Journal of Artificial Intelligence and Software Engineering
Vol. 5, No. 3, September 2025, pp. 1119~1132
E-ISSN: 2777-001X, DOI: 10.30811/jaise.v5i3.7559

Application of K-Means Clustering Algorithm for Disease Grouping at Blessing Dental Care Clinic

Cintiya Aulya Fransiska[1], Dafid[2]
[1,2] Sistem Informasi, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang, Kota Palembang – Sumatera Selatan, 30113, Indonesia

Informasi Artikel

Diterima : 7 Agustus 2025
Revisi : 17 Agustus 2025
Publikasi : 30 September 2025

Kata Kunci:

Algoritma K-Means,
Clustering,
CRISP-DM,
Data Mining,
Python.

ABSTRAK

Klinik Blessing Dental Care merupakan sebuah klinik yang menyediakan pelayanan praktek dokter gigi dan praktek dokter umum yang bertempat di Palembang. Klinik ini menawarkan perawatan umum dan perawatan gigi yang dikelola oleh dokter yang berpengalaman di bidangnya. Fokus data yang digunakan merupakan data rekam medis, khususnya dari praktek umum. Pengelompokkan data yang besar menjadi beberapa kelompok berdasarkan karakteristik pola yang serupa dengan pemafaatan algoritma K-Means Clustering pada data mining CRISP-DM dipilih lebih efektif dalam menangani berbagai keluhan penyakit yang bervariasi melalui proses Clustering. Hasil penelitian menunjukkan bahwa bentuk cluster 1 sebanyak 220 data dominan kategori penyakit gangguan pernapasan, cluster 2 sebanyak 335 data dominan kategori penyakit gangguan kardiovaskular, cluster 3 sebanyak 584 data dominan kategori penyakit gangguan kardiovaskular, cluster 4 sebanyak 363 data dominan penyakit gangguan pernapasan, cluster 5 sebanyak 70 data dominan penyakit gangguan pernapasan, cluster 6 sebanyak 254 data dominan penyakit kardiovaskular dan cluster 7 sebanyak 165 data dominan penyakit THT. Pada cluster 7 dengan nilai SSE yang diperoleh 3189,16 menujukkan penurunan makin kecil dan pola menyebar mulai optimal dengan kencederungan pola lebih menyebar.

ABSTRACT

Blessing Dental Care Clinic is a clinic that provides dental practice services and general practitioner practices located in Palembang. This clinic offers general care and dental care managed by experienced doctors in their fields. The focus of the data used is medical record data, especially from general practice. Grouping large data into several groups based on similar pattern characteristics by utilizing the K-Means Clustering algorithm in CRISP-DM data mining was chosen to be more effective in handling various complaints of various diseases through the Clustering process. The results showed that the form of cluster 1 was 220 dominant data in the respiratory disease category, cluster 2 was 335 dominant data in the cardiovascular disease category, cluster 3 was 584 dominant data in the cardiovascular disease category, cluster 4 was 363 dominant data in respiratory disease, cluster 5 was 70 dominant data in respiratory disease, cluster 6 was 254 dominant data in cardiovascular disease and cluster 7 was 165 dominant data in ENT disease. In cluster 7 with an SSE value of 3189.16, the decrease is getting smaller and the spread pattern is starting to be optimal with a tendency for the pattern to be more spread out.

***Penulis Koresponden**

Email: cintiyaulya14@mhs.mdp.ac.id

Cara sitasi IEEE:

C.A. Fransiska & I. Dafid, "Application of K-Means Clustering Algorithm for Disease Grouping at Blessing Dental Care Clinic," *Journal of Artificial Intelligence and Software Engineering (J-AISE)*, vol. 5, no. 3, pp. 1119-1132, September 2025, doi: 10.30811/jaise.v5i3.7559

1. PENDAHULUAN

Pada era kemajuan digital saat ini, memungkinkan cepatnya proses pengumpulan data dan teknologi penyimpanan untuk menghimpun jumlah data suatu organisasi yang sangat luas. Dengan terciptanya efisiensi pengumpulan data yang besar tentunya memerlukan teknologi yang mendukung penemuan informasi baru dalam dataset. Evaluasi dan penyebaran hasil akan membantu meningkatkan pelayanan, meningkatkan kualitas layanan kesehatan pada klinik. Penerapan metode CRISP-DM (Cross Industry Standard Process for Data Mining) sebagai pendekatan sistematis dalam mengubah data rekam medis menjadi bentuk pengetahuan yang baru, dengan proses pengelompokkan penyakit dilakukan dengan menggunakan algoritma K-Means Clustering.

Data mining merupakan sebuah teknologi ekstrasi data dalam penyimpanan besar yang digunakan dalam sebuah organisasi untuk mengubah data mentah menjadi informasi pendukung dalam proses identifikasi pola terhadap suatu data seperti contoh, jenis penyakit yang paling sering terjadi pada organisasi guna mengantisipasi lonjakan kasus penyakit tertentu, sebagai pendukung pengambilan keputusan, dan meningkatkan efisiensi strategi bagi suatu organisasi. Penyakit merupakan penyebab kematian tertinggi di Indonesia, hal ini mencerminkan pentingnya perubahan dalam pola hidup dan peningkatan kesadaran kesehatan masyarakat.

Penyediaan obat yang tepat merupakan aspek krusial dalam upaya pencegahan dan pengobatan penyakit, peran serta individu untuk menjaga pola hidup sehat dan mengakses pelayanan kesehatan secara teratur sangat penting untuk mendukung upaya klinis dalaam menangani penyakit yang sering terjadi [1]. Hal ini dapat membantu instalasi kesehatan dalam mempersiapkan kebutuhan persediaan obat preventif untuk penyakit tertentu. Perencanaan akan kebutuhan obat-obatan merupakan salah satu aspek penting untuk pengelolaan obat-obatan, perencanaan akan kebutuhan obat-obatan yang tepat dapat membuat pengadaan obat-obatan menjadi lebih efektif dan efisien sehingga ketersediaan obat dapat cukup sesuai dengan kebutuhan [2]. Dengan menggunakan metode clustering tersebut diharapakan hasil dari pengelompokkan penyakit yang terjadi di instalasi kesehatan akan lebih tepat dan cepat guna melakukan penanganan terhadap penyakit yang sering terjadi [3].

Ketidakefisienan dalam pengelolaan dan analisis data dapat menyebabkan berbagai tantangan, seperti lambatnya respon terhadap tren penyakit yang menyebabkan kurang tepatnya perencanaan sumber data. Identifikasi pola dan cluster penyakit dari data pasien yang tersedia masih menjadi tantangan tersendiri. Analisis data pasien diperlukan untuk melakukan pengelompokkan penyakit agar dapat membantu dalam penanganan dan alokasi sumber daya supaya instalasi kesehatan dapat melakukan pencegahan dan pengendalian penyebaran penyakit [4].

Pada klinik Blessing Dental Care, data rekam medis pasien khususnya pada data dari praktek umum sistem operasional dirancang untuk menyimpan, mengelola, dan mengakses data pasien untuk mencatat riwayat kunjungan pasien, diagnosis, tindakan medis, dan resep obat secara terstruktur. Namun, semakin bertambahnya jumlah kunjungan pasien membuat klinik perlu melakukan pengecekan yang cepat untuk melihat diagnosis yang pernah dialami pasien sebelumnya serta menampilkan data statistik penyakit yang lebih akurat.

Dengan berbagai macam diagnosis yang telah tersimpan, proses pengecekan diagnosis yang dialami pasien sebelumnya memperoleh data berdasarkan karakteristik tertentu serta menganalisis tren penyakit mengalami kesulitan dikarenakan keterbatasan dalam menampilkan statistik pengelompokkan penyakit yang terjadi. Dalam mengatasi hal tersebut pemanfaatan algoritma K-Means Clustering pada data mining dipilih karena efektif dalam mengelompokkan data besar menjadi beberapa kelompok berdasarkan karakteristik pola yang serupa tanpa memerlukan label awal sehingga cocok untuk menangani data diagnosis yang beragam diklinik.

Dalam penelitian ini, fokus data yang digunakan merupakan data dari praktek dokter umum saja, tidak termasuk data dari perawatan gigi. Penelitian pada praktek umum menangani berbagai keluhan penyakit yang bervariasi sehingga data yang tersedia lebih cocok untuk dilakukan proses clustering. Diharapkan pengelolaan data pasien dapat memberikan solusi untuk memahami dan meningkatkan kualitas perawatan yang diberikan kepada pasien berdasarkan diagnosis penyakit yang dialami dapat membantu proses pengambilan keputusan untuk mengatasi permasalahan analisa klinik terhadap tren penyakit dan kebutuhan pasien.

Pendekatan IT dengan menggunakan teknik data mining dengan metode tersebut digunakan untuk membagi data menjadi subset data berdasarkan kesamaan yang telah ditentukan sebelumnya, sehingga

berdasarkan data rekam medis pasien yang tesimpan dapat ditemukan informasi baru [5]. Dalam konteks ini data mining dengan menggunakan metode K-Means Clustering dapat menjadi solusi yang efektif untuk mengelola data rekam medis di rumah sakit maupun klinik [6]. Pentingnya proses pencarian pola untuk mengelompokkan penyakit secara efisien sehingga dapat diperoleh suatu informasi yang tidak diketahui sebelumnya diharapkan dapat berguna dalam menunjang pengambilan keputusan [7]. Upaya ini dapat berkontribusi terhadap kualitas pelayanan kesehatan secara keseluruhan.

Dengan dilakukan identifikasi data berdasarkan pengelompokkan penyakit berdasarkan faktor pengelompokkan lain seperti usia, jenis kelamin, status kawin, pekerjaan, riwayat alergi, dan diagnosa untuk mengetahui tren penyakit dengan menggunakan bantuan algoritma *K-Means Clustering*. Membantu dalam menghubungkan pengecekan diagnosis dengan faktor lain yang berpengaruh pada diagnosis penyakit yang terjadi di Blessing Dental Care. Dengan mempercepat proses pengecekan diagnosa melalui pengelompokkan ciri-ciri yang serupa melalui hasil penelitian. Dengan penelitian ini bertujuan untuk mendukung tindakan cepat dalam penanganan penyakit dengan melihat potensi penyakit yang paling sering terjadi, dengan pengelompokkan jenis penyakit dari data yang digunakan agar sebuah instalasi kesehatan dapat mempersiapkan kebutuhan persediaan obat preventif untuk penyakit tertentu. Membantu peningkatan penanganan penyakit yang terjadi untuk mencegah ketidakefisienan penanganan penyakit.

2. METODE

2.1 Metodologi

Penggunaan metode data mining digunakan untuk pengelolaan data berskala besar, data dengan teknik ini digunakan untuk menghasilkan suatu pengetahuan yang baru melalui proses identifikasi dan pengelompokkan data berdasarkan karakteristik yang sama [8], dengan melalui proses eksplorasi dan analisis dengan menggali data mendukung fungsionalitas [9]. Menggunakan pendekatan data mining CRISP-DM sebagai langkah-langkah dalam melakukan penelitian. CRISP-DM memiliki standar data mining sebagai pemecahan masalah yang umum bagi penelitian, metode ini terdiri dari enam fase sebagai langkah-langkah melakukan penelitian yaitu *Bussines Understanding*, Data *Understanding*, Data *Preparation*, *Modeling*, *Evaluation*, *Deployment*[10].

1. *Bussines Understanding*
   Melakukan identifikasi permasalahan melalui proses wawancara dan pengumpulan data. Saat ini, klinik menghadapi tantangan dalam menganalisis tren penyakit untuk peningkatan kualitas perawatan serta pengecekan terhadap tipe golongan diagnosis yang dialami.
2. Data *Understanding*
   Proses pengumpulan data dan evaluasi dari data yang dapat digunakan dalam pengelompokkan data, dengan menggunakan data rekam medis pasien praktek umum yang didapat sebanyak 2.019 record dengan jumlah 12 atribut yaitu Nama_Pasien, Alamat, Nama_ortu/suami/istri, Pendidikan, Tanggal_kunjungan, Tahun_lahir, Jenis_kelamin, Status_kawin, Pekerjaan, Riwayat_alergi, Keluhan, Diagnosa. Data yang diperoleh merupakan data pasien dari tahun 2024 dari bulan Januari sampai Desember, penjelasan mengenai atribut dalam data terdapat pada tabel berikut.

Table 1. Deskripsi Atribut

No.	Atribut	Tipe Data	Deskripsi Atribut	Domain Value
1.	Nama_pasien	*String*	Keterangan nama lengkap pasien yang berkunjung ke klinik, digunakan sebagai identitas utama dalam pencatatan.	Nama pasien contoh “Dian”, “Reno”.
2.	Alamat	*String*	Informasi tempat tinggal pasien, untuk keperluan administratif untuk mengetahui wilayah pasien yang terbanyak.	Wilayah tempat tinggal seperti “Alang-alang lebar”, “Ilir timur II”
3.	Nama_ortu/suami/istri	*String*	Informasi anggota keluarga terdekat pasien yang mendampingi pasien.	Nama pendamping pasien contoh “Tina”, “Bayu”
4.	Pendidikan	*String*	Tingkat pendidikan terakhir pasien.	Seperti “S1”, “SMA”, “SD”
5.	Tanggal_kunjungan	*Date*	Untuk melihat waktu kunjungan pasien.	Format (DD/MM/YY).
6.	Tahun_lahir	*Number*	Menghitung usia pasien, dimana dapat digunakan untuk keperluan analisa penyakit berdasarkan usia.	Berupa angka biasanya terdiri dari 4 angka, contoh 1999
7.	Jenis_kelamin	*String*	Terdiri dari “Laki-laki” dan “Perempuan”, data ini dapat berperan dalam analisis penyakit berdasarkan gender.	Genre pasien contoh “Laki-laki”, “Perempuan”
8.	Status_kawin	*String*	Penyakit dapat disebabkan oleh gaya hidup maupun hubungan sosial pasien sehari-hari.	Status kawin seperti “Belum kawin”, “Kawin”, “Cerai”.

9.	Pekerjaan	*String*	Aktivitas kerja dapat menjadi faktor penyakit yang dialami.	Aktivitas yang sering dilakukan “Guru”, “Dokter”
10.	Riwayat_alergi	*String*	Informasi toleransi pasien terhadap bahan tertentu. Penting untuk keamanan pasien, mencegah kontak terhadap alergi.	Informasi alergi pasien, bisa ada maupun tidak. Jika ada contoh nya biasa berupa alergi “Seafood”, “Susu”, dll.
11.	Keluhan	*String*	Gejala kesehatan yang dialami pasien saat berkunjung untuk pemeriksaan.	Tanda-tanda ketika mengalami sakit contoh “Pusing”, “Mual”, dll.
12.	Diagnosa	*String*	Hasil pemeriksaan dari keluhan yang dialami pasien, menentukan jenis penyakit.	Nama penyakit dalam bahasa medis, yang di derita seperti “Ispa”, “Mumps”, dll.

3. Data *Preparation*

Sumber data dipersiapkan agar data sesuai dengan model data yang dibutuhkan, transformasi data meliputi proses seleksi dan pembersihan data dengan menggunakan tools *Google Colab*. Data yang sudah diidentifikasi akan dilakukan proses seleksi, pembersihan, dan diubah menjadi data baru yang telah disesuaikan dengan kebutuhan algoritma yang akan digunakan.

a. Data *Selection* (Pemilihan Data)

Proses memilih atribut dari kolom dan baris yang relevan untuk keperluan analisis, berikut atribut yang digunakan pada proses pemilihan data. Dengan populasi rentan menjadi sasaran yang memiliki alergi yang dirasakan berpengaruh terhadap karakteristik jenis kelamin, usia, alergi, dan status sosial [11]. Terdiri dari Tahun_lahir, Jenis_kelamin, Status_kawin, Pekerjaan, Riwayat_alergi, Diagnosa yang merupakan atribut cenderung berpengaruh terhadap kemungkinan diagnosa. Proses pembersihan dilakukan dengan menghapus terlebih dahulu kolom atribut yang tidak digunakan, penulis menggunakan tools *Google Colab*.

b. Data *Cleaning* (Pembersihan Data)

Pembersihan data yang digunakan untuk menghilangkan atribut yang tidak digunakan agar data tetap efisien untuk dilanjutkan ke penerapan algoritma. Proses pembersihan dilakukan dimana dilakukan hapus baris data dari beberapa baris pada data terdapat nilai *null*. Langkah ini melibatkan proses pembersihan data dengan menghapus atribut yang tidak digunakan, data yang tidak konsisten, dan perbaikan kesalahan penulisan [4].

c. Data *Transformation* (Transformasi Data)

Data yang sudah dilakukan seleksi kemudian diubah agar data sesuai dengan algoritma yang akan digunakan, seperti Tahun_lahir, Jenis_kelamin, Pekerjaan, Diagnosa. Hal ini dilakukan untuk menghindari ketidak konsistenan data, adapun hasil yang didapatkan setelah melalui proses transformasi. Mengubah atribut Tahun_lahir menjadi Usia, data yang diambil merupakan data tahun 2024. Maka untuk mengubah nya dilakukan pengurangan dari tahun data dengan tahun dari data pasien, Atribut Jenis_kelamin diubah kedalam bentuk angka untuk memudahkan proses pengelompokkan.

Table 2 Atribut Jenis_Kelamin

Atribut	Kode Numerik
Laki-Laki	1
Perempuan	2

Pasien memiliki berbagai macam pekerjaan yang berbeda-beda, dikarenakan terdapat banyaknya data maka dilakukan pengelompokkan berdasarkan kategori [12]. Dengan dilakukan mengelompokkan pekerjaan dengan karakteristik yang serupa menjadi masing-masing kategori [13]. Kemudian kategori tersebut diubah ke dalam bentuk numerik/penomoran.

Table 3 Atribut Kategori_pekerjaan

Atribut	Kode Numerik
Pendidikan	1
Fisik Berat	2
Fisik Aktif	3
Kantoran/Profesional	4
Rumah Tangga	5
Freelance	6
Industri Kreatif	7
Tidak Bekerja	8
Lainnya	9

Atribut status_kawin diubah menjadi tipe angka agar mempermudah proses pengelompokkan.

Table 4 Atribut Status_Kawin

Atribut	Kode Numerik
Kawin	1
Kawin (Cerai Mati)	2
Cerai	3
Belum Kawin	4

Riwayat alergi yang beragam juga menjadi faktor dilakukan pengelompokkan berdasarkan karakteristik alergi yang serupa. Dengan mengubah kedalam beberapa kelompok kategori alergi yang berikutnya diubah dalam tipe angka.

Table 5 Atribut Kategori_alergi

Atribut	Kode Numerik
Alergi Obat	1
Alergi Makanan/Minuman	2
Alergi Lingkungan	3
Alergi Penyakit	4
Alergi Tindakan Medis	5
Tidak ada	6

Atribut terakhir merupakan atribut diagnosis yang di kelompokkan berdasarkan karakteristik yang serupa [14], kemudian diubah ke tipe numerik. Atribut-atribut yang telah dipilih merupakan data yang akan digunakan untuk mendukung proses penentuan pengelompokkan pada penelitian ini.

Table 6 Atribut Kategori_diagnosis

Atribut	Kode Numerik
Penyakit Infeksi & Inflamasi	1
Gangguan Saluran Cerna/Reproduksi	2
Gangguan Darah & Hematologi	3
Gangguan Autoimun & Reumatik	4
Gangguan Pernapasan	5
Gangguan Kardiovaskular	6
Gangguan Saraf & Psikologis	7
Gangguan Kulit & Rambut	8
Kanker/Neoplasma	9
Gangguan Muskuloskeletal/Ortopedi	10
Gangguan Saraf/Neurologis	11
Gangguan Mata	12
Gangguan Kulit & Alergi	13
Gangguan Endokrin/Metabolik	14
Bedah Minor/Prosedural	15
Penyakit Lambung & Refluks	16
Kondisi Akut	17
Penyakit Sistemik	18
Cedera Ringan	19
Penyakit Limfatik/Infeksi Kelenjar Getah Bening	20
Tumor Jinak	21
Penyakit THT	22
Penyakit Gigi & Mulut	23
Reumatologi/Autoimun	24
Lainnya	25

3. HASIL DAN PEMBAHASAN

3.1 *Modeling*

Metode *K-Means clustering* menggambarkan proses pengelompokan titik-titik informasi ke dalam 2 kelompok ataupun lebih sehingga titik-titik informasi yang tercantum di dalam kelompok yang sama lebih mirip satu sama lain [6]. Pengelompokkan dengan k-means dinilai lebih efektif untuk penemuan pola dalam data medis dikarenakan kebutuhan untuk membuat perhitungan berulang pada dataset lengkap disetiap siklus sebelum mencapai hasil yang berkualitas [15]. Ada beberapa tahap yang harus dilakukan dalam melakukan *cluster* antara lain [9] :

1. Memilih jumlah *cluster* K yang optimal.
   Dengan menggunakan metode optimasi *cluster*, metode *elbow* yang merupakan sebuah teknik visual untuk menentukan jumlah *cluster* (K) yang membentuk siku pada suatu titik. Nilai K-Means yang didasari oleh total jarak kuadrat antara setiap titik data dengan *centroid cluster* dikarenakan SSE

mengukur seberapa jauh data berada dari pusat *cluster* maka, dilanjutkan dengan hitungan SSE untuk masing-masing K dengan rumus seperti berikut [7].

$$\text{SSE} = \sum_{i=1}^{k} \sum_{xi \in Ci} D(x_i, C_i)^2$$

Rumus SSE untuk cluster = 2

$$SSE_{K=2} = \sum_{i=1}^{200} \|x_i - \mu_0\|^2 + \sum_{j=1}^{164} \|x_j - \mu_1\|^2 = 8607$$

Rumus SSE untuk cluster = 3

$$SSE_{K=3} = \sum_{i=1}^{110} \|x_i - \mu_0\|^2 + \sum_{j=1}^{130} \|x_j - \mu_1\|^2 + \sum_{k=1}^{124} \|x_k - \mu_2\|^2 = 7210$$

Rumus SSE untuk cluster = 4

$$SSE_{K=4} = \sum_{i=1}^{80} \|x_i - \mu_0\|^2 + \sum_{j=1}^{90} \|x_j - \mu_1\|^2 + \sum_{k=1}^{100} \|x_k - \mu_2\|^2 + \sum_{l=1}^{94} \|x_l - \mu_3\|^2 = 6057$$

Rumus SSE untuk cluster = 5

$$SSE_{K=5} = \sum_{i=1}^{60} \|x_i - \mu_0\|^2 + \sum_{j=1}^{70} \|x_j - \mu_1\|^2 + \sum_{k=1}^{75} \|x_k - \mu_2\|^2 + \sum_{l=1}^{80} \|x_l - \mu_3\|^2 + \sum_{m=1}^{79} \|x_m - \mu_4\|^2 = 4306$$

Rumus SSE untuk cluster = 6

$$SSE_{K=6} = \sum_{i=1}^{45} \|x_i - \mu_0\|^2 + \sum_{j=1}^{55} \|x_j - \mu_1\|^2 + \sum_{k=1}^{50} \|x_k - \mu_2\|^2 + \sum_{l=1}^{60} \|x_l - \mu_3\|^2 + \sum_{m=1}^{58} \|x_m - \mu_4\|^2 + \sum_{n=1}^{58} \|x_n - \mu_5\|^2 = 3714$$

Rumus SSE untuk cluster = 7

$$SSE_{K=7} = \sum_{i=1}^{40} \|x_i - \mu_0\|^2 + \sum_{j=1}^{45} \|x_j - \mu_1\|^2 + \sum_{k=1}^{45} \|x_k - \mu_2\|^2 + \sum_{l=1}^{50} \|x_l - \mu_3\|^2 + \sum_{m=1}^{45} \|x_m - \mu_4\|^2 + \sum_{n=1}^{45} \|x_n - \mu_5\|^2 + \sum_{0=1}^{44} \|x_o - \mu_6\|^2 = 3189$$

Rumus SSE untuk cluster = 8

$$SSE_{K=8} = \sum_{i=1}^{35} \|x_i - \mu_0\|^2 + \sum_{j=1}^{40} \|x_j - \mu_1\|^2 + \sum_{k=1}^{40} \|x_k - \mu_2\|^2 + \sum_{l=1}^{40} \|x_l - \mu_3\|^2 + \sum_{m=1}^{42} \|x_m - \mu_4\|^2 + \sum_{n=1}^{40} \|x_n - \mu_5\|^2 + \sum_{0=1}^{42} \|x_o - \mu_6\|^2 + \sum_{p=1}^{41} \|x_p - \mu_7\|^2 = 2799$$

Rumus SSE untuk cluster = 9

$$SSE_{K=9} = \sum_{i=1}^{30} \|x_i - \mu_0\|^2 + \sum_{j=1}^{35} \|x_j - \mu_1\|^2 + \sum_{k=1}^{35} \|x_k - \mu_2\|^2 + \sum_{l=1}^{35} \|x_l - \mu_3\|^2 + \sum_{m=1}^{35} \|x_m - \mu_4\|^2 + \sum_{n=1}^{35} \|x_n - \mu_5\|^2 + \sum_{0=1}^{35} \|x_o - \mu_6\|^2 + \sum_{p=1}^{35} \|x_p - \mu_7\|^2 + \sum_{q=1}^{32} \|x_q - \mu_8\|^2 = 2547$$

Rumus SSE untuk cluster = 10

$$SSE_{K=10} = \sum_{i=1}^{28} \|x_i - \mu_0\|^2 + \sum_{j=1}^{30} \|x_j - \mu_1\|^2 + \sum_{k=1}^{30} \|x_k - \mu_2\|^2 + \sum_{l=1}^{30} \|x_l - \mu_3\|^2 + \sum_{m=1}^{30} \|x_m - \mu_4\|^2 + \sum_{n=1}^{30} \|x_n - \mu_5\|^2 + \sum_{0=1}^{30} \|x_o - \mu_6\|^2 + \sum_{p=1}^{30} \|x_p - \mu_7\|^2 + \sum_{q=1}^{30} \|x_q - \mu_8\|^2 + \sum_{r=1}^{28} \|x_r - \mu_9\|^2 = 2451$$

Table 7 Tabel Hasil SSE

No.	Jumlah Cluster (K)	SSE (Inertia)
1.	2	8607
2.	3	7210
3.	4	6057
4.	5	4306
5.	6	3714
6.	**7**	**3189**
7.	8	2799
8.	9	2547
9.	10	2451

Pada cluster 2 sampai 5 nilai SSE turun tapi masih tergolong tinggi dan silhouette score belum maksimum dan ada kemungkinan beberapa pola data masih sangat menyebar dan berkemungkinan beberapa pola data bercampur, pada cluster 6 nilai SSE hampir mendekati nilai optimal namun cenderung kurang menyebar. Pada cluster 7 pola menyebar mulai optimal, sedangkan pada cluster 8 sampai 10 perhitungan SSE mengalami penurunan yang kurang signifikan sedangkan perolehan nilai k yang terlalu kecil menyebabkan pola penyebaran terlalu berjauhan dengan titik awal.

Perolehan nilai dari tersebut menujukkan banyaknya jumlah anggota yang didapatkan dari masing-masing *cluster*, semakin besar K, SEE akan semakin kecil dikarenakan lebih banyak centroid sehingga jarak rata-rata semakin pendek, data yang tersebar luas menghasilkan SSE besar, serta *centroid* yang optimal menghasilkan SSE minimal sehingga total jarak kuadrat semua titik data ke *centroid* nya menjadi paling kecil.Penurunan SSE mulai terlihat pada titik K tertentu dengan ditandai perolehan perhitungan cluster yang terjadi berikutnya tidak mengalami peningkatan nilai yang signifikan, pada tabel 7 tersebut menunjukkan K = 7 merupakan jumlah paling optimal dengan kecenderungan pola lebih menyebar dibanding nilai cluster lainnya.

2. Menentukan jumlah *cluster* dengan memilih K titik pusat/*centroid* awal secara acak dari dataset. Berdasarkan data yang dipilih didapatkan titik pusat *cluster* yang akan digunakan sebagai titik awal proses pengelompokkan *cluster* dengan disesuaikan berdasarkan k optimal.

Table 8 Titik Pusat Cluster

Atribut	C1	C2	C3	C4	C5	C6	C7
Baris ke-	226	1807	739	1273	381	215	626
Jenis_kelamin	2	1	2	1	2	2	2
Status_kawin	4	4	1	1	1	4	1
Usia	9	19	51	49	48	15	48
Kategori_pekerjaan	8	1	5	4	4	1	5
Kategori_alergi	6	6	6	6	1	6	6
Kategori_diagnosis	5	6	6	5	5	6	22

Pemilihan titik awal dilakukan dengan memilih satu *centroid* awal acak lalu memilih berikutnya yang jauh dari sebelumnya untuk meminimalkan adanya kesamaan antar cluster.

3. Penghitungan jarak *cluster* dengan *centroid* terdekat dengan menggunakan perhitungan jarak *Euclidean Distance* yang dirumuskan sebagai berikut :

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

Dengan contoh perhitungan jarak data 1 ke pusat cluster C1 sebagai berikut:

a. Hitung jarak data 1 ke pusat cluster C1:

$$d(P1, C1) = \sqrt{(2-2)^2 + (4-4)^2 + (33-9)^2 + (4-8)^2 + (6-6)^2 + (2-5)^2}$$

$$= \sqrt{0 + 0 + 576 + 16 + 0 + 9} = \sqrt{601} \approx 24{,}52$$

b. Hitung jarak data 1 ke pusat cluster C2:

$$d(P1, C2) = \sqrt{(2-1)^2 + (4-4)^2 + (33-19)^2 + (4-1)^2 + (6-6)^2 + (2-6)^2}$$
$$= \sqrt{1+0+196+9+0+16} = \sqrt{222} \approx 14{,}8$$

c. Hitung jarak data 1 ke pusat cluster C3:

$$d(P1, C3) = \sqrt{(2-2)^2 + (4-1)^2 + (33-51)^2 + (4-5)^2 + (6-6)^2 + (2-6)^2}$$
$$= \sqrt{0+9+324+1+0+16} = \sqrt{350} \approx 18{,}7$$

d. Hitung jarak data 1 ke pusat cluster C4:

$$d(P1, C4) = \sqrt{(2-1)^2 + (4-1)^2 + (33-49)^2 + (4-4)^2 + (6-6)^2 + (2-5)^2}$$
$$= \sqrt{1+9+256+0+0+9} = \sqrt{275} \approx 16{,}6$$

e. Hitung jarak data 1 ke pusat cluster C5:

$$d(P1, C5) = \sqrt{(2-2)^2 + (4-1)^2 + (33-48)^2 + (4-4)^2 + (6-1)^2 + (2-5)^2}$$
$$= \sqrt{0+9+225+0+25+9} = \sqrt{268} \approx 16{,}4$$

f. Hitung jarak data 1 ke pusat cluster C6:

$$d(P1, C6) = \sqrt{(2-2)^2 + (4-4)^2 + (33-15)^2 + (4-1)^2 + (6-6)^2 + (2-6)^2}$$
$$= \sqrt{0+0+324+9+0+16} = \sqrt{349} \approx 18{,}7$$

g. Hitung jarak data 1 ke pusat cluster C7:

$$d(P1, C7) = \sqrt{(2-2)^2 + (4-1)^2 + (33-48)^2 + (4-5)^2 + (6-6)^2 + (2-22)^2}$$
$$= \sqrt{0+9+225+1+0+400} = \sqrt{635} \approx 25{,}2$$

Dilanjutkan ke seluruh data 2, 3, 4,…, n. Dari data tersebut maka diperoleh pula jarak masing-masing data ke pusat *cluster* yang tergolong.

Table 9 Perolehan Jarak Terdekat

No.	Jarak Terdekat	Cluster
1	1,36	6
2	1,41	6
3	0,80	3
4	0,79	3
5	1,74	3
6	1,03	1
7	1,31	3
8	1,09	3
9	0,59	6
10	0,56	3
…	…	…
1986	0,7	4
1987	1,13	4
1988	2,23	7
1989	1,43	7
1990	1,05	6
1991	1,01	4

Semakin dekat jarak diperoleh maka semakin dekat data tersebut ke pusat *cluster*.

4. Pengelompokkan jarak yang terdekat dengan data antara pusat *cluster*.

Setiap data ditempatkan ke cluster dengan pusat (centroid) terdekat, berdasarkan ukuran jarak tertentu. Data yang memiliki jarak yang berdekatan akan digabungkan dalam satu kelompok pusat cluster.

Table 10 Pengelompokkan Jarak Iterasi ke-1

No.	**C1**	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**	**Cluster**
P1	24,51	14,89	18,70	16,58	16,37	18,68	25,19	2
P2	22,71	13,37	20,85	18,70	18,41	17,02	27,20	2
P3	42,40	32,78	5	4,69	7,14	36,68	21,21	4
P4	44,38	34,74	5,38	5,83	8,18	38,65	21,58	3
P5	22,58	13,67	20,71	18,60	18,30	17,26	27,09	2
P6	5	11,13	39,54	37,56	36,90	9,11	41,89	1
P7	19,69	11,09	23,34	21,26	20,85	14,49	28,28	2
P8	64,17	54,32	22,20	24,12	25,59	58,29	31,40	3
P9	15,58	5,09	29,69	27,80	27,27	8,06	29,06	2
P10	36,35	26,36	6,08	4,24	5,91	30,29	16,30	4
…	…	…	…	…	…	…	…	…
P1986	31,81	21,79	11,78	10,29	10,72	25,69	14,49	4
P1987	29,56	19,77	15,71	14,42	14,52	23,49	14,28	7
P1988	39,77	30,11	14,31	14,62	15,32	33,76	4,89	7
P1989	48,60	39,17	19,15	20,39	21,23	42,70	6	7

P1990	17,46	7,74	25,21	23,21	22,75	11,44	27,98	2
P1991	52,17	42,41	10,14	12,32	14,24	46,38	19,89	3

Setelah didapatkan pengelompokkan iterasi tersebut maka dilakukan perhitungan pusat *cluster* sampai perhitungan pusat *cluster* baru, proses dilanjutkan dari iterasi ke-1 dengan menggunakan langkah yang serupa hingga tidak ada perubahan data yang diperoleh dalam tahapan iterasi *cluster* berikutnya.

Table 11 Pengelompokkan Jarak Iterasi ke-4

No.	C1	C2	C3	C4	C5	C6	C7	Cluster
P1	24,51	14,89	18,70	16,58	16,37	18,68	25,19	2
P2	22,71	13,37	20,85	18,70	18,41	17,02	27,20	2
P3	42,40	32,78	5	4,69	7,14	36,68	21,21	4
P4	44,38	34,74	5,38	5,83	8,18	38,65	21,58	3
P5	22,58	13,67	20,71	18,60	18,30	17,26	27,09	2
P6	5	11,13	39,54	37,56	36,90	9,11	41,89	1
P7	19,69	11,09	23,34	21,26	20,85	14,49	28,28	2
P8	64,17	54,32	22,20	24,12	25,59	58,29	31,40	3
P9	15,58	5,09	29,69	27,80	27,27	8,06	29,06	2
P10	36,35	26,36	6,08	4,24	5,91	30,29	16,30	4
…	…	…	…	…	…	…	…	…
P1986	31,81	21,79	11,78	10,29	10,72	25,69	14,49	4
P1987	29,56	19,77	15,71	14,42	14,52	23,49	14,28	7
P1988	39,77	30,11	14,31	14,62	15,32	33,76	4,89	7
P1989	48,60	39,17	19,15	20,39	21,23	42,70	6	7
P1990	17,46	7,74	25,21	23,21	22,75	11,44	27,98	2
P1991	52,17	42,41	10,14	12,32	14,24	46,38	19,89	3

Perhitungan yang dihasilkan dari proses iterasi sebelumnya sampai pada iterasi ke-4 mendapatkan hasil yang serupa, proses perhitungan iterasi dihentikan sampai ke iterasi ke-4.

5. Perhitungan pusat *cluster* baru berdasarkan data *cluster* yang telah ditemukan sebelumnya dengan menggunakan rumus sebagai berikut:

$$C(i) = \frac{x_1 + x_2 + x_{\dots} + x_{\dots}}{\sum x} \dots (2)$$

Dengan perhitungan yang dihasilkan diperlihatkan pada tabel berikut:

Table 12 Pusat *Cluster* Baru

Atribut	C1	C2	C3	C4	C5	C6	C7
Jenis_kelamin	0,36	-1,16	0,85	-1,16	-0,39	0,85	0,1
Status_kawin	1,1	1,12	-0,74	-0,9	-0,5	1,12	-0,62
Usia	-1,23	-0,96	0,77	0,6	0,3	-0,7	0,5
Kategori_pekerjaan	1,7	-1,2	0,3	0,07	0,09	-1,0	0,26
Kategori_alergi	0,2	0,19	0,17	0,17	-5,10	0,2	0,17
Kategori_diagnosis	-0,13	-0,05	-0,25	-0,29	0,13	-0,14	2,004

Dari data tersebut didapatkan data yang paling mendekati pusat *cluster* berdasarkan iterasi yang dilakukan dari keseluruhan data sampai tidak terjadinya perubahan data.

3.2 *Evaluation*

Pada proses dari data iterasi yang didapatkan sebelumnya, karakteristik atribut yang dihasilkan dari masing-masing *cluster* tersebut menunjukkan.

a. *Cluster* 1 (C1): Didominasi oleh pasien perempuan dengan status belum kawin, berusia kisaran 9 tahun dengan belum adanya pekerjaan dan tidak ada riwayat alergi yang diderita. Dengan kecenderungan diagnosis yang paling dominan merupakan kategori gangguan pernapasan.
b. *Cluster* 2 (C2): Didominasi oleh pasien laki-laki dengan status belum kawin, berusia kisaran 19 tahun dengan pekerjaan dalam bidang pendidikan dan tidak ada riwayat alergi yang diderita. Dengan kecenderungan diagnosis yang paling dominan merupakan kategori penyakit gangguan kardiovaskular.
c. *Cluster* 3 (C3): Didominasi oleh pasien perempuan dengan status kawin, berusia kisaran 51 tahun dengan pekerjaan dalam bidang rumah tangga dan tidak ada riwayat alergi yang diderita. Dengan kecenderungan diagnosis yang paling dominan merupakan kategori penyakit gangguan kardiovaskular.
d. *Cluster* 4 (C4): Didominasi oleh pasien laki-laki dengan status kawin, berusia kisaran 49 tahun dengan pekerjaan dalam bidang kantoran/profesional dan tidak ada riwayat alergi yang diderita. Dengan kecenderungan diagnosis yang paling dominan merupakan kategori penyakit gangguan pernapasan.
e. *Cluster* 5 (C5): Didominasi oleh pasien perempuan dengan status kawin, berusia kisaran 48 tahun dengan pekerjaan dalam bidang kantoran/profesional dan riwayat alergi obat yang diderita.

Dengan kecenderungan diagnosis yang paling dominan merupakan kategori penyakit gangguan pernapasan.

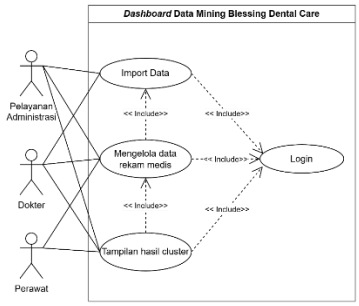
f. *Cluster* 6 (C6): Didominasi oleh pasien perempuan dengan status kawin, berusia kisaran 15 tahun dengan pekerjaan dalam bidang pendidikan dan tidak ada riwayat alergi yang diderita. Dengan kecenderungan diagnosis yang paling dominan merupakan kategori penyakit gangguan kardiovaskular.

g. *Cluster* 7 (C7): Didominasi oleh pasien perempuan dengan status kawin, berusia kisaran 48 tahun dengan pekerjaan dalam bidang rumah tangga dan tidak ada riwayat alergi yang diderita. Dengan kecenderungan diagnosis yang paling dominan merupakan kategori penyakit THT.

3.3 *Deployment*

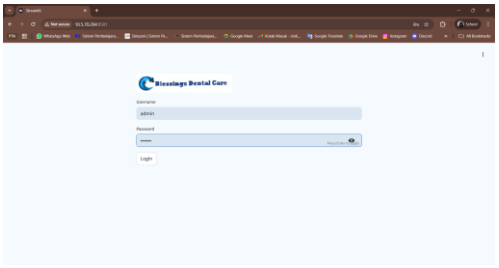
Tahapan pengembangan sistem untuk menerapkan model yang digunakan agar dapat diterapkan pada organisasi untuk mendukung analisa tren penyakit, dengan menggunakan metodologi *waterfall* untuk pendekatan yang lebih sistematis. Adapun tahapan yang dilakukan sebagai berikut [16]:

1. Analisis Kebutuhan (*Requirements*)
   Bentuk analisis digambarkan melalui diagram *use case* yang berfungsi untuk mendefinisikan proses yang akan dilakukan oleh sistem dan komponen-komponennya yang didapatkan berdasarkan informasi yang telah dikumpulkan diperoleh batasan sistem [17].



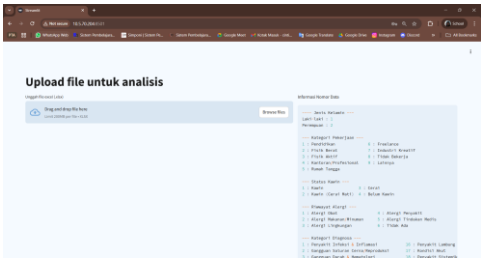
Gambar 1 Diagram *Use Case*

2. Desain (*Design*)
   Rancangan tampilan penggunaan sistem digambarkan melalui kardinalitas antar himpunan entitas pada sistem.
3. Implementasi (*Implementation*)
   Pengembangan bentuk rancangan antarmuka ke dalam bentuk program menggunakan kode program *python*.
   a. Tampilan Login
      Halaman awal sebelum *user* mengakses halaman *dashboard*, memasukkan *username* dan *password* untuk memastikan aplikasi diakses oleh pengguna yang terdaftar atau pengguna yang diberikan akses.



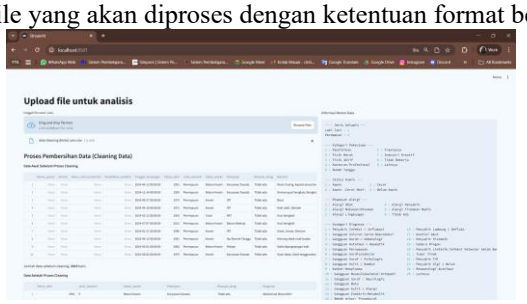
Gambar 2 Tampilan *Login*

   b. Tampilan *Dashboard* Analisis



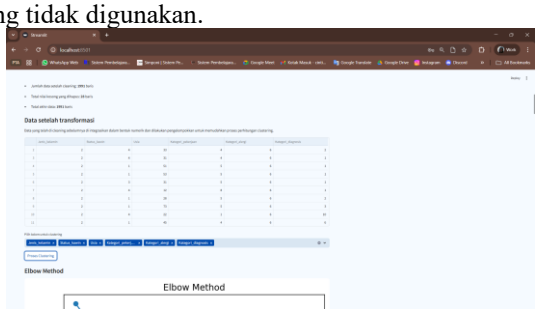
Gambar 3 Tampilan *Upload File*

Memasukkan file yang akan diproses dengan ketentuan format berupa file *excel*.



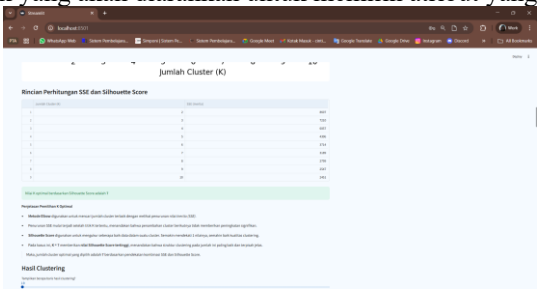
Gambar 4 Tampilan Data *Cleaning*

Data yang dipilih akan dilakukan pembersihan data dengan menghilangkan nilai *null* dan atribut yang tidak digunakan.



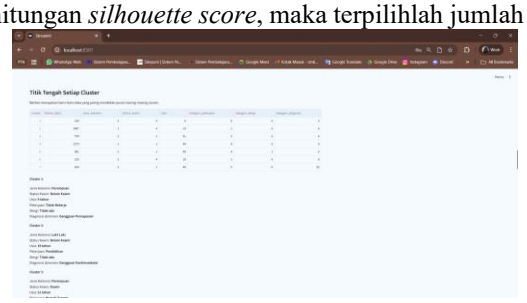
Gambar 5 Tampilan Data Transformasi

Kemudian dari proses *cleaning* data setelah cleaning dilanjutkan ke proses transformasi ke tipe numerik yang akan diarahkan untuk memilih atribut yang di *cluster*.



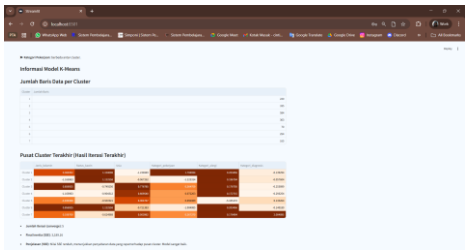
Gambar 6 Tampilan Pemilihan K Optimal

Dilakukan perhitungan untuk pemilihan nilai K optimal dengan menggunakan metode *elbow* dan perhitungan *silhouette score*, maka terpilihlah jumlah *cluster* yang digunakan.

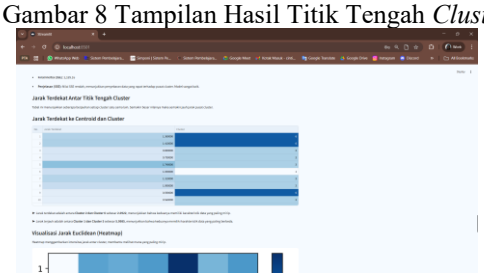


Gambar 7 Tampilan Perolehan Titik Tengah *Cluster*

Saat dilakukan proses *clustering* diperoleh analisis perolehan titik tengah *cluster* berserta dengan deskripsi yang didapatkan berdasarkan informasi penomoran data.

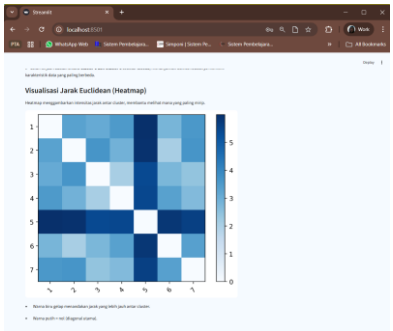


Gambar 8 Tampilan Hasil Titik Tengah *Cluster*



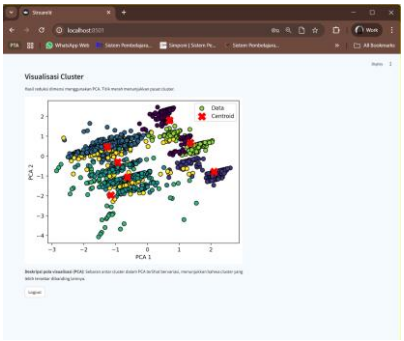
Gambar 9 Tampilan Jarak Terdekat *Centroid* ke *Cluster*

Jumlah *cluster* yang dipilih dengan memilih titik pusat awal dari dataset dan memperhitungkan jarak *cluster* dengan *centroid* terdekat dan visualisasi jarak *euclidean* (*Heatmap*).



Gambar 10 Tampilan Visualisasi Jarak *Euclidean* (*Heatmap*)

Bentuk visualisasi dilanjutkan dengan hasil reduksi menggunakan PCA dan deskripsi dari sebaran pola visualisasi.



Gambar 11 Tampilan Visualisasi PCA

4. Verifikasi (*Verification*)
   Program yang telah dibuat akan melalui proses pengujian kelayakan memastikan sistem yang telah dibuat berjalan tanpa *error* dengan dilakukan uji coba.

Table 13 Uji Coba *Dashboard*

No.	Kasus	Hasil yang Diharapkan	Hasil Uji Coba

1	Memasukkan *username* dan *password* yang telah terdaftar, lalu klik "*Login*"	Berhasil melakukan proses *login*, dilanjutkan ke halaman *Dashboard*.	Berhasil
2	Memasukkan *username* dan *password* yang belum terdaftar, lalu klik "*Login*"	Pengguna gagal melakukan *login* dan tetap di halaman login, sistem menampilkan pesan kesalahan "*username* dan *password* salah"	Berhasil
3	Masukkan file, klik "*browser file*" atau *drag* file yang akan di analisis dengan type ".xlsx".	File berhasil di tambahkan dan langsung dilakukan proses cleaning dilanjutkan ke proses transformasi kemudian penentuan jumlah *cluster*.	Berhasil
4	Klik "Proses C*lustering*".	File dilanjutkan ke tahap analisis berdasarkan algoritma *k-means clustering*. Analisis yang dihasilkan meliputi pemilihan K optimal dengan S, perolehan titik tengah setiap *cluster*, perolehan jarak terdekat *centroid* ke *cluster,* Informasi model *k-means*, Perolehan jarak euclidean, Visualisasi *heatmap*, Visualisasi PCA. Setiap perolehan terdapat penjelasan.	Berhasil

5. Pemeliharaan (*Maintenance*)
   Akan dilakukan pengecekan sistem berkala untuk memastikan sistem terupdate agar tidak terjadi *error* diluar uji coba sebelumnya.

4. KESIMPULAN

Hasil dari penerapan algoritma k-means clustering yang dihasilkan berdasarkan dari pemrosesan pengelompokkan data rekam medis pasien dari bentuk 7 cluster yang didapatkan sebanyak 220 data pasien pada cluster 1 dengan tren diagnosis dominan merupakan kategori penyakit gangguan pernapasan, 335 data pasien pada cluster 2 dengan diagnosis dominan merupakan kategori penyakit gangguan kardiovaskular, 584 data pasien pada cluster 3 dengan diagnosis dominan merupakan kategori penyakit gangguan kardiovaskular, 363 data pasien pada cluster 4 dengan diagnosis dominan merupakan kategori penyakit gangguan pernapasan, 70 data pasien pada cluster 5 dengan diagnosis dominan merupakan kategori penyakit gangguan pernapasan, 254 data pasien pada cluster 6 dengan diagnosis dominan merupakan kategori penyakit kardiovaskular, 165 data pasien pada cluster 7 dengan diagnosis dominan merupakan kategori penyakit THT dari total 1.991 data yang sudah di cleaning.

Tampilan hasil dari proses penerapan yang dihasilkan dari dashboard yang telah dibuat terdapat proses transformasi dari data rekam medis yang sudah tidak terdapat nilai null. Proses transformasi yang dilakukan merupakan hasil pengelompokkan data menjadi beberapa kategori, dan pemilihan nilai k optimal dengan menggunakan metode elbow beserta perhitungan silhouette score dilanjutkan dengan proses penentuan titik tengah per cluster secara acak, informasi model k-means yang dihasilkan seperti jumlah baris data yang didapatkan per cluster dan jumlah iterasi yang dihasilkan, jarak euclidean titik tengah cluster, serta bentuk visualisasi dalam bentuk heatmap dan PCA. Setiap hasil yang ditampilkan disertakan dengan deskripsi, dimana hasil ini juga menunjukkan penjelasan mengenai karakteristik data yang paling mirip antar cluster.

REFERENSI

[1] H. Dilawati, H. Widianto, dan A. Kuswiadji, "Klasterisasi Data Rekam Medis Pasien Menggunakan Metode K-Means Clustering Di Rumah Sakit Widodo Ngawi," *BIOS J. Teknol. Inf. dan Rekayasa Komput.*, vol. 5, no. 2, hal. 5–8, 2024, [Daring]. Tersedia pada: https://bios.sinergis.org/bios/article/view/134

[2] L. 'Izzah dan A. Jananto, "Penerapan Algoritma K-Means Clustering Untuk Perencanaan Kebutuhan Obat Di Klinik Citra Medika," *J. Ilm. Komput.*, vol. 18, no. 1, hal. 69, 2022, doi: 10.35889/progresif.v18i1.769.

[3] Nadhila, Marsono, dan J. Halim, "Penerapan Data Mining Untuk Pengelompokan Penyakit Yang Sering Terjadi Pada Pasien RSUD (Rumah Sakit Umum Daerah) Kota Langsa Menggunakan Metode K-Means Clustering," *J. Cybertech*, no. September, hal. 1–12, 2020, [Daring]. Tersedia pada: www.trigunadharma.ac.id

[4] E. A. Herdiaman, A. Sudiarjo, dan M. Hikmatyar, "KLASTERISASI PASIEN PADA RSUD CIAMIS MENGGUNAKAN METODE K-MEANS," *JITET (Jurnal Inform. dan Tek. Elektro Ter.*, vol. 12, no. 3, 2024, [Daring]. Tersedia pada: https://journal.eng.unila.ac.id/index.php/jitet/article/view/5124/2092

[5] A. Ali, "Klasterisasi Data Rekam Medis Pasien Menggunakan Metode K-Means Clustering di Rumah Sakit Anwar Medika Balong Bendo Sidoarjo," *MATRIK J. Manajemen, Tek. Inform. dan Rekayasa Komput.*, vol. 19, no. 1, hal. 186–195, 2019, doi: 10.30812/matrik.v19i1.529.

[6] W. Purba, G. A. Sembiring, A. Saputra, M. T. Turnip, dan B. J. I. Manihuruk, "PENERAPAN DATA MINING UNTUK

PENGELOLAAN DATA REKAM MEDIS MENGGUNAKAN METODE K-MEANS CLUSTERING PADA RUMAH SAKIT ROYAL PRIMA MEDAN," *J. TEKINKOM*, vol. 6, no. 1, hal. 158–168, 2023, doi: 10.37600/tekinkom.v6i1.857.

[7] O. J. Harmaja, H. Halawa, W. S. Hulu, dan S. Loi, "Implementasi Algoritma K-Means Clustering Untuk Pengelompokkan Penyakit Pasien Pada Puskesmas Pulo Brayan," *Sains dan Teknol.*, vol. 5, no. 1, hal. 150–157, 2023, doi: https://doi.org/10.55338/saintek.v5i1.1306.

[8] C. A. Sugianto, A. H. Rahayu, dan A. Gusman, "Algoritma K-Means untuk Pengelompokkan Penyakit Pasien pada Puskesmas Cigugur Tengah," *Jt. (Journal Inf. Technol.*, vol. 2, no. 2, hal. 39–44, 2020, doi: 10.47292/joint.v2i2.30.

[9] F. Kurnia, I. Fahmi, E. Wahyudi, dan G. E. S. Mige, "Penerapan Algoritma K-Means Untuk Pengelompokan Diagnosa Penyakit Mata Berdasarkan Rentang Usia," *J. SPEKTRO*, vol. 2, no. 1, hal. 10–17, 2019, [Daring]. Tersedia pada: https://ejurnal.undana.ac.id/index.php/spektro/article/view/1373/1092

[10] K. Rahayu, L. Novianti, dan M. Kusnandar, "Implementation Data Mining With K-Means Algorithm For Clustering Distribution Rabies Case Area In Palembang City," *J. Phys. Conf. Ser.*, vol. 1500, no. 1, 2020, doi: 10.1088/1742-6596/1500/1/012121.

[11] A. E. Clarke, S. J. Elliott, Y. St. Pierre, L. Soller, S. La Vieille, dan M. Ben-Shoshan, "Demographic characteristics associated with food allergy in a Nationwide Canadian Study," *Allergy, Asthma Clin. Immunol.*, vol. 17, no. 1, hal. 1–7, 2021, doi: 10.1186/s13223-021-00572-z.

[12] Widiastuti, "Klasifikasi jenis pekerjaan kantor yang di lakukan mahasiswa pada praktik kerja lapangan," *J. Pendidik. Manaj. PERKANTORAN*, vol. 5, no. 1, hal. 109–117, 2020, doi: 10.17509/jpm.v4i2.18008.

[13] I. L. Organization, "The International Standard Classification of Occupations." [Daring]. Tersedia pada: https://isco-ilo.netlify.app/en/isco-08/

[14] W. H. Organization, "International Statistical Classification of Diseases and Related Health Problems 10th Revision." [Daring]. Tersedia pada: https://icd.who.int/browse10/2010/en

[15] F. Cirett-Galán, R. T. Peralta, dan O. F. G. Mora, "K-Means Cluster Analysis to Support Diabetic Patient Care," *Res. Sq.*, 2023.

[16] A. A. Wahid, "Analisis Metode Waterfall Untuk Pengembangan Sistem Informasi," *J. Ilmu-ilmu Inform. dan Manaj. STMIK*, vol. 1, 2020, [Daring]. Tersedia pada: https://www.researchgate.net/profile/Aceng-Wahid/publication/346397070_Analisis_Metode_Waterfall_Untuk_Pengembangan_Sistem_Informasi/links/5fbfa91092851c933f5d76b6/Analisis-Metode-Waterfall-Untuk-Pengembangan-Sistem-Informasi.pdf

[17] L. Setiyani, "Desain Sistem : Use Case Diagram Pendahuluan," *LPPM STIMK ROSMA/ Pros. Semin. Nas. Inov. Adopsi Teknol.*, hal. 246–260, 2021, [Daring]. Tersedia pada: https://journal.uii.ac.id/AUTOMATA/article/view/19517