Jurnal Pseudocode 11(2) September 2024, hal. 79-88

ISSN 2355-5920 | e-ISSN 2655-1845
https://ejournal.unib.ac.id/pseudocode/index

Penilaian Pembayaran Kredit dengan *Logistic Regression* dan *Random Forest* pada Home Credit

Titin Yulianti[a,*], Amanda Hasna Cahyana[a], Muhamad Komarudin[a], Yessi Mulyani[a], Hery Dian Septama[a]

[a]*Prodi Teknik Informatika, Jurusan Teknik Elektro, Fakultas Teknik, Universitas Lampung*

Informasi Naskah:
Diterima: 30 Agustus 2024/ Direview: 31 Agustus 2024/ Direvisi: 17 September 2024/ Disetujui Terbit: 25 September 2024
DOI: 10.33369/pseudocode.11.2.79-88
*Korespondensi: titin.yulianti@eng.unila.ac.id

Abstract

Global economic growth has resulted in increasingly complex societal needs. Financial institutions aim to provide solutions to address these demands. However, the presence of problem loans poses a significant risk, prompting the use of classification techniques in data mining to mitigate this issue. This study develops a model to predict customers' credit payment capabilities, helping financial institutions avoid problematic loans. The research employs the SMOTE resampling technique to address class imbalance and enhance credit assessments. Findings indicate that the model utilizing SMOTE outperforms the one without it in terms of AUC. Among the two machine learning algorithms analyzed that are Logistic regression and random forest, the Random forest model with SMOTE demonstrates the best performance, achieving an accuracy of 90%, precision of 92%, recall of 88%, F1-score of 90%, and an AUC of 0.97. The optimal model identifies ten key features influencing credit repayment assessments: normalized scores from external sources, customer number change periods, previous installment payment counts, customer age, registration duration, credit application periods at the bureau, identity document update timelines, information update periods at the bureau, and employment duration. Additionally, the study creates visual dashboards to enhance the credit repayment assessment process.

Keywords: Prediction; Logistic Regression; Random Forest; Credit; Repayment Capabilities.

1. Pendahuluan

Uang dibutuhkan dalam memenuhi beragam kebutuhan, mulai dari kebutuhan primer hingga tersier. Namun, adanya keterbatasan finansial dapat menyebabkan terhambatnya pemenuhan kebutuhan. Lembaga keuangan, seperti Home Credit, memberikan fasilitas untuk memenuhi kebutuhan kompleks masyarakat melalui pemberian layanan kredit. Meskipun kredit dapat menjadi sumber pendapatan bagi lembaga keuangan, namun kredit yang bermasalah justru menjadi risiko bagi lembaga keuangan. Data statistik perbankan Indonesia menunjukkan rasio kredit bermasalah jenis konsumsi pada Februari 2023 yang telah mencapai 5,45%. Angka tersebut mengalami kenaikan sebesar 1.54% dari Desember 2022 dan telah melewati ambang batas yang ditentukan oleh Peraturan Bank Indonesia. PT Home Credit Indoensia dalam laporannya menyebutkan bahwa salah satu langkah penanganan risiko pembiayaan/kredit yaitu dengan perbaikan proses *credit scoring* dengan mempertimbangkan informasi eksternal yang berfokus pada pembuatan profil pelanggan yang akurat [1]. Oleh karena itu, analisis pemberian kredit dengan menggunakan data mining diperlukan sebagai mitigasi risiko terkait penunggakan dalam pembayaran.

Penelitian ini mengacu kepada beberapa penelitian sebelumnya yang relevan dan dapat digunakan sebagai referensi. Penelitian sejenis pernah dilakukan untuk klasifikasi dalam diagnosis penyakit hepatitis dengan menggunakan model klasifikasi *Logistic Regression* dan *Naive Bayes*. Penelitian tersebut memprediksi kondisi pasien dan hasilnya menunjukkan bahwa model yang dibangun dengan algoritma *Logistic Regression* memperoleh nilai akurasi dan AUC lebih tinggi dalam melakukan diagnosis penyakit hepatitis [2]. *Logistic regression* juga digunakan dalam penelitian yang melakukan pengelompokkan *Tweet* ke dalam beberapa kelas emosi yang telah ditentukan. Performa algoritma Logistic Regression dibandingkan dengan Random Forest yang makan pada keduanya dikenakan *SMOTE* untuk mengatasi ketidakseimbangan data. Hasil penelitian ini menunjukkan bahwa dengan menggunakan model klasifikasi *Logistic Regression* dan *Naive Bayes* [3].

SMOTE merupakan salah satu metode yang banyak digunakan untuk mengatasi ketidakseimbangan kelas pada data yang memberikan hasil klasifikasi yang lebih baik. Pada penelitian yang mengklasifikasikan kemungkinan pelanggan *churn*, dilakukan perbandingan performa *Logistic regression* dengan penggunaan *SMOTE* dan tanpa *SMOTE*. Hasilnya menunjukkan bahwa nilai akurasi model yang dibangun menggunakan *Logistic Regression* dengan *SMOTE* lebih baik dalam memprediksi kemungkinan pelanggan *churn* [4]. Selain itu, terdapat penelitian lain yang melakukan prediksi kegagalan pembayaran dalam pengembalian pinjaman menggunakan *Logistic Regression*, *SVM*, *Decision Tree*, dan *Random Forest* dengan *SMOTE*. Hasil penelitian tersebut

menyimpulkan bahwa masalah ketidakseimbangan data dapat diatasi menggunakan *SMOTE* dengan algoritma terbaik adalah *Random Forest* [5].

Selain itu, penelitian untuk diagnosis pasien penyakit hati dengan teknik *SMOTE* juga pernah dilakukan. Penelitian tersebut dibangun dengan menggunakan model klasifikasi *Naïve Bayes*, *KNN*, *Random Forest*, dan *SVM* [6]. Hasilnya menunjukkan bahwa model Random Forest memperoleh nilai akurasi dan *AUC* lebih tinggi dalam melakukan diagnosis penyakit hati. Penelitian lain untuk menilai pengajuan kredit calon nasabah menggunakan *KNN*, *Random Forest*, *SVM*, dan *Multilayer Perceptron* (MLP) dengan *SMOTE* yang juga menunjukkan hasil model yang dibangun dengan *Random Forest* lebih baik dibandingkan algoritma lainnya [7].

Hasil penelitian yang menggunakan metode Chi-Square untuk memilih atribut penting dari dataset application train/test, bureau, dan previous application menunjukkan bahwa model yang dikembangkan dengan algoritma Random Forest mencapai tingkat akurasi yang lebih tinggi dalam menilai pengajuan kredit nasabah [8].

Berdasarkan penelitian-penelitian tersebut diketahui bahwa model yang paling efektif yaitu *logistic regression* dan *random forest* tergantung dari data yang digunakan. Sementara itu, metode yang banyak digunakan untuk mengatasi ketidakseimabngan data dan menunjukkan peningkatan pada hasil klasifikasi yaitu metode *SMOTE*. Dari Oleh karena itu, penelitian ini bertujuan untuk menerapkan *SMOTE* dalam penilaian kemampuan pembayaran dengan menggunakan data Home Credit *Default Risk* dari Kaggle [9]. Model dibangun dengan *logistic regression* dan *random forest* untuk menentukan model yang paling efektif dalam penilaian kemampuan pembayaran kredit. Kemudian dilakukan pembuatan dashboard untuk visualisasi atribut-atribut yang mempengaruhi hasil penilaian kemampuan pembayaran kredit dengan model *machine learning* terbaik menggunakan Google Looker Studio.

2. Metodologi Penelitian

CRISP-DM, yang merupakan singkatan dari *Cross Industry Standard Process for Data Mining*, adalah sebuah standar yang diakui untuk data mining yang digunakan untuk menganalisis strategi penyelesaian masalah di berbagai sektor. Penerapan CRISP-DM melibatkan enam langkah, yaitu pemahaman bisnis, pemahaman data, persiapan data, pemodelan, evaluasi, dan penerapan. Proses klasifikasi untuk menilai kemampuan pembayaran kredit nasabah menggunakan metode CRISP-DM dapat dilihat pada Gambar 1.

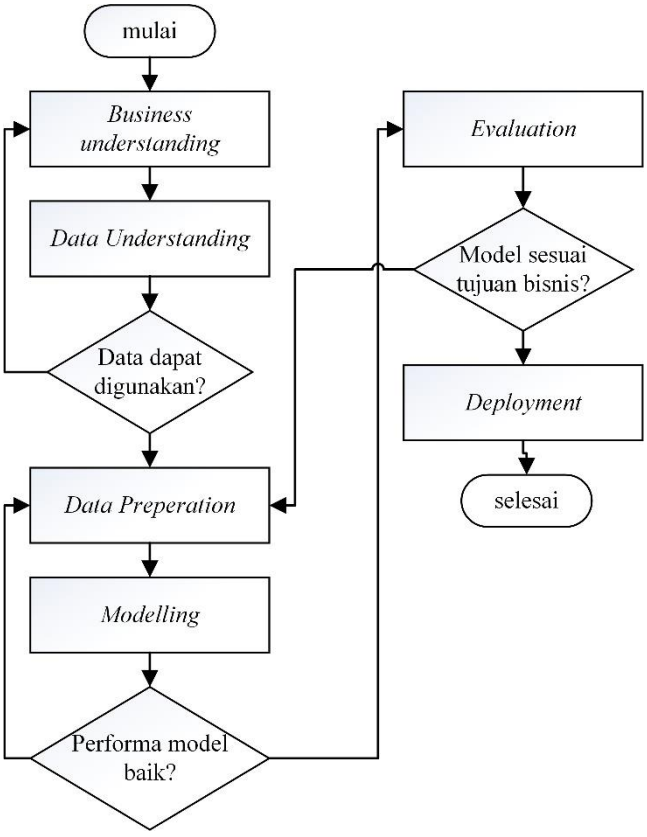
Tahap pertama dalam *business undertanding* bertujuan untuk mengidentifikasi tujuan bisnis serta tujuan dari proses *data mining* yang akan dilakukan. Selanjutnya, dalam tahap pemahaman data, dilakukan pengumpulan, deskripsi, eksplorasi, dan evaluasi kualitas data. Pada tahap persiapan data, aktivitas yang dilakukan meliputi penanganan *missing value* dan *outliers*, normalisasi data, dan pengurangan data untuk mempersiapkan dataset agar siap diproses oleh komputer. Pengurangan data adalah proses yang mengurangi dimensi dataset sambil tetap mempertahankan hasil analisis yang optimal. Dalam tahap ini, fitur atau atribut yang memiliki hubungan rendah terhadap akurasi model akan dihilangkan atau direduksi [10].

Pengurangan dimensi dapat dilakukan melalui proses seleksi fitur. Dalam penelitian ini, metode yang digunakan untuk pemilihan fitur adalah korelasi. Korelasi adalah metode statistik yang digunakan untuk mengukur sejauh mana hubungan antara satu variabel dengan variabel lainnya, tanpa mempertimbangkan apakah salah satu variabel bergantung pada yang lainnya [11].

Proses pemilihan fitur menggunakan metode korelasi menghasilkan nilai dalam kisaran -1 sampai 1. Nilai yang positif mengindikasikan adanya hubungan searah antara dua variabel, sedangkan nilai negatif menunjukkan adanya hubungan yang berlawanan di antara keduanya. Nilai *correlation* antara dua variabel, X dan Y, dihitung dengan persamaan matematis berikut :

$$r = \frac{\sum_{i=1}^{n}(X_i - \hat{X})(Y_i - \widehat{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \hat{X})^2 (Y_i - \widehat{Y})^2}} \quad (1)$$

X_i dan Y_i merupakan nilai sampel ke-i dari X dan Y, sementara $\hat{X}$ dan $\hat{Y}$ adalah rata-rata sampel X dan Y, jumlah sampe dinyatakan dengan n.



Gbr. 1 Diagram alir *CRISP-DM*.

Tahapan dalam pemodelan mencakup pemilihan algoritma machine learning, pembagian data, serta proses pembuatan model. Dalam penelitian ini, algoritma yang digunakan adalah *Logistic regression* dan *Random forest.*

Setelah itu, langkah berikutnya adalah membagi data. Atribut yang telah dipilih pada tahap persiapan data akan dibagi menjadi dua variabel baru. Pembagian ini dilakukan

dalam dua skenario: yang pertama tanpa menerapkan SMOTE, dan yang kedua dengan penerapan SMOTE.

SMOTE adalah metode *oversampling* yang dirancang untuk mengatasi masalah ketidakseimbangan data dengan menghasilkan data sintetis dari kelas yang kurang representatif. Proses ini melibatkan pengurangan vektor fitur dari kelas minoritas dengan nilai dari *nearest neighbor* di kelas yang sama. Selanjutnya, selisih yang dihasilkan dikalikan dengan angka acak antara 0 hingga 1. Hasil perkalian ini kemudian ditambahkan ke vektor fitur awal untuk menghasilkan vektor baru [12]. Berikut adalah persamaan matematis proses SMOTE :

$$X_{new} = X_i + (\hat{X}_l - X_i)\,\delta \quad (2)$$

X_i adalah vektor dari fitur pada kelas minoritas, $\hat{X}_l$ *k-nearest neigbors* untuk X_i, dan δ koefisien dengan nilai antara 0 dan 1.

Data tersebut selanjutnya digunakan untuk melatih dan menguji model dengan menggunakan *logistic regression* dan *random forest*. *Logistic regression* adalah model linier yang umum digunakan dalam klasifikasi. Berbeda dari *linear regression* yang fokus pada prediksi variabel numerik, *logistic regression* digunakan untuk klasifikasi variabel respons yang bersifat kategorikal dengan memanfaatkan fungsi logistik. Algoritma ini berfungsi untuk menganalisis pola hubungan antara sekumpulan variabel independen dan variabel respons.

Variabel independen dapat berupa data kategorik atau numerik, sementara variabel respons bersifat kategorikal. Oleh karena itu, logistic regression sangat cocok untuk diterapkan pada variabel dependen yang bersifat kategorikal atau biner. Algoritma ini mampu menggali hubungan antara variabel dependen dan independen, yang dapat berupa atribut nominal, ordinal, atau rasio. [13].

Logistic function yang digunakan dalam algoritma *logistic regression* ini didapatkan dengan mengganti nilai Y pada *linear function* dengan nilai Y pada *sigmoid function* yang mengubah bentuk *odds* menjadi logaritma. Berikut adalah rumus *sigmoid function* yang didapatkan:

$$P = \frac{1}{1+e^{-Y}} \quad (3)$$

Setelah mendapatkan *sigmoid function*, penyetaraan nilai dilakukan dengan mensubstitusikan nilai Y pada *linear function* ke dalam *sigmoid function* untuk menghasilkan probabilitas.

$$P = \frac{1}{1+e^{-(b_0+b_1x+b_2x_2+\cdots+b_px_p)}} \quad (4)$$

Algoritma kedua yaitu *Random Forest* yang merupakan algoritma *supervised learning* dan termasuk ke dalam model *ensemble*. Model *ensemble* merupakan model yang menggabungkan beberapa metode lain untuk meningkatkan nilai akurasi hasil proses klasifikasi.

Random Forest terdiri dari sejumlah decision tree yang bersama-sama membentuk sebuah hutan klasifikasi. Setiap pohon keputusan berfungsi berdasarkan nilai vektor acak yang diambil sebagai sampel secara independen dan merata di seluruh pohon dalam hutan. Saat melakukan klasifikasi, masing-masing pohon keputusan memberikan suara untuk kelas yang paling banyak didukung [14].

Random Forest adalah hasil pengembangan dari metode *Classification and Regression Tree* (*CART*) yang menerapkan teknik *bagging* serta seleksi fitur acak. Proses klasifikasi dengan metode *Random Forest* terdiri dari dua tahap. Tahap pertama melakukan pembentukan 'k' pohon untuk membentuk hutan acak, sementara tahap kedua melakukan klasifikasi dengan menggunakan hutan acak yang telah terbentuk. Proses pembangunan pohon keputusan menggunakan metode *CART* dilakukan dengan menggunakan perhitungan *information gain*. *Information gain* digunakan untuk memilih atribut pada setiap node dari pohon-pohon klasifikasi. Berikut adalah rumus perhitungan *information gain :*

$$Gain\,(S,A) = Entropy\,(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} * Entropy(S_i) \quad (5)$$

S adalah himpunan kasus dan A adalah fitur, $|S_i|$ Proporsi S_i terhadap S, sementara $|S|$ menyatakan jumlah kasus dalam S. Perhitungan *information gain* melibatkan perhitungan *entropy* sehingga untuk menentukan information gain nilai entropy harus dihitung terlebih dahulu dengan rumus :

$$Entropy\,(S) = \sum_{i=1}^{n} p_i \log_2(p_i) \quad (6)$$

dengan p_i adalah proporsi Proporsi S_i terhadap S.

Setelah pembuatan pohon keputusan selesai, proses pengambilan keputusan dilakukan dengan melakukan penghitungan suara untuk setiap hasil klasifikasi pada tiap pohon dan pemilihan hasil akhir dilakukan dengan memilih kelas yang paling banyak dipilih pada setiap pohon [14].

Kemudian, model *machine learning* yang berhasil dibuat akan diukur perfomanya pada tahap *evaluation* dengan parameter, *accuracy*, *precision*, *recall*, *F1-score*, dan *AUC*. Model dengan hasil evaluasi terbaik akan digunakan dalam proses prediksi dengan data baru. Terakhir, pada tahap *deployment*, hasil pembelajaran model digunakan untuk melakukan prediksi pada data pengajuan nasabah yang kemudian disebarkan melalui pembuatan *dashboard* menggunakan Google Looker Studio.

3. Hasil dan Pembahasan

3.1. *Bussiness Understanding*

Home Credit merupakan sebuah perusahaan keuangan yang menyediakan layanan pemberian kredit konsumen. Perusahaan ini memberikan kesempatan pada calon nasabah untuk mengajukan pinjaman kredit dengan jangka waktu terntentu. Kesempatan tersebut menjadi tantangan utama yang dihadapi Home Credit karena jika terjadi kekeliruan dalam

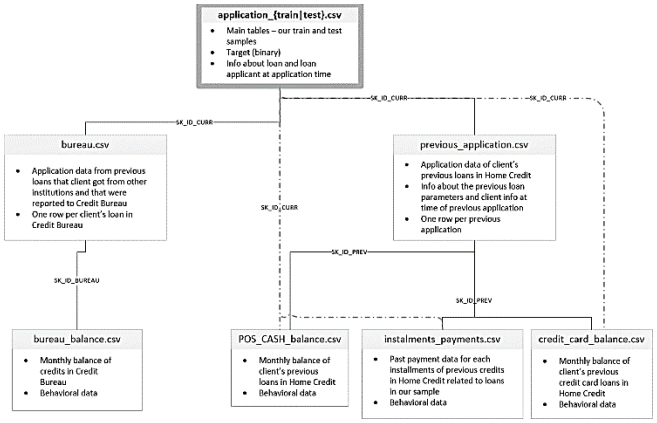
identifikasi calon nasabah maka dapat menyebabkan risiko kredit bermasalah. Dalam menghadapi masalah ini, Home Credit menggunakan *data mining* untuk mengidentifikasi atribut-atribut yang berpengaruh terhadap penilaian kemampuan pembayaran kredit.

Penelitian ini bertujuan mengembangkan model klasifikasi menggunakan algoritma *machine learning*, yaitu *Logistic Regression* dan *Random Forest* untuk memprediksi kemampuan pembayaran nasabah. Selain itu, penelitian ini juga akan menghasilkan *dashboard* visualisasi dengan Google Looker Studio yang dapat membantu Home Credit dalam melihat faktor-faktor yang memengaruhi penilaian kredit secara lebih efisien. Pendekatan ini diharapkan dapat membantu keberlanjutan bisnis Home dan mengurangi risiko terjadinya kredit bermasalah.

3.2. *Data Understanding*

Pada tahap ini dilakukan beberapa langkah yang meliputi pengumpulan, pendeskripsian, eksplorasi, dan pemeriksaan kualitas data. Data yang digunakan merupakan data telekomunikasi dan transaksi calon nasabah Home Credit yang diperoleh dari sumber internal Home Credit dan biro kredit. Penelitian ini menggabungkan tujuh dataset yang tersedia yaitu *bureau*, *bureau balance*, data *application train/test*, *POS CASH balance*, *credit card balance*, *installments payments*, dan *previous application*.

Ketujuh data tersebut memiliki relasi yang dihubungkan melalui atribut ID yang bernama SK_ID_CURR, SK_ID_BUREAU, dan SK_ID_PREV. Hal tersebut yang memungkinkan penggunaan informasi dari semua dataset untuk keputusan klasifikasi dengan lebih baik. Relasi antara tiap dataset tersebut ditampilkan pada Gambar 2.



Gbr. 2 Sumber data dan relasi antar-dataset [8].

Pada tahapan *data understanding*, dilakukan beberapa kegiatan seperti pengumpulan, pendeskripsian, eksplorasi, dan pemeriksaan.

Proses selanjutnya adalah pendeskripsian data yang digunakan untuk memberikan pemahaman lebih baik tentang gambaran dan kondisi dataset dalam penelitian.

1) Pada dataset *application train* terdapat 307.511 *record* dan 122 atribut, dengan 65 atribut bertipe data float, 41 atribut bertipe data integer, dan 16 atribut bertipe data *object*. Sementara itu, dataset *application test* terdapat 48.744 *record* dan 121 atribut. Dataset ini menjelaskan mengenai data nasabah yang mengajukan pinjaman kredit.
2) Dataset *bureau* terdapat 1.716.428 *record* dan 17 atribut dengan 8 atribut bertipe data float, 6 atribut bertipe data integer, dan 3 atribut bertipe data *object*. Dataset *bureau* menjelaskan mengenai pinjaman kredit calon nasabah sebelumnya yang ada pada lembaga keuangan lain dan telah dilaporkan ke biro kredit.
3) Dataset *bureau balance* memiliki 27.299.925 *record* dan 3 atribut dengan 2 atribut bertipe integer dan 1 atribut bertipe data *object*. Dataset tersebut menjelaskan mengenai saldo bulanan pinjaman kredit sebelumnya di biro kredit.
4) Dataset *previous application* terdapat 1.670.214 *record* dan 37 atribut dengan 15 atribut bertipe data float, 6 atribut bertipe data integer, dan 16 atribut bertipe data *object*. Dataset *previous application* menjelaskan mengenai pengajuan pinjaman kredit sebelumnya yang telah dilakukan calon nasabah pada Home Credit.
5) Dataset *POS cash balance, credit card balance*, dan *installments payments* menjelaskan informasi riwayat kredit nasabah yang pernah ada di Home Credit yang memuat informasi kartu kredit dan pembayaran cicilan. Pada *POS cash balance* terdapat 10.001.358 *record* dan 8 atribut dengan 2 atribut bertipe data float, 5 atribut bertipe data integer, dan 1 atribut bertipe data *object*.
6) Dataset *credit card balance* terdapat 3.840.312 *record* dan 23 atribut dengan 15 atribut bertipe data float, 7 atribut bertipe data integer, dan 1 atribut bertipe data *object*.
7) Dataset *installments payments* terdapat 13.605.401 *record* dan 8 atribut dengan 5 atribut bertipe data float dan 3 atribut bertipe data integer.

Tahapan selanjutnya adalah eksplorasi data yang dilakukan dengan menggunakan visualisasi untuk membantu memahami hubungan, menemukan pola, dan menampilkan data agar dapat lebih mudah diinterpretasikan. Seluruh dataset divisualisasikan pada bagian ini, kecuali dataset application test. Hal tersebut dilakukan karena dataset tersebut tidak memuat label TARGET dan hanya akan digunakan untuk prediksi setelah model terbaik dipilih. Kemudian, tahapan pemeriksaan kualitas data dilakukan dengan mengecek missing values. Dari semua dataset hanya dataset bureau balance yang tidak terdapat missing values. Jumlah missing values pada masing-masing dataset lainnya adalah 73 atribut untuk dataset application train, 7 atribut untuk dataset bureau, 16 atribut untuk dataset previous application, 2 atribut untuk dataset POS Cash balance serta installments payments, dan 9 atribut untuk credit card balance.

Selain melakukan pengecekan *missing values*, proses pemeriksaan kualitas data juga dilakukan dengan mengecek nilai kunci penggabungan yang duplikat pada setiap record. Hal tersebut perlu dilakukan untuk mengoptimalkan kinerja *machine learning* dalam mempelajari pola-pola setiap atribut saat melakukan klasifikasi. Proses pengurangan *record* dilakukan dengan mengurutkan data berdasarkan salah satu atribut dan hanya mengambil satu record pada setiap ID yang

duplikat untuk tiap dataset.

Hasil dari pengurangan record yang telah dilakukan memberikan perubahan dimensi pada setiap dataset. Perubahan dimensi pada setiap dataset yang mengalami proses pengurangan record dapat dilihat pada tabel I.

Tabel I Hasil pengurangan record setiap dataset

No	Nama Dataset	Jumlah Record Setelah Penghapusan
1.	Bureau	305.811
2.	Bureau Balance	817.395
3.	Previous Application	338.857
4.	POS cash balance.	936.325
5.	Credit Card Balance	104.307
6.	Instalments Payments	997.752

3.3. *Data Preparation*

Proses pertama yang dilakukan pada data preparation adalah pengintegrasian data. Proses integrasi data ini dilakukan dengan menggabungkan seluruh sumber data yang dimiliki menjadi sebuah DataFrame, seperti yang ditunjukkan oleh tabel II. Proses pengintegrasian data dilakukan dengan left join keenam dataset dengan dataset application train. Pengintegrasian ini dilakukan dengan menggunakan ID penggabungan yang ada pada tiap dataset seperti pada Tabel II.

Tabel II integrasi *application train* pada setiap dataset

No	**Nama Dataset**	**ID yang Digunakan**	**Dimensi sesudah pengintegrasian**
1	*Bureau, dan Bureau Balance*	SK_ID_CURR, SK_ID_BUREAU	DF1 : 307.511 baris dan 140 atribut
2	*Previous Application*	SK_ID_CURR	DF2 : 307.511 baris dan 176 atribut
3	*Credit Card Balance*	SK_ID_PREV	DF3 : 307.511 baris dan 197 atribut
4	*POS cash balance.*	SK_ID_PREV	DF4 : 307.511 baris dan 204 atribut
5	*Instalments Payments*	SK_ID_PREV	DF5 : 307.511 baris dan 210 atribut

Selanjutnya adalah proses penghapusan data duplikat yang dilakukan untuk mengurangi dimensi data sehingga pemakaian memori akan lebih efisien. Selain itu, proses perubahan nama pada atribut dengan nama duplikat juga dilakukan untuk menghindari ambiguitas. Dalam penelitian ini proses penghapusan data duplikat dilakukan setelah proses penggabungan data dilakukan, hasil akhir dari proses penghapusan data duplikat dan nama kolom duplikat disimpan ke dalam variabel dfjoin yang memiliki dimensi 307.511 *record* dan 209 atribut.

Pada tahapan *data preparation* juga dilakukan penanganan pada atribut yang memiliki format tidak sesuai. Dalam penelitian ini proses untuk menangani masalah ketidaksesuaian format dilakukan pada atribut yang bertipe data *object*. Seluruh atribut bertipe data object memiliki beberapa nilai *unique* di dalamnya. Dari 40 atribut yang bertipe data *object*, terdapat 12 atribut yang memiliki nilai XNA. Nilai XNA tersebut merujuk pada suatu nilai yang tidak diketahui isinya. Oleh karena itu, pada penelitian ini nilai XNA diisi dengan nilai *nan* yang menandakan bahwa nilai tersebut tidak ditentukan atau ditandai sebagai nilai hilang.

Tahapan *data preparation* selanjutnya adalah penanganan *missing values*. Proses penanganan *missing values* dilakukan pada 155 atribut. Pada penelitian ini, penanganan *missing values* dilakukan dengan dua cara, yaitu dengan penghapusan atribut yang memiliki *missing values* dan dengan pengisian nilai pada atribut yang memiliki *missing values*. Proses penghapusan dilakukan pada atribut dengan jumlah *missing values* yang lebih dari 50%. Pada penelitian ini terdapat 74 atribut dengan *missing values* lebih dari 50% sehingga dimensi DataFrame berubah menjadi 307.511 baris dan 135 atribut. Penanganan missing values yang kedua dilakukan dengan melakukan pengisian nilai pada atribut yang memiliki *missing values*. Terdapat 81 atribut yang mengalami proses imputasi. Proses imputasi dilakukan dengan menggunakan nilai median, mean, dan modus. Pengisian nilai dengan median dan mean dilakukan ketika atribut bertipe data integer sedangkan pengisian nilai dengan modus dilakukan ketika atribut bertipe data *object*. Imputasi atribut dengan nilai dengan *mean* atau rata-rata dilakukan ketika data memiliki distribusi normal atau tidak memiliki *outliers*. Pengisian nilai *missing values* dengan cara ini dapat menjaga pusat massa data tetap konstan karena tidak mengubah varians dari dataset asli sehingga karakteristik sebaran data dapat dipertahankan. Sementara itu, imputasi atribut dengan nilai *median* dilakukan ketika data memiliki distribusi miring atau mengandung *outliers*. Pengisian nilai dengan *median* dipengaruhi oleh perubahan dalam urutan nilai sehingga nilai ini tidak terpengaruh oleh nilai-nilai ekstrem pada *outliers* yang dapat mengubah bentuk data.

Kemudian tahapan penanganan atribut yang memiliki *outliers* dilakukan, tahapan ini akan mengisi nilai *outliers* menggunakan *upper bound* dan *lower bound*. Pengisian nilai menggunakan *upper bound* dilakukan ketika nilai berada di atas ambang batas sedangkan *lower bound* dilakukan ketika nilai berada di bawah ambang batas. Dalam penelitian ini penanganan atribut yang memiliki *outliers* hanya dilakukan pada 30 atribut bertipe data numerik. Sementara atribut numerik lain tidak dilakukan penanganan karena beberapa atribut merupakan nilai yang telah dinormalisasi dan beberapa atribut memiliki rentang nilai yang tidak terlalu tinggi. Selanjutnya proses pengodean data dilakukan karena pada algoritma *machine learning* hanya dapat menganalisis nilai input yang berupa bilangan numerik bukan *object*. Oleh karena itu, atribut dengan tipe data object harus mengalami

pengodean menjadi bilangan numerik agar dapat dianalisis. Penelitian ini menggunakan *dummy encoding* untuk mengurangi penggunaan dimensi data karena pada dummy *encoding* tidak semua *unique values* mengalami pengodean menjadi atribut baru, tetapi akan dikurangi satu *unique value* yang akan dijadikan sebagai referensi. Hasil dari *dummy encoding* tersebut akan mengubah dimensi *DataFrame* yang telah mengalami *cleaning* menjadi 307.511 baris dan 326 atribut. Dari 326 atribut yang telah mengalami pengodean, 104 atribut memiliki tipe data numerik dan 222 atribut memiliki tipe data boolean.

Proses normalisasi pada penelitian ini juga dilakukan dengan melakukan *scaling* pada atribut yang bertipe data numerik. Proses *scaling* dilakukan agar tiap variabel numerik memiliki skala atau jangkauan nilai yang sama. *Scaling* dilakukan menggunakan *MinMAxScaler*() yang mengubah skala nilai pada setiap atribut menjadi 0 sampai dengan 1. Proses *scaling* hanya dilakukan untuk atribut yang bertipe data numerik dengan mengecualikan beberapa atribut yang nilainya telah dinormalisasi.

Proses terakhir dalam *data preparation* adalah *Feature selection* yang dilakukan untuk memilih atribut paling relevan dan signifikan dalam memprediksi variabel TARGET. Dalam melakukan pemilihan *feature*, kunci penggabungan yang nilainya *unique* pada setiap *record* perlu dihapus karena atribut tersebut hanya sebagai pengidentifikasi tiap *record* sehingga tidak memberikan informasi signifikan. Proses pemilihan atribut tersebut diambil dari atribut dengan nilai korelasi tertinggi, yaitu nilai yang mendekati 1 dan -1. Penelitian ini melakukan pemilihan *feature* dengan menerapkan *threshold* nilai korelasi, yaitu hanya mengambil nilai korelasi yang berada di atas 0.04 dan di bawah -0.04. Hasil atribut yang terpilih sebagai *feature* dalam pembuatan model penilaian kemampuan membayar kredit dapat dilihat pada tabel III. Jumlah atribut yang akan dipilih untuk pemodelan adalah 24 atribut.

Tabel III Atribut yang terpilih sebagai *feature* pemodelan

No	Nama Atribut	Nilai Korelasi
1.	EXT_SOURCE_2	-0.160
2.	EXT_SOURCE_3	-0.156
3.	NAME_EDUCATION_TYPE_Higher education	-0.057
4.	CREDIT_ACTIVE_Closed	-0.048
5.	NAME_INCOME_TYPE_Pensioner	-0.046
6.	DAYS_EMPLOYED	-0.045
7.	DAYS_FIRST_DRAWING	-0.042
8.	AMT_PAYMENT_INSTALMENTS	-0.040
9.	DAYS_REGISTRATION	0.042
10.	OCCUPATION_TYPE_Laborers	0.043
11.	FLAG_DOCUMENT_3	0.044
12.	REG_CITY_NOT_LIVE_CITY	0.044
13.	DAYS_CREDIT_UPDATE	0.045
14.	FLAG_EMP_PHONE	0.046
15.	NAME_EDUCATION_TYPE_Secondary / secondary special	0.050
16.	REG_CITY_NOT_WORK_CITY	0.051
17.	DAYS_ID_PUBLISH	0.051
18.	CODE_GENDER_M	0.055
19.	DAYS_LAST_PHONE_CHANGE	0.055
20.	DAYS_CREDIT	0.056
21.	NAME_INCOME_TYPE_Working	0.057
22.	REGION_RATING_CLIENT	0.059
23.	REGION_RATING_CLIENT_W_CITY	0.061
24.	DAYS_BIRTH	0.078

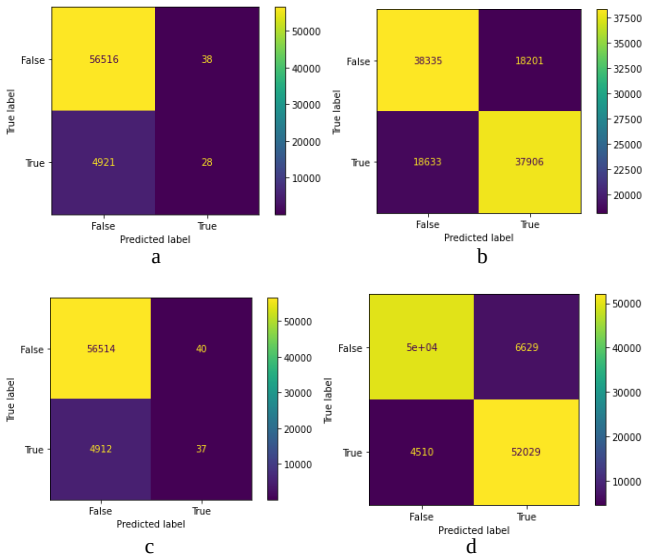
3.4. *Modeling*

Algoritma yang digunakan dalam membuat model penilaian kemampuan pembayaran kredit adalah algoritma *Logistic Regression* dan *Random Forest*. Pada masing-masing algoritma diterapkan dua kondisi yang berbeda, yaitu dengan menerapkan *SMOTE* dan tanpa menerapkan *SMOTE*. Hal tersebut dilakukan untuk melihat perbandingan perfoma kedua model algoritma dengan dua kondisi yang diterapkan. Proses selanjutnya dalam modeling adalah pembagian data. Data yang telah diolah pada data *pereparation* akan dibagi menjadi dua bagian, yaitu data *training* dan *testing*. Pembagian data dilakukan dengan membagi data sebanyak 80% untuk *training* yang digunakan dalam melatih model dan 20% untuk *testing* yang digunakan dalam menguji kinerja model yang telah dibuat. Hasil pembagian data tersebut dapat dilihat pada tabel IV.Proses pertama yang dilakukan pada *data preparation* adalah pengintegrasian data.

Tabel IV Pembagian data *training* dan *testing*

Keterangan	*Training*		*Testing*	
Kondisi	**Tanpa *SMOTE***	**Dengan *SMOTE***	**Tanpa *SMOTE***	**Dengan *SMOTE***
Target 1	19.876	226.147	4.949	56.539
Target 0	226.132	226.150	56.554	56.536
Jumlah	246.008	452.297	61.503	113.075

Hasil pemodelan dengan kedua algoritma *machine learning* ditampilkan pada gambar 3. Gambar tersebut menunjukkan hasil prediksi model tanpa menerapkan metode *SMOTE* dan dengan menerapkan metode *SMOTE* dalam melakukan klasifikasi. Hasil prediksi tersebut ditampilkan ke dalam *confussion matrix* yang menunjukkan hasil data prediksi dengan data sebenarnya. Gambar 3 menunjukkan bahwa model *Random Forest* selalu memiliki jumlah data yang diprediksi secara tepat lebih besar. Hal tersebut terlihat dari nilai *True Negative* dan *True Positive Random Forest* yang lebih besar baik tanpa *SMOTE* maupun dengan *SMOTE*. Kedua nilai tersebut menunjukkan jumlah data yang berhasil diprediksi model dengan tepat dan sesuai data sebenarnya.



Gbr. 3 *Confussion Matrix* (a. *Logistic Regression* tanpa *SMOTE*, b. *Logistic Regression* dengan *SMOTE*, c. *Random Forest* tanpa *SMOTE*, d. *Random Forest* dengan *SMOTE*).

3.5. *Evaluation*

Evaluasi diakukan untuk menilai seberapa baik model *machine learning* yang telah dibangun dalam menilai kemampuan pembayaran kredit nasabah. Adapun metrik evaluasi yang digunakan yaitu *accuracy*, *precision*, *recall*, dan *F1-Score*. Selain itu, *Area Under the ROC Curve* atau *AUC* juga digunakan dalam evaluasi yang menampilkan tingkat keberhasilan dengan memisahkan data positif.

Hasil pengukuran kinerja tanpa proses *SMOTE* menunjukkan bahwa model klasifikasi yang dibangun tanpa menerapkan metode *SMOTE* dengan menggunakan *Logistic Regression* memiliki nilai *accuracy*, *precision*, *recall*, *F1-score* sama dengan model yang dibangun dengan algoritma *Random Forest*. Namun, pada evaluasi nilai *AUC* dari *Logistic Regression* memiliki nilai yang lebih besar, yaitu 0.73. Nilai *AUC* model tersebut masuk ke dalam kategori *fair classification*. Jika dilihat dari *Confusion Matrix*, model *Random Forest* memiliki nilai jumlah prediksi yang benar atau sesuai dengan nilai aslinya pada klasifikasi lebih banyak dibandingkan model *Logistic Regression*.

Tabel V Evaluasi performa model *Logistic Regression (LR)* dan *Random Forest (RF)*

Model	*Accuracy*	*Precision*	*Recall*	*F1 Score*	*AUC*
Tanpa *SMOTE*					
LR	92%	92%	100%	96%	0.73
RF	92%	92%	100%	96%	0.72
Dengan *SMOTE*					
LR	67%	67%	68%	68%	0.74
RF	90%	92%	88%	90%	0.97

Kemudian pada hasil pengukuran kinerja dengan proses *SMOTE* juga menunjukkan bahwa model klasifikasi yang dibangun dengan menerapkan metode *SMOTE* dan menggunakan *Random Forest* memiliki nilai *accuracy*, *precision*, *recall*, *F1-score* lebih tinggi dibandingkan model yang dibangun dengan algoritma *Logistic Regression*. Pada evaluasi dengan nilai *AUC*, model klasifikasi yang dibangun dengan *Random Forest* juga memiliki nilai yang lebih besar, yaitu 0.97. sehingga masuk ke dalam kategori *excellent classification*. Sementara, model Logistic Regression hanya memiliki nilai *AUC* sebesar 0.74.

Sementara itu, evaluasi terkait waktu pemodelan dan prediksi masing-masing metode menunjukkan bahwa dengan SMOTE membutuhkan waktu lebih lama dibandingkan tanpa SMOTE.

Tabel VII Pengukuran waktu pemodelan dan prediksi

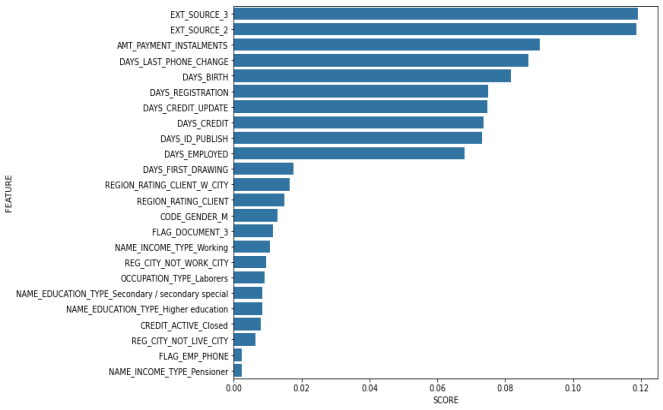
Model	*Waktu Pemodelan*	*Waktu Prediksi*
Tanpa *SMOTE*		
LR	4.9 detik	0.1 detik
RF	50.7 detik	1.7 detik
Dengan *SMOTE*		
LR	15.2 detik	0.3 detik
RF	322.7 detik	11.1 detik

Dari hasil evaluasi kinerja pemodelan disimpulkan bahwa model *Random Forest Classifier* merupakan model terbaik. Meskipun waktu yang diperlukan dalam pemodelan pada algoritma *random forest* lebih lama, hal tersebut bukan berarti dapat langsung menunjukkan bahwa model tersebut memiliki kualitas yang rendah. Lamanya waktu pemodelan bisa terjadi karena proses pembelajaran yang lebih mendalam dan kompleks terhadap fitur terutama pada data yang berukuran besar. Hal tersebut dapat dilihat dari nilai *AUC Random Forest* yang lebih besar dari algoritma *Logistic Regression* baik tanpa *SMOTE* atau dengan *SMOTE*. Pada algoritma *Random Forest* dengan *SMOTE*, nilai *AUC* yang dihasilkan

merupakan nilai paling besar. Oleh karena itu, model algoritma *Random Forest* yang menerapkan metode *SMOTE* dengan atribut-atribut pada proses pemilihan *feature* dapat diterapkan dalam melakukan penilaian kemampuan pembayaran kredit nasabah di Home Credit.

3.6. *Deployment*

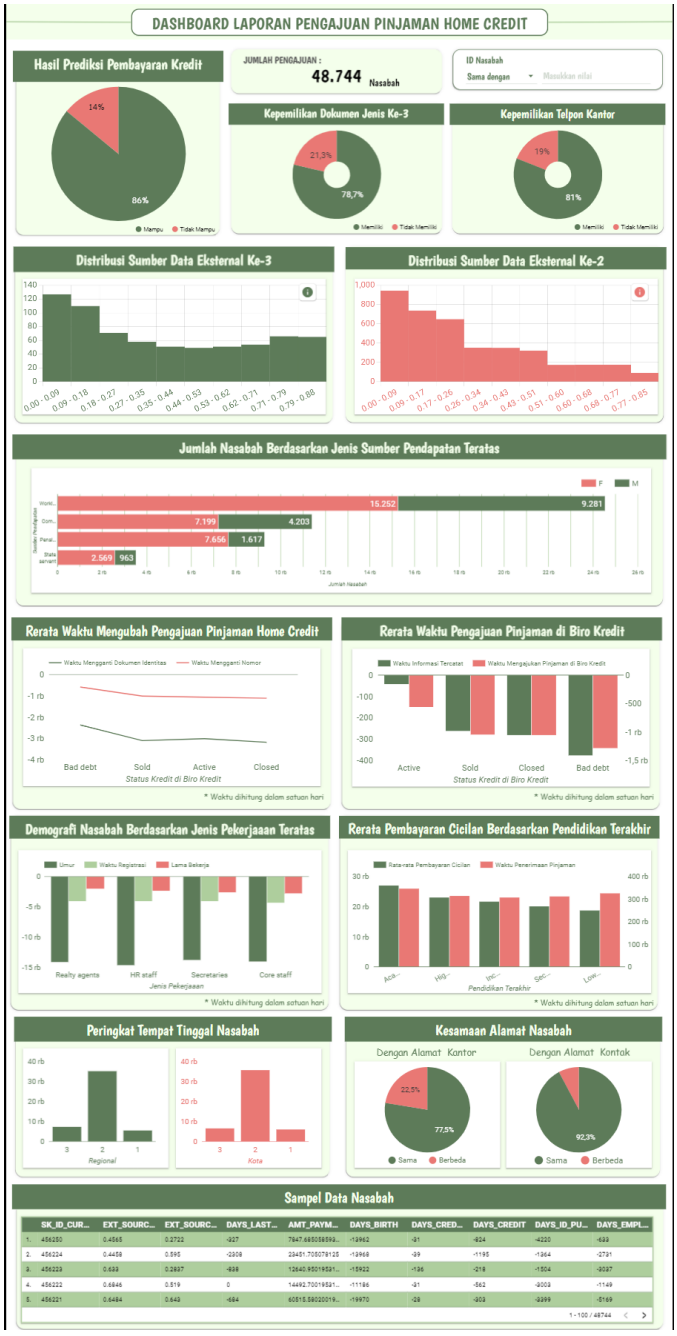
Proses pertama *deployment* dilakukan dengan pemilihan *feature importance* dalam mengklasifikasi kemampuan pembayaran kredit. Tahapan ini membantu Home Credit dalam mengetahui faktor-faktor yang paling signifikan dalam menentukan kemampuan nasabahnya untuk membayar kredit. Dari model *Random Forest* yang dibangun menggunakan 24 atribut dan dipilih sepuluh atribut paling berkontribusi dalam menentukan hasil prediksi penilaian kemampuan pembayaran kredit yang diambil berdasarkan skor *feature importance* atribut yang berada di atas 0.05 seperti yang terlihat pada Gambar 4. Sepuluh *feature* tersebut adalah atribut skor yang dinormalisasi dari sumber data eksternal ke-3 (EXT SOURCE 3), skor yang dinormalisasi dari sumber data eksternal ke-2 (EXT SOURCE 2),), jumlah pembayaran cicilan pada kredit sebelumnya di Home Credit (AMT PAYMENT INSTALMENTS), rentang waktu nasabah mengganti nomor DAYS LAST PHONE CHANGE), usia nasabah (DAYS BIRTH), waktu registrasi (DAYS REGISTRATION), waktu informasi nasabah diperbarui oleh biro kredit (DAYS CREDIT UPDATE), rentang waktu mengajukan kredit di Biro Kredit (DAYS CREDIT), rentang waktu nasabah mengganti dokumen identitas (DAYS ID PUBLISH), dan lama waktu nasabah bekerja (DAYS EMPLOYED). Dengan memfokuskan pada sepuluh fitur terbaik, diharapkan dapat lebih memberikan representasi faktor-faktor yang memengaruhi penilaian kemampuan pembayaran kredit.



Gbr. 4 *Feature Importance.*

Selanjutnya adalah proses *deployment* dengan *dashboard* yang dilakukan untuk menampilkan hasil prediksi dari klasifikasi. Pengembangan *dashboard* dilakukan dengan menggunakan dataset *application test* yang memiliki 121 atribut tanpa atribut TARGET. Namun, sebelum melakukan prediksi dengan data tersebut, dataset application test harus di-cleaning dengan proses yang sama seperti saat melakukan training pada data aplication train. Hla tersebut dilakukan agar dataset yang akan diprediksi memiliki dimensi yang sama dengan model Random Forest. Hasil akhir dari DataFrame yang digunakan memeililki dimensi 48.744 record dan atribut.

Proses selanjutnya adalah melakukan prediksi dengan model Random Forest. Hasil prediksi yang telah dilakukan dikembangkan ke dalam sebuah dashboardseperti pada Gambar 5. Dashboard tersebut memberikan informasi mengenai penilaian kelayakan nasabah yang mengajukan pinjaman. Pada gambar 5 dapat dilihat bahwa dashboard yang dikembangkan tersebut memiliki fungsi kontrol berupa filter lanjutan yang dapat digunakan untuk melakukan pencarian ID nasabah dari nasabah. Apabila ID yang diinputkan terdapat pada database maka akan ditampilkan informasi dari nasabah yang memuat hasil prediksi pembayaran, status kepemilikan dokumen, nilai external source, sumber pendapatan, dan informasi lain dari nasabah tersebut.



Gbr. 5 *Dashboard* laporan pengajuan pinjaman di Home Credit.

Berdasarkan hasil visualisasi pada *dashboard* laporan pengajuan pinjaman di Home Credit, terdapat beberapa saran dan rekomendasi yang dapat diterapkan dalam melakukan penilaian kemampuan pembayaran pinjaman nasabah. Pertama, Home Credit dapat lebih mempertimbangkan nilai sumber eksternal ke-3 dan ke-2 dalam proses penilaian kemampuan pembayaran. Kedua atribut ini menjelaskan skor yang dinormalisasi dari sumber data eksternal. Nasabah dengan nilai yang tinggi pada kedua atribut ini cenderung memiliki prediksi pembayaran kredit yang lebih baik. Kemudian, Home Credit dapat mempertimbangkan nasabah yang memiliki nilai pembayaran cicilan dalam jumlah besar pada kredit sebelumnya. Hal tersebut menunjukkan kemampuan nasabah yang tetap dapat mengembalikan pinjaman kreditnya meskipun kredit yang diberikan bernilai besar.

Selanjutnya, Home Credit dapat mempertimbangkan nasabah berdasarkan informasi waktu penggantian dokumen yang dilakukan. Jika nasabah sering melakukan penggantian dokumen pengajuan maka jarak waktu pergantian dokumen akan semakin dekat dengan waktu pengajuan. Hal tersebut dapat menunjukkan bahwa nasabah memberikan pinjaman berdasarkan informasi yang tidak dapat dipercaya. Oleh karena itu, pilihlah nasabah yang memiliki jangka waktu penggantian dokumen paling lama. Selain itu, Home Credit perlu mempertimbangkan usia dan lama durasi nasabah telah bekerja. Kedua faktor tersebut memberikan informasi tentang stabilitas keuangan dan kemampuan nasabah untuk membayar kembali pinjaman yang akan diberikan. Oleh karena itu, pilihlah nasabah dengan usia dan durasi bekerja yang semakin besar karena cenderung memiliki kemampuan mengembalikan pinjaman kredit.

Home Credit juga harus mempertimbangkan nasabah yang telah melakukan pendaftaran akun dari waktu lama. Waktu pendaftaran yang lama tersebut dapat memungkinkan Home Credit untuk memiliki pemahaman yang lebih baik tentang nasabah. Kemudian, Home Credit juga dapat mempertimbangkan informasi yang didapatkan dari biro kredit, seperti jarak waktuantara pengajuan pinjaman di biro kredit dengan pengajuan di Home Credit. Semakin besar jarak pengajuan yang dimiliki nasabah akan meningkatkan prediksi kemampuan nasabah dalam melakukan pembayaran kredit karena waktu pembayaran pinjaman di biro kredit akan semakin cepat selesai. Selain itu, waktu yang menjelaskan kapan pembaruan informasi dalam biro kredit juga menjadi pertimbangan yang penting. Jika waktu pembaruan informasi yang dilaporkan biro kredit semakin baru maka data tersebut akan semakin memperlihatkan perilaku nasabah dalam melakukan pinjaman di biro kredit.

4. Kesimpulan

Penerapan SMOTE untuk menangani ketidakseimbangan kelas berdampak positif pada peningkatan evaluasi kemampuan pembayaran kredit. Ini terlihat dari nilai AUC yang lebih tinggi dibandingkan tanpa metode SMOTE, baik untuk algoritma Logistic Regression maupun Random Forest, yaitu 0.74 dan 0.97. Hasil evaluasi menunjukkan bahwa model Random Forest Classifier yang menggunakan SMOTE menghasilkan kinerja terbaik, dengan akurasi 90%, precision 92%, recall 88%, F1-score 90%, dan AUC 0.97. Atribut yang diidentifikasi berdasarkan pentingnya fitur dalam model Random Forest Classifier meliputi skor dinormalisasi dari sumber data eksternal ke-3, skor dari sumber data eksternal ke-2, periode nasabah mengganti nomor, jumlah cicilan sebelumnya di Home Credit, usia nasabah, waktu registrasi, durasi pengajuan kredit di biro kredit, periode nasabah mengganti dokumen identitas, waktu pembaruan informasi oleh biro kredit, serta lama kerja nasabah. Selain itu, sebuah dashboard yang menampilkan hasil prediksi dari pengujian aplikasi dengan algoritma Random Forest Classifier berhasil dibuat menggunakan Google Looker Studio.

Referensi

[1] PT Home Credit Indonesia, "Laporan_Keberlanjutan_2022_PT_Home_Credit_Indonesia_(BHS)_Clean.pdf," 2022. Accessed: Apr. 23, 2024. [Online]. Available: https://www.google.com/url?sa=t&source=web&rct=j&opi=89978449&url=https://www.homecredit.co.id/static/pdf/Laporan_Keberlanjutan_2022_PT_Home_Credit_Indonesia_(BHS)_Clean.pdf&ved=2ahUKEwiZuK_SjcmIAxULzjgGHY74L30QFnoECBgQAQ&usg=AOvVaw0ApoRVitQ2AE9h1dh8T6L6

[2] Amrin and O. Pahlevi, "Implementasi Algoritma Klasifikasi Logistic Regression dan Naïve Bayes untuk Diagnosa Penyakit Hepatitis," *Jurnal Teknik Komputer AMIK BSI*, vol. 8, pp. 162–167, Jul. 2022, doi: 10.31294/jtk.v4i2.

[3] W. O. Simanjuntak, A. B. P. Negara, and R. Septriana, "Perbandingan Algoritma Logistic Regression dan Random Foret (Studi Kasus : Klasifikasi Emosi Tweet)," *Jurnal Aplikasi dan Riset Informatika*, vol. 2, pp. 160–164, Agustus 2023, doi: 10.26418/juara.v2i1.69682.

[4] M. F. Mujaddid, S. Al-Faraby, and Adiwijaya, "Analisis Churn Prediction Menggunakan Metode Logistic Regression dan SMOTE (Synthetic Minority Over-sampling Technique) Pada Perusahaan Telekomunikasi," *Universitas Telkom*, vol. 4, pp. 5046–5054, 2017.

[5] L. Zhu, D. Qiu, D. Ergu, C. Ying, and K. Liu, "A Study on Predicting Loan Default Based on the Random Forest Algorithm," presented at the 7th International Conference on Information Technology and Quantitative Management (ITQM 2019), Granada: Elseiver B.V., 2019, pp. 503–513.

[6] M. A. Khadija and N. A. Setiawan, "Detecting Liver Disease Diagnosis by Combining SMOTE,Information Gain Attribute Evaluation and Ranker," : *Jurnal Ilmiah Teknologi dan Informas*, vol. 9, pp. 13–17, Jun. 2020.

[7] M. I. C. Rachmatullah, "Penerapan SMOTE untuk Meningkatan Kinerja Klasifikasi Penilaian Kredit," *Jurnal Riset Komputer)*, vol. 10, pp. 302–309, Feb. 2023, doi: 10.30865/jurikom.v10i1.5612.

[8] H. D. Septama, T. Yulianti, D. Budiyanto, S. M. Mulyadi, and A. H. Cahyana, "A Comparative Analysis of Machine Learning Algorithms for Credit Risk Scoring using Chi-Square Feature Selection," in *2023 International Conference on Converging Technology in Electrical and Information Engineering (ICCTEIE)*, Bandar Lampung, Indonesia: IEEE, Oct. 2023, pp. 32–37. doi: 10.1109/ICCTEIE60099.2023.10366576.

[9] A. Montoya, KrillOdintsov, and M. Kotek, "Home Credit Default Risk." Kaggle, 2018. Accessed: Dec. 01, 2022. [Online]. Available: https://kaggle.com/competitions/home-credit-default-risk

[10] D. Barapatre and V. A, “Data Preparation on Largae Datasets for Data Science,” *Asian J Pharm Clin Res*, vol. 10, no. 13, p. 485, Apr. 2017, doi: 10.22159/ajpcr.2017.v10s1.20526.

[11] Michael J de Smith, *A Comprehensive Handbook of Statistical Concepts, Techniques and Software Tools*. London: The Winchelsea Press, 2018.

[12] E. Sutoyo and M. A. Fadlurrahman, “Penerapan SMOTE untuk Mengatasi Imbalance Class dalam Klasifikasi Television Advertisement Performance Rating Menggunakan Artificial Neural Network,” *Jurnal Edukasi dan Penelitian Informatika*, vol. 6, pp. 379–385, 2020.

[13] S. R. A. Nirwana, “Application of Multinomial Logistic Regression to Determine Factors That Affect the Study Program Selection in Department of Mathematics FMIPA UNM,” *Journal of Mathematics and Statistics*, vol. 01, 2015.

[14] J. Han, M. Kamber, and P. Jian, *Data Mining: Concepts and Techniques*, Third Edition. Waltham: Morgan Kaufmann, 2012.