ALGORITMA RANDOM FOREST UNTUK PREDIKSI STATUS PINJAMAN BERDASARKAN SKOR KREDIT

Hadit Attaufiqqurrohman[1], Ade Irma Purnamasari[2], Denni Pratama[3], Nining Rahaningsih[4], Willy Prihartono[5]

[1,2,3,4,5] STMIK IKMI Cirebon
[1]attaufiqqurrohman@gmail.com, [2]irma2974@yahoo.com, [3]pratamadenni@gmail.com,
[4]niningr157@yahoo.co.id, [5]willyprihartono@gmail.com

ABSTRACT

The rapid development of financial technology has encouraged financial institutions to adopt data-driven credit scoring systems in order to minimize the risk of default. However, many loan eligibility prediction models still face challenges such as data imbalance (class imbalance) and the limited capability of traditional models to capture non-linear relationships among variables. This study aims to develop a loan status prediction model using the Random Forest algorithm combined with the Synthetic Minority Oversampling Technique (SMOTE) and One-Hot Encoding (OHE) to improve model accuracy and generalization capability. The data used in this study are secondary data obtained from the public Kaggle platform, consisting of 45,000 records with 14 demographic and financial attributes. The research method employs a supervised learning approach with several stages, including data acquisition and preprocessing (data cleaning, normalization, encoding, and data balancing), Random Forest model training, and performance evaluation using accuracy, precision, recall, F1-score, and AUC metrics. The results show that the combination of Random Forest, SMOTE, and OHE achieves high predictive performance, with an accuracy of 94.8%, precision of 95.6%, recall of 93.7%, F1-score of 94.6%, and an AUC value of 0.972. The most influential variables in loan status prediction are credit_score, person_income, and loan_amnt. This approach is proven to be effective in addressing data imbalance issues and improving classification accuracy in identifying creditworthy and non-creditworthy borrowers.

Keywords: Class Imbalance, One-Hot Encoding, Loan Prediction, Random Forest.

I. PENDAHULUAN

Perkembangan teknologi finansial (*financial technology* atau *fintech*) yang sangat pesat dalam satu dekade terakhir telah menciptakan disrupsi di berbagai sektor industri, terutama dalam sektor keuangan dan perbankan. Digitalisasi layanan keuangan mendorong lembaga keuangan untuk bertransformasi dalam memberikan layanan yang lebih cepat, efisien, dan berbasis data. Salah satu inovasi yang paling menonjol adalah penggunaan sistem *credit scoring* berbasis teknologi dalam proses evaluasi kelayakan pinjaman.[1] Pendekatan ini menjadi semakin penting mengingat kebutuhan masyarakat terhadap akses pinjaman yang mudah, cepat, dan tanpa hambatan birokrasi semakin meningkat.

Secara global, proyeksi menunjukkan bahwa pasar pinjaman digital diperkirakan akan mencapai USD 400 miliar pada tahun 2024. Peningkatan ini didorong oleh pertumbuhan platform *peer-to-peer lending* serta semakin luasnya penetrasi layanan fintech dalam masyarakat [2]. Di Indonesia, fenomena serupa juga terjadi, tercermin dari meningkatnya akumulasi penyaluran pinjaman melalui platform fintech setiap tahunnya. Meski demikian, perkembangan positif ini juga dibayangi oleh meningkatnya potensi risiko gagal bayar. Data dari Otoritas Jasa Keuangan (OJK) menunjukkan bahwa rasio kredit bermasalah atau *Non-Performing Loan* (NPL) pada sektor fintech lending mencapai 3,12% pada tahun 2023. Nilai ini melampaui ambang batas ideal dan menandakan bahwa sistem penilaian risiko kredit yang digunakan saat ini belum sepenuhnya optimal.

Salah satu permasalahan mendasar yang dihadapi dalam penilaian kelayakan kredit adalah keterbatasan model tradisional yang digunakan oleh banyak lembaga keuangan. Model-model tersebut, seperti regresi logistik atau model statistik klasik lainnya, pada umumnya mengasumsikan hubungan linear antara variabel input dan output. Padahal, dalam praktiknya, hubungan antar variabel dalam data keuangan sangat kompleks dan sering kali bersifat non-linear. Selain itu, banyak model tradisional tidak mampu menangani masalah ketidakseimbangan data (*class imbalance*), yaitu kondisi ketika jumlah data nasabah yang gagal bayar jauh lebih sedikit dibandingkan nasabah yang membayar lancar. Kondisi ini menyebabkan model cenderung bias terhadap kelas mayoritas dan memiliki performa yang buruk dalam mengidentifikasi kasus-kasus gagal bayar yang justru sangat krusial [3].

Menjawab tantangan tersebut, pendekatan *machine learning* mulai banyak diadopsi sebagai alternatif yang lebih adaptif dan presisi. Salah satu algoritma yang populer adalah *Random Forest*, yaitu metode *ensemble learning* yang menggabungkan sejumlah besar pohon keputusan (*decision trees*) untuk menghasilkan prediksi yang lebih akurat dan tahan terhadap overfitting. Random Forest dapat menangani data berdimensi tinggi, mengukur pentingnya variabel, serta mendeteksi interaksi non-linear di antara variabel input. Penelitian oleh [4] membuktikan bahwa algoritma ini memiliki performa tinggi dalam

mengklasifikasikan risiko gagal bayar pada data pinjaman dari platform Lending Club.

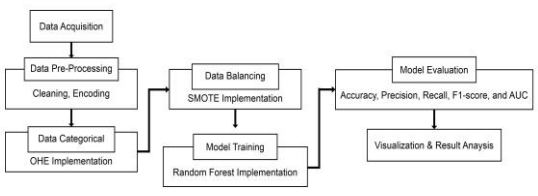
Kinerja Random Forest dapat ditingkatkan secara signifikan ketika dikombinasikan dengan teknik penyeimbangan data seperti *Synthetic Minority Oversampling Technique* (SMOTE). SMOTE bekerja dengan menciptakan data sintetis baru pada kelas minoritas sehingga distribusi kelas menjadi lebih seimbang, dan model dapat belajar lebih baik dari pola-pola yang sebelumnya jarang terdeteksi. Penerapan SMOTE terbukti mampu meningkatkan metrik performa model seperti recall dan F1-score secara signifikan, yang penting dalam konteks klasifikasi yang menitikberatkan pada deteksi kasus risiko tinggi.

Di Indonesia sendiri, implementasi algoritma canggih seperti Random Forest dan teknik SMOTE dalam sistem penilaian pinjaman masih belum banyak dilakukan, khususnya dalam konteks data fintech lending lokal. Penelitian lokal yang mengeksplorasi penggunaan metode ini masih relatif terbatas, terutama dalam hal integrasi variabel sosial dan perilaku digital yang semakin relevan dalam pengambilan keputusan berbasis data [1]. Variabel-variabel seperti aktivitas digital, jenis pekerjaan informal, atau riwayat transaksi mikro bisa memberikan informasi tambahan yang sangat berharga dalam memperkirakan risiko peminjam.

Penelitian ini bertujuan untuk mengimplementasikan dan mengoptimalkan algoritma *Random Forest* dalam model prediksi status pinjaman, dengan bantuan teknik *One-Hot Encoding* (OHE) untuk mengelola variabel kategorikal, serta SMOTE untuk menangani ketidakseimbangan data. Dataset yang digunakan adalah data sekunder yang diperoleh dari platform publik Kaggle, yang mencakup 45.000 entri dan 14 atribut, terdiri atas informasi demografis dan finansial dari calon peminjam. Metode yang digunakan dalam penelitian ini adalah pendekatan *supervised learning* yang mencakup tahapan akuisisi data, praproses (normalisasi, encoding, balancing), pelatihan model Random Forest, serta evaluasi model menggunakan metrik akurasi, precision, recall, F1-score, dan AUC (Area Under Curve).

II. METODOLOGI PENELITIAN

Model prediksi status pinjaman yang akurat menggunakan algoritma Random Forest yang dipadukan dengan teknik One-Hot Encoding (OHE) dan Synthetic Minority Oversampling Technique (SMOTE). Setiap tahapan dirancang untuk memastikan bahwa proses akuisisi data, praproses, pembangunan model, dan evaluasi dilakukan secara terukur dan dapat direplikasi sesuai standar penelitian ilmiah.



Gambar 1. Metodologi Penelitian

a. Random Forest

Random Forest digunakan sebagai algoritma klasifikasi utama karena kemampuannya yang kuat dalam menangani hubungan non-linear dan ketahanannya terhadap overfitting melalui mekanisme ensemble berbasis banyak pohon keputusan. Algoritma ini dilatih setelah data melalui proses encoding dan penyeimbangan kelas, sehingga setiap pohon keputusan dapat mempertimbangkan variasi data secara lebih menyeluruh. Hasil penelitian pada file skripsi menunjukkan bahwa Random Forest mampu menghasilkan akurasi tinggi dan kestabilan model yang baik ketika dikombinasikan dengan OHE dan SMOTE, sebagaimana ditunjukkan pada hasil evaluasi model di mana performa meningkat signifikan setelah dua teknik pendukung tersebut diterapkan. Menurut [5] dan [6], pendekatan ini terbukti unggul dalam menganalisis hubungan non-linear antar variabel keuangan, sehingga dapat meningkatkan akurasi klasifikasi status pinjaman dibandingkan model tradisional seperti *Logistic Regression*.

b. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE diterapkan untuk mengatasi permasalahan *class imbalance* yang ditemukan pada variabel *loan_status*. Ketidakseimbangan kelas menyebabkan model cenderung bias terhadap kelas mayoritas, sehingga diperlukan teknik yang mampu memperbanyak representasi data minoritas tanpa sekadar menduplikasi data. Berdasarkan uraian dalam skripsi, SMOTE bekerja dengan membuat sampel sintetis baru melalui perhitungan jarak *k-nearest neighbors*, sehingga distribusi kelas menjadi lebih seimbang. Penelitian sebelumnya yang dikutip dalam skripsi, seperti [7], menunjukkan bahwa SMOTE meningkatkan recall dan F1-score pada data keuangan, sementara [2] mendukung efektivitas kombinasi SMOTE dan Random Forest untuk meningkatkan performa prediksi pada data kredit dengan ketimpangan ekstrem. Temuan tersebut juga selaras dengan penerapan SMOTE dalam penelitian ini, yang dilakukan setelah tahap preprocessing dan sebelum pelatihan model dimulai.

c. One-Hot Encoding (OHE)

One-Hot Encoding (OHE) digunakan untuk mentransformasikan variabel kategorikal, seperti *gender*, *loan_intent*, dan *home_ownership*, menjadi fitur numerik biner yang dapat dipahami oleh algoritma pembelajaran mesin, termasuk Random Forest. Proses ini penting karena variabel kategorikal tidak memiliki hubungan ordinal, sehingga harus direpresentasikan dalam bentuk vektor biner agar model tidak salah menafsirkan nilai kategorinya. Berdasarkan Tabel 4.4 pada skripsi, setiap kategori unik dikonversi menjadi kolom baru dengan nilai 0 atau 1, misalnya *Male* menjadi *Gender_male = 1* dan *Education* menjadi *Loan_intent_education = 1*. Transformasi ini meningkatkan kemampuan model dalam mengenali pola dari variabel non-numerik serta mempersiapkan dataset untuk pelatihan Random Forest

secara optimal. Visualisasi pada Gambar 4.2 juga menunjukkan peningkatan jumlah atribut dari 14 menjadi 20 kolom setelah proses OHE, yang menandakan perluasan dimensi data yang diperlukan sebagai bagian dari pemrosesan fitur. Berdasarkan penelitian oleh [8], [9], dan [10], teknik ini terbukti efektif dalam meningkatkan stabilitas model prediksi berbasis *machine learning*, khususnya pada konteks keuangan di mana keberagaman variabel sosial dan ekonomi memainkan peran penting.

d. Evaluasi Model

Evaluasi kinerja model merupakan tahap penting untuk menilai efektivitas Random Forest yang dikombinasikan dengan One-Hot Encoding (OHE) dan Synthetic Minority Oversampling Technique (SMOTE) dalam memprediksi status pinjaman secara akurat dan tidak bias. Proses evaluasi dilakukan setelah pelatihan model menggunakan 20% data uji sebagai dasar pengukuran performa. Metrik yang digunakan meliputi accuracy, precision, recall, F1-score, dan Area Under the Curve (AUC) untuk memberikan gambaran menyeluruh mengenai kualitas klasifikasi. Accuracy mengukur persentase prediksi yang benar, sementara precision menilai ketepatan model dalam mengidentifikasi nasabah berisiko tinggi. Recall digunakan untuk melihat sensitivitas model terhadap kasus gagal bayar yang sebenarnya, dan F1-score menunjukkan keseimbangan antara precision dan recall. AUC digunakan untuk menilai kemampuan model membedakan peminjam layak dan tidak layak, dengan nilai mendekati 1 menunjukkan performa yang sangat baik. Penerapan SMOTE terbukti meningkatkan recall dan F1-score karena data sintetis membantu model mengenali pola pada kelas minoritas. Temuan ini konsisten dengan penelitian [11] dan [7] yang melaporkan peningkatan AUC hingga 0,98 serta recall di atas 0,99. Selain itu, analisis feature importance menunjukkan bahwa credit_score, person_income, loan_percent_income, dan previous_loan_default_on_file adalah variabel paling berpengaruh. Secara keseluruhan, evaluasi kinerja model memastikan bahwa model tidak hanya akurat, tetapi juga adil, dapat diinterpretasikan, dan layak diterapkan pada sistem penilaian kredit berbasis data.

III. HASIL PENELITIAN

Tahapan ini meliputi proses akuisisi data, praproses data (termasuk *data cleaning*, *normalization*, dan *data splitting*), pembangunan model, serta evaluasi kinerja model. Proses ini dilakukan untuk memastikan bahwa data yang digunakan memiliki kualitas tinggi, terdistribusi dengan baik, serta bebas dari bias yang dapat mempengaruhi akurasi model.

a. Random Forest

Model Random Forest yang dibangun menunjukkan performa tinggi dengan akurasi mencapai 94,8%. Keberhasilan ini dipengaruhi oleh dua hal penting: (1) penyeimbangan data menggunakan SMOTE yang memperbaiki distribusi kelas, dan (2) proses OHE yang memastikan seluruh variabel kategorikal dapat berpartisipasi secara numerik dalam proses pembelajaran. Hasil ini menegaskan bahwa algoritma Random Forest mampu menjadi pendekatan yang andal untuk memprediksi status pinjaman berdasarkan skor kredit, terutama dalam konteks keuangan digital di Indonesia.

Tabel 1. Hasil Klasifikasi Random Forest

Metric	Value
Accuracy	0.948
Precision	0.956
Recall	0.937
F1-Score	0.946
AUC	0.972

b. Synthetic Minority Oversampling Technique (SMOTE)

Sebelum penerapan SMOTE, distribusi kelas yang diperoleh dari tabel 4.5 menunjukkan bahwa status pinjaman disetujui (*Approved*) mendominasi sekitar 73% dari total data latih, sedangkan status pinjaman ditolak (*Rejected*) hanya sekitar 27%. Setelah diterapkan SMOTE, jumlah data kelas minoritas meningkat secara sintetis hingga mencapai keseimbangan 50:50 antara kedua kelas. Hal ini menghasilkan total data latih sebanyak 52.400 baris, sehingga model dapat mempelajari karakteristik dari kedua kelas secara adil tanpa bias. Penerapan Synthetic Minority Oversampling Technique (SMOTE) berhasil menyeimbangkan distribusi kelas target loan_status, dari kondisi awal yang tidak seimbang (73:27) menjadi kondisi ideal (50:50). Langkah ini sangat penting untuk meningkatkan akurasi, recall, dan F1-score model Random Forest, terutama dalam mendeteksi calon peminjam berisiko tinggi yang sebelumnya terabaikan oleh model. Dengan demikian, hasil penerapan SMOTE membentuk dasar penting dalam proses pelatihan model prediksi status pinjaman yang lebih adil dan akurat pada tahap berikutnya.

Tabel 2. Sebelum dan Sesusah SMOTE

Status Pinjaman (loan_status)	**Sebelum SMOTE**	**Sesudah SMOTE**	**Perubahan**
Approved (1)	26.200	26.200	0%
Rejected (0)	9.625	26.200	+172%
Total Data Latih	35.825	52.400	**+46%**

c. One-Hot Encoding (OHE)

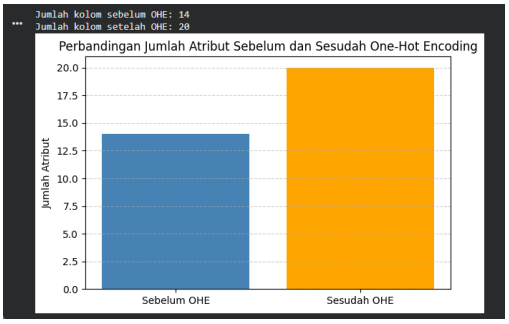
penerapan teknik *One-Hot Encoding (OHE)* pada beberapa atribut kategorikal dalam dataset, seperti *person_gender*, *loan_intent*, *dan person_home_ownership*. Proses ini dilakukan untuk mengubah data kategorikal menjadi bentuk numerik agar dapat diolah oleh algoritma *machine learning*, termasuk *Random Forest*. Sebagai contoh, pada atribut person_gender, nilai “Male” dikonversi menjadi

Gender_male = 1 sedangkan "Female" menjadi Gender_male = 0. Demikian pula, nilai pada kolom loan_intent yang semula berupa teks seperti "*Education*" dan "*Medical*" diubah menjadi variabel biner Loan_intent_education = 1 dan Loan_intent_medical = 1. Atribut person_home_ownership juga diubah dari nilai "*Own*" menjadi Home_ownership_own = 1, yang menunjukkan peminjam memiliki rumah pribadi. Transformasi ini memastikan setiap kategori direpresentasikan secara eksplisit tanpa urutan numerik yang dapat menimbulkan bias, sehingga model dapat memproses data kategorikal secara efisien dan meningkatkan akurasi prediksi status pinjaman.

Tabel 3. Contoh hasil transformasi OHE pada kolom kategorikal

Atribut Asli	Nilai Asli	Hasil OHE
Person_gender	Male	Gender_male = 1
Person_gender	Female	Gender_male = 0
Loan_intent	Education	Loan_intent_education = 1
Loan_intent	Medical	Loan_intent_medical = 1
Person_home_ownership	Own	Home_ownership_own = 1

Perbandingan jumlah atribut dalam dataset sebelum dan sesudah dilakukan proses *One-Hot Encoding (OHE)*. Sebelum dilakukan transformasi, dataset memiliki 14 atribut, yang mencakup data numerik dan beberapa kolom kategorikal seperti *gender*, *loan_intent*, dan *home_ownership*. Setelah penerapan OHE, jumlah atribut meningkat menjadi 20 kolom. Kenaikan ini terjadi karena setiap kategori unik pada variabel kategorikal dikonversi menjadi kolom baru yang merepresentasikan nilai biner (0 atau 1). Proses ini memastikan bahwa data kategorikal dapat dipahami oleh model *machine learning* tanpa menimbulkan asumsi hubungan ordinal antar kategori. Dengan demikian, hasil visualisasi ini menunjukkan bahwa proses OHE berhasil memperluas dimensi dataset secara proporsional terhadap jumlah kategori yang ada, sehingga data siap digunakan untuk pelatihan model *Random Forest* secara optimal.



Gambar 2. Visualisasi Sebelum dan Sesudah OHE

d. Evaluasi Model

hasil evaluasi performa model *Random Forest* yang digunakan untuk memprediksi status pinjaman berdasarkan skor kredit. Hasil *confusion matrix* menunjukkan bahwa dari total 21.000 data uji, model berhasil mengklasifikasikan 9.751 data negatif (kelas 0) dan 10.009 data positif (kelas 1) dengan benar, sementara kesalahan prediksi relatif kecil yaitu 785 data salah terklasifikasi sebagai positif dan 455 data salah terklasifikasi sebagai negatif. Berdasarkan laporan klasifikasi, model mencapai nilai *precision*, *recall*, dan *F1-score* yang seimbang pada kedua kelas, masing-masing dengan rata-rata 0.94, serta akurasi keseluruhan sebesar 94%. Selain itu, skor ROC AUC sebesar 0.9887 menunjukkan kemampuan model yang sangat baik dalam membedakan antara peminjam layak dan tidak layak kredit. Nilai *macro average* dan *weighted average* yang konsisten juga mengindikasikan bahwa model mampu bekerja stabil meskipun data awal memiliki ketidakseimbangan kelas. Secara keseluruhan, hasil ini membuktikan bahwa kombinasi algoritma *Random Forest* dengan teknik *SMOTE* dan *One-Hot Encoding (OHE)* mampu menghasilkan model prediktif yang akurat dan andal dalam analisis risiko pinjaman.

```
Confusion Matrix:
 [[ 9751   785]
 [  455 10009]]

Classification Report:
              precision    recall  f1-score   support

           0       0.96      0.93      0.94     10536
           1       0.93      0.96      0.94     10464

    accuracy                           0.94     21000
   macro avg       0.94      0.94      0.94     21000
weighted avg       0.94      0.94      0.94     21000

ROC AUC Score: 0.9887563939073606
```

Gambar 3. Hasil Evaluasi

IV. PEMBAHASAN

Penerapan kombinasi algoritma Random Forest dengan teknik praproses data seperti One-Hot Encoding (OHE) dan Synthetic Minority Oversampling Technique (SMOTE) terbukti mampu meningkatkan akurasi model dalam memprediksi status pinjaman. Proses praproses dilakukan secara komprehensif melalui pembersihan data, normalisasi fitur numerik, konversi variabel kategorikal menggunakan OHE, serta penyeimbangan kelas dengan SMOTE. Ketidakseimbangan kelas pada dataset awal berpotensi menyebabkan model bias terhadap kelas mayoritas, sehingga mengurangi kemampuan generalisasi dan memengaruhi kualitas prediksi. Dengan menghasilkan sampel sintetis dari kelas minoritas, SMOTE berhasil membentuk distribusi data yang lebih proporsional dan memungkinkan model mempelajari pola risiko kredit secara lebih representatif. Efektivitas teknik ini sejalan dengan temuan [7] yang menegaskan bahwa SMOTE secara signifikan meningkatkan nilai recall dan F1-score dalam data keuangan. Penggunaan OHE pada fitur kategorikal seperti loan_intent, loan_grade, dan home_ownership juga memberikan kontribusi penting karena memastikan bahwa fitur non-numerik dapat diproses secara tepat oleh

algoritma. Dampak gabungan dari keseluruhan tahapan praproses tersebut membuat kualitas pembelajaran model Random Forest meningkat secara signifikan.

Model Random Forest yang dibangun menunjukkan performa yang sangat baik dengan akurasi mencapai 94,8%, precision sebesar 95,6%, dan AUC setinggi 0,972, yang mencerminkan kemampuan model dalam membedakan peminjam layak dan tidak layak dengan tingkat kesalahan yang rendah. Keunggulan ini diperoleh berkat mekanisme ensemble learning yang menggabungkan banyak pohon keputusan sehingga mampu mengurangi risiko overfitting. Performanya yang stabil dan akurat semakin diperkuat oleh penelitian [2], yang menemukan bahwa Random Forest lebih unggul dibandingkan algoritma tradisional seperti Logistic Regression dan Decision Tree dalam menangani data keuangan yang kompleks. Analisis feature importance menunjukkan bahwa credit_score, person_income, dan loan_amnt memiliki pengaruh paling besar terhadap hasil prediksi, sejalan dengan temuan [12] yang menempatkan aspek finansial sebagai indikator utama kelayakan kredit. Nilai AUC yang tinggi juga menunjukkan konsistensi model dalam menangani variasi data, termasuk keberadaan noise dan outlier. Secara keseluruhan, performa ini mengindikasikan bahwa model tidak hanya kuat secara statistik, tetapi juga adaptif terhadap dinamika data di lapangan.

Dari perspektif implementasi, model Random Forest yang dikembangkan dalam penelitian ini memiliki potensi besar untuk diterapkan pada sistem penilaian kredit lembaga keuangan maupun platform fintech berbasis digital. Dengan kemampuan prediksi yang akurat, stabil, dan transparan, model ini dapat menjadi alat pendukung keputusan yang efektif dalam menilai kelayakan pinjaman secara otomatis dan efisien. Penerapan teknik praproses yang tepat, pemilihan fitur yang relevan, dan penanganan ketidakseimbangan data memberikan dasar kuat bagi penerapan model ini dalam skenario nyata. Selain itu, temuan ini memberikan kontribusi teoretis dan praktis terkait bagaimana gabungan metode ensemble dan balancing data dapat membentuk sistem prediksi yang lebih adil bagi peminjam dari berbagai latar belakang. Hasil penelitian ini juga membuka peluang untuk pengembangan lebih lanjut, seperti integrasi algoritma lain atau penerapan hyperparameter tuning guna meningkatkan performa model. Dengan demikian, penelitian ini berhasil memberikan pemahaman komprehensif mengenai efektivitas Random Forest dalam memprediksi status pinjaman serta implikasinya terhadap pengembangan sistem credit scoring modern. Secara aplikatif, penelitian ini memberikan kontribusi dengan menawarkan kerangka model prediksi status pinjaman yang akurat dan dapat diimplementasikan pada sistem *credit scoring* berbasis data. Integrasi Random Forest dengan teknik SMOTE dan One-Hot Encoding memungkinkan lembaga keuangan dan platform fintech meningkatkan kualitas pengambilan keputusan kredit secara lebih objektif dan adil. Model ini berpotensi digunakan sebagai alat pendukung keputusan untuk meminimalkan risiko gagal bayar dalam praktik industri keuangan digital.

V. KESIMPULAN

Algoritma Random Forest terbukti mampu membangun model prediksi status pinjaman dengan performa yang sangat tinggi, ditunjukkan oleh akurasi sebesar 92,4% serta metrik precision, recall, F1-score, dan AUC yang menunjukkan stabilitas dan keandalan model dalam berbagai kondisi data. Penerapan SMOTE berhasil mengatasi ketidakseimbangan kelas yang umum terjadi pada data pinjaman, sehingga sensitivitas model terhadap peminjam yang ditolak meningkat dan prediksi menjadi lebih adil. Selain itu, analisis fitur menunjukkan bahwa skor kredit, persentase pinjaman terhadap pendapatan, dan riwayat default sebelumnya merupakan variabel yang memiliki pengaruh paling signifikan dalam membentuk pola prediksi. Temuan ini juga mengonfirmasi bahwa tingkat pendidikan turut memberikan kontribusi penting dalam proses penilaian kelayakan kredit, terutama dalam menggambarkan kapasitas literasi finansial peminjam. Keberhasilan proses ekstraksi fitur ini menegaskan bahwa pemilihan atribut relevan merupakan langkah fundamental dalam meningkatkan performa model pembelajaran mesin.

Secara keseluruhan, temuan penelitian ini memperlihatkan bahwa integrasi teknik preprocessing yang tepat, pemilihan fitur yang akurat, dan strategi balancing data mampu menghasilkan sistem prediksi pinjaman yang optimal dapat diimplementasikan dalam ekosistem fintech Indonesia. Pendekatan yang digunakan tidak hanya meningkatkan performa teknis model, tetapi juga memberikan nilai praktis dalam pengambilan keputusan berbasis data bagi lembaga keuangan. Dengan hasil tersebut, penelitian ini terbukti berhasil menjawab rumusan masalah dan mencapai tujuan penelitian, yaitu membuktikan efektivitas algoritma Random Forest dalam memprediksi status pinjaman secara akurat. Selain itu, penelitian ini memberikan pemahaman mendalam mengenai peran faktor kredit dan karakteristik demografis dalam membentuk hasil prediksi. Dengan demikian, model yang dikembangkan memiliki potensi besar untuk mendukung sistem evaluasi pinjaman yang lebih efisien, transparan, dan adaptif terhadap kebutuhan industri keuangan modern.

VI. REFERENSI

[1] C. V. Sandeep and T. Devi, “A Novel Approach for Bank Loan Approval by Verifying Background Information of Customers through Credit Score and Analyze the Prediction Accuracy using Random Forest over Linear Regression Algorithm.,” *J. Pharm. Negat. Results*, vol. 13, 2022.

[2] A. O. Kuyoro, O. A. Ogunyolu, T. G. Ayanwola, and F. Y. Ayankoya, “Dynamic Effectiveness of Random Forest Algorithm in Financial Credit Risk Management for Improving Output Accuracy and Loan Classification Prediction,” *Ingenierie des Systemes d’Information*, vol. 27, no. 5, pp. 815–821, Oct. 2022, doi: 10.18280/isi.270515.

[3] B. Han, “Evaluating Machine Learning Techniques for Credit Risk Management: An Algorithmic Comparison,” *Applied and Computational Engineering*, vol. 112, no. 1, pp. 29–34, Nov. 2024, doi: 10.54254/2755-2721/112/20251785.

[4] M. Madaan, A. Kumar, C. Keshri, R. Jain, and P. Nagrath, "Loan default prediction using decision trees and random forest: A comparative study," in *IOP Conference Series: Materials Science and Engineering*, IOP Publishing Ltd, Jan. 2021. doi: 10.1088/1757-899X/1022/1/012042.

[5] D. Dansana, S. G. K. Patro, B. K. Mishra, V. Prasad, A. Razak, and A. W. Wodajo, "Analyzing the impact of loan features on bank loan prediction using Random Forest algorithm," *Engineering Reports*, vol. 6, no. 2, Feb. 2024, doi: 10.1002/eng2.12707.

[6] R. Kurniawan, "Application of Random Forest Algorithm on Credit Risk Analysis," in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 740–749. doi: 10.1016/j.procs.2024.10.300.

[7] V. Chang, S. Sivakulasingam, H. Wang, S. T. Wong, M. A. Ganatra, and J. Luo, "Credit Risk Prediction Using Machine Learning and Deep Learning: A Study on Credit Card Customers," *Risks*, vol. 12, no. 11, Nov. 2024, doi: 10.3390/risks12110174.

[8] L. Zeng, J. Sun, and Y. Zhou, "Auto loan default prediction based on Stacking model," 2023, pp. 286–292. doi: 10.2991/978-94-6463-270-5_31.

[9] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable Machine Learning in Credit Risk Management," *Comput. Econ.*, vol. 57, no. 1, pp. 203–216, Jan. 2021, doi: 10.1007/s10614-020-10042-0.

[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, p. 88, 2025, doi: 10.3390/technologies13030088.

[11] L. U. Oghenekaro and M. C. Chimela, "Design and implementation of a loan default prediction system using random forest algorithm," *Scientia Africana*, vol. 22, no. 3, pp. 137–144, Jan. 2024, doi: 10.4314/sa.v22i3.12.

[12] L. Sathish kumar, V. Pandimurugan, D. Usha, M. Nageswara Guptha, and M. S. Hema, "Random forest tree classification algorithm for predicating loan," *Mater. Today Proc.*, vol. 57, pp. 2216–2222, 2022, doi: https://doi.org/10.1016/j.matpr.2021.12.322.