JIKOMTI:
Jurnal Ilmiah Ilmu Komputer dan Teknologi Informasi
e- ISSN: 3089-2996
Vol. 2 No. 2, Des 2025

Analisis Preferensi Pengguna terhadap Genre Film Menggunakan Eksplorasi Data pada Dataset *MovieLens*

Rodhiyah Desviana[1], Verdi Yasin[2]
[1]Universitas Lancang Kuning, Pekanbaru
[2]STMIK Jayakarta, Jakarta

E-mail:
rodhiyahdesviana.study@gmail.com[1*], verdiyasin29@gmail.com[2]

Abstract

This study aims to explore user preferences for movie genres based on rating behavior in the MovieLens dataset during the period of 2023 to 2024. Using exploratory data analysis (EDA) techniques, this research examines the distribution of ratings, genre popularity in terms of average and total ratings, and behavior of the most active users. The results of the study indicate that Drama, Action, and Comedy are the most frequently watched and rated genres by users in the MovieLens dataset. However, genres such as Film-Noir, War, and Western have the highest average ratings, despite having relatively fewer ratings. These findings suggest that user preferences are influenced not only by the popularity of genres but also by the satisfaction level with films in those genres. Furthermore, analysis of the most active users reveals variations in individual preferences, which can serve as a foundation for developing more personalized and accurate recommendation systems.

Keywords*: data science; exploratory data analysis; movie genres; movielens; user preferences*

Abstrak

Penelitian ini bertujuan untuk mengeksplorasi preferensi pengguna terhadap genre film berdasarkan data rating pengguna dari dataset MovieLens selama periode 2023 hingga 2024. Dengan menggunakan teknik *Exploratory Data Analysis* (EDA), penelitian ini menganalisis pola distribusi rating, genre favorit berdasarkan rata-rata dan jumlah rating, serta kecenderungan pengguna paling aktif dalam memberi ulasan. Hasil penelitian menunjukkan bahwa genre Drama, *Action*, dan *Comedy* merupakan genre yang paling sering ditonton dan dinilai oleh pengguna pada dataset MovieLens. Namun, genre *Film-Noir*, *War*, dan *Western* memiliki rata-rata rating tertinggi, meskipun jumlah ratingnya relatif lebih sedikit. Temuan ini mengindikasikan bahwa preferensi pengguna tidak hanya dipengaruhi oleh popularitas genre, tetapi juga oleh tingkat kepuasan terhadap film dalam genre tersebut. Selain itu, analisis terhadap pengguna paling aktif menunjukkan adanya variasi preferensi individu yang dapat menjadi dasar pengembangan sistem rekomendasi yang lebih personal dan akurat.

Kata kunci: data sains; analisis data eksplorasi; genre film; *movielens*; preferensi pengguna

1. PENDAHULUAN

Di era digital saat ini, konsumsi media hiburan, khususnya film, mengalami peningkatan yang signifikan seiring dengan hadirnya berbagai *platform streaming* dan layanan distribusi konten daring. Salah satu tantangan utama yang dihadapi oleh penyedia layanan tersebut adalah memahami preferensi

pengguna secara mendalam untuk memberikan rekomendasi film yang relevan dan personal. Genre film menjadi salah satu dimensi utama dalam pemetaan selera penonton, karena setiap individu cenderung memiliki ketertarikan terhadap genre tertentu seperti drama, aksi, komedi, romantis, horor, atau dokumenter.

Untuk mengkaji preferensi pengguna terhadap genre film, diperlukan pendekatan berbasis data yang mampu mengekstrak pola dan kecenderungan dari perilaku pengguna. Dataset MovieLens yang disediakan oleh GroupLens Research menjadi salah satu sumber data terbuka yang banyak digunakan dalam penelitian sistem rekomendasi. Dataset ini mencakup jutaan rating film yang diberikan oleh ribuan pengguna, serta dilengkapi dengan informasi genre dan waktu rating, sehingga sangat ideal untuk dianalisis menggunakan pendekatan eksplorasi data.

Beberapa penelitian sebelumnya telah memanfaatkan teknik pembelajaran mesin untuk mengklasifikasikan atau memprediksi genre film. Metode LSTM digunakan untuk memprediksi genre film di IMDb Indonesia yang dapat menghasilkan prediksi yang akurat dengan nilai RMSE di bawah 2,50 serta mampu menangkap pola dan tren dari data genre film di IMDb Indonesia [1]. Selanjutnya metode LSTM digunakan untuk mengklasifikasikan multi kelas dalam memprediksi genre film berdasarkan sinopsis. Hasil penelitian menunjukkan bahwa metode LSTM mampu mencapai tingkat akurasi 98% dan loss 5%[2]. Metode CNN digunakan untuk menganalisis preferensi genre film di kalangan Gen Z dan menunjukkan hasil bahwa metode tersebut dapat digunakan untuk memahami preferensi pengguna berdasarkan data visual dan teks, memberikan wawasan baru dalam analisis genre film[3]. Penelitian selanjutnya menggunakan algoritma *Naïve Bayes* dan *Random Forest* untuk menganalisis genre film, dan menunjukkan bahwa *Random Forest* memberikan akurasi yang lebih tinggi (84,1%) dibandingkan *Naïve Bayes* (68,2%)[4]. Algoritma *K-Means* digunakan untuk menganalisis preferensi penonton anime berdasarkan genre film yang mana hasil evaluasi menggunakan *confusion matrix* menunjukkan akurasi model sebesar 95%, sedangkan nilai *silhouette score* sebesar 0,285 yang mengindikasikan pemisahan *cluster* yang memadai[5].

Penelitian lain berfokus pada pengembangan sistem rekomendasi, di antaranya adalah penggunaan model *Neural Network* untuk mengevaluasi kinerja sistem rekomendasi berbasis *deep learning* pada dataset MovieLens. Model *Neural Network* dirancang untuk mempelajari hubungan kompleks antara pengguna dan film dengan memanfaatkan fitur-fitur seperti *embedding* pengguna dan film, serta lapisan *dense* untuk memprediksi rating. Penelitian ini menunjukkan bahwa sistem rekomendasi berbasis *deep learning* performa lebih baik dibandingkan metode tradisional[6].

Penelitian selanjutnya menggunakan metode *content-based filtering* untuk menganalisis elemen film seperti genre, sutradara, aktor utama, dan deskripsi dalam pembuatan sistem rekomendasi film Indonesia dengan tujuan untuk memberikan rekomendasi yang relevan kepada *user* berdasarkan daftar film yang telah mereka tonton dan sukai. Hasil pengujian menunjukkan bahwa teknik ini efektif dalam memberikan rekomendasi yang sesuai dengan preferensi pengguna dan menghasilkan kepuasan pengguna yang tinggi[7]. Terdapat penelitian lain yang menggunakan metode *K-Means Clustering* untuk menganalisis data dari berbagai kelompok preferensi film di Netflix sebagai rekomendasi film. Hasil penelitian ini menyimpulkan bahwa film-film *short movie* hingga film-film layar lebar yang bergenre *action* (33%) dan *comedy* (20%) berada di *cluster* 0. Film-film layar lebar hingg film biografi atau dokumenter yang bergenre *action* (37%) dan drama (19%) berada di *cluster* 1, dan yang berada di *cluster* 2 adalah film-film *series* atau *short movie* yang bergenre drama (32%), *crime* (24%), dan *action* (22%)[8].

Namun, sebagian besar penelitian tersebut lebih menitikberatkan pada model prediktif dan optimalisasi algoritma, bukan pada analisis eksploratif mendalam terhadap preferensi pengguna berdasarkan genre film secara deskriptif. Penelitian ini menawarkan

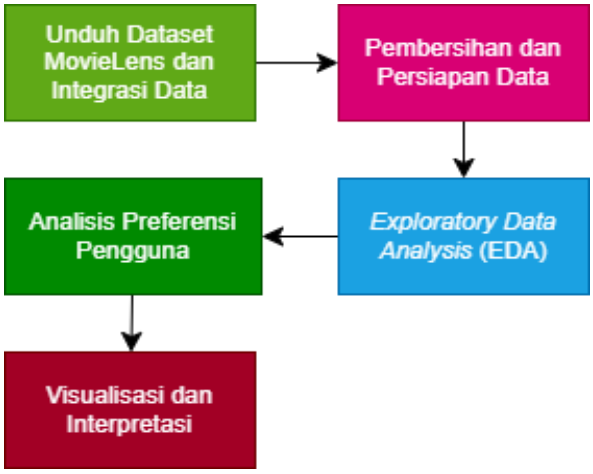
pendekatan yang berbeda dengan mengedepankan analisis eksploratif berbasis visualisasi dan statistik deskriptif, bukan prediksi semata.

Penelitian ini menggali preferensi pengguna terhadap genre film melalui EDA, dengan memanfaatkan informasi movie dan distribusi rating yang kaya dari MovieLens. Dengan demikian, penelitian ini diharapkan dapat memberikan kontribusi nyata dalam pengembangan sistem rekomendasi film yang lebih kontekstual dan personal, serta memperkaya pemahaman akademik mengenai perilaku penonton film digital.

Tujuan dari penelitian ini adalah untuk melakukan eksplorasi data terhadap preferensi pengguna terhadap genre film menggunakan dataset MovieLens dengan pendekatan *Exploratory Data Analysis* (EDA), guna mengidentifikasi pola rating, distribusi genre favorit.

2. METODE

Penelitian ini merupakan penelitian kuantitatif deskriptif yang bertujuan untuk menggambarkan pola preferensi pengguna terhadap genre film berdasarkan data dari MovieLens. Fokus penelitian adalah eksploratif, yaitu mendeskripsikan fenomena berdasarkan data yang tersedia tanpa menguji hipotesis formal. Data dalam penelitian ini diolah menggunakan bahasa pemograman *Python* melalui *Jupyter Notebook* dan menggunakan beberapa *library* di antaranya: *pandas, matplotlib, seaborn*. Tahapan dalam penelitian dapat dilihat pada Gambar 1 berikut.



Gambar 1. Tahapan Penelitian

2.1. Unduh Dataset MovieLens dan Integrasi Data

Tahap awal dari penelitian ini dimulai dengan mengunduh dataset MovieLens dari situs GroupLens[9]. Dataset MovieLens yang digunakan adalah *MovieLens Beliefs Dataset* 2024 yang memiliki beberapa *file* .csv. Penelitian ini hanya menggunakan dua *file* .csv yaitu ratings_for_additional_users dan movies. *File* movies.csv memiliki 3 kolom (movieId, title, genres) dan 105.071 data *record*. Sedangkan *file* ratings_for_additional_users.csv memiliki 4 kolom (userId, movieId, rating, tstamp) dengan jumlah data *record* sebanyak 4.185.688 data dan rentang waktu mulai dari tahun 1997 hingga 2024.

Karena perangkat yang digunakan untuk pengolahan data memiliki spesifikasi yang terbatas, maka data rating yang akan digunakan terlebih dahulu disaring (*filtering*) untuk mengurangi kompleksitas dan ukuran dataset. Data rating yang diambil dari *file* ratings_for_additional_users.csv adalah tahun 2023-2024. Hasil *filter* kemudian disimpan menjadi *file* .csv baru dengan nama filtered_ratings_2023_2024.csv. Setelah data disaring, selanjutnya adalah menggabungkan data filtered_ratings_2023_2024.csv dan movies.csv berdasarkan movieId.

2.2. Pembersihan dan Persiapan Data

Tahapan ini dilakukan agar data yang digunakan benar-benar bersih dan siap untuk digunakan. Pada tahapan pembersihan data,

dilakukan penghapusan data rating yang tidak valid, yakni menghapus data yang ratingnya -1.0. Kemudian menghapus data film yang tidak memiliki genre (no genres listed). Data setelah pembersihan dapat dilihat pada Gambar 2 berikut.

	userId	movieId	rating	tstamp	title	genres
0	393217	1	3.5	2023-01-25 19:45:46	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
1	393217	6	4.0	2023-02-07 21:17:19	Heat (1995)	Action|Crime|Thriller
2	393217	16	3.5	2023-01-25 19:50:40	Casino (1995)	Crime|Drama
3	393217	17	4.5	2023-01-25 19:49:45	Sense and Sensibility (1995)	Drama|Romance
4	393217	32	3.0	2023-01-25 19:46:01	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)	Mystery|Sci-Fi|Thriller
...	...	...	...	...	...	...
959216	393205	278702	2.0	2023-01-26 18:20:14	Glass Onion: A Knives Out Mystery (2022)	Comedy|Crime|Mystery
959217	393205	279562	3.0	2023-01-26 18:24:15	Hellraiser (2022)	Horror|Mystery|Thriller
959218	393205	279812	2.5	2023-01-25 17:08:16	The Menu (2022)	Comedy|Horror
959219	393205	286897	4.0	2024-02-08 15:36:05	Spider-Man: Across the Spider-Verse (2023)	Action|Adventure|Animation|Sci-Fi
959220	393205	290213	2.5	2024-02-08 15:34:46	The Equalizer 3 (2023)	Action|Crime|Thriller

Gambar 2. Dataset setelah pembersihan

Hasil data setelah pembersihan dan dapat digunakan adalah sebanyak 818.433 data dengan 6 kolom yakni userId, movieId, rating, tstamp, title, dan genres.

2.3. *Exploratory Data Analysis* (EDA)

Tahapan selanjutnya adalah eksplorasi data. Pada tahapan ini yang dilakukan membuat visualisasi distribusi rating pengguna, dan analisis genre. Visualisasi rating pengguna bertujuan untuk melihat kecenderungan rating yang diberikan pengguna.

Analisis genre bertujuan untuk melihat rata-rata dan jumlah rating per genre. Analisis genre dilakukan dengan memecah data genre film terlebih dahulu, karena satu film bisa memiliki banyak genre. Selanjutnya menghitung rata-rata rating untuk tiap genre, kemudian menghitung jumlah rating untuk tiap genre.

Rating pengguna dan analisis genre divisualisasikan dalam bentuk *bar chart*.

2.4. Analisis Preferensi Pengguna

Tujuan dari analisis ini adalah untuk mengidentifikasi kecenderungan dan pola preferensi pengguna terhadap berbagai genre film berdasarkan data rating dalam dataset MovieLens. Dengan mengevaluasi rating yang diberikan oleh pengguna pada film dengan genre tertentu, analisis ini bertujuan untuk mengetahui genre mana yang paling disukai maupun kurang diminati oleh pengguna secara umum. Informasi ini dapat dimanfaatkan dalam sistem rekomendasi film, pengembangan konten yang sesuai dengan selera mayoritas, serta strategi pemasaran berbasis genre favorit pengguna.

Pada analisis preferensi pengguna ini dilakukan penentuan 5 pengguna paling aktif berdasarkan jumlah rating yang diberikan, menghitung dan menampilkan 3 genre favorit dari 5 pengguna aktif berdasarkan jumlah rating. Visualisai hasil dalam bentuk *bar chart*.

2.5. Visualisasi dan Interpretasi

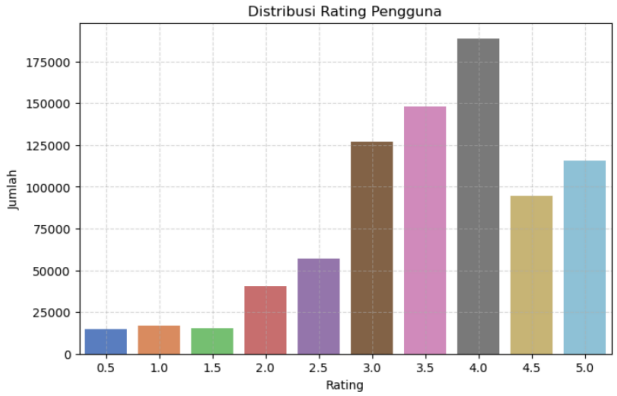
Visualisasi digunakan untuk menyajikan distribusi rating pengguna terhadap film dalam bentuk grafik yang mudah dipahami, seperti histogram atau diagram batang per genre. Melalui visualisasi ini, pola preferensi menjadi lebih terlihat, termasuk persebaran rating, kecenderungan terhadap genre tertentu, serta potensi anomali atau ketidakseimbangan data. Interpretasi dari hasil visualisasi membantu memberikan wawasan yang lebih jelas dan mendalam mengenai perilaku pengguna dalam memberikan rating berdasarkan genre film yang mereka tonton.

3. HASIL DAN PEMBAHASAN

Pada tahap ini, dilakukan eksplorasi data untuk menganalisis preferensi pengguna terhadap genre film berdasarkan dataset MovieLens. Analisis difokuskan pada distribusi rating yang diberikan oleh pengguna berdasarkan genre film. Tujuan dari tahap ini adalah untuk mengidentifikasi pola penilaian pengguna serta mengetahui genre mana yang paling disukai. Berikut adalah hasil dan pembahasannya:

3.1. Distribusi Rating Pengguna

Visualisasi distribusi rating pengguna dapat dilihat pada Gambar 3.

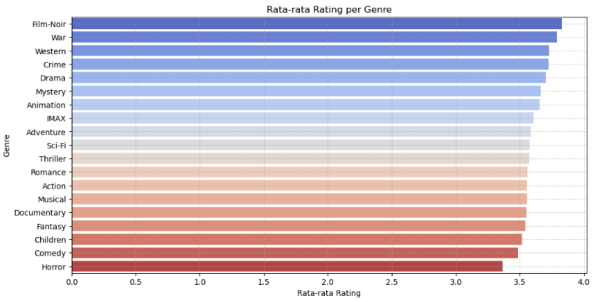


Gambar 3. Distribusi Rating Pengguna

Hasil analisis menunjukkan bahwa distribusi rating pengguna cenderung pada nilai 3.5 – 4.0. Hal ini menunjukkan bahwa sebagian besar pengguna cenderung memberikan penilaian positif terhadap film yang mereka tonton.

3.2. Rata-rata Rating per Genre

Rata-rata rating per genre didapat dengan cara mengambil nilai rata-rata dari setiap genre berdasarkan rating yang diberikan pengguna. Rata-rata rating digunakan untuk mengukur seberapa tinggi nilai yang diberikan pengguna secara rata-rata untuk genre tertentu.



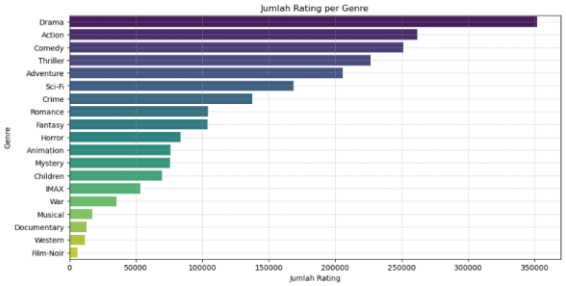
Gambar 4. Rata-rata Rating per Genre

Gambar 4 menampilkan rata-rata rating yang diberikan pengguna terhadap setiap genre film. Dari grafik tersebut terlihat bahwa genre *Film-Noir* memiliki rata-rata rating tertinggi yaitu sebesar 3,83, diikuti oleh *War* dan *Western*. Meskipun jumlah penonton atau rating pada genre-genre ini relatif rendah, nilai rata-rata yang tinggi menunjukkan bahwa film dalam genre tersebut memiliki kualitas yang diapresiasi oleh penonton setianya.

Sebaliknya, genre yang lebih populer seperti *Comedy* dan *Horror* justru memiliki rata-rata rating yang lebih rendah, yang mengindikasikan adanya perbedaan selera atau kepuasan pengguna terhadap film dalam genre tersebut. Rata-rata rating ini mencerminkan persepsi pengguna terhadap kualitas konten, terlepas dari seberapa sering genre tersebut ditonton.

3.3. Jumlah Rating per Genre

Jumlah rating per genre digunakan untuk mengukur seberapa sering genre tertentu ditonton dan dinilai oleh pengguna. Jumlah rating per genre divisualisasikan dalam bentuk *bar chart* seperti yang ditampilkan pada Gambar 5 berikut.



Gambar 5. Jumlah Rating per Genre

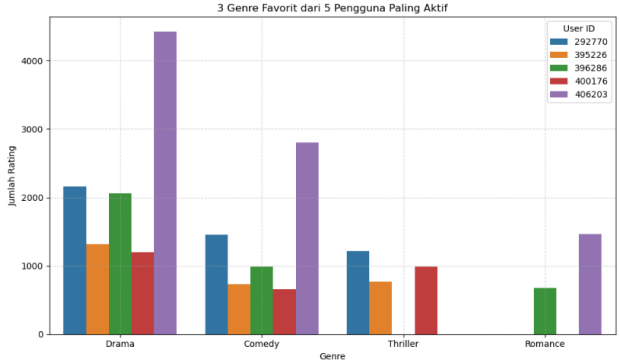
Pada Gambar 5 terlihat bahwa genre Drama menempati urutan teratas dengan jumlah rating lebih dari 350.000, diikuti oleh *Action*, *Comedy*, dan *Thriller*. Angka ini menunjukkan bahwa genre-genre tersebut merupakan yang paling sering ditonton dan dinilai oleh pengguna, menandakan tingkat popularitas yang tinggi.

Di sisi lain, genre seperti *Film-Noir*, *Western*, dan *Documentary* memiliki jumlah rating yang jauh lebih sedikit, yang kemungkinan disebabkan oleh jumlah film yang terbatas atau segmentasi penonton yang lebih spesifik. Grafik ini menggambarkan tingkat eksposur atau jangkauan tiap genre di kalangan pengguna secara umum.

3.4. Preferensi Pengguna

Analisis preferensi pengguna bertujuan untuk mengidentifikasi genre film yang paling disukai oleh pengguna, khususnya mereka yang aktif dalam memberikan rating. Selain itu, analisis ini juga dapat memberikan wawasan

mengenai pola minat individu terhadap genre tertentu, yang berguna dalam memahami karakteristik pengguna aktif dan mendukung sistem rekomendasi yang lebih personal. Hasil analisis preferensi pengguna terhadap genre film disajikan dalam bentuk grafik seperti yang terlihat pada Gambar 6.



Gambar 6. Genre Favorit dari 5 Pengguna

Berdasarkan grafik, terlihat bahwa Drama menjadi genre yang paling dominan di antara seluruh pengguna, dengan jumlah rating tertinggi diberikan oleh pengguna dengan ID 406203, yaitu sebanyak 4.421 rating. Genre *Comedy* dan *Thriller* juga muncul secara konsisten dalam tiga besar pada hampir semua pengguna, menunjukkan bahwa ketiganya merupakan genre yang paling sering ditonton dan dinilai oleh pengguna aktif.

Di samping itu, genre *Romance* muncul sebagai salah satu preferensi utama bagi dua pengguna, yaitu 396286 dan 406203, dengan jumlah rating yang cukup tinggi dibandingkan genre lainnya di luar tiga besar.

Temuan ini menunjukkan bahwa pengguna paling aktif cenderung memiliki preferensi terhadap genre-genre umum seperti Drama, *Comedy*, dan *Thriller*, meskipun terdapat variasi individu yang menunjukkan minat terhadap genre lain seperti *Romance*.

4. KESIMPULAN

Penelitian ini berhasil mengeksplorasi preferensi pengguna terhadap genre film menggunakan data rating MovieLens dari tahun 2023–2024. Hasil analisis menunjukkan bahwa genre drama, *action*, dan *comedy* yang paling sering ditonton dan dinilai oleh pengguna pada dataset MovieLens. Namun, genre *Film-Noir*, *War*, dan *Western* memiliki rata-rata rating tertinggi. Berdasarkan hasil tersebut, dapat disimpulkan bahwa preferensi pengguna terhadap genre film menunjukkan adanya perbedaan antara genre yang populer berdasarkan jumlah rating dan genre yang memiliki tingkat kepuasan tinggi berdasarkan rata-rata rating. Oleh karena itu, penelitian selanjutnya disarankan untuk menggabungkan analisis preferensi dengan faktor-faktor tambahan seperti demografi pengguna, waktu pemberian rating, serta analisis konten dan sentimen ulasan film agar dapat memberikan pemahaman yang lebih mendalam dan rekomendasi yang lebih personal serta akurat.

5. DAFTAR PUSTAKA

[1] D. Novi, M. Tampubolon, V. Vincensia Hulu, R. Oktavianus Sipahutar, and O. Sihombing, "Analisis Prediksi Genre Film Pada Internet Movie Database Indonesia Menggunakan Metode Long Short Term Memory," *Jurnal TEKINKOM*, vol. 6, no. 2, p. 2023, 2023, doi: 10.37600/tekinkom.v6i2.925.

[2] R. Amelia and D. B. Santoso, "Prediksi Genre Film Dengan Klasifikasi Multi Kelas Sinopsis Menggunakan Jaringan LSTM," *Journal of Information Technology and Computer Science (INTECOMS)*, vol. 6, no. 2, pp. 771–779, 2023.

[3] J. Kindly Susanto, B. Nugraha, A. Hidayat, S. Eka Prastya, and U. Sari Mulia Banjarmasin, "Movie Genre Product Convolutional Neural Network Impact For Gen Z," *INNOVATIVE: Journal Of Social Science Research*, vol. 4, 2024.

[4] N. Novenrodumetasa, I. M. A. D. Suarjaya, and I. M. S. Raharja, "Analisis Genre Film Berdasarkan Data Subtitle," 2023.

[5] N. Zalzabila and R. Prathivi, "Analisis Preferensi Penonton Anime berbasiskan Genre Film menggunakan Metode K-

Means," *KESATRIA: Jurnal Penerapan Sistem Informasi (Komputer & Manajemen)*, vol. 6, no. 1, pp. 235–241, 2025.

[6] D. Laras and H. Hasrullah, "Analisis Kinerja Sistem Rekomendasi Film Berbasis Deep Learning Menggunakan Model Neural Network Pada Dataset Movielens," *Jurnal Locus Penelitian dan Pengabdian*, vol. 4, no. 1, pp. 1047–1054, Jan. 2025, doi: 10.58344/locus.v4i1.3768.

[7] A. Zakharia, A. D. Ulhaq, A. B. Suryono, N. C. Nugroho, D. F. Hafith, and N. D. A. Gusmao, "Sistem Rekomendasi Film Indonesia Menggunakan Metode Content-Based Filtering," *Jurnal Ilmu Komputer dan Pendidikan*, vol. 2, pp. 671–678, 2024, [Online]. Available: https://journal.mediapublikasi.id/index.php/logic

[8] C. Budihartanti *et al.*, "Pengelompokkan Film Pada Platform Netflix Menggunakan Metode K-Means Clustering Sebagai Rekomendasi Film," *Journal of Information System Research (JOSH)*, vol. 5, no. 4, pp. 1392–1402, 2024, doi: 10.47065/josh.v5i4.5482.

[9] GroupLens, "https://grouplens.org/datasets/movielens/ml_belief_2024/."