Nama : Izzan Muhammad Fa'iz (1103210126)

Kelas : TK4505

UTS III Clustering (UTSClustering)

1. Jika K-Means menghasilkan nilai Silhouette Score rendah (misalnya 0.3) walaupun Elbow Method menyarankan K=5 sebagai titik optimal, maka inkonsistensi ini bisa disebabkan oleh bentuk distribusi data yang tidak sesuai asumsi K-Means, yaitu cluster non-spherical atau tidak seimbang dalam kepadatan dan ukuran. K-Means mengasumsikan bahwa klaster memiliki bentuk bundar (isotropik) dan ukuran yang seragam. Terdapat juga solusi alternatif untuk validasi cluster yaitu:
    - Gap Statistic, yang membandingkan inertia model dengan distribusi acak untuk menentukan seberapa 'berarti' struktur klaster.
    - Bootstrapping Stability, yaitu dengan membuat sub-sample dari data dan mengukur konsistensi label clustering.

    Metode ini membantu memahami apakah hasil klaster yang diperoleh benar-benar stabil dan tidak muncul secara kebetulan. Distribusi data yang non-spherical, seperti klaster berbentuk bulan sabit atau dengan variasi densitas tinggi, membuat KMeans tidak dapat memisahkan klaster secara optimal, sehingga nilai silhouette menjadi rendah walaupun inertia (total jarak dalam klaster) tampak minimum pada elbow method.

2. Ketika dataset mengandung fitur numerik seperti Quantity dan UnitPrice serta fitur kategorikal high-cardinality seperti Description, preprocessing menjadi kunci untuk menyelaraskan skala dan menjaga struktur informasi. Langkah preprocessing efektif yaitu seperti Standardisasi (contohnya Z-Score) pada fitur numerik agar berada pada skala yang seragam. Lalu untuk representasi teks menggunakan:
    - TF-IDF Vectorization, yang menyeimbangkan frekuensi kata umum dan unik.
    - UMAP (Uniform Manifold Approximation and Projection) untuk mereduksi dimensi hasil TF-IDF ke bentuk yang kompak.

    Adapun resiko One-Hot Encoding untuk Description sangat tinggi yaitu:
    - Dapat menghasilkan ribuan kolom karena banyaknya deskripsi unik.

- Meningkatkan sparsity dan dimensionality, menyebabkan algoritma seperti KMeans atau DBSCAN menjadi tidak efisien dan overfitting.
- Tidak mempertahankan semantik antar kata/produk.

Sebaliknya, TF-IDF dan embedding berdimensi rendah mampu mempertahankan struktur semantik dan hubungan konteks antar produk, yang jauh lebih berguna untuk mengelompokkan pola belanja yang bermakna.

3. Model DBSCAN dikenal sangat sensitif terhadap pemilihan nilai parameter epsilon (eps), yang menentukan radius maksimum agar titik data dapat dianggap sebagai bagian dari satu klaster. Dalam dataset yang tidak seimbang, seperti ketika 90% pelanggan berasal dari satu negara tertentu seperti UK, kesalahan dalam menetapkan nilai epsilon dapat menyebabkan klasterisasi yang gagal atau terlalu banyak noise. Untuk mengatasi hal ini, pendekatan yang paling umum adalah dengan membuat grafik k-distance, yaitu grafik yang memplot jarak ke tetangga ke-k (dengan k = MinPts). Titik tekukan (elbow) dari grafik ini dapat digunakan sebagai estimasi epsilon yang optimal. Secara alternatif, kuartil ke-3 atau ke-4 dari distribusi jarak antar titik juga dapat digunakan sebagai parameter epsilon awal dalam proses automasi. Sementara itu, nilai MinPts tidak boleh tetap, tetapi perlu disesuaikan dengan kerapatan lokal. Di daerah yang sangat padat, MinPts yang lebih besar akan membantu memfilter outlier, sedangkan di daerah jarang, MinPts yang lebih kecil akan tetap mampu mengidentifikasi klaster kecil yang valid. Strategi adaptif ini penting agar DBSCAN dapat memisahkan klaster padat dari noise secara efektif, terutama dalam dataset transaksi ritel dengan variasi kerapatan regional.
4. Jika hasil clustering menunjukkan overlap signifikan antara "high-value customers" dan "bulk buyers" (keduanya memiliki total pengeluaran tinggi), maka pendekatan unsupervised murni tidak cukup untuk memisahkan klaster yang konseptualnya berbeda namun datanya mirip. Terdapat juga dua solusi lanjutan yang bisa digunakan yaitu:
    - Semi-supervised clustering, seperti Constrained K-Means (COP-KMeans), dengan penambahan must-link dan cannot-link constraints berdasarkan domain knowledge.
    - Metric learning, seperti menggunakan Mahalanobis distance yang mengukur jarak antar titik berdasarkan korelasi fitur, bukan sekadar Euclidean distance.

Namun, penggunaan teknik seperti Mahalanobis atau embedding seringkali mengurangi interpretabilitas dari hasil klaster bagi pemangku kepentingan bisnis. Oleh karena itu, perlu dijaga keseimbangan antara akurasi teknis dan keterbacaan bisnis, misalnya dengan tetap melabeli hasil akhir menggunakan segmen bisnis seperti "frequent-low", "rare-high", dsb.

5. Untuk mengidentifikasi pola pembelian periodik, fitur temporal dari kolom InvoiceDate dapat dirancang menjadi berbagai bentuk seperti hari dalam seminggu, jam pembelian, atau representasi siklikal menggunakan transformasi sinus dan kosinus terhadap waktu. Fitur-fitur ini mampu menangkap kebiasaan pelanggan, seperti kecenderungan melakukan pembelian pada pagi hari atau menjelang akhir pekan. Namun, penting untuk memperhatikan risiko data leakage ketika menggunakan fitur agregasi waktu seperti rata-rata pembelian bulanan atau total pembelian mingguan. Jika agregasi ini dihitung berdasarkan keseluruhan data sebelum pembagian data ke dalam train-test, maka informasi dari masa depan dapat bocor ke model pelatihan, menghasilkan nilai AUC atau metrik evaluasi yang terlalu tinggi namun tidak realistis. Untuk mencegah hal ini, sangat disarankan untuk melakukan split data berdasarkan waktu terlebih dahulu (*temporal split*), dan memastikan bahwa semua fitur yang digunakan pada saat pelatihan hanya berasal dari periode sebelumnya. Selain itu, penggunaan fitur lag seperti pembelian dalam 7 hari terakhir perlu dilakukan dengan hati-hati karena bisa memperkenalkan noise jika perilaku pembelian pelanggan sangat bervariasi. Fitur-fitur ini sebaiknya digunakan hanya jika terdapat pola siklus pembelian yang kuat dan stabil.