

JURNAL STATISTIKA DAN SAINS DATA
Volume 4 Edisi 1, Oktober, 2026, 189-199
https://journal.student.uny.ac.id/index.php/jssd
https://dx.doi.org/10.21831/jssd.v4i1.24109

OPTIMALISASI SISTEM REKOMENDASI FILM BERBASIS DEMOGRAFI PENGGUNA LAYANAN NETFLIX

Condoleeza Martince Indasi Menanti, Universitas Cenderawasih, Indonesia
Aryanto*, Universitas Cenderawasih, Indonesia
*e-mail: aryanto@fmipa.uncen.ac.id

Abstrak. Netflix sebagai layanan *streaming* global memanfaatkan sistem rekomendasi untuk meningkatkan kepuasan dan keterlibatan pengguna. Umumnya, sistem ini mengandalkan algoritma berbasis konten atau kolaboratif. Namun, pendekatan tersebut sering mengabaikan data demografis yang juga memengaruhi preferensi tontonan, seperti usia, jenis kelamin, dan negara. Penelitian ini bertujuan membangun Sistem Rekomendasi Film Berbasis Demografi dengan pendekatan *K-Means Clustering*. Data diambil dari dataset pengguna Netflix di Kaggle. Setelah melalui tahap pra-pemrosesan, dilakukan segmentasi pengguna berdasarkan atribut demografis. Setiap cluster dianalisis untuk mengetahui genre film yang dominan, yang kemudian digunakan sebagai dasar rekomendasi. Evaluasi sistem dilakukan melalui analisis kesesuaian rekomendasi dan survei pengguna. Hasil dari penelitian ini diharapkan memberikan rekomendasi yang lebih relevan dengan karakteristik pengguna serta mampu mengatasi masalah *cold-start*. Pendekatan ini diharapkan dapat dikembangkan lebih lanjut untuk membentuk sistem rekomendasi yang lebih inklusif dan adaptif terhadap keragaman pengguna global.

Kata Kunci : Sistem Rekomendasi, Demografi, K-Means, Cluster, Netflix

Abstract. *Netflix, as a global streaming service, uses recommendation systems to enhance user satisfaction and engagement. These systems typically rely on content-based or collaborative algorithms. However, such approaches often overlook demographic factors like age, gender, and country, which significantly influence viewing preferences. This study aims to develop a demographic-based movie recommendation system using K-Means Clustering. The data is sourced from a publicly available Netflix user dataset on Kaggle. After preprocessing, users are segmented based on demographic attributes. Each cluster is analyzed to determine the most preferred movie genres, which form the basis for tailored recommendations. System evaluation is carried out through relevance analysis and user feedback surveys. The results of this study are expected to provide recommendations that are more relevant to user characteristics and capable of overcoming cold-start problems. This approach is expected to be further developed to form a recommendation system that is more inclusive and adaptive to the diversity of global users.*

Keywords: Recommender System, Demographics, , K-Means, Cluster, Netflix

Copyright © 2026, Jurnal Statistika dan Sains Data, Online ISSN: 3025-9649

PENDAHULUAN

Kemajuan teknologi digital serta pesatnya perkembangan internet telah memberikan dampak yang luar biasa terhadap berbagai aspek kehidupan, salah satunya dalam bidang hiburan. Kini, masyarakat global dapat dengan mudah mengakses berbagai bentuk hiburan tanpa harus bergantung pada media konvensional. Salah satu bentuk transformasi yang paling menonjol adalah kemunculan dan popularitas layanan streaming digital. Di antara berbagai platform yang tersedia, Netflix menjadi salah satu layanan yang paling banyak digunakan di seluruh dunia. Platform ini menyediakan ribuan pilihan judul film dan serial yang dapat diakses oleh pengguna kapan saja dan di mana saja, tanpa batasan geografis, serta melayani beragam pengguna dari berbagai latar belakang sosial, budaya, dan usia. Untuk mempertahankan minat pengguna serta meningkatkan intensitas keterlibatan (*engagement*) mereka terhadap layanan yang disediakan, Netflix mengintegrasikan fitur canggih berupa sistem rekomendasi yang telah terbukti menjadi salah satu elemen kunci dalam meningkatkan pengalaman pengguna di platform tersebut.

Sistem rekomendasi sendiri merupakan salah satu aplikasi praktis dari kecerdasan buatan (*artificial intelligence*) yang dirancang secara khusus untuk memberikan saran atau usulan konten kepada pengguna berdasarkan preferensi atau kecenderungan perilaku yang telah ditunjukkan oleh pengguna tersebut pada masa lalu. Tujuan utama dari pengembangan dan penerapan sistem ini adalah untuk membantu pengguna dalam menemukan item, produk, atau konten yang sesuai dengan selera dan kebutuhannya di tengah melimpahnya informasi dan konten yang tersedia secara digital. Sistem ini bekerja dengan cara mengumpulkan, memproses, serta menganalisis berbagai jenis data yang berkaitan dengan aktivitas pengguna, seperti histori pencarian konten, aktivitas menonton sebelumnya, pemberian rating, maupun bentuk interaksi lain yang terjadi saat pengguna menggunakan layanan platform tersebut (Ariyus, Maningga, Sembiring, 2024).

Salah satu bentuk implementasi sistem rekomendasi yang banyak digunakan adalah sistem berbasis demografi (*demographic-based recommender system*). Sistem penyaringan demografis membagi pengguna ke dalam kelas-kelas berdasarkan data demografi dan preferensi mereka. Setiap kelas ini terdiri dari kelompok pengguna yang memiliki karakteristik dan pola perilaku serupa, yang digunakan sebagai dasar dalam memberikan rekomendasi. Rekomendasi yang dihasilkan bersifat stereotipikal karena didasarkan pada asumsi umum mengenai kelompok tersebut. Sistem ini membuat keputusan berdasarkan kecenderungan kolektif dari kelompok pengguna yang memiliki minat yang sama terhadap suatu produk. Dengan pendekatan ini, sistem dapat memberikan rekomendasi dengan cepat dan efisien, cukup dengan melihat item-item yang disukai oleh kelompok demografis terkait(Gangothri, Saranya, Venkataraman, 2015).

Sebelum melakukan pengkajian lebih jauh terkait penerapan sistem rekomendasi berbasis demografi, perlu diketahui bahwa sejumlah studi terdahulu telah menunjukkan bagaimana data demografis dapat dimanfaatkan dalam pengembangan sistem rekomendasi. Sebagai contoh, terdapat sebuah penelitian yang memanfaatkan pendekatan demographic filtering untuk mengevaluasi perilaku dari 50 orang pengemudi transportasi umum. Berdasarkan hasil klasifikasi yang dilakukan, diperoleh informasi bahwa 23 orang pengemudi dikategorikan sebagai pengemudi aman, 18 orang sebagai semi-agresif, dan sisanya 9 orang diklasifikasikan sebagai pengemudi agresif. Dari hasil ini, disimpulkan bahwa para pengemudi yang termasuk dalam kategori aman dan semi-agresif memiliki kualitas yang cukup baik dalam mengemudi, sedangkan mereka yang tergolong agresif disarankan untuk mengikuti pelatihan lanjutan guna meningkatkan kualitas dan keselamatan berkendara (Safrizal, 2024).

Copyright © 2026, Jurnal Statistika dan Sains Data, Online ISSN: 3025-9649

Studi lainnya yang turut mendukung pentingnya pemanfaatan data demografi dalam sistem rekomendasi adalah penelitian yang mengembangkan sistem rekomendasi mobil berdasarkan informasi demografis pengguna. Meskipun pendekatan ini dinilai menarik, hasil akhir dari sistem tersebut belum mencapai performa yang optimal. Salah satu penyebab utama dari keterbatasan tersebut adalah karena pemilihan parameter demografis yang digunakan tidak sepenuhnya relevan, serta terdapat kekurangan dalam teknik klasifikasi terhadap nilai-nilai dari parameter demografi yang ada. Di samping itu, variasi dalam metode diskretisasi maupun pengkodean data juga memengaruhi seberapa besar tingkat kemiripan antar pengguna yang bisa diidentifikasi oleh sistem, dan hal ini berdampak langsung pada akurasi hasil rekomendasi yang diberikan (Maharani & Gunawan, 2014).

Berdasarkan pemaparan dari beberapa hasil penelitian sebelumnya tersebut, dapat ditarik kesimpulan awal bahwa informasi demografis yang dimiliki oleh pengguna dapat dimanfaatkan secara efektif sebagai salah satu dasar untuk membentuk sistem rekomendasi yang adaptif. Dalam skala yang lebih luas, terutama dalam konteks global di mana Netflix melayani jutaan pengguna dari berbagai wilayah geografis dan latar belakang budaya, integrasi data demografi ke dalam sistem rekomendasi menjadi semakin penting. Pendekatan ini tidak hanya akan meningkatkan tingkat akurasi dari hasil rekomendasi yang diberikan oleh sistem, tetapi juga berpotensi besar untuk menyesuaikan saran konten secara lebih relevan terhadap kondisi sosial dan nilai-nilai budaya pengguna. Harapannya, hasil dari penelitian ini nantinya dapat memberikan kontribusi dalam pengembangan sistem rekomendasi film berbasis demografi yang lebih efektif dan adaptif, serta mampu mengevaluasi performa sistem rekomendasi berbasis demografi jika dibandingkan dengan sistem rekomendasi konvensional lainnya, baik dari aspek akurasi maupun tingkat kepuasan pengguna terhadap rekomendasi yang diterima.

METODE

Data

Data yang akan digunakan dalan penelitian ini adalah dataset yang diunduh dari link https://www.kaggle.com/datasets/smayanj/netflix-users-database di situs Kaggle dengan judul "Netflix User Database" yang menyediakan kumpulan data terkait pengguna Netflix. Data ini mencangkup beberapa atribut, yakni *User_ID, Name, Age, Country, Subscription_Type, Watch_Time_Hours, Favorite_Genre, Last_Login*. Namun karena penelitian ini berbasis demografi, maka hanya akan berfokus menggunakan atribut *Age, Country,* Dan *Favorite_Genre*. Oleh Karena itu data akan melalui tahap Preprocessing sebelum masuk ke tahap pengolahan data. Proses pengerjaan data pada penelitian ini dilakukan dengan menggunakan Pyton.

Pra-pemrosesan Data

Pada tahap ini, dilakukan penanganan *missing values* menggunakan metode imputasi, kemudian *data cleaning* atau penghapusan data, *encoding* terhadap data kategorikal (misalnya *one-hot encoding* untuk negara), serta penskalaan data numerik (seperti usia) menggunakan *MinMaxScaler* atau *StandardScaler*, dengan rumus persamaan:

$$MiniMaxScaler = x' = \frac{x - x_{min}}{x_{max} - x_{min}} \quad \ldots(1)$$

$$StandardScaler = z = \frac{x - \mu}{\sigma} \quad \ldots(2)$$

di mana x' adalah nilai yang telah diskalakan dengan *MinMaxScaler*, x adalah nilai asli, x_{min} dan x_{max} adalah nilai minimum dan maksimum dalam dataset, z adalah nilai yang telah distandarisasi dengan *StandardScaler*, μ adalah rata-rata dari dataset, dan σ adalah standar deviasi dari dataset (Geron, 2019).

Copyright © 2026, Jurnal Statistika dan Sains Data, Online ISSN: 3025-9649

Segmentasi pengguna

Pada tahap ini, algoritma *K-Means Clustering* akan digunakan untuk mengelompokkan pengguna berdasarkan atribut demografi. *K-Means Clustering* merupakan salah satu metode sederhana untuk melakukan partisi sekumpulan data menjadi *K* cluster yang berbeda dan tidak tumpang-tindih(Muhima, dkk, 2021.). Algoritma ini bekerja dengan cara berikut:

1. Inisialisai: yaitu menentukan nilai *k*, di mana nilai *k* merepresentasikan jumlah cluster dan pilih *k* titik acak sebagai *centroid*(pusat) awal cluster.
2. Penugasan cluster: Menetapkan setiap titik ke cluster dengan jarak terdekat. Pada tahap ini juga, kita perlu menghitung jarak setiap titik dengan *centroid*. Rumus yang digunakan untuk menghitung jarak setiap titik dengan *centroid* yakni *Euclidean Distance*, yakni sebagai berikut:
   $d(x,y) = \sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$ …(3)
   Dimana *d(x,y)* adalah jarak antar objek x dan y, *n* adalah dimensi dari data, *Xi* adalah koordinat objek *X* pada dimensi data *n*, dan *Yi* adalah koordinat dari objek *Y* pada dimensi data *n*
3. Pembaruan titik *centroid*: Setelah semua data point ditugaskan, *centroid* diperbarui dengan menghitung rata-rata dari semua data point dalam cluster tersebut dengan menggunakan *Euclidean Distance* lagi.
4. Iterasi(Pengulangan): Ulangi Langkah 2 dan 3 hingga jika tidak ada perubahan dalam penugasan cluster atau *centroid* maka pengklasteran selesai. (Muhima, dkk, 2021., Wahyudi, dkk, 2020.)

Untuk jumlah cluster optimal ditentukan menggunakan *Elbow Method* dan *Silhouette Score*[4]. *Elbow Method* digunakan dalam menentukan clusster optimal dengan memberikan informasi berupa grafik yang menunjukkan nilai *within cluster sum square* (WCSS) dan nilai k yang mengalami penurunan (Syakur et al., 2018)

$wcss = \sum_{K=1}^{K}\sum_{Xi\in S_K}(X_i - C_K)^2$ …(4)

dengan X_i merupakan data ke−i dan C_K merupakan centroid clusster ke−i(Dalam Devi, et al, 2023).

Silhouette Score digunakan untuk menilai seberapa baik tiap data cocok dengan clusternya dibandingkan dengan clusster lain. Nilainya berkisar dari -1 sampai 1, di mana:

- Nilai mendekati 1: data sangat sesuai dengan clusternya
- Nilai mendekati 0: data berada di batas antara dua cluster
- Nilai negatif: data kemungkinan salah cluster

Shilouette Score dapat dicari menggunakan persamaan berikut:

$$s(i) = \frac{b(i)-a(i)}{max\{a(i),b(i)\}} \quad \ldots(5)$$

Dimana *a(i)* adalah rata-rata jarak antara data iii dan semua titik dalam klaster yang sama sedangkan *b(i)* merupakan rata-rata jarak antara data iii dan semua titik pada klaster terdekat berikutnya

Nilai rata-rata dari semua *s(i)* digunakan sebagai *Silhouette Score* rata-rata untuk klastering tersebut, dengan persamaan sebagai berikut:

$$S = \frac{1}{n}\sum_{i=1}^{n} s(i) \quad \ldots(6)$$

Kemudian hasil visualisasi grafik dari *Elbow Method* dan *Shilouette Score* akan digunakan untuk membantu menentukan jumlah cluster yang optimal.

Analisis preferensi genre

Pada tahap ini, dihitung rata-rata tingkat preferensi terhadap tiap genre film untuk setiap cluster demografis, dan diidentifikasi genre film yang paling dominan di setiap kelompok

Copyright © 2026, Jurnal Statistika dan Sains Data, Online ISSN: 3025-9649

pengguna. Kemudian tabel hasil analisis preferensi genre disajikan untuk memberikan gambaran tentang genre yang paling disukai oleh setiap cluster demografis (Aggarwal, C. C. , 2016).

Pengembangan Sistem Rekomendasi

Pada tahap ini, dibentuk profil rekomendasi untuk tiap cluster berdasarkan genre yang paling disukai. Sebagai contoh, jika pengguna dalam cluster "Remaja di *USA*" paling menyukai genre *Action* dan *Sci-Fi*, maka sistem akan merekomendasikan genre tersebut kepada pengguna baru dengan profil demografis serupa .(Ricci, F., Rokach, L., & Shapira, B. 2015)

Evaluasi Sistem

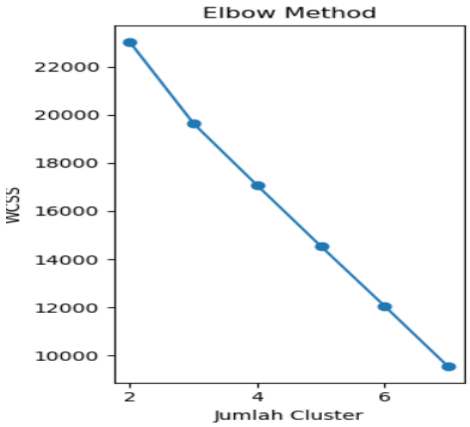
Evaluasi dilakukan secara kualitatif dengan menganalisis kesesuaian rekomendasi dengan karakteristik demografi pengguna. Selain itu, evaluasi kuantitatif sederhana dapat dilakukan melalui survei kecil yang mengumpulkan umpan balik terhadap relevansi rekomendasi dari pengguna secara acak.(Gunawardana, A., & Shani, G., 2015)

HASIL DAN PEMBAHASAN

Hasil

Visualisasi grafik yang ditampilkan pada bagian berikut ini merupakan hasil dari proses analisis data menggunakan metode *Elbow* dan *Silhouette Score*, yang diimplementasikan melalui pemrograman dengan bantuan perangkat lunak Python. Kedua grafik ini digunakan untuk mengevaluasi kualitas dan efektivitas dari proses klasterisasi data dalam penelitian ini. Grafik *Elbow Method* berfungsi untuk menentukan jumlah klaster yang optimal dengan mengamati titik di mana penurunan nilai WCSS (*Within-Cluster Sum of Squares*) mulai melambat secara signifikan, sementara grafik *Silhouette Score* digunakan untuk mengukur konsistensi dan pemisahan antar klaster yang terbentuk.

Dengan menggunakan Python sebagai alat bantu utama dalam pengolahan dan visualisasi data, kita dapat secara sistematis menilai performa dari model klasterisasi yang digunakan, serta memastikan bahwa jumlah cluster yang dipilih memberikan hasil segmentasi yang paling representatif terhadap karakteristik data pengguna.



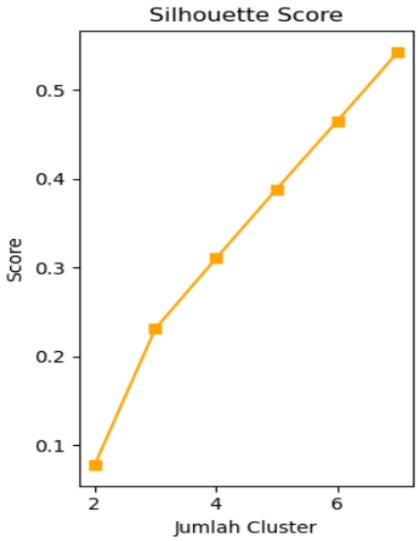
Gambar 1. Grafik *Elbow Method*

Grafik yang ditampilkan pada gambar 1 menggambarkan hubungan antara jumlah klaster yang digunakan dalam proses segmentasi data (ditampilkan pada sumbu X) dengan nilai *Within-Cluster Sum of Squares* atau WCSS (ditampilkan pada sumbu Y), yang merupakan ukuran total variasi dalam setiap klaster. Secara umum, grafik ini menunjukkan pola penurunan nilai WCSS seiring bertambahnya jumlah klaster, yang menandakan bahwa semakin banyak klaster yang digunakan, maka tingkat keseragaman atau homogenitas data dalam masing-masing klaster menjadi semakin tinggi.

Copyright © 2026, Jurnal Statistika dan Sains Data, Online ISSN: 3025-9649

Hal ini terjadi karena dengan menambahkan jumlah klaster, data dapat dikelompokkan dengan lebih spesifik dan sesuai dengan karakteristik yang serupa. Namun demikian, grafik tersebut tidak menunjukkan adanya titik *"elbow"* atau sudut tajam yang mencolok, yaitu titik di mana penurunan WCSS mulai melambat secara signifikan dan dianggap sebagai jumlah klaster optimal. Ketiadaan titik *elbow* yang jelas menyulitkan proses penentuan jumlah klaster terbaik jika hanya mengandalkan metode ini.

Meskipun demikian, penurunan WCSS tetap dapat diamati dan terlihat cukup konsisten atau linier, dengan indikasi bahwa penurunan yang lebih signifikan terjadi pada jumlah klaster ke-3 hingga ke-4. Oleh karena itu, meskipun metode ini memberikan gambaran umum tentang efisiensi segmentasi, interpretasinya tetap memerlukan pertimbangan tambahan dari metode evaluasi lain, sehingga perlu dilanjutkan dengan penggunaan *Silhouette Score*.



Gambar 2. Grafik *Shilouette score*

Sementara itu, Grafik pada gambar 2 menunjukkan nilai *Silhouette Score* untuk setiap jumlah cluster yang diuji. Metrik ini digunakan untuk menilai seberapa baik suatu objek dikelompokkan dalam clusternya, dengan nilai yang lebih tinggi menunjukkan pemisahan antar cluster yang lebih jelas dan kualitas segmentasi yang lebih baik. Terlihat bahwa nilai *Silhouette Score* meningkat secara bertahap seiring bertambahnya jumlah cluster, dengan nilai tertinggi tercapai saat jumlah cluster mencapai 7.

Hal ini mengindikasikan bahwa segmentasi paling optimal terjadi pada titik tersebut. Namun, karena peningkatannya tidak terlalu signifikan dan cenderung bertahap, maka jumlah cluster 6 atau 7 dapat dianggap ideal. Dalam penelitian ini, dipilih 6 clusster karena dirasa cukup representatif dan masih konsisten dengan hasil evaluasi grafik *Elbow Method*, serta memudahkan dalam interpretasi dan implementasi sistem rekomendasi.

Tabel 1. Hasil Clustering Pengguna berdasarkan rentang usia dengan menggunakan *K-Means*

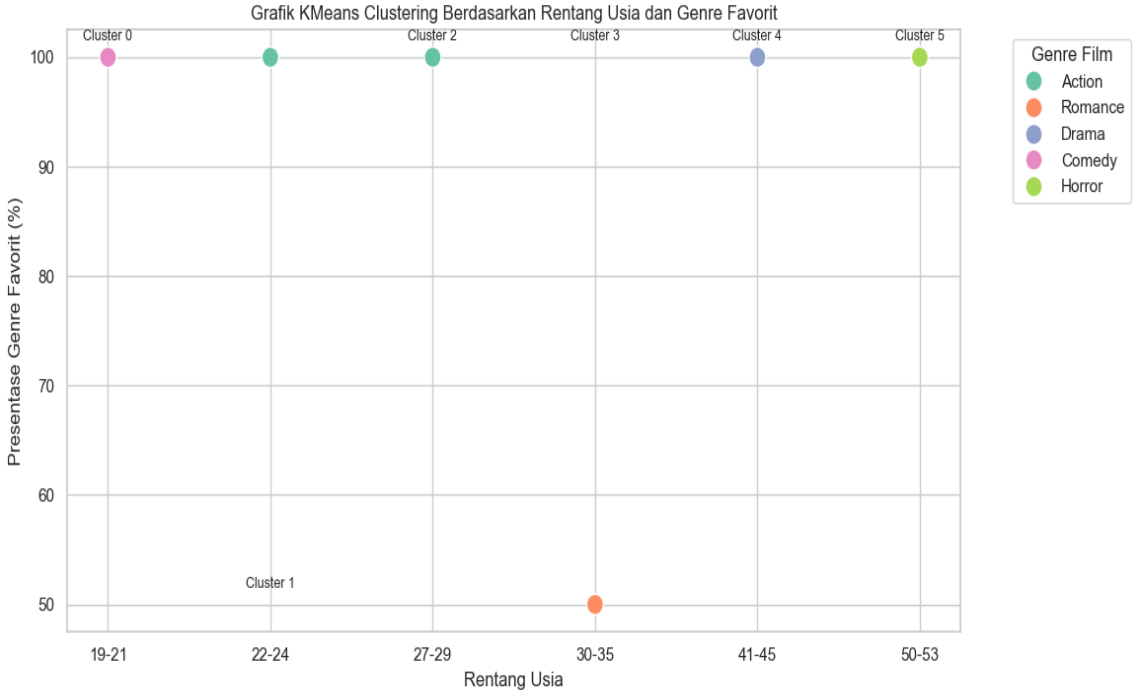
Cluster	Rentang Usia	Genre Favorit	Presentase%
0	22-24	Action	100
1	30-35	Romance	50
2	41-45	Drama	100
3	19-21	Comedy	100
4	27-29	Action	100
5	50-53	Horror	100

Copyright © 2026, Jurnal Statistika dan Sains Data, Online ISSN: 3025-9649

Hasil dari proses analisis klaster berdasarkan rentang usia menunjukkan bahwa preferensi pengguna terhadap genre film cenderung berbeda-beda sesuai dengan kelompok usia mereka. Pada klaster pengguna dengan rentang usia 19–21 tahun, genre yang paling disukai dan mendominasi secara keseluruhan adalah *Comedy*, dengan tingkat preferensi mencapai 100%. Artinya, seluruh pengguna dalam kelompok usia ini secara konsisten menunjukkan minat yang kuat terhadap film bergenre komedi. Sementara itu, kelompok usia 22–24 tahun dan 27–29 tahun sama-sama menunjukkan ketertarikan yang sangat tinggi terhadap genre *Action*, dengan proporsi kesukaan yang juga mencapai 100%, menandakan bahwa genre aksi menjadi pilihan utama bagi pengguna dalam kelompok usia muda dewasa.

Di sisi lain, pengguna yang berada dalam kelompok usia 41–45 tahun secara keseluruhan memilih genre *Drama* sebagai genre favorit mereka, juga dengan persentase sebesar 100%. Hal serupa terjadi pada pengguna berusia 50–53 tahun yang menunjukkan dominasi penuh terhadap genre *Horror*, menggambarkan bahwa pengguna pada usia yang lebih dewasa memiliki kecenderungan minat terhadap genre yang lebih emosional atau menegangkan.

Satu-satunya klaster yang tidak menunjukkan dominasi tunggal terhadap satu genre adalah kelompok usia 30–35 tahun. Dalam kelompok ini, hanya 50% pengguna yang memilih genre *Romance*, sementara sisanya tampaknya memiliki minat yang lebih tersebar terhadap genre lain. Hal ini mengindikasikan adanya keberagaman preferensi dalam klaster ini, sehingga tidak ada satu genre yang benar-benar menonjol. Temuan ini menunjukkan bahwa semakin heterogen usia pengguna, maka semakin beragam pula pola konsumsi konten mereka. Secara visual, perbedaan preferensi antar klaster ini dapat diamati lebih lanjut pada Gambar 3 yang menyajikan hasil segmentasi menggunakan K-Means berdasarkan usia terhadap genre film favorit.



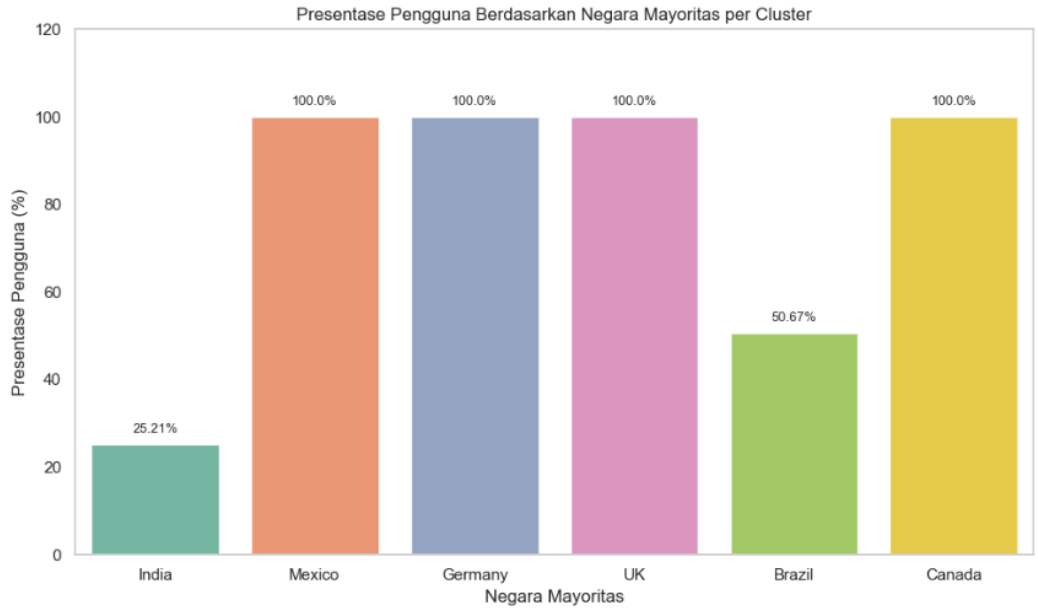
Gambar 3. Grafik hasil segmentasi *KMeans* Clustering Berdasarkan Rentang Usia terhadap Genre Favorit.

Copyright © 2026, Jurnal Statistika dan Sains Data, Online ISSN: 3025-9649

Tabel 2. Hasil Clustering Pengguna berdasarkan Mayoritas Negara asal pengguna dengan menggunakan *K-Means*

Cluster	Negara Mayoritas	Jumlah pengguna	Presentase%
0	India	2505	25.21
1	*Mexico*	2493	100
2	*Germany*	2547	100
3	*UK*	2592	100
4	Brazil	2503	50.67
5	*Canada*	2490	100

Sementara itu, segmentasi berdasarkan negara mayoritas dalam masing-masing cluster memperlihatkan bahwa sebagian besar cluster memiliki komposisi negara yang homogen. Cluster 1, 2, 3, dan 5 secara eksklusif didominasi oleh pengguna dari satu negara, yaitu *Mexico*, *Germany, UK,* dan *Canada* dengan persentase 100%. Namun, cluster 0 dan 4 menunjukkan karakteristik yang lebih beragam. Cluster 0 memiliki mayoritas pengguna dari India sebesar 25,21%, menandakan bahwa cluster ini terdiri dari berbagai negara. Cluster 4 juga menunjukkan keberagaman, dengan hanya 50,67% pengguna berasal dari Brazil. Secara keseluruhan, hasil ini mengindikasikan bahwa preferensi genre film sangat dipengaruhi oleh rentang usia, sedangkan karakteristik berdasarkan negara sebagian besar terklaster secara jelas, meskipun terdapat beberapa kelompok dengan komposisi negara yang lebih bervariasi. Secara visualisasi, dapat dilihat pada chat berikut.



Gambar 4. *Chart Bar* hasil *KMeans* clustering berdasarkan asal Negara pengguna.

Pembahasan

Setelah menghasilkan cluster diatas, maka akan dibentuk sistem rekomendasi film dengan menggunakan demografi pengguna, yaitu rentang usia dan negara asal pengguna. Hasil rekomendasi nya tertera dalam tabel berikut.

Copyright © 2026, Jurnal Statistika dan Sains Data, Online ISSN: 3025-9649

Tabel. 3 Hasil Sistem Rekomendasi Film berbasis Demografi Pengguna Netflix

Cluster	Rentang Usia (Tahun)	Rekomendasi Genre Film	Presentase Penonton Genre(%)	Negara Mayoritas Penonton Genre	Presentase Penontom dari Negara tsb(%)
0	22–24	*Action*	100	India	25.21
1	30–35	*Romance*	50	*Mexico*	100
2	41–45	*Drama*	100	*Germany*	100
3	19–21	*Comedy*	100	*UK*	100
4	27–29	*Action*	100	Brazil	50.67
5	50–53	*Horror*	100	*Canada*	100

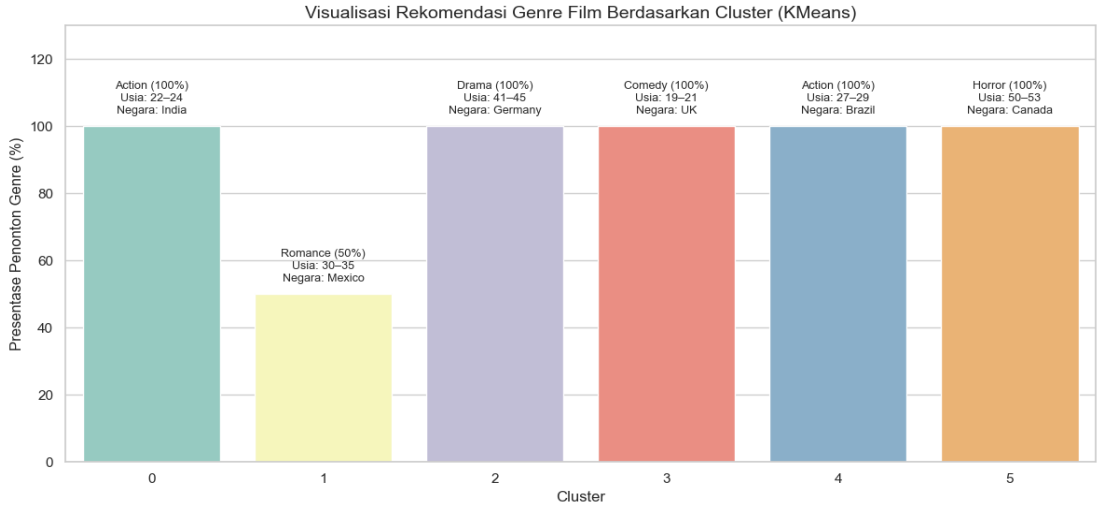
Berdasarkan hasil sistem rekomendasi berbasis cluster, setiap kelompok pengguna memiliki karakteristik demografis dan preferensi genre film yang berbeda-beda. Pengguna yang termasuk dalam Cluster 0 umumnya berada pada rentang usia 22–24 tahun dengan genre favorit *Action*, yang disukai oleh seluruh anggota cluster (100%). Negara mayoritas dari cluster ini adalah India, meskipun hanya mencakup sekitar 25,21% dari total anggota cluster. Sehingga sistem secara otomatis dapat merekomendasikan film bergenre *"Action"* bagi pengguna yang berusia 22-24 tahun maupun bagi para pengguna yang berlokasi di India.

Sementara itu, Cluster 1 terdiri dari pengguna berusia 30–35 tahun, dengan genre *Romance* sebagai pilihan terbanyak meski hanya mencapai 50% dari keseluruhan. Negara dengan representasi tertinggi dalam cluster ini adalah *Mexico*, yang mencakup seluruh anggota (100%). Sehingga sistem secara otomatis dapat merekomendasikan film bergenre *"Romance"* bagi pengguna yang berusia 30-35 tahun maupun bagi para pengguna yang berlokasi di *Mexico*.

Pada Cluster 2, pengguna didominasi oleh usia 41–45 tahun dan memiliki preferensi kuat terhadap genre *Drama*, dengan tingkat kesukaan sebesar 100%. Negara mayoritas dalam cluster ini adalah *Germany*, juga mencakup 100% pengguna. Sehingga sistem secara otomatis dapat merekomendasikan film bergenre *"Drama"* bagi pengguna yang berusia 41–45tahun maupun bagi para pengguna yang berlokasi di *Germany*. Cluster 3 berisi pengguna yang masih muda, dengan usia 19–21 tahun, dan menunjukkan minat penuh terhadap genre *Comedy* (100%). Semua anggota cluster ini berasal dari *UK*, menjadikannya sebagai negara dominan dalam kelompok tersebut. Sehingga sistem secara otomatis dapat merekomendasikan film bergenre *"Comedy"* bagi pengguna yang berusia 19–21 tahun maupun bagi para pengguna yang berlokasi *UK*.

Selanjutnya, Cluster 4 mencakup pengguna dengan usia 27–29 tahun, di mana genre *Action* kembali menjadi favorit seluruh anggota (100%). Negara yang paling banyak muncul dalam cluster ini adalah Brazil, dengan persentase sebesar 50,67%. Sehingga sistem secara otomatis dapat merekomendasikan film bergenre *"Action"* bagi pengguna yang berusia 27–29 tahun maupun bagi para pengguna yang berlokasi di Brazil. Terakhir, Cluster 5 terdiri dari pengguna berusia 50–53 tahun, yang secara keseluruhan menyukai genre *Horror* (100%). Negara yang mendominasi kelompok ini adalah *Canada*, yang juga mencakup 100% dari anggotanya. Sehingga sistem secara otomatis dapat merekomendasikan film bergenre *"Horror"* bagi pengguna yang berusia 50–53 tahun maupun bagi para pengguna yang berlokasi di *Canada.* Secara visualisasi, hasil fungsi sistem rekomendasi dapat dilihat pada gambar *Chart Bar* berikut.

Copyright © 2026, Jurnal Statistika dan Sains Data, Online ISSN: 3025-9649



Gambar 5. *Chart Bar* hasil Sistem Rekomendasi Film

SIMPULAN

Penelitian ini berhasil mengembangkan sebuah sistem rekomendasi film yang didasarkan pada data demografis pengguna Netflix, dengan menerapkan algoritma *K-Means Clustering* sebagai metode segmentasi utama. Tujuan dari pendekatan ini adalah untuk mengelompokkan pengguna ke dalam klaster-klaster yang memiliki kesamaan karakteristik, khususnya berdasarkan usia dan asal negara, sehingga dapat diidentifikasi pola preferensi genre film yang dominan dalam tiap kelompok pengguna tersebut. Dengan menggunakan atribut demografis ini, sistem rekomendasi mampu memberikan saran film yang lebih kontekstual dan relevan sesuai dengan latar belakang pengguna.

Hasil proses klasterisasi menunjukkan bahwa variabel usia memberikan pengaruh yang cukup signifikan terhadap preferensi genre film, di mana kelompok usia tertentu cenderung menyukai genre film yang spesifik. Sementara itu, data negara pengguna juga turut memberikan kontribusi dalam membentuk klaster, meskipun dalam banyak kasus ditemukan bahwa pengelompokan berdasarkan negara cenderung lebih homogen. Analisis ini diperkuat dengan penggunaan metode evaluasi seperti *Elbow Method* dan *Silhouette Score*, yang secara konsisten menunjukkan bahwa pembagian terbaik dari data pengguna adalah ke dalam enam klaster yang berbeda.

Setiap klaster yang dihasilkan menunjukkan kecenderungan yang konsisten terhadap genre tertentu, seperti misalnya dominasi genre Action pada kelompok usia muda, atau genre Drama dan Horror pada kelompok usia yang lebih tua. Temuan ini mendukung asumsi bahwa preferensi hiburan sangat dipengaruhi oleh faktor-faktor demografis. Sistem rekomendasi berbasis demografi ini memiliki keunggulan dalam menangani masalah *cold-start*, yakni kondisi di mana pengguna baru belum memiliki riwayat aktivitas atau penilaian yang cukup untuk digunakan dalam sistem rekomendasi berbasis perilaku.

Secara keseluruhan, pendekatan demografis dalam sistem rekomendasi ini terbukti efektif dalam meningkatkan akurasi dan relevansi saran film yang diberikan kepada pengguna. Hasil dari penelitian ini memberikan kontribusi penting bagi pengembangan sistem rekomendasi di masa depan, dan menunjukkan potensi besar pendekatan ini untuk digunakan sebagai solusi alternatif atau pelengkap bagi sistem berbasis konten (*content-based filtering*) dan sistem kolaboratif (*collaborative filtering*), terutama dalam ekosistem layanan streaming global seperti Netflix yang memiliki basis pengguna yang sangat beragam.

Copyright © 2026, Jurnal Statistika dan Sains Data, Online ISSN: 3025-9649

DAFTAR PUSTAKA

Aggarwal, C. C. (2016). *Recommender Systems: The Textbook*. Springer.

Ariyus D., Manongga D., Sembiring I.. 2024. Pengantar Sistem Rekomendasi Teori Dan Implementasi. Purbalingga. Eureka Media Aksara.

Devi A. A., Dharmawan K., Tastrawati T.. 2023. Pengelompokan Saham Menggunakan K-Means Dalam Pembentukan Portofolio Optimal . E-Jurnal Matematika Vol. 12(4), November 2023, Pp. 302-308 Issn: 2303-1751.Doi: Https://Doi.Org/10.24843/Mtk.2023.V12.I04.P433

Gangothri V., Saranya S., Venkataraman D.. 2015. A Comprehensive Review Of Recommender Systems. International Journal Of Applied Engineering Research Number 6 (Pp. 13909-13919). Research India Publications. Coimbatore

Géron, A. (2019). *Hands-On Machine Learning With Scikit-Learn, Keras, And Tensorflow* (2nd Ed.). O'reilly Media.

Gunawardana, A., & Shani, G. (2015). *Evaluating Recommender Systems*. In Ricci, F., Rokach, L., & Shapira, B. (Eds.), *Recommender Systems Handbook* (Pp. 265–308). Springer.

Jain, A. K. (2010). Data Clustering: 50 Years Beyond K-Means. *Pattern Recognition Letters*, 31(8), 651–666.

Kaggle. (N.D.). *Netflix User Data*. Retrieved From Https://Www.Kaggle.Com/Datasets/Smayanj/Netflix-Users-Database

Maharani H., Gunawan A. F..(2014). Sistem Rekomendasi Mobil Berdasarkan *Demographic* Dan *Content-Based Filtering*. Jurnal Telematika, Vol. 9 No. 2. Bandung: Institut Teknologi Harapan Bangsa.

Muhima R. R., Kurniawan M., Wardhana R. S., Yudhana A., Sunardi, Rahmawati M. W., & Yuliastu E. G.. 2021. "*Kupas Tuntas Algoritma Clustering: Konsep Perhitungan Manual Dan Program"*. Yogyakarta: Andi.

Ricci, F., Rokach, L., & Shapira, B. (2015). *Recommender Systems Handbook*. Springer.

Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid To The Interpretation And Validation Of Cluster Analysis. *Journal Of Computational And Applied Mathematics*, 20, 53–65.

Safrizal. (2024). Sistem Rekomendasi Demografic Filtering Untuk Penilaian Perilaku Dan Kompetensi Pengemudi Transportasi Umum Nasional. Jurnal Umj. Jakarta : Universitas Muhammadiyah Jakarta.

Syakur, M. A., Khotimah, B. K., Rochman, E. M. S., & Satoto, B. D. (2018). Integration K-Means Clustering Method And Elbow Method For Identification Of The Bestcustomer Profile Cluster. Iop Conference Series: Materials Science And Engineering, 336(1). Https://Doi.Org/10.1088/1757899x/336/1/012017

Wahyudi M., Masitha, Saragih R., Solikhun.. 2020. "*Data Mining: Penerapan Algoritma K-Means Clustering Dan K-Medoids Clustering*". Jakarta : Yayasan Kita Menulis.

Copyright © 2026, Jurnal Statistika dan Sains Data, Online ISSN: 3025-9649