LAPORAN PRAKTIKUM DATA MINING
ANALISIS KARAKTERISTIK PADA PLATFORM SPOTIFY MENGGUNAKAN ALGORITMA K-MEANS

Rizki Adrian Bennovry 1), Muhammad Kaisar Firdaus 2), Deyvan Lexefal 3), Salwa Naqwadisa Madinna 4).
Program Studi Sains Data, Jurusan Sains, Institut Teknologi Sumatera
Email : rizki.121450073@student.itera.ac.id 1),
muhammad.121450135@student.itera.ac.id 2), deyvan.121450148@student.itera.ac.id 3),
salwa.121450157@student.itera.ac.id 4)

Abstrak

Penelitian ini bertujuan untuk menggali potensi dan aplikasi metode K-Means Clustering dalam analisis data, dengan fokus pada pemahaman pola-pola tersembunyi dalam kumpulan data yang kompleks. K-Means Clustering adalah salah satu algoritma yang umum digunakan dalam machine learning dan data science untuk mengelompokkan data menjadi kelompok-kelompok homogen berdasarkan karakteristik yang serupa. Dalam konteks ini, kami mengeksplorasi efektivitas algoritma K-Means Clustering dalam memahami struktur internal dari data yang mencakup berbagai dimensi. Penelitian ini menggunakan pendekatan eksperimental dengan menerapkan K-Means Clustering pada dataset yang mencakup berbagai variabel. Analisis tersebut memberikan wawasan mendalam tentang perbedaan dan kemiripan antara observasi, memungkinkan identifikasi pola-pola yang mungkin tidak terdeteksi secara langsung. Kami juga mempertimbangkan variasi parameter K untuk mengevaluasi dampaknya terhadap hasil clustering. Hasil penelitian ini memberikan kontribusi penting terhadap pemahaman kita tentang kemampuan K-Means Clustering dalam mengatasi tantangan analisis data yang kompleks. Implikasi temuan ini dapat memberikan panduan bagi praktisi data science dan peneliti untuk menerapkan algoritma ini secara efektif dalam berbagai konteks. Kesimpulannya, penelitian ini menyoroti potensi K-Means Clustering sebagai alat yang berguna untuk mengungkap struktur dalam data multidimensional, dengan aplikasi yang luas dalam berbagai domain, termasuk pengelompokan konsumen, pengelolaan inventaris, dan pemrosesan sinyal.

Kata kunci : Clustering, Algoritma, K-Means

1. Pendahuluan

a. Latar Belakang

Dalam era digital ini, volume data yang dihasilkan terus meningkat secara eksponensial di berbagai sektor, termasuk bisnis, ilmu pengetahuan, dan teknologi. Peningkatan ini membawa tantangan baru dalam pemahaman dan ekstraksi informasi berharga dari data yang besar dan kompleks. Dalam konteks ini, metode clustering menjadi kritis untuk mengelompokkan data dan mengungkap pola-pola yang mungkin tersembunyi.
Semakin maju sebuah perkembangan teknologi maka semakin banyak sebuah inovasi, contohnya adalah perkembangan tentang kemajuan dalam bidang musik. Musik sudah menjadi hal yang banyak digemari pada semua kalangan usia, banyak nya jenis atau genre yang beredar membuat banyak orang sering untuk mendengarkan musik dan ditambah dengan kemajuan teknologi yang membuat orang mudah untuk mengakses dari manapun dan kapanpun.
Salah satu contoh platform yang dapat di akses dengan mudah adalah Spotify, dimana aplikasi ini memudahkan orang orang untuk mengakses musik yang bisa di akses gratis dan berbayar. Pada penelitian ini kami mencari sebuah pengelompokan berdasarkan genre musik yang sering di dengar oleh banyak orang dengan menggunakan algoritma pemrograman K-Means clustering pada bahasa pemrograman python.

b. Tujuan

Tujuan dari algoritma K-Means adalah mengelompokkan data ke dalam kelompok-kelompok yang homogen berdasarkan karakteristik yang serupa. Beberapa tujuan khusus dari algoritma K-Means meliputi:

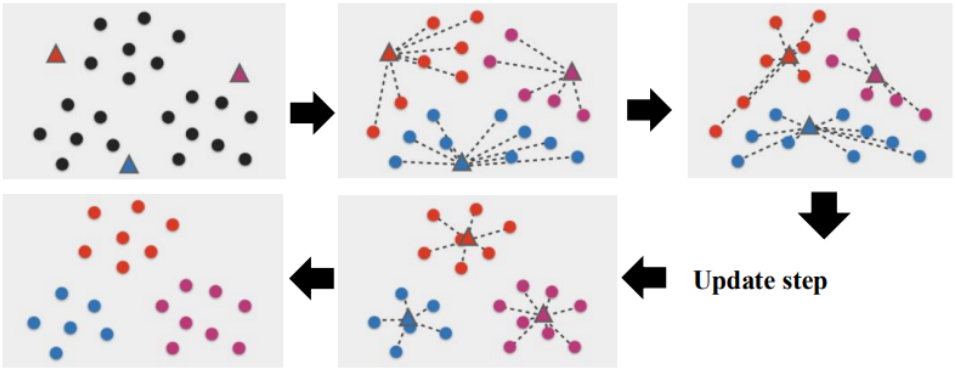
1. Pengelompokan Data yang Efisien
   Algoritma K-Means bertujuan untuk mengelompokkan data ke dalam kelompok-kelompok yang saling eksklusif dan saling berbeda berdasarkan propertinya.
   Memberikan kemampuan untuk mengatasi data yang tidak berlabel, sehingga memungkinkan pengelompokan data tanpa memerlukan informasi sebelumnya tentang kategori atau kelas.
2. Optimasi Pusat Kluster
   Algoritma berusaha untuk mengoptimalkan posisi pusat kluster agar jarak antara titik data di dalam kluster seminimal mungkin.
   Mencapai keseimbangan antara variabilitas internal dalam kluster dan perbedaan antar kluster.

3. Meminimalkan Jarak Antara Pusat Kluster dan Titik Data
   Menetapkan titik data ke kluster yang memiliki pusat terdekat, dengan mengukur jarak menggunakan metrik tertentu seperti Euclidean distance.
   Tujuan utamanya adalah meminimalkan varians dalam setiap kluster.
4. Penentuan Jumlah Kluster yang Optimal (K)
   Algoritma berupaya menentukan jumlah kluster yang optimal (K) dengan cara menghitung varians atau inverse kluster dan meminimalkannya.
   Menentukan jumlah kluster yang sesuai dengan struktur internal data tanpa memerlukan informasi sebelumnya.
5. Skalabilitas dan Kecepatan
   Merupakan tujuan algoritma K-Means untuk bekerja dengan cepat dan efisien, sehingga dapat menangani dataset besar dengan waktu komputasi yang masuk akal.

Melalui pencapaian tujuan-tujuan ini, algoritma K-Means membantu mengorganisir data ke dalam kelompok-kelompok yang bermakna, mendukung pengambilan keputusan, dan memfasilitasi analisis lebih lanjut terhadap struktur data.

c. Konsep Dasar

Pada tahun 1967, J. MacQueen memperkenalkan Algoritma K-Means sebagai salah satu metode clustering yang sangat umum digunakan untuk mengelompokkan data berdasarkan ciri-ciri bersama yang serupa. Clustering ini menghasilkan kelompok data yang disebut sebagai cluster, di mana data di dalamnya mempunyai ciri-ciri serupa. Dalam proses clustering, data dikelompokkan ke dalam cluster-cluster berdasarkan kemiripan antar data. Prinsip dasar dari clustering adalah memaksimalkan kesamaan antar anggota dalam satu cluster sementara meminimumkan kesamaan antar anggota cluster yang berbeda. Tujuan utama dari Algoritma K-Means adalah menemukan kelompok atau cluster pada data, dengan jumlah cluster yang ditentukan oleh variabel K.
Proses kerja Algoritma K-Means dapat dijelaskan sebagai berikut: pertama, tentukan jumlah cluster K yang diinginkan untuk pengelompokan data. Kemudian, inisialisasikan pusat cluster K atau centroid secara acak dengan memilih K titik data secara acak sebagai pusat massa awal. Setelah itu, setiap titik data ditetapkan ke pusat cluster terdekat berdasarkan jarak Euclidean. Titik data yang paling dekat dengan pusat massa tertentu dianggap sebagai bagian dari cluster tersebut. Selanjutnya, hitung ulang pusat klaster dengan mengambil rata-rata semua titik data yang telah ditetapkan ke klaster tersebut. Langkah-langkah 3 dan 4 diulang hingga pusat klaster berhenti bergerak atau iterasi mencapai batas yang ditentukan, menandakan konvergensi algoritma.



Gambar 1. Algoritma K-Means

Gambar 1 mengilustrasikan ide dasar K-Means yang bertujuan untuk mengurangi jumlah jarak kuadrat antara titik data dan pusat kluster yang ditugaskan. Proses ini dilakukan dengan mengalokasikan ulang titik data secara berulang ke pusat massa terdekat dan menggeser posisi pusat klaster ke pusat dari titik-titik yang telah ditetapkan. Hasilnya adalah pembentukan klaster yang lebih padat dan terpisah, mempromosikan pengelompokan yang lebih efisien dan efektif.

2. Metode

2.1. Data

Data yang digunakan pada penelitian ini bernama spotify.csv yang diberikan oleh asisten praktikum Data Mining. Dataset "spotify.csv" terdiri dari 32,833 entri dan 23 kolom yang menyajikan informasi terkait berbagai lagu dalam platform Spotify. Setiap baris mewakili sebuah lagu, sementara kolom-kolomnya memberikan rincian yang bervariasi tentang propertis dan atribut-atribut lagu tersebut.

Beberapa kolom kunci dalam dataset ini mencakup informasi identifikasi seperti "track_id," "track_name," dan "track_artist" yang memberikan detail tentang judul dan artis lagu. Selain itu, dataset mencantumkan skor popularitas lagu dengan kolom "track_popularity."

Informasi terkait album juga terdapat dalam dataset, dengan kolom seperti "track_album_id," "track_album_name," dan "track_album_release_date" yang memberikan detail tentang album tempat lagu tersebut terdapat. Selanjutnya, terdapat informasi tentang playlist, dengan kolom "playlist_name," "playlist_id," "playlist_genre,"

dan "playlist_subgenre" yang memberikan konteks tentang playlist di mana lagu-lagu tersebut termasuk.
Dataset ini juga memuat atribut audio dari lagu-lagu, seperti "danceability," "energy," "loudness," "speechiness," "acousticness," dan sebagainya, yang memberikan gambaran tentang karakteristik audio dari masing-masing lagu. Atribut-atribut ini dapat digunakan untuk analisis lebih lanjut terkait preferensi musik dan pengelompokan lagu berdasarkan properties tertentu.
Selain itu, dataset ini juga mencakup kolom seperti "tempo" dan "duration_ms" yang memberikan informasi tentang tempo dan durasi lagu secara bersamaan. Keseluruhan, dataset ini menyediakan dasar yang kaya dan beragam untuk analisis data terkait preferensi musik dan karakteristik audio pada platform Spotify.

2.2. Metode

Sebelum melakukan clustering dengan algoritma K-Means, kami melakukan data understanding and preparation terlebih dahulu untuk mengetahui sekaligus meningkatkan kualitas data, dimana memuat beberapa langkah sebagai berikut :

1. Introduce Dataset
   Melihat beberapa awal baris pada dataset agar memudahkan pengenalan data. Lalu memuat informasi dari semua variabel di dataset agar dapat mengetahui dataset yang digunakan jumlah dari dimensi data dan jumlah serta tipe data dari masing-masing variabel
2. Mengatasi Missing Value
   Pada setiap dataset tidak menutup kemungkinan terdapat sebuah missing value atau nilai kosong pada sebuah variabel di dataset. Missing value dapat dihapus atau dibiarkan tergantung kebutuhan, pada pemrograman K-Mean direkomendasikan untuk di hapus untuk menghindari dari kesalahan prediksi dan menghilangkan bias agar dapat meningkatkan akurasi dari algoritma dan dapat memberikan hasil pengelompokkan yang akurat.
3. Korelasi Variabel Numerik
   Setelah melihat informasi dataset, jika terdapat variabel yang bertipe sejenis numerik maka dapat melihat korelasi antar variabel numerik yang berguna untuk melihat hubungan antara variabel numerik yang satu dengan yang lainnya. Pada proses ini terdapat cara yang mempermudah pengguna untuk menganalisis yaitu dengan cara memvisualisasikan korelasi pada matriks heatmap yang dapat menghasilkan visualisasi yang mudah di pahami.

4. Outliers
   Outliers, atau pencilan, merujuk kepada observasi atau titik data yang secara signifikan berbeda dari sebagian besar data dalam sebuah set. Outliers dapat

mempengaruhi hasil analisis statistik dan merusak asumsi-asumsi yang mendasari beberapa metode statistik. Identifikasi dan penanganan outliers menjadi penting dalam analisis data untuk memastikan hasil yang akurat dan representatif.

5. Normalize

   Normalize atau Normalisasi adalah proses mengubah nilai-nilai dalam suatu dataset sehingga mereka dapat dibandingkan secara relatif, terlepas dari skala aslinya. Tujuan normalisasi adalah untuk memastikan bahwa berbagai atribut atau fitur dalam dataset memiliki dampak yang setara dalam analisis atau pemodelan, terutama dalam konteks machine learning atau analisis statistik. Normalisasi membantu mengatasi perbedaan skala yang signifikan antara variabel, sehingga mencegah variabel dengan skala besar mendominasi pengaruh dalam pemodelan.

3. Hasil & Pembahasan

3.1 Data Understanding dan Data Preparation

Data Understanding digunakan untuk memeriksa dan mengidentifikasi pola pada masalah data serta, memahami struktur dan karakteristik data. Data Preparation memastikan persiapan data sebelum menerapkan model.

- **Mendeskripsikan data**

```
[ ] df.head() # Menampilkan 5 baris pertama dari dataset untuk preview
```

	track_id	track_name	track_artist	track_popularity	track_album_id	track_album_name	track_album_release_date	playlist_name	
0	6f807x0ima9a1j3VPbc7VN	I Don't Care (with Justin Bieber) - Loud Luxur...	Ed Sheeran	66	2oCs0DGTsRO98Gh5ZSl2Cx	I Don't Care (with Justin Bieber) [Loud Luxury...	2019-06-14	Pop Remix	37i9dQZF1
1	0r7CVbZTWZgbTCYdfa2P31	Memories - Dillon Francis Remix	Maroon 5	67	63rPSO264uRjW1X5E6cWv6	Memories (Dillon Francis Remix)	2019-12-13	Pop Remix	37i9dQZF1
2	1z1Hg7Vb0AhHDiEmnDE79l	All the Time - Don Diablo Remix	Zara Larsson	70	1HoSmj2eLcsrR0vE9gThr4	All the Time (Don Diablo Remix)	2019-07-05	Pop Remix	37i9dQZF1
3	75FpbthrwQmzHlBJLuGdC7	Call You Mine - Keanu Silva Remix	The Chainsmokers	60	1nqYsOef1yKKuGOVchbsk6	Call You Mine - The Remixes	2019-07-19	Pop Remix	37i9dQZF1
4	1e8PAfcKUYoKkxPhrHqw4x	Someone You Loved - Future Humans Remix	Lewis Capaldi	69	7m7vv9wlQ4i0LFuJiE2zsQ	Someone You Loved (Future Humans Remix)	2019-03-05	Pop Remix	37i9dQZF1

Setelah melakukan input data, lalu menampilkan hasil beberapa baris pertama dari suatu DataFrame, termasuk semua baris dan kolom yang ada. Hal ini memberikan gambaran tentang data terstruktur dan nilai-nilai awalnya.

- **Melihat Informasi Mengenai Data**

```
df.info() # Menampilkan informasi tentang dataset

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 32833 entries, 0 to 32832
Data columns (total 23 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   track_id                  32833 non-null  object
 1   track_name                32828 non-null  object
 2   track_artist              32828 non-null  object
 3   track_popularity          32833 non-null  int64
 4   track_album_id            32833 non-null  object
 5   track_album_name          32828 non-null  object
 6   track_album_release_date  32833 non-null  object
 7   playlist_name             32833 non-null  object
 8   playlist_id               32833 non-null  object
 9   playlist_genre            32833 non-null  object
 10  playlist_subgenre         32833 non-null  object
 11  danceability              32833 non-null  float64
 12  energy                    32833 non-null  float64
 13  key                       32833 non-null  int64
 14  loudness                  32833 non-null  float64
 15  mode                      32833 non-null  int64
 16  speechiness               32833 non-null  float64
 17  acousticness              32833 non-null  float64
 18  instrumentalness          32833 non-null  float64
 19  liveness                  32833 non-null  float64
 20  valence                   32833 non-null  float64
 21  tempo                     32833 non-null  float64
 22  duration_ms               32833 non-null  int64
dtypes: float64(9), int64(4), object(10)
memory usage: 5.8+ MB
```

Dengan melihat informasi mengenai data, didapatkan informasi ringkas yang mencakup beberapa aspek penting dari Data Frame, seperti jumlah kolom, nama kolom, tipe data dan apakah ada nilai-nilai yang hilang. Pada keseluruhan variabel terdapat 3 jenis tipe data yaitu object, float64, int64. Hal ini sangat penting karena dapat membantu dalam pengambilan keputusan untuk pembersihan data.

- **Memeriksa Adanya Missing Value**

```
df.isnull().sum() # Cek nilai yang hilang (missing values) dalam dataset

track_id                  0
track_name                5
track_artist              5
track_popularity          0
track_album_id            0
track_album_name          5
track_album_release_date  0
playlist_name             0
playlist_id               0
playlist_genre            0
playlist_subgenre         0
danceability              0
energy                    0
key                       0
loudness                  0
mode                      0
speechiness               0
acousticness              0
instrumentalness          0
liveness                  0
valence                   0
tempo                     0
duration_ms               0
dtype: int64
```

Pada data ini dilakukan pengecekan adanya missing value dalam data frame dan terdapat missing value yang terletak pada variabel track_name, track_artist, track_album_name yang berjumlah 5.

- **Menghapus Data Kosong**

```
[ ] df.isnull().sum() # Cek kembali nilai yang hilang (missing values) dalam dataset

track_id                  0
track_name                0
track_artist              0
track_popularity          0
track_album_id            0
track_album_name          0
track_album_release_date  0
playlist_name             0
playlist_id               0
playlist_genre            0
playlist_subgenre         0
danceability              0
energy                    0
key                       0
loudness                  0
mode                      0
speechiness               0
acousticness              0
instrumentalness          0
liveness                  0
valence                   0
tempo                     0
duration_ms               0
dtype: int64
```

Setelah itu, melakukan penghapusan seluruh data yang memiliki missing value agar tidak ada lagi nilai yang memiliki missing value dalam data frame. Penghapusan missing value ini dapat berguna untuk analisis dan membangun suatu model.

- **Memeriksa Korelasi**

```
df.corr() # Mengidentifikasi korelasi antara variabel numerik dalam dataset
```

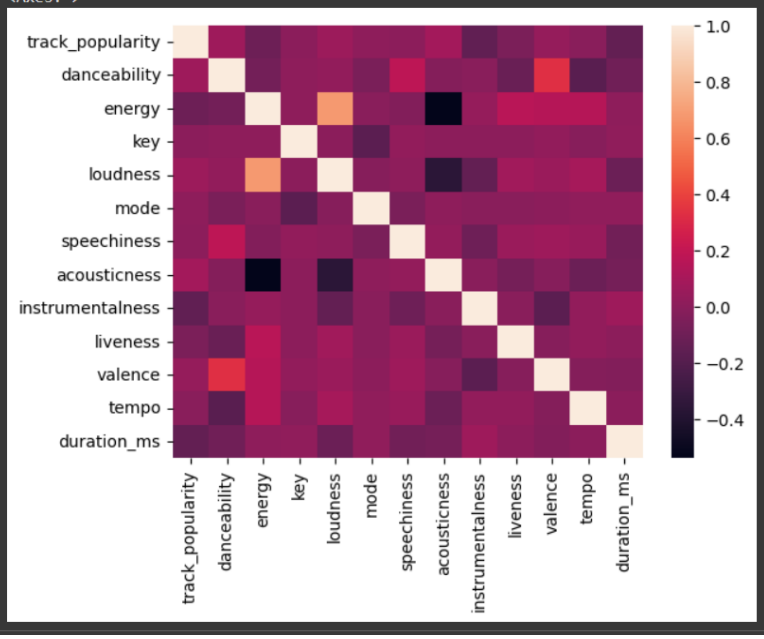
```
<ipython-input-214-aea4b5ebf735>:1: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid c
  df.corr() # Mengidentifikasi korelasi antara variabel numerik dalam dataset
```

	track_popularity	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumentalness	liveness	valence	tempo	duration_ms
track_popularity	1.000000	0.064754	-0.108984	-0.000405	0.057717	0.010553	0.007067	0.085042	-0.150003	-0.054593	0.033278	-0.005538	-0.143634
danceability	0.064754	1.000000	-0.086074	0.011771	0.025351	-0.058711	0.181808	-0.024515	-0.008658	-0.123899	0.330538	-0.184132	-0.096922
energy	-0.108984	-0.086074	1.000000	0.009972	0.676662	-0.004778	-0.032184	-0.539732	0.033282	0.161317	0.151050	0.150072	0.012560
key	-0.000405	0.011771	0.009972	1.000000	0.000920	-0.173981	0.022462	0.004378	0.006022	0.002834	0.019933	-0.013316	0.015141
loudness	0.057717	0.025351	0.676662	0.000920	1.000000	-0.019242	0.010313	-0.361646	-0.147823	0.077589	0.053411	0.093761	-0.115039
mode	0.010553	-0.058711	-0.004778	-0.173981	-0.019242	1.000000	-0.063446	0.009399	-0.006760	-0.005485	0.002567	0.014339	0.015576
speechiness	0.007067	0.181808	-0.032184	0.022462	0.010313	-0.063446	1.000000	0.026168	-0.103385	0.055337	0.064756	0.044649	-0.089432
acousticness	0.085042	-0.024515	-0.539732	0.004378	-0.361646	0.009399	0.026168	1.000000	-0.006881	-0.077247	-0.016833	-0.112782	-0.081553
instrumentalness	-0.150003	-0.008658	0.033282	0.006022	-0.147823	-0.006760	-0.103385	-0.006881	1.000000	-0.005505	-0.175406	0.023303	0.063256
liveness	-0.054593	-0.123899	0.161317	0.002834	0.077589	-0.005485	0.055337	-0.077247	-0.005505	1.000000	-0.020432	0.020887	0.006197
valence	0.033278	0.330538	0.151050	0.019933	0.053411	0.002567	0.064756	-0.016833	-0.175406	-0.020432	1.000000	-0.025639	-0.032292
tempo	-0.005538	-0.184132	0.150072	-0.013316	0.093761	0.014339	0.044649	-0.112782	0.023303	0.020887	-0.025639	1.000000	-0.001347
duration_ms	-0.143634	-0.096922	0.012560	0.015141	-0.115039	0.015576	-0.089432	-0.081553	0.063256	0.006197	-0.032292	-0.001347	1.000000

Pada hasil korelasi didapatkan nilai 1 yang artinya terdapat korelasi yang kuat atau korelasi positif sempurna , sedangkan korelasi yang bernilai negatif artinya terdapat korelasi dengan arah terbalik atau korelasi negatif sempurna.

```
sns.heatmap(df.corr()) # Menampilkan heatmap untuk visualisasi korelasi antar variabel numerik
```

```
<ipython-input-215-c6a7eba8a8d4>:1: FutureWarning: The default value of numeric_only in DataFrame.cor
  sns.heatmap(df.corr()) # Menampilkan heatmap untuk visualisasi korelasi antar variabel numerik
<Axes: >
```



Didapatkan visualisasi berdasarkan matriks korelasi yang telah dicari sebelumnya. Yang dimana, visualisasi ini memberikan gambaran yang lebih jelas tentang korelasi antar kolom dalam data frame. Pada warna sel yang lebih terang menunjukkan korelasi yang lebih tinggi, sementara sel yang lebih gelap menunjukkan korelasi yang lebih rendah.

- **Menghapus Outliers**

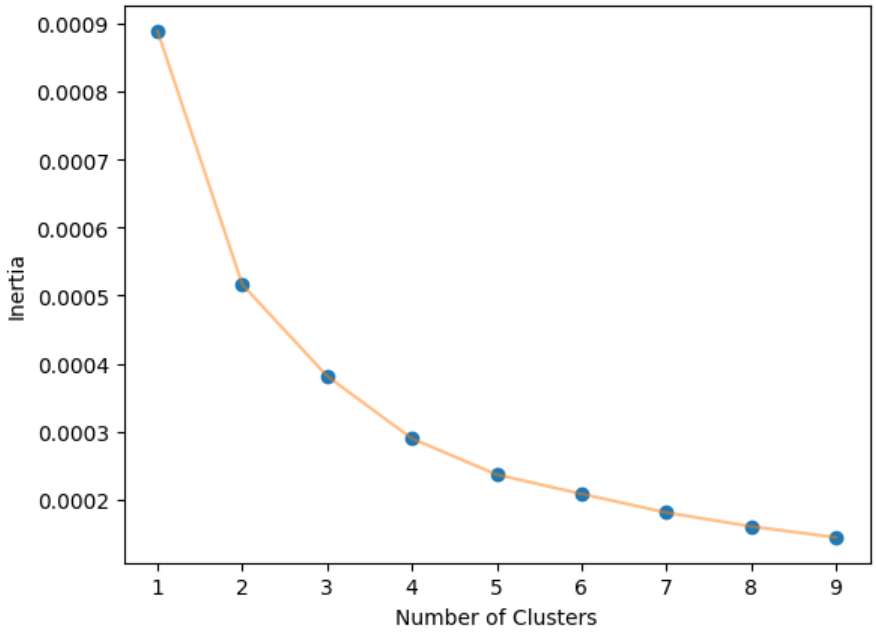
```
print(cleaned_data.describe()) # Statistik deskriptif dari dataset setelah menghapus outliers
```

```
       track_popularity  danceability        energy           key  \
count      19223.000000  19223.000000  19223.000000  19223.000000
mean          45.449462      0.665450      0.715206      5.349477
std           25.065336      0.135364      0.150405      3.604297
min            0.000000      0.266000      0.199000      0.000000
25%           28.000000      0.575000      0.613000      2.000000
50%           49.000000      0.679000      0.729000      6.000000
75%           65.000000      0.765000      0.833000      8.000000
max           99.000000      0.979000      0.999000     11.000000

           loudness          mode   speechiness  acousticness  \
count  19223.000000  19223.000000  19223.000000  19223.000000
mean      -6.084896      0.573011      0.084542      0.141706
std        2.280973      0.494653      0.060777      0.153921
min      -13.454000      0.000000      0.022400      0.000002
25%       -7.367000      0.000000      0.040500      0.019800
50%       -5.762000      1.000000      0.059600      0.081300
75%       -4.464000      1.000000      0.109000      0.217000
max        0.642000      1.000000      0.268000      0.614000

       instrumentalness      liveness       valence         tempo  \
count      19223.000000  19223.000000  19223.000000  19223.000000
mean           0.000632      0.163430      0.535133    119.177623
std            0.001841      0.100606      0.223885     24.841750
min            0.000000      0.015000      0.029200     50.454000
25%            0.000000      0.090800      0.361000     99.774500
50%            0.000002      0.124000      0.538000    119.967000
75%            0.000126      0.219000      0.711000    131.988000
max            0.012000      0.480000      0.981000    184.474000

         duration_ms
count   19223.000000
```

Dilakukan penghapusan outlier yang telah dilakukan pengecekan pada proses deskripsi data agar data yang akan diolah lebih efektif. Setelah itu, menampilkan ringkasan statistik deskriptif dari data frame yang telah dibersihkan.

3.2 Elbow Method

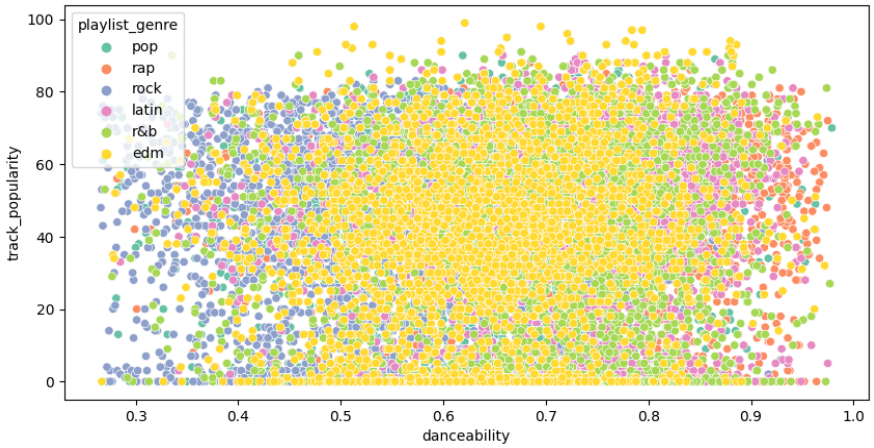
Elbow method digunakan dalam analisis cluster untuk menentukan jumlah cluster yang optimal dalam suatu data untuk keseimbangan yang baik antara kompleksitas model dengan variasi dalam data.



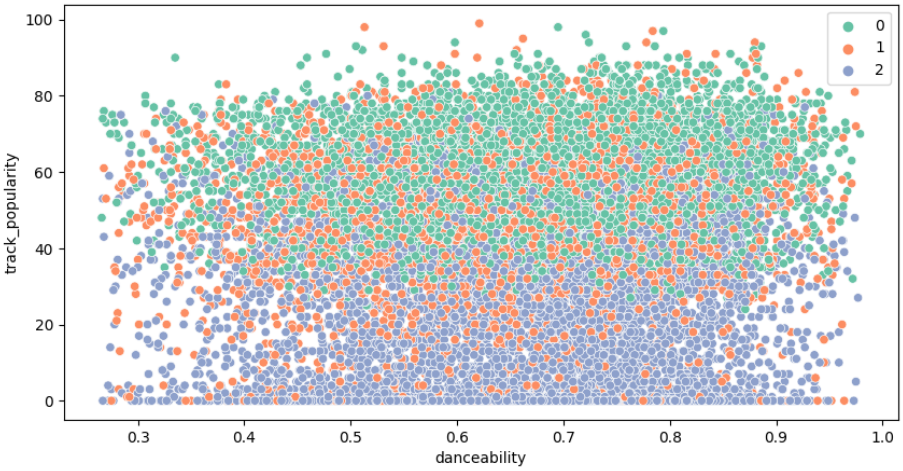
Grafik elbow diatas menunjukkan nilai inertia yang optimal berubah seiring dengan peningkatan jumlah klaster dari 1 hingga 9. Terlihat pada garis mengalami patahan yang membentuk elbow atau siku pada saat k=5, maka dengan menggunakan metode ini diperoleh k optimal saat berada di k=5.

3.3 K-Means Clustering

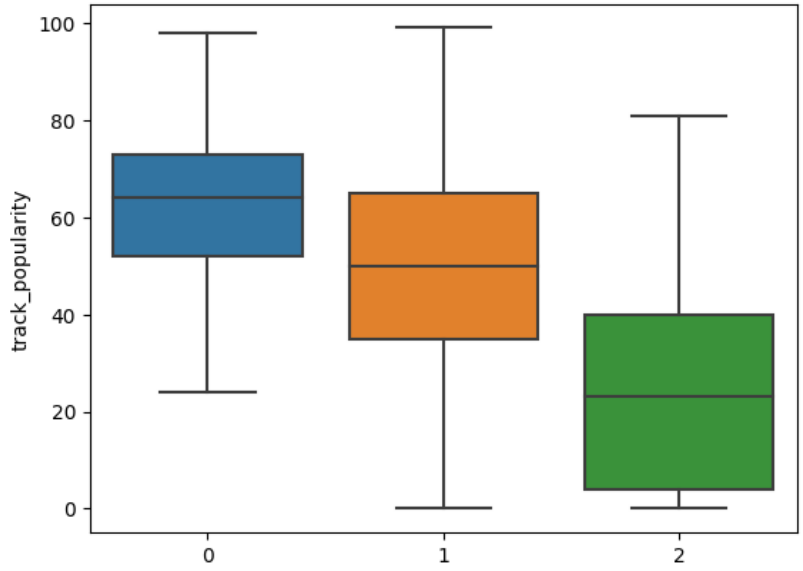
K-Means merupakan algoritma clustering yang digunakan untuk mengelompokkan data ke dalam cluster yang bertujuan untuk mendapatkan sejumlah cluster yang isinya sebisa mungkin homogen atau mungkin berbeda satu sama lain, sehingga batas antar cluster akan menjadi lebih jelas.



Hasil visualisasi ini memungkinkan data terklasifikasi ke dalam klaster yang berbeda dengan menggunakan warna yang berbeda. Sumbu x dan y yang mewakili kolom-kolom tertentu dalam data frame, dengan memberikan warna berdasarkan kolom kategori playlist_genre.



Hasil visualisasi diatas menunjukkan titik-titik data yang berdekatan sudah dikelompokkan data terklasifikasi ke dalam cluster yang berbeda dengan menggunakan warna yang berbeda. Dengan menggunakan kolom "danceability" dan "track_popularity" sebagai sumbu x dan y, dapat dilihat pola sebaran yang memberikan informasi tentang karakteristik masing-masing.



Hasil dari diagram boxplot diatas, mengidentifikasi hubungan antara cluster dan variabel track popularitas dengan sebaran track popularitas di setiap kategori dan dari hasil visualisasi boxplot tersebut, terlihat bahwa tidak terdapat outlier pada beberapa atribut. Hal ini akan membuat clustering menjadi efektif.

4. Kesimpulan

Penelitian menggunakan dataset ‘spotify’ yang terdiri dari 32.8333 entri dan 23 kolom, menyajikan informasi terkait lagu-lagu di platform Spotify. Dataset mencakup berbagai informasi, termasuk identifikasi lagu, skor popularitas, informasi album, dan atribut audio seperti danceability, energy, loudness dan sebagainya. Sebelum menerapkan algoritma K-Means untuk clustering, dilakukan tahapan data understanding and preparation. Langkah-langkahnya melibatkan pemeriksaan awal dataset, penanganan missing value, analisis korelasi variabel numerik, identifikasi dan penanganan outliers, serta normalisasi data. Proses normalisasi dilakukan untuk membandingkan nilai-nilai dalam dataset secara relatif, dengan dilakukan analisis menggunakan metode Elbow untuk menentukan jumlah cluster optimal dalam algoritma K-Means. Hasilnya menunjukkan bahwa k=5 adalah jumlah optimal cluster. Penerapan K-Means clustering menghasilkan visualisasi data terklasifikasi ke dalam klaster yang berbeda. Pola sebaran karakteristik lagu, seperti danceability dan track popularity, dapat diamati. Diagram boxplot juga menunjukkan bahwa tidak ada outlier pada beberapa atribut, meningkatkan efektivitas clustering. Secara keseluruhan, penelitian ini menggunakan pendekatan analisis data yang komprehensif untuk memahami preferensi musik dan karakteristik audio pada platform Spotify melalui algoritma K-Means. Meskipun relatif tidak rumit, K-Means cukup efektif untuk melakukan clustering. Kelemahan dari K-Means yaitu memerlukan perkiraan berapa banyak cluster yang secara alamiah sudah ada di dalam data.

References

Modul Praktikum Data Mining 4

"\/." 2023. YouTube. https://www.academia.edu/24914918/ANALISIS_BIVARIAT_DATA_NUMERIK_DAN_NUMERIK_Uji_Korelasi_dan_Regres.

,M.Sc., Dios K. 2020. *Pengenalan Machine Learning dengan Python*. Vol. 290. Jakarta: Elex Media Komputindo.

"Analisis Korelasi: Memahami Definisi, Tujuan, Fungsi, dan Contoh Analisis Korelasi." n.d. Ruang Jurnal. Accessed November 29, 2023. https://ruangjurnal.com/analisis-korelasi-memahami-definisi-tujuan-fungsi-dan-contoh-analisis-korelasi/.

Aulia, Tita E. 2022. "Cara Mendeteksi dan Menangani Outlier." Pacmann. https://pacmann.io/blog/cara-mendeteksi-dan-menangani-outlier-saat-melakukan-data-analysis.

"Bagaimana Cara Menghadapi Missing Value bagi Pemula?" 2022. Blog - Algoritma Data Science School. https://blog.algorit.ma/missing-value/.

"BASIS DATA." n.d. Repository Unikom. Accessed November 29, 2023. https://repository.unikom.ac.id/51040/1/Materi%204%20-%20Pengantar%20Normalisasi%20Data%20%5BSBD%20-%202017%5D.pdf.

"Cara Menangani Missing Value pada Project Data Science." 2022. Algoritma Data Science School. https://algorit.ma/blog/missing-value-2022/.

"Cara Mengatasi Missing Value Pada Dataset." 2021. Yasya Indra Blog. https://www.yasyaindra.com/2021/08/mengatasi-missing-value.html.

“Data Mining dengan Teknik Clustering Menggunakan Algoritma K-Means pada Data Transaksi Superstore.” n.d. SNIA. Accessed November 29, 2023. http://repository.unjani.ac.id/repository/817cc637bc33e2218345e70a6929dab6.pdf.

“Data Normalisasi: Referensi Komponen - Azure Machine Learning.” 2023. Microsoft Learn. https://learn.microsoft.com/id-id/azure/machine-learning/component-reference/normalize-data?view=azureml-api-2.

“Dataset dan Variabel • Nural Learning.” n.d. Nural Learning. Accessed November 29, 2023. https://learn.nural.id/course/data-analyses/regresi-linier-di-R/dataset-dan-variabel.

Díaz, Raúl. 2022. “.” . - YouTube. https://blog.restatolahdata.id/mengukur-hubungan-variabel-dengan-analisis-korelasi/.

“How to Handling Missing Value Using Imputation in R.” 2020. MATIIN LAUGIWA PRAWIRA PUTRA. https://laugiwa.medium.com/how-to-handling-missing-value-using-imputation-in-r-ab7af6314b20.

Hussein, Saddam. 2021. “K-means Clustering: Pengertian, Metode Algoritma, Beserta Contoh.” geospasialis. https://geospasialis.com/k-means-clustering/.

“Normalisasi Data: Pengertian, Tujuan, dan Metodenya.” 2022. Trivusi. https://www.trivusi.web.id/2022/09/normalisasi-data.html.

“Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Untuk Menentukan Strategi Promosi Mahasiswa Baru Universitas Bina.” n.d. Open Journal Systems. Accessed November 29, 2023. http://eprints.binadarma.ac.id/15448/1/10-Article%20Text-33-2-10-20210219.pdf.

Rohman, Muhammad A. n.d. “Teknik untuk menghilangkan outlier dalam pembersihan data -.” Sekolah Stata. Accessed November 29, 2023. https://sekolahstata.com/teknik-untuk-menghilangkan-outlier-dalam-pembersihan-data/.

Wichmann, Nick. 2021. “,.” , - YouTube. https://media.neliti.com/media/publications/268954-normalisasi-database-dan-migrasi-databas-df741ba4.pd.