
13	12	12	12			T
45	38	40	38		T	
….	….	….	….	….	….	….
23	29	28	23	T		
17	12	13	12		T	
39	34	35	34		T	
24	23	23	23			T
55	62	61	55	T		
			Index	544	451	30

Langkah berikutnya perlu ditentukan kembali titik klaster baru yang dihitung dengan mencari nilai rata-rata berdasarkan dari data anggota masing-masing kelompok klaster. Berikut adalah perhitungan untuk penentuan nilai titik klaster baru dari proses iterasi pertama:

C1 baru ke 2 (*Trestbps*) $= \frac{70.018}{544} = 129$

C1 baru ke 2 (*Chol*) $= \frac{113.653}{544} = 209$

C2 baru ke 2 (*Trestbps*) $= \frac{61.048}{451} = 135$

C2 baru ke 2 (*Chol*) $= \frac{131.090}{451} = 291$

C3 baru ke 2 (*Trestbps*) $= \frac{3.836}{30} = 128$

C3 baru ke 2 (*Chol*) $= \frac{7.407}{30} = 247$

Tabel III Titik Pusat Awal Klaster Baru ke-2

Titik pusat awal cluster	Trestbps	Chol
Cluster Baru ke -1	129	209
Cluster Baru ke-2	135	291

Titik pusat awal cluster	Trestbps	Chol
Cluster Baru ke-3	128	247

Nilai titik klaster baru tersebut digunakan kembali untuk perhitungan iterasi ke-2, dengan langkah yang sama seperti sebelumnya, maka dibawah ini adalah contoh untuk perhitungan *Euclidean Distance* terhadap data ke-1 dan data ke-1025 pada proses iterasi kedua:

Data Ke- 1

$$D1(c1) = \sqrt{(125-129)^2+(212-209)^2} = 5$$

$$D1(c2) = \sqrt{(125-135)^2+(212-291)^2} = 79$$

$$D1(c3) = \sqrt{(125-128)^2+(212-247)^2} = 35$$

Data Ke- 1025

$$D1025(c1) = \sqrt{(120-129)^2+(188-209)^2} = 23$$

$$D1025(c2) = \sqrt{(120-135)^2+(188-291)^2} = 104$$

$$D1025(c3) = \sqrt{(120-128)^2+(188-247)^2} = 59$$

Tabel IV Euclidean Distance dan Matriks Kelompok Data Iterasi 2

C1	C2	C3	Jarak Terpendek	C1	C2	C3
5	79	35	5	T		
13	88	46	13	T		
39	117	75	39	T		
20	89	48	20	T		
86	4	48	4		T	
48	55	28	28			T
110	35	72	35		T	
86	25	53	25		T	
41	44	8	8			T
77	14	40	14		T	
….	….	….	….	….	….	….
17	70	29	17	T		
49	34	11	11			T
69	30	33	30		T	
49	45	19	19			T
23	104	59	23	T		
			Index	403	299	323

Langkah berikutnya perlu ditentukan kembali titik klaster baru yang dihitung dengan mencari nilai rata-rata berdasarkan dari data anggota masing-masing kelompok klaster. Berikut adalah perhitungan untuk penentuan nilai titik klaster baru untuk proses iterasi ketiga:

C1 baru ke 3 (*Trestbps*) $= \frac{52.414}{403} = 130$

C1 baru ke 3 (*Chol*) $= \frac{80.382}{403} = 199$

C2 baru ke 3 (*Trestbps*) $= \frac{41.255}{299} = 138$

C2 baru ke 3 (*Chol*) $= \frac{91.956}{299} = 308$

C3 baru ke 3 (*Trestbps*) $= \frac{41.233}{323} = 128$

C3 baru ke 3 (*Chol*) $= \frac{79.812}{323} = 247$

Tabel V Titik Pusat Awal Klaster Baru ke-3

Titik pusat awal cluster	Trestbps	Chol
Cluster Baru ke -1	130	199
Cluster Baru ke-2	138	308
Cluster Baru ke-3	128	247

Nilai titik klaster baru tersebut digunakan kembali untuk perhitungan iterasi ke-3, dengan langkah yang sama seperti sebelumnya, maka dibawah ini adalah contoh untuk perhitungan *Euclidean Distance* terhadap data ke-1 dan data ke-1025 pada proses iterasi ketiga:

Data Ke- 1

$$D1(c1) = \sqrt{(125\text{-}130)^2+(212\text{-}199)^2} = 14$$

$$D1(c2) = \sqrt{(125\text{-}138)^2+(212\text{-}308)^2} = 96$$

$$D1(c3) = \sqrt{(125\text{-}128)^2+(212\text{-}247)^2} = 35$$

Data Ke- 1025

$$D1025(c1) = \sqrt{(120\text{-}130)^2+(188\text{-}199)^2} = 15$$

$$D1025(c2) = \sqrt{(120\text{-}138)^2+(188\text{-}308)^2} = 121$$

$$D1025(c3) = \sqrt{(120\text{-}128)^2+(188\text{-}247)^2} = 60$$

Tabel VI Euclidean Distance dan Matriks Kelompok Data Iterasi 3

C1	C2	C3	Jarak Terpendek	C1	C2	C3
14	96	35	14	T		
11	105	46	11	T		
30	134	75	30	T		
18	105	49	18	T		
95	14	48	14		T	
57	71	28	28			T
120	26	72	26		T	
94	29	53	29		T	
51	61	8	8			T
87	27	39	27		T	

C1	C2	C3	Jarak Terpendek	C1	C2	C3
….	….	….	….	….	….	….
24	87	29	24	T		
59	51	11	11			T
78	43	33	33			T
58	60	19	19			T
15	121	60	15	T		
			Index	359	261	405

Langkah berikutnya perlu ditentukan kembali titik klaster baru yang dihitung dengan mencari nilai rata-rata berdasarkan dari data anggota masing-masing kelompok klaster menggunakan rumus berikut dibawah ini:

C1 baru ke 4 (*Trestbps*) $= \frac{46.384}{359} = 129$

C1 baru ke 4 (*Chol*) $= \frac{70.451}{359} = 196$

C2 baru ke 4 (*Trestbps*) $= \frac{36.097}{261} = 138$

C2 baru ke 4 (*Chol*) $= \frac{81.615}{261} = 313$

C3 baru ke 4 (*Trestbps*) $= \frac{52.421}{405} = 129$

C3 baru ke 4 (*Chol*) $= \frac{100.084}{405} = 247$

Tabel VII Titik Pusat Awal Klaster Baru ke-4

Titik pusat awal cluster	Trestbps	Chol
Cluster Baru ke -1	129	196
Cluster Baru ke-2	138	313
Cluster Baru ke-3	129	247

Nilai titik klaster baru tersebut digunakan kembali untuk perhitungan iterasi ke-4, dengan langkah yang sama seperti sebelumnya, maka dibawah ini adalah contoh untuk perhitungan *Euclidean Distance* terhadap data ke-1 dan data ke-1025 pada proses iterasi keempat:

Data Ke- 1

$$D1(c1) = \sqrt{(125-129)^2+(212-196)^2} = 16$$

$$D1(c2) = \sqrt{(125-138)^2+(212-313)^2} = 102$$

$$D1(c3) = \sqrt{(125-129)^2+(212-247)^2} = 35$$

Data Ke- 1025

$$D1025(c1) = \sqrt{(120-129)^2+(188-196)^2} = 12$$

$$D1025(c2) = \sqrt{(120-138)^2+(188-313)^2} = 126$$

$$D1025(c3) = \sqrt{(120-129)^2+(188-247)^2} = 60$$

Tabel VIII Euclidean Distance dan Matriks Kelompok Data Iterasi 4

C1	C2	C3	Jarak Terpendek	C1	C2	C3
16	102	35	16	T		
13	110	45	13	T		
27	139	75	27	T		
20	110	48	20	T		
98	19	48	19		T	
59	75	29	29			T
123	25	73	25		T	
98	32	52	32		T	
54	66	10	10			T
90	31	40	31		T	
….	….	….	….	….	….	….
27	92	28	27	T		
62	56	12	12			T
81	47	34	34			T
61	65	21	21			T
12	126	60	12	T		
			Index	345	240	440

Langkah berikutnya perlu ditentukan kembali titik klaster baru yang dihitung dengan mencari nilai rata-rata berdasarkan dari data anggota masing-masing kelompok klaster menggunakan rumus berikut dibawah ini:

C1 baru ke 5 (*Trestbps*) $= \frac{44.316}{345} = 128$

C1 baru ke 5 (*Chol*) $= \frac{67.324}{345} = 195$

C2 baru ke 5 (*Trestbps*) $= \frac{32.963}{240} = 137$

C2 baru ke 5 (*Chol*) $= \frac{75.815}{240} = 316$

C3 baru ke 5 (*Trestbps*) $= \frac{57.623}{440} = 131$

C3 baru ke 5 (*Chol*) $= \frac{109.011}{440} = 248$

Tabel IX Titik Pusat Awal Klaster Baru ke-5

Titik pusat awal cluster	Trestbps	Chol
Cluster Baru ke -1	128	195
Cluster Baru ke-2	137	316
Cluster Baru ke-3	131	248

Nilai titik klaster baru tersebut digunakan kembali untuk perhitungan iterasi ke-5, dengan langkah yang sama seperti sebelumnya, maka dibawah ini adalah contoh untuk perhitungan *Euclidean Distance* terhadap data ke-1 dan data ke-1025 pada proses iterasi kelima:

1. Data Ke- 1

$$D1(c1) = \sqrt{(125\text{-}128)^2 + (212\text{-}195)^2} = 17$$

$$D1(c2) = \sqrt{(125-137)^2+(212-316)^2} = 105$$

$$D1(c3) = \sqrt{(125-131)^2+(212-248)^2} = 36$$

2. Data Ke- 1025

$$D1025(c1) = \sqrt{(120-128)^2+(188-195)^2} = 11$$

$$D1025(c2) = \sqrt{(120-137)^2+(188-316)^2} = 129$$

$$D1025(c3) = \sqrt{(120-131)^2+(188-248)^2} = 61$$

Tabel X Euclidean Distance dan Matriks Kelompok Data Iterasi 5

C1	C2	C3	Jarak Terpendek	C1	C2	C3
17	105	36	17	T		
14	113	46	14	T		
27	142	75	27	T		
21	113	48	21	T		
99	22	47	22		T	
60	77	31	31			T
124	23	72	23		T	
99	35	50	35		T	
55	69	11	11			T
91	34	39	34		T	
….	….	….	….	….	….	….
28	95	28	28			T
63	59	12	12			T
82	49	34	34			T
62	68	22	22			T
11	129	61	11	T		
			Index	346	230	449

Langkah berikutnya perlu ditentukan kembali titik klaster baru yang dihitung dengan mencari nilai rata-rata berdasarkan dari data anggota masing-masing kelompok klaster menggunakan rumus berikut dibawah ini:

C1 baru ke 6 (*Trestbps*) $= \frac{44.296}{346} = 128$

C1 baru ke 6 (*Chol*) $= \frac{67.549}{346} = 195$

C2 baru ke 6 (*Trestbps*) $= \frac{31.553}{220} = 137$

C2 baru ke 6 (*Chol*) $= \frac{73.019}{220} = 317$

C3 baru ke 6 (*Trestbps*) $= \frac{59.093}{449} = 132$

C3 baru ke 6 (*Chol*) $= \frac{111.582}{449} = 249$

Tabel XI Titik Pusat Awal Klaster Baru ke-6

Titik pusat awal cluster	Trestbps	Chol
Cluster Baru ke -1	128	195
Cluster Baru ke-2	137	317
Cluster Baru ke-3	132	249

Nilai titik klaster baru tersebut digunakan kembali untuk perhitungan iterasi ke-6, dengan langkah yang sama seperti sebelumnya, maka dibawah ini adalah contoh untuk perhitungan *Euclidean Distance* terhadap data ke-1 dan data ke-1025 pada proses iterasi kelima:

Data Ke- 1

$$D1(c1) = \sqrt{(125\text{-}128)^2+(212\text{-}195)^2} = 17$$

$$D1(c2) = \sqrt{(125\text{-}137)^2+(212\text{-}317)^2} = 106$$

$$D1(c3) = \sqrt{(125\text{-}132)^2+(212\text{-}249)^2} = 37$$

Data Ke- 1025

$$D1025(c1) = \sqrt{(120\text{-}128)^2+(188\text{-}195)^2} = 11$$

$$D1025(c2) = \sqrt{(120\text{-}137)^2+(188\text{-}317)^2} = 131$$

$$D1025(c3) = \sqrt{(120\text{-}132)^2+(188\text{-}249)^2} = 62$$

Tabel XII Euclidean Distance dan Matriks Kelompok Data Iterasi 6

C1	C2	C3	Jarak Terpendek	C1	C2	C3
17	106	37	17	T		
14	115	46	14	T		
27	144	76	27	T		
21	115	48	21	T		
99	23	46	23		T	
60	79	32	32			T
124	23	72	23		T	
99	36	50	36		T	
54	71	12	12			T
91	35	39	35		T	
….	….	….	….	….	….	….
28	97	29	28	T		
63	61	12	12			T
82	50	34	34			T
61	69	22	22			T
11	131	62	11	T		
			Index	355	215	455

Langkah berikutnya perlu ditentukan kembali titik klaster baru yang dihitung dengan mencari nilai rata-rata berdasarkan dari data anggota masing-masing kelompok klaster menggunakan rumus berikut dibawah ini:

C1 baru ke 7 (*Trestbps*) $= \frac{45.400}{355} = 128$

C1 baru ke 7 (*Chol*) $= \frac{69.547}{355} = 196$

C2 baru ke 7 (*Trestbps*) $= \frac{29.566}{215} = 138$

C2 baru ke 7 (*Chol*) $= \frac{68.777}{215} = 320$

C3 baru ke 7 (*Trestbps*) $= \frac{59.936}{445} = 132$

C3 baru ke 7 (*Chol*) $= \frac{113.826}{445} = 250$

Tabel XIII Titik Pusat Awal Klaster Baru ke-7

Titik pusat awal cluster	Trestbps	Chol
Cluster Baru ke -1	128	196
Cluster Baru ke-2	138	320
Cluster Baru ke-3	132	250

Nilai titik klaster baru tersebut digunakan kembali untuk perhitungan iterasi ke-7, dengan langkah yang sama seperti sebelumnya, maka dibawah ini adalah contoh untuk perhitungan *Euclidean Distance* terhadap data ke-1 dan data ke-1025 pada proses iterasi kelima:

Data Ke- 1

$$D1(c1) = \sqrt{(125-128)^2+(212-196)^2} = 16$$

$$D1(c2) = \sqrt{(125-138)^2+(212-320)^2} = 109$$

$$D1(c3) = \sqrt{(125-132)^2+(212-250)^2} = 39$$

Data Ke- 1025

$$D1025(c1) = \sqrt{(120-128)^2+(188-196)^2} = 11$$

$$D1025(c2) = \sqrt{(120-138)^2+(188-320)^2} = 133$$

$$D1025(c3) = \sqrt{(120-132)^2+(188-250)^2} = 63$$

Tabel XIV Euclidean Distance dan Matriks Kelompok Data Iterasi 7

C1	C2	C3	Jarak Terpendek	C1	C2	C3
16	109	39	16	T		
14	117	48	14	T		
28	146	77	28	T		
21	117	50	21	T		
99	26	44	26		T	
59	81	32	32			T
123	24	70	24		T	

C1	C2	C3	Jarak Terpendek	C1	C2	C3
98	38	48	38		T	
54	73	12	12			T
90	37	37	37			T
….	….	….	….	….	….	….
28	99	30	28	T		
62	63	10	10			T
81	53	33	33			T
61	71	22	22			T
11	133	63	11	T		
			Index	355	208	462

Langkah berikutnya perlu ditentukan kembali titik klaster baru yang dihitung dengan mencari nilai rata-rata berdasarkan dari data anggota masing-masing kelompok klaster menggunakan rumus berikut dibawah ini:

C1 baru ke 8 (*Trestbps*) $= \frac{45.400}{355} = 128$

C1 baru ke 8 (*Chol*) $= \frac{69.547}{355} = 196$

C2 baru ke 8 (*Trestbps*) $= \frac{28.628}{208} = 138$

C2 baru ke 8 (*Chol*) $= \frac{66.784}{208} = 321$

C3 baru ke 8 (*Trestbps*) $= \frac{60.874}{462} = 132$

C3 baru ke 8 (*Chol*) $= \frac{115.819}{462} = 251$

Tabel XV Titik Pusat Awal Klaster Baru ke-8

Titik pusat awal cluster	Trestbps	Chol
Cluster Baru ke -1	128	196
Cluster Baru ke-2	138	321
Cluster Baru ke-3	132	251

Nilai titik klaster baru tersebut digunakan kembali untuk perhitungan iterasi ke-8, dengan langkah yang sama seperti sebelumnya, maka dibawah ini adalah contoh untuk perhitungan *Euclidean Distance* terhadap data ke-1 dan data ke-1025 pada proses iterasi kelima:

1. Data Ke- 1

$$D1(c1) = \sqrt{(125-128)^2+(212-196)^2} = 16$$

$$D1(c2) = \sqrt{(125-138)^2+(212-321)^2} = 110$$

$$D1(c3) = \sqrt{(125-132)^2+(212-251)^2} = 39$$

2. Data Ke- 1025

$$D1025(c1) = \sqrt{(120-128)^2+(188-196)^2} = 11$$

$$D1025(c2) = \sqrt{(120-138)^2+(188-321)^2} = 134$$

$$D1025(c3) = \sqrt{(120-132)^2+(188-251)^2} = 64$$

Tabel XVI Euclidean Distance dan Matriks Kelompok Data Iterasi 8

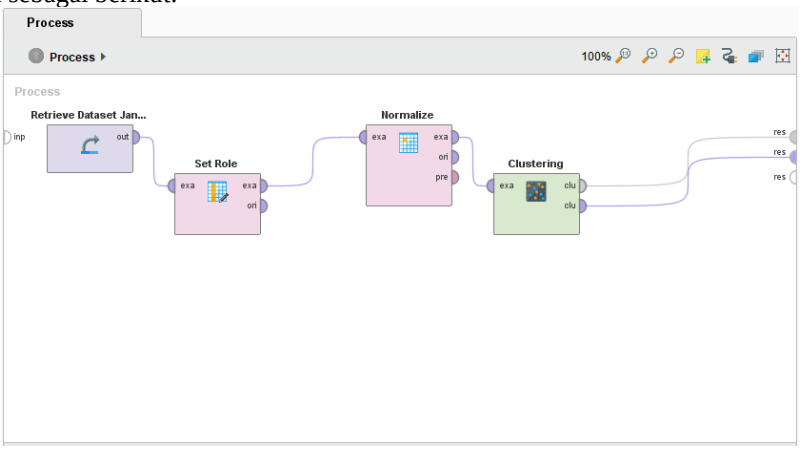
C1	C2	C3	Jarak Terpendek	C1	C2	C3
16	110	39	16	T		
14	118	48	14	T		
28	147	78	28	T		
21	119	50	21	T		
99	27	44	27		T	
59	82	32	32			T
123	24	70	24		T	
98	39	48	39		T	
54	74	12	12			T
90	38	37	37			T
….	….	….	….	….	….	….
28	100	31	28	T		
62	64	10	10			T
81	54	33	33			T
61	73	22	22			T
11	134	64	11	T		
			Index	355	208	462

Tabel XVII Kelompok Klaster

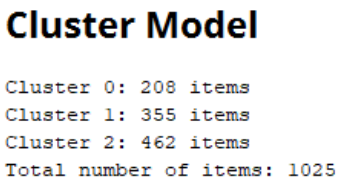
Age	Kategori Klaster
52	low
53	low
70	low
61	low
62	medium
58	high
58	medium
55	medium
46	high
54	high
….	….
59	low
60	high

Age	Kategori Klaster
47	high
50	high
54	low

Pada proses ini metode klasterisasi dengan algoritma K-Means diterapkan untuk pembentukan kelompok klaster dengan keakurasian yang tepat . Dalam penelitian ini menggunakan pengujian perhitungan dengan *tools* Rapid Miner, hasil pengujian yang didapat dengan menggunakan *tools* Rapid Miner adalah dengan tahapan langkah - langkah sebagai berikut:



Gambar 2 Proses Rapid Miner Dengan Algoritma K-Means



Gambar 3 Cluster Model dari 1025 record data

Gambar 4 Grafik Scatter Plot dari Klaster yang terbentuk

V. CONCLUSION

Dari hasil pengujian yang telah dilakukan dalam penelitian ini, maka dapat diambil suatu kesimpulan yaitu melalui beberapa tahapan didapatkan hasil bahwa proses *clustering* dengan algoritma K-Means berhenti pada iterasi ke-8, karena posisi objek dari masing – masing *cluster* sudah tidak berubah dan mendapatkan nilai yang optimal. Metode klasterisasi tersebut di proses dengan algoritma K-Means yang dimana hasilnya juga menunjukkan sebuah wawasan baru yaitu pengelompokkan tingkat risiko penyakit jantung berdasarkan usia dengan 3 *cluster*. *Cluster* 1 merupakan kategori usia dengan tingkat resiko penyakit jantung cukup rendah atau *Low* yaitu 355 dari 1025 kategori usia yang diuji, kemudian *cluster* 2 adalah kategori usia dengan tingkat resiko penyakit jantung sedang atau *Medium* yaitu 208 dari 1025 kategori usia yang diuji, dan terakhir adalah *cluster* 3 merupakan kategori usia dengan tingkat kategori usia cukup tinggi atau *High* yaitu 462 dari 1025 kategori usia yang diuji. Pengujian menggunakan *tools* RapidMiner juga dapat menghasilkan wawasan yang serupa yaitu masing-masing klaster memilik anggota kelompok klaster sesuai dengan perhitungan manual seperti Cluster_0 pada Rapid Miner memiliki 208 anggota klaster yang merepresentasikan klaster *Medium*, Cluster_1 memiliki 355 anggota kelompok klaster sebagai representasi klaster *Low*, dan Cluster_2 memiliki 462 anggota klaster yang sesuai dengan representasi klaster *High*.

REFERENCES

[1] A. Rohman, “Komparasi Metode Klasifikasi Data Mining Untuk Prediksi Penyakit Jantung,” *Neo Tek.*, vol. 2, no. 2, pp. 21–28, 2017, doi: 10.37760/neoteknika.v2i2.766.

[2] C. S. Nur Aeni Widiastuti, Stefanus Santosa, “ALGORITMA KLASIFIKASI DATA MINING NAÏVE BAYES BERBASIS PARTICLE SWARM OPTIMIZATION UNTUK DETEKSI PENYAKIT JANTUNG,” *Pseudocode*, vol. 1, no. 1, pp. 11–14, 2014, doi: 10.1038/nmeth.f.284.

[3] D. P. Utomo and M. Mesran, “Analisis Komparasi Metode Klasifikasi Data Mining dan Reduksi Atribut Pada Data Set Penyakit Jantung,” *J. Media Inform. Budidarma*, vol. 4, no. 2, pp. 437–444, 2020, doi: 10.30865/mib.v4i2.2080.

[4] D. M. U. Atmaja and R. Mandala, “Analisa Judul Skripsi untuk Menentukan Peminatan Mahasiswa Menggunakan Vector Space Model dan Metode K-Nearest Neighbor,” *IT Soc.*, vol. 4, no. 2, pp. 1–6, 2020, doi: 10.33021/itfs.v4i2.1182.

[5] D. Marini *et al.*, “Pembangunan Sistem Informasi Biaya Proyek pada PT. Skyline Semesta Menggunakan Metode Earned Value Management (EVM),” *Semin. Nas. Apl. Teknol. Inf.*, pp. 11–2018, 2018.

[6] D. Haryadi, “Penerapan Algoritma K-Means Clustering Pada Produksi Perkebunan Kelapa Sawit Menurut Provinsi,” *J. Informatics Commun. Technol.*, vol. 3, no. 1, pp. 50–64, 2021.

[7] D. Haryadi and R. Mandala, “Prediksi Harga Minyak Kelapa Sawit Dalam Investasi Dengan Membandingkan Algoritma Naïve Bayes, Support Vector Machine dan K-Nearest Neighbor,” *IT Soc.*, vol. 4, no. 1, pp. 28–38, 2019, doi: 10.33021/itfs.v4i1.1181.

[8] D. Aprilla, C, B. Donny, Aji, A. Lia, and W. I, Wayan, Simri, “DATA MINING dengan RAPID MINER,” *Produk-produk perangkat lunak gratis dan bersifat open source yang demikian banyak jumlahnya, telah memudahkan kita dalam melakukan proses Pengolah. dan Anal. data. Dalam melakukan Anal. terhadap data mining, RapidMiner merupakan salah sat*, pp. 1–128, 2013.