

EVALUASI PERBANDINGAN TEKNIK OVERSAMPLING TERHADAP KINERJA RANDOM FOREST DALAM MENDETEKSI TRANSAKSI HARI SPESIAL DI E-COMMERCE

Rizal Muslim Sinaga, Ali Afr Rahman S. Effendi,
Najwa Latifah Hasibuan, Fatimah Asro Harahap, Fanny Ramadhani

Ilmu Komputer, Universitas Negeri Medan

Jalan Willem Iskandar, Pasar V Medan Estate, Percut Sei Tuan, Deli Serdang, Indonesia
sinaga.rizal456@gmail.com

ABSTRAK

Perkembangan pesat e-commerce, khususnya pada event-event spesial seperti Harbolnas dan Black Friday, mengarah pada permasalahan ketidakseimbangan data dalam klasifikasi transaksi. Pada hari biasa, jumlah transaksi lebih banyak dibandingkan dengan hari spesial, yang menyebabkan model machine learning cenderung bias. Penelitian ini bertujuan untuk membandingkan teknik-teknik oversampling seperti Random Oversampling, SMOTE, ADASYN, dan Borderline-SMOTE untuk menyeimbangkan data dalam klasifikasi transaksi e-commerce. Metode yang digunakan dalam penelitian ini adalah pendekatan kuantitatif, di mana dataset transaksi e-commerce diproses dengan berbagai teknik oversampling untuk memperbaiki ketidakseimbangan kelas. Data yang telah diproses kemudian digunakan untuk pelatihan model klasifikasi dengan algoritma Random Forest. Evaluasi dilakukan dengan metrik Precision, Recall, dan F1-Score untuk mengukur kinerja model dalam mengklasifikasikan data yang telah disesuaikan. Hasil penelitian menunjukkan bahwa teknik SMOTE memberikan hasil yang lebih baik dalam meningkatkan keseimbangan antara Precision, Recall, dan F1-Score, dibandingkan dengan teknik oversampling lainnya. Penelitian ini menyarankan penggunaan teknik SMOTE untuk meningkatkan akurasi dalam klasifikasi transaksi e-commerce pada hari spesial.

Kata kunci : *e-commerce, ketidakseimbangan data, SMOTE, ADASYN, Random Forest, oversampling.*

1. PENDAHULUAN

Perkembangan teknologi digital telah mengubah pola konsumsi masyarakat, menjadikan e-commerce sebagai pilihan utama dalam bertransaksi. Selama pandemi, sektor ini mencatat pertumbuhan signifikan hingga 91% menurut Kemkominfo, seiring meningkatnya penggunaan aplikasi belanja online untuk memenuhi kebutuhan sehari-hari [1]. Fenomena yang menonjol dalam e-commerce adalah lonjakan transaksi saat momen spesial seperti Harbolnas, bulan Ramadhan, promo tanggal kembar (11.11, 12.12), dan perayaan lainnya. Pada hari-hari tersebut, perilaku konsumen berubah akibat pengaruh diskon besar, flash sale, dan bundling produk yang memicu pembelian impulsif [2].

Namun, data transaksi e-commerce sering kali tidak seimbang antara hari spesial dan hari biasa. Jumlah transaksi pada hari spesial cenderung lebih sedikit dibanding hari biasa, sehingga menimbulkan tantangan dalam penerapan algoritma machine learning. Model klasifikasi cenderung lebih mengenali pola dari kelas mayoritas (hari biasa), dan kesulitan mendeteksi kelas minoritas (hari spesial).

Untuk mengatasi permasalahan ini, digunakan teknik oversampling seperti Random Oversampling, SMOTE, ADASYN, dan Borderline-SMOTE yang dapat meningkatkan representasi kelas minoritas [3][4][15]. Penelitian ini bertujuan membandingkan efektivitas teknik-teknik tersebut dalam meningkatkan performa klasifikasi pada data transaksi e-commerce yang tidak seimbang. Algoritma Random Forest digunakan sebagai model utama karena kemampuannya yang andal dalam menangani data

besar, toleran terhadap noise, serta efektif dalam proses klasifikasi melalui sistem voting dari banyak pohon keputusan [6]. Seluruh proses dilakukan menggunakan bahasa pemrograman Python.

2. TINJAUAN PUSTAKA

2.1. Penelitian Terdahulu

Hari spesial seperti promo tanggal kembar terbukti berpengaruh signifikan terhadap perilaku pembelian impulsif konsumen di e-commerce. Penelitian menunjukkan bahwa berbagai promosi yang ditawarkan oleh Shopee pada tanggal-tanggal kembar, seperti gratis ongkir, diskon, cashback, hingga flash sale setiap jam, mampu mendorong pengguna untuk melakukan pembelian secara spontan tanpa perencanaan [7]. Pembelian impulsif sering dipicu oleh promo besar, khususnya saat special event day seperti tanggal-tanggal kembar [8]. Beberapa faktor seperti motivasi hedonik, citra merek, dan promosi juga memiliki pengaruh besar terhadap pembelian impulsif [9]. Peran promosi dalam meningkatkan penjualan di Shopee telah banyak dibahas [10], sementara efektivitas strategi pemasaran Shopee dalam menciptakan impulse buying, terutama pada hari spesial, juga telah diteliti [11]. Promo tanggal kembar seperti 11.11 dan 12.12 diketahui berkontribusi hingga 73% dalam mendorong pembelian, yang menunjukkan bahwa diskon besar menjadi faktor utama pengambilan keputusan konsumen [2].

Meski dampak promosi dan hari spesial terhadap perilaku konsumen telah banyak diteliti, tantangan teknis dalam klasifikasi transaksi e-commerce,

khususnya ketidakseimbangan data antara Hari Biasa dan Hari Spesial, masih jarang dibahas. Penelitian ini mengatasi masalah tersebut dengan menerapkan teknik oversampling (SMOTE, ADASYN, Borderline-SMOTE) pada algoritma Random Forest, serta menguji efektivitasnya menggunakan Python

2.2. E - Commerce

E-commerce merujuk pada aktivitas jual beli yang dilakukan secara elektronik melalui jaringan internet [12]. Seiring perkembangan teknologi, berbagai platform marketplace telah berkembang pesat di Indonesia, seperti Lazada, Shopee, dan Tokopedia. Setiap platform memiliki kelebihan dan kekurangan masing-masing, termasuk perbedaan dalam strategi promosi dan pendekatan pemasaran yang digunakan untuk menarik konsumen [13].

2.3. Python

Python adalah bahasa pemrograman yang menggunakan sistem *interpreter*, yang berarti kode dijalankan langsung baris per baris tanpa perlu dikompilasi terlebih dahulu. Bahasa ini dapat dijalankan di berbagai sistem operasi seperti Windows, Linux, dan lainnya [14]. Python juga merupakan salah satu bahasa pemrograman yang sangat populer karena sifatnya yang serbaguna. Selain digunakan untuk pengembangan perangkat lunak umum, Python juga banyak digunakan dalam pengembangan sistem berbasis *Machine Learning* dan *Deep Learning* [15].

2.4. Machine Learning Dan Klasifikasi

Machine learning merupakan bagian dari data mining yang digunakan untuk mengekstraksi pola atau informasi penting dari kumpulan data berukuran besar. Proses ini dilakukan secara otomatis dan sering kali diterapkan untuk mendukung pengambilan keputusan berbasis data. Data yang digunakan umumnya bersifat historis dan kompleks, sehingga diperlukan pendekatan sistematis untuk menemukan keteraturan atau hubungan antar data [16].

Salah satu teknik utama dalam machine learning adalah klasifikasi, yaitu proses untuk mengelompokkan data ke dalam kelas-kelas tertentu berdasarkan kemiripan fitur atau karakteristik. Klasifikasi bertujuan untuk membedakan antara objek yang termasuk dalam kelas berbeda dan biasanya digunakan dalam permasalahan supervised learning. Berbagai algoritma dapat digunakan dalam proses ini, seperti decision tree, Naïve Bayes, k-nearest neighbor, dan support vector machines, yang masing-masing memiliki keunggulan dalam konteks dan jenis data tertentu [16].

2.5. Random Forest

Random Forest (RF) adalah metode klasifikasi berbasis komputasi yang terdiri dari sekumpulan pohon keputusan (decision trees) yang menghasilkan prediksi berdasarkan sistem suara terbanyak (majority

vote). Metode ini membentuk banyak pohon tanpa melakukan pemangkasan (pruning) dan menggunakan Indeks Gini sebagai dasar pemilihan fitur pada setiap simpul internal pohon keputusan [6].

2.6. Permasalahan Data Tidak Seimbang

Ketidakseimbangan data merupakan tantangan kritis dalam pengembangan model klasifikasi, khususnya pada kasus transaksi e-commerce seperti klasifikasi transaksi Hari Spesial. Dalam kondisi seperti ini, jumlah data dari kelas minoritas sangat sedikit dibandingkan kelas mayoritas. Ketidakseimbangan ini membuat model cenderung mempelajari pola dari kelas mayoritas secara berlebihan, sehingga mengabaikan karakteristik penting dari kelas minoritas. Akibatnya, performa model dalam mengklasifikasikan kelas minoritas menurun, menyebabkan tingkat *false negative* yang tinggi dan potensi kegagalan dalam mendeteksi kejadian penting yang jarang muncul. Diantika menjelaskan bahwa data minoritas kerap dianggap sebagai noise atau outlier oleh model, sehingga memperbesar kemungkinan terjadinya kesalahan klasifikasi terhadap data tersebut [16]. Oleh karena itu, diperlukan strategi khusus yang mampu menyeimbangkan distribusi antar kelas sebelum model dilatih, agar performa klasifikasi menjadi lebih adil dan akurat untuk semua kelas yang ada.

2.7. Teknik Oversampling

Ketidakseimbangan data merupakan tantangan utama dalam klasifikasi transaksi e-commerce, khususnya pada data Hari Spesial yang jarang muncul. Teknik oversampling seperti SMOTE terbukti meningkatkan akurasi hingga 97,5% pada prediksi diabetes [3]. Namun, SMOTE memiliki keterbatasan dalam menangani wilayah sulit klasifikasi, sehingga ADASYN yang lebih adaptif dinilai lebih efektif dalam kondisi ketidakseimbangan ekstrem [4]. Random Oversampling (ROS), meski sederhana, mampu meningkatkan akurasi hingga 96,9% dalam klasifikasi phishing [16], sementara Borderline-SMOTE yang menekankan data di dekat batas keputusan menunjukkan performa tinggi dengan F1-score 81,16% dan recall 85,25% dalam klasifikasi ujaran kebencian [5]. Oleh karena itu, pemilihan teknik oversampling yang tepat, seperti kombinasi ADASYN dan Borderline-SMOTE, penting untuk meningkatkan deteksi kelas minoritas secara efektif.

2.8. Metode Evaluasi Model

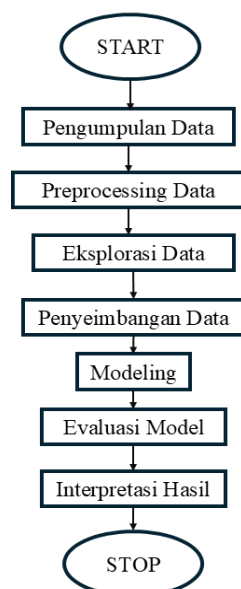
Evaluasi model klasifikasi dilakukan menggunakan metrik seperti akurasi, presisi, recall, F1-score, dan ROC-AUC. Akurasi mengukur proporsi prediksi yang benar, namun pada data tidak seimbang, metrik seperti recall dan F1-score lebih informatif karena mampu menunjukkan performa terhadap kelas minoritas. ROC-AUC digunakan untuk mengukur kemampuan model dalam membedakan antar kelas secara keseluruhan. Penggunaan kombinasi metrik ini

penting agar evaluasi model menjadi lebih objektif dan menyeluruh, terutama pada kasus klasifikasi dengan distribusi kelas yang tidak merata [16][5].

3. METODE PENELITIAN

Penelitian ini menggunakan pendekatan kuantitatif dengan metode analisis data untuk mengidentifikasi pola dan tren dalam data transaksi e-commerce secara objektif dan sistematis. Algoritma utama yang digunakan adalah Random Forest, algoritma machine learning berbasis pohon keputusan yang populer karena kestabilan prediksinya, kemampuannya mengurangi overfitting, serta efisiensinya dalam menangani data berskala besar dan kompleks. Fokus penelitian ini adalah membandingkan efektivitas beberapa teknik oversampling dalam menangani ketidakseimbangan data antara Hari Spesial dan Hari Biasa pada transaksi e-commerce, yaitu SMOTE (Synthetic Minority Over-sampling Technique), ADASYN (Adaptive Synthetic Sampling), dan Borderline-SMOTE.

3.1. Tahapan Penelitian



Gambar 1. Diagram Alur penelitian

Gambar 1 berikut menjelaskan tahapan penelitian dari tahap awal hingga evaluasi hasil model. Penjelasan tiap tahapan dijabarkan sebagai berikut:

- Pengumpulan Data** : Tahap awal melibatkan akuisisi dataset transaksi e-commerce. Dataset ini mencakup atribut-atribut seperti jumlah halaman produk yang dikunjungi, durasi kunjungan pengguna, bounce rate, nilai halaman, indikator Hari Spesial, bulan transaksi, jenis pengunjung, dan indikator akhir pekan. Informasi ini digunakan sebagai fitur untuk proses klasifikasi transaksi.
- Preprocessing Data** : Tahap ini mencakup pembersihan data (menghapus duplikasi dan menangani nilai yang hilang), encoding variabel kategorikal ke bentuk numerik, serta pelabelan

data berdasarkan kategori Hari Spesial dan Hari Biasa.

- Eksplorasi Data** : Analisis deskriptif dilakukan untuk memahami distribusi data awal, seperti jumlah transaksi pada Hari Spesial dan Hari Biasa, serta preferensi kategori produk berdasarkan waktu transaksi. Tujuannya adalah mengidentifikasi potensi bias dan struktur dasar data.
- Penyeimbangan Data** : Tahap ini menjadi inti dari penelitian. Untuk mengatasi ketidakseimbangan kelas, digunakan tiga teknik oversampling, yaitu SMOTE, ADASYN, dan Borderline-SMOTE. Masing-masing teknik menghasilkan data sintetis untuk kelas minoritas (Hari Spesial) dengan pendekatan berbeda. Proses ini bertujuan untuk meningkatkan kemampuan model dalam mengenali transaksi yang jarang terjadi.
- Modeling** : Model klasifikasi dibangun menggunakan algoritma Random Forest. Model dilatih menggunakan data hasil oversampling dari setiap teknik untuk dibandingkan performanya.
- Evaluasi Model** : Evaluasi dilakukan menggunakan metrik akurasi, precision, recall, F1-score, dan confusion matrix. Hasil evaluasi dibandingkan antar teknik oversampling untuk menilai mana yang paling efektif dalam meningkatkan performa model pada data tidak seimbang.
- Interpretasi Hasil** : Hasil evaluasi dan perbandingan dianalisis lebih lanjut untuk memahami pengaruh masing-masing teknik oversampling terhadap performa model klasifikasi. Analisis juga mencakup dampak terhadap kemampuan model dalam mengenali transaksi pada Hari Spesial serta potensi bias yang ditimbulkan.

3.2. Algoritma Random Forest

Metode utama yang digunakan dalam penelitian ini adalah algoritma Random Forest. Random Forest merupakan algoritma klasifikasi berbasis *ensemble learning* yang terdiri atas sekumpulan pohon keputusan. Setiap pohon dibentuk dari subset data dan subset fitur yang dipilih secara acak. Hasil dari semua pohon digabungkan dengan metode voting untuk menentukan prediksi akhir.

3.3. Pehitungan Indeks Gini

Untuk menentukan pemisahan terbaik di setiap node dalam pohon, digunakan Indeks Gini sebagai kriteria pengukuran impuritas (*impurity*). Formula perhitungannya adalah sebagai berikut:

$$Gini(S) = 1 - \sum_{i=1}^c p_i^2 \quad (1)$$

Dimana:

- $Gini(S)$ adalah tingkat ketidakmurnian (impurity) pada suatu simpul (node).
- P_i adalah proporsi sampel dalam kelas
- c adalah jumlah kelas dalam dataset.

Kualitas pemisahan (split) dihitung menggunakan rumus:

$$Gini_{split} = \sum_{i=1}^k \frac{n_i}{n} Gini(S_i) \quad (2)$$

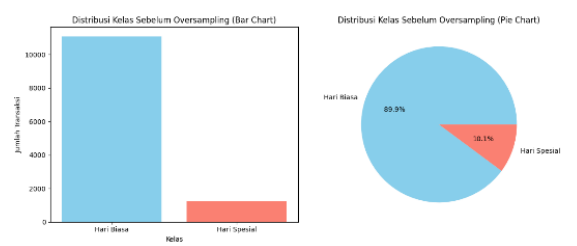
Dimana:

- $Gini_{split}$ adalah skor ketidakmurnian setelah pembagian data
- n_i adalah jumlah sampel dalam subset
- n adalah jumlah total sampel pada node sebelum pembagian.
- k adalah jumlah subset setelah pembagian.

4. HASIL DAN PEMBAHASAN

4.1. Hasil Penelitian

Dataset penelitian ini terdiri dari 12.329 data transaksi e-commerce dengan atribut seperti jumlah halaman dikunjungi, durasi kunjungan, bounce rate, nilai halaman, Hari Spesial, bulan, jenis pengunjung, dan akhir pekan. Terdapat dua kelas: Hari Biasa (11.078 data) dan Hari Spesial (1.251 data), dengan rasio distribusi 11.078:1.251 yang menunjukkan ketimpangan signifikan antar kelas.



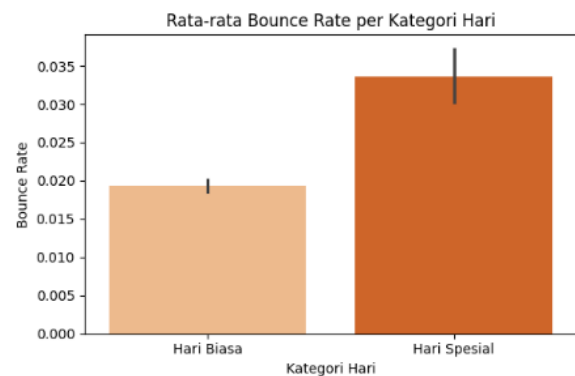
Gambar 2. Visualisasi Distribusi Kelas

Berikut penjelasan dari Gambar 2:

- Bar Chart (kiri) memperlihatkan perbandingan jumlah transaksi antara kelas *Hari Biasa* dan *Hari Spesial*. Terlihat bahwa jumlah transaksi pada *Hari Biasa* mendominasi secara signifikan dibandingkan *Hari Spesial*.
- Pie Chart (kanan) menampilkan proporsi distribusi kelas dalam bentuk persentase. *Hari Biasa* memiliki porsi sebesar 89.9%, sedangkan *Hari Spesial* hanya 10.1%, yang menunjukkan ketidakseimbangan kelas (imbalanced dataset).
- Ketidakseimbangan ini berpotensi menyebabkan model pembelajaran mesin menjadi bias terhadap kelas mayoritas (*Hari Biasa*) dan mengabaikan kelas minoritas (*Hari Spesial*), terutama jika tidak dilakukan penanganan seperti oversampling atau undersampling.
- Oleh karena itu, sebelum proses pelatihan model dilakukan, penting untuk melakukan teknik penyeimbangan kelas agar performa model terhadap kedua kelas dapat lebih adil dan akurat.

Setelah preprocessing, data penelitian terdiri dari 12.205 baris yang telah dibersihkan dari duplikat, missing values, outlier, dan dikonversi ke format yang sesuai. Kolom *SpecialDay* telah di-encode menjadi 0 (*Hari Biasa*) dan 1 (*Hari Spesial*), dengan distribusi 11.377 data *Hari Biasa* dan 828 *Hari Spesial*.

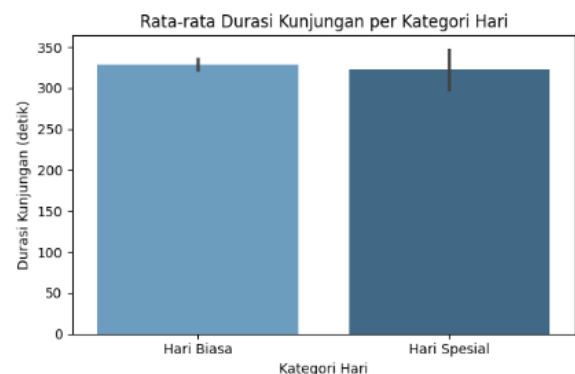
Ketimpangan ini menunjukkan adanya class imbalance, yang perlu ditangani untuk mencegah bias model.



Gambar 3. Rata-rata Bounce Rate per Kategori Hari

Berikut penjelasan dari Gambar 3:

- Gambar 3 menunjukkan bahwa bounce rate pada *Hari Spesial* lebih tinggi (sekitar 0.034) dibandingkan dengan *Hari Biasa* (sekitar 0.019).
- Hal ini mengindikasikan bahwa pada *Hari Spesial*, meskipun promosi meningkat, lebih banyak pengguna yang meninggalkan platform setelah melihat sedikit konten.
- Kenaikan bounce rate ini bisa disebabkan oleh ekspektasi pengguna terhadap promo yang tidak sesuai atau kurangnya pengalaman pengguna yang optimal.
- Artinya, promosi saja tidak cukup pengalaman navigasi dan kualitas penawaran juga sangat berperan dalam mempertahankan pengguna.



Gambar 4. Rata-rata Durasi Kunjungan per Kategori Hari

Berikut penjelasan dari Gambar 4:

- Gambar 4 memperlihatkan bahwa rata-rata durasi kunjungan pengguna pada *Hari Biasa* sedikit lebih tinggi (sekitar 330 detik) dibandingkan dengan *Hari Spesial* (sekitar 320 detik).
- Ini menunjukkan bahwa pada *Hari Biasa*, pengguna cenderung lebih sabar atau tertarik untuk menjelajahi konten di platform.
- Meskipun promosi banyak terjadi pada *Hari Spesial*, pengguna mungkin lebih cepat dalam membuat keputusan, entah itu langsung membeli atau langsung keluar.

- d. Perbedaan ini bisa menjadi sinyal bahwa motivasi kunjungan pada Hari Spesial dan Hari Biasa berbeda, dan hal ini penting untuk strategi pemasaran berbasis waktu.

Sebelum oversampling, dataset menunjukkan ketidakseimbangan signifikan antara kelas Hari Biasa sebanyak 11.377 transaksi dan Hari Spesial sebanyak 828 transaksi. Ketimpangan ini dapat memengaruhi performa model, khususnya dalam mendeteksi Hari Spesial. Untuk mengatasinya, digunakan empat teknik oversampling yaitu SMOTE, ADASYN, Borderline SMOTE, dan Random Over Sampler yang menambahkan data sintesis pada kelas minoritas agar distribusi kelas menjadi seimbang.

Tabel 1. Distribusi Kelas Sebelum dan Sesudah Oversampling

Teknik Oversampling	Hari Biasa	Hari Spesial	Total Data
Sebelum Oversampling	11,377	828	12,205
RandomOverSampler (ROS)	11,377	11,377	22,754
SMOTE	11,377	11,377	22,754
ADASYN	11,377	11,377	22,754
Borderline-SMOTE	11,377	11,377	22,754

Berikut penjelasan dari Tabel 1:

- Sebelum oversampling, data sangat tidak seimbang (Hari Biasa: 11.377; Hari Spesial: 828), yang berisiko membuat model bias ke kelas mayoritas.
- Setelah oversampling (SMOTE, ADASYN, Borderline-SMOTE, ROS), jumlah data Hari Spesial disamakan menjadi 11.377.
- Keempat teknik berhasil menciptakan distribusi data yang seimbang, sehingga model diharapkan lebih akurat dalam mendeteksi Hari Spesial dan mengurangi bias.

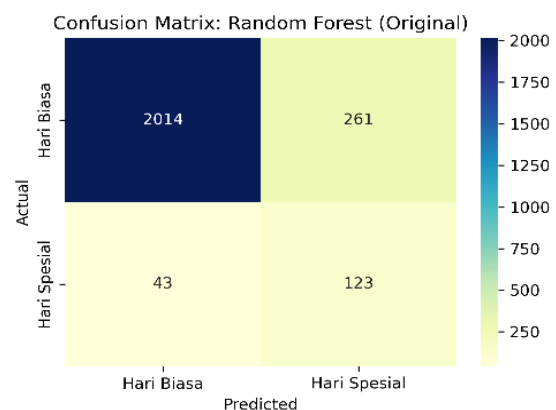
Tabel 2. Hasil Evaluasi Model Random Forest dengan Teknik Oversampling

Model	Accuracy	Precision	Recall	F1-Score
Random Forest (Original)	87.55%	32.03%	74.10%	44.73%
Random Forest (Random Oversampling)	92.63%	42.39%	23.49%	30.23%
Random Forest (SMOTE)	92.09%	41.82%	41.57%	41.69%
Random Forest (ADASYN)	91.72%	39.29%	39.76%	39.52%
Random Forest (Borderline-SMOTE)	92.01%	40.52%	37.35%	38.87%

Berikut penjelasan dari Tabel 2:

- Random Forest (Original) memiliki akurasi 87.55%, precision 32.03%, recall 74.10%, dan F1-Score 44.73%, menunjukkan ketidakseimbangan metrik.

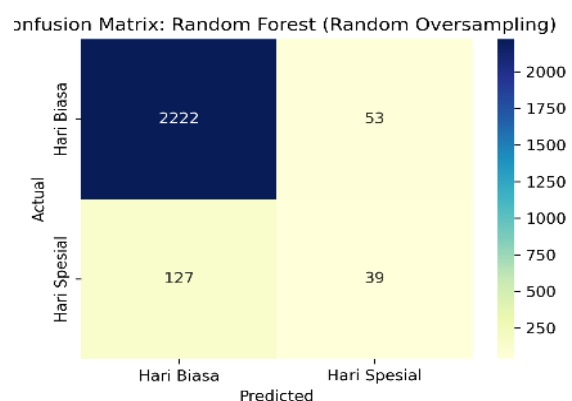
- Random Oversampling meningkatkan akurasi menjadi 92.63% dan precision 42.39%, tetapi recall menurun ke 23.49%, dengan F1-Score 30.23%.
- SMOTE menghasilkan akurasi 92.09%, precision 41.82%, recall 41.57%, dan F1-Score 41.69%, paling seimbang di antara semua teknik.
- ADASYN mencatat akurasi 91.72%, precision 39.29%, recall 39.76%, dan F1-Score 39.52%, cukup baik namun sedikit di bawah SMOTE.
- Borderline-SMOTE mencapai akurasi 92.01%, precision 40.52%, recall 37.35%, dan F1-Score 38.87%, performa stabil tapi kurang unggul dibanding SMOTE.



Gambar 5. Confusion Matrix Random Forest (Original)

Berikut penjelasan dari Gambar 5:

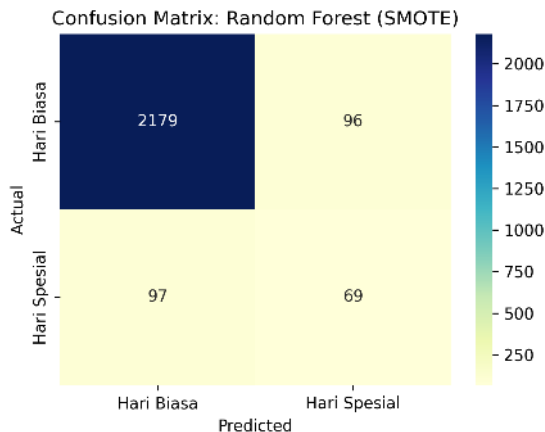
- Model tanpa penanganan ketidakseimbangan data (original) memiliki kinerja tinggi dalam mengklasifikasikan Hari Biasa (2014 benar dari total 2275), namun kesulitan dalam mengenali Hari Spesial (hanya 123 benar dari 166).
- False negative pada Hari Spesial cukup tinggi (43), yang artinya model sering mengklasifikasikan Hari Spesial sebagai Hari Biasa.
- Hal ini menunjukkan bahwa model bias terhadap kelas mayoritas. Biasa.



Gambar 6. Confusion Matrix Random Forest (Random Oversampling)

Berikut penjelasan dari Gambar 6:

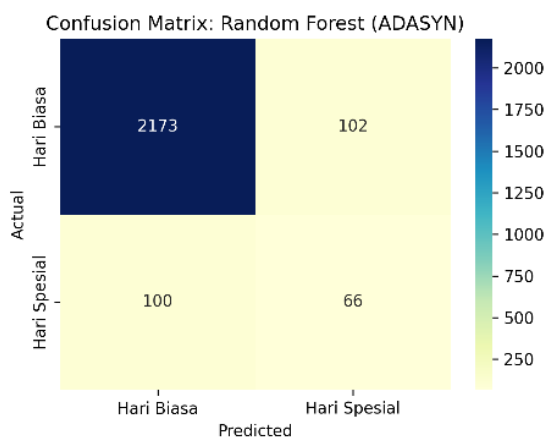
- Metode random oversampling meningkatkan jumlah data pada kelas minoritas secara acak.
- Model tetap sangat akurat untuk Hari Biasa (2222 benar), namun kemampuan mengenali Hari Spesial masih rendah (hanya 39 benar dari 166).
- False negative masih tinggi (127), menunjukkan oversampling acak belum cukup membantu mengenali kelas minoritas.



Gambar 7. Confusion Matrix Random Forest (SMOTE)

Berikut penjelasan dari Gambar 7:

- SMOTE berhasil meningkatkan deteksi Hari Spesial menjadi 69 kasus benar (dari 166), lebih baik dari random oversampling.
- False negative menurun menjadi 97, sementara false positive juga menurun menjadi 96.
- Performa menjadi lebih seimbang meskipun masih dominan mengenali Hari Biasa.

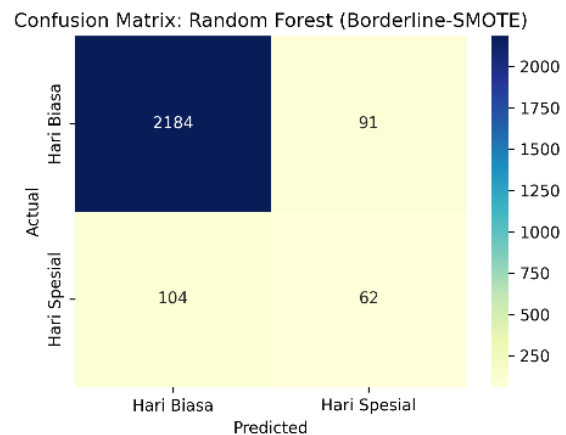


Gambar 8. Confusion Matrix Random Forest (ADASYN)

Berikut penjelasan dari Gambar 8:

- ADASYN menghasilkan 66 prediksi benar untuk Hari Spesial, sedikit lebih rendah dari SMOTE namun false negative lebih kecil (100).
- Model masih cenderung lebih akurat untuk Hari Biasa (2173 benar).

- ADASYN cukup seimbang, dengan distribusi kesalahan yang lebih terkontrol.

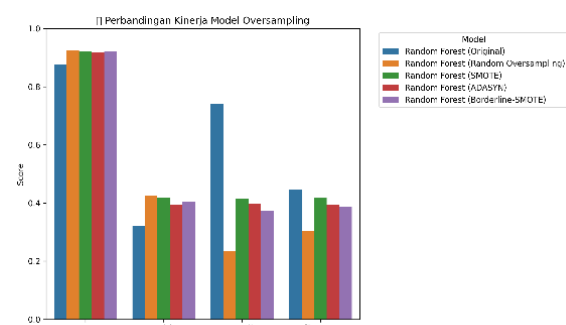


Gambar 9. Confusion Matrix Random Forest (Borderline-SMOTE)

Berikut penjelasan dari Gambar 9:

- Borderline-SMOTE menghasilkan performa serupa ADASYN dengan 62 prediksi benar untuk Hari Spesial, dan false negative 104.
- Model sangat baik dalam mengenali Hari Biasa (2184 benar).
- Borderline-SMOTE fokus pada sampel minoritas yang berada di area batas, namun hasilnya masih belum signifikan jauh lebih baik.

Berdasarkan evaluasi, SMOTE menunjukkan keseimbangan terbaik antara precision dan recall, menandakan kemampuannya menghasilkan sampel minoritas yang representatif. Meski recall belum ideal, SMOTE unggul dalam F1-score dibanding ADASYN yang cenderung menghasilkan noise karena variasi sampel yang tinggi. Sementara itu, Random Oversampling menghasilkan precision tertinggi, namun recall terendah, menunjukkan model sering benar tapi banyak melewati Hari Spesial..



Gambar 10. Perbandingan Kinerja Model Oversampling

Berikut penjelasan dari Gambar 10:

- SMOTE menunjukkan performa terbaik secara keseluruhan dengan keseimbangan precision, recall, dan F1-score tertinggi.

- b. ADASYN memiliki recall yang baik namun precision lebih rendah, kemungkinan akibat noise dari sampel sintetis.
- c. Random Oversampling menghasilkan precision tertinggi, tetapi recall paling rendah, menunjukkan model terlalu konservatif terhadap kelas minoritas.
- d. Borderline-SMOTE memiliki performa sedang, tapi tidak melebihi SMOTE.
- e. Secara umum, teknik sintetis seperti SMOTE dan ADASYN unggul dalam recall, meskipun ada penurunan di precision dibanding Random Oversampling.

4. 2. Pembahasan

Penelitian ini menunjukkan bahwa ketidakseimbangan kelas dalam data transaksi e-commerce berdampak signifikan terhadap performa model klasifikasi. Model Random Forest tanpa oversampling memiliki akurasi tinggi, tetapi precision untuk kelas minoritas (Hari Spesial) sangat rendah, menandakan bias terhadap kelas mayoritas (Hari Biasa). Hal ini menjadi tantangan karena transaksi pada Hari Spesial penting dalam strategi bisnis seperti promosi dan personalisasi.

Untuk mengatasi masalah ini, digunakan empat teknik oversampling: Random Oversampling, SMOTE, ADASYN, dan Borderline SMOTE. Meskipun Random Oversampling mudah diterapkan dan meningkatkan akurasi, teknik ini gagal meningkatkan recall, yang mengindikasikan bahwa model masih kesulitan mengenali pola pada kelas minoritas.

SMOTE memberikan performa terbaik dengan F1-score tertinggi. Dengan mensintesis data berdasarkan tetangga terdekat dalam kelas minoritas, SMOTE meningkatkan recall tanpa banyak mengorbankan precision. Namun, ada risiko menghasilkan data yang tidak realistis jika distribusi awal kompleks atau mengandung outlier.

ADASYN, yang lebih adaptif dan fokus pada sampel sulit, menunjukkan performa yang mendekati SMOTE namun sedikit lebih rendah. Borderline SMOTE, yang menambahkan data sintetis di sekitar batas kelas, memberikan keseimbangan precision dan recall yang cukup baik, meskipun masih di bawah SMOTE. Fokus yang terlalu besar pada perbatasan kelas juga meningkatkan risiko overfitting.

Berdasarkan hasil evaluasi seperti confusion matrix dan metrik lainnya, pemilihan teknik oversampling yang tepat sangat penting dalam menangani data tidak seimbang. Temuan ini menunjukkan bahwa kualitas data sintetis harus diperhatikan, terutama dalam konteks bisnis yang membutuhkan akurasi tinggi untuk pengambilan keputusan.

Kelebihan penelitian ini meliputi penggunaan berbagai teknik oversampling dan evaluasi komparatif yang menyeluruh. Namun, keterbatasannya terletak pada hanya menggunakan satu algoritma klasifikasi, yaitu Random Forest. Penelitian lanjutan dapat

mengeksplorasi algoritma lain seperti XGBoost, LightGBM, atau neural network, serta mempertimbangkan parameter tuning dan penghapusan outlier untuk hasil yang lebih optimal.

Secara keseluruhan, penelitian ini memberikan kontribusi penting dalam pengembangan model klasifikasi pada data tidak seimbang, khususnya di ranah e-commerce, dengan potensi peningkatan strategi bisnis berbasis data.

5. KESIMPULAN DAN SARAN

Penelitian ini menegaskan bahwa ketidakseimbangan kelas berdampak signifikan terhadap kinerja model klasifikasi transaksi e-commerce, terutama dalam mengenali Hari Spesial. Di antara teknik oversampling yang diuji, SMOTE memberikan hasil paling seimbang dalam hal precision dan recall, serta menghasilkan F1-score tertinggi. Oleh karena itu, SMOTE direkomendasikan sebagai metode penanganan data tidak seimbang dalam konteks ini. Untuk pengembangan lebih lanjut, penelitian dapat diarahkan pada kombinasi teknik oversampling dengan algoritma klasifikasi lain serta pengujian pada dataset yang lebih kompleks guna meningkatkan generalisasi model.

DAFTAR PUSTAKA

- [1] D. E. B. Jabat, L. L. Tarigan, M. Purba, dan M. Purba, "Pemanfaatan Platform E-Commerce Melalui Marketplace," *Jurnal Ekonomi, Bisnis Dan Teknologi*, vol. 2, no. 2, pp. 16–21, 2022.
- [2] T. M. Putra, N. Nurhayati, dan S. Pohan, "Pengaruh Iklan Tangkal Kembar terhadap Perilaku Impulsive Buying pada Platform E-Commerce di Indonesia: Kajian Literatur Kualitatif," *Jurnal Penelitian Inovatif*, vol. 4, no. 4, pp. 2625–2638, 2024. doi: <https://doi.org/10.54082/jupin.999>
- [3] A. Nugroho dan D. Harini, "Teknik Random Forest untuk Meningkatkan Akurasi Data Tidak Seimbang," *JSITIK: Jurnal Sistem Informasi dan Teknologi Informasi Komputer*, vol. 2, no. 2, pp. 128–140, 2024. doi: <https://doi.org/10.53624/jsitik.v2i2.379>
- [4] M. Aqsha dan N. Sunusi, "Performa Klasifikasi Data Tidak Seimbang Dengan Pendekatan Machine Learning (Studi Kasus: Diabetes Indian Pima)," *Jurnal Matematika UNAND*, vol. 12, no. 2, pp. 176–193, 2023. doi: <https://doi.org/10.25077/jmua.12.2.176-193.2023>
- [5] R. Ridwan, E. H. Hermaliani, dan M. Ernawati, "Penerapan metode SMOTE untuk mengatasi imbalanced data pada klasifikasi ujaran kebencian," *Computer Science (Co-Science)*, vol. 4, no. 1, pp. 80–88, 2024. doi: <https://doi.org/10.31294/coscience.v4i1.2990>
- [6] S. Amaliah, M. Nusrang, dan A. Aswi, "Penerapan Metode Random Forest Untuk

- Klasifikasi Varian Minuman Kopi di Kedai Kopi Konijiwa Bantaeng," *VARIANSI: Journal of Statistics and Its Application on Teaching and Research*, vol. 4, no. 3, pp. 121–127, 2022. doi: <https://doi.org/10.35580/variansiunm31>
- [7] N. Rahayu, D. Rakhmawati, and I. G. A. Arista, "Pengaruh live streamer dan promo tanggal kembar terhadap impulsive buying (studi kasus pengguna e-commerce Shopee di Kota Mataram)," *Ganec Swara*, vol. 18, no. 1, pp. 263–268, Mar. 2024.
- [8] L. Ariyanti dan S. Setyo Iriani, "Pengaruh Promosi Penjualan Dan Electronic Word of Mouth Terhadap Pembelian Impulsif Pada Saat Special Event Day (Studi Pada Konsumen Shopee Jawa Timur)," *SIBATIK JOURNAL*, vol. 1, no. 5, pp. 491–508, 2022. doi: <https://doi.org/10.54443/sibatik.v1i5.58>
- [9] A. N. Siburian dan N. Anggrainie, "Pengaruh Hedonic Shopping Motivation, Brand Image, Brand Ambassador, Diskon, Harga dan Sales Promotion terhadap Pembelian Implusif Pada e-Commerce Tiktok Shop Dimasa Pandemi Covid-19," *Jurnal Mirai Management*, vol. 7, no. 3, pp. 176–191, 2022. doi: <https://doi.org/10.37531/mirai.v7i3.2820>
- [10] Z. Alamin, R. Missouri, S. Sutriawan, F. Fathir, dan K. Khairunnas, "Perkembangan E-commerce: Analisis Dominasi Shopee sebagai Primadona Marketplace di Indonesia," *J-ESA (Jurnal Ekonomi Syariah)*, vol. 6, no. 2, pp. 120–131, 2023. doi: <https://doi.org/10.52266/jesa.v6i2.2484>
- [11] R. Roisah dan W. Dianita, "Pengaruh Promosi Penjualan Dan Shopping Enjoyment Terhadap Pembelian Impulsif Pada E-Commerce Shopee," *Service Management Triangle: Jurnal Manajemen Jasa*, vol. 5, no. 1, pp. 9–19, 2023. doi: <https://doi.org/10.51977/jsj.v5i1.1204>
- [12] R. Hermiati, A. Asnawati, dan I. Kanedi, "Pembuatan E-Commerce Pada Raja Komputer Menggunakan Bahasa Pemrograman PHP dan Database MySQL," *Jurnal Media Infotama*, vol. 17, no. 1, pp. 54–66, 2021. doi: <https://doi.org/10.37676/jmi.v17i1.1317>
- [13] M. H. Setiawan, R. Komarudin, dan D. N. Kholifah, "Pengaruh Kepercayaan, Tampilan Dan Promosi Terhadap Keputusan Pemilihan Aplikasi Marketplace," *Jurnal Infortech*, vol. 4, no. 2, p. 141, 2022. doi: <https://doi.org/10.31294/infortech.v4i2.13913>
- [14] M. R. A. Fitra, A. A. S. Effendi, dan F. Ramadhani, "Implementasi Python Dalam Pengolahan Data Pribadi Mahasiswa Ilmu Komputer Angkatan 23 Pada Universitas Negeri Medan Menggunakan Struktur Data Linked List," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 9, no. 1, pp. 51–58, 2025. doi: <https://doi.org/10.36040/jati.v9i1.12164>
- [15] M. R. S. Alfarizi, M. Z. Al-farish, M. Taufiqurrahman, G. Ardiansah, dan M. Elgar, "Penggunaan Python Sebagai Bahasa Pemrograman untuk Machine Learning dan Deep Learning," *KARIMAH TAUHID*, vol. 2, no. 1, pp. 1–6, 2023. doi: <https://doi.org/10.30997/karimahtauhid.v2i1.7518>
- [16] S. Diantika, "Penerapan teknik random oversampling untuk mengatasi imbalance class dalam klasifikasi website phishing menggunakan Algoritma LightGBM," *JATI (Jurnal Mahasiswa Teknik Informatika)*, vol. 7, no. 1, pp. 19–25, 2023. doi: <https://doi.org/10.36040/jati.v7i1.6006>