**1. Judul**

**Analisis Klasterisasi Data Kesehatan untuk Identifikasi Risiko Stroke**

**2. Deskripsi Permasalahan**

**Permasalahan atau Tujuan Analisis:**  
Tujuan dari analisis ini adalah untuk mengelompokkan pasien berdasarkan kemiripan karakteristik kesehatan mereka guna mengidentifikasi kelompok dengan risiko stroke yang lebih tinggi.

**Alasan Pemilihan Metode:**  
Metode klasterisasi dipilih karena data ini tidak memiliki label target yang jelas. Oleh karena itu, metode *unsupervised learning* seperti K-Means sangat sesuai untuk menemukan pola atau kelompok tersembunyi dalam data. Data yang digunakan cukup kompleks karena mengandung fitur numerik dan kategorikal, serta memiliki variasi nilai yang tinggi.

**Deskripsi Singkat Metode:**  
K-Means adalah algoritma pembelajaran tak terawasi yang membagi data ke dalam sejumlah klaster berdasarkan kedekatan nilai fitur. Proses iteratif ini meminimalkan jarak antara anggota klaster dan pusat klasternya (centroid), sehingga menghasilkan pembagian kelompok yang optimal.

**3. Deskripsi Dataset**

Dataset yang digunakan merupakan dataset stroke dari sektor kesehatan yang terdiri dari **5110 observasi (baris)** dan **11 fitur utama**, seperti usia, jenis kelamin, tekanan darah, indeks massa tubuh (*BMI*), dan kadar glukosa.

**Fitur yang paling relevan** untuk analisis ini dipilih sebagai berikut:

* age: usia pasien
* avg\_glucose\_level: rata-rata kadar glukosa dalam darah
* bmi: indeks massa tubuh pasien

**4. Tahapan Pengolahan Data**

**Preprocessing:**

* Penanganan nilai hilang pada kolom bmi menggunakan imputasi.
* Encoding fitur kategorikal menggunakan label encoding.
* Normalisasi fitur numerik menggunakan z-score agar skala data seragam.

**Training dan Evaluasi Model:**

* Data dibagi berdasarkan fitur numerik yang relevan (age, avg\_glucose\_level, bmi).
* Model K-Means dijalankan dengan jumlah klaster optimal **k=3** (berdasarkan metode elbow).
* Evaluasi dilakukan menggunakan **Silhouette Score**.

**Visualisasi:**

* Hasil klaster divisualisasikan dalam bentuk 2D menggunakan PCA.
* Distribusi stroke ditampilkan per klaster.
* Diagram batang menunjukkan rata-rata nilai fitur pada tiap klaster.

**5. Hasil dan Analisis**

**Metode Evaluasi:**  
Silhouette Score digunakan untuk menilai kualitas pemisahan klaster. Nilai Silhouette Score sebesar **0.2676** menunjukkan bahwa pemisahan klaster cukup baik, walau belum sempurna.

**Rata-rata Fitur per Klaster:**

|  |  |  |  |
| --- | --- | --- | --- |
| Cluster | Age | Avg Glucose Level | BMI |
| 0 | 0.218 | -0.370 | 0.148 |
| 1 | -1.447 | -0.263 | -0.998 |
| 2 | 0.778 | 2.154 | 0.549 |

**Distribusi Stroke per Klaster:**

|  |  |
| --- | --- |
| Cluster | Proporsi Stroke |
| 0 | 4.36% |
| 1 | 0.33% |
| 2 | 13.22% |

**Analisis Hasil:**

* **Klaster 2** merupakan kelompok dengan risiko stroke tertinggi (**13.22%**) dan memiliki nilai rata-rata usia, glukosa, dan BMI yang juga tertinggi.
* **Klaster 1** memiliki risiko stroke paling rendah (**0.33%**) dengan nilai usia dan BMI yang sangat rendah.
* **Klaster 0** menunjukkan risiko stroke menengah (**4.36%**) dengan karakteristik usia dan BMI sedang, namun kadar glukosa di bawah rata-rata.

**Kelebihan:**

* Model mampu mengelompokkan individu berdasarkan ciri-ciri kesehatan penting secara cukup terpisah.

**Kelemahan:**

* Nilai Silhouette Score < 0.3 menunjukkan bahwa ada overlap antar klaster, sehingga pemisahan belom optimal namun sudah cukup baik

**6. Kesimpulan**

Analisis klasterisasi pada data kesehatan menunjukkan bahwa pasien dengan kadar glukosa tinggi, usia tua, dan BMI tinggi (Cluster 2) memiliki risiko stroke yang paling tinggi. Hal ini konsisten dengan pengetahuan medis bahwa faktor-faktor tersebut berkontribusi besar terhadap risiko stroke.

**Insight penting:**

* **Kadar glukosa dan usia** adalah dua fitur yang sangat berpengaruh dalam membedakan klaster risiko stroke.
* Model klasterisasi dapat menjadi alat bantu awal dalam mengidentifikasi kelompok risiko tinggi stroke