Nama : Faiz Hibatullah
Kelas : TK-45-G05
NIM : 1103210172

Analisa Clustering Model

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. Pada kasus pertama, meskipun elbow method menunjukkan K optimal = 5, silhouette score yang rendah (0.3) mengindikasikan masalah. Hal utamanya adalah karena algoritma K-Means mengasumsikan cluster berbentuk sferis dengan ukuran yang relatif seragam. Bila bentuk data sebenarnya tidak sferis, misalnya memanjang atau tidak memiliki batas yang jelas, jarak Euclidean yang digunakan tidak mampu menangkap perbedaan antar cluster secara tepat. Ini menyebabkan:
   - Banyak data berada di batas antar cluster atau terjadi overlapping.
   - Outlier yang terdistribusi secara tidak merata turut berkontribusi menurunkan silhouette score.

   Untuk mengatasi inkonsistensi ini, beberapa strategi validasi alternatif bisa dipakai:
   - Gap Statistic: Metode ini menghitung dispersion dalam cluster dan membandingkannya dengan data yang dihasilkan secara acak. Dengan begitu, gap statistic bisa mengungkap apakah pemisahan cluster benar-benar signifikan atau hanya hasil kebetulan.

- Validasi Bootstrapping: Dengan mengulang clustering pada sampel-sampel data yang berbeda, kita bisa menilai apakah pembagian cluster konsisten. Jika cluster tidak stabil antar sampel, berarti struktur cluster belum optimal.

Distribusi data non-sferis merupakan akar masalah karena K-Means yang berbasis Euclidean tidak bisa menangkap variansi berbeda pada tiap arah, sehingga meskipun secara global inersia bisa menurun, perbedaan dalam cluster tidak terlihat jelas.

2. Dalam dataset yang mengandung fitur numerik seperti Quantity dan UnitPrice serta fitur kategorikal berdimensi tinggi seperti Description, sangat penting untuk menyamakan skala antar fitur dan memberikan representasi yang bermakna kepada teks sebelum clustering dilakukan. Fitur numerik sebaiknya dinormalisasi atau distandarisasi sehingga perhitungan jarak antar data tidak didominasi oleh perbedaan skala. Sementara itu, untuk kolom Description, metode seperti TF-IDF dapat mengubah teks menjadi representasi numerik dengan memberikan bobot berdasarkan pentingnya kata dalam konteks dokumen. Selain itu, teknik embedding menggunakan model seperti Word2Vec atau GloVe, yang kemudian direduksi dimensinya dengan pendekatan seperti UMAP, dapat menghasilkan representasi yang lebih padat dan bermakna secara semantik. Penggunaan one-hot encoding pada fitur Description bisa menghasilkan vektor yang sangat panjang dan sparse karena tingginya jumlah kategori, sehingga menimbulkan masalah terkait dengan curse of dimensionality, yang akhirnya membuat perhitungan jarak tidak efektif.

3. Dalam penerapan DBSCAN pada data transaksi yang tidak seimbang—misalnya, ketika mayoritas data berasal dari pelanggan dari UK—penentuan parameter epsilon (ε) sangat krusial. Salah satu cara adaptif untuk menetapkan nilai ε adalah dengan membuat k-distance graph, di mana setiap titik data diplot berdasarkan jarak ke tetangga ke-k (dengan k biasanya sesuai dengan nilai MinPts). Grafik tersebut biasanya menunjukkan "titik siku" atau perubahan tajam yang bisa dijadikan acuan untuk memilih epsilon yang tepat. Selain itu, statistik seperti kuartil ketiga dari distribusi jarak k-tetangga dapat digunakan untuk mengotomasi pemilihan nilai ε secara lebih adaptif terhadap variabilitas kepadatan data. Karena kepadatan data bisa berbeda pada masing-masing wilayah, penting pula untuk menyesuaikan nilai MinPts secara regional agar cluster padat maupun area dengan kepadatan rendah dapat dipisahkan secara konsisten.

4. Ketika analisis pasca-clustering menunjukkan terdapat overlap signifikan antara cluster “high-value customer” dan “bulk-buyers” (misalnya, pada basis total pengeluaran), maka pendekatan berikut bisa membantu:
   - Constrained Clustering (Semi-Supervised): Dengan memasukkan informasi domain seperti must-link atau cannot-link constraints, clustering dapat diarahkan agar memperhatikan batas-batas yang menurut bisnis memang harus dipisah. Ini membantu mengurangi overlap pada kedua cluster tersebut.
   - Metric Learning dengan Mahalanobis Distance: Pendekatan ini mempelajari cara terbaik untuk menimbang fitur berdasarkan relevansi antar data. Dengan demikian, jarak yang dihitung pun lebih mencerminkan perbedaan penting antar cluster.

   Namun, tantangan utama dari metode non-Euclidean seperti ini adalah:
   - Interpretabilitas:
     Jarak yang dihasilkan tidak langsung mudah dimengerti secara intuitif, yang bisa

menyulitkan komunikasi hasil kepada pihak bisnis yang mengharapkan hasil yang transparan.

Sehingga, meskipun secara teknis pendekatan ini dapat memperbaiki pemisahan cluster, harus ada keseimbangan antara akurasi dan kemudahan interpretasi.

5. Untuk mengoptimalkan pemanfaatan atribut InvoiceDate dalam konteks identifikasi pola pembelian periodik, kita dapat merancang fitur temporal seperti:
   - Hari dalam Seminggu: Menunjukkan apakah pembelian terjadi pada hari kerja atau akhir pekan.
   - Jam Pembelian: Menandai waktu transaksi (pagi, siang, sore, malam) untuk melihat kecenderungan pola pembelian dalam kurun waktu harian.

   Namun, perlu diingat bahwa ketika melakukan agregasi temporal, misalnya menghitung rata-rata pembelian bulanan, harus hati-hati untuk menghindari data leakage. Jika agregasi dilakukan tanpa time-based cross-validation, informasi dari masa depan bisa secara tidak sengaja masuk ke dalam model, sehingga mengakibatkan overfitting. Penggunaan lag features, seperti mengambil data transaksi 7 hari sebelumnya, bisa membantu menyorot tren jangka pendek, tetapi hal ini juga berisiko menambahkan noise apabila pola pembelian sangat fluktuatif. Oleh karena itu, penerapan validasi berbasis waktu menjadi sangat penting agar model hanya belajar dari informasi yang seharusnya tersedia, sehingga pola yang diekstrak mencerminkan tren jangka panjang yang sebenarnya tanpa terdistorsi oleh fluktuasi acak.