

Analisis Preferensi Penonton Anime berbasis *Genre* Film menggunakan Metode K-Means

Niken Zalzabila¹, Rastri Prathivi²

^{1,2}Universitas Semarang, Indonesia

E-mail: niken100974@gmail.com¹, vivi@usm.ac.id²

Abstract

This study aims to analyze anime audience preferences based on genres using the K-Means clustering algorithm. The dataset consists of 100 popular anime titles with features such as ratings, votes, and genres. The research steps include data preprocessing, clustering with the Elbow method to determine the optimal number of clusters, and applying the K-Means algorithm. The clustering results revealed four clusters with unique characteristics, highlighting differences in popularity and genre preferences. Evaluation using the Confusion Matrix shows a model accuracy of 95%, while the Silhouette score of 0.285 indicates adequate cluster separation. These findings are expected to provide insights for streaming platforms to deliver more personalized and relevant anime recommendations to viewers.

Keywords: *audience preferences, anime, K-Means clustering, genre, streaming platforms.*

Abstrak

Penelitian ini bertujuan untuk menganalisis preferensi penonton anime berdasarkan genre menggunakan algoritma K-Means clustering. Dataset yang digunakan mencakup 100 judul anime terpopuler dengan fitur seperti rating, votes, dan genre. Tahapan penelitian meliputi preprocessing data, pengelompokan data menggunakan metode Elbow untuk menentukan jumlah cluster optimal, serta penerapan algoritma K-Means. Hasil clustering menghasilkan empat cluster dengan karakteristik unik, mencakup perbedaan dalam popularitas dan preferensi genre. Evaluasi menggunakan Confusion Matrix menunjukkan akurasi model sebesar 95%, sedangkan nilai Silhouette score sebesar 0,285 mengindikasikan pemisahan cluster yang memadai. Temuan penelitian ini diharapkan dapat memberikan wawasan untuk platform streaming dalam menyediakan rekomendasi anime yang lebih personal dan relevan bagi penonton.

Kata Kunci: *preferensi penonton, anime, K-Means clustering, genre, platform streaming.*

1. Pendahuluan

Anime, jenis hiburan populer di Jepang, memiliki banyak penonton karena menyediakan berbagai *genre* yang dapat dinikmati oleh berbagai kelompok usia [1]. Dengan beragam *genre* yang ditawarkan, mulai dari aksi, petualangan, komedi, hingga *drama* dan fiksi ilmiah, anime telah menarik audiens dari berbagai latar belakang dan usia. Menurut laporan dari Grand View Research, pasar anime global diperkirakan mencapai nilai \$25,46 miliar pada tahun 2020 dan diproyeksikan tumbuh dengan CAGR 9,5% dari 2021 hingga 2028 [2]. Pertumbuhan yang signifikan ini menunjukkan pentingnya memahami preferensi penonton anime untuk industri hiburan dan *platform streaming*.

Perkembangan anime sebagai fenomena budaya global telah menarik perhatian akademisi. Strategi manajemen dari brand Uniqlo dan strategi kolaborasi mereka dengan program anime terbukti sangat berhasil. Kolaborasi ini tidak hanya menarik penggemar

anime tetapi juga konsumen yang mencari kenyamanan dalam pakaian sehari-hari, sehingga meningkatkan penjualan produk Uniqlo secara signifikan [3]. Penelitian oleh Pratama (2020) juga menganalisis bagaimana anime telah menjadi produk budaya transnasional yang mempengaruhi industri media di berbagai negara. Studi ini menunjukkan bahwa preferensi penonton anime tidak hanya dipengaruhi oleh faktor lokal tetapi juga oleh tren global [4].

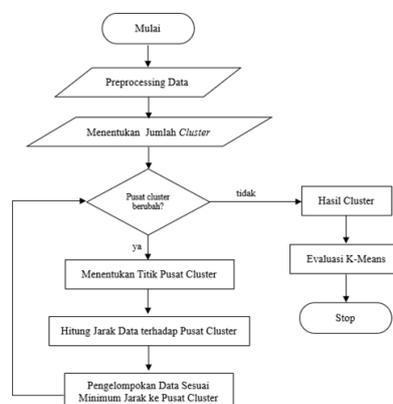
Dalam era big data dan personalisasi konten, pemahaman yang mendalam tentang preferensi penonton menjadi sangat penting bagi platform streaming dan studio produksi anime. Anime, bentuk hiburan populer asal Jepang, semakin banyak tersedia melalui layanan streaming di seluruh dunia untuk memenuhi permintaan yang terus meningkat [5]. Popularitas anime mendorong banyak layanan streaming untuk menyertakan anime dalam katalog mereka. Faktor-faktor seperti interaktivitas dan kekayaan konten turut memengaruhi persepsi manfaat (*perceived usefulness*), sementara persepsi manfaat, persepsi kemudahan penggunaan, harga, dan pengaruh sosial memengaruhi kemauan untuk berlangganan (*willingness to subscribe*). Sebaliknya, ketersediaan alternatif gratis memiliki pengaruh negatif terhadap kemauan untuk berlangganan [6].

Metode tradisional dalam menganalisis preferensi penonton seringkali mengandalkan survei dan fokus grup, yang memiliki keterbatasan dalam skala dan objektivitas. Di sisi lain, pendekatan berbasis data menggunakan teknik machine learning dapat memberikan wawasan yang lebih komprehensif dan objektif [7]. Salah satu metode yang potensial untuk digunakan dalam konteks ini adalah algoritma K-Means *clustering*.

K-Means *clustering* adalah teknik unsupervised learning yang dapat mengelompokkan data ke dalam *cluster* berdasarkan kesamaan karakteristik [8] [9]. Dalam konteks preferensi anime, K-Means dapat digunakan untuk mengidentifikasi pola-pola dalam pilihan *genre* penonton, memungkinkan pemahaman yang lebih baik tentang segmen audiens dan tren yang muncul. Penelitian ini bertujuan untuk menganalisis preferensi penonton anime berdasarkan *genre* film menggunakan metode K-Means. Mengacu kepada penikmat anime yang jumlahnya tidak sedikit, maka perlu adanya pengelompokan terlebih dahulu sebelum memberikan rekomendasi agar hasil daftar rekomendasi menjadi lebih akurat [10]. Dengan memanfaatkan dataset yang berisi informasi tentang peringkat anime dan *genre*-nya, penelitian ini akan mengeksplorasi bagaimana penonton dapat dikelompokkan berdasarkan preferensi *genre* mereka. Hasil dari analisis ini diharapkan dapat memberikan wawasan berharga bagi industri anime, *platform streaming*, dan peneliti di bidang media dan hiburan [11].

2. Metodologi Penelitian

Penelitian ini menggunakan dataset yang terdiri dari 100 judul anime terpopuler untuk menganalisis preferensi penonton berdasarkan *genre* film.



Gambar 1. Metode Penelitian yang akan Dilakukan

Metode yang digunakan adalah K-Means *clustering*, yang merupakan teknik *unsupervised learning* untuk mengelompokkan data ke dalam *cluster* berdasarkan kesamaan karakteristik [12]. Dalam penelitian ini, fitur utama yang digunakan untuk *clustering* meliputi *rating*, *votes*, tahun rilis, *durasi*, dan *genre*. Pendekatan ini bertujuan untuk menemukan pola dalam data yang memungkinkan segmentasi audiens berdasarkan preferensi *genre* anime.

2.1. Preprocessing Data

Tahap preprocessing bertujuan untuk memastikan data siap digunakan dalam algoritma K-Means *clustering*. Langkah-langkahnya meliputi :

- 1) **Menghapus data dengan nilai NaN / Kosong** : Data dengan nilai NaN atau kosong dihapus.
- 2) **Konversi Data Numerik** : Semua data numerik disiapkan agar dapat diproses, termasuk menghapus tanda koma pada kolom *votes* untuk memastikan format data konsisten
- 3) **Encoding Genre menggunakan One-Hot Encoding** : Kolom *genre* dikonversi menjadi data numerik menggunakan metode *One-Hot Encoding*, menghasilkan kolom biner (0 atau 1) untuk setiap *genre*
- 4) **Pemilihan Fitur** : Fitur *rating*, *votes*, dan *genre* dipilih sebagai variabel utama karena relevan untuk mengukur popularitas dan kepuasan anime.
- 5) **Normalisasi Fitur** : Fitur yang dipilih di-scaling agar memiliki rentang nilai yang sama, menggunakan metode standardisasi, sehingga algoritma K-Means dapat bekerja lebih optimal.

2.2. Clustering Menggunakan K-Means

Data yang telah diproses dikelompokkan menggunakan algoritma K-Means dengan langkah berikut:

- 1) **Menentukan Jumlah Cluster (K)** : Metode Elbow digunakan untuk menentukan jumlah *cluster* optimal. Grafik SSE (Sum of Squared Errors) diplot untuk berbagai nilai K, dan titik "elbow" dipilih sebagai jumlah *cluster* yang optimal. *Elbow method* melibatkan pemetaan nilai SSE pada berbagai jumlah *cluster* dan mencari titik "siku" (elbow) pada grafik. Titik ini adalah tempat penurunan SSE mulai melambat, menandakan bahwa menambah jumlah *cluster* setelah titik ini memberikan manfaat yang lebih sedikit [13].
- 2) **Penerapan K-Means Clustering** : Setelah nilai K ditentukan, data dikelompokkan ke dalam K *cluster* berdasarkan fitur seperti *rating*, *votes*, dan *genre*. Setiap *cluster* merepresentasikan kelompok anime dengan karakteristik serupa.

2.3. Evaluasi Hasil Clustering

Hasil *clustering* dievaluasi dengan dua metode utama *Silhouette score* untuk mengukur kualitas *clustering* dan *Confusion matrix* untuk mengukur akurasi pengelompokan. Berikut adalah penjelasan dan rumus dari masing-masing metode evaluasi.

1) Silhouette score

Silhouette score digunakan untuk mengukur seberapa baik titik-titik data dalam setiap *cluster*. Metode ini mengukur seberapa dekat setiap titik data dalam suatu *cluster* dan seberapa jauh jarak titik data tersebut dari *cluster* lain. Nilai *Silhouette score* berkisar antara -1 hingga 1, dengan nilai mendekati 1 menunjukkan bahwa data telah dikelompokkan dengan baik dan memiliki pemisahan yang jelas antar *cluster* [12]. *Silhouette score* untuk setiap titik data *i* dihitung dengan rumus sebagai berikut :

$$s(j) = \frac{b(j) - a(j)}{\max[a(j), b(j)]} \quad (1)$$

Keterangan :

s = nilai silhouette.

$a(j)$ = rata-rata jarak dari objek j dengan objek yang berada di *cluster* berbeda.

$b(j)$ = rata-rata jarak dari objek j dengan seluruh objek yang berada di *cluster* yang sama.

Silhouette score total diperoleh dengan menghitung rata-rata *Silhouette score* dari semua titik data :

$$\text{Silhouette Score} = \frac{1}{n} \sum_{i=1}^n s(i) \quad (2)$$

dengan n adalah jumlah total titik data. Nilai *Silhouette score* yang mendekati 1 menunjukkan bahwa titik-titik data berada di *cluster* yang benar, sementara nilai mendekati -1 menunjukkan bahwa titik mungkin lebih sesuai berada di *cluster* lain.

2) Confusion matrix

Confusion matrix adalah tabel yang menyatakan klasifikasi jumlah data uji yang benar dan jumlah data uji yang salah [14]. *Confusion matrix* digunakan untuk mengevaluasi hasil *clustering* dengan cara membandingkan label asli (jika tersedia) dengan hasil pengelompokan. Matriks ini menunjukkan jumlah titik data yang dikelompokkan dengan benar atau salah dalam setiap *cluster*. Rumus *confusion matrix* untuk menghitung accuracy, precision, dan recall seperti berikut :

$$\text{accuracy} = \frac{TP+TN}{\text{total}} \quad (3)$$

$$\text{precision} = \frac{TP}{TP+FP} \quad (4)$$

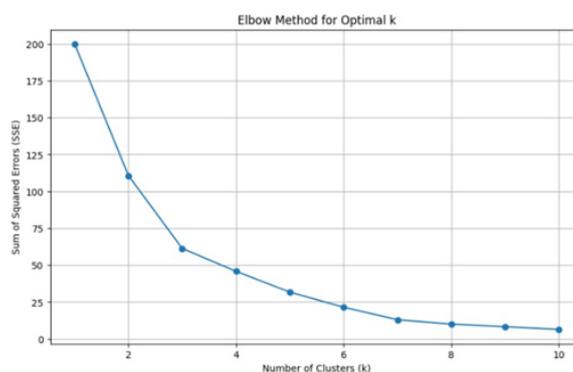
$$\text{recal} = \frac{TP}{TP+FN} \quad (5)$$

2.4. Analisis Cluster

Analisis dilakukan untuk memahami karakteristik setiap *cluster*. Menganalisis rata-rata dan distribusi variabel (*rating*, *votes*, *genre*) untuk memahami perbedaan antar *cluster*. Menggunakan hasil analisis untuk melihat perbedaan mendalam antara klaster satu dengan klaster lainnya.

3. Hasil dan Pembahasan

Setelah menerapkan algoritma K-Means pada dataset 100 judul anime terpopuler, data dikelompokkan berdasarkan *rating* dan jumlah *votes*. Hasil *Elbow method* (Gambar 2) menunjukkan bahwa titik optimal terdapat pada $K=4$, di mana penurunan SSE mulai melambat. Ini menunjukkan bahwa membagi data menjadi 4 *cluster* sudah cukup untuk menghasilkan pengelompokan yang baik tanpa menambah kompleksitas.



Gambar 2. Perhitungan *Elbow method* Menghitung Jumlah Cluster

Setelah proses *clustering*, data *genre* digabungkan dengan hasil *cluster*, menghasilkan informasi tentang *rating*, jumlah *votes*, dan *genre* untuk setiap anime. Rata-rata *rating* dan jumlah *votes* dihitung untuk setiap *cluster*, memberikan gambaran awal mengenai performa tiap *cluster* berdasarkan penilaian penonton dan popularitasnya.

3.1. Hasil *Cluster*

Hasil *clustering* disimpan dalam kolom baru "*cluster*" pada dataset, menunjukkan *cluster* masing-masing anime. Dataset terbagi menjadi empat *cluster* dengan karakteristik berbeda, mencakup perbedaan *rating*, *votes*, dan *genre*, yang mencerminkan segmen penonton dengan preferensi berbeda. Hasil *clustering* diringkas dalam tabel di bawah ini.

Tabel 1. Hasil *Clustering*

<i>Cluster</i>	Rata-rata <i>Rating</i>	Rata-rata <i>votes</i>	<i>Genre</i>	Karakteristik Utama
0	8.45	2.444	<i>Animation, Action, Adventure, Comedy</i>	Anime dengan <i>rating</i> tinggi namun tingkat popularitas sedang; cenderung menarik minat penonton yang lebih niche.
1	7.80	12.701	<i>Action, Adventure, Drama, Fantasy</i>	Anime yang sangat populer dengan kualitas tinggi ; banyak diperbincangkan dan memiliki audiens yang luas.
2	8.36	403.429	<i>Animation, Family, Adventure, Drama</i>	Anime cukup populer dengan <i>rating</i> tinggi; memiliki daya tarik bagi keluarga dan penonton muda, tetapi tidak setinggi <i>rating cluster</i> 1.
3	8.04	30.458	<i>Animation, Drama, Action, Adventure</i>	Anime berkualitas tinggi dengan <i>rating</i> yang baik, namun kurang populer; menunjukkan bahwa meskipun memiliki kualitas, mungkin kurang dikenal oleh masyarakat umum.

Berdasarkan hasil *clustering* dari Table 1, anime dikelompokkan ke dalam empat *cluster* dengan karakteristik yang berbeda. *Cluster* 0 memiliki *rating* tinggi (8.45), namun dengan jumlah *votes* yang relatif lebih rendah (2.444), menunjukkan anime dengan kualitas tinggi tetapi lebih populer di kalangan penonton niche. *Cluster* 1 adalah anime yang sangat populer (*rating* 7.80 dan 12.701 *votes*), mencakup *genre* *action, adventure, drama, dan fantasy*, dan memiliki audiens yang luas. *Cluster* 2 memiliki *rating* 8.36 dan *votes* 403.429, menunjukkan anime yang cukup populer dengan daya tarik kuat untuk keluarga dan penonton muda. Sementara itu, *Cluster* 3 memiliki *rating* 8.04 dan 30.458 *votes*, menunjukkan anime berkualitas tinggi tetapi kurang populer di kalangan masyarakat umum.

Cluster 2 memiliki *rating* tertinggi (8.37) dengan jumlah *votes* yang sangat besar (403,429), menunjukkan bahwa anime dalam *cluster* ini sangat populer dan disukai oleh banyak penonton. *Cluster* 1, meskipun memiliki *rating* yang lebih rendah (7.80), memiliki jumlah *votes* yang sangat tinggi (12,701), menandakan popularitasnya di kalangan penonton. Di sisi lain, *Cluster* 0 memiliki *rating* yang paling tinggi (8.45), namun jumlah *votes* yang relatif lebih rendah dibandingkan *cluster* lainnya (2,444), yang menunjukkan bahwa meskipun kualitas anime dalam *cluster* ini tinggi, jumlah penontonya tidak sebesar di *cluster* yang lebih populer. *Cluster* 3 memiliki *rating* yang sedikit lebih rendah (8.04), tetapi masih menunjukkan *votes* yang cukup besar (30,458), menandakan audiens yang setia.

3.2. Evaluasi K-Means

Evaluasi hasil *clustering* yang dilakukan dengan menggunakan *Confusion matrix* menunjukkan performa klasifikasi dari model *clustering*. Dalam penelitian ini, penulis memetakan hasil *clustering* ke *genre* asli anime untuk mengevaluasi akurasi model dan hasil evaluasi menunjukkan akurasi sebesar 95% dengan nilai *Silhouette score* yang diperoleh adalah 0,285.

Tabel 2. Hasil *Confusion Matrix*

<i>Actual</i>	<i>Predicted</i>	
	<i>True</i>	<i>False</i>
<i>True</i>	63 (TP)	6 (FN)
<i>False</i>	6 (FP)	189 (TN)

Tabel 2 menunjukkan hasil evaluasi model menggunakan *confusion matrix*, yang membandingkan prediksi model dengan nilai sebenarnya :

- True Positive* (TP) : Jumlah kasus di mana model benar memprediksi kelas positif sebanyak 63.
- False Negative* (FN) : Jumlah kasus di mana model salah memprediksi kelas negatif, padahal sebenarnya kelas positif sebanyak 6.
- False Positive* (FP) : Jumlah kasus di mana model salah memprediksi kelas positif, padahal sebenarnya kelas negatif sebanyak 6.
- True Negative* (TN) : Jumlah kasus di mana model benar memprediksi kelas negatif sebanyak 189.

Berdasarkan rumus *confusion matrix* untuk menghitung *accuracy*, *precision*, dan *recall* penulis mendapatkan hasil :

$$accuracy = \frac{63 + 189}{264} = \frac{252}{264} = 0.9545$$

$$precision = \frac{63}{63 + 6} = \frac{63}{69} = 0.913$$

$$recall = \frac{63}{63 + 6} = \frac{63}{69} = 0.913$$

Tabel *accuracy*, *precision* dan *recall* dalam persentase ditampilkan dalam Tabel 3 berikut :

Table 3. Akurasi, Precision dan Recall dari K-Means

K-Means	<i>Accuracy</i>	<i>Precision</i>	<i>Recal</i>
	0,95	0,91	0,91

4. Kesimpulan

Penelitian ini menggunakan algoritma K-Means untuk mengelompokkan anime berdasarkan *rating* dan *votes*. Hasil *Elbow method* menunjukkan bahwa pembagian data menjadi 4 *cluster* (K=4) menghasilkan pengelompokan yang optimal. Setiap *cluster* memiliki karakteristik berbeda, dengan *Cluster 2* sebagai yang paling populer dan disukai penonton, sementara *Cluster 0* memiliki *rating* tinggi namun lebih niche. Evaluasi menggunakan *Confusion matrix* menunjukkan akurasi 95% dan *Silhouette score* 0,285, menandakan model *clustering* berhasil dengan baik. Hasil ini memberikan wawasan berguna untuk platform streaming dalam merancang rekomendasi yang lebih personal bagi penonton anime.

Daftar Pustaka

- [1] G. P. Brahmantha, E. Utami, And A. Yaqin, "Klasifikasi *Genre* Anime Berdasarkan Sinopsis Menggunakan Algoritma K-Nearest Neighbors," *Jurnal Manajemen*

- Informatika & Sistem Informasi (Misi)*, Vol. 7, No. 1, Pp. 15–24, 2024, Doi: 10.36595/Misi.V7i1.2
- [2] Grand View Research, “Anime Market Size, Share & Trends Analysis Report By Type (T.V., Movie, Video, Internet Distribution, Merchandising, Music), By Region, And Segment Forecasts, 2021 – 2028.” Accessed: Oct. 02, 2024. [Online]. Available: <https://www.grandviewresearch.com/industry-analysis/anime-market>
- [3] Z. Muttaqin, C. Savitri, S. Suroso, And K. M. Gu, “Strategi Kolaborasi Uniqlo Dengan Program Anime Terhadap Hasil Penjualan: Studi Pada Uniqlo Dengan Program Anime,” *Akademik: Jurnal Mahasiswa Humanis*, Vol. 4, No. 3, Pp. 1056–1069, 2024, Doi: <https://doi.org/10.37481/jmh.v4i3.995>.
- [4] A. D. Pratama And A. Puspitasari, “Diplomasi Budaya Anime Sebagai Upaya Penguatan Soft Power Jepang Periode 2014-2018,” *Balcony*, Vol. 4, No. 1, Pp. 11–23, 2020.
- [5] M. Sakuma, “The Future Of Fansubs: Facing The Advent Of Legal Anime On Streaming Platforms,” *Skase Journal Of Translation And Interpretation*, Vol. 16, No. 1, Pp. 40–56, 2023.
- [6] J. Aurima, S. Susaldi, N. Agustina, A. Masturoh, R. Rahmawati, And M. Tresiana Monika Madhe, “Faktor-Faktor Yang Berhubungan Dengan Kejadian Stunting Pada Balita Di Indonesia,” *Open Access Jakarta Journal Of Health Sciences*, Vol. 1, No. 2, Pp. 43–48, Nov. 2021, Doi: 10.53801/Oajjhs.V1i3.23.
- [7] Y. Deldjoo, T. Di Noia, And F. A. Merra, “A Survey On Adversarial Recommender Systems: From Attack/Defense Strategies To Generative Adversarial Networks,” *Acm Computing Surveys (Csur)*, Vol. 54, No. 2, Pp. 1–38, 2021.
- [8] A. Saxena *Et Al.*, “A Review Of Clustering Techniques And Developments,” *Neurocomputing*, Vol. 267, Pp. 664–681, 2017, Doi: <https://doi.org/10.1016/j.neucom.2017.06.053>.
- [9] A. Yudhistira And R. Andika, “Pengelompokan Data Nilai Siswa Menggunakan Metode K-Means Clustering,” *Journal Of Artificial Intelligence And Technology Information (Jaiti)*, Vol. 1, No. 1, Pp. 20–28, Feb. 2023, Doi: 10.58602/Jaiti.V1i1.22.
- [10] I. Jayaperwira, A. T. Wibowo, And D. Nurjanah, “Anime Rekomendasi Menggunakan Collaborative Filtering,” *Eproceedings Of Engineering*, Vol. 10, No. 3, 2023.
- [11] D. Arthur And S. Vassilvitskii, “K-Means++: The Advantages Of Careful Seeding.”
- [12] R. Rahmattullah, Indwiarti, And A. A. Rohmawati, “Clustering Harga Rumah: Perbandingan Model K-Means Dan Gaussian Mixture Model,” *E-Proceeding Of Engineering*, Vol. 10, No. 3, Pp. 3441–3449, Jun. 2023.
- [13] L. Ardiansyah And S. A. Awalludin, “Implementation Of The K-Mean Algorithm To Determine The Level Of Student Satisfaction With The Online Learning Uhamka System (Olu),” *Jurnal Pembelajaran Dan Matematika Sigma (Jpms)*, Vol. 9, No. 1, Pp. 162–171, May 2023, Doi: 10.36987/Jpms.V9i1.4121.
- [14] D. Normawati And S. A. Prayogi, “Implementasi Naïve Bayes Classifier Dan Confusion Matrix Pada Analisis Sentimen Berbasis Teks Pada Twitter,” 2021.