JDMSI, Vol. 7, No. 2, 2026, 12-21, ISSN: 2745-8485

PENERAPAN METODE K-MEANS CLUSTERING UNTUK PENGELOMPOKAN FILM BERDASARKAN RATING DAN DURASI

Pandega Lirma Arga[1], M aflah Robbani[2], Rizki Agustina Saputri[3], Anggun Nida Suftiani[4], M Diva Aji Pratama[5]

[12345]*Program Studi Sistem Informasi Universitas Muhammadiyah Brebes*
[12345]*Alamat: Jl. Pangeran Diponegoro, Samping Fly Over Kretek, Paguyangan (kec. Paguyangan, Kabupaten Brebes, Jawa Tengah 52276)*
Email: divapengin@gmail.com , riskiagustina0808@gmail.com, anggunnidasuftiani@gmail.com, divaji@gmail.com,

ABSTRAK

Banyaknya jumlah film yang tersedia di berbagai platform membuat proses pengelompokan film berdasarkan karakteristiknya menjadi penting, baik untuk kebutuhan rekomendasi maupun analisis tren perfilman. Penelitian ini bertujuan untuk menerapkan algoritma K-Means Clustering dalam mengelompokkan film berdasarkan dua atribut, yaitu rating IMDb dan durasi film. K-Means merupakan salah satu algoritma clustering non-hierarki yang bekerja dengan mengelompokkan objek ke dalam k kelompok berdasarkan kedekatan jarak Euclidean terhadap titik pusat (centroid) tertentu. Penelitian ini menggunakan enam sampel film dengan rating tinggi (8,5 ke atas) yang dikelompokkan ke dalam tiga cluster (k=3). Data dinormalisasi menggunakan metode min-max sebelum diproses agar kedua atribut memiliki skala yang setara. Hasil penelitian menunjukkan bahwa algoritma K-Means berhasil mengelompokkan film ke dalam tiga cluster yang berbeda karakteristik durasinya, yaitu cluster film berdurasi pendek-sedang, cluster film dengan rating sangat tinggi berdurasi sedang, dan cluster film epik berdurasi sangat panjang. Proses iterasi konvergen setelah dua kali iterasi, menunjukkan bahwa algoritma K-Means efektif dan efisien digunakan untuk pengelompokan data film berskala kecil.

Kata Kunci: K-Means, Clustering, Data Mining, Film, Rating, Durasi

ABSTRACT

The large number of movies available across various platforms makes it important to group movies based on their characteristics, both for recommendation purposes and film trend analysis. This study aims to apply the K-Means Clustering algorithm to group movies based on two attributes, namely IMDb rating and movie duration. K-Means is a non-hierarchical clustering algorithm that works by grouping objects into k clusters based on the Euclidean distance to a certain centroid. This study uses six sample movies with high ratings (8.5 and above) grouped into three clusters (k=3). The data is normalized using the min-max method before processing so that both attributes have an equal scale. The results show that the K-Means algorithm successfully grouped the movies into three clusters with different duration characteristics, namely a cluster of short-to-medium duration movies, a cluster of very high-rated movies with medium duration, and a cluster of epic movies with very long duration. The iteration process converged after two iterations, indicating that the K-Means algorithm is effective and efficient for clustering small-scale movie data.

Keywords: *K-Means, Clustering, Data Mining, Movies, Rating, Duration*

1. PENDAHULUAN

Industri perfilman terus berkembang pesat dengan jumlah judul film yang semakin banyak diproduksi setiap tahunnya. Platform seperti IMDb mencatat ratusan ribu judul film dengan berbagai variasi rating, genre, dan durasi. Banyaknya jumlah film ini menyulitkan proses analisis maupun pengelompokan film secara manual berdasarkan karakteristiknya, terutama ketika pihak yang berkepentingan (misalnya platform streaming atau peneliti perfilman) ingin memahami pola pengelompokan film berdasarkan atribut kuantitatif seperti rating dan durasi.

Data mining menyediakan berbagai teknik untuk menggali pola tersembunyi dalam data berskala besar, salah satunya adalah teknik clustering (pengelompokan). Clustering bertujuan mengelompokkan sekumpulan objek ke dalam beberapa kelompok (cluster) sedemikian rupa sehingga objek dalam satu cluster memiliki kemiripan karakteristik yang tinggi, sementara objek pada cluster berbeda memiliki karakteristik yang berbeda pula.

Salah satu algoritma clustering yang paling populer digunakan karena kesederhanaan dan efisiensinya adalah K-Means. Algoritma ini telah banyak diterapkan pada berbagai domain permasalahan, seperti pengelompokan data obat-obatan (Gustientiedina et al., 2019) dan penentuan predikat kelulusan mahasiswa (Sari et al., 2018). Keberhasilan penerapan K-Means pada berbagai kasus tersebut mendorong penelitian ini untuk menerapkan algoritma yang sama pada kasus pengelompokan film berdasarkan rating dan durasi.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk menerapkan algoritma K-Means Clustering dalam mengelompokkan film berdasarkan atribut rating dan durasi, sehingga dapat diketahui pola pengelompokan film yang terbentuk secara objektif berdasarkan perhitungan matematis.

1.1 Rumusan Masalah

Berdasarkan latar belakang di atas, rumusan masalah dalam penelitian ini adalah bagaimana menerapkan algoritma K-Means Clustering untuk mengelompokkan film berdasarkan rating dan durasi ke dalam beberapa cluster yang bermakna.

1.2 Tujuan Penelitian

Penelitian ini bertujuan untuk menerapkan algoritma K-Means Clustering pada data film berdasarkan rating dan durasi, serta menganalisis karakteristik masing-masing cluster yang terbentuk dari hasil perhitungan.

1.3 Manfaat Penelitian

Penelitian ini diharapkan dapat menjadi referensi penerapan algoritma K-Means pada data non-transaksional seperti data film, serta membantu memahami pola pengelompokan film berdasarkan atribut rating dan durasi yang dapat dimanfaatkan misalnya untuk kebutuhan sistem rekomendasi film.

2. TINJAUAN PUSTAKA

2.1 Data Mining dan Clustering

Data mining adalah proses penggalian pola, informasi, dan pengetahuan yang bermanfaat dari sekumpulan data berukuran besar. Salah satu teknik dalam data mining adalah clustering, yaitu proses pengelompokan sekumpulan objek data ke dalam beberapa kelompok (cluster) berdasarkan tingkat kemiripan karakteristiknya, tanpa memerlukan label kelas sebelumnya (unsupervised learning).

2.2 Algoritma K-Means

K-Means merupakan algoritma clustering non-hierarki yang mengelompokkan data ke dalam k cluster, di mana setiap data akan menjadi anggota cluster dengan titik pusat (centroid) terdekat. Kedekatan antara data dan centroid umumnya dihitung menggunakan rumus jarak Euclidean. Algoritma K-Means bekerja secara iteratif, yaitu memperbarui posisi centroid pada setiap iterasi berdasarkan rata-rata anggota cluster, hingga posisi centroid tidak lagi berubah (konvergen).

Algoritma K-Means telah banyak diterapkan dalam berbagai penelitian karena konsepnya yang sederhana namun efektif, seperti pada pengelompokan data obat-obatan (Gustientiedina et al., 2019) dan penentuan jumlah cluster optimal menggunakan metode Elbow (Maori & Evanita, 2023).

2.3 Penelitian Terkait

Penelitian Sari et al. (2018) menerapkan K-Means untuk menentukan predikat kelulusan mahasiswa berdasarkan nilai akademik, menunjukkan bahwa algoritma ini juga dapat diterapkan pada data non-transaksional berskala kecil hingga besar. Penelitian Maori & Evanita (2023) juga menekankan pentingnya penentuan jumlah cluster (k) yang tepat, misalnya menggunakan metode Elbow, agar hasil pengelompokan lebih optimal.

3. METODE PENELITIAN

3.1 Objek dan Sumber Data Penelitian

Objek penelitian ini adalah enam film dengan rating IMDb tinggi (8,5 ke atas), yaitu 12 Angry Men, Whiplash, The Shawshank Redemption, Fight Club, The Godfather Part II, dan The Lord of the Rings: The Return of the King. Jumlah sampel dibatasi pada enam film dengan tujuan agar proses perhitungan manual algoritma K-Means dapat ditunjukkan secara rinci dan mudah ditelusuri setiap langkahnya. Data rating dan durasi masing-masing film diperoleh dari situs IMDb (Internet Movie Database) sebagai sumber data utama, karena IMDb merupakan basis data film yang datanya diperbarui dan diverifikasi secara berkelanjutan, berbeda dengan platform ensiklopedia terbuka seperti Wikipedia yang berpotensi memiliki data tidak konsisten karena dapat disunting oleh siapa saja.

3.2 Atribut yang Digunakan

Penelitian ini menggunakan dua atribut numerik sebagai dasar pengelompokan, yaitu rating IMDb (skala 1-10) dan durasi film dalam satuan menit. Kedua atribut ini dipilih karena bersifat kuantitatif dan tersedia secara konsisten untuk seluruh film di IMDb, sehingga memudahkan proses perhitungan jarak Euclidean pada algoritma K-Means.

3.3 Data Penelitian

Data mentah keenam film yang digunakan dalam penelitian ini ditampilkan pada Tabel 1.

Tabel 1. Data Rating dan Durasi Film

Kode	Judul Film	Rating IMDb	Durasi (menit)
F1	12 Angry Men	9,0	96
F2	Whiplash	8,5	106
F3	The Shawshank Redemption	9,3	142
F4	Fight Club	8,8	139
F5	The Godfather Part II	9,0	202
F6	LOTR: The Return of the King	9,0	201

3.4 Normalisasi Data

Karena rentang nilai atribut rating (8,5-9,3) jauh lebih kecil dibandingkan atribut durasi (96-202 menit), maka sebelum dilakukan perhitungan jarak Euclidean, kedua atribut terlebih

dahulu dinormalisasi menggunakan metode min-max agar berada pada rentang [0,1] yang setara. Rumus normalisasi min-max yang digunakan adalah sebagai berikut:

$$x' = (x - min(x)) / (max(x) - min(x))$$

dengan x' adalah nilai hasil normalisasi, x adalah nilai asli, min(x) adalah nilai minimum, dan max(x) adalah nilai maksimum pada atribut tersebut.

3.5 Langkah-Langkah Algoritma K-Means

Langkah-langkah penerapan algoritma K-Means pada penelitian ini adalah sebagai berikut:

1. Menentukan jumlah cluster (k). Pada penelitian ini ditentukan k=3 berdasarkan pertimbangan bahwa data memiliki tiga kelompok durasi yang tampak berbeda secara visual (pendek, sedang, dan panjang).

2. Menentukan titik pusat (centroid) awal secara acak/manual sejumlah k dari data yang ada.

3. Menghitung jarak setiap data ke seluruh centroid menggunakan rumus jarak Euclidean:

$$d(x,c) = \sqrt{[\ (x1-c1)^2 + (x2-c2)^2\ ]}$$

4. Mengelompokkan setiap data ke cluster dengan centroid terdekat (jarak terkecil).

5. Menghitung ulang posisi centroid baru berdasarkan rata-rata seluruh anggota masing-masing cluster.

6. Mengulangi langkah 3-5 hingga posisi centroid tidak lagi berubah (konvergen), yang menandakan proses clustering telah selesai.

4. HASIL DAN PEMBAHASAN

4.1 Hasil Normalisasi Data

Berdasarkan data pada Tabel 1, nilai minimum dan maksimum masing-masing atribut adalah rating minimum 8,5 (Whiplash) dan maksimum 9,3 (The Shawshank Redemption), sedangkan durasi minimum 96 menit (12 Angry Men) dan maksimum 202 menit (The Godfather Part II). Hasil normalisasi min-max kedua atribut ditampilkan pada Tabel 2.

Tabel 2. Hasil Normalisasi Data

Kode	Judul Film	Rating Ternormalisasi	Durasi Ternormalisasi
F1	12 Angry Men	0,6250	0,0000
F2	Whiplash	0,0000	0,0943
F3	The Shawshank Redemption	1,0000	0,4340
F4	Fight Club	0,3750	0,4057
F5	The Godfather Part II	0,6250	1,0000
F6	LOTR: The Return of the King	0,6250	0,9906

4.2 Penentuan Centroid Awal

Centroid awal ditentukan dengan memilih tiga data yang memiliki karakteristik durasi berbeda secara signifikan, yaitu data F1 (durasi terpendek), F3 (durasi sedang), dan F5 (durasi terpanjang), sebagaimana ditampilkan pada Tabel 3.

Tabel 3. Centroid Awal

Centroid	Sumber Data	Rating Ternormalisasi	Durasi Ternormalisasi
C1	F1 - 12 Angry Men	0,6250	0,0000
C2	F3 - The Shawshank Redemption	1,0000	0,4340
C3	F5 - The Godfather Part II	0,6250	1,0000

4.3 Iterasi 1: Perhitungan Jarak Euclidean

Contoh perhitungan jarak data F1 (12 Angry Men) ke centroid C1:

$$d(F1,C1) = \sqrt{[(0,6250-0,6250)^2 + (0,0000-0,0000)^2]} = \sqrt{0} = 0,0000$$

Perhitungan yang sama dilakukan untuk seluruh data terhadap ketiga centroid. Hasil lengkap jarak Euclidean iterasi 1 dan hasil pengelompokan ditampilkan pada Tabel 4.

Tabel 4. Hasil Perhitungan Jarak Euclidean - Iterasi 1

Kode	Jarak ke C1	Jarak ke C2	Jarak ke C3	Cluster Terpilih
F1	0,0000	0,5735	1,0000	Cluster 1
F2	0,6321	1,0561	1,1004	Cluster 1
F3	0,5735	0,0000	0,6790	Cluster 2
F4	0,4765	0,6256	0,6448	Cluster 1
F5	1,0000	0,6790	0,0000	Cluster 3
F6	0,9906	0,6711	0,0094	Cluster 3

Berdasarkan Tabel 4, terbentuk pengelompokan sementara: Cluster 1 beranggotakan F1, F2, F4; Cluster 2 beranggotakan F3; dan Cluster 3 beranggotakan F5, F6. Selanjutnya, posisi centroid diperbarui berdasarkan rata-rata anggota masing-masing cluster, sebagaimana ditampilkan pada Tabel 5.

Tabel 5. Centroid Baru Setelah Iterasi 1

Centroid	Anggota Cluster	Rating Ternormalisasi	Durasi Ternormalisasi
C1	F1, F2, F4	0,3333	0,1667
C2	F3	1,0000	0,4340
C3	F5, F6	0,6250	0,9953

4.4 Iterasi 2: Verifikasi Konvergensi

Menggunakan centroid baru dari Tabel 5, dilakukan kembali perhitungan jarak Euclidean untuk seluruh data. Hasil perhitungan iterasi 2 ditampilkan pada Tabel 6.

Tabel 6. Hasil Perhitungan Jarak Euclidean - Iterasi 2

Kode	Jarak ke C1	Jarak ke C2	Jarak ke C3	Cluster Terpilih
F1	0,3359	0,5735	0,9953	Cluster 1
F2	0,3411	1,0561	1,0965	Cluster 1
F3	0,7183	0,0000	0,6751	Cluster 2
F4	0,2426	0,6256	0,6404	Cluster 1
F5	0,8829	0,6790	0,0047	Cluster 3
F6	0,8740	0,6711	0,0047	Cluster 3

Hasil pengelompokan pada iterasi 2 (Tabel 6) menunjukkan hasil yang sama persis dengan iterasi 1, yaitu Cluster 1 = {F1, F2, F4}, Cluster 2 = {F3}, dan Cluster 3 = {F5, F6}. Karena tidak terjadi perubahan anggota cluster maupun posisi centroid dibandingkan iterasi sebelumnya, maka proses iterasi dinyatakan konvergen dan algoritma K-Means berhenti pada iterasi ke-2.

4.5 Hasil Akhir Pengelompokan

Tabel 7. Hasil Akhir Pengelompokan Film

Cluster	Anggota Film	Karakteristik
Cluster 1	12 Angry Men, Whiplash, Fight Club	Durasi pendek-sedang (96-139 menit), rating tinggi bervariasi (8,5-9,0)

Cluster	Anggota Film	Karakteristik
Cluster 2	The Shawshank Redemption	Durasi sedang (142 menit), namun rating tertinggi di antara seluruh sampel (9,3)
Cluster 3	The Godfather Part II, LOTR: The Return of the King	Durasi sangat panjang (>200 menit), tergolong film epik/serial panjang

Berdasarkan Tabel 7, algoritma K-Means berhasil membentuk tiga cluster dengan karakteristik yang berbeda secara jelas. Cluster 1 didominasi oleh film-film dengan durasi pendek hingga sedang, mencerminkan film-film yang lebih ringkas namun tetap memiliki rating tinggi. Cluster 2 hanya beranggotakan satu film, yaitu The Shawshank Redemption, yang menonjol karena memiliki rating tertinggi (9,3) di antara seluruh sampel meskipun durasinya berada pada rentang sedang — hal ini menunjukkan bahwa algoritma K-Means berhasil menangkap keunikan film tersebut relatif terhadap film lain dengan durasi serupa (seperti Fight Club) namun rating yang lebih rendah. Cluster 3 beranggotakan dua film dengan durasi sangat panjang (di atas 200 menit), yaitu The Godfather Part II dan The Lord of the Rings: The Return of the King, yang keduanya memang dikenal sebagai film-film berdurasi panjang bergenre epik.

Hasil ini menunjukkan bahwa meskipun atribut rating memiliki rentang nilai yang sempit, proses normalisasi min-max berhasil membuat atribut tersebut tetap berkontribusi secara proporsional dalam pembentukan cluster, tidak didominasi sepenuhnya oleh atribut durasi yang memiliki rentang nilai lebih besar. Algoritma K-Means pada penelitian ini juga terbukti efisien karena hanya membutuhkan dua kali iterasi untuk mencapai konvergensi, yang menunjukkan bahwa data memiliki struktur pengelompokan yang cukup jelas dan terpisah.

5. KESIMPULAN DAN SARAN

5.1 Kesimpulan

Berdasarkan hasil penelitian, dapat disimpulkan bahwa algoritma K-Means Clustering berhasil diterapkan untuk mengelompokkan film berdasarkan rating dan durasi ke dalam tiga cluster yang memiliki karakteristik berbeda: Cluster 1 (durasi pendek-sedang), Cluster 2 (rating sangat tinggi dengan durasi sedang), dan Cluster 3 (durasi sangat panjang/epik). Proses iterasi konvergen setelah dua kali iterasi, menunjukkan algoritma bekerja secara efisien pada data berskala kecil. Penelitian ini juga menunjukkan pentingnya proses normalisasi data sebelum

penerapan K-Means, khususnya ketika atribut yang digunakan memiliki rentang nilai (skala) yang berbeda jauh, seperti rating (skala 1-10) dan durasi (dalam menit).

5.2 Saran

Penelitian selanjutnya disarankan untuk: (1) menggunakan jumlah sampel film yang lebih besar agar hasil pengelompokan lebih representatif secara statistik; (2) menentukan jumlah cluster (k) yang optimal menggunakan metode Elbow atau Silhouette Score, bukan hanya berdasarkan observasi visual; (3) menambahkan atribut lain seperti jumlah votes, genre, atau tahun rilis untuk memperkaya hasil analisis pengelompokan; serta (4) mengimplementasikan algoritma ini ke dalam sebuah aplikasi atau dashboard interaktif agar dapat digunakan untuk mendukung sistem rekomendasi film secara nyata.

DAFTAR PUSTAKA

Gustientiedina, G., Adiya, M. H., & Desnelita, Y. (2019). Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan. Jurnal Nasional Teknologi dan Sistem Informasi, 5(1), 17–24. https://doi.org/10.25077/teknosi.v5i1.2019.17-24

Maori, N. A., & Evanita, E. (2023). Metode Elbow dalam Optimasi Jumlah Cluster pada K-Means Clustering. Simetris: Jurnal Teknik Mesin, Elektro dan Ilmu Komputer, 14(2), 277–288. https://doi.org/10.24176/simet.v14i2.9630

Sari, V. N., Yupianti, Y., & Maharani, D. (2018). Penerapan Metode K-Means Clustering Dalam Menentukan Predikat Kelulusan Mahasiswa Untuk Menganalisa Kualitas Lulusan. JURTEKSI (Jurnal Teknologi dan Sistem Informasi), 4(2), 133–140.

IMDb (Internet Movie Database). (2026). Rating dan Durasi Film. Diakses Juli 2026, dari https://www.imdb.com