Laporan Proyek Data Mining:

Segmentasi "Vibe" Musik Menggunakan Clustering pada Dataset Spotify

Penulis: Justin Matthew T

NIM: A11.2023.14877

Tanggal: 24 Oktober 2025

1. Tujuan Utama Analisis

Tujuan utama dari analisis ini adalah untuk melakukan segmentasi (clustering) pada katalog lagu Spotify berdasarkan karakteristik audio intrinsiknya. Analisis ini menggunakan gabungan dua teknik *Unsupervised Learning*: **Clustering** sebagai tujuan utama, dan **Dimension Reduction** sebagai metode pendukung.

Secara spesifik, **Dimension Reduction (PCA)** digunakan terlebih dahulu untuk menyederhanakan 7 fitur audio yang kompleks menjadi 2 komponen utama. Setelah itu, **Clustering (K-Means)** diterapkan pada komponen yang telah disederhanakan tersebut untuk menemukan segmen "vibe" lagu yang bermakna dan dapat divisualisasikan.

Manfaat bisnis dari analisis ini adalah untuk meningkatkan pengalaman pengguna dan strategi konten melalui:

1. **Rekomendasi Musik yang Lebih Baik:** Dengan memahami "vibe" lagu (misalnya, "Energi Tinggi" vs. "Akustik Santai"), kita dapat merekomendasikan lagu yang secara musikal mirip, bahkan dari artis atau genre yang berbeda.
2. **Pembuatan Playlist Otomatis:** Memungkinkan pembuatan *playlist* yang lebih cerdas dan kohesif (misalnya, "Playlist Fokus" atau "Playlist Pesta") secara otomatis.
3. **Wawasan Konten:** Memberi pemahaman yang lebih dalam tentang "resep" audio dari berbagai segmen musik di platform kami.

2. Deskripsi Data dan Eksplorasi Awal

Download Dataset : https://www.kaggle.com/datasets/zaheenhamidani/ultimate-spotify-tracks-db

Link Colab :
https://colab.research.google.com/drive/1CPE5_yT7bLUTB5aeUlRQv9JjEWo1tO44#scrollTo=_uRg5IyJRrH8

Dataset yang dipilih adalah "Spotify Tracks DB" (SpotifyFeatures.csv), sebuah database komprehensif yang berisi metadata dan fitur audio yang diekstraksi oleh Spotify untuk ribuan lagu.

- **Atribut Kunci: Analisis ini berfokus pada 7 fitur audio numerik utama yang mendefinisikan "rasa" dari sebuah lagu:**
    - **acousticness:** Ukuran kepercayaan (0.0-1.0) apakah lagu tersebut akustik.
    - **danceability:** Ukuran (0.0-1.0) seberapa cocok sebuah lagu untuk menari, berdasarkan elemen seperti tempo, stabilitas ritme, dan kekuatan ketukan.

- **energy:** Ukuran (0.0-1.0) yang mewakili intensitas dan aktivitas. Lagu yang energik terasa cepat, keras, dan berisik.
- **instrumentalness:** Ukuran (0.0-1.0) yang memprediksi apakah sebuah lagu tidak mengandung vokal.
- **liveness:** Ukuran (0.0-1.0) yang mendeteksi kehadiran penonton dalam rekaman (menunjukkan lagu tersebut direkam *live*).
- **loudness:** Tingkat kenyaringan suara secara keseluruhan dalam desibel (dB), umumnya berkisar antara -60 hingga 0 dB.
- **speechiness:** Ukuran (0.0-1.0) yang mendeteksi keberadaan kata-kata yang diucapkan dalam sebuah lagu. Nilai tinggi menunjukkan konten non-musik (misal: podcast, audiobook).

- **Tujuan:** Menggunakan 7 fitur audio ini untuk mengelompokkan lagu ke dalam segmen-segmen yang homogen secara audio, tanpa menggunakan informasi genre atau artis.
- **Ukuran Data:** Dataset penuh berisi ratusan ribu entri. Untuk efisiensi analisis dan kecepatan pemodelan, sampel acak sebanyak 20.000 lagu diambil untuk proyek ini, karena keterbatasan computing power, yang aslinnya dataset ini terdapat 232.725 Lagu.

3. Persiapan Data (Data Cleaning & Feature Engineering)

Langkah-langkah persiapan data berikut telah dilakukan sebelum pemodelan:

1. **Pembersihan:** Data dibersihkan dari entri duplikat berdasarkan track_id dan baris dengan nilai fitur audio yang hilang (NaN) telah dihapus.
2. **Seleksi Fitur:** Hanya 7 fitur audio yang relevan (disebutkan di atas) yang dipilih sebagai *input* untuk model.
3. **Standardisasi (Scaling):** Ini adalah langkah krusial. Semua 7 fitur audio distandardisasi menggunakan StandardScaler (dari Scikit-learn). Ini mengubah skala data sehingga memiliki rata-rata 0 dan standar deviasi 1, memastikan fitur seperti loudness (rentang -60 s/d 0) tidak mendominasi fitur seperti danceability (rentang 0 s/d 1).

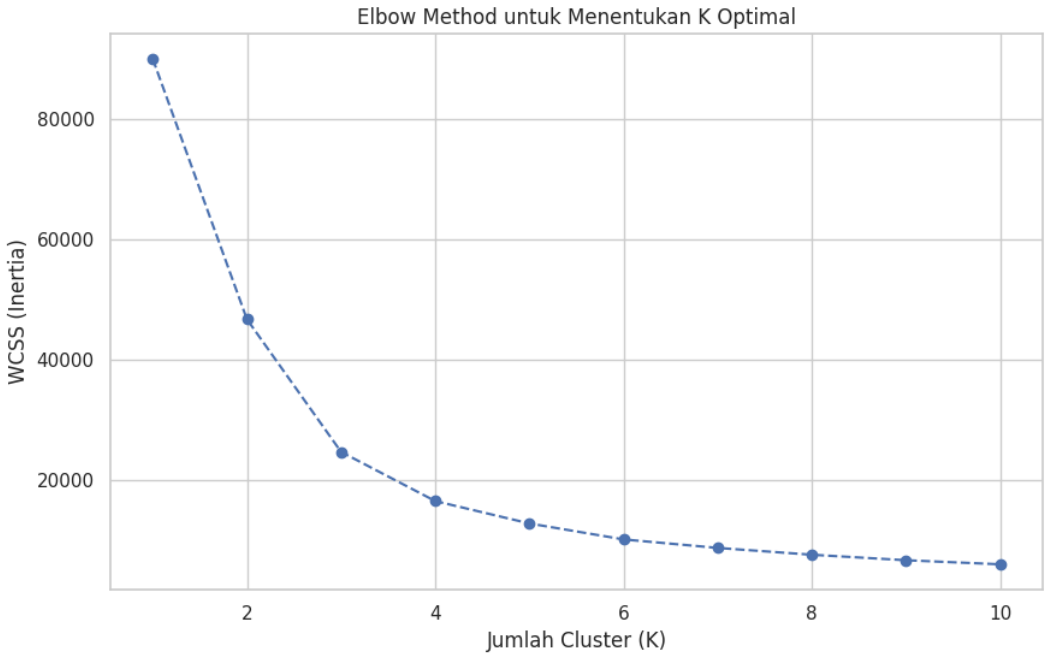
4. Ringkasan Pelatihan Model

Untuk mendapatkan hasil clustering yang optimal dan dapat divisualisasikan, kami menggunakan pendekatan dua langkah:

1. **Dimension Reduction (PCA):** 7 fitur audio yang telah distandardisasi diringkas menjadi 2 "Komponen Utama" (PCA) untuk memungkinkan visualisasi 2D. 2 komponen ini berhasil menangkap **66.30%** dari total variasi data.
2. **Clustering (K-Means):** Model K-Means diterapkan pada 2 komponen PCA tersebut.

Kami melatih **3 variasi model** untuk menemukan jumlah cluster (K) yang optimal, didukung oleh dua metode:

1. **Eksperimen 1 (Elbow Method):** Plot "Elbow Method"menunjukkan bahwa WCSS (inertia) mulai melandai secara signifikan setelah K=4 atau K=5. Ini mengindikasikan bahwa 4, 5, atau 6 adalah jumlah cluster yang potensial.



2. **Eksperimen 2, 3, 4 (Silhouette Score):** Kami melatih model K-Means dengan K=4, K=5, dan K=6 dan membandingkan **Silhouette Score**-nya (metrik yang mengukur seberapa baik cluster terpisah; lebih tinggi lebih baik).

Model (Variasi)	**Silhouette Score**
K-Means dengan K=4	**0.4798 (BEST)**
K-Means dengan K=5	0.4507
K-Means dengan K=6	0.4287

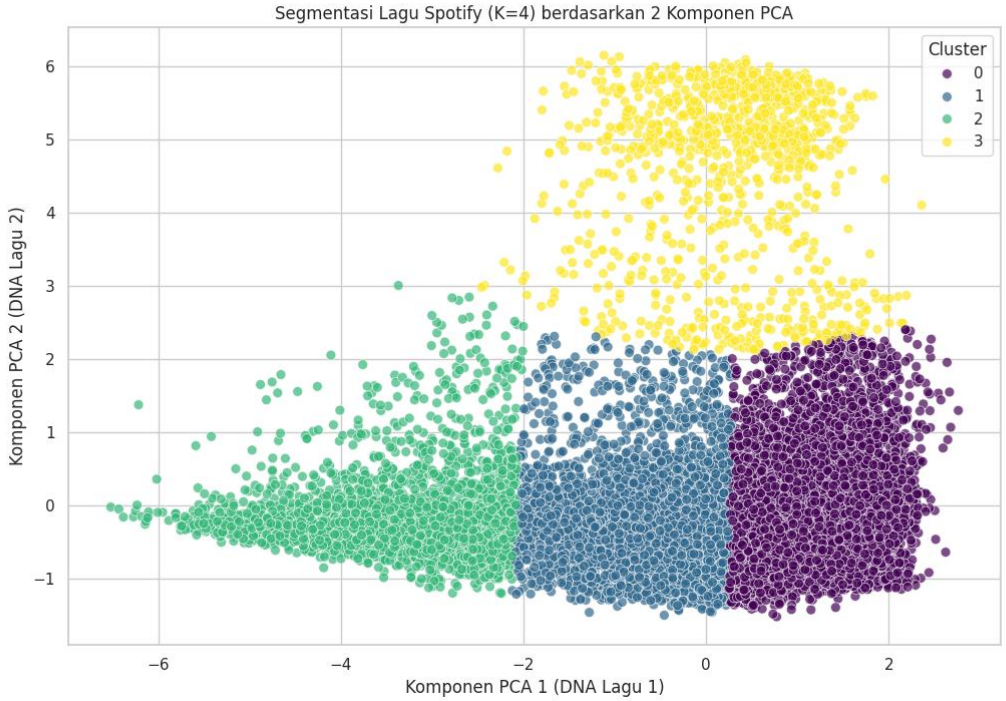
5. Rekomendasi Model Final

Berdasarkan perbandingan di atas, kami merekomendasikan **Model K-Means dengan K=4** sebagai model final.

Alasan: Model ini memberikan **Silhouette Score tertinggi (0.4798)**, yang menunjukkan pemisahan cluster yang paling optimal secara statistik. Selain itu, 4 cluster ini menghasilkan 4 "persona" lagu yang sangat berbeda, jelas, dan dapat ditindaklanjuti secara bisnis, seperti yang akan dijelaskan di bawah.

6. Temuan Utama dan Wawasan (Key Findings)

Analisis berhasil mengidentifikasi 4 segmen (cluster) "vibe" musik yang berbeda dari 20.000 lagu.



Profil audio rata-rata untuk setiap cluster adalah sebagai berikut:

cluster_label	acousticness	danceability	energy	instrumentalness	liveness	loudness	speechiness
0	0.124195531	0.623630605	0.73836009	0.038587227	0.200991002	-6.024179298	0.103229606
1	0.539906433	0.542638344	0.427977089	0.180438039	0.164529901	-10.70484926	0.066576771
2	0.895569432	0.306712781	0.143565869	0.580703384	0.155902428	-20.73804106	0.050170689
3	0.728573632	0.551012948	0.666997211	0.003048721	0.741231574	-11.43844522	0.782036653

- **Cluster 0: "Energi Tinggi / Lagu Pesta Mainstream"**
    - **Profil:** energy (0.74) dan loudness (-6.02) sangat tinggi, danceability (0.62) tinggi. acousticness (0.12) sangat rendah.
    - **Wawasan:** Ini adalah segmen lagu pop, EDM, dan *dance* yang paling komersial dan energik di platform.

- **Cluster 1: "Akustik Santai / Chill Vibes"**
    - **Profil:** acousticness (0.54) sedang-tinggi, namun energy (0.43) dan loudness (-10.70) rendah-sedang.
    - **Wawasan:** Segmen ini mewakili musik *easy listening*, pop akustik, dan folk. Cocok untuk *playlist* santai.
- **Cluster 2: "Instrumental Tenang / Klasik & Ambient"**
    - **Profil:** acousticness (0.90) dan instrumentalness (0.58) sangat tinggi. energy (0.14) dan loudness (-20.74) sangat rendah.
    - **Wawasan:** Ini adalah segmen musik yang paling tenang dan sunyi, didominasi oleh instrumen tanpa vokal. Sangat ideal untuk *playlist* fokus, belajar, atau tidur.
- **Cluster 3: "Spoken Word / Rekaman Live"**
    - **Profil:** speechiness (0.78) dan liveness (0.74) sangat tinggi, jauh di atas cluster lain.
    - **Wawasan:** Temuan yang menarik. Model ini berhasil mengisolasi konten non-Musik, seperti *podcast*, *stand-up comedy*, atau rekaman konser *live* ke dalam segmennya sendiri.

7. Kelemahan dan Langkah Selanjutnya

Meskipun analisis ini berhasil, ada beberapa keterbatasan dan peluang untuk perbaikan di masa depan:

- **Kelemahan Model:**
    1. **PCA:** Penggunaan PCA untuk visualisasi (mengurangi 7 fitur menjadi 2) menyebabkan hilangnya sebagian informasi variasi data (sekitar **33.70%** variasi hilang). Cluster yang terlihat tumpang tindih di plot 2D mungkin sebenarnya sangat terpisah di ruang 7-dimensi.
    2. **"Spoken Word":** Cluster 3 (Spoken Word) mungkin terlalu beragam (mencampur podcast, komedi, dan konser).
- **Rekomendasi Langkah Selanjutnya:**
    1. **Analisis Genre:** Melakukan analisis lebih dalam pada Cluster 0, 1, dan 2 untuk melihat distribusi genre di dalamnya. Apakah "Energi Tinggi" didominasi Pop atau Hip-Hop? (Seperti yang kita diskusikan!)
    2. **Sub-Cluster:** Mengambil Cluster 3 ("Spoken Word") dan melakukan *clustering* ulang (sub-clustering) hanya pada data tersebut untuk memisahkannya lebih lanjut (misalnya, memisahkan Podcast dari Konser Live).
    3. **Clustering pada Dimensi Penuh:** Melakukan clustering pada 7 fitur audio asli (tanpa PCA) dan membandingkan hasilnya, meskipun tidak dapat divisualisasikan secara langsung.