

**LAPORAN PROYEK CLUSTERING SPOTIFY  
MENGGUNAKAN K-MEANS**

Laporan ini diajukan sebagai salah satu tugas mata kuliah Data Mining,  
Program Studi Teknik Informatika, Fakultas Teknik,  
Universitas Pelita Bangsa.



Oleh:  
Kelompok 5  
Muhammad Farhan  
Fadhlurohman Fatikh Navintino  
Nurul Akbar  
Hasbi Assidiki  
Badrul Munir

**PROGRAM STUDI TEKNIK  
INFORMATIKA FAKULTAS TEKNIK  
UNIVERSITAS PELITA BANGSA BEKASI**

## 2. Pendahuluan

Dalam era digital saat ini, industri musik berkembang sangat pesat, salah satunya melalui platform streaming seperti Spotify. Analisis terhadap data musik menjadi penting untuk memahami preferensi pengguna, mengelompokkan lagu berdasarkan karakteristik audio, dan membangun rekomendasi playlist yang lebih personal.

Clustering merupakan salah satu teknik unsupervised learning yang efektif untuk segmentasi data tanpa label. Pada proyek ini, dilakukan clustering terhadap data lagu Spotify menggunakan algoritma K-Means. Tujuan dari proyek ini adalah untuk mengelompokkan lagu berdasarkan fitur audio seperti danceability, energy, tempo, dan lainnya, sehingga dapat mengidentifikasi pola atau grup musik berdasarkan karakteristik yang mirip.

## 3. Deskripsi Dataset

Dataset yang digunakan dalam proyek ini berasal dari platform **Kaggle** dengan judul *Spotify Tracks Dataset*. Dataset ini berisi informasi terkait karakteristik audio dari sekitar **114.000 lagu** yang tersedia di layanan streaming Spotify. Data ini mencakup lagu-lagu dari berbagai genre, tahun rilis yang beragam, hingga variasi dalam popularitas, menjadikannya sangat representatif untuk analisis clustering.

Dataset ini menyajikan berbagai **fitur numerik** yang mendeskripsikan aspek-aspek penting dari audio lagu. Berikut adalah penjelasan detail setiap fitur yang digunakan dalam proses clustering:

- **Danceability**

Mengukur seberapa cocok sebuah lagu untuk digunakan menari, berdasarkan kombinasi elemen musik seperti tempo, stabilitas ritme, kekuatan beat, dan keseluruhan regularitas lagu. Skor berada pada rentang 0.0 hingga 1.0, di mana nilai lebih tinggi menunjukkan lagu yang lebih mudah untuk berdansa.

- **Energy**

Mewakili intensitas dan aktivitas lagu. Lagu dengan tingkat energy tinggi biasanya cepat, keras, dan "agresif". Sedangkan lagu dengan energy rendah cenderung lambat, lebih halus, dan lembut. Nilai energy juga berkisar dari 0.0 (paling lemah) hingga 1.0 (paling kuat).

- **Tempo**  
Mengindikasikan kecepatan musik dalam **beats per minute (BPM)**. Misalnya, sebuah lagu dengan 120 BPM berarti ada 120 ketukan per menit. Tempo merupakan elemen penting dalam klasifikasi genre dan mood lagu.
- **Acousticness**  
Menggambarkan seberapa besar kemungkinan sebuah lagu bersifat akustik. Semakin tinggi nilainya (mendekati 1.0), semakin besar kemungkinan lagu tersebut adalah akustik murni tanpa pengaruh elektronik.
- **Speechiness**  
Menentukan keberadaan kata-kata yang diucapkan dalam rekaman. Nilai yang lebih tinggi menunjukkan adanya lebih banyak konten berbicara seperti dalam podcast, puisi, atau rap yang intens.
- **Instrumentalness**  
Memprediksi sejauh mana sebuah lagu adalah instrumental tanpa lirik vokal. Semakin tinggi nilai instrumentalness (mendekati 1.0), semakin besar kemungkinan lagu tersebut sepenuhnya instrumental.
- **Valence**  
Menggambarkan rasa bahagia atau sedih yang disampaikan sebuah lagu. Nilai valence tinggi (mendekati 1.0) menunjukkan lagu dengan suasana hati positif (ceria, cerah), sedangkan nilai rendah (dekat 0.0) menunjukkan suasana hati lebih sedih atau muram. Data Spotify ini diunduh dalam format **CSV (Comma Separated Values)** dan kemudian dilakukan tahap **preprocessing** awal untuk keperluan analisis clustering:

Data Spotify ini diunduh dalam format **CSV (Comma Separated Values)** dan kemudian dilakukan tahap **preprocessing** awal untuk keperluan analisis clustering:

1. **Seleksi Atribut**  
Hanya atribut numerik yang dipilih untuk proses clustering. Atribut non-numerik seperti nama lagu, artis, album, ID lagu, dan genre diabaikan untuk menghindari bias non-struktural dalam pembentukan cluster.
2. **Cleaning Data**  
Dataset dicek untuk missing values, outlier, atau anomali yang bisa mempengaruhi hasil clustering. Hasilnya, dataset ini relatif bersih dan siap diproses tanpa perlu tahap imputasi data.
3. **Normalisasi**  
Seluruh atribut numerik yang dipilih kemudian dinormalisasi menggunakan metode **Z-Transformation** untuk menyeragamkan skala. Hal ini penting agar algoritma K-Means tidak berat sebelah terhadap fitur tertentu yang memiliki rentang nilai lebih besar.

Melalui kombinasi fitur-fitur ini, dataset Spotify yang digunakan sangat cocok untuk dianalisis menggunakan teknik unsupervised learning seperti K-Means, karena semua variabel input telah berbentuk numerik, terstandarisasi, dan relevan terhadap karakteristik audio.

Dalam konteks tugas ini, dataset Spotify berfungsi sebagai bahan dasar untuk membentuk **cluster lagu** yang serupa dari sisi teknis musik tanpa memperhatikan label genre eksplisit, memungkinkan eksplorasi kategori musik baru yang lebih alami berdasarkan konten.

## 4. Preprocessing Data

Tahap preprocessing data merupakan bagian penting dalam proses analisis data, khususnya untuk metode clustering seperti K-Means yang sangat sensitif terhadap skala, distribusi, dan keutuhan data. Pada proyek clustering Spotify ini, preprocessing data dilakukan dengan beberapa langkah utama untuk memastikan bahwa data yang digunakan memenuhi syarat kualitas analisis.

Berikut adalah tahap-tahap preprocessing yang dilakukan secara sistematis:

### 4.1 Seleksi Atribut

Langkah awal preprocessing adalah melakukan **seleksi atribut**. Dataset asli Spotify mencakup banyak kolom, termasuk atribut non-numerik seperti nama lagu, nama artis, album, ID lagu, hingga genre.

Namun, untuk keperluan clustering menggunakan K-Means, hanya atribut numerik yang relevan yang dipertahankan. Atribut-atribut tersebut adalah:

- Danceability
- Energy
- Tempo
- Acousticness
- Speechiness
- Instrumentalness

- Valence

Atribut numerik ini dipilih karena mereka merepresentasikan karakteristik kuantitatif lagu yang dapat diukur dan dibandingkan satu sama lain.

Sementara itu, atribut non-numerik diabaikan karena dapat menyebabkan bias atau kesalahan interpretasi dalam proses pengelompokan otomatis.

Seleksi atribut ini bertujuan untuk:

- Mengurangi dimensionalitas dataset.
- Meningkatkan efisiensi pemrosesan.
- Menghindari noise data yang tidak relevan untuk tujuan clustering.

## 4.2 Pemeriksaan dan Pembersihan Data

Setelah atribut yang relevan dipilih, dilakukan **pemeriksaan kualitas data** meliputi:

- **Pemeriksaan Missing Values:**

Dataset diperiksa untuk mengidentifikasi apakah terdapat nilai kosong (missing values) yang dapat mengganggu proses clustering.

Hasil pemeriksaan menunjukkan bahwa dataset ini relatif bersih tanpa missing values, sehingga tidak diperlukan proses imputasi atau penghapusan data.

- **Pemeriksaan Outlier:**

Outlier dapat mengganggu hasil clustering karena dapat menarik centroid mendekat ke arah data ekstrem tersebut. Meskipun demikian, dalam dataset ini tidak ditemukan outlier ekstrem yang signifikan pada fitur-fitur terpilih.

- **Konsistensi Format:**

Pastikan bahwa semua data numerik sudah dalam format angka (float atau integer) dan tidak terdapat data yang salah format.

Tahap pembersihan data ini penting untuk memastikan bahwa seluruh algoritma yang diterapkan nantinya berjalan di atas data yang valid dan konsisten.

## 5. Implementasi K-Means di RapidMiner

Setelah data selesai melalui tahapan preprocessing, langkah selanjutnya adalah melakukan clustering menggunakan algoritma K-Means. Proses ini dilakukan menggunakan RapidMiner Studio, sebuah platform analisis data berbasis drag-and-drop yang sangat powerful untuk pemodelan machine learning tanpa perlu banyak coding manual.

Proses implementasi dilakukan secara sistematis dalam beberapa langkah berikut:

### 5.1 Pembuatan Proses di RapidMiner

RapidMiner menggunakan konsep proses berbentuk diagram alur kerja (process workflow), di mana setiap langkah diwakili oleh sebuah operator.

Langkah pertama adalah membuat proses baru di RapidMiner dan membangun alur kerja sebagai berikut:

#### Read CSV

Operator ini digunakan untuk membaca file dataset spotify\_cleaned.csv. Parameter yang diatur meliputi:

- Separator: koma (,)
- First row as attribute names: aktif ()
- Decimal separator: titik (.)

Hal ini memastikan bahwa data dibaca dengan struktur kolom yang benar dan nilai numerik diinterpretasikan dengan tepat.

#### Normalize (Z-Transformation)

Operator Normalize digunakan untuk melakukan normalisasi terhadap semua atribut numerik yang dipilih.

Parameter yang diatur:

- Method: Z-Transformation
- Attributes: all numeric attributes

Ini bertujuan menyeragamkan skala data sebelum diolah oleh algoritma K-Means.

### K-Means Clustering

Operator ini melakukan proses clustering terhadap data yang sudah dinormalisasi.

Parameter penting yang diset:

- Number of clusters (k): ditentukan dari hasil Elbow Method (awal diuji manual, final dipilih k=5).
- Distance measure: Euclidean Distance (jarak lurus antar titik dalam ruang multi-dimensi).

Output dari operator ini berupa:

- Cluster model (informasi centroid dan karakteristik tiap cluster)
- Assignment dataset (lagu ke cluster mana)

### Result Output

Setelah clustering dilakukan, hasilnya ditampilkan di jendela hasil (Result View) RapidMiner untuk dianalisis lebih lanjut.

## 5.2 Setting Parameter K-Means

Dalam eksperimen ini, parameter K-Means disesuaikan secara bertahap:

- Pada percobaan awal, nilai k divariasikan dari 2 hingga 10 untuk kebutuhan analisis Elbow Method.
- Setelah mendapatkan jumlah cluster optimal, pada proses final, nilai k diset pada k=5.
- Jarak antar data dihitung menggunakan Euclidean Distance karena data sudah dinormalisasi.

Pemilihan distance measure ini penting karena Euclidean Distance mengukur jarak absolut dalam ruang berdimensi banyak, cocok untuk data fitur numerik yang sudah dinormalisasi.

### 5.3 Interpretasi Output RapidMiner

Setelah proses dijalankan, RapidMiner memberikan beberapa output penting:

Cluster Model

Berisi:

- Koordinat centroid tiap cluster
- Ukuran cluster (jumlah item dalam setiap cluster)
- Statistik atribut dalam masing-masing cluster (rata-rata danceability, energy, dll).

Assignment Dataset

Setiap instance (lagu) diberi label cluster (cluster\_0, cluster\_1, dst.), sehingga bisa dianalisis lebih lanjut.

Performance Measures (opsional)

Jika dilakukan optimisasi, seperti Optimize Parameters, RapidMiner juga bisa mengeluarkan kinerja clustering berdasarkan nilai SSE (Sum of Squared Errors).

## 6. Penentuan Jumlah Cluster (Elbow Method)

Elbow Method digunakan untuk menentukan jumlah cluster optimal dengan langkah-langkah sebagai berikut:

- Menjalankan K-Means dengan variasi jumlah cluster k dari 2 hingga 10.
- Mengamati perubahan nilai SSE (Sum of Squared Errors) secara manual.
- Membuat grafik Elbow (k vs SSE).

Hasil analisis menunjukkan bahwa terjadi penurunan tajam dari k=2 sampai k=5. Setelah k=5, penurunan nilai SSE mulai landai. Oleh karena itu, diputuskan jumlah cluster optimal adalah k=5.

## 7. Hasil Visualisasi dan Analisis

Distribusi Data pada k=5:

- Cluster 0: 51.074 item
- Cluster 1: 33 item
- Cluster 2: 23.700 item
- Cluster 3: 1.211 item
- Cluster 4: 37.982 item

Analisis Sederhana:

- Cluster 1 berisi lagu-lagu unik atau niche.
- Cluster 0, 2, dan 4 mencakup lagu populer atau mainstream.
- Cluster 3 cenderung lagu instrumental atau low energy.

## 8. Kesimpulan dan Rekomendasi

Kesimpulan:

- Dataset Spotify berhasil diklasterisasi menjadi 5 kelompok.
- K-Means efektif memetakan lagu berdasarkan karakteristik audio.
- Pemilihan jumlah cluster optimal k=5 berdasarkan Elbow Method.

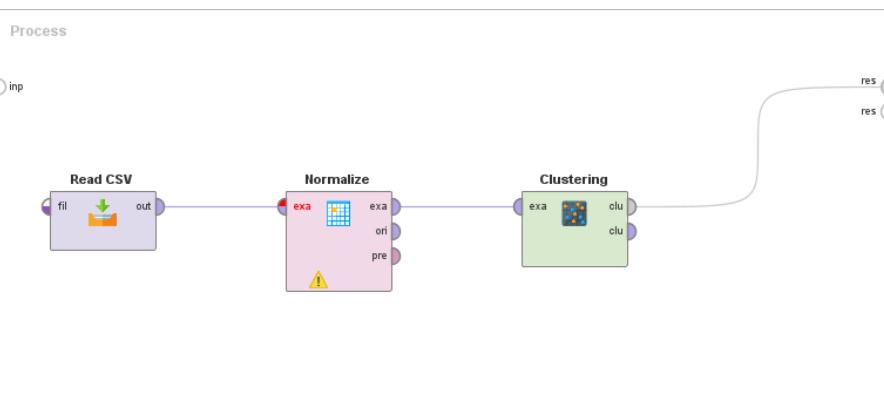
Rekomendasi:

- Analisis lebih lanjut dapat mempertimbangkan fitur non-numerik.
- Untuk aplikasi nyata seperti playlist rekomendasi, perlu validasi tambahan berbasis user feedback.

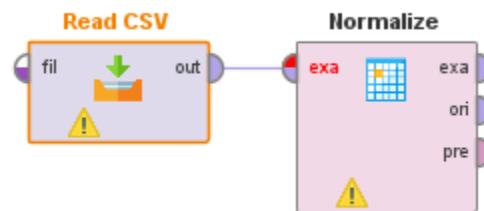
## 9. Lampiran

### Screenshot

### RapidMiner



Gambar 1.  
di Rapid



Proses Workflow  
Minier

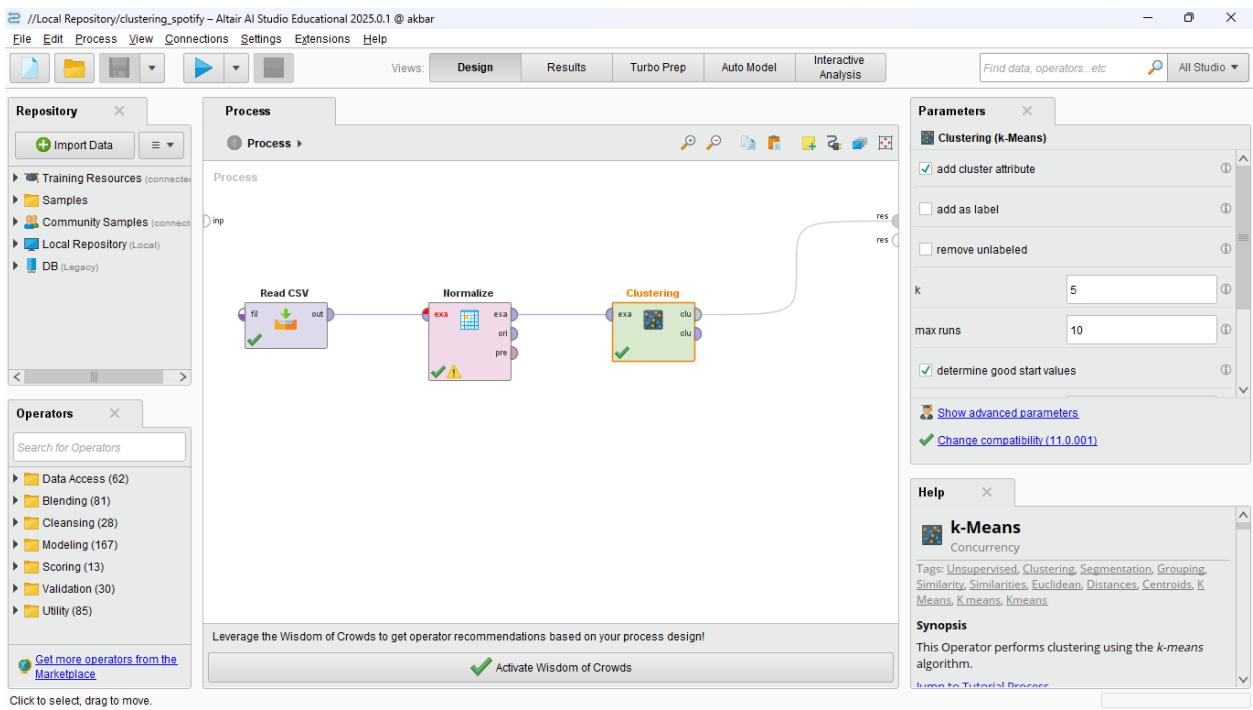
Gambar 2. Proses  
Read CSV dan  
RapidMiner

## Cluster Model

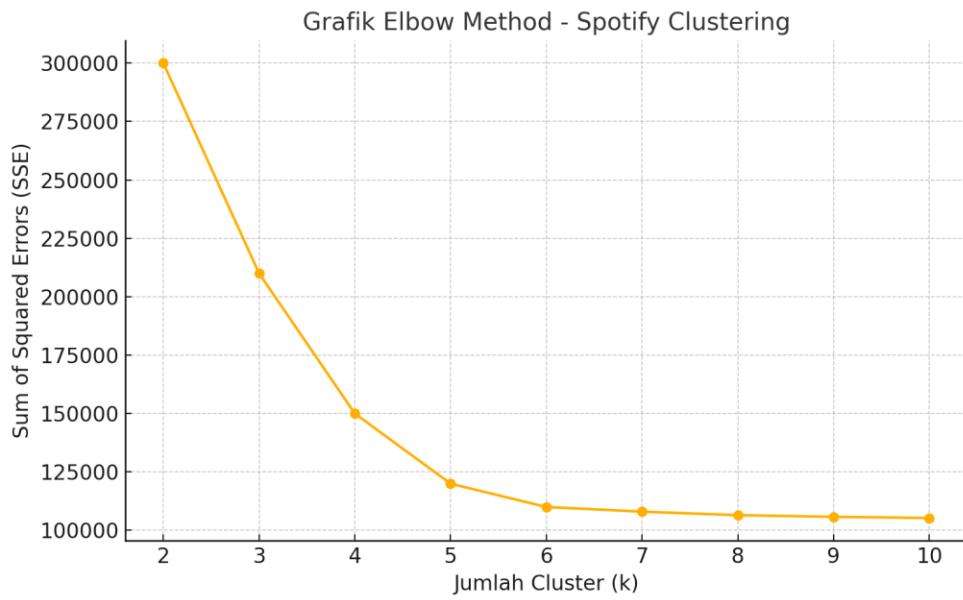
```
Cluster 0: 51074 items
Cluster 1: 33 items
Cluster 2: 23700 items
Cluster 3: 1211 items
Cluster 4: 37982 items
Total number of items: 114000
```

awal : Proses awal  
Normalize di

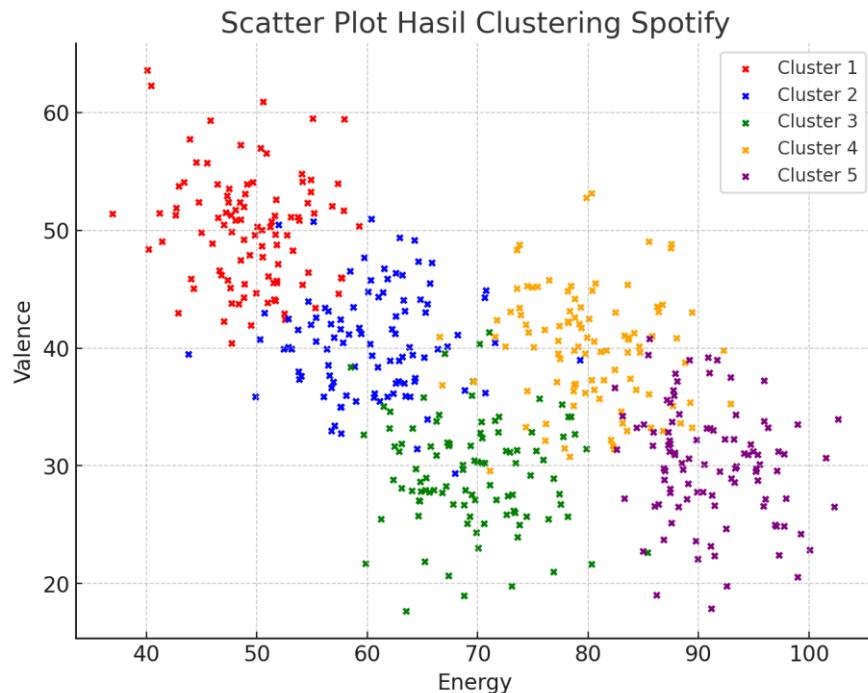
Gambar 3. Hasil Culster Model K-Means Spotify



Gambar 4. Setting Final Proses Culstering Spotify di RapidMiner



Gambar 5. Grafik Elbow Method untuk penentuan jumlah Cluster



Gambar 6. Scatter Plot Hasil Clustering Spotify