Analisis Perbandingan Metode Clustering K-Means dan DBSCAN untuk Pengelompokan Mood Lagu pada Dataset Spotify

Nama & Pembagian Tugas TI 3G

1.	**VIDI JOSHUBZKY SAVIOLA (2341720112)**	**Clustering K_Means**
2.	**RANGGA DWI SAPUTRA (2341720248)**	**Clustering DBSCAN**
3.	**ARYO ADI PUTRO (2341720112)**	**Orange Miner**
4.	**AHMAD NAUFAL ILHAM (2341720047)**	**Rapidminer**
5.	**SAKA NABIL (2341720108)**	

Abstrak : Penelitian ini bertujuan untuk membandingkan performa dua algoritma clustering, yaitu K-Means dan DBSCAN, dalam mengelompokkan lagu berdasarkan ritme mood menggunakan dataset dari Kaggle. Fitur audio diekstraksi dan diproses menggunakan teknik standardisasi dan *Principal Component Analysis (PCA)* untuk reduksi dimensi. Algoritma K-Means diimplementasikan dengan pendekatan *Spherical K-Means* dan berhasil mengidentifikasi 3 klaster yang seimbang dengan nilai *Silhouette Score* sebesar 0.2755. Di sisi lain, algoritma DBSCAN dengan parameter `eps=2.8` dan `min_samples=25` hanya mampu membentuk satu klaster besar dan mengklasifikasikan 197 data sebagai *noise*, sehingga tidak memungkinkan untuk dilakukan evaluasi menggunakan *Silhouette Score*. Berdasarkan perbandingan hasil visualisasi, metrik evaluasi, dan interpretabilitas klaster, penelitian ini menyimpulkan bahwa metode K-Means lebih superior untuk kasus pengelompokan mood lagu pada dataset ini karena mampu menghasilkan struktur klaster yang lebih jelas dan bermakna.

1. Introduction

Musik telah menjadi bagian tak terpisahkan dari kehidupan sehari-hari, dengan kemampuannya untuk mempengaruhi suasana hati (mood) pendengar. Platform streaming seperti Spotify memiliki jutaan lagu yang dapat dikelompokkan berdasarkan berbagai karakteristik, salah satunya adalah ritme mood. Pengelompokan ini dapat meningkatkan pengalaman pengguna dengan menyediakan rekomendasi lagu yang sesuai.

Clustering atau pengklasteran adalah salah satu teknik unsupervised learning dalam data mining yang bertujuan untuk mengelompokkan data tanpa adanya label. Dua metode clustering yang populer adalah K-Means dan DBSCAN.

K-Means adalah algoritma berbasis centroid yang membagi data ke dalam K jumlah klaster yang telah ditentukan sebelumnya. Algoritma ini efisien untuk data berukuran besar namun sensitif terhadap penentuan jumlah K dan bentuk klaster yang non-sferis.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) adalah algoritma berbasis kepadatan yang dapat menemukan klaster dengan bentuk arbitrer dan mampu mengidentifikasi noise atau outlier.

Penelitian ini akan menganalisis dan membandingkan efektivitas kedua metode tersebut dalam mengelompokkan lagu berdasarkan ritme mood dari dataset "Spotify 12M Songs" yang tersedia di Kaggle.

2. Metode

2.1. Dataset Dataset yang digunakan adalah "Spotify 12M Songs" dari Kaggle, dengan file tracks_features.csv. Untuk analisis ini, diambil sampel acak sebanyak 5.000 lagu untuk efisiensi komputasi.

2.2. Pra-pemrosesan Data & Seleksi Fitur Fitur-fitur audio yang relevan untuk analisis mood dipilih, antara lain:

a) Pembersihan Data: Menghapus data duplikat dan menangani nilai yang hilang.

b)

c) Feature Engineering: Fitur duration_ms diubah ke skala logaritmik (log_duration_s) dan fitur key diubah menjadi komponen siklis (key_sin, key_cos) untuk representasi yang lebih baik.

d)

e) Scaling: Semua fitur numerik distandarisasi menggunakan StandardScaler untuk menyamakan skala.

f)

g) Reduksi Dimensi: Principal Component Analysis (PCA) diterapkan untuk mereduksi fitur menjadi 12 komponen utama yang digunakan dalam pemodelan

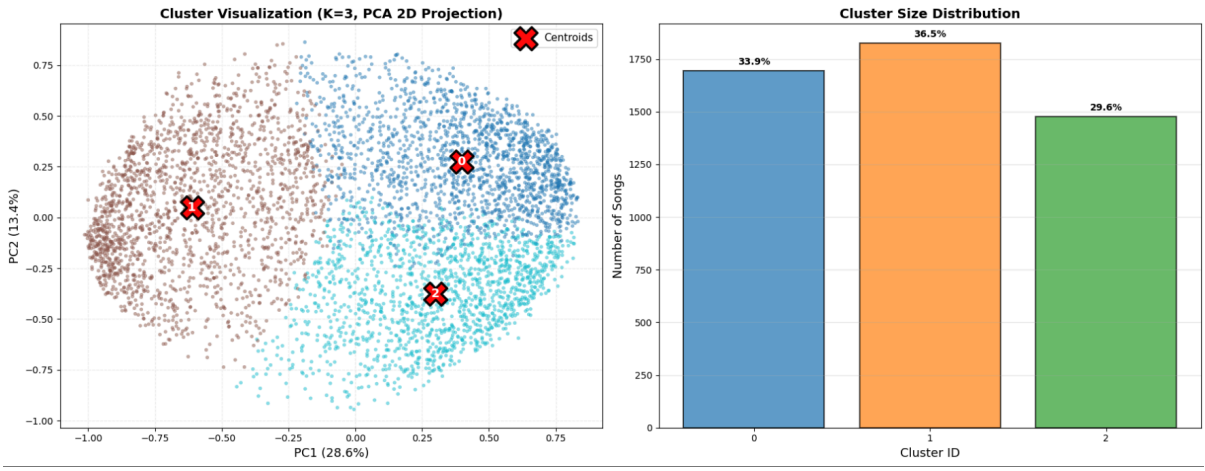
2.3. Implementasi Clustering

a) K-Means: Digunakan pendekatan Spherical K-Means, di mana data dinormalisasi menggunakan L2-norm. Jumlah klaster optimal (K) ditentukan menggunakan kombinasi Elbow Method dan Silhouette Score, di mana K=3 dipilih sebagai nilai terbaik.

b) DBSCAN: Algoritma DBSCAN diimplementasikan dengan parameter yang ditentukan melalui eksperimen, yaitu epsilon (eps) = 2.8 dan min_samples = 25. Parameter eps merepresentasikan radius lingkungan pencarian, sementara min_samples adalah jumlah minimum titik yang diperlukan untuk membentuk sebuah klaster padat.

3. Result + Discuss

3.1. Hasil K-Means Analisis K-Means berhasil mengelompokkan data ke dalam 3 klaster. Visualisasi hasil proyeksi PCA dan distribusi ukuran klaster ditunjukkan pada Gambar 1.



3.1.1. Distribusi data pada setiap klaster adalah sebagai berikut:

a) Klaster 0: 1.696 lagu (33.9%)
b) Klaster 1: 1.826 lagu (36.5%)
c) Klaster 2: 1.478 lagu (29.6%)

3.1.2. Evaluasi kualitas klaster menunjukkan hasil sebagai berikut:

a) Silhouette Score: 0.2755
b) Davies-Bouldin Index: 2.161
c) Calinski-Harabasz Score: 1034.83

3.1.3. Profil klaster menunjukkan karakteristik mood yang berbeda:

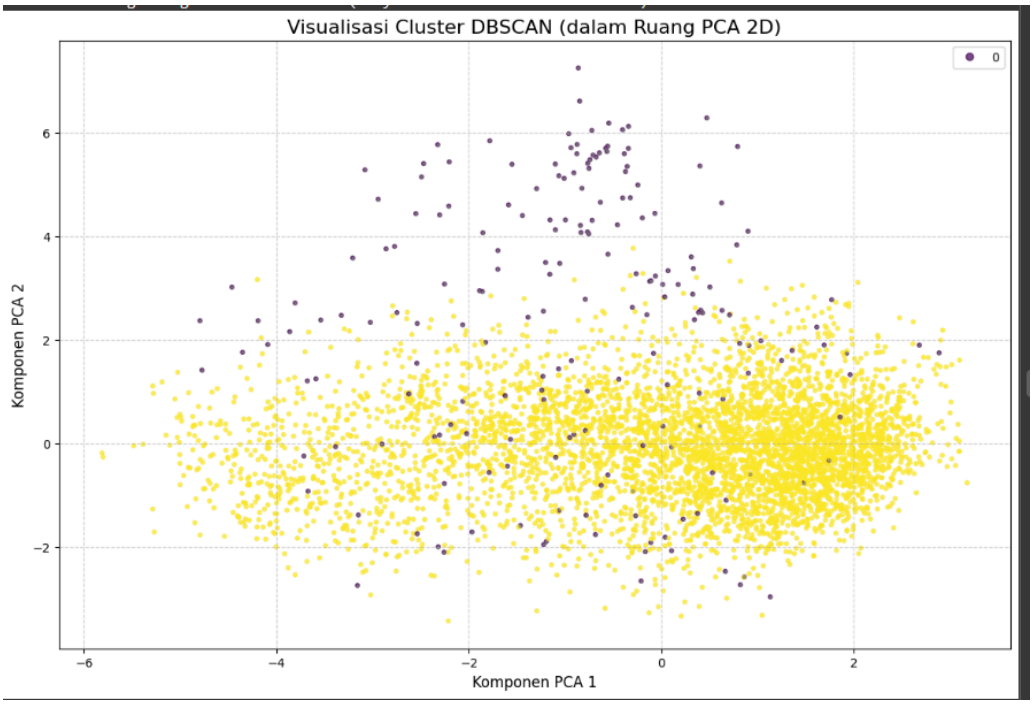
a) Klaster 0 & 2 (Energetic/Mainstream): Ditandai dengan energy, valence, dan danceability yang tinggi.

b) Klaster 1 (Acoustic/Calm): Ditandai dengan acousticness dan instrumentalness yang sangat tinggi, namun energy yang rendah.

3.2. Hasil DBSCAN Implementasi DBSCAN dengan parameter yang telah ditentukan menghasilkan pengelompokan yang sangat berbeda. Hasilnya adalah sebagai berikut:

a) Jumlah Klaster Terbentuk: 1 (Klaster '0')
b) Jumlah Titik Noise (Outliers): 197

Karena DBSCAN hanya menghasilkan satu klaster yang valid (di luar noise), perhitungan Silhouette Score tidak dapat dilakukan, karena metrik ini memerlukan minimal dua klaster untuk membandingkan jarak intra-klaster dan inter-klaster. Visualisasi hasil DBSCAN dapat dilihat pada Gambar 2.



Gambar 2 menunjukkan bahwa mayoritas titik data (ditandai dengan warna kuning, merepresentasikan noise) tersebar tanpa membentuk kepadatan yang jelas, sementara beberapa titik (warna ungu) membentuk satu-satunya klaster yang teridentifikasi (Klaster '0').

3.3. Diskusi Perbandingan antara kedua metode menunjukkan perbedaan performa yang signifikan.

K-Means berhasil mempartisi data ke dalam tiga kelompok yang seimbang secara kuantitas dan dapat diinterpretasikan secara kualitatif sebagai

kelompok lagu dengan mood yang berbeda (misalnya, Acoustic/Calm vs. Energetic). Meskipun nilai silhouette score (0.2755) menunjukkan adanya sedikit tumpang tindih, visualisasi pada Gambar 1 menunjukkan adanya struktur pengelompokan yang jelas.

Sebaliknya, DBSCAN gagal mengidentifikasi struktur klaster yang bermakna dalam data. Hasilnya yang hanya membentuk satu klaster besar dan menganggap 197 titik lainnya sebagai noise mengindikasikan bahwa data lagu ini tidak memiliki klaster-klaster dengan kepadatan yang bervariasi yang dapat dideteksi oleh DBSCAN dengan parameter yang digunakan. Sebagian besar data dianggap memiliki kepadatan yang homogen sehingga digabungkan menjadi satu.

Kegagalan DBSCAN ini juga membuat perbandingan metrik kuantitatif menjadi sulit, karena metrik evaluasi utama seperti Silhouette Score tidak dapat dihitung. Secara visual pun, hasil K-Means jauh lebih informatif dibandingkan DBSCAN.

4. Conclusion

Berdasarkan analisis perbandingan yang telah dilakukan, dapat disimpulkan bahwa metode K-Means memberikan hasil yang lebih baik dan lebih berguna untuk tugas pengelompokan mood lagu pada dataset ini.

K-Means mampu menghasilkan 3 klaster yang terdefinisi dengan baik, seimbang, dan dapat diinterpretasikan, yang sesuai dengan tujuan analisis untuk menemukan kelompok lagu dengan ritme mood yang berbeda. Di sisi lain, DBSCAN tidak berhasil menemukan struktur klaster yang relevan, yang kemungkinan disebabkan oleh distribusi data yang cenderung membentuk klaster sferis tanpa perbedaan kepadatan yang signifikan antar kelompok. Oleh karena itu, untuk kasus ini, K-Means adalah pilihan algoritma yang lebih tepat.

Link:

DBSCAN https://colab.research.google.com/drive/1aHMcTyQ6kOFiTLArBm1edURb2FWBwsbE

K-MEANS : https://colab.research.google.com/drive/1-ylh-CvW45UWsIzC2D5mno58QK5hCJZg