IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 4, No. 2 Juli 2025
https://subset.id/index.php/IJCSR

Komparasi XGBoost dan C4.5 dalam Klasifikasi Risiko Kredit untuk Kelayakan Pinjaman di India

Rastra Ardiansyah Pora[1 b)], Dina Maulina*[2 a)], Ninik Trihartanti[3c)]

[1]Informatika, [2] Manajemen Informatika, [3] Sistem Informasi
Fakutas Ilmu Komputer
Universitas Amikom Yogyakarta, Indonesia

Author Emails
[a)] Corresponding author: dina.m@amikom.ac.id
[b)]rastraapora@students.amikom.ac.id
[c)]ninik.t@amikom.ac.id

Abstract. The digital transformation has driven significant progress in the financial sector, particularly in credit and lending services. Although application processes have become more accessible, credit risk remains a major challenge that must be effectively managed. This study aims to compare the performance of the XGBoost and C4.5 algorithms in predicting and classifying credit risk to determine loan eligibility. The methodology follows the stages of Knowledge Discovery in Data (KDD), including data selection, preprocessing, transformation, and model development. The evaluation was conducted using 10-Fold Cross Validation with varying data ratios (90:10, 80:20, 70:30, and 60:40). Algorithm performance was assessed using accuracy, precision, recall, and F1-score metrics. The results indicate that the C4.5 algorithm outperforms XGBoost, both before and after hyperparameter optimization using GridSearchCV. At the 80:20 data split, C4.5 achieved an accuracy of 96.20% with an AUC of 0.9730. These findings suggest that C4.5 is more effective in handling large-scale data and provides efficient classification outcomes, while XGBoost remains a competitive alternative after optimization. This study offers strategic recommendations for financial institutions to enhance the accuracy of credit risk analysis based on applicant data.

Keywords **:** *XGBoost, C4.5, Credit Risk, Loan Eligibility, Machine Learning, Classification, Hyperparameter Optimization*

Abstraksi. Transformasi digital telah mendorong kemajuan signifikan di sektor keuangan, khususnya dalam layanan kredit dan pinjaman. Meskipun proses pengajuan kini semakin mudah, risiko kredit tetap menjadi tantangan utama yang harus diantisipasi secara efektif. Penelitian ini bertujuan untuk membandingkan kinerja algoritma XGBoost dan C4.5 dalam memprediksi dan mengklasifikasikan risiko kredit guna menentukan kelayakan pinjaman. Metode yang digunakan mengikuti tahapan Knowledge Discovery in Data (KDD), meliputi seleksi data, pra-pemrosesan, transformasi, dan pembentukan model. Pengujian dilakukan menggunakan 10-Fold Cross Validation dengan variasi rasio data melalui empat pengujian. Evaluasi kinerja algoritma didasarkan pada metrik akurasi, presisi, recall, dan F1-score. Hasil penelitian menunjukkan bahwa algoritma C4.5 unggul dibandingkan XGBoost, baik sebelum maupun setelah optimasi hyperparameter menggunakan GridSearchCV. Pada rasio 80:20, C4.5 mencatat akurasi 96,20% dengan AUC sebesar 0,9730. Temuan ini menyimpulkan bahwa C4.5 lebih efektif dalam menangani data berukuran besar dan menghasilkan klasifikasi yang efisien, sementara XGBoost tetap merupakan alternatif kompetitif pasca optimasi. Studi ini memberikan rekomendasi strategis bagi lembaga keuangan dalam meningkatkan akurasi analisis risiko kredit berbasis data.

Kata Kunci : XGBoost, C4.5, Risiko Kredit, Kelayakan Pinjaman, Machine Learning, Klasifikasi, Optimasi Hyperparameter

IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 4, No. 2 Juli 2025
https://subset.id/index.php/IJCSR

PENDAHULUAN

Pada era digitalisasi yang semakin pesat, sektor keuangan mengalami perkembangan signifikan, terutama dalam hal kredit dan pinjaman. Layanan pinjaman kini tidak hanya terbatas pada lembaga keuangan tradisional, tetapi juga meluas ke platform *fintech* yang menawarkan kemudahan akses bagi pemohon. Walaupun prosesnya mudah, risiko kredit masih menjadi tantangan yang perlu diantisipasi dengan baik. Risiko kredit merupakan ketidakmampuan yang dialami oleh individu, perusahaan, institusi, atau lembaga dalam memenuhi kewajiban yang telah disepakati bersama sesuai aturan yang berlaku, di mana kewajiban tersebut tidak diselesaikan tepat waktu, baik sebelum maupun sesudah tenggat waktu yang ditentukan[1].

Seiring dengan perkembangan yang pesat, peningkatan risiko kredit menjadi hal yang tidak terhindarkan. Menurut data Bank Indonesia, total kredit yang disalurkan oleh perbankan mulai meningkat, dengan nilai kredit mencapai Rp 4.246,6 triliun pada Oktober 2016, yang menunjukkan pertumbuhan sebesar 7,4% dibandingkan dengan tahun 2015[2]. Sedangkan menurut data Otoritas Jasa Keuangan (OJK), *outstanding* pinjaman online dari Maret 2022 hingga Maret 2023 menunjukkan kredit tidak lancar (30-90 hari) meningkat Rp3,58 triliun (71,13% YoY), mayoritas dari pinjaman perseorangan sebesar Rp3,3 triliun dan sisanya dari badan usaha, sementara kredit macet di atas 90 hari pada Maret 2023 mencapai Rp1,43 triliun dari total pinjaman yang belum terbayar[3]. Saat ini, metode tradisional dalam menilai kelayakan kredit sering tidak mampu menangkap kompleksitas pola pengajuan pinjaman yang dinamis[4]. Pendekatan berbasis data dan machine learning diharapkan dapat memberikan akurasi lebih tinggi dalam analisis risiko kredit. Algoritma seperti *XGBoost* dan *C4.5* menjadi pilihan utama dalam membangun model prediksi yang lebih akurat dan andal dalam mengelompokkan data lama dan baru[3]. *XGBoost* dikenal efektif dalam menanggulangi kasus *machine learning* berskala besar dan sangat baik dalam menganalisis data, sementara algoritma *C4.5* menawarkan kemudahan interpretasi hasil[5].

Penelitian ini bertujuan untuk membantu lembaga keuangan, baik perbankan maupun *fintech*, dalam memilih algoritma yang paling efektif dan efisien untuk memprediksi dengan mengklasifikasikan risiko kredit, sehingga dapat mengurangi potensi kerugian akibat debitur yang gagal membayar pinjaman. Diharapkan dengan pemilihan algoritma yang tepat, lembaga keuangan dapat meningkatkan ketepatan dalam proses pengambilan keputusan terkait pemberian pinjaman, yang pada akhirnya dapat meningkatkan profitabilitas dan mengurangi risiko operasional. Selain itu penelitian ini juga bertujuan memberikan wawasan bagi pengembang aplikasi atau sistem terkait yang membutuhkan analisis risiko kredit berbasis *machine learning*, sehingga dapat diimplementasikan dalam sistem penilaian kredit mereka. Batasan dalam penelitian ini yaitu data yang digunakan dalam penelitian ini terbatas pada dataset pemohon pinjaman di negara India, meliputi variabel-variabel seperti usia, pengalaman kerja, status perkawinan, kepemilikan rumah, penghasilan tahunan dan lain-lain. Pemilihan negara India sebagai dataset dalam penelitian ini didasarkan pada beberapa pertimbangan ilmiah dan praktis. Dataset tersebut bersifat terbuka (*open access*) dan tersedia di *platform* seperti *Kaggle*, dengan struktur data yang lengkap dan siap untuk dianalisis, sehingga sangat mendukung prinsip transparansi, replikasi, dan validasi dalam penelitian ilmiah. Selain itu, dataset ini memuat atribut-atribut yang umum digunakan dalam analisis kelayakan pinjaman seperti usia, penghasilan, status pernikahan, pengalaman kerja, dan kepemilikan aset yang sifatnya cukup universal dan relevan dalam membangun model klasifikasi risiko.

Algoritma menggunakan split data dengan pembagian nilai *split* rasio sebesar 90:10, 80:20, 70:30, dan 60:40. Penelitian berfokus pada perbandingan pengujian kinerja model berdasarkan akurasi *recall*, *F1-score*, dan *confusion matrix*.

TINJAUAN PUSTAKA

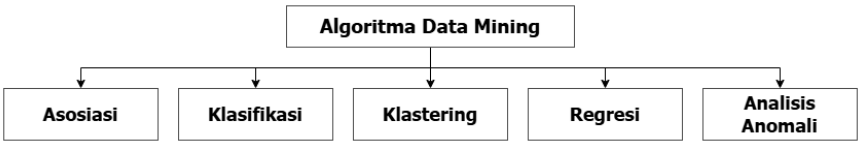
Penelitian ini mengacu pada berbagai literatur akademik dan studi terdahulu yang berkaitan dengan penerapan algoritma *C4.5* dan *XGBoost*. Sejumlah penelitian sebelumnya dijadikan referensi untuk mendukung analisis perbandingan kinerja kedua algoritma dalam mengklasifikasikan risiko kredit. Eksplorasi penggunaan algoritma *XGBoost* dalam menganalisis risiko kredit sektor keuangan. Menggunakan metode KDD dan data dari *German Credit Dataset* yang telah diseimbangkan dengan *SMOTE, XGBoost* menunjukkan performa tinggi dengan akurasi 83% dan AUC 0,918. Temuan ini menegaskan efektivitas *XGBoost* dalam menangani masalah kredit bermasalah akibat seleksi pelanggan yang kurang akurat[2]. Penerapan algoritma C4.5 untuk memprediksi risiko kredit di PT. WOM Finance.

IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 4, No. 2 Juli 2025
https://subset.id/index.php/IJCSR

Dengan data pelanggan dan enam atribut utama, C4.5 mampu mengklasifikasikan nasabah ke dalam kategori "lancar" atau "tidak lancar" dengan akurasi 79%. Atribut jumlah pinjaman menjadi penentu utama dalam pembentukan pohon keputusan. Proses dilakukan dengan Weka dan menghasilkan model yang bermanfaat bagi lembaga pembiayaan dalam menilai risiko nasabah[3]. Perbandingan *XGBoost* dan *Random Forest* dalam klasifikasi keputusan kredit menggunakan data dari *Kaggle. XGBoost* unggul dengan skor akurasi, presisi, *recall,* dan *F1-score* sempurna (1.0) pada dua ukuran dataset. *Random Forest* meski tinggi akurasi, memiliki *F1-score* lebih rendah pada dataset kecil. Penelitian menyarankan teknik *oversampling/undersampling* untuk data tidak seimbang dan menegaskan keandalan *XGBoost* dalam klasifikasi kredit[4]. Fokus Perbandingan *XGBoost* dan *AdaBoost* untuk analisis risiko kredit di *platform P2P Lending*. Setelah reduksi fitur dari 136 menjadi 34, evaluasi menunjukkan *XGBoost* (AUC 0,92) lebih unggul dari *AdaBoost* (AUC 0,89). Studi ini menyoroti pentingnya pemanfaatan *machine learning* dalam mempercepat dan mengamankan proses persetujuan pinjaman di *fintech*[5]. Optimalisasi algoritma C4.5 dengan *Forward Selection* dan *Stratified Sampling* untuk prediksi kelayakan kredit. Dengan *RapidMiner* sebagai alat bantu, akurasi meningkat menjadi 79,11%, lebih baik dibandingkan tanpa optimasi. Penelitian ini menunjukkan bahwa pemilihan fitur dan sampling yang tepat dapat meningkatkan performa model prediktif dalam sistem penilaian kredit[6]. Sedangkan dalam penelitian yang dilakukan penulis disini fokus pada membandingkan kinerja *XGBost* dengan C.4.5, memberikan perspektif berbeda dalam konteks algoritma. Dataset yang digunakan mengambil dari kaggle untuk kelayakan pinjaman. Prediksi dengan mengklasifikasikan kelas 0 dan 1 (Layak, tidak layak) untuk melihat resiko kredit yang dihasilkan dari masing-masing algoritma. Penelitian ini menggunakan *oversampling, Hyperparameter tuningGridSearchCV*, dalam menganalisis data dengan menggunakan *Phyton* dan *Tool Google Colab.*

Kreditur adalah pihak yang memberikan kredit atau pinjaman, baik kepada individu, perusahaan, maupun pemerintah, dalam konteks perbankan[7]. Pemberi pinjaman biasanya berasal dari bank atau lembaga keuangan non-bank, seperti perusahaan kartu kredit maupun investor pribadi. Tujuan kreditur menawarkan pinjaman untuk mendapatkan keuntungan berupa bunga atau imbal hasil dari dana yang dipinjamkan. Risiko kredit merupakan ketidakmampuan yang dialami oleh individu, perusahaan, institusi, atau lembaga dalam memenuhi kewajiban yang telah disepakati bersama sesuai aturan yang berlaku, di mana kewajiban tersebut tidak diselesaikan tepat waktu, baik sebelum maupun sesudah tenggat waktu yang ditentukan[8]. Untuk mengurangi risiko kredit ini, kreditur sering kali melakukan analisis profil seperti diversifikasi portofolio pinjaman dan peningkatan pengawasan pinjaman yang ketat pada pemohon yang berpotensi mengalami kesulitan keuangan, guna mengurangi risiko kerugian finansial[9]

Data mining adalah proses analisis data berskala besar untuk menemukan pola tersembunyi yang relevan dan berguna, dengan bantuan teknik statistik dan komputasi. Tujuannya adalah menghasilkan informasi yang dapat mendukung pengambilan keputusan strategis dalam konteks bisnis[10]. Data mining memiliki karakteristik utama, yaitu mampu mengungkap pola tersembunyi, bekerja secara efektif pada data berukuran besar, dan mendukung pengambilan keputusan strategis[11]. Jenis-jenis algoritma dalam data mining yang sering diterapkan untuk menyelesaikan berbagai permasalahan dan menggali pengetahuan baru dapat dikelompokkan ke dalam beberapa kategori pada gambar 1[12]:



GAMBAR 1 Jenis-jenis Algoritma Data Mining

Knowledge Discovery in Database (KDD) merupakan alur yang harus dilalui untuk menemukan informasi dan pola-pola yang berguna dalam data[13]. Dalam proses tersebut terdapat serangkaian tahapan yang berfungsi untuk mengolah data mentah menjadi informasi yang bernilai. Pohon keputusan merupakan salah satu metode klasifikasi yang terkenal karena kemudahannya dalam diinterpretasi oleh manusia [14]. Pohon keputusan juga adalah model klasifikasi yang dibuat menggunakan karakteristik pohon. Caranya, membagi kelompok besar menjadi kecil dengan variabel target lebih banyak disimpan data pada tabel tersebut dan diubah menjadi model pohon[15]. Metode pohon keputusan memiliki kelebihan berupa kemudahan pemahaman, visualisasi skenario yang menyeluruh, serta efisiensi

IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 4, No. 2 Juli 2025
https://subset.id/index.php/IJCSR

waktu dalam pembuatan dan penyelesaiannya[16]. Algoritma C4.5 adalah algoritma yang digunakan untuk membentuk pohon keputusan, di mana pohon keputusan merupakan metode yang efektif untuk melakukan prediksi atau klasifikasi[17]. Dilakukan dengan membagi kumpulan data besar menjadi kelompok-kelompok data yang lebih kecil melalui serangkaian aturan keputusan. Algoritma C4.5 mengkonstruksi pohon keputusan dari data yang dilatih, yang berupa kasus-kasus dalam basis data[7]. XGBoost (*Extreme Gradiant Boosting*) adalah *tree boosting* yang dapat diskalakan yang telah terbukti sebagai metode machine learning yang sangat efektif dan banyak digunakan untuk melakukan prediksi nilai suatu data. XGBoost memanfaatkan pohon keputusan sebagai model utama dan menggunakan teknik boosting untuk meningkatkan performa model secara keseluruhan.

METODE PENELITIAN

Dataset yang digunakan dalam penelitian ini adalah dataset publik milik Yamin Hossain yang berjudul "Applicant Details for Loan Approval in India". Dataset ini dapat diunduh melalui situs Kaggle dengan nama “Yamin Hossain” dan dipilih karena memiliki fitur yang lengkap dan relevan dengan objek penelitian, seperti informasi usia, pendapatan tahunan, pengalaman kerja, status pernikahan, kepemilikan rumah, dan risiko gagal bayar pinjaman. Dataset ini cocok digunakan dalam pemodelan kelayakan pinjaman karena mencakup atribut-atribut penting untuk analisis risiko kredit. Dataset ini memiliki sekitar 100.000 data dengan ukuran file 7.55 MB dalam format *comma-separated values* (CSV).

Pengambilan dataset

Dataset yang digunakan dalam penelitian ini adalah dataset publik milik Yamin Hossain yang berjudul "*Applicant Details for Loan Approval*" yang berisi informasi mengenai pemohon pinjaman di India. Dataset ini dapat diunduh melalui situs Kaggle dengan nama “Yamin Hossain”, data ini dipilih karena memiliki atribut yang lengkap dan relevan dengan objek penelitian, seperti informasi usia, pendapatan tahunan, pengalaman kerja, status pernikahan, kepemilikan rumah, dan risiko gagal bayar pinjaman. Dataset ini cocok digunakan dalam pemodelan klasifikasi untuk prediksi kelayakan pinjaman karena mencakup atribut-atribut penting untuk analisis risiko kredit.

Data preprocessing dan transformation

Proses *preprocessing* mencakup pembersihan data, seperti mengisi nilai yang hilang, perubahan nama, serta memperbaiki data yang tidak konsisten atau tidak valid[18]. Sementara itu, data *transformation* bertujuan untuk mengubah data ke format yang sesuai agar bisa diproses oleh model [9]. Seperti mengubah data kategori menjadi angka melalui encoding dan menormalisasi data numerik agar lebih mudah diproses oleh model, kedua prosesnya dilakukan menggunakan tool.

Pemodelan klasifikasi algoritma XGBoost dan C4.5

Klasifikasi merupakan salah satu metode dalam data mining yang digunakan untuk mengelompokkan data ke dalam kategori atau kelas berdasarkan atribut tertentu. Secara garis besar, klasifikasi bertujuan untuk memprediksi kelas atau label dari data yang belum diketahui, dengan memanfaatkan informasi yang diperoleh dari data yang telah diklasifikasikan sebelumnya[10].
Dalam tahap pemodelan dilakukan terlebih dahulu data *splitting* yang merupakan pembagian data train dan data test dalam rasio sebelum dilakukannya pemodelan algortima[11].Penelitian ini menggunakan dua algoritma, yaitu *XGBoost* dan *C4.5*, yang masing-masing diterapkan dengan pendekatan *Ensemble Machine Learning* dan *Decision Tree.* Berikut penjelasan dari dua algoritma yang dipakai dan pendekatannya:

a. *XGBoost (Ensemble Machine Learning)*
   Konsep dasar dari *ensemble learning* adalah gabungan atau kombinasi dari beberapa model yang baik dalam melakukan prediksi atau klasifikasi yang dapat menghasilkan output yang lebih baik dibandingkan dengan menggunakan satu model basis tetap[12]. Algoritma *XGBoost* adalah model yang menggabungkan beberapa model *tree ensemble* untuk melakukan optimasi model yang kuat[13].

IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 4, No. 2 Juli 2025
https://subset.id/index.php/IJCSR

b. *C4.5 (Decision Tree)*
Konsep dasar dari *decision tree* adalah seperangkat aturan yang membagi data populasi besar dan heterogen menjadi populasi lebih kecil dengan variabel target yang lebih dominan, di mana data dalam pohon keputusan biasanya disajikan dalam format tabel berupa atribut dan struktur data[19]. Algoritma *C4.5* menggunakan teknik tersebut dengan elemen seperti *Entropy* dan *Gain*[20].

K-Fold Cross Validation (10-fold)

K-Fold Cross Validation adalah metode evaluasi model dalam *Machine Learning* yang membagi dataset menjadi k kelompok (*fold*) dengan ukuran yang sama, pada setiap iterasi, k-1 *fold* digunakan untuk pelatihan (*training*) dan 1 *fold* untuk validasi. Proses ini berulang hingga semua *fold* digunakan sebagai validasi, lalu rata-rata hasil evaluasi dihitung untuk menilai kinerja model secara keseluruhan[21]. Penelitian ini menerapkan metode *cross validation* 10 *fold* untuk mengevaluasi model, dengan membagi dataset menjadi sepuluh kelompok. Digunakannya teknik ini karena banyak diterapkan peneliti karena didapati mengurangi bias yang didapatkan didalam pengambilan sebuah sampel[14].

Hyperparameter tuning menggunakan GridSearchCV (10 fold)

Hyperparameter berfungsi untuk mengatur berbagai aspek dalam algoritma machine learning dan dapat memberikan dampak yang signifikan terhadap model yang dihasilkan serta performanya[15]. Di penelitian ini akan menggunakan *GridSearchCV* dengan 10 *fold* untuk memaksimalkan dan juga meningkatkan akurasi dan keandalam model. Metode *hyperparameter* tuning ini memungkinkan untuk melakukan pemindaian pada sejumlah *hyperparameter* yang dipilih. Pada *GridSearchCV*, berbagai kombinasi *hyperparameter* diterapkan pada model, dan kinerja setiap kombinasi di evaluasi dengan menggunakan validasi silang. Kombinasi hyperparameter dengan performa terbaik akan dipilih sebagai *hyperparameter* terbaik untuk model[16]. Pada Tabel 1 merupakan parameter-parameter dan jangkauan nilai dari *XGBoost* dan *C4.5* yang akan dipakai untuk mencari rekomendasi yang terbaik.

TABEL 1. *Hyperparameter* dan nilai jangkauan (*XGBoost* dan *C4.5*)

Algoritma	Parameter dan Nilai Jangkauan
XGBoost	*max_depth* = 3, 5, 7 *learning_rate* = 0.01, 0.1, 0.2 *n_estimators* = 100, 200, 300 *subsample* = 0.5, 0.7, 1.0
C4.5	*max_depth* = None, 3, 5, 7, 10, 15, 20 *min_samples_split* = 2, 5, 10, 15, 20 *min_samples_leaf* = 1, 2, 4, 6, 10

Evaluasi hasil

Setelah tahap pemodelan klasifikasi pada algoritma *XGBoost* dan *C4.5* selesai akan masuk ke tahap berikutnya yaitu evaluasi hasil. Dalam tahap evaluasi hasil model, dilakukan pengukuran menggunakan beberapa metrik evaluasi yang mencakup akurasi, presisi, *recall* (sensitivitas), dan *F1-Score*. Setelah mendefinisikan metrik evaluasi, model diuji menggunakan data pengujian yang telah dipisahkan sebelumnya[17].

IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 4, No. 2 Juli 2025
https://subset.id/index.php/IJCSR

HASIL DAN PEMBAHASAN

Pengambilan dataset

Pengambilan dataset "*Applicant_Details*" yang digunakan memiliki jumlah 100.000 data dengan ukuran file 7.55 MB dalam format *csv* yang terdiri dari 13 atribut, dimana 12 atribut yaitu *Applicant_ID, Annual_Income, Applicant_Age, Work_Experience, Marital_Status, House_Ownership, Vehicle_Ownership(car), Occupation, Residence_City, Residence_State, Years_in_Current_Employment,* dan *Years_in_Current_Residence* yang merupakan variabel independen dengan 1 atribut yaitu *Loan_Default_Risk* yang merupakan variabel dependen yang akan digunakan untuk perbandingan pemodelan klasifikasi algoritma *XGBoost* dan *C4.5* dalam memprediksi kelayakan pinjaman.

Hasil data preprocessing dan transformation dataset

Setelah dilakukan pengambilan dataset, data akan diolah menggunakan *tool Google Colab* dengan bahasa *python* dan dilakukannya data *preprocessing* serta data *transformation* sebelum dilakukan pemodelan. Tampilan hasil data ditunjukkan pada Tabel 2.

TABEL 2. Hasil *preprocessing* dan *transformation* dataset

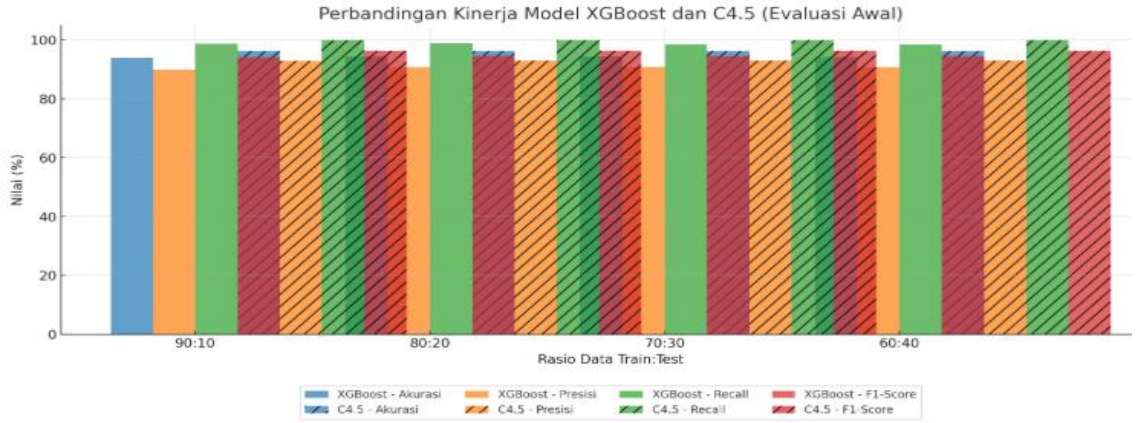
Applicant Age	*Work Experience*	*Maritial Status*	…	*Years in Current Residence*	*Loan Default Risk*	*Annual Income IDR*
76	0	1	…	12	0	1931531000
37	18	1	…	11	0	1851870600
66	8	1		12	0	278298385
…	…	…	…	…	…	…
66	11	1		14	0	1183227160
52	6	1	…	14	0	833598345
58	9	1	…	13	0	774556770

Pada Tabel 2 menampilkan hasil akhir dataset yang siap digunakan untuk pemodelan. Perubahan nya meliputi pembersihan data, penghapusan atribut, perubahan nama atribut, penskalaan pada pendapatan tahunan dari INR menjadi IDR, diubahnya nilai atribut dari string menjadi angka dan penyimbangan data. Semua dilakukan agar data bersih, hanya terdapat atribut yang penting, mudah terbaca,variabel nya terhubung dan tidak *overfitting* saat dilakukan pada pemodelan.

Hasil pemodelan klasifikasi algoritma XGBoost dan C4.5

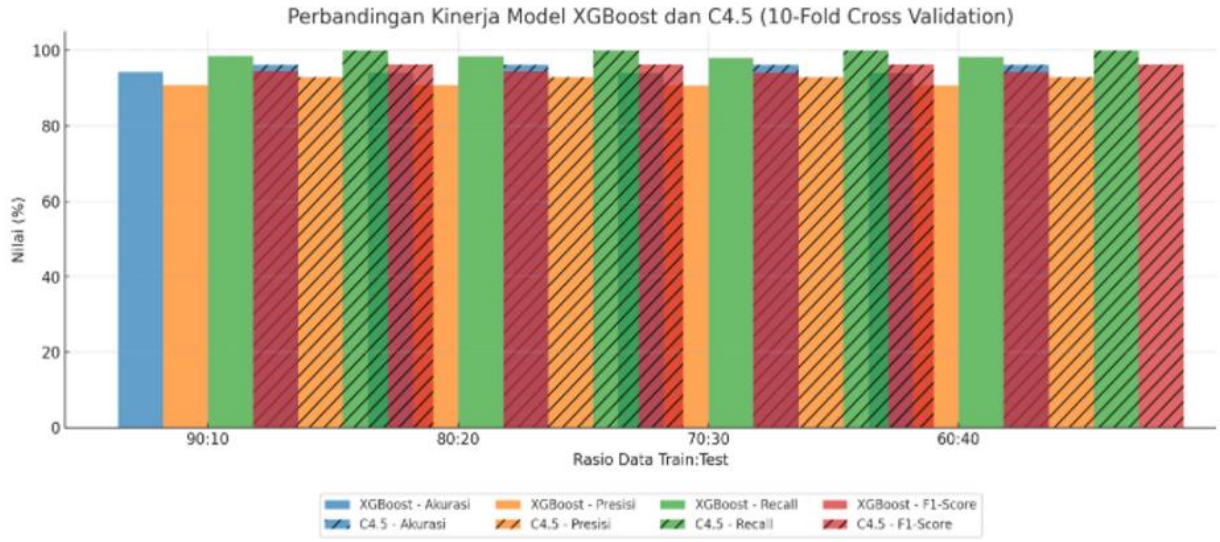
Pada tahap ini, data yang telah melewati *preprocessing* dan *transformation* dibagi menjadi dua bagian: data latih (*train*) dan data uji (*test*), dengan beberapa perbandingan rasio pembagian, yaitu 90:10, 80:20, 70:30, dan 60:40. Dalam pembagian 90:10, 90% dari dataset digunakan untuk melatih model, sementara 10% sisanya digunakan untuk pengujian. Demikian pula, pada rasio 80:20, 70:30, dan 60:40, data latih mencakup mayoritas dataset, sedangkan data uji mencakup sisanya. Berikut hasil pemodelan biasa pada algoritma *XGBoost* dengan perbandingan rasio ditunjukkan pada gambar grafik 2 dibawah ini,

IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 4, No. 2 Juli 2025
https://subset.id/index.php/IJCSR



GAMBAR 2. Grafik Perbandingan Kinerja XGBoost dan C4.5 Evaluasi Awal

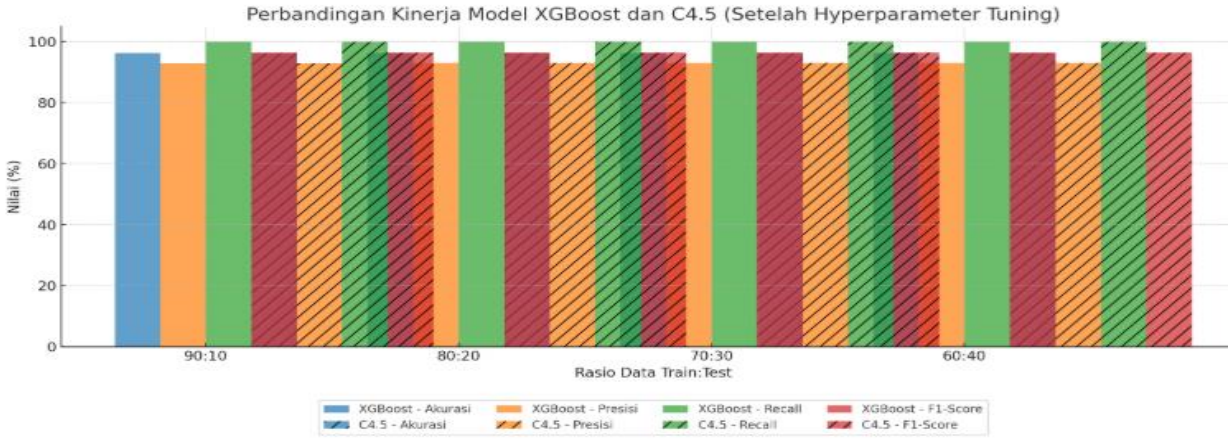
Grafik histogram pada gambar 2 yang membandingkan metrik evaluasi model (Akurasi, Presisi, Recall, dan F1-Score) antara algoritma *XGBoost* dan C4.5 pada berbagai rasio data latih dan uji (90:10 hingga 60:40). Grafik ini memudahkan untuk melihat bahwa C4.5 secara konsisten unggul pada semua metrik di tiap rasio. *XGBoost* memiliki performa baik tetapi sedikit di bawah C4.5, terutama pada presisi dan *F1-Score*. Pada Gambar Grafik 3 berikut ini hasil evaluasi *10-Fold Cross Validation* menunjukkan perbandingan kinerja model *XGBost* dengan C4.5



GAMBAR 3. Grafik Perbandingan Kinerja XGBoost dan C4.5 (10-Fold Cross Validation)

Dari grafik 3 diatas dapat disimpulkan jika C4.5 tetap unggul di semua metrik dan semua rasio data. Performa C4.5 sangat konsisten, dengan hanya sedikit perbedaan antar rasio. XGBoost menunjukkan penurunan kecil saat data latih lebih sedikit (rasio lebih kecil), tetapi masih kompetitif. Gambar 4 adalah grafik histogram dari hasil evaluasi setelah hyperparameter tuning dengan GridSearchCV

IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 4, No. 2 Juli 2025
https://subset.id/index.php/IJCSR



GAMBAR 4. Grafik evaluasi setelah hyperparameter tuning

Dari gambar grafik 4 tersebut terlihat jika *XGBoost* mengalami peningkatan performa signifikan dibanding sebelum tuning, dengan akurasi menyamai C4.5 di beberapa rasio. C4.5 tetap unggul dengan hasil yang lebih stabil dan konsisten di semua metrik. Grafik ini memperlihatkan bahwa setelah optimasi, performa kedua algoritma sangat berdekatan, namun C4.5 sedikit lebih unggul dari sisi konsistensi.

KESIMPULAN

Fokus utama dari penelitian ini adalah mengevaluasi dan membandingkan kinerja algoritma XGBoost dan C4.5, bukan membangun model yang spesifik untuk wilayah geografis tertentu. Oleh karena itu, meskipun data berasal dari India, penelitian ini tetap valid dalam konteks evaluasi algoritma. Namun demikian, dalam penelitian dapat disadari bahwa penggunaan data dari negara lain dapat membawa potensi bias, mengingat perbedaan kondisi sosial, ekonomi dalam bidang kurs, budaya setempat, dan geografis dari negara tersebut yang dapat memengaruhi nilai dan pola pemohon pinjaman.

Penggunaan data ini juga memberikan gambaran awal tentang bagaimana model-model dapat berpotensi diadaptasi ke berbagai konteks global di masa mendatang, tentu juga dengan penyesuaian pada struktur dan distribusi datanya agar bisa berhasil.

Hasil penelitian menunjukkan bahwa algoritma *C4.5* unggul dibandingkan *XGBoost*, baik sebelum maupun setelah optimasi *hyperparameter* dengan *GridSearchCV*. Pada tahap akhir di rasio 80:20*, C4.5* mencapai akurasi tertingi 96,20% dengan AUC 0.9730, walaupun nilai *C4.5* sebelum dan sesudah *tuning* hampir sama yang menandakan hasil optimal pada *C4.5*, sementara *XGBoost* memperoleh akurasi yang hampir setara setelah optimasi. Kesimpulan menunjukkan bahwa algoritma *C4.5* lebih unggul untuk memproses data besar dan memberikan hasil prediksi yang efisien, sementara *XGBoost* dapat menjadi alternatif kompetitif setelah dilakukan optimasi.

DAFTAR PUSTAKA

[1] E. Komalasari and G. S. Manda, "THE Pengaruh Risiko Kredit Dan Risiko Operasional Terhadap Kinerja Perbankan Pada Bank BUMN Yang terdaftar di BEI Periode 2012-2020," *J. Ilm. Ekon. dan Bisnis*, vol. 15, no. 1, pp. 89–95, 2022, doi: 10.51903/e-bisnis.v15i1.657.

[2] I. M. Sari, S. Siregar, and I. Harahap, "Manajemen Risiko Kredit bagi Bank Umum," *Semin. Nas. Teknol. Komput. Sains 2020*, vol. 1, no. 1, pp. 553–557, 2020.

[3] A. A. Saputra, B. N. Sari, and C. Rozikin, "Penerapan Algoritma Extreme Gradient Boosting (Xgboost) Untuk Analisis Risiko Kredit," *J. Ilm. Wahana Pendidik.*, vol. 10, no. 7, pp. 27–36, 2024, doi: https://doi.org/10.5281/zenodo.10960080.

[4] S. E. H. Yulianti, O. Soesanto, and uana Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," *J. Math. Theory Appl.*, vol. 4, no. 1, pp. 21–26, 2022, doi: 10.31605/jomta.v4i1.1792.

[5] N. Handayani, H. Wahyono, J. Trianto, and D. S. Permana, "Prediksi Tingkat Risiko Kredit dengan Data Mining

Menggunakan Algoritma Decision Tree C.45," *JURIKOM (Jurnal Ris. Komputer)*, vol. 8, no. 6, p. 198, 2021, doi: 10.30865/jurikom.v8i6.3643.

[6] J. M. A. S. Dachi and P. Sitompul, "Analisis Perbandingan Algoritma XGBoost dan Algoritma Random Forest Ensemble Learning pada Klasifikasi Keputusan Kredit," *J. Ris. Rumpun Mat. Dan Ilmu Pengetah. Alam*, vol. 2, no. 2, pp. 87–103, 2023, doi: 10.55606/jurrimipa.v2i2.1470.

[7] R. D. Mendrofa, M. H. Siallagan, J. Amalia, and D. P. Pakpahan, "Credit Risk Analysis With Extreme Gradient Boosting and Adaptive Boosting Algorithm," *J. Inf. Syst. Hosp. Technol.*, vol. 5, no. 1, pp. 1–7, 2023, doi: 10.37823/insight.v5i1.233.

[8] A. Amalia, A. Zaidah, and I. N. Isnainiyah, "Prediksi Kualitas Udara Menggunakan Algoritma K-Nearest Neighbor," *JIPI (Jurnal Ilm. Penelit. dan Pembelajaran Inform.*, vol. 07, no. 02, pp. 496–507, 2022, doi: 0.29100/jipi.v7i2.2843.

[9] F. Putra, H. F. Tahiyat, R. M. Ihsan, R. Rahmaddeni, and L. Efrizoni, "Penerapan Algoritma K-Nearest Neighbor Menggunakan Wrapper Sebagai Preprocessing untuk Penentuan Keterangan Berat Badan Manusia," *MALCOM Indones. J. Mach. Learn. Comput. Sci.*, vol. 4, no. 1, pp. 273–281, 2024, doi: 10.57152/malcom.v4i1.1085.

[10] D. A. Manalu and G. Gunadi, "Implementasu Metode Data Mining K-Means CLustering Terhadap Data Pembayaran Transaksi Menggunakan Bahasa Pemrograman Python Pada CV Digital Dimensi," *Infotech J. Technol. Inf.*, vol. 8, no. 1, pp. 175–305, 2022, doi: 10.37365/jti.v8i1.131.

[11] G. Gumelar, Q. Ain, R. Marsuciati, S. Agustanti Bambang, A. Sunyoto, and M. Syukri Mustafa, "Kombinasi Algoritma Sampling dengan Algoritma Klasifikasi untuk Meningkatkan Performa Klasifikasi Dataset Imbalance," *SISFOTEK Sist. Inf. dan Teknol.*, pp. 250–255, 2021.

[12] S. Hidayat *et al.*, "Optimalisasi Model Ensemble Learning dengan Augmentasi dan SMOTE pada Sistem Pendeteksi Kualitas Buah," *JTIM J. Teknol. Inf. dan Multimed.*, vol. 6, no. 1, pp. 27–36, 2024, doi: 10.35746/jtim.v6i1.406.

[13] S. Suwarno and R. Kusnadi, "Comparative Analysis of SVM, XGBoost and Neural Network on Hate Speech Classification," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 5, no. 5, pp. 896–903, 2021, doi: 10.29207/resti.v5i5.3506.

[14] T. Ridwansyah, "Implementasi Text Mining Terhadap Analisis Sentimen Masyarakat Dunia Di Twitter Terhadap Kota Medan Menggunakan K-Fold Cross Validation Dan Naïve Bayes Classifier," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 2, no. 5, pp. 178–185, 2022, doi: 10.30865/klik.v2i5.362.

[15] A. Hakim Arif and A. Solichin, "Hyperparameter Optimization on Ensemble Regression Tree for Lip Coloring Simulation," *J. Tek. Inform. dan Sist. Inf.*, vol. 8, no. 2, pp. 297–310, 2022, doi: 10.28932/jutisi.v8i2.4611.

[16] I. Muhamad Malik Matin, "Hyperparameter Tuning Menggunakan GridsearchCV pada Random Forest untuk Deteksi Malware," *Multinetics*, vol. 9, no. 1, pp. 43–50, 2023, doi: 10.32722/multinetics.v9i1.5578.

[17] F. Aziz, P. Ishak, and S. Abasa, "Klasifikasi Depresi Menggunakan Support Vector Machine: Pendekatan Berbasis Data Text Mining," *J. Pharm. Appl. Comput. Sci.*, vol. 2, no. 2, pp. 33–38, 2024, doi: 10.59823/jopacs.v2i2.53.

[18] S. P. Dewi, N. Nurwati, and E. Rahayu, "Penerapan Data Mining Untuk Prediksi Penjualan Produk Terlaris Menggunakan Metode K-Nearest Neighbor," *Build. Informatics, Technol. Sci.*, vol. 3, no. 4, pp. 639–648, 2022, doi: 10.47065/bits.v3i4.1408.

[19] J. E. C. Candra and M. F. E. Prasetyo, "Pemanfaatan Sensor Fingerprint Untuk Kendali dan Keamanan Sepeda Motor Berbasis Arduino," *J. Desain Dan Anal. Teknol.*, vol. 2, no. 1, pp. 66–74, 2023, doi: 10.58520/jddat.v2i1.22.

[20] H. Hendri, "Implementasi Data Mining Dengan Metode C4.5 Untuk Prediksi Mahasiswa Penerima Beasiswa," *Indones. J. Comput. Sci.*, vol. 10, no. 2, pp. 312–321, 2021, doi: 10.33022/ijcs.v10i2.3013.

[21] S. Sudarto and K. Kusrini, "Klasifikasi Tsunami Gempa Bumi Dengan Teknik Stacking Ensemble Machine Learning," *JIP (Jurnal Inform. Polinema)*, vol. 10, no. 4, 2023, doi: 10.33795/jip.v10i4.5655.