



# Pengelompokkan Film Pada Platform Netflix Menggunakan Metode K-Means Clustering Sebagai Rekomendasi Film

Cahyani Budihartanti<sup>1,\*</sup>, Cyrill Ilario Ifaru<sup>2</sup>, Alfia Zahra<sup>2</sup>, Muhammad Hardithya Aenuddin<sup>2</sup>,  
Eri Nur Setiawan<sup>2</sup>, Cahyadi Anugrah Maulana<sup>2</sup>, Riky Putra Pratama<sup>2</sup>

<sup>1</sup>Fakultas Teknologi Informasi, Program Studi Sistem Informasi, Universitas Nusa Mandiri, Jakarta  
Jl. Raya Jatiwaringin No.2, RT.8/RW.13, Cipinang Melayu, Kec. Makasar, Kota Jakarta Timur, Daerah Khusus Ibukota  
Jakarta, Indonesia

<sup>2</sup>Fakultas Teknik dan Informatika, Program Studi Sistem Informasi, Universitas Bina Sarana Informatika, Depok  
Jl. Margonda Raya No.8, Pondok Cina, Kecamatan Beji, Kota Depok, Jawa Barat, Indonesia

Email: <sup>1,\*</sup>cahyani.cbh@nusamandiri.ac.id, <sup>2</sup>cyrill.ilario2@gmail.com, <sup>3</sup>alfia.zahra99@gmail.com,

<sup>4</sup>muhammadadhi2803@gmail.com, <sup>5</sup>erinurs2604@gmail.com, <sup>6</sup>cahyadi.am@gmail.com, <sup>7</sup>rikypupratama97@gmail.com

Email Penulis Korespondensi: cahyani.cbh@nusamandiri.ac.id

Submitted: 03/07/2024; Accepted: 31/07/2024; Published: 31/07/2024

**Abstrak**–Netflix telah menjadi ikon utama dalam perkembangan teknologi di dunia hiburan. Kemajuan dalam teknologi informasi memperkenalkan kita untuk mempermudah melakukan akses ribuan judul film dan acara TV. Netflix adalah layanan berlangganan yang dapat mempermudah pengguna menonton film dan acara TV tanpa iklan dalam perangkat yang terhubung ke Internet. Dengan semakin banyaknya judul film dan acara TV membuat penonton kesulitan dalam menentukan pilihan film yang akan ditonton. Metode K-Means Clustering digunakan untuk menganalisis data dari berbagai kelompok preferensi film di Netflix. Hasil analisis ini dapat membantu pengguna memilih rekomendasi film berdasarkan rating dan genre yang mereka inginkan. Dengan cara ini, pengguna akan lebih mudah dalam membuat pilihan untuk menonton film dan acara TV. Dataset yang diambil berasal dari kaggle sebanyak 9957 data. Hasil dari penelitian dengan menggunakan metode K-means Clustering dapat disimpulkan bahwa terdapat 3 cluster, dimana Cluster 0 merupakan film – film short movie hingga film – film layar lebar dengan rata – rata bergenre action 33% dan comedy 20%, cluster 1 merupakan film – film layar lebar hingga film biografi atau documenter dengan rata – rata bergenre action 37% dan drama 19%, dan cluster 2 merupakan film – film series atau short movie dengan rata – rata bergenre drama 32%, crime 24% dan action 22%.

**Kata Kunci:** Data Mining, K-Means Clustering, Netflix, Film, RapidMiner

**Abstract**–Netflix has become a major icon in technological developments in the world of entertainment. Advances in information technology allow us to make it easier to access thousands of film and TV show titles. Netflix is a subscription service that makes it easier for users to watch movies and TV shows without ads on devices connected to the Internet. With the increasing number of film and TV show titles, it makes it difficult for viewers to decide which film to watch. The K-Means Clustering method is used to analyze data from various film preference groups on Netflix. The results of this analysis can help users choose film recommendations based on the rating and genre they want. In this way, it will be easier for users to make choices for watching movies and TV shows. The dataset taken came from Kaggle with 9957 data. The results of research using the K-means Clustering method can be concluded that there are 3 clusters, where Cluster 0 is short films to feature films with an average of 33% action genre and 20% comedy, cluster 1 is films - feature films to biographical or documentary films with an average of 37% action genre and 19% drama, and cluster 2 is series or short films with an average of 32% drama genre, 24% crime and 22% action.

**Keywords:** Data Mining, K-Means Clustering, Netflix, Film, RapidMiner

## 1. PENDAHULUAN

Perkembangan teknologi informasi sudah merubah secara drastis kaidah kita menikmati hiburan. Layanan streaming seperti Netflix kini menjelma sebagai lambang utama peralihan ini. Kemajuan dalam teknologi informasi mempermudah masyarakat untuk melakukan akses ribuan acara TV dan film hanya dalam rumah dengan nyaman. Melalui era revolusi digital, Netflix dan platform serupa sudah menjadi sumber utama untuk memenuhi kebutuhan hiburan kita.

Netflix merupakan suatu contoh platform streaming video terbesar di dunia, yang dibangun di Amerika Serikat sejak tahun 1997. Awal mula, platform ini fokus pada penyewaan DVD dengan pengiriman melalui pos. Akan tetapi, sejak tahun 2007, Netflix mengalami pertumbuhan yang cepat dan melapangkan layanannya menuju streaming video online. Netflix adalah layanan berlangganan yang memungkinkan anggotanya menonton film dan acara TV tanpa iklan melalui perangkat yang terhubung ke Internet. Dengan Netflix, pengguna bisa melihat tayangan favorit mereka di mana saja dan kapanpun mereka mau melalui smartphone, smart TV, tablet, PC, dan laptop [1].

Keterlibatan konsumen yang bertambah besar, bersamaan dengan hadirnya tantangan baru untuk memberikan konten yang sesuai dengan kemauan mereka, mendorong kebutuhan akan analisis data yang lebih mendalam. Penelitian ini mengungkapkan bahwa persepsi tentang popularitas film dapat bervariasi secara subjektif tergantung pada selera individu, kelompok, dan bagaimana data film dikelola di layanan Netflix. Dengan semakin banyaknya judul film dan acara TV membuat penonton kesulitan dalam menentukan pilihan film yang akan ditonton. Oleh karena itu diperlukan pengelompokkan film ke dalam beberapa cluster agar dapat menjadi rekomendasi kepada para pengguna platform Netflix.

Data Mining adalah proses analisis suatu data guna mendeteksi pola – pola pada sekumpulan data. Data mining adalah cabang atas kecerdasan buatan (Artificial Intelligence). Pada data mining, ada banyak metode yang bisa digunakan sebagai alternatif untuk menyelesaikan masalah, tergantung pada pemanfaatannya, termasuk prediksi, klarifikasi, klastering, dan estimasi [2].

Dalam data mining, clustering adalah teknik yang digunakan untuk menganalisis data dengan tujuan membagi data menjadi beberapa kelompok. Proses clustering ini mengelompokkan sekumpulan data ke pada kelompok-kelompok yang mempunyai kesamaan di dalam satu kelompok, sementara data yang tidak memiliki kesamaan dimasukkan ke dalam kelompok yang berbeda [3].

Algoritma K-means kerap dipilih untuk pengelompokkan data karena kemudahan dan efisiensinya dalam mengelompokkan data ke dalam cluster-cluster yang signifikan. Algoritma ini diakui oleh IEEE sebagai salah satu algoritma data mining terkemuka karena keunggulannya dalam menganalisis data secara efektif [4].

Dalam penelitian ini, metode data mining yang akan dipakai untuk menyelesaikan masalah adalah metode KDD (Knowledge Data Discovery). KDD adalah pendekatan penting untuk menggali informasi baru yang tersembunyi dalam data mentah. Ini merupakan proses sistematis yang bertujuan untuk menemukan pola-pola berharga atau pengetahuan baru dari data yang tersedia [5].

Tujuan dari penelitian ini adalah melakukan analisa rekomendasi film populer yang banyak disukai oleh pengguna dengan metode clustering.

Pada penelitian terdahulu yang dilakukan oleh Ramadhan pada tahun 2022, dimana penelitian yang mengangkat dengan tema Rekomendasi film Netflix menggunakan k-means clustering dengan collaborative filtering dan collaborative filtering, hasil pengujian dari penelitian ini menunjukkan bahwa k-means clustering dengan collaborative filtering mendapatkan hasil terbaik dengan prediksi terbaik sebesar 2.8466, mendapatkan nilai MAE sebesar 0.5029, dan nilai RMSE sebesar 0.6354[6].

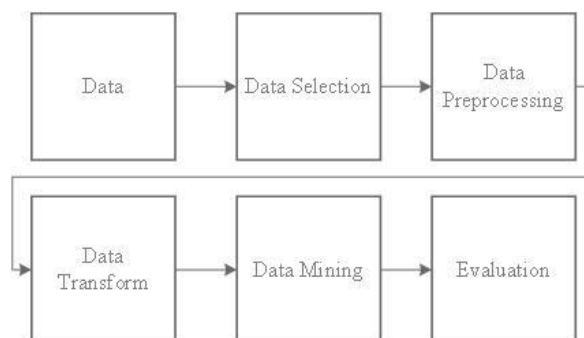
Penelitian mengenai rekomendasi film pada platform Netflix juga dilakukan oleh Mukhsinin dkk pada tahun 2024, dimana pada penelitian tersebut menerapkan metode Decision Tree, penelitian menggunakan dua dataset utama, movies\_df dan ratings\_df, penelitian melibatkan langkah-langkah pengumpulan dan penggabungan data, penentuan fitur dan variabel target, pembagian data, pelatihan model, serta evaluasi. penentuan fitur dan variabel target, pembagian data, pelatihan model, serta evaluasi. Hasilnya mencakup evaluasi model Decision Tree dengan metrik akurasi, precision, recall, dan F1-score untuk setiap kategori rating, serta visualisasi grafik batang tentang jumlah rating film dan presentase rating dari 1-5 [7].

Penelitian mengenai Rekomendasi film juga dilakukan oleh Billah dkk, dimana penelitian ini bertujuan untuk mengembangkan sistem rekomendasi film menggunakan kombinasi dari Collaborative Filtering, PCA, dan K-Means. Metode PCA diterapkan pada data agar waktu yang dibutuhkan saat proses clustering lebih cepat. Rata-rata kompleksitas waktu yang dihasilkan adalah 1.061282 [8]

## 2. METODOLOGI PENELITIAN

### 2.1 Tahapan Penelitian

Tahapan penelitian merupakan gambaran alur proses yang dilakukan pada penelitian. Adapun tahapan penelitian dapat dilihat berikut:



**Gambar 1.** Tahapan Penelitian

Pada gambar 1 merupakan langkah – langkah dalam tahapan KDD (Knowledge Data Discovery) yang digunakan dalam penelitian, yakni sebagai berikut :

1. Data

Langkah pertama adalah mengumpulkan data melalui bermacam sumber yang valid dengan tujuan penelitian atau proses bisnis yang hendak dilakukan. Dataset ini didapatkan melalui sumber data Kaggle, dengan total 9957 dataset yang terkumpul.

2. Data Selection



Menentukan juga menetapkan data mana yang mau dipakai pada proses klastering dan proses KDD. Data selection adalah pengambilan data dari sebuah dataset yang nantinya akan diolah [9].

3. **Data Preprocessing**  
Bersihkan dan ubah data agar siap untuk dianalisis [10]. Mempersiapkan dan membersihkan data untuk proses pengklasteran dan analisis.
4. **Data Transformation**  
Transformasi data sesuai dengan format yang dibutuhkan adalah Langkah penting dalam pengolahan data mining, dimana data yang tidak sesuai dengan format akan disingkirkan dan menyisakan data yang sesuai untuk digunakan [11].
5. **Data Mining**  
Tahapan utama dari KDD ialah menggunakan teknik dan algoritma data mining untuk mengidentifikasi pola, hubungan, atau informasi berharga dari data yang telah dipersiapkan sebelumnya.
6. **Evaluation**  
Tahap evaluasi ini menghasilkan model dan pola melalui teknik data mining, yang selanjutnya dievaluasi untuk menilai sejauh mana tujuan telah tercapai.

## 2.2 Metode Penelitian

Metode penelitian yang dipakai pada penelitian ini adalah jenis penelitian sekunder dimana sumber data yang tidak langsung memberikan data kepada pengumpul data [12]. Dataset yang digunakan pada penelitian ini merupakan jenis public dataset dimana dataset diperoleh dari kaggle.

## 2.3 Tinjauan Pustaka

1. **Data Mining**  
Data mining ialah proses penggunaan teknik-teknik pembelajaran mesin untuk secara otomatis menganalisis dan mengekstraksi pengetahuan dari data [13]. Perkembangan data dimulai pada tahun 1960-an dengan sistem pengelolaan data yang masih primitif. Pada tahun 1970-an, pengelolaan data menjadi lebih terstruktur dengan adopsi struktur tabel relasional. Pada periode ini, penggunaan query, antarmuka pengguna (user interface), online transaction processing (OLTP), serta penerimaan, pengelolaan, dan penyimpanan data telah mengadopsi teknologi yang lebih canggih [14]. Walaupun berakar pada pembelajaran mesin dan statistika, data mining juga meluas ke beragam bidang ilmu komputer dan disiplin lainnya seperti jaringan, lingkungan, biologi, keuangan, dan bidang lainnya [15]. Oleh karena itu, data mining mencakup berbagai disiplin ilmu seperti pembelajaran mesin, kecerdasan buatan, statistik, dan manajemen basis data [16]. Karakteristik data mining, yaitu sebagai berikut [17]:
  - a. Data mining melibatkan penemuan informasi yang tersembunyi dan pola-pola data yang data yang sebelumnya tidak dikenali.
  - b. Data mining umumnya menggunakan data dalam skala besar untuk menghasilkan hasil yang lebih dapat diandalkan.
  - c. Data mining bermanfaat untuk membuat keputusan yang penting, terutama dalam strategi.
2. **RapidMiner**  
RapidMiner adalah platform perangkat lunak yang kuat untuk ilmu data dan pembelajaran mesin. Ini menyediakan beragam alat untuk persiapan data, pemodelan, evaluasi, dan implementasi [18]. RapidMiner menerima penghargaan dari Rexer Analytics Data Miner Survey sebagai salah satu perangkat lunak dengan rating yang paling memuaskan. Di dalamnya terdapat aplikasi yang berdiri sendiri dalam analisis data dan sebagai mesin data mining yang mencakup proses seperti perubahan data, memuat data, visualisasi data, dan metode permodelan data [19]. Rapid Miner ini memiliki antarmuka GUI (Graphical User Interface) yang mempermudah penggunaannya. Rapid Miner merupakan software sumber terbuka yang ditulis menggunakan bahasa pemrograman Java dan dilisensikan di bawah GNU Public License, sehingga dapat digunakan pada berbagai sistem operasi [15].
3. **Clustering**  
Clustering adalah proses memilah informasi dalam suatu kumpulan menjadi sejumlah kelompok di mana tingkat kesamaan informasinya pada suatu kelompok lebih tinggi dari kesamaan informasi tersebut atas informasi pada kelompok lain [20]. Tujuan clustering yakni mengelompokkan sejumlah objek atau data ke dalam kelompok atau cluster sedemikian rupa sehingga setiap cluster memuat data yang serupa atau memiliki kesamaan sebanyak mungkin [21]. Tujuan utamanya adalah agar titik-titik yang berada dalam satu cluster (intracluster) lebih mirip satu sama lain daripada dengan titik-titik dari cluster lain. Sebaliknya, titik data dari cluster yang berbeda (intercluster) memiliki perbedaan yang lebih jelas dibandingkan dengan titik data di dalam cluster yang sama [22]. Dengan metode ini, kita dapat mengidentifikasi apakah ada perbedaan yang signifikan antara cluster yang terbentuk, menjelajahi profil dan karakteristik dari setiap cluster yang dihasilkan, serta mengevaluasi hubungan antara masing-masing objek dalam setiap cluster [18].
4. **K-Means**  
Banyak aplikasi dalam Data Mining memanfaatkan algoritma K-Means karena algoritma klastering ini sering digunakan dan populer, meskipun terdapat beberapa algoritma lain yang juga dipakai. K-Means adalah



algoritma clustering yang bersifat iteratif. Tujuannya adalah guna mengelompokkan data ke dalam sejumlah kluster yang sudah dipilih sebelumnya [23]. Algoritma ini mulai dengan menetapkan nilai awal kluster (K) secara acak sebagai centroid, yaitu titik pusat dari setiap kluster. Selanjutnya, algoritma menggunakan rumus Euclidean untuk menghitung jarak setiap data ke masing-masing centroid guna menentukan centroid terdekat. Selanjutnya, data dikelompokkan ke dalam kluster yang mempunyai centroid terdekat dengan masing-masing data. Proses ini diulang sampai nilai centroid stabil, menunjukkan bahwa tidak ada lagi perubahan dalam klasifikasi data [7]. K- Means hanya dapat mengolah data yang memiliki atribut numerik [22]. Langkah-langkah pengelompokan data adalah [24][25]:

- a. Pilih jumlah kluster
- b. Inisialisasi awal dan pusat kluster dilakukan secara random.
- c. Setiap data ditempatkan ke pusat kluster terdekat berdasarkan jarak antar obyek. Pada tahap ini jarak dihitung dengan menentukan kemiripan atau ketidakmiripan data dengan metode jarak Euclidean (Euclidean Distance) dengan rumus seperti pada persamaan 1 :

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

dimana :

d(x, y)= ukuran ketidakmiripan

x<sub>i</sub> = (x<sub>1</sub>, x<sub>2</sub>, ..., x<sub>i</sub> ) yaitu variabel data

y<sub>i</sub>= (y<sub>1</sub>, y<sub>2</sub>, ..., y<sub>i</sub>) yaitu variabel pada titik pusat.

- d. Hitung pusat kluster yang baru dengan keanggotaan yang baru dengan cara menghitung rata-rata obyek pada kluster. Penghitungan bisa juga dengan menggunakan median.
  - e. Hitung kembali jarak tiap objek dengan pusat kluster yang baru, hingga kluster tidak berubah, maka proses pengklasteran selesai.
5. Netflix

Netflix ialah sebuah platform streaming video yang memungkinkan pengguna menonton acara televisi, film, dan macam-macam program lainnya secara online melalui internet. Layanan ini menggunakan teknologi streaming yang memungkinkan pengguna untuk menonton video secara langsung tanpa perlu menanti proses pengunduhan. Netflix telah tumbuh sebagai suatu platform streaming terbesar di dunia dengan anggota dari berbagai negara dan jumlah keanggotaan mencapai jutaan. Sebagai hasilnya, Netflix berkomitmen untuk meningkatkan terus kualitas layanan dan produk mereka, yang bisa diakses secara fleksibel dan mudah melalui platform inovatif yang populer di seluruh dunia [19]. Bukan hanya menayangkan film atau series saja, namun juga memproduksi sendiri film atau series dengan tim produksi mereka [26].

### 3. HASIL DAN PEMBAHASAN

#### 3.1 Data

Langkah pertama ialah mengumpulkan data melalui bermacam sumber yang valid dengan tujuan penelitian atau proses bisnis yang ingin dilancarkan. Dataset ini diperoleh dari Kaggle dengan total sebanyak 9957 data. Data awal pada proses data mining memiliki atribut seperti “Title, Year, Certificate, Duration, Genre, Rating, Description, Stars, Votes”. Atribut ini akan mempengaruhi proses data mining serta membantu mengurangi informasi yang tidak relevan atau noise, sehingga memudahkan interpretasi hasil dan pengambilan keputusan.

Tabel 1. Dataset

Title	Year	Certificate	Duration	Genre	Rating	Description	Stars	Votes
Cobra Kai	(2018-2021)	TV-14	30 min	Action, Comedy, Drama	8.5	Decades after their 1984 All Valley Karate Tournament, a middle-aged Daniel LaRusso and Johnny Lawrence find themselves martial-arts rivals again.	['Ralph Macchio, 'William Zabka, 'Courtney Henggeler, 'Xolo Maridueña]	177,031



Title	Year	Certificate	Duration	Genre	Rating	Description	Stars	Votes
The Crown	(2016â€“)	TV-MA	58 min	Biography, Drama, History	8.7	Follows the political rivalries and romance of Queen Elizabeth II's reign and the events that shaped the second half of the twentieth century.	['Claire Foy, ', 'Olivia Colman, ', 'Imelda Staunton, ', 'Matt Smith']	199,885
Better Call Saul	(2015â€“2022)	TV-MA	46 min	Crime, Drama	8.9	The trials and tribulations of criminal lawyer Jimmy McGill before his fateful run-in with Walter White and Jesse Pinkman.	['Bob Odenkirk, ', 'Rhea Seehorn, ', 'Jonathan Banks, ', 'Patrick Fabian']	501,384
Devil in Ohio	-2022	TV-MA	356 min	Drama, Horror, Mystery	5.9	When a psychiatrist shelters a mysterious cult escapee, her world is turned upside down as the girl's arrival threatens to tear her own family apart.	['Emily Deschanel, ', 'Sam Jaeger, ', 'Gerardo Celasco, ', 'Madeleine Arthur']	9,773
Cyberpunk: Edgerunners	(2022â€“)	TV-MA	24 min	Animation, Action, Adventure	8.6	A Street Kid trying to survive in a technology and body modification-obsessed city of the future. Having everything to lose, he chooses to stay alive by becoming an Edgerunner, a Mercenary outlaw also known as a Cyberpunk.	['Zach Aguilar, ', 'Kenichiro Ohashi, ', 'Emi Lo, ', 'Aoi Yã»ki']	15,413
The Sandman	(2022â€“)	TV-MA	45 min	Drama, Fantasy, Horror	7.8	Upon escaping after decades of imprisonment by a mortal wizard,	['Tom Sturridge, ', 'Boyd Holbrook, ', 'Patton Oswalt, ',	116,358



Title	Year	Certificate	Duration	Genre	Rating	Description	Stars	Votes
Rick and Morty	(2013â€“)	TV-MA	23 min	Animation, Adventure, Comedy	9.2	Dream, the personification of dreams, sets about to reclaim his lost equipment. An animated series that follows the exploits of a super scientist and his not-so-bright grandson. A high school chemistry teacher diagnosed with inoperable lung cancer turns to manufacturing and selling methamphetamine in order to secure his family's future.	'Vivienne Acheampong'] ['Justin Roiland, ', 'Chris Parnell, ', 'Spencer Grammer, ', 'Sarah Chalke']	502,16
Breaking Bad	(2008â€“2013)	TV-MA	49 min	Crime, Drama, Thriller	9.5	After an experimental gene therapy turns them into monsters, three twenty-somethings band together to hunt down the scientist responsible and force him to make them human again.	['Bryan Cranston, ', 'Aaron Paul, ', 'Anna Gunn, ', 'Betsy Brandt']	1,831,340
The Imperfects	(2022â€“)	TV-MA	45 min	Action, Adventure, Drama	6.3		['Morgan Taylor Campbell, ', 'Italia Ricci, ', 'Rhianna Jagpal, ', 'Itachi Godoy']	3,123

### 3.2 Data Selection

Tahapan kedua adalah memilah data yang kosong, terisi sebagian, dan yang tidak diperlukan. Dari 9957 data kami membutuhkan 500 sample maka dari itu dari 9957 data diolah menggunakan pengambilan sampel acak sederhana sehingga mendapatkan hasil 500 data sesuai yang ingin digunakan dalam penelitian.

### 3.3 Data Preprocessing

Tahapan Ketiga adalah memilah atribut yang digunakan dalam algoritma K – Means Clustering. Dari data set yang telah diperoleh maka untuk memenuhi clustering atribut yang akan digunakan “Title, Year, Duration, Rating, Votes, Genre”, dapat dilihat pada tabel 2 data Preprocessing

Tabel 2. Data Preprocessing

Title	Year	Duration	Rating	Votes	Genre
A Boy Called Christmas	2021	106	67	21273	Adventure



Title	Year	Duration	Rating	Votes	Genre
A Knight's Tale	2001	132	69	186888	Action
A Man Apart	2003	109	60	47628	Action
A Prayer Before Dawn	2017	116	68	23669	Action
A Very Long Engagement	2004	133	76	73958	Drama
...	...	...	...	...	...
Young Royals	2021	50	83	30076	Drama
Your Name.	2016	106	84	262352	Animation
You're Next	2011	95	65	99371	Horror
Zero Dark Thirty	2012	157	74	299829	Drama
Zodiac	2007	157	77	536068	Crime

**3.4 Data Transformation**

Tahapan keempat ini kami mentransformasi isi dari atribut tittle, year, votes dan genre dengan kriteria sebagai berikut.

1. Pada tabel 3, merupakan transformasi data yang diambil dari Judul film yang disusun sesuai dengan cara berurutan dengan cara ascending. Pada tabel 3, dapat dilihat terdapat 500 data film yang akan digunakan dalam penelitian ini.

**Tabel 3.** Data Judul Film

Title	Tranformasi Title
A Boy Called Christmas	1
A Knight's Tale	2
A Man Apart	3
A Prayer Before Dawn	4
A Very Long Engagement	5
...	...
You're Next	498
Zero Dark Thirty	499
Zodiac	500

2. Pada tabel 4, merupakan transformasi data yang diambil dari tahun film yang disusun sesuai dengan cara berurutan dengan cara ascending yang dimulai dari tahun terkecil sampai terbesar. Pada tabel 4 terbagi menjadi 4 kategori tahun.

**Tabel 4.** Data Tahun Film

Year	Tranformasi Year
2000 - 2006	1
2007 - 2012	2
2013 - 2018	3
2019 - 2024	4

3. Pada tabel 5, merupakan transformasi data yang diambil dari votes film, dimana terbagi menjadi 3 kategori, yaitu: 1 untuk votes <100.000, 2 untuk votes >100.000 & <1.000.000, dan 3 untuk votes >1.000.000.

**Tabel 5.** Data Vote

Votes	Tranformasi Votes
<100.000	1
>100.000 & <1.000.000	2
>1.000.000	3

4. Pada tabel 6, merupakan transformasi data yang diambil dari Genre film yang terbagi menjadi 14 kategori.

**Tabel 6.** Data Genre Film

Genre	Tranformasi Genre
Action	1
Adventure	2
Animation	3
Biography	4
Comedy	5
Crime	6
Documentary	7

Genre	Tranformasi Genre
Drama	8
Fantasy	9
Game-Show	10
Horror	11
Mystery	12
Sci-Fi	13
Thriller	14

5. Setelah mentranformasi isi atribut tittle, year, votes dan genre, maka didapatkan hasil data set seperti berikut supaya dapat diolah oleh algoritma K – Means Clustering, yang dapat dilihat pada tabel 7.

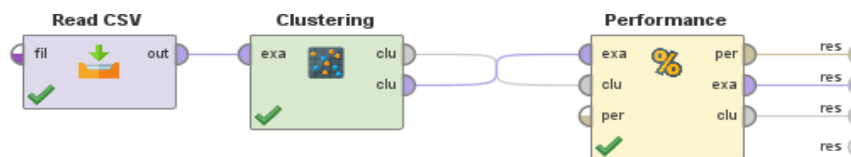
**Tabel 7.** Transform Dataset

Title	Year	Duration	Rating	Genre	Votes
1	4	106	67	1	2
2	1	132	69	2	1
3	1	109	60	1	1
4	3	116	68	1	1
5	1	133	76	1	8
...	...	...	...	...	...
496	4	50	83	1	8
497	3	106	84	2	3
498	2	95	65	1	11
499	3	157	74	2	8
500	2	157	77	2	6

Pada tabel 7, merupakan bentuk transform dataset dengan atribut tittle, year, duration, rating, genre, dan votes.

### 3.5 Data Mining

Pada tahap kelima ini, kita akan mengeksekusi data memakai algoritma clustering K-Means dengan bantuan software RapidMiner. Hal pertama yang dikerjakan pada algoritma K-means adalah menentukan jumlah cluster (K Cluster). Lalu, menentukan centroid secara acak disesuaikan dengan jumlah K cluster yang sudah ditentukan. Dalam penelitian ini, nilai K cluster yang digunakan ialah 3 K cluster sehingga nantinya akan dikelompokkan sebanyak 3 cluster. Berikut pengolahan dataset menggunakan rapidminer dengan algoritma K – Means Clustering, yang dapat dilihat pada gambar 2.



**Gambar 2.** Pengolahan Dataset.

Pada gambar 2 merupakan pengolahan dataset menggunakan rapidminer.

### 3.6 Evaluation

Tahap keenam ini merupakan hasil dari data mining yang telah dilakukan menggunakan rapidminer dengan algoritma K – Means Clustering.

## Cluster Model

```
Cluster 0: 316 items
Cluster 1: 134 items
Cluster 2: 50 items
Total number of items: 500
```

**Gambar 3.** Model Cluster

Gambar 3, merupakan hasil pengolahan data set menggunakan Rapidminer, dimana pengolahan menghasilkan 3(tiga) cluster model, Cluster 0 sebanyak 316 items, Cluster 1 sebanyak 134 items dan Cluster 2 sebanyak 50 items.

**Tabel 8.** Model Cluster

Attribute	Cluster 0	Cluster 1	Cluster 2
Year	3.209	3.200	3.480

Attribute	Cluster 0	Cluster 1	Cluster 2
Duration	102.620	137.358	50.500
Rating	63.351	71.097	74.940
Genre	1.285	1.485	1.160
Votes	4.044	4.112	5.260

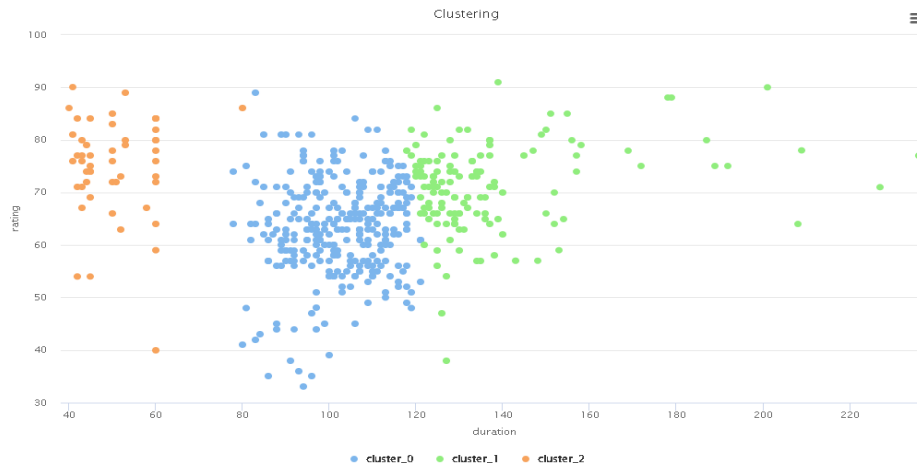
Pada Tabel 8, merupakan model cluster dari 5 attribute yang menghasilkan 3 cluster, yaitu cluster 0, cluster 1 dan cluster 2.

### PerformanceVector

```
PerformanceVector:
Avg. within centroid distance: 56.357
Avg. within centroid distance_cluster_0: 38.622
Avg. within centroid distance_cluster_1: 106.838
Avg. within centroid distance_cluster_2: 33.153
```

**Gambar 4.** Performance Vector

Pada gambar 4, merupakan hasil Performance vector hasil olah data menggunakan Rapidminer.



**Gambar 5.** Hasil Visual dari Clustering RapidMiner

Gambar 5 merupakan hasil dari pengelolaan menggunakan rapidminer, yang terbagi menjadi 3 cluster, yaitu:

- Cluster 0 (Biru) adalah film-film yang berdurasi dari 78 menit hingga 121 menit, rating dari 33 hingga 89. Berikut list film dari cluster\_0.

**Tabel 9.** Cluster\_0

Title	Year	Duration	Rating	Votes	Genre	Cluster
American Psycho	2000	102	76	598117	Crime	cluster_0
The Emperor's New Groove	2000	78	74	203092	Animation	cluster_0
Jurassic Park III	2001	92	59	321937	Action	cluster_0
Legally Blonde	2001	96	64	216974	Comedy	cluster_0
Serendipity	2001	90	68	113554	Comedy	cluster_0
...	...	...	...	...	...	...
The Man from Toronto	2022	110	58	43867	Action	cluster_0
The Sea Beast	2022	115	71	36407	Animation	cluster_0
The Tinder Swindler	2022	114	71	64227	Documentary	cluster_0
The Weekend Away	2022	89	56	23932	Crime	cluster_0
Windfall	2022	92	57	22612	Crime	cluster_0

- Cluster 1 (Hijau) adalah film-film yang berdurasi dari 118 menit hingga 236 menit, rating dari 38 hingga 91. Berikut list film dari cluster\_1.

**Tabel 10.** Cluster\_1

Title	Year	Duration	Rating	Votes	Genre	Cluster
Gladiator	2000	155	85	1481531	Action	cluster_1
A Knight's Tale	2001	132	69	186888	Action	cluster_1
Spirited Away	2001	125	86	75118	Animation	cluster_1



Title	Year	Duration	Rating	Votes	Genre	Cluster
The Lord of the Rings: The Fellowship of the Ring	2001	178	88	1844055	Action	cluster_1
Spider-Man	2002	121	74	809904	Action	cluster_1
...	...	...	...	...	...	...
Major	2022	150	82	25431	Action	cluster_1
Purple Hearts	2022	122	67	32797	Drama	cluster_1
RRR	2022	187	80	107203	Action	cluster_1
The Bubble	2022	126	47	23324	Comedy	cluster_1
The Woman in the House	2022	208	64	44195	Comedy	cluster_1

3. Cluster 2 (orange) adalah film-film yang berdurasi dari 40 menit hingga 80 menit, rating dari 40 hingga 90. Berikut list film dari cluster\_2.

Tabel 11. Cluster\_2

Title	Year	Duration	Rating	Votes	Genre	Cluster
NCIS	2003	60	78	145896	Action	cluster_2
Grey's Anatomy	2005	41	76	303617	Drama	cluster_2
Hell's Kitchen	2005	42	71	20887	Game-Show	cluster_2
NCIS: Los Angeles	2009	43	67	53571	Action	cluster_2
Blue Bloods	2010	43	77	40037	Crime	cluster_2
...	...	...	...	...	...	...
Archive 81	2022	60	73	48309	Drama	cluster_2
First Kill	2022	60	64	25843	Drama	cluster_2
Pieces of Her	2022	52	63	25208	Crime	cluster_2
Resident Evil	2022	60	40	39033	Action	cluster_2
Vikings: Valhalla	2022	51	72	37571	Action	cluster_2

#### 4. KESIMPULAN

Setelah melakukan penelitian mengenai rekomendasi film di Netflix menggunakan metode K-Means Clustering, dapat disimpulkan bahwa Cluster 0 merupakan film – film short movie hingga film – film layar lebar dengan rata – rata bergenre action 33% dan comedy 20%, cluster 1 merupakan film – film layar lebar hingga film biografi atau documenter dengan rata – rata bergenre action 37% dan drama 19%, dan cluster 2 merupakan film – film series atau short movie dengan rata – rata bergenre drama 32%, crime 24% dan action 22%. Dengan ini dapat memudahkan para pengguna supaya dapat memilih film dengan preferensi yang sedang diinginkan dan terdapat juga rating dan genre sebagai bahan pertimbangan dalam memilih. Metode K-Means Clustering efektif digunakan untuk mengelompokkan film berdasarkan preferensi pengguna. Dengan cara ini, pengguna dapat diberikan rekomendasi film yang lebih relevan dan sesuai dengan minat.

#### REFERENCES

- [1] I. Fitrianti, A. Voutama, and Y. Umaidah, "Clustering Film Populer Pada Aplikasi Netflix Dengan Menggunakan Algoritma K-Means Dan Metode CRISP-DM," *JTSI (Jurnal Teknol. Sist. Informasi)*, vol. 4, no. 2, pp. 301–311, 2023.
- [2] A. S. Lubis, T. Tugiono, and H. Hafizah, "Data Mining Estimasi Biaya Produksi Ikan Kembung Rebus Dengan Regresi Linier Berganda," *J. Sist. Inf. Triguna Dharma (JURSI TGD)*, vol. 1, no. 6, pp. 888–897, 2022, doi: 10.53513/jursi.v1i6.5732.
- [3] S. Setyaningtyas, B. I. Nugroho, and Z. Arif, "Tinjauan Pustaka Sistematis Pada Data Mining: Studi Kasus Algoritma K-Means Clustering," *J. Teknoif Tek. Inform. Inst. Teknol. Padang*, vol. 10, no. 2, pp. 52–61, 2022.
- [4] N. Suarna, N. Hidayah, and W. Prihartono, "Pengelompokan Data Film Pada Netflix Menggunakan Algoritma K-Means Clustering," *JATI (Jurnal Mhs. Tek. Inform., vol. 7, no. 6, pp. 3834–3842, 2023.*
- [5] Khoirunnisa Hamidah and A. Voutama, "Analisis Faktor Tingkat Kebahagiaan Negara Menggunakan Data World Happiness Report dengan Metode Regresi Linier," *Explor. IT J. Keilmuan dan Apl. Tek. Inform., vol. 15, no. 1, pp. 1–7, 2023, doi: 1 https://doi.org/10.35891/explorit.*
- [6] M. T. M. Ramadhan and E. B. Setiawan, "Netflix Movie Recommendation System Using Collaborative Filtering With K-Means Clustering Method on Twitter," *J. Media Inform. Budidarma*, vol. 6, no. 4, pp. 2056–2063, 2022, doi: 10.30865/mib.v6i4.4571.
- [7] 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 Indones. J. Mach. Learn. Comput. Sci., vol. 4, no. 2, pp. 570–579, 2024, doi: 10.57152/malcom.v4i2.1255.*
- [8] M. Billah, M. A. Zartesyia, and D. S. Prasvita, "Penerapan Collaborative Filtering, PCA dan K-Means dalam Pembangunan Sistem Rekomendasi Ongoing dan Upcoming Film Animasi Jepang," in *Senamika (Seminar Nasional Mahasiswa Ilmu Komputer dan Aplikasinya)*, 2021, no. April, pp. 579–587, [Online]. Available:



<https://conference.upnvj.ac.id/index.php/senamika/article/view/1318/1089>.

- [9] F. Alghifari and D. Juardi, "Penerapan Data Mining Pada Penjualan Makanan Dan Minuman Menggunakan Metode Algoritma Naïve Bayes," *J. Ilm. Inform.*, vol. 9, no. 2, pp. 75–81, 2021, doi: 10.33884/jif.v9i02.3755.
- [10] N. Novitasari, N. D. Nuris, and R. Herdiana, "Penerapan Algoritma K-Means untuk Clustering Data Jumlah Penduduk Miskin Berdasarkan Kota/Kabupaten di Jawabarot menggunakan Rapidminer," *J. Inform. Terpadu*, vol. 9, no. 1, pp. 68–73, 2023, doi: 10.54914/jit.v9i1.660.
- [11] A. Amna et al., *Data Mining, Pertama*. Padang, Sumatera Barat: PT Global Eksekutif Teknologi, 2023.
- [12] P. R. Andini, "Pengaruh Konten pada Official Akun TikTok Ruangguru terhadap Prestasi Belajar Followers," *J. Janaloka*, vol. 1, no. 1, pp. 15–24, 2023, [Online]. Available: <https://journals.usm.ac.id/index.php/janaloka/article/view/7022>.
- [13] H. Pratiwi, S. S. Handajani, and R. Respatiwilan, *Clustering dan Penerapannya Dalam Bidang Seismologi*. Yogyakarta: Deepublish, 2021.
- [14] U. Sa'adah, M. Y. Rochayani, D. W. Lestari, and D. A. Lusia, *Kupas Tuntas Algoritma Data Mining dan Implementasi Menggunakan R*. Universitas Brawij, 2021.
- [15] B. G. Sudarsono, M. I. Leo, A. Santoso, and F. Hendrawan, "Analisis Data Mining Data Netflix Menggunakan Aplikasi Rapid Miner," *JBASE - J. Bus. Audit Inf. Syst.*, vol. 4, no. 1, pp. 13–21, 2021, doi: 10.30813/jbase.v4i1.2729.
- [16] A. Z. Siregar, "Implementasi Metode Regresi Linier Berganda Dalam Estimasi Tingkat Pendaftaran Mahasiswa Baru," *Kesatria J. Penerapan Sist. Inf. (Komputer dan Manajemen)*, vol. 2, no. 3, pp. 133–137, 2021, [Online]. Available: <https://tunasbangsa.ac.id/pkm/index.php/kesatria/article/view/73>.
- [17] J. Hutagalung, *Kombinasi K-Means Clustering dan Metode Moora*. Yogyakarta: Deepublish, 2021.
- [18] M. R. Nahjan, N. Heryana, and A. Voutama, "Implementasi Rapidminer Dengan Metode Clustering K-Means Untuk Analisa Penjualan Pada Toko OJ Cell," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 7, no. 1, pp. 101–104, 2023, doi: 10.36040/jati.v7i1.6094.
- [19] F. Zafira, B. Irawan, and A. Bahtiar, "Penerapan Data Mining Untuk Estimasi Stok Barang Dengan Metode K-Means Clustering," *JATI (Jurnal Mhs. Tek. Inform.)*, vol. 8, no. 1, pp. 156–161, 2024, doi: 10.36040/jati.v8i1.8319.
- [20] I. Nuryani and D. Darwis, "Analisis Clustering Pada Pengguna Brand Hp Menggunakan Metode K-Means," in *Prosiding Seminar Nasional Ilmu Komputer*, 2021, vol. 1, no. 1, pp. 190–211.
- [21] L. Muflikhah, D. E. P. Ratnawati, and R. R. Mardi, *Data Mining*. Malang: UB Press, 2018.
- [22] D. Jollyta, W. Ramdhan, and M. Zarlis, *Konsep Data Mining dan Penerapan*. Deepublish, 2020.
- [23] R. T. Vulandari, *Data Mining Teori dan Aplikasi Rapidminer*. Yogyakarta: Gava Media, 2017.
- [24] M. Wahyudi, M. Masitha, R. Saragih, and S. Solikhun, *Data Mining : Penerapan Algoritma K-Means Clustering dan K-Medoids Clustering, Pertama.*, vol. 01. Yayasan Kita Menulis, 2020.
- [25] S. Witanto, D. E. Ratnawati, and S. Anam, "Pengelompokan Fungsi Aktif Senyawa Data SMILES (Simplified Molecular Input Line Entry System) Menggunakan Metode K-Means Dengan Inisialisasi Pusat Klaster Menggunakan Metode Heuristic O (N LogN)," *J. Pengemb. Teknol. Inf. dan Ilmu Komput.*, vol. 3, no. 1, pp. 702–707, 2019.
- [26] I. V. O. Situmeang, "Terpaan Media Streaming Netflix Dan Kualitas Aplikasi Netflix Terhadap Keputusan Penggunaan (Survei Pada Gen Z Di Perumahan Citra Gran, Cibubur)," *IKRAITH-HUMANIORA J. Sos. dan Hum.*, vol. 6, no. 1, pp. 160–174, 2022, [Online]. Available: <https://journals.upi-yai.ac.id/index.php/ikraith-humaniora/issue/archive>.