Analisa Model Clustering

Nama : Atanasius Pradiptha Sampurno
Kelas : TK-45-02
NIM : 1103213036

1. Jika algoritma K-Means menghasilkan nilai silhouette score rendah (0.3) meskipun elbow method menunjukkan K=5 sebagai optimal pada dataset ini, faktor apa yang menyebabkan inkonsistensi ini? Bagaimana strategi validasi alternatif (misal: analisis gap statistic atau validasi stabilitas cluster via bootstrapping) dapat mengatasi masalah ini, dan mengapa distribusi data non-spherical menjadi akar masalahnya?

   Jawaban:

   Nilai silhouette score rendah meskipun K=5 optimal menurut elbow method bisa disebabkan oleh distribusi data non-spherical yang tidak cocok dengan K-Means. Metode alternatif seperti gap statistic dan bootstrapping lebih cocok untuk evaluasi clustering data dengan bentuk tidak sferis.

2. Dalam dataset dengan campuran fitur numerik (Quantity, UnitPrice) dan kategorikal high-cardinality (Description), metode preprocessing apa yang efektif untuk menyelaraskan skala dan merepresentasikan fitur teks sebelum clustering? Jelaskan risiko menggunakan One-Hot Encoding untuk Description, dan mengapa teknik seperti TF-IDF atau embedding berdimensi rendah (UMAP) lebih robust untuk mempertahankan struktur cluster!

   Jawaban:

   Fitur numerik seperti Quantity dan UnitPrice perlu dinormalisasi atau distandarisasi. Untuk Description dengan high-cardinality, One-Hot Encoding kurang efektif, sementara TF-IDF atau embedding seperti UMAP lebih cocok karena mengurangi dimensi dan mempertahankan informasi penting.

3. Hasil clustering dengan DBSCAN sangat sensitif terhadap parameter epsilon—bagaimana menentukan nilai optimal epsilon secara adaptif untuk memisahkan cluster padat dari noise pada data transaksi yang tidak seimbang (misal: 90% pelanggan dari UK)? Jelaskan peran k-distance graph dan kuartil ke-3 dalam automasi parameter, serta mengapa MinPts harus disesuaikan berdasarkan kerapatan regional!

   Jawaban:

   Menentukan epsilon yang optimal dapat dilakukan dengan menganalisis k-distance graph, di mana perubahan tajam menunjukkan batas cluster. Kuartil ke-3 membantu menentukan epsilon, sementara MinPts harus disesuaikan berdasarkan kerapatan lokal data.

4. Jika analisis post-clustering mengungkapkan overlap signifikan antara cluster "high-value customers" dan "bulk buyers" berdasarkan total pengeluaran, bagaimana teknik semi-supervised (contoh: constrained clustering) atau integrasi metric learning (Mahalanobis distance) dapat memperbaiki pemisahan cluster? Jelaskan tantangan dalam mempertahankan interpretabilitas bisnis saat menggunakan pendekatan non-Euclidean!

   Jawaban:

   Jika ada tumpang tindih antara cluster, semi-supervised clustering atau metric learning dengan Mahalanobis distance bisa membantu memperbaiki pemisahan cluster. Tantangan utama adalah interpretabilitas bisnis yang menurun dengan metrik non-Euclidean.

5. Bagaimana merancang temporal features dari InvoiceDate (misal: hari dalam seminggu, jam pembelian) untuk mengidentifikasi pola pembelian periodik (seperti transaksi pagi vs. malam)? Jelaskan risiko data leakage jika menggunakan agregasi temporal (misal: rata-rata pembelian bulanan) tanpa time-based cross-validation, dan mengapa lag features (pembelian 7 hari sebelumnya) dapat memperkenalkan noise pada cluster!

Jawaban:

Fitur seperti hari dalam seminggu dan jam pembelian bisa mengidentifikasi pola pembelian periodik. Namun, penggunaan agregasi temporal tanpa time-based cross-validation bisa menyebabkan data leakage, sementara lag features bisa memperkenalkan noise dalam clustering.