KLASTERISASI KARAKTERISTIK FILM NETFLIX MENGGUNAKAN METODE FUZZY C-MEANS

Ismail Amirul Huda, April Lia Hananto, Tukino, Agustia Hananto
Sistem Informasi, Universitas Buana Perjuangan Karawang
Jalan Ronggo Waluyo Sirnabaya, Puseurjaya, Telukjambe Timur, Karawang, Jawa Barat
si22.ismailhuda@mhs.ubpkarawang.ac.id

ABSTRAK

Kemajuan teknologi informasi telah mendorong peningkatan penggunaan platform streaming seperti Netflix yang menghasilkan data film dalam jumlah besar dengan karakteristik yang beragam. Akan tetapi, data tersebut umumnya masih berada dalam kondisi mentah dan tidak terstruktur, sehingga menyulitkan proses analisis dalam mengidentifikasi pola karakteristik film secara optimal. Penelitian ini bertujuan untuk melakukan pengelompokan data film Netflix agar lebih terorganisasi dan mudah dianalisis dengan menggunakan teknik klasterisasi. Metode yang diterapkan adalah Fuzzy C-Means (FCM), yang memungkinkan setiap data memiliki derajat keanggotaan pada lebih dari satu klaster sehingga mampu merepresentasikan kedekatan karakteristik data secara lebih fleksibel. Tahapan penelitian meliputi preprocessing data melalui penghapusan nilai kosong, normalisasi menggunakan metode Min-Max, penentuan jumlah klaster, serta proses klasterisasi hingga mencapai kondisi konvergen. Hasil penelitian menunjukkan bahwa data film dapat dikelompokkan ke dalam tiga klaster dengan karakteristik yang berbeda, yaitu film berkualitas tinggi, film menengah, dan film komersial. Evaluasi menggunakan Silhouette Score menghasilkan nilai sebesar 0,362 yang menunjukkan bahwa kualitas klasterisasi berada pada kategori cukup baik. Dengan demikian, metode Fuzzy C-Means mampu menghasilkan pengelompokan data yang representatif serta mendukung proses analisis dan pengambilan keputusan berbasis data.

Kata kunci : *Fuzzy C-Means, Klasterisasi, Data Mining, Film Netflix, Preprocessing Data*

1. PENDAHULUAN

Kemajuan teknologi informasi pada era digital telah membawa perubahan signifikan dalam cara masyarakat mengakses dan menikmati hiburan, khususnya melalui platform layanan streaming film seperti Netflix yang menyediakan berbagai pilihan tayangan secara daring dengan jumlah koleksi yang sangat besar dan terus berkembang [1]. Fenomena ini tidak hanya meningkatkan kemudahan akses bagi pengguna, tetapi juga menghasilkan akumulasi data dalam skala besar yang mencerminkan beragam karakteristik film, seperti durasi, tahun rilis, genre, serta tingkat popularitas atau penerimaan penonton terhadap setiap judul yang tersedia [2]. Namun, besarnya *volume* dan kompleksitas data tersebut justru menimbulkan permasalahan dalam proses analisis, karena data yang tersedia umumnya masih bersifat mentah, tidak terstruktur, dan memiliki karakteristik yang saling tumpang tindih, sehingga sulit untuk diinterpretasikan secara langsung tanpa melalui proses pengolahan yang sistematis [3]. Kondisi ini menyebabkan potensi informasi yang terkandung dalam data belum dapat dimanfaatkan secara optimal, khususnya dalam mengidentifikasi pola dan kecenderungan karakteristik film yang dapat mendukung pengambilan keputusan berbasis data. Dalam konteks ini, data mining menjadi pendekatan yang relevan untuk mengekstraksi pengetahuan dari data berskala besar, salah satunya melalui teknik klasterisasi yang bertujuan untuk mengelompokkan data berdasarkan tingkat kemiripan karakteristik tertentu sehingga terbentuk struktur data yang lebih terorganisasi dan mudah dianalisis [4][5]. Meskipun metode klasterisasi seperti K-Means banyak digunakan karena kesederhanaan dan efisiensinya, pendekatan ini memiliki keterbatasan mendasar, yaitu hanya mengakomodasi keanggotaan tunggal di mana setiap data dipaksa berada dalam satu klaster secara mutlak, sehingga kurang mampu merepresentasikan realitas data film yang memiliki sifat ambigu dan multidimensi, seperti satu film yang dapat memiliki lebih dari satu genre atau tingkat popularitas yang berada di antara beberapa kategori [6]. Keterbatasan tersebut menunjukkan adanya kebutuhan akan metode yang mampu menangani ketidakpastian dan derajat kedekatan antar data secara lebih fleksibel. Fuzzy C-Means hadir sebagai alternatif yang menawarkan konsep keanggotaan ganda dengan memberikan nilai derajat keanggotaan pada setiap data terhadap beberapa klaster sekaligus, sehingga mampu merepresentasikan hubungan antar data secara lebih realistis dan adaptif terhadap kompleksitas data [7]. Dengan karakteristik tersebut, metode ini dinilai lebih sesuai untuk digunakan dalam proses klasterisasi data film yang tidak memiliki batasan kategori yang tegas. Berdasarkan permasalahan dan kebutuhan tersebut, penelitian ini bertujuan untuk menerapkan metode Fuzzy C-Means dalam melakukan klasterisasi data film Netflix guna menghasilkan pengelompokan yang lebih representatif, terstruktur, dan informatif, sehingga dapat membantu dalam mengungkap pola karakteristik film secara lebih komprehensif serta mendukung proses analisis data dan pengambilan keputusan berbasis informasi, sekaligus memberikan kontribusi akademis dalam pengembangan penerapan

metode klasterisasi pada data berskala besar di bidang data mining dan sistem informasi.

2. TINJAUAN PUSTAKA

2.1. Data dan Pengolahan Data

Data merupakan representasi fakta objektif dalam bentuk numerik, teks, maupun simbol yang berfungsi sebagai fondasi utama dalam menghasilkan informasi bernilai guna. Dalam penelitian berbasis klasterisasi, penggunaan data kuantitatif atau numerik menjadi dominan karena memungkinkan dilakukannya analisis matematis untuk mengukur tingkat kedekatan antar objek berdasarkan atribut yang dimiliki[8]. Kualitas dan kelengkapan data sangat menentukan akurasi hasil analisis, sehingga diperlukan proses pengolahan awal (*preprocessing*) sebelum memasuki tahap inti[1]. Proses ini mencakup pembersihan data dari ketidakkonsistenan, pemilihan atribut yang relevan, serta normalisasi untuk menyamakan skala pengukuran . Tahapan pengolahan ini krusial untuk memastikan data berada dalam kondisi siap uji, sehingga metode klasterisasi yang diterapkan dapat menghasilkan pengelompokan objek yang optimal dan *valid* secara ilmiah.

2.2. Data Mining

Data mining merupakan tahapan krusial dalam proses *Knowledge Discovery in Databases* (KDD) yang berfokus pada upaya menemukan pola, hubungan, atau kecenderungan tertentu dari sekumpulan data melalui pendekatan komputasi dan analisis matematis[9]. Implementasi data mining memungkinkan proses ekstraksi informasi dari basis data berskala besar dilakukan secara sistematis, sehingga menghasilkan struktur informasi yang lebih terorganisasi dibandingkan proses analisis manual. Dalam praktiknya, data mining mencakup berbagai teknik analisis seperti klasifikasi, regresi, asosiasi, dan klasterisasi, di mana pemilihan teknik tersebut sangat bergantung pada tujuan penelitian serta karakteristik data yang digunakan[10]. Sebagai dasar konseptual, data mining menjadi jembatan penting untuk mentransformasi data mentah menjadi pengetahuan yang memiliki nilai guna bagi pengambilan keputusan.

2.3. Klasterisasi

Klasterisasi merupakan metode *unsupervised learning* dalam data mining yang digunakan untuk mengelompokkan data ke dalam beberapa klaster berdasarkan tingkat kemiripan karakteristik antar objek tanpa memerlukan label kelas sebelumnya[11]. Prinsip utama teknik ini adalah membentuk kelompok dengan tingkat keseragaman (homogenitas) tinggi di dalam klaster, namun memiliki perbedaan yang kontras antar klaster[6]. Penentuan kemiripan antar data umumnya dihitung menggunakan ukuran jarak tertentu, seperti jarak Euclidean, yang mengharuskan penggunaan atribut kuantitatif agar dapat diproses secara matematis[12]. Berdasarkan sifat keanggotaannya, metode ini terbagi menjadi dua jenis, yaitu klasterisasi tegas (*hard clustering*) dan klasterisasi lunak (*soft clustering*). Pada hard clustering, setiap data terikat secara mutlak pada satu kelompok tunggal, sementara soft clustering memungkinkan suatu data memiliki derajat keanggotaan pada lebih dari satu klaster secara simultan[13]. Pendekatan *soft clustering* dinilai lebih representatif untuk menangani dataset dengan batas antar kelompok yang tidak tegas atau saling beririsan, sehingga menjadi landasan pemilihan metode dalam penelitian ini[14].

2.4. Metode Fuzzy C-Means

Fuzzy C-Means (FCM) merupakan teknik klasterisasi dalam pendekatan *soft clustering* yang memungkinkan setiap data memiliki tingkat keanggotaan pada lebih dari satu klaster. Berbeda dengan metode klasterisasi tegas[15], FCM menghasilkan nilai keanggotaan pada interval 0 hingga 1 yang merepresentasikan derajat kedekatan suatu data terhadap pusat klaster. Proses algoritma ini bersifat iteratif, diawali dengan penentuan jumlah klaster dan pembentukan matriks derajat keanggotaan awal. Pusat klaster diperbarui secara berulang berdasarkan bobot nilai keanggotaan hingga mencapai kondisi konvergen, yaitu saat perubahan nilai antar iterasi lebih kecil dari batas toleransi yang ditetapkan. Perhitungan pusat klaster dilakukan menggunakan persamaan (1):

$$v_j = \frac{\sum_{i-1}^{n} (\mu_{ij})^m x_i}{\sum_{i=1}^{n} (\mu_{ij})^m} \quad (1)$$

Keterangan: v_j adalah pusat klaster ke- j, μ_{ij} adalah derajat keanggotaan data ke- i pada klaster ke- j, m adalah parameter pembobot, x_i adalah data ke- i, dan n adalah jumlah data.

Secara konseptual, FCM berupaya meminimalkan fungsi objektif yang menggambarkan total jarak antara data dan pusat klaster menggunakan jarak Euclidean pada atribut numerik. Parameter pembobot m memengaruhi tingkat keluwesan pembentukan klaster, di mana nilai yang lebih besar menghasilkan batas antarkelompok yang lebih samar[16]. Dalam penelitian ini, FCM diterapkan pada data film Netflix untuk memberikan gambaran pengelompokan yang lebih fleksibel dan informatif dibandingkan metode konvensional.

2.5. Preprocessing Data

Preprocessing data merupakan tahapan krusial untuk mempersiapkan data mentah agar berada dalam kondisi siap uji sebelum dilakukan analisis lebih lanjut[17]. Proses ini bertujuan untuk mengatasi berbagai permasalahan data seperti nilai kosong (*missing values*), ketidakkonsistenan, serta perbedaan skala antar atribut yang dapat memengaruhi validitas hasil pengelompokan[18]. Salah satu tahapan utama dalam preprocessing pada penelitian ini adalah normalisasi data menggunakan metode Min-Max

Normalization[19]. Normalisasi dilakukan untuk menyetarakan rentang nilai setiap atribut sehingga tidak terjadi dominasi nilai pada atribut tertentu yang memiliki skala lebih besar dalam proses klasterisasi. Perhitungan normalisasi Min-Max dilakukan dengan menggunakan persamaan (2):

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (2)$$

Keterangan: X' adalah nilai hasil normalisasi, X adalah nilai data awal, X_{min} adalah nilai minimum pada atribut, dan X_{max} adalah nilai maksimum pada atribut. Melalui tahapan ini, data akan bertransformasi ke dalam rentang nilai yang seragam, sehingga algoritma Fuzzy C-Means dapat memberikan hasil klasterisasi yang lebih optimal dan akurat.

2.6. Film Netflix Sebagai Objek Penelitian

Objek utama dalam penelitian ini adalah dataset film Netflix yang bersumber dari platform Kaggle, yang mencakup berbagai atribut informasi seperti durasi, tahun rilis, genre, serta atribut numerik lainnya. Pemilihan data film Netflix didasarkan pada keberagaman karakteristik konten yang tersedia, sehingga memungkinkan terbentuknya pola-pola tertentu yang representatif saat dilakukan proses klasterisasi. Mengingat pertumbuhan pesat industri layanan streaming, data ini menjadi sumber informasi yang relevan untuk dianalisis guna memahami struktur dan kecenderungan karakteristik film secara mendalam. Dalam prosesnya, dilakukan tahap seleksi atribut untuk menentukan variabel numerik yang sesuai dengan kebutuhan komputasi metode Fuzzy C-Means. Langkah seleksi ini krusial agar hasil pengelompokan lebih terarah dan mampu mencerminkan karakteristik data secara akurat, sehingga dapat memberikan gambaran informatif mengenai pola sebaran film berdasarkan variabel yang dianalisis.

2.7. Penelitian Terkait

Berbagai penelitian terdahulu telah memanfaatkan teknik data mining untuk menganalisis data film dengan pendekatan yang beragam. Suraya [1] menggunakan metode Exploratory Data Analysis (EDA) dan visualisasi data yang menunjukkan dominasi konten film pada platform Netflix sebesar 69,7% serta menghasilkan pengelompokan kategori berdasarkan periode waktu, yaitu lama hingga modern. Namun, pendekatan ini masih terbatas pada analisis deskriptif sehingga belum mampu mengungkap pola pengelompokan data secara mendalam. Selanjutnya, Rahman [6] menerapkan metode K-Means dalam kerangka Knowledge Discovery in Database (KDD) dan berhasil mengelompokkan 2.108 data film Amazon Prime ke dalam enam klaster utama, meskipun metode ini masih bergantung pada keanggotaan tunggal yang cenderung menyederhanakan kompleksitas data. Penelitian oleh Andrew [5] membandingkan metode K-Means dan Agglomerative Hierarchical Clustering (AHC), di mana K-Means menghasilkan pemisahan klaster yang lebih tegas, sedangkan AHC menunjukkan struktur pengelompokan yang lebih variatif, namun keduanya belum mampu merepresentasikan tingkat kedekatan data secara fleksibel. Selain itu, Heikal [20] menggunakan K-Means untuk mengidentifikasi tiga segmen pengguna dengan karakteristik berbeda, tetapi pendekatan ini masih menghadapi keterbatasan dalam menangani data yang memiliki sifat ambigu. Di sisi lain, Mukhsinin [21] menerapkan algoritma Decision Tree untuk klasifikasi rating film dengan tingkat akurasi sebesar 39,47%, yang menunjukkan bahwa pendekatan klasifikasi belum sepenuhnya optimal dalam menangkap hubungan kompleks antar data. Sementara itu, Sudarsono [2] menggunakan algoritma Apriori dan menunjukkan efektivitas RapidMiner dalam mengidentifikasi genre film populer, namun lebih berfokus pada hubungan asosiasi dibandingkan pengelompokan karakteristik data secara menyeluruh.

Berdasarkan berbagai penelitian tersebut, dapat disimpulkan bahwa sebagian besar pendekatan yang digunakan masih memiliki keterbatasan dalam merepresentasikan kompleksitas dan ketidakpastian karakteristik data film, terutama yang bersifat multidimensi dan saling tumpang tindih. Hal ini menunjukkan adanya kebutuhan akan metode yang mampu mengakomodasi derajat kedekatan antar data secara lebih fleksibel. Oleh karena itu, penerapan metode Fuzzy C-Means menjadi relevan untuk dikaji lebih lanjut, karena mampu memberikan representasi keanggotaan ganda yang lebih adaptif dalam proses klasterisasi, khususnya pada data film streaming seperti Netflix yang memiliki karakteristik tidak tegas dan dinamis.

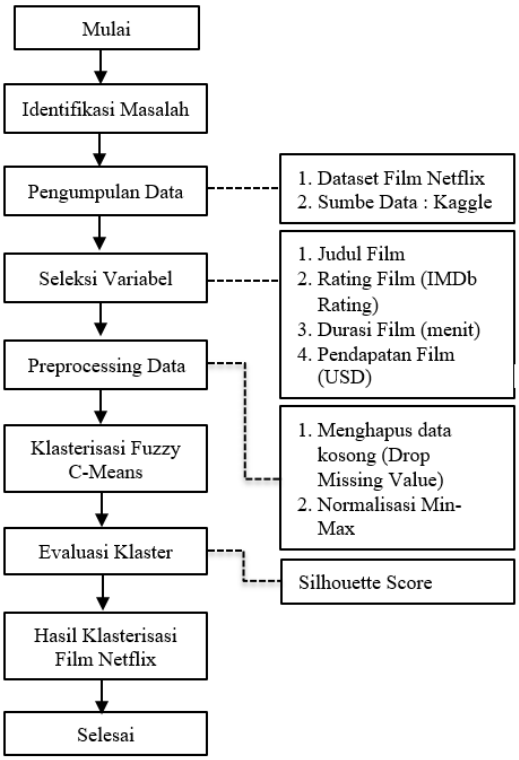
3. METODE PENELITIAN

3.1. Objek Penelitian

Objek penelitian ini adalah data film pada platform Netflix yang diperoleh dari dataset terbuka Kaggle. Dataset ini dipilih karena memiliki karakteristik atribut yang beragam, sehingga relevan untuk diuji menggunakan metode Fuzzy C-Means (FCM). Sebanyak 1.000 data film digunakan dalam penelitian ini dengan fokus pada tiga variabel utama: rating IMDb (imdb_rating), durasi film (duration_minutes), dan pendapatan film (box_office_revenue). Ketiga variabel tersebut dipilih karena mampu merepresentasikan aspek kualitas, lama penayangan, serta tingkat popularitas film secara kuantitatif.

3.2. Prosedur Penelitian

Prosedur penelitian disusun secara sistematis untuk menjamin alur kerja yang terarah, mulai dari tahap awal hingga evaluasi hasil. Alur penelitian secara lengkap ditunjukkan pada Gambar 1.



Gambar 1. Flowchart alur penelitian

3.3. Identifikasi Masalah

Tahap ini dilakukan untuk menentukan fokus penelitian berdasarkan kondisi data film Netflix yang tersedia. Dataset memiliki karakteristik yang beragam sehingga diperlukan pendekatan yang dapat mengelompokkan data berdasarkan tingkat kemiripan karakteristiknya. Penelitian ini memanfaatkan metode Fuzzy C-Means sebagai teknik klasterisasi untuk memperoleh gambaran karakteristik pada masing-masing kelompok film.

3.4. Pengumpulan Data

Tahap pengumpulan data dilakukan dengan memperoleh dataset film Netflix dari situs Kaggle dalam format Comma *Separated Values* (CSV). Dataset tersebut berisi informasi mentah mengenai profil film seperti judul, durasi, rating, dan pendapatan yang akan digunakan sebagai dasar analisis pada tahap penelitian selanjutnya.

3.5. Seleksi Variabel

Pemilihan atribut dilakukan berdasarkan relevansi terhadap karakteristik film. Atribut yang digunakan meliputi judul film sebagai identitas data, rating film (IMDb rating) sebagai representasi kualitas, durasi film (menit) sebagai aspek teknis, serta pendapatan film (USD) sebagai aspek komersial. Kombinasi atribut tersebut digunakan untuk menggambarkan karakteristik film dalam proses klasterisasi.

Dataset yang digunakan dalam penelitian ini diperoleh dari Kaggle dengan jumlah data sebanyak 1040 data. Data tersebut terdiri dari beberapa atribut yang merepresentasikan karakteristik film. Selanjutnya, data akan melalui tahap preprocessing untuk memastikan kualitas data sebelum dilakukan.

3.6. Preprocessing Data

Pada tahap ini dilakukan proses preprocessing untuk mempersiapkan data sebelum klasterisasi, yang meliputi penghapusan data yang memiliki nilai kosong (missing value) serta normalisasi data. Penanganan data kosong dilakukan dengan cara menghapus baris data yang memiliki nilai kosong pada atribut yang digunakan, karena jumlahnya relatif sedikit sehingga tidak mempengaruhi keseluruhan data secara signifikan. Selanjutnya, dilakukan normalisasi menggunakan metode Min-Max Normalization untuk menyetarakan rentang nilai setiap atribut ke dalam skala 0 hingga 1, sehingga tidak terdapat atribut yang memiliki pengaruh lebih besar dalam proses perhitungan jarak pada klasterisasi.

3.7. Klasterisasi Fuzzy C-Means

Data yang telah siap digunakan sebagai input untuk proses pengelompokan. Algoritma FCM melakukan proses iterasi untuk menentukan pusat klaster dan derajat keanggotaan setiap data film terhadap masing-masing klaster. Proses ini dilakukan secara berulang hingga mencapai kondisi konvergen atau perubahan nilai yang berada di bawah batas toleransi.

3.8. Evaluasi Klaster

Evaluasi dilakukan untuk menilai kualitas hasil pengelompokan yang dihasilkan. Proses ini dilakukan dengan menguji beberapa jumlah klaster yang berbeda, kemudian masing-masing hasil klasterisasi dievaluasi menggunakan metode Silhouette Score. Metode ini digunakan untuk mengukur tingkat kedekatan data dalam satu klaster serta perbedaan antar klaster, sehingga dapat diketahui kualitas pengelompokan yang dihasilkan. Berdasarkan nilai evaluasi tersebut, ditentukan jumlah klaster yang paling optimal untuk digunakan dalam proses klasterisasi.

3.9. Hasil Klasterisasi Film Netflix

Proses klasterisasi dilakukan menggunakan metode Fuzzy C-Means untuk mengelompokkan data berdasarkan tingkat kemiripan karakteristik yang dimiliki. Metode ini digunakan karena mampu memberikan derajat keanggotaan pada setiap data terhadap masing-masing klaster, sehingga hasil pengelompokan menjadi lebih fleksibel. Proses klasterisasi dilakukan terhadap data yang telah melalui tahap preprocessing dengan jumlah klaster yang telah ditentukan sebelumnya berdasarkan hasil evaluasi. Hasil dari proses ini berupa pengelompokan data ke dalam beberapa klaster yang memiliki karakteristik tertentu.

4. HASIL DAN PEMBAHASAN

4.1. Deskripsi Data

Dataset yang digunakan dalam penelitian ini diperoleh dari Kaggle dengan jumlah data sebesar 1040 data. Dataset tersebut berisi informasi terkait karakteristik film yang mencakup judul film, rating film (IMDb), durasi film dalam satuan menit, serta pendapatan film dalam satuan dolar Amerika Serikat (USD). Data yang diperoleh masih dalam kondisi awal (data mentah) sehingga perlu dilakukan proses preprocessing sebelum digunakan dalam tahap klasterisasi.

Tabel 1. Data awal film netflix

No	Judul Film	Rating Film (IMDb)	Durasi (Menit)	Pendapatan (USD)
1	Dragon Legend	-	35	-
2	Fire Mystery	4.4	111	52,104,449
3	A Empire	5.6	98	391,629,433
4	Mystery Quest	6.4	116	2,016,406
5	Hero Empire	-	108	-
6	Kingdom Day	4.6	111	2,736,647
7	Hero Journey	4.9	104	16,638,609
8	Mission Dream	5.9	82	7,674,756
…….	…….	……..	…….	…….
1040	Adventure King	7.9	104	26,772,080

Tabel 1, menunjukkan data awal yang digunakan dalam penelitian. Data tersebut masih mengandung nilai kosong sehingga perlu dilakukan preprocessing sebelum digunakan pada tahap klasterisasi.

4.2. Preprocessing Data

Pada tahap ini dilakukan proses preprocessing untuk mempersiapkan data sebelum klasterisasi, yang meliputi penghapusan data yang mempunyai nilai kosong (missing value) serta normalisasi data. Penanganan data kosong dilakukan dengan cara menghapus baris data yang mempunyai nilai kosong pada atribut yang digunakan. Setelah melakukan proses tersebut, jumlah data berkurang sebanyak 752 data dari 1040 data menjadi 288 data.

Tabel 2. Data dengan missing value

No	Judul Film	Rating Film (IMDb)	Durasi (Menit)	Pendapatan (USD)
1	Dragon Legend	-	35	-
2	Strom Warrior	3.3	37	-
3	A Empire	5.6	98	391,629,433
4	Mystery Quest	6.4	116	2,016,406
5	Hero Empire	-	108	-
6	Kingdom Day	4.6	111	2,736,647
7	Hero Journey	4.9	104	16,638,609
8	Mission Dream	5.9	82	7,674,756
…….	…….	……..	…….	…….
752	Adventure King	7.9	104	26,772,080

Tabel 2, menunjukkan contoh data yang memiliki nilai kosong pada beberapa atribut. Data tersebut tidak dapat digunakan dalam proses klasterisasi karena dapat mempengaruhi hasil pengelompokan.

Tabel 3. Data setelah preprocessing

No	Judul Film	Rating Film (IMDb)	Durasi (Menit)	Pendapatan (USD)
1	Princess Phoenix	7.6	116	7,278,095
2	Fire Mystery	4.4	111	52,104,449
3	A Empire	5.6	98	391,629,433
4	Mystery Quest	6.4	116	2,016,406
5	First Hero	6.7	452	55,547,817
6	Kingdom Day	4.6	111	2,736,647
…….	…….	……..	…….	…….
288	Adventure King	7.9	104	26,772,080

Tabel 3, menunjukkan data yang telah melalui proses preprocessing, di mana seluruh atribut telah selesai sehingga data siap digunakan dalam proses klasterisasi. Setelah dilakukan proses penghapusan data yang mempunyai nilai kosong, jumlah data berkurang dari 1040 data menjadi 288 data.

Tabel 4. Jumlah data sebelum dan sesudah preprocessing

Keterangan	Jumlah Data
Data Awal	1040
Setelah Preprocessing	288

Tabel 4, menunjukkan perubahan jumlah data setelah dilakukan penghapusan data kosong. Data yang digunakan dalam proses selanjutnya merupakan data yang telah bersih dan siap untuk dianalisis. Selanjutnya dilakukan normalisasi menggunakan metode Min-Max untuk menyetarakan rentang nilai setiap atribut ke dalam skala 0 hingga 1, sehingga tidak terdapat atribut yang memiliki pengaruh lebih besar dalam proses perhitungan jarak pada klasterisasi.

Table 5. Data hasil normalisasi min max

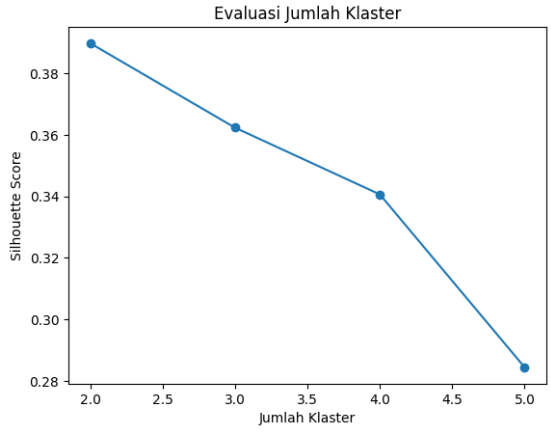
No	Judul Film	Rating Film (IMDb)	Durasi (Menit)	Pendapatan (USD)
1	Princess Phoenix	0.736264	0.184028	0.003568
2	Fire Mystery	0.384615	0.180556	0.025628
3	A Empire	0.912088	0.203125	0.000619
4	Mystery Quest	0.494505	0.185764	0.003002
5	First Hero	0.725275	0.244792	0.004599
6	Kingdom Day	0.417582	0.206597	0.000998
…..	…….	……..	….……	……..…
288	Adventure King	0.406593	0.175347	0.001333

4.3. Penentuan Jumlah Klasster

Penentuan jumlah klaster dilakukan dengan menguji beberapa nilai k, yaitu k = 2 sampai k = 5. Setiap hasil klasterisasi dievaluasi menggunakan metode Silhouette Score untuk mengetahui kualitas pengelompokan yang dihasilkan.

Tabel 6. Hasil evaluasi jumlah klaster

Jumlah Klaster (k)	Silhouette Score
2	0.389
3	0.362
4	0.340
5	0.284



Gambar 2. Grafik evaluasi jumlah klaster

Tabel 6, menunjukkan hasil evaluasi untuk setiap jumlah klaster yang diuji. Nilai Silhouette Score tertinggi diperoleh pada k=2, namun jumlah klaster tersebut dianggap kurang mampu mewakili variasi data secara lebih rinci.

Gambar 2, menunjukkan bahwa nilai Silhouette Score mengalami penurunan seiring bertambahnya jumlah klaster. Meskipun k=2 memiliki nilai tertinggi, dipilih k=3 karena masih memiliki nilai yang cukup baik dan mampu memberikan pengelompokan yang lebih representatif.

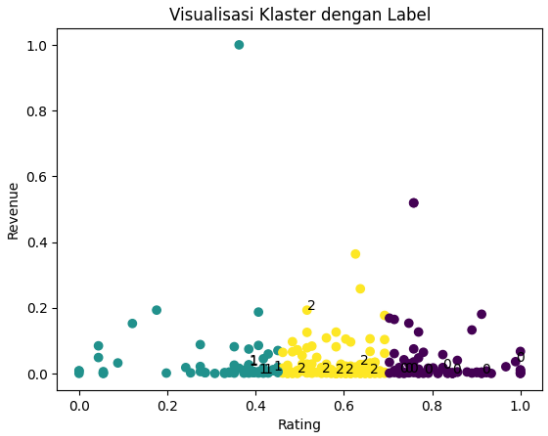
4.4. Hasil Klasterisasi

Proses klasterisasi dilakukan menggunakan metode Fuzzy C-Means dengan jumlah klaster sebanyak 3. Hasil klasterisasi berupa pengelompokan data ke dalam tiga kelompok berdasarkan kesamaan karakteristik.

Tabel 7. Hasil klasterisasi

No	Judul Film	Rating Film (IMDb)	Durasi (Menit)	Pendapatan (USD)	Cluster
1	Princess Phoenix	7.6	116	7,278,095	0
2	Fire Mystery	4.4	114	52,104,449	1
3	Quest Queen	9.2	127	1,286,089	0
4	Dark Princess	5.4	117	6,128,075	2
5	Dragon King	7.5	151	9,373,734	0
6	Warrior War	4.7	129	2,057,069	1
….	…..	…….	…….	….……	……
288	Mystery Quest	6.4	116	2,016,406	2

Tabel 7, menunjukkan hasil klasterisasi, di mana setiap data telah memiliki label klaster yang menunjukkan kelompoknya masing-masing. Label klaster ini menjadi dasar dalam proses analisis untuk mengetahui karakteristik dari setiap kelompok data.



Gambar 3. Visualisasi hasil klasterisasi

Gambar 3, menunjukkan penyajian data berdasarkan atribut yang digunakan dalam proses klasterisasi. Setiap warna merepresentasikan klaster yang berbeda-beda. Terlihat bahwa data komprehensif menjadi beberapa kelompok sesuai dengan hasil pengelompokan menggunakan metode Fuzzy C-Means.

4.5. Analisis Klaster

Analisis klaster dilakukan untuk menentukan karakteristik dari setiap kelompok data yang dihasilkan. Analisis ini dilakukan dengan perhitungan nilai rata-rata dari setiap atribut pada masing-masing klaster.

Tabel 8. Rata-rata tiap klaster

Cluster	Rating Film (IMDb)	Durasi (menit)	Pendapatan (USD)
0	8.35	113	69,000,000
1	6.20	125	49,000,000
2	3.74	126	87,000,000

Berdasarkan Tabel 8, terlihat bahwa setiap klaster memiliki karakteristik yang berbeda. Klaster 0 memiliki nilai rating tertinggi yaitu sebesar 8,35 dengan durasi rata-rata 113 menit serta pendapatan yang cukup tinggi, sehingga menunjukkan kelompok film dengan kualitas yang baik. Klaster 1 memiliki nilai rating sedang yaitu sebesar 6,20 dengan durasi paling panjang serta pendapatan paling rendah, sehingga merepresentasikan kelompok film dengan karakteristik menengah. Sementara itu, klaster 2 memiliki nilai rating paling rendah yaitu sebesar 3,74 namun memiliki pendapatan tertinggi, yang menunjukkan kelompok film yang bersifat komersial.

Tabel 9. Penamaan klaster

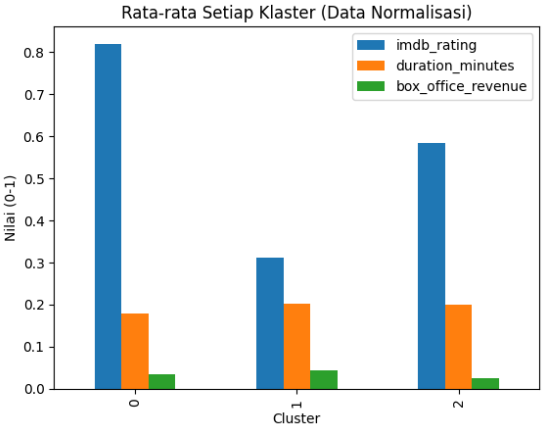
Cluster	Nama Klaster
0	Film Berkualitas Tinggi
1	Film Menengah
2	Film Komersial

Hasil klasterisasi menunjukkan adanya perbedaan karakteristik yang cukup jelas antar kelompok data. Klaster dengan nilai rating tinggi dan pendapatan yang juga tinggi menunjukkan bahwa film dengan kualitas yang baik cenderung memiliki performa komersial yang baik. Hal ini mengindikasikan bahwa kualitas film menjadi salah satu faktor yang mempengaruhi tingkat pendapatan.

Di sisi lain, terdapat klaster yang memiliki nilai rating rendah namun memiliki pendapatan tertinggi. Fenomena ini menunjukkan bahwa tidak semua film dengan kualitas rendah mengalami kegagalan secara komersial. Faktor lain seperti strategi pemasaran, popularitas aktor, atau tren pasar dapat mempengaruhi tingginya pendapatan film tersebut.

Sementara itu, klaster dengan nilai sedang pada seluruh atribut menunjukkan kelompok film yang tidak menonjol baik dari segi kualitas maupun pendapatan. Kelompok ini merepresentasikan film dengan karakteristik umum yang cenderung berada di tengah.

Hasil ini menunjukkan bahwa metode Fuzzy C-Means mampu mengelompokkan data berdasarkan pola yang tidak hanya bergantung pada satu atribut, tetapi mempertimbangkan keseluruhan karakteristik data, sehingga menghasilkan pengelompokan yang lebih fleksibel dan informatif.



Gambar 4. Grafik rata-rata setiap klaster

Gambar 4, menunjukkan perbandingan nilai rata-rata setiap atribut pada masing-masing klaster setelah dilakukan normalisasi. Penggunaan data hasil normalisasi bertujuan untuk menghindari perbedaan skala antar atribut, sehingga setiap atribut dapat dibandingkan secara seimbang dalam satu grafik.

Berdasarkan grafik tersebut, terlihat bahwa Klaster 0 memiliki nilai tertinggi pada atribut rating, yang menunjukkan bahwa kelompok ini terdiri dari film dengan kualitas yang baik. Klaster 1 menunjukkan nilai yang relatif sedang pada seluruh atribut, sehingga menggambarkan kelompok film dengan karakteristik menengah. Sementara itu, Klaster 2 memiliki nilai rating yang paling rendah, namun menunjukkan nilai yang tinggi pada atribut pendapatan, yang mengindikasikan bahwa kelompok ini terdiri dari film yang memiliki performa komersial tinggi meskipun kualitasnya relatif rendah.

Perbedaan pola pada grafik ini menunjukkan bahwa setiap klaster memiliki karakteristik yang berbeda secara jelas. Hal ini menegaskan bahwa metode klasterisasi yang digunakan mampu mengelompokkan data berdasarkan pola yang terbentuk dari kombinasi atribut yang digunakan.

4.6. Evaluasi

Evaluasi klaster dilakukan untuk menilai kualitas hasil pengelompokan data yang telah dilakukan menggunakan metode Fuzzy C-Means. Pada penelitian ini, evaluasi dilakukan menggunakan metode Silhouette Score, yang digunakan untuk mengukur tingkat kedekatan data dalam satu klaster serta perbedaan antar klaster.

Berdasarkan hasil pengujian yang telah dilakukan pada beberapa jumlah klaster, diperoleh nilai Silhouette Score tertinggi pada k=2 sebesar 0,389. Namun, dalam penelitian ini dipilih k=3 sebagai jumlah klaster yang digunakan karena dianggap mampu memberikan pengelompokan yang lebih representatif dan mudah diinterpretasikan. Pada jumlah klaster k=3, diperoleh nilai Silhouette Score sebesar 0,362**.**

Nilai Silhouette Score sebesar 0,362 menunjukkan bahwa hasil klasterisasi memiliki kualitas yang cukup baik dalam memisahkan data ke dalam kelompok yang berbeda. Meskipun tidak mencapai nilai yang sangat tinggi, hasil ini tetap dapat digunakan untuk analisis karena masih menunjukkan adanya struktur pengelompokan yang jelas pada data.

5. KESIMPULAN DAN SARAN

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa metode Fuzzy C-Means berhasil digunakan untuk mengelompokkan data film ke dalam tiga klaster berdasarkan karakteristik yang dimiliki. Proses preprocessing dengan penghapusan data kosong dan normalisasi data mampu meningkatkan kualitas data sehingga hasil klasterisasi menjadi lebih optimal. Hasil pengelompokan menunjukkan adanya perbedaan karakteristik yang jelas pada setiap klaster, yaitu kelompok film berkualitas tinggi, film menengah, dan film komersial, di mana terdapat fenomena bahwa film dengan kualitas rendah tetap dapat memiliki pendapatan yang tinggi. Evaluasi menggunakan Silhouette Score menghasilkan nilai sebesar 0,362 yang menunjukkan bahwa hasil klasterisasi memiliki kualitas yang cukup baik. Untuk penelitian selanjutnya, disarankan untuk menggunakan jumlah data yang lebih besar serta menambahkan atribut lain yang relevan agar hasil klasterisasi menjadi lebih akurat dan representatif, serta dapat mempertimbangkan penggunaan metode clustering lainnya sebagai pembanding untuk meningkatkan kualitas analisis.

DAFTAR PUSTAKA

[1] p. W. Suraya, “eksplorasi dan visualisasi data konten netflix berdasarkan perbandingan film dan acara tv yang dibuat oleh negara di dunia,” 2025.

[2] b. G. Sudarsono, m. I. Leo, a. Santoso, and f. Hendrawan, “analisis data mining data netflix menggunakan aplikasi rapid miner,” jbase - journal of business and audit information systems, vol. 4, no. 1, apr. 2021, doi: 10.30813/jbase.v4i1.2729.

[3] a. A. Alzahra et al., “klasifikasi genre dan rating konten di netflix : wawasan dari analisis chi-square dan clustering,” jatisi (jurnal teknik informatika dan sistem informasi), vol. 12, no.1, no. 1, pp. 1–8, mar. 2025.

[4] d. A. Manalu and g. Gunadi, “implementasi metode data mining k-means clustering terhadap data pembayaran transaksi menggunakan bahasa pemrograman python pada cv digital dimensi,” infotech: journal of technology information, vol. 8, no. 1, pp. 43–54, jun. 2022, doi: 10.37365/jti.v8i1.131.

[5] n. Andrew, “clustering data film dan series netflix menggunakan metode k-means dan agglomerative hierarchical clustering,” versi cetak, 2023. [online]. Available: www.kaggle.com

[6] n. F. Rahman et al., “penerapan algoritma k-means untuk klasterisasi film pada platform amazon prime video,” 2025.

[7] h. Syukron, m. Fauzi fayyad, f. Junita fauzan, y. Ikhsani, and u. Rizkya gurning, “malcom: indonesian journal of machine learning and computer science comparison k-means k-medoids and fuzzy c-means for clustering customer data with lrfm model "perbandingan k-means k-medoids dan fuzzy c-means untuk pengelompokan data pelanggan dengan model lrfm,” vol. 2, pp. 76–83, 2022.

[8] m. Djaka permana, a. Lia hananto, e. Novalia, b. Huda, and t. Paryono, “klasterisasi data jamaah umrah pada tanurmutmainah tour menggunakan algoritma k-means,” jurnal komtekinfo, pp. 15–20, feb. 2023, doi: 10.35134/komtekinfo.v10i1.332.

[9] r. Muliono and z. Sembiring, “data mining clustering menggunakan algoritma k-means untuk klasterisasi tingkat tridarma pengajaran dosen,” 2019.

[10] d. Puspita sari, s. Shofia hilabi, and agustia hananto, “penerapan data mining metode k-nearest neighbor untuk memprediksi kelulusan siswa sekolah menengah pertama,” smartics journal, vol. 9, no. 1, pp. 14–19, mar. 2023, doi: 10.21067/smartics.v9i1.8088.

[11] a. Rohmatullah, “klasterisasi data pertanian di kabupaten lamongan menggunakan algoritma k-means dan fuzzy c means,” jurnal ilmiah teknosains, no. 2, p. 1, 2020.

[12] i. Ferdiansyah, b. Huda, a. Hananto, and u. Buana perjuangan karawang, “analisis clustering menggunakan metode k-means pada kemiskinan di jawa timur tahun 2020,” innovative: journal of social sciense research, vol. 4, no.3, no. 1, may 2024.

[13] p. Apriyani, a. R. Dikananda, and i. Ali, “penerapan algoritma k-means dalam klasterisasi kasus stunting balita desa tegalwangi,” hello world jurnal ilmu komputer, vol. 2, no. 1, pp. 20–33, mar. 2023, doi: 10.56211/helloworld.v2i1.230.

[14] i. Virgo, s. Defit, and y. Yuhandri, "klasterisasi tingkat kehadiran dosen menggunakan algoritma k-means clustering," jurnal sistim informasi dan teknologi, pp. 23–28, mar. 2020, doi: 10.37034/jsisfotek.v2i1.17.

[15] y. Nataliani, "klasterisasi kinerja karyawan menggunakan algoritma fuzzy c-means," aiti: jurnal teknologi informasi, vol. 17, no. Agustus, pp. 118–129, 2020.

[16] h. Firdaus, "analisa cluster menggunakan k-means dan fuzzy c-means dalam pengelompokan provinsi menurut data intesitas bencana alam di indonesia tahun 2017-2021," 2021.

[17] a. Agung, a. Daniswara, i. Kadek, and d. Nuryana, "data preprocessing pola pada penilaian mahasiswa program profesi guru," journal of informatics and computer science, vol. 05, 2023.

[18] a. Gunawan, a. Hermawan, and d. Avianto, "bulletin of computer science research analisis pengaruh preprocessing data dan hyperparameter tuning pada backpropagation neural network dalam klasifikasi stroke," 2026, doi: 10.47065/bulletincsr.v6i2.956.

[19] l. Fisch et al., "deepmriprep: voxel-based morphometry preprocessing via deep neural networks," nat. Comput. Sci., vol. 6, no. 3, pp. 250–259, jan. 2026, doi: 10.1038/s43588-026-00953-7.

[20] j. Heikal, "customer segmentation using k-means clustering on top of 4 platform streaming in indonesia," jurnal cendekia ilmiah, vol. 4, no. 4, 2025.

[21] d. A. Mukhsinin, m. Rafliansyah, s. A. Ibrahim, r. Rahmaddeni, and d. Wulandari, "implementasi algoritma decision tree untuk rekomendasi film dan klasifikasi rating pada platform netflix," malcom: indonesian journal of machine learning and computer science, vol. 4, no. 2, pp. 570–579, mar. 2024, doi: 10.57152/malcom.v4i2.1255.