ISSN (Online) 2747-0563
Volume 4 Tahun 2024

Klasifikasi Kegagalan Pembayaran Kredit Nasabah Bank dengan Algoritma XGBoost

Fransiska Aloysia Putri Prasetya[1], Paulina H. Prima Rosa[2*]

[1]Program Studi Informatika, Universitas Sanata Dharma
[1]fransiskaprasetya2612@gmail.com

[2] Program Studi Informatika, Universitas Sanata Dharma

**Corresponding author email*: rosa@usd.ac.id

***Abstrak*— Dengan menggunakan algoritma XGBoost, penelitian ini bertujuan untuk mengimplementasikan algoritma XGBoost sebagai Pembangunan model awal dan model yang diperbaiki menggunakan hyperparameter tuning untuk klasifikasi kegagalan pembayaran kredit nasabah bank, menggunakan data yang berasal dari Kaggle. Proses seleksi fitur dilakukan berdasar nilai information gain dengan nilai variasi batas bawah information gain 0.002, 0.004, 0.006, dan 0.01. Validasi model dilakukan dengan metode K-Fold Cross Validation (dengan nilai K = 3, 5, dan 10). Dilakukan juga penyetelan hiperparameter. Hasil penelitian menunjukan bahwa algoritma XGBoost memiliki akurasi yang cukup baik mencapai 93.10% dengan menggunakan 7 hyperparameter tuning, 14 atribut yang paling relevan berdasarkan hasil penggunaan information gain dengan k-fold 10 setelah data dikenai proses balancing menggunakan teknik SMOTE terhadap data latih dan 88.46% terhadap data uji yang telah melalui tahap preprocessing dengan penggunaan information gain sebagai kriteria seleksi fitur serta 7 nilai hyperparameter tuning terbaik. Hal ini mengindikasikan bahwa XGBoost efektif digunakan dalam klasifikasi kegagalan pembayaran kredit nasabah bank ketika data sudah seimbang dan melalui beberapa proses seperti penggunaan information gain untuk menentukan atribut yang paling informatif atau relevan dalam memprediksi kelas atau label target, dan juga penggunaan 7 hyperparameter tuning yang dapat membantu menaikan akurasi algoritma.**

Kata Kunci— **XGBoost, Kegagalan pembayaran kredit, Klasifikasi, *Balancing*, SMOTE, *Cross-validation*.**

I. PENDAHULUAN

Sistem perbankan memainkan peran penting dalam mendukung pertumbuhan ekonomi melalui pemberian kredit kepada individu atau entitas bisnis untuk memenuhi kebutuhan konsumtif atau produktif mereka. Kebutuhan produktif mencakup peningkatan dan ekspansi aktivitas usaha, sementara kebutuhan konsumtif, seperti membeli rumah, dapat diakomodasi melalui fasilitas pinjaman seperti Kredit Kepemilikan Rumah (KPR) [1]

XGBoost (eXtreme Gradient Boosting) merupakan algoritma ensemble learning yang memanfaatkan teknik boosting. XGBoost digunakan untuk klasifikasi dan regresi, dengan membangun serangkaian pohon keputusan yang lemah secara berturut-turut, dan mengkombinasikan hasilnya untuk meningkatkan kinerja model. Selain itu penerapan algoritma XGBoost juga, telah terbukti dalam beberapa contoh penelitian menggunakan teknik klasifikasi mendapatkan hasil akurasi yang sangat baik[2].

Data yang digunakan pada penelitian ini bersifat imbalanced, dimana klasifikasi pada dataset yang tidak seimbang dapat menimbulkan masalah karena algoritma pembelajaran sensitif terhadap jumlah instance pelatihan. Dengan demikian, diperlukan algoritma untuk memodifikasi set pelatihan dengan meningkatkan ukuran kelas minoritas, entah dengan menduplikasi atau melakukan interpolasi pada instansi minoritas. Salah satu metode oversampling yang terkenal adalah algoritma *SMOTE* [3] Metode ini bahkan menginspirasi berbagai pendekatan baru dalam pembelajaran mesin terawasi, termasuk klasifikasi multilabel, pembelajaran inkremental, pembelajaran semi-terawasi, dan lain-lain. *SMOTE* adalah standar untuk mengatasi masalah data tidak seimbang [4]

Seleksi fitur adalah metode untuk mengurangi jumlah atribut yang digunakan dalam analisis data. Tujuannya adalah untuk mempertahankan atribut yang paling relevan dan menghindari atribut yang berlebihan, sehingga dapat mempercepat proses klasifikasi dan meningkatkan akurasi hasil dari algoritme klasifikasi [5]. Metode ini akan melalui proses komputasi untuk mendapatkan atribut-atribut yang paling berpengaruh terhadap dataset ini.

Pada penelitian sebelumnya yang dilakukan oleh [6] yang membandingkan hasil akurasi algoritma *Random Forest* dan *XGBoost* yang diimplementasikan untuk menentukan kelayakan pemberian kredit. Maka didapatkan hasil akurasi tertinggi dari ke-2 perbandingan algoritma yaitu menggunakan algoritma *XGBoost* karena nilai *AUC* yang dihasilkan paling tinggi. Berbagai penelitian telah dilakukan untuk membantu mengklasisifikasi kegagalan pembayaran kredit nasabah bank salah satunya yaitu yang dilakukan oleh [7] dengan membandingkan empat metode klasifikasi yaitu : *Random Forest*, *KNN*, *SVM*, dan *MLP*, dengan penerapan *SMOTE* pada dataset menghasilkan peningkatan kinerja. Penelitian selanjutnya yang dilakukan oleh [8] dengan menggunakan algoritma *XGBoost* dan *Logistic Regression* mendapatkan hasil dimana *XGBoost* menghasilkan rata-rata *akurasi* 85,51%; *F1 Score* 83,81%; *precision* 83,80% dan *recall* 84,04%

Penelitian ini diharapkan dapat menghasilkan model klasifikasi yang lebih akurat dan efisien dalam memprediksi kegagalan pembayaran kredit nasabah bank. Dengan menerapkan algoritma XGBoost, penelitian ini juga bertujuan untuk meningkatkan performa model melalui penggunaan teknik seleksi fitur seperti Information Gain, metode validasi K-Fold Cross Validation, dan tuning hyperparameter dengan RandomizedSearchCV. Hasil yang dicapai dari penelitian ini diharapkan dapat memberikan kontribusi yang signifikan terhadap manajemen risiko kredit dan membantu bank dalam mengambil keputusan yang lebih tepat terkait penilaian kredit nasabah.

ISSN (Online) 2747-0563
Volume 4 Tahun 2024

II. DASAR TEORI

A. Penambangan Data

Penambangan Data adalah proses semi otomatis yang menggunakan teknik statistik, matematika, kecerdasan buatan, dan pembelajaran mesin untuk mengekstraksi dan mengidentifikasi informasi potensial yang bermanfaat yang tersimpan dalam database besar. Ini merupakan bagian dari proses KDD (*Knowledge Discovery in Databases*) yang meliputi tahap-tahap seperti pemilihan data, pra-pengolahan, transformasi, data mining, dan evaluasi hasil. KDD juga dikenal secara umum sebagai penemuan pengetahuan dalam basis data [9].

B. Normalisasi *Min-Max*

Normalisasi *Min-Max* bertujuan untuk merubah skala fitur pada data sehingga nilai-nilai fitur tersebut berada dalam rentang tertentu, biasanya antara 0 dan 1. Kelebihan dari metode atau pendekatan ini yaitu mempertahankan semua relasi nilai data secara akurat, hal tersebut tidak mengakibatkan potensi distorsi (*bias*) pada data [10].

C. Penyeimbangan Data

Ketidakseimbangan data mempengaruhi bagaimana klasifikasi berfungsi. Klasifikasi pada dataset yang tidak seimbang dapat menimbulkan masalah karena algoritma pembelajaran sensitif terhadap jumlah instance pelatihan. SMOTE adalah metode untuk menyamakan jumlah sampel data pada kelas minoritas dengan memilih dan menyesuaikan data sampel sehingga jumlahnya sejajar dengan kelas mayoritas [11].

D. K-Fold Cross Validation

K-Fold cross validation adalah metode yang digunakan untuk menilai seberapa baik model yang telah dibuat dengan menggunakan data yang ada. Proses pembuatan model menggunakan data pelatihan, sementara pengujian model dilakukan dengan menggunakan data pengujian [12]. Tujuan K-Fold Cross Validation dalam klasifikasi yaitu untuk mengukur seberapa baik performa model klasifikasi secara umum dengan membagi data menjadi beberapa bagian (fold), metode ini memungkinkan untuk melatih dan menguji model menggunakan setiap bagian data secara bergantian.

E. Extremly Gradient Boosting (XGBoost)

XGBoost adalah sebuah implementasi dari metode *Gradient Boosting* dengan mengadopsi formulasi model yang lebih terstruktur untuk mengatasi masalah overfitting. Pengembangan XGBoost dilakukan oleh T. Chen dan C. Guestrin pada tahun 2016, yang merupakan library gradient boosting terdistribusi yang telah dioptimalkan untuk desain yang lebih efisien, fleksibel dan partable. XGBoost bekerja untuk melakukan klasifikasi, sehingga menghasilkan akurasi yang tinggi dan kinerja model yang unggul [13].
Langkah-langkah algoritma Extreme Gradient Boosting[14]:

1. Membuat prediksi awal

$$f_0(x) = \left(\sum_{i=1}^{n} yi \ / \mathrm{n}\right) \quad (1)$$

2. Perhitungan eror atau residual

$$\hat{Y} = y - f_0(X) \quad (2)$$

3. Pelatihan model
   (1) Membuat struktur pohon dengan memisahkan data ke dalam dua partisi melalui berbagai kemungkinan split.
   (2) Menghitung nilai similarity dan gain pada setiap pohon yang terbentuk melalui nilai split optimal.

$$Similarity = \quad (3)$$

$$\left(\sum \hat{y}_1\right)^2 / \sum \left[Previous\ f_1 \,.(1 - Previous\ f_1\ (x)] + \lambda\right.$$

$$Gain = \left(Left_{Similarity} + Right_{Similarity}\right) - root_{similarity} \quad (4)$$

(3) Melakukan pemisahan ulang pada pohon yang memiliki nilai gain maksimal hingga mencapai batas max_depth untuk membentuk pohon secara menyeluruh.
(4) Setelah pembentukan pohon, lakukan pruning (pemangkasan) untuk mengurangi ukuran pohon keputusan dengan menghapus atau memotong bagian-bagian yang kurang efektif untuk klasifikasi.
(5) Menghitung *output value*

$$Output\ Value = \left(\sum \hat{y}_1\right) / \sum [F_{i-1} \times (1 - F_{i-1})] - \lambda \quad (5)$$

(6) Menghitung prediksi dari model, dimana setiap titik data melewati pohon keputusan terakhir untuk menghasilkan h1(x). selanjutnya, menghitung prediksi f1(x) dan nilai residu.

$$F_n(x) = \quad (6)$$

$$1/1 + EXP\ (-x)([h_0(x)/1 - h_0(x)] + \sum_{i=1}^{n}[\eta \times h_i(x)]\ )$$

(7) Melakukan Langkah-langkah yang sama kembali (Langkah ke-3 Pelatihan Model) untuk membangun pohon lain
(8) Membuat prediksi akhir

F. Confusion Matrix

Confusion matrix adalah suatu teknik yang menampilkan hasil melalui tabel matriks untuk menghitung tingkat akurasi, recall, presisi, dan eror [15].

TABEL I
CONFUSION MATRIX

Prediksi	**Nilai Asli**	
	Positive	*Negative*
Positive	TP	FP
Negative	FN	TN

G. Pemyetelan Hiperparameter

Proses penyetelan hiperparameter *(hyperparameter tuning*) menggunakan random search cross validation adalah teknik hyperparameter tuning yang digunakan untuk mengoptimalkan kinerja model machine learning dengan mencoba berbagai

ISSN (Online) 2747-0563
Volume 4 Tahun 2024

kombinasi hyperparameter secara acak. RandomizedSearchCV adalah suatu kelas yang termasuk dalam library scikit-learn di Phyton dan dapat dipergunakan dalam hyperparameter tuning, untuk memberikan keseimbangan antara efisiensi dan performa, serta membantu dalam membangun model machine learning yang lebih baik dan lebih andal. Dimana hyperparameter tuning yang digunakan seperti diuraikan dalam Tabel II.

TABEL II

HIPERPARAMETER

Hiperparameter	Fungsi Hiperparameter
n_estimators	Menentukan jumlah pohon keputusan yang akan dibangun
max_depth	Kedalaman pohon
Min_child_weight	Menentukan jumlah minimum total bobot yang diperlukan di dalam suatu cabang pohon.
eta (learning_rate)	Membantu mempersingkat langkah dalam pembaruan model
Gamma	Parameter yang mengontrol besarnya pemangkasan pada setiap langkah pemisahan
subsample	Bagian atau pecahan dari keseluruhan data training yang akan digunakan untuk menyesuaikan (fit) setiap pohon dalam model
colsample_bylevel	Bagian atau proporsi dari keseluruhan fitur (kolom) yang akan digunakan saat membangun setiap level pohon dalam model

Cara Kerja penyetelan hiperparameter adalah sebagai berikut:

1. Tentukan ruang parameter hiperparameter yang ingin dicari.
2. Tentukan jumlah iterasi Random Search (misalnya, n_iter = 50).
3. Untuk setiap iterasi, pilih kombinasi hiperparameter secara acak dari ruang parameter.
4. Untuk setiap kombinasi hiperparameter yang dipilih, gunakan K-Fold cross-validation untuk mengevaluasi model:
   - Dataset dibagi menjadi K subset (folds).
   - Model dilatih pada K-1 subset dan divalidasi pada subset yang tersisa. Proses ini diulang K kali, setiap kali dengan subset validasi yang berbeda.
   - Hasil evaluasi pada setiap fold dirata-rata untuk mendapatkan skor kinerja untuk kombinasi hyperparameter tersebut.
5. Ulangi langkah 3 dan 4 untuk semua iterasi.
6. Pilih kombinasi hiperparameter dengan skor kinerja terbaik dari semua iterasi.

III. METODOLOGI PENELITIAN

A. *Sumber Data*

Data yang digunakan dalam penelitian ini diperoleh dari platform Kaggle (https://www.kaggle.com/datasets/nikhil1e9/loan-default/data) dalam bentuk file .xls. Data yang digunakan dalam penelitian ini berjumlah 255.348 data dan 18 kolom atribut data. Dimana kolom atau atribut "Default" merupakan kelas label yang bernilai 1 dan 0 untuk menyatakankan nasabah yang gagal membayar (nilai 1) dan berhasil membayar kredit (nilai 0). Jumlah nasabah yang gagal membayar kredit sebanyak 29.653 nasabah, sedangkan yang berhasil membayar kredit sebanyak 225.695 nasabah. Tabel III berikut ini mendeskripsikan atribut-atribut yang terdapat dalam dataset.

TABEL III

DESKRIPSI ATRIBUT DALAM DATASET

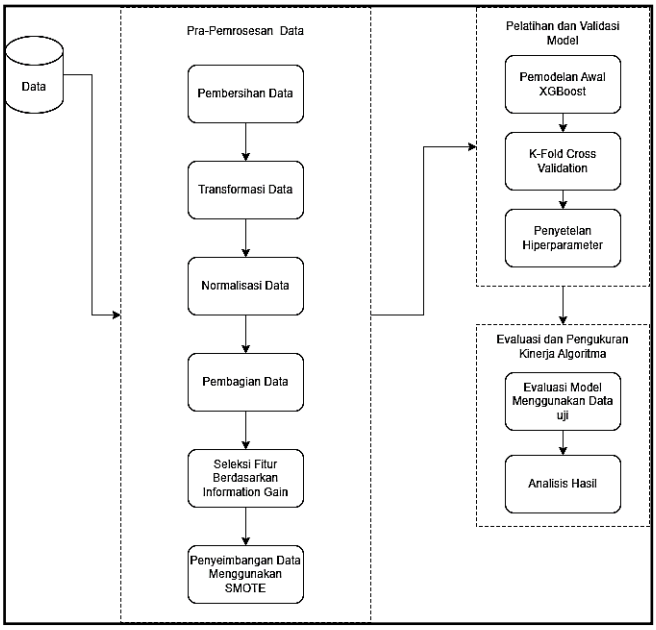
No	Nama Atribut	Tipe Data	Deskripsi Atribut
1.	*LoanID*	*String*	Nomor identifikasi unik untuk setiap peminjam
2.	*Age*	*Integer*	Menunjukkan usia peminjam
3.	*Income*	*Integer*	Merupakan pendapatan tahunan dari peminjam
4.	*LoanAmount*	*Integer*	Jumlah uang yang dipinjam oleh peminjam
5.	*CreditScore*	*Integer*	Skor kredit pinjaman
6.	*MonthsEmployed*	*Integer*	Menunjukkan berapa lama peminjam telah bekerja di tempat kerja saat mengajukan pinjaman
7.	*NumCreditLines*	*Integer*	Jumlah total garis kredit atau kartu kredit yang dimiliki oleh peminjam
8.	*InterestRate*	*Float*	Tingkat suku bunga yang diterapkan pada pinjaman
9.	*LoanTerm*	*Integer*	Jangka waktu atau durasi pinjaman dalam satuan bulan
10.	*DTIRatio*	*Float*	Rasio antara total kewajiban finansial bulanan (termasuk cicilan pinjaman) dengan total pendapatan bulanan
11.	*Education*	*String*	Tingkat pendidikan dari peminjam
12.	*EmploymentType*	*String*	Jenis pekerjaan atau status pekerjaan dari peminjam
13.	*MaritalStatus*	*String*	Status perkawinan dari peminjam
14.	*HasMortgage*	*String*	Menunjukkan apakah peminjam memiliki hipotek atau pinjaman rumah saat ini
15.	*HasDependents*	*String*	Menunjukkan apakah peminjam memiliki tanggungan atau anggota keluarga lain yang bergantung pada pendapatan mereka
16.	*LoanPurpose*	*String*	Tujuan penggunaan pinjaman
17.	*HasCoSigner*	*String*	Menunjukkan apakah ada penjamin atau *co-signer* yang mendukung

ISSN (Online) 2747-0563
Volume 4 Tahun 2024

No	Nama Atribut	Tipe Data	Deskripsi Atribut
			peminjam dalam pinjaman ini
18.	*Default*	*Integer*	Variabel target dalam *dataset*

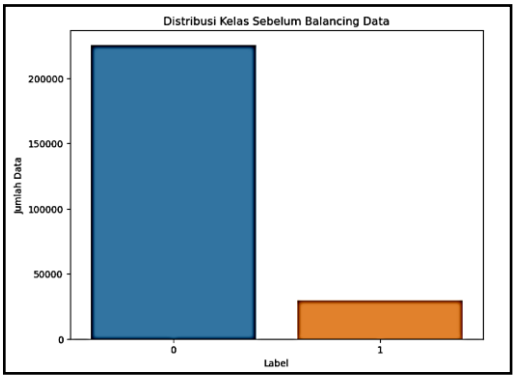
B. *Langkah Penelitian*

Gambar 1 berikut ini merupakan diagram alur penelitian yang dilakukan.

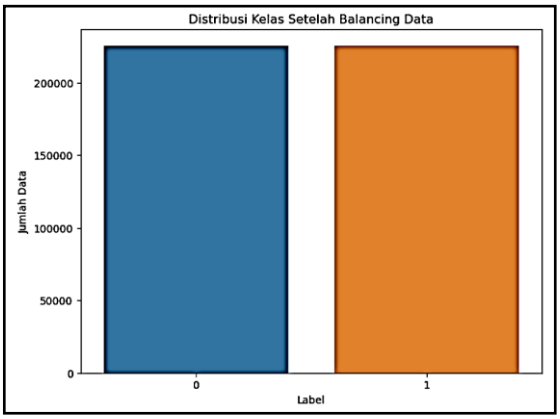


Gbr. 1 Diagram alur penelitian

Penelitian dimulai dengan pra-pemrosesan yang mencakup pembersihan data, transformasi data, normalisasi data, pembagian data, seleksi fitur, dan penyeimbangan data. Dalam tahap pembersihan data dilakukan pembersihan data dari nilai yang hilang (*missing values*). Karena dari pengecekan terhadap dataset tidak ditemukan data yang hilang maka tidak perlu dilakukan apapun. Dalam tahap transformasi data, atribut yang bertipe kategorikal diubah menjadi tipe data numerik dengan menggunakan teknik *encoding*. Dalam tahap normalisasi data, diterapkan metode *min-max normalization* sehingga atribut-atribut yang memiliki rentang nilai yang berbeda-beda bisa disetarakan. Dalam tahap pembagian data, data dibagi terlebih dahulu menggunakan train_test_split dengan skala 80 : 20 dimana 80% merupakan data dijadikan datavlatih dan 20% data menjadi data uji. Proses seleksi fitur dilakukan menggunakan *information gain* sebagai kriteria seleksi untuk mengidentifikasi atribut-atribut yang paling informatif berdasarkan pada kelas tertentu. Seleksi fitur dilakukan berdasar nilai *information gain*. Dilakukan beberapa eksprimen untuk mencari ambang batas information gain guna mendapatkan model dengan akurasi yang baik. Sebagaimana tampak dalam Gambar 2, jumlah data yang dikategorikan gagal melakukan pembayaran kredit lebih sedikit dibandingkan dengan yang berhasil melakukan pembayaran kredit. Untuk itu perlu dilakukan proses penyeimbangan data. Implementasi proses penyeimbangan data menggunakan teknik *Synthetic Minority Over-sampling Technique* (SMOTE). Dengan teknik tersebut, maka distribusi data menjadi seperti dalam Gambar 3.



Gbr. 2 Distibusi data sebelum penyeimbangan



Gbr. 3 Distibusi data setelah penyeimbangan

Tahap pelatihan model menggunakan data latih yang diambil dari 80% dataset dilakukan berdasarkan skenario yang terdiri dari 2 tahap utama sebagai berikut:

a. Pembangunan model awal (*initial model*)

Pada tahap ini, model awal dibangun dengan parameter default dari XGBoost dalam Python dan diuji dengan menerapkan beberapa eksperimen seleksi fitur dan k-fold cross validation. Eksperimen seleksi fitur dilakukan dengan 4 nilai batas bawah information gain yaitu 0.002, 0.004, 0.006, dan 0.01. Pada setiap pemilihan batas bawah information gain, atribut dengan nilai information gain >= batas bawah dipilih untuk digunakan dalam prroses selanjutnya, sementara atribut dengan nilai information gain < batas bawah tidak dipergunakan dalam proses selanjutnya. Sedangkan eksperimen dengan k-fold cross validation dilakukan dengan 3 variasi nilai k yaitu 3, 5, dan 10. Tujuan dari pengujian ini adalah untuk mencari kombinasi parameter yang menghasilkan nilai akurasi tertinggi. Tabel IV berikut ini menggambarkan variasi eksperimen pembagunan model awal tersebut.

TABEL IV

VARIASI EKSPERIMEN PEMBANGUAN MODEL AWAL

Batas Bawah *Information Gain*	K-Fold Cross Validation
0.002	3

ISSN (Online) 2747-0563
Volume 4 Tahun 2024

Batas Bawah *Information Gain*	**K-Fold Cross Validation**
	5
	10
0.004	3
	5
	10
0.002	3
	5
	10
0.01	3
	5
	10

b. Pengujian Model dengan Penyetelan Hiperparameter
Pada tahap ini, model terbaik yang didapat dari pembangunan model awal diperbaiki dengan melalukan penyetelan hiperparameter untuk memperoleh model dengan akurasi yang lebih baik. Berdasarkan penelitian Yulianti et.al (2022), dilakukan penyetelan 7 parameter seperti tercantum dalam Tabel V.

TABEL V
NILAI HIPERPARAMETER

Hyperparameter	**Random Search Values**
n_estimators	400, 300, 200, 100
max_depth	8, 7, 6, 5, 4
Min_child_weight	0, 1, 2, 3, 4, 5, 6, 7
eta (learning_rate)	0.3, 0.2, 0.1, 0.05, 0.025
Gamma	0, 0.1, 0.2, 0.3, 0.4, 1, 1.5, 2
subsample	1, 0.75, 0.5, 0.15
colsample_bylevel	0.1, 0.2, 0.25, 1.0

Setelah model terbaik diperoleh dari 2 jenis pelatihan dengan menggunakan data latih di atas, langkah selanjutnya adalah mengevaluasi model tersebut terhadap data uji yang diambil dari 20% dataset untuk memprediksi akurasi model terhadap dataset yang berbeda dari data latih.

IV. HASIL DAN PEMBAHASAN

A. *Pembangunan Model Awal*

Tabel VI berikut ini merupakan daftar fitur hasil seleksi berdasar information gain dengan nilai batas bawah 0.002, 0.004, 0.006, dan 0.01. Fitur-fitur terpilih kemudian dikombinasikan dengan 3 variasi nilai fold untuk melihat akurasinya. Tabel VII merangkum hasil eksperimen tersebut. Terlihat bahwa akurasi tertinggi model awal ini sebesar 92.94% diperoleh dari seleksi fitur yang menggunakan nilai batas bawah information gain = 0.002 dan menggunakan k-fold cross validation dengan nilai k = 10. Fitur-fitur yang relevan atau memiliki kontribusi paling signifikan yaitu *Age*, *HasCoSigner*, *InterestRate*, *HasMortgage*, *HasDependents*, *Income*, *MaritalStatus*, *MonthsEmployed*, *EmploymentType*, *NumCreditLines*, *LoanPurpose*, *Education*, *LoanAmount* dan *LoanTerm.*

TABEL VI
FITUR TERPILIH BERDASAR INFORMATION GAIN

Nilai Batas Bawah *Information Gain*	**Fitur Terpilih**
0.002	Age HasCoSigner InterestRate HasMortgage HasDependents Income MaritalStatus MonthsEmployed EmploymentType NumCreditLines LoanPurpose Education LoanAmount LoanTerm
0.004	Age HasCoSigner InterestRate HasMortgage HasDependents Income MaritalStatus MonthsEmployed EmploymentType NumCreditLines LoanPurpose Education
0.006	Age HasCoSigner InterestRate HasMortgage HasDependents Income MaritalStatus
0.01	Age HasCoSigner InterestRate

TABEL VII
AKURASI MODEL AWAL DENGAN VARIASI BATAS BAWAH INFORMATION GAIN DAN K-FOLD

Batas Bawah *Information Gain*	***K-fold Cross Validation***	**Akurasi *XGBoost***
0.002	3	92.42%
	5	92.69%
	10	92.94%
0.004	3	91.31%
	5	91.85%
	10	91.69%
0.006	3	79.41%
	5	79.67%
	10	79.68%
0.01	3	66.60%
	5	66.87%
	10	66.96%

ISSN (Online) 2747-0563
Volume 4 Tahun 2024

B. *Penyetelen Hiperparameter*

Model awal terbaik yang diperoleh dari eksperimen di atas yaitu model yang diperoleh dari seleksi fitur yang menggunakan nilai batas bawah information gain = 0.002 dan menggunakan k-fold cross validation dengan nilai k = 10 selanjutnya dikenai penyetelan hiperparameter sesuai skenario dalam Tabel IV. Dengan teknik random search cross validation pada Python, diperoleh nilai hiperparameter terbaik sebagaimana tercantum dalam Tabel VII. Akurasi dari model yang telah melalui penyetelan hiperparameter adalah 93.10%, meningkat 0,16% dibandingkan model awal.

TABEL VIII

HIPERPARAMETER TERBAIK

Hiperparameter	***Random Search Values***	**Nilai Hiperparameter Terbaik**
n_estimators	400, 300, 200, 100	400
max_depth	8, 7, 6, 5, 4	6
Min_child_weight	0, 1, 2, 3, 4, 5, 6, 7	1
eta (learning_rate)	0.3, 0.2, 0.1, 0.05, 0.025	0.2
Gamma	0, 0.1, 0.2, 0.3, 0.4, 1, 1.5, 2	0.3
subsample	1, 0.75, 0.5, 0.15	0.75
colsample_bylevel	0.1, 0.2, 0.25, 1.0	1.0

Model terbaik yang diperoleh dari penyetelan hiperparameter selanjutnya dipergunakan untuk mengklasifikasikan data uji yang telah disisihkan sebanyak 20% dari dataset. Hasil akurasi yang diperoleh sebesar 88.46%.

Secara keseluruhan, akurasi model klasifikasi yang dibangun dari model awal hingga penyetelan parameter sangat baik, namun akurasi klasifikasi terhadap data uji masih di bawah 90%. Ekspl.

V. KESIMPULAN

Berdasarkan penelitian yang telah dilakukan, maka didapatkan simpulan sebagai berikut :

- Metode Extreme Gradient Boosting (XGBoost) berhasil dipergunakan untuk melakukan klasifikasi kegagalan pembayaran kredit nasabah bank. Model yang dibangun menggunakan algoritma Extreme Gradient Boosting (XGBoost) dengan menggunakan 14 atribut yang dikenai proses penyeimbangan data dengan teknik SMOTE serta dilakukan penyetelan hiperparameter menghasilkan akurasi terbaik sebesar 93.10% pada nilai k-fold = 10.
- Model evaluation terhadap data uji yang disisihkan sebesar 20% mendapatkan hasil akurasi yang cukup baik sebesar 88.46%.

Penelitian lanjutan untuk lebih meningkatkan akurasi dapat dicoba lakukan dengan menggabungkan algoritma XGBoost dengan algoritma boosting lainnya seperti Catboost, AdaBoost, dan sebagainya.

REFERENSI

[1] A. Nursyahriana, M. Hadjat, and I. Tricahyadinata, "Analisis Faktor Penyebab Terjadinya Kredit Macet," *FORUM EKONOMI*, vol. 19, no. 1, 2017.

[2] E. H. Yulianti, O. Soesanto, and Y. Sukmawaty, "Penerapan Metode Extreme Gradient Boosting (XGBOOST) pada Klasifikasi Nasabah Kartu Kredit," *JOMTA Journal of Mathematics: Theory and Applications*, vol. 4, no. 1, 2022.

[3] N. Soonthornphisaj, T. Sira-Aksorn, and P. Suksankawanich, "Social media comment management using smote and random forest algorithms," *International Journal of Networked and Distributed Computing*, vol. 6, no. 4, pp. 204–209, 2018.

[4] A. Fernández, S. García, F. Herrera, and N. V Chawla, "SMOTE for Learning from Imbalanced Data: Progress and Challenges, Marking the 15-year Anniversary," 2018.

[5] S. Hidayatul, A. Aini, Y. A. Sari, and A. Arwan, "Seleksi Fitur Information Gain untuk Klasifikasi Penyakit Jantung Menggunakan Kombinasi Metode K-Nearest Neighbor dan Naïve Bayes," 2018. [Online]. Available: http://j-ptiik.ub.ac.id

[6] S. F. N. Halim and U. Azmi, "Analisis Perbandingan Klasifikasi dan Penerapan SMOTE Dalam Imbalanced Data pada Credit Card Default."

[7] M. Ibnu and C. Rachmatullah, "Penerapan SMOTE untuk Meningkatan Kinerja Klasifikasi Penilaian Kredit," *Jurnal Riset Komputer)*, vol. 10, no. 1, pp. 2407–389, 2023, doi: 10.30865/jurikom.v10i1.5612.

[8] A. Yaqin, "Penilaian Kredit Menggunakan Algoritma XGBoost dan Logistic Regression," *Jurnal Informatika: Jurnal Pengembangan IT*, vol. 8, no. 1, pp. 4–10, 2023.

[9] C. Zai, "IMPLEMENTASI DATA MINING SEBAGAI PENGOLAHAN DATA," 2022.

[10] W. Li and Z. Liu, "A method of SVM with normalization in intrusion detection," in *Procedia Environmental Sciences*, Elsevier B.V., 2011, pp. 256–262. doi: 10.1016/j.proenv.2011.12.040.

[11] A. Nikmatul Kasanah, Muladi, and U. Pujianto, "Penerapan Teknik SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Objektivitas Berita Online Menggunakan Algoritma KNN," *Terakreditasi SINTA Peringkat 2*, vol. 3, pp. 196–201, 2019.

[12] I. Ayu Made Supartini, I. Komang Gde Sukarsa, and I. Gusti Ayu Made Srinadi, "ANALISIS DISKRIMINAN PADA KLASIFIKASI DESA DI KABUPATEN TABANAN MENGGUNAKAN METODE K-FOLD CROSS VALIDATION," vol. 6, no. 2, pp. 106–115, 2017.

[13] R. D. Mendrofa, M. H. Siallagan, J. Amalia, and D. P. Pakpahan, "Credit Risk Analysis With Extreme Gradient Boosting and Adaptive Boosting Algorithm," *Journal of Information System,Graphics, Hospitality and Technology*, vol. 5, no. 1, pp. 1–7, Mar. 2023, doi: 10.37823/insight.v5i1.233.

[14] J. Iskandar, V. C. Mawardi, and J. Hendryli, "Analisis Media Sosial Penyedia Layanan Internet Menggunakan Algoritma XGBOOST," 2022.

[15] M. R. Givari, R. Mochamad, and Y. U. Sulaeman[2], "Perbandingan Algoritma SVM, Random Forest Dan XGBoost Untuk Penentuan Persetujuan Pengajuan Kredit," vol. 16, no. 1, 2022, [Online]. Available: https://journal.uniku.ac.id/index.php/ilkom