

Penerapan Algoritma Xgboost Untuk Analisis Dan Klasifikasi Risiko Kredit Berbasis Data Mining

Application of XGBoost Algorithm for Credit Risk Analysis and Classification Based on Data Mining

Sigit Frastio¹, Mas Ayoe Elhias^{b,2}

^{a,b}Universitas Potensi Utama, Jl. KL. Yos Sudarso Km. 6,5 No. 3-A, Tanjung Mulia, Medan, 20241, Indonesia
sigitfrasetio78@gmail.com¹, masayoe02@gmail.com @gmail.com 2

ABSTRAK

Penelitian ini bertujuan untuk menganalisis dan mengklasifikasikan risiko kredit berbasis data mining menggunakan algoritma Extreme Gradient Boosting (XGBoost). Data mining, sebagai pendekatan dalam eksplorasi pola dan hubungan antara karakteristik nasabah dengan kemungkinan gagal bayar (default), memiliki potensi untuk meningkatkan akurasi prediksi risiko kredit. Rumusan masalah dalam penelitian ini mencakup penerapan data mining untuk mengidentifikasi pola dan hubungan atribut nasabah dengan risiko gagal bayar, implementasi algoritma XGBoost dalam klasifikasi risiko kredit, pemahaman terhadap risiko Non-Performing Loan (NPL) yang dapat mengganggu stabilitas keuangan bank, serta pemahaman tentang karakteristik dan kualitas data yang digunakan dalam pembangunan model. Tujuan penelitian ini adalah untuk menerapkan data mining dalam mengeksplorasi pola risiko gagal bayar nasabah, mengembangkan model klasifikasi berbasis XGBoost, mengevaluasi kinerja algoritma XGBoost dalam prediksi risiko kredit, serta mengidentifikasi fitur-fitur yang paling berpengaruh terhadap klasifikasi risiko kredit. Hasil yang diperoleh menunjukkan bahwa algoritma XGBoost berhasil mencapai tingkat akurasi sebesar 87%, presisi 82%, recall 93%, dan F1-Score 87%. Penelitian ini diharapkan dapat memberikan kontribusi dalam pengambilan keputusan pemberian kredit yang lebih akurat dan mengurangi potensi kerugian bagi lembaga keuangan.

Kata Kunci : Data Mining, XGBoost, Risiko Kredit, Klasifikasi, Non-Performing Loan

ABSTRACT

This study aims to analyze and classify credit risk based on data mining using the Extreme Gradient Boosting (XGBoost) algorithm. Data mining, as an approach for exploring patterns and relationships between customer characteristics and the likelihood of default, has the potential to improve the accuracy of credit risk prediction. The research questions include the application of data mining to identify patterns and relationships between customer attributes and the risk of default, implementation of the XGBoost algorithm in credit risk classification, understanding of Non-Performing Loan (NPL) risk that can disrupt bank financial stability, and comprehension of the characteristics and quality of data used in model development. The objectives of this study are to apply data mining to explore patterns of default risk, develop an XGBoost-based classification model, evaluate the performance of the XGBoost algorithm in credit risk prediction, and identify the most influential features in credit risk classification. The results show that the XGBoost algorithm achieved an accuracy of 87%, precision of 82%, recall of 93%, and an F1-Score of 87%. This study is expected to contribute to more accurate credit decision-making and reduce potential losses for financial institutions. Restaurants are a gathering place for all groups to come and enjoy the food served by the restaurant.

Keywords: Data Mining, XGBoost, Credit Risk, Classification, Non-Performing Loan

Info Artikel :

Disubmit: 11 September 2025

Direview: 10 Oktober 2025

Diterima : 3 November 2025

Copyright © 2025 – senadimu. All rights reserved.

1. PENDAHULUAN

Dalam industri perbankan, penyaluran kredit merupakan salah satu kegiatan utama yang berperan penting dalam menghasilkan keuntungan. Namun, aktivitas ini juga mengandung risiko yang tinggi, terutama risiko kredit atau risiko kegagalan pembayaran dari debitur. Oleh karena itu, penting bagi lembaga keuangan untuk memiliki sistem yang dapat membantu dalam mengidentifikasi dan mengevaluasi kemungkinan seorang calon nasabah mengalami gagal bayar. Salah satu pendekatan yang digunakan untuk memodelkan risiko tersebut adalah *Probability of Default* (PD) Model, yaitu model prediktif yang bertujuan untuk memperkirakan probabilitas bahwa seorang peminjam akan mengalami default dalam jangka waktu tertentu [1].

Seiring berkembangnya teknologi dan ketersediaan data yang semakin besar, pendekatan data mining menjadi alternatif yang menjanjikan untuk membangun model prediksi risiko kredit. Data mining memungkinkan eksplorasi pola tersembunyi dalam data historis debitur guna memprediksi kemungkinan gagal bayar berdasarkan atribut tertentu, seperti usia, pendapatan, riwayat pembayaran, dan status pekerjaan. Dalam beberapa tahun terakhir, algoritma seperti *Support Vector Machine* (SVM), *Decision Tree* (C4.5), dan *Naïve Bayes* telah digunakan secara luas dalam penelitian untuk menganalisis dan mengklasifikasikan risiko kredit dengan tingkat akurasi yang [2].

Penelitian oleh Marco et al. (2024) yang membandingkan kinerja algoritma SVM dan Naïve Bayes dalam memprediksi risiko kredit menunjukkan bahwa SVM memiliki akurasi lebih tinggi yaitu sekitar 88%, dibandingkan dengan Naïve Bayes yang hanya mencapai sekitar 75%. Sementara itu, Sari (2025) menggunakan algoritma Decision Tree C4.5 dan memperoleh akurasi sebesar 67,48% dalam mengklasifikasikan data nasabah berdasarkan risiko gagal bayar. Penelitian lain oleh [3] menggunakan algoritma yang sama pada platform RapidMiner dan berhasil memperoleh akurasi yang lebih tinggi, yaitu mencapai 97,78%, menunjukkan potensi besar dari penerapan metode klasifikasi ini.

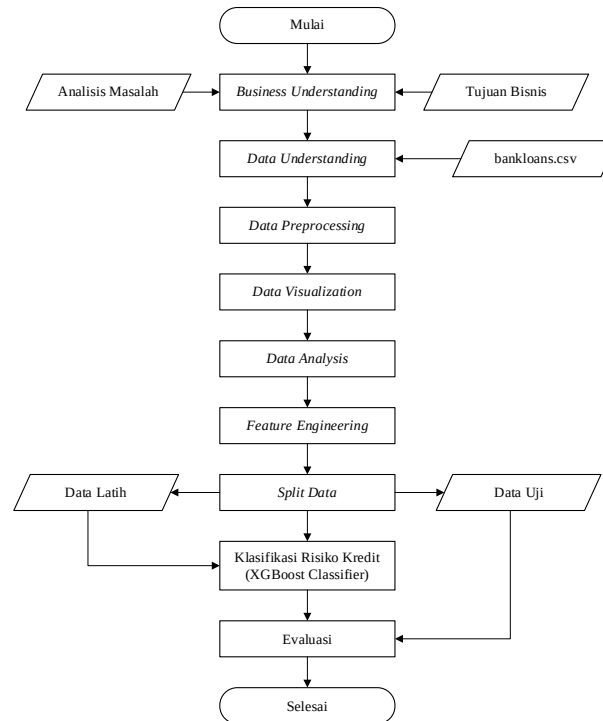
Meskipun model klasik masih banyak digunakan, tren terkini menunjukkan adanya pergeseran ke metode yang lebih kompleks seperti *ensemble learning* dan *deep learning*. [1] mengeksplorasi penggunaan algoritma LightGBM dan XGBoost yang menunjukkan performa unggul dalam menangani data besar dan kompleks dalam kasus prediksi default. Sementara itu, [4] mengusulkan arsitektur gabungan antara *Convolutional Neural Network* (CNN) dan *Transformer* yang menunjukkan hasil lebih baik dibanding model-model tradisional dalam memprediksi perilaku gagal bayar nasabah.

Selain akurasi, aspek keadilan (*fairness*) dan *interpretabilitas model* menjadi perhatian utama dalam pengembangan model risiko kredit. Penelitian oleh [5] menunjukkan bahwa tahapan preprocessing, seperti penggunaan *Truncated SVD*, dapat mempengaruhi keadilan hasil model. Hal ini penting mengingat sistem keuangan harus mematuhi prinsip non-diskriminatif dan transparan. Lebih lanjut, [6] meneliti penggunaan teknik SHAP (*SHapley Additive exPlanations*) dalam menjelaskan keputusan model PD dan menemukan bahwa model yang dapat dijelaskan memiliki potensi lebih besar untuk diadopsi secara luas oleh lembaga keuangan karena meningkatkan kepercayaan pengguna dan regulator[7].

Dengan mempertimbangkan kondisi tersebut, penelitian ini berfokus pada analisis dan klasifikasi risiko kredit berbasis data mining untuk membangun model *Probability of Default* yang andal dan akurat. Tujuan utama penelitian ini adalah membantu bank dalam pengambilan keputusan pemberian kredit serta meminimalkan potensi kerugian finansial akibat gagal bayar. Melalui pemanfaatan data historis nasabah dan penerapan algoritma klasifikasi yang tepat, diharapkan model yang dibangun mampu mendukung manajemen risiko kredit secara lebih efektif dan efisien.

2. METODE PENELITIAN

Metode penelitian ini disusun secara sistematis untuk mendukung analisis dan klasifikasi risiko kredit berbasis data mining. Diagram alur berikut menjelaskan secara visual tahapan metodologi penelitian ini dapat dilihat pada gambar.1.



Gambar. 1 Tahapan Penelitian Analisis dan Klasifikasi Risiko Kredit (*XGBoost Classifier*)

Tahapan metodologi penelitian dalam proses klasifikasi risiko kredit menggunakan algoritma *XGBoost Classifier* [8]. Proses dimulai dari tahap Analisis Masalah pada bagian *Business Understanding* untuk mengidentifikasi permasalahan dan menentukan tujuan bisnis yang ingin dicapai. Selanjutnya dilakukan *Data Understanding* dengan memahami dataset yang digunakan (misalnya *bankloans.csv*), termasuk karakteristik dan struktur datanya. Tahap berikutnya adalah *Data Preprocessing*, yang meliputi pembersihan data, penanganan nilai hilang, deteksi dan penanganan *outlier*, serta normalisasi variabel jika diperlukan. Setelah itu, dilakukan *Data Visualization* untuk memberikan gambaran visual mengenai pola dan distribusi data, diikuti dengan *Data Analysis* untuk mengeksplorasi hubungan antar variabel[9].

Tahap *Feature Engineering* dilakukan untuk membuat, memodifikasi, atau memilih fitur yang relevan guna meningkatkan kinerja model. Data yang telah siap kemudian dibagi (*Split Data*) menjadi data latih dan data uji. Model Klasifikasi Risiko Kredit kemudian dibangun menggunakan algoritma *XGBoost Classifier* dengan memanfaatkan data latih, dan selanjutnya dievaluasi menggunakan data uji pada tahap Evaluasi [10]. Proses ini berakhir dengan tahap Selesai, di mana hasil evaluasi model dapat digunakan sebagai dasar pengambilan keputusan terkait risiko kredit.

1) Dataset Penelitian

Penelitian ini memanfaatkan data sekunder yang diperoleh dari platform *Kaggle*,

khususnya dataset *bankloans.csv* yang berisi informasi historis terkait profil dan risiko kredit nasabah. Seluruh proses pengolahan dan analisis data dilaksanakan secara *offline* menggunakan perangkat komputer pribadi peneliti. Kegiatan penelitian meliputi tahapan pemahaman data, pra-pemrosesan, analisis, *feature engineering*, pembangunan model klasifikasi, hingga evaluasi model, yang dilakukan di lingkungan kerja peneliti dengan dukungan perangkat lunak *Python* dan pustaka *machine learning* yang relevan [11].

	age	ed	employ	address	income	debtinc	creddebt	othdebt	default
0	41	3	17	12	176	9.3	11.359392	5.008608	1.0
1	27	1	10	6	31	17.3	1.362202	4.000798	0.0
2	40	1	15	14	55	5.5	0.856075	2.168925	0.0
3	41	1	15	14	120	2.9	2.658720	0.821280	0.0
4	24	2	2	0	28	17.3	1.787436	3.056564	1.0

Gambar. 2 Dataset Risiko Kredit

Dataset ini berisi 851 entri data yang mencakup berbagai variabel terkait profil nasabah dan informasi keuangan yang relevan untuk analisis risiko kredit. Variabel-variabel tersebut dapat dilihat pada tabel. 1.

Tabel. 1 Deskripsi Dataset Penelitian

No	Nama Variabel	Deskripsi	Tipe Data	Keterangan Nilai
1	<i>age</i>	Usia individu (dalam tahun)	Numerik (<i>Integer</i>)	Semakin tinggi nilai menunjukkan usia lebih tua
2	<i>ed</i>	Tingkat pendidikan	Kategorikal (<i>Integer</i>)	Angka merepresentasikan level pendidikan tertentu
3	<i>employ</i>	Lama bekerja (dalam tahun)	Numerik (<i>Integer</i>)	Semakin tinggi nilai menunjukkan pengalaman kerja lebih lama
4	<i>address</i>	Lama tinggal di alamat saat ini (dalam tahun)	Numerik (<i>Integer</i>)	Menggambarkan stabilitas domisili
5	<i>income</i>	Pendapatan tahunan (dalam ribuan satuan mata uang)	Numerik (<i>Float</i>)	Semakin tinggi nilai menunjukkan pendapatan lebih besar
6	<i>debtinc</i>	Rasio utang terhadap pendapatan	Numerik (<i>Float</i>)	Mengukur proporsi utang terhadap pendapatan
7	<i>creddebt</i>	Jumlah utang kartu kredit	Numerik (<i>Float</i>)	Diukur dalam satuan mata uang
8	<i>othdebt</i>	Jumlah utang lainnya	Numerik (<i>Float</i>)	Diukur dalam satuan mata uang
9	<i>default</i>	Status gagal bayar	Kategorikal (<i>Biner</i>)	1 = gagal bayar, 0 = tidak gagal bayar

Dataset yang ditampilkan berisi data pinjaman bank dengan beberapa variabel penting yang menjelaskan kondisi finansial dan demografis peminjam. Kolom *age* menunjukkan usia peminjam, *ed* mencerminkan tingkat pendidikan, dan *employ* mengindikasikan lama masa kerja dalam tahun. Kolom *address* merepresentasikan lama tinggal di alamat saat ini dalam tahun, sedangkan *income* menunjukkan penghasilan tahunan. Selain itu, terdapat variabel *debtinc* yang merupakan rasio hutang terhadap penghasilan, *creddebt* yang menggambarkan jumlah hutang kartu kredit, dan *othdebt* sebagai total hutang lainnya. Variabel target *default* menunjukkan status kegagalan pembayaran pinjaman, dengan nilai 1 berarti peminjam gagal

bayar dan 0 berarti tidak gagal bayar. Data ini berguna untuk analisis risiko kredit dan dapat digunakan dalam pemodelan prediktif untuk memprediksi kemungkinan gagal bayar berdasarkan atribut-atribut tersebut.

2) Pengolahan Data

Pada tahapan ini pengolahan data dilakukan untuk memastikan data telah clear atau bersih dari kesalahan, duplikasi dan data outlier, terlebih dahulu dilakukan pemahaman bisnis dalam bidang keuangan. Salah satu tantangan utama yang dihadapi lembaga perbankan adalah meningkatnya risiko kredit bermasalah (*Non-Performing Loan/NPL*) yang dapat mengganggu stabilitas keuangan dan profitabilitas bank. Risiko ini timbul ketika nasabah gagal memenuhi kewajiban pembayaran kredit tepat waktu sesuai perjanjian. Proses penilaian kelayakan kredit yang tidak optimal dapat menyebabkan pemberian kredit kepada debitur berisiko tinggi, yang pada akhirnya meningkatkan kerugian finansial dan biaya operasional akibat penanganan kredit macet.

Tujuan dari penelitian ini adalah membangun model klasifikasi risiko kredit berbasis *data mining* yang mampu memprediksi kemungkinan nasabah gagal bayar (*default*) pada saat pengajuan pinjaman. Model ini akan menggunakan algoritma *XGBoost Classifier* untuk menghasilkan prediksi yang akurat dengan memanfaatkan variabel-variabel seperti usia, tingkat pendidikan, lama bekerja, pendapatan tahunan, rasio utang terhadap pendapatan, dan komponen utang lainnya.

3) Data Understanding

Pada tahapan ini dilakukan untuk memahami karakteristik data yang digunakan dalam penelitian, termasuk struktur, tipe variabel, serta makna dari setiap atribut yang tersedia. Dataset yang digunakan pada penelitian ini berisi informasi demografis, riwayat pekerjaan, kondisi keuangan, dan status gagal bayar nasabah.

4) Data Preprocessing

Tahap *data preprocessing* merupakan proses penting sebelum data digunakan untuk pemodelan. Tujuan utamanya adalah memastikan kualitas data agar analisis yang dilakukan menghasilkan model yang akurat dan andal. Pada tahap ini, dilakukan beberapa langkah utama, yaitu *data cleaning*, *outlier detection*, dan *deal with outliers*.

5) Data Visualization

Tahap visualisasi data dilakukan untuk memahami pola distribusi, hubungan antar variabel, dan perbedaan karakteristik antar kelompok pada dataset. Beberapa teknik visualisasi yang digunakan meliputi (*Histogram of Ages*, *Histogram of Credit Debt (creddebt)*, *Box Plot for Income by Default Status*, *Scatter Plot of Income vs. Debt-to-Income Ratio*, *Bar Plot for Average Income by Education Level*, *Pie Chart Showing the Proportion of Default vs. Non-Default Cases*). Teknik ini digunakan atas dasar kebutuhan penelitian.

6) Data Analysis

Tahap *data analysis* bertujuan untuk menjawab pertanyaan penelitian yang telah dirumuskan, dengan memanfaatkan data yang telah melalui proses pembersihan, penanganan *outlier*, dan visualisasi. Analisis ini dilakukan untuk menggali hubungan, perbedaan, dan pola tertentu yang dapat memberikan wawasan dalam memahami faktor-faktor yang memengaruhi risiko *default* pada nasabah, seperti (perbedaan rata-rata usia, hubungan tingkat pendidikan, pengaruh lama *Debt-to-Income Ratio*, hubungan jumlah hutang dengan risiko Default).

7) Feature Engineering

Tahap *feature engineering* dilakukan untuk membuat, mengubah, atau menggabungkan fitur-fitur dalam dataset agar dapat meningkatkan kemampuan model dalam menangkap pola dan hubungan antar variabel. Langkah ini sangat penting karena kualitas fitur seringkali lebih menentukan performa model dibanding hanya mengandalkan algoritma.

8) Split Data

Dataset dibagi menjadi dua bagian, yaitu 80% data latih (*training set*) dan 20% data uji

(*testing set*). Proporsi 80:20 dipilih karena memberikan keseimbangan antara jumlah data yang cukup untuk pelatihan dan jumlah data yang memadai untuk evaluasi.

9) Klasifikasi Risiko Kredit (*XGBoost Classifier*)

Tahap ini, dilakukan pembangunan model klasifikasi untuk memprediksi risiko kredit nasabah menggunakan algoritma XGBoost Classifier (*Extreme Gradient Boosting*). XGBoost merupakan salah satu algoritma *ensemble learning* berbasis pohon keputusan (*decision tree*) yang menggabungkan banyak model pohon secara bertahap, di mana setiap pohon baru memperbaiki kesalahan yang dibuat oleh pohon sebelumnya.

Formulasi Fungsi Objektif, XGBoost meminimalkan fungsi objektif berikut:

$$\mathcal{L}(\phi) = \sum_{i=1}^n l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \quad (1)$$

Fungsi regularisasi didefinisikan sebagai:

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^T w_j^2 \quad (2)$$

Pendekatan *Taylor Orde* Kedua, untuk efisiensi komputasi, fungsi kerugian didekati menggunakan pengembangan *Taylor orde* kedua:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^n \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t) \quad (3)$$

dengan

$$g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}} \text{ (gradient)} \quad (4)$$

$$h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2} \text{ (hessian)} \quad (5)$$

Nilai *gain* yang digunakan untuk memilih titik pemisahan terbaik dihitung dengan

$$\text{Gain} = \frac{1}{2} \left[\frac{(\sum_{i \in L} g_i)^2}{\sum_{i \in L} h_i + \lambda} + \frac{(\sum_{i \in R} g_i)^2}{\sum_{i \in R} h_i + \lambda} - \frac{(\sum_{i \in L \cup R} g_i)^2}{\sum_{i \in L \cup R} h_i + \lambda} \right] - \gamma \quad (6)$$

Melalui kombinasi pendekatan *gradient boosting* dan regularisasi ini, XGBoost mampu menghasilkan model prediksi yang akurat, efisien, dan memiliki kemampuan generalisasi yang baik untuk tugas klasifikasi risiko kredit [12].

10) Evaluasi Model

Setelah pelatihan, model dievaluasi menggunakan data uji. Metrik yang digunakan meliputi *Accuracy*, *Precision*, *Recall*, dan *F1 Score*, serta laporan klasifikasi (*classification report*) yang memberikan gambaran detail performa pada masing-masing kelas. Evaluasi ini bertujuan menilai seberapa baik model dalam memprediksi risiko kredit.

Setelah model dibuat, dilakukan evaluasi menggunakan *test set*. Model digunakan untuk memprediksi label, lalu dibandingkan dengan label asli. Hasilnya diringkas dalam: 1. True Positives (TP), 2. False Positives (FP), 3. True Negatives (TN), 4. False Negatives (FN). Dari *confusion matrix*, dihitunglah metrik evaluasi, Akurasi: Proporsi prediksi yang benar dari keseluruhan data. Presisi: Seberapa banyak prediksi positif yang benar-benar positif. Recall: Seberapa banyak data positif yang berhasil terprediksi dengan benar. F1-Score: Harmonik rata-rata dari presisi dan recall, berguna untuk data tidak seimbang.

Metrik ini digunakan untuk menilai performa klasifikasi model secara komprehensif. Menandai akhir dari proses klasifikasi, di mana model telah selesai dibuat dan dievaluasi. Selanjutnya, hasil dapat dianalisis, disimpulkan, dan digunakan untuk pengambilan keputusan atau pengembangan lebih lanjut.

A. Tinjauan Pustaka

1) Gap Penelitian

Informasi tentang GAP penelitian sangat dibutuhkan sebelum memulai penelitian, pada bab ini akan diuraikan penelitian terdahulu dengan bidang dan topik yang hampir sama, dalam bagian ini juga nantinya mendukung penjelasan dari *state of the art* terhadap bidang ilmu informatika.

Penelitian dengan judul “*Credit scoring: Does XGBoost outperform logistic regression? A test on Italian SMEs*” Penelitian ini membandingkan kinerja XGBoost dengan regresi logistik dalam penilaian kredit pada UKM Italia. Hasil penelitian menunjukkan bahwa XGBoost dapat memberikan kinerja yang lebih baik dalam konteks ini [9].

Penelitian dengan judul “*Advanced user credit risk prediction model using LightGBM, XGBoost and TabNet with SMOTEENN*” Penelitian ini mengembangkan model prediksi risiko kredit pengguna canggih menggunakan LightGBM, XGBoost, dan TabNet dengan SMOTEENN untuk menangani data tidak seimbang. Pendekatan ini menunjukkan potensi dalam meningkatkan akurasi prediksi risiko kredit pada pengguna canggih [10].

Penelitian dengan judul “*A novel XGBoost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified focal loss function*” Penelitian ini mengusulkan ekstensi baru dari XGBoost untuk menangani masalah ketidakseimbangan kelas dalam data skor kredit dengan menggabungkan link nilai ekstrem umum dan fungsi loss fokus yang dimodifikasi. Pendekatan ini bertujuan untuk meningkatkan kinerja model dalam menghadapi data yang tidak seimbang [12].

Penelitian dengan judul “*A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting*” Penelitian ini mengembangkan pendekatan machine learning yang kuat untuk analisis risiko kredit menggunakan dataset pinjaman besar dengan memanfaatkan deep learning dan XGBoost. Pendekatan ini menunjukkan potensi dalam meningkatkan akurasi prediksi risiko kredit pada dataset besar [13].

Penelitian dengan judul “*Enhanced credit score prediction using ensemble deep learning model*” Penelitian ini mengusulkan model deep learning ensemble yang ditingkatkan untuk prediksi skor kredit, menggabungkan XGBoost, LightGBM, dan TabNet. Pendekatan ini bertujuan untuk meningkatkan akurasi prediksi skor kredit dengan memanfaatkan kekuatan masing-masing model [14].

Meskipun beberapa penelitian telah mengusulkan metode untuk menangani ketidakseimbangan kelas, tantangan ini masih menjadi masalah utama dalam prediksi risiko kredit. Model seperti XGBoost sering dianggap sebagai “black box”, sehingga sulit untuk menjelaskan keputusan model kepada pemangku kepentingan. Sebagian besar penelitian berfokus pada dataset tertentu atau wilayah geografis tertentu, sehingga hasilnya mungkin tidak dapat digeneralisasi ke konteks lain [15].

2) Risiko Kredit (*Credit Risk*)

Risiko kredit adalah potensi kerugian yang dihadapi oleh lembaga keuangan akibat kegagalan nasabah dalam memenuhi kewajiban pembayaran pinjaman [16]. Dalam konteks perbankan, risiko ini diukur dengan tiga parameter utama: *Probability of Default* (PD), *Loss Given Default* (LGD), dan *Exposure at Default* (EAD). Ketiga parameter ini digunakan untuk menghitung *Expected Credit Loss* (ECL), yang merupakan estimasi kerugian yang diharapkan akibat default. ECL dihitung dengan rumus:

$$ECL = PD \times LGD \times EAD \quad (7)$$

3) Model Pengukuran Risiko Kredit

Terdapat dua pendekatan utama dalam mengukur risiko kredit, Pendekatan Tradisional: Menggunakan model statistik klasik seperti Altman Z-Score dan model logistik untuk menghitung PD berdasarkan rasio keuangan dan variabel makroekonomi. Pendekatan Modern:

Menggunakan teknik data mining dan machine learning untuk membangun model PD yang lebih akurat dan adaptif terhadap perubahan data.

4) *Probability of Default (PD)*

PD adalah estimasi probabilitas bahwa seorang nasabah akan gagal memenuhi kewajiban pembayaran pinjaman dalam periode tertentu. PD digunakan dalam berbagai analisis kredit dan kerangka manajemen risiko untuk menilai potensi kerugian akibat default. Menurut definisi dalam literatur, PD mencakup kemungkinan bahwa peminjam tidak dapat memenuhi pembayaran pokok atau bunga sesuai dengan ketentuan yang telah disepakati.

Dalam beberapa tahun terakhir, pendekatan berbasis data mining telah banyak digunakan untuk membangun model PD yang lebih akurat. Metode seperti *Random Forest*, *XGBoost*, dan *LightGBM* telah terbukti efektif dalam klasifikasi risiko kredit. Misalnya, penelitian oleh Yu et al. (2024) menunjukkan bahwa kombinasi *LightGBM* dengan teknik SMOTEENN mampu meningkatkan akurasi prediksi risiko kredit pada data nasabah bank komersial.

Selain itu, pendekatan berbasis *Gaussian Mixture Models (GMM)* juga diperkenalkan oleh Arian et al. (2020) sebagai alternatif dalam klasifikasi skor kredit, dengan hasil yang kompetitif dibandingkan metode konvensional [17].

5) *Algoritma XGBoost (Extreme Gradient Boosting)*

XGBoost (Extreme Gradient Boosting) adalah algoritma pembelajaran mesin berbasis ensemble yang menggabungkan banyak model pohon keputusan (*decision trees*) untuk memprediksi output. *XGBoost* menggunakan pendekatan *boosting* yang mengoptimalkan model secara iteratif dengan memberikan bobot lebih pada data yang lebih sulit diprediksi. Berikut adalah formulasi dasar dari algoritma *XGBoost* dalam konteks machine learning dan data mining [17].

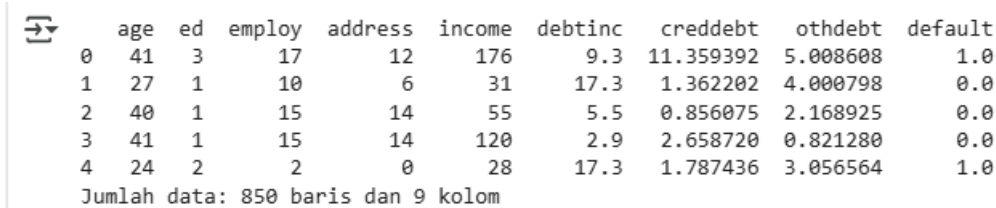
6) Data Mining dalam Analisis Risiko Kredit

Data mining adalah proses menemukan pola atau informasi yang berguna dari kumpulan data besar menggunakan teknik statistik, matematika, dan algoritma pembelajaran mesin [18]. Dalam konteks risiko kredit, data mining digunakan untuk menganalisis data nasabah dan memprediksi kemungkinan gagal bayar [19]. Teknik data mining untuk klasifikasi risiko kredit, *Random Forest*: Metode *ensemble* yang membangun banyak pohon keputusan untuk meningkatkan akurasi klasifikasi [20]. *XGBoost*: Algoritma *boosting* yang efisien dan efektif dalam menangani data tidak seimbang. *LightGBM*: Implementasi *gradient boosting* yang dirancang untuk kecepatan dan efisiensi. *TabNet*: Model *deep learning* yang dirancang khusus untuk data tabular [21].

3. HASIL DAN PEMBAHASAN

A. *Data Understanding*

Tahap *Data Understanding* bertujuan untuk memahami karakteristik data yang digunakan dalam penelitian, termasuk struktur, tipe variabel, serta makna dari setiap atribut yang tersedia. *Dataset* yang digunakan pada penelitian ini berisi informasi demografis, riwayat pekerjaan, kondisi keuangan, dan status gagal bayar nasabah dapat dilihat pada gambar



	age	ed	employ	address	income	debtinc	creddebt	othdebt	default
0	41	3	17	12	176	9.3	11.359392	5.008608	1.0
1	27	1	10	6	31	17.3	1.362202	4.000798	0.0
2	40	1	15	14	55	5.5	0.856075	2.168925	0.0
3	41	1	15	14	120	2.9	2.658720	0.821280	0.0
4	24	2	2	0	28	17.3	1.787436	3.056564	1.0

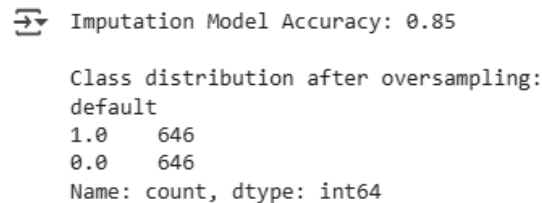
Jumlah data: 850 baris dan 9 kolom

Gambar.3 Tampilan Dataset Risiko Kredit

B. Data Preprocessing

Untuk memastikan kualitas data sebelum tahap pemodelan, dilakukan dua proses utama, yaitu imputasi nilai target yang hilang dan penyeimbangan distribusi kelas. Proses imputasi bertujuan untuk melengkapi data target yang kosong dengan memanfaatkan model regresi logistik, sedangkan penyeimbangan kelas dilakukan menggunakan teknik SMOTE (*Synthetic Minority Over-sampling Technique*) untuk menghindari bias model terhadap kelas mayoritas.

Gambar. 4 menampilkan akurasi model imputasi yang diperoleh serta distribusi kelas target setelah proses *oversampling*.

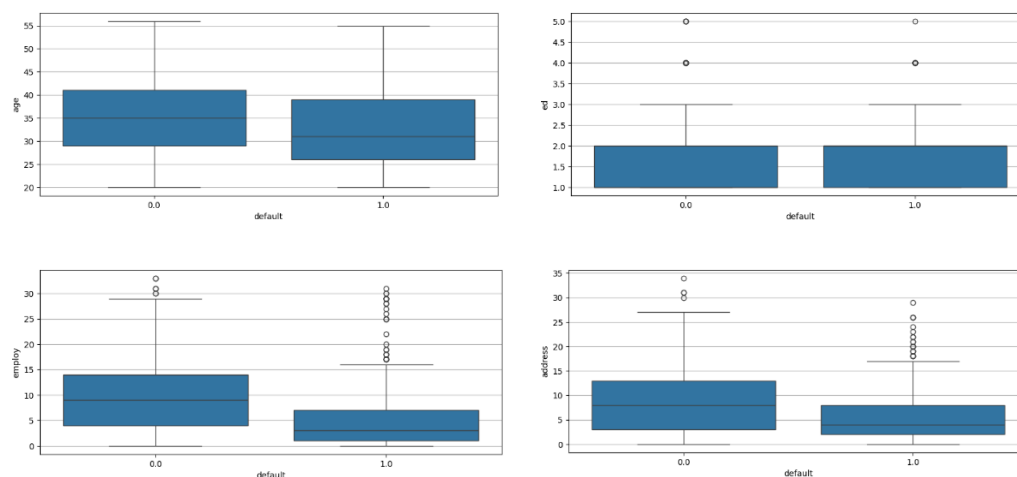


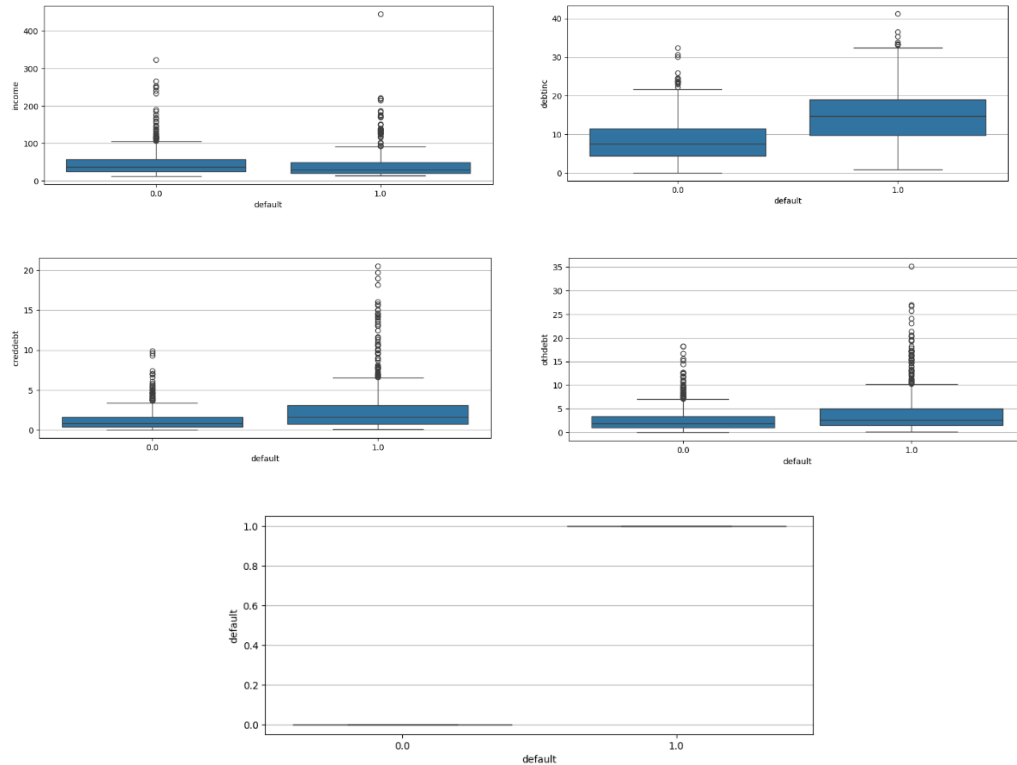
Gambar. 4 Hasil Imputasi Data dan Oversampling pada Dataset

Hasil evaluasi model imputasi menunjukkan nilai akurasi sebesar 0,85 pada data uji, yang mengindikasikan bahwa model regresi logistik yang digunakan mampu memprediksi nilai target yang hilang dengan tingkat ketepatan yang tinggi.

C. Outlier Detection and Explanation

Proses visualisasi data menggunakan *box plot* untuk mendeteksi keberadaan *outlier* pada setiap variabel numerik, dengan membandingkan distribusi nilai antara kelompok default dan non-default. Proses ini memanfaatkan perulangan untuk membuat *box plot* pada seluruh kolom yang terdapat dalam *df_combined*, dengan variabel default sebagai pembeda kelompok pada sumbu X dan variabel numerik sebagai sumbu Y.



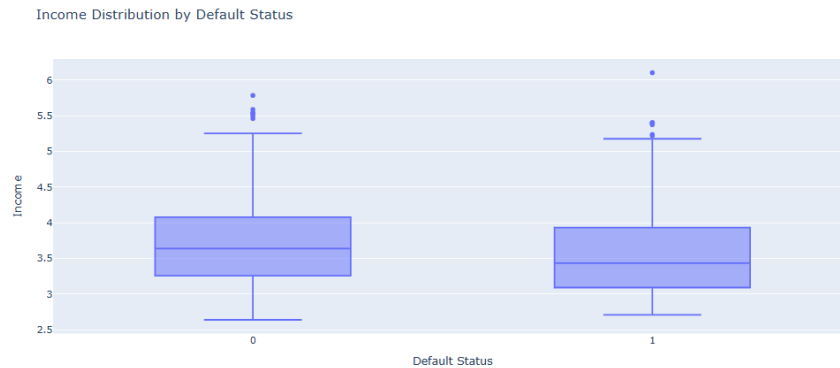


Gambar.5 Boxplot Variabel Prediktor terhadap Status Gagal Bayar untuk Analisis *Outlier*

Berdasarkan rangkaian boxplot, terlihat pada gambar IV.3, distribusi berbagai variabel terhadap status gagal bayar (*default*) dengan pola dan perbedaan yang menarik. Pada variabel *age* (umur), distribusi pada kedua kelompok status gagal bayar menunjukkan pola yang cukup mirip, namun terdapat perbedaan median, di mana kelompok *default* = 1 (gagal bayar) memiliki median umur sekitar 31 tahun, sedikit lebih rendah dibandingkan *default* = 0 (tidak gagal bayar) yang median umurnya sekitar 35 tahun. Rentang antar kuartil (IQR) pada kedua kelompok relatif sempit, menandakan konsistensi umur responden dalam setiap kategori, dan tidak tampak adanya outlier signifikan. Temuan ini mengindikasikan bahwa umur dapat menjadi salah satu faktor pembeda dalam risiko gagal bayar, di mana nasabah yang lebih muda cenderung memiliki proporsi gagal bayar lebih tinggi dibandingkan yang lebih tua.

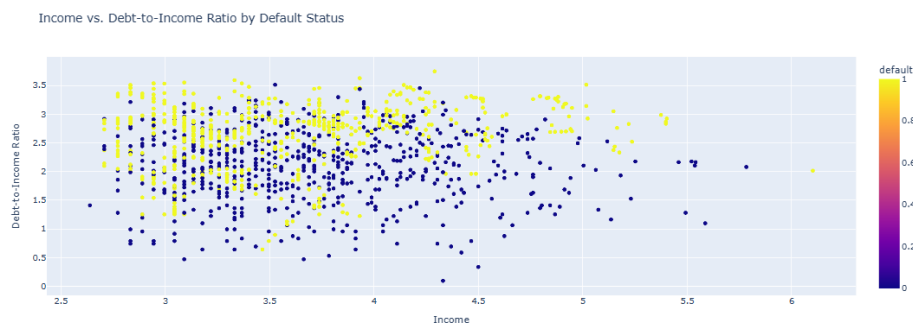
D. Data Visualization

Data visualisasi untuk mengetahui pola distribusi, hubungan antar variabel, dan perbedaan karakteristik antar kelompok pada dataset. Dari visualisasi data yang dilakukan dapat diketahui pola hubungan antara variabel. *Box plot* digunakan untuk membandingkan distribusi pendapatan antara kelompok nasabah yang *default* dan yang *non-default*. Grafik ini menampilkan median, kuartil, dan potensi pencilan pada masing-masing kelompok.



Gambar. 6 Distribusi Pendapatan Berdasarkan Status Gagal Bayar (Default Status)

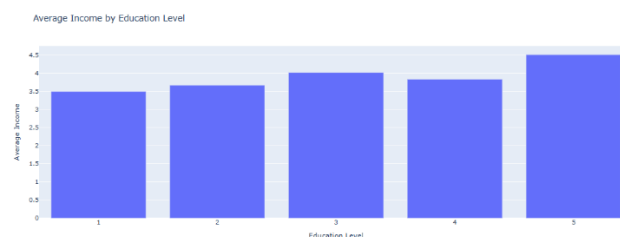
Hubungan antara pendapatan dengan rasio utang terhadap pendapatan. Pola sebaran titik dapat memberikan gambaran apakah terdapat korelasi tertentu di antara kedua variabel ini, serta mengidentifikasi nasabah dengan rasio utang yang ekstrem.



Gambar. 7 Hubungan Pendapatan dan Rasio Hutang terhadap Pendapatan Berdasarkan Status Gagal Bayar

Gambar.7 menunjukkan hubungan antara pendapatan (Income) dan rasio hutang terhadap pendapatan (Debt-to-Income Ratio) pada nasabah, dengan warna titik yang merepresentasikan status gagal bayar (default), di mana warna kuning menunjukkan gagal bayar (default = 1) dan biru menunjukkan tidak gagal bayar (default = 0).

Untuk mengetahui pola rata-rata pendapatan pada setiap tingkat pendidikan. Visualisasi data dilakukan dengan menggunakan grafik batang.



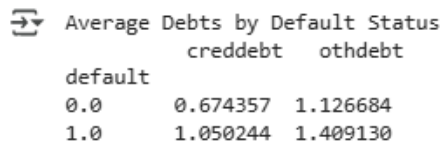
Gambar. 8 Hubungan Rata-rata Pendapatan Berdasarkan Tingkat Pendidikan

Gambar. 8 menggambarkan rata-rata pendapatan (*Average Income*) pada berbagai tingkat pendidikan (*Education Level*). Terlihat adanya tren peningkatan pendapatan seiring dengan bertambahnya tingkat pendidikan. Pendidikan level 1 memiliki rata-rata pendapatan terendah, sekitar 3,5, sementara level 5 memiliki rata-rata pendapatan tertinggi, sekitar 4,5. Meskipun terdapat sedikit fluktuasi pada level 3 dan 4, pola umumnya menunjukkan bahwa semakin tinggi tingkat pendidikan, semakin besar potensi pendapatan yang diperoleh. Hal ini mengindikasikan bahwa tingkat pendidikan berperan penting dalam meningkatkan peluang pendapatan yang lebih tinggi, yang pada gilirannya dapat berpengaruh positif terhadap kemampuan membayar kewajiban kredit.

E. Data Analysis

Tujuan data analysis dalam penelitian ini yaitu, untuk menjawab pertanyaan penelitian yang telah dirumuskan, dengan memanfaatkan data yang telah melalui proses pembersihan, penanganan *outlier*, dan visualisasi. Untuk menjawab pertanyaan Keterkaitan Jumlah Utang dengan Risiko *Default*

Pertanyaan ini penting untuk penilaian risiko kredit, yaitu apakah nasabah dengan jumlah utang yang lebih tinggi memiliki kecenderungan lebih besar untuk gagal bayar.



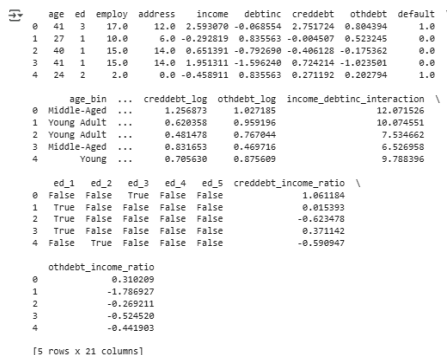
Average Debts by Default Status			
	creddebt	othdebt	
default			
0.0	0.674357	1.126684	
1.0	1.050244	1.409130	

Gambar. 9 Hasil Analisis Data Antara Jumlah Utang dengan Risiko *Default*

Data ini menunjukkan perbandingan rata-rata hutang kartu kredit (*creddebt*) dan hutang lainnya (*othdebt*) antara nasabah yang tidak gagal bayar (*default* = 0) dan yang gagal bayar (*default* = 1). Nasabah yang gagal bayar memiliki rata-rata *creddebt* sebesar 1,05, lebih tinggi dibandingkan dengan nasabah yang tidak gagal bayar yang hanya sebesar 0,67. Pola serupa juga terlihat pada *othdebt*, di mana kelompok gagal bayar memiliki rata-rata 1,41, sedangkan kelompok tidak gagal bayar hanya 1,13. Perbedaan ini mengindikasikan bahwa tingkat hutang yang lebih tinggi, baik pada kartu kredit maupun jenis hutang lainnya, berkorelasi dengan peningkatan risiko gagal bayar. Temuan ini menegaskan bahwa total beban hutang merupakan faktor penting yang perlu diperhatikan dalam penilaian risiko kredit.

F. Feature Engineering

Hasil dari proses *feature engineering* dilakukan untuk membuat, mengubah, atau menggabungkan fitur-fitur dalam dataset agar dapat meningkatkan kemampuan model dalam menangkap pola dan hubungan antar variabel.



	age	ed	employ	address	income	debtinc	creddebt	othdebt	default
0	41	3	17.0	12.0	2.593070	-0.068554	2.751724	0.804394	1.0
1	27	1	10.0	6.0	-0.292819	0.835563	-0.004507	0.522245	0.0
2	40	1	15.0	14.0	0.451391	-0.795098	-0.406128	-0.175362	0.0
3	41	1	15.0	14.0	1.951311	-1.596240	0.724214	-1.023501	0.0
4	24	2	2.0	0.0	-0.458911	0.835563	0.271192	0.202794	1.0

	age_bin	...	creddebt_log	othdebt_log	income_debtinc_interaction
0	Middle-Aged	...	1.256873	1.027185	12.071526
1	Young Adult	...	0.620358	0.959196	10.074551
2	Young Adult	...	0.481478	0.767944	7.534662
3	Middle-aged	...	0.831653	0.469716	6.526958
4	Young	...	0.705630	0.875609	9.788396

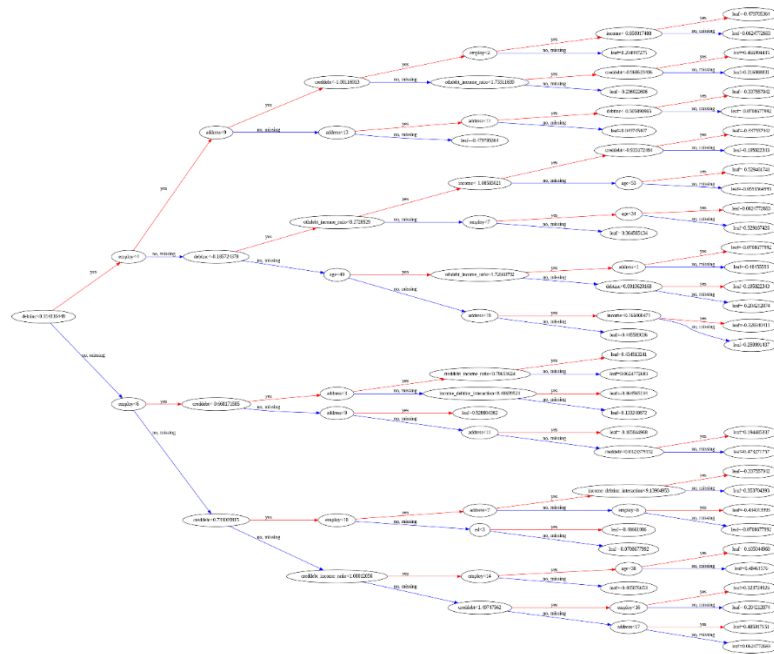
	ed_1	ed_2	ed_3	ed_4	ed_5	creddebt_income_ratio
0	False	False	True	False	False	1.061104
1	True	False	False	False	False	0.015393
2	True	False	False	False	False	-0.623478
3	True	False	False	False	False	0.371142
4	False	True	False	False	False	-0.590947

	othdebt_income_ratio
0	0.310209
1	-1.786927
2	-0.269211
3	-0.524520
4	-0.441903

[5 rows x 21 columns]

Gambar. 10 Data Nasabah dengan Variabel Asli dan Turunan

dilakukan pembangunan model klasifikasi untuk memprediksi risiko kredit nasabah menggunakan algoritma XGBoost *Classifier (Extreme Gradient Boosting)*. XGBoost merupakan salah satu algoritma ensemble learning berbasis pohon keputusan (*decision tree*) yang menggabungkan banyak model pohon secara bertahap, di mana setiap pohon baru memperbaiki kesalahan yang dibuat oleh pohon sebelumnya.



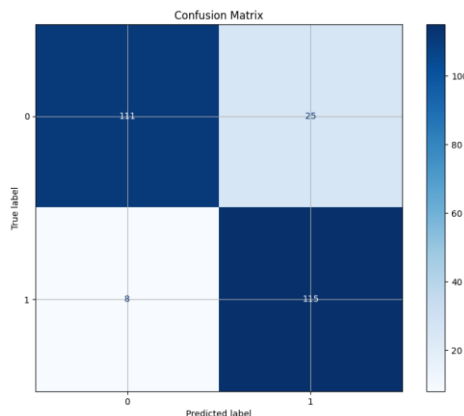
Gambar. 11 menunjukkan representasi decision tree yang digunakan untuk menganalisis dan mengklasifikasikan risiko kredit berdasarkan sejumlah variabel seperti debtinc (rasio utang terhadap pendapatan), employ (lama bekerja), creddebt (rasio hutang kartu kredit), address (lama tinggal di alamat), age (usia), dan income (pendapatan). Pohon keputusan ini bekerja dengan memecah data menjadi cabang-cabang berdasarkan nilai ambang batas tertentu pada setiap atribut, menghasilkan jalur yang berujung pada keputusan atau prediksi risiko kredit (ya = risiko tinggi, tidak = risiko rendah). Setiap simpul (node) menampilkan kriteria pemisahan, sedangkan cabang merah menunjukkan kondisi “ya” dan cabang biru menunjukkan kondisi “tidak atau data hilang”. Model seperti ini memudahkan interpretasi karena memberikan aturan yang jelas untuk mengklasifikasikan calon peminjam berdasarkan profil keuangannya.

Setelah pelatihan, model dievaluasi menggunakan data uji. Metrik yang digunakan meliputi *Accuracy*, *Precision*, *Recall*, dan *F1 Score*, serta laporan klasifikasi (*classification report*) yang memberikan gambaran detail performa pada masing-masing kelas. Evaluasi ini bertujuan menilai seberapa baik model dalam memprediksi risiko kredit. Setelah model dibuat, dilakukan evaluasi menggunakan test set. Model digunakan untuk memprediksi label, lalu dibandingkan dengan label asli. Hasilnya adalah sebagai berikut:

Accuracy: 0.87				
Precision: 0.82				
Recall: 0.93				
F1 Score: 0.87				
	precision	recall	f1-score	support
0.0	0.93	0.82	0.87	136
1.0	0.82	0.93	0.87	123
accuracy			0.87	259
macro avg	0.88	0.88	0.87	259
weighted avg	0.88	0.87	0.87	259

Gambar. 12 Output Model dalam Melakukan Learning Data dan Uji Model

Dari gambar. 12 menunjukkan hasil evaluasi model menunjukkan bahwa tingkat akurasi keseluruhan mencapai 87%, artinya dari 259 data yang diuji, sekitar 87% prediksi sesuai dengan kondisi sebenarnya. Untuk kelas 0.0 (tidak berisiko), model memiliki precision sebesar 0.93, yang berarti 93% dari prediksi “tidak berisiko” benar adanya, dan recall sebesar 0.82, yang berarti model mampu mengenali 82% dari seluruh data tidak berisiko. Sebaliknya, untuk kelas 1.0 (berisiko), nilai precision adalah 0.82, sehingga 82% dari prediksi “berisiko” benar, dan recall sebesar 0.93, yang menunjukkan model berhasil mendeteksi 93% dari seluruh nasabah berisiko. Nilai *F1-score* untuk kedua kelas sama-sama berada di angka 0.87, menandakan keseimbangan yang baik antara precision dan recall. Nilai macro average dan weighted average yang hampir sama (sekitar 0.88 untuk precision dan recall) menunjukkan bahwa model tidak terlalu bias terhadap salah satu kelas, meskipun jumlah data pada masing-masing kelas berbeda. Secara keseluruhan, performa ini mengindikasikan bahwa model cukup andal dalam mengklasifikasikan risiko kredit dengan keseimbangan yang baik antara deteksi risiko dan ketepatan prediksi.

Gambar. 13 Diagram *Confusion Matrix*

Gambar. 13 menggambarkan kinerja model dalam mengklasifikasikan risiko kredit ke dalam dua kelas, yaitu 0 (tidak berisiko) dan 1 (berisiko). Pada kotak kiri atas terdapat 111 data yang merupakan *True Negative (TN)*, yaitu data nasabah yang sebenarnya tidak berisiko dan berhasil diprediksi dengan benar sebagai tidak berisiko. Kotak kanan atas menunjukkan 25 data *False Positive (FP)*, yaitu data nasabah yang sebenarnya tidak berisiko tetapi diprediksi model sebagai berisiko. Kotak kiri bawah berisi 9 data *False Negative (FN)*, yaitu nasabah yang sebenarnya berisiko namun diprediksi tidak berisiko. Sementara itu, kotak kanan bawah berisi 114 data *True Positive (TP)*, yaitu nasabah yang benar-benar berisiko dan diprediksi dengan tepat sebagai berisiko. Dari pola ini, terlihat bahwa model memiliki keseimbangan yang cukup baik antara kemampuan mendeteksi nasabah berisiko (Recall

tinggi) dan ketepatan prediksi risiko (Precision tinggi), yang tercermin dari nilai akurasi keseluruhan sebesar 87%.

4. KESIMPULAN

Berdasarkan hasil analisis dan klasifikasi risiko kredit berbasis data mining menggunakan algoritma XGBoost, diperoleh nilai akurasi sebesar 87%, presisi 82%, recall 93%, dan F1-Score 87%. Nilai akurasi yang cukup tinggi ini menunjukkan bahwa model mampu memprediksi risiko kredit dengan tingkat ketepatan yang baik. Tingginya nilai recall, terutama untuk kategori nasabah berisiko tinggi, mengindikasikan bahwa model efektif dalam mengidentifikasi calon debitur yang berpotensi gagal bayar, sehingga dapat membantu lembaga keuangan meminimalisasi kerugian. Namun, nilai presisi yang sedikit lebih rendah dibanding recall mengindikasikan adanya kemungkinan terjadinya false positive, yaitu nasabah yang sebenarnya layak mendapatkan kredit namun diklasifikasikan sebagai berisiko. Hal ini perlu menjadi perhatian karena dapat berdampak pada peluang bisnis yang hilang. Mengingat profil risiko nasabah dan kondisi ekonomi dapat berubah seiring waktu, evaluasi ulang model secara berkala diperlukan untuk memastikan performa tetap optimal. Model ini dapat diintegrasikan ke dalam *credit scoring system* lembaga keuangan, di mana hasil prediksi risiko kredit digunakan sebagai salah satu indikator untuk menyetujui atau menolak pengajuan kredit. Lakukan analisis terhadap nasabah yang termasuk dalam kategori FP dan FN untuk mengetahui pola kesalahan *model*. Hasil analisis ini dapat digunakan untuk perbaikan seleksi fitur dan *preprocessing data*. Gunakan *k-fold cross-validation* agar evaluasi model tidak bergantung pada satu set data uji saja, sehingga hasilnya lebih mencerminkan kemampuan model di berbagai skenario data.

UCAPAN TERIMA KASIH

Puji syukur penulis ucapkan kehadiran Allah SWT yang telah memberikan rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan penelitian ini yang merupakan syarat untuk mendapatkan gelar sarjana Komputer, Fakultas Teknik dan Ilmu Komputer, Universitas Potensi Utama. Dan mengucapkan terimakasih yang sebesar-besarnya kepada Kaprodi Sistem Informasi Ibu Mas Ayoe Elhias Nst, M.Kom., dan Universitas Potensi Utama Medan yang membantu dalam penerbitan karya ilmiah ini.

REFERENSI

- [1] Zhu, M., et al. (2024). Credit Default Prediction Using Ensemble Learning Methods. arXiv:2402.17979. <https://arxiv.org/abs/2402.17979>
- [2] Marco, Y., et al. (2024). Klasifikasi Risiko Kredit Menggunakan Algoritma SVM dan Naïve Bayes. Universitas Brawijaya. <https://repository.ub.ac.id/id/eprint/234851>
- [3] Rohmah, R., & Kurniawati, Y. (2024). Analisis Risiko Kredit Menggunakan C4.5 di RapidMiner. Edustem Journal. <https://e-journal.ivet.ac.id/index.php/EDUSTEM/article/view/3493>
- [4] Wang, Y., et al. (2024). Credit Default Prediction with CNN and Transformer. arXiv:2412.18222. <https://arxiv.org/abs/2412.18222>
- [5] Wang, Z., Jiang, C., Ding, Y., Lyu, X., & Liu, Y. (2018). A novel behavioral scoring model for estimating probability of default over time in peer-to-peer lending. *Electronic Commerce Research and Applications*, 27, 74-82.
- [6] Sari, M. M. (2025). Analisis Klasifikasi Risiko Kredit dengan Decision Tree C4.5. *Jurnal Multiple*. <https://journal.institercom-edu.org/index.php/multiple/article/view/840>

- [7] Yu, C., Jin, Y., Xing, Q., Zhang, Y., Guo, S., & Meng, S. (2024, July). Advanced user credit risk prediction model using lightgbm, xgboost and tabnet with smoteenn. In 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS) (pp. 876-883). IEEE.
- [8] SMITI, Salima, SOUI, Makram, et GHEDIRA, Khaled. Tri-XGBoost model improved by BLSmote-ENN: an interpretable semi-supervised approach for addressing bankruptcy prediction. *Knowledge and information systems*, 2024, vol. 66, no 7, p. 3883-3920.
- [9] Zedda, S. (2024). Credit scoring: Does XGboost outperform logistic regression? A test on Italian SMEs. *Research in International Business and Finance*, 70, 102397.
- [10] Yu, C., Jin, Y., Xing, Q., Zhang, Y., Guo, S., & Meng, S. (2024, July). Advanced user credit risk prediction model using lightgbm, xgboost and tabnet with smoteenn. In 2024 IEEE 6th International Conference on Power, Intelligent Computing and Systems (ICPICS) (pp. 876-883). IEEE.
- [11] Hamim, M., El Mouden, I., Ouzir, M., Moutachauik, H., & Hain, M. (2021). A novel dimensionality reduction approach to improve microarray data classification. *IIUM Engineering Journal*, 22(1), 1-22.
- [12] Mushava, J., & Murray, M. (2022). A novel XGBoost extension for credit scoring class-imbalanced data combining a generalized extreme value link and a modified focal loss function. *Expert Systems with Applications*, 202, 117233.
- [13] Petropoulos, A., Siakoulis, V., Stavroulakis, E., & Klamargias, A. (2019). A robust machine learning approach for credit risk analysis of large loan level datasets using deep learning and extreme gradient boosting. *IFC Bulletins chapters*, 49.
- [14] Xing, Q., Yu, C., Huang, S., Zheng, Q., Mu, X., & Sun, M. (2024). Enhanced credit score prediction using ensemble deep learning model. *arXiv preprint arXiv:2410.00256*.
- [15] Kim, H. H., & Swanson, N. (2013). Mining big data using parsimonious factor and shrinkage methods (No. 2013-16). Working paper.
- [16] Sharma, R., Schommer, C., & Vivarelli, N. (2020, October). Building up explainability in multi-layer perceptrons for credit risk modeling. In 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA) (pp. 761-762). IEEE.
- [17] Ha, V. S., Lu, D. N., Choi, G. S., Nguyen, H. N., & Yoon, B. (2019, February). Improving credit risk prediction in online peer-to-peer (P2P) lending using feature selection with deep learning. In 2019 21st International Conference on Advanced Communication Technology (ICACT) (pp. 511-515). IEEE.
- [18] Nimmala, R. (2022). Harnessing Machine Learning for Comparative Analysis of Bottom-Up and Top-Down Approaches in Climate Credit Risk Modeling. *Journal of Technological Innovations*, 3(2).
- [19] Arora, N., & Kaur, P. D. (2020). A Bolasso based consistent feature selection enabled random forest classification algorithm: An application to credit risk assessment. *Applied Soft Computing*, 86, 105936.
- [20] Saavedra, C. A. P. B., da Costa, J. P. V., Diniz, F. B., Albuquerque, P. H. M., & Kimura, H. (2024). Comparing lifetime estimates of probability of default for refinancing operations with survival analysis and ensemble methods. *Communications in Statistics: Case Studies, Data Analysis and Applications*, 10(1), 1-26.
- [21] Kim, H. H. (2022). A dynamic analysis of household debt using a self-organizing map. *Empirical Economics*, 62(6), 2893-2919.