K-Means

1. Cara kerja algoritma K-Means

K-Means adalah algoritma clustering yang bertujuan untuk membagi data menjadi sejumlah k cluster yang ditentukan sebelumnya, berdasarkan kedekatan jarak antara titik data dan centroid cluster. Algoritma ini bekerja dengan iteratif untuk meminimalkan variasi dalam cluster (intra-cluster variance). Langkah-langkah utama dalam algoritma K-Means adalah sebagai berikut.

Inisialisasi:

- Pilih jumlah cluster (k) yang diinginkan.
- Inisialisasi centroid awal. Ini bisa dilakukan secara acak atau dengan metode k-means++ yang memilih centroid lebih cerdas untuk memperbaiki konvergensi.

Assignment Step (Penugasan):

- Untuk setiap titik data, hitung jarak antara titik data tersebut dan setiap centroid.
- Tugaskan titik data ke cluster yang centroid-nya paling dekat (dengan jarak terkecil).

Update Step (Pembaruan):

- Setelah semua titik data ditugaskan ke cluster, hitung ulang posisi centroid dengan mengambil rata-rata dari semua titik data yang berada dalam cluster tersebut.

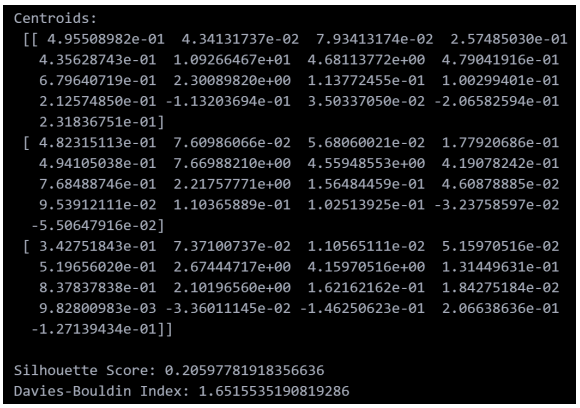
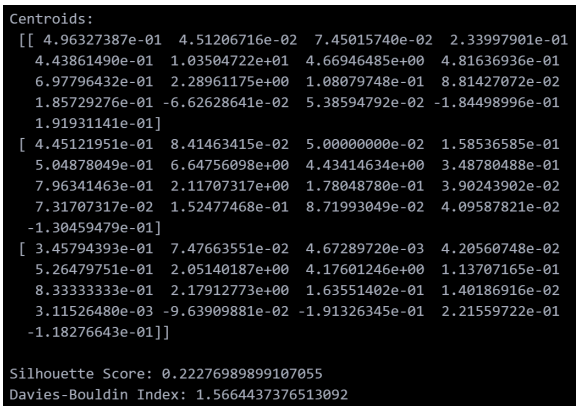
Iterasi:

- Ulangi langkah Assignment dan Update sampai centroid tidak berubah lagi (konvergensi), atau hingga jumlah iterasi maksimum tercapai.

Hasil Akhir:

- Algoritma berhenti, dan hasil akhir adalah pembagian data menjadi k cluster dengan centroid yang optimal berdasarkan minimisasi variasi dalam cluster.

4. Hasil Evaluasi Model

From Scratch	From Scikit-Learn
	

Hasil centroid dari implementasi K-Means scratch menunjukkan distribusi centroid yang berbeda dari yang dihasilkan oleh sklearn. Hal ini mungkin disebabkan oleh perbedaan dalam metode inisialisasi centroid, cara pengukuran jarak, atau konvergensi yang digunakan dalam implementasi tersebut.

Model K-Means yang diimplementasikan from scratch menghasilkan Silhouette Score sebesar 0.206 dan Davies-Bouldin Index sebesar 1.652. Silhouette Score yang rendah menunjukkan bahwa jarak antara data dalam cluster relatif dekat, tetapi jarak antara data di cluster yang berbeda cukup jauh, namun masih rendah, menunjukkan bahwa hasil clustering tidak terlalu baik. Davies-Bouldin Index yang lebih tinggi mengindikasikan bahwa rata-rata rasio jarak intra-cluster terhadap inter-cluster adalah lebih tinggi, yang berarti ada ruang untuk perbaikan dalam segregasi cluster.

Sebaliknya, model K-Means yang diimplementasikan menggunakan scikit-learn menghasilkan Silhouette Score sebesar 0.223 dan Davies-Bouldin Index sebesar 1.566. Meskipun Silhouette Score-nya masih rendah, ia menunjukkan sedikit peningkatan dibandingkan dengan implementasi from scratch, dan Davies-Bouldin Index yang lebih rendah menunjukkan bahwa model scikit-learn sedikit lebih baik dalam mengelompokkan data dengan membedakan cluster.

Hal tersebut terjadi karena implementasi from scratch mungkin memiliki keterbatasan dalam hal optimasi, seperti konvergensi centroid atau penanganan kasus khusus, yang dapat mempengaruhi hasil akhir. Metode inisialisasi centroid juga dapat mempengaruhi performa model; dalam hal ini, scikit-learn menggunakan metode K-Means++ yang lebih canggih untuk inisialisasi centroid, yang secara umum memberikan hasil yang lebih stabil dan lebih baik dibandingkan dengan inisialisasi acak atau sederhana.

5. *Improvement*

Untuk mencapai hasil clustering yang lebih baik, penting untuk melakukan optimasi konvergensi algoritma K-Means dengan memastikan bahwa centroid benar-benar mencapai posisi stabil dan tidak bergerak secara signifikan antara iterasi. Penanganan kasus khusus juga dapat dilakukan dengan memeriksa dan menangani kasus khusus seperti cluster kosong atau fitur dengan variansi sangat kecil yang mungkin mempengaruhi hasil clustering.

Selain itu, evaluasi hasil clustering harus dilakukan secara menyeluruh dengan mempertimbangkan untuk menggunakan berbagai metode tambahan seperti Dunn Index dan Elbow Method. Jika memungkinkan, lakukan cross-validation pada data untuk memastikan bahwa hasil clustering konsisten dan tidak bergantung pada subset data tertentu. Analisis mendalam terhadap cluster yang dihasilkan juga dapat membantu memastikan bahwa hasil clustering mencerminkan struktur data yang diharapkan dan memberikan wawasan tambahan tentang efektivitas model.