**LAPORAN PROGRAM CLUSTERING CARDIO TRAIN MENGGUNAKAN ALGORITMA KMEANS**

A yellow logo with a cartoon character

Description automatically generated

Dosen Pengampu:

Nelly Indriani Widiastuti, M.T.

Disusun oleh kelompok 4 :

Asep Yaman Suyaman (10122004)

Tito Muhammad Athoriq (10122010)

Bambang Firman Fatoni (10122027)

Dendi Ramdhani (10122035)

**PROGRAM STUDI TEKNIK INFORMATIKA**

**FAKULTAS TEKNIK DAN ILMU KOMPUTER**

**UNIVERSITAS KOMPUTER INDONESIA**

**2024**

Pendahuluan

Algoritma K-Means adalah salah satu metode clustering yang paling populer dan banyak digunakan dalam analisis data. Clustering adalah teknik pembelajaran tanpa pengawasan yang bertujuan untuk mengelompokkan data yang mirip satu sama lain ke dalam kelompok yang sama. Dalam konteks K-Means, setiap kelompok disebut "cluster," dan tujuan algoritma ini adalah untuk membagi data ke dalam K cluster yang berbeda.

Algoritma K-Means bekerja dengan cara berikut:

1. Inisialisasi: Memilih K titik secara acak sebagai centroid awal dari cluster.
2. Assignment Step: Menentukan setiap titik data ke centroid terdekat, sehingga membentuk K cluster.
3. Update Step: Menghitung ulang centroid dari setiap cluster sebagai rata-rata dari semua titik data dalam cluster tersebut.
4. Iterasi: Mengulangi langkah 2 dan 3 hingga centroid tidak lagi berubah atau perubahan sangat kecil, menunjukkan konvergensi.

Kelebihan dan Kekurangan Algoritma K-Means

Kelebihan:

* Sederhana dan Cepat: K-Means mudah diimplementasikan dan biasanya lebih cepat dibandingkan dengan metode clustering lainnya, terutama pada dataset besar.
* Efisien: Algoritma ini memiliki kompleksitas komputasi O(n), yang membuatnya efisien untuk data dalam jumlah besar.

Kekurangan:

* Pemilihan K: Menentukan jumlah cluster (K) yang optimal bisa menjadi tantangan dan biasanya memerlukan metode tambahan seperti Elbow Method atau Silhouette Analysis.
* Sensitif terhadap Inisialisasi: Hasil akhir bisa berbeda tergantung pada pemilihan centroid awal, meskipun hal ini bisa diatasi dengan melakukan beberapa kali inisialisasi.
* Tidak Cocok untuk Semua Bentuk Cluster: K-Means cenderung bekerja dengan baik pada data yang memiliki cluster berbentuk bulat dan dengan ukuran yang sama.

Algoritma K-Means memiliki berbagai aplikasi dalam berbagai bidang, di antaranya:

1. Segmentasi Pelanggan: Digunakan oleh perusahaan untuk mengelompokkan pelanggan berdasarkan karakteristik demografis atau perilaku pembelian, sehingga dapat menawarkan produk atau layanan yang lebih sesuai dengan kebutuhan masing-masing kelompok.
2. Pengenalan Pola: Dalam bidang pengenalan wajah atau pengenalan tulisan tangan, K-Means dapat digunakan untuk mengelompokkan data gambar ke dalam kelompok yang serupa.
3. Analisis Pasar: Perusahaan dapat menggunakan K-Means untuk mengelompokkan produk atau layanan berdasarkan pola penjualan atau preferensi pelanggan.
4. Image Compression: Algoritma ini juga dapat digunakan untuk kompresi gambar dengan mengelompokkan pixel gambar ke dalam beberapa cluster warna.
5. Pendeteksian Anomali: Dalam sistem keamanan siber, K-Means dapat membantu mengidentifikasi aktivitas mencurigakan dengan mengelompokkan data aktivitas jaringan.

Dengan kemampuannya yang luas dan efisiensi yang tinggi, algoritma K-Means tetap menjadi alat yang sangat berguna dalam analisis data dan pemrosesan informasi.

Teori-Teori yang Relevan

1. Teori Vektor dan Geometri Euclidean: K-Means didasarkan pada jarak Euclidean untuk menentukan seberapa mirip satu titik data dengan titik lainnya. Dalam ruang vektor, jarak Euclidean antara dua titik A dan B dengan koordinat (x1,y1) dan (x2​,y2​) didefinisikan sebagai:

Algoritma K-Means meminimalkan jumlah total jarak kuadrat antara setiap titik data dan centroid cluster terdekatnya.

1. Teori Optimasi: Proses K-Means dapat dilihat sebagai upaya untuk meminimalkan fungsi objektif, yaitu jumlah total variansi dalam cluster. Fungsi objektif ini biasanya dinyatakan sebagai:

di mana adalah jarak Euclidean antara titik data x dan centroid dan Ci​ adalah cluster ke-i. Algoritma ini berusaha untuk menemukan posisi optimal centroid yang meminimalkan J.

1. Teori Probabilitas dan Statistik: Meskipun K-Means adalah metode deterministik, pendekatan probabilistik seperti K-Means++ dapat digunakan untuk inisialisasi centroid. K-Means++ meningkatkan hasil K-Means dengan memilih centroid awal berdasarkan distribusi probabilitas, sehingga meningkatkan peluang mendapatkan hasil clustering yang lebih baik dan stabil.
2. Teori Konvergensi: K-Means biasanya dijamin untuk konvergen, meskipun tidak selalu menuju solusi global optimal. Konvergensi terjadi karena setiap langkah algoritma menurunkan fungsi objektif J, dan karena jumlah kemungkinan partisi data adalah terbatas, algoritma ini akhirnya berhenti.

Analisa

Pertama-tama masukkan terlebih dahulu data yang akan di-kluster kan, disini kita menggunakan dataset “Cardio Train” yang memiliki fitur jenis kelamin, tinggi badan, berat badan, tingkat kolesterol, glukosa, perokok, meminum alkohol, dan aktif berlari

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **gender** | **height** | **weight** | **cholesterol** | **gluc** | **smoke** | **alco** | **active** |
| 2 | 168 | 62.0 | 1 | 1 | 0 | 0 | 1 |
| 1 | 156 | 85.0 | 3 | 1 | 0 | 0 | 1 |
| 1 | 165 | 64.0 | 3 | 1 | 0 | 0 | 0 |
| 2 | 169 | 82.0 | 1 | 1 | 0 | 0 | 1 |
| 1 | 156 | 56.0 | 1 | 1 | 0 | 0 | 0 |
| 1 | 151 | 67.0 | 2 | 2 | 0 | 0 | 0 |
| 1 | 157 | 93.0 | 3 | 1 | 0 | 0 | 1 |
| 2 | 178 | 95.0 | 3 | 3 | 0 | 0 | 1 |
| 1 | 158 | 71.0 | 1 | 1 | 0 | 0 | 1 |
| 1 | 164 | 68.0 | 1 | 1 | 0 | 0 | 0 |
| 1 | 169 | 80.0 | 1 | 1 | 0 | 0 | 1 |
| 2 | 173 | 60.0 | 1 | 1 | 0 | 0 | 1 |
| 2 | 165 | 60.0 | 1 | 1 | 0 | 0 | 0 |
| 1 | 158 | 78.0 | 1 | 1 | 0 | 0 | 1 |
| 2 | 181 | 95.0 | 1 | 1 | 1 | 1 | 1 |
| 2 | 172 | 112.0 | 1 | 1 | 0 | 0 | 0 |
| 1 | 170 | 75.0 | 1 | 1 | 0 | 0 | 0 |
| 1 | 158 | 52.0 | 1 | 3 | 0 | 0 | 1 |
| 1 | 154 | 68.0 | 1 | 1 | 0 | 0 | 0 |
| 2 | 162 | 56.0 | 1 | 1 | 1 | 0 | 1 |

Langkah selanjutnya lakukan inisialisasi

1. Tentukan K, pada kasus ini diasumsikan K nya adalah 3
2. Lalu pilih secara acak data yang akan dijadikan centroid, dalam kasus ini kita pilih data ke-1, ke-10, dan ke-20
3. Cari jarak antara setiap data dengan centroidnya

|  |  |  |  |
| --- | --- | --- | --- |
| jarak ke centroid 1 | jarak ke centroid 2 | jarak ke centroid 3 | cluster |
| 0,0 | 7,3 | 8,5 | 1 |
| 26,0 | 18,9 | 29,7 | 2 |
| 4,4 | 4,6 | 8,9 | 1 |
| 20,0 | 14,9 | 26,9 | 2 |
| 13,5 | 14,4 | 6,2 | 3 |
| 17,8 | 13,1 | 15,7 | 2 |
| 33,0 | 26,1 | 37,4 | 2 |
| 34,6 | 30,6 | 42,3 | 2 |
| 13,5 | 6,8 | 15,6 | 2 |
| 7,3 | 0,0 | 12,3 | 2 |
| 18,1 | 13,0 | 25,0 | 2 |
| 5,4 | 12,1 | 11,7 | 1 |
| 3,7 | 8,1 | 5,2 | 1 |
| 18,9 | 11,7 | 22,4 | 2 |
| 35,5 | 32,0 | 43,4 | 2 |
| 50,2 | 44,7 | 56,9 | 2 |
| 13,2 | 9,2 | 20,7 | 2 |
| 14,3 | 17,2 | 6,2 | 3 |
| 15,3 | 10,0 | 14,5 | 2 |
| 8,5 | 12,3 | 0,0 | 3 |

Lalu buat centroid yang baru dengan mencari rata-rata dari fitur yang memiliki cluster yang sama.

Berikut data yang memiliki cluster 1 menurut hasil perhitungan yang pertama

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | gender | height | weight | cholesterol | gluc | smoke | alco | active |
| 2 | 168 | 62,00 | 1 | 1 | 0 | 0 | 1 |
| 1 | 165 | 64,00 | 3 | 1 | 0 | 0 | 0 |
| 2 | 173 | 60,00 | 1 | 1 | 0 | 0 | 1 |
| 2 | 165 | 60,00 | 1 | 1 | 0 | 0 | 0 |
| Rata-rata | 1,75 | 167,75 | 61,5 | 1,5 | 1 | 0 | 0 | 0,5 |

Lakukan langkah tersebut ke semua centroid yang baru, lalu hitung kembali jarak data ke centroid yang baru.

Lakukan tahap tersebut sampai centroid yang baru sama dengan centroid sebelumnya atau tidak terjadi perubahan yang besar pada centroidnya.