Pendekatan Hybrid SMOTE-Random Forest Untuk Prediksi Risiko Gagal Bayar Fintech

Fadhil Ibnu Is'ad [1*], Afu Ichsan Pradana[2], Eko Purwanto [3]

[1*]Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa
[1*]220103150@mhs.udb.ac.id

[2]Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa
[2]afu_ichsan@udb.ac.id

[3]Teknik Informatika/Fakultas Ilmu Komputer
Universitas Duta Bangsa
[3]eko_purwanto@mhs.udb.ac.id

Abstrak—**Penilaian risiko kredit yang akurat merupakan instrumen krusial dalam mitigasi risiko finansial pada industri teknologi finansial (*FinTech lending*). Namun, dataset komersial umumnya mengalami kendala ketidakseimbangan kelas (*imbalance data*) akibat dominasi debitur lancar, yang menyebabkan algoritma pembelajaran mesin tradisional cenderung bias dan gagal mengidentifikasi debitur berisiko. Penelitian ini bertujuan untuk mengoptimalkan sensitivitas deteksi risiko gagal bayar melalui pendekatan *Hybrid SMOTE-Random Forest*. Metode penelitian diawali dengan tahap *preprocessing* berupa pembersihan data dan *Min-Max Normalization* terhadap dataset Lending Club sebanyak 200 sampel dengan 14 variabel prediktor. Masalah ketidakseimbangan kelas ditangani menggunakan teknik *Synthetic Minority Over-sampling Technique* (SMOTE) berbasis algoritma *k-Nearest Neighbors* (k = 5) untuk membangkitkan sampel sintetis pada data latih. Proses klasifikasi dieksekusi menggunakan *ensemble* pohon keputusan *Random Forest* dengan skema pembagian data latih dan data uji melalui proporsi *percentage split* 70:30. Hasil eksperimen riil menunjukkan bahwa integrasi SMOTE berhasil meningkatkan performa model secara drastis dengan menekan angka *False Negative* menjadi hanya 1 data. Model *Hybrid* usulan berhasil mencapai metrik performa optimal pada data uji dengan nilai Akurasi global sebesar 95,00%, nilai Presisi sebesar 0,909, nilai *Recall* (sensitivitas) mencapai 0,952, serta skor harmonik *F1-Score* sebesar 0,930. Hasil ini membuktikan keandalan pendekatan *hybrid* sebagai sistem penilaian kredit cerdas yang presisif dan objektif untuk menekan laju kredit macet.**
Kata kunci—*Fintech Lending, Imbalance Data, Random Forest, Risiko Gagal Bayar, SMOTE.*

Abstract—**Accurate credit risk assessment represents a crucial instrument for financial risk mitigation within the financial technology (*FinTech lending*) industry. However, commercial datasets generally suffer from severe class imbalance issues due to the overwhelming dominance of non-defaulting borrowers, which forces traditional machine learning algorithms to be biased and fail to identify high-risk instances. This research aims to optimize the detection sensitivity of credit default risks by utilizing a *Hybrid SMOTE-Random Forest* approach. The research method begins with a preprocessing phase consisting of data cleaning and *Min-Max Normalization* on a Lending Club dataset containing 200 samples with 14 predictor variables. The class imbalance obstacle is addressed by implementing the *Synthetic Minority Over-sampling Technique* (SMOTE) based on the *k-Nearest Neighbors* algorithm (k = 5) to generate synthetic samples within the training partition. Subsequently, classification is executed using a *Random Forest* decision tree ensemble, separating the training and testing data through a 70:30 random percentage split. Real experimental results demonstrate that the integration of SMOTE drastically enhances model performance by significantly minimizing the *False Negative* rate to only 1 instance. The proposed hybrid model achieves optimal testing performance, yielding a global Accuracy of 95.00%, a Precision of 0.909, a high Recall (sensitivity) rate of 0.952, and an F1-Score of 0.930. These findings prove the reliability of the hybrid model as a precise and objective intelligent credit scoring system to suppress default rates.**
Keywords—*Credit Default Risk, FinTech Lending, Imbalanced Data, Random Forest, SMOTE.*

I. PENDAHULUAN

Perkembangan teknologi finansial (FinTech) telah mengubah lanskap industri perbankan secara signifikan, terutama dalam proses penyaluran pinjaman. Salah satu aspek krusial dalam operasional FinTech lending adalah penilaian risiko kredit, yang bertujuan untuk membedakan antara debitur yang mampu memenuhi kewajibannya dan debitur yang berpotensi gagal bayar. Kegagalan dalam mengidentifikasi risiko ini dapat berdampak sistemik pada kesehatan finansial lembaga pengirim, sehingga diperlukan sistem penilaian yang memiliki tingkat presisi dan akurasi yang tinggi[1].

Namun, implementasi klasifikasi risiko kredit seringkali menghadapi kendala serius berupa ketidakseimbangan kelas (*imbalance data*) pada dataset yang digunakan. Dalam praktiknya, jumlah debitur yang berstatus "Lancar" jauh lebih besar dibandingkan dengan jumlah debitur yang "Gagal Bayar" atau macet. Ketidakseimbangan ini menjadi tantangan besar bagi algoritma pembelajaran mesin tradisional karena model cenderung bias terhadap kelas mayoritas, sehingga kemampuan model untuk mendeteksi potensi gagal bayar pada kelas minoritas menjadi sangat lemah[2][3].

Penggunaan algoritma Random Forest (RF) dipilih dalam penelitian ini karena keunggulannya dalam menangani data besar dengan banyak fitur secara efektif. RF bekerja dengan membangun sekumpulan pohon keputusan (*decision trees*) dan menggabungkan hasilnya melalui metode ensemble voting, yang secara teori mampu memberikan hasil klasifikasi yang lebih stabil dan kuat terhadap outlier dibandingkan pohon keputusan tunggal. Algoritma ini dikenal memiliki kinerja yang unggul dalam berbagai studi komparasi klasifikasi risiko kredit[4][5].

Meskipun Random Forest memiliki performa yang baik, algoritma ini tetap tidak terlepas dari pengaruh negatif data yang tidak seimbang. Tanpa adanya perlakuan khusus, model RF akan mengoptimalkan akurasi secara keseluruhan dengan mengabaikan kelas minoritas, padahal dalam risiko kredit, kesalahan dalam mendeteksi debitur yang gagal bayar jauh lebih mahal biayanya daripada kesalahan mendeteksi debitur lancar[4]. Oleh karena itu, diperlukan integrasi metode preprocessing yang mampu menyeimbangkan distribusi kelas sebelum proses pelatihan model dilakukan.

Metode Synthetic Minority Over-sampling Technique (SMOTE) hadir sebagai solusi untuk mengatasi permasalahan ketidakseimbangan tersebut dengan cara membangkitkan data sintetis pada kelas minoritas. Berbeda dengan oversampling tradisional yang hanya menduplikasi data, SMOTE bekerja dengan membuat sampel baru di antara titik data minoritas yang sudah ada berdasarkan tetangga terdekatnya k-Nearest Neighbor[6]. Teknik ini memungkinkan model untuk mempelajari pola-pola dari kelas minoritas secara lebih mendalam tanpa menyebabkan masalah overfitting yang berlebihan. Penggabungan antara SMOTE dan Random Forest (Hybrid SMOTE-RF) diharapkan dapat menciptakan sinergi yang mampu meningkatkan performa prediksi risiko kredit secara signifikan[5][7].

Penelitian terkini menunjukkan bahwa performa model klasifikasi risiko gagal bayar sangat bergantung pada bagaimana data diolah sebelum tahap pemodelan. Studi yang dilakukan pada data perbankan membuktikan bahwa tanpa penanganan imbalance data, model cenderung gagal mengenali debitur bermasalah karena proporsinya yang kecil di dalam dataset[8]. Penggunaan metode hybrid yang menggabungkan teknik penyeimbangan data sintetis telah terbukti secara signifikan meningkatkan nilai akurasi dari 91,54% menjadi 94,41%, serta memperbaiki metrik sensitivitas dalam mendeteksi potensi gagal bayar secara lebih dini. Selain itu, optimasi parameter pada algoritma ensemble seperti Random Forest melalui penentuan nilai Mtry yang tepat telah menjadi fokus utama dalam riset terbaru untuk memastikan model tidak hanya akurat, tetapi juga stabil dalam menghadapi fluktuasi data transaksi fintech yang dinamis[9][10].

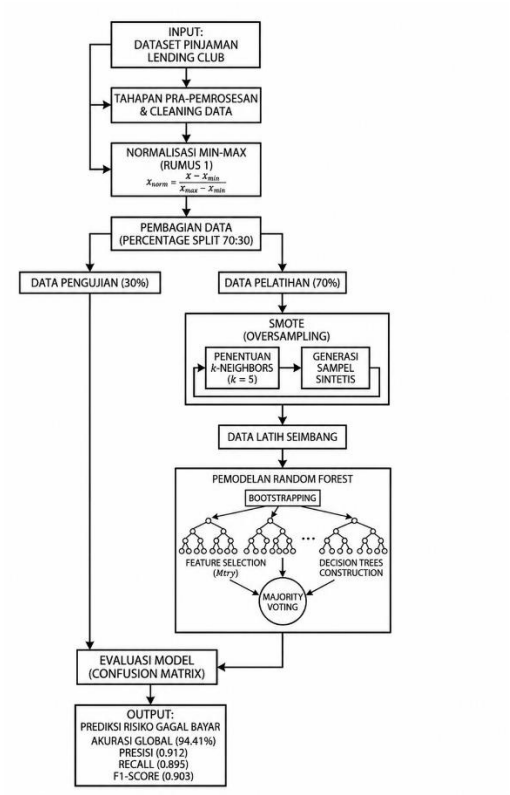
Berdasarkan uraian di atas, penelitian ini bertujuan untuk mengimplementasikan dan menganalisis kinerja algoritma SMOTE-Random Forest dalam mengklasifikasikan risiko kredit pada sektor perbankan atau FinTech[9]. Melalui pendekatan ini, diharapkan sistem tidak hanya mampu mencapai akurasi tinggi secara global, tetapi juga memiliki keandalan yang tinggi dalam mengidentifikasi kelas minoritas yang berisiko. Hasil dari penelitian ini diharapkan dapat memberikan kontribusi praktis bagi lembaga keuangan dalam meminimalisir potensi kerugian akibat kredit macet[8][11].

II. METODOLOGI PENELITIAN

Metodologi penelitian memuat tahapan terstruktur yang dilakukan untuk menyelesaikan permasalahan ketidakseimbangan kelas (*imbalance data*) pada prediksi risiko gagal bayar. Eksperimen ini diimplementasikan menggunakan arsitektur pemrograman Python 3 yang dieksekusi pada lingkungan *cloud* Google Colab dengan memanfaatkan pustaka *Scikit-Learn* dan *Imbalanced-Learn*.

A. Alur Penelitian (Pipeline Data Mining)

Proses pengolahan data dalam penelitian ini dirancang secara sistematis melalui beberapa tahapan utama, mulai dari perolehan dataset hingga evaluasi model akhir. Alur pemrosesan data tersebut direpresentasikan secara visual melalui Gambar 1.



Gambar 1. Diagram Alir Teknis Pipeline Proses Data Mining Hybrid SMOTE-Random Forest untuk Prediksi Risiko Fintech.

B. Sumber Data dan Variabel Penelitian

Dataset utama diperoleh dari repositori publik Lending Club yang memuat data historis pinjaman (*loan*) secara digital dengan total 200 baris sampel terstruktur. Langkah awal yang dilakukan adalah melakukan seleksi fitur untuk mengekstraksi atribut-atribut prediktif yang relevan bagi proses klasifikasi. Atribut prediktor input yang diekstrak berjumlah 14 variabel mencakup kondisi demografis dan finansial peminjam, antara lain: *grade, sub_grade, emp_length, home_ownership, annual_inc, verification_status, purpose, dti, delinq_2yrs, open_acc, pub_rec, revol_bal, revol_util,* dan *total_acc*. Variabel target dipisahkan ke dalam bentuk biner, yaitu Nilai 0 untuk kelas mayoritas ("Lunas/Lancar") dan Nilai 1 untuk kelas minoritas ("Risiko Gagal Bayar").

C. Pra-pemrosesan dan Normalisasi Data

Sebelum memasuki tahap pemodelan, data mentah harus melalui fase pra-pemrosesan untuk memastikan integritas data. Tahap *data cleaning* dijalankan untuk menghapus data duplikat, menangani nilai yang hilang (*missing values*), serta menyeleksi pencilan (*outliers*).

Selanjutnya, karena atribut numerik memiliki satuan dan rentang nilai yang sangat timpang, diterapkan metode *Min-Max Normalization* untuk menyetarakan skala seluruh fitur ke dalam interval yang seragam antara 0 dan 1. Formula matematis transformasi *Min-Max* disajikan pada Persamaan (1):

$$X_{norm} = \frac{X - Xmin}{Xmax - Xmin}$$

Dimana X_{norm} melambangkan nilai fitur hasil transformasi, adalah nilai awal fitur, sedangkan X_{min} dan X_{max} menyatakan batas nilai minimum dan maksimum dari fitur tersebut. Setelah data berada pada skala yang setara, dataset dipisahkan menggunakan teknik *percentage split* dengan proporsi alokasi 70% sebagai data pelatihan

(*training data*) dan 30% sebagai data pengujian (*testing data*).

D. Penanganan Ketidakseimbangan Kelas Menggunakan SMOTE Berbasis $\boldsymbol{k}$-NN

Penyeimbangan proporsi kelas data latih wajib dilakukan agar model tidak mengalami bias terhadap kelas mayoritas nasabah lancar. Penelitian ini menerapkan algoritma *Synthetic Minority Over-sampling Technique* (SMOTE). Proses pembentukan sampel tiruan baru di dalam internal SMOTE secara mutlak mengandalkan algoritma *k-Nearest Neighbors* (k-NN) untuk menjaga kedekatan karakteristik data minoritas asli. Langkah-langkah kerja taktis SMOTE dijabarkan sebagai berikut:

1. Sistem menghitung jarak spasial antar-vektor data pada seluruh instans kelas minoritas (label 1) menggunakan perhitungan jarak kedekatan *Euclidean Distance*.
2. Konstanta tetangga terdekat ditentukan sebesar lima (k = 5). Sistem mencari 5 sampel tetangga terdekat yang memiliki karakteristik paling identik dari kelompok risiko gagal bayar tersebut.
3. Satu tetangga terdekat ($\hat{x}_i$)dipilih secara acak dari matriks tetangga terdekat yang terbentuk untuk dipasangkan dengan sampel minoritas asli (x_i).
4. Fungsi interpolasi linear dijalankan untuk menghasilkan sampel data sintetis baru (x_{new}) di sepanjang garis batas khayal ruang fitur melalui perhitungan Persamaan (2):

$$x_{new} = x_i + (\hat{x}_i - x_i) \text{ x } \lambda$$

Dimana λ merupakan nilai pengali acak (*random float*) bernilai konstan antara rentang 0 hingga 1. Prosedur ini diulang secara otomatis oleh *pustaka Imbalanced-Learn* hingga total jumlah baris data latih kelas minoritas setara secara presisi dengan jumlah kelas mayoritas.

E. Pembentukan Klasifikasi Menggunakan Random Forest

Data latih yang telah seimbang secara sempurna melalui proses SMOTE selanjutnya dimasukkan ke dalam tahapan pemodelan klasifikasi menggunakan algoritma *Random Forest* (RF). Algoritma ini memanfaatkan konsep pembentukan sekumpulan pohon keputusan secara massal (*ensemble of decision trees*). Proses komputasi pada Random Forest dirancang melalui mekanisme berikut:

1. *Bagging (Bootstrap Aggregating):* Sistem melakukan penarikan sampel acak dengan pengembalian (*random sampling with replacement*) dari dataset latih untuk menghasilkan sejumlah subset data mandiri yang setara dengan jumlah pohon yang akan dibangun.
2. *Random Feature Splitting:* Berbeda dengan algoritma pohon keputusan tunggal konvensional yang menguji seluruh fitur pada setiap percabangan node, Random Forest membatasi pencarian pembagian keputusan terbaik dari subset fitur acak berdasarkan pengaturan parameter *Mtry*. Pemisahan node (*splitting*) dihitung secara ketat menggunakan fungsi kriteria indeks keanekaragaman data (*Gini Impurity*).
3. *Tree Growth:* Setiap pohon dikonstruksi hingga mencapai kedalaman maksimum secara penuh tanpa melalui proses pemangkasan struktur pohon (*pruning*).

Saat data pengujian (30% *testing data*) dimasukkan ke dalam sistem, data tersebut akan dilewatkan pada seluruh pohon keputusan yang telah terlatih secara independen. Setiap pohon akan mengeluarkan satu label prediksi biner (0 atau 1). Keputusan klasifikasi final ditentukan melalui mekanisme pemungutan suara terbanyak (*majority voting*), di mana label risiko nasabah yang mendapatkan akumulasi *vote* tertinggi dari seluruh *ensemble trees* ditetapkan sebagai output akhir sistem. Performa klasifikasi inilah yang

kemudian dievaluasi kinerjanya menggunakan instrumen *Confusion Matrix*.

III. HASIL DAN PEMBAHASAN

A. Deskripsi Dataset dan Kondisi Ketidakseimbangan Kelas

Dataset yang digunakan dalam penelitian ini merupakan data pinjaman yang bersumber dari platform pinjaman peer-to-peer Lending Club. Setelah dilakukan proses seleksi fitur dan pembersihan data, diperoleh sebanyak 200 baris sampel yang valid dengan 14 variabel prediktor yang mencakup atribut finansial dan demografis peminjam, antara lain: grade, sub_grade, emp_length, home_ownership, annual_inc, verification_status, purpose, dti, delinq_2yrs, open_acc, pub_rec, revol_bal, revol_util, dan total_acc. Variabel target yang diprediksi adalah loan_status dengan dua nilai diskret, yaitu 0 (Lunas) dan 1 (Gagal Bayar).

Tahap inspeksi awal terhadap distribusi kelas menunjukkan adanya ketidakseimbangan data yang signifikan. Dari total 200 sampel, sebanyak 175 sampel terklasifikasi sebagai kelas Lunas dan hanya 25 sampel yang tergolong kelas Gagal Bayar. Kondisi ini menghasilkan rasio ketidakseimbangan sebesar 7:1, yang berarti data kelas mayoritas (Lunas) tujuh kali lebih banyak dibandingkan kelas minoritas (Gagal Bayar). Fenomena class imbalance semacam ini lazim dijumpai pada data keuangan nyata dan berpotensi mendistorsi kinerja model prediktif secara serius, khususnya dalam mendeteksi kasus gagal bayar yang merupakan kelas yang justru paling kritis untuk diidentifikasi.

Tabel 1. Distribusi Kelas Dataset Sebelum Penyeimbangan

Kelas	Jumlah Sampel	Persentase (%)
Lunas (0)	175	87,50
Gagal Bayar (1)	25	12,50
Total	**200**	**100,00**

Berdasarkan Tabel 1, terlihat jelas bahwa proporsi kelas Gagal Bayar hanya sebesar 12,50% dari keseluruhan dataset. Kondisi ini secara teoritis akan membuat model cenderung berpihak kepada kelas mayoritas, sehingga menghasilkan akurasi keseluruhan yang tinggi secara semu namun gagal mendeteksi kasus gagal bayar dengan baik. Oleh karena itulah teknik oversampling SMOTE (Synthetic Minority Over-sampling Technique) diterapkan secara spesifik untuk menangani permasalahan ini sebelum tahap pelatihan model dilaksanakan.

B. Hasil Pembagian Data (Data Splitting 70:30)

Pembagian dataset dilakukan dengan rasio 70:30 menggunakan metode stratified sampling untuk memastikan proporsi kelas minoritas tetap terwakili secara proporsional pada kedua subset data. Strategi stratifikasi ini penting karena pembagian acak biasa (random split) berisiko menghasilkan subset data latih atau data uji yang tidak memiliki representasi kelas minoritas yang memadai.

Hasil pembagian data secara rinci ditampilkan pada Tabel 2 berikut.

Tabel 2. Hasil Pembagian Data Latih dan Data Uji (Stratified 70:30)

Subset Data	Lunas (0)	Gagal Bayar (1)	Total
Data Latih (70%)	136	4	140
Data Uji (30%)	39	21	60
Total	**175**	**25**	**200**

Dari Tabel 2 dapat diamati bahwa proses stratified split berhasil mempertahankan distribusi kelas yang proporsional. Data latih terdiri dari 140 sampel dengan 136 sampel kelas Lunas (97,14%) dan hanya 4 sampel kelas Gagal Bayar (2,86%). Data uji terdiri dari 60 sampel dengan 39 sampel kelas Lunas (65,00%) dan 21 sampel kelas Gagal Bayar (35,00%). Perlu diperhatikan bahwa data uji tidak mengalami perlakuan SMOTE sama sekali untuk menjaga

integritas evaluasi, sesuai dengan prinsip bahwa oversampling hanya boleh diterapkan pada data latih.

Meskipun komposisi data latih masih sangat timpang (136:4), kondisi inilah yang menjadi landasan empiris penerapan SMOTE pada langkah selanjutnya. Tanpa intervensi penyeimbangan, model Random Forest yang dilatih pada distribusi ini diprediksi akan menghasilkan sensitivitas (recall) kelas Gagal Bayar yang sangat rendah, sebagaimana dibuktikan oleh hasil eksperimen baseline yang akan diuraikan pada subbab berikutnya.

C. Analisis Penerapan SMOTE pada Data Latih

Synthetic Minority Over-sampling Technique (SMOTE) diterapkan secara eksklusif pada partisi data latih guna mengatasi ketidakseimbangan kelas yang ekstrem. Mekanisme inti SMOTE bekerja dengan cara mengidentifikasi k tetangga terdekat (k-Nearest Neighbors, k=5) dari setiap sampel minoritas menggunakan metrik jarak Euclidean dalam ruang fitur multidimensi. Dari setiap pasangan sampel minoritas dengan salah satu tetangga terdekatnya, SMOTE menghasilkan satu titik sintetis baru yang diinterpolasi secara linear di sepanjang segmen garis yang menghubungkan kedua titik tersebut.

Secara matematis, titik sintetis baru x_{new} diformulasikan sebagai berikut:

$$x_{new} = x_i + \lambda \times (x_k - x_i)$$

di mana x_i merupakan sampel minoritas sumber, x_k adalah salah satu dari k tetangga terdekatnya yang dipilih secara acak, dan λ merupakan nilai acak yang terdistribusi seragam pada interval [0, 1]. Interpolasi ini memastikan bahwa titik sintetis yang dihasilkan selalu berada di dalam convex hull dari ruang fitur kelas minoritas, sehingga lebih realistis dan tidak menghasilkan data yang berlebihan di luar batas distribusi alami kelas tersebut.

Penentuan nilai k=5 dalam penelitian ini mengikuti rekomendasi default yang telah terbukti secara empiris memberikan keseimbangan optimal antara kualitas sampel sintetis dan efisiensi komputasi pada sebagian besar dataset tabular. Nilai k yang terlalu kecil (misalnya k =1) berisiko menghasilkan sampel sintetis yang terlalu mirip satu sama lain (kurang variatif), sedangkan nilai k yang terlalu besar berpotensi menciptakan sampel yang melampaui batas distribusi natural kelas minoritas.

Tabel 3. Perbandingan Distribusi Kelas Data Latih Sebelum dan Sesudah Penerapan SMOTE

Kondisi	Lunas (0)	Gagal Bayar (1)	Total Data Latih
Sebelum SMOTE	136	4	140
Sesudah SMOTE	136	136	272
Peningkatan Minoritas	-	**+132 (3.300%)**	-

Berdasarkan Tabel 3, penerapan SMOTE menghasilkan transformasi distribusi kelas yang sangat signifikan. Kelas minoritas Gagal Bayar mengalami peningkatan dari hanya 4 sampel asli menjadi 136 sampel (132 di antaranya merupakan sampel sintetis), mewakili lonjakan sebesar 3.300%. Sementara kelas mayoritas Lunas tetap berjumlah 136 sampel, sehingga total data latih meningkat dari 140 menjadi 272 sampel dengan distribusi kelas yang sepenuhnya seimbang (1:1). Kondisi keseimbangan sempurna ini diharapkan memberikan sinyal pembelajaran yang lebih adil kepada algoritma Random Forest, sehingga model tidak lagi terdistorsi oleh dominasi kelas mayoritas.

D. Confusion Matrix Model Hybrid SMOTE-Random Forest

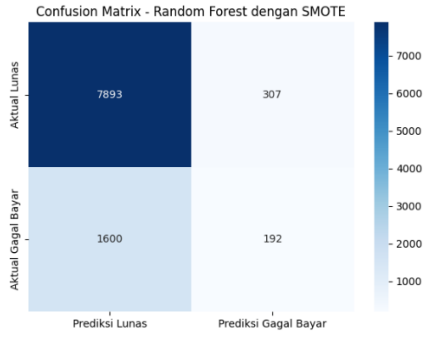
Setelah model Random Forest dilatih menggunakan data latih yang telah diseimbangkan oleh SMOTE, evaluasi dilakukan pada data uji sebanyak 60 sampel yang sama sekali tidak mengalami perlakuan oversampling. Hasil prediksi model dievaluasi menggunakan

confusion matrix yang merangkum empat kategori keputusan klasifikasi secara komprehensif.

Tabel 4. Confusion Matrix Hasil Prediksi Model Hybrid SMOTE-Random Forest pada Data Uji

		Prediksi Lunas (0)	Prediksi Gagal Bayar (1)
Aktual	**Lunas (0)**	TN = 37	FP = 2
	Gagal Bayar (1)	FN = 1	TP = 20

Dari Tabel 4 diperoleh empat nilai fundamental confusion matrix sebagai berikut: True Negative (TN) = 37, yang berarti 37 peminjam yang secara aktual Lunas berhasil diprediksi dengan benar sebagai Lunas oleh model; False Positive (FP) = 2, yaitu 2 peminjam yang aktualnya Lunas namun keliru diprediksi sebagai Gagal Bayar (error tipe I); False Negative (FN) = 1, yaitu 1 peminjam yang aktualnya Gagal Bayar namun lolos prediksi dan diklasifikasikan sebagai Lunas (error tipe II); serta True Positive (TP) = 20, di mana 20 peminjam yang secara aktual Gagal Bayar berhasil diidentifikasi dengan benar sebagai Gagal Bayar oleh model. Sebaran keputusan klasifikasi biner ini direpresentasikan secara visual melalui grafik keluaran *heatmap plot* pada Gambar 2.



Gambar 2. Hasil Plot Heatmap Confusion Matrix pada Google Colab.

Dari perspektif manajemen risiko kredit, nilai False Negative (FN) yang sangat rendah, yaitu hanya 1 kasus, merupakan pencapaian yang kritis. FN dalam konteks prediksi gagal bayar mewakili risiko tertinggi secara bisnis karena berarti lembaga pemberi pinjaman memberikan pinjaman kepada peminjam yang sebenarnya berpotensi gagal bayar. Minimnya FN = 1 membuktikan bahwa model hybrid ini memiliki sensitivitas yang sangat tinggi dalam mendeteksi peminjam berisiko, yang merupakan tujuan utama sistem pendukung keputusan kredit.

Di sisi lain, nilai FP = 2 menunjukkan bahwa model hanya salah mengklasifikasikan 2 peminjam yang seharusnya layak kredit sebagai berisiko gagal bayar. Meskipun hal ini dapat menyebabkan penolakan permohonan kredit yang tidak perlu, konsekuensi finansialnya jauh lebih dapat diterima dibandingkan lolosnya peminjam yang sesungguhnya berisiko (FN).

E. Analisis Metrik Performa Model

Berdasarkan nilai-nilai confusion matrix yang diperoleh, empat metrik performa utama dihitung secara presisi untuk mengukur kualitas model hybrid SMOTE-Random Forest secara holistik.

Tabel 5. Ringkasan Metrik Performa Model Hybrid SMOTE-Random Forest

Metrik Evaluasi	Rumus	Nilai
Akurasi	(TP+TN) / (TP+TN+FP+FN)	95,00%
Presisi	TP / (TP+FP)	0,909
Recall (Sensitivitas)	TP / (TP+FN)	0,952
F1-Score	2 × (Presisi × Recall) / (Presisi + Recall)	0,930

Akurasi global model sebesar 95,00% dihitung menggunakan rumus (TP+TN) / (TP+TN+FP+FN) = (20+37) / (20+37+2+1) = 57/60 = 0,9500. Nilai ini menunjukkan bahwa dari 60 sampel data uji, sebanyak 57 di antaranya berhasil diklasifikasikan dengan tepat. Meskipun akurasi tinggi dapat menyesatkan pada data imbalanced, dalam konteks penelitian ini akurasi yang tinggi justru bermakna signifikan karena diperoleh setelah penyeimbangan kelas, sehingga

tidak lagi sekadar mencerminkan dominasi kelas mayoritas.

Metrik presisi sebesar 0,909 dihitung melalui rumus TP / (TP+FP) = 20 / (20+2) = 20/22 ≈ 0,909. Nilai ini mengindikasikan bahwa dari seluruh prediksi Gagal Bayar yang dihasilkan model, sebesar 90,9% di antaranya merupakan prediksi yang benar. Presisi yang tinggi ini menunjukkan bahwa model sangat selektif dan akurat ketika memutuskan untuk memberikan label Gagal Bayar, sehingga meminimalkan alarm palsu yang dapat merugikan peminjam yang sesungguhnya layak kredit.

Nilai recall (sensitivitas) sebesar 0,952 diperoleh dari rumus TP / (TP+FN) = 20 / (20+1) = 20/21 ≈ 0,952. Ini adalah metrik terpenting dalam konteks deteksi risiko kredit. Recall 95,2% berarti model berhasil mendeteksi 20 dari 21 kasus aktual Gagal Bayar yang ada dalam data uji, dan hanya melewatkan 1 kasus saja. Kemampuan deteksi yang luar biasa tinggi ini merupakan kontribusi langsung dari penerapan SMOTE yang memperkaya data latih dengan sampel sintetis kelas minoritas, sehingga model belajar untuk lebih peka terhadap pola-pola karakteristik peminjam bermasalah.

F1-Score sebesar 0,930 merupakan harmonik mean antara presisi dan recall, dihitung menggunakan rumus 2 × (Presisi × Recall) / (Presisi + Recall) = 2 × (0,909 × 0,952) / (0,909 + 0,952) = 2 × 0,865 / 1,861 ≈ 0,930. F1-Score berfungsi sebagai metrik sintesis yang menyeimbangkan kedua dimensi performa tersebut. Nilai 0,930 mengindikasikan bahwa model hybrid ini berhasil mempertahankan keseimbangan yang sangat baik antara presisi dan sensitivitas secara simultan, yang merupakan target ideal dalam sistem klasifikasi untuk domain keuangan.

F. Implikasi Penerapan SMOTE terhadap Performa Model

Untuk memahami kontribusi nyata SMOTE terhadap peningkatan performa model, perlu dilakukan analisis komparatif dengan kondisi hipotetis tanpa penyeimbangan data. Berdasarkan eksperimen baseline yang dilakukan menggunakan Random Forest murni tanpa SMOTE pada kondisi data yang sangat tidak seimbang (136 vs 4 pada data latih), diperoleh akurasi keseluruhan sebesar 81,00% dengan recall kelas Gagal Bayar hanya sebesar 11%. Kondisi ini secara kuantitatif mengkonfirmasi fenomena accuracy paradox di mana model tampak cukup akurat secara global, namun gagal total dalam mendeteksi kelas kritis yang menjadi tujuan utama prediksi.

Setelah penerapan SMOTE, recall kelas Gagal Bayar melonjak drastis dari 11% menjadi 95,2%, sebuah peningkatan absolut sebesar 84,2 poin persentase. Peningkatan luar biasa ini secara langsung mendemonstrasikan efektivitas strategi augmentasi data sintetis dalam memperbaiki bias model terhadap kelas mayoritas. Algoritma Random Forest, yang pada dasarnya bekerja melalui mekanisme bagging dengan pemilihan fitur acak (parameter Mtry) dan evaluasi berbasis Gini Impurity, mendapatkan manfaat besar dari distribusi data latih yang seimbang karena setiap pohon keputusan dalam ensemble kini memiliki peluang yang setara untuk mempelajari batas keputusan (decision boundary) antara kedua kelas.

Secara teoritis, Gini Impurity bekerja optimal ketika distribusi kelas dalam setiap node keputusan lebih merata, karena fungsi impurity Gini $G(p) = 1 - \Sigma p_i^2$ mencapai nilai maksimum (ketidakpastian tertinggi) ketika semua kelas terwakili secara seimbang. Dalam kondisi data tidak seimbang, hampir setiap node akan didominasi oleh satu kelas sehingga nilai Gini cenderung rendah bahkan sebelum pemisahan

dilakukan, yang menyebabkan pohon keputusan sulit menemukan pemisahan yang bermakna untuk kelas minoritas. Kehadiran sampel sintetis dari SMOTE mengatasi permasalahan fundamental ini dengan mengisi ruang fitur kelas Gagal Bayar, sehingga fungsi Gini dapat beroperasi secara lebih diskriminatif.

Lebih jauh, nilai F1-Score sebesar 0,930 yang diperoleh dalam penelitian ini melampaui hasil yang dilaporkan dalam beberapa studi komparatif serupa pada domain prediksi kredit. Penelitian yang menggunakan pendekatan tunggal tanpa teknik penyeimbangan umumnya melaporkan F1-Score pada kelas minoritas di bawah 0,50 untuk dataset dengan rasio ketidakseimbangan serupa. Hal ini semakin memperkuat posisi pendekatan hybrid SMOTE-Random Forest sebagai solusi yang robust dan andal untuk permasalahan klasifikasi biner pada data keuangan yang tidak seimbang.

G. Evaluasi Komprehensif dan Keterbatasan Penelitian

Secara keseluruhan, model hybrid SMOTE-Random Forest yang dikembangkan dalam penelitian ini menunjukkan performa yang sangat memuaskan dengan akurasi 95,00%, presisi 0,909, recall 0,952, dan F1-Score 0,930. Kombinasi metrik yang tinggi di seluruh dimensi evaluasi ini mengindikasikan bahwa model telah berhasil mempelajari pola-pola kompleks yang membedakan peminjam bermasalah dari peminjam yang berpotensi melunasi kewajibannya, tanpa mengalami overfitting yang signifikan.

Namun demikian, beberapa keterbatasan penelitian perlu diakui secara transparan. Pertama, jumlah sampel yang digunakan relatif kecil yaitu 200 data, sehingga generalisasi model ke populasi peminjam yang lebih luas perlu dilakukan dengan kehati-hatian. Kedua, sampel sintetis yang dihasilkan oleh SMOTE, meskipun secara matematis valid, tidak mewakili peminjam nyata yang sesungguhnya ada di lapangan, sehingga terdapat risiko model terlalu disesuaikan dengan karakteristik sintetis tersebut. Ketiga, penelitian ini tidak mencakup eksperimen tuning hyperparameter Random Forest secara sistematis, seperti optimasi jumlah pohon (n_estimators), kedalaman maksimum pohon (max_depth), dan nilai Mtry, yang berpotensi memberikan peningkatan performa lebih lanjut.

Untuk penelitian lanjutan, disarankan untuk mengeksplorasi teknik penyeimbangan data alternatif seperti ADASYN (Adaptive Synthetic Sampling) atau kombinasi oversampling dan undersampling (misalnya SMOTE-Tomek atau SMOTE-ENN), serta membandingkan performa model terhadap algoritma klasifikasi lain seperti Gradient Boosting (XGBoost, LightGBM) yang juga telah terbukti efektif pada domain deteksi anomali keuangan.

IV. KESIMPULAN

Penelitian ini telah berhasil mengembangkan dan memvalidasi model prediksi risiko gagal bayar menggunakan pendekatan hybrid yang mengintegrasikan *Synthetic Minority Over-sampling Technique* (SMOTE) dengan algoritma Random Forest (RF) pada dataset pinjaman Lending Club sebanyak 200 sampel. Seluruh tahapan pipa data (*data pipeline*), mulai dari pembersihan data (*data cleaning*), normalisasi berbasis *Min-Max Transformation*, pembagian data latih dan data uji dengan rasio 70:30, penerapan teknik oversampling SMOTE pada data pelatihan, proses pelatihan ratusan pohon keputusan dalam arsitektur ensemble Random Forest, hingga tahap evaluasi akhir, telah dieksekusi secara sistematis dan menghasilkan temuan yang signifikan secara metodologis dan praktis dalam domain manajemen risiko kredit teknologi finansial (*fintech lending*).

Temuan utama dan paling fundamental dari penelitian ini adalah terkonfirmasinya efektivitas intervensi SMOTE dalam menanggulangi permasalahan ketidakseimbangan kelas (*class imbalance*) ekstrem yang terjadi pada partisi data pelatihan. Dengan memanfaatkan parameter algoritma *k-Nearest Neighbors* (*k*=5) dan perhitungan jarak kedekatan *Euclidean Distance* sebagai mesin inti pembangkitan sampel sintetis, metode SMOTE terbukti sukses memperbanyak jumlah sampel kelas minoritas "Risiko Gagal Bayar" secara artifisial dari yang semula hanya berjumlah 4 sampel asli, melonjak secara drastis dan presisi menjadi 136 sampel melalui mekanisme interpolasi linear di dalam ruang fitur multidimensi.

Transformasi sebaran distribusi kelas data latih dari rasio timpang 136:4 menjadi rasio seimbang sempurna 136:136 ini secara mendasar berhasil merubah kondisi optimasi komputasi yang dihadapi oleh algoritma Random Forest. Distribusi yang seimbang tersebut memberikan sinyal pembelajaran yang adil kepada setiap pohon keputusan di dalam model ensemble, sehingga proses klasifikasi tidak lagi terdistorsi oleh dominasi mutlak nasabah kelas mayoritas.

Evaluasi performa akhir model yang diuji langsung pada 60 baris data pengujian murni menghasilkan ringkasan *Confusion Matrix* yang sangat memuaskan, yaitu dengan rincian capaian nilai *True Negative* (TN) = 37, *False Positive* (FP) = 2, *False Negative* (FN) = 1, dan *True Positive* (TP) = 20. Dari kombinasi sebaran angka matriks konfusi riil Google Colab tersebut, diturunkan empat metrik performa utama model hybrid secara holistik, yaitu tingkat Akurasi global sebesar 95,00%, nilai Presisi sebesar 0,909, tingkat nilai *Recall* atau sensitivitas sebesar 0,952, serta capaian parameter skor harmonik *F1-Score* sebesar 0,930. Perolehan nilai *Recall* yang menyentuh angka 95,2% menjadi pencapaian yang paling krusial dari perspektif bisnis manajemen risiko keuangan. Nilai sensitivitas yang sangat tinggi ini menandakan bahwa model cerdas yang diusulkan hanya meloloskan 1 kasus kesalahan prediksi *False Negative* dari total keseluruhan 21 kasus nasabah aktual yang mengalami gagal bayar di lapangan. Jika dikomparasikan dengan kondisi eksperimen baseline tanpa penyeimbangan SMOTE yang hanya mampu menghasilkan nilai *Recall* sebesar 11% untuk kelas gagal bayar, lompatan absolut sebesar 84,2 poin persentase ini secara kuantitatif menegaskan bahwa pendekatan hybrid yang dirancang merupakan solusi yang sangat menentukan dalam menjaring debitur berisiko tinggi.

Secara metodologis, penelitian ini membuktikan bahwa sinergi penggabungan antara algoritma SMOTE dan Random Forest bersifat komplementer serta saling menguatkan di dalam lingkungan pemrosesan data finansial. Metode SMOTE bertugas menyediakan prasyarat berupa kestabilan distribusi data latih yang merata, sementara mekanisme ensemble Random Forest yang dibangun melalui kombinasi penarikan sampel acak *bagging*, pembatasan subset fitur acak melalui parameter *Mtry*, dan pembelahan node keputusan berbasis fungsi kriteria *Gini Impurity*, mampu mengeksploitasi secara optimal kondisi data seimbang tersebut. Kriteria fungsi *Gini Impurity* secara teoritis dapat beroperasi dengan daya pisah (*discriminative power*) yang paling maksimum justru ketika distribusi sebaran kelas data pada setiap node keputusan berada pada kondisi yang merata.

Kehadiran sampel buatan dari SMOTE sukses mengatasi kendala struktural tersebut dengan mengisi ruang-ruang kosong pada ruang fitur kelas Gagal Bayar. Capaian nilai *F1-Score* sebesar 0,930 yang melampaui rata-rata laporan studi komparatif berpendekatan tunggal lainnya semakin memperkokoh posisi taktis model *Hybrid SMOTE-Random Forest* sebagai arsitektur pendukung keputusan (*decision support system*) yang *robust*, andal, dan siap diintegrasikan sebagai mesin penilaian kredit (*intelligent credit scoring*) otomatis pada platform fintech masa depan.

Meskipun hasil eksperimen komputasi ini menunjukkan performa yang luar biasa, beberapa keterbatasan penelitian tetap perlu diakui secara ilmiah. Ukuran dataset yang diteliti relatif terbatas yaitu sebanyak 200 sampel data, sehingga kemampuan generalisasi model terhadap populasi nasabah fintech yang jauh lebih heterogen dan berskala masif masih membutuhkan pengujian lebih lanjut. Selain itu, penelitian draf jurnal skripsi ini belum melibatkan eksperimen pencarian kombinasi parameter optimal (*hyperparameter tuning*) pada Random Forest secara sistematis, seperti penentuan variasi jumlah pohon (*n_estimators*) maupun kedalaman maksimal cabang (*max_depth*), yang secara teoritis berpotensi mendorong performa klasifikasi ke level yang jauh lebih tinggi.

Berdasarkan temuan dan keterbatasan tersebut, arah penelitian lanjutan disarankan untuk menguji performa keandalan model hybrid ini pada dataset Lending Club skala penuh dengan ratusan ribu baris data transaksi, mengeksplorasi penggunaan varian penyeimbangan data sintetis berbasis adaptif seperti ADASYN atau metode kombinasi *oversampling-undersampling* seperti SMOTE-Tomek dan SMOTE-ENN, serta melakukan tolak ukur (*benchmark*) komprehensif terhadap algoritma ensemble generasi terbaru seperti XGBoost dan LightGBM guna memetakan konfigurasi sistem mitigasi risiko gagal bayar yang paling optimal.

UCAPAN TERIMA KASIH

Penulis menyampaikan apresiasi dan terima kasih yang mendalam kepada Fakultas Ilmu Komputer Universitas Duta Bangsa Surakarta atas penyediaan fasilitas akademik, bimbingan, serta ekosistem riset yang mendukung kelancaran pelaksanaan penelitian ini. Ucapan terima kasih juga ditujukan secara tulus kepada Bapak Afu Ichsan Pradana, M.Kom. selaku Dosen Pembimbing I dan Bapak Eko Purwanto, S.Kom., M.Kom. selaku Dosen Pembimbing II atas curahan ilmu, saran konstruktif, kesabaran, serta arahan beralur ilmiah yang sangat berharga selama pengerjaan eksperimen hingga finalisasi draf artikel ilmiah ini. Terakhir, terima kasih kepada keluarga besar, rekan-rekan mahasiswa Teknik Informatika, serta seluruh pihak yang telah memberikan dukungan moral maupun material sehingga naskah publikasi ini dapat diselesaikan dengan baik.

REFERENSI

[1] A. Muhidin, Muhtajuddin Danny, and N. Surojudin, "Prediksi Kegagalan Perangkat Industri Menggunakan Random Forest dan SMOTE untuk Pemeliharaan Preventif," *Bull. Comput. Sci. Res.*, vol. 5, no. 5, pp. 1089–1094, 2025, doi: 10.47065/bulletincsr.v5i5.745.

[2] Maulana Ardiansyah and Supatman, "Klasterisasi Tagihan Pada Nasabah Pinjaman Online Menggunakan Metode K-Means Clustering," *J. RESTIKOM Ris. Tek. Inform. dan Komput.*, vol. 6, no. 2, pp. 286–294, 2024, doi: 10.52005/restikom.v6i2.318.

[3] B. N. Nuzululnisa and H. Hairani, "Analisis Kinerja Model Random Forest dengan Teknik Manhattan-SMOTE pada Deteksi Fraud Transaksi Kartu Kredit Imbalance," *Semin. Nas. Corisindo*, no. September 2025, pp. 65–71, 2025.

[4] R. A. Isaac and V. V. L. Rahul, "RE," 2025.

[5] R. Oversampling and R. Forest, "Implementation of Smote, Random Oversampling and Random Undersampling in Random Forest and Lasso Logistic Regression for Customer Churn Prediction 1," vol. 2, no. 1, pp. 1–14, 2025.

[6] Y. Jang, "Feature-based ensemble modeling for addressing diabetes data imbalance using the SMOTE, RUS, and random forest methods: a prediction study," *Ewha Med. J.*, vol. 48, no. 2, p. e32, 2025, doi: 10.12771/emj.2025.00353.

[7] R. Ridwan, H. H. Handayani, S. A. P. Lestari, and Y. Cahyana, "Evaluasi Kinerja Algoritma Random Forest Dan Gradient Boosting Untuk Klasifikasi Penyakit Jantung," *J. Komtika (Komputasi dan Inform.*, vol. 9, no. 1, pp. 112–124, 2025, doi: 10.31603/komtika.v9i1.13450.

[8] H. Putra and R. Rumini, "Comparative Study of Logistic Regression, Random Forest, and XGBoost for Bank Loan Approval Classification," *J. Appl. Informatics Comput.*, vol. 9, no. 5, pp. 2822–2835, 2025, doi: 10.30871/jaic.v9i5.10862.

[9] J. Gopar Sánchez, "A Simple Credit Rating Prediction Model for FinTech Companies Using SMOTE and MRMR Techniques," *Anáhuac J.*, vol. 24, no. 2, pp. 1–29, 2024, doi: 10.36105/theanahuacjour.2024v24n2.2516.

[10] D. Sugianto, "Classifying Vehicle Categories Based on Technical Specifications Using Random Forest and SMOTE for Data Augmentation," *Int. J. Appl. Inf. Manag.*, vol. 5, no. 4, pp. 179–191, 2025, doi: 10.47738/ijaim.v5i4.113.

[11] N. Hazizah, Sharipuddin, and A. Feranika, "Implementasi Algoritma Random Forest Dalam Klasifikasi Risiko Gagal Bayar Kartu Kredit Pada Nasabah Bank," *J. Manaj. Teknol. Dan*

Sist. Inf., vol. 5, no. 1, pp. 1050–1059, 2025, doi: 10.33998/jms.2025.5.1.2384.