K-Means

Adinda Putri - 13523071

K-Means merupakan algoritma yang dapat mengelompokkan data ke beberapa cluster (clustering) berdasarkan kemiripannya. K-means digunakan untuk unlabeled data dan bertujuan untuk menemukan pola dari data-data tersebut. Contohnya, K-Means dapat digunakan untuk melakukan segmentasi customers ke dalam cluster seperti “Frequent Buyers” atau “Big Spenders” berdasarkan riwayat pembeliannya. Algoritma ini akan mengelompokkan data ke k clusters. Untuk mengukur kemiripan data-data, digunakan berbagai metrik jarak seperti euclidean, manhattan, atau minkowski.

Cara Kerja

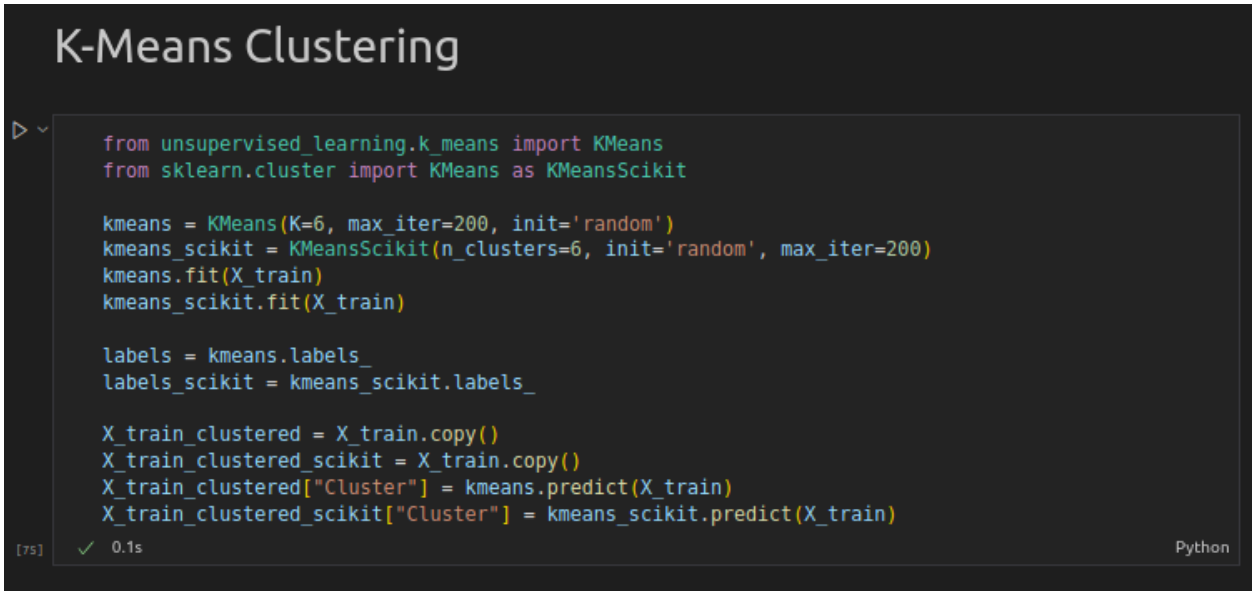
Berikut cara kerja K-Means untuk pemilihan centroid secara acak.

1. **Initialization:** Tentukan nilai *k* dan pilih *k* cluster centroids secara acak
2. **Assignment Step:** Setiap data point dikelompokkan ke centroid terdekat untuk membentuk clusters.
3. **Update Step:** Setelah assignment, hitung ulang setiap centroid dari tiap cluster dengan cara menghitung nilai rata-rata datanya.
4. **Repeat:** Langkah 1-3 diulangi hingga tidak ada centroid yang berubah pada update step atau jumlah iterasi sudah tercapai (pre-defined max_iter)

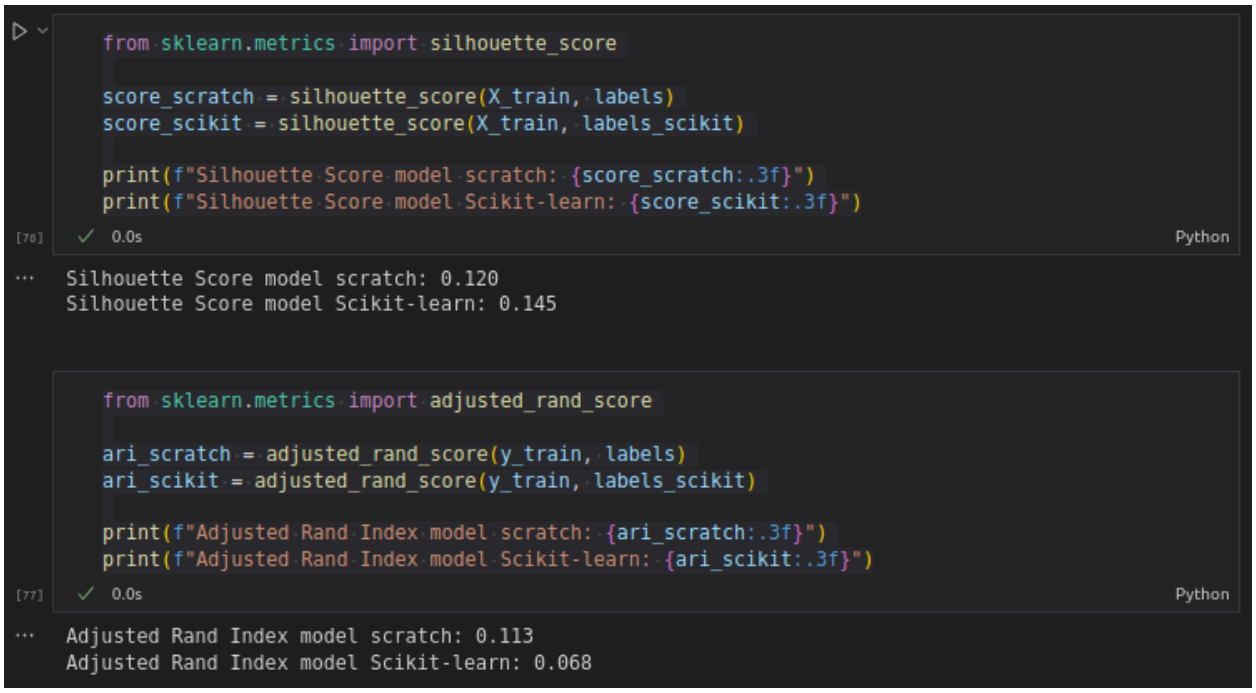
Sementara itu, berikut cara kerja K-Means++:

1. **Initialization:** Pilih satu centroid secara acak dari data points. Untuk *k-1* centroid cluster lainnya, pilih berdasarkan probabilitas yang proporsional terhadap jarak kuadrat antara data point dengan centroid terdekat.
2. **Clustering:** Kelompokkan data point ke centroid terdekat
3. **Update:** Hitung ulang centroid tiap cluster dengan menghitung rata-rata nilai datanya.
4. **Repeat:** Langkah 2-3 diulangi hingga tidak ada centroid yang berupa pada update step atau jumlah iterasi sudah tercapai (pre-defined max_iter)

Perbandingan hasil model dari scratch dengan dari Scikit-Learn:



Gambar 1. Inisialisasi, Training, dan Prediksi Masing-Masing Model
Sumber: Penulis



Gambar 2. Perbandingan Silhouette Score dan ARI Score Masing-Masing Model
Sumber: Penulis

Evaluasi yang dilakukan pada kedua model adalah silhouette score dan adjusted rand index. Silhouette score mengukur seberapa baik sebuah titik data cocok dalam cluster-nya dibandingkan dengan cluster lain. ARI mengukur kesamaan antara cluster yang ditemukan oleh model dengan label kelas asli (ground truth). Perbandingan hasil evaluasi menunjukkan bahwa implementasi K-Means dari scratch berhasil memberikan hasil yang bisa dibilang kompetitif. Perbedaan ini wajar karena K-Means bersifat stokastik dari inisialisasi centroid awal, namun hasil menunjukkan bahwa implementasi dari model scratch sudah benar.

Referensi:

[1] *K-Means Clustering from Scratch - Machine Learning with Python*, YouTube, 2020. [Daring]. Tersedia: https://www.youtube.com/watch?v=4b5d3muPQmA. [Diakses: 2 September 2025].

[2] *K-means Clustering Introduction*, GeeksforGeeks, 2024. [Daring]. Tersedia: https://www.geeksforgeeks.org/machine-learning/k-means-clustering-introduction/. [Diakses: 2 September 2025].

[3] *K-Means from Scratch*, Python Engineer, 2022. [Daring]. Tersedia: https://www.python-engineer.com/courses/mlfromscratch/12_kmeans/. [Diakses: 2 September 2025].

[4] *K-Means Algorithm*, GeeksforGeeks, 2024. [Daring]. Tersedia: https://www.geeksforgeeks.org/machine-learning/ml-k-means-algorithm/. [Diakses: 2 September 2025].