**IMPLEMENTASI DATA MINING UNTUK ANALISIS TREND LAGU DI SPOTIFY PADA TAHUN 2023 BERDASARKAN KARAKTERISTIK AUDIO MENGGUNAKAN METODE K-MEANS CLUSTERING**

Disusun untuk Memenuhi Tugas Besar Mata Kuliah Data Mining



Disusun Oleh:

Kelompok 3 Data Mining-A

|  |  |
| --- | --- |
| Trici Ayunda | 2311522017 |
| Vania Zhafira Zahra | 2311523007 |
| Suci Nurhaliza | 2311521009 |
| Muhammad Diaz Ananda S | 2311521015 |
| Ikhwan Hamidi | 2311521003 |

Dosen Pengampu:

Aina Hubby Aziira, M.Eng.

Dwi Welly Sukma Nirad, S.Kom, M.T.

**DEPARTEMEN SISTEM INFORMASI**

**FAKULTAS TEKNOLOGI INFORMASI**

**UNIVERSITAS ANDALAS**

**2025**

# KATA PENGANTAR

Puji syukur kami panjatkan ke hadirat Allah Swt. atas rahmat dan karunia-Nya sehingga kami dapat menyelesaikan laporan ini tepat waktu. Tanpa pertolongan-Nya, tentu kami tidak dapat menyelesaikan laporan ini dengan baik. Shalawat serta salam semoga senantiasa tercurah kepada Nabi Muhammad Saw., suri teladan umat manusia, yang syafaatnya selalu kita harapkan di hari kemudian.

Kami juga mengucapkan syukur kepada Allah Swt. atas limpahan nikmat dan sehat, baik itu berupa fisik maupun akal pikiran sehingga kami mampu menyusun laporan ini sebagai tugas besar Mata Kuliah *Data Mining*, dengan judul **“Implementasi Data Mining untuk Analisis Tren Lagu di Spotify pada Tahun 2023 Berdasarkan Karakteristik Audio Menggunakan Metode K-Means Clustering”**.

Ucapan terima kasih yang sebesar-besarnya kami sampaikan kepada Ibu Aina Hubby Aziira, M.Eng. dan Ibu Dwi Welly Sukma Nirad, S.Kom, M.T selaku dosen Mata Kuliah *Data Mining* yang telah membimbing, memberikan arahan, serta mendukung kami dalam menyelesaikan tugas ini.

Kami menyadari bahwa laporan ini masih jauh dari kata sempurna dan masih banyak terdapat kesalahan serta kekurangan di dalamnya. Oleh karena itu, kami sangat mengharapkan kritik dan saran yang membangun dari para pembaca guna perbaikan di masa mendatang. Kemudian, apabila terdapat banyak kesalahan pada laporan ini kami mohon maaf yang sebesar-besarnya.

Demikian, semoga laporan ini dapat bermanfaat bagi pembaca maupun penulis sendiri. Terima kasih.

Padang, 2025

Penulis

# 

# DAFTAR ISI

[KATA PENGANTAR 2](#_Toc201314267)

[DAFTAR ISI 3](#_Toc201314268)

[DAFTAR GAMBAR 5](#_Toc201314269)

[DAFTAR RUMUS 6](#_Toc201314270)

[DAFTAR TABEL 7](#_Toc201314271)

[BAB I PENDAHULUAN 1](#_Toc201314272)

[**1.1** **Latar belakang** 1](#_Toc201314273)

[**1.2 Rumusan Masalah** 1](#_Toc201314274)

[**1.3 Batasan Masalah** 2](#_Toc201314275)

[**1.4 Tujuan** 2](#_Toc201314276)

[**1.5 Manfaat** 2](#_Toc201314277)

[BAB II METODOLOGI PENELITIAN 4](#_Toc201314278)

[**2.1** **Spotify** 4](#_Toc201314279)

[**2.2** **Data Mining** 4](#_Toc201314280)

[**2.2.1** **Pengertian Data Mining** 4](#_Toc201314281)

[**2.2.2** **Tujuan Data Mining** 4](#_Toc201314282)

[**2.2.3** **Manfaat Data Mining** 5](#_Toc201314283)

[**2.3** **Clustering** 5](#_Toc201314284)

[**2.3.1** **K-means** 6](#_Toc201314285)

[**2.4** **Objek Penelitian** 7](#_Toc201314286)

[**2.5** **Metode Penelitian** 8](#_Toc201314287)

[**2.6** **Pengumpulan Data** 8](#_Toc201314288)

[**2.6.2** **Atribut** 8](#_Toc201314289)

[**2.7** **Flowchart Penelitian** 12](#_Toc201314290)

[BAB III PEMBAHASAN 15](#_Toc201314291)

[**3.1** **Perhitungan Manual** 15](#_Toc201314292)

[**3.1.1** **Perhitungan K-means** 16](#_Toc201314293)

[**3.2** **Implementasi** 25](#_Toc201314294)

[**3.2.1** **Import Library** 25](#_Toc201314295)

[**3.2.2** **Data Gathering** 25](#_Toc201314296)

[**3.2.3** **Data Preparation** 26](#_Toc201314297)

[**3.2.4** **Data Analysis** 28](#_Toc201314298)

[**3.2.5** **Implementasi K-Means** 30](#_Toc201314299)

[BAB IV PENUTUP 42](#_Toc201314300)

[**4.1** **Kesimpulan** 42](#_Toc201314301)

[**4.2** **Saran** 42](#_Toc201314302)

[DAFTAR PUSTAKA 43](#_Toc201314303)

# DAFTAR GAMBAR

[Gambar 1 Flowchart Penelitian 12](#_Toc201313569)

[Gambar 2 Korelasi Semua Fitur Numerik 14](#_Toc201313570)

[Gambar 3 Metode Elbow untuk Jumlah klaster 16](#_Toc201313571)

[Gambar 4 Import Library 23](#_Toc201313572)

[Gambar 5 Data Gathering 23](#_Toc201313573)

[Gambar 6 Data Shape 23](#_Toc201313574)

[Gambar 7 Data Info 24](#_Toc201313575)

[Gambar 8 Data Null 25](#_Toc201313576)

[Gambar 9 Penanganan Data NUll 26](#_Toc201313577)

[Gambar 10 Re-Check Data NULL 26](#_Toc201313578)

[Gambar 11 Data Analysis 27](#_Toc201313579)

[Gambar 12 Memilih Atribut 28](#_Toc201313580)

[Gambar 13 Pengecekan Duplikat 29](#_Toc201313581)

[Gambar 14 Standarisasi Data Fitur Audio 29](#_Toc201313582)

[Gambar 15 Menentukan Jumlah Cluster 30](#_Toc201313583)

[Gambar 16 Menentukan Centroid 31](#_Toc201313584)

[Gambar 17 Pelatihan Model K-Means 31](#_Toc201313585)

[Gambar 18 Penentuan Label Cluster 32](#_Toc201313586)

[Gambar 19 Distribusi Cluster 10 Data 32](#_Toc201313587)

[Gambar 20 Visualisasi 33](#_Toc201313588)

[Gambar 21 Visualisasi 34](#_Toc201313589)

[Gambar 22 Pelatihan Model K-Means 35](#_Toc201313590)

[Gambar 23 Hasil Cluster 36](#_Toc201313591)

[Gambar 24 Jumlah Anggota Tiap Cluster 36](#_Toc201313592)

[Gambar 25 Mengubah Centroid 36](#_Toc201313593)

[Gambar 26 Visualisasi 39](#_Toc201313594)

# DAFTAR RUMUS

[Rumus 1 Euclidean distance 17](#_Toc201313701)

[Rumus 2 Centroid 18](#_Toc201313702)

# DAFTAR TABEL

[Table 1 Data Perhitungan Algoritma K-Means 14](#_Toc201313604)

[Table 2 Data Centroid Iterasi Pertama 17](#_Toc201313605)

[Table 3 Hasil Perhitungan Iterasi Pertama 18](#_Toc201313606)

[Table 4 Hasil Klaster Iterasi Pertama 18](#_Toc201313607)

[Table 5 Data Centroid Iterasi Kedua 19](#_Toc201313608)

[Table 6 Hasil Perhitungan Iterasi Kedua 20](#_Toc201313609)

[Table 7 Hasil Klaster Iterasi Kedua 21](#_Toc201313610)

# BAB I PENDAHULUAN

## **Latar belakang**

Musik telah menjadi bagian penting dalam kehidupan sehari-hari, di mana masyarakat cenderung memilih lagu berdasarkan genre dan preferensi pribadi. Dengan perkembangan teknologi, layanan streaming musik seperti Spotify telah memberikan kemudahan bagi pengguna dalam mengakses berbagai macam lagu secara digital. Sejak diluncurkan pada tahun 2008, Spotify telah menjadi salah satu platform musik terbesar yang menyediakan fitur gratis maupun berbayar bagi penggunanya.

Seiring bertambahnya jumlah pengguna, pemahaman tentang tren lagu yang populer di suatu periode menjadi penting, baik bagi industri musik, kreator konten, maupun pengiklan. Salah satu cara untuk menganalisis tren lagu adalah dengan menggunakan teknik Data Mining, khususnya metode K-Means Clustering, yang mampu mengelompokkan lagu berdasarkan karakteristik audio tertentu.

Dalam beberapa penelitian sebelumnya, metode K-Means telah digunakan dalam berbagai kasus klasifikasi, seperti segmentasi pelanggan berdasarkan perilaku belanja, pengelompokan mahasiswa berdasarkan nilai akademik, serta identifikasi pola penyakit berdasarkan data medis. Namun, studi yang mengaplikasikan metode K-Means untuk mengelompokkan lagu berdasarkan karakteristik audio pada Spotify masih terbatas. Oleh karena itu, penelitian ini bertujuan untuk menganalisis tren lagu populer di Spotify pada tahun 2023 berdasarkan karakteristik audio menggunakan metode K-Means Clustering.

## **1.2 Rumusan Masalah**

Berdasarkan latar belakang di atas, rumusan masalah dalam laporan ini adalah sebagai berikut:

1. Bagaimana karakteristik audio dari lagu-lagu populer di Spotify pada tahun 2023?
2. Bagaimana metode K-Means Clustering dapat digunakan untuk mengelompokkan lagu berdasarkan karakteristik audio?
3. Apa saja pola yang dapat diidentifikasi dari hasil analisis clustering terhadap lagu-lagu populer di Spotify?

## **1.3 Batasan Masalah**

Agar penelitian lebih terarah, beberapa batasan yang diterapkan dalam penelitian ini adalah:

1. **Sumber Data**

Data yang digunakan dalam penelitian ini berasal dari dataset "spotify 2023" yang diperoleh dari *Kaggle*.

1. **Variabel yang dianalisis**

Analisis hanya dilakukan berdasarkan karakteristik audio yang tersedia dalam dataset dengan memilih beberapa fitur yang relevan.

1. **Lingkup Geografis**

Analisis tidak membedakan tren musik berdasarkan negara atau wilayah tertentu, melainkan melihat pola secara umum di seluruh pengguna *Spotify*.

## **1.4 Tujuan**

Adapun tujuan dari penelitian adalah sebagai berikut:

1. Menganalisis karakteristik audio dari lagu-lagu populer di Spotify pada tahun 2023.
2. Mengimplementasikan metode K-Means Clustering untuk mengelompokkan lagu berdasarkan karakteristik audio.
3. Mengidentifikasi pola atau kecenderungan dalam tren lagu populer berdasarkan hasil clustering.

## **1.5 Manfaat**

Adapun manfaat yang diharapkan dari penelitian ini adalah:

1. **Bagi Industri Musik**

Membantu industri musik dalam memahami tren lagu berdasarkan karakteristik audio untuk strategi pemasaran yang lebih efektif.

1. **Bagi Pengguna Spotify**

Memberikan wawasan tentang pola lagu populer, sehingga pengguna dapat menemukan lagu yang sesuai dengan preferensi mereka.

1. **Bagi Peneliti di Bidang Data Mining**

Menambah referensi terkait penerapan metode K-Means Clustering dalam analisis tren musik.

1. **Bagi Pengembang Aplikasi Musik**

Memberikan wawasan tentang bagaimana karakteristik audio memengaruhi popularitas lagu, yang dapat digunakan dalam pengembangan fitur rekomendasi musik yang lebih baik.

# 

# BAB II METODOLOGI PENELITIAN

## **Spotify**

Spotify adalah layanan streaming musik digital yang diluncurkan pada tahun 2008 oleh perusahaan Spotify AB asal Swedia. Platform ini menyediakan akses terhadap jutaan lagu dari berbagai genre, artis, dan era secara daring. Pengguna dapat menikmati musik secara gratis dengan iklan atau melalui langganan premium tanpa iklan. Spotify juga menyediakan berbagai fitur seperti playlist personalisasi, statistik audio, serta Application Programming Interface (API) yang memungkinkan pengambilan data lagu untuk analisis lebih lanjut. Dalam penelitian ini, Spotify menjadi sumber utama data berupa metadata dan karakteristik audio dari lagu-lagu yang dirilis pada tahun 2023.

## **Data Mining**

Penelitian ini menggunakan penerapan data mining yang mempunyai pengertian, tujuan dan manfaat data mining. Hal tersebut menjadi alasan untuk menerapkan data mining pada penelitian. Data akan diolah dengan mengikuti tahap data mining dan mendapatkan hasil yang mengacu pada tujuan serta manfaat data mining.

### **Pengertian Data Mining**

Data Mining adalah proses untuk menemukan pola atau informasi tersembunyi dari sejumlah besar data secara otomatis atau semi-otomatis, dengan tujuan mendapatkan pengetahuan yang bermanfaat. Proses ini mencakup analisis statistik, *machine learning,* dan teknik *database* untuk menghasilkan hubungan, pola atau tren dalam data.

### **Tujuan Data Mining**

Tujuan utama dari data mining adalah untuk mengidentifikasi pola tersembunyi yang berguna dan memiliki makna dalam kumpulan data yang besar. Dalam konteks penelitian ini, tujuan data mining adalah sebagai berikut:

* Mengelompokkan lagu berdasarkan kesamaan fitur audio menggunakan metode K-Means Clustering.
* Mengidentifikasi pola umum dalam karakteristik audio lagu-lagu populer di Spotify pada tahun 2023.
* Memberikan wawasan kepada industri musik, kreator konten, dan pengiklan mengenai tren yang sedang berlangsung.

### **Manfaat Data Mining**

Metode *data mining* telah banyak diterapkan dalam kehidupan nyata untuk menyelesaikan berbagai permasalahan manusia. Beberapa manfaat dari implementasi *data mining* menurut Adinugroho & Sari (2018) antara lain sebagai berikut:

1. Penginderaan Jauh Pengolahan citra satelit dalam penginderaan jarak jauh merupakan salah satu hasil implementasi metode klasifikasi dalam *data mining*.
2. Segmentasi Konsumen *Data mining* dapat digunakan untuk mengelompokkan konsumen berdasarkan karakteristik demografis, perilaku, maupun preferensi mereka, sehingga memudahkan pengambilan keputusan yang lebih tepat sasaran.

Di dunia bisnis, *data mining* telah lama dimanfaatkan untuk meningkatkan efisiensi dan keuntungan. Pola belanja konsumen, misalnya, dapat dianalisis dan dipetakan menggunakan berbagai teknik seperti *clustering*, *association rule mining*, dan *classification*.

## **Clustering**

Clustering merupakan metode untuk mengelompokkan objek-objek ke dalam beberapa grup berdasarkan informasi yang mencerminkan keterkaitan antar objek tersebut. Tujuan utamanya adalah untuk meningkatkan kemiripan antar anggota dalam satu kelompok, sekaligus mengurangi kemiripan antara kelompok yang berbeda. Dengan demikian, data dibagi ke dalam sejumlah klaster yang masing-masing berisi objek-objek dengan karakteristik yang serupa.

### **K-means**

Algoritma K-Means merupakan teknik *data mining* yaitu metode *clustering* atau pengelompokkan yang proses pemodelannya dilakukan tanpa supervisi/pembelajaran serta metode pengelompokkan datanya dilakukan secara partisi. Pada metode yang digunakan dalam algoritma K-Means, data akan dikelompokkan menjadi beberapa kelompok, dan setiap kelompok memiliki ciri-ciri yang mirip satu sama lain, namun berbeda dengan kelompok lainnya.

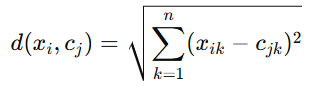
Hal ini bertujuan untuk meminimalisir perbedaan di dalam satu *cluster* dan memaksimalkan perbedaan antara *cluster* yang berbeda (Rizki Muliono dan Zulfikar Sembiring, 2019).

Istilah dalam Algoritma K-Means:

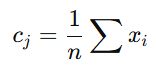
1. Cluster: Merupakan suatu kelompok atau grup.
2. Centroid: Titik pusat yang digunakan untuk menentukan jarak (*Euclidean Distance*).
3. Iterasi: Proses pengulangan yang akan berhenti ketika hasil iterasi telah mencapai konvergensi.

Langkah-Langkah Algoritma K-Means:

1. Menentukan jumlah *k-cluster* yang ingin dibentuk.
2. Membangkitkan nilai acak untuk pusat *cluster* awal (*centroid*) sebanyak *k-cluster*.
3. Menghitung jarak setiap data input terhadap masing-masing *centroid* menggunakan rumus jarak *Euclidean Distance* hingga ditemukan jarak terdekat dari setiap data terhadap *centroid*. Berikut adalah persamaan *Euclidean Distance*:



1. Mengklasifikasikan setiap data berdasarkan kedekatannya dengan *centroid* (jarak terkecil).
2. Memperbarui nilai *centroid*. Nilai *centroid* baru diperoleh dari rata-rata data dalam *cluster* yang bersangkutan menggunakan rumus:



1. Melakukan perulangan dari langkah 2 hingga 5 sampai anggota tiap *cluster* tidak ada yang berubah.
2. Jika langkah 6 telah terpenuhi, maka nilai rata-rata pusat *cluster* (μ j​) pada iterasi terakhir akan digunakan sebagai parameter untuk menentukan klasifikasi data.

## **Objek Penelitian**

Objek Penelitian ini berfokus pada analisis trend lagu di Spotify pada tahun 2023 berdasarkan karakteristik audio menggunakan metode K-Means Clustering dengan menggunakan data “Spotify 2023”. Variabel yang akan diteliti adalah karakteristik audio dari lagu-lagu yang tersedia di platform Spotify pada tahun 2023. Data yang digunakan mencakup berbagai atribut audio yang menentukan karakteristik sebuah lagu, seperti danceability, energy, valence. Variabel ini akan digunakan untuk mengelompokkan lagu-lagu berdasarkan kesamaan karakteristik audio menggunakan metode K-Means Clustering.

Pemilihan variabel ini didasarkan pada asumsi bahwa karakteristik audio suatu lagu memiliki pengaruh besar terhadap popularitas dan tren musik di Spotify. Dengan memahami pola dari lagu-lagu yang memiliki kesamaan fitur audio, dapat diidentifikasi genre atau gaya musik yang dominan dalam tren tahun 2023.

## **Metode Penelitian**

Metode penelitian ini menggunakan studi literatur pada dataset lagu-lagu Spotify tahun 2023 dengan pendekatan yang memanfaatkan sumber-sumber tertulis yang relevan dan terpercaya. Penelitian ini bertujuan untuk mengumpulkan informasi, menganalisis data, dan menarik kesimpulan mengenai tren lagu berdasarkan karakteristik audio. Proses analisis dilakukan dengan menerapkan algoritma *K-Means Clustering* untuk mengelompokkan lagu-lagu ke dalam klaster berdasarkan kesamaan fitur audio seperti tempo, energi, valensi, dan lainnya.

## **Pengumpulan Data**

Tahapan pengumpulan data dalam proses penelitian merupakan tahapan penting yang berperan dalam pengambilan informasi yang diperlukan untuk menjawab pertanyaan penelitian dan mencapai tujuan penelitian serta memberikan manfaat sesuai dengan kebutuhan.

1. **Sumber Data**

Dataset yang digunakan dalam penelitian ini adalah spotify\_2023, yang diperoleh dari platform Kaggle. Dataset ini berisi informasi mengenai lagu-lagu yang dirilis atau populer di Spotify selama tahun 2023, termasuk berbagai karakteristik audio seperti tempo, *danceability, energy, valence, loudness, acousticness, instrumentalness, dan speechiness.*

Dataset ini dipilih karena menyediakan data yang cukup lengkap untuk melakukan analisis tren musik berdasarkan karakteristik audio. Dengan menggunakan dataset ini, penelitian dapat mengelompokkan lagu-lagu ke dalam beberapa klaster berdasarkan kemiripan fitur audio menggunakan metode K-Means Clustering.

### **Atribut**

Atribut dalam data merujuk pada kolom atau fitur yang mewakili karakteristik atau variabel tertentu dari data yang sedang dianalisis. Setiap atribut memberikan informasi spesifik yang digunakan untuk memahami dan menganalisis fenomena yang sedang diteliti. Dalam proses analisis data lagu Spotify tahun 2023, digunakan sejumlah atribut yang merepresentasikan informasi penting terkait karakteristik lagu, artis, waktu perilisan, hingga performa lagu di berbagai platform musik digital. Penjelasan masing-masing atribut yang digunakan dalam dataset adalah sebagai berikut:

1. track\_name

Merupakan nama atau judul dari lagu yang terdapat dalam dataset. Atribut ini menjadi identitas utama dari setiap entri lagu yang dianalisis.

1. artist(s)\_name

Berisi nama penyanyi atau grup musik yang membawakan lagu. Dalam beberapa kasus, satu lagu dapat dibawakan oleh lebih dari satu artis sehingga nama-nama artis dapat digabung dalam satu entri.

1. artist\_count

Menunjukkan jumlah artis yang berkontribusi pada satu lagu. Atribut ini memberikan informasi apakah lagu tersebut merupakan kolaborasi atau dibawakan oleh artis tunggal.

1. released\_year, released\_month, released\_day

Ketiga atribut ini menjelaskan tanggal rilis lagu secara rinci, meliputi tahun, bulan, dan hari. Informasi ini penting untuk menganalisis tren lagu berdasarkan waktu perilisan dan pola distribusinya sepanjang tahun.

1. in\_spotify\_playlists

Merupakan jumlah playlist di platform Spotify yang memasukkan lagu tersebut. Semakin tinggi angkanya, semakin populer atau banyak didengar lagu tersebut oleh pengguna Spotify.

1. in\_spotify\_charts

Menunjukkan berapa kali lagu muncul di tangga lagu (charts) Spotify. Atribut ini menggambarkan tingkat popularitas dan eksistensi lagu dalam sistem pemeringkatan platform tersebut.

1. streams

Mengacu pada jumlah total pemutaran lagu di Spotify. Nilai ini sering kali disajikan dalam format teks atau angka besar yang perlu dibersihkan (pre-processing) sebelum dianalisis secara numerik.

1. in\_apple\_playlists dan in\_apple\_charts

Dua atribut ini berfungsi seperti atribut Spotify, namun berlaku untuk platform Apple Music. Informasi ini dapat digunakan untuk membandingkan performa lagu di berbagai platform streaming.

1. in\_deezer\_playlists dan in\_deezer\_charts

Menunjukkan jumlah playlist dan kemunculan lagu pada tangga lagu di platform Deezer. Sama seperti sebelumnya, data ini membantu mengukur eksposur lagu di luar Spotify.

1. in\_shazam\_charts

Menggambarkan seberapa sering lagu dikenali dan muncul di tangga lagu Shazam. Atribut ini menarik karena menunjukkan lagu-lagu yang sering dicari oleh publik menggunakan fitur pencarian audio.

1. bpm

Merupakan singkatan dari *beats per minute*, yaitu jumlah ketukan per menit pada sebuah lagu. Atribut ini menunjukkan tempo atau kecepatan lagu. Lagu dengan BPM tinggi cenderung energik dan cepat, sedangkan BPM rendah cenderung lambat dan tenang.

1. key

Merujuk pada tangga nada utama dari lagu, seperti C, D minor, G#, dan sebagainya. Atribut ini menggambarkan warna musikalitas lagu dan bisa berpengaruh pada suasana yang ditimbulkan.

1. mode

Mengindikasikan modus lagu, yaitu mayor atau minor. Lagu dengan modus mayor biasanya terdengar ceria dan positif, sementara modus minor cenderung bernuansa sedih atau melankolis.

1. danceability\_%

Menunjukkan tingkat keterdansa-an sebuah lagu, yaitu seberapa cocok lagu tersebut untuk digunakan dalam aktivitas menari. Nilai ini diukur dalam bentuk persentase (0–100%), di mana semakin tinggi nilainya, semakin mudah lagu tersebut ditarikan.

1. valence\_%

Menggambarkan nuansa emosional dari sebuah lagu. Nilai valence tinggi menunjukkan lagu yang positif dan ceria, sedangkan nilai rendah menunjukkan lagu yang suram, sedih, atau serius.

1. energy\_%

Mengukur intensitas dan semangat dari sebuah lagu. Lagu dengan energi tinggi biasanya terdengar keras, cepat, dan padat instrumen, sementara lagu dengan energi rendah cenderung lembut dan tenang.

1. acousticness\_%

Menunjukkan seberapa besar unsur akustik yang dimiliki lagu tersebut. Lagu dengan nilai akustik tinggi kemungkinan besar menggunakan instrumen akustik dan minim pemrosesan digital

1. instrumentalness\_%

Mengindikasikan kemungkinan sebuah lagu bersifat instrumental, yaitu tanpa vokal. Nilai tinggi pada atribut ini menunjukkan lagu yang cenderung murni musik, tanpa suara manusia

1. liveness\_%

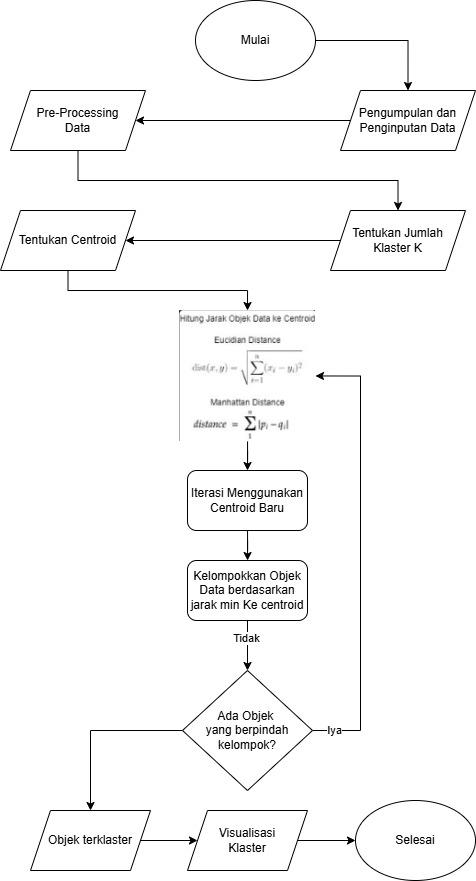
Merepresentasikan kemungkinan bahwa lagu direkam dalam kondisi live (langsung). Nilai tinggi biasanya menunjukkan adanya elemen penonton, tepuk tangan, atau suasana konser.

1. speechiness\_%

Mengukur seberapa banyak elemen “berbicara” dalam lagu tersebut. Lagu dengan nilai speechiness tinggi umumnya memiliki konten berupa pidato, rap, atau spoken word, bukan melodi atau nyanyian.

Atribut-atribut tersebut kemudian diproses lebih lanjut dalam tahapan *pre-processing* dan digunakan sebagai dasar untuk mengelompokkan lagu menggunakan algoritma **K-Means Clustering**, terutama berdasarkan karakteristik audionya. Hal ini bertujuan untuk menemukan pola-pola tersembunyi dan tren umum dari lagu-lagu populer di Spotify pada tahun 2023.

## **Flowchart Penelitian**



Gambar 1 Flowchart Penelitian

1. Mulai, merupakan tahap awal dari proses.
2. Pengumpulan dan Pengunduhan Data

Pada tahap ini, dataset *"spotify\_2023"* yang diperoleh dari Kaggle dikumpulkan dan diunduh untuk diproses.

1. Pre-processing Data

Untuk memeriksa apakah ada nilai yang hilang (*missing value*) dan data duplikat pada dataset.

1. Pemrosesan Data

Tentukan jumlah klaster K. Pada tahap ini, jumlah klaster yang ingin dibentuk (K) ditentukan terlebih dahulu sebelum proses klasterisasi dimulai.

1. Menentukan Centroid

Setelah jumlah klaster ditentukan, centroid awal untuk setiap klaster harus ditentukan. Centroid awal ini dapat dipilih secara acak atau menggunakan metode inisialisasi tertentu.

1. Hitung Jarak Objek Data ke Centroid

Untuk setiap objek data, jarak antara objek data tersebut dengan setiap centroid dihitung menggunakan metrik jarak tertentu, misalnya:

* Euclidean Distance: Jarak Euclidean antara objek data *x* dan centroid *y* dihitung dengan rumus:



* Manhattan Distance: Jarak Manhattan antara objek data *x* dan centroid *y* dihitung dengan rumus:



1. Iterasi Menggunakan Centroid Baru

Setelah jarak setiap objek data ke centroid dihitung, objek data akan dikelompokkan ke dalam *cluster* dengan centroid terdekat. Kemudian, centroid baru untuk setiap klaster dihitung kembali menggunakan nilai rata-rata dari anggota klaster tersebut. Proses ini diulang secara iteratif hingga centroid konvergen (tidak ada perubahan anggota klaster).

1. Kelompokkan Objek Data Berdasarkan Jarak Minimum ke Centroid

Pada setiap iterasi, setiap objek data akan dikelompokkan ke dalam *cluster* dengan centroid terdekat berdasarkan jarak minimum yang dihitung sebelumnya.

1. Ada Objek yang Berpindah Kelompok?

Jika masih ada objek data yang berpindah ke klaster lain setelah iterasi terakhir, maka proses iterasi menggunakan centroid baru dilanjutkan kembali. Namun, jika tidak ada objek data yang berpindah klaster, maka proses clustering dianggap konvergen dan selesai.

1. Visualisasi Klaster

Sesuai dengan algoritma yang dilakukan. Penggunaan diagramnya bisa berbeda, namun tujuannya tetap sama, yaitu membuat laporan dalam bentuk gambar yang mudah dipahami.

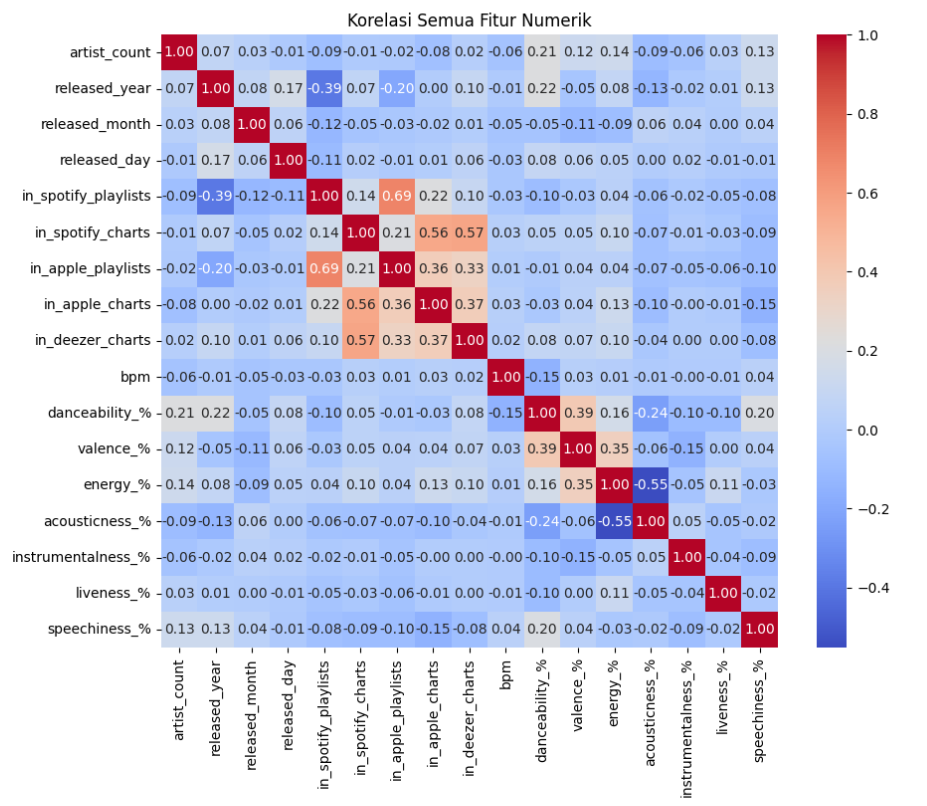
1. Selesai, tahap akhir dari proses.

# 

# BAB III PEMBAHASAN

## **Perhitungan Manual**

Berdasarkan hasil visualisasi matriks korelasi antar fitur numerik pada dataset, kelompok kami memilih tiga fitur utama, yaitu *danceability\_%*, *valence\_%*, dan *energy\_%* untuk dieksekusi lebih lanjut menggunakan metode K-Means Clustering. Pemilihan ketiga fitur ini didasarkan pada korelasi yang cukup kuat di antara mereka, seperti korelasi antara *valence\_%* dengan *energy\_%* sebesar 0.35, dan *valence\_%* dengan *danceability\_%* sebesar 0.39.



Gambar 2 Korelasi Semua Fitur Numerik

Ketiganya merepresentasikan karakteristik audio yang berkaitan erat dengan emosi, semangat, dan daya tarik dari sebuah lagu, yang sangat relevan dalam menentukan preferensi pendengar di Spotify. Dengan menggunakan fitur-fitur ini dalam metode K-Means Clustering, diharapkan dapat terbentuk pengelompokan lagu yang mencerminkan tren karakteristik audio yang dominan pada tahun 2023, sehingga mendukung tujuan utama penelitian dalam memahami pola popularitas lagu berdasarkan fitur audio secara lebih mendalam.

Sebagai data sampel perhitungan kami mengambil 10 baris data teratas seperti pada tabel dibawah ini.

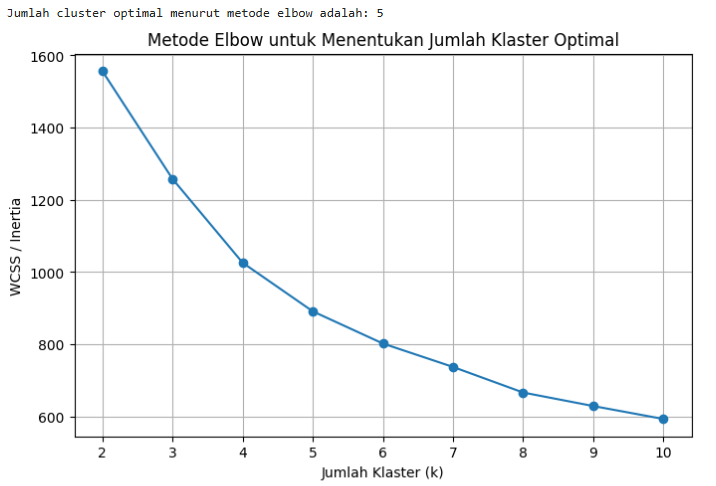
Table 1 Data Perhitungan Algoritma K-Means

|  |  |  |  |
| --- | --- | --- | --- |
| **Data** | **Danceability\_% (A)** | **Valence\_% (B)** | **Energy\_% (C)** |
| 1 | 80 | 89 | 83 |
| 2 | 71 | 61 | 74 |
| 3 | 51 | 32 | 53 |
| 4 | 55 | 58 | 72 |
| 5 | 65 | 23 | 80 |
| 6 | 92 | 66 | 58 |
| 7 | 67 | 83 | 76 |
| 8 | 67 | 26 | 71 |
| 9 | 85 | 22 | 62 |
| 10 | 81 | 56 | 48 |

### **Perhitungan K-means**

Algoritma K Means merupakan salah satu metode *partial clustering* berbasis titik pusat (*centroid*) dimana proses *clustering* dilakukan dengan meminimalkan jarak jumlah kuadrat antara data dengan masing-masing pusat klaster. Algoritma K-Means dalam penerapannya memerlukan tiga parameter yang seluruhnya ditentukan pengguna yaitu jumlah klaster k, inisialisasi klaster, dan jarak.

Dalam proses *clustering*, penentuan jumlah klaster yang optimal merupakan tahap penting untuk menghasilkan pengelompokan data yang representatif dan bermakna. Oleh karena itu, kami menggunakan metode Elbow untuk menentukan jumlah klaster terbaik berdasarkan nilai *Within-Cluster Sum of Squares* (WCSS) seperti yang terlihat pada gambar berikut ini:



Gambar 3 Metode Elbow untuk Jumlah klaster

Berdasarkan grafik Elbow yang dihasilkan, terlihat adanya “tekukan” atau titik siku pada nilai *k* = 5. Titik ini menunjukkan perubahan signifikan terhadap penurunan WCSS sebelum mulai melandai, sehingga jumlah klaster yang paling optimal adalah 5. Dengan demikian, dalam implementasi metode K-Means Clustering untuk menganalisis tren lagu di Spotify tahun 2023 berdasarkan karakteristik audio (*danceability\_%*, *valence\_%*, dan *energy\_%*), kelompok kami menetapkan jumlah klaster sebanyak lima untuk mendapatkan segmentasi lagu yang lebih informatif dan sesuai dengan distribusi data yang ada.

Berikut langkah-langkah melakukan clustering dengan menggunakan algoritma k-means:

1. Menentukan nilai K sebanyak jumlah klaster atau kelompok yang diinginkan. Jumlah klaster yang diambil pada data ini adalah sebanyak 5 kelompok yang terdiri dari:
2. **Lagu akustik-tenang (Klaster 1)**

Lagu-lagu dalam klaster ini cenderung memiliki nilai energy dan danceability yang rendah dengan suasana yang lebih lembut dan relaks. Biasanya menggunakan instrumen akustik atau minim efek elektronik. Cocok untuk didengarkan saat belajar, membaca, atau ketika ingin ketenangan.

1. **Lagu energik-tari (Klaster 2)**

Klaster ini terdiri dari lagu-lagu dengan nilai danceability dan energy yang tinggi. Lagu-lagu ini memiliki irama cepat dan beat yang kuat, cocok untuk suasana pesta, olahraga, atau aktivitas fisik lainnya. Sering kali masuk dalam genre EDM, dance pop, atau hip-hop energik.

1. **Lagu mellow-emosional (Klaster 3)**

Lagu-lagu pada klaster ini memiliki energy yang relatif sedang hingga rendah dan valence rendah. Umumnya menyampaikan emosi mendalam seperti kesedihan, kerinduan, atau introspeksi diri. Cocok untuk suasana melankolis atau reflektif.

1. **Lagu upbeat-ceria (Klaster 4)**

Lagu-lagu ini memiliki kombinasi nilai danceability dan valence yang tinggi. Karakteristiknya ceria, penuh semangat, dan meningkatkan mood. Umumnya cocok diputar saat suasana hati sedang baik atau ingin membangkitkan semangat positif.

1. **Lagu eksperimental atau unik (Klaster 5)**

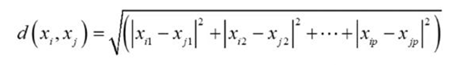
Lagu dalam kelompok ini tidak sepenuhnya sesuai dengan ciri khas empat klaster lainnya. Bisa jadi karena komposisi yang tidak biasa, kombinasi atribut yang ekstrem, atau aransemen yang inovatif. Lagu-lagu ini memiliki gaya khas yang menonjol dibanding lagu pada klaster lainnya.

1. Setelah menentukan nilai K sebanyak jumlah cluster yang diinginkan, lalu pilih sebanyak K data dari set data sebagai pusat klaster (*centroid*) secara random. Disini data random yang diambil yaitu pada baris ke 2,6, 8 dan 10 seperti Tabel 3.2 dibawah

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data | Centroid | A | B | C |
| 1 | 1 | 80 | 89 | 83 |
| 2 | 2 | 71 | 61 | 74 |
| 6 | 3 | 92 | 66 | 58 |
| 8 | 4 | 67 | 26 | 71 |
| 10 | 5 | 81 | 56 | 48 |

Table 2 Data Centroid Iterasi Pertama

1. Menghitung jarak antara objek dengan masing-masing centroid menggunakan rumus euclidean distance



Rumus 1 Euclidean distance

Setelah melakukan perhitungan untuk mencari jarak antar centroid, lalu kelompokkan data tersebut berdasarkan jarak terdekat dengan centroid nya dengan menggunakan nilai minimum kelima centroid tersebut.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data** | **A** | **B** | **C** | **Centroid** | | | | | **Min** | **Klas- ter** |
| **C1** | **C2** | **C3** | **C4** | **C5** |
| 1 | 80 | 89 | 83 | 0,0000 | 30,7571 | 36,0278 | 65,4370 | 48,1144 | 0,0000 | **1** |
| 2 | 71 | 61 | 74 | 30,7571 | 0,0000 | 26,8701 | 35,3553 | 28,3019 | 0,0000 | **2** |
| 3 | 51 | 32 | 53 | 70,6399 | 41,0122 | 53,4977 | 24,8193 | 38,7427 | 24,8193 | **4** |
| 4 | 55 | 58 | 73 | 41,0609 | 16,3095 | 40,7185 | 34,2345 | 36,1248 | 16,3095 | **2** |
| 5 | 65 | 23 | 80 | 67,7495 | 38,9358 | 55,3353 | 9,6954 | 48,6724 | 9,6954 | **4** |
| 6 | 92 | 66 | 58 | 36,0278 | 26,8701 | 0,0000 | 48,9285 | 17,9165 | 0,0000 | **3** |
| 7 | 67 | 83 | 76 | 15,9374 | 22,4499 | 35,1852 | 57,2189 | 41,3401 | 15,9374 | **1** |
| 8 | 67 | 26 | 71 | 65,4370 | 35,3553 | 48,9285 | 0,0000 | 40,3113 | 0,0000 | **4** |
| 9 | 85 | 22 | 62 | 70,3918 | 43,1393 | 44,7325 | 20,5183 | 36,9865 | 20,5183 | **4** |
| 10 | 81 | 56 | 48 | 48,1144 | 28,3019 | 17,9165 | 40,3113 | 0,0000 | 0,0000 | **5** |

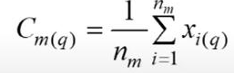
Table 3 Hasil Perhitungan Iterasi Pertama

Pada Tabel 3 terlihat bahwa jika suatu data dekat dengan centroid pertama atau nilai minimum nya pada centroid pertama maka akan dikelompokkan menjadi cluster 1 begitupun untuk data yang dekat pada centroid 2 dan 3 maka masing-masing nya akan masuk ke dalam cluster 2 dan cluster 3. Berikut hasil dari klaster pada Tabel 4 dibawah

Table 4 Hasil Klaster Iterasi Pertama

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **DATA** | **C1** | **C2** | **C3** | **C4** | **C5** |
| 1 | 1 |  |  |  |  |
| 2 |  | 1 |  |  |  |
| 3 |  |  |  | 1 |  |
| 4 |  | 1 |  |  |  |
| 5 |  |  |  | 1 |  |
| 6 |  |  | 1 |  |  |
| 7 | 1 |  |  |  |  |
| 8 |  |  |  | 1 |  |
| 9 |  |  |  | 1 |  |
| 10 |  |  |  |  | 1 |

1. Setelah mendapatkan kelompok pada iterasi pertama selanjutnya kita akan mencari pada iterasi kedua ataupun selanjutnya hingga tidak ada lagi objek yang berpindah cluster.
2. Pada iterasi kedua ini, tentukan terlebih dahulu centroid baru dengan menggunakan rumus sebagai berikut:



Rumus 2 Centroid

Cara melakukan perhitungan untuk menentukan centroid nya dengan rumus tersebut sebagai berikut:

1. Pada tabel 4 terlihat bahwa data yang masuk pada centroid 1 sebanyak 2 yaitu data ke-1 dan 7. Jadi untuk mencari centroid baru nya dengan membagi data ke 1 dan 7 dibagi dengan dua.
2. Pada tabel 4 untuk data yang masuk ke dalam centroid 2 sebanyak 2 terdapat pada data ke- 2, dan 4. Jadi untuk mencari centroid baru nya dengan membagi data ke 2, dan 4 dibagi dengan dua.
3. Pada tabel 4 terlihat bahwa data yang masuk pada centroid 3 hanya 1 pada baris 6. Jadi datanya tetap.
4. Pada tabel 4 untuk data yang masuk ke dalam centroid 4 sebanyak 4 buah terdapat pada data ke-3,5,8, dan 9. Jadi untuk mencari centroid baru nya dengan membagi data ke-3,5,8, dan 9 dibagi dengan empat.
5. Pada tabel 4 terlihat bahwa data yang masuk pada centroid 5 hanya 1 pada baris 10. Jadi datanya tetap.

Setelah mencari perhitungan tersebut maka akan dapat hasil perhitungan yang ada pada Tabel 5 berikut

Table 5 Data Centroid Iterasi Kedua

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Data | Centroid | A | B | C |
| 1,7 | 1 | 73,5 | 86 | 79,5 |
| 2,4 | 2 | 63 | 59,5 | 73 |
| 6 | 3 | 92 | 66 | 58 |
| 3,5,8,9 | 4 | 65,5 | 25,75 | 66,5 |
| 10 | 5 | 81 | 56 | 48 |

1. Setelah mendapatkan centroid baru, melakukan perhitungan kembali dengan rumus euclidean distance dan didapatkan hasil seperti Tabel 3.6 dibawah:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Data** | **A** | **B** | **C** | **Centroid** | | | | | **Min** | **Klas- ter** |
| **C1** | **C2** | **C3** | **C4** | **C5** |
| 1 | 80 | 89 | 83 | 7,9687 | 35,4859 | 36,0278 | 66,9557 | 48,1144 | 7,9687 | **1** |
| 2 | 71 | 61 | 74 | 25,7196 | 8,2006 | 26,8701 | 36,4563 | 28,3019 | 8,2006 | **2** |
| 3 | 51 | 32 | 53 | 64,2223 | 36,0590 | 53,4977 | 20,7741 | 38,7427 | 20,7741 | **4** |
| 4 | 55 | 58 | 73 | 34,3875 | 8,2006 | 40,3609 | 34,3593 | 35,4401 | 8,2006 | **2** |
| 5 | 65 | 23 | 80 | 63,5728 | 37,2189 | 55,3353 | 13,7863 | 48,6724 | 13,7863 | **4** |
| 6 | 92 | 66 | 58 | 34,7059 | 33,2904 | 0,0000 | 48,9343 | 17,9165 | 0,0000 | **3** |
| 7 | 67 | 83 | 76 | 7,9687 | 24,0260 | 35,1852 | 58,0522 | 41,3401 | 7,9687 | **1** |
| 8 | 67 | 26 | 71 | 60,9467 | 33,7972 | 48,9285 | 4,7500 | 40,3113 | 4,7500 | **4** |
| 9 | 85 | 22 | 62 | 67,3387 | 44,8470 | 44,7325 | 20,3608 | 36,9865 | 20,3608 | **4** |
| 10 | 81 | 56 | 48 | 44,1418 | 31,0040 | 17,9165 | 38,6984 | 0,0000 | 0,0000 | **5** |

Table 6 Hasil Perhitungan Iterasi Kedua

Pada Tabel 6 terlihat bahwa hasil klaster yang diperoleh sama dengan hasil klaster pada iterasi pertama. Hal ini menunjukkan bahwa proses pengelompokan telah berjalan dengan stabil dan konsisten, serta tidak terjadi perubahan signifikan pada iterasi berikutnya. Dengan demikian, hasil klaster tersebut dapat dianggap valid dan dapat digunakan untuk analisis lebih lanjut, sebagaimana ditampilkan pada Tabel 7 di bawah ini:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **DATA** | **C1** | **C2** | **C3** | **C4** | **C5** |
| 1 | 1 |  |  |  |  |
| 2 |  | 1 |  |  |  |
| 3 |  |  |  | 1 |  |
| 4 |  | 1 |  |  |  |
| 5 |  |  |  | 1 |  |
| 6 |  |  | 1 |  |  |
| 7 | 1 |  |  |  |  |
| 8 |  |  |  | 1 |  |
| 9 |  |  |  | 1 |  |
| 10 |  |  |  |  | 1 |

Table 7 Hasil Klaster Iterasi Kedua

Kesimpulan :

Berdasarkan hasil perhitungan K-Means yang telah mencapai konvergensi pada iterasi kedua, setiap lagu dari sampel data dapat dikelompokkan ke dalam lima klaster utama yang merepresentasikan suasana dan karakteristik audio. Proses ini bertujuan untuk mengelompokkan lagu-lagu di Spotify tahun 2023 berdasarkan atribut seperti tempo, energi, danceability, dan parameter lainnya yang menggambarkan suasana atau nuansa lagu.

Dari hasil pengelompokan yang ditampilkan pada Tabel 3.7, terlihat bahwa Klaster 4 merupakan klaster yang paling banyak berisi lagu (4 lagu), sedangkan klaster lainnya terdiri dari 1 hingga 2 lagu. Hal ini menunjukkan bahwa karakteristik lagu yang tergolong dalam klaster 4 lebih umum atau lebih dominan dalam dataset yang dianalisis. Masing-masing lagu dikaitkan ke klaster berdasarkan nilai minimum jarak (inertia) terhadap centroid, sehingga memberikan hasil pengelompokan yang optimal dan representatif.

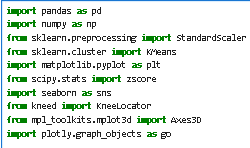
Secara keseluruhan, hasil klasterisasi ini memberikan gambaran yang jelas mengenai pengelompokan tren lagu berdasarkan ciri audio. Dengan pendekatan ini, analisis tren musik dapat dilakukan dengan lebih sistematis, membantu dalam memahami preferensi pendengar dan karakteristik musik yang populer di tahun 2023.

## **Implementasi**

Implementasi dilakukan dengan menggunakan bahasa pemrograman python pada jupyter notebook. Adapun tahapan-tahapan yang dilakukan adalah sebagai berikut:

### **Import Library**

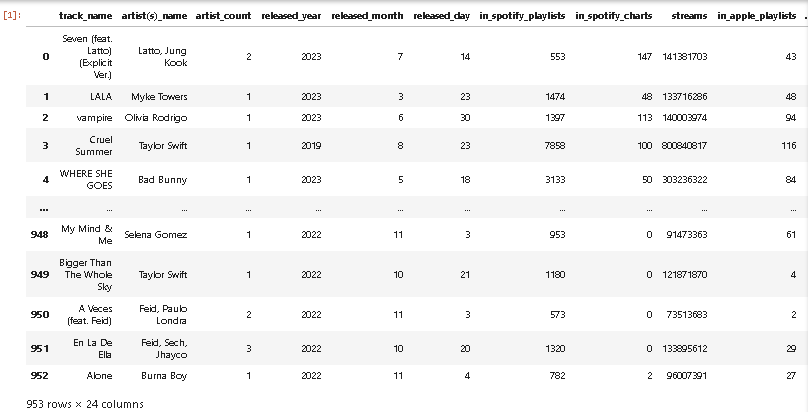
Melakukan import library yang akan digunakan dan dibutuhkan selama implementasi. Adapun library-library yang digunakan adalah seperti yang terdapat pada gambar dibawah ini.



Gambar 4 Import Library

### **Data Gathering**

Mengumpulkan dan menampilkan data yang dibutuhkan untuk dianalisis. Dalam implementasi dataset yang digunakan yaitu dataset “Spotify 2023” dengan menggunakan library pandas untuk menampilkan dataset. Dataset ini memiliki 953 baris yang terdiri dari 24 kolom seperti yang terdapat pada gambar dibawah ini.



Gambar 5 Data Gathering

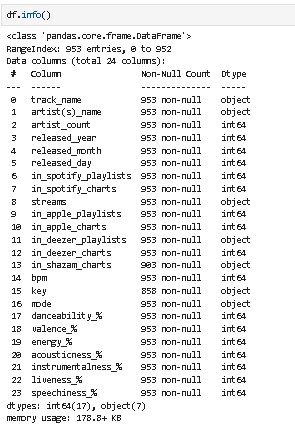
### **Data Preparation**

Menampilkan informasi ukuran (dimensi) dari data frame yang akan digunakan. dengan kode dan tampilan hasil seperti gambar dibawah ini.



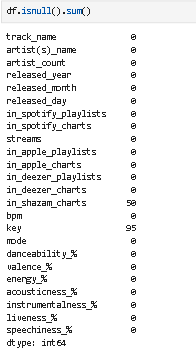
Gambar 6 Data Shape

Menampilkan ringkasan informasi struktur data frame dari setiap kolom yang ada, tipe data setiap kolom, jumlah data yang tidak kosong, penggunaan memori dan informasi lainnya seperti yang terdapat pada gambar dibawah ini.



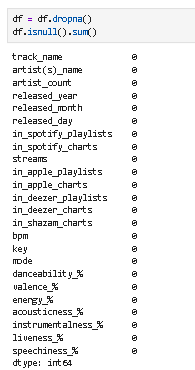
Gambar 7 Data Info

Mengecek dan menghitung jumlah nilai null yang terdapat pada data frame. Pada dataset yang digunakan diketahui pada kolom in\_shazam\_charts terdapat 50 nilai null dan 95 nilai null pada kolom key, hasil yang diperoleh seperti pada gambar di bawah ini.



Gambar 8 Data Null

Melakukan penanganan data nilai null dengan cara menghapus baris yang terdapat nilai null. Setelah dilakukan penghapusan nilai null, ditampilkan kembali jumlah nilai null setelah dilakukan penanganan missing value. Hasil data yang diperoleh setelah penghapusan nilai null seperti pada gambar dibawah ini.



Gambar 9 Penanganan Data NUll

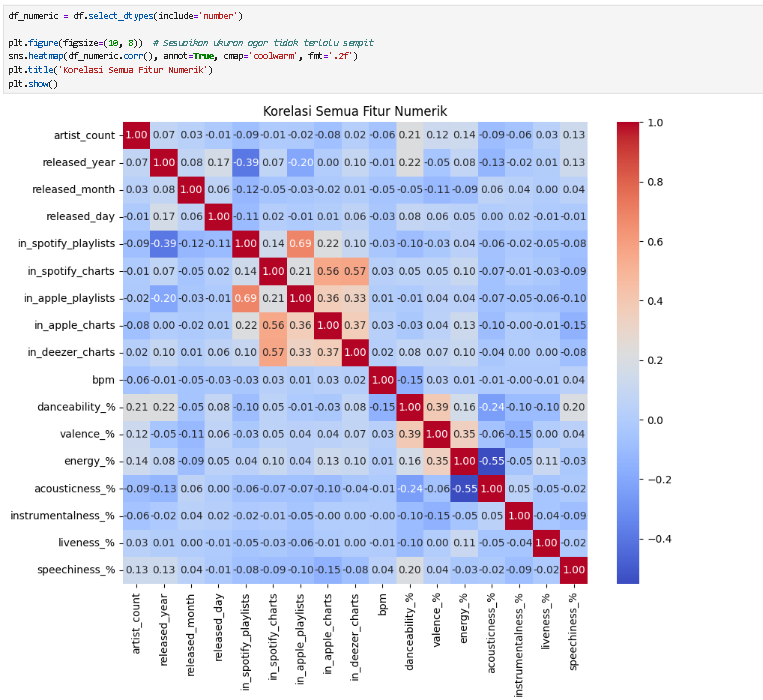
Melakukan pengecekan dan menghitung jumlah data duplikat. Pada dataframe yang digunakan tidak ditemukan data duplikat, seperti yang terlihat pada gambar dibawah ini.



Gambar 10 Re-Check Data NULL

### **Data Analysis**

Menampilkan Heatmap untuk mengevaluasi korelasi antar semua fitur numerik yang terdapat pada dataset. Heatmap ini menunjukkan nilai korelasi Pearson antar fitur numerik. Warna merah menunjukkan korelasi positif, biru untuk negatif. Nilai mendekati 1 atau -1 menunjukkan hubungan kuat.



Gambar 11 Data Analysis

Berdasarkan evaluasi Heatmap dilakukan pemilihan 3 fitur audio yang memiliki korelasi saling berkaitan antara fitur-fitur audio lainnya. Pemilihan fitur danceability, valence, dan energy didasarkan pada keterkaitan tematik (musikal) dan korelasi sedang antar fitur, yang menghindari multikolinearitas. 3 fitur audio yang dipilih adalah Danceability, Valence, dan Energy. Alasan pemilihan 3 fitur ini dipilih adalah:

1. Danceability vs Valence (0.39)

Terdapatkorelasi positif sebesar 0.39 antara danceability dan valence. Artinya, semakin tinggi tingkat keceriaan atau mood positif (valence) dalam sebuah lagu, cenderung semakin tinggi pula tingkat kemudahannya untuk ditarikan (danceability).

1. Danceability vs Energy (0.30)

Danceability memiliki korelasi positif sebesar 0.30 dengan energy. Ini menunjukkan bahwa lagu yang lebih energik (lebih cepat, keras, atau intens) cenderung juga lebih mudah untuk ditarikan.

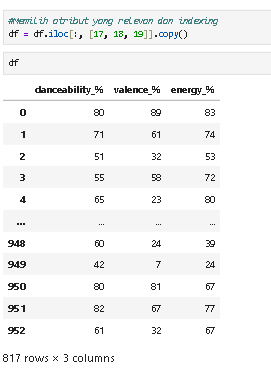
1. Valence vs Energy (0.35)

Korelasi antara valence dan energy sebesar 0.35 menunjukkan bahwa lagu dengan suasana yang lebih ceria atau positif cenderung juga memiliki tingkat energi yang lebih tinggi.

Ketiga fitur tersebut saling berkorelasi secara positif, namun tidak terlalu kuat (sedang) tanpa membuat model terlalu redundant. Hal ini mengindikasikan bahwa masing-masing fitur membawa informasi unik, dan sangat cocok digunakan bersama-sama dalam analisis clustering.

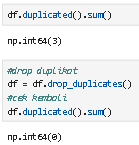
### **Implementasi K-Means**

1. Mengimplementasikan pemilihan 3 fitur yang telah dianalisis sebelumnya. Fitur yang dipilih merupakan kolom 17, 18, dan 19 pada dataframe. Selanjutnya, Dataframe ke-3 fitur ditampilkan yang terdiri dari 817 baris dan 3 kolom. Seperti yang terlihat pada gambar dibawah ini.



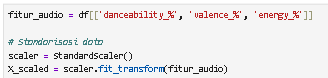
Gambar 12 Memilih Atribut

1. Melakukan pengecekan duplikat hanya untuk 3 fitur yang dipilih. Pada 3 fitur ditemukan 3 data duplikat, sehingga dilakukan penanganan dengan menghapus data duplikat tersebut. kemudian kembali mengecek data duplikat setelah dilakukan penanganan data duplikat.



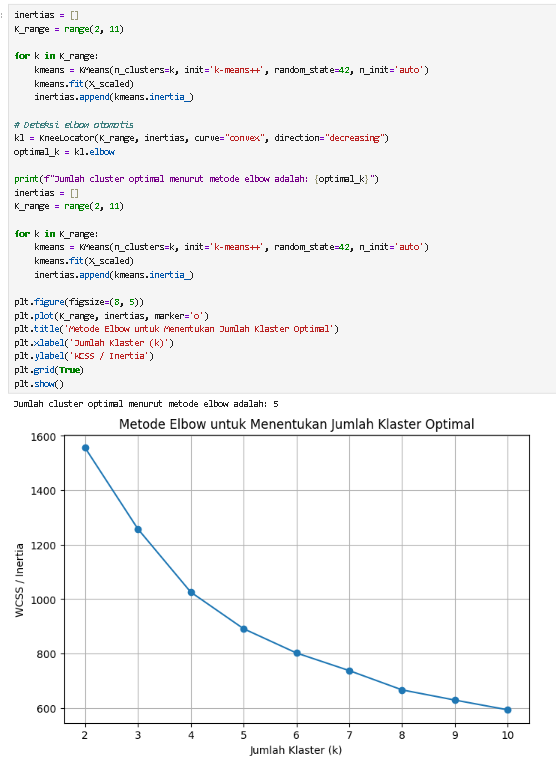
Gambar 13 Pengecekan Duplikat

1. Melakukan standarisasi data fitur audio yang dipilih sebelum dilakukan analisis. proses standarisasi menggunakan StandardScaler dari pustaka Scikit-Learn. Standarisasi ini bertujuan untuk mengubah skala data agar memiliki rata-rata (mean) 0 dan standar deviasi 1, sehingga setiap fitur memiliki kontribusi yang setara dalam analisis.



Gambar 14 Standarisasi Data Fitur Audio

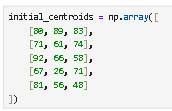
1. Menentukan jumlah cluster optimal algoritma K-Means menggunakan metode Elbow. Pertama, dilakukan iterasi K-Means untuk berbagai nilai k (jumlah klaster) dari 2 hingga 10, dan dihitung nilai inertia (WCSS / Within-Cluster Sum of Squares) untuk masing-masing k. Nilai inertia ini menunjukkan seberapa baik data dikelompokkan; semakin kecil inertia, semakin baik. Selanjutnya, digunakan fungsi KneeLocator untuk mendeteksi titik siku (elbow point) pada grafik inertia. Titik siku ini mengindikasikan jumlah cluster optimal, yaitu saat penurunan inertia mulai melambat secara signifikan. Berdasarkan grafik Elbow dan hasil perhitungan otomatis, jumlah cluster optimal ditentukan sebanyak 5 klaster, yang akan digunakan dalam proses clustering selanjutnya. Seperti yang terdapat pada gambar dibawah ini.



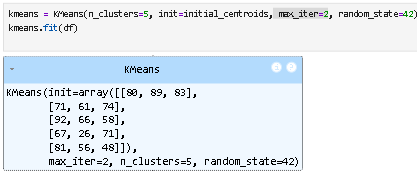
Gambar 15 Menentukan Jumlah Cluster

#### Implementasi Perhitungan Manual K-Means Menggunakan 10 Data Teratas, Penentuan Centroid Manual, dan Menggunakan Data Asli

1. Menentukan manual centroid dengan menggunakan data ke 1, 2, 6, 8, dan 10 dengan jumlah cluster k=5. Nilai data centroid seperti pada gambar dibawah ini.

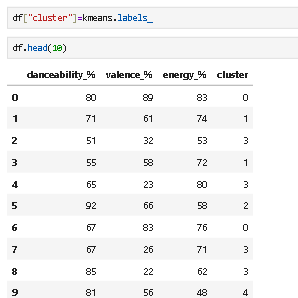


Gambar 16 Menentukan Centroid

1. Pelatihan model K-Means dengan jumlah klaster sebanyak 5, berdasarkan hasil penentuan jumlah cluster optimal sebelumnya pada metode Elbow. Pada implementasi ini digunakan centroid awal yang ditentukan secara manual melalui parameter init. Penggunaan centroid awal yang eksplisit ini bertujuan untuk mengamati secara langsung bagaimana model membentuk klaster dari titik awal tertentu sesuai dengan perhitungan manual yang dilakukan. Parameter max\_iter diatur menjadi 2, yang berarti proses iterasi hanya dilakukan sebanyak dua kali sesuai perhitungan manual dilakukan sampai iterasi ke-2. Selain itu, random\_state=42 digunakan agar hasil eksperimen tetap konsisten dan dapat direproduksi. Dengan konfigurasi ini, model akan langsung menggunakan centroid yang diberikan dan menghentikan iterasi setelah dua langkah pembaruan centroid, tanpa melakukan pengulangan inisialisasi (karena n\_init secara otomatis dianggap satu oleh Scikit-Learn saat init diberikan secara manual). Proses lengkapnya seperti yang terdapat pada gambar dibawah ini.

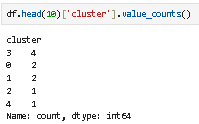
Gambar 17 Pelatihan Model K-Means

1. Setelah model K-Means berhasil dilatih, langkah selanjutnya adalah melakukan penentuan label klaster untuk setiap data. Hal ini dilakukan dengan mengambil atribut labels dari objek K-Means, yang berisi hasil pengelompokan (klaster) dari masing-masing baris data. Label klaster tersebut kemudian disimpan ke dalam kolom baru bernama "cluster" pada DataFrame utama. Dengan cara ini, setiap data akan memiliki informasi tambahan mengenai klaster mana yang menjadi representasinya seperti yang terdapat pada gambar dibawah ini.



Gambar 18 Penentuan Label Cluster

1. Analisis distribusi klaster untuk mengetahui distribusi jumlah data per klaster pada subset DataFrame yang terdiri dari 10 baris pertama. Dengan menggunakan fungsi .(value\_counts()) pada kolom “Cluster” adalah analisis statistik dari label klaster tersebut, yaitu menghitung berapa banyak data yang masuk ke masing-masing klaster, khususnya hanya dari 10 data pertama yang ditampilkan dengan .head(10). Hasilnya menunjukkan bahwa klaster 3 mendominasi dengan 4 data, diikuti oleh klaster 0 dan 1 masing-masing dengan 2 data, serta klaster 2 dan 4 yang masing-masing memiliki 1 data seperti yang terlihat pada gambar dibawah ini.

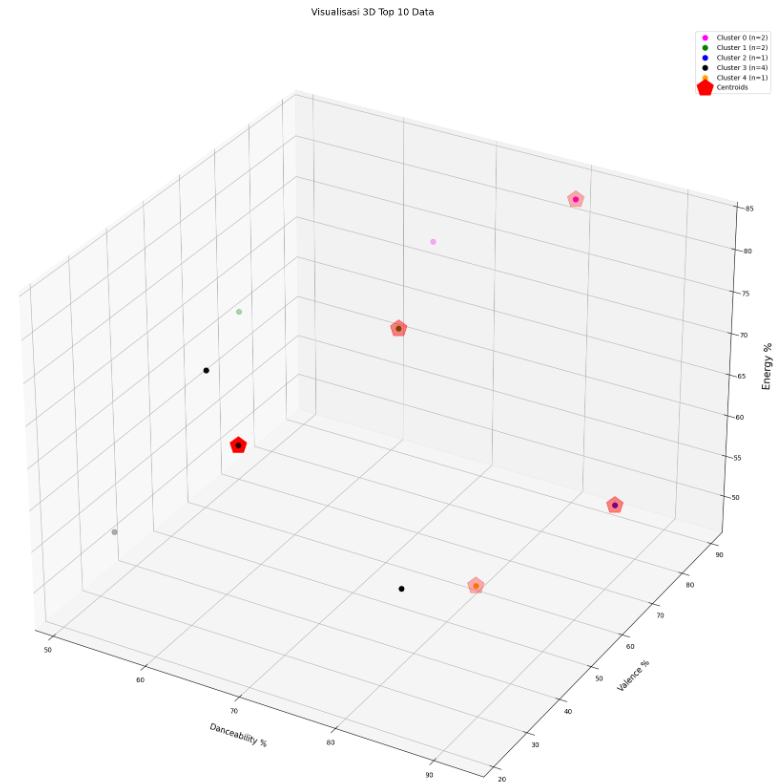


Gambar 19 Distribusi Cluster 10 Data

1. Visualisasi 3D hasil clusterisasi pada 10 data pertama setelah pemodelan dan perhitungan dengan K-Means. Seperti yang terlihat pada gambar dibawah ini.



Gambar 20 Visualisasi



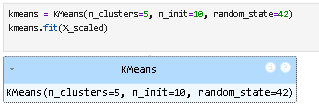
Gambar 21 Visualisasi

Visualisasi diatas menampilkan representasi 3D dari cluster trend lagu untuk 10 lagu teratas (Top 10 Data) berdasarkan fitur audio *Danceability*, *Valence*, dan *Energy*. Setiap titik kecil merepresentasikan satu lagu, yang telah dikelompokkan ke dalam klaster yang sama seperti pada dataset penuh. Warna titik menandakan klaster masing-masing lagu, sedangkan simbol pentagon merah besar menunjukkan posisi centroid dari masing-masing klaster berdasarkan seluruh data.

Tujuan dari visualisasi ini adalah untuk melihat kecenderungan atau sebaran lagu-lagu terpopuler terhadap klaster-klaster yang telah terbentuk sebelumnya. Dengan membandingkan posisi lagu-lagu ini terhadap centroid, kita dapat memahami karakteristik umum dari lagu populer, misalnya apakah mereka cenderung memiliki *Danceability* tinggi, *Valence* rendah, atau *Energy* sedang hingga tinggi. Visualisasi ini juga memperlihatkan bahwa lagu-lagu populer tersebar di beberapa klaster, menandakan variasi gaya atau suasana dalam daftar teratas.

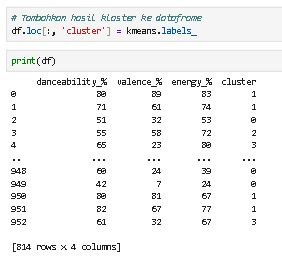
#### Implementasi K-Means Menggunakan Seluruh Data dengan Standarisasi Data

1. Proses pelatihan model K-Means dengan parameter yang disesuaikan berdasarkan hasil identifikasi jumlah cluster optimal sebelumnya. Model ditentukan untuk membentuk 5 klaster (n\_clusters=5) dengan nilai n\_init=10, artinya algoritma akan dijalankan sebanyak 10 kali dengan inisialisasi centroid yang berbeda secara acak, dan hasil terbaik (dengan inersia terendah) akan dipilih. Parameter random\_state=42 digunakan untuk menjamin reprodusibilitas atau hasil yang konsisten setiap kali kode dijalankan ulang. Tidak adanya parameter max\_iter berarti model akan menggunakan nilai default dari max\_iter, yaitu 300 iterasi maksimum, yang memungkinkan proses K-Means berjalan hingga konvergen secara otomatis. Tahapan ini bertujuan untuk memperoleh klasterisasi yang stabil dan optimal berdasarkan fitur yang telah distandarisasi sebelumnya. Kode pelatihan model K-Means seperti yang terdapat pada gambar dibawah ini.

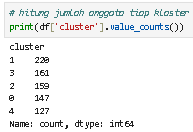


Gambar 22 Pelatihan Model K-Means

1. Hasil klasterisasi yang diperoleh dari algoritma K-Means ditambahkan ke dalam DataFrame utama dengan menambahkan kolom baru bernama cluster. Nilai-nilai pada kolom ini berasal dari atribut labels\_ milik objek KMeans, yang berisi informasi mengenai klaster keberapa setiap data dikategorikan berdasarkan hasil pemodelan. Dengan adanya kolom ini, setiap data dapat diidentifikasi berdasarkan kelompok atau segmen yang memiliki karakteristik audio yang mirip, seperti danceability, valence, dan energy. Penambahan hasil klaster ke dalam DataFrame ini bertujuan untuk mempermudah analisis lebih lanjut terhadap masing-masing klaster, seperti memahami pola dominan, melakukan visualisasi, hingga pengambilan keputusan berbasis segmen. Kemudian dilanjutkan dengan cetak perhitungan jumlah anggota di setiap cluster seperti pada gambar dibawah ini.



Gambar 23 Hasil Cluster



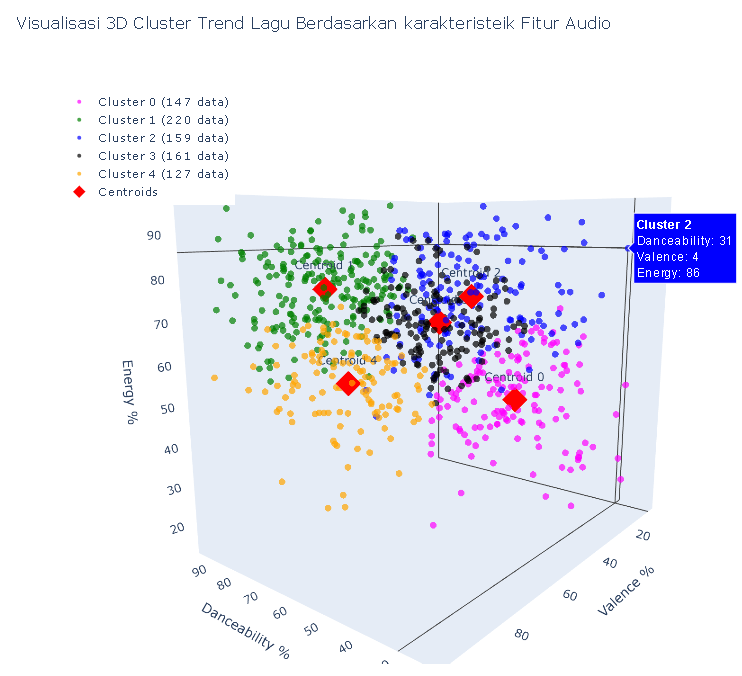
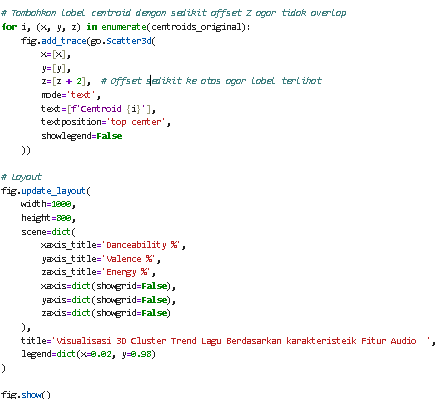
Gambar 24 Jumlah Anggota Tiap Cluster

1. Melanjutkan proses untuk mengubah koordinat pusat klaster (centroids) hasil algoritma *K-Means* dari skala data yang telah dinormalisasi kembali ke skala asli data. Pertama, centroids\_scaled diperoleh dari atribut cluster\_centers\_ milik objek kmeans, yang merepresentasikan posisi pusat cluster pada data yang telah diskalakan (menggunakan StandardScaler). Selanjutnya, nilai-nilai ini dikembalikan ke bentuk aslinya melalui fungsi inverse\_transform() dari objek scaler. Hasilnya disimpan dalam variabel centroids\_original, yang berisi koordinat pusat klaster dalam skala data awal sebelum normalisasi. Proses ini penting agar hasil klasterisasi dapat diinterpretasikan secara lebih intuitif dan relevan terhadap data asli.



Gambar 25 Mengubah Centroid

1. Visualisasi 3D Cluster Trend Lagu Berdasarkan karakteristik Fitur Audio berdasarkan hasil proses K-Means untuk melihat persebaran data cluster.



Gambar 26 Visualisasi

Visualisasi diatas merupakan representasi 3D clustering terhadap tren lagu berdasarkan tiga fitur audio utama, yaitu Danceability, Valence, dan Energy. Setiap titik dalam grafik mewakili satu lagu, yang telah dikelompokkan ke dalam lima klaster menggunakan algoritma *K-Means*. Warna yang berbeda menunjukkan klaster yang berbeda, sedangkan simbol berlian merah menandai posisi centroid atau pusat masing-masing klaster.

Tujuan visualisasi ini adalah untuk menunjukkan bagaimana lagu-lagu dikelompokkan berdasarkan kemiripan karakteristik audio. Misalnya, lagu dalam klaster 2 (berwarna biru) memiliki nilai *Energy* yang tinggi, namun *Danceability* dan *Valence* yang relatif rendah, sebagaimana ditunjukkan dalam kotak informasi di sisi kanan. Visualisasi ini membantu dalam memahami pola umum atau tren dari lagu-lagu dalam dataset berdasarkan aspek-aspek emosional dan ritmisnya.

# BAB IV PENUTUP

## **Kesimpulan**

Penelitian ini bertujuan untuk menganalisis tren lagu-lagu populer di Spotify pada tahun 2023 dengan melihat karakteristik audio lagu menggunakan metode K-Means Clustering. Hasil penelitian menunjukkan bahwa karakteristik audio lagu dapat dikelompokkan berdasarkan fitur seperti danceability, valence, dan energy. Dengan menggunakan K-Means Clustering, lagu-lagu tersebut terbagi ke dalam lima kelompok yang mewakili suasana atau genre tertentu, seperti lagu yang akustik dan tenang, lagu energik untuk menari, lagu mellow yang emosional, lagu ceria yang upbeat, dan lagu-lagu eksperimental yang unik.

Untuk menentukan jumlah kelompok yang paling optimal, digunakan metode Elbow, yang menunjukkan bahwa lima kelompok adalah pilihan terbaik. Pengelompokan ini membantu memberikan gambaran lebih jelas tentang tren musik berdasarkan karakteristik audio yang ada, seperti hubungan antara mood atau keceriaan (valence) dengan tingkat energi (energy), serta kemudahan untuk menari (danceability). Hasil pengelompokan ini memberi wawasan tentang pola musik yang lebih umum dan tren musik yang dominan di spotify pada tahun 2023.

Hasil dari penelitian ini dapat memberikan informasi yang sangat berguna bagi industri musik, pembuat konten, dan advertiser untuk merancang strategi pemasaran yang lebih efektif. Selain itu, hasil ini juga bisa digunakan untuk meningkatkan fitur rekomendasi musik yang lebih personal di platform seperti Spotify.

## **Saran**

Pada penelitian ini, seharusnya kami bisa mengeksplorasi lebih banyak fitur audio yang dapat mempengaruhi tren lagu, seperti tempo, loudness, atau acousticness. Penambahan fitur-fitur ini dapat memberikan gambaran yang lebih mendalam tentang karakteristik audio lagu dan membantu meningkatkan hasil pengelompokan.

Selain itu, penggunaan teknik clustering lainnya, seperti DBSCAN atau Agglomerative Hierarchical Clustering, bisa menjadi pilihan untuk melihat apakah ada hasil yang lebih stabil atau dapat menangani outlier dengan lebih baik.Serta pengujian model menggunakan dataset yang lebih besar atau dari tahun-tahun sebelumnya juga akan memberikan perspektif yang lebih luas tentang pola dan perubahan tren musik.

# DAFTAR PUSTAKA