Journal of Digital Ecosystem for Natural Sustainability (JoDENS)
Vol.5, No.2, Desember 2025, pp. 86~92
e-ISSN: 2798-6179
DOI : 10.63643/jodens.v5i2.326

Prediksi Kelompok Usia Pengguna Netflix Menggunakan Metode Random Forest Berdasarkan Analisis Genre Tontonan dan Perilaku Pengguna

Abelina Stevie Maria Trafin [1], Masparudin [2]*, Eka Lia Febrianti [3], Ilwan Syafrinal[4]
[1,2,3,4]*Teknik Perangkat Lunak, Universitas Universal*
**Corresponding author E-mail: masparudin.mahmud@gmail.com*

Article Info

Article history:

Received 01-12-2025
Revised 16-12-2025
Accepted 28-12-2025

Keyword:

Netflix, Random Forest, SMOTE, Klasifikasi Usia, Perilaku Pengguna

ABSTRACT

The accuracy of user demographics, particularly age, on video streaming platforms is often compromised by the widespread practice of shared accounts. This study addresses this challenge by implicitly classifying user age groups (Youth, Young Adult, Adult, Middle-Aged, Senior) based solely on behavioral data, including viewing genre frequency, sentiment analysis of reviews, and expenditure patterns. The core methodology employs a Random Forest Classifier optimized with SMOTE (Synthetic Minority Over-sampling Technique) to mitigate the severe class imbalance present in the dataset. The initial Baseline Model performed poorly, achieving only 40,13% accuracy and failing to identify minority classes. After implementing SMOTE and hyperparameter tuning, the Final Model demonstrated significant improvement, achieving an Accuracy of 79,26%. The engineered feature, Spend per Person, was identified as the most dominant predictor, validating the approach of using economic factors to differentiate genuine individual usage. Crucially, the model showed exceptional reliability in detecting sensitive age segments, such as Youth (F1-Score 0,88) and Seniors (F1-Score 0,75). This research provides an effective data-driven solution for enhancing age-based content personalization and parental control features.

I. PENDAHULUAN

Saat ini, industri digital berkembang dengan pesat, terutama layanan *platform streaming video*. Salah satu platform *streaming* yang paling populer adalah Netflix [1]. Netflix menyediakan beragam jenis tontonan yang relevan dan dapat memberikan rekomendasikan tontonan kepada pengguna berdasarkan riwayat tontonan dan juga pilihan genre [2]. Memberikan layanan terbaik menjadi faktor yang penting untuk menarik perhatian pengguna, terlebih dengan merekomendasikan konten konten yang relevan bagi pengguna[1].

Namun, sistem rekomendasi ini masih memiliki keterbatasan mengenai detail data pengguna terutama untuk data usia. Keterbatasan sistem rekomendasi ini disebabkan dari data pengguna yang tidak relevan. Keterbatasan rekomendasi sangat berdampak bagi kelompok usia yang masih di bawah umur dan juga sangat tidak relevan bagi beberapa kelompok usia lainnya. Hal ini bisa terjadi ketika pengguna mendaftarkan akun *email* atau data diri dari anggota keluarganya, dan bisa saja pengguna menggunakan akun berbagi (*shared account*), yang biasanya terjadi dalam penelitian klasifikasi pengguna media sosial [3]. Data-data yang tidak relevan ini membuat sistem kesulitan dalam mengkategorikan tayangan mana yang layak untuk disajikan sesuai dengan kelompok usia tertentu [4]. Hal ini dapat menghambat keakuratan penyelarasan data penonton berdasarkan pilihan genre secara *online* [5].

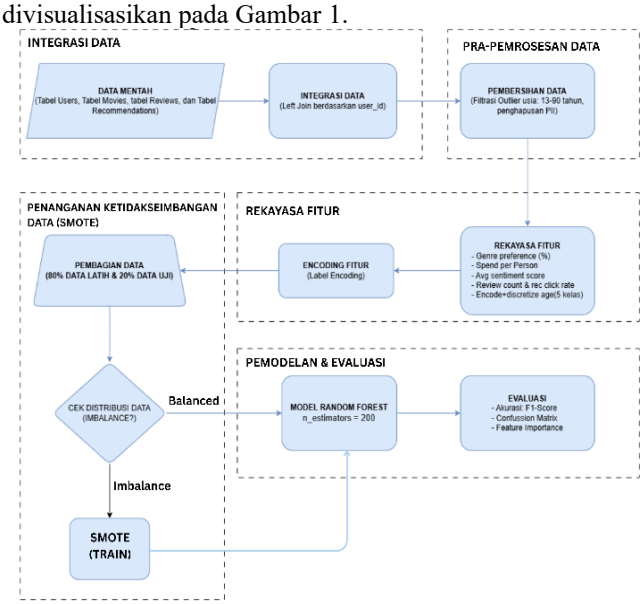
Di sisi lain, Netflix memiliki data yang sangat banyak, yang dimana data ini bisa diolah untuk mengetahui pengelompokkan usia pengguna berdasarkan genre tontonan dan perilaku pengguna [2], [6]. Namun, analisis awal terhadap data perilaku ini juga menunjukkan adanya

ketidakseimbangan kelas *(imbalanced data)* yang ekstrem di mana kelompok usia tertentu lebih dominan [7], [8]. Oleh karena itu, penelitian ini bertujuan untuk memprediksi kelompok usia pengguna dari genre tontonan dan perilaku pengguna, sekaligus mengatasi masalah ketidakseimbangan data yang dapat menyebabkan bias model[7], [8].

Agar mendapat hasil yang akurat dalam menganalisis data perilaku pengguna yang rumit dan berskala besar, penelitian ini menggunakan metode Random Forest yang dioptimasi dengan teknik Synthetic Minority Over-sampling Technique (SMOTE)[2]. Metode Random Forest merupakan algoritma *ensemble learning* yang bisa menangani data dengan jumlah banyak dan kompleks [2], [9]. Metode ini digabungkan dengan SMOTE untuk memastikan model dapat mempelajari pola dari semua kelompok usia secara merata, sehingga mampu menunjukkan keakuratan prediksi dan pengolahan data yang beraspek tinggi[7].

II. METODE

Penelitian ini mengikuti kerangka kerja sistematis (*framework*) yang dirancang untuk mengubah data mentah yang terfragmentasi menjadi model prediksi kelompok usia pengguna Netflix yang akurat. Tahapan penelitian divisualisasikan pada Gambar 1.



Gambar 1. Diagram Alur penelitian (*Research Flowchart*)

Secara umum, metode penelitian ini terdiri dari 5 fase utama, yaitu integrasi data (*Data Integration*), pra-pemprosesan data (*Pre-processing*), rekayasa fitur (*Feature Engineering*), penanganan ketidakseimbangan data (SMOTE) (*Handling Imbalance Data*), pemodelan dan evaluasi (*Modeling and Evalution*).

A. Integrasi Data (*Data Integration*)

Sumber data yang digunakan dalam penelitian ini adalah data sekunder dari repositori Kaggle dengan judul "Netflix 2025: User Behavior". Dataset ini dapat diakses melalui tautan https://www.kaggle.com/datasets/sayeeduddin/netflix-2025user-behavior-dataset-210k-records dan merepresentasikan aktivitas pengguna pada platform streaming Netflix [2], [6]. Data diketahui dalam kondisi terpisah, menjadi empat tabel entitas yang berbeda: Tabel Users, Tabel Movies, tabel Reviews, dan Tabel Recommendations. Agar mendapatkan profil pengguna yang utuh (360-degree user view), maka dilakukan tahapan Integrasi Horizontal menggunakan teknik *Left Join*. Atribut user_id digunakan sebagai kunci utama (primary key) untuk menggabungkan tabel users, reviews dan Recommendations. Di sisi lain, tabel Movies dikombinasikan dengan tabel Reviews menggunakan movie_id untuk menghubungkan pilihan genre yang ditonton oleh pengguna[4], [5].

B. Pra-pemprosesan Data (*Data Preprocessing*)

Data hasil pencampuran masih mengandung noise, inkonsistensi format, dan data sensitif. Tahapan ini bertujuan untuk membersihkan data agar memadai untuk dijadikan input model (model-ready). Dilakukan pembersihan otomatis untuk anomali yang terjadi pada data numerik, seperti koma(,) dan notasi ilmiah(misa: 5,7E + 15), dengan memastikan seluruh nilai kembali ke format *floating point* yang sesuai[4].

Agar sesuai dengan etika penambangan data (data mining ethics), atribut yang menandung informasi identitas pribadi(PII) dihapus guna menjaga anonimitas pengguna dan mencegah *overfitting* model[10]. Atribut yang meliputi: First Name, Last Name, Email, and User_ID.

Selanjutnya untuk menjaga validitas penelitian dan mencegah bias pada model prediksi, maka dilakukan Domain-based Filtering pada atribut usia (age) untuk menghilangkan impossible value (nilai negatif) dan extreme outlier (nilai usia > 90 tahun). batasan usia yang valid ditetapkan pada rentang 13 hingga 90 tahun[11], [12].

C. Rekayasa Fitur (*Feature Engineering*)

Tahapan ini difokuskan pada pengambilan dan menciptakan fitur baru (*derived features*) untuk mendukung klasifikasi perilaku. Variabel target awal berupa, usia numerik (age) diubah menjadi data kategorikal ordinal (age_group) untuk mengklasifikasikan pengguna ke dalam segmen pasar yang relevan[11], [4].Pembagian kelas usia ini dibagi ke dalam lima kelompok utama, yaitu Remaja yang mencakup rentang usia 13 hingga 17 tahun, Dewasa Muda untuk pengguna berusia 18 hingga 25 tahun, serta kelompok Dewasa yang berada pada rentang 26 hingga 35 tahun. Selanjutnya, kategori Paruh Baya mencakup usia 36 hingga 50 tahun, sedangkan kelompok Senior mencakup pengguna berusia 51 tahun ke atas. Pembagian ini digunakan sebagai dasar klasifikasi dalam proses analisis dan pemodelan. Setelah pengelompokan, atribut age (numerik) dihapus dari dataset untuk mencegah kebocoran data *(data leakage)*.

Rasio Pengeluaran per Orang (*Spend per Person*) merupakan fitur baru yang diciptakan dengan membagi *Monthly Spend* dengan *Household Size*[4]. Fitur ini bertujuan

mengukur daya beli riil individu, membedakan antara pengguna lajang dengan pengeluaran tinggi dan pengguna berkeluarga dengan pengeluaran terbagi. Selanjutnya, fitur perilaku (*Behavioral Features*) menyaring pola interaksi pengguna, meliputi rata-rata sentimen ulasan (*avg_sentiment*), frekuensi ulasan (*review_count*), dan rasio klik terhadap rekomendasi (*rec_click_rate*) [13], [14]. Selain itu, proses encoding dilakukan dengan mengubah seluruh atribut kategorikal nominal, seperti *Country, Primary Device*, dan *Genre*, menjadi format numerik menggunakan metode *Label Encoding* agar dapat diproses oleh algoritma Random Forest[15].

D. Penanganan Ketidakseimbangan Data (SMOTE)

Analisis awal terhadap distribusi data mengungkapkan adanya ketidakseimbagan kelas yang signifikan, ini dapat menyebabkan model lebih condong ke kelas mayoritas[7]. Untuk menghindari masalah ini, maka diterapkan teknik *Synthetic Minority Over-sampling Technique* (SMOTE)[16], [17]. SMOTE bekerja dengan membuat sample baru untuk kelas minoritas berdasarkan tetangga terdekat *(k-nearest neighbors)* di ruang fitur, sehingga distribusi kelas pada data latih menjadi seimbang (proporsional)[9].

E. Pemodelan dan Evaluasi

Model dibangun menggunakan algoritma *Random Forest Classifier*. Algoritma ini dipilih karena kemampuannya menangani data tabular dengan dimensi tinggi dan mampu memodelkan hubungan non-linear antar fitur melalui mekanisme *ensemble*[2], [9], [15]. Dataset dibagi menggunakan metode *Hold-out Split* dengan rasio 80% data latih dan 20% data uji *(testing set)*. Lalu dilakukan *Grid Search Cross-Validation* (CV=3) pada data latih untuk optimasi parameter Random Forest: n_estimators (100, 200), *max_depth* (None, 20, 30), dan *min_samples_split* (2, 5, 10). Kinerja dibandingkan antara Baseline Model (*Random Forest* tanpa SMOTE) dan SMOTE Model (*Random Forest* dengan SMOTE)[16].

Kinerja model diukur berdasarkan metrik-metrik berikut. Akurasi (Accuracy) digunakan untuk melihat proporsi prediksi yang benar secara keseluruhan, sedangkan *Confusion Matrix* memberikan gambaran detail mengenai distribusi prediksi pada setiap kelas. Selain itu, *F1-Score* (Makro/Tertimbang) digunakan sebagai metrik utama untuk mengevaluasi keseimbangan antara *Precision* dan *Recall* pada kelas minoritas [1], [18]. Analisis Feature Importance juga dilakukan untuk mengidentifikasi variabel yang berkontribusi terbesar melalui perhitungan Gini Importance.

III. HASIL DAN PEMBAHASAN

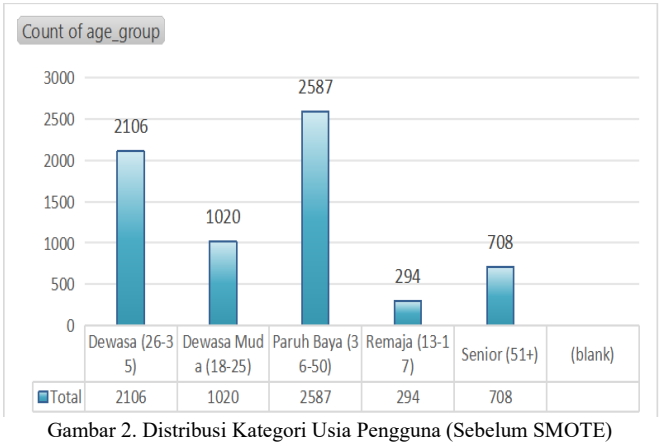
A. Pra-pemrosesan Data dan Statistik Deskriptif

Tahap awal penelitian dimulai dengan penggabungan empat tabel entitas terpisah (*Users, Movies, Reviews, Recommendations*) yang menghasilkan total 10.301 baris data mentah. Berdasarkan hasil eksplorasi data (*Exploratory Data Analysis*), terdapat beberapa anomali yang terdeteksi mengganggu keakuratan hasil prediksi [4], [6]. Temuan dan tindakan perbaikan yang dilakukan dirangkum dalam Tabel 1.

Tabel 1. Ringkasan Pra-pemrosan Data

Jenis Anomali	Temuan	Tindakan Perbaikan
Kesalahan Format Numerik	Muncul notasi ilmiah (5,71E+15) & desimal koma (,)	Konversi otomatis ke format float standar Python
Data Outlier (Usia)	Teridentifikasi usia negatif & ekstrem (>100 tahun)	Filtrasi domain-based (valid: 13-90 tahun)
Atribut Sensitif	Keberadaan kolom First Name, Last Name, Email	Penghapusan kolom untuk perlindungan privasi
Pengurangan Sampel	Data tidak valid/ekstrem yang dieliminasi berjumlah 3.586 baris	Eliminasi sampel (Sisa dataset bersih : 6715 baris)

Tabel 1 memperlihatkan secara rinci langkah-langkah kritis yang diambil untuk menjamin kualitas data sebelum masuk ke tahap pemodelan. Keputusan untuk mengeliminasi 3.586 baris data, yang setara dengan sekitar 35% dari total dataset awal, merupakan langkah vital untuk meminimalisir noise yang dapat mendistorsi proses pembelajaran algoritma Random Forest. Pengurangan volume data ini secara spesifik menargetkan anomali ekstrem yang tidak masuk akal secara statistik, seperti usia negatif atau format angka yang korup akibat kesalahan sistem. Selain itu, penghapusan atribut identitas pribadi (PII) dilakukan tidak hanya untuk mematuhi standar etika privasi data mining, tetapi juga untuk memastikan model murni mempelajari pola perilaku (behavioral patterns) dan bukan menghafal identitas unik pengguna yang dapat menyebabkan overfitting. Dengan data yang bersih, model dipaksa untuk mencari hubungan kausalitas yang sebenarnya antara genre tontonan dan usia. Setelah data dibersihkan, variabel target age dikelompokkan ke dalam lima kategori. Analisis pembagian kelas mengunkpkan ketidakseimbangan data yang cukup ekstrem *(imbalanced dataset)*[7], [8], seperti terlihat pada Gambar 2 berikut:



Gambar 2. Distribusi Kategori Usia Pengguna (Sebelum SMOTE)

Berdasarkan gambar 2, dataset didominasi oleh kelas Paruh Baya(36-50 tahun) dengan 2.587 pengguna, sedangkan kelas Remaja(13-17 tahun) hanya memiliki 294 pengguna. Ketidakseimban

gan rasio 1:8 ini berpotensi menyebabkan bias model jika tidak dilakukan teknik resampling[15].

B. Evaluasi Kinerja Model Klasifikasi

Pengujian model dilakukan dalam dua skenario utama untuk membuktikan efektivitas metode. yang diusulkan. Skenario pertama, yaitu Model Baseline (Tanpa Penyeimbangan Data), pada tahap eksperimen pertama algoritma Random Forest dilatih menggunakan dataset asli yang telah melalui pra-pemrosesan namun belum dilakukan penanganan ketidakseimbangan kelas (imbalanced class handling)[2], [15]. Hasil evaluasi model baseline ini disajikan secara lengkap pada tabel 2.

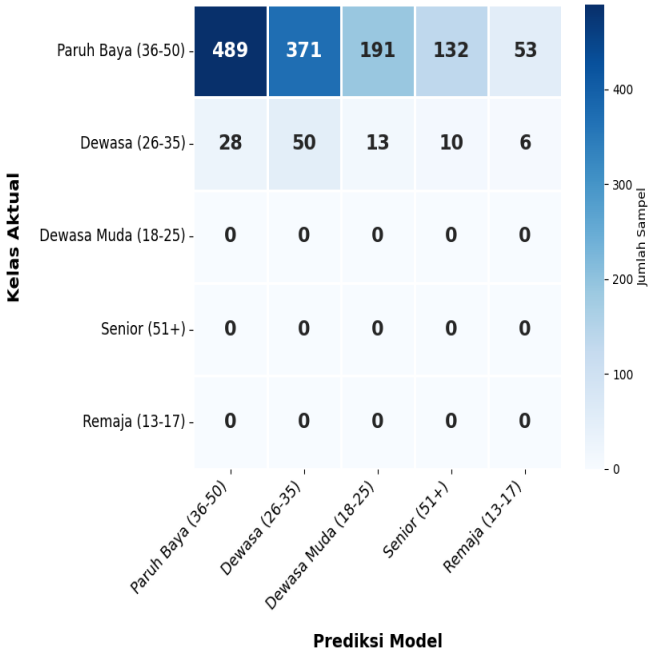
Tabel 2. Laporan Klasifikasi Model Baseline (Sebelum Optimasi)

Kategori Usia	Precision	Recall	F1-Score	Support
Paruh Baya (36–50)	0.38	0.40	0.39	1236
Dewasa (26–35)	0.11	0.47	0.18	107
Dewasa Muda (18–25)	0.00	0.00	0.00	0
Senior (51+)	0.00	0.00	0.00	0
Remaja (13–17)	0.00	0.00	0.00	0
Accuracy		0.40		1343
Macro Average	0.10	0.17	0.11	1343
Weighted Average	0.32	0.40	0.35	1343

Berdasarkan hasil evaluasi pada tabel 2 terlihat jelas bahwa model baseline mengalami bias mayoritas (majority class bias). Nilai akurasi total yang tercatat sebesar 40,13% menutupi kelemahan fundamental model dalam mengenali kelas minoritas. Hal ini terbukti model gagal total dalam mendeteksi kelas Dewasa Muda (18-25), Senior (51+), dan Remaja (13-17), yang ditunjukkan oleh nilai True Positive (0) pada diagonal Confusion Matrix.

Secara praktis, angka ini menunjukkan bahwa model gagal mendeteksi satu pun pengguna dari kelompok usia minoritas ini. Sebagian besar pengguna dari kelas ini salah diprediksi sebagai Paruh Baya atau Dewasa. Rendahnya nilai *F1-Score* pada kelas minoritas mengonfirmasi bahwa model *baseline* tidak layak digunakan untuk segmentasi pengguna muda dan lansia[3], [12], [13].

Secara lebih mendalam, fenomena *zero-recall* atau nilai 0.00 pada kolom *Recall* untuk kelas Dewasa Muda, Senior, dan Remaja di Tabel 2 mengindikasikan kegagalan fatal model dalam mengenali karakteristik unik kelompok tersebut. Dalam konteks industri *streaming*, ini berarti algoritma benar-benar 'buta' terhadap preferensi pengguna muda dan lansia. Model baseline terjebak dalam fenomena *accuracy paradox*, di mana ia terlihat memiliki akurasi 40% hanya karena menebak hampir semua data sebagai kelas mayoritas (Paruh Baya). Model ini tidak mempelajari fitur pembeda (diskriminan) antar kelas, melainkan hanya bermain aman dengan probabilitas statistik kelas terbanyak. Jika model mentah ini diterapkan pada sistem rekomendasi nyata, dampaknya adalah rekomendasi konten yang sangat bias dan tidak relevan bagi 60% populasi pengguna lainnya, yang berpotensi menurunkan retensi pelanggan secara signifikan.



Gambar 3 *Confusion Matrix base Model* (Sebelum SMOTE)

Analisis terhadap *Confusion Matrix* tabel 2 menunjukkan kegagalan deteksi total pada tiga kelas minoritas: Dewasa Muda (18-25), Senior (51+), dan Remaha (13-17). Nilai True Positif (sel diagonal) pada ketiga kelas tersebut adalah nol (0). Ini berarti model gagal mendeteksi satu pun pengguna dari kelompok usia minoritas ini. Sebagian besar data minoritas tersebut keliru diprediksi sebagai Paruh Baya, yang merupakan kelas mayoritas pada dataset asli. Secara praktis, model tidak layak digunakan untuk memprofilkan segmen Remaja dan Senior, yang sering menjadi target utama dalam strategi penargetan konten dan kontrol orang tua (parental control)[11], [12].

Rendahnya performa ini mengkonfirmasi hipotesis bahwa ketidakseimbangan distribusi data menjadi hambatan utama dalam klasifikasi demografi ini[7], [8]. Model cenderung "malas" dan bias memprediksi semua pengguna ke dalam kelas mayoritas demi meminimalkan error rata-rata, namun mengorbankan akurasi pada segmen pengguna spesifik[9], [19]. Oleh karena itu, penerapan teknik penyeimbang data, seperti SMOTE, adalah langkah mutlak yang diperlukan pada tahap eksperimen selanjutnya[16], [17].

Setelah kegagalan pada Model Baseline yang disebabkan oleh bias kelas mayoritas (Skenario 1), dilakukan langkah

perbaikan krusial yaitu penanganan ketidakseimbangan data menggunakan SMOTE. Teknik ini bertujuan menyeimbangkan distribusi kelas dengan menghasilkan sampel sintetis pada kelas minoritas, sehingga model memiliki data yang cukup untuk mempelajari pola pada semua kelompok usia[19].

Tabel 3. Laporan Klasifikasi Model Final

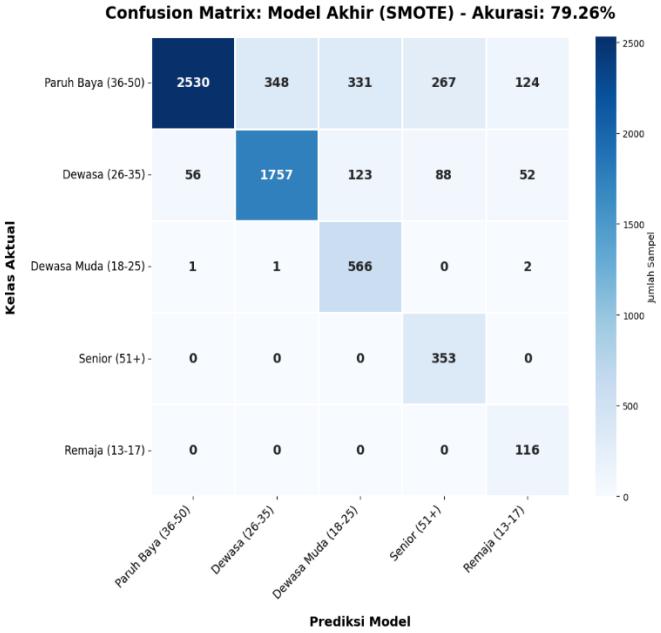
Kategori Usia	Precision	Recall	F1-Score	Keterangan
Remaja (13–17)	0.46	1.00	0.63	Baik
Senior (51+)	0.59	1.00	0.74	Baik
Dewasa Muda (18–25)	0.56	0.99	0.71	Baik
Dewasa (26–35)	0.84	0.85	0.84	Sangat Baik
Paruh Baya (36–50)	0.98	0.71	0.82	Sangat Baik
Accuracy			0.80	–
Weighted Average	0.91	0.66	0.77	Sangat Baik

Berdasarkan tabel 3, dapat dilihat bahwa model yang dikembangkan menunjukkan peningkatan performa yang sangat signifikan antar kelas usia, dengan akurasi keseluruhan mencapai 79,26%. Nilai ini menunjukkan bahwa model tidak lagi hanya mengandalkan prediksi kelas mayoritas, tetapi telah mampu mengenali pola perilaku pengguna dari berbagai kelompok usia dengan lebih seimbang[2], [15].

Model menunjukkan kinerja terbaik pada kelas Remaja (13–17 tahun) dan Senior (51+), yang tercermin dari nilai prediksi benar yang sangat tinggi. Pada kelas Remaja, seluruh data dapat diprediksi dengan benar (*True Positive* = 116), sedangkan pada kelas Senior seluruh data juga berhasil dikenali secara tepat (*True Positive* = 353). Hal ini menunjukkan bahwa model mampu membedakan kelompok usia ekstrem dengan sangat baik, kemungkinan besar karena perbedaan pola perilaku yang cukup kontras dibandingkan kelompok usia produktif.

Kelas Dewasa Muda (18–25 tahun) juga menunjukkan performa yang sangat baik dengan jumlah *True Positive* sebesar 566 dari total 570 data. Tingginya nilai ketepatan pada kelas ini mengindikasikan bahwa model berhasil mempelajari karakteristik khas pengguna dewasa muda, seperti intensitas interaksi dan pola konsumsi konten digital.Sebaliknya, tingkat kesalahan terbesar masih terjadi pada kelas Dewasa (26–35 tahun) dan Paruh Baya (36–50 tahun). Kesalahan klasifikasi silang antar kedua kelompok ini cukup tinggi, yang terlihat dari banyaknya data Dewasa yang diprediksi sebagai Paruh Baya dan sebaliknya. Fenomena ini dapat dimaklumi mengingat kedua kelompok usia tersebut memiliki kemiripan profil ekonomi, preferensi konten, dan perilaku konsumsi, sehingga batas klasifikasi menjadi lebih kabur dibandingkan kelas usia lain[14], [10].

Secara keseluruhan, hasil ini menunjukkan bahwa model telah berhasil mengurangi bias mayoritas dan mampu melakukan segmentasi usia secara lebih akurat, terutama pada kelompok usia yang sebelumnya sulit dikenali pada model baseline [1], [6], [18].



Gambar 4. Confusion Matrix Model Akhir

Berdasarkan Gambar 4, dapat dilihat bahwa model menunjukkan kemampuan klasifikasi yang sangat baik, khususnya dalam mengenali kelompok usia ekstrem, yaitu Remaja (13–17 tahun) dan Senior (51+). Seluruh data Remaja (13–17 tahun) yang diuji (116) dikelompokkan dengan benar dan baik. Menampilkan kemampuan model untuk mengenali pola remaja dengan akurat dan tidak ditemukan kesalahan dalam pengelompokkan. Dari 353 data Senior (51+) berhasil diprediksi dengan tepat dan tidak ditemukan *false positif* dan *false negative*. Dewasa Muda (18–25) dari 570 data, terdapat 566 data yang diprediksi dengan benar. Terdapat kesalahan sangat kecil dan tersebar ke kelas Remaja dan Paruh Baya. Dengan ini, model dapat menampilkan pola perilaku dewasa muda dapat dipelajari dengan baik.

C. Analisis Signifikansi Fitur (*Feature Importance*)

Analisis *Feature Importance* (Gini Importance) mengungkap fakta menarik bahwa variabel ekonomi dan perilaku jauh lebih dominan dibandingkan variabel demografis statis lainnya. Fitur Spend per Person (Rasio Pengeluaran per Orang) muncul sebagai prediktor terkuat. Temuan ini memvalidasi hipotesis bahwa 'daya beli' adalah proksi (perwakilan) terbaik untuk usia. Kelompok Paruh Baya dan Dewasa cenderung memiliki daya beli mandiri yang stabil dan seringkali menanggung biaya langganan untuk anggota keluarga lain (Household Size besar), sedangkan Remaja dan Dewasa Muda (Mahasiswa) cenderung memiliki pengeluaran individu yang lebih rendah atau bergantung pada akun berbagi.

Selain faktor ekonomi, fitur perilaku Review Count (Frekuensi Ulasan) dan Avg Sentiment (Sentimen Rata-rata) menunjukkan pola generasi yang distingtif. Hasil analisis data

memperlihatkan bahwa generasi yang lebih muda (Remaja dan Dewasa Muda) jauh lebih vokal dan aktif meninggalkan jejak digital berupa ulasan atau rating dibandingkan kelompok Senior yang cenderung menjadi *passive viewers*. Perbedaan gaya interaksi inilah yang ditangkap oleh Random Forest untuk memisahkan usia pengguna meskipun mereka menonton film yang sama. Terakhir, fitur Rec_Click_Rate membedakan tingkat literasi digital; kelompok usia muda cenderung lebih responsif terhadap fitur rekomendasi algoritma dibandingkan kelompok usia lanjut yang lebih sering menggunakan fitur pencarian manual. Kombinasi antara jejak finansial (kemampuan bayar) dan jejak interaksi (gaya penggunaan aplikasi) inilah yang memungkinkan model memprediksi usia dengan akurasi tinggi tanpa memerlukan data tanggal lahir yang eksplisit.

D. Pembahasan Temuan Penelitian

Model *Random Forest* yang dioptimasi dengan SMOTE memberikan solusi yang solid untuk memprediksi kelompok usia pengguna Netflix dengan akurasi 79.26%.

Keberhasilan model ini membuktikan bahwa pendekatan berbasis data perilaku (*behavioral data-driven*) efektif untuk menyimpulkan demografi pengguna secara implisit, mengatasi masalah ketidakakuratan data usia yang diinput secara langsung (misalnya, pada skenario *shared account*).

Signifikansi fitur perilaku dalam model ini menguatkan argumentasi bahwa prediksi demografi dapat dilakukan secara implisit dari jejak digital pasif, seperti genre tontonan dan perilaku interaksi.

IV. KESIMPULAN

Berdasarkan hasil analisis, pengujian model, dan pembahasan temuan, penelitian mengenai klasifikasi kelompok usia pengguna layanan streaming video dapat dilakukan secara efektif menggunakan pendekatan berbasis data perilaku. Penelitian ini menunjukkan bahwa informasi demografi pengguna, terlebih usia, bisa diprediksi secara tidak langsung melalui pola interaksi dan aktivitas online pengguna, sehingga mampu menekan ketidakakuratan data demografi yang sering terjadi akibat penggunaan data yang tidak valid.

Selain itu, permasalahan imbalanced data menjadi kelemahan utama pada model baseline dengan tingkat akurasi 40,31%, terlebih pada ketidakmampuan model dalam membedakan kelas minoritas seperti Remaja dan Senior, berhasil diatasi melalui penerapan metode SMOTE dan optimasi *hyperparameter*. Penerapan tersebut menghasilkan peningkatan performa yang signifikan pada model akhir, dengan akurasi mencapai 79.26%.

Model *Random Forest Classifier* yang telah dioptimasi juga menampilkan keakuratan tinggi dalam memprediksi kelompok usia ekstrem, dengan mencapai F1-Score sebesar 0.80 pada kelas Remaja (13–17 tahun) dan 0.75 pada kelas Senior (51+). Temuan ini berperan penting terhadap pengembangan fitur kontrol orang tua dan strategi pemasaran yang lebih efektif.

Lebih lanjut, hasil analisis fitur mengimplikasikan bahwa faktor ekonomi yang digambarkan oleh fitur *Spend per Person* menjadi prediktor paling dominan dalam membedakan kelompok usia, khususnya dalam konteks layanan streaming yang menerapkan skema akun keluarga. Tidak hanya itu, faktor perilaku seperti *avg_sentiment* dan *rec_impression_count* juga menunjukkan kontribusi yang memberikan dampak signifikan. Ini menandakan bahwa perbedaan gaya bahasa dalam ulasan serta tingkat eksplorasi aplikasi menjadi bentuk perbedaan perilaku yang kuat antar generasi.

DAFTAR PUSTAKA

[1] V. Shelake, S. Fernandes, and S. Shrungare, “AI-Driven Personalized Movie Recommendations: A Content and Sentiment-Aware Model for Streaming and Digital Entrepreneurship,” *Aptisi Trans. Technopreneursh.*, vol. 7, no. 2, Apr. 2025, doi: 10.34306/att.v7i2.550.

[2] A. S. Salsabila, C. A. Sari, and E. H. Rachmawanto, “Classification of Movie Recommendation on Netflix Using Random Forest Algorithm,” *Adv. Sustain. Sci. Eng. Technol.*, vol. 6, no. 3, p. 02403016, Jul. 2024, doi: 10.26877/asset.v6i3.676.

[3] M. Gollapalli *et al.*, “Machine Learning Approach to Users’ Age Prediction: A Telecom Company Case Study in Saudi Arabia,” *Math. Model. Eng. Probl.*, vol. 10, no. 5, pp. 1619–1629, Oct. 2023, doi: 10.18280/mmep.100512.

[4] F. Qiu and Y. Cui, “An analysis of user behavior in online video streaming,” in *Proceedings of the international workshop on Very-large-scale multimedia corpus, mining and retrieval*, New York, NY, USA: ACM, Oct. 2010, pp. 49–54. doi: 10.1145/1878137.1878149.

[5] B. Veloso, B. Malheiro, J. C. Burguillo, and ..., “Improving On-line Genre-based Viewer Profiling,” *TVX2017 Work. ...*, 2017, [Online]. Available: http://www.open-access.bcu.ac.uk/id/eprint/4829%0Ahttp://www.open-access.bcu.ac.uk/4829/1/WS4p1- Jeremy Foss.pdf

[6] B. H. Hayadi, “Clustering Netflix Shows Based on Features Using K-means and Hierarchical Algorithms to Identify Content Patterns,” *Int. J. Appl. Inf. Manag.*, vol. 5, no. 2, pp. 98–110, Jul. 2025, doi: 10.47738/ijaim.v5i2.102.

[7] A. Kulkarni, D. Chong, and F. A. Batarseh, “Foundations of data imbalance and solutions for a data democracy,” in *Data Democracy*, Elsevier, 2020, pp. 83–106. doi: 10.1016/B978-0-12-818366-3.00005-8.

[8] Anju Fauziah and Julan Hernadi, “Klasifikasi Data Tak Seimbang Menggunakan Algoritma Random Forest dengan SMOTE dan SMOTE-ENN,”

Teknomatika J. Inform. dan Komput., vol. 17, no. 2, pp. 38–47, Mar. 2025, doi: 10.30989/teknomatika.v17i2.1530.

[9] A. M. A. Rahim, Inggrid Yanuar Risca Pratiwi, and Muhammad Ainul Fikri, "Klasifikasi Penyakit Jantung Menggunakan Metode Synthetic Minority Over-Sampling Technique Dan Random Forest Clasifier," *Indones. J. Comput. Sci.*, vol. 12, no. 5, Nov. 2023, doi: 10.33022/ijcs.v12i5.3413.

[10] K. De Bock and D. Van den Poel, "Predicting Website Audience Demographics forWeb Advertising Targeting Using Multi-Website Clickstream Data," *Fundam. Informaticae*, vol. 98, no. 1, pp. 49–70, Jan. 2010, doi: 10.3233/FI-2010-216.

[11] L. S. R. and U. K., "Age Group Classification and Gender Prediction using Facial Skin Texture Analysis," *Int. J. Comput. Appl.*, vol. 186, no. 53, pp. 20–26, Dec. 2024, doi: 10.5120/ijca2024924208.

[12] R. Chew, C. Kery, L. Baum, T. Bukowski, A. Kim, and M. Navarro, "Predicting Age Groups of Reddit Users Based on Posting Behavior and Metadata: Classification Model Development and Validation," *JMIR Public Heal. Surveill.*, vol. 7, no. 3, p. e25807, Mar. 2021, doi: 10.2196/25807.

[13] Z. Anwer, S. Qureshi, S. M. Zeeshan Iqbal, A. Zia, and S. Anwer, "Predicting user behavior on video streaming by using watch-time duration analysis," *Knowledge-Based Syst.*, vol. 332, p. 114779, Jan. 2026, doi: 10.1016/j.knosys.2025.114779.

[14] E. M. Khan, M. S. H. Mukta, M. E. Ali, and J. Mahmud, "Predicting Users' Movie Preference and Rating Behavior from Personality and Values," *ACM Trans. Interact. Intell. Syst.*, vol. 10, no. 3, pp. 1–25, Sep. 2020, doi: 10.1145/3338244.

[15] S. Mahimkar and D. G. S. K. Lagan Goel, "Predictive Analysis of TV Program Viewership Using Random Forest Algorithms," *IJRAR-International J. Res. Anal. Rev. (IJRAR), E-ISSN 2348-1269, P-ISSN 2349*, vol. 5138, no. October 2021, pp. 309–322, 2021.

[16] V. Oktaviani, N. Rosmawarni, and M. P. Muslim, "Perbandingan Kinerja Random Forest Dan Smote Random Forest Dalam Mendeteksi Dan Mengukur Tingkat Stres Pada Mahasiswa Tingkat Akhir," *Inform. J. Ilmu Komput.*, vol. 20, no. 1, pp. 43–49, Apr. 2024, doi: 10.52958/iftk.v20i1.9158.

[17] M. Umer *et al.*, "Scientific papers citation analysis using textual features and SMOTE resampling techniques," *Pattern Recognit. Lett.*, vol. 150, pp. 250–257, Oct. 2021, doi: 10.1016/j.patrec.2021.07.009.

[18] P. Khant and B. Tidke, "Multimodal Approach to Recommend Movie Genres Based on Multi Datasets," *Indian J. Sci. Technol.*, vol. 16, no. 30, pp. 2304–2310, Aug. 2023, doi: 10.17485/IJST/v16i30.1238.

[19] N. Istiqamah and M. Rijal, "Klasifikasi Ulasan Konsumen Menggunakan Random Forest dan SMOTE," *J. Syst. Comput. Eng.*, vol. 5, no. 1, pp. 66–77, Jan. 2024, doi: 10.61628/jsce.v5i1.1061.