Institut Riset dan Publikasi Indonesia (IRPI)
MALCOM: Indonesian Journal of Machine Learning and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/malcom
Vol. 4 Iss. 4 October 2024, pp: 1568-1577
ISSN(P): 2797-2313 | ISSN(E): 2775-8575

Grouping of ENT Disease Data Using K-Means Clustering Algorithm

Pengelompokan Data Penyakit THT Menggunakan Algoritma K-Means Clustering

Minta Ito Hutagalung[1*], Sriani[2]

[1,2]Program Studi Ilmu Komputer, Universitas Islam Negeri Sumatera Utara, Indonesia

E-Mail: [1]mintaitohutagalung2231@gmail.com, [2]sriani@uinsu.ac.id

Received Aug 27th 2024; Revised Oct 6th 2024; Accepted Oct 12th 2024
Corresponding Author: Minta Ito Hutagalung

Abstract

Disturbed body balance can cause various types of diseases, including diseases of the nose, ears, and throat (ENT). ENT diseases are considered dangerous because they attack several vital human organs that function to hear, breathe, and swallow. Bacterial or viral infections are often the cause of this disease, but some are caused by abnormalities in cell growth that can develop into tumors or cancer. The main problem in this study is how to group ENT disease patient data based on age using the K-Means Clustering method. This study was conducted at Gunung Tua Hospital, involving 51 patient data processed using Python. The results showed that patients could be divided into three groups (3 clusters) based on age and type of disease, with cluster 0 consisting of elderly patients who were more susceptible to diseases such as sinusitis with a total of 10 data, then cluster 1 consisting of younger patients with milder disease diagnoses with a total of 19 data. Meanwhile, cluster 2 includes patients with varying age ranges and more diverse diagnoses, such as OMSK and Allergic Rhinitis with a total of 22 data. The results of the study showed that the K-Means Clustering algorithm was able to group patients well, and evaluation using DBI produced a value of 0.90, which indicates that the quality of the resulting cluster is quite good.

Keyword: Data Mining, Elbow Method, ENT Diseases, K-Means Clustering, Machine Learning

Abstrak

Keseimbangan tubuh yang terganggu dapat menyebabkan berbagai jenis penyakit, termasuk penyakit pada Hidung, Telinga, dan Tenggorokan (THT). Penyakit THT dianggap berbahaya karena menyerang beberapa organ vital manusia yang berfungsi untuk mendengar, bernapas, dan menelan. Infeksi bakteri atau virus sering menjadi penyebab penyakit ini, namun ada juga yang disebabkan oleh kelainan pertumbuhan sel yang dapat berkembang menjadi tumor atau kanker. Permasalahan utama dalam penelitian ini adalah bagaimana mengelompokkan data pasien penyakit THT berdasarkan umur menggunakan Algoritma K-Means Clustering. Penelitian ini dilakukan di RSUD Gunung Tua, dengan melibatkan 51 data pasien yang diolah menggunakan Python. Hasil penelitian menunjukkan bahwa pasien dapat dibagi menjadi tiga kelompok (3 cluster) berdasarkan usia dan jenis penyakit, dengan cluster 0 terdiri dari pasien usia lanjut yang lebih rentan terhadap penyakit seperti sinusitis dengan total 10 data, kemudian cluster 1 terdiri dari pasien yang lebih muda dengan diagnosis penyakit yang lebih ringan dengan total 19 data. Sedangkan cluster 2 mencakup pasien dengan rentang umur yang bervariasi dan diagnosis yang lebih beragam, seperti OMSK dan Rhinitis Alergi dengan total 22 data. Hasil penelitian menunjukkan bahwa algoritma K-Means Clustering mampu mengelompokkan pasien dengan baik, dan evaluasi menggunakan DBI menghasilkan nilai 0.90, yang menandakan bahwa kualitas cluster yang dihasilkan cukup baik.

Kata Kunci: K-Means Clustering, Metode Siku, Pembelajaran Mesin, Penambangan Data, Penyakit THT

1. PENDAHULUAN

Keseimbangan tubuh tidak dapat dipertahankan yang menyebabkan terjadinya penyakit. Terdapat beragam jenis penyakit yang dialami seseorang, di antaranya penyakit pada hidung atau penyakit Hidung, Telinga, dan Tenggorokan (THT). Sangat banyak keluhan penyakit THT terutama pada hidung yang ditemui dalam kehidupan sehari-hari, seperti sinusitis dan polip hidung. Sinus biasanya terletak di saluran pernapasan manusia, tepatnya di rongga hidung, dan menyebabkan masalah pada sistem pernapasan karena cairan menumpuk pada organ sinus [1]. Saat ini penyakit THT menjadi penyakit yang cukup banyak diderita oleh masyarakat di dunia termasuk di Indonesia, hal ini dikarenakan di Indonesia penderita penyakit THT berjumlah

sekitar 190-230 per 1000 penduduk atau sekitar 38.4% [3]. Selain itu, berdasarkan data yang diperoleh dari Puskesmas Kabupaten Bekasi bahwa penyakit saluran pernapasan menduduki 10 penyakit terbanyak di tahun 2022. Penyakit saluran pernapasan tersebut seperti acute upper respiratory infection unspecified dengan persentase 16.95% atau sekitar 47.709 penduduk dan acute nasopharyngitis (common cold) dengan persentase 9,52% atau sekitar 26.801 penduduk [2].

Penyakit THT merupakan penyakit yang disebabkan adanya gangguan pada organ telinga, hidung, dan tenggorokan. Karena organ tersebut sangat vital, maka gangguan yang terjadi pada ketiga organ tersebut atau salah satunya dapat berbahaya. Penyakit ini dapat menyerang berbagai usia, mulai dari bayi, anak-anak, hingga orang dewasa. Telinga merupakan organ yang berfungsi untuk mendengar dan mekanisme keseimbangan [3]. Hidung merupakan organ yang berfungsi sebagai indra penciuman dan bagian terluar dari sistem pernapasan [4]. Tenggorokan merupakan organ yang di dalamnya terdapat saluran pita suara, menelan, dan bernapas. Penyakit THT termasuk penyakit yang berbahaya karena menyerang beberapa organ vital manusia. Penyakit di sekitar hidung, telinga, dan tenggorokan biasanya disebabkan oleh infeksi kuman, tetapi banyak pula yang diakibatkan oleh kelainan perkembangan sel tubuh, yang kemudian menjadi tumor atau kanker . Penyakit THT seringkali dianggap remeh oleh sebagian masyarakat, sehingga kurangnya penanganan dan akibatnya akan membuat penyakit sebelumnya lebih parah atau menimbulkan penyakit-penyakit yang lain . Hal tersebut terjadi karena kurangnya informasi yang ada. Seringkali kita telah mengetahui bahwa tubuh kita mengalami gangguan kesehatan berdasarkan gejala-gejala yang dirasakan, namun belum mengetahui pasti penyakit apa yang sedang menyerang tubuh kita serta bagaimana cara mengobatinya [5].

Permasalahan utama dalam penelitian ini adalah kurangnya pemahaman masyarakat tentang penyakit THT, yang kerap diabaikan meskipun berpotensi berbahaya. Tanpa pengetahuan yang memadai tentang gejala dan penanganannya, penyakit THT sering tidak ditangani serius, yang dapat memperburuk kondisi pasien. Untuk itu, diperlukan analisis data untuk mengelompokkan jenis penyakit THT berdasarkan pola tertentu agar informasi yang tersembunyi dapat diungkap [6]. Untuk mengatasi masalah ini, diperlukan analisis data yang dapat mengelompokkan penyakit THT berdasarkan kemiripan karakteristiknya agar informasi yang selama ini tersembunyi dapat diperoleh dan dipahami [7] dengan melakukan pengelompokan penyakit THT menggunakan Algoritma K-Means Clustering yang dimana merupakan salah satu metode pada data mining yang digunakan untuk mengelompokan data menjadi beberapa cluster [8]. Oleh karena itu, Algoritma K-Means Clustering dipilih sebagai solusi untuk mengelompokkan data penyakit THT dalam beberapa klaster berdasarkan pola atau kesamaan karakteristik tertentu [9]. K-Means sangat cocok untuk kasus ini karena mampu mengelompokkan data secara non-hirarki, di mana tujuan utamanya adalah meminimalkan jarak antar-objek dalam setiap klaster dan memaksimalkan jarak antar-klaster, sehingga kelompok penyakit yang terbentuk lebih terstruktur dan mudah dianalisis [10].

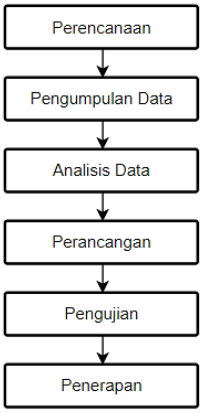
Tinendung (2023), menggunakan data bayi balita di posyandu dalam clustering atau pengelompokan status staunting menggunakan algoritma k- means dengan atribut atau coloumn yang digunakan adalah nama, usia, berat dan tinggi, berat badan umur (bb/u), tinggi badan umur (tb/u), berat badan tinggi badan (bb/tb) maka didapatkan hasil bahwa sebanyak 31 anak menderita status stunting atau cluster 1, cluster 0 atau status normal ada sebanyak 43 anak dan cluster 2 atau status pertumbuhan cepat dengan jumlah anak sebanyak 27 maka dapat disimpulkan bahwa kurang lebih 30% anak menderita status stunting dari 101 data [11]. Kemudian penelitian Yudistira (2023), Algoritma K-Means clustering menunjukan bahwa berdasarkan hasil cluster data siswa menggunakan dataset siswa dalam satu semester, maka didapatkan cluster 0 berjumlah 59 siswa, cluster 1 berjumlah 94 siswa, dan cluster 2 berjumlah 1 siswa. Hasil pengujian menggunakan elbow method maka jumlah cluster yang baik yang digunakan adalah 3 cluster, sehingga dalam penelitian ini menggunaka 3 cluster yaitu cluster 0, cluster 1, dan cluster 2 [12].

Pa (2023) juga memperoleh hasil diagnosa, usia, kelompok yang memiliki himpunan/nilai paling tinggi dan paling banyak data hasil diagnosa pasien, yaitu pada Cluster 1 berjumlah 825 data hasil diagnosa pasien Badan Penyelenggara Jaminan Sosial (BPJS) yang menggunakan BPJS pada Jenis penyakit Demam Tinggi dan hasil diagnosanya adalah Asma, kemudian usia adalah>60 Tahun. Pada cluster 3 pada group jenis penyakit adalah dengan hasil diagnosa, kemudian usia,yang memiliki himpunan/nilai paling tinggi dan paling menengah data hasil diagnosa pasien, yaitu pada Cluster 1 berjumlah 519 data hasil diagnosa pasien, kelompok data pasien hasil diagnosa yang menggunakan BPJS pada Jenis penyakit Demam Tinggi dan hasil diagnosanya adalah Anemia, kemudian usia yang digunakan adalah>60 Tahun [13].

Perbedaan utama penelitian ini dengan studi sebelumnya adalah fokus pada pengelompokan penyakit THT berdasarkan usia pasien di RSU Gunung Tua, yang belum dilakukan sebelumnya. Sementara studi Pa (2023) mengelompokkan penyakit umum seperti Demam Tinggi dan Anemia berdasarkan penggunaan layanan BPJS, penelitian ini bertujuan mengidentifikasi pola spesifik penyakit THT sesuai usia, yang dapat membantu rumah sakit dalam perencanaan penanganan yang lebih tepat. Dengan menggunakan K-Means Clustering dan tools yang digunakan ialah Google Colabs berbasis Python [14], penelitian ini menawarkan kebaruan dalam pengelompokan data penyakit THT secara efisien dan terotomasi. [15].

2. METODOLOGI PENELITIAN

Dalam melakukan penelitian dibutuhkan suatu proses penyelidikan yang tersusun secara sistematis yang ditujukan pada penyampaian informasi dalam menyelesaikan suatu permasalahan. Oleh karena itu dibutuhkan kerangka penelitian agar penelitian tersebut terarah dan beraturan untuk mencapai penelitian yang sangat baik. Atau dengan kata lain ini merupakan gambaran umum terkait alur penelitian yang akan dilakukan. Perancangan dalam kerangka kerja penelitian ini dilakukan dengan tahap-tahap yang ditunjukkan pada gambar 1.



Gambar 1. Kerangka Kerja Penelitian

2.1. Perencanaan

Pada tahap ini merupakan awal untuk menentukan permasalahan sebelum melakukan penelitian pada objek penelitian. Dengan mencari sumber informasi masalah pada objek penelitian untuk mencari penyelesaian berkaitan dengan permasalahan. Sehingga dapat menguraikan masalah dan memudahkan langkah dalam menyelesaikan masalah dan didalam perencanaan ini juga peneliti sudah merencanakan tujuannya.

2.2. Pengumpulan Data

Studi literatur dilakukan untuk mengumpulkan pengetahuan dari berbagai macam sumber literatur berupa buku-buku, jurnal dan karya ilmiah lainnya yang berkaitan dengan topik yang penulis angkat. Dan wawancara merupakan teknik pengumpulan data yang akan dilakukan peneliti. Wawancara yang dilakukan langsung dengan narasumber selaku perawat THT di Rumah Sakit Umum Daerah (RSUD) Gunung Tua.

2.3. Preprocessing Data

Preprocessing Data adalah upaya mendapatkan data siap pakai. Sebelum proses data mining dapat dilaksanakan, perlu dilakukan proses cleaning pada data yang menjadi fokus Knowledge Discovery In Database (KDD) [16]. Preprocessing data meliputi transformasi, seleksi, dan normalisasi data. Sebagai contoh penerapan normalisasi data menggunakan Min-Max Scaler, ditunjukkan pada persamaan 1.

$$X' = \frac{X - Xmin}{Xmax - Xmin} \quad (1)$$

Dimana:

X : Nilai aktual dari suatu kriteria
X_{min} : Nilai minimum dari kriteria
X_{max} : Nilai maksimum dari kriteria
X' : Nilai normalisasi

2.4. Perancangan

Berdasarkan analisis yang dilakukan, maka penulis akan menggali data untuk pengelompokan dengan kriteria (variabel) yang telah ditentukan. Adapun kriteria yang diambil dari penelitian tersebut adalah umur pada pasien. Kemudian dari data tersebut dimasukan kedalam Microsoft Excel dan diolah menggunakan Python dengan tools Jupyter Notebook. Adapun *flowchart* algoritma K-Means ditunjukkan pada gambar 2.

Teknik menghitung jarak yang digunakan adalah Euclidean Distance dimana digunakan untuk menggambarkan tingkat kemiripan antara dua objek atau lebih. Semakin dekat jaraknya, semakin banyak objek yang termasuk dalam kelompok yang sama, berikut rumus jarak euclidean distance pada persamaan 2.

$$D = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2)$$

Dimana:

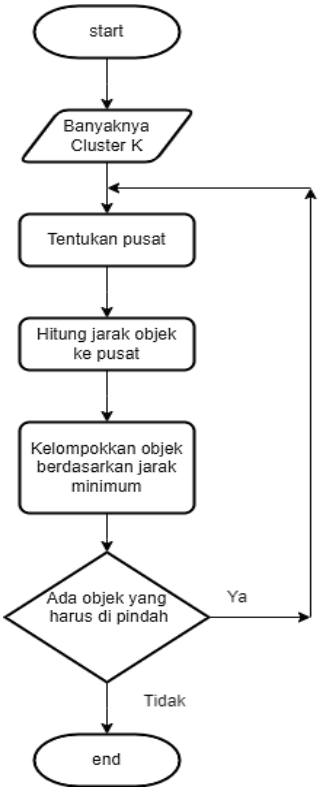
D : Euclidean Distance
x_i : Atribut pada data pertama
y_i : Atribut pada data kedua

Adapun rumus yang digunakan algoritma K-Means ditunjukkan pada persamaan 3.

$$d_{ik} = \sqrt{\text{£}_j^m} \ (c_{ij} - c_{kj}) \quad (3)$$

Dimana :

C_{ik} : Pusat cluster
C_{kj} : Data



Gambar 2. Flowchart K-Means Clustering

2.5. Pengujian

Pada tahap ini, peneliti melakukan pengujian setelah mendapatkan data untuk mengetahui apakah data yang telah diolah sesuai dengan yang diharapkan. Data yang diujikan dimasukkan dan diolah menggunakan python dengan tools Jupyter Notebook [17].

2.6 Penerapan

Penerapan algortima K-Means clustering pada penyakit THT untuk mengelompokan data ke dalam kelompok yang sama berdasarkan umur dan jenis penyakit. Data ini kemudian dimasukkan ke dalam format yang sesuai untuk analisis clustering. Penerapan algoritma ini diharapkan dapat mempermudah dalam membuat diagnosis yang lebih akurat tentang jenis penyakit pada THT [18].

3. HASIL DAN PEMBAHASAN

Penelitian ini akan membahas pengelompokan data penyakit Telinga, Hidung, dan Tenggorokan (THT) menggunakan Algoritma K-Means clustering. Dengan pendekatan ini, diharapkan dapat diidentifikasi kelompok-kelompok penyakit yang memiliki kesamaan karakteristik, sehingga dapat membantu dalam proses diagnosis dan pengambilan keputusan medis yang lebih tepat. K-Means akan membagi data ke dalam beberapa cluster yang optimal, berdasarkan kemiripan data dalam setiap kelompok.

3.1. Representasi Data

Penelitian ini menggunakan data pasien THT sebanyak 51 data, yang mencakup berbagai informasi penting seperti nama, usia, diagnosis dan Lama Perawatan. Data tersebut digunakan untuk menganalisis pola penyakit dan melakukan pengelompokan berdasarkan kesamaan karakteristik dengan Algoritma K-Means clustering, yang dapat dilihat pada Tabel 1.

Table 1. Data Pasien THT 2023

No	Nama	Umur	Diagnosa	Lama Perawatan (Hari)
1	FRL	58	Rhintis Alergic	4
2	PH	59	Sinusitis	3
3	PD	11	OMSK	3
4	P	56	OMSK	5
5	DH	30	Impacted serumen	3
6	PN	14	Rhintis Alergic	4
7	Z	5	Impacted serumen	3
8	AAR	5	OMA	4
9	ZH	50	Sinusitis	5
10	AR	12	OMA	3
…	…	…	…	…
51	SLW	45	Impacted serumen	3

3.2. Transformasi Data

Dalam penerapan metode clustering, langkah pertama yang dilakukan adalah mengubah data menjadi bentuk numerik dengan menggunakan kode-kode yang telah ditentukan. Proses ini dapat dijelaskan lebih lanjut pada kelompok di bawah ini:

Kelompok Usia, meliputi pengelompokan data pasien berdasarkan usia, ada 4 kategori yaitu :

1. <19 tahun, mencakup pasien yang berusia dibawah 19 tahun
2. 20-39 tahun mencakup pasien berusia antara 20-39 tahun
3. 40-59 tahun, mencakup pasien berusia antara 40-59 tahun
4. >60 tahun, mencakup pasien berusia diatas 60 tahun

Pengelompokan ini bertujuan untuk mempermudah analisis pola penyakit THT yang terjadi pada berbagai kelompok usia. Kemudian jenis diagnosa yang diberikan kepada pasien dengan penyakit THT. Diagnosa yang tercantum meliputi:

1. Impacted Serum, kondisi di mana terdapat penumpukan cairan di telinga.
2. OMA (Otitis Media Akut), infeksi telinga tengah yang sering terjadi pada anak-anak.
3. OMSK (Otitis Media Sekunder), infeksi telinga yang terjadi sebagai komplikasi dari infeksi lain.
4. Rhinitis Alergik, peradangan pada hidung akibat reaksi alergi.
5. Sinusitis, peradangan pada sinus yang dapat menyebabkan nyeri dan kesulitan bernapas.

Deskripsi ini bertujuan untuk memberikan gambaran mengenai variasi penyakit yang umum terjadi pada pasien dengan keluhan THT. Selanjutnya lama perawatan yang diberikan kepada pasien berdasarkan kategori waktu. Lama perawatan yang tercantum adalah 2-5 hari.

Setelah kriteria di atas dikodekan, langkah berikutnya adalah mentransformasikan data sesuai dengan kriteria tersebut untuk dihitung menggunakan metode clustering. Proses ini bertujuan untuk mempermudah pengelompokan data berdasarkan kesamaan atau perbedaan yang teridentifikasi dari hasil transformasi, sehingga pembentukan cluster menjadi lebih efektif dan akurat. Data yang telah ditransformasikan akan diproses lebih lanjut untuk menentukan pola atau kelompok yang relevan.

3.3. Normalisasi Data

Normalisasi menggunakan metode Min-Max Scaling, yang mengubah nilai setiap parameter ke rentang [0, 1]. Proses ini penting untuk mencegah bias dalam analisis, karena perbedaan skala antarparameter dapat mempengaruhi hasil clustering. Dengan normalisasi, setiap parameter memberikan kontribusi yang seimbang, sehingga hasil pengelompokan lebih akurat dan adil. Merujuk dari persamaan 1, hasil normalisasi dapat ditunjukkan pada tabel 2.

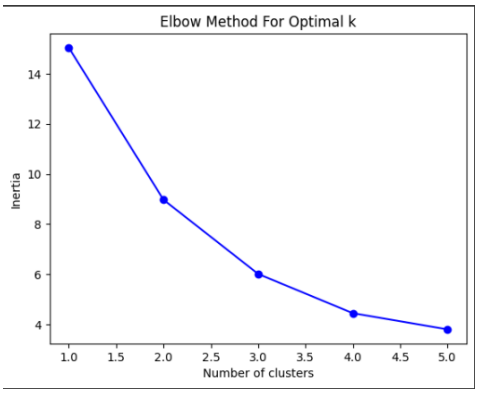
Table 2. Normalisasi Data

No	Inisial	X	Y	Z
1	FRL	0,66	0,75	0,66
2	PH	0,66	1,00	0,33

No	Inisial	X	Y	Z
3	PD	0,00	0,50	0,33
4	P	0,66	0,50	1,00
5	DH	0,33	0,00	0,33
6	PN	0,00	0,75	0,66
7	Z	0,00	0,00	0,33
8	AAR	0,00	0,25	0,66
9	ZH	0,66	1,00	1,00
10	AR	0,00	0,25	0,33
11	RUH	0,00	0,00	0,33
12	MS	0,00	0,75	1,00
13	SP	0,66	0,75	0,66
14	KN	0,66	0,75	0,66
15	MS	1,00	0,00	0,33
16	NS	0,00	0,75	1,00
17	RL	0,33	0,50	1,00
18	HD	1,00	0,50	0,33
19	YA	0,00	0,25	1,00
20	NZD	0,00	1,00	0,66

3.4. Penerapan Algoritma K-Means Clustering

Dalam penelitian ini, Algoritma K-Means Clustering diterapkan untuk mengelompokkan pasien dengan penyakit THT. Salah satu tantangan utama dalam penggunaan K-Means Clustering adalah menentukan jumlah cluster yang optimal. Untuk mengatasi masalah ini, peneliti menggunakan Algoritma Elbow, yang membantu mengidentifikasi titik di mana penambahan cluster tidak lagi memberikan peningkatan signifikan dalam variasi total yang dijelaskan oleh model [19]. Dengan pendekatan ini, jumlah cluster yang optimal dapat dipilih secara lebih tepat, sehingga analisis pengelompokan menjadi lebih akurat dan efektif. Grafik Metode Elbow dapat dilihat pada gambar 3.



Gambar 3. Grafik Metode Elbow

Berdasarkan gambar 3, penurunan grafik mulai melambat setelah k = 3. Ini menunjukkan bahwa tiga cluster merupakan jumlah optimal untuk data Penyakit THT yang telah dinormalisasi. Dengan memilih k = 3, kita dapat memastikan bahwa cluster yang terbentuk tidak hanya meminimalkan variasi dalam data, tetapi juga mencegah terjadinya overfitting akibat penggunaan jumlah cluster yang terlalu banyak. Setelah jumlah cluster ditentukan, langkah selanjutnya adalah menetapkan centroid awal secara acak yang akan digunakan dalam inisialisasi centroid. Pada kasus ini, centroid awal diambil dari hasil normalisasi data, dengan C0 berasal dari data ke-1, C1 dari data ke-5, dan C2 dari data ke-12. Dapat dilihat pada tabel 3.

Table 3. Centroid Awal

Centroid Awal	Umur	Diagnosa	Lama Perawatan
Centroid 0	0,666	0,750	0,666
Centroid 1	0,333	0,000	0,333
Centroid 2	0,000	0,750	1,000

Setelah menetapkan centroid awal, langkah berikutnya adalah menghitung jarak antara setiap centroid dengan data menggunakan rumus Euclidean Distance seperti yang ditunjukkan di bawah ini.

Iterasi 1:

1. Menghitung jarak centroid terdekat data ke 1 pada centroid 0 dengan nilai atribut (0,66667; 0,75; 0,66667)

$$C0 = \sqrt{(0{,}66667 - 0{,}66667)^2 + (0{,}75 - 0{,}75)^2 + (0{,}66667 - 0{,}66667)^2} = 0$$

2. Menghitung jarak centroid terdekat data ke 1 pada centroid 1 dengan nilai atribut (0,33333; 0; 0,33333)

$$C1 = \sqrt{(0{,}33333 - 0{,}66667)^2 + (0 - 0{,}75)^2 + (0{,}33333 - 0{,}66667)^2} = 0{,}885850502$$

3. Menghitung jarak centroid terdekat data ke 1 pada centroid 2 dengan nilai atribut (0;0,75;1)

$$C2 = \sqrt{(0 - 0{,}66667)^2 + (0{,}75 - 0{,}75)^2 + (1 - 0{,}66667)^2} = 0{,}745357483$$

Proses ini dilakukan sampai pada data terakhir. Dari hasil perhitungan jarak centroid terdekat di atas, diperoleh hasil Iterasi 1 yang ditampilkan pada tabel 4.

Table 4. Hasil Iterasi 1

Data	Jarak Centroid Itrasi 1			Cluster
	C0	C1	C2	
1	0,0000	0,8858	0,7453	0
2	0,41667	1,0540	0,9753	0
3	0,78617	0,6009	0,7120	1
4	0,41666	0,8975	0,7120	0
5	0,88585	0,0000	1,0573	1
6	0,66667	0,8858	0,3333	2
7	1,05738	0,3333	1,0034	1
8	0,83333	0,5335	0,6009	1
9	0,41666	1,2472	0,7120	0
10	0,89753	0,4166	0,8333	1
…	…	…	…	
51	0,71200	1,1055	0,4166	2

Iterasi ini terus dilakukan sampai tidak terjadi perubahan pengelompokkan data maka proses pengolahan data akan dihentikan. Hasil dari pengolahan data didapatkan cluster 0 sebanyak 10 sampel, cluster 1 sebanyak 19 sampel dan cluster 2 sebanyak 22 sampel. Dapat dilihat pada tabel 5.

Table 5. Hasil Clustering

No	Cluster	Jumlah Cluster
1	C0	10
2	C1	19
3	C2	22

3.5 Penerapan K-Means Clustering pada python

Beberapa bagian proses dari pengelompokan data dengan mengunakan python programming dapat dilihat pada step berikut:

1. Import Data
   Pada tahap ini data yang akan di import berupa data pasien THT yang berjumlah 51 data dengan 4 atribut yaitu: Inisial, Umur, Diagnosa dan Lama Perawatan.

2. Seleksi Data
   Pada tahap seleksi data, variabel yang akan digunakan dalam proses clustering dipilih secara cermat. Dalam hal ini, variabel 'Inisial' akan dihapus karena tidak relevan untuk proses clustering

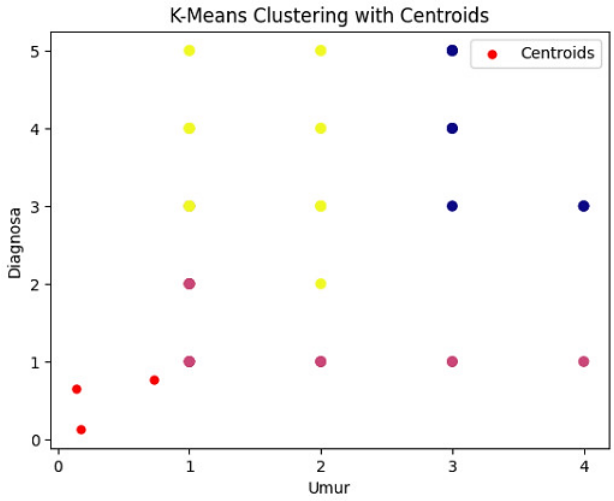
3. Normalisasi Data
   Tujuannya adalah agar data memiliki nilai atribut yang konsisten untuk meningkatkan efisiensi, akurasi, dan interpretasi model analitik.

4. Metode Elbow
   Gambar 3 menunjukkan hasil dari metode Elbow dalam menentukan jumlah kluster optimal untuk analisis K-Means Clustering. Dari grafik, terlihat bahwa nilai inersia menurun seiring bertambahnya jumlah kluster, tetapi penurunan tersebut mulai melambat setelah mencapai sekitar 3 kluster. Titik siku pada grafik menandakan bahwa 3 kluster adalah jumlah yang optimal, di mana model tetap efektif tanpa menjadi terlalu kompleks. Dengan demikian, penggunaan metode Elbow memberikan panduan yang

jelas untuk memilih jumlah kluster yang tepat, yang akan meningkatkan akurasi analisis data dan efektivitas pengelompokan penyakit THT berdasarkan usia pasien.

5. K-Means Label
   Berdasarkan grafik di atas, terlihat bahwa penurunan mulai melambat setelah K = 3, yang menandakan bahwa tiga cluster adalah jumlah yang optimal untuk data penyakit THT. Berikut adalah program yang digunakan untuk proses melabelkan clustering dengan K = 3.

6. Proses Clustering
   Berikut adalah program yang digunakan untuk proses pembuatan grafik clustering dengan K = 3, seperti ditunjukkan pada gambar 4.



Gambar 4. Output Grafik Clustering

Gambar 4 menyajikan hasil dari penerapan Algoritma K-Means Clustering pada data penyakit THT, dengan fokus pada dua variabel utama: umur dan diagnosa pasien. Grafik ini menunjukkan distribusi pasien yang dikelompokkan berdasarkan karakteristik usia dan jenis diagnosa yang berbeda, di mana setiap titik mewakili pasien dengan warna yang menunjukkan kluster yang berbeda. Centroid yang ditandai dengan warna merah berfungsi sebagai titik referensi untuk setiap kluster, mengindikasikan rata-rata dari pasien dalam kelompok tersebut. Hasil grafik ini menggambarkan adanya pola distribusi penyakit berdasarkan umur, yang dapat membantu dalam identifikasi karakteristik penyakit yang umum terjadi pada kelompok usia tertentu.

Penggunaan K-Means Clustering dalam analisis ini memungkinkan pengelompokan data yang lebih sederhana dan terstruktur, sehingga memfasilitasi pemahaman yang lebih baik mengenai hubungan antara umur dan diagnosa. Secara keseluruhan, grafik ini memberikan wawasan yang signifikan untuk meningkatkan penanganan penyakit THT. Dapat disimpulkan dataset terbagi menjadi 3 cluster yaitu cluster 0 (Terdiri dari pasien yang rentang umurnya lebih tua dan diagnosis penyakit tertentu) dan jumlah cluster 0 sebanyak 10 data. cluster 1 (pasien dengan umur lebih muda dengan diagnosis lebih ringan) dan jumlah cluster 1 sebanyak 19 data. Dan cluster 2 (Pasien dengan rentang umur yang bervariasi, tetapi didiagnosis dengan kondisi yang lebih beragam) dan jumlah cluster 2 sebanyak 22 data.

Selanjutnya adalah menghitung hasil DBI untuk mengevalusi hasil Cluster dimana hasil evaluasi cluster didapat adalah 0.90 yang menunjukkan bahwa hasil cluster cukup baik, meskipun nilainya tidak mendekati nol. Biasanya, nilai DBI berkisar antara 0 hingga tak terbatas. Semakin rendah nilai DBI, semakin baik kualitas cluster [20].

4. KESIMPULAN

Berdasarkan hasil clustering menggunakan algoritma K-Means dengan jumlah k=3, dimana Cluster 0, yaitu terdiri dari pasien yang umumnya lebih tua dengan diagnosis penyakit THT tertentu, seperti Rhinitis Alergi atau Sinusitis dengan total 10 data. Kemudian Cluster 1 Meliputi pasien yang lebih muda dengan diagnosis penyakit yang lebih ringan, seperti Impaksi Serumen atau Otitis Media Akut (OMA), dengan total 19 data. Mencakup pasien dengan rentang usia yang lebih beragam serta diagnosis penyakit yang lebih variatif, seperti Otitis Media Supuratif Kronik (OMSK) dan Rhinitis Alergi, dengan total 22 data. Dari hasil tersebut, dapat disimpulkan bahwa Cluster 2 memiliki jumlah pasien terbanyak, yaitu 22 data.

Dalam evaluasi cluster, DBI digunakan untuk mengevaluasi hasil cluster yang digunakan seperti k-means clustering. Nilai ini dihitung berdasarkan rasio antara jarak rata-rata antar anggota cluster dengan jarak antar cluster. Nilai yang dihasilkan adalah 0.90, yang menunjukkan bahwa cluster yang dihasilkan memiliki kualitas yang cukup baik. Semakin rendah nilai DBI (mendekati nol), semakin baik hasil clustering, karena cluster yang terbentuk memiliki jarak antar cluster yang besar dan variasi dalam cluster yang kecil.

Hasil penelitian menunjukkan bahwa algoritma K-Means Clustering mampu mengelompokkan pasien dengan baik, dan evaluasi menggunakan DBI menghasilkan nilai 0,90, yang menandakan bahwa kualitas cluster yang dihasilkan cukup baik. Namun, kelemahan penelitian ini terletak pada terbatasnya jumlah variabel yang digunakan dalam clustering, seperti hanya mempertimbangkan usia dan diagnosis. Untuk penelitian selanjutnya, disarankan untuk memasukkan variabel tambahan, seperti riwayat medis dan tingkat keparahan penyakit, guna menghasilkan cluster yang lebih akurat dan mendalam.

REFERENSIS

[1] M. U. Nuha, A. Fatahillah, and S. Setiawani, "Analisis Numerik Aliran Udara pada Rongga Hidung akibat Penyakit Sinusitis menggunakan Metode Volume Hingga," *Limits J. Math. Its Appl.*, vol. 19, no. 2, p. 217, 2022, doi: 10.12962/limits.v19i2.13683.

[2] R. Pramudita and F. Hibatullah, "Sistem Diagnosis Penyakit THT Berbasis Website Menggunakan Rapid Application Development," *Inf. Syst. Educ. Prof. J. Inf. Syst.*, vol. 9, no. 1, p. 109, 2024, doi: 10.51211/isbi.v9i1.2945.

[3] A. Y. Labolo, A. Anas, B. Betrisandi, and W. Yunus, "Penerapan Metode Fuzzy Mamdani Untuk Mendeteksi Penyakit Telinga Pada Puskesmas Marisa," *Simtek J. Sist. Inf. dan Tek. Komput.*, vol. 7, no. 1, pp. 69–73, 2022, doi: 10.51876/simtek.v7i1.126.

[4] B. N. Rahman, R. Maulana, and F. Utaminingrum, "Sistem Pendeteksi Penyakit Sinusitis berdasarkan Kondisi Ingus dan Suhu Tubuh menggunakan Support Vector Machine (SVM)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 6, no. 2, pp. 545–551, 2022, [Online]. Available: http://j-ptiik.ub.ac.id

[5] M. P. Sari and A. Desiani, "Diagnosa Penyakit THT (Telinga, Hidung, Tenggorokan) menggunakan Metode Certainty Factor pada Sistem Pakar," *J. Artif. Intell. Softw. Eng.*, vol. 3, no. 1, p. 7, 2023, doi: 10.30811/jaise.v3i1.3902.

[6] E. Yolanda, "Penerapan Algoritma K-Means Clustering Untuk Pengelompokan Data Pasien Rehabilitasi Narkoba," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 1, pp. 182-`191, 2023, doi: 10.30865/klik.v4i1.1107.

[7] J. Dongga, A. ` Sarungallo, N. Koru, and G. Lante, "Implementasi Data Mining Menggunakan Algoritma Apriori Dalam Menentukan Persediaan Barang (Studi Kasus: Toko Swapen Jaya Manokwari)," *G-Tech J. Teknol. Terap.*, vol. 7, no. 1, pp. 119–126, 2023, doi: 10.33379/gtech.v7i1.1938.

[8] H. A. Ulvi and M. Ikhsan, "Comparison of K-Means and K-Medoids Clustering Algorithms for Export and Import Grouping of Goods in Indonesia," *J. dan Penelit. Tek. Inform.*, vol. 8, no. 3, pp. 1641–1655, 2024, [Online]. Available: https://doi.org/10.33395/sinkron.v8i3.13815

[9] S. Syahidatul Helma *et al.*, "Clustering pada Data Fasilitas Pelayanan Kesehatan Kota Pekanbaru Menggunakan Algoritma K-Means," *Puzzle Res. Data Technol. Fak. Sains dan Teknol.*, vol. 1, no. November, p. 4, 2019.

[10] Z. Huang, H. Zheng, C. Li, and C. Che, "Application of Machine Learning-Based K-means Clustering for Financial Fraud Detection," *Acad. J. Sci. Technol.*, vol. 10, no. 1, pp. 33–39, 2024, doi: 10.54097/74414c90.

[11] I. S. Tinendung and I. Zufria, "Pengelompokan Status Stunting Pada Anak Menggunakan Algoritma K-Means Clustering," vol. 7, pp. 2014–2023, 2023, doi: 10.30865/mib.v7i4.6908.

[12] A. Yudhistira and R. Andika, "Pengelompokan Data Nilai Siswa Menggunakan Algoritma K-Means Clustering," *J. Artif. Intell. Technol. Inf.*, vol. 1, no. 1, pp. 20–28, 2023, doi: 10.58602/jaiti.v1i1.22.

[13] L. W. Pa, "Penerapan Data Mining Pengelompokan Hasil Diagnosa Pasien BPJS Berdasarkan Usia Menggunakan Metode Clustering (Studi Kasus: RSU Bidadari Binjai)," *J. Inf. Technol.*, vol. 2, no. 1, pp. 8–14, 2022, doi: 10.32938/jitu.v2i1.1036.

[14] I. Amelia and F. M. Sarimole, "Analisis Sentimen Tanggapan Pengguna Media Sosial X Terhadap Program Beasiswa KIP-Kuliah dengan Menggunakan Algoritma Support Vector Machine (SVM)," vol. 5, no. 3, pp. 2994–3003, 2024.

[15] N. M. A. Mahar, Vihi Atina, and Nugroho Arif Sudibyo, "Pemodelan Prediksi Kelulusan Mahasiswa Dengan Metode Naïve Bayes Di Uniba," *J. Manaj. Inform. dan Sist. Inf.*, vol. 6, no. 2, pp. 148–158, 2023, doi: 10.36595/misi.v6i2.875.

[16] N. W. Utami and A. A. I. I. Paramitha, "Penerapan Data Mining Untuk Mengetahui Pola Pemilihan Program Studi Di Stmik Primakara Menggunakan Algoritma K-Means Clustering," *J. Teknol. Inf. dan Komput.*, vol. 7, no. 4, pp. 456–463, 2021, doi: 10.36002/jutik.v7i4.1540.

[17] S. R. Pratama and A. H. Mirza, "Penerapan Data Mining Untuk Memprediksi Tingkat Inflasi

Menggunakan Metode Regresi Linier Berganda Pada BPS," *Bina Darma Conf. Comput. Sci.*, pp. 245–255, 2021.

[18] R. Nuraini, "Implementasi Euclidean Distance dan Segmentasi K-Means Clustering Pada Identifikasi Citra Jenis Ikan Nila," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 3, no. 1, pp. 1–8, 2022.

[19] L. Pamungkas, N. A. Dewi, and N. A. Putri, "Classification of Student Grade Data Using the K-Means Clustering Method," *J. Sisfokom (Sistem Inf. dan Komputer)*, vol. 13, no. 1, pp. 86–91, 2024, doi: 10.32736/sisfokom.v13i1.1983.

[20] B. M. Liu *et al.*, "Association of the Drug Burden Index (DBI) exposure with outcomes: A systematic review," *J. Am. Geriatr. Soc.*, vol. 72, no. 2, pp. 589–603, 2024, doi: 10.1111/jgs.18691.