Penerapan Algoritma K-Means Untuk Mengelompokkan Film Berdasarkan Rating Dan Popularitas Di Aplikasi Netflix

Liana Putri[1*], Putri Puspita Sari[2], Shilvani Laratika Hawia[3], Pujianto[4]

[1]Informatika/Teknik Dan Komputer
Universitas Baturaja
[1*]lianaputri201999@gmail.com

[2]Informatika/Teknik Dan Komputer
Universitas Baturaja
[2]putripuspitasari0226@gmail.com

[3]Informatika/Teknik Dan Komputer
Universitas Baturaja
[3]shilvani036@gmail.com

[4]Informatika/Teknik Dan Komputer
Universitas Baturaja
[4]pujianto.mail@gmail.com

Abstrak— ***Perkembangan platform* streaming *digital, seperti Netflix, menghasilkan data film dalam jumlah besar yang dapat dimanfaatkan untuk memperoleh informasi melalui teknik* data mining*. Salah satu teknik yang banyak digunakan adalah* clustering*, yaitu proses mengelompokkan data berdasarkan tingkat kemiripan karakteristik tanpa memerlukan label kelas. Penelitian ini bertujuan menerapkan algoritma K-Means untuk mengelompokkan data film berdasarkan atribut* Popularity *dan* Vote Average *menggunakan Orange Data Mining. Dataset yang digunakan berasal dari The Movie Database (TMDB) dengan jumlah awal 9.827 data film. Setelah melalui tahapan* preprocessing*, diperoleh 1.991 data yang layak dianalisis. Proses* clustering *dilakukan dengan jumlah* cluster *sebanyak empat menggunakan algoritma K-Means. Hasil penelitian menunjukkan bahwa algoritma K-Means mampu mengelompokkan data film ke dalam empat* cluster *yang memiliki karakteristik berbeda berdasarkan tingkat popularitas dan penilaian pengguna. Visualisasi menggunakan* Data Table *dan* Scatter Plot *mempermudah proses interpretasi hasil pengelompokan sehingga dapat memberikan informasi mengenai karakteristik setiap kelompok film. Hasil penelitian ini diharapkan dapat menjadi dasar dalam pengembangan sistem rekomendasi film serta mendukung proses analisis data pada platform* streaming*.***
Kata kunci— Data mining*, K-Means, clustering, Orange Data Mining, TMDB.*

Abstract— ***The rapid growth of digital streaming platforms, such as Netflix, has generated large volumes of movie data that can be analyzed using data mining techniques. One of the most widely used techniques is clustering, which groups data based on the similarity of their characteristics without requiring predefined class labels. This study aims to implement the K-Means algorithm to cluster movie data based on the Popularity and Vote Average attributes using Orange Data Mining. The dataset was obtained from The Movie Database (TMDB), consisting of 9,827 movie records. After the preprocessing stage, 1,991 valid records were selected for analysis. The clustering process was performed using the K-Means algorithm with four clusters. The results show that the K-Means algorithm successfully grouped the movie data into four clusters with different characteristics based on popularity and user ratings. The visualization generated through the Data Table and Scatter Plot widgets facilitates the interpretation of clustering results and provides insights into the characteristics of each movie group. The findings of this study are expected to support the development of movie recommendation systems and assist data analysis on digital streaming platforms.***
Keywords— Data mining, K-Means, clustering, Orange Data Mining, TMDB.

I. PENDAHULUAN

Industri perfilman mengalami perkembangan yang pesat seiring dengan meningkatnya penggunaan platform *streaming* digital, seperti Netflix, yang menyediakan berbagai pilihan film dengan karakteristik yang beragam. Data film yang tersedia tidak hanya mencakup judul dan genre, tetapi juga berbagai atribut seperti tingkat popularitas (*popularity*), penilaian pengguna (*vote average*), jumlah suara (*vote count*), serta informasi pendukung lainnya. Banyaknya data tersebut memberikan peluang untuk memperoleh informasi yang bermanfaat melalui penerapan teknik analisis data sehingga dapat membantu pengguna maupun penyedia layanan dalam memahami karakteristik film secara lebih sistematis [2], [10].

Salah satu pendekatan yang banyak digunakan untuk menganalisis data berukuran besar adalah *data mining*. *Data mining* merupakan proses menggali informasi atau pola yang tersembunyi dari sekumpulan data sehingga menghasilkan

pengetahuan yang dapat dimanfaatkan dalam pengambilan keputusan [2]. Salah satu teknik dalam *data mining* adalah *clustering*, yaitu proses mengelompokkan objek berdasarkan tingkat kemiripan karakteristik tanpa memerlukan label kelas sebelumnya. Teknik ini termasuk ke dalam metode *unsupervised learning* dan banyak diterapkan dalam berbagai bidang karena mampu menemukan struktur alami dari suatu kumpulan data [1], [7].

Algoritma K-Means merupakan salah satu algoritma *clustering* yang paling banyak digunakan karena memiliki proses komputasi yang sederhana, efisien, serta mampu mengelompokkan data numerik dalam jumlah besar dengan baik [3], [12]. Proses pengelompokan dilakukan dengan membagi data ke dalam beberapa *cluster* berdasarkan kedekatan setiap objek terhadap pusat *cluster* (*centroid*) menggunakan ukuran jarak tertentu. Karakteristik tersebut menjadikan algoritma K-Means sebagai salah satu metode yang banyak dimanfaatkan dalam analisis data pada berbagai bidang penelitian [4].

Beberapa penelitian terdahulu telah menerapkan algoritma K-Means pada pengelompokan data film. Muhajir dan Sari [5] menerapkan algoritma K-Means yang dipadukan dengan DBSCAN untuk mengelompokkan data film. Hasil penelitian menunjukkan bahwa metode tersebut mampu menghasilkan pengelompokan film berdasarkan karakteristik data sehingga dapat membantu proses analisis dan rekomendasi film.

Penelitian lain yang dilakukan oleh Siregar [9] menunjukkan bahwa algoritma K-Means mampu mengelompokkan data penayangan film berdasarkan karakteristik yang dimiliki sehingga mempermudah proses analisis dan pengambilan keputusan.

Meskipun demikian, penelitian-penelitian sebelumnya masih berfokus pada atribut atau pendekatan tertentu sehingga interpretasi karakteristik setiap *cluster* belum dijelaskan secara rinci. Selain itu, pemanfaatan perangkat lunak Orange Data Mining sebagai media visualisasi hasil *clustering* pada data film juga masih terbatas. Oleh karena itu, penelitian ini dilakukan untuk menerapkan algoritma K-Means dalam mengelompokkan film berdasarkan atribut *Popularity* dan *Vote Average* menggunakan Orange Data Mining sehingga diperoleh gambaran karakteristik setiap *cluster* secara lebih jelas.

Dataset yang digunakan dalam penelitian ini berasal dari The Movie Database (TMDB) yang berisi 9.827 data film dengan berbagai atribut [8]. Setelah melalui tahapan *preprocessing*, diperoleh 1.991 data yang layak dianalisis. Selanjutnya dilakukan proses *clustering* menggunakan algoritma K-Means untuk mengelompokkan film berdasarkan tingkat popularitas dan penilaian pengguna. Hasil penelitian diharapkan dapat memberikan informasi mengenai karakteristik setiap kelompok film serta menjadi dasar dalam pengembangan sistem rekomendasi berbasis konten pada platform *streaming*.

II. METODOLOGI PENELITIAN

A. Dataset

Penelitian ini menggunakan dataset film yang diperoleh dari The Movie Database (TMDB)**.** Dataset terdiri atas 9.827 data film dengan beberapa atribut, seperti *Title, Genre, Popularity, Vote Average*, dan *Vote Count*. Data tersebut kemudian melalui tahap *preprocessing* sehingga diperoleh 1.991 data yang memenuhi kriteria untuk dianalisis menggunakan algoritma K-Means. Dataset TMDB dipilih karena menyediakan informasi film yang lengkap dan sering digunakan sebagai sumber data dalam penelitian terkait analisis maupun sistem rekomendasi film [8].

B. Alur Kerja Penelitian

Penelitian ini dilakukan menggunakan perangkat lunak Orange Data Mining sebagai media analisis dan visualisasi data. Orange menyediakan berbagai *widget* yang memudahkan proses *preprocessing*, penerapan algoritma *clustering*, hingga visualisasi hasil analisis [11]. Alur penelitian dimulai dari proses pemuatan dataset, dilanjutkan dengan *preprocessing*, pemilihan atribut, proses *clustering* menggunakan algoritma K-Means, kemudian

diakhiri dengan visualisasi hasil pengelompokan menggunakan *Data Table* dan *Scatter Plot*.

C. Tahapan Preprocessing Data

Sebelum proses *clustering* dilakukan, dataset terlebih dahulu melalui beberapa tahapan *preprocessing* untuk meningkatkan kualitas data yang akan dianalisis. Tahapan tersebut dijelaskan sebagai berikut.

1. Pemuatan Data (File)

Dataset film dalam format CSV dimuat ke dalam Orange Data Mining menggunakan *widget* File. Tahap ini bertujuan untuk membaca seluruh data beserta atribut yang tersedia sehingga dapat diproses pada tahapan berikutnya.

2. Pengaturan Domain (Edit Domain)

Tahap ini dilakukan untuk menyesuaikan tipe data setiap atribut sesuai dengan karakteristiknya. Atribut numerik ditetapkan sebagai data numerik, sedangkan atribut teks diklasifikasikan sebagai data kategorikal atau *string*. Pengaturan domain diperlukan agar proses analisis dapat berjalan dengan baik.

3. Penanganan Nilai Hilang (Impute)

Tahap *Impute* digunakan untuk menangani data yang memiliki nilai kosong (*missing value*). Proses ini bertujuan agar data yang digunakan dalam analisis tetap lengkap dan tidak memengaruhi hasil pengelompokan.

4. Seleksi Atribut (Select Columns)

Pada tahap ini dilakukan pemilihan atribut yang digunakan dalam proses *clustering*. Dari beberapa atribut yang tersedia, penelitian ini menggunakan atribut Popularity dan Vote Average karena keduanya dianggap mampu merepresentasikan tingkat popularitas dan penilaian pengguna terhadap film.

5. Penerapan Algoritma K-Means

Dataset yang telah melalui tahap *preprocessing* kemudian diproses menggunakan algoritma K-Means dengan jumlah *cluster* sebanyak empat (k = 4)**.** Algoritma ini mengelompokkan data berdasarkan kedekatan setiap objek terhadap pusat *cluster* (*centroid*) menggunakan jarak Euclidean sehingga objek yang memiliki karakteristik serupa berada dalam kelompok yang sama [3], [12].

Secara matematis, perhitungan jarak Euclidean dirumuskan sebagai berikut.

$$d(x,y) = \sqrt{\left\{ \sum_{\{i=1\}}^{\{n\}} (x_i - y_i)^2 \right\}}$$

dengan:
(d(x,y)) = jarak antara dua objek,
(x_i) = nilai atribut objek pertama,
(y_i) = nilai atribut objek kedua,
(n) = jumlah atribut yang digunakan.

D. Paramater Clustering

Parameter yang digunakan pada proses *clustering* ditunjukkan pada Tabel 1.

Tabel 1. Parameter Algoritma K-Means

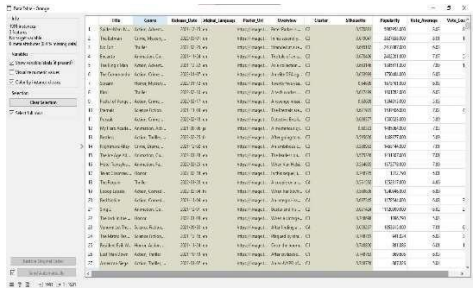
Paramater	**Nilai**
Algoritma	K-Means
Jumlah Cluster (k)	4
Metode Jarak	Euclidean Distance
Atribut	Popularity, Vote Average
Perangkat Lunak	Orange Data Mining

Parameter tersebut digunakan sebagai dasar dalam proses pengelompokan data film sehingga diperoleh kelompok film dengan karakteristik yang berbeda berdasarkan tingkat popularitas dan penilaian pengguna.

III. HASIL DAN PEMBAHASAN

A. Hasil Clustering

Pada tahap ini dilakukan proses *clustering* menggunakan algoritma K-Means terhadap 1.991 data film yang telah melalui tahap *preprocessing*. Proses pengelompokan dilakukan menggunakan dua atribut utama, yaitu Popularity dan Vote Average, dengan jumlah *cluster* sebanyak empat. Hasil pengelompokan ditampilkan melalui *widget* Data Table pada Orange Data Mining sehingga setiap data film dapat diketahui keanggotaannya pada masing-masing *cluster*.

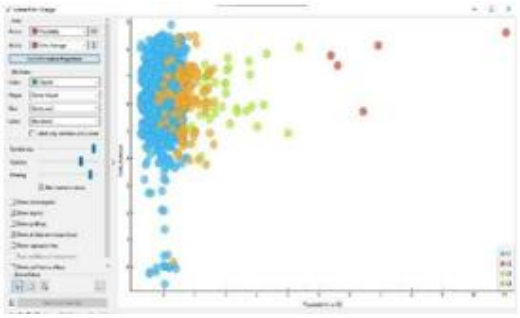


Gambar 1. Tampilan Data Table Hasil Clustering pada Orange Data Mining (1.991 Instance)

Berdasarkan hasil tersebut, setiap film berhasil dikelompokkan ke dalam salah satu dari empat *cluster*. Pengelompokan ini menunjukkan bahwa algoritma K-Means mampu memisahkan data berdasarkan tingkat kemiripan nilai *Popularity* dan *Vote Average* sehingga karakteristik setiap kelompok dapat dianalisis lebih lanjut.

B. Visualisasi Scatter Plot

Visualisasi hasil *clustering* menggunakan Scatter Plot bertujuan untuk memperlihatkan penyebaran data berdasarkan atribut *Popularity* dan *Vote Average*. Setiap warna menunjukkan kelompok (*cluster*) yang berbeda sehingga pola penyebaran data dapat diamati secara visual.



Gambar 2. Scatter Plot Hasil K-Means Clustering (Popularity vs Vote Average)

Berdasarkan Gambar tersebut, terlihat bahwa sebagian besar data berada pada rentang nilai popularitas rendah hingga sedang. Beberapa data dengan tingkat popularitas tinggi membentuk kelompok yang terpisah sehingga menunjukkan adanya perbedaan karakteristik dibandingkan kelompok lainnya. Visualisasi ini memperlihatkan bahwa algoritma K-Means mampu membentuk kelompok dengan distribusi yang relatif jelas.

C. Analisis Karakteristik Setiap Cluster

Hasil *clustering* menghasilkan empat kelompok film dengan karakteristik yang berbeda.

1. Cluster 0

   Cluster ini didominasi oleh film dengan nilai *Popularity* rendah dan *Vote Average* sedang. Karakteristik tersebut menunjukkan bahwa film dalam kelompok ini memiliki tingkat popularitas yang relatif rendah meskipun masih memperoleh penilaian yang cukup baik dari pengguna.

2. Cluster 1

   Cluster ini terdiri atas film dengan nilai *Popularity* dan *Vote Average* yang tinggi. Film dalam kelompok ini dapat dikategorikan sebagai film populer dengan kualitas yang baik berdasarkan penilaian pengguna.

3. Cluster 2

   Cluster ini berisi film yang memiliki nilai *Popularity* sangat tinggi. Kelompok ini menunjukkan karakteristik film yang memperoleh perhatian besar dari pengguna sehingga dapat dikategorikan sebagai film *blockbuster*.

4. Cluster 3

   Cluster ini memiliki nilai *Popularity* menengah dengan *Vote Average* yang relatif stabil. Film pada kelompok ini termasuk film yang cukup dikenal oleh pengguna, tetapi belum mencapai tingkat popularitas seperti pada Cluster 2.

D. Pembahasan

Hasil penelitian menunjukkan bahwa algoritma K-Means mampu mengelompokkan data film berdasarkan atribut *Popularity* dan *Vote Average* dengan baik. Penggunaan dua atribut tersebut menghasilkan empat kelompok film yang memiliki karakteristik berbeda sehingga mempermudah proses identifikasi pola data.

Hasil penelitian ini sejalan dengan penelitian Muhajir dan Sari [5] yang menunjukkan bahwa algoritma K-Means mampu mengelompokkan data film berdasarkan karakteristik tertentu. Selain itu, penelitian Siregar [9] juga menunjukkan bahwa algoritma K-Means mampu mengelompokkan data penayangan film

berdasarkan karakteristik yang dimiliki sehingga mempermudah proses analisis dan pengambilan keputusan.

Dengan memanfaatkan Orange Data Mining [11], proses *clustering* menjadi lebih mudah dilakukan karena seluruh tahapan mulai dari *preprocessing* hingga visualisasi hasil dapat dilakukan secara terintegrasi. Oleh karena itu, hasil pengelompokan ini berpotensi dimanfaatkan sebagai dasar dalam pengembangan sistem rekomendasi film maupun analisis karakteristik film pada platform *streaming*.

IV. KESIMPULAN

Berdasarkan hasil penelitian, dapat disimpulkan bahwa algoritma K-Means berhasil diterapkan untuk mengelompokkan data film berdasarkan atribut *Popularity* dan *Vote Average* menggunakan Orange Data Mining. Melalui tahapan *preprocessing*, sebanyak 9.827 data film berhasil diseleksi menjadi 1.991 data yang layak untuk dianalisis. Proses *clustering* dengan jumlah empat *cluster* menghasilkan kelompok film yang memiliki karakteristik berbeda berdasarkan tingkat popularitas dan penilaian pengguna. Hasil visualisasi menggunakan *Data Table* dan *Scatter Plot* menunjukkan bahwa proses pengelompokan mampu memisahkan data ke dalam kelompok yang relatif jelas sehingga memudahkan interpretasi karakteristik setiap *cluster*. Dengan demikian, penerapan algoritma K-Means dapat menjadi salah satu pendekatan yang efektif dalam menganalisis data film serta memberikan informasi yang bermanfaat sebagai dasar pengembangan sistem rekomendasi maupun analisis karakteristik film pada platform *streaming*. Penelitian selanjutnya disarankan menggunakan atribut yang lebih beragam serta membandingkan algoritma K-Means dengan metode *clustering* lainnya untuk memperoleh hasil pengelompokan yang lebih optimal.

REFERENSI

[1] C. C. Aggarwal and C. K. Reddy, *Data Clustering: Algorithms and Applications*. Boca Raton, FL, USA: CRC Press, 2013.

[2] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2012.

[3] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, Jun. 2010.

[4] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: John Wiley & Sons, 2005.

[5] M. Muhajir and A. A. P. Sari, "Implementasi Metode Improved K-Means dengan Algoritma DBSCAN untuk Pengelompokan Film," *Unisda Journal of Mathematics and Computer Science (UJMC)*, vol. 6, no. 1, pp. 30–38, 2020.

[6] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987.

[7] P. N. Tan, M. Steinbach, and V. Kumar, *Introduction to Data Mining*. Boston, MA, USA: Pearson Addison-Wesley, 2006.

[8] The Movie Database (TMDB), "TMDB API Documentation." [Online]. Available: https://developer.themoviedb.org/. [Accessed: Apr. 2026].

[9] A. I. N. Siregar, "Analisa Clustering untuk Mengelompokkan Data Penayangan Film Bioskop Menggunakan Algoritma K-Means," *Internal (Jurnal Ilmu Komputer dan Teknologi Informasi)*, vol. 2, no. 2, pp. 80–87, 2024.

[10] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical Machine Learning Tools and Techniques*, 4th ed. Cambridge, MA, USA: Morgan Kaufmann, 2016.

[11] J. Demšar *et al*., "Orange: Data Mining Toolbox in Python," *Journal of Machine Learning Research*, vol. 14, pp. 2349–2353, 2013.

[12] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, vol. 28, no. 1, pp. 100–108, 1979.