HIERARCHICAL CLUSTERING

KELOMPOK 6

PERKENALAN KELOMPOK:

ADAM MAHABAYU MUHIBBULLOH
234311002

AFIQ GALUH SETYA RAMADHANI
234311004

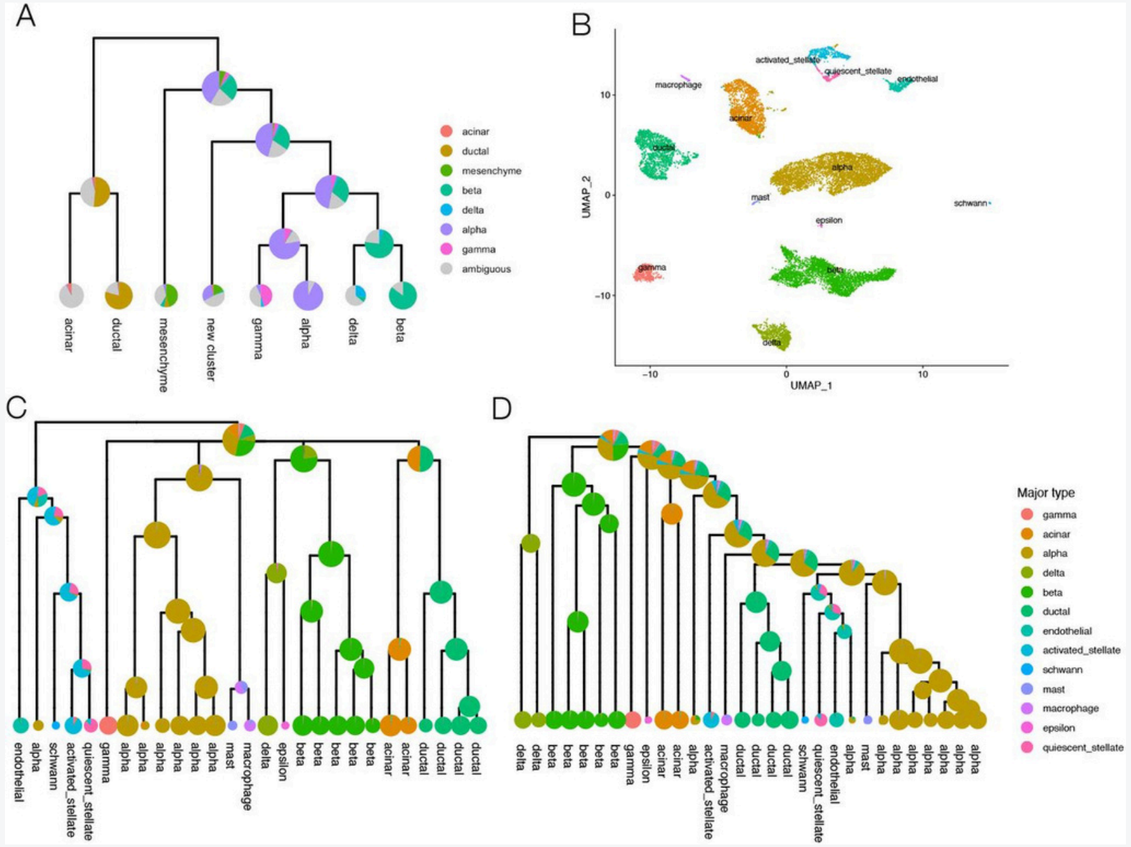
MOHAMMAD DIMAS BAHRUL IKHWANI
2343110017

APA ITU HIERARCHICAL CLUSTERING?

HIERARCHICAL CLUSTERING ADALAH METODE UNSUPERVISED LEARNING (PEMBELAJARAN TANPA PENGAWASAN).

Tujuannya adalah untuk membangun sebuah hierarki (struktur seperti pohon) dari sekumpulan data.

- Tidak perlu menentukan jumlah cluster (k) di awal.
- Hasil utamanya adalah Dendrogram, sebuah diagram pohon yang memvisualisasikan seluruh proses penggabungan.
- Metode paling umum adalah "Agglomerative" (dari bawah ke atas).



METODE AGGLOMERATIVE (BOTTOM-UP / DARI BAWAH KE ATAS)

Cara Kerja:

Mulai

Setiap titik data dianggap sebagai satu cluster individual. Jika Anda punya 300 data, Anda mulai dengan 300 cluster.

Hitung & Gabung

Algoritma mencari dua cluster yang paling mirip/dekat (berdasarkan metrik jarak dan kriteria keterkaitan), lalu menggabungkannya menjadi satu cluster baru.

Ulangi

Proses ini diulangi terus-menerus menggabungkan dua cluster terdekat hingga pada akhirnya hanya tersisa satu cluster besar yang berisi seluruh titik data.

HIPOTESA FUNCTION

DALAM HIERARCHICAL CLUSTERING, TIDAK ADA "FUNGSI HIPOTESA"

Mengapa? Tujuan kita bukan untuk memprediksi sebuah nilai (seperti harga rumah atau mpg). Tujuan kita adalah untuk menemukan struktur atau pola tersembunyi di dalam data itu sendiri.

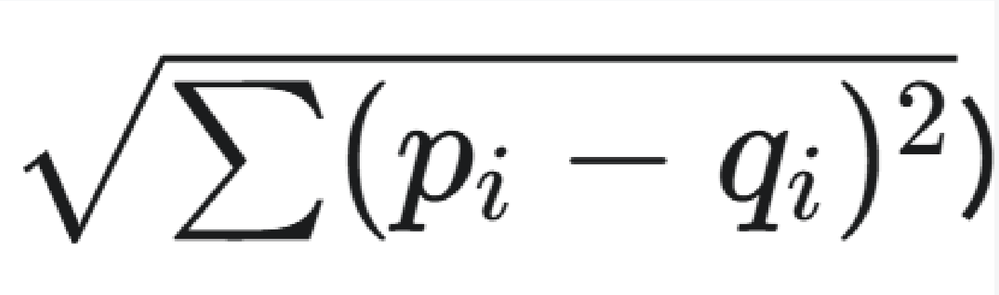
Apa Padanannya? Konsep yang paling dekat dengan "hipotesa" atau hasil dari model ini adalah Dendrogram itu sendiri.

- Dendrogram adalah "Hipotesa" Visual: Dendrogram adalah hipotesa lengkap dari model tentang seluruh struktur hierarkis data.
- Ini menunjukkan titik data mana yang paling mirip (digabung di level rendah/bawah) dan mana yang paling berbeda (baru digabung di level sangat tinggi/atas).
- Tidak "melatih" sebuah fungsi h(x). Membangun sebuah struktur (Dendrogram) yang merepresentasikan data. Dengan "memotong" dendrogram ini pada ketinggian tertentu, Anda bisa mendapatkan sejumlah k cluster

COST FUNCTION

DALAM HIERARCHICAL CLUSTERING, TIDAK ADA "FUNGSI BIAYA" GLOBAL YANG DIOPTIMALKAN.

Mengapa? Algoritma ini bersifat greedy (rakus). Pada setiap langkah, ia hanya membuat keputusan terbaik secara lokal (menggabungkan dua cluster terdekat) tanpa mempedulikan apakah keputusan itu akan optimal secara global di akhir.



Apa Padanannya? "Biaya" atau "keputusan" di setiap langkah penggabungan ditentukan oleh dua hal:

1. Metrik Jarak (Distance Metric):
   - Ini adalah "biaya" level terendah: Bagaimana mengukur jarak antara dua titik data?
   - Pilihan umum adalah Jarak Euclidean (jarak garis lurus).

2. Kriteria Keterkaitan (Linkage Criterion):
   - Inilah konsep yang paling dekat dengan "Fungsi Biaya".
   - Ini mendefinisikan cara mengukur jarak antara dua cluster (yang mungkin berisi banyak titik). "Biaya" untuk menggabungkan dua cluster dihitung berdasarkan kriteria ini.

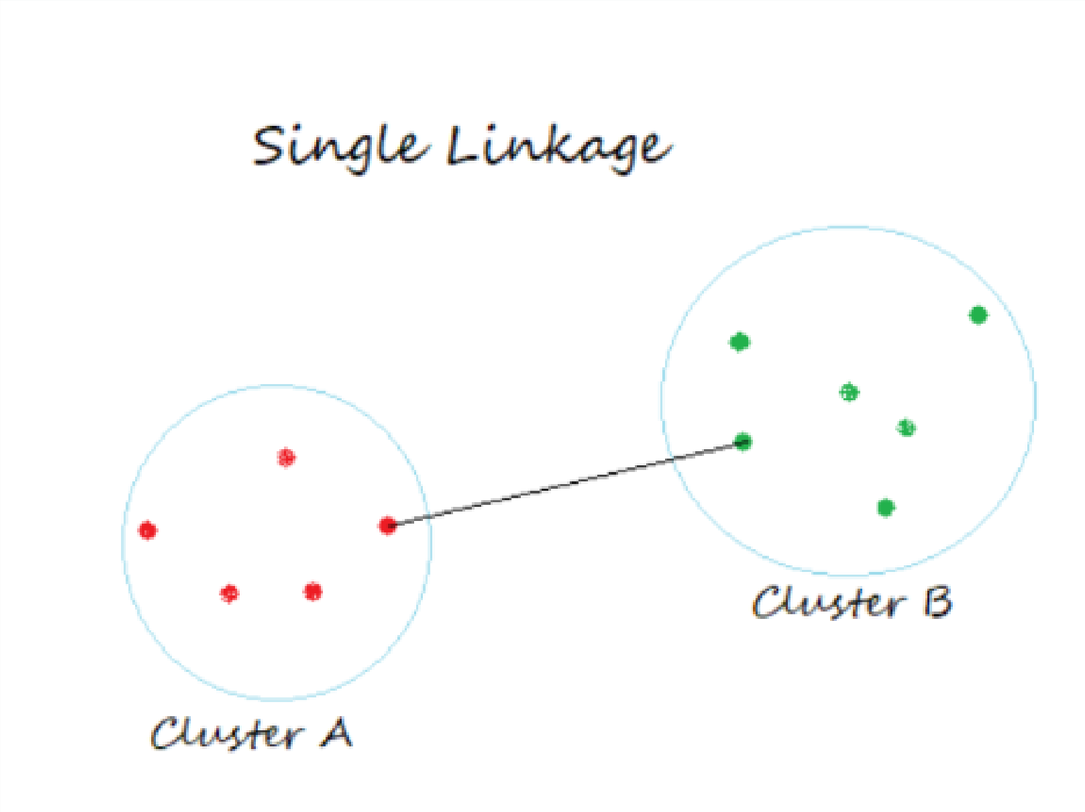
COST FUNCTION (LANJUT)

Pilihan Kriteria Linkage:

- Single Linkage: Jarak antara dua cluster = jarak antara dua titik terdekat dari masing-masing cluster. Cenderung menghasilkan cluster yang panjang seperti rantai.
- Complete Linkage: Jarak antara dua cluster = jarak antara dua titik terjauh dari masing-masing cluster. Cenderung menghasilkan cluster yang bulat dan padat.
- Average Linkage: Jarak antara dua cluster = jarak rata-rata dari semua pasangan titik di kedua cluster.
- Ward's Linkage (Yang Kita Gunakan): Ini adalah yang paling mirip dengan "meminimalkan fungsi biaya". Ward's Linkage menggabungkan dua cluster sedemikian rupa sehingga peningkatan total varians internal (within-cluster sum of squares) menjadi seminimal mungkin. Ia mencoba menemukan cluster yang paling padat dan bulat.

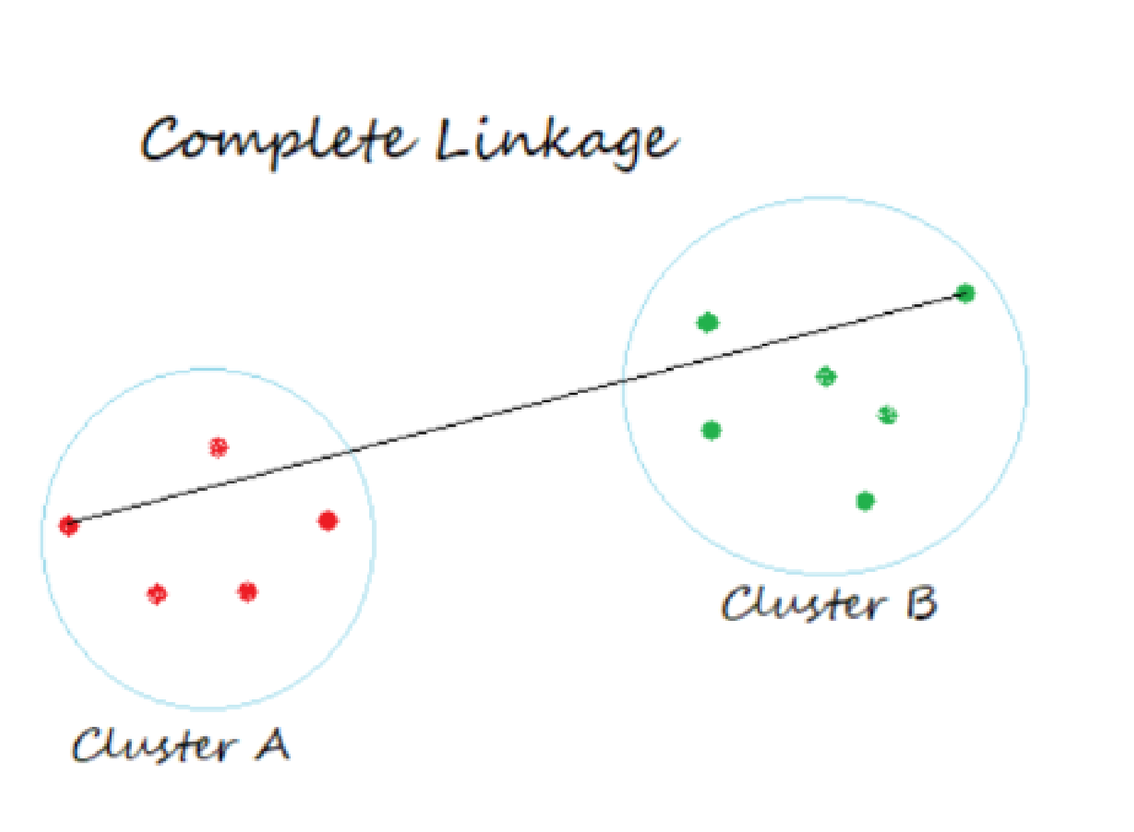
CONTOH VISUALISASI

Single Linkage: Jarak antara dua cluster = jarak antara dua titik terdekat dari masing-masing cluster. Cenderung menghasilkan cluster yang panjang seperti rantai.



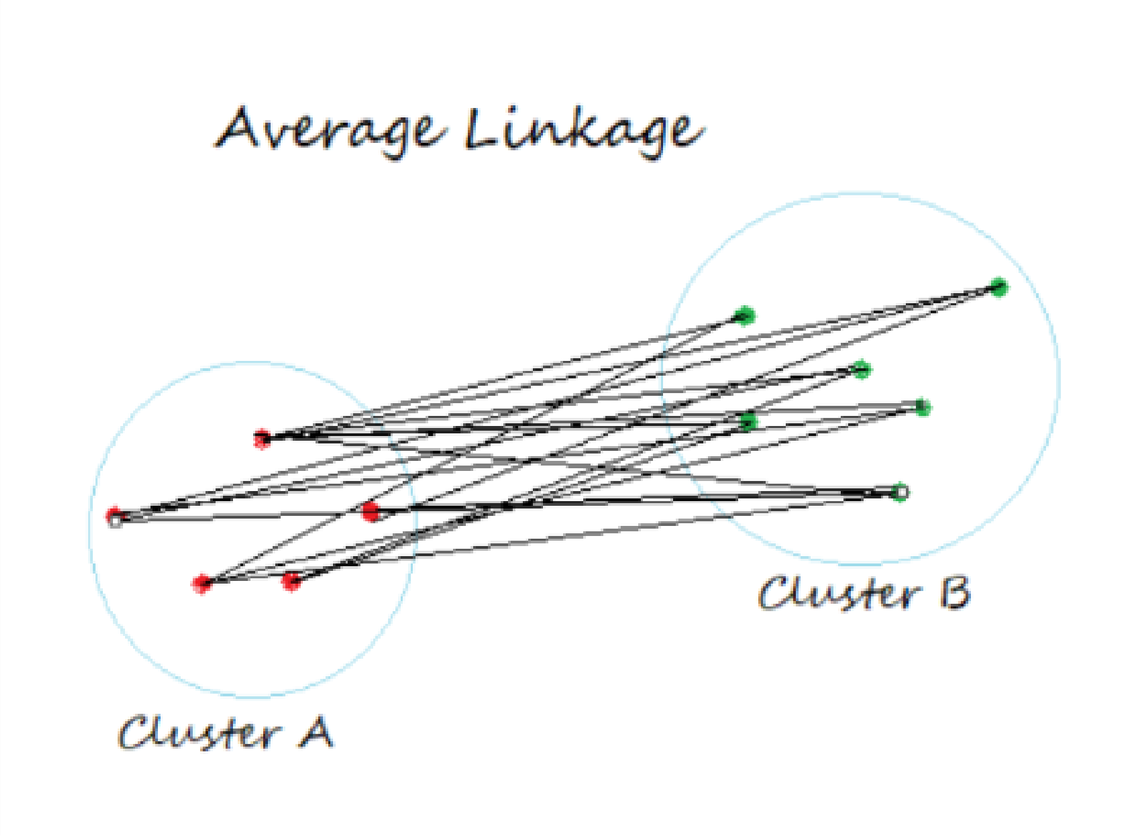
CONTOH VISUALISASI

Complete Linkage: Jarak antara dua cluster = jarak antara dua titik terjauh dari masing-masing cluster. Cenderung menghasilkan cluster yang bulat dan padat.



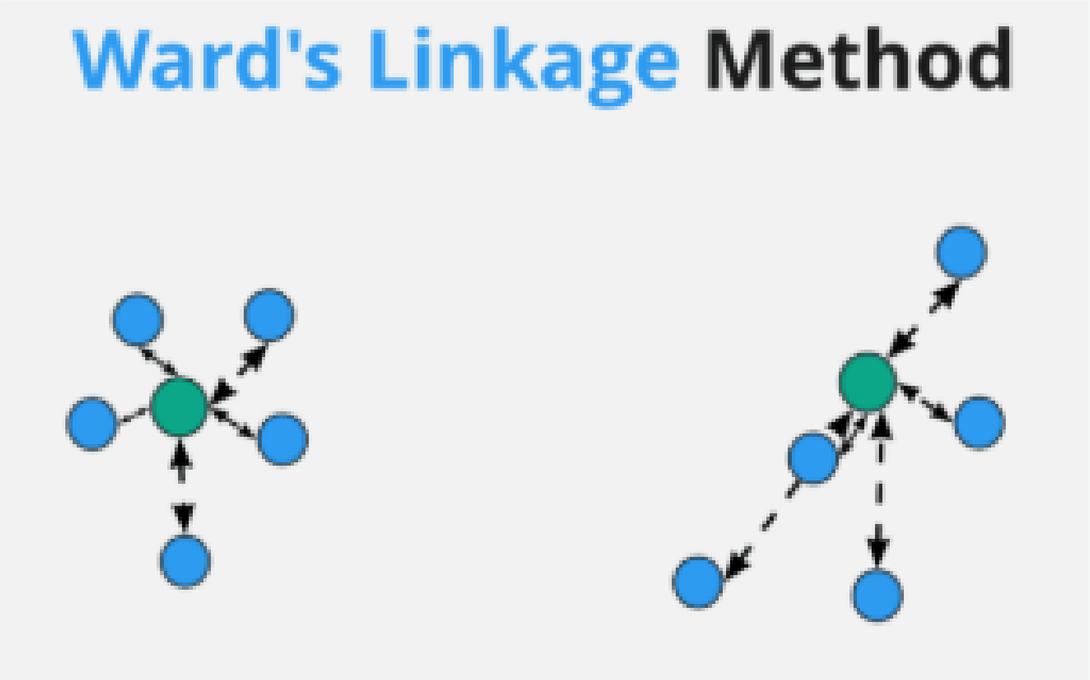
CONTOH VISUALISASI

Average Linkage: Jarak antara dua cluster = jarak rata-rata dari semua pasangan titik di kedua cluster.



CONTOH VISUALISASI

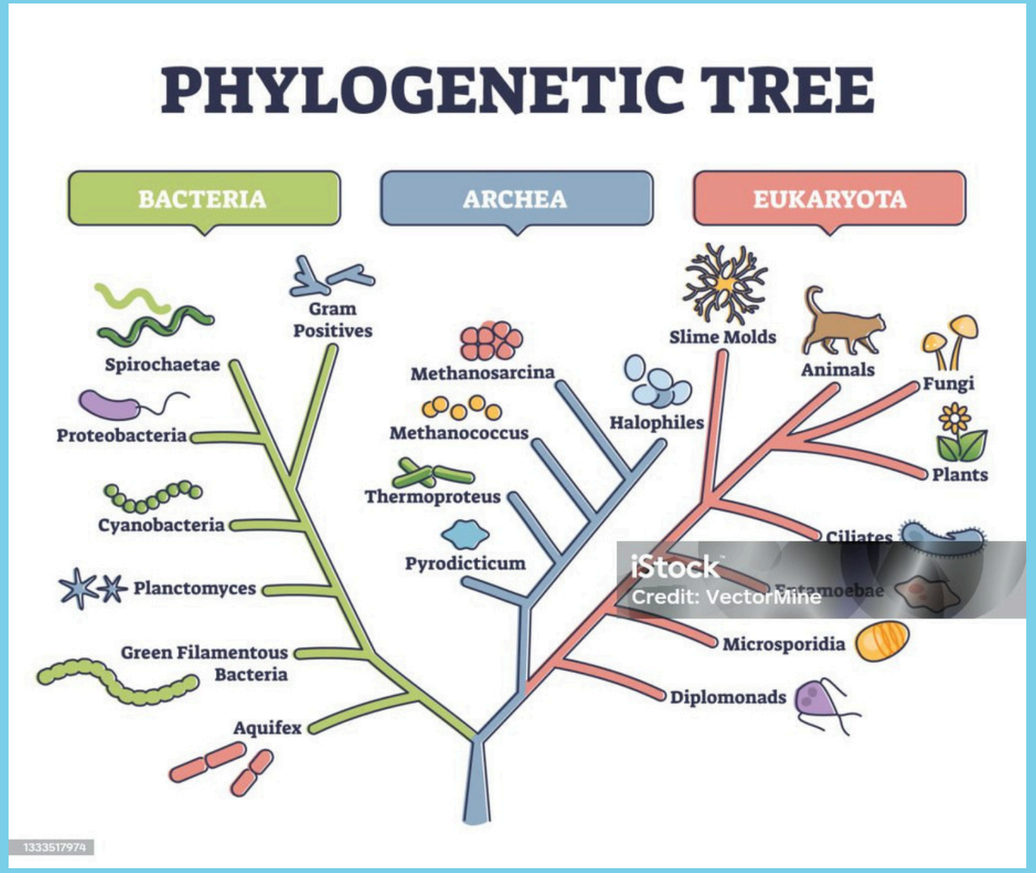
Ward's Linkage (Yang Kita Gunakan): Ini adalah yang paling mirip dengan "meminimalkan fungsi biaya". Ward's Linkage menggabungkan dua cluster sedemikian rupa sehingga peningkatan total varians internal (within-cluster sum of squares) menjadi seminimal mungkin. Ia mencoba menemukan cluster yang paling padat dan bulat.



CONTOH KASUS

BIOLOGI (TAKSONOMI GENETIK)

Ini adalah contoh paling klasik. Digunakan untuk membangun "pohon keluarga" evolusi (pohon filogenetik). Spesies dengan kesamaan genetik yang tinggi akan digabungkan di level rendah (cabang yang berdekatan), sementara spesies yang berbeda (misal: manusia dan bakteri) baru akan bertemu di level yang sangat tinggi (akar pohon).



CONTOH DARI : DIVISIVE

KESIMPULAN

Berdasarkan kode program yang dijalankan, data Auto-MPG berhasil dibersihkan, distandardisasi, dan dianalisis menggunakan metode hierarchical clustering dengan pendekatan Ward’s linkage. Dendrogram yang dihasilkan menunjukkan bahwa data mobil secara alami membentuk dua kelompok besar, namun ketika dendrogram dipotong pada level jarak yang lebih rendah, kelompok besar tersebut terpecah menjadi beberapa subkelompok, sehingga sangat memungkinkan untuk menghasilkan tiga cluster. Hal ini terjadi karena setiap cabang dalam dendrogram mewakili tingkat kemiripan antar data, dan subcabang yang bergabung pada jarak yang rendah merupakan kelompok data yang sangat mirip. Dengan demikian, proses clustering dengan tiga cluster tetap valid karena dendrogram memperlihatkan bahwa satu kelompok besar (warna oranye) dapat terpisah menjadi dua subcluster, sementara kelompok besar lainnya (warna hijau) tetap menjadi satu cluster. Secara keseluruhan, analisis ini menunjukkan bahwa mobil-mobil dalam dataset dapat dikelompokkan menjadi tiga karakter utama berdasarkan kemiripan fitur seperti berat, konsumsi bahan bakar, dan spesifikasi mesin.


THANK YOU
VERY MUCH!