Jurnal Ilmu Komputer dan Sistem Informasi (JIRSI) **e-ISSN 2830-3954**
Volume: 5, Nomor: 3, September 2026: 805-820 **p-ISSN 2830-6031**
https://jurnal.unity-academy.sch.id/index.php/jirsi/index

Penerapan Algoritma K-Means Untuk Klasterisasi Pasien Berdasarkan Data Penyakit

Application of the K-Means Algorithm for Patient Clustering Based on Disease Data

Ghufron Makmun Lbs*[1], Abdul Halim Hasugian[2]
[1,2]Universitas Islam Negeri Sumatera Utara
E-mail: [1]ghufronlubis1@gmail.com, [2]abdulhalimhasugian@uinsu.ac.id

Abstrak

Puskesmas Sibuhuan memiliki data kunjungan pasien yang terus bertambah setiap bulan, namun pemanfaatan data tersebut masih terbatas pada penyusunan laporan sehingga informasi mengenai karakteristik kelompok pasien belum dimanfaatkan secara optimal. Penelitian ini bertujuan untuk menerapkan algoritma K-Means dalam proses klasterisasi pasien serta mengidentifikasi pola karakteristik pasien berdasarkan variabel usia, jenis kelamin, dan jenis penyakit. Penelitian menggunakan metode kuantitatif dengan pendekatan data mining. Data yang digunakan merupakan data sekunder berupa 1000 data kunjungan pasien Puskesmas Sibuhuan periode Januari 2026. Setelah dilakukan pembersihan data, sebanyak 104 data dengan missing value dihapus sehingga diperoleh 896 data yang digunakan untuk proses klasterisasi. Tahapan penelitian meliputi preprocessing data, normalisasi menggunakan metode Min-Max, penentuan jumlah cluster menggunakan metode elbow, proses klasterisasi menggunakan algoritma K-Means, serta evaluasi hasil klasterisasi menggunakan Silhouette Coefficient. Hasil penelitian menunjukkan bahwa jumlah cluster optimal adalah tiga cluster, yaitu Cluster 0 sebanyak 423 pasien dengan penyakit terbanyak sistem pernapasan, pencernaan, dan peredaran darah, Cluster 1 sebanyak 320 pasien dengan penyakit terbanyak gejala, tanda, dan temuan klinis abnormal, dan kehamilan, persalinan, dan masa nifas, dan Cluster 2 sebanyak 153 pasien dengan penyakit infeksi dan parasit, endokrin, nutrisi dan metabolik serta gangguan mental dan perilaku. Evaluasi menghasilkan nilai Silhouette Coefficient sebesar 0,6384, yang menunjukkan bahwa hasil klasterisasi memiliki kualitas yang baik. Hasil penelitian ini diharapkan dapat menjadi informasi pendukung bagi Puskesmas Sibuhuan dalam mengidentifikasi karakteristik kelompok pasien sehingga dapat mendukung perencanaan pelayanan kesehatan yang lebih tepat sasaran.

***Kata kunci**: Klasterisasi, K-Means, Pasien, Puskesmas Sibuhuan*

Abstract

Sibuhuan Community Health Center (Puskesmas Sibuhuan) experiences a continuous increase in patient visit data every month. However, the utilization of these data is still limited to administrative reporting, resulting in the underutilization of information regarding patient group characteristics. This study aims to implement the K-Means algorithm for patient clustering and to identify patient characteristic patterns based on age, gender, and disease variables. This research employed a quantitative method with a data mining approach. The data used were secondary data consisting of 1,000 patient visit records from Puskesmas Sibuhuan during the January 2026 period. Following data cleaning, 104 records containing missing values were removed, resulting in 896 records used for the clustering process. The research stages included data preprocessing, Min-Max normalization, determining the optimal number of clusters using the Elbow Method, clustering using the K-Means algorithm, and evaluating the clustering results using the Silhouette Coefficient. The results showed that the optimal number of clusters was three, Cluster 0 consisted of 423 patients, predominantly

diagnosed with diseases of the respiratory, digestive, and circulatory systems. Cluster 1 comprised 320 patients, with the most prevalent diagnoses involving symptoms, signs, and abnormal clinical findings, as well as pregnancy, childbirth, and the puerperium. Cluster 2 consisted of 153 patients, predominantly diagnosed with infectious and parasitic diseases, endocrine, nutritional and metabolic diseases, and mental and behavioral disorders. The evaluation produced a Silhouette Coefficient value of 0.6384, indicating that the clustering results were of good quality. The findings of this study are expected to provide supporting information for Puskesmas Sibuhuan in identifying patient group characteristics and to assist in planning more targeted healthcare services.

***Keywords**: Clustering, K-Means, Patients, Puskesmas Sibuhuan*

1. PENDAHULUAN

Perkembangan teknologi informasi telah membawa pengaruh yang besar terhadap bidang kesehatan. Transformasi digital mendorong institusi pelayanan kesehatan untuk beradaptasi guna meningkatkan efektivitas, efisiensi, dan kualitas pelayanan. Salah satu implementasinya adalah teknologi informasi untuk mendukung pengelolaan pelayanan kesehatan, seperti administrasi puskesmas, penggunaan rekam medis elektronik, *telemedicine*, penggunaan sistem informasi manajemen rumah sakit (SIMRS), serta pengambilan keputusan yang didukung oleh data[1]. Seiring dengan penerapan rekam medis digital dan sistem informasi kesehatan, data pasien yang mencakup identitas, riwayat kunjungan, hingga diagnosis penyakit terus bertambah. Namun, besarnya volume data tersebut belum dimanfaatkan secara optimal dan masih sering hanya berfungsi sebagai arsip, sehingga belum mampu mendukung pengambilan keputusan secara efektif. Kondisi ini menunjukkan bahwa ketersediaan data yang melimpah belum diimbangi dengan pemanfaatannya sebagai aset strategis dalam pelayanan kesehatan [2].

Permasalahan tersebut juga terjadi di Puskesmas Sibuhuan, Kecamatan Barumun, Kabupaten Padang Lawas, Provinsi Sumatera Utara. Sebagai fasilitas pelayanan kesehatan tingkat pertama, Puskesmas Sibuhuan setiap bulannya mengelola data rekam medis pasien dalam jumlah yang besar. Namun, pemanfaatan data tersebut masih terbatas pada kegiatan pencatatan dan pelaporan rutin, akan tetapi belum digunakan untuk menggali pola penyakit maupun karakteristik pasien yang dapat mendukung perencanaan program kesehatan. Menurut World Health Organization [3], data yang dikumpulkan melalui sistem informasi kesehatan memiliki peran penting dalam mendukung analisis, perencanaan, manajemen, dan pengambilan keputusan. Oleh karena itu, pemanfaatan data kesehatan dapat memberikan informasi yang berguna untuk penyusunan program kesehatan, serta perencanaan kebutuhan obat dan alat kesehatan secara lebih efektif.

Untuk menyelesaikan permasalahan tersebut, dibutuhkan pendekatan teknologi yang bisa mengolah data kesehatan menjadi informasi yang bermanfaat. Salah satu pendekatan yang relevan adalah data *mining*, yaitu menggali informasi serta mengidentifikasi pola dari sekumpulan data berukuran besar. Dalam penelitian ini digunakan metode *clustering*, yaitu teknik mengelompokkan data berdasarkan tingkat kesamaan karakteristiknya. Teknik ini bersifat *unsupervised* sehingga sesuai untuk mengidentifikasi pola data kesehatan yang belum memiliki label [4]. K-Means adalah salah satu algoritma yang dapat digunakan untuk

clustering, dipilih karena mudah diimplementasikan dan memiliki efisiensi komputasi yang baik. K-Means mengelompokkan data berdasarkan jarak terdekat terhadap *centroid* yang diperbarui secara iteratif hingga mencapai kondisi konvergen[5].

Penerapan algoritma K-Means pada data kesehatan telah dilakukan dalam berbagai penelitian dengan objek dan karakteristik data yang beragam. Aldiyatna *et al.* [6], menerapkan algoritma K-Means pada dataset pasien penyakit diare di Puskesmas Beber dan memperoleh tiga *cluster* dengan kualitas *clustering* yang baik berdasarkan nilai *Davies-Bouldin Index* (DBI). Pada penelitian Juleha *et al.* [7], penyakit menular dan tidak menular dikelompokkan berdasarkan wilayah pasien sehingga menghasilkan dua kelompok wilayah dengan tingkat risiko penyakit yang memiliki perbedaan. Adapun, Adawiyah *et al.* [8] memanfaatkan algoritma K-Means dalam mengelompokkan rekomendasi metode kontrasepsi berdasarkan karakteristik pasien.

Penerapan algoritma K-Means juga digunakan pada penelitian Rosida dan Wijaya [9], untuk mengelompokkan kasus HIV/AIDS di Provinsi Jawa Barat dan memperoleh jumlah *cluster* optimal sebanyak dua berdasarkan nilai DBI. Selain itu, Wala *et al.* [10] menerapkan K-Means untuk pengelompokan pasien penyakit jantung berdasarkan berbagai karakteristik medis sehingga menghasilkan empat *cluster* yang mampu menggambarkan tingkat risiko pasien. Temuan pada penelitian terdahulu mengindikasikan bahwa algoritma K-Means efektif dalam mengungkap pola yang terbentuk pada kumpulan data, yang dapat menjadi dasar dalam menunjang proses pengambilan keputusan pada sektor pelayanan kesehatan.

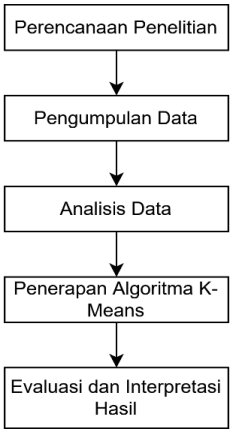
Meskipun algoritma K-Means telah banyak diterapkan pada data kesehatan, penelitian terdahulu menggunakan fokus dan kombinasi variabel yang berbeda, seperti penyakit tertentu, wilayah tempat tinggal, maupun karakteristik medis pasien. Penelitian yang mengombinasikan variabel usia, jenis kelamin, dan jenis penyakit berdasarkan klasifikasi ICD-10 pada data rekam medis di tingkat puskesmas masih terbatas. Kombinasi ketiga variabel tersebut memungkinkan terbentuknya profil klaster yang lebih komprehensif karena mencakup aspek demografis dan klinis secara bersamaan. Selain itu, penerapan algoritma K-Means pada data rekam medis Puskesmas Sibuhuan belum pernah dilakukan.

Berdasarkan pemaparan sebelumnya, penelitian ini bertujuan mengimplementasikan algoritma K-Means untuk mengelompokkan pasien berdasarkan data penyakit dan kemiripan karakteristiknya dan mengidentifikasi pola karakteristik pada setiap *cluster*. Hasil yang diperoleh dari penelitian ini diharapkan bisa digunakan sebagai referensi dalam penerapan data *mining*, terutama algoritma K-Means pada klasterisasi data kesehatan, serta memberikan informasi dan gambaran karakteristik kelompok pasien yang dapat dimanfaatkan oleh Puskesmas untuk mendukung penyusunan program kesehatan, pengelolaan sumber daya, dan peningkatan kualitas pelayanan kesehatan.

2. METODOLOGI PENELITIAN

Kerangka penlitian disusun untuk memberikan alur penelitian, mulai dari perencanaan, pengumpulan dan analisis data, klasterisasi dengan algoritma K-Means, hingga evaluasi dan interpretasi hasil. Pendekatan kuantitatif digunakan

dalam penelitian ini dengan menerapkan metode data *mining* yang diimplementasikan dengan bahasa pemrograman Python. Kerangka penelitian disajikan pada Gambar 1.



Gambar 1. Kerangka Penelitian

2.1. Perencanaan Penelitian

Tahapan awal mengidentifikasi permasalahan penelitian, penentuan metode yang digunakan, serta studi literatur yang relevan dengan topik penelitian. Referensi tersebut menjadi dasar dalam penyusunan metode penelitian, pemilihan algoritma, dan pelaksanaan penelitian secara sistematis.

2.2. Pengumpulan Data

Selanjutnya, dilakukan pengumpulan data sekunder berupa rekam medis atau laporan kunjungan pasien Puskesmas Sibuhuan, Kabupaten Padang Lawas, periode Januari 2026. Dataset yang diperoleh terdiri atas 1.000 baris data pasien.

2.3. Analisis Data

Selanjutnya, dataset dianalisis untuk memastikan kualitasnya sebelum digunakan dalam pemodelan algoritma. Analisis diawali dengan tahapan *preprocessing* sebagai berikut:

1. Seleksi Data, memilih dan menetapkan tiga variabel utama dari dataset (Usia, Jenis Kelamin, dan Jenis Penyakit)
2. Membersihkan Data, untuk menangani *missing value* agar kualitas data tetap terjaga. Dari 1.000 data awal, terdapat 104 data dengan *missing value* yang dihapus sehingga tersisa 896 data untuk dianalisis.
3. Transformasi Data, variabel jenis kelamin diubah menggunakan *Label Encoding* menjadi 0 dan 1, sedangkan jenis penyakit dikategorikan berdasarkan klasifikasi ICD-10 dari WHO, kemudian setiap kategori diberi kode numerik menggunakan *Label Encoding*.
4. Normalisasi Data, menggunakan metode *Min-Max Scaler* untuk menyamakan rentang nilai setiap variabel agar tidak mendominasi perhitungan jarak pada algoritma K-Means.

2.4. Penerapan Algoritma K-Means

Selanjutnya algoritma K-Means diterapkan dengan menentukan jumlah *cluster* menggunakan metode *Elbow*. *Centroid* awal ditentukan dari data hasil

preprocessing. Setiap data dikelompokkan berdasarkan jarak *Euclidean* terdekat terhadap *centroid*, kemudian *centroid* diperbarui berdasarkan rata-rata anggota masing-masing *cluster*. Proses tersebut dilakukan secara iteratif hingga konvergen, yaitu tidak terdapat perubahan keanggotaan *cluster*.

2.5. Evaluasi dan Interpretasi Hasil

Setelah hasil didapatkan, mengevaluasi kualitas *cluster* menggunakan *Silhouette Coefficient*. selanjutnya melakukan interpretasi hasil terhadap karakteristik masing-masing *cluster* sehingga diperoleh informasi mengenai karakteristik kelompok pasien yang terbentuk.

3. HASIL DAN PEMBAHASAN

Hasil penelitian disajikan mulai dari deskripsi dataset, *preprocessing*, penentuan jumlah *cluster* dengan metode Elbow, klasterisasi menggunakan algoritma K-Means, mengevaluasi kualitas *cluster* dengan *Silhouette Coefficient*, serta interpretasi karakteristik pada setiap *cluster*.

3.1. Deskripsi Dataset

Dataset yang digunakan berasal dari data sekunder, yaitu rekam medis atau laporan kunjungan pasien yang diperoleh dari Puskesmas Sibuhuan, Kabupaten Padang Lawas. Dataset penelitian terdiri atas 1.000 data pasien periode Januari 2026 dengan atribut nama pasien, usia, jenis kelamin, jenis penyakit, dan alamat. Untuk menjaga kerahasiaan identitas pasien, nama pasien direpresentasikan dalam bentuk kode.

Atribut yang digunakan pada proses klasterisasi meliputi usia, jenis kelamin, dan jenis penyakit. Sementara itu, variabel alamat tidak digunakan dalam proses klasterisasi, tetapi dimanfaatkan sebagai informasi pendukung pada tahap interpretasi hasil untuk memberikan gambaran karakteristik setiap *cluster*. Tabel 1 menyajikan sebagian data sebagai representasi dari keseluruhan dataset yang digunakan dalam penelitian.

Tabel 1. Deskripsi Dataset

No.	Kode Pasien	Usia	Jenis Kelamin	Jenis Penyakit	Alamat
1	P0001	10	Laki-Laki	Paranoid schizophrenia (F20.0)	Hasahatan Jae
2	P0002	15	Perempuan	Essential hypertension (I10)	Psr Sibuhuan
3	P0003	20	Perempuan	Chronic laryngitis (J37.0)	Sibuhuan Jae
4	P0004	13	Laki-Laki	Acute upper respiratory infectio (J06.9)	Psr Sibuhuan
5	P0005	3	Perempuan	Influenza due to identified avian infliuenza virus (J09)	Psr Sibuhuan
...	.....	....	.....	.....	.....
996	P0996	15	Perempuan	Cough (R05)	Sibuhuan Jae

997	P0997	35	Perempuan	Fever of other and unknown origin (R50)	Sipagabu
998	P0998	46	Perempuan	Gastritis (K29)	Tano Bato
999	P0999	4	Laki-Laki	Dyspepsia (K30)	Sitarolo Julu
1000	P1000	48	Laki-Laki	Headache (R51)	Bulu Sonik

3.2. Transformasi Data

Transformasi dilakukan pada variabel jenis kelamin diubah menjadi representasi numerik melalui teknik label *encoding*, sedangkan untuk variabel jenis penyakit terdapat sebanyak 165 variasi jenis penyakit dalam dataset yang dikelompokkan ke dalam kategori penyakit berdasarkan klasifikasi ICD-10 yang ditetapkan oleh *World Health Organization*. Pengelompokan ini dilakukan untuk menyederhanakan variasi jenis penyakit yang cukup banyak ke dalam kelompok penyakit yang lebih terstruktur berdasarkan klasifikasi yang digunakan. Selanjutnya, setiap kategori penyakit direpresentasikan dalam bentuk numerik menggunakan teknik label *encoding* agar dapat diolah dalam proses klasterisasi menggunakan algoritma K-Means.

Tabel 2. Transformasi Jenis Kelamin

Kode	Jenis Kelamin
0	Perempuan
1	Laki-laki

Tabel 3. Kategorisasi dan Transformasi Jenis Penyakit

Kode	Kode ICD-10	Kategori Penyakit ICD-10
0	A00–B99	Bab I - Penyakit infeksi dan parasit
1	E00–E90	Bab IV - Penyakit endokrin, nutrisi, dan metabolik
2	F00-F99	Bab V – Gangguan mental dan perilaku
3	I00-I99	Bab IX - Penyakit sistem peredaran darah
4	J00-J99	Bab X - Penyakit sistem pernapasan
5	K00-K93	Bab XI - Penyakit sistem pencernaan
6	L00-L99	Bab XII - Penyakit kulit dan jaringan subkutan
7	M00-M99	Bab XIII - Penyakit sistem muskuloskeletal dan jaringan ikat
8	N00-N99	Bab XIV - Penyakit sistem genitourinari
9	O00-O99	Bab XV - Kehamilan, persalinan, dan masa nifas
10	R00-R99	Bab XVIII - Gejala,tanda,dan temuan klinis/lab abnormal

3.3. Normalisasi Data

Proses normalisasi data dengan menerapkan metode *Min-Max Scaler*. Normalisasi bertujuan untuk menyetarakan skala nilai pada atribut numerik, sehingga skala antarvariabel dapat diseragamkan dan tidak terdapat variabel yang memiliki pengaruh lebih besar terhadap perhitungan jarak pada algoritma K-Means.

Pada penelitian ini, normalisasi diterapkan pada variabel usia karena memiliki rentang nilai yang cukup besar dibandingkan variabel lainnya. Melalui metode ini, nilai usia ditransformasikan ke dalam rentang 0 hingga 1 tanpa mengubah hubungan atau distribusi relatif antar data. Data normalisasi selanjutnya dijadikan sebagai masukan pada proses klasterisasi menggunakan algoritma K-Means. Sebagai contoh, penerapan metode *Min-Max Scaler* pada data pertama untuk variabel usia ditunjukkan sebagai berikut.

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

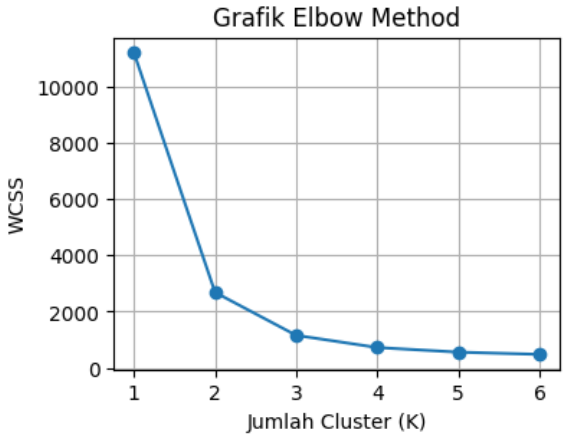
$$X' = \frac{34 - 1}{88 - 1} = \frac{33}{87} = 0{,}37931$$

Tabel 4. Normalisasi Data

No.	Kode pasien	Usia	Jenis Kelamin	Kategori Penyakit
1	P0001	0,37931	1	2
2	P0002	0,59770	0	3
3	P0003	0,29885	0	4
4	P0004	0,13793	1	4
5	P0005	0,02299	0	4
....	.....	.....	...	...
996	P0996	0,16092	0	10
997	P0997	0,39080	0	10
998	P0998	0,51724	0	5
999	P0999	0,03448	1	5
1000	P1000	0,54023	1	10

3.4. Penentuan Jumlah Cluster

Jumlah *cluster* yang optimal ditentukan melalui metode *elbow* dengan menghitung nilai *Within-Cluster Sum of Squares* (WCSS) pada beberapa jumlah *cluster* dan memvisualisasikannya dalam bentuk grafik. Nilai K yang dipilih sebagai jumlah *cluster* optimal berada pada titik ketika penurunan WCSS tidak lagi signifikan, yang ditandai dengan terbentuknya pola menyerupai siku (*elbow)*.



Gambar 2. Metode Elbow

Gambar 2. diatas, dapat dilihat nilai WCSS menurun tajam hingga K=3. Setelah nilai tersebut, kurva mulai membentuk pola yang lebih landai, Hal tersebut mengindikasikan bahwa penambahan jumlah *cluster* selanjutnya tidak menghasilkan penurunan nilai WCSS yang tajam. Dengan demikian, jumlah *cluster* yang digunakan ditetapkan sebanyak tiga.

3.5. Penerapan Algoritma K-Means

Tahapan berikutnya penentuan *centroid* awal yang merupakan titik pusat awal dalam algoritma K-Means. *Centroid* dipilih dari data hasil *preprocessing* yang memiliki karakteristik berbeda dan digunakan sebagai acuan untuk menghitung jarak antar data menggunakan *Euclidean Distance*. Untuk mempermudah penyajian proses perhitungan, contoh perhitungan manual dilakukan menggunakan 40 data pertama sebagai representasi dari dataset, sedangkan implementasi K-Means melalui Google Colab menggunakan bahasa pemrograman Python dilakukan terhadap seluruh dataset yang berjumlah 1.000 data pasien.

Tabel 5. Centroid Awal

Centroid Awal	Usia	Jenis Kelamin	Kategori Penyakit
Centroid 0	0,29885	0	4
Centroid 1	0,31034	0	9
Centroid 2	0,58621	0	1

Selanjutnya *Euclidean Distance* digunakan untuk mengukur tingkat kedekatan antar data terhadap *centroid* sehingga dapat dikelompokkan pada *cluster* berdasarkan jarak terdekat dengan rumus:

$$d = \sqrt{(\chi_{data} - \chi_{centroid})^2 + (y_{data} - y_{centroid})^2 + \cdots + (Z_{data} - Z_{centroid})^2} \quad (2)$$

Iterasi 1

K=3

C0 = (0,29885; 0; 4)

C1 = (0,31034; 0; 9)

C2 = (0,58621; 0; 1)

a. Menghitung jarak centroid terdekat pada data ke 3.

$$C0 = \sqrt{(0{,}29885 - 0{,}29885)^2 + (0-0)^2 + (4-4)^2} = 0$$

$$C1 = \sqrt{(0{,}29885 - 0{,}31034)^2 + (0-0)^2 + (4-9)^2} = 5{,}00001$$

$$C2 = \sqrt{(0{,}29885 - 0{,}58621)^2 + (0-0)^2 + (4-1)^2} = 3{,}01373$$

b. Menghirung jarak centroid terdekat pada data ke 11.

$$C0 = \sqrt{(0{,}31034 - 0{,}29885)^2 + (0-0)^2 + (4-4)^2} = 5{,}00001$$

$$C1 = \sqrt{(0{,}31034 - 0{,}31034)^2 + (0-0)^2 + (4-9)^2} = 0$$

$$C2 = \sqrt{(0{,}31034 - 0{,}58621)^2 + (0-0)^2 + (4-1)^2} = 8{,}00476$$

c. Menghitung jarak centroid terdekat pada data ke 35.

$$C0 = \sqrt{(0{,}58621 - 0{,}29885)^2 + (0-0)^2 + (4-4)^2} = 3{,}01373$$

$$C1 = \sqrt{(0{,}58621 - 0{,}31034)^2 + (0-0)^2 + (4-9)^2} = 8{,}00476$$

$$C2 = \sqrt{(0{,}58621 - 0{,}58621)^2 + (0-0)^2 + (4-1)^2} = 0$$

Berdasarkan hasil perhitungan jarak terhadap *centroid* terdekat, diperoleh hasil pengelompokan data pada iterasi 1 yang dapat dilihat pada Tabel 6.

Tabel 6. Hasil Iterasi 1

Data ke-	Jarak Centroid pada Iterasi 1			Cluster
	C0	C1	C2	
1	2.23752	7.07140	1.42927	2
2	1.04370	6.00688	2.00003	0
3	0	5.00001	3.01373	0
4	1.01286	5.10193	3.19389	0
5	0.27586	5.00825	3.05241	0
...	.....	.....	.....	...
35	3.01373	8.00475	0	2
36	7.07111	2.23610	10.05335	1
37	4.00041	1.00106	7.00377	1
38	1.04705	5.10777	3.16236	0
39	6.01198	1.06550	9.00047	1
40	0.09195	5.00065	3.00636	0

Kemudian dilanjutkan ke iterasi berikutnya. Sebelum melakukan iterasi ke-2, *centroid* diperbarui berdasarkan nilai rata-rata setiap atribut dari seluruh anggota pada setiap *cluster.*

Tabel 7. Centroid Baru

Centroid Awal	Usia	Jenis Kelamin	Kategori Penyakit
Centroid 0	0,33005	0,61905	4,66667
Centroid 1	0,36398	0,33333	9,91667
Centroid 2	0,27586	0,71429	1,28571

Selanjutnya dilakukan perhitungan pada iterasi ke-2 menggunakan *centroid* hasil pembaruan dari iterasi sebelumnya. Tahapan yang dilakukan sama seperti pada iterasi 1, yaitu mengukur jarak antara setiap data dan *centroid* menggunakan metode *Euclidean Distance*. dan mengelompokkannya pada *cluster* jarak terdekat. Hasil iterasi ke-2 tidak menunjukkan perubahan keanggotaan *cluster*. Kondisi tersebut menandakan bahwa algoritma K-Means telah konvergen, sehingga iterasi dihentikan dan hasil pengelompokan ditetapkan sebagai hasil akhir klasterisasi.

Tabel 8. Hasil Iterasi 2

Data ke-	Jarak Centroid pada Iterasi 2			Cluster
	C0	C1	C2	
1	2,69626	7,98694	0,72391	2
2	1,81215	6,98618	2,08668	0
3	0,91030	5,92641	2,80680	0
4	1,10033	6,01263	2,78294	0
5	1,12982	6,01889	2,87358	0
...	.....	.....	.....	...

35	3,72738	8,92567	0,82955	2
36	6,44889	1,03641	9,72845	1
37	3,41043	1,92863	6,73274	1
38	0,81037	5,98951	3,08646	0
39	5,39787	0,78649	8,80917	1
40	0,70984	5,97203	2,96590	0

3.6. Implementasi Menggunakan Google Colab

Implementasi K-Means pada penelitian ini dilakukan menggunakan bahasa pemrograman Python. Tahapan ini meliputi proses *import* dataset, *preprocessing*, penentuan jumlah *cluster*, pembentukan *cluster* menggunakan K-Means, visualisasi hasil *cluster*, serta evaluasi kualitas *cluster*.

1. Import Dataset

Tahap pertama pada implementasi algoritma di Google Colab adalah mengimpor dataset penelitian ke dalam pemrograman Python.

```
dataset = pd.read_excel('/content/drive/MyDrive/Tugas Akhir/dataset.xlsx')
print(dataset[['kode pasien', 'usia', 'jenis kelamin', 'kategori penyakit', 'Alamat']])

Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.moun
    kode pasien  usia  jenis kelamin  kategori penyakit         Alamat
0         P0001    34              1                2.0  Hasahatan Jae
1         P0002    53              0                3.0   Psr Sibuhuan
2         P0003    27              0                4.0   Sibuhuan Jae
3         P0004    13              1                4.0   Psr Sibuhuan
4         P0005     3              0                4.0   Psr Sibuhuan
..          ...   ...            ...                ...            ...
995       P0996    15              0               10.0   Sibuhuan Jae
996       P0997    35              0               10.0       Sipagabu
997       P0998    46              0                5.0      Tano Bato
998       P0999     4              1                5.0  Sitarolo Julu
999       P1000    48              1               10.0     Bulu Sonik

[1000 rows x 5 columns]
```

Gambar 3. Import Dataset

Dataset yang digunakan berupa *file* Microsoft Excel (.xlsx) yang berisi 1.000 data kunjungan pasien Puskesmas Sibuhuan.

2. *Preprocessing* Data

Selanjutnya data diperiksa guna memastikan kelengkapan data dan kesesuaian data yang dapat memengaruhi hasil pengelompokan.

```
kode pasien            0
usia                   0
jenis kelamin          0
kategori penyakit    104
dtype: int64

Menghapus Missing Value:
kode pasien          0
usia                 0
jenis kelamin        0
kategori penyakit    0
dtype: int64
Jumlah data sekarang: 896
```

Gambar 4. Penanganan *Missing Value*

Dari 1.000 data kunjungan pasien yang diperoleh, terdapat 104 data yang memiliki *missing value*. Data tersebut dihapus sehingga diperoleh 896 data yang digunakan dalam proses klasterisasi.

Selanjutnya melakukan tahapan normalisasi pada atribut numerik supaya setiap variabel memiliki nilai dengan skla yang seragam, sehingga perhitungan jarak antar data dapat dilakukan secara lebih optimal.

```
dataset['usia_normalisasi'] = scaler.fit_transform(dataset[['usia']])
print(dataset[['kode pasien', 'usia_normalisasi', 'jenis kelamin', 'kategori penyakit']])
```

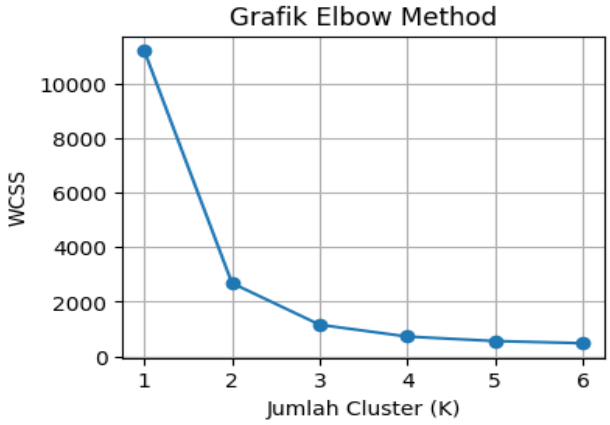
	kode pasien	usia_normalisasi	jenis kelamin	kategori penyakit
0	P0001	0.379310	1	2.0
1	P0002	0.597701	0	3.0
2	P0003	0.298851	0	4.0
3	P0004	0.137931	1	4.0
4	P0005	0.022989	0	4.0
..	...	...	...	...
995	P0996	0.160920	0	10.0
996	P0997	0.390805	0	10.0
997	P0998	0.517241	0	5.0
998	P0999	0.034483	1	5.0
999	P1000	0.540230	1	10.0

[896 rows x 4 columns]

Gambar 5. Normalisasi Data

3. Metode *Elbow*

Menentukan jumlah *cluster* optimal dengan metode *Elbow*, lalu nilai *Within Cluster Sum of Squares* (WCSS) dianalisis pada berbagai nilai K. Jumlah *cluster* optimal ditetapkan pada titik siku (*elbow point*) yang mulai melandai.



Gambar 6. Grafik Metode Elbow

Berdasarkan nilai WCSS yang diperoleh, dapat dilihat mengalami penurunan seiring dengan bertambahnya jumlah *cluster* (K). Penurunan yang mulai mengecil setelah K=3 menunjukkan bahwa penambahan jumlah K berikutnya tidak menghasilkan peningkatan yang tajam, sehingga dipilih tiga *cluster* yang merupakan jumlah *cluster* optimal.

4. Proses K-Means *Clustering*

Setelah jumlah *cluster* ditentukan, proses selanjutnya dilanjutkan dengan penerapan algoritma K-Means untuk melakukan pengelompokan data. Berikut ini akan ditampilkan hasil *cluster* yang diproses oleh algoritma K-Means.

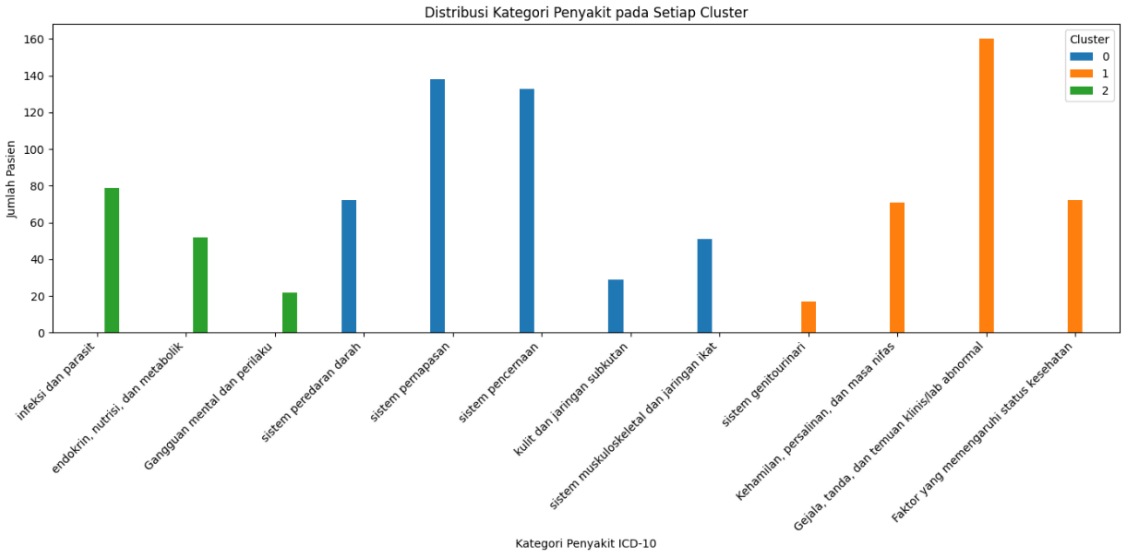
jumlah_cluster

	Cluster	Jumlah Data
0	0	423
1	1	320
2	2	153

Gambar 7. Jumlah Data Setiap *Cluster*

5. Visualisasi Hasil *Cluster*

Tahap selanjutnya adalah memvisualisasikan hasil pengelompokan data menggunakan *Bar Chart*. Visualisasi ini bertujuan untuk memberikan gambaran mengenai distribusi data pada setiap *cluster*, hingga hasil klasterisasi dapat dianalisis secara lebih jelas.



Gambar 8. Visualisasi *Bar Chart*

3.7. Evaluasi Hasil Cluster

Evaluasi *cluster* menggunakan metode *Silhouette Coefficient* untuk menghitung tingkat kemiripan data pada *cluster* dan perbedaan antar *cluster*. Nilai *Silhouette Coefficient* pada antara -1 hingga 1, dengan nilai yang semakin mendekati 1 mengindikasikan kualitas pengelompokan yang semakin baik.

```
    'Metode Evaluasi': ['Silhouette Score'],
    'Nilai': round(silhouette, 4)
})
evaluasi
```

	Metode Evaluasi	Nilai
0	Silhouette Score	0.6384

Gambar 9. Evaluasi Menggunakan *Sillhouette Coefficient*

3.8. Interpretasi Hasil

Selanjutnya dilakukan analisis terhadap masing-masing *cluster* untuk mengidentifikasi karakteristik kelompok pasien yang terbentuk. Hasil analisis ini diharapkan dapat memberikan informasi yang bermanfaat dalam mendukung perencanaan pelayanan kesehatan di Puskesmas Sibuhuan.

CLUSTER 0

	kode pasien	usia_normalisasi	jenis kelamin	kategori penyakit	Alamat
1	P0002	0.597701	0	3.0	Psr Sibuhuan
2	P0003	0.298851	0	4.0	Sibuhuan Jae
3	P0004	0.137931	1	4.0	Psr Sibuhuan
4	P0005	0.022989	0	4.0	Psr Sibuhuan
11	P0012	0.379310	1	4.0	Psr Sibuhuan
...	...	...	...	...	...
986	P0987	0.310345	0	7.0	Pasir Julu
989	P0990	0.298851	0	5.0	Sipagabu
990	P0991	0.402299	0	5.0	Psr Sibuhuan
997	P0998	0.517241	0	5.0	Tano Bato
998	P0999	0.034483	1	5.0	Sitarolo Julu

423 rows × 5 columns

Gambar 10. Hasil Klasterisasi Anggota Cluster 0

Cluster 0 terdiri atas 423 pasien dan merupakan cluster dengan jumlah anggota terbanyak dibandingkan cluster lainnya. Cluster ini didominasi oleh pasien perempuan sebanyak 257 pasien, sedangkan pasien laki-laki berjumlah 166 pasien. Berdasarkan penyakit, sebagian besar pasien berada pada penyakit sistem pernapasan dan penyakit sistem pencernaan, masing-masing sebanyak 138 dan 133 pasien. Selain itu, Cluster ini didominasi pasien dari Pasar Sibuhuan sebanyak 204 pasien, dengan penyakit sistem pencernaan sebagai yang terbanyak, yaitu 66 pasien. Selanjutnya, penyakit sistem pernapasan dominan di Janji Lobi (10 dari 24 pasien), Mompang (7 dari 15 pasien), dan Sibuhuan Jae (8 dari 14 pasien), sedangkan Sibuhuan Julu didominasi penyakit sistem pencernaan (7 dari 18 pasien).

CLUSTER 1

	kode pasien	usia_normalisasi	jenis kelamin	kategori penyakit	Alamat
6	P0007	0.666667	1	8.0	Sibuhuan Julu
10	P0011	0.310345	0	9.0	Psr Sibuhuan
12	P0013	0.379310	0	11.0	Janj Lobi
14	P0015	0.264368	1	11.0	Sibuhuan Jae
20	P0021	0.252874	0	11.0	Psr Sibuhuan
...	...	...	...	...	...
993	P0994	0.367816	0	11.0	Psr Sibuhuan
994	P0995	0.482759	0	10.0	Sibuhuan Julu
995	P0996	0.160920	0	10.0	Sibuhuan Jae
996	P0997	0.390805	0	10.0	Sipagabu
999	P1000	0.540230	1	10.0	Bulu Sonik

320 rows × 5 columns

Gambar 11. Hasil Klasterisasi Anggota Cluster 1

Cluster 1 terdiri atas 320 pasien dan merupakan cluster dengan jumlah anggota terbanyak kedua setelah Cluster 0. Cluster ini didominasi oleh pasien perempuan sebanyak 214 pasien, sedangkan pasien laki-laki 106 pasien. Berdasarkan penyakit, sebagian besar pasien berada pada kategori gejala, tanda, dan temuan klinis/lab abnormal sebanyak 160 pasien, diikuti kategori faktor yang memengaruhi status kesehatan dan kontak dengan pelayanan kesehatan sebanyak 72 pasien, sedangkan kasus kehamilan, persalinan, dan masa nifas sebanyak 71 pasien. Selain itu, Cluster

ini didominasi pasien dari Pasar Sibuhuan sebanyak 118 pasien, dengan penyakit gejala, tanda, dan temuan klinis/lab abnormal sebagai yang terbanyak, yaitu 51 pasien. Pasar Sibuhuan juga menjadi wilayah dengan kasus kehamilan, persalinan, dan masa nifas terbanyak, yaitu 29 pasien.

CLUSTER 2

	kode pasien	usia_normalisasi	jenis kelamin	kategori penyakit	Alamat
0	P0001	0.379310	1	2.0	Hasahatan Jae
5	P0006	0.137931	1	0.0	Psr Sibuhuan
7	P0008	0.252874	1	2.0	Tano Bato
8	P0009	0.252874	0	2.0	Psr Sibuhuan
9	P0010	0.287356	1	2.0	Binabo Jae
...	...	...	...	...	...
948	P0949	0.390805	0	1.0	Purbatua
952	P0953	0.011494	0	0.0	Janji Lobi
954	P0955	0.586207	0	0.0	Sitarolo Julu
982	P0983	0.126437	0	0.0	Sibuhuan Jae
992	P0993	0.298851	0	0.0	Psr Sibuhuan

153 rows × 5 columns

Gambar 12. Hasil Klasterisasi Anggota Cluster 2

Cluster 2 terdiri atas 153 pasien dan merupakan *cluster* dengan jumlah anggota paling sedikit dibandingkan *Cluster* 0 dan *Cluster* 1. *Cluster* ini didominasi oleh pasien perempuan sebanyak 90 pasien, sedangkan pasien laki-laki 63 pasien. Berdasarkan penyakit, sebagian besar pasien berada pada penyakit Infeksi dan Parasit sebanyak 79 pasien, diikuti penyakit endokrin, nutrisi, dan metabolik sebanyak 52 pasien, dan gangguan mental dan perilaku 22 pasien. Selain itu, *Cluster* ini didominasi pasien dari Pasar Sibuhuan sebanyak 83 pasien, dengan penyakit infeksi dan parasit yaitu 39 pasien. Selanjutnya, penyakit yang sama di Janji Lobi (6 dari 11 pasien) dan Bulu Sonik (3 dari 4 pasien), sedangkan Arse Simatorkis didominasi penyakit endokrin, nutrisi, dan metabolik (2 dari 4 pasien). Sementara itu, Saba Riba memiliki jumlah yang sama antara penyakit infeksi dan parasit serta gangguan mental dan perilaku, masing-masing sebanyak 2 dari 4 pasien.

KESIMPULAN

Berdasarkan penerapan algoritma K-Means yang dilakukan pada dataset pasien Puskesmas Sibuhuan, dapat disimpulkan bahwa algoritma K-Means berhasil diterapkan untuk melakukan proses klasterisasi data pasien berdasarkan variabel usia, jenis kelamin, dan jenis penyakit. Proses klasterisasi diawali dengan tahap *preprocessing* data yang meliputi transformasi variabel, pemberian label pada data kategorikal, serta normalisasi usia dengan metode *Min-Max* untuk menyetarakan skala nilai setiap atribut. Selanjutnya menentukan jumlah *cluster* yang optimal dengan menerapkan *Elbow Method* sebelum proses pembentukan *cluster* dengan menerapkan algoritma K-Means. Hasil klasterisasi menghasilkan tiga *cluster*, yaitu Cluster 0 sebanyak 423 pasien, Cluster 1 sebanyak 320 pasien, dan Cluster 2 sebanyak 153 pasien. Evaluasi kualitas hasil pengelompokan melalui metode *Silhouette Coefficient* menghasilkan nilai sebesar 0,6384, yang memperlihatkan bahwa hasil *cluster* yang terbentuk tergolong baik. Setiap *cluster* memiliki karakteristik pasien yang berlainan berdasarkan kombinasi variabel yang

digunakan, sehingga dapat memberikan gambaran mengenai pola kelompok pasien di Puskesmas Sibuhuan. Informasi tersebut diharapkan mampu mendukung untuk penyusunan program kesehatan serta pengambilan keputusan untuk meningkatkan kualitas pelayanan kesehatan di Puskesmas Sibuhuan. Untuk penelitian selanjutnya dapat mengembangkan hasil penelitian ini dengan menambahkan variabel yang relevan untuk memperoleh karakteristik *cluster* yang lebih komprehensif, menggunakan dataset dengan jumlah yang lebih besar untuk menghasilkan gambaran klasterisasi yang lebih menyeluruh, serta membandingkan algoritma K-Means dengan metode *clustering* lainnya, seperti *Hierarchical Clustering*, DBSCAN, dan K-Medoids, guna menentukan metode yang memberikan hasil pengelompokan terbaik.

DAFTAR PUSTAKA

[1] P. Maulaningrum, Siti Mujanah, and Achmad Yanu Alif Fianto, "Transformasi Digital di Sektor Kesehatan Tinjauan Literatur tentang Penerapan Teknologi Informasi dalam Manajemen Pelayanan," *Jurnal Ilmu Manajemen, Ekonomi dan Kewirausahaan*, vol. 5, no. 1, pp. 494–503, Jun. 2025, doi: 10.55606/jimek.v5i1.6399.

[2] J. A. Noyari, A. Aprillia, R. Ginting Munthe, A. Sutarman, and E. Kallas, "Optimasi Kinerja Sistem Informasi Manajemen Kampus Menggunakan Teknik Data Mining," *Jurnal MENTARI: Manajemen, Pendidikan dan Teknologi Informasi*, vol. 3, no. 1, pp. 52–63, 2024, doi: 10.33050/mentari.v3i1.

[3] World Health Organization, "Toolkit for Analysis and Use of Routine Health Facility Data: Integrated Health Services Analysis: District and Facility Level," Geneva, 2023.

[4] W. Aulia, A. Putera Utama Siahaan, L. Marlina, and M. Iqbal, "K-Means Clustering Algorithm Analysis For Grouping Patient Medical Record Data Based On Disease Type," *Informatika dan Sains*, vol. 14, no. 04, pp. 832–843, 2024, doi: 10.54209/infosains.v14i04.

[5] G. M. M. Sujak, H. N. Rofiq, and F. I. Tawakal, "Implementasi K-Means Clustering untuk Optimalisasi Anggaran Penyakit Tidak Menular," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 1, pp. 67–74, Nov. 2024, doi: 10.57152/malcom.v5i1.1597.

[6] K. Aldiyatna, N. Rahaningsih, and R. D. Dana, "Penerapan Data Mining untuk Clustering Penyakit Diare Menggunakan Algoritma K-Means (Studi Kasus: Puskesmas Beber)," *Jurnal Mahasiswa Teknik Informatika*, vol. 8, no. 3, pp. 3124–3131, 2024.

[7] M. Juleha, R. Limia Budiarti, and N. Kahar, "Analisis Penyakit Menular dan Tidak Menular Berdasarkan Wilayah Tempat Tinggal Menggunakan Metode K-Means dan Fuzzy C-Means (Studi Kasus: Puskesmas Bayung Lencir)," in *Seminar Nasional Informatika (SENATIKA)*, Jambi, 2025, pp. 262–268.

[8] Q. Adawiyah, S. Defit, and Sumijan, "Penerapan Algoritma K-Means Clustering untuk Mengelompokkan Rekomendasi Metode Kontrasepsi Berbasis Machine Learning di Puskesmas," *Jurnal KomtekInfo*, vol. 11, no. 4, pp. 300–305, Sep. 2024, doi: 10.35134/komtekinfo.v11i4.563.

[9] W. Rosida and Y. A. Wijaya, "Klasterisasi Penyakit HIV/AIDS di Jawa Barat Menggunakan Algoritma K-Means Clustering," *Blend Sains Jurnal Teknik*, vol. 1, no. 4, pp. 306–315, Mar. 2023, doi: 10.56211/blendsains.v1i4.235.

[10] J. Wala and R. Umar, "Implementasi K-Means Clustering pada Pengelompokan Pasien Penyakit Jantung," *Jurnal Informatika Sunan Kalijaga)*, vol. 9, no. 3, pp. 205–216, 2024.