

## **Penerapan Ensemble Learning Untuk Klasifikasi Preferensi Film Pada Dataset MovieLens Dengan Penanganan Kelas Tidak Seimbang**

**Rodhiyah Desviana<sup>1</sup>, Susandri<sup>2</sup>, Syafrial<sup>3</sup>**

<sup>1,2,3</sup>Program Studi Magister Ilmu Komputer Universitas Lancang Kuning <sup>1,2,3</sup>Jl. Yos Sudarso KM. 8 Rumbai, Pekanbaru, Riau, telp.

0811 753 2015 e-mail: <sup>1</sup>rodhiyahdesviana.study@gmail.com,  
<sup>2</sup>susandri@unilak.ac.id, <sup>3</sup>syafrial.ab@gmail.com

### **Abstrak**

*Sistem rekomendasi film semakin dibutuhkan seiring bertambahnya jumlah konten digital dan meningkatnya kebutuhan pengguna akan rekomendasi yang sesuai dengan preferensi mereka. Penelitian ini berfokus pada klasifikasi preferensi pengguna terhadap film menggunakan data MovieLens, dengan pendekatan ensemble learning untuk meningkatkan performa prediksi. Salah satu tantangan utama dalam klasifikasi ini adalah ketidakseimbangan kelas antara data pengguna yang menyukai dan tidak menyukai film, yang dapat memengaruhi akurasi dan kestabilan model. Untuk mengatasi hal tersebut, digunakan teknik SMOTE (Synthetic Minority Over-sampling Technique) guna menyeimbangkan distribusi kelas secara efektif, serta dilakukan tuning hiperparameter menggunakan GridSearchCV untuk mendapatkan konfigurasi terbaik pada model. Model ensemble yang diuji meliputi Random Forest, XGBoost, dan Voting Classifier sebagai kombinasi dari kedua model tersebut. Hasil eksperimen menunjukkan bahwa Voting Classifier memberikan kinerja terbaik secara rata-rata, dengan skor ROC AUC sebesar 0.6466 dan F1 Score sebesar 0.5934. Performa terbaik secara individual juga dicapai oleh Voting Classifier pada test size 0.1 dan random state 2024, dengan ROC AUC sebesar 0.6540 dan F1 Score sebesar 0.5966. Temuan ini menunjukkan bahwa kombinasi SMOTE dan model ensemble yang dioptimalkan dapat secara signifikan meningkatkan akurasi klasifikasi preferensi pengguna, serta memberikan kontribusi praktis yang berarti bagi pengembangan sistem rekomendasi film yang lebih akurat dan responsif terhadap kebutuhan pengguna.*

**Kata Kunci:** Ensemble Learning, Klasifikasi Preferensi Film, Dataset MovieLens, SMOTE, Kelas Tidak Seimbang

### **Abstract**

*Movie recommendation systems are increasingly important as the volume of digital content grows and users demand personalized suggestions aligned with their preferences. This study focuses on classifying user preferences for movies using the MovieLens dataset, employing ensemble learning approaches to improve prediction performance. A major challenge addressed is the class imbalance between users who like and dislike movies, which can affect model accuracy and stability. To overcome this, the Synthetic Minority Over-sampling Technique (SMOTE) was applied to effectively balance class distribution, alongside hyperparameter tuning via GridSearchCV to optimize model configurations. The ensemble models evaluated include Random Forest, XGBoost, and a Voting Classifier that combines both. Experimental results demonstrate that the Voting Classifier achieved the best average performance, with a ROC AUC score of 0.6466 and an F1 Score of 0.5934. The best individual performance was also attained by the Voting Classifier at a test size of 0.1 and random state 2024, achieving a ROC AUC of 0.6540 and F1 Score of 0.5966. These findings indicate that the integration of SMOTE and optimized ensemble models significantly enhances the accuracy of user preference classification and provides valuable practical insights for developing more accurate and responsive movie recommendation systems.*

**Keywords:** Ensemble Learning, Film Preference Classification, MovieLens Dataset, SMOTE, Imbalanced Classes

## 1. PENDAHULUAN

Sistem rekomendasi telah menjadi komponen krusial dalam berbagai *platform* digital, termasuk layanan *streaming* film. Seiring bertambahnya jumlah konten, pengguna semakin membutuhkan bantuan sistem cerdas untuk menemukan film yang sesuai dengan preferensi mereka. Salah satu pendekatan yang umum digunakan dalam sistem rekomendasi adalah memanfaatkan data historis berupa rating film yang diberikan oleh pengguna. Ketika data preferensi ini digunakan dalam pendekatan klasifikasi, muncul tantangan berupa ketidakseimbangan kelas antara data pengguna yang menyukai dan tidak menyukai film. Ketidakseimbangan ini dapat menyebabkan penurunan performa model prediktif, terutama dalam mengenali kelas minoritas, serta menimbulkan bias dalam rekomendasi yang dihasilkan. Oleh karena itu, penanganan terhadap data tidak seimbang menjadi aspek penting dalam pengembangan sistem rekomendasi yang adil dan akurat.

Penelitian sebelumnya melakukan analisis sentimen menggunakan model CNN dan LSTM untuk mengklasifikasikan teks serta menggunakan teknik seperti TF-IDF, Word2Vec, dan SMOTEN untuk penyeimbangan data dalam menganalisis ulasan film. Hasil penelitian menunjukkan CNN mencapai akurasi yang sangat tinggi (98,56%) sedangkan LSTM tidak jauh berbeda (98,53%). Hasil F1-Score kedua model juga menunjukkan kinerja yang namun LSTM memiliki nilai yang lebih tinggi (78,92%) dibanding CNN (77,87%). Pendekatan analisis sentimen sangat efektif dalam mengekstraksi wawasan berharga dari ulasan film dan sangat membantu dalam membuat keputusan yang tepat tentang film yang akan ditonton[1]. Penelitian lain menggunakan kombinasi teknik augmentasi data dan *ensemble* yang terbaru maupun yang inovatif, termasuk penggunaan jaringan *adversarial generative* (*generative adversarial networks*, GANs) untuk mengatasi masalah keseimbangan kelas. Evaluasi dilakukan terhadap sembilan metode augmentasi dan sembilan metode ensemble pada sejumlah dataset benchmark yang umum digunakan, guna mengidentifikasi kombinasi terbaik yang mampu meningkatkan akurasi dan kebal terhadap ketidakseimbangan data. Hasil menunjukkan bahwa kombinasi metode augmentasi tradisional seperti SMOTE dan ROS dengan ensemble secara signifikan meningkatkan kinerja klasifikasi dan lebih efisien secara komputasi dibandingkan penggunaan GANs[2]. Penanganan ketidakseimbangan data juga dilakukan dalam membangun model deteksi dini untuk memprediksi peristiwa di pasar modal Indonesia dengan menggunakan lima model ensemble learning yaitu Random Forest, ExtraTrees, CatBoost, XGBoost, dan LightGBM. Hasil penelitian menunjukkan bahwa model ExtraTrees dengan SMOTE-ENN memiliki G-Mean tertinggi untuk prediksi satu hari yakni sebesar 96,88%, dan model LightGBM dengan SMOTE adalah model yang terbaik untuk prediksi lima hari dengan nilai G-Mean sebesar 89,21%[3].

Penelitian selanjutnya menggabungkan Naïve Bayes, Random Forest, SMOTE, dan XGBoost untuk mengeksplorasi dan meningkatkan akurasi dalam klasifikasi sentimen terkait pemindahan Ibu Kota Indonesia. Hasil penelitian menunjukkan kombinasi Random Forest dengan SMOTE efektif dalam meningkatkan akurasi prediksi sentimen, menyeimbangkan dataset, dan meningkatkan deteksi kelas minoritas dengan tingkat akurasi mencapai 91,25%[4]. Teknik penyeimbangan data juga dilakukan dalam penelitian menganalisis dampak teknik penyeimbangan data pada dataset diabetes PIMA dengan membandingkan XGBoost dan Random Forest. Terdapat enam skema penyeimbangan yang digunakan pada penelitian tersebut yaitu RUS, UPS, SMOTE, ADASyn, SMOTE-Tomek, dan SMOTEN. Pada penelitian ini, kombinasi XGBoost dan

SMOTE-Tomek menunjukkan kinerja terbaik, sementara Random Forest dan SMOTEN menunjukkan hasil yang baik[5]. Selanjutnya metode ensemble learning digunakan bersama dengan teknik Voting Classifier yakni menggabungkan beberapa model klasifikasi seperti Random Forest Classifier, Logistic Regression, dan XGBoost Classifier untuk mendeteksi sindrom ovarium polikistik atau Polycystic Ovary Syndrome (PCOS) yakni kelainan hormonal pada wanita. Hasil penelitian menunjukkan metode yang digunakan memberikan kinerja yang lebih baik dengan tingkat akurasi sebesar 95%, precision 100%, recall 85%, F1-Score 92%, dan AUC sebesar 94,34%[6].

Penelitian yang pernah dilakukan berikutnya menggunakan teknik oversampling seperti SMOTE, ADASyn, dan Random Oversampling untuk meningkatkan performa model XGBoost Classifier dalam memprediksi penyakit AIDS. Hasil menunjukkan teknik Random Oversampling dengan XGBoost Classifier memberikan performa terbaik dengan tingkat akurasi 94,44%, precision 90,72%, recall 98,74%, dan F1-Score sebesar 94,65%[7]. Model Random Forest dan XGBoost juga telah digunakan pada penelitian untuk mendeteksi penipuan E-Commerce menggunakan augmentasi data CGAN (Conditional Generative Adversarial Network). Hasil penelitian menunjukkan bahwa penggunaan augmentasi data dengan CGAN secara signifikan meningkatkan kinerja kedua model. Penerapan augmentasi meningkatkan hasil recall menjadi 95,15% untuk Random Forest dan 95,22% untuk XGBoost, dengan F1-score masing-masing 97,47% dan 97,42%, serta accuracy masing-masing 97,50% dan 97,42%. XGBoost menunjukkan sedikit keunggulan dalam precision dan recall dibandingkan Random Forest, terutama pada dataset yang telah diaugmentasi[8].

Penelitian sebelumnya telah mengembangkan model ensemble learning multistep terintegrasi yang menggabungkan berbagai teknik ensemble dengan metode penyeimbangan data untuk mendeteksi transaksi penipuan kartu kredit. Model tersebut menunjukkan performa yang sangat baik dan berpotensi digunakan dalam sistem deteksi penipuan berskala besar[9]. Penelitian lain mengevaluasi algoritma Random Forest dan XGBoost dengan berbagai teknik upsampling (SMOTE, ADASYN, GNUS) pada dataset dengan tingkat ketidakseimbangan yang bervariasi, dan menemukan bahwa XGBoost yang dikombinasikan dengan SMOTE secara konsisten memberikan performa terbaik[10]. Selain itu, terdapat pendekatan hybrid yang menggabungkan oversampling dan undersampling dengan optimasi Bayesian pada XGBoost untuk mengatasi masalah ketidakseimbangan kelas. Pendekatan ini terbukti mampu meningkatkan performa klasifikasi jika dibandingkan dengan metode lain seperti RUSBoost dan CatBoost[11].

Metode AdaBoost dan XGBoost telah digunakan dalam penelitian analisis sentimen pada dataset IMDb yang terdiri dari 50.000 ulasan film. Hasilnya menunjukkan bahwa XGBoost memberikan performa lebih baik dibandingkan AdaBoost setelah dilakukan praproses teks dan vektorisasi menggunakan TF-IDF[12]. Sementara itu, penelitian lain mengembangkan algoritma rekomendasi berbasis ensemble learning dengan menggabungkan pendekatan berbasis konten, KNN, dan SVD. Evaluasi pada dataset TMDB5000 dan MovieLens menunjukkan bahwa metode ensemble mampu memberikan rekomendasi yang lebih dipersonalisasi, meskipun performa keseluruhannya sedikit lebih rendah dibandingkan beberapa metode individual[13].

Meskipun metode yang digunakan dalam beberapa penelitian tersebut berbeda dengan pendekatan yang diusulkan dalam studi ini, konsep dasar penggunaan ensemble learning dan teknik penyeimbangan data tetap memberikan kontribusi penting dalam

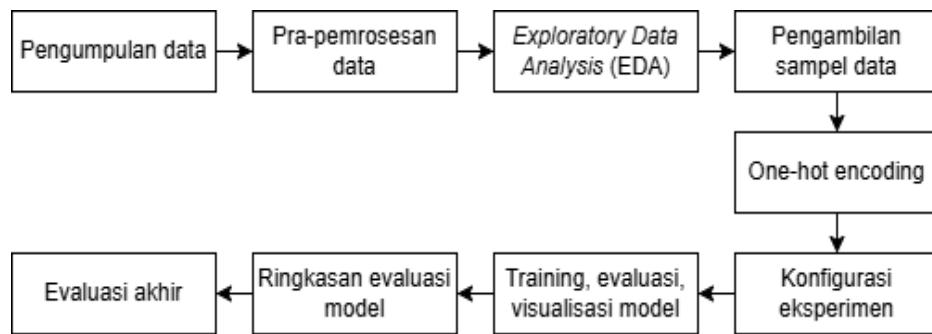
mengatasi masalah ketidakseimbangan kelas di berbagai domain aplikasi. Pendekatan ini menunjukkan fleksibilitas dan efektivitas dalam meningkatkan performa model klasifikasi, sehingga sangat relevan untuk diterapkan dalam konteks sistem rekomendasi film yang juga menghadapi masalah ketidakseimbangan kelas.

Penelitian ini bertujuan untuk mengkaji penerapan ensemble learning dalam klasifikasi preferensi film pada dataset MovieLens yang memiliki karakteristik kelas tidak seimbang. Meskipun sudah banyak penelitian yang membahas teknik ensemble dan penyeimbangan data, masih terbatas studi yang secara khusus menguji kombinasi ini dalam konteks sistem rekomendasi berbasis MovieLens. Oleh karena itu, penelitian ini akan membandingkan kinerja beberapa model ensemble, yaitu Random Forest, XGBoost, dan Voting Classifier, dengan menerapkan optimasi hyperparameter menggunakan Grid Search CV serta teknik penyeimbangan data menggunakan SMOTE. Diharapkan pendekatan ini dapat meningkatkan akurasi dan keandalan klasifikasi, serta menghasilkan sistem rekomendasi yang lebih adil dan efektif dibandingkan model tunggal tanpa penanganan khusus terhadap ketidakseimbangan kelas.

## 2. METODE PENELITIAN

Alur kerja penelitian ini dimulai dengan tahapan pengumpulan data, yaitu mengambil dataset MovieLens dari situs GroupLens. Setelah itu, dilakukan pra-pemrosesan data yang mencakup pembersihan data, penggabungan data, dan pembuatan fitur biner. Tahapan selanjutnya adalah *Exploratory Data Analysis* (EDA), dilakukan untuk mendapatkan wawasan mengenai distribusi dataset. Kemudian dilakukan pengambilan sampel data dan proses *one-hot encoding* untuk memisahkan kolom genre ke dalam format representasi biner. Setelah data siap, tahap berikutnya adalah konfigurasi eksperimen, yaitu menentukan parameter seperti *test size* dan *random state* yang akan digunakan dalam proses pelatihan model. Selanjutnya dilakukan proses pelatihan (*training*), evaluasi, dan visualisasi model untuk mengukur kinerja model menggunakan metrik yang relevan serta menampilkan hasil dalam bentuk visual untuk mempermudah interpretasi. Setelah semua eksperimen dilakukan, dilanjutkan dengan tahap penyusunan ringkasan hasil evaluasi, yakni membandingkan performa dari masing-masing model berdasarkan metrik yang diperoleh. Tahap terakhir adalah evaluasi akhir, yang bertujuan untuk menentukan model terbaik secara rata-rata maupun berdasarkan performa individu.

Penelitian ini dilakukan dengan beberapa tahapan utama seperti yang dapat dilihat pada Gambar 1.



**Gambar 1.** Tahapan Penelitian

## 2.1. Pengumpulan Data

Dataset yang digunakan dalam penelitian ini adalah MovieLens Beliefs Dataset 2024 yang diunduh dari situs GroupLens[14] dan hanya menggunakan dua buah file CSV, yakni file ratings\_for\_additional\_user.csv yang memiliki empat buah kolom (userId, movieId, rating, tstamp), berisi 4.185.688 data rating dengan rentang waktu mulai dari tahun 1997 hingga 2024, dan file movies.csv yang memiliki tiga buah kolom (movieId, title, genres) dan 105.071 data film. Karena perangkat yang digunakan kurang memadai untuk memproses data rating yang sangat besar, maka terlebih dahulu dilakukan penyaringan pada data rating dan menyimpan hasil penyaringan ke dalam file csv baru. Data rating yang digunakan hanya data yang berada pada rentang tahun 2023 hingga 2024, dan hasilnya disimpan dengan nama filtered\_ratings\_2023\_2024.csv. Jumlah data rating setelah dilakukan penyaringan adalah 975.584 record.

Selanjutnya adalah melakukan import library dan import dataset. Bahasa pemrograman yang digunakan dalam penelitian ini adalah Bahasa Python melalui Jupyter Notebook dan library yang digunakan antara lain: pandas, numpy, matplotlib, seaborn, sklearn, xgboost, imblearn. Import library digunakan untuk mengakses library yang akan digunakan. Import dataset digunakan untuk mengambil data rating dan data movie yang akan dianalisis dalam program Python, dan melihat apakah data yang akan digunakan sudah bersih atau belum. Data keseluruhan dapat dilihat pada Gambar 2.a dan Gambar 2.b berikut.

	userId	movieId	rating	tstamp		movieId	title	genres
0	393217	1	3.5	2023-01-25 19:45:46		0	1	Toy Story (1995) Adventure Animation Children Comedy Fantasy
1	393217	6	4.0	2023-02-07 21:17:19		1	2	Jumanji (1995) Adventure Children Fantasy
2	393217	16	3.5	2023-01-25 19:50:40		2	3	Grumpier Old Men (1995) Comedy Romance
3	393217	17	4.5	2023-01-25 19:49:45		3	4	Waiting to Exhale (1995) Comedy Drama Romance
4	393217	32	3.0	2023-01-25 19:46:01		4	5	Father of the Bride Part II (1995) Comedy
...	...	...	...	...		...	...	...
975579	393205	278702	2.0	2023-01-26 18:20:14		105066	262135	Becoming Cousteau (2021) Documentary
975580	393205	279562	3.0	2023-01-26 18:24:15		105067	262137	Death Goes to School (1953) Mystery
975581	393205	279812	2.5	2023-01-25 17:08:16		105068	262139	Street of Shadows (1953) Crime Drama
975582	393205	286897	4.0	2024-02-08 15:36:05		105069	262141	The Accusation (2021) Drama
975583	393205	290213	2.5	2024-02-08 15:34:46		105070	262143	Eugénie Grandet (2021) Drama

a. Data Rating

b. Data Movie

**Gambar 2. Data Rating dan Data Movie**

Namun data tersebut masih belum bersih, data rating yang masih memiliki rating -1.0 dan juga terdapat data movie yang tidak memiliki genre. Untuk itu akan dilakukan pembersihan data pada tahapan pra-pemrosesan data.

## 2.2. Pra-Pemrosesan Data

Tahapan pra-pemrosesan data dilakukan untuk membersihkan data, menggabungkan data, dan pembuatan fitur biner. Pembersihan data dimulai dengan menyaring data rating yang memiliki rating -1.0. Setelah penyaringan, jumlah data rating menjadi 823.722 record. Selanjutnya melakukan penyaringan data movie yang tidak memiliki genre. Setelah penyaringan, jumlah data movie menjadi 96.351 record. Langkah berikutnya adalah penggabungan data rating dan data movie berdasarkan movieId. Jumlah data setelah

penggabungan adalah 818.956 record dengan enam buah kolom yaitu userId, movieId, rating, tstamp, title, genres. Hasil penggabungan data dapat dilihat pada Gambar 4.

	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
...	...	...	...	...	...	...
818951	393205	278702	2.0	2023-01-26 18:20:14	Glass Onion: A Knives Out Mystery (2022)	Comedy Crime Mystery
818952	393205	279562	3.0	2023-01-26 18:24:15	Hellraiser (2022)	Horror Mystery Thriller
818953	393205	279812	2.5	2023-01-25 17:08:16	The Menu (2022)	Comedy Horror
818954	393205	286897	4.0	2024-02-08 15:36:05	Spider-Man: Across the Spider-Verse (2023)	Action Adventure Animation Sci-Fi
818955	393205	290213	2.5	2024-02-08 15:34:46	The Equalizer 3 (2023)	Action Crime Thriller

**Gambar 3.** Dataset Setelah penggabungan

Setelah penggabungan data, langkah selanjutnya adalah membuat fitur biner “Liked” berdasarkan rating. Rating  $\geq 4$  masuk ke dalam kategori “Liked” sedangkan rating dibawah 4 masuk ke kategori “Not Liked”. Setelah membuat fitur biner, lanjut ke pembuatan label untuk ditambahkan ke dalam kolom dataset dengan ketentuan biner 0 dilabeli “Not Liked” dan biner 1 diberi label “Liked”. Hasil dataset setelah pembuatan fitur biner dan label dapat dilihat pada Gambar 5.

	userId	movieId	rating	tstamp	title	genres	liked	liked_label
0	393217	1	3.5	2023-01-25 19:45:46	Toy Story (1995)	Adventure Animation Children Comedy Fantasy	0	Not Liked
1	393217	6	4.0	2023-02-07 21:17:19	Heat (1995)	Action Crime Thriller	1	Liked
2	393217	16	3.5	2023-01-25 19:50:40	Casino (1995)	Crime Drama	0	Not Liked
3	393217	17	4.5	2023-01-25 19:49:45	Sense and Sensibility (1995)	Drama Romance	1	Liked
4	393217	32	3.0	2023-01-25 19:46:01	Twelve Monkeys (a.k.a. 12 Monkeys) (1995)	Mystery Sci-Fi Thriller	0	Not Liked
...	...	...	...	...	...	...	...	...
818951	393205	278702	2.0	2023-01-26 18:20:14	Glass Onion: A Knives Out Mystery (2022)	Comedy Crime Mystery	0	Not Liked
818952	393205	279562	3.0	2023-01-26 18:24:15	Hellraiser (2022)	Horror Mystery Thriller	0	Not Liked
818953	393205	279812	2.5	2023-01-25 17:08:16	The Menu (2022)	Comedy Horror	0	Not Liked
818954	393205	286897	4.0	2024-02-08 15:36:05	Spider-Man: Across the Spider-Verse (2023)	Action Adventure Animation Sci-Fi	1	Liked
818955	393205	290213	2.5	2024-02-08 15:34:46	The Equalizer 3 (2023)	Action Crime Thriller	0	Not Liked

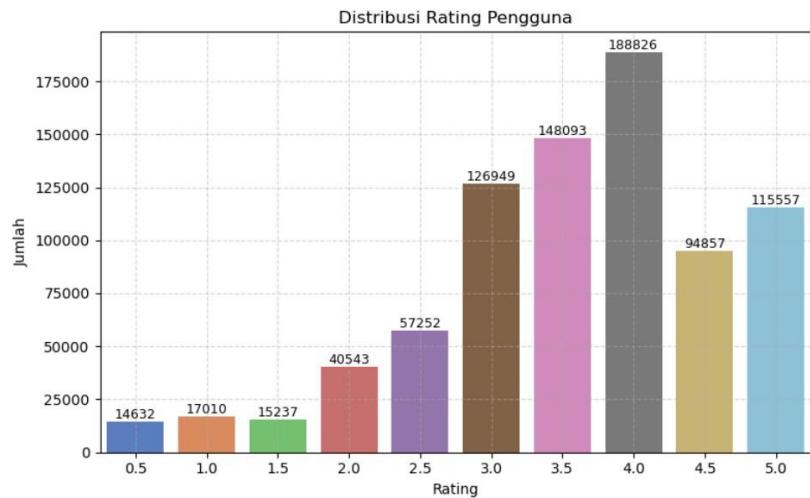
**Gambar 4.** Fitur Biner dan Pembuatan Label

### 2.3. Exploratory Data Analysis (EDA)

Tahapan eksplorasi data dilakukan untuk melihat distribusi data secara menyeluruh pada dataset. Visualisasi distribusi pada penelitian ini antara lain visualisasi distribusi rating pengguna untuk melihat rating yang diberikan oleh pengguna, distribusi

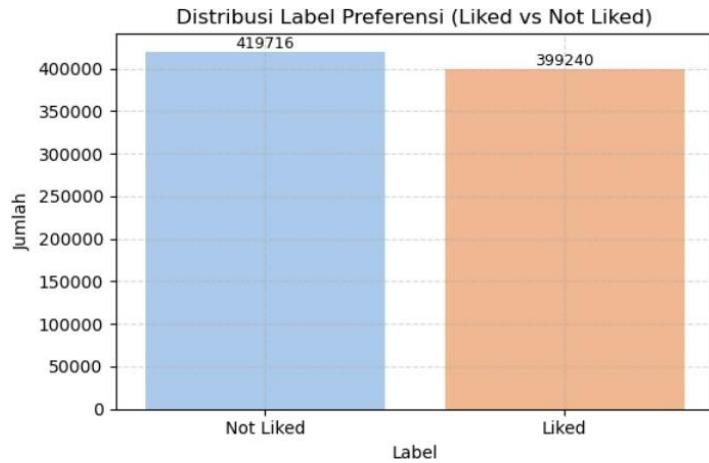
label preferensi Liked dan Not Liked untuk melihat berapa banyak pengguna yang menyukai atau tidak menyukai sebuah film, serta distribusi genre film untuk melihat genre film yang disukai pengguna.

Rating yang paling banyak diberikan pengguna adalah rating 4.0 sebesar 188.826, kemudian diikuti dengan rating 3.5 sebesar 148.093, dan rating 3.0 sebesar 126.949. Sementara itu rating 0.5 menjadi rating yang paling sedikit diberikan pengguna yakni sebesar 14.632. Visualisasi distribusi rating pengguna dapat dilihat pada Gambar 6.



**Gambar 5.** Distribusi Rating Pengguna

Distribusi label preferensi Liked dan Not Liked dapat dilihat pada Gambar 7 berikut.

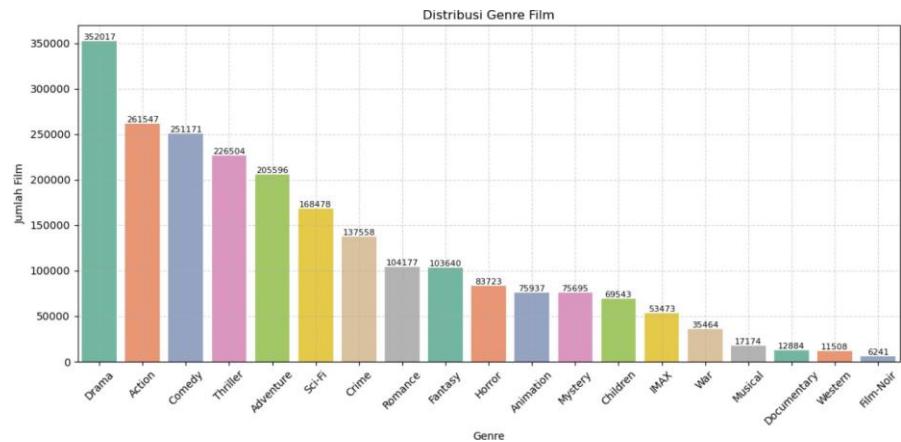


**Gambar 6.** Distribusi Liked dan Not Liked

Grafik distribusi label preferensi menunjukkan bahwa data terdiri dari dua kelas, yaitu Not Liked (419.716) dan Liked (399.240). Meskipun terdapat perbedaan jumlah, distribusinya relatif mendekati seimbang. Untuk memastikan bahwa proporsi antar kelas tetap konsisten selama proses pelatihan dan pengujian, pada tahapan selanjutnya akan dilakukan *stratified sampling* dalam proses pembagian data. Dengan pendekatan ini, distribusi label pada data latih dan data uji mencerminkan kondisi sebenarnya dari

keseluruhan dataset, sehingga dapat meminimalkan potensi bias model terhadap kelas mayoritas.

Visualisasi distribusi genre film menyajikan data tentang berbagai genre film dalam bentuk visual yang menunjukkan seberapa populer setiap genre, seperti pada Gambar 8 berikut.



**Gambar 7.** Distribusi Genre Film

Grafik distribusi genre film berdasarkan preferensi pengguna menunjukkan bahwa film bergenre Drama merupakan yang paling banyak disukai dengan total 352.017 interaksi positif, diikuti oleh Action (261.547), dan Comedy (251.171). Sebaliknya, film dengan genre Documentary (12.884), Western (11.508), dan Film-Noir (6.241) memiliki jumlah interaksi positif yang jauh lebih rendah, yang dapat mengindikasikan bahwa genre-genre tersebut kurang diminati oleh pengguna dalam dataset ini.

#### 2.4. Pengambilan Sampel Data

Pengambilan sampel data dalam penelitian ini dilakukan secara acak (*random sampling*) sebanyak 50.000 data dengan distribusi label sebanyak 25.809 kelas “*Not Liked*” dan 24.191 kelas “*Liked*”. Gambar 9 menunjukkan distribusi label pada dataset setelah dilakukan *random sampling*.

	userid	movieid	rating	tstamp	title	genres	liked	liked_label
82408	395965	58559	5.0	2023-04-13 17:28:26	Dark Knight, The (2008)	Action Crime Drama  MAX	1	Liked
672606	367330	250362	3.0	2023-11-13 19:03:22	Censor (2021)	Horror	0	Not Liked
42032	394872	901	5.0	2023-03-11 15:48:06	Funny Face (1957)	Comedy Musical	1	Liked
706153	373999	8368	3.5	2024-02-23 13:51:38	Harry Potter and the Prisoner of Azkaban (2004)	Adventure Fantasy  MAX	0	Not Liked
602934	87053	288773	1.0	2024-04-13 05:55:03	Anatomy of a Fall (2023)	Crime Drama Mystery	0	Not Liked
...	...	...	...	...	...	...	...	...
69896	395670	186631	4.0	2023-11-18 12:42:33	Everybody Knows (2018)	Drama	1	Liked
392596	404641	40815	4.0	2023-11-24 05:33:45	Harry Potter and the Goblet of Fire (2005)	Adventure Fantasy Thriller  MAX	1	Liked
199081	398942	2	1.0	2023-09-20 10:07:10	Jumanji (1995)	Adventure Children Fantasy	0	Not Liked
335456	402739	202439	5.0	2023-09-30 23:27:20	Parasite (2019)	Comedy Drama	1	Liked
123111	396765	112183	4.0	2023-05-05 23:02:17	Birdman: Or (The Unexpected Virtue of Ignorance...)	Comedy Drama	1	Liked

**Gambar 8.** Sampel Acak Dataset

## 2.5. One-hot Encoding

Tahapan one-hot encoding dalam penelitian ini digunakan untuk memisahkan kolom genre, menggabungkan fitur dan menjadi beberapa kolom biner yang mana masing-masing kolom mewakili satu genre. Jika genre ada pada film kolom akan diisi dengan 1, dan jika tidak ada kolom akan diisi dengan 0. Hasilnya dapat dilihat pada Gambar 10.

userId	movieId	Action	Adventure	Animation	Children	Comedy	Crime	Documentary	Drama	...	Film-Noir	Horror	IMAX	Musical	Mystery	Romance	Sci-Fi	Thriller	War	Western
82408	395965	58559	1	0	0	0	0	1	0	1	...	0	0	1	0	0	0	0	0	0
672606	367330	250362	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0
42032	394872	901	0	0	0	0	1	0	0	0	...	0	0	0	1	0	0	0	0	0
706153	373999	8368	0	1	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0
602934	87053	288773	0	0	0	0	0	1	0	1	...	0	0	0	0	1	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
69896	395670	186631	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0
392596	404641	40815	0	1	0	0	0	0	0	0	...	0	0	1	0	0	0	1	0	0
199081	398942	2	0	1	0	1	0	0	0	0	0	...	0	0	0	0	0	0	0	0
335456	402739	202439	0	0	0	0	1	0	0	1	...	0	0	0	0	0	0	0	0	0
123111	396765	112183	0	0	0	1	0	0	1	0	...	0	0	0	0	0	0	0	0	0

**Gambar 9. One-hot Encoding Kolom Genre**

## 2.6. Konfigurasi Eksperimen

Pada tahap ini, ditentukan nilai-nilai parameter eksperimen utama, yaitu *test size* dan *random state*. Kedua parameter ini digunakan dalam proses pembagian data menjadi data latih dan data uji. Beberapa kombinasi nilai digunakan untuk menguji kestabilan model terhadap variasi pembagian data. Tahapan ini membentuk dasar dari proses eksperimen berulang (*looping*) yang dilakukan selanjutnya. Eksperimen dilakukan dengan mengubah nilai *test size* dan *random state* secara sistematis. Nilai *test size* yang digunakan adalah 0.1, 0.2, 0.3, 0.4, dan 0.5, sedangkan *random state* divariasikan antara angka 42 dan 2024.

## 2.7. Training, Evaluasi, dan Visualisasi Model

Tahapan ini merupakan inti dari proses eksperimen yang dilakukan secara berulang (*looping*) untuk setiap kombinasi *test size* dan *random state*. Pada setiap iterasi eksperimen, data latih diseimbangkan terlebih dahulu menggunakan metode SMOTE untuk mengatasi potensi ketimpangan kelas yang mungkin muncul akibat pengambilan sampel secara acak. Model yang digunakan adalah Random Forest dan XGBoost, yang masing-masing dioptimalkan menggunakan Grid Search. Hasil terbaik dari masing-masing model kemudian digabungkan ke dalam Voting Classifier. Evaluasi dilakukan menggunakan metrik Accuracy, Precision, Recall, F1-Score, dan ROC AUC, serta divisualisasikan dalam bentuk ROC Curve, Precision-Recall Curve, dan Confusion Matrix. Visualisasi ini membantu dalam memahami performa model secara komprehensif pada tiap kombinasi eksperimen.

## 2.8. Ringkasan Evaluasi Model

Hasil evaluasi dari semua konfigurasi eksperimen kemudian dirangkum dalam bentuk tabel untuk memudahkan analisis dan perbandingan. Tabel ini mencakup metrik evaluasi untuk setiap kombinasi model, *test size*, dan *random state*, serta parameter terbaik hasil Grid Search. Data ringkasan ini juga disimpan ke dalam file CSV untuk dokumentasi dan keperluan analisis lanjutan.

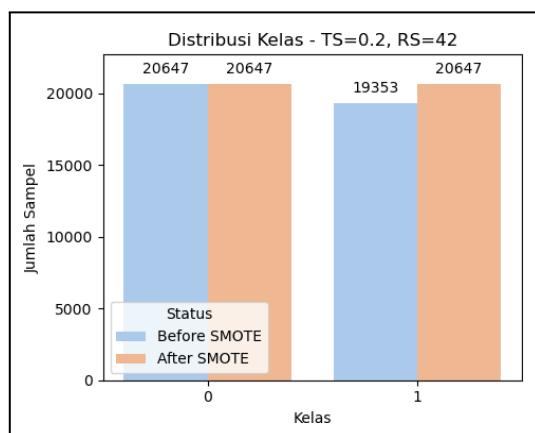
## 2.9. Evaluasi Akhir

Tahapan ini bertujuan untuk menyimpulkan model terbaik berdasarkan hasil evaluasi pada tahap sebelumnya. Model dengan performa paling konsisten dan metrik tertinggi (terutama ROC AUC dan F1-Score) dipilih sebagai model final. Model terbaik ditentukan berdasarkan rata-rata nilai ROC AUC dan F1 Score dari beberapa pengujian yang dilakukan sebelumnya. Selain itu, hasil evaluasi digunakan untuk mengkaji apakah pendekatan seperti SMOTE dan ensemble learning memberikan dampak signifikan terhadap peningkatan performa model. Evaluasi ini menjadi dasar dalam merekomendasikan strategi pemodelan yang optimal untuk data sejenis.

## 3. HASIL DAN PEMBAHASAN

### 3.1 Hasil

Data hasil random sampling masih menunjukkan adanya ketidakseimbangan kelas antara label Not Liked dan Liked. Untuk mengatasi hal ini, digunakan metode SMOTE (Synthetic Minority Over-sampling Technique) pada data latih sebelum proses pelatihan model. SMOTE bekerja dengan menambahkan data sintetis pada kelas minoritas hingga jumlahnya seimbang dengan kelas mayoritas. Dengan penerapan SMOTE, distribusi kelas dalam data latih menjadi seimbang, sehingga model dapat belajar secara adil dari kedua kelas. Visualisasi distribusi kelas sebelum dan sesudah SMOTE ditampilkan pada Gambar 11.



**Gambar 10.** Distribusi Kelas Sebelum dan Sesudah SMOTE

Gambar 11 menunjukkan distribusi label pada data latih sebelum dan sesudah dilakukan oversampling menggunakan SMOTE dengan konfigurasi test size sebesar 0.2. Terlihat bahwa sebelum dilakukan SMOTE, terdapat ketidakseimbangan jumlah antara kelas Not Liked (20.647) dan Liked (19.353). Setelah proses SMOTE, jumlah kedua kelas menjadi seimbang, masing-masing sebanyak 20.647 data.

Dengan distribusi kelas yang seimbang pada data latih, model machine learning dapat belajar secara adil dari kedua kelas. Hal ini penting untuk mencegah bias terhadap kelas mayoritas dan meningkatkan performa model dalam mengenali kedua kelas secara seimbang. Visualisasi serupa juga ditunjukkan pada konfigurasi test size lainnya, dengan pola distribusi yang konsisten.

Model dievaluasi menggunakan berbagai kombinasi parameter test size dan random state untuk mengamati konsistensi performa. Tiga model yang dibandingkan dalam penelitian ini adalah Random Forest, XGBoost, dan Voting Classifier. Penentuan parameter terbaik dilakukan menggunakan teknik GridSearchCV, yang secara otomatis mencari kombinasi parameter optimal untuk setiap model berdasarkan metrik ROC AUC. Tabel 1 berikut merupakan ringkasan hasil evaluasi beberapa kombinasi terbaik berdasarkan nilai ROC AUC:

**Tabel 1.** Ringkasan Evaluasi Model

Model	Test Size	Random State	ROC AUC	Accuracy	F1 Score	Best Params
Voting Classifier	0.4	42	<b>0.6482</b>	0.6054	0.5980	{'rf': {'max_depth': 10, 'n_estimators': 100}, 'xgb': {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100}}
Voting Classifier		2024	0.6540	<b>0.6076</b>	0.5966	{'rf': {'max_depth': 10, 'n_estimators': 100}, 'xgb': {'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100}}
XGBoost	0.1	2024	0.6496	0.6070	<b>0.6025</b>	{'learning_rate': 0.1, 'max_depth': 3, 'n_estimators': 100}
Random Forest	0.4	2024	0.6453	0.6020	0.5871	{'max_depth': 10, 'n_estimators': 100}

Dari hasil evaluasi tersebut, Voting Classifier consistently menunjukkan performa paling stabil dan unggul pada sebagian besar kombinasi parameter, baik dari segi ROC AUC maupun F1 Score. Hal ini menunjukkan bahwa pendekatan ensemble yang menggabungkan Random Forest dan XGBoost berhasil meningkatkan kemampuan generalisasi model. Parameter terbaik yang ditemukan untuk masing-masing model adalah Random Forest: max\_depth=10 dengan n\_estimators=100, XGBoost: learning\_rate=0.1, max\_depth=3, dan n\_estimators=100, sedangkan Voting Classifier adalah kombinasi dari parameter terbaik Random Forest dan XGBoost.

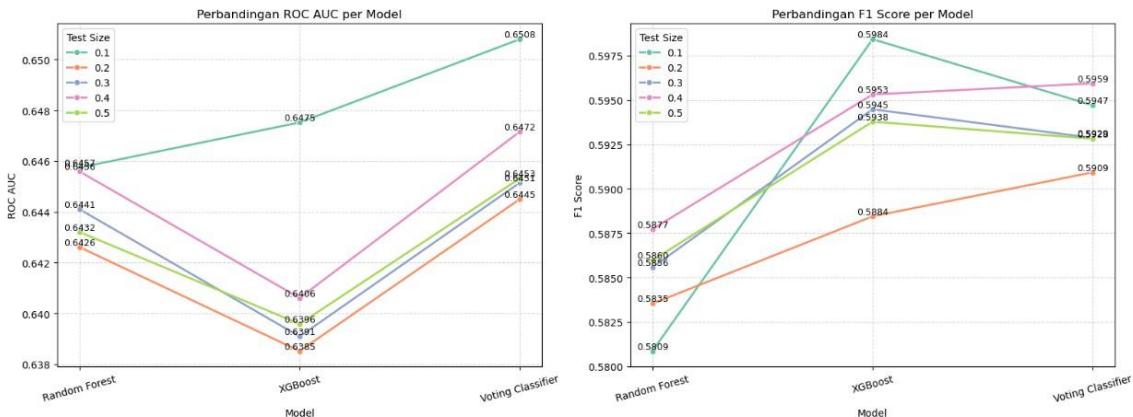
Setelah melakukan eksperimen dengan berbagai kombinasi parameter test size dan random state, dilakukan ringkasan hasil evaluasi untuk setiap model berdasarkan rata-rata skor ROC AUC dan F1 Score. Tabel 2 memperlihatkan nilai rata-rata performa dari tiga model yang diuji, yaitu Random Forest, XGBoost, dan Voting Classifier.

**Tabel 2.** Rata-rata Skor ROC dan F1 Score per Model

Model	ROC AUC	F1 Score
Random Forest	0.644251	0.584729
Voting Classifier	0.646587	0.593449
XGBoost	0.641063	0.594089

Dari tabel tersebut, terlihat bahwa model Voting Classifier memberikan nilai rata-rata ROC AUC tertinggi, menunjukkan performa terbaik dalam hal kemampuan membedakan kelas positif dan negatif. Namun, untuk metrik F1 Score, yang mengukur keseimbangan antara precision dan recall, model XGBoost menunjukkan nilai tertinggi, mengindikasikan kemampuan yang lebih baik dalam mempertahankan keseimbangan antara *false positive* dan *false negative*.

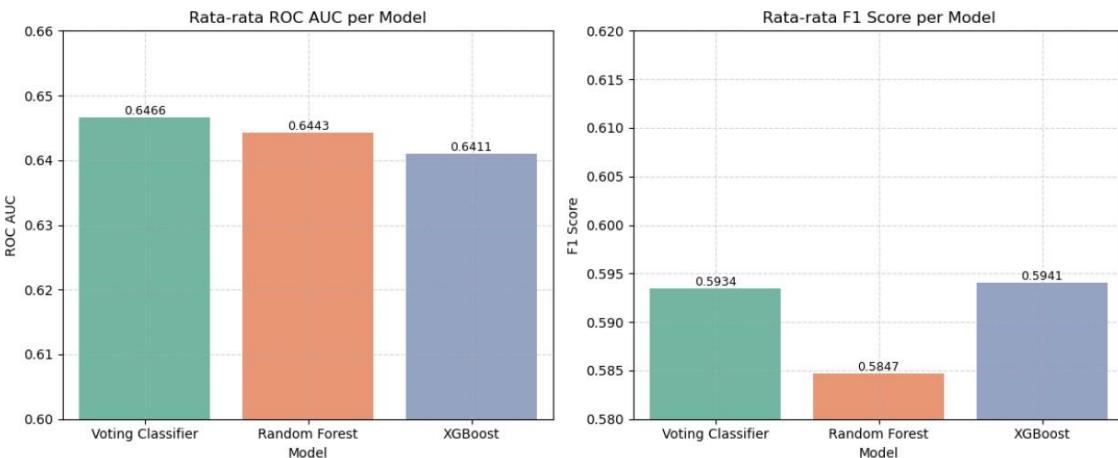
Visualisasi grafik perbandingan skor ROC AUC dan F1 Score pada Gambar 12 memperlihatkan tren ini secara jelas. Voting Classifier unggul pada ROC AUC di hampir semua *test size*, sedangkan XGBoost unggul pada F1 Score di *test size* 0.1, 0.3, dan 0.5. Random Forest secara konsisten berada di bawah kedua model tersebut untuk kedua metrik.



**Gambar 11.** Grafik Perbandingan ROC AUC dan F1 Score Table 4 menentukan nilai R

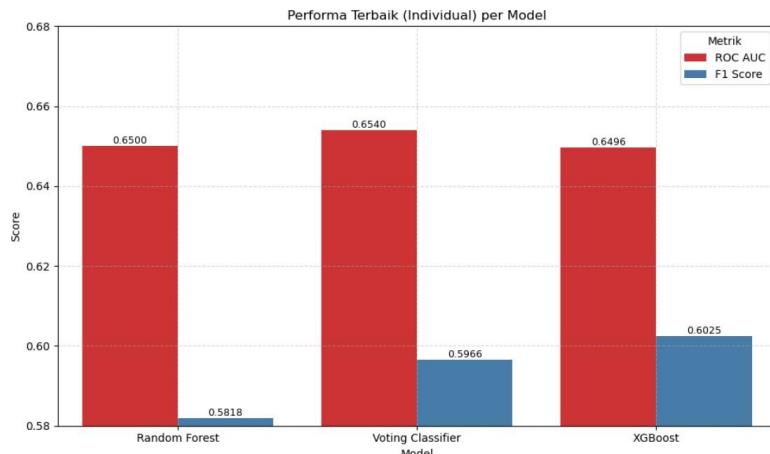
Grafik ini memberikan gambaran mengenai kekuatan masing-masing model dalam konteks evaluasi yang berbeda, sehingga pemilihan model dapat disesuaikan dengan prioritas metrik yang ingin dioptimalkan dalam aplikasi sebenarnya.

Berdasarkan rata-rata performa dari seluruh eksperimen, model terbaik secara keseluruhan adalah Voting Classifier, dengan rata-rata ROC AUC sebesar 0.6466 dan rata-rata F1 Score sebesar 0.5934. Model ini menggabungkan keunggulan Random Forest dan XGBoost sehingga mampu memberikan performa yang lebih stabil dan unggul dibandingkan model individu. Grafik rata-rata performa seluruh eksperimen dapat dilihat pada Gambar 13.



**Gambar 12.** Grafik Rata-rata Performa Seluruh Eksperimen

Hasil terbaik secara individual juga ditemukan pada Voting Classifier dengan *test size* 0.1 dan random state 2024, yang mencapai ROC AUC 0.6540 dan F1 Score 0.5966. Grafik performa terbaik pada *test size* 0.1 dapat dilihat pada Gambar 14.



**Gambar 13.** Performa Terbaik pada Test Size 0.1

### 3.2 Pembahasan

Penelitian ini menunjukkan bahwa model Voting Classifier merupakan model terbaik secara rata-rata dengan nilai ROC AUC sebesar 0.6466 dan F1 Score sebesar 0.5934. Hasil terbaik secara individual juga diperoleh oleh Voting Classifier pada test size 0.1 dan random state 2024 dengan ROC AUC 0.6540 dan F1 Score 0.5966. Model XGBoost unggul pada beberapa kondisi pengujian, khususnya untuk F1 Score pada test size 0.1, 0.3, dan 0.5. Dalam penelitian ini, metode SMOTE diterapkan untuk mengatasi ketidakseimbangan kelas pada data pelatihan. Penggunaan SMOTE membantu memperbaiki representasi kelas minoritas dengan mensintesis sampel baru sehingga model memiliki kesempatan belajar yang lebih baik terhadap kelas minoritas. Hal ini berkontribusi pada peningkatan performa, terutama pada metrik F1 Score yang sensitif terhadap keseimbangan presisi dan recall. Namun, penggunaan SMOTE juga perlu diperhatikan agar tidak menimbulkan overfitting, terutama pada data dengan sampel yang terbatas.

Performa model ensemble Voting Classifier yang lebih baik dibandingkan model tunggal seperti Random Forest dan XGBoost mengindikasikan bahwa penggabungan hasil prediksi dari beberapa model dapat meningkatkan kemampuan klasifikasi. Hal ini didukung oleh perbaikan keseimbangan antara sensitivitas dan presisi, yang tercermin pada nilai ROC AUC dan F1 Score. Namun, variasi performa model juga dipengaruhi oleh ukuran test size dan nilai random state, yang menunjukkan pentingnya pemilihan parameter yang tepat dalam eksperimen. Meskipun Voting Classifier unggul di ROC AUC, model XGBoost ternyata menunjukkan hasil F1 Score yang lebih tinggi pada beberapa ukuran test size tertentu, menunjukkan bahwa model tunggal masih memiliki potensi unggul pada metrik tertentu dalam kondisi tertentu.

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan. Selain keterbatasan terkait variasi model dan parameter, keterbatasan perangkat keras menjadi faktor pembatas utama dalam proses eksperimen. Kapasitas komputasi yang terbatas mengharuskan pengambilan sampel data yang relatif kecil, sehingga model dilatih dan diuji hanya pada subset data. Kondisi ini dapat mempengaruhi performa model serta kestabilan hasil evaluasi. Oleh karena itu, hasil penelitian ini mungkin belum sepenuhnya merepresentasikan performa model jika diaplikasikan pada dataset yang jauh lebih besar.

atau dalam kondisi nyata dengan data yang lengkap. Untuk itu, disarankan agar penelitian lanjutan dilakukan dengan sumber daya komputasi yang lebih memadai sehingga memungkinkan pemrosesan data dalam skala penuh dan menghasilkan evaluasi yang lebih komprehensif. Selain itu, penelitian selanjutnya disarankan untuk menggunakan sampel data yang lebih besar agar model dapat dilatih dan diuji dengan lebih representatif serta mendapatkan hasil yang lebih stabil dan akurat. Eksplorasi lebih lanjut terhadap teknik balancing data yang berbeda, termasuk variasi SMOTE atau metode oversampling lainnya, serta variasi model ensemble juga dapat dilakukan guna meningkatkan performa model lebih jauh.

#### 4. KESIMPULAN

Model Voting Classifier menunjukkan performa terbaik secara keseluruhan dengan rata-rata ROC AUC sebesar 0.6466 dan F1 Score sebesar 0.5934, serta performa terbaik individual pada test size 0.1 dan random state 2024. Model ini mampu menggabungkan keunggulan model Random Forest dan XGBoost untuk meningkatkan klasifikasi. Hasil ini mengimplikasikan bahwa pendekatan ensemble dapat menjadi strategi efektif dalam meningkatkan performa prediksi pada data tidak seimbang. Penggunaan teknik SMOTE dalam penelitian ini berperan penting dalam menangani ketidakseimbangan data dan meningkatkan performa model klasifikasi, khususnya pada metrik F1 Score. Namun, pemilihan parameter dan ukuran data sangat mempengaruhi hasil yang diperoleh.

Keterbatasan perangkat keras yang dimiliki dalam penelitian ini membatasi ukuran sampel data yang dapat diproses, sehingga eksperimen dilakukan pada subset data yang relatif kecil. Hal ini berpotensi mempengaruhi performa model dan kestabilan hasil evaluasi yang diperoleh. Oleh karena itu, untuk mendapatkan hasil yang lebih representatif dan valid, disarankan agar penelitian selanjutnya menggunakan perangkat keras dengan kapasitas komputasi yang lebih besar sehingga dapat mengakomodasi pemrosesan data dalam skala penuh.

Disarankan agar pengujian selanjutnya dilakukan dengan sampel data yang lebih besar dan kapasitas perangkat keras ditingkatkan agar proses pelatihan dan evaluasi dapat berjalan optimal. Selain itu, penelitian selanjutnya dapat mengeksplorasi teknik balancing data yang lebih variatif, termasuk modifikasi SMOTE atau metode *oversampling* lainnya, serta model *ensemble* lain guna memperoleh hasil yang lebih baik.

Penelitian ini membuka peluang untuk pengembangan model klasifikasi yang lebih kuat dengan pemanfaatan *ensemble learning* dan pengelolaan data tidak seimbang secara efektif. Dengan dukungan perangkat dan data yang memadai, hasil yang lebih optimal dapat dicapai di masa depan.

#### UCAPAN TERIMAKASIH

Penulis mengucapkan terima kasih kepada dosen pembimbing atas arahan dan masukan yang diberikan selama proses penelitian ini. Ucapan terima kasih juga disampaikan kepada rekan yang turut membantu dalam proses diskusi dan penulisan, serta berkontribusi sebagai penulis ketiga dalam penelitian ini.

#### **DAFTAR PUSTAKA**

- [1] S. Utami, K. M. Lhaksmana, and Y. Sibaroni, "Deep Learning and Imbalance Handling on Movie Review Sentiment Analysis," *SinkrOn*, vol. 7, no. 3, pp. 1894-1907, Jul. 2023, doi: 10.33395/sinkron.v8i3.12770.
- [2] A. A. Khan, O. Chaudhari, and R. Chandra, "A review of ensemble learning and data augmentation models for class imbalanced problems: Combination, implementation and evaluation," Jun. 15, 2024, *Elsevier Ltd.* doi: 10.1016/j.eswa.2023.122778.
- [3] P. A. R. Mukhlasin, A. Fitrianto, A. M. Soleh, and W. Z. A. Wan Muhamad, "Ensemble learning with imbalanced data handling in the early detection of capital markets," *Journal of Accounting and Investment*, vol. 24, no. 2, pp. 600-617, May 2023, doi: 10.18196/jai.v24i2.17970.
- [4] Y. Irawan, R. Wahyuni, and R. Ordila, "Comparative Analysis of Machine Learning Algorithms with SMOTE and Boosting Techniques in Accuracy Improvement," *The Indonesian Journal of Computer Science*, vol. 13, no. 5, 2024, doi: 10.33022/ijcs.v13i5.4386.
- [5] F. O. Aghware *et al.*, "Effects of Data Balancing in Diabetes Mellitus Detection: A Comparative XGBoost and Random Forest Learning Approach," *NIPES Journal of Science and Technology Research*, vol. 7, no. 1, pp. 1-11, Mar. 2025, doi: 10.37933/nipes/7.1.2025.1.
- [6] T. Permatasari Djaka and A. S. Winarsih, "Analisis Kinerja Ensemble Learning dan Algoritma Tunggal dalam Klasifikasi Sindrom Ovarium Polikistik Menggunakan Random Forest, Logistic Regression, dan XGBoost," *Infotekmesin*, vol. 16, no. 01, 2025, doi: 10.35970/infotekmesin.v16i1.2504.
- [7] D. F. Wicaksono, R. S. Basuki, and D. Setiawan, "Peningkatan Performa Model Machine Learning XGBoost Classifier melalui Teknik Oversampling dalam Prediksi Penyakit AIDS," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 8, no. 2, p. 736, Apr. 2024, doi: 10.30865/mib.v8i2.7501.
- [8] S. Sarmini, S. Sunardi, and A. Fadlil, "Performa Random Forest dan XGBoost pada Deteksi Penipuan E-Commerce Menggunakan Augmentasi Data CGAN," *Building of Informatics, Technology and Science (BITS)*, vol. 6, no. 3, pp. 1919-1931, Dec. 2024, doi: 10.47065/bits.v6i3.6430.
- [9] M. A. Talukder, M. Khalid, and M. A. Uddin, "An integrated multistage ensemble machine learning model for fraudulent transaction detection," *J Big Data*, vol. 11, no. 1, Dec. 2024, doi: 10.1186/s40537-024-00996-5.
- [10] M. Imani, A. Beikmohammadi, and H. R. Arabnia, "Comprehensive Analysis of Random Forest and XGBoost Performance with SMOTE, ADASYN, and GNUS Under Varying Imbalance Levels," *Technologies (Basel)*, vol. 13, no. 3, Mar. 2025, doi: 10.3390/technologies13030088.
- [11] P. Zhang, Y. Jia, and Y. Shang, "Research and application of XGBoost in imbalanced data," *Int J Distrib Sens Netw*, vol. 18, no. 6, Jun. 2022, doi: 10.1177/15501329221106935.
- [12] I. Gusti Ayu Nandia Lestari, N. Made Rai Masita Dewi, K. Gita Meiliana, and I. Komang Agus Ady Aryanto, "Effectiveness of AdaBoost and XGBoost Algorithms in Sentiment Analysis of Movie Reviews," 2025. [Online]. Available: <http://jurnal.polibatam.ac.id/index.php/JAIC>

- [13] W. Fang, Y. Sha, M. Qi, and V. S. Sheng, "Movie Recommendation Algorithm Based on Ensemble Learning," *Intelligent Automation and Soft Computing*, vol. 34, no. 1, pp. 609–622, 2022, doi: 10.32604/iasc.2022.027067.
- [14] GroupLens, "[https://grouplens.org/datasets/movielens/ml\\_belief\\_2024/](https://grouplens.org/datasets/movielens/ml_belief_2024/)."



**Prosiding- SEMASTER: Seminar Nasional Teknologi Informasi & Ilmu Komputer** is licensed under a [Creative Commons Attribution International \(CC BY-SA 4.0\)](https://creativecommons.org/licenses/by-sa/4.0/)