

TUGAS GROUP PROJECT (TGP) #2

Mata Kuliah: Analitika Data & Diagnostik

Dosen: Prof. Arif Djunaidy

Tugas 2 – Analisis Klaster dan Analisis Penciran

A. Analisis Klaster

Judul Tugas:

Perbandingan Kinerja dan Validitas Algoritme Klustering pada Data Kompleks dengan Distribusi Tidak Seragam

Tujuan:

- Menerapkan algoritme klaster partisional, hierarkis, dan berbasis densitas (K-Means, Agglomerative, dan DBSCAN).
- Mengevaluasi hasil klaster menggunakan berbagai metrik validitas internal dan eksternal.
- Menunjukkan pengaruh parameter (misal: jumlah klaster, eps, minPts) terhadap struktur klaster yang terbentuk.

Dataset:

1. Kelompok 1 s.d. 4

Mall Customers Dataset (Kaggle) — data segmentasi pelanggan berdasarkan perilaku belanja.

Sumber data: https://www.kaggle.com/datasets/shwetabh123/mall-customers?utm_source=chatgpt.com

2. Kelompok 5 s.d. 8

Wholesale Customer Data (UCI) — data pembelian pelanggan B2B.

Sumber data: <https://archive.ics.uci.edu/ml/datasets/wholesale+customers>

3. Kelompok 9 s.d. 12

Credit Card Users Dataset (Kaggle) — perilaku transaksi pengguna kartu kredit.

Sumber data: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud?utm_source=chatgpt.com

Tugas dan Langkah Analisis:

1. Eksplorasi Data:

- Lakukan normalisasi dan analisis distribusi variabel.
- Visualisasikan data dalam 2D/3D dengan PCA atau t-SNE.

2. Implementasi Klustering:

- Jalankan **K-Means**, **Hierarchical (Ward's & Complete Linkage)**, dan **DBSCAN**.
- Gunakan parameter bervariasi untuk menunjukkan perubahan struktur klaster.

3. Evaluasi dan Validasi Klaster:

- Gunakan **SSE**, **Silhouette Coefficient**, **Dunn Index**, dan **Entropy/Purity**.
- Tampilkan **similarity matrix**

4. Analisis Hasil:

- Jelaskan bentuk klaster yang dihasilkan (jumlah, kepadatan, overlapping).
- Bandingkan kekuatan dan kelemahan setiap algoritme pada dataset yang sama.
- Kaitkan hasil dengan **teori jenis klaster** (well-separated, density-based, contiguity-based).

5. Luaran yang Dikumpulkan:

- Laporan analisis (PDF) dengan narasi hasil visual dan tabel evaluasi.
- Kode Python (Jupyter Notebook).
- Rekaman Presentasi kelompok (maks. 15 menit).

6. Kriteria Penilaian:

| Aspek | Bobot |
|------------------------------------------|-------|
| Kualitas eksplorasi & prapemrisesan data | 20% |
| Ketepatan penerapan 3 algoritme klaster | 30% |
| Analisis validitas & evaluasi hasil | 30% |
| Kedalaman interpretasi & visualisasi | 20% |

B. Analisis Pencilan (Outlier Analysis)

Judul Tugas:

Deteksi Anomali Multidimensi Menggunakan Pendekatan Statistik, Jarak, dan Densitas

Tujuan Tugas:

- Mengimplementasikan berbagai pendekatan deteksi pencilan seperti **Grubbs' Test**, **k-NN Distance**, dan **Local Outlier Factor (LOF)**.
- Membandingkan efektivitas metode statistik, berbasis jarak, dan berbasis densitas dalam mendekripsi anomali.
- Menganalisis trade-off antara **true detection rate** dan **false alarm rate** (Base Rate Fallacy)

Dataset:

1. Kelompok 1 s.d. 4

Network Intrusion Detection Dataset (KDD99 atau NSL-KDD).

Sumber data:

https://www.kaggle.com/datasets/hassan06/nslkdd?utm_source=chatgpt.com

2. Kelompok 5 s.d. 8

Credit Card Fraud Dataset (Kaggle).

Sumber data: https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud?utm_source=chatgpt.com

3. Kelompok 9 s.d. 12

Sensor Fault Detection Dataset (NASA/CMAPSS subset).

Sumber data: https://data.nasa.gov/dataset/cmapss-jet-engine-simulated-data/resource/5224bcd1-ad61-490b-93b9-2817288accb8?utm_source=chatgpt.com

Langkah Analisis:

1. **Eksplorasi dan Persiapan Data:**
 - Analisis distribusi atribut, deteksi skewness, dan normalisasi.
 - Visualisasikan data 2D/3D untuk menunjukkan potensi pencilan.
2. **Implementasi Tiga Pendekatan:**
 - **Statistik:** Uji Grubbs atau metode Likelihood-based (mixture model M & A).
 - **Distance-based:** gunakan k-NN atau Mahalanobis distance.
 - **Density-based:** gunakan **LOF (Local Outlier Factor)**.
3. **Evaluasi Hasil:**
 - Hitung metrik **Precision, Recall, F1, AUC ROC** untuk mendeteksi outlier.
 - Analisis **false alarm rate (FAR)** dan **detection rate (DR)** seperti dijelaskan dalam kuliah
4. **Analisis Komparatif:**
 - Tampilkan perbedaan jumlah dan distribusi pencilan hasil tiap metode.
 - Diskusikan skenario di mana metode statistik gagal namun LOF berhasil, dan sebaliknya.
5. **Luaran:**
 - Laporan (PDF) lengkap berisi teori singkat, eksperimen, hasil evaluasi, dan analisis.
 - Kode Python (.ipynb).
 - Visualisasi interaktif (matplotlib/plotly).
6. **Kriteria Penilaian:**

| Aspek | Bobot |
|-----------------------------------------------|-------|
| Pemilihan & penyipam dataset | 15% |
| Implementasi 3 metode deteksi | 35% |
| Evaluasi metrik & visualisasi | 25% |
| Analisis interpretatif (base rate, trade-off) | 25% |

C. Catatan Tambahan:

- (1) Semua analisis **harus dapat direproduksi** dengan kode Python terbuka (tanpa library komersial).
- (2) Gunakan **matplotlib, seaborn, scikit-learn, dan scipy**.
- (3) Diskusikan hasil menggunakan pendekatan **analitik dan konseptual**, sesuai dengan teori yang dijelaskan dalam kuliah
- (4) Hasil akhir sebaiknya menunjukkan **pemahaman mendalam antara teori dan praktik**, bukan sekadar menjalankan algoritme.

(5) TGP #2 akan memberikan **kontribusi nilai sebesar 25%** dari keseluruhan nilai mata kuliah

- TGP #1 (Preprocessing dan EDA) 20%
- **TGP #2 (Clustering Analysis) 25%**
- TGP #3 (Association Analysis) 15%
- Evaluasi Tengah Semester 20%
- Evaluasi Akhir Semester 20%

-----oooOooo-----