Prediksi Kelayakan Pemberian Kredit Menggunakan Metode Random Forest

Noval Firmansah, Uce Indahyanti dan Ade Eviyanti

Universitas Muhammadiyah Sidoharjo
Jl. Mojopahit No.666 B, Sidowayah, Celep, Kec. Sidoarjo, Kabupaten Sidoarjo, Jawa Timur 61215,
E-mail: novalfirmansah2811@gmail.com, uceindahyanti@umsida.ac.id, adeeviyanti@umsida.ac.id

Abstrak

Penelitian ini didasarkan pada penggunaan data dari Kaggle, yang terdiri dari 32,581 baris dan 12 kolom, untuk mengembangkan model prediksi kelayakan kredit dengan menggunakan metode *Random Forest*. Tujuan penelitian adalah mengidentifikasi faktor-faktor yang memengaruhi kelayakan kredit dan mengembangkan model yang akurat dalam memprediksi apakah seorang peminjam layak atau tidak menerima kredit. Metode *Random Forest*, sebuah teknik *ensemble learning* yang menggabungkan sejumlah besar pohon keputusan, digunakan sebagai pendekatan utama dalam pembangunan model prediksi. Penggunaan *Random Forest* membantu mengurangi *overfitting*, meningkatkan akurasi prediksi, serta memberikan kemampuan untuk mengukur pentingnya masing-masing fitur dalam kelayakan kredit. Selain itu, penelitian ini melibatkan serangkaian langkah-langkah pra-pemrosesan data, termasuk imputasi *missing value* dan penanganan *outlier*, serta pembagian dataset menjadi data latih dan data uji. Hasil penelitian menunjukkan bahwa model mencapai akurasi sebesar 93,28% dengan parameter terbaik, yaitu 'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, dan 'n_estimators': 100. Penelitian ini memberikan kontribusi penting dalam pemahaman tentang kelayakan kredit dan pengembangan model prediksi yang dapat digunakan oleh lembaga keuangan untuk pengambilan keputusan kredit yang lebih tepat.

Kata kunci: Kelayakan Kredit; *Outlier*; Model Prediksi; *Random Forest.*

Pendahuluan

Saat ini, hampir semua sektor memerlukan kemudahan dalam mengelola informasi yang dimiliki, termasuk dalam bidang pengkreditan. Pembelian dengan sistem kredit sudah menjadi hal umum dalam masyarakat. Proses pengajuan kredit menjadi lebih mudah dan terjangkau, sehingga menyebabkan peningkatan jumlah konsumen yang mengajukan kredit [1]. Situasi seperti ini menyebabkan masyarakat kadang-kadang tidak lagi mempertimbangkan kemampuan keuangan mereka. Akibatnya, pihak pembiayaan menghadapi dampak signifikan, terutama ketika semakin banyak konsumen yang tidak mampu membayar cicilan atau angsuran, yang disebut sebagai "Kredit Macet".

Dengan banyaknya jumlah konsumen yang mengajukan kredit akan menimbulkan penumpukan data pengajuan kredit. Hal ini berdampak besar pada bagian *Credit Analyst* selaku pihak yang menentukan kelayakan pemberian kredit sepeda motor dan memegang data para customer [2]. Penggunaan data mining untuk klasifikasi kelayakan pemberian kredit akan sangat membantu pihak perusahaan dimana klasifikasi itu sendiri untuk membedakan kelas data yang layak dan tidak layak untuk melakukan kredit [3]. Dengan diklasifikasikannya data akan mempermudah dan mempercepat kinerja dari bagian *Credit Analyst.*

Penelitian sejenis pernah dilakukan oleh Syafi'i, Odi Nurdiawan, Gifthera Dwilestari pada tahun 2022 dengan judul "Penerapan *Machine Learning* Untuk Menentukan Kelayakan Kredit Menggunakan Metode *Support Vektor Machine*" Penelitian ini menggunakan algoritma *Support Vector Machine* untuk menilai kelayakan kredit. Hasil dari *Performance Vector* menunjukkan bahwa prediksi kredit lancar sebanyak 130 kasus dengan benar, prediksi kredit macet sebanyak 72 kasus dengan benar, prediksi kredit lancar yang sebenarnya macet sebanyak 41 kasus, dan prediksi kredit macet yang sebenarnya macet sebanyak 332 kasus.

Berdasarkan data tersebut, tingkat akurasi dari *performance vector* algoritma *Support Vector Machine* adalah sebesar 80.34%. Hal ini menggambarkan kemampuan algoritma untuk melakukan prediksi dengan tepat dalam kasus kelayakan kredit [4].

DOI : http://dx.doi.org/10.32409/jikstik.22.4.3515

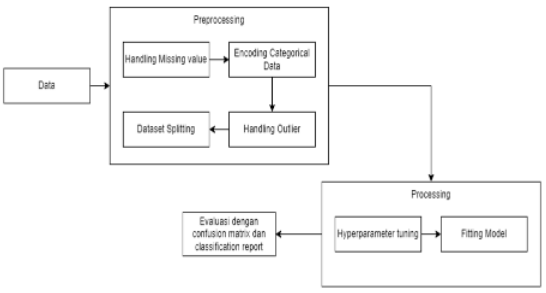
Dalam perbandingan antara penelitian ini dengan penelitian sebelumnya, terdapat kesenjangan yang signifikan yang perlu dipertimbangkan. Penelitian sebelumnya telah menggunakan algoritma *Support Vector Machine* (SVM) untuk menilai kelayakan kredit, sementara peneliti telah mengadopsi pendekatan yang berbeda dengan menerapkan metode *Random Forest* yang dikombinasikan dengan *Grid Search* CV untuk mengoptimalkan *hyperparameter*. *Grid Search* CV merupakan bagian dari modul scikit-learn yang melakukan validasi untuk lebih dari satu model serta menyediakan *hyperparameter* masing-masing secara otomatis dan sistematis [5].

Metode Penelitian

Dalam penelitian ini, peneliti menerapkan metodologi yang komprehensif untuk mengkaji dan meningkatkan kinerja model prediksi kelayakan kredit. Metodologi ini melibatkan serangkaian tahapan penting, dimulai dari *preprocessing* data, di mana peneliti mengatasi masalah *missing value*, melakukan *encoding* pada data kategoris, dan mengelola *outlier* untuk memastikan data yang digunakan dalam analisis adalah yang paling representatif. Selanjutnya, dataset dibagi menjadi subset pelatihan dan pengujian untuk memvalidasi model. Proses penting lainnya adalah *hyperparameter tuning* dengan metode *Grid Search* CV untuk mengoptimalkan model.

Tuning parameter merupakan proses penyesuaian parameter pada model *machine learning* untuk meningkatkan performanya [6]. Setelah proses tuning, peneliti melakukan *fitting* model menggunakan algoritma *Random Forest* yang telah dioptimalkan. Terakhir, dalam tahap evaluasi, metrik-metrik kualitas seperti *classification report* dan *confusion matrix* digunakan untuk menganalisis performa model.

Classification report adalah sebuah report sederhana yang hanya dengan sekali *coding*, kalian sudah akan mengetahui nilai *precission, recall, f1 score,* akurasi, rata-rata akurasi makro dan rata-rata akurasi terbeban [7]. *Confusion matrix* digunakan untuk memperoleh nilai *precision, recall*, dan *accuracy*. Nilai *Confusion matrix* biasanya ditunjukkan dalam satuan persen (%) [8].



Gambar 1: Tahapan Penelitian

Data

Data yang digunakan dalam penelitian ini diperoleh dari sumber eksternal, yaitu situs web Kaggle. Dataset ini terdiri dari sebanyak 32,581 baris data yang mencakup 12 atribut yang relevan untuk analisis kelayakan kredit yaitu: person_age, person_income, person_home_ownership, person_emp_length, loan_intent, loan_grade, loan_amnt, loan_int_rate, loan_status, loan_percent_income, cb_person_default_on_file, cb_person_cred_hist_length

person_age	person_income	person_home_ownership	person_emp_length
22	59000	RENT	123.0
21	9600	OWN	5.0
25	9600	MORTGAGE	1.0
23	65500	RENT	4.0
24	54400	RENT	8.0
...	...	...	...
57	53000	MORTGAGE	1.0
54	120000	MORTGAGE	4.0
65	76000	RENT	3.0
56	150000	MORTGAGE	5.0
66	42000	RENT	2.0

Gambar 2: Sampel Dataset 1-4

loan_intent	loan_grade	loan_amnt	loan_int_rate	loan_s
PERSONAL	D	35000	16.02	
EDUCATION	B	1000	11.14	
MEDICAL	C	5500	12.87	
MEDICAL	C	35000	15.23	
MEDICAL	C	35000	14.27	
...	...	...	...	
PERSONAL	C	5800	13.16	
PERSONAL	A	17625	7.49	
HOMEIMPROVEMENT	B	35000	10.99	
PERSONAL	B	15000	11.48	
MEDICAL	B	6475	9.99	

Gambar 3: Sampel Dataset 5-8

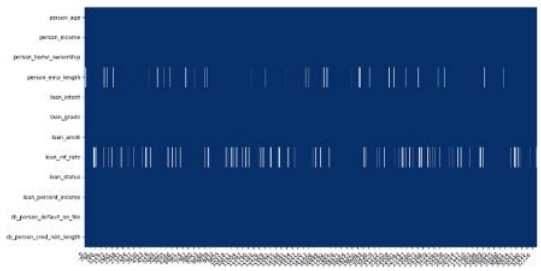
loan_status	loan_percent_income	cb_person_default_on_file	cb_person_cred_hist_length
1	0.59	Y	3
0	0.10	N	2
1	0.57	N	3
1	0.53	N	2
1	0.55	Y	4
...	...	...	...
0	0.11	N	30
0	0.15	N	19
1	0.46	N	28
0	0.10	N	26
0	0.15	N	30

Gambar 4: Sampel Dataset 9-12

Preprocessing

Handling Missing Value

Dalam tahapan ini, penelitian ini memfokuskan pada penanganan nilai-nilai yang hilang dalam dataset yang digunakan. Visualisasi dalam Gambar 2 menggambarkan distribusi nilai-nilai yang hilang dalam dataset yang digunakan dalam penelitian ini.



Gambar 5: Visualisasi Missing Value

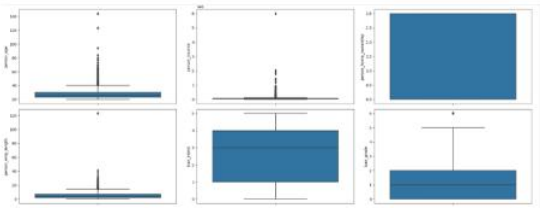
Keberadaan nilai-nilai yang hilang dalam data dapat memengaruhi keakuratan dan keandalan analisis. Untuk mengatasi permasalahan ini, penelitian ini menerapkan teknik imputasi mean. Teknik ini melibatkan perhitungan nilai rata-rata dari setiap kolom yang mengandung data yang hilang, dan kemudian mengganti nilai-nilai yang hilang dengan nilai rata-rata tersebut. Pendekatan ini memungkinkan kita untuk menjaga integritas dataset sambil meminimalkan dampak dari nilai-nilai yang hilang pada hasil analisis. Teknik imputasi mean yang digunakan dalam penelitian ini bertujuan untuk memastikan bahwa dataset yang digunakan dalam pemodelan adalah lengkap dan siap untuk proses selanjutnya.

Encoding Categorical Data

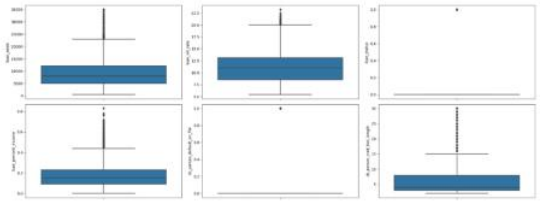
Pada tahap *encoding categorical data*, peneliti menerapkan metode Label *Encoder* untuk mengatasi variabel kategoris dalam dataset. Pendekatan ini membantu mengubah data kategoris menjadi format yang dapat diolah oleh algoritma machine learning [9]. Label *Encoder* mengassign label numerik unik untuk setiap nilai kategori dalam kolom yang sesuai. Hasilnya adalah transformasi data yang lebih sesuai dengan proses pemodelan. Ini memungkinkan peneliti untuk memasukkan informasi kategoris ke dalam model tanpa perlu menyusun multiple kolom baru yang bisa memengaruhi dimensi dataset secara signifikan. Metode ini adalah salah satu langkah penting dalam persiapan data yang memungkinkan peneliti untuk meningkatkan akurasi dan kinerja model.

Handling Outlier

Pada tahapan penanganan *outlier* dalam penelitian ini, peneliti menerapkan metode *Z-Score* untuk mengidentifikasi dan mengatasi data yang berada di luar ambang batas -3 hingga +3. Pendekatan ini membantu peneliti dalam mengevaluasi sejauh mana titik data individu berbeda dari rata-rata dan standar deviasi populasi.



Gambar 6: Visualisasi Handling Outlier 1-6



Gambar 7: Visualisasi Handling Outlier 7-12

Data yang melebihi batas tersebut dianggap sebagai *outlier* dan diperlakukan sesuai dengan langkah-langkah yang diperlukan, termasuk pemrosesan lebih lanjut atau penghapusan. Penggunaan metode *Z-Score* memungkinkan peneliti untuk menjaga integritas dataset sambil mengidentifikasi potensi anomali yang dapat memengaruhi hasil analisis secara signifikan. Langkah ini merupakan bagian integral dari persiapan data yang berfokus pada memastikan bahwa model prediksi kelayakan kredit yang dikembangkan oleh peneliti beroperasi pada data yang berkualitas dan terbebas dari gangguan yang mungkin terjadi akibat adanya *outlier*.

Dataset Splitting

Dalam tahap ini, dataset yang digunakan dalam penelitian ini dibagi menjadi dua subset utama: data pelatihan dan data tes. Data *splitting* atau pemisahan data adalah metode membagi data menjadi dua bagian atau lebih yang membentuk subhimpunan data. Umumnya, data *splitting* memisahkan dua bagian, bagian pertama digunakan untuk mengevaluasi atau uji data dan data lainnya digunakan untuk melatih model [10]. Peneliti memilih untuk membagi dataset menjadi 3 variasi pembagian. Diantaranya, diberikan pada tabel 1 yaitu:

Tabel 1: Pembagian Rasio Dataset

Rasio Data Tes	Rasio Data Pelatihan
90%	10%
80%	20%
75%	25%

Penggunaan rasio yang bervariasi ini bertujuan untuk memastikan bahwa model yang dikembangkan memiliki sejumlah data yang cukup untuk pelatihan yang memadai, sekaligus memungkinkan evaluasi yang solid terhadap performa model yang dihasilkan. Data pelatihan digunakan untuk melatih model dengan pola yang terkandung dalam dataset, sementara data tes digunakan untuk menguji sejauh mana model mampu melakukan prediksi yang akurat pada data yang tidak terlibat dalam proses pelatihan. Pembagian dataset ini adalah langkah penting dalam pengembangan model prediksi kelayakan kredit dan membantu menilai kinerja model secara obyektif.

Processing

Dalam tahap ini, peneliti menggunakan algoritma *Random Forest* untuk membangun model prediksi kelayakan kredit. *Random Forest* adalah salah satu jenis algoritma *Machine Learning* yang sangat berguna dalam tugas klasifikasi seperti yang dihadapi dalam penelitian ini. Algoritma ini bekerja dengan cara menggabungkan beberapa pohon keputusan (*decision trees*) menjadi satu model yang kuat.

Pohon keputusan adalah model yang mengambil keputusan berdasarkan serangkaian aturan. *Random Forest* mengambil langkah lebih lanjut dengan membuat banyak pohon keputusan secara acak. Masing-masing pohon ini memberikan prediksi independen, dan kemudian hasil dari semua pohon digabungkan untuk menghasilkan prediksi akhir.

Keuntungan *Random Forest* termasuk kemampuannya mengatasi *overfitting* (pemodelan yang terlalu mendetail pada data pelatihan) dan memiliki performa yang baik dalam menghadapi dataset besar dengan banyak atribut seperti yang digunakan dalam penelitian ini. Melalui proses ini, peneliti akan melakukan penyetelan *hyperparameter* menggunakan metode *Grid Search* CV untuk menemukan konfigurasi optimal dari model *Random Forest* yang akan digunakan dalam menganalisis kelayakan kredit.

Tabel 2: Hyperparameter Tuning

Hyperparameter	Value
n_estimators	10, 50, 100
max_depth	None, 10, 20, 30
min_samples_split	2, 5, 10
min_samples_leaf	1, 2, 4

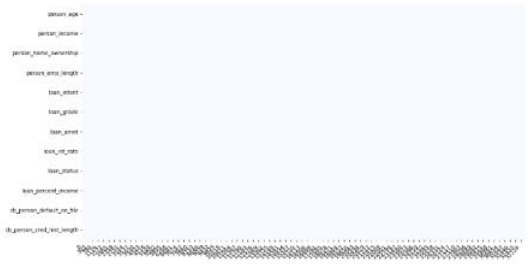
Setelah tahap pengolahan data yang mencakup penanganan nilai-nilai yang hilang, encoding data kategoris, penanganan *outlier* dengan metode *Z-Score*, dan pembagian dataset menjadi beberapa rasio perbaningan antara data tes dengan data pelatihan, peneliti melanjutkan ke tahap pemrosesan model. Dalam tahap ini, peneliti menjalankan proses *hyperparameter* tuning menggunakan *Grid Search Cross-Validation* (Grid Search CV). Peneliti telah menyusun daftar *hyperparameter* yang akan dioptimalkan sebagai tabel 2 berikut.

Proses *tuning* ini bertujuan untuk menemukan kombinasi *hyperparameter* terbaik yang akan meningkatkan performa model *Random Forest* dalam memprediksi kelayakan kredit. Selanjutnya, setelah mendapatkan konfigurasi *hyperparameter* terbaik, peneliti melakukan *fitting* model menggunakan *Random Forest* dengan *hyperparameter* yang telah dioptimalkan. Hasilnya adalah model prediksi kelayakan kredit yang siap untuk evaluasi lebih lanjut. Tahap pemrosesan ini memainkan peran kunci dalam menghasilkan model yang akurat dan andal dalam memprediksi kelayakan kredit, dan proses tuning hyperparameter memastikan bahwa model tersebut ditingkatkan secara signifikan.

Hasil dan Pembahasan

Hasil Preprocessing

Dalam tahap *preprocessing*, peneliti telah dengan cermat mengatasi nilai-nilai yang hilang dalam dataset. Metode yang diterapkan adalah imputasi mean, yang memungkinkan peneliti untuk mengisi nilai-nilai yang kosong dengan rerata dari atribut yang bersangkutan. Hasil dari tindakan ini adalah pemulihan integritas data dalam dataset yang awalnya mengandung sejumlah nilai yang tidak lengkap. Dengan demikian, dataset telah berhasil dipersiapkan dengan lebih baik untuk langkah-langkah analisis selanjutnya dalam penelitian ini.



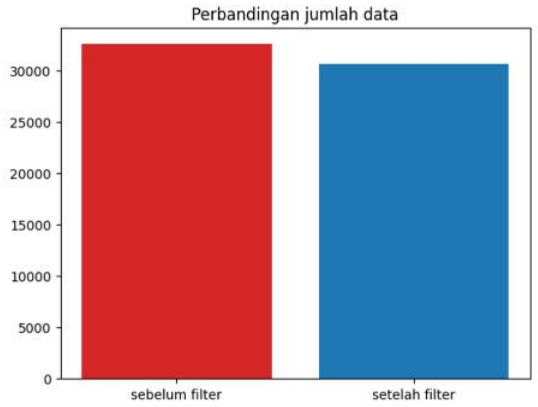
Gambar 8: Hasil Impute Missing Value

Langkah selanjutnya dalam tahap *preprocessing* adalah penanganan *outlier* menggunakan metode *Z-Score*. Proses ini bertujuan untuk mengidentifikasi titik-titik data yang ekstrem atau anomali yang mungkin ada dalam dataset. Hasil dari langkah ini adalah pemfilteran data, mengurangi jumlah baris dalam dataset dari 32,581 menjadi 30,649.

Setelah berhasil menyelesaikan imputasi dan penanganan *outlier*, peneliti melakukan pembagian dataset menjadi dua kelompok utama, yaitu data train dan data tes. Data train, yang terdiri dari 24,519 baris, akan digunakan untuk melatih model prediksi kelayakan kredit, sementara data tes, yang

berjumlah 6,130 baris, akan digunakan untuk menguji kinerja model.

Pembagian ini dilakukan dengan beberapa rasio data tes dan data pelatihan. Dengan cara ini, peneliti dapat memastikan bahwa model yang dikembangkan dapat diuji dengan baik dan objektif, dan dapat memberikan hasil yang akurat dalam menilai kelayakan kredit para peminjam.



Gambar 9: Hasil Filter Outlier

Hasil *Processing*

Dalam tahap "*Processing*," peneliti menggunakan algoritma *Random Forest* untuk membangun model prediksi kelayakan kredit. Model ini telah melewati serangkaian penyetelan *hyperparameter* dengan metode Grid SearchCV untuk mencari konfigurasi terbaik.

Tabel 3: Hasil Akurasi Tiap Rasio Dataset

Rasio Data Tes	Rasio Data Pelatihan	Akurasi
90%	10%	93,28%
80%	20%	93,07%
75%	25%	93,21%

Hasil dari tahap ini adalah pengembangan model yang sangat akurat dalam memprediksi kelayakan kredit. Dari beberapa rasio perbandingan yang telah dilakukan uji coba, model Random Forest memiliki tingkat akurasi tertinggi mencapai 93,28%, yang mengindikasikan kemampuan yang sangat baik dalam mengklasifikasikan peminjam menjadi kategori "kredit lancar" dan "kredit macet.". Adapun masing-masing akurasi hasil uji coba yang telah dilakukan adalah terdapat pada tabel 3 berikut.

Selain akurasi yang tinggi, parameter terbaik dari masing-masing rasio perbandingan data yang diidentifikasi oleh penelitian ini adalah seperti pada tabel 4 sebagai berikut. Konfigurasi ini optimal dalam menghasilkan model yang handal dalam menganalisis kelayakan kredit.

Tabel 4: Hasil Tuning

Rasio Data Tes	Rasio Data Pelatihan	Hyperparameter	Value
90%	10%	n_estimators	100
		max_depth	30
		min_samples_split	2
		min_samples_leaf	1
80%	20%	n_estimators	100
		max_depth	None
		min_samples_split	2
		min_samples_leaf	1
75%	25%	n_estimators	100
		max_depth	30
		min_samples_split	2
		min_samples_leaf	1

Evaluasi

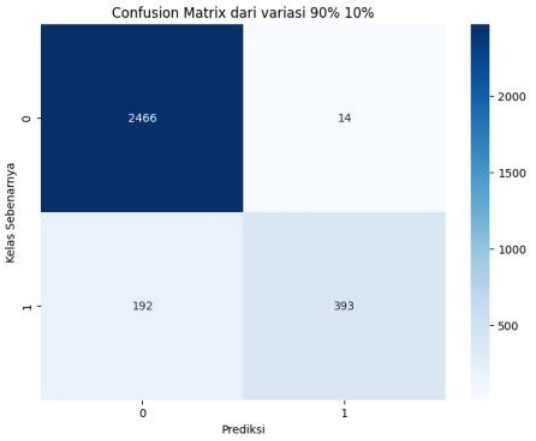
Setelah model kelayakan kredit menggunakan algoritma *Random Forest* telah berhasil dikembangkan, peneliti melakukan evaluasi untuk mengukur sejauh mana model tersebut dapat memprediksi kelayakan peminjam dengan akurat. Evaluasi dilakukan dengan merujuk pada *Confusion Matrix* yang menggambarkan hasil prediksi model. *Confusion Matrix* menghasilkan metrik-metrik penting berikut:

True Positive (TP): Sebanyak 2466 kasus berhasil diprediksi sebagai "kredit lancar."

False Positive (FP): Terdapat 14 kasus yang salah diprediksi sebagai "kredit lancar."

False Negative (FN): Terdapat 192 kasus yang salah diprediksi sebagai "kredit macet."

True Negative (TN): Sebanyak 393 kasus berhasil diprediksi sebagai "kredit macet."



Gambar 10: Visualisasi Confusion Matrix

Dengan informasi ini, beberapa metrik evaluasi telah dihitung, termasuk akurasi, presisi, recall (sensitivitas), F1-score, dan lainnya.

Laporan Klasifikasi dari variasi dataset 90 10:

	precision	recall	f1-score	support
0.0	0.9278	0.9944	0.9599	2480
1.0	0.9656	0.6718	0.7923	585
accuracy			0.9328	3065
macro avg	0.9467	0.8331	0.8761	3065
weighted avg	0.9350	0.9328	0.9279	3065

Gambar 11: Classification Report

Sesuai laporan klasifikasi dari gambar di atas yaitu *precision* 94,67%, dan *recall* 83,31%. Hasil evaluasi ini akan memberikan pandangan yang lebih komprehensif tentang kinerja model dan seberapa baik model ini dalam memprediksi kelayakan kredit peminjam. Dengan tingkat akurasi sebesar 93,28% dan parameter terbaik yang telah diidentifikasi, model ini diharapkan dapat menjadi alat yang efektif dalam pengambilan keputusan kredit.

Penutup

Dalam penelitian ini, peneliti telah melaksanakan serangkaian langkah yang komprehensif dalam upaya mengembangkan model prediksi kelayakan kredit yang handal. Dengan menggunakan algoritma *Random Forest* dan melalui proses prapemrosesan data yang cermat, peneliti berhasil mencapai tingkat akurasi sebesar 93,28% dalam memprediksi apakah seorang peminjam layak atau tidak menerima kredit. Selain itu, parameter terbaik untuk model ini telah diidentifikasi, yaitu 'max_depth': 30, 'min_samples_leaf': 1, 'min_samples_split': 2, dan 'n_estimators': 100, yang menunjukkan konfigurasi optimal untuk analisis kelayakan kredit.

Hasil dari penelitian ini memiliki potensi untuk memberikan dampak positif yang signifikan dalam dunia keuangan dan pengambilan keputusan kredit. Dengan model ini, lembaga keuangan dan pemberi pinjaman dapat memperoleh alat yang kuat untuk menilai kelayakan peminjam secara lebih objektif dan berdasarkan data. Model ini juga dapat membantu dalam mengurangi risiko kredit yang mungkin terjadi akibat pengambilan keputusan yang kurang tepat.

Saran pengembangan penelitian selanjutnya bisa dikembangkan dengan penggunaan metode ensemble learning lainnya untuk komparasi hasil terbaik dengan *Random Forest*.

Daftar Pustaka

[1] B. Prasojo dan E. Haryatmi, "Analisa Prediksi Kelayakan Pemberian Kredit Pinjaman dengan Metode Random Forest", Jurnal Nasional Teknologi dan Sistem Informasi, vol. 7, no. 2, doi: 10.25077/teknosi.v7i2.2021.79-89, 2021.

[2] D. Prasetyo Tarigan dan A. Wantoro, "Sistem Pendukung Keputusan Pemberian Kredit Mobil dengan Fuzzy Tsukamoto (Studi Kasus : Pt Clipan Finance)", TELEFORTECH Journal of Telematics and Information Technology, Vol.1, No. 1, DOI: 10.33365/tft.v1i1.870, 2020.

[3] M. Rizki, M. Isnaini, H. Umam dan M. L. Hamzah, "Aplikasi Data Mining Dengan Metode CHAID Dalam Menentukan Status Kredit", Jurnal Sains, Teknologi dan Industri, vol. 18, no. 1, pp. 29–33, 2020.

[4] Syafi'i, O. Nurdiawan dan G. Dwilestari, "Penerapan Machine Learning Untuk Menentukan Kelayakan Kredit Menggunakan Metode Support Vektor Machine", Jurnal Sistem Informasi dan Manajemen, vol. 10, no. 2, 2022.

[5] Z. M. E. Darmawan dan A. Fauzan Dianta, "Implementasi Optimasi Hyperparameter GridSearchCV Pada Sistem Prediksi Serangan Jantung Menggunakan SVM", Teknologi: Jurnal Ilmiah Sistem Informasi, vol. 13, no. 1, 2023.

[6] Obey Al Farobi, "Implementasi Metode Support Vector Machine (Svm) Untuk Mengetahui Respon Masyarakat Indonesia Terhadap pemberian vaksin sinovac ", Skripsi, Fakultas Sains dan Teknologi Universitas Islam Negeri Syarif Hidayatullah Jakarta, 2021.

[7] U. L. Yuhana and A. Purwarianti, "Tuning Hyperparameter pada Gradient Boosting", Jurnal Edukasi dan Penelitian Informatika (JEPIN), vol. 8, no. 1, 2022.

[8] W. I. Rahayu, C. Prianto, and E. A. Novia, "Perbandingan Algoritma K-Means Dan Naïve Bayes Untuk Memprediksi Prioritas Pembayaran Tagihan Rumah Sakit Berdasarkan Tingkat Kepentingan Pada Pt. Pertamina (Persero)", Jurnal Teknik Informatika, vol. 13, no. 2, 2021.

[9] W. D McGinnis, C. Siu, A. and H. Huang, "Category Encoders: a scikit-learn-contrib package of transformers for encoding categorical data", The Journal of Open Source Software, vol. 3, no. 21, doi: 10.21105/joss.00501, 2018.

[10] S. Wahyu Iriananda, R. P. Putra dan K. S. Nugroho, " Analisis Sentimen dan Analisis Data Eksploratif Ulasan Aplikasi Marketplace Google Playstore", CIASTECH 2021 – >Kesiapan Indonesia Dalam Menghadapi Krisis Energi Global, ISSN Cetak : 2622-1276, Universitas Widyagama, Malang, 2021.