Jurnal Riset Multidisiplin Edukasi
https://journal.hasbaedukasi.co.id/index.php/jurmie
Halaman : 1051-1059

Clustering Data Diabetes Menggunakan Algoritma K-Means

Kresna Bayu Prasetyo[1], Dewi Oktafiani[2]
Stmik Amikom Surakarta Sukoharjo Indonesia [1,2]
Email: kresna.10380@mhs.amikomsolo.ac.id[1], dewioktafiani@dosen.amikomsolo.ac.id[2]

Informasi

Volume : 3
Nomor : 1
Bulan : Januari
Tahun : 2026
E-ISSN : 3062-9624

Abstract

Diabetes is one of the chronic diseases whose prevalence continues to increase globally, including in Indonesia. This study employed the K-Means Clustering method with k = 3 to group diabetes patients based on their risk factors. The analysis identified three main groups: young patients in good health, obese patients with high risk, and elderly patients with moderate risk. The findings indicate a strong correlation between glucose levels, body mass index (BMI), age, and the risk level of diabetes. Comparison with the outcome data shows that the group consisting of obese patients with high glucose levels has a greater likelihood of developing diabetes.

Keyword: *Algoritma, K-Means, Diabetes*

Abstrak

Diabetes merupakan salah satu penyakit kronis yang prevalensinya terus meningkat secara global termasuk di Indonesia. Penelitian ini menggunakan metode K-Means Clustering dengan k = 3 untuk mengelompokkan pasien diabetes berdasarkan faktor risiko yang dimiliki. Dari hasil analisis, terbentuk tiga kelompok utama, yaitu pasien muda dengan kondisi sehat, pasien obesitas dengan risiko tinggi, dan pasien usia tua dengan risiko sedang. Analisis menunjukkan adanya hubungan yang cukup kuat antara kadar glukosa, indeks massa tubuh (BMI), usia, dan tingkat risiko diabetes. Perbandingan dengan data outcome memperlihatkan bahwa kelompok obesitas dengan kadar glukosa tinggi memiliki peluang lebih besar untuk mengembangkan diabetes.

Kata Kunci: *Algoritma, K-Means, Diabetes.*

A. PENDAHULUAN

Diabetes adalah penyakit gangguan metabolik yang ditandai oleh kenaikan glukosa darah akibat penurunan sekresi insulin oleh sel beta pankreas dan atau resistensi insulin [1] yang merupakan salah satu penyakit kronis yang prevalensinya terus meningkat secara global termasuk di Indonesia. Penyakit ini ditandai dengan tingginya kadar glukosa dalam darah akibat gangguan pada sekresi atau fungsi insulin. Faktor risiko diabetes meliputi usia, jenis kelamin, indeks massa tubuh (IMT), tekanan darah, dan kadar gula darah sewaktu. Analisis terhadap data pasien diabetes menjadi penting untuk memahami pola dan karakteristik penyakit ini, sehingga dapat membantu dalam pengambilan keputusan klinis dan pengembangan strategi pencegahan serta penanganan yang lebih efektif [2].

Salah satu metode yang efektif untuk menganalisis data pasien adalah dengan menggunakan teknik clustering. Clustering adalah metode penting dalam analisis data yang digunakan untuk mengelompokkan berbagai objek ke dalam cluster atau kelompok tertentu berdasarkan karakteristik yang serupa. Pada proses clustering, objek-objek di dalam satu cluster memiliki kemiripan yang lebih tinggi dibandingkan dengan objek-objek yang berada di cluster lain [3] . Algoritma K-Means merupakan salah satu metode clustering yang paling umum digunakan karena kesederhanaan dan efisiensinya dalam mengelompokkan data numerik. Algoritma ini bekerja dengan membagi data ke dalam beberapa klaster berdasarkan jarak terdekat antara data dan titik pusat (centroid) yang ditentukan secara interaktif [4].

Dalam konteks data pasien diabetes, penerapan algoritma K-Means dapat membantu mengidentifikasi kelompok pasien dengan karakteristik serupa, seperti tingkat risiko komplikasi atau respons terhadap pengobatan tertentu. Sebagai contoh, penelitian yang dilakukan di Puskesmas Lappae [5] menggunakan metode K-Means untuk mengelompokkan data pasien berdasarkan variabel usia, jenis kelamin, tekanan darah, kadar gula darah, dan IMT. Hasilnya menunjukkan dua klaster utama: kelompok berisiko tinggi dengan jumlah 584 pasien dan kelompok berisiko rendah dengan 456 pasien. Analisis ini membantu dalam memahami distribusi faktor risiko di antara pasien dan dapat menjadi dasar untuk intervensi medis yang lebih tepat sasaran.

Dengan demikian, penerapan algoritma K-Means dalam pengelompokan data diabetes tidak hanya memberikan wawasan mendalam mengenai pola dan karakteristik pasien, tetapii juga berpotensi meningkatkan kualitas layanan kesehatan melalui strategi penanganan yang lebih terfokus dan efisien. K-Means terbukti lebih efisien dan menjadi pilihan yang tepat dalam menangani data berukuran besar. Oleh karena itu, pemilihan algoritma clustering yang tepat harus disesuaikan dengan karakteristik data dan tujuan analisis yang ingin dicapai.

TINJAUAN PUSTAKA

Beberapa penelitian sebelumnya mengenai deteksi penyakit diabetes dengan KMeans clustering oleh Praja et al. [6] Menunjukkan bahwa hasil pengujian metode KMeans adalah 73,438 %. Penelitian lainnya yang dilakukan oleh Gestavito et al. [7] Mengenai pengelompokan tingkat risiko penyakit diabetes melitus menggunakan Algoritma K-Means Clustering menghasilkan bahwa penyebaran data yang seragam dalam klaster, memberikan pemahaman mendalam tentang pola data. Hasil tersebut dapat mendukung pemahaman dan penanganan lebih lanjut terhadap penyakit diabetes. Penelitian serupa yang dilakukan oleh

Satriatama et al. [8] analisis klaster menggunakan K-Means mampu menunjukkan adanya dua kelompok pasien.

Tak hanya dalam mendeteksi penyakit diabetes saja, metode K-Means terbukti mampu menjadi salah satu solusi untuk menyajikan visualisasi informasi mengenai penyakit, seperti penelitian yang dilakukan oleh Anggraini et al. [9] Pada penelitian tersebut menunjukkan bahwa penggunaan 2 cluster menjadi cluster terbaik dengan nilai Silhouette Coefficient menunjukkan hasil dengan nilai SC yaitu 0,646. Penelitian lain yang memanfaatkan K-Means Clustering seperti penelitian yang dilakukan oleh Laksono et al. [10] Yang mana metode data mining K-Means clustering dapat membantu mengorganisir dan menganalisis data rekam medis dengan lebih efektif, sehingga dapat meningkatkan kualitas pelayanan.

B. METODE PENELITIAN

Sumber Data

Dataset pada penelitian ini berasal dari platform Kaggle. Tujuan utama dataset ini adalah untuk melakukan prediksi diagnostik terhadap kemungkinan seorang pasien mengidap diabetes berdasarkan sejumlah parameter diagnostik yang tersedia. Data disimpan dalam format CSV, sehingga mempermudah proses pengolahan dan analisis menggunakan Google Colab.

Sumber : (https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset/data)

Preprocessing Data

Sebelum dilakukan analisis dan implementasi model machine learning, dataset ini melalui beberapa tahap preprocessing agar dataset siap digunakan untuk proses analisis lebih lanjut sebagai berikut:

Pemisahan Fitur dan Target

Dataset terdiri dari 8 atribut dan 1 label target (Outcome), yang mengindikasikan apakah seseorang menderita diabetes (1) atau tidak (0). Fitur-fitur yang digunakan dalam penelitian ini meliputi:

a) Pregnancies: Jumlah kehamilan
b) Glucose: Kadar glukosa dalam darah
c) Blood Pressure: Tekanan darah
d) Skin Thickness: Ketebalan lipatan kulit
e) Insulin: Kadar insulin dalam darah

f) BMI: Indeks Massa Tubuh
g) Diabetes Pedigree Function: Faktor genetik diabetes
h) Age: Usia pasien

Pembersihan Data (Data Cleaning)

Kolom Outcome dihapus pada tahap awal karena metode clustering tidak memerlukan label dalam pembentukan cluster.

Normalisasi Data

Proses normalisasi dilakukan pada seluruh variabel numerik menggunakan metode MinMaxScaler untuk menyamakan skala setiap fitur dalam rentang 0 hingga 1. Langkah ini bertujuan agar tidak ada fitur yang memiliki pengaruh berlebihan terhadap perhitungan jarak dalam proses clustering. Normalisasi juga penting untuk mencegah fitur dengan nilai skala besar, seperti kadar glukosa dan BMI, mendominasi hasil pengelompokan data.

Penentuan Jumlah Cluster Optimal

a) Elbow Method digunakan untuk menentukan jumlah cluster optimal dengan mengidentifikasi titik siku pada grafik inertia.
b) Silhouette Score digunakan untuk mengevaluasi kualitas pembentukan cluster dan memastikan pemisahan antar cluster berjalan baik.

Pemilihan Metode Cluster

Berdasarkan karakteristik dataset yang bersifat numerik dan tujuan penelitian untuk mengelompokkan pasien berdasarkan kemiripan atribut medis, dipilih algoritma K-Means Clustering. Algoritma ini efisien, mudah diimplementasikan, dan cocok untuk dataset berukuran sedang. Pada penelitian ini digunakan nilai k = 3 yang diperoleh dari analisis sebelumnya.

C. HASIL DAN PEMBAHASAN

Hasil penelitian ini menyajikan penerapan metode K-Means Clustering pada dataset diabetes, mencakup penentuan jumlah cluster optimal, karakteristik setiap cluster, visualisasi hasil pengelompokan, evaluasi model, serta analisis pembahasan temuan.

Hasil Penentuan Jumlah Cluster Optimal

Proses penentuan jumlah cluster optimal dilakukan menggunakan Elbow Method dan Silhouette Score.

- Elbow Method menunjukkan titik siku pada k = 3, yang mengindikasikan jumlah cluster optimal.

- Nilai Silhouette Score sebesar 0,1815 menunjukkan bahwa struktur cluster masih lemah, namun tetap mampu memberikan pemisahan awal antar kelompok risiko.

Kombinasi hasil kedua metode ini menjadi dasar pemilihan jumlah cluster k = 3 untuk proses clustering selanjutnya.

Hasil Clustering dengan K-Means

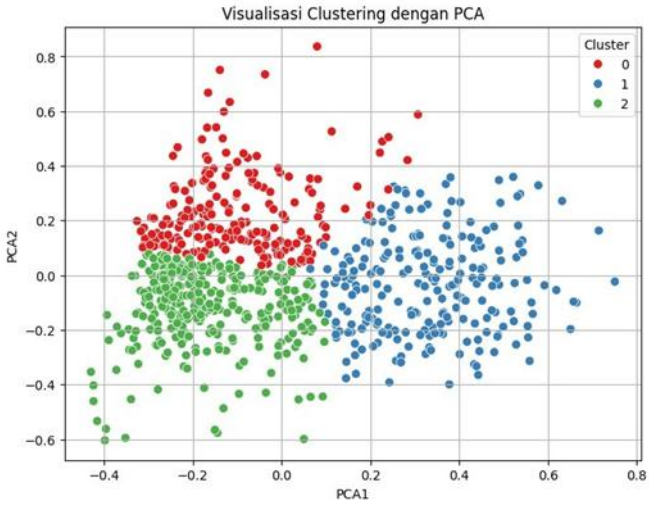
Dengan nilai k = 3, algoritma K-Means Clustering menghasilkan tiga kelompok pasien sebagai berikut:

1. Cluster 0 (Muda & Sehat, Risiko Rendah)
    a) Usia rata-rata: 26 tahun
    b) Glukosa: 104,3 mg/dL (rendah)
    c) BMI: 28,7 (normal)
    d) Kadar insulin rendah
2. Cluster 1 (Obesitas & Risiko Tinggi)
    a) Usia rata-rata: 28,9 tahun
    b) Glukosa: 136,3 mg/dL (tertinggi)
    c) BMI: 36,7 (obesitas)
    d) Insulin: 171,97 µU/mL (tertinggi)
3. Cluster 2 (Usia Tua, Risiko Sedang)
    a) Usia rata-rata: 46,7 tahun
    b) Glukosa: 130,4 mg/dL (sedang)
    c) BMI: 32,5 (moderat)
    d) Kehamilan rata-rata: 7,67

Berdasarkan hasil cluster, Cluster 0 mudah diidentifikasi sebagai pasien berisiko rendah, sementara Cluster 1 dan 2 lebih sulit dibedakan, sehingga masih ada pasien berisiko tinggi atau sedang yang masuk ke cluster lain. Hal ini menunjukkan model lebih akurat mengenali pasien sehat dibanding pasien berisiko tinggi.

Visualisasi Hasil Clustering

Untuk mempermudah pemahaman hasil clustering, data divisualisasikan menggunakan Principal Component Analysis (PCA). PCA berfungsi mereduksi dimensi data sehingga dapat ditampilkan dalam dua dimensi. Gambar berikut menunjukkan distribusi cluster yang terbentuk setelah pengelompokan data menggunakan algoritma K-Means.



Gambar 1. Visualisasi PCA

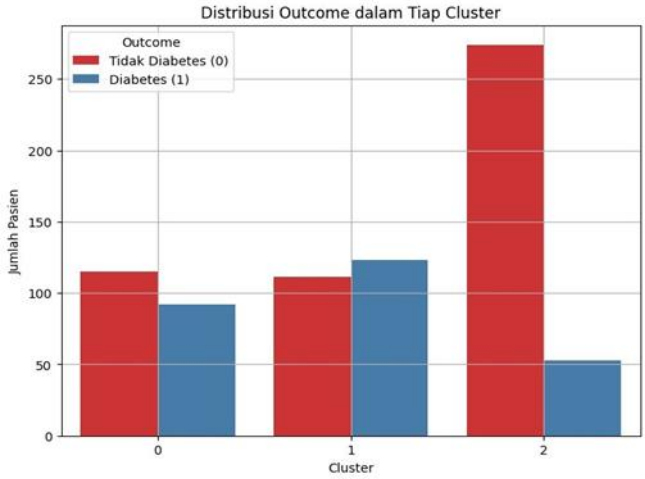
Dari hasil visualisasi diatas, dapat dilihat bahwa ke-3 cluster terpisah dengan jelas, menunjukan bahwa K-Means berhasil mengelompokan data dengan baik berdasarkan fitur fitur utama seperti glukosa BMI dan usia.

Perbandingan Hasil Clustering dengan Outcome

Kolom Outcome tidak digunakan dalam proses pembentukan cluster karena algoritma K-Means merupakan metode unsupervised learning. Kolom tersebut dihapus pada tahap awal preprocessing dan hanya ditambahkan kembali setelah proses clustering selesai untuk tujuan evaluasi dan interpretasi hasil.

Untuk mengetahui hubungan antara hasil clustering dengan label asli (Outcome), dilakukan analisis distribusi Outcome pada masing-masing cluster. Hasil analisis menunjukkan:

- Cluster 0 → mayoritas Outcome = 0 (tidak diabetes) → risiko rendah.
- Cluster 1 → mayoritas Outcome = 1 (diabetes) → risiko tinggi.
- Cluster 2 → distribusi Outcome seimbang → risiko sedang.



Gambar 2. Perbandingan Outcome

Visualisasi perbandingan distribusi Outcome pada setiap cluster menggunakan countplot, yang memperlihatkan perbedaan distribusi yang cukup signifikan, khususnya pada Cluster 1 yang didominasi oleh pasien dengan diabetes.

Evaluasi Model

Pemilihan jumlah cluster optimal dilakukan menggunakan Elbow Method, yang menunjukkan k = 3 sebagai titik ideal karena setelah k = 3, penurunan nilai SSE mulai melambat. Hasil Silhouette Score sebesar 0,1815 menunjukkan kualitas pemisahan cluster yang cukup baik. Kedua metrik ini konsisten mendukung penggunaan k = 3, sehingga pengelompokan data pasien ke dalam tiga cluster dapat dianggap valid.

Pembahasan

Analisis menunjukkan bahwa glukosa, BMI, dan usia menjadi faktor pembeda utama antar cluster.

a) Cluster 1 mengindikasikan kelompok obesitas dengan glukosa tinggi dan risiko tertinggi mengidap diabetes, sejalan dengan penelitian sebelumnya yang menegaskan obesitas & hiperglikemia sebagai faktor utama perkembangan diabetes.
b) Cluster 0 mewakili kelompok muda dengan metabolisme sehat, risiko rendah.
c) Cluster 2 mencerminkan kelompok usia lanjut dengan risiko sedang, di mana faktor usia berperan besar dalam peningkatan risiko.

Hasil clustering ini dapat digunakan untuk pengelompokan pasien guna keperluan medis seperti program pencegahan dan pengendalian diabetes. Namun, terdapat keterbatasan:

a) Dataset hanya memuat variabel tertentu sehingga faktor lain yang berpotensi memengaruhi risiko diabetes tidak terakomodasi.

b) K-Means efektif untuk data numerik berbasis jarak, tetapi kurang optimal mendeteksi bentuk cluster kompleks; algoritma lain seperti DBSCAN atau Gaussian Mixture Model (GMM) dapat dieksplorasi untuk analisis lanjutan.

D. KESIMPULAN

Berdasarkan penerapan algoritma K-Means Clustering dengan k = 3 pada dataset diabetes, penelitian ini berhasil mengelompokkan pasien menjadi tiga kategori utama, yaitu:

a) pasien muda dengan kondisi metabolisme sehat dan risiko rendah.
b) pasien obesitas dengan kadar glukosa tinggi yang memiliki risiko tertinggi terhadap diabetes.
c) pasien usia lanjut dengan risiko sedang.

Hasil analisis mengonfirmasi hubungan yang signifikan secara pola clustering antara kadar glukosa, indeks massa tubuh (BMI), dan usia terhadap tingkat risiko diabetes. Perbandingan hasil clustering dengan label *Outcome* memperlihatkan bahwa kelompok obesitas dengan glukosa tinggi memiliki peluang terbesar untuk mengidap diabetes.

Temuan ini memberikan dasar yang kuat bagi perancangan strategi pencegahan dan intervensi medis yang lebih terarah pada kelompok berisiko tinggi. Untuk penelitian selanjutnya, disarankan memasukkan variabel tambahan dan mempertimbangkan algoritma clustering alternatif guna meningkatkan akurasi dan ketepatan pengelompokan.

E. DAFTAR PUSTAKA

[1] M. Sukmadani Rusdi, “Hipoglikemia Pada Pasien Diabetes Melitus,” JSSCR, vol. 2, no. 2, pp. 83–90, Aug. 2020, doi: 10.37311/jsscr.v2i2.4575.

[2] Hartanti, J. K. Pudjibudojo, L. Aditama, and R. P. Rahayu, “Pencegahan dan Penanganan Diabetes Mellitus,” Fak. Psikol. Univ. Surabaya, pp. 1–96, 2013.

[3] P. Rahayu et al., Buku Ajar Data Mining, vol. 1, no. January 2024. 2018.

[4] M. D. Kurniawan, B. Priyatna, and F. Nurapriani, “Implementasi Algoritma K-Means Untuk Klasterisasi Data Obat Puskesmas Kotabaru,” vol. 7, 2023.

[5] M. Fadliansyah, “PROGRAM STUDI INFORMATIKA FAKULTAS TEKNIK UNIVERSITAS MUHAMMADIYAH MAKASSAR 2024”.

[6] A. Praja, C. Lubis, and D. E. Herwindiati, “DETEKSI PENYAKIT DIABETES DENGAN METODE FUZZY C-MEANS CLUSTERING DAN K-MEANS CLUSTERING,” Computatio, vol. 1, no. 1, p. 15, Apr. 2017, doi: 10.24912/computatio.v1i1.233.

[7] R. Gestavito, A. I. Hadiana, and F. R. Umbara, “PENGELOMPOKAN TINGKAT RISIKO

PENYAKIT DIABETES MELITUS MENGGUNAKAN ALGORITMA K-MEANS CLUSTERING".

[8] A. E. Satriatama et al., "Analisis Klaster Data Pasien Diabetes untuk Identifikasi Pola dan Karakteristik Pasien," JTEKSIS, vol. 5, no. 3, pp. 172–182, Jul. 2023, doi: 10.47233/jteksis.v5i3.828.

[9] R. Anggraini, E. Haerani, J. Jasril, and I. Afrianty, "Pengelompokkan Penyakit Pasien Menggunakan Algoritma K-Means," Jur. Ris. Kom., vol. 9, no. 6, p. 1840, Dec. 2022, doi: 10.30865/jurikom.v9i6.5145.

[10] B. Laksono, Y. Syahidin, and Y. Yunengsih, "Implementasi Data Mining Klasterisasi Data Pasien Rawat Inap dengan Algoritma K-Means Clustering," JTSIA, vol. 7, no. 2, pp. 621–627, Apr. 2024, doi: 10.32493/jtsi.v7i2.39354.