Analisa Clustering Model

Nama: Arif Al Imran

Kelas: TK-45-G05

NIM: 1103210193

Soal:

1. Jika algoritma K-Means menghasilkam nilai silhouette score rendah (0.3) meskipun elbow method menunjukkan K=5 sebagai optimal pada dataset ini, factor apa yang menyebabkan inkonsistensi ini? Bagaimana strategi validasi alternatif (misal: analisis gap statistic atauu validasi stabilitas cluster via bootstrapping) dapat mengatasi masalah ini, dan mengapa distribusi data non-spherical menjadi akar masalahnya?
2. Dalam dataset dengan campuran fitur numerik (Quantity, UnitPrice) dan kategorikal high-cardinality (Description), metode preprocessing apa yang efektif untuk menyelaraskan skala dna merepresentasikan fitur teks sebelum clustering? Jelaskan risiko mengguunakan One-Hot Encoding untuuk Description, dan mengapa Teknik seperti TF-IDF atau embedding berdimensi rendah (UMAP) lebih robust untuk mempertahankan struktur cluster
3. Hasil clustering dengan DBSCAN sangat sensitive terhadap parameter epsilon. Bagaimana menetukan 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!
4. Jika analisis post-clustering mengungkapkan overlap signifikan antara cluster "high-value customer" 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-Ecuclidean!
5. Bagaimana merancang temporal features dari InvoiceDate (misal: hari dalam seminggu. Jam pembelian) untuk mengidentifikasi pola pembelian periodic (seperti transaksi pagi vs malam)? Jelaskan risiko data leakage jika menggunakan agregasi temporal (misal: rata-erata pembellian bulanan) tanpa time-based cross-validation, dan mengapa lag features (pembelian 7 hari sebelumnya) dapat memperkenalkan noise pada cluster!

Jawab:

1. Inkonsistensi antara Silhouette Score dan Elbow Method

Nilai silhouette score rendah (0.3) meskipun elbow method menunjukkan K=5 sebagai optimal menunjukkan adanya inkonsistensi dalam struktur cluster.

Faktor Penyebab Utama:

- K-Means mengasumsikan cluster berbentuk sferis dengan ukuran seragam
- Data sebenarnya memiliki bentuk non-sferis, memanjang, atau tanpa batas yang jelas
- Jarak Euclidean tidak mampu menangkap perbedaan antar cluster secara tepat
- Banyak data berada di batas antar cluster atau terjadi overlapping
- Outlier terdistribusi tidak merata, menurunkan silhouette score

Strategi Validasi Alternatif:

- **Gap Statistic:** Mengukur dispersion dalam cluster dan membandingkannya dengan distribusi acak, mengungkap apakah pemisahan cluster benar-benar signifikan
- **Validasi Bootstrapping:** Mengulang clustering pada berbagai sampel data untuk menilai konsistensi pembagian cluster, mendeteksi ketidakstabilan struktur

Akar Masalah - Distribusi Data Non-Sferis:

- K-Means berbasis jarak Euclidean tidak mampu menangkap variansi berbeda pada berbagai arah
- Meskipun secara global inersia menurun pada K=5, perbedaan dalam cluster tidak terlihat jelas
- Bentuk geometris asli cluster tidak sesuai dengan asumsi dasar algoritma K-Means

2. Preprocessing untuk Dataset dengan Fitur Campuran

Untuk dataset dengan fitur numerik (Quantity, UnitPrice) dan kategorikal high-cardinality (Description), diperlukan pendekatan preprocessing yang tepat.

Metode Preprocessing Efektif:

- **Fitur Numerik:** Normalisasi atau standardisasi untuk menyeragamkan skala
- **Fitur Kategorikal (Description):**
  - TF-IDF: Mengubah teks menjadi representasi numerik dengan pembobotan berdasarkan pentingnya kata
  - Word embeddings (Word2Vec, GloVe) dengan reduksi dimensi UMAP untuk representasi semantik yang padat

Risiko One-Hot Encoding untuk Description:

- Menghasilkan vektor sangat panjang dan sparse karena tingginya jumlah kategori
- Memunculkan curse of dimensionality yang membuat perhitungan jarak tidak efektif
- Memboroskan memori dan meningkatkan waktu komputasi

- Berisiko overfitting karena dimensi terlalu tinggi

Keunggulan TF-IDF dan Embeddings:

- Menyediakan representasi berbasis makna yang lebih robust
- Mengurangi dimensi secara signifikan tanpa kehilangan informasi penting
- Mempertahankan kedekatan semantik antar item yang mirip
- Meningkatkan kualitas clustering dengan mempertahankan struktur data yang bermakna

3. Penentuan Parameter Optimal DBSCAN untuk Data Tidak Seimbang

DBSCAN sensitif terhadap parameter epsilon (ε), terutama pada data transaksi tidak seimbang seperti 90% pelanggan dari UK.

Metode Penentuan Epsilon Secara Adaptif:

- **K-distance Graph:** Plot jarak setiap titik data ke tetangga ke-k (k = MinPts)
- Identifikasi "titik siku" (titik belok) dalam grafik sebagai kandidat nilai epsilon
- Titik siku menandai transisi dari titik dalam cluster padat ke titik noise/outlier

Peran Kuartil Ke-3 dalam Automasi Parameter:

- Menggunakan kuartil ke-3 dari distribusi jarak k-tetangga sebagai dasar penentuan epsilon
- Mengatasi variabilitas kepadatan dengan pendekatan statistik yang adaptif
- Membantu mengotomatisasi pemilihan epsilon tanpa inspeksi visual

Penyesuaian MinPts Berdasarkan Kerapatan Regional:

- Daerah dengan kepadatan berbeda memerlukan nilai MinPts yang berbeda
- Cluster padat memerlukan MinPts lebih tinggi untuk menghindari chaining effect
- Area sparse memerlukan MinPts lebih rendah untuk mendeteksi cluster kecil
- Pendekatan adaptive DBSCAN dapat menerapkan parameter yang bervariasi secara spasial

4. Mengatasi Overlap Cluster dengan Teknik Lanjutan

Saat terjadi overlap signifikan antara cluster "high-value customer" dan "bulk-buyers", teknik lanjutan dapat membantu memperbaiki pemisahan.

Teknik Semi-Supervised (Constrained Clustering):

- Memasukkan informasi domain sebagai must-link atau cannot-link constraints
- Mengarahkan algoritma clustering untuk memperhatikan batasan bisnis yang relevan
- Membantu mengurangi overlap dengan memanfaatkan pengetahuan domain

Integrasi Metric Learning (Mahalanobis Distance):

- Mempelajari cara optimal untuk menimbang fitur berdasarkan relevansinya
- Menghasilkan jarak yang lebih merefleksikan perbedaan penting antar cluster
- Mampu menangkap struktur korelasi dalam data yang tidak terlihat dengan jarak Euclidean

Tantangan Interpretabilitas dengan Pendekatan Non-Euclidean:

- Jarak yang dihasilkan tidak mudah dipahami secara intuitif
- Sulit menjelaskan kepada stakeholder bisnis mengapa dua entitas dikelompokkan bersama
- Kompleksitas metrik pembelajaran bisa mengaburkan insight bisnis yang sederhana
- Memerlukan trade-off antara akurasi teknis dan kemudahan interpretasi bisnis

5. Perancangan Temporal Features untuk Analisis Pola Pembelian

Untuk mengidentifikasi pola pembelian periodik dari atribut InvoiceDate, perlu dirancang fitur temporal yang tepat.

Perancangan Fitur Temporal:

- **Hari dalam Seminggu:** Mengidentifikasi pola berbeda antara hari kerja vs akhir pekan
- **Jam Pembelian:** Mengelompokkan transaksi berdasarkan waktu (pagi, siang, sore, malam)
- **Seasonality:** Mengekstrak pola musiman (bulanan, kuartalan) dari data
- **Interval antar Pembelian:** Mengukur frekuensi dan keteraturan transaksi pelanggan

Risiko Data Leakage dalam Agregasi Temporal:

- Jika agregasi (seperti rata-rata pembelian bulanan) dilakukan tanpa time-based cross-validation
- Informasi masa depan bisa masuk ke model, menyebabkan overfitting
- Evaluasi performa menjadi tidak realistis karena menggunakan informasi yang seharusnya tidak tersedia
- Menghasilkan model dengan kemampuan generalisasi rendah untuk data baru

Masalah dengan Lag Features:

- Fitur lag (seperti pembelian 7 hari sebelumnya) dapat memperkenalkan noise jika pola pembelian fluktuatif
- Tidak semua pelanggan memiliki pola pembelian yang konsisten
- Musiman jangka pendek dapat mengaburkan tren jangka panjang yang lebih penting
- Validasi berbasis waktu menjadi krusial untuk memastikan model belajar dari informasi yang seharusnya tersedia, menghindari distorsi oleh fluktuasi acak