PRAKTIKUM DATA MINING

PENERAPAN K-MEANS CLUSTERING PADA DATASET SPOTIFY

DISUSUN OLEH:
Kelompok 3

1.	Mujadid Choirus Surya	121450015
2.	Pramudya Wibowo	121450030
3.	A Rafi Paringgom Iwari	121450039
4.	Veni Zahara Kartika	121450075
5.	M. Faqih	121450120

PROGRAM STUDI SAINS DATA
FAKULTAS SAINS
INSTITUT TEKNOLOGI SUMATERA
2023/2024

LAPORAN PRAKTIKUM DATA MINING

PENERAPAN K-MEANS CLUSTERING PADA DATASET SPOTIFY

Mujadid Choirus Surya[1)], Pramudya Wibowo[2)], A Rafi Paringgom Iwari[3)], Veni Zahara Kartika[4)], M.Faqih[5)]

Program Studi Sains Data, Jurusan Sains, Institut Teknologi Sumatera

Email : mujadid.121450015@student.itera.ac.id[1)], pramudya.121450030@student.itera.ac.id[2)], arafi.121450039@student.itera.ac.id[3)], veni.121450075@student.itera.ac.id[4)], mfaqih.121450120@student.itera.ac.id[5)]

Abstrak

Industri musik menghadapi perubahan signifikan, terutama dengan kemajuan teknologi dan ketersediaan data yang melimpah. Penerapan algoritma klasterisasi menjadi penting untuk memahami struktur dan pola di dalam dataset tersebut. Penelitian ini diharapkan dapat memberikan kontribusi positif dalam konteks pengembangan konten, pemasaran, dan strategi bisnis di industri musik. Hasil dari penerapan clustering KMeans 2D terbentuk menjadi tiga cluster yakni label 0, label 1, dan label 2 berdasarkan kolom 'instrumentalness' dan 'track_popularity'. Sedangkan Hasil dari clustering 3D terbentuk menjadi tiga cluster yakni label 0, label 1, dan label 2 berdasarkan kolom ‘danceability’, 'instrumentalness' dan 'track_popularity'. Kedua metode KMeans secara 2D ataupun 3D ternyata mampu menemukan insight menarik terhadap pengelolaan strategi pemasaran yang lebih baik, berdasarkan kondisi dari masing-masing pendengar dan penikmat pemusik.

Kata Kunci : Clustering, KMeans, Data Preparation, dan Data Mining**.**

1. Pendahuluan

Dalam era transformasi digital saat ini, industri musik menghadapi perubahan signifikan, terutama dengan kemajuan teknologi dan ketersediaan data yang melimpah. Platform musik daring, seperti Spotify, menyimpan jumlah besar data yang mencerminkan preferensi dan kebiasaan mendengarkan pengguna. Dalam konteks ini, penerapan algoritma klasterisasi menjadi penting untuk memahami struktur dan pola di dalam dataset tersebut. Penelitian ini berfokus pada pemanfaatan metode k-means dalam mengelompokkan data dari Spotify, dengan tujuan mengidentifikasi kelompok pengguna berdasarkan karakteristik musik yang serupa.

Laporan ini mengeksplorasi potensi algoritma k-means dalam menggolongkan preferensi musik pengguna Spotify. Eksplorasi ini tidak hanya memberikan wawasan mendalam tentang kesamaan dalam perilaku mendengarkan, tetapi juga membuka peluang untuk meningkatkan personalisasi konten musik. Melalui pendekatan ini, diharapkan penelitian ini akan memberikan kontribusi berharga bagi pemahaman industri musik modern dan menginspirasi pengembangan strategi personalisasi yang lebih efektif.

Dengan memahami dinamika klaster pada data Spotify, penelitian ini diharapkan dapat memberikan kontribusi positif dalam konteks pengembangan konten, pemasaran, dan strategi bisnis di industri musik. Hasil penelitian diharapkan dapat membantu pengambilan keputusan yang lebih baik, serta memberikan landasan untuk pengembangan inovatif dalam penyajian musik secara daring.

2. Metode

2.1 Pengertian Data Mining

Data mining adalah tentang menemukan fitur-fitur penting dari data yang digunakan dan menemukan model yang menggambarkan kelas data atau konsep data. Data mining menggunakan berbagai teknik pembelajaran komputer untuk menganalisis dan mengekstrak pengetahuan secara otomatis. Data mining memungkinkan Anda melakukan proses berulang dan interaktif untuk mencari model dan pola baru, dan juga untuk menemukan hubungan antara nilai atribut dalam suatu data. Data mining adalah serangkaian proses yang digunakan untuk menemukan nilai dari koleksi data berupa pengetahuan yang tidak diketahui secara manual.

2.2 Deskripsi Dataset

Dataset yang digunakan berasal dari kaggle,kita menggunakan dataset spotify,Dataset spotify adalah kumpulan data yang berisi informasi tentang lagu-lagu yang ada di Spotify. Dataset ini mencakup informasi seperti (track_id, track_name, track_artist, track popularity, track_album_id, track_album_name, track_album_release_date, playlist_name, playlist_id, playlist_genree, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration_ms).

Dataset Spotify tersedia dalam berbagai ukuran, mulai dari beberapa ribu lagu hingga jutaan lagu. Dataset yang lebih besar biasanya lebih komprehensif dan dapat memberikan informasi yang lebih akurat tentang tren musik.

2.3 Clustering

Clustering dapat diartikan sebagai proses pengorganisasian kelompok-kelompok data ke dalam kelompok-kelompok yang berbeda sedemikian rupa sehingga objek-objek yang serupa menjadi anggota suatu cluster dan objek-objek yang berbeda menjadi anggota cluster yang lain. Setiap cluster berisi data yang semirip mungkin. Ukuran kemiripan biasanya dihitung berdasarkan jarak. Jarak dalam cluster dijaga sekecil mungkin, dan jarak antar cluster dipilih sebesar mungkin. Jadi di satu cluster harus sama dan berbeda di cluster lain. Definisi ini mengasumsikan bahwa terdapat beberapa parameter penting yang mewakili persamaan atau ketidaksamaan antar cluster[1].

2.4 Algoritma K-means

K-Means merupakan salah satu teknik clustering data non-hierarki yang mengelompokkan data dalam bentuk satu atau lebih klaster/grup. K-Means merupakan

algoritma clustering yang memiliki metode partisi berbasis pusat selain algoritma k-Medoids berbasis objek.

Berikut langkah-langkah algoritma K-Means yaitu.

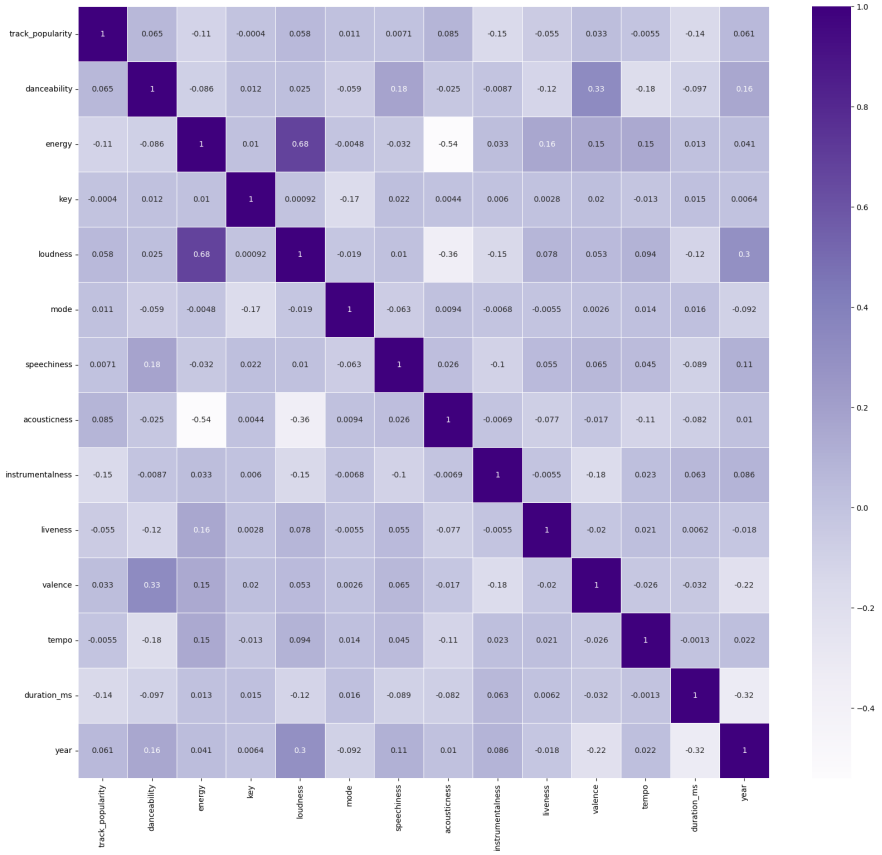
1. Tentukan jumlah cluster (k) pada dataset sebagai nilai centroid.
2. Hitung jarak antara data Anda dan pusat cluster menggunakan rumus jarak Euclidean.
3. Pusat cluster baru ditentukan ketika seluruh data telah ditugaskan ke cluster terdekat.
4. Proses penentuan pusat cluster dan penempatan data ke dalam cluster diulang terus menerus hingga nilai centroid tidak berubah lagi.

3. Hasil

3.1. Data Preprocessing, Understanding, dan Preparation

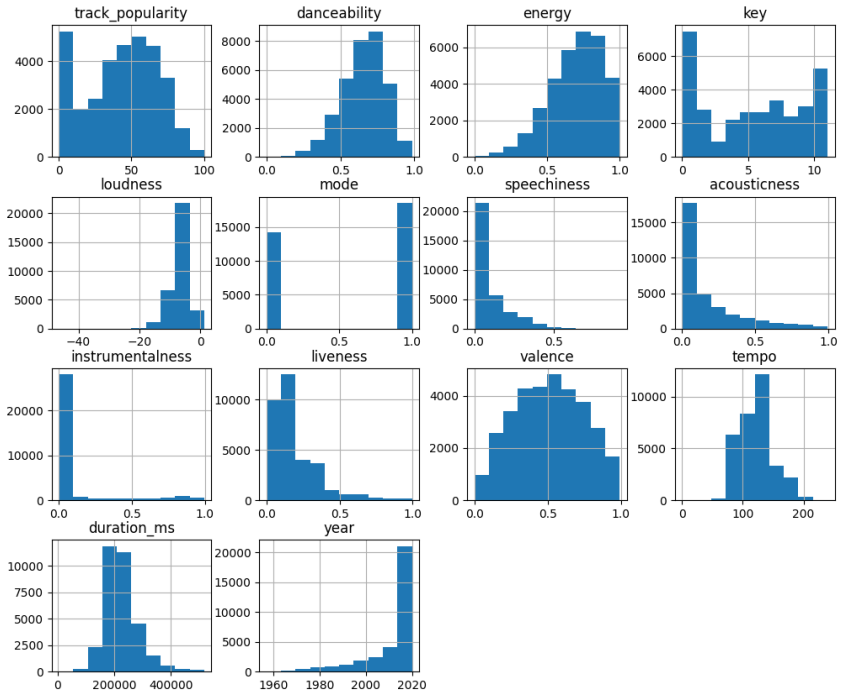
Pada proses ini, dataset spotify akan terlebih dahulu dilakukan data preprocessing yakni pengecekkan terhadap data duplicate dan missing value. Hasilnya ternyata didalam dataset spotify tidak terdapat data duplicate namun terdapat missing value pada fitur track_name, track_artist, dan track_album_name sebanyak 5 buah sehingga seluruh missing value akan dihapus untuk memberikan kualitas data yang baik. Selain itu, dilakukan beberapa penghapusan kolom atribut yang dianggap tidak terlalu penting dalam proses clustering K-Means yakni kolom track_id, track_album_id, dan playlist_id.

Selanjutnya pada tahap data understanding dan data preparation dilakukan konversi tipe data menjadi data datetime dan menambahkan kolom baru yakni kolom year untuk mempermudah melihat insight menarik dari data. Selain itu, dilakukan summary data statistika dan hubungan antar setiap atribut dengan atribut lainnya seperti pada **Gambar 3.1.1**.



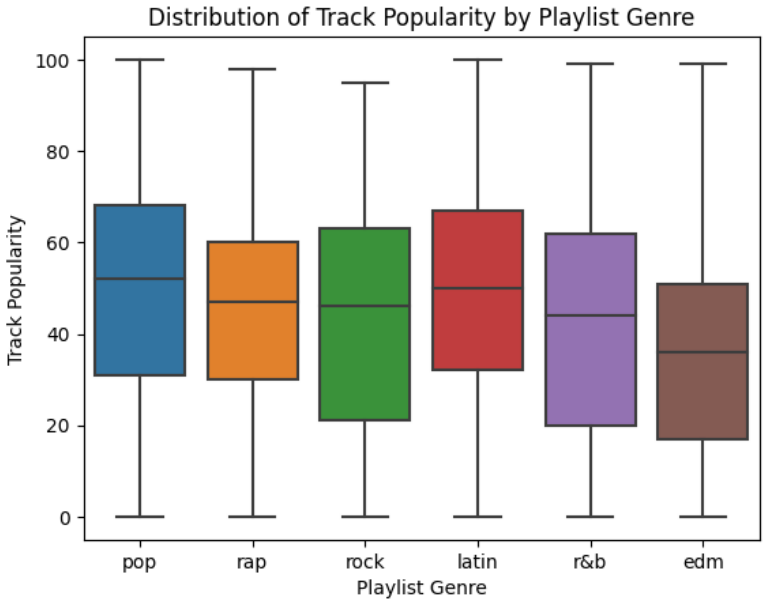
Gambar 3.1.1. Korelasi antara atribut

Hasil dari plot korelasi pada **Gambar 3.1.1** yakni warna gelap ungu menunjukkan hubungan positif antara variabel yang berarti korelasinya kuat, sedangkan jika warna yang muncul semakin terang maka korelasi menunjukkan hubungan negatif antara variabel yang berarti korelasinya lemah. Hubungan korelasi antara variabel yang dimiliki oleh dataset ini terdiri dari nilai skala -1 hingga 1. Nilai -1 mengartikan bahwa hubungan variabel memiliki hubungan terbalik yang artinya jika salah satu variabel naik maka variabel lainnya akan turun sedangkan nilai 1 mengartikan bahwa hubungan variabel memiliki korelasi yang erat dan saling terhubung sehingga jika salah satu variabel naik maka variabel lainnya akan ikut naik. Selain itu, jika korelasi menunjukkan nilai 0 maka kedua variabel tidak memiliki hubungan dan tidak saling mempengaruhi.



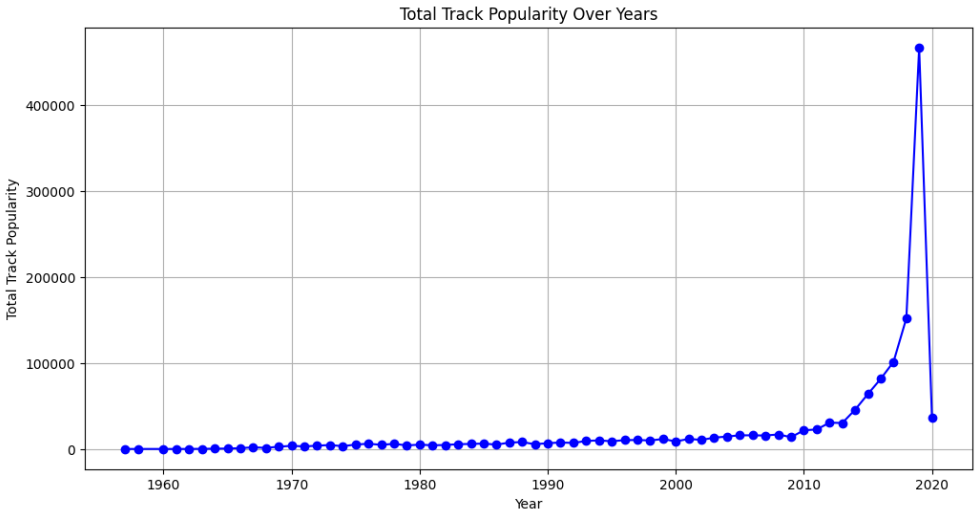
Gambar 3.1.2. Histogram variabel numerik

Pada **Gambar 3.1.2** yang merupakan histogram diperuntukkan untuk variabel numerik, terdapat 14 histogram yang berarti dataset spotify memiliki variabel dengan tipe data numerik sebanyak 14 atribut, diantaranya yakni track_popularity, danceability, energy, key, loudness, mode, speechiness, acousticness, instrumentalness, liveness, valence, tempo, duration_ms, dan year.



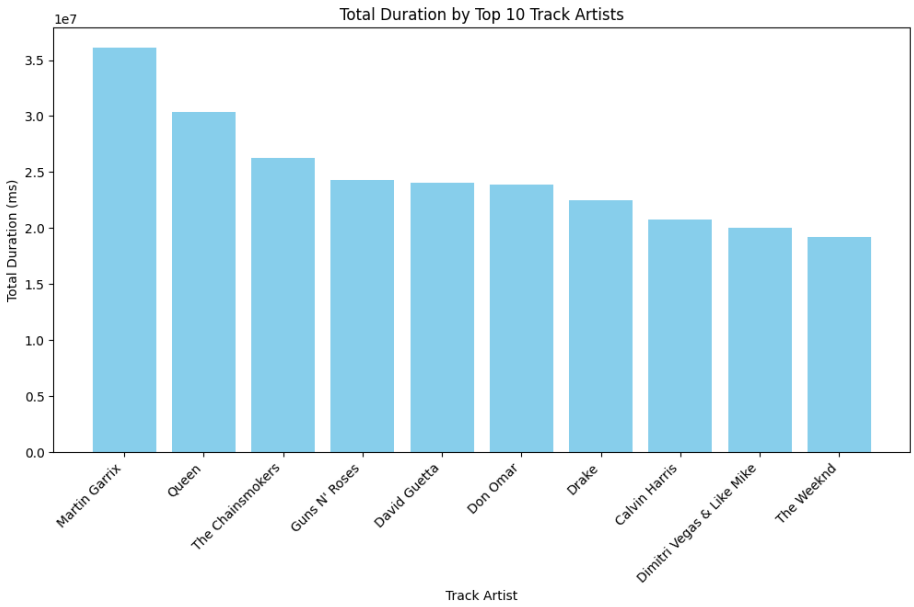
Gambar 3.1.3. Boxplot track popularity by playlist genre

Selanjutnya dilakukan visualisasi boxplot untuk mengetahui sebaran distribusi dari track popularity berdasarkan playlist genre. Terlihat pada **Gambar 3.1.3** sebaran playlist genre paling banyak dimiliki genre rock dan r&b kemudian nilai median dari setiap genre yakni pop, rap, rock, latin, r&b, dan edm berada disekitar nilai 40 hingga 60.



Gambar 3.1.4. Time series track popularity

Pada **Gambar 3.1.4** terlihat plot time series terhadap track popularity setiap tahunnya. Sejak tahun 1960 hingga 2010, popularitas yang dihasilkan dari track cukup rendah sedangkan ketika tahun 2010 popularitas dari track music mulai meningkat cukup tajam dengan puncaknya berada di tahun 2018 hingga 2019 dengan total popularitas yang dihasilkan lebih dari 400000. Hal ini dapat terjadi karena adanya teknologi informasi yang berkembang secara besar-besaran sehingga tidak adanya batas budaya akibat arus globalisasi yang menyeluruh.

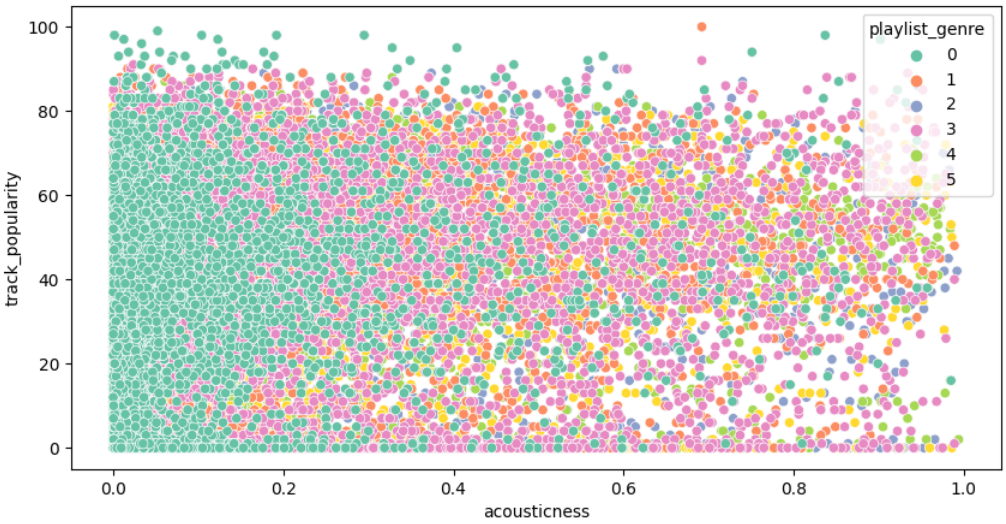


Gambar 3.1.5. Bar plot total duration by top 10 track artists

Selanjutnya pada **Gambar 3.1.5** terlihat bar plot yang memiliki sebaran data dari 10 top artist berdasarkan total durasinya. Bar yang dihasilkan sudah berurut secara menurun dari artist yang memiliki total durasi tertinggi hingga terendah yakni Martin Garix, Queen, The Chainsmokers, Guns N'Roses, David Guetta, Don Omar, Drake, Calvin Harris, Dimitri Vegas & Like Mike, dan The weeknd.

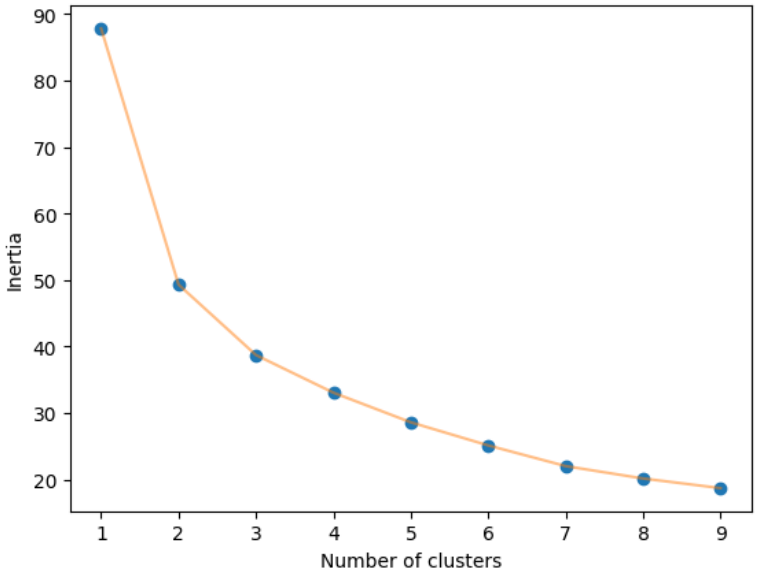
3.2. K-Means 2D

Sebelum melakukan penerapan K-Means, pada dataset spotify dilakukan proses label encoding pada kolom dengan nilai kategorik dengan tujuan agar pemrosesan clustering dapat dilakukan dengan baik.



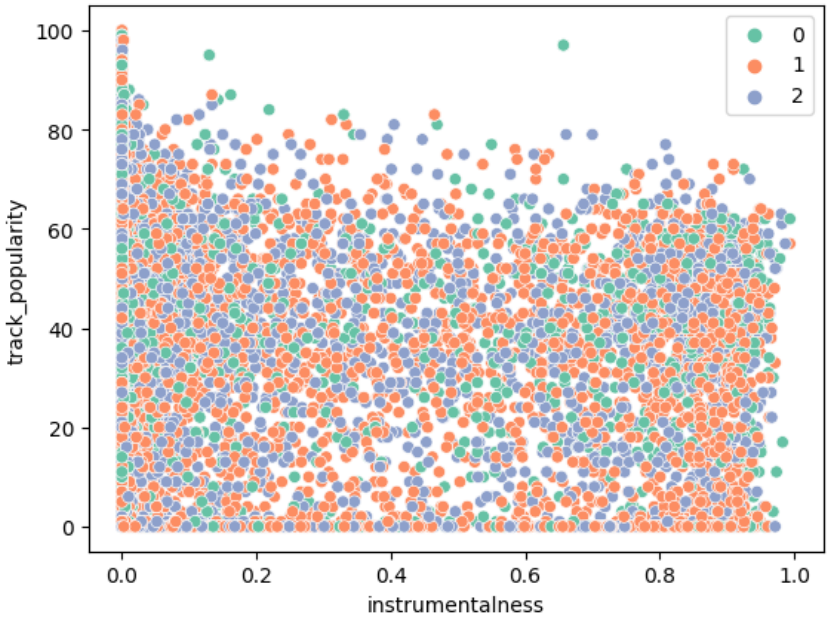
Gambar 3.2.1. Scatterplot acousticness dan track popularity

Pada tahap selanjutnya, dilakukan visualisasi terhadap kolom acousticness sebagai sumbu x dan kolom track popularity sebagai sumbu y seperti pada **Gambar 3.2.1**. Plot ini menunjukkan sebaran data antara data acousticness dan track popularity dengan titik data yang berwarna merepresentasikan kolom playlist genre. Terlihat bahwa sebaran data playlist genre berwarna hijau lebih mendominasi diikuti sebaran data berwarna pink. Sebaran data berwarna hijau merepresentasikan data berdasarkan genre pop sedangkan sebaran data berwarna pink merepresentasikan data berdasarkan genre latin. Hal ini berarti dalam konteks data antara kolom acousticness dan track popularity berdasarkan genrenya lebih didominasi oleh genre pop dan latin.



Gambar 3.2.2. Elbow plot

Selanjutnya dilakukan penerapan clustering K-Means dengan tahap awal menentukan banyaknya kelompok cluster yang akan terbentuk. Untuk mengetahui banyaknya kelompok cluster yang terbentuk dapat dicari tahu melalui plot elbow seperti pada **Gambar 3.2.2** dengan mengambil nilai K-optimal**.** Hasil dari elbow curve menunjukkan angka 3 sebagai banyaknya kelompok yang akan terbentuk akibat lengkungan yang cukup tajam pada titik tersebut.

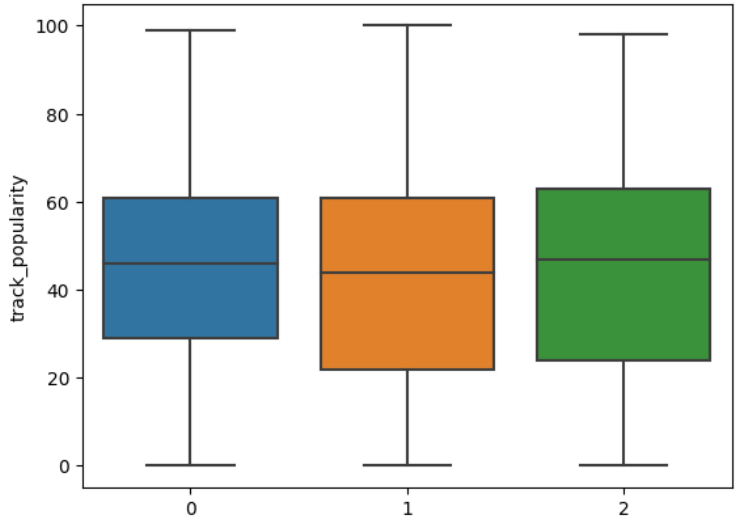


Gambar 3.2.3. Clustering K-Means

Penerapan metode K-Means dilakukan terhadap kolom-kolom tertentu saja yakni kolom 'instrumentalness' dan 'track_popularity' dengan menggunakan nilai K optimal = 3, yang berasal dari hasil grafik elbow curve. Hasil clusterisasi terdapat pada **Gambar 3.2.3**, yang menunjukkan pola kelompok-kelompok yang jelas dengan label seperti berikut:

- Label 0 direpresentasikan warna hijau
- Label 1 direpresentasikan warna orange
- Label 2 direpresentasikan warna ungu

Pemberian label pada dataset ini memiliki beberapa insight menarik, salah satunya yakni untuk menemukan strategi pemasaran yang lebih baik terhadap kondisi dari masing-masing pendengar dan pemusik untuk memperoleh keuntungan yang signifikan.

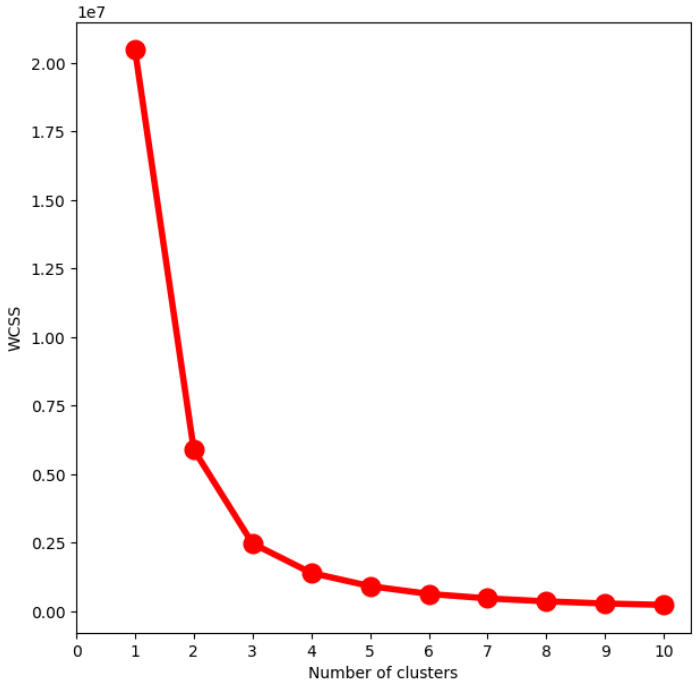


Gambar 3.2.4. Boxplot track popularity

Selain itu, pada **Gambar 3.2.4** dilakukan visualisasi dalam bentuk boxplot terhadap hasil clusterisasi K-Means sebelumnya. Boxplot tersebut menunjukkan hasil sebaran data yang cukup stabil dengan rata-rata nilai median berkisar pada 40 hingga 60 tanpa adanya outlier.

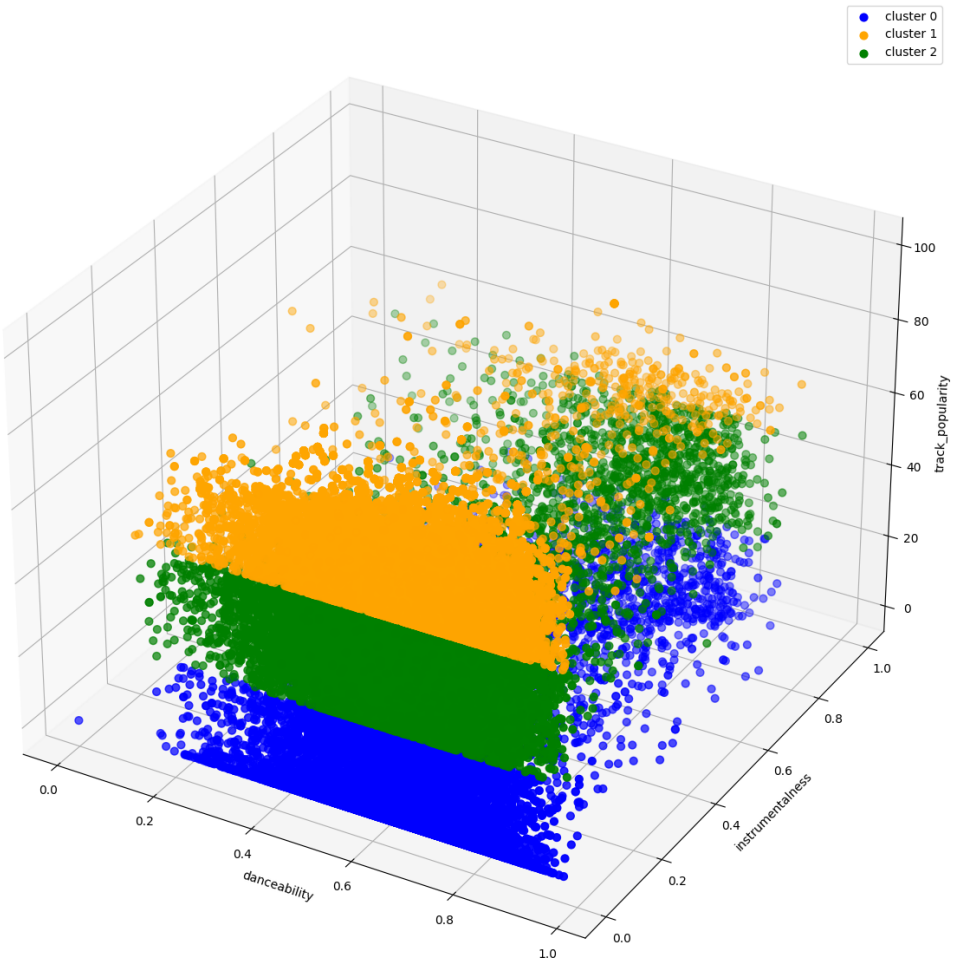
3.3. K-Means 3D

Penerapan metode K-Means 3D dapat dilakukan dengan salah satu cara yakni, mengambil kolom danceability sebagai sumbu x, kolom instrumentalness sebagai sumbu y, dan kolom track_popularity sebagai sumbu z.



Gambar 3.3.1. WCSS curve

Tahap awal yang dapat dilakukan adalah mencari nilai K-optimal menggunakan metode WCSS atau Within-Cluster Sum of Square. WCSS adalah jumlah kuadrat jarak antara setiap titik dan pusat massa dalam sebuah cluster. Saat kita memplot WCSS dengan nilai K, plotnya tampak seperti Siku. Pada **Gambar 3.3.1** terlihat bahwa nilai K-optimal yang didapat untuk melakukan clustering K-Means 3D adalah 3, sehingga akan terbentuk 3 kelompok clusterisasi berdasarkan kolom danceability sebagai sumbu x, kolom instrumentalness sebagai sumbu y, dan kolom track_popularity sebagai sumbu z.



Gambar 3.3.2 Clustering K-Means 3D

Hasil yang diperoleh dari penerapan metode K-Means 3D terhadap kolom-kolom tertentu saja menggunakan nilai K optimal = 3, yang berasal dari hasil grafik WCSS yakni terdapat pada **Gambar 3.3.2**. Plot ini menunjukkan pola kelompok-kelompok yang jelas dengan label seperti berikut:

- Label 0 direpresentasikan warna biru: Label ini menunjukkan kelompok dengan kisaran nilai danceability dan instrumentalness 0.2 hingga 1.0 dengan nilai track popularity yang rendah.
- Label 1 direpresentasikan warna orange: Label ini menunjukkan kelompok dengan kisaran nilai danceability dan instrumentalness 0.2 hingga 1.0 dengan nilai track popularity yang sedang.
- Label 2 direpresentasikan warna hijau: Label ini menunjukkan kelompok dengan kisaran nilai danceability dan instrumentalness 0.2 hingga 1.0 dengan nilai track popularity yang tinggi.

Pemberian label pada dataset ini memiliki beberapa insight menarik, salah satunya yakni untuk menemukan strategi pemasaran yang lebih baik terhadap kondisi dari masing-masing pendengar dan pemusik untuk memperoleh keuntungan yang signifikan.

4. **Kesimpulan**
   Berdasarkan hasil penerapan algoritma clustering pada dataset spotify didapatkan pengolompokan data menggunakan metode k-means Clustering 2D dan 3D. Hasil dari clustering 2D terbentuk menjadi tiga cluster yakni label 0, label 1, dan label 2 berdasarkan kolom 'instrumentalness' dan 'track_popularity'. Sedangkan Hasil dari clustering 3D terbentuk menjadi tiga cluster yakni label 0, label 1, dan label 2 berdasarkan kolom ‘danceability’, 'instrumentalness' dan 'track_popularity'. Dengan menerapkan metode clustering K-Means pada dataset Spotify, informasi yang didapat memberikan wawasan berharga bagi berbagai pihak seperti penyedia layanan streaming musik untuk meningkatkan pengalaman para pendengar musik dan memahami dinamika popularitas lagu.

Referensi

[1] A. Rohmah, F. Sembiring dan A. E. , “IMPLEMENTASI ALGORITMA K-MEANS CLUSTERING ANALYSIS,” *SISMATIK (Seminar Nasional Sistem Informasi dan Manajemen Informatika),* pp. 290-298, 2021.

Lampiran

Berikut merupakan lampiran link drive yang berisi code dengan format ipynb dan dataset yang digunakan, link youtube berisi video presentasi, dan logbook :

- Link skrip pemrograman dan dataset
  Link pemrograman:
  https://colab.research.google.com/drive/1eMLtuM7JUIn4WK8o3FvdMsebGx7oDj-F?usp=sharing#scrollTo=YmAgtwdLlmMz

  Link dataset:
  https://drive.google.com/file/d/1tf76Cj6LIBgy8mtHcoECfbLAYHsuxD-T/view?usp=drive_link

- Link Youtube:
  https://youtu.be/L6zgA6qScwQ