Bintang Nabiil Lukman
1103223126
TK-46-06

Analisa – ClusteringUTS

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:

   Ketidaksesuaian antara hasil elbow method dan nilai silhouette score yang rendah (misalnya 0.3) menunjukkan bahwa meskipun variansi inersia berkurang optimal pada $K=5$, kualitas pemisahan cluster tetap buruk. Hal ini bisa terjadi karena asumsi dasar K-Means adalah bahwa data terdistribusi secara spherical dan seimbang, padahal dalam praktik dataset bisa memiliki bentuk non-spherical (misalnya spiral, elongated, atau cluster dengan kepadatan berbeda). Distribusi seperti ini membuat K-Means sulit menangkap batas alami antar kelompok, sehingga anggota cluster menjadi tumpang tindih dan nilai silhouette score menurun.

   Sebagai strategi validasi alternatif, metode seperti Gap Statistic membandingkan within-cluster dispersion dengan distribusi acak sebagai baseline, sehingga dapat mengidentifikasi K optimal yang lebih sesuai dengan struktur data alami. Selain itu, bootstrapping stabilitas cluster (mengukur seberapa sering data point tetap dalam cluster yang sama saat dilakukan sampling ulang) memberikan insight apakah cluster benar-benar konsisten atau hanya artefak dari noise atau struktur data yang tidak cocok dengan K-Means. Kombinasi dari metode ini memberikan validasi yang lebih robust dan data-driven, mengurangi ketergantungan pada visual elbow curve saja. Dengan demikian, untuk data yang tidak berbentuk bulat atau memiliki variansi antar cluster yang tinggi, pendekatan ini lebih akurat dalam mencerminkan struktur sebenarnya.

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:

Untuk dataset yang mencampurkan fitur numerik seperti Quantity dan UnitPrice dengan fitur kategorikal high-cardinality seperti Description (misalnya ribuan jenis produk), preprocessing harus menyelaraskan skala sekaligus merepresentasikan informasi semantik dari teks. Fitur numerik sebaiknya dinormalisasi (misal: StandardScaler atau MinMaxScaler) agar tidak mendominasi jarak dalam ruang fitur saat clustering.

Untuk Description, penggunaan One-Hot Encoding sangat tidak disarankan karena dapat menghasilkan matriks yang sangat besar dan sparse, yang pada gilirannya memperburuk performa dan memperbesar jarak antar titik (curse of dimensionality). Sebaliknya, teknik seperti TF-IDF memberikan representasi teks yang lebih informatif dengan menekankan kata unik yang membedakan produk. Bahkan, teknik lanjutan seperti dimensionality reduction (misalnya UMAP, PCA, atau embedding seperti Word2Vec/BERT) dapat menangkap struktur semantik dan distribusi yang lebih padat, sehingga lebih sesuai untuk clustering. Teknik ini membantu menjaga struktur global dan lokal dari data, yang penting untuk deteksi pola atau segmentasi pelanggan berdasarkan produk yang mereka beli.

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:

DBSCAN adalah algoritma yang efektif untuk data dengan cluster padat dan noise, tetapi sangat sensitif terhadap parameter epsilon (ε). Salah satu pendekatan adaptif untuk memilih epsilon adalah melalui analisis k-distance graph. Dalam metode ini, kita menghitung jarak ke k-nearest neighbor (misalnya k = MinPts - 1) untuk setiap titik, lalu menyortir dan memplotnya. Titik "elbow" atau kenaikan tajam dalam grafik menunjukkan nilai epsilon optimal, karena mencerminkan perbedaan signifikan antara titik dalam cluster padat dan titik outlier.

Untuk mengautomasi, epsilon bisa dipilih berdasarkan kuartil ke-3 (Q3) atau percentile ke-95 dari distribusi k-distance, yang membantu menyeimbangkan sensitivitas terhadap outlier sambil tetap menjaga struktur utama. Nilai MinPts harus disesuaikan berdasarkan kerapatan lokal data; misalnya, pada data transaksi di mana 90% pelanggan berasal dari UK, region ini bisa memiliki kerapatan tinggi dan memerlukan MinPts lebih

besar dibanding region minoritas. Dengan demikian, parameter DBSCAN harus disetel secara berbasis distribusi lokal, bukan secara global, agar pemisahan cluster tetap valid pada data yang tidak seimbang.

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:

   Ketika hasil clustering menunjukkan overlap signifikan antara segmen seperti “high-value customers” dan “bulk buyers”, ini menandakan bahwa fitur seperti total pengeluaran tidak cukup membedakan perilaku mereka. Dalam kasus seperti ini, metode semi-supervised clustering seperti constrained clustering (misalnya: COP-KMeans atau seeded KMeans) dapat membantu. Dengan memasukkan prior knowledge berupa must-link (dua pelanggan harus dalam satu cluster) dan cannot-link (dua pelanggan tidak boleh dalam satu cluster), algoritma dapat diarahkan untuk lebih akurat memisahkan segmen yang serupa namun memiliki nuansa berbeda.

   Selain itu, metric learning seperti Mahalanobis distance memungkinkan pembelajaran jarak antar data yang mempertimbangkan korelasi antar fitur, sehingga bisa memisahkan cluster yang secara Euclidean tampak tumpang tindih. Namun, pendekatan ini memperkenalkan tantangan baru: interpretabilitas bisnis menjadi lebih sulit, karena batas antar cluster ditentukan oleh transformasi linier kompleks, bukan berdasarkan fitur asli seperti ‘Quantity’ atau ‘Invoice Frequency’. Oleh karena itu, meskipun akurasi segmentasi meningkat, pengguna bisnis harus diberikan visualisasi atau penjelasan tambahan untuk memahami hasil clustering tersebut.

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:

   Merancang fitur temporal dari InvoiceDate sangat penting untuk mendeteksi pola pembelian periodik. Fitur seperti hari dalam seminggu, jam pembelian, atau apakah

transaksi terjadi pada akhir pekan bisa mengungkap tren seperti preferensi belanja di pagi hari atau pengaruh promosi harian. Transformasi semacam ini sebaiknya direpresentasikan dalam bentuk numerik atau siklik (misal: menggunakan sin/cos untuk hari ke-7), agar kompatibel dengan algoritma clustering.

Namun, perlu kehati-hatian saat menggunakan agregasi seperti rata-rata pembelian bulanan karena hal ini bisa menyebabkan data leakage, yakni model "mengintip" ke masa depan jika tidak menggunakan time-based cross-validation. Validasi harus menjaga urutan waktu agar fitur agregasi benar-benar mencerminkan informasi historis yang valid saat evaluasi.

Sementara itu, penggunaan lag features (misalnya: total pembelian 7 hari sebelumnya) juga harus dipertimbangkan secara hati-hati. Meski berguna untuk menangkap pola musiman, lag dapat memperkenalkan noise, terutama jika frekuensi pembelian pelanggan tidak teratur. Kombinasi waktu belanja dan frekuensi historis memang memberi nilai tambah, tetapi harus dikontrol secara eksplisit untuk menjaga kestabilan hasil clustering.