ROUTERS: Jurnal Sistem dan Teknologi Informasi
Vol. 4, No. 1, Februari 2026, pp. 53-63
ISSN: 2962-1224, DOI: 10.25181/rt.v5i1.4713

Analisis Pengelompokan Lagu Terpopuler Spotify Menggunakan Algoritma K-Means Berdasarkan Popularitas

Arko Fernanda Wibawa [1], Juwita Valentiya [2]
[1, 2] Teknologi Rekayasa Internet, Politeknik Negeri Lampung

INFORMASI ARTIKEL

Diterima 31 Desember 2025
Direvisi 27 Februari 2026
Diterbitkan 28 Februari 2026

Kata kunci:

Spotify;
K-means;
Popularitas streaming;
Transformasi logaritmik;
Segmentasi lagu

ABSTRAK

Perkembangan platform *streaming* membuat katalog musik menjadi sangat besar sehingga pola popularitas lagu sulit dipahami hanya dari satu indikator. Indikator total pemutaran (*streams*) dapat menunjukkan capaian popularitas kumulatif, tetapi tidak selalu mencerminkan momentum pemutaran terkini. Kondisi ini membuat pihak terkait seperti platform *streaming*, artis, dan label membutuhkan segmentasi lagu berdasarkan pola popularitas yang lebih informatif. Penelitian ini menerapkan pendekatan data mining dengan clustering K-Means untuk mengelompokkan lagu-lagu terpopuler Spotify berdasarkan indikator popularitas *streaming*. Kontribusi penelitian ini adalah membangun kerangka segmentasi lagu populer berbasis kombinasi indikator total pemutaran, pemutaran harian, serta rasio pemutaran harian terhadap total pemutaran untuk menangkap kekuatan tren terkini. Metode yang digunakan meliputi pembersihan data, imputasi nilai hilang, transformasi logaritmik untuk mengurangi ketimpangan distribusi, rekayasa fitur rasio, standardisasi fitur, pelatihan model K-Means, pemilihan jumlah cluster menggunakan Elbow Method dan Silhouette Score, serta evaluasi menggunakan Inertia, Silhouette Score, Calinski–Harabasz Index, dan Davies–Bouldin Index. Hasil menunjukkan bahwa model final dengan k = 4 menghasilkan Inertia 2673,011 dan Silhouette Score 0,364835, serta membentuk empat segmen lagu yang dapat diinterpretasikan. Cluster 0 merepresentasikan lagu *super-trending* (rasio harian tertinggi), cluster 1 merepresentasikan lagu populer lama dengan aktivitas harian rendah, cluster 2 merepresentasikan *mega hits* dengan total *streams* sangat tinggi dan aktivitas harian masih kuat, serta cluster 3 merepresentasikan lagu populer dengan performa stabil. Segmentasi ini memberikan wawasan praktis untuk prioritas promosi, kurasi *playlist*, dan analisis tren.

Clustering Analysis Of Spotify Most Streamed Songs Using The K-Means Algorithm Based On Popularity

ARTICLE INFO

Received December 31, 2025
Revised February 27, 2026
Published February 28, 2026

Keyword:

Spotify;
K-Means;
streaming popularity;
logarithmic transformation;
song segmentation

ABSTRACT

The rapid growth of music streaming platforms has created very large catalogs, making popularity patterns difficult to understand using a single indicator. Total streams reflect cumulative success, but they do not always represent current listening momentum. This situation motivates the need for song segmentation based on more informative popularity patterns to support decision-making for streaming platforms, artists, and labels. This study applied a data mining approach using K-Means clustering to group Spotify most-streamed songs based on streaming popularity indicators. The main contribution was a segmentation framework that combined total streams, daily streams, and a daily-to-total streams ratio to better capture current momentum. The method included data cleaning, missing value imputation, logarithmic transformation to reduce skewness, feature engineering of a ratio variable, feature standardization, K-Means training, cluster number selection using the elbow method and Silhouette Score, and evaluation using Inertia, Silhouette Score, the Calinski–Harabasz Index, and the Davies–Bouldin Index. The final model with k = 4 achieved an Inertia of 2673.011 and a Silhouette Score of 0.364835 and produced four interpretable segments. Cluster 0 represented super-trending songs with the highest daily-to-total ratio, cluster 1 represented legacy popular songs with low daily activity, cluster 2 represented mega hits with extremely high total streams and still strong daily activity, and cluster 3 represented consistently performing songs with stable daily streams. These segments provided practical insights for promotion prioritization, playlist curation, and trend interpretation.

Corresponding Author:

Arko Fernanda Wibawa, Politeknik Negeri Lampung,
Bandar Lampung, Indonesia
Email: arkofernanda31@gmail.com

1. PENDAHULUAN

Perkembangan teknologi informasi dan internet telah mengubah cara masyarakat mengakses musik, dari media fisik menuju layanan digital berbasis *streaming*. Spotify menjadi salah satu platform yang menyediakan katalog musik sangat besar dengan jutaan lagu yang dapat diakses kapan saja [1], [2]. Di satu sisi, kondisi ini memberi kebebasan bagi pendengar untuk menemukan musik baru, tetapi di sisi lain menimbulkan tantangan dalam memahami pola konsumsi musik dan tren popularitas yang terjadi pada skala besar [2].

Popularitas lagu pada platform *streaming* umumnya direpresentasikan melalui indikator kuantitatif seperti total jumlah pemutaran (*streams*) dan jumlah pemutaran harian (*daily*) [3]. Total *streams* menggambarkan akumulasi popularitas sepanjang waktu, sedangkan *daily streams* lebih menggambarkan kekuatan momentum saat ini [3]. Dua lagu dengan total *streams* mirip dapat memiliki *daily streams* yang berbeda, yang berarti tingkat "kepanasan" tren juga berbeda [3]. Oleh

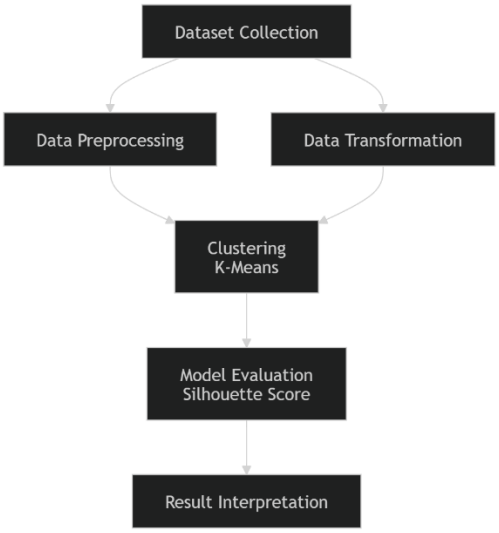
karena itu, analisis yang menggabungkan beberapa indikator popularitas dapat menghasilkan pemahaman yang lebih kaya dibandingkan penggunaan satu indikator saja [3].

Pendekatan data mining dapat digunakan untuk mengekstraksi pola tersembunyi dari data dalam jumlah besar. Salah satu pendekatan yang relevan adalah *unsupervised learning*, khususnya *clustering*, karena mampu mengelompokkan objek tanpa label [1], [2]. Dalam konteks penelitian ini, objek yang dianalisis adalah lagu-lagu terpopuler Spotify, sedangkan dasar pengelompokan adalah indikator popularitas *streaming* [3]. K-Means dipilih karena sederhana, mudah diimplementasikan, dan efisien untuk data numerik berukuran sedang [1], [4]. Namun, K-Means memiliki keterbatasan seperti sensitif terhadap skala data, *outlier*, serta kebutuhan menentukan jumlah *cluster* yang optimal [2]. Karena itu, penelitian ini menggunakan standardisasi fitur serta evaluasi jumlah *cluster* dengan kombinasi Elbow Method dan Silhouette Score [3].

Beberapa penelitian sebelumnya menggunakan K-Means pada domain musik, baik untuk pengelompokan lagu maupun untuk mendukung sistem rekomendasi [1], [4], [2]. Penelitian-penelitian tersebut menunjukkan bahwa *clustering* dapat membantu membentuk segmen yang dapat diinterpretasikan untuk mendukung pengambilan keputusan [1], [3]. Meski demikian, masih diperlukan analisis yang fokus pada pengelompokan lagu terpopuler berdasarkan indikator popularitas *streaming* agar dapat membedakan lagu yang kuat secara historis, lagu yang sedang naik, dan lagu yang performanya stabil [3].

2. METODE

Pada bagian ini dijelaskan alur penelitian secara ringkas, yang ditunjukkan pada Gambar 1, mulai dari persiapan dataset, preprocessing dan pembentukan fitur, pemodelan *clustering* menggunakan K-Means, penentuan jumlah *cluster* (Elbow dan Silhouette), hingga evaluasi kualitas *cluster* menggunakan beberapa indeks validasi [3], [9].



Gambar 1. Bagan alur penelitian

2.1. Desain Penelitian

Penelitian ini menggunakan pendekatan data mining dengan *unsupervised learning* untuk membentuk segmentasi lagu berdasarkan pola popularitas *streaming*. Metode utama yang digunakan adalah *clustering* K-Means yang umum digunakan dalam pengelompokan data musik dan sistem rekomendasi untuk membentuk segmen yang dapat diinterpretasikan [1], [4], [2].

2.2. Dataset

Dataset yang digunakan adalah Spotify Most Streamed Songs (file: Spotify most streamed.csv). Dataset berisi 2.500 baris dan 3 kolom awal, yaitu Artist and Title, Streams, dan Daily. Kolom *Streams* dan *Daily* masih berupa teks dengan pemisah ribuan, serta kolom *Daily* memiliki sejumlah kecil nilai hilang, sehingga diperlukan tahap pembersihan dan konversi data sebelum pemodelan.

2.3. Exploratory Data Analysis

Tahap EDA dilakukan untuk memahami struktur data, tipe data, dan kualitas data. Pemeriksaan mencakup: (1) ukuran dataset dan tipe data tiap kolom, (2) pengecekan nilai hilang, serta (3) ringkasan statistik deskriptif pada variabel numerik setelah konversi untuk melihat sebaran dan potensi ketimpangan nilai.

2.4. Data Cleaning dan Preprocessing

Tahapan preprocessing disusun agar data siap digunakan pada algoritma berbasis jarak. Pertama, nilai pada *Streams* dan *Daily* dibersihkan dari karakter non-angka (misalnya koma pemisah ribuan), lalu dikonversi ke numerik dan disimpan sebagai *Streams_num* dan *Daily_num*. Kedua, kolom *Artist* and *Title* dipisahkan menggunakan delimiter " - " menjadi *Artist* dan *Title*. Jika judul tidak terdeteksi, *Title* diisi dengan "Unknown Title". Kolom *Artist* dan *Title* tidak digunakan sebagai fitur model, namun dipertahankan untuk kebutuhan interpretasi hasil *cluster*.

2.5. Fitur yang Digunakan dalam Clustering

Fitur inti yang digunakan sebagai input clustering adalah *log_streams*, *log_daily*, dan *daily_to_streams*. Pemilihan fitur ini konsisten dengan tujuan penelitian yang berfokus pada indikator popularitas *streaming*.

2.6. Standarisasi Fitur

Karena K-Means sensitif terhadap perbedaan skala, fitur distandardisasi agar setiap variabel berkontribusi sebanding pada jarak Euclidean [2]. Standardisasi menggunakan z-score:

$$z = (x - \mu) / \sigma \quad (1)$$

dengan μ adalah rata-rata dan σ adalah simpangan baku. Praktik standardisasi dan evaluasi *clustering* dengan metrik validasi juga umum dibahas pada literatur evaluasi *clustering* [5].

2.7. Algoritma K-Means

K-Means membagi data menjadi *k* cluster dengan meminimalkan jarak kuadrat titik data terhadap centroid cluster. Jarak yang digunakan adalah Euclidean:

$$d(x, c) = \text{sqrt}(\Sigma (xj - cj)^2) \quad (2)$$

K-Means dipilih karena sederhana, efisien, serta banyak digunakan dalam studi pengelompokan musik dan rekomendasi berbasis *clustering* [1], [4].

2.8. Penentuan Jumlah Cluster

Jumlah cluster ditentukan dengan Elbow Method menggunakan Inertia untuk beberapa nilai *k*. Titik *elbow* dipilih saat penurunan Inertia mulai melandai. Silhouette Score untuk membandingkan kualitas pemisahan *cluster* pada beberapa *k*. Nilai lebih tinggi menunjukkan pemisahan *cluster* lebih baik. Pendekatan Elbow dan Silhouette merupakan cara yang lazim digunakan dalam studi *clustering* lagu Spotify maupun evaluasi segmentasi berbasis K-Means [3], [5].

2.9. Evaluasi Model Clustering

Kualitas *clustering* dievaluasi menggunakan beberapa metrik agar penilaian tidak bergantung pada satu indikator saja [5], yaitu:

1. Inertia (semakin rendah semakin baik).
2. Silhouette Score (semakin tinggi semakin baik).
3. Calinski–Harabasz Index (semakin tinggi semakin baik).
4. Davies–Bouldin Index (semakin rendah semakin baik).

2.10. Lingkungan Implementasi

Eksperimen dilakukan menggunakan Python untuk pengolahan data, transformasi, standardisasi, pelatihan model K-Means, serta visualisasi evaluasi (grafik Elbow dan Silhouette) dan visualisasi sebaran *cluster*.

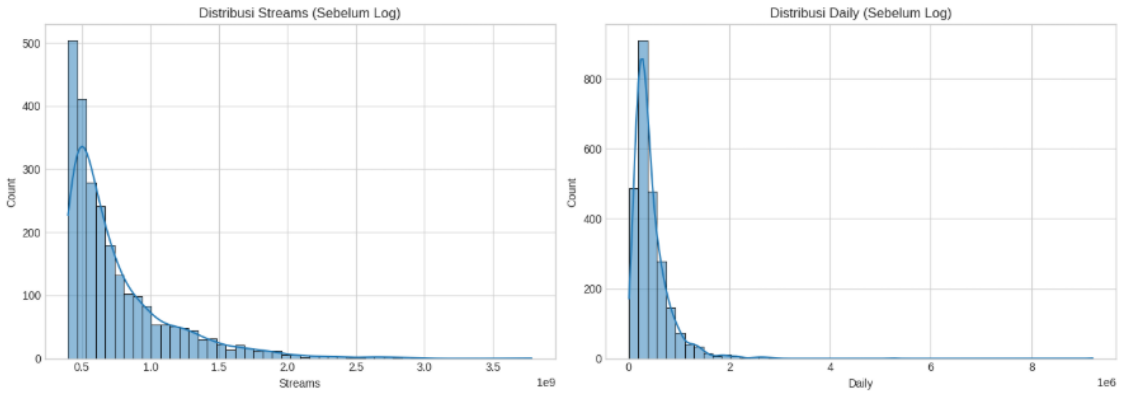
3. HASIL DAN PEMBAHASAN

Pada bagian ini disajikan hasil pemodelan *clustering* pada dataset Spotify *most streamed* serta pembahasan interpretasi *cluster* yang terbentuk. Hasil mencakup perbandingan model *baseline* dan model final, proses pemilihan jumlah *cluster*, serta profil karakteristik tiap *cluster* berdasarkan indikator popularitas *streaming*.

3.1. Ringkasan Eksplorasi Data

Tahap eksplorasi data dilakukan untuk memahami karakteristik awal variabel yang digunakan dalam pemodelan. Analisis ini penting untuk mengidentifikasi pola distribusi, potensi ketimpangan nilai, serta memastikan kesiapan data sebelum diterapkan pada algoritma berbasis jarak seperti K-Means.

Setelah data dikonversi ke numerik, sebaran *Streams_num* dan *Daily_num* menunjukkan ketimpangan (*right-skewed*), yaitu sebagian kecil lagu memiliki nilai streaming yang jauh lebih tinggi dibandingkan mayoritas lagu. Pola ketimpangan ini dapat dilihat pada Gambar 2.



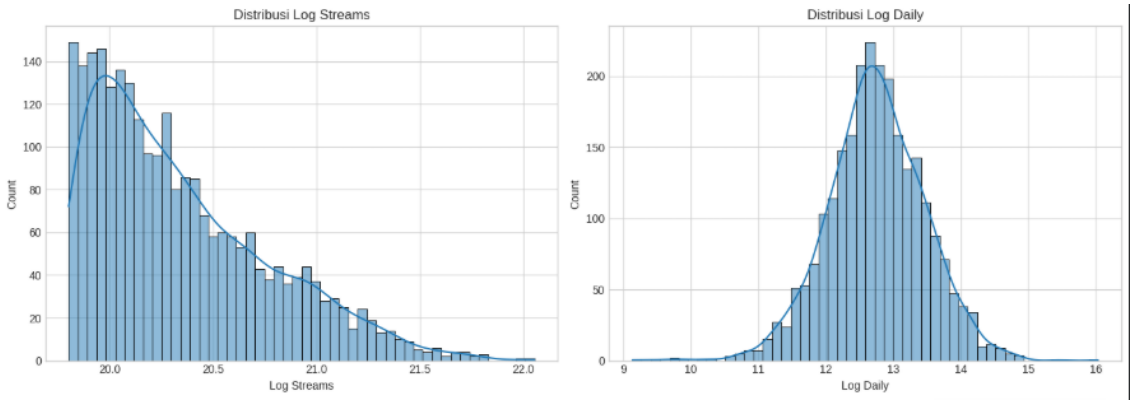
Gambar 2. Distribusi Streams dan Daily sebelum log

Pada Gambar 2, distribusi data memiliki ekor panjang di sisi kanan yang menandakan keberadaan nilai ekstrem dengan skala jauh di atas rata-rata. Sebaran seperti ini umum ditemui pada data popularitas dan dapat memengaruhi perhitungan jarak pada metode berbasis *centroid*, sehingga transformasi logaritmik digunakan untuk mengurangi dominasi nilai ekstrem [2]. Selain itu, jumlah nilai hilang pada *Daily* tergolong sangat kecil dan telah ditangani pada tahap *preprocessing*, sehingga data siap digunakan untuk pemodelan.

Distribusi yang sangat miring seperti ini umum ditemui pada data popularitas digital, karena hanya sebagian kecil lagu yang mencapai tingkat pemutaran sangat tinggi, sementara sebagian besar lainnya berada pada kisaran menengah atau rendah. Ketimpangan tersebut

berpotensi memengaruhi perhitungan jarak Euclidean pada metode berbasis centroid, karena nilai ekstrem dapat mendominasi pembentukan cluster [2].

Untuk mengurangi dominasi nilai ekstrem dan menstabilkan variansi, dilakukan transformasi logaritmik pada kedua variable. Hasil transformasi ditunjukkan pada Gambar 3.



Gambar 3. Distribusi Streams dan Daily Sesudah log

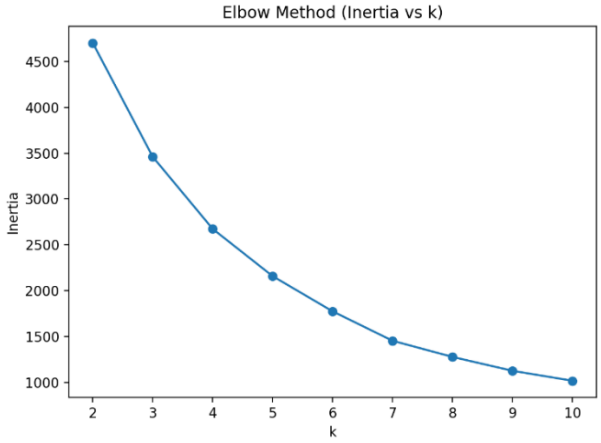
Pada Gambar 3, distribusi menjadi lebih seimbang dan mendekati simetris dibandingkan distribusi awal. Transformasi ini membantu memastikan bahwa setiap fitur memberikan kontribusi yang lebih proporsional dalam proses *clustering*.

3.2. Hasil Clustering Menggunakan K-Means (baseline)

Model baseline digunakan sebagai pembanding awal dengan menetapkan jumlah cluster $k = 3$ tanpa optimasi lebih lanjut. Model dilatih menggunakan tiga fitur hasil *preprocessing*, yaitu *log_streams*, *log_daily*, dan *daily_to_streams* yang telah distandardisasi. Hasil evaluasi model *baseline* menunjukkan Silhouette Score sebesar 0,448924 dan Inertia sebesar 3459,676. Nilai Silhouette tersebut mengindikasikan pemisahan antar *cluster* cukup baik, namun masih terdapat tumpang tindih antar kelompok sehingga segmentasi yang terbentuk masih relatif kasar. *Baseline* ini digunakan sebagai titik awal sebelum penentuan jumlah *cluster* yang lebih representatif, sebagaimana praktik umum pada studi *clustering* lagu Spotify [3].

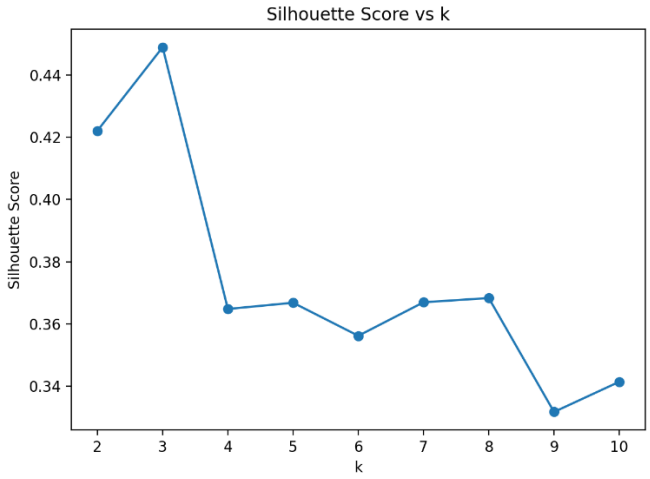
3.3. Penentuan Jumlah Cluster (model final)

Untuk memperoleh jumlah *cluster* yang lebih representatif, dilakukan penentuan k menggunakan kombinasi Elbow Method dan Silhouette Score. Elbow Method dihitung pada rentang $k = 2$ sampai $k = 10$ dan divisualisasikan pada Gambar 4.



Gambar 4. Grafik Elbow Method (Inertia vs *K*)

Pada Gambar 5 grafik menunjukkan penurunan Inertia yang cukup tajam dari k = 2 ke k = 3 dan dari k = 3 ke k = 4, kemudian mulai melandai setelah k = 4. Pola ini mengindikasikan adanya titik "tekukan" (*elbow*) di sekitar k = 4, sehingga penambahan cluster setelah nilai tersebut tidak memberikan pengurangan Inertia yang signifikan.



Gambar 5. Grafik Silhouette Score vs k

Pemilihan k pada praktiknya sering mempertimbangkan trade-off antara kualitas separasi dan kebutuhan interpretasi segmen, karena metrik validasi dapat memberikan sinyal yang tidak selalu identik untuk setiap kandidat k [5]. Berdasarkan pertimbangan Elbow dan kestabilan Silhouette, yang masing-masing ditunjukkan pada Gambar 4 dan Gambar 5, untuk tujuan segmentasi yang lebih rinci, dipilih k = 4 sebagai jumlah cluster untuk model final [3], [5].

3.4. Hasil Clustering menggunakan K-Means (model final)

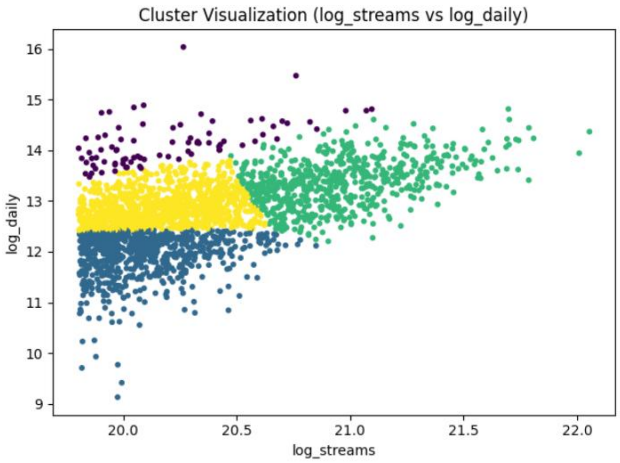
Model final dilatih menggunakan k = 4 dengan fitur *log_streams*, *log_daily*, dan *daily_to_streams* yang telah distandardisasi. Hasil evaluasi menunjukkan Silhouette Score sebesar 0,364835 dan Inertia sebesar 2673,011. Selain itu, diperoleh Calinski–Harabasz sebesar 1502,451 dan Davies–Bouldin sebesar 0,858136, seperti yang terangkum dalam Tabel 1. Penggunaan beberapa metrik ini bertujuan memberi perspektif lebih lengkap mengenai kekompakan dan pemisahan *cluster* [5].

Tabel 1. Evaluasi Model Clustering K-Means pada Dataset Spotify

Model	k	Inertia (lower better)	Silhouette (higher better)	Calinski-Harabasz (higher better)	Davies-Bouldin (lower better)
Baseline K-Means	3	3459.676	0.448924	1458.04	**0.84534**
Final K-Means	4	2673.011	0.364835	1502.451	**0.858136**

Dari hasil pada Tabel 1, model final menunjukkan *cluster* yang lebih rapat (Inertia lebih rendah) dan struktur *cluster* secara keseluruhan sedikit membaik (Calinski–Harabasz meningkat). Namun, Silhouette Score menurun dibanding *baseline* dan Davies–Bouldin sedikit meningkat. Hal ini menunjukkan adanya *trade-off* saat segmentasi dibuat lebih rinci, sehingga kualitas separasi antar *cluster* sedikit berkurang meskipun *cluster* menjadi lebih kompak [5]. Model final tetap dipilih karena

menghasilkan empat segmen yang lebih informatif untuk analisis lanjutan dan interpretasi bisnis/kurasi [3], [6], seperti yang diilustrasikan dalam grafik pada Gambar 6.



Gambar 6. Visualisasi Clustering (*log_streams* vs *log_daily*)

Pada Gambar 6 terlihat masing-masing *cluster* menempati area yang relatif berbeda meskipun terdapat sedikit tumpang tindih pada area nilai *log_streams* yang berdekatan. Kondisi *overlap* seperti ini masih lazim terjadi pada data popularitas yang memiliki transisi gradual antar kelompok, terutama ketika indikator yang dipakai bersifat agregat [3], [2].

3.5. Interpretasi cluster pada lagu Spotify most streamed

Interpretasi cluster dilakukan dengan menghitung profil tiap cluster menggunakan rata-rata dan median pada *Streams_num*, *Daily_num*, *log_streams*, *log_daily*, serta *daily_to_streams*. Ringkasan profil cluster dapat dilihat pada Tabel 2.

Tabel 2. Profile Cluster (Model Final k = 4)

Metric	Cluster 0	Cluster 1	Cluster2	Cluster 3
n	81	758	640	**1021**
Streams_mean	6,31E+08	5,29E+08	1,19E+09	**5,65E+08**
Streams_median	5,29E+08	5,09E+08	1,19E+09	**5,65E+08**
Daily_median	1.583.092	164.320,9	703.779,4	**403.917,2**
Log_streams_mean	20,20602	20,08656	20,94099	**20,16750**
Log_daily_mean	14,15154	11,93533	13,35643	**12,85917**
Ratio_mean	0,002585	0,000314	0,000557	**0,000706**
Ratio_median	0,002053	0,000304	0,000514	**0,000645**

Berdasarkan profil pada Tabel 2 , interpretasi setiap cluster adalah:

1. Cluster 0 (super-trending)
   Cluster 0 memiliki rasio *daily_to_streams* tertinggi (sekitar 0,2585%) dan *daily_mean* paling tinggi. Ini menunjukkan kelompok lagu yang sedang sangat aktif diputar setiap hari, sehingga dapat dipandang sebagai lagu dengan momentum tren yang sangat kuat. Segmentasi berbasis aktivitas terkini seperti ini relevan untuk kebutuhan rekomendasi dan kurasi *playlist* yang adaptif terhadap tren [1], [6].

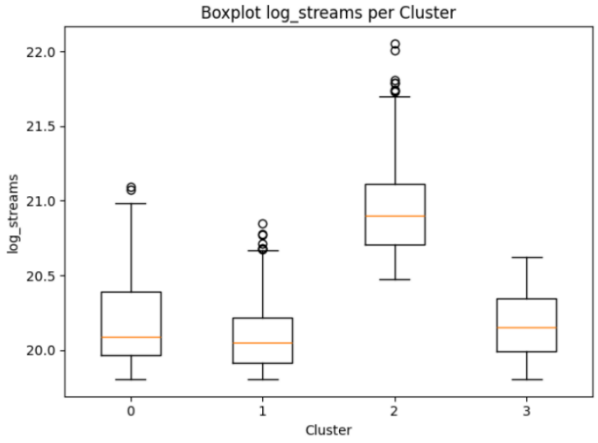
2. Cluster 1 (popular historis, aktivitas harian rendah)
   Cluster 1 memiliki *daily_mean* paling rendah dan rasio *daily_to_streams* paling kecil (sekitar 0,0314%), meskipun total *streams* relatif tinggi. Kelompok ini menunjukkan lagu-lagu yang

telah mengumpulkan *streams* besar secara historis namun momentum harian menurun, sehingga lebih cocok diposisikan sebagai katalog "*evergreen/legacy*" dibanding konten promosi agresif [4], [6].

3. Cluster 2 (mega hits)
   Cluster 2 memiliki total *streams* tertinggi (sekitar 1,30 miliar) dan *daily_mean* cukup tinggi. Kelompok ini merepresentasikan lagu-lagu yang sangat kuat secara kumulatif dan masih sering diputar, sehingga dapat dikategorikan sebagai *mega hits* yang tetap relevan. Pengelompokan lagu "hit besar" seperti ini sering menjadi komponen penting pada strategi rekomendasi dan penempatan konten utama [1], [4], [7].

4. Cluster 3 (konsisten stabil)
   Cluster 3 memiliki jumlah anggota terbesar, dengan total *streams* menengah-tinggi dan aktivitas harian yang stabil. Rasio *daily_to_streams* berada di tingkat menengah, sehingga cluster ini menggambarkan lagu-lagu yang performanya konsisten. Kelompok seperti ini cocok untuk *playlist* tematik atau rekomendasi yang menekankan kestabilan konsumsi, bukan hanya puncak tren sesaat [5], [8].

Untuk memperjelas perbedaan distribusi popularitas total antar *cluster*, digunakan boxplot *log_streams* per *cluster* seperti pada Gambar 7.



Gambar 7. Boxplot log_streams per Cluster

Pada Gambar 7, boxplot menunjukkan Cluster 2 cenderung memiliki *log_streams* lebih tinggi, sedangkan Cluster 1 memiliki median *log_streams* lebih rendah dan dispersi lebih kecil. Perbandingan distribusi seperti ini membantu memperjelas karakteristik tiap *cluster* di luar rata-rata saja, karena data popularitas sering memiliki sebaran yang lebar [5].

3.6. Implikasi Hasil Clustering

Hasil segmentasi ini dapat dimanfaatkan dalam beberapa konteks. Pertama, Cluster 0 dapat diprioritaskan untuk strategi promosi dan penempatan pada *playlist* utama karena berisi lagu yang sedang sangat aktif diputar, sejalan dengan praktik sistem rekomendasi yang menekankan adaptasi terhadap tren [1], [9], [10]. Kedua, Cluster 2 dapat dipandang sebagai katalog inti *mega hits* yang penting dipertahankan eksposurnya karena memiliki nilai kumulatif sangat tinggi sekaligus masih aktif diputar [1], [4], [7]. Ketiga, Cluster 3 cocok untuk *playlist* tematik yang membutuhkan lagu dengan performa stabil dan konsisten, sehingga dapat meningkatkan kenyamanan pengalaman pengguna dalam rekomendasi jangka menengah [8], [11]. Keempat, Cluster 1 dapat dimanfaatkan sebagai katalog populer historis yang tetap relevan, namun biasanya tidak memerlukan promosi agresif dan lebih tepat untuk konteks nostalgia atau *long-tail content* [6], [9].

Selain itu, segmentasi berbasis indikator *streaming* ini dapat melengkapi pendekatan rekomendasi yang berfokus pada fitur audio. Beberapa penelitian menunjukkan pemodelan berbasis fitur audio dapat digunakan untuk prediksi atau rekomendasi, namun indikator *streaming* menangkap refleksi langsung perilaku pendengar dan dinamika tren [12], [10].

3.7. Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan. Pertama, fitur yang digunakan hanya indikator kuantitatif agregat (*total streams* dan *daily streams* beserta turunannya), sehingga faktor lain seperti genre, tahun rilis, wilayah pendengar, strategi promosi, maupun fitur audio detail belum dimasukkan. Padahal, fitur audio sering digunakan pada penelitian lain untuk memodelkan popularitas atau membangun rekomendasi yang lebih kaya [12], [13]. Kedua, metode *clustering* yang digunakan hanya K-Means yang sensitif terhadap skala dan *outlier* serta mengasumsikan bentuk *cluster* relatif "bulat", sehingga algoritma lain seperti DBSCAN atau Gaussian Mixture Model belum dieksplorasi. Studi komparatif pada berbagai domain menekankan pentingnya membandingkan beberapa metode *clustering* untuk memastikan struktur *cluster* tidak terlalu bergantung pada asumsi satu algoritma [14], [15]. Penelitian lanjutan dapat menambah fitur audio dan metadata, membandingkan algoritma *clustering*, serta menguji penerapan segmentasi pada skenario rekomendasi personal yang lebih adaptif [11], [10], [16].

4. KESIMPULAN

Penelitian ini berhasil melakukan pengelompokan lagu terpopuler Spotify menggunakan algoritma K-Means berdasarkan indikator popularitas *streaming*. Data mentah terlebih dahulu diproses melalui konversi numerik, penanganan nilai hilang, transformasi logaritmik, serta pembentukan fitur rasio *daily_to_streams*, kemudian distandardisasi agar sesuai untuk pemodelan berbasis jarak. Penentuan jumlah cluster menggunakan kombinasi Elbow Method dan Silhouette Score menghasilkan pilihan $k = 4$ sebagai konfigurasi yang paling representatif untuk segmentasi yang lebih rinci. Hasil clustering membentuk empat kelompok utama, yaitu Cluster *super-trending* dengan aktivitas harian dan rasio *daily_to_streams* tertinggi, Cluster lagu populer historis dengan aktivitas harian rendah, Cluster *mega hits* dengan total *streams* tertinggi dan tetap aktif diputar, serta Cluster lagu dengan performa stabil yang memiliki popularitas menengah-tinggi dan pemutaran harian konsisten. Segmentasi ini memberikan wawasan yang dapat dimanfaatkan untuk mendukung kurasi *playlist*, prioritas promosi, serta analisis tren popularitas lagu berbasis data. Ke depan, penelitian dapat dikembangkan dengan menambahkan fitur audio dan metadata lain serta membandingkan beberapa algoritma *clustering* agar hasil segmentasi lebih komprehensif.

Ucapan Terima Kasih

Penulis menyampaikan apresiasi dan terima kasih kepada Ibu Agiska Ria Supriyatna, S.Si., M.T.I. serta Ibu Dian Ayu Afifah, S.Si., M.Sc. atas pendampingan dan arahan yang diberikan sepanjang proses penelitian. Saran, koreksi, dan wawasan yang beliau berikan sangat membantu penulis dalam menyusun dan menyempurnakan penelitian ini.

DAFTAR PUSTAKA

[1] S. Mukhopadhyay, A. Kumar, D. Parashar, and M. Singh, "Enhanced Music Recommendation Systems: A Comparative Study of Content-Based Filtering and K-Means Clustering Approaches," *RIA*, vol. 38, no. 1, pp. 365–376, Feb. 2024, doi: 10.18280/ria.380138.

[2] A. Parthasarathy, "Music Recommendation Using Machine Learning Algorithms," vol. 11, no. 11, 2023. doi: http://doi.one/10.1729/Journal.36826.

[3] N. Rohman and A. Wibowo, "Clustering Of Popular Spotify Songs In 2023 Using K-Means Method And Silhouette Coefficient," *pilar*, vol. 20, no. 1, pp. 18–24, Apr. 2024, doi: 10.33480/pilar.v20i1.4937.

[4] B. Daga, H. Kadam, S. Shrungare, S. Fernandes, and A. Johsnon, "Music Recommendation System," *IJCTT*, vol. 71, no. 5, pp. 26–36, May 2023, doi: 10.14445/22312803/IJCTT-V71I5P105.

[5] M. Gagolewski, M. Bartoszuk, and A. Cena, "Are Cluster Validity Measures (In)valid?," *Information Sciences*, vol. 581, pp. 620–636, Dec. 2021, doi: 10.1016/j.ins.2021.10.004.
[6] H. Magadum, H. K. Azad, H. Patel, and R. H. R, "Music Recommendation Using Dynamic Feedback And Content-Based Filtering," *Multimedia Tools and Applications*, vol. 83, no. 32, pp. 77469–77488, Sep. 2024, doi: 10.1007/s11042-024-18636-8.
[7] R. Kumar and Rakesh, "Music Recommendation System Using Machine Learning," in *2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N)*, Greater Noida, India: IEEE, Dec. 2022, pp. 572–576. doi: 10.1109/ICAC3N56670.2022.10074362.
[8] L. E. Ekemeyong Awong and T. Zielinska, "Comparative Analysis of the Clustering Quality in Self-Organizing Maps for Human Posture Classification," *Sensors*, vol. 23, no. 18, p. 7925, Sep. 2023, doi: 10.3390/s23187925.
[9] M. Garanayak, S. K. Nayak, Sangeetha K., T. Choudhury, and Shitharth S., "Content and Popularity-Based Music Recommendation System:," *International Journal of Information System Modeling and Design*, vol. 13, no. 7, pp. 1–14, Dec. 2022, doi: 10.4018/ijismd.315027.
[10] L. Liu *et al.*, "Personalized Music Recommendation Algorithm Based On Machine Learning," *Multimedia Systems*, vol. 31, no. 3, p. 166, Jun. 2025, doi: 10.1007/s00530-025-01749-x.
[11] M. G. Galety, R. Thiagarajan, R. Sangeetha, L. K. B. Vignesh, S. Arun, and R. Krishnamoorthy, "Personalized Music Recommendation model based on Machine Learning," in *2022 8th International Conference on Smart Structures and Systems (ICSSS)*, Chennai, India: IEEE, Apr. 2022, pp. 1–6. doi: 10.1109/ICSSS54381.2022.9782288.
[12] J. S. Gulmatico, J. A. B. Susa, M. A. F. Malbog, A. Acoba, M. D. Nipas, and J. N. Mindoro, "SpotiPred: A Machine Learning Approach Prediction of Spotify Music Popularity by Audio Features," in *2022 Second International Conference on Power, Control and Computing Technologies (ICPC2T)*, Raipur, India: IEEE, Mar. 2022, pp. 1–5. doi: 10.1109/ICPC2T53885.2022.9776765.
[13] J. T. Anthony, G. E. Christian, V. Evanlim, H. Lucky, and D. Suhartono, "The Utilization of Content Based Filtering for Spotify Music Recommendation," in *2022 International Conference on Informatics Electrical and Electronics (ICIEE)*, Yogyakarta, Indonesia: IEEE, Oct. 2022, pp. 1–4. doi: 10.1109/ICIEE55596.2022.10010097.
[14] M. Alizade, R. Kheni, S. Price, B. C. Sousa, D. L. Cote, and R. Neamtu, "A Comparative Study of Clustering Methods for Nanoindentation Mapping Data," *Integr Mater Manuf Innov*, vol. 13, no. 2, pp. 526–540, Jun. 2024, doi: 10.1007/s40192-024-00349-3.
[15] B. A. Hassan, N. B. Tayfor, A. A. Hassan, A. M. Ahmed, T. A. Rashid, and N. N. Abdalla, "From A-to-Z review of clustering validation indices," *Neurocomputing*, vol. 601, p. 128198, Oct. 2024, doi: 10.1016/j.neucom.2024.128198.
[16] A. Singh and S. Gupta, "Recommendation System Algorithms For Music Therapy," in *2023 13th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, Noida, India: IEEE, Jan. 2023, pp. 138–143. doi: 10.1109/Confluence56041.2023.10048894.