PENGELOMPOKAN DATA FILM PADA *NETFLIX* MENGGUNAKAN ALGORITMA *K-MEANS CLUSTERING*

Nana Suarna, Nurni Hidayah, Willy Prihartono
Teknik Informatika, STMIK IKMI Cirebon
Jalan Perjuangan 10B, Karyamulya, Kesambi, Kota Cirebon, Indonesia
nurnihidayah20@gmail.com

ABSTRAK

Dalam era digital yang berkembang sangat pesat, adanya platform *streaming* telah membuka pintu bagi ribuan film dan serial TV untuk diakses oleh jutaan pelanggan di seluruh dunia. *Netflix* sebagai salah satu pemimpin dalam industri layanan streaming, telah merevolusi cara kita mengkonsumsi berbagai konten. Permasalahan dalam penelitian ini bahwa persepsi tentang "film terpopuler" dapat bervariasi secara subyektif tergantung pada preferensi individu, kelompok, atau faktor budaya serta pengelolaan data film dilayanan streaming seperti *Netflix* didalam konteks Informatika. Tujuan penelitian ini untuk mengelompokkan *dataset* film *Netflix* menggunakan algoritma *K-Means* untuk mengidentifikasi pola atau kesamaan antara film-film berdasarkan atribut-atribut tertentu serta mengevaluasi struktur dan jarak antar klaster yang dihasilkan. Analisis data film akan membantu dalam mengidentifikasi tren yang relevan. Metode penelitian ini *Knowledge Discovery in Database* (KDD) dengan Algoritma *K-Means Clustering* menggunakan *dataset* yang diperoleh dari *Kaggle* sebagai data acuan. Hasil dari penelitian ini dalam menggunakan metode KDD dan Algoritma *K-Means Clustering*, menunjukkan 2 cluster yang memiliki nilai rata-rata centroid yang berbeda. Dari kedua cluster tersebut menemukan film-film popularitas dalam *dataset* film *Netflix* pada *cluster* 0 yaitu dengan ciri-ciri ideal rata-rata atribut *Hidden Gem Score* sebesar 8.034, atribut *Runtime* sebesar 0.793 dan atribut *IMDb Score* sebesar 7.148. Dari hasil penelitian tersebut dapat membantu pengguna dalam menemukan film-film yang sesuai dengan minat dan preferensi pengguna.

Kata kunci : *Algoritma K-Means, Film, Netflix*

1. PENDAHULUAN

Dalam dekade terakhir, perkembangan teknologi informasi telah mengubah secara dramatis cara kita mengonsumsi hiburan. Layanan streaming seperti *Netflix* telah menjadi simbol utama dari perubahan ini. Kemajuan teknologi informasi telah memungkinkan kita dengan mudah mengakses ribuan judul film dan acara televise dari kenyamanan rumah kita sendiri. Dalam konteks revolusi digital ini, *Netflix* dan platform serupa telah menjadi tempat kita memuaskan kehausan akan hiburan. Dengan penggunaan yang semakin luas, *Netflix* telah menjadi salah satu pemain utama dalam industri hiburan global. Menghadirkan konten berkualitas tinggi kepada jutaan pelanggan diseluruh dunia. Hal ini menyoroti betapa pentingnya pengelolaan konten dalam menghadapi pertumbuhan pesat ini. Keterlibatan pengguna yang semakin besar, bersama dengan tantangan menyajikan konten yang sesuai dengan preferensi mereka, telah memunculkan kebutuhan akan analisis data yang lebih mendalam.

Permasalahan yang ditemukan dalam penelitian ini bahwa persepsi tentang "film terpopuler" dapat bervariasi secara subyektif tergantung pada preferensi individu, kelompok serta pengelolaan data film dilayanan streaming seperti *Netflix* didalam konteks Informatika. Pertumbuhan eksponensial dalam jumlah konten film yang tersedia di platform ini telah menciptakan tantangan besar. Salah satu permasalahan utama adalah bagaimana mengorganisasi dan mengelompokkan data film sehingga pengguna dapat dengan mudah menavigasi dan menemukan konten yang sesuai dengan preferensi mereka. Sementara *Netflix* telah melakukan upaya besar untuk mencari film popular yang berbeda, masih ada permasalahan yang lebih dalam yang perlu dipecahkan. Data film mencakup beragam atribut, seperti *Title, Hidden Gem Score, Runtime,* dan *IMDb Score*. Mengorganisasi atribut-atribut ini dalam skala besar dan memahami hubungan dengan preferensi pengguna adalah tantangan yang sangat kompleks. Dengan terus bertambahnya konten, platform seperti *Netflix* harus dapat mengelola dan menyajikan data untuk jutaan pengguna dengan cepat dan efisien. Ini mengharuskan penggunaan metode analisis data yang canggih untuk memproses data secara real-time dan memberikan hasil relevan.

Penelitian terdahulu menurut Fitrianti dengan judul Clustering Film Populer Pada Aplikasi *Netflix* Dengan Menggunakan Algoritma *K-Means* Dan Metode *CRISP- DM*, menyimpulkan dalam penelitian tersebut metode K-Means mampu untuk mengelompokkan data film *Netflix* dengan menggunakan 3 cluster yaitu terendah, tertinggi dan sedang. Dengan hasilnya, cluster 1 menjadi cluster yang ciri-ciri ideal dalam mengelompokkan film popular di Netflix yaitu dengan nilai rata-rata pada atribut rating sebesar 8.571, atribut durasi sebesar 166.036, dan atribut votes sebesar 1.157.600,464 [1].

Tujuan dari penelitian ini adalah untuk menganalisis dan memahamii hubungan antara film dengan Negara produksinya di*dataset Netflix* serta

menerapkan metode klastering pada *dataset* yang di peroleh dari *Kaggle* sebagai data pabrik. Selain tujuan utama yang telah disebutkan sebelumnya, penelitian ini juga bertujuan untuk mengeksplorasi potensi hubungan antara atribut-atribut pada *dataset* film Netflix. Selain itu, penelitian ini berupaya untuk mengidentifikasi dan mengetahui *film* atau judul-judul yang paling popular diproduksi di *Netflix*. Signifikan penelitian ini juga terletak dalam kontribusinya pada pengembangan ilmu data dan analisis data. Penggunaan algoritma *K-Means clustering* dalam konteks analisis data film adalah aplikasi yang menarik, dan hasil penelitian ini dapat membuka pintu untuk pengembangan lebih lanjut dalam analisis data dan metode klastering.

Algoritma *k-Means* sering digunakan dalam pengelompokkan data karena sifatnya yang sederhana dan efisien, serta telah diakui sebagai salah satu dari algoritma *data mining* teratas oleh IEEE karena keunggulannya dalam menganalisis data [2].

Clustering adalah proses mengelompokkan data ke dalam beberapa kelompok di mana data dalam satu kelompok memiliki kesamaan yang lebih besar dengan data lain dalam kelompok yang sama daripada dengan data dalam kelompok lain. [3].

K-Means clustering merupakan sebuah metode yang digunakan untuk mengelompokkan karakteristik objek dalam data besar dengan tujuan untuk merangkum pengolahan objek yang luas, sehingga mempermudah deskripsi sifat atau karakteristik dari setiap kelompoknya [4].

Metode penelitian ini menggunakan pendekatan *Knowledge Discovery in Database*s (KDD). *Knowledge discovery in database* (KDD) adalah proses yang mencakup pengumpulan dan penggunaan data historis dengan tujuan menemukan pola, keteraturan, atau hubungan dalam kumpulan data yang besar [5].

2. TINJAUAN PUSTAKA

2.1. Film

Film adalah produksi visual bergerak yang mengisahkan cerita atau kejadian tertentu yang bisa dinikmati melalui tayangan di bioskop atau televisi. Sebagai media komunikasi, film menggunakan elemen audio visual untuk menyampaikan pesan kepada penontonnya. Dianggap sebagai salah satu bentuk media massa, film memiliki kemampuan untuk menyampaikan banyak informasi dalam durasi waktu yang terbatas karena sifatnya yang bersifat audio visual. [6].

2.2. Netflix

Netflix merupakan platform streaming video yang paling populer secara global dan sedang menerapkan teknologi analisis data dan *machine learning* untuk meningkatkan pengalaman pengguna. [1].

2.3. Hidden Gem Score

Hidden gem adalah istilah dalam bahasa Inggris yang secara literal berarti "permata tersembunyi". Namun, istilah ini juga digunakan untuk menggambarkan sesuatu yang memiliki nilai yang tinggi, namun belum mendapatkan pengakuan luas atau tersembunyi dari perhatian publik. *Hidden gem* bisa merujuk pada suatu lokasi, karya seni, produk, atau bahkan individu yang memiliki kualitas atau nilai unik yang belum banyak diketahui oleh banyak orang. Di platform media sosial, "*hidden gem*" seringkali muncul sebagai tag atau keterangan yang diberikan oleh pengguna untuk membagikan penemuan atau rekomendasi terkait sesuatu yang dianggap memiliki nilai atau keunikannya yang masih belum banyak dikenal oleh orang lain. [7].

2.4. IMDb Score

Internet Movie Database (IMDb) telah menjadi salah satu referensi utama untuk menemukan informasi yang terkait dengan sebuah film. IMDb juga menyediakan penilaian terhadap film-film yang terdaftar di dalam basis data mereka. Peringkat film-film ini diberikan oleh para pengamat serta individu yang memiliki akun di platform tersebut. Salah satu faktor dalam menentukan popularitas suatu film adalah jumlah orang yang memberikan penilaian atau ulasan terhadap film tersebut. [8].

2.5. K-Means

Secara prinsip, *K-Means* adalah suatu teknik pengelompokan data *non-hirarkis* (sekatan) yang bertujuan untuk membagi data menjadi dua kelompok atau lebih [9]. *K-Means* adalah salah satu dari berbagai algoritma dalam bidang data mining yang dapat digunakan untuk mengelompokkan atau melakukan pengelompokan (*clustering*) terhadap data [10]. Algoritma *k-means clustering* melakukan dua tugas utama, yakni:

1. Menentukan nilai terbaik untuk titik pusat K atau centroid dengan proses iteratif (perulangan).
2. Menetapkan setiap titik data ke pusat k terdekat. Titik-titik data yang dekat dengan pusat-k tertentu, kemudian dibuatkan sebuah kluster [11].

2.6. Davies Bouldin Index (DBI)

Dalam menentukan kluster terbaik menggunakan penilaian DBI, *Davis Bouldien Index* berperan sebagai ukuran untuk meninjau dan mengevaluasi hasil dari algoritma pengelompokan. DBI yang memberikan hasil minimal di dalam suatu kluster dianggap sebagai indikator dari skema pengelompokkan yang optimal [12].

3. METODE PENELITIAN

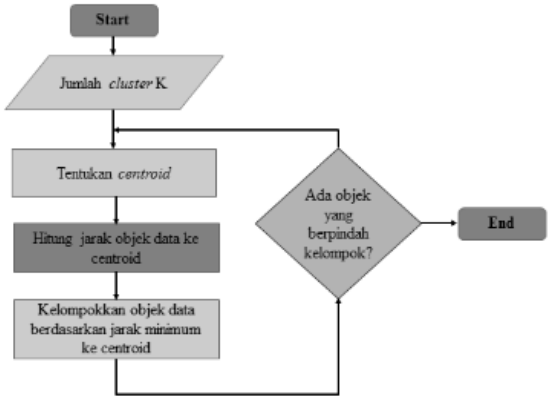
Metode yang digunakan yaitu *Knowledge Discovery in Database* (KDD) dan Algoritma *K-Means Clustering* yang bertujuan untuk menganalisis

dan mengelompokkan dari sumber data yang tersedia. Pendekatan ini memungkinkan identifikasi atribut yang relevan, klasterisasi data, serta penggunaan kriteria evaluasi yang sesuai.

3.1. *K-Means Clustering*

Metode K-Means Clustering termasuk dalam teknik partisi yang membagi objek ke dalam k klaster terpisah. Pada K-Means, setiap objek harus termasuk ke dalam klaster tertentu, tetapi pada setiap iterasi, objek dapat pindah ke klaster lain yang berbeda.

Proses pengelompokan data menggunakan algoritma k-means clustering ini dilakukan melalui serangkaian langkah-langkah seperti Gambar 1.

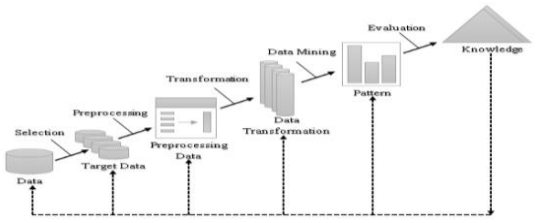


Gambar 1. Diagram Alir Proses *K-Means Clustering*[13]

Dari Gambar 1, urutan atau langkah dalam proses *K-Means Clustering* yaitu menentukan jumlah cluster, selanjutnya menentukan centroid, selanjutnya hitung jarak objek data ke centroid, selanjutnya kelompokkan objek data berdasarkan jarak minimum ke centroid, selanjutnya lakukan perulangan tersebut.

3.2. *Knowledge Discovery in Database* (KDD)

Melalui metode KDD, penelitian ini akan menghasilkan pengetahuan yang berakar dari *database* yang tersedia, membantu menjawab pertanyaan penelitian, serta mendukung pencapaian tujuan penelitian secara efektif. Tahapan KDD dapat dilihat dari Gambar 2.



Gambar 2. Tahapan *Knowledge Discovery in Database* [14].

Dari Gambar 2, dijelaskan bahwa langkah-langkah dalam tahapan KDD yaitu sebagai berikut:

1. Data
   Tahapan pertama adalah pengumpulan data dari berbagai sumber yang relevan dengan tujuan penelitian atau proses bisnis yang ingin dijalankan. *Dataset* diperoleh dari data yang bersumber dari *Kaggle*. Data yang diperoleh sebanyak 9425 *dataset* film *Netflix* pada tahun 2015-2021.
2. Data Selection
   Memilih dan menentukan data mana yang akan digunakan dalam klaster dan proses KDD. Tahapan ini berfokus menyeleksi data yang relevan dan memiliki kualitas yang memadai untuk proses selanjutnya. Menghapus atau menyeleksi data yang bersifat tidak relevan, menangani nilai yang hilang atau null, dan memastikan data dalam format sesuai.
3. Data Prepocessing
   Membersihkan dan mempersiapkan data untuk pengklasteran dan analisis. Melibatkan langkah-langkah seperti menghapus atribut atau memilih atribut yang akan digunakan.
4. Data Transformation
   Merubah data sesuai dengan format ekstensi data menjadi satu langkah krusial dalam pengolahan data mining. Hal ini diperlukan karena beberapa metode dalam data mining membutuhkan format khusus agar dapat diolah secara efektif [2].
5. Data Mining
   Tahap inti dari KDD, dimana teknik dan algoritma data mining diterapkan untuk mengidentifikasi pola, hubungan, atau informasi yang berguna dari data yang telah dipersiapkan sebelumnya.
   Algoritma *K-Means* melibatkan dua tahap utama, yaitu identifikasi lokasi pusat *cluster* dan penentuan anggota dari masing-masing *cluster*. Proses *clustering* dimulai dengan pengidentifikasian data yang akan dikelompokkan, Xij (dengan i=1 hingga n; j=1 hingga m, dimana n adalah jumlah data yang akan dikelompokkan dan m adalah jumlah variable). Pada awal iterasi, pusat dari setiap *cluster* ditetapkan secaca acak atau bebas, Ckj (dengan k=1 hingga k; j-1 hingga m). Kemudian, jarak antara setiap data dan setiap pusat *cluster* dihitung[9]. Proses penjumlahan landasan algoritma *K-Means* dapat dilihat ketentuan dibawah ini :
   a. Menentukkan jumlah *cluster* yang akan dibuat pada bentuk kelompok dan tetapkan *cluster* K.
   b. Memakai jarak *Euclidean* selanjutnya dijumlahkan setiap data ke pusat *cluster*

$$d(i,k) = \sqrt{\sum_{i}^{m} (C_i j - C_k j)^2}$$

   c. Mengelompokkan data pada *cluster* dengan jarak terpendek dengan persamaan

$$min \sum_{k}^{i} -a_i k- = \sqrt{\sum_{i}^{m} (C_i j - C_i j)^2}$$

d. Menjumlahkan pusat *cluster* yang baru menggunakan persamaan

$$C_{kj} = \frac{\sqrt{\sum_{k}^{i} X_{ij}}}{P}$$

Dengan : Xij? Kluster ke k p = banyaknya anggota kluster ke – k. Ulangi langkah kedua sampai dengan seterusnya sehingga sudah tidak ada lagi data yang berpindah ke kluster yang lain.

6. Evaluation

Tahap evaluasi ini menghasilkan pola dan model dari teknik data mining, yang kemudian dievaluasi untuk menilai pencapaian tujuan. Jika hasilnya tidak sesuai, beberapa opsi dapat dipertimbangkan seperti menggunakan hasil sebagai umpan balik untuk memperbaiki proses data mining lainnya atau menerima hasilnya sebagai temuan yang tak terduga yang mungkin bermanfaat. Alur informasi yang dihasilkan dari data mining perlu ditampilkan dengan jelas, mudah dimengerti oleh pihak yang berkepentingan, dan melibatkan pemeriksaan untuk memeastikan kesesuaian dengan fakta atau hipotesis yang telah dievaluasi sebelumnya dalam proses *Knowledge Discovery in Database* (KDD).

Dalam penelitian ini, analisis data film pada *Netflix* menggunakan algoritma *K-Means* diimplementasikan melalui perangkat lunak *RapidMiner* versi 10.2. Penentuan jumlah *cluster* (K) dalam algoritma ini dipilih secara acak tanpa melebihi jumlah data yang ada. Eksperimen dilakukan dengan mengujii pengelompokkan sebanyak 9 kali, mulai dari K = 2 hingga K = 10. Tujuan dari pengujian ini adalah untuk menemukan *cluster* terbaik yang dapat ditemukan, yang tercermin dari nimai DBI (*Davies-Bouldin Index*) yang terendah.

4. HASIL DAN PEMBAHASAN

Hasil penelitian yang dilakukan dalam pembahasan ini yaitu akan menguraikan proses bagaimana pengelompokkan dataset film Netflix dengan proses pengujian menggunakan RapidMiner versi 10.2.

4.1. *Data*

Dataset yang diperoleh dari Kaggle terdiri dari 29 atribut dengan jumlah data 9425 dataset dari tahun 2015-2021. Kemudian dalam tahap data mining terseleksi menjadi 4 atribut dengan jumlah data 1401 dataset dan dalam rentang waktu 2019-2021. Data tersebut diolah menggunakan aplikasi RapidMiner. Atribut yang digunakan dijelaskan pada Tabel 1.

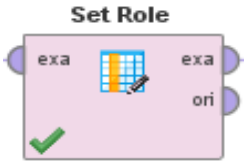
Tabel 1. Penjelasan Atribut Data Film *Netflix*

No.	Atribut	Type Data	Keterangan
1	*Title*	*Polynominal*	Judul dari film
2	*Hidden Gem Score*	*Real*	Sebuah nilai yang mungkin seberapa terembunyi atau tidak terkenalnya sebuah konten
3	*Runtime*	*Polynominal*	Durasi atau panjang konten media
4	*IMDb Score*	*Real*	Skor atau penilaian yang diberikan oleh pengguna di situs IMDb

Dari Tabel 1 telah dijelaskan atribut-atribut yang akan digunakan dalam penelitian ini.

4.2. *Data Selection*

Tahapan *selection* digunakan untuk menyeleksi atau memilih data yang akan diolah. Data yang diolah yaitu data film *Netflix*. Tahapan ini dilakukan untuk memilih data yang akan diproses di *RapidMiner*, tahapan ini dilakukan di Microsoft *Excel* dan di *RapidMiner*. Pada Microsoft *Excel* menghapus data yang bernilai null, memilih tahun dari 2019-2021. Pada *RapidMiner* menggunakan operator *Set Role* untuk menentukan id pada *dataset*, dapat dilihat pada Gambar 3.



Gambar 3. Operator *Set Role* Pada *RapidMiner*

Parameter pada operator *Set Role* yang digunakan dapat dilihat pada Gambar 2 dengan menggunakan atribut *name* yaitu *Title* menjadi target *role* yaitu id dapat dilihat pada Tabel 2 berikut.

Tabel 2. *Parameter* operator *Set Role*

No.	*Parameter*	Isi
1.	*attribute name*	*Title*
2.	target *role*	Id

Dari hasil pembacaan operator *Set Role* didapat informasi pada Tabel 3 berikut.

Tabel 3. Hasil Pembacaan Operator *Set Role*

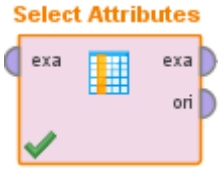
No.	Uraian	Keterangan
1	*Record*	1401
2	*Special Attribute*	1
3	*Reguler Atribute*	13
4	*Attribute :*	
	Genre	*Polynominal*
	Tags	*Polynominal*
	Languages	*Polynominal*
	Series or Movie	*Polynominal*
	Hidden Gem Score	*Real*
	Country Availability	*Polynominal*
	Runtime	*Polynominal*

No.	Uraian	Keterangan
	Actors	*Polynominal*
	View Rating	*Polynominal*
	IMDb Score	*Real*
	Netflix Release Date	*Date*
	Summary	*Polynominal*
	IMDb Votes	*Integer*

Karena *dataset* sebelum dimasukkan ke *RapidMiner* sudah terseleksi beberapa di *Excel* maka yang tadinya 29 atribut menjadi 14 atribut saat dimasukkan *RapidMiner*, dikarenakan mempunyai nilai yang sama dan nilai *null*.

4.3. *Data Preprocessing*

Sebelum melaksanakan *Data Mining*, langkah awal yang penting adalah menjalani serangkaian proses, termasuk membersihkan duplikasi data, memeriksa inkonsistensi dalam data, dan memperbaiki kesalahan yang mungkin muncul, seperti kesalahan cetak (*tipografi*). Selain itu, proses *enrichment* data dilakukan untuk memperkaya data yang telah ada dengan menambahkan informasi tambahan yang relevan dan diperlukan dalam Kontak Penemuan Pengetahuan (KDD), termasuk data atau *informs* eksternal yang dapat meningkatkan kualitas dan keberagaman data [3]. Dalam tahap ini diperoleh data sebanyak 1401 *record* pada tahun 2019-2021 dengan 4 atribut yang akan digunakan yaitu *Title*, *Hidden Gem Score*, *Runtime* dan *IMDb Score*. Berikut adalah proses *processing* pada *RapidMiner* menggunakan operator *Select Attributes* dapat dilihat pada Gambar 4 berikut.



Gambar 4. Operator *Select Attributes* pada *RapidMiner*

Operator *Select Attributes* dalam *RapidMiner* digunakan untuk memilih, menghapus, atau mengatur atribut dalam *dataset*. Ini memungkinkan mengubah tipe data, dan melakukan transformasi lainnya pada atribut sebelum proses analisis data. *Parameter* yang digunakan pada operator *Select Attributes* dapat dilihat pada Tabel 4 berikut.

.

Tabel 4. *Parameter* pada Operator *Select Attributes*

No.	***Parameter***	***Value***
1	*Type*	*include attributes*
2	*attribute filter type*	*a subset*
3	*select subset*	*select attributes:*
		Hidden Gem Score
		IMDb Score
		Runtime
		Title

Hasil dari penggunaan operator *Select Attributes* diperoleh informasi pada Tabel 5 berikut.

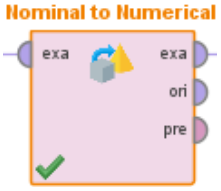
Tabel 5. Hasil Operator *Select Attributes*

No.	Nama Atribut	Jenis Data	Missing
1	*Hidden Gem Score*	*Real*	0
2	*IMDb Score*	*Real*	0
3	*Runtime*	*Polynominal*	0
4	*Title*	*Polynominal*	0

Dari hasil *result* dari statistik *dataset* 4 atribut seperti tampak pada Tabel 5, diketahui bahwa tidak ada atribut yang memiliki nilai *missing*. Untuk memeriksa konsisten atau tidak konsistennya *dataset* yang digunakan diperiksa per-*record* secara langsung dan menunjukkan bahwa *dataset* memiliki data yang konsisten terhadap nilainya.

4.4. *Data Transformation*

Pada tahap ini peneliti merubah data *non-numeric* menjadi *numeric*. Tahap penelitian ini menggunakan operator *Nominal to Numerical* dapat ditampilkan pada Gambar 5 berikut.



Gambar 5. Operator *Nominal to Numerical* pada RapidMiner

Parameter yang digunakan pada operator *Nominal to Numerical* yaitu pada attribute filter type memilih subset dan attributes yang akan dijadikan Numerik yaitu atribut *Runtime* dan *Title* dapat dilihat pada Tabel 6 berikut.

Tabel 6. *Parameter* pada Operator *Nominal to Numerical*

No.	***Parameter***	Isi
1	*attribut filter type*	*Subset*
2	*Attributes*	*Select Attributes*
		Runtime
3	*coding type*	*unique integers*

Dikarenakan atribut-atribut sebelumnya sudah mengalami *selection* dan hanya terpilih 4 atribut, dan atribut *Title* menjadi id, atribut *Hidden Gem Score* dan atribut *IMDb Score* sudah berbentuk numerik. Maka hanya atribut *Runtime* yang berbentuk polynominal diubah menjadi numerik. Hasil dari penggunaan operator Nominal to Numerical dapat dilihat pada Gambar 6 Berikut.

Gambar 6. Hasil Penggunaan Operator *Nominal to Numerical*

4.5. *Data Mining*

Langkah ini berfungsi untuk menentukan metode atau teknik yang paling sesuai dalam mengidentifikasi pola atau informasi menarik dalam data yang telah dipilih. Dalam penelitian ini, proses data mining diimplementasikan dengan menggunakan metode *Clustering*, dengan memanfaatkan Algoritma *K-Means*. Pada tahap ini juga peneliti melakukan proses *clustering*, menguji dan mengevaluasi hasil dari proses tersebut dalam platform *RapidMiner*. Operator yang digunakan adalah operator *Clustering (K-Means)* dapat dilihat pada Gambar 7 berikut.

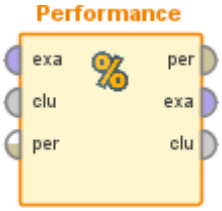
Gambar 7. Operator *Clustering* pada *RapidMiner*

Parameter yang digunakan pada *k-Means* atau *Clustering* dapat dilihat pada Tabel 7 berikut.

Tabel 7. *Parameter* pada Operator Clustering

No.	***Parameter***	Isi
1	Add *cluster* attribute	Digunakan
2	K	2
3	*max runs*	10
4	*determine good star values*	Digunakan
5	*measure types*	*MixedMeasured*
6	*Mixed measured*	*MixedMeasured*
7	*Max optimization*	100

Pada saat menggunakan operator *clustering* kita perlu mengetahui berapa klaster yang terbaik dalam *dataset* ini, lalu dalam penelitian ini menggunakan performance *Davies Bouldin Index* (DBI) dapat dilihat pada Gambar 8 berikut.



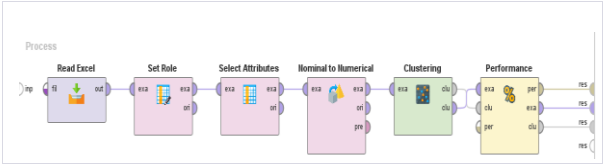
Gambar 8 Operator *Performance* untuk *Davies Bouldin Index* pada *RapidMiner*

Parameter yang digunakan pada operator *Performance* dapat dilihat pada Tabel 8 berikut.

Tabel 8. *Parameter* pada Operator *Performance*

No.	***Parameter***	Isi
1	*Main criterion*	*Davies Bouldin*
2	*Normalize*	Digunakan
3	*Maximize*	Digunakan

Pada operator *Performance*, parameter *normalize* dan *maximize* digunakan karena jika hanya mencentang *normalize* saja akan menjadi *minus*, dan menggunakan atau mencentang *maximize* hanya menghilangkan *minus*nya saja tidak merubah *cluster* atau lainnya.



Gambar 9. Tahapan *Clustering* Pada *RapidMiner*

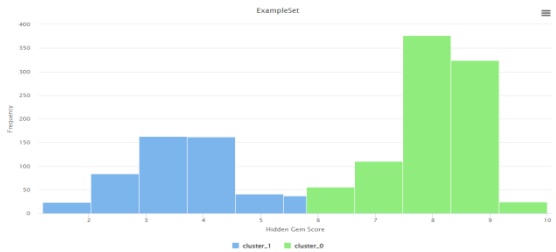
Pada Gambar 9 merupakan tahapan-tahapan clustering pada RapidMiner yang dapat dijelaskan sebagai berikut:

1. Memasukkan dataset dalam format *Excel* pada operator *Read Excel.*
2. Mengklasifikasikan sebuah atribut sebagai atribut khusus atau standard, disini memilih atribut *Title* menjadi id.
3. Memilih atribut yang akan di*cluster*kan yaitu *Title*, *Hidden Gem Score*, *Runtime*, dan *IMDb Score*.
4. Merubah *NonNumeric* menjadi *Numeric.*
5. Mencoba mengklasterkan beberapa kali dengan DBI untuk menemukan klaster yang terbaik yaitu mencoba cluster 2 sampai dengan cluster 10.
6. Menambahkan operator performance untuk dapat melihat nilai DBI tersebut.

Tabel 9. Nilai DBI Setiap *Cluster*

Cluster	**Nilai DBI**
K2	0.187
K3	0.304
K4	0.379
K5	0.353
K6	0.344
K7	0.346
K8	0.363
K9	0.349
K10	0.324

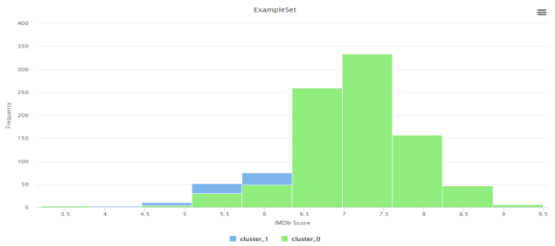
Dari Tabel 9, mengacu pada prinsip DBI, nilai yang dianggap baik dalam data mining adalah semakin kecil atau mendekati nol. Dalam Tabel 4.2 nilai DBI yang paling baik adalah cluster 2, dengan nilai DBI yan mendekati nol yaitu 0.187. Oleh karena itu, pengelompokkan dilakukan dengan menggunakan 2 cluster.



Gambar 10. *Cluster* pada Atribut *Hidden Gem Score*

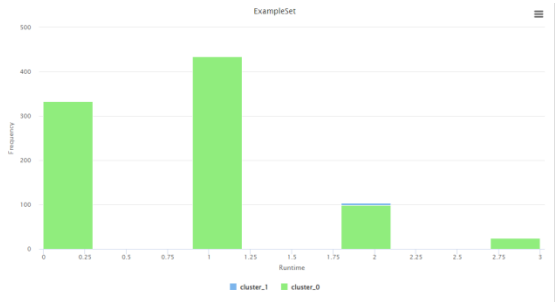
Penelitian ini berhasil mengimplementasikan teknik *clustering*, dari proses tersebut menghasilkan 3 visualisasi data dalam bentuk histogram pada Gambar 10, Gambar 11, dan Gambar 12.

Bentuk grafik yang ditujukan pada Gambar 10 ini merupakan visualiasasi dari hasil pemodelan *cluster* pada atribut *Hidden Gem Score*. Dari pengujian diatas dapat disimpulkan bahwa *cluster* 0 memiliki jumlah rata-rata sebanyak 8.034 yang dilambangkan dengan warna hijau dan *cluster* 1 dengan rata-rata sebanyak 3.668 yang dilambangkan dengan warna biru.



Gambar 11. *Cluster* pada Atribut *IMDb Score*

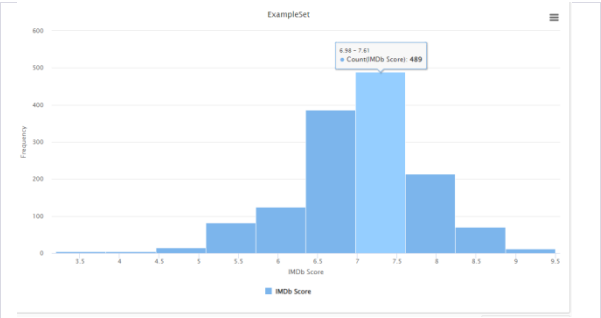
Pada Gambar 11 merupakan hasil visualisasi dari *cluster* pada atribut *IMDb Score*. Dari pengujian yang telah dilakukan software *RapidMiner*, dapat disimpulkan bahwa *cluster* 0 memiliki jumlah rata-rata sebanyak 7.148 yang dilambangkan dengan warna hijau dan *cluster* 1 dengan rata-rata berjumlah 6.817 yang dilambangkan dengan warna biru.



Gambar 12. Cluster pada Atribut Runtime

Pada Gambar 12 merupakan hasil visualisasi dari *cluster* pada atribut *Runtime*. Dari pengujian yang telah dilakukan software *RapidMiner*, dapat disimpulkan bahwa *cluster* 0 memiliki jumlah rata-rata sebanyak 0.793 yang dilambangkan dengan warna hijau dan *cluster* 1 memiliki rata-rata dengan jumlah 1.016 yang dilambangkan dengan warna biru. Salah satu visualisasi data film *Netflix* berdasarkan *IMDb Score* yang diperoleh dapat diperlihatkan pada Gambar 12 berikut.

Pada Gambar 13, jumlah data film *Netflix* 1401 *dataset* paling banyak memberikan *rating* atau *score* 6.98 - 7.61 dengan jumlah 489 nilai. Algoritma *K-Means Clustering* dapat dimanfaatkan untuk mengelompokkan data film *Netflix* dengan atribut *Title, Hidden Gem Score, Runtime dan IMDb Score* dengan hasil *Cluster Model* dan jumlah anggota setiap *cluster* dapat dilihat pada Gambar 14 berikut.



Gambar 13. Visualisasi Data Film *Netflix* Berdasarkan *IMDb Score*

Cluster Model

```
Cluster 0: 891 items
Cluster 1: 510 items
Total number of items: 1401
```

Gambar 14. Hasil Jumlah Item dari 2 *Cluster*

Dari Gambar 14, menunjukkan bahwa data telah berhasil dikelompokkan menjadi 2 klaster yaitu *Cluster* 0 dengan jumlah 891 anggota dan *Cluster* 1 berjumlah 510 anggota. Jumlah *item* yang berbeda menunjukkan adanya variasi dalam sifat atau karakteristik item-item tersebut. Klaster dengan jumlah yang lebih tinggi menunjukkan adanya lebih banyak film. Selanjutnya analisis nilai rata-rata *centroid* pada setiap *cluster* dari atribut yang telah ditentukan. Hasil dari nilai rata-rata atribut dapat dilihat dari Gambar 15 berikut.

Attribute	cluster_0	cluster_1
Runtime	0.793	1.016
Hidden Gem Score	8.034	3.665
IMDb Score	7.148	6.817

Gambar 15. Hasil Rata-rata *Centroid* dari Setiap *Cluster*

Dari Gambar 14 dan Gambar 15 dapat disimpulkan bahwa setiap *cluster* dalam kategori film berdasarkan *Title*, *Hidden Gem Score*, *Runtime* dan *IMDb Score* dapat dianalisis sebagai berikut :

1) Cluster 0 : Film *Netflix* dengan jumlah *Title*, *Hidden Gem Score*, *Runtime* dan *IMDb Score* tinggi.
2) Cluster 1 : Film *Netflix* dengan jumlah *Title*, *Hidden Gem Score*, *Runtime* dan *IMDb Score* rendah.

Dengan dihasilkannya *cluster-cluster* tersebut ditemukannya film popular yaitu dengan ciri-ciri ideal dalam mengelompokkan data film *Netflix*, dengan nilai rata-rata atribut *Hidden Gem Score* sebesar 8.034, atribut *Runtime* sebesar 0.793 Dan atribut *IMDb Score* sebesar 7.158 yang terpadat pada

Cluster 0. Beberapa title yang terdapat pada *cluster* 0 atau film-film popular yaitu pada Gambar 16 berikut.

Gambar 16. *Title* Film *Cluster* 0

Pada Gambar 16 merupakan beberapa *Title* pada *Cluster* 0 yang artinya *Title-title* pada film popular di data *Netflix*. Selama penelitian, jarak antar klaster diukur dati titik pusat klaster terdekat dan terjauh untuk memahami distribusi atau perbedaan antar klaster berdasarkan atribut-atribut yang dianalisis dalam *dataset Netflix*. Jumlah rata-rata dalam jarak *centroid* setiap *cluster* pada Tabel 10.

Tabel 10. Jarak *Centroid* setiap *Cluster*

Cluster	Rata-rata Jarak Centroid
K Pusat	0.611
K0	0.542
K1	0.733

Pada Tabel 10 dapat dilihat bahwa jarak antar *cluster* terhadap titik pusat *cluster* pada data film *Netflix* pada *cluster* 0 dengan rata-rata jarak 0.542 dan *cluster* 1 rata-rata jarak 0.733.

5. KESIMPULAN DAN SARAN

Dari serangkaian riset yang telah dilakukan, terdapat beberapa kesimpulan yang dapat diambil. Pertama, penggunaan algoritma K-Means pada data film *Netflix* berhasil menghasilkan dua klaster: klaster rendah dan tinggi. Kedua, melalui *metode Knowledge Discovery in Database* (KDD) dan evaluasi menggunakan nilai *Davies Bouldin Index* (DBI), hasil optimal diperoleh pada klaster 0 dengan nilai DBI 0.187. Klaster ini mengelompokkan film berdasarkan atribut *Title, Hidden Gem Score, Runtime,* dan *IMDb Score*. Ketiga, klaster 0 menunjukkan film-film dengan nilai *Hidden Gem Score* tinggi (8.034), durasi *Runtime* yang panjang (0.793 jam), dan *IMDb Score* yang tinggi (7.148), mengindikasikan popularitas dan kualitas film yang tinggi. Jarak rata-rata antar klaster terhadap titik pusat *cluster* adalah 0.542 untuk klaster 0 dan 0.733 untuk klaster 1. Saran untuk penelitian selanjutnya mencakup penggunaan hasil clustering ini dalam membantu pengguna menemukan film sesuai preferensi mereka, serta saran untuk melakukan perbandingan dengan metode atau algoritma *clustering* lainnya guna meningkatkan keberagaman dan keakuratan analisis. Penelitian ini memberikan wawasan dalam pengelompokkan film *Netflix* dan memberikan potensi penggunaan hasil klastering untuk membantu pengguna dalam menemukan konten yang sesuai. Meskipun hasilnya cukup mengesankan, ada ruang bagi penelitian lebih lanjut untuk memperbaiki dan memperluas analisis dengan mempertimbangkan penggunaan metode atau algoritma clustering yang berbeda.

DAFTAR PUSTAKA

[1] I. Fitrianti, A. Voutama, and Y. Umaidah, "Clustering Film Populer pada Aplikasi Netflix dengan Menggunakan Algoritma K-Means dan Metode CRISP-DM," *J. Teknol. Sist. ...*, vol. 4, no. 2, pp. 301–311, 2023, [Online]. Available: https://jurnal.mdp.ac.id/index.php/jtsi/article/view/4929%0Ahttps://jurnal.mdp.ac.id/index.php/jtsi/article/download/4929/1546

[2] G. Gustientiedina, M. H. Adiya, and Y. Desnelita, "Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan," *J. Nas. Teknol. dan Sist. Inf.*, vol. 5, no. 1, pp. 17–24, Apr. 2019, doi: 10.25077/teknosi.v5i1.2019.17-24.

[3] M. Iqbal, "KLASTERISASI DATA JAMAAH UMROH PADA AULIYA TOUR & TRAVEL MENGGUNAKAN METODE K-MEANS CLUSTERING," *JURTEKSI (Jurnal Teknol. dan Sist. Informasi)*, vol. 5, no. 2, pp. 97–104, Jun. 2019, doi: 10.33330/jurteksi.v5i2.352.

[4] S. Paembonan and H. Abduh, "Penerapan Metode Silhouette Coefficient untuk Evaluasi Clustering Obat," *PENA Tek. J. Ilm. Ilmu-Ilmu Tek.*, vol. 6, no. 2, p. 48, 2021, doi: 10.51557/pt_jiit.v6i2.659.

[5] I. Nasution, A. Perdana Windarto, and M. Fauzan, "Penerapan Algoritma K-Means Dalam Pengelompokan Data Penduduk Miskin Menurut Provinsi," *Technol. Sci. (BITS*, vol. 2, no. 2, pp. 76–83, 2020, [Online]. Available: https://www.bps.go.id.

[6] W. Witanti, F. R. Umbara, U. Jenderal, and A. Yani, "PERAMALAN GENRE FILM TERPOPULER BERDASARKAN DATASET MYMOVIE MENGGUNAKAN METODE AUTOREGRESSIVE INTEGRATED," vol. 1861, no. 9, pp. 610–617, 2023.

[7] R. Pasys, "Arti Kata 'Hidden Gem' yang Populer, Pengertian dan Contoh Kalimatnya," Grids kids. Accessed: Dec. 04, 2023. [Online]. Available:

https://kids.grid.id/read/473957593/arti-kata-hidden-gem-yang-populer-pengertian-dan-contoh-kalimatnya

[8] G. Amadeo, "9 Film Indonesia dengan Rating Terbaik Menurut IMDb," idn times. Accessed: Dec. 04, 2023. [Online]. Available: https://www.idntimes.com/hype/entertainment/gregorius-amadeo-1/9-film-indonesia-dengan-rating-terbaik-menurut-imdb

[9] L. Sinaga, A. Ahmad, and M. Safii, "PENERAPAN DATA MINING PADA JUMLAH PELANGGAN PERUSAHAAN AIR BERSIH MENURUT PROVINSI MENGGUNAKAN METODE K-MEANS CLUSTERING," *J. Resist. (Rekayasa Sist. Komputer)*, vol. 2, no. 2, pp. 119–125, 2019, doi: 10.31598/jurnalresistor.v2i2.418.

[10] R. K. Dinata, S. Safwandi, N. Hasdyna, and N. Azizah, "Analisis K-Means Clustering pada Data Sepeda Motor," *INFORMAL Informatics J.*, vol. 5, no. 1, p. 10, 2020, doi: 10.19184/isj.v5i1.17071.

[11] Trivusi, "K-Means Clustering: Pengertian, Cara Kerja, Kelebihan, dan Kekurangannya," trivusi.web.id. Accessed: Dec. 23, 2023. [Online]. Available: https://www.trivusi.web.id/2022/06/algoritma-kmeans-clustering.html?m=1

[12] A. Febrian, Nana Suarna, and Gifthera Dwilestari, "Penerapan Algoritma K-Means Untuk Mengelompokkan Data Pengiriman Paket Di Kantor Pos Cirebon," *J. Teknol. Technoscientia*, vol. 15, no. 1, pp. 23–27, 2022, doi: 10.34151/technoscientia.v15i1.3858.

[13] Haris Kurniawan, Sarjon Defit, and Sumijan, "Data Mining Menggunakan Metode K-Means Clustering Untuk Menentukan Besaran Uang Kuliah Tunggal," *J. Appl. Comput. Sci. Technol.*, vol. 1, no. 2, pp. 80–89, Dec. 2020, doi: 10.52158/jacost.v1i2.102.

[14] T. Hartati and Y. Arie Wijaya, "ANALISIS DATA LALU LINTAS JARINGAN DI KANTOR CANGEHGAR CYBER OPERATION CENTER MENGGUNAKAN ALGORITMA K-MEANS NETWORK TRAFFIC DATA ANALYSIS AT CANGEHGAR CYBER OPERATION CENTER OFFICE USING K-MEANS ALGORITHM," *J. Ilm. NERO*, vol. 7, no. 1, p. 2022, 2022.