Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 639−647
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

Clustering Data Pasien Berdasarkan Usia di Puskesmas Menerapkan Metode K-Means

Alifia Herlin Lutfiannisa, Maimunah*, Pristi Sukmasetya

Fakultas Teknik, Teknik Informatika, Universitas Muhammadiyah Magelang, Magelang
Jl. Tidar No.21, Magersari, Kec. Magelang Sel., Kota Magelang, Jawa Tengah, Indonesia
Email: [1]herlynellin@gmail.com, [2,*]maimunah@unimma.ac.id, [3]pristi.sukmasetya@ummgl.ac.id
Email Penulis Korespondensi: maimunah@unimma.ac.id
Submitted: **04/01/2024;** Accepted: **25/01/2024;** Published: **27/01/2024**

Abstrak−Penelitian ini bertujuan untuk melakukan clustering data penyakit berdasarkan usia pasien menggunakan metode K-Means di Puskesmas Tlogomulyo Kabupaten Temanggung yang mengalami kesulitan dalam mengelola data penyakit yang tidak teratur. Puskesmas sebagai pusat pelayanan kesehatan primer membutuhkan analisis data untuk mengidentifikasi pola atau kelompok penyakit yang sejenis berdasarkan usia. Metode K-Means Clustering digunakan dalam pengolahan data pasien untuk memahami distribusi penyakit dan membantu pengambilan keputusan terkait pencegahan, penanganan, serta perencanaan perawatan kesehatan yang tepat. Hasil clustering menunjukkan adanya 2 cluster, di mana cluster pertama didominasi oleh kasus skabies pada usia 10-20 tahun, sementara cluster kedua memiliki prevalensi tinggi ISPA pada pasien berusia sekitar 50-55 tahun. Evaluasi menggunakan Silhouette Coefficient menunjukkan pembentukan 2 cluster sebagai yang paling optimal dengan nilai 0.4374. Temuan ini memberikan wawasan penting untuk pengembangan strategi penanggulangan penyakit yang lebih efektif berdasarkan karakteristik dan profil kesehatan masing-masing cluster.

Kata Kunci: Data Mining; Clustering Pasien; K-Means

Abstract−This research aims to perform clustering of disease data based on patient age using the K-Means method at the Tlogomulyo Community Health Center in Temanggung Regency, which faces challenges in managing irregular health data. As a primary healthcare centre, the health facility requires data analysis to identify patterns or groups of diseases based on age. The K-Means clustering method is employed in processing patient data to understand the distribution of diseases and aid decision-making regarding prevention, treatment, and healthcare planning. The clustering results reveal two clusters where the first cluster is dominated by scabies cases in the 10-20 age group, while the second cluster exhibits a high prevalence of Acute Respiratory Infection (ARI) in patients around the age of 51. Evaluation using the Silhouette Coefficient indicates that forming 2 clusters is the most optimal, with a value of 0.44. These findings provide crucial insights for the development of more effective disease management strategies based on the characteristics and health profiles of each cluster at the Tlogomulyo Community Health Center.

Keywords: Data Mining; Patient Clustering; K-Means

1. PENDAHULUAN

Puskesmas adalah lembaga pelayanan kesehatan primer yang memiliki peran sebagai inti pembangunan kesehatan, sebagai pusat peran serta masyarakat dalam bidang kesehatan, dan menyelenggarakan pelayanan kesehatan awal yang menyeluruh, terpadu, dan berkelanjutan pada suatu wilayah tertentu [1][2]. Puskesmas berperanan penting dalam menyediakan pelayanan kesehatan kepada masyarakat. Organisasi yang berinteraksi langsung dengan masyarakat mempunyai tanggung jawab untuk mengumpulkan dan menyebarkan informasi tentang penyakit. Informasi ini sangat penting untuk membuat keputusan mengenai pilihan pengobatan dan perencanaan layanan kesehatan. Dalam kasus flu, rumah sakit harus memberikan informasi mengenai gejala dan tindakan pencegahan [3]. Penting untuk menganalisis data penyakit berdasarkan usia untuk memahami distribusi penyakit di fasilitas kesehatan, karena individu dengan usia berbeda dapat menunjukkan penyakit yang berbeda-beda karena perbedaan fisik, mental, dan sistem kekebalan [4].

Puskesmas Tlogomulyo yang terletak di Kecamatan Tlogomulyo Kabupaten Temanggung adalah salah satu instansi pemerintah di bidang kesehatan. Dalam penanganan data penyakit yang dikumpulkan oleh puskesmas, sering kali terdapat kesulitan dalam mengidentifikasi pola atau keterkaitan antar penyakit yang ada karena data yang banyak dan tidak beraturan. Oleh karena itu, diperlukan pengolahan data penyakit untuk membantu menunjukan informasi tentang pola dan hubungan antara usia pasien dan jenis penyakit yang paling sering terjadi. Berdasarkan informasi tersebut, pihak puskesmas dapat mengambil langkah-langkah atau kebijakan seperti diadakannya penyuluhan dalam mengantisipasi pencegahan penyakit dan menjaga ketersediaan stok obat-obatan[5][6].

Untuk mendapatkan informasi yang akurat, maka data pasien perlu dianalisis dengan menggunakan metode data mining. Menurut [7], Data mining mengacu pada proses mengekstraksi pengetahuan dari sekumpulan besar data untuk mengungkap pola yang dapat dikembangkan lebih lanjut. Ini melibatkan penggunaan teknik komputasi dan algoritma seperti statistik, pembelajaran mesin, pengenalan pola, dan pemrosesan basis data. Dalam proses data mining, data yang tidak terstruktur atau tidak terorganisir diubah menjadi informasi yang bermakna dan berguna.

Metode yang akan diimplementasikan adalah metode K-Means clustering. Clustering yaitu proses menggabungkan kumpulan data yang berjumlah besar ke dalam beberapa kelas atau cluster berdasarkan

Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 639−647
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

karakteristik masing-masing [8][9]. K-Means adalah teknik clustering yang digunakan untuk menggabungkan sekumpulan objek ke dalam cluster berdasarkan kesamaan fiturnya. Algoritma ini mempertimbangkan sejumlah centroid yang telah ditentukan dan menugaskan setiap objek ke cluster yang centroidnya paling dekat dengannya. Proses ini berlanjut secara iteratif hingga centroid tidak lagi berubah secara signifikan. Setelah algoritma konvergen, setiap objek diklasifikasikan sebagai anggota cluster yang pusat massanya paling dekat dengannya [10][11]. K-Means adalah salah satu metode clustering yang efektif dan populer dalam analisis data. Dalam konteks pusat kesehatan masyarakat, pengelompokan data membantu mengidentifikasi pola atau kelompok penyakit yang serupa berdasarkan usia pasien.

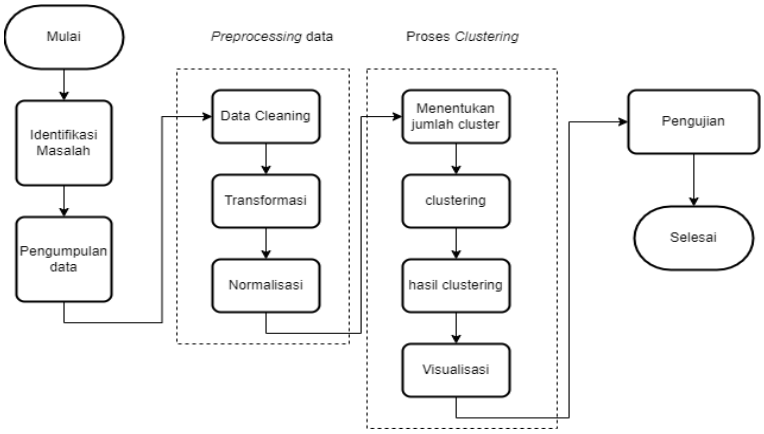
Beberapa penelitian tentang pengolahan data pasien puskesmas telah dilakukan, diantaranya tentang clustering data puskesmas menggunakan Metode K-Means Clustering [12][13], Algoritma K-Means Clustering dipergunakan untuk mengelompokkan 45 puskesmas di Kabupaten Banyuwangi berdasarkan target cakupan Imunisasi Dasar Lengkap (IDL). Penelitian tersebut menggunakan data imunisasi bayi, balita, dan anak-anak tahun 2016-2017. Hasil penelitian diperoleh tiga cluster puskesmas dengan status imunisasi berbeda, yaitu cukup, kurang, dan sangat baik. Cluster pertama berisi 19 puskesmas, cluster kedua berisi 24 puskesmas, dan cluster ketiga berisi 2 puskesmas.

Selain itu, ada penelitian lain tentang pengelompokan data puskesmas adalah [14], Puskesmas Bahorok mengambil keputusan untuk mengkategorikan penyakit berdasarkan usia pasien. Hal ini menghasilkan analisis terhadap 20 titik data yang dibagi menjadi 3 kelompok. Penelitian lain juga dilakukan menggunakan metode K-Medoids [15], penelitian tersebut dilakukan untuk mengelompokkan dan mengetahui penyakit yang paling sering dialami oleh pasien. Kemudian ada juga clustering data catatan medis untuk identifikasi penyakit endemik menggunakan Fuzzy C-Means [5]. Dari penelitian sebelumnya, dapat disimpulkan bahwa metode K-Means sangat efektif dalam menangani data dengan jumlah variabel yang besar dan jumlah cluster yang kecil. Maka dari itu diterapkan algoritma K-Means Clustering untuk mengkategorikan penyakit berdasarkan usia pasien di Puskesmas Tlogomulyo.

2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Ada beberapa langkah yang diterapkan dalam penelitian clustering dengan metode K-Means pada data pasien di Puskesmas Tlogomulyo. Tahapan penelitian antara lain identifikasi masalah, pengumpulan data, preprocessing data, proses clustering, dan yang terakhir pengujian. sebagaimana yang ditampilkan dalam Gambar 1.



Gambar 1. Tahapan Penelitian

2.2 Identifikasi Masalah

Puskesmas menghadapi tantangan dalam penanganan data penyakit karena ketidakteraturan dan banyaknya informasi. Akibatnya, sulit untuk mengidentifikasi pola atau korelasi antar penyakit yang ada. Oleh karena itu, pengolahan data penyakit diperlukan untuk membantu mengungkap informasi mengenai jenis penyakit yang paling sering muncul dan hubungan antara usia pasien dengan penyakit tersebut[16].

2.3 Pengumpulan Data

Pengumpulan data dilaksanakan untuk memperoleh informasi yang dibutuhkan untuk mencapai tujuan atau target penelitian. Pengumpulan data yang dilakukan pada penelitian ini berupa data sekunder yang didapat langsung dari Puskesmas Tlogomulyo Temanggung. Data tersebut berupa data register pasien bulan Januari dan Februari dengan jumlah data sebanyak 1415 data pada tahun 2022, yang digunakan untuk menganalisis perubahan yang mungkin terjadi pada awal tahun tersebut. Data yang digunakan terdiri dari 4 atribut yang berisikan desa, diagnosa, umur dan jenis kelamin pasien.

Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 639−647
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

2.4 Preprocessing Data

Langkah preprocessing data merupakan langkah yang penting dalam proses clusterisasi karena menentukan hasil cluster [17]. Dari data yang telah dimiliki, akan diolah menggunakan metode K-Means Clustering untuk mengetahui pengelompokkan data sehingga akan menampilkan hasil dari clustering.

a. Data Cleaning
   Sebelum melakukan proses data mining, pertama perlu dilakukannya proses data cleaning. Proses data cleaning meliputi menghilangkan duplikasi data, mengoreksi data yang tidak konsisten, dan memperbaiki kesalahan pada data seperti kesalahan penulisan [18] [19]. Pada proses ini dilakukan imputasi data. Imputasi data adalah proses penggantian atau pengisian nilai yang kosong atau hilang dalam dataset dengan nilai yang dapat diperkirakan, seperti menggunakan rerata, median, atau metode prediksi berbasis statistik atau machine learning.
b. Data Transfomasi
   Proses transformasi data sangat penting karena memungkinkan konversi format data asli menjadi bentuk yang sesuai yang dapat dianalisis atau dimodelkan lebih lanjut. Ketika data awal berupa teks atau kata, seperti nama diagnosis, maka perlu melalui proses inisialisasi data untuk mengubahnya menjadi bentuk numerik. Transformasi data membantu meningkatkan pemahaman data, mengurangi bias, meningkatkan kualitas data, dan memastikan bahwa data disiapkan untuk tahap selanjutnya dalam proses analisis.
c. Normalisasi
   Tahap selanjutnya adalah normalisasi. Normalisasi juga disebut sebagai proses pengubahan data ke dalam bentuk standar atau 'normal' agar memudahkan proses dan evaluasi data. Pada penelitian ini, Min-Max Normalization digunakan sebagai pendekatan sederhana dimana data dapat disesuaikan ke dalam rentang yang telah ditetapkan sebelumnya dengan batas yang telah ditentukan sebelumnya. Berikut adalah rumus minmax [12].

$$X_n = \frac{(X_0 - X_{min})(Maxvalue - Minvalue)}{(X_{max} - X_{min}) + Minvalue} \quad (1)$$

   Teknik tersebut akan menghasilkan data dengan kisaran nilai antara 0 hingga 1. Teknik ini berguna untuk standarisasi nilai dalam rentang yang ditargetkan. Pendekatan ini akan digunakan untuk normalisasi data pasien puskesmas berdasarkan seluruh dataset, yang akan mempermudah dengan metode K-Means Clustering.

2.5 Algoritma K-Means

Algoritma K-Means merupakan algoritma pengelompokkan data berdasarkan titik pusat cluster (centroid) terdekat dengan setiap data [20] [21]. K-Means adalah algoritma pengelompokan yang banyak digunakan untuk kumpulan data besar yang hanya bekerja dengan data numerik. Algoritma dimulai dengan memilih titik pusat massa secara acak dan menentukan jumlah cluster yang diinginkan. Kemudian data tersebut dikelompokkan ke dalam cluster-cluster berdasarkan kriteria yang telah ditentukan, dimana setiap cluster mempunyai titik pusat yang disebut centroid. Proses ini diulangi hingga titik pusat massa akhir yang optimal ditemukan. Berikut adalah langkah – langkah untuk menerapkan algoritma K-Means pada clustering data penyakit berdasarkan usia [22]:

a. Menentukan Jumlah Cluster (k)
   1. Jumlah cluster (k) merupakan parameter yang harus ditentukan sebelum menjalankan algoritma K-Means.
   2. Dapat menggunakan metode silhouette score untuk menentukan jumlah cluster yang optimal.
b. Inisialisasi Pusat Cluster
   1. Pusat cluster awal dapat diinisialisasi secara acak atau menggunakan metode lain.
   2. Pada kasus ini, lebih baik menginisialisasi pusat cluster menggunakan rentang usia yang ada dalam dataset sebagai panduan awal.
c. Iterasi Algoritma K-Means
   Mulai iterasi algoritma dengan langkah – langkah seperti:
   1. Menghitung jarak antar setiap titik data dan pusat cluster menggunakan matrik jarak Euclidean atau metrik jarak lainnya.

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (2)$$

   2. Menentukan cluster terdekat untuk setiap titik data berdasarkan jarak terkecil.
   3. Menghitung pusat baru untuk setiap cluster dengan mengambil nilai rata-rata dari semua titik data dalam cluster tersebut.
   4. Ulangi langkah – langkah ini hingga konvergensi, yaitu ketika pusat cluster tidak berubah secara signifikan antara iterasi.

2.6 Pengujian Hasil Clustering

Setelah proses algoritma K-Means selesai dan telah konvergen, lankah selanjutnya adalah melakukan evaluasi terhadap hasil clustering yang telah diperoleh. Evaluasi ini dilakukan dengan melihat pola atau karakteristik

Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 639–647
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

penyakit yang muncul dalam setiap cluster berdasarkan usia pasien [23]. Dengan demikian dapat dipahami bagaimana distribusi penyakit yang berbeda pada kelompok usia yang berbeda dan mengidentifikasi pola khusus yang mungkin muncul dalam setiap cluster. Untuk membantu analisis hasil pengelompokan dari algoritma K-Means, dapat menggunakan visualisasi. Salah satu cara yang efektif adalah dengan membuat plot yang menggambarkan hubungan antara usia dan cluster yang telah dibentuk. Dengan visualisasi ini, dapat dengan jelas melihat bagaimana distribusi usia pasien di setiap cluster dan apakah ada pola tertentu yang menonjol. Plot usia versus cluster ini akan memudahkan kita dalam mengidentifikasi apakah ada tren atau pola khusus dalam distribusi penyakit berdasarkan usia di setiap cluster. Hasil cluster yang sudah terbentuk akan dianalisis dengan menggunakan silhouette coefficient. Silhouette coefficient digunakan untuk mengukur seberapa baik suatu objek ditempatkan dalam cluster tertentu berdasarkan jarak antara objek tersebut dengan objek di cluster yang sama dibandingkan dengan objek di cluster lain. Silhouette coefficient akan memberikan skor antara -1 hingga 1, di mana skor yang lebih tinggi menunjukkan bahwa objek tersebut ditempatkan dengan baik dalam clusternya dan memiliki jarak yang cukup jauh dengan cluster lain. Dengan menganalisis silhouette coefficient ini, dapat menilai kekuatan dan kualitas dari setiap cluster yang terbentuk. Sehingga, informasi ini dapat digunakan untuk memvalidasi hasil clustering, mengidentifikasi apakah pengelompokan yang diperoleh sudah optimal, serta untuk memahami sejauh mana objek-objek dalam data ditempatkan dengan tepat dalam cluster-cluster yang ada. Analisis menggunakan silhouette coefficient ini akan memberikan wawasan tambahan dalam memahami struktur dan karakteristik dari data yang telah dikelompokkan. Rumus skala silhouette coefficient dapat dilihat pada Tabel 1.

Tabel 1. Skala Silhoutte Coefficient

Skala	Keterangan
0.7 < SC <= 1	Strong Structure Medium
0.5 < SC <= 0.7	Medium Structure
0.25 < SC <= 0.5	Weak Structure
SC <= 0.25	No Structure

3. HASIL DAN PEMBAHASAN

Pada penelitian ini menggunakan dataset sebanyak 1415 data, dengan 4 atribut yang digunakan terdiri dari desa, diagnosa, umur, dan jenis kelamin pasien. Dataset penelitian dapat dilihat pada Gambar 2

No	Tgl Daftar	Tgl Periksa	No Reg	Nama	Tgl Lahir	Umur	Jenis Kelamin	Desa	Jenis Kujungan	Diagnosis	Jenis Pasien
1	26-02-\n2021	03-01-\n2022	21.00.000305	SITI AZIFAH	18-\n08-\n1997	25,8	P	Losari	Kunjungan umum (BP Umum)	Abses	lama
2	25-10-\n2019	03-01-\n2022	19.00.002228	MUSLIH KHUDIN	15-\n05-\n2017	6,1	L	Langgeng	Kunjungan umum (BP Umum)	ISPA	lama
3	16-01-\n2021	03-01-\n2022	21.00.000109	MUHAMMAD FARID ATTHOILLAH	07-\n07-\n2015	7,9	L	Legoksari	Kunjungan umum (BP Umum)	Vulnus Laceratum (\nLaserasi / Luka Robek\n)	lama
4	20-04-\n2015	03-01-\n2022	15.00.006416	RISKI SETIAWAN	06-\n09-\n2007	15,7	L	Gedegan	Kunjungan umum (BP Umum)	Skabies	baru
5	30-12-\n2021	03-01-\n2022	21.00.002519	CHOERUN	13-\n08-\n1968	54,8	L	Balerejo	Kunjungan umum (BP Umum)	Lipoma	lama
...	...	...	...	...	...	...	...	...	...	...	...
1411	12-\n11-\n2021	26-02-\n2022	21.00.002151	TAAT	11-\n12-\n1995	27,5	L	Losari	Kunjungan umum (BP Umum)	Sehat	baru
1412	02-\n02-\n2018	26-02-\n2022	18.00.000209	munjaemah	12-\n08-\n1966	56,8	P	Jragan	Kunjungan umum (BP Umum)	Otitis Media Akut (O M A)	baru
1413	27-\n09-\n2019	26-02-\n2022	19.00.002030	DICKI FEBRIAN	18-\n03-\n2019	4,2	L	Losari	KIA Ibu / Anak	ISPA	lama
1414	20-\n04-\n2015	26-02-\n2022	15.00.002014	RIZKY RAMADANI IMAWATI	30-\n12-\n1999	23,4	P	Tanjungsari	KIA Poli KB	Suntik	baru
1415	21-\n09-\n2016	26-02-\n2022	16.00.007524	dewi indah sartika	13-\n09-\n1989	33,7	P	Tanjungsari	KIA Poli KB	Suntik	baru

Gambar 2. Dataset Penelitian

3.1 Preprocessing Data

a. Data Cleaning
Cleaning data pada tahap ini yang dilakukan yaitu pengecekan missing value dan imputasi data dengan mengisi nilai-nilai yang hilang atau NaN (Not a Number) dari beberapa tabel yang kosong di setiap atribut yang ada dengan tujuan untuk menghindari kesalahan perhitungan agar hasil pengelompokan menjadi akurat. Gambar 3 adalah cleaning dataset yang dilakukan

```
value = df['Desa'].mode()[0]
df['Desa']= df['Desa'].fillna(value)
```

```
Desa            50
Jenis Kujungan   6
Diagnosis       41
Jenis Pasien    42
dtype: int64
```

Gambar 3. Cleaning Dataset

Pada Gambar 3 terdapat baris kode yang digunakan untuk imputation. Imputation adalah teknik untuk menghitung kembali nilai yang hilang dalam suatu data. Dengan menggunakan potongan kode tersebut, nilai-

Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 639−647
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

nilai kosong dalam kolom, akan diisi dengan nilai mode dari kolom tersebut. Imputasi data atribut desa bersifat kategorikal, penggunaan mode sebagai metode pengisian nilai yang hilang menjadi relevan karena mode merepresentasikan nilai yang paling sering muncul dalam distribusi atribut desa. Berbeda dengan data numerik yang menggunakan mean atau median, data kategorikal tidak memiliki nilai tengah yang dapat dihitung, sehingga mode menjadi pilihan yang lebih sesuai untuk menggantikan nilai yang kosong dengan nilai yang paling umum muncul dalam atribut desa, memberikan representasi yang tepat terhadap karakteristik unik dari data kategorikal tersebut.

b. Transformasi
Tahapan berikutnya adalah transformasi data, yaitu mengubah format atribut menjadi tipe data numerik (tipe data integer untuk angka). Pada tahap ini, transformasi dilakukan dengan merubah isi variable desa dan diagnosis menjadi numerik agar dapat dilakukan K-Means clustering. Dikarenakan algoritma ini membutuhkan representasi numerik yang kontinu untuk mengukur jarak antara titik data. Transformasi data dapat dilihat pada Gambar 4

	Umur	Desa	Diagnosis
0	25.8	31	1
1	6.1	29	67
2	7.9	30	143
3	15.7	18	119
4	54.8	8	86

Gambar 4. Transformasi Data

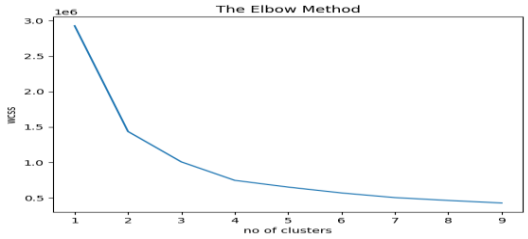
c. Normalisasi
Setelah transformasi, kemudian tahap normalisasi. Normalisasi data dilakukan untuk menyesuaikan nilai-nilai data ke dalam suatu rentang yang ditentukan, memudahkan perhitungan seperti kesamaan atau operasi clustering. Salah satu metode normalisasi yang umum digunakan adalah metode Min-Max, sebuah teknik sederhana dalam proses skala nilai berdasarkan batas-batas yang ditetapkan. Tujuan utama normalisasi adalah agar variabel-variabel dengan skala yang berbeda dapat diakomodasi dengan baik dalam analisis, meningkatkan konvergensi algoritma, dan memastikan kontribusi setiap variabel seimbang dalam proses perhitungan. Gambar 5 adalah normalisasi data

```
array([[0.27008929, 0.64583333, 0.00689655],
       [0.05022321, 0.60416667, 0.46206897],
       [0.0703125 , 0.625     , 0.9862069 ],
       ...,
       [0.02901786, 0.64583333, 0.46206897],
       [0.24330357, 0.83333333, 0.86206897],
       [0.35825893, 0.83333333, 0.86206897]])
```

Gambar 5. Normalisasi Data

3.2 Proses Clustering

a. Menentukan Jumlah Cluster
Dalam menentukan jumlah cluster, menggunakan Elbow Method. Tujuan dari proses ini adalah untuk menemukan jumlah cluster yang paling optimal dengan menghitung Within-cluster Sum of Squares (WCSS) dari iterasi 1 sampai 10. Komponen K-Means dalam WCSS ini menggambarkan total jarak kuadrat antara setiap titik data dengan pusat cluster masing-masing. Ketika WCSS dipetakan dengan nilai K, grafiknya akan menunjukkan sebuah “siku” atau elbow saat jumlah cluster meningkat. Berdasarkan analisis dengan dengan Elbow Method dan WCSS, diperoleh bahwa 2 cluster adalah pilihan optimal untuk implementasi K-Means seperti pada Gambar 6

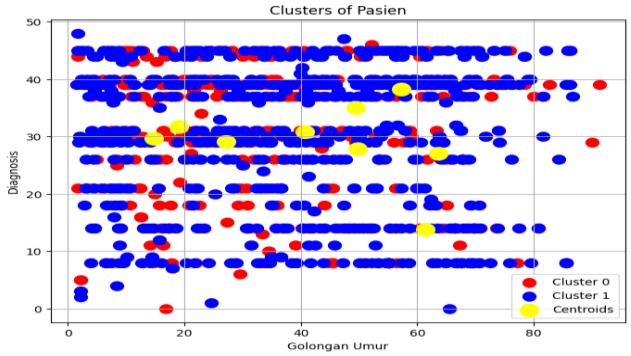


Gambar 6. Menentukan Jumlah Cluster

b. Clustering
Dalam langkah ini dilakukan proses K-Means clustering dataset. Hasil scatter plot dapat dilakukan dengan melihat label cluster berdasarkan warna plot. Terdapat dua cluster data yaitu cluster 0 dan cluster 1, yang

Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 639–647
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

direpresentasikan oleh titik biru dan merah. Titik kuning yang diberi label “Centroids”, mewakili titik pusat dari masing-masing cluster. Titik biru (Cluster 1) lebih tersebar di seluruh kelompok usia dan tingkat diagnosis. Sedangkan titik merah (Cluster 2) terutama terkonsentrasi di bagian bawah plot, menunjukkan tingkat diagnosis yang lebih rendah di berbagai kelompok usia. Plot dapat dilihat pada Gambar 7



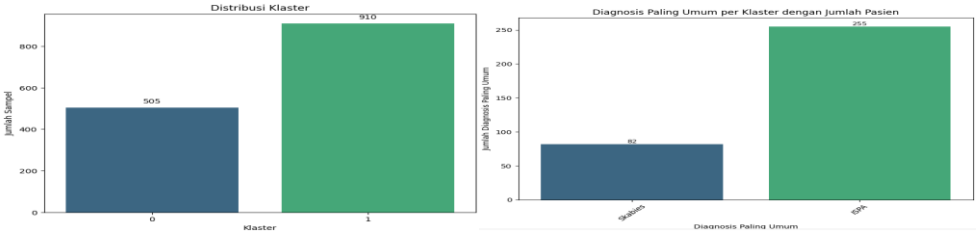
Gambar 7. Scatter Plot

c. Hasil Clustering
Setelah proses clustering selesai, hasil cluster disajikan seperti pada Gambar 8.

	No	Tgl Daftar	Tgl Periksa	No Reg	Nama	Tgl Lahir	Umur	Jenis Kelamin	Desa	Jenis Kujungan	Diagnosis	Jenis Pasien	klaster
0	1	26-02-\n2021	03-01-\n2022	21.00.000305	SITI AZIFAH	18-\n08-\n1997	25.8	P	Losari	Kunjungan umum (BP Umum)	Abses	lama	1
1	2	25-10-\n2019	03-01-\n2022	19.00.002228	MUSLIH KHUDIN	15-\n05-\n2017	6.1	L	Langgeng	Kunjungan umum (BP Umum)	ISPA	lama	1
2	3	16-01-\n2021	03-01-\n2022	21.00.000109	MUHAMMAD FARID ATTHOILLAH	07-\n07-\n2015	7.9	L	Legoksari	Kunjungan umum (BP Umum)	Vulnus Laceratum (Laserasi / Luka Robek)	lama	0
3	4	20-04-\n2015	03-01-\n2022	15.00.006416	RISKI SETIAWAN	06-\n09-\n2007	15.7	L	Gedegan	Kunjungan umum (BP Umum)	Skabies	baru	0
4	5	30-12-\n2021	03-01-\n2022	21.00.002519	CHOERUN	13-\n08-\n1968	54.8	L	Balerejo	Kunjungan umum (BP Umum)	Lipoma	lama	0
5	6	15-11-\n2019	03-01-\n2022	19.00.002390	SRIYAMI	22-\n03-\n1960	63.2	P	Tlogomulyo	Kunjungan umum (BP Umum)	Gastritis	lama	1
6	7	03-01-\n2022	03-01-\n2022	22.00.000002	retno setyo p	05-\n10-\n1968	54.7	P	Manding	Kunjungan umum (BP Umum)	Konjungtivitis Alergika	baru	1
7	8	08-02-\n2011	03-01-\n2022	14.00.001794	ALDO KRISTIANTO	17-\n08-\n1996	26.8	L	Tlogomulyo	Kunjungan umum (BP Umum)	Konjungtivitis Alergika	baru	1
8	9	08-02-\n2011	03-01-\n2022	13.00.000255	PONIAH	18-\n07-\n1973	49.9	P	Pagersari	Kunjungan umum (BP Umum)	Epilepsi	baru	1

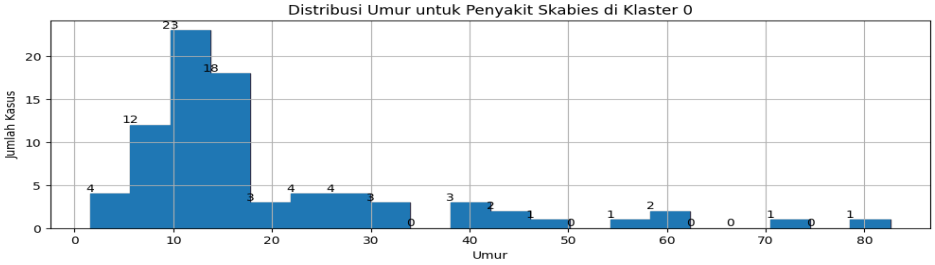
Gambar 8. Hasil Clustering

Pada analisis clustering, menghasilkan 2 cluster dengan masing-masing memiliki jumlah data 505 pada cluster 0 dan 910 pada cluster 1. Hasil analisis clustering menunjukkan bahwa setiap kelompok atau cluster memiliki pola keseragaman dalam distribusi usia. Namun, perbedaan yang signifikan antara cluster tersebut tercermin dalam jenis diagnosis yang paling umum dialami oleh anggota dari setiap cluster. Pada cluster 0, skabies merupakan diagnosis terbanyak dengan total 82 data, sedangkan pada cluster 1, ISPA menjadi diagnosis terbanyak dengan jumlah data mencapai 255. Visualilasi jumlah data dapat dilihat pada Gambar 9.



Gambar 9. Bar Plot Jumlah Data dan Diagnosis Cluster 0 dan 1

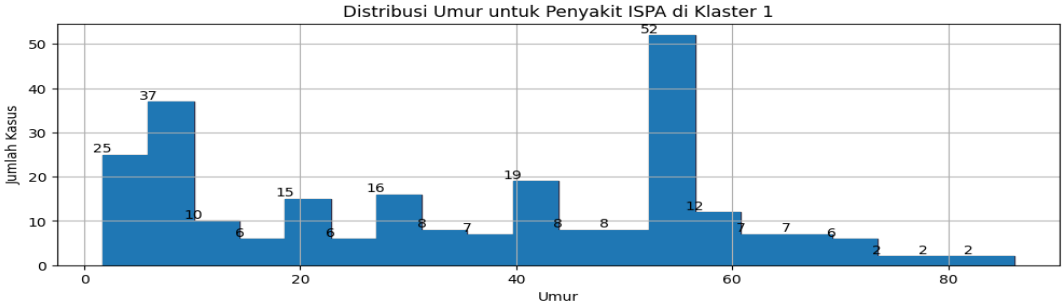
d. Visualisasi
Setelah dilakukan proses clustering yang menghasilkan 2 cluster yaitu cluster 0 dan cluster 1, selanjutnya hasil dari cluster tersebut divisualisasikan dalam bentuk bar plot seperti pada Gambar 10 dan Gambar 11



Gambar 10. Bar Plot Distribusi Umur Penyakit Cluster 0

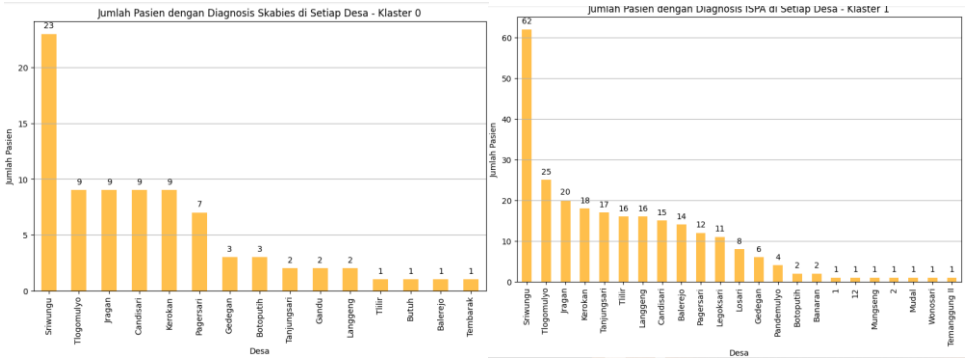
Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 639–647
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

Dari gambar tersebut, dapat disimpulkan bahwa pada cluster 0 terdapat prevalensi atau angka kejadian yang signifikan dari diagnosis skabies di rentang usia 10 hingga 20 tahun. Temuan ini menunjukkan kecenderungan bahwa kelompok individu dalam rentang usia tersebut memiliki tingkat kejadian skabies yang lebih tinggi dibandingkan dengan kelompok usia lainnya.



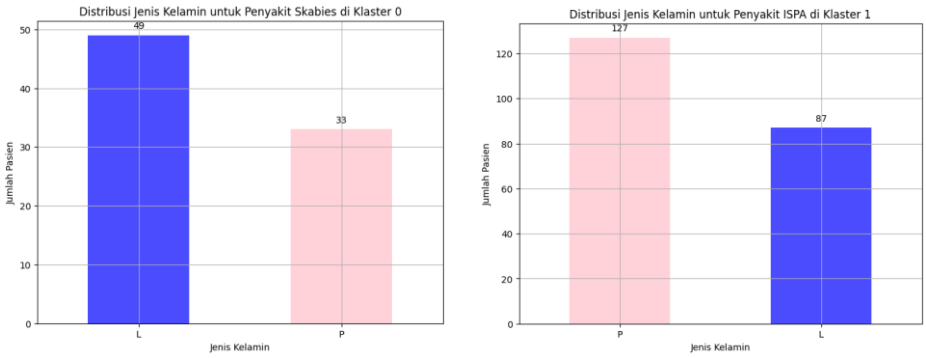
Gambar 11. Bar Plot Distribusi Umur Penyakit Cluster 1

Dan pada gambar 11, terungkap bahwa pada cluster 1 terdapat kecenderungan yang mencolok terkait dengan diagnosis Infeksi Saluran Pernapasan Akut (ISPA), dengan tingkat kejadian yang lebih tinggi terfokus pada pasien dengan usia sekitar 50 hingga 55 tahun. Hasil ini mengindikasikan bahwa penyakit ISPA lebih banyak diderita oleh usia dewasa. Hasil dari clustering ini juga menunjukan bahwa Desa Sriwungu menjadi wilayah terbanyak dari kedua jenis penyakit tersebut. Pada cluster 0, penyakit skabies banyak dialami oleh pasien dari wilayah Sriwungu dengan total 23 dari 82 penderita skabies. Demikian juga dengan penderita ISPA pada cluster 1 yang banyak dialami oleh pasien dari daerah Sriwungu, total pasien tersebut ada 62 dari 255 penderita ISPA. Visualisasi jumlah desa dengan penderita terbanyak dapat dilihat pada Gambar 12



Gambar 12. Bar Plot Desa dengan Penderita Terbanyak Dari Cluster 0 Dan 1

Gambar 13 adalah bar plot yang memperlihatkan distribusi jenis kelamin pada pasien yang didiagnosis mengalami skabies dan ISPA dalam cluster 0 dan cluster 1. Pada cluster 0, penyakit skabies terbanyak terjadi pada pasien laki-laki dengan jumlah 49, sedangkan pasien perempuan sebanyak 33. Sementara itu, pada cluster 1, penyakit ISPA dominan dialami oleh pasien perempuan dengan jumlah 127, sedangkan pasien laki-laki mencapai 87 seperti yang terlihat pada Gambar 13



Gambar 13. Bar Plot Jenis Kelamin dengan Penderita Terbanyak Dari Cluster 0 Dan 1

3.3 Pengujian

Setelah menerapkan algoritma K-Means untuk analisis clustering, langkah berikutnya adalah mengevaluasi kualitas model dengan menggunakan Silhouette Coefficient. Tujuan dari evaluasi ini adalah untuk mengetahui

Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 639–647
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

kualitas dari suatu objek pada suatu cluster dan membantu kita menilai seberapa baik kualitas keseluruhan dari model tersebut. Silhouette Coefficient memberikan indikasi tentang seberapa baik definisi cluster dan seberapa baik objek-objek berada dalam cluster yang sesuai. Gambar 14 menunjukan hasil dari pengujian pada kualitas cluster diperoleh nilai Silhouette Coefficient sebesar 0,4374.

```
silhouette = silhouette_score(X, y_kmeans)
silhouette

0.43740591584286675
```

Gambar 14. Hasil Silhouette Coefficient

4. KESIMPULAN

Hasil penelitian tentang Clustering Data Penyakit Berdasarkan Usia dengan Metode K-Means menunjukkan adanya 2 cluster yang dapat diidentifikasi dari total 1415 data penyakit yang dianalisis. Cluster pertama terdiri dari 505 data, yang secara signifikan didominasi oleh kasus penyakit skabies (82 data). Lebih spesifik, penyakit ini umumnya dialami oleh pasien dengan rentang usia 10 sampai 20 tahun, dengan jumlah penderita laki-laki lebih banyak dengan jumlah 49 pasien, dan kejadian yang lebih tinggi terdapat di desa Sriwungu, mencapai lebih dari 20 kasus dalam cluster ini. Sementara itu, cluster kedua terdiri dari 910 data, menunjukkan prevalensi yang lebih tinggi dari penyakit Infeksi Saluran Pernapasan Akut (ISPA) dengan jumlah kasus mencapai 255 data. Analisis lebih lanjut menunjukkan bahwa ISPA paling sering dijumpai pada pasien dengan usia 50 hingga 55 tahun, berjenis kelamin perempuan dengan jumlah 224 pasien, dan lebih banyak terjadi di desa Sriwungu, dengan jumlah lebih dari 60 kasus. Dari hasil analisis Silhouette Coefficient pada Clustering Data Penyakit Berdasarkan Usia dengan menggunakan Metode K-Means menunjukkan bahwa pembentukan 2 cluster dengan nilai optimal yaitu sebesar 0.4374. Nilai Silhouette Coefficient (SC) dengan nilai mendekati 1 dianggap sebagai nilai yang paling optimal karena menunjukkan efektivitas yang baik dalam pembentukan cluster. Berdasarkan hasil nilai pengujian yang dihasilkan diperoleh bahwa nilai cluster termasuk kedalam cluster yang lemah. Hasil ini memberikan wawasan yang berharga terkait distribusi penyakit dalam kelompok usia tertentu dan dapat menjadi dasar untuk pengembangan strategi penanggulangan penyakit yang lebih efektif.

REFERENCES

[1] R. A. Farissa, R. Mayasari, and Y. Umaidah, “Perbandingan Algoritma K-Means dan K-Medoids Untuk Pengelompokkan Data Obat dengan Silhouette Coefficient di Puskesmas Karangsambung,” J. Appl. Informatics Comput., vol. 5, no. 2, pp. 109–116, 2021, doi: 10.30871/jaic.v5i1.3237.
[2] D. Ariyanto, “Data Mining Menggunakan Algoritma K-Means untuk Klasifikasi Penyakit Infeksi Saluran Pernafasan Akut,” J. Sistim Inf. dan Teknol., vol. 4, pp. 13–18, 2022, doi: 10.37034/jsisfotek.v4i1.117.
[3] A. Y. Simanjuntak, I. S. E. S. Simatupang, and A. Anita, “Implementasi Data Mining Menggunakan Metode Naã• Ve Bayes Classifier Untuk Data Kenaikan Pangkat Dinas Ketenagakerjaan Kota Medan,” J. Sci. Soc. Res., vol. 5, no. 1, p. 85, 2022, doi: 10.54314/jssr.v5i1.804.
[4] N. Klakotskaya, P. Laurson, A. V. Libek, and A. Kikas, “Assessment of the Aim Characteristics of Strawberry (Fragaria × Ananassa) Cultivars in Estonia by Using the K-Means Clustering Method,” Horticulturae, vol. 9, no. 1, 2023, doi: 10.3390/horticulturae9010104.
[5] D. Andreswari, R. Efendi, and K. Prastio, “Clustering Data Rekam Medis untuk Penentuan Penyakit Endemi di Daerah Kabupaten Bengkulu Selatan dengan Mengimplementasikan Metode FUZZY C-MEANS,” J. Rekursif, vol. 11, no. 1, pp. 42–52, 2023, doi: 10.33369/rekursif.v11i1.23274.
[6] D. Gustian and M. S. Al-Farits, “Data Mining Untuk Melihat Minat Belajar Siswa Menerapkan Metode K-Means,” J. Inf. Syst. Res., vol. 4, no. 3, pp. 775–784, 2023, doi: 10.47065/josh.v4i3.3218.
[7] D. A. I. C. Dewi and D. A. K. Pramita, “Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali,” Matrix J. Manaj. Teknol. dan Inform., vol. 9, no. 3, pp. 102–109, 2019, doi: 10.31940/matrix.v9i3.1662.
[8] T. D. Harjanto, A. Vatresia, and R. Faurina, “Analisis Penetapan Skala Prioritas Penanganan Balita Stunting Menggunakan Metode DBSCAN Clustering (Studi Kasus Data Dinas Kesehatan Kabupaten Lebong),” Rekursif J. Inform., vol. 9, no. 1, pp. 30–42, 2021, doi: 10.33369/rekursif.v9i1.14982.
[9] U. I. N. Ar-raniry, F. Tarbiyah, D. A. N. Keguruan, P. Studi, and P. Teknologi, “Perancangan Media Pembelajaran Bahasa Pemrograman Python Menggunakan Aplikasi Scratch Untuk Siswa Sekolah Menengah,” p. 72, 2022.
[10] A. Praja, C. Lubis, and D. E. Herwindiati, “Deteksi Penyakit Diabetes Dengan Metode Fuzzy C-Means Clustering Dan K-Means Clustering,” Comput. J. Comput. Sci. Inf. Syst., vol. 1, no. 1, p. 15, 2017, doi: 10.24912/computatio.v1i1.233.
[11] A. Triono, A. S. Budi, and R. Abdillah, “Implementasi peretasan sandi vigenere chipher menggunakan bahasa pemrograman python,” vol. 1, no. 1, pp. 1–9, 2023.
[12] A. Chusyairi and P. Ramadar Noor Saputra, “Pengelompokan Data Puskesmas Banyuwangi Dalam Pemberian Imunisasi Menggunakan Metode K-Means Clustering,” Telematika, vol. 12, no. 2, pp. 139–148, 2019, doi: 10.35671/telematika.v12i2.848.
[13] A. Maulana, A. Nazir, R. M. Candra, S. Sanjaya, and F. Syafria, “Clustering Vaksinasi Penyakit Mulut dan Kuku Menggunakan Algoritma K-Means,” J. Inf. Syst. Res., vol. 4, no. 3, pp. 894–902, 2023, doi: 10.47065/josh.v4i3.3363.

Journal of Information System Research (JOSH)
Volume 5, No. 2, Januari 2024, pp 639–647
ISSN 2686-228X (media online)
https://ejurnal.seminar-id.com/index.php/josh/
DOI 10.47065/josh.v5i2.4755

[14] B. Serasi Ginting and M. Simanjuntak, "Pengelompokan Penyakit Pada Pasien Berdasarkan Usia Dengan Metode K-Means Clustering (Studi Kasus : Puskesmas Bahorok)," Algoritm. J. Ilmu Komput. dan Inform., vol. 6341, no. November, p. 2, 2021.
[15] D. U. Iswavigra, S. Defit, and G. W. Nurcahyo, "Data Mining dalam Pengelompokan Penyakit Pasien dengan Metode K-Medoids," J. Inf. dan Teknol., vol. 3, pp. 181–189, 2021, doi: 10.37034/jidt.v3i4.150.
[16] M. L. Kolling et al., "Data mining in healthcare: Applying strategic intelligence techniques to depict 25 years of research development," Int. J. Environ. Res. Public Health, vol. 18, no. 6, pp. 1–21, 2021, doi: 10.3390/ijerph18063099.
[17] B. H. Prakoso, E. Rachmawati, D. R. P. Mudiono, V. Vestine, and G. E. J. Suyoso, "Klasterisasi Puskesmas dengan K-Means Berdasarkan Data Kualitas Kesehatan Keluarga dan Gizi Masyarakat," J. Buana Inform., vol. 14, no. 01, pp. 60–68, 2023, doi: 10.24002/jbi.v14i01.7105.
[18] A. D. Andini and T. Arifin, "Implementasi Algoritma K-Medoids Untuk Klasterisasi Data Penyakit Pasien Di Rsud Kota Bandung," J. Responsif Ris. Sains dan Inform., vol. 2, no. 2, pp. 128–138, 2020, doi: 10.51977/jti.v2i2.247.
[19] B. Wira, A. E. Budianto, and A. S. Wiguna, "Implementasi Metode K-Medoids Clustering untuk Mengetahui Pola Pemilihan Program Studi," J. Terap. Sains Teknol., vol. 1, no. 3, pp. 54–69, 2019.
[20] N. Purba, P. Poningsih, and H. S. Tambunan, "Penerapan Algoritma K-Means Clustering Pada Penyebaran Penyakit Infeksi Saluran Pernapasan Akut (ISPA) di Provinsi Riau," J. Inf. Syst. Res., vol. 2, no. 3, pp. 220–226, 2021, [Online]. Available: http://ejurnal.seminar-id.com/index.php/josh/article/view/736.
[21] V. Herlinda and D. Darwis, "Analisis Clustering Untuk Recredesialing Fasilitas Kesehatan Menggunakan Metode Fuzzy C-Means," Darwis, Dartono, vol. 2, no. 2, pp. 94–99, 2021, doi: 10.33365/jtsi.v2i2.890.
[22] P. Zong, J. Jiang, and J. Qin, "Study of high-dimensional data analysis based on clustering algorithm," 15th Int. Conf. Comput. Sci. Educ. ICCSE 2020, no. Iccse, pp. 638–641, 2020, doi: 10.1109/ICCSE49874.2020.9201656.
[23] M. A. V. Ideal, "Classification of Patient Complaints against Patient Medical Record Data Using the K Means Method," J. Sistim Inf. dan Teknol., vol. 5, pp. 1–6, 2022, doi: 10.37034/jsisfotek.v5i1.151.