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

Analisis Performa Logistic Regression dan Random Forest dalam Klasifikasi Kelayakan Penerimaan Kredit

Andreas Adrian[1,a)], Ike Verawati [2,a)]

[a)] Program Studi Informatika
Fakultas Ilmu Komputer
Universitas Amikom Yogyakarta

Author Emails
[1)] Corresponding author: andre46@students.amikom.ac.id
[2)]ikeverawati@amikom.ac.id

Abstract. *Credit eligibility determination is a crucial process in the banking and financial industry. It greatly affects the financial institutions involved and may even lead to an unhealthy financial condition if there are errors in credit eligibility decisions. Machine learning emerges as a solution to minimize such errors. To improve accuracy and efficiency in credit eligibility classification, this study focuses on the implementation of two machine learning models: Logistic Regression and Random Forest Classifier. Logistic Regression was chosen for its ability to identify linear relationships between input and output variables, while Random Forest Classifier offers advantages in handling complex and non-linear datasets. The main objective of this study is to compare the performance of these two models in credit eligibility classification. The comparison was carried out through several stages, including Literature Review, Data Acquisition which involves utilizing a publicly available banking dataset from Kaggle, EDA, Pre-Processing, Modeling, Evaluation, and Analysis of Model Evaluation. The dataset used contains financial information of customers. The performance comparison in this study employed accuracy, precision, recall, F1-score, and AUC-ROC metrics to evaluate each model. The results of this study show that the Random Forest model outperforms Logistic Regression, achieving an Accuracy score of 0.95, Precision of 0.93, Recall of 0.98, and F1-score of 0.96. The AUC score, which measures how well the model distinguishes between classes 1 and 0, reached 0.98. The findings of this research are expected to provide valuable recommendations for the banking industry in selecting the most appropriate model for credit eligibility assessment.*

Keywords **:**
Classification, Credit Eligibility, Logistic Regression, Machine Learning, Random Forest Classifier.

Abstraksi. Penentuan kelayakan penerimaan kredit merupakan proses yang sangat penting dalam industri perbankan dan keuangan. Hal ini sangat berpengaruh bagi badan keuangan tersebut, bahkan dapat menyebabkan kondisi finansial badan keuangan tersebut tidak sehat karena kesalahan dalam keputusan kelayakan kredit. Machine learning hadir untuk meminimalisir kesalahan tersebut. Untuk meningkatkan akurasi dan efisiensi dalam klasifikasi kelayakan kredit, penelitian ini berfokus pada penerapan dua model machine learning, yaitu Logistic Regression dan Random Forest Classifier. Logistic Regression dipilih karena kemampuannya dalam mengidentifikasi hubungan linear antara variabel input dan output, sedangkan Random Forest Classifier memiliki keunggulan dalam menangani dataset yang kompleks dan non-linear. Tujuan utama dari penelitian ini adalah untuk membandingkan performa kedua model tersebut dalam tugas klasifikasi kelayakan kredit. Perbandingan dilakukan dengan tahapan Studi Literatur, Akuisisi Data (Pengumpulan data) yang mengambil dataset perbankan public di kaggle, EDA, Pre-Processing, Modelling, Evaluasi, dan Analisis Evaluasi Model. Dataset yang akan digunakan mencakup informasi data finansial dari nasabah. Perbandingan performa pada penelitian ini menggunakan matrix akurasi, precision, recall, F1-Score dan AUC-ROC untuk mengevaluasi kinerja masing-masing model. Penelitian ini menghasilkan bahwa model random forest lebih unggul dengan skor Akurasi 0.95, Presisi 0.93, Recall 0.98 dan F1 Score 0.96. Skor AUC yang digunakan untuk melihat seberapa baik model dalam membedakan class 1 dan 0 mencapai 0.98. Hasil penelitian ini diharapkan mampu memberikan rekomendasi yang bermanfaat bagi industri perbankan dalam memilih model yang paling tepat untuk penilaian kelayakan kredit.

Kata Kunci : Kelayakan Kredit, Klasifikasi, Logistic Regression, Machine Learning, Random Forest Classifier.

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

Peningkatan akses terhadap layanan keuangan, terutama dalam bentuk pinjaman, merupakan faktor penting dalam mendorong pertumbuhan ekonomi, terutama di negara negara berkembang[1]. Lembaga keuangan seperti Bank dan Koperasi merupakan organisasi yang melayani jasa simpan dan pinjam. Bank adalah lembaga keungan yang memiliki fungsi pokok untuk memberi kredit dan jasa[2]. Bank juga berperan penting dalam menjembatani kebutuhan modal kerja dan investasi bagi UMKM[3]. Lembaga keuangan yang menyediakan jasa simpan pinjam juga masih memiliki banyak permasalah. Contohnya seperti pada saat ada lonjakan nasabah yang mengajukan pinjaman, waktu proses menjadi lama, prosedur terlalu rumit hingga penggunaan SOP yang tidak konsisten[4]. Di sisi lain, ada beberapa lembaga keuangan yang memiliki keterbatasan modal untuk memberikan pinjaman kepada anggotanya[5]. Oleh karena itu, pengelolaan risiko kredit tetap menjadi tantangan utama bagi banyak lembaga keuangan dalam menjaga kesehatan finansialnya. Dalam menghadapi tantangan tersebut, teknologi machine learning muncul sebagai solusi potensial. Teknologi ini telah digunakan di berbagai bidang. mulai dari kesehatan yang mulai menerapkan teknologi ini untuk prediksi penyakit stroke pada pasien[6], pada sektor pertanian yang digunakan untuk penentuan jenis buah mangga[7], dan juga di gunakan dalam bidang teknologi sendiri dengan menerapkan sebuah model machine learning untuk klasifikasi email phising[8].

Dalam penelitian ini, Logistic Regression dan Random Forest dipilih sebagai dua model yang akan dibandingkan dalam klasifikasi kelayakan penerimaan kredit. Logistic Regression sering digunakan dalam analisis prediktif karena kesederhanaannya. Model ini bekerja dengan baik ketika terdapat hubungan linier antara variabel-variabel independen dengan hasil yang diinginkan[9],[10]. Di sisi lain, Random Forest merupakan algoritma yang berbasis pohon keputusan dan mampu menangani data yang lebih kompleks. Algoritma ini dikenal karena akurasinya yang tinggi dan kemampuannya dalam mengatasi overfitting, serta kemampuannya dalam menangani data yang bersifat non-linier[11]. Dengan membandingkan kedua model ini menggunakan dataset yang memiliki karakteristik 50% linear dan 50% non-linear melalui pemilihan fitur, penelitian ini bertujuan untuk mengetahui model mana yang lebih unggul dalam konteks klasifikasi kelayakan kredit.

Sebagai solusi atas permasalahan yang dihadapi lembaga keuangan dalam pengelolaan risiko kredit, penerapan machine learning, khususnya melalui algoritma Logistic Regression dan Random Forest, diharapkan dapat meningkatkan efisiensi dan akurasi dalam proses penilaian kelayakan kredit. Dengan menggunakan kedua model ini, koperasi dapat mengoptimalkan pengambilan keputusan terkait pemberian pinjaman, mempercepat proses pengolahan data nasabah, serta meminimalisir risiko kredit bermasalah. Hasil analisis komparatif ini akan membantu lembaga keuangan memilih model yang paling sesuai untuk diterapkan dalam meningkatkan kinerja keuangan serta menjaga stabilitas operasional mereka.

Tujuan yang akan dicapai oleh peneliti dalam penelitiannya adalah untuk menganalisis dan mengetahui model machine learning terbaik yang telah diuji dengan metode evaluasi dan dataset yang telah di tetapkan yang nantinya diharapkan mampu menjadi opsi yang dapat digunakan lembaga keuangan dalam kasus simpan pinjam untuk mengatasi permasalahannya. Adapun batasan Masalah dalam penilitian ini meliputi model machine learning yang akan diuji pada penelitian ini yaitu Logistic Regression dan Random Forest Classifier, Metode Evaluasi yang dilakukan pada kedua model machine learning menggunakan matrix accuration, precision, recall, F1-Score dan AUC-ROC, dataset yang digunakan merupakan dataset open source yang berada di internet, dataset yang digunakan di atur sedemikian rupa agar bersifat 50% linear dan 50% non-linear

Penelitian ini diharapkan dapat memberikan beberapa manfaat antara lain memberikan informasi tentang model terbaik dalam kasus penilaian kelayakan kredit. Serta memberikan opsi algoritma machine learning yang unggul kepada lembaga keuangan yang melayani simpan pinjam dalam menilai kelayakan kredit pada nasabahnya.

TINJAUAN PUSTAKA

Peneliti telah melakukan studi literatur sebelum memulai penelitian ini. Peneliti mengacu pada beberapa penelitian yang relevan dengan topik peneliti seperti penelitian Utiarahman *et al*.[12] dimana peneliti membandingkan 4 algoritma machine learning yaitu Regresi Logistik, Decision Tree, KNN dan SVM . Hasil dari penelitian ini menyebutkan bahwa Decision Tree merupakan algoritma terbaik dengan hasil akurasi 99.3%, presisi 0.97-1.00, recall 0.98-1.00 dan f1-score 0.98-1.00. Namun, fitur yang dipilih dari dataset yang digunakan pada penelitian ini kurang jelas

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

Handayani, *et al* [13] yang berusaha untuk mendapatkan performa algoritma random forest tertinggi dengan manganalisis pembagian data train dan data test yang digunakan. Hasil dari penelitian ini random forest mendapatkan performa tertinggi dengan akurasi 88.6%, presisi 88.1%, recall 88.6% dan f1-score 88.2% pada saat menggunakan skenario pembagian data train dan test 90:10. Namun dataset yang digunakan peneliti dalam penelitian ini masih dataset yang bersifat linear

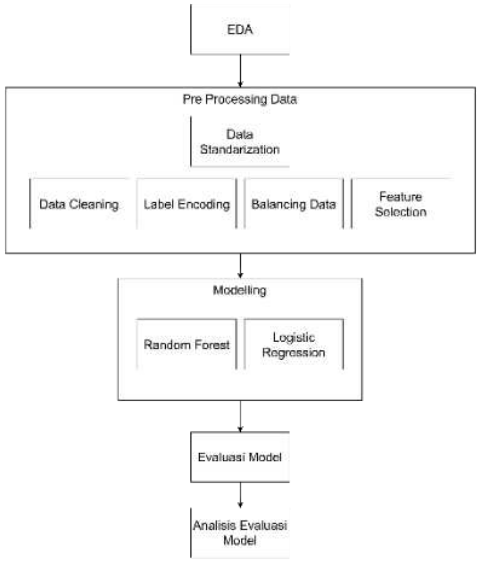
Nugroho [14] yang membandingkan 5 algoritma machine learning yaitu Logistic Regression, Random Forest, KNN, SGD dan Neural Network. Penelitian menggunakan data hasil pemeriksaan penyakit kardiovaskular sejumlah 425.195 data. Hasil dari penelitian ini menunjukkan bahwa algoritma Neural Network merupakan algoritma dengan kinerja terbaik yaitu dengan nilai akurasi sebesar 89.46%, AUC 0.873, presisi 0.877 dan recall 0.896. namun pemilihan fitur yang dipilih serta metode yang dilakukan untuk memilih fitur oleh peniliti kurang jelas

Akbar [15] dimana penelitiannya membandingkan 3 algoritma machine learning yaitu Regresi Logistik, Decision Tree, dan Random Forest. Penelitian menggunakan dataset sebanyak 319.795 data. Penelitian ini menerapkan metode k-fold pada feature selection. Yang dimana metode tersebut bersifat rumit dan kurang efisien dalam menemukan fitur yang berkorelasi rendah. Penggunaan cara manual dalam pemilihan feature berdasarkan heatmap akan lebih efisien jika dilakukan

Maulina [16] yang membandingkan 2 algoritma machine learning, random forest classifier dan SVM. Random Forest memperoleh skor akurasi yang lebih tinggi yaitu 95% jika dibandingkan dengan SVM yang hanya mendapatkan 91% .Pada penelitian ini, peneliti menggunakan teknik SelectKBest menggunakan uji ANOVA-F yang dimana teknik ini hanya akan memilih feature yang berhubungan linear dengan class yang sudah dipilih. Penggunaan teknik feature selection berdasarkan heatmap akan menghasilkan opsi fitur yang dapat dipilih lebih beragam dan dapat disesuaikan dengan tujuan penelitian. Serta Gustian [17] yang juga membandingkan 3 algoritma machine learning yaitu logistic regression, SVM, dan random forest classifier. Hasil dari penelitian ini yaitu nilai akurasi dari SVM merupakan yang tertinggi dari ke 3 model tersebut. Tetapi penelitian ini menggunakan matriks evaluasi yang sangat terbatas untuk perbandingannya. Penelitian ini hanya membandingkan berdasarkan akurasi saja. Akan lebih baik jika setidaknya perbandingan dilakukan dengan matriks evaluasi standar (akurasi, presisi, recall dan AUC-ROC.

METODE PENELITIAN

Metode penelitian ini terdiri atas beberapa tahapan yaitu EDA, Pre Processing, Modelling, Evaluasi dan Analisis Evaluasi Model sebagaimana dapat dilihat pada gambar 1.



GAMBAR 1. Metode Penelitian

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

Pada penelitian ini data yang digunakan merupakan dataset public yang didapatkan dari situs kaggle. Dataset ini terdiri dari 32.586 baris data dalam bentuk file csv. Dan terdiri dari 13 fitur. Fitur merupakan sejumlah atribut yang menangkap karakteristik dari object data tersebut[18]. antara lain customer_id, customer_age, customer_income, home_ownership, employment_duration, loan_intent, loan_grade, loan_amnt, loan_int_rate, term_years, historical_default, cred_hist_length, Current_loan_status. berikut merupakan sampel data yang terdapat pada dataset.

A. EDA

Exploratory Data Analysis (EDA) merupakan proses di mana peneliti secara menyeluruh memeriksa fitur-fitur yang terdapat pada dataset, mengidentifikasi adanya anomali, serta mempelajari setiap fitur mulai dari arti, format, hingga seberapa penting fitur tersebut terhadap penelitian yang dilakukan. Pada tahap ini, tipe dari masing-masing fitur juga diperiksa dengan cermat untuk memastikan kesesuaiannya. Pemeriksaan ini sangat penting karena hasilnya akan menjadi dasar bagi langkah-langkah lanjutan, seperti proses data cleaning maupun data standardization, sehingga data yang digunakan dalam pemodelan nantinya benar-benar berkualitas dan siap diolah.

B. Pre-Processing

Ditahap ini dilakukan data cleaning, label encoding, balancing data dan pemilihan fitur.

a. Standarization dilakukan untuk menyeragamkan format data agar dapat diolah oleh algoritma machine learning dan agar pada saat dilakukan encoding tidak terlalu banyak data baru yang tercipta.
b. Data cleaning dilakukan untuk menghilangkan data n/a.
c. Label encoding dilakukan untuk mengubah fitur yang mempunyai tipe kategorikal ke bentuk numerik.
d. Balancing data dilakukan agar jumlah kelas 0 dan 1 menjadi seimbang. Hal ini bertujuan untuk meningkatkan kinerja model[19]. Balancing data pada penelitian ini menggunakan 2 teknik undersampling. Random undersampling dan Nearmiss v2.
e. Feature Selection, pemilihan fitur pada penelitian ini menggunakan dasar korelasi yang dihasilkan oleh heatmap dan juga grafik korelasi. Dengan demikian karakteristik dataset akan bersifat 50% linear dan 50% non linear. Konsep Linear pada dataset adalah ketika antar variabel/fitur pada dataset memiliki korelasi yang konstan. Prinsip utama dari konsep ini adalah *input:output = contant* [20]. Apabila di tunjukkan dengan grafik, maka data data tersebut akan membentuk sebuah garis lurus. Konsep Non-Linear pada sebuah dataset adalah ketika hubungan antara variabel tidak selalu konstan[20]. Dalam konsep ini, prinsip utama yang ada pada konsep linear tidak lagi berlaku. Yang artinya bahwa hubungan antara kedua variabel (input dan output) lebih kompleks[21]. Apabila digambarkan pada sebuah grafik maka konsep tidak akan membentuk garis lurus.

C. Modelling

Pada tahap modeling, setelah dataset melalui proses pre-processing sehingga siap digunakan, peneliti membangun dua model machine learning yaitu Logistic Regression dan Random Forest. Dataset kemudian dibagi menjadi data latih dan data uji dengan rasio 80:20 untuk mendukung performa model [22], Setelah itu dilakukan proses training pada data latih hingga diperoleh model akhir yang siap digunakan. Model tersebut kemudian diuji kembali pada data uji untuk memastikan kinerja, sekaligus dibandingkan performanya dalam tugas klasifikasi kelayakan kredit.

D. Evaluasi Model

Hasil performa dari masing-masing model akan dianalisis pada tahap ini dengan menggunakan beberapa matriks evaluasi, yaitu akurasi, presisi, recall, F1-Score, serta AUC-ROC. Kelima metode evaluasi ini dipilih karena mampu memberikan gambaran yang lebih komprehensif mengenai kualitas prediksi model. Akurasi digunakan untuk melihat persentase prediksi yang benar, presisi digunakan untuk mengetahui seberapa banyak prediksi positif yang tepat, dan recall digunakan untuk mengukur seberapa baik model menemukan seluruh kasus positif yang sebenarnya. F1-Score merupakan kombinasi harmonis antara presisi dan recall yang membantu menilai keseimbangan performa model. Selain itu, AUC-ROC digunakan untuk mengukur seberapa baik model dalam membedakan antara nasabah yang layak kredit dan yang tidak layak kredit [23]. Berbagai metrik tersebut juga sangat membantu dalam tahap tuning model agar kinerjanya semakin optimal [24].

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

E. Analisis Evaluasi Model

Pada tahap ini, performa dari kedua model akan dibandingkan secara menyeluruh dengan tujuan untuk menemukan model dengan performa terbaik. Perbandingan dilakukan berdasarkan hasil pengukuran metrik evaluasi yang telah diperoleh sebelumnya, seperti akurasi, presisi, recall, F1-Score, dan AUC-ROC. Setiap nilai metrik dianalisis untuk melihat keunggulan dan kelemahan masing-masing model dalam mengklasifikasikan data nasabah. Proses ini tidak hanya membandingkan angka, tetapi juga mempertimbangkan konsistensi model dalam menentukan kelayakan kredit.

HASIL DAN PEMBAHASAN

A. EDA

Pada tahap ini, peneliti mempelajari berbagai hal yang memengaruhi penelitian dari dataset yang digunakan, termasuk memahami fitur-fitur yang tersedia dan tipe data dari masing-masing fitur. Berdasarkan Tabel 1, dapat dilihat bahwa dataset memiliki beberapa kolom utama seperti "customer_id", "customer_age", "customer_income", "home_ownership", "employment_duration", "loan_intent", "loan_grade", "loan_amnt", "loan_int_rate", "term_years", "historical_default", "cred_hist_length", dan "Current_loan_status". Setiap kolom memiliki jumlah data yang berbeda-beda, terlihat dari nilai Non-Null Count.

TABEL 1. Output informasi dataset

#	Column	Non-Null Count	Dtype
0	customer_id	32583 non-null	float64
1	customer_age	32586 non-null	int64
2	customer_income	32586 non-null	object
3	home_ownership	32586 non-null	object
4	employment_duration	31691 non-null	float64
5	loan_intent	32586 non-null	object
6	loan_grade	32586 non-null	object
7	loan_amnt	32585 non-null	object
8	loan_int_rate	29470 non-null	float64
9	term_years	32586 non-null	int64
10	historical_default	11849 non-null	object
11	cred_hist_length	32586 non-null	int64
12	Current_loan_status	32582 non-null	object

Beberapa kolom seperti "employment_duration", "loan_int_rate", "historical_default", dan "Current_loan_status" memiliki jumlah Non-Null Count yang tidak penuh, menunjukkan adanya missing value yang perlu ditangani pada tahap pre-processing. Selain itu, ditemukan juga beberapa fitur yang bertipe object padahal seharusnya dapat dikonversi menjadi numerik, misalnya pada "customer_income" dan "loan_amnt". Hal ini menunjukkan adanya ketidaksesuaian format data yang harus diperbaiki agar dapat diproses lebih lanjut oleh model machine learning. Analisis awal ini membantu peneliti memahami karakteristik dataset serta langkah-langkah yang diperlukan untuk membersihkan dan menyiapkan data sebelum masuk ke tahap pemodelan.

B. Pre Processing

Tahap ini terbagi menjadi 5 tahapan penting, yaitu data standarization, data cleaning, label encoding, balancing data dan feature selection. Detail dari kelima tahapan tersebut dapat dilihat pada poin setiap tahapan.

A. *Data Standarization*

Fitur "loan_amnt", berformat object tetapi tidak bisa langsung di ubah menjadi numeric karena memiliki simbol "£", ",", dan ".". begitu juga dengan fitur "customer_income" yang memiliki simbol ",". Maka perlu dilakukan beberapa tahap untuk menghilangkan simbol simbol tersebut. Fitur "customer_income" dan "loan_amnt" di ubah

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

yang semula bertipe data object menjadi numerik. Serta perbaikan penamaan fitur pada fitur terakhir yang semula "Current_loan_status" menjadi "current_loan_status".

B. *Data Cleaning*

Tahap ini bertujuan untuk menghilangkan missing value. Data missing value dapat dilihat di tabel 2

TABEL 2. Output informasi missing value

Nama Fitur	Jumlah Missing Value
customer_id	3
customer_age	0
customer_income	0
home_ownership	0
employment_duration	895
loan_intent	0
loan_grade	0
loan_amnt	0
loan_int_rate	3115
term_years	0
historical_default	20737
cred_hist_length	0
current_loan_status	4

Dari tabel 2 dapat dilihat bahwa terdapat beberapa missing value di berbagai macam fitur. Peneliti menggunakan library pandas untuk menghilangkan missing value yang berjumlah sedikit jika dibandingkan dengan total data. Handling missing value pada penelitian ini dilakukan dengan 2 cara. Drop n/a dan mengganti nilai n/a dengan rata rata (mean) dari sebuah fitur. Dilihat dari tabel 2 terdapat fitur yang memiliki missing value yang sangat rendah yaitu "current_loan_status". Dan juga terdapat fitur yang tidak digunakan yaitu "customer_id". Maka dengan demikian bisa dilakukan proses drop n/a pada "current_loan_status" dan drop fitur pada fitur "customer_id".

Untuk fitur employment duration, peneliti menggunakan nilai 4,8 yang merupakan rata rata nilai dari fitur tersebut untuk mengisi missing value. Fitur "loan_int_rate" juga di isi dengan nilai rata rata yaitu 11. Oleh karena fitur "historical_default" bertipe data object, maka peneliti menggunakan modus (nilai yang sering muncul) untuk mengisi missing value. Berdasarkan penggunaan method ".value_counts" pada library pandas, didapatkan bahwa value "Y" merupakan modus dengan kemunculan nya yang berjumlah 6127. Sedangkan "N" hanya berjumlah 5717.

C. *Label Encoding*

Fitur kategorikal berbentuk object, seperti "home_ownership," "loan_intent," "loan_grade," "historical_default," dan "current_loan_status," masih belum diubah ke dalam format numerik. Padahal, algoritma machine learning hanya dapat memproses data dalam bentuk numerik [25]. Oleh karena itu, diperlukan proses transformasi fitur untuk mengonversi data kategorikal menjadi representasi numerik agar dapat digunakan dalam model machine learning. Salah satu metode yang umum digunakan adalah label encoding, di mana setiap kategori unik dalam fitur dikodekan dengan nilai numerik yang berbeda. Metode ini memungkinkan model untuk memahami perbedaan antara kategori tanpa harus mengubah struktur data yang ada.

Value dari fitur "home_ownership" berubah setelah melalui tahap ini menjadi seperti pada tabel 3

TABEL 3. Fitur "home_ownership" encoding

Sebelum label encoding	Setelah label encoding
MORTGAGE	0
OTHER	1
OWN	2
RENT	3

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

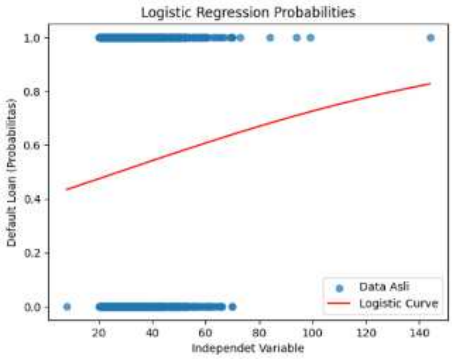
Tabel 3 menunjukkan hasil proses label encoding yang diterapkan pada fitur "home_ownership". Sebelumnya, fitur ini memiliki nilai kategorikal berupa "MORTGAGE", "OTHER", "OWN", dan "RENT" yang tidak dapat langsung diproses oleh model machine learning. Setelah melalui tahap label encoding, setiap kategori tersebut diubah menjadi representasi numerik, yaitu "MORTGAGE" menjadi 0, "OTHER" menjadi 1, "OWN" menjadi 2, dan "RENT" menjadi 3. Hal tersebut juga dilakukan pada beberapa fitur yang bertipe data object lainnya seperti fitur "loan_intent", "loan_grade", "historical_default", dan "current_loan_status". Transformasi ini bertujuan untuk mempermudah algoritma dalam membaca serta memproses data secara efisien tanpa mengubah makna dari setiap kategori. Dengan demikian, data yang semula bersifat teks kini telah siap digunakan dalam tahap pemodelan dan evaluasi lebih lanjut..

D. *Balancing Data*

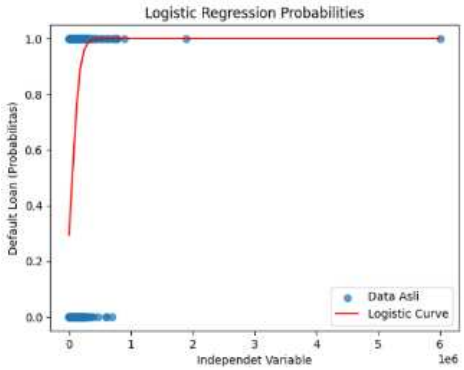
Terdapat ketidakseimbangan pada dataset yang digunakan. Class 1 berjumlah 25.742 sedangkan class 0 hanya berjumlah 6.839. Hal ini dapat menurunkan performa model. Oleh karena itu peneliti menerapkan 2 teknik undersampling, yaitu Random undersampling dan Nearmiss v2 untuk menangani imbalance data dan menjaga agar jumlah data pada dataset tidak terlalu banyak sehingga berpengaruh pada waktu pelatihan yang lama[26]. Kedua teknik undersampling tersebut mengasilkan jumlah data pada class 1 berubah menjadi sama dengan class 0 yaitu 6839. Dengan demikian, data menjadi lebih proporsional dan siap digunakan untuk pemodelan tanpa bias terhadap kelas mayoritas.

E. *Feature Selection*

Peneliti menggunakan dasar korelasi di heatmap dan juga grafik korelasi antar variabel independen dan dependen (probability plot). Berdasarkan korelasi fitur yang telah dilakukan, Peneliti mengambil fitur yang memiliki korelasi >0 dengan variabel y. Dengan parameter tersebut maka diperoleh 8 fitur yang lolos seleksi yaitu, "customer_age", "customer_income", "employment_duration", "loan_amnt", "term_years", "cred_hist_length", "loan_intent_numeric", "historical_default_numeric". Untuk membuat sifat dataset menjadi 50% linear dan 50% non linear, maka peneliti melakukan seleksi fitur sekali lagi dengan grafik korelasi antara 1 variabel x dengan variabel y. Hal ini tidak mungkin dapat dilakukan dengan nilai variabel y yang merupakan binary (0 dan 1), maka peneliti menggunakan bantuan dari model logistic regression untuk menampilkan grafik probabilitas korelasi antara 2 variabel tersebut. Proses ini dilakukan satu per satu untuk tiap fitur yang sudah lolos seleksi di tahap sebelumnya.



GAMBAR 2. Hasil korelasi antara fitur "customer_age" dengan y



GAMBAR 3. Hasil korelasi antara fitur "customer_income" dengan y

Gambar 2 dan Gambar 3 merupakan salah satu contoh korelasi. Grafik pada gambar 2 menunjukkan bahwa independent variable yang merupakan fitur "cutomer_age" memiliki korelasi linear dengan dependent variable yaitu "current_loan_status_numeric". Berbanding terbalik dengan contoh lainnya yang dapat dilihat di Grafik pada Gambar 3 yang menunjukkan bahwa independent variable yang merupakan fitur "cutomer_income" memiliki korelasi non-linear dengan dependent variable yaitu "current_loan_status_numeric".

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

TABEL 4. Klasifikasi Fitur Berdasarkan Karakteristik

Fitur yang bersifat linear dengan variabel y	Fitur yang bersifat non-linear dengan variabel y
customer_age	customer_income
term_years	employment_duration
cred_hist_length	loan_amnt
loan_intent_numeric	historical_default_numeric

Dari grafik masing-masing fitur yang telah di teliti, dapat disimpulkan bahwa fitur yang bersifat linear dan fitur yang bersifat non-linear dengan variabel y dapat dilihat pada tabel 4. Dari sifat masing masing fitur terhadap variabel y dapat disimpulkan bahwa fitur yang bersifat linear berjumlah 4 dan non-linear berjumlah 4. Maka dengan demikian, sifat dataset 50% linear dan 50% non linear sudah tercapai.

C. Modelling

Tahap modelling terdiri dari splitting data, inisialisasi model, dan training. Rasio data train dan test pada penelitian ini menggunakan ratio yang sering dipakai di penelitian sebelumnya dan dapat meningkatkan kinerja model yaitu rasio 80:20. Dengan demikian maka data train menjadi berjumlah 10.942 dan data test berjumlah 2.736 dengan proporsi 50:50 tiap class nya. Komposisi ini digunakan di kedua dataset. Baik yang menggunakan teknik random undersampling maupun Nearmiss v2. Pada tahap inisialisasi model dan training, peneliti menggunakan paramater training default dari masing masing model machine learning.

D. Evaluasi

Evaluasi model terbagi menjadi 3 tahapan. Tahapan pertama yaitu melihat performa model dengan matriks evaluasi yang meliputi Akurasi, Presisi, Recall dan F1-Score. Hasil dari tahapan ini dapat dilihat pada tabel 5.

TABEL 5. Output Hasil Evaluasi Model Logistic Regression dan Random Forest Classifier

Model Machine Learning	Akurasi	Presisi	Recall	F1-Score
Logistic Regression (Random Undersampling)	0.77	0.75	0.83	0.78
Random Forest (Random Undersampling)	0.85	0.81	0.91	0.86
Logistic Regression (Nearmiss V2 Undersampling)	0.88	0.87	0.90	0.88
Random Forest (Nearmiss V2 Undersampling)	0.95	0.93	0.98	0.96

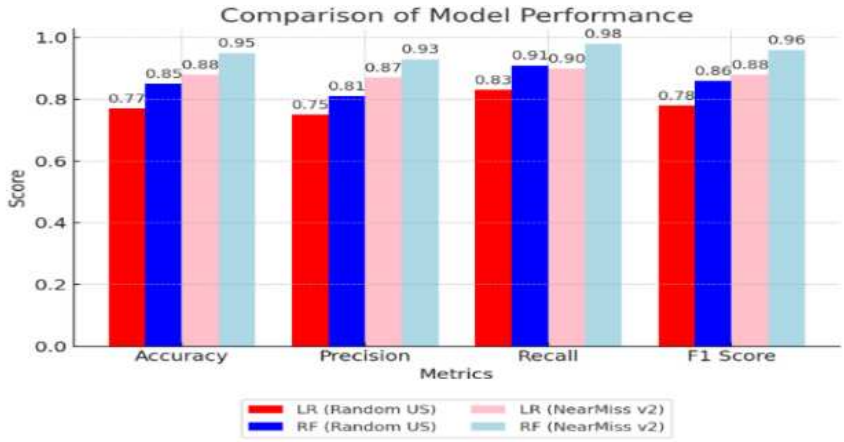
Dari tabel 5 dapat dilihat bahwa model Logistic Regression memiliki performa yang cukup baik ketika menggunakan teknik undersampling Nearmiss v2. Skor Akurasi yang didapatkan oleh model Logistic Regression yaitu 0.88, Presisi 0.87, Recall 0.90 dan F1 Score 0.88. Sedangkan model Random Forest Classifier meraih skor terbaik ketika menggunakan teknik undersampling Nearmiss v2. Skor Akurasinya mencapai 0.95, Presisi 0.93, Recall 0.98 dan F1 Score 0.96.

Tahap terakhir pada evaluasi model adalah menampilkan dan melihat grafik AUC-ROC dari masing masing model. Peneliti menemukan bahwa ketika teknik undersampling Nearmiss v2 di terapkan, model logistic regression mengalami peningkatan yang signifikan. Model ini dapat membedakan class 0 dan 1 dengan baik ditunjukkan dengan skor ROC yang mencapai 0.94. Model Random Forest juga mengalami peningkatan. Dengan penggunaan Nearmiss V2 undersampling, model ini sangat baik dalam membedakan class 0 dan 1 ditunjukkan dengan skor ROC nya yang mencapai 0.98.

E. Analisis Evaluasi Model

Hasil evaluasi dari kedua model dapat dilihat perbandingan nya melalui grafik matriks evaluasi.

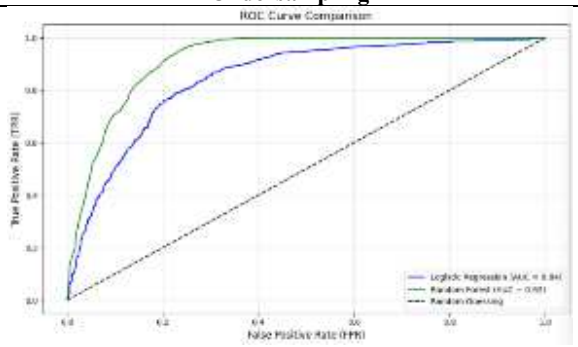
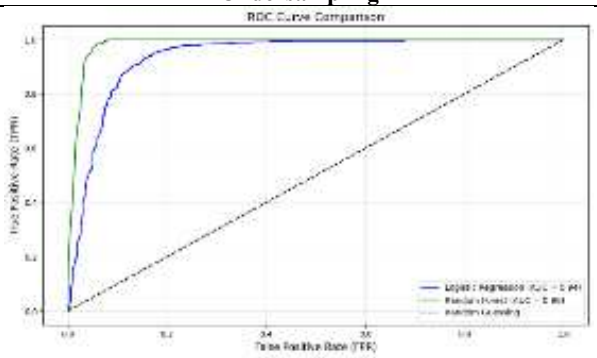
IJCSR: The Indonesian Journal of Computer Science Research
E-ISSN 2963-9174 **DOI prefix** 10.37905
Volume 4, No. 2 Juli 2025
https://subset.id/index.php/IJCSR



GAMBAR 4 Perbandingan performa kedua model

Dari grafik pada gambar 4 dapat disimpulkan bahwa model Random Forest lebih unggul dengan skor tertinggi Akurasi 0.95, Presisi 0.93, Recall 0.98 dan F1 Score 0.96 saat menggunakan teknik undersampling Nearmiss V2. Sedangkan model Logistic Regression yang hanya mendapatkan skor performa tertinggi menggunakan Nearmiss V2 dengan detail skor yaitu Akurasi 0.88, Presisi 0.87, Recall 0.90 dan F1 Score 0.88. Tingkat konsistensi kedua model dalam membedakan prediksi ke kedua class, dapat dilihat pada grafik perbandingan AUC-ROC dibawah.

TABEL 6. Output Hasil Grafiik AUC-ROC Kedua Model pada teknik undersampling yang berbeda

Grafik AUC-ROC kedua model pada Random Undersampling	**Grafik AUC-ROC kedua model pada Nearmiss V2 Undersampling**
	

Dari grafik pada tabel 6, terlihat bahwa model random forest classifier lebih unggul dalam membedakan kelas positif dan negatif pada kedua teknik undersampling dengan score tertinggi AUC 0.98, sementara model logistic regression hanya mendapatkan AUC tertinggi sebesar 0.94. Hal ini disebabkan oleh sifat dataset yang telah di buat bersifat 50% linear dan 50% non linear. Dengan grafik ini terlihat bahwa random forest dapat menangani hampir semua jenis data sedangkan logistic regression kurang mampu dalam menghadapi data data yang bersifat non-linear.

TABEL 7. Waktu Pelatihan Model

Model	**Waktu Pelatihan**
Logistic regression	121 ms
Random Forest	1.2 s
Logistic Regression (Nearmiss V2)	173 ms
Random Forest (Nearmiss V2)	1.35 s

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

Selain dari matriks evaluasi dan AUC-ROC, peneliti juga mempertimbangkan waktu yang digunakan untuk melatih masing masing model. Berdasarkan Tabel 7, dapat disimpulkan bahwa ketika kedua model dilatih dengan dataset yang menghasilkan performa terbaik model yaitu dataset yang menggunakan teknik undersampling Nearmiss V2, model Logistic Regression memerlukan waktu hanya 173 milisecond untuk menyelesaikan pelatihan sedangkan Random Forest memerlukan waktu lebih lama dari waktu pelatihan logistic regression yaitu 1.35 detik.

KESIMPULAN

Kesimpulan dari penelitian ini adalah bahwa model Logistic Regression memerlukan waktu 173 ms untuk melakukan pelatihan dan mendapatkan skor Akurasi 0.88, Presisi 0.87, Recall 0.90, F1 Score 0.88, serta AUC 0.94 yang menunjukkan kemampuan membedakan class 1 dan 0. Sementara itu, model Random Forest Classifier memerlukan waktu 1.35 detik untuk pelatihan dan memperoleh skor Akurasi 0.95, Presisi 0.93, Recall 0.98, F1 Score 0.96, serta AUC 0.98. Dari kedua model yang dibandingkan, dapat disimpulkan bahwa Random Forest merupakan model terbaik dalam kasus penilaian kelayakan kredit dengan dataset yang digunakan karena unggul di segala lini kecuali waktu pelatihan dibandingkan dengan model Logistic Regression

TINJAUAN PUSTAKA

[1] M. W. Batubara, "Peran Koperasi Syariah Dalam Meningkatkan Perekonomian dan Kesejahteraan Masyarakat Di Indonesia," *J. Ilm. Ekon. Islam*, vol. 7, no. 03, pp. 1494–1498, 2021, [Online]. Available: http://jurnal.stie-aas.ac.id/index.php/jiedoi:http://dx.doi.org/10.29040/jiei.v7i3.2878

[2] K. C. C. Merentek, "Analisis Kinerja Keuangan Antara Bank Negara Indonesia (Bni) Dan Bank Mandiri Menggunakan Metode Camel," *J. Ris. Ekon. Manajemen, Bisnis dan Akunt.*, vol. 1, no. 3, pp. 645–652, 2013.

[3] S. B. Wilarjo, "Pengertian, Peranan, dan Perkembangan Bank Syariah di Indonesia," *Igarss 2014*, vol. 2, no. 1, pp. 1–5, 2014.

[4] M. Syafriansyah, "Analisis Sistem Dan Prosedur Pemberian Kredit Pada Koperasi Simpan Pinjam Sentosa Di Samarinda," *Ilmu Adm. Bisnis*, vol. 3, no. 1, pp. 83–93, 2015, [Online]. Available: http://www.bi.go.id/

[5] R. D. Perkasa and W. N. Sulistiani, "Peran dan Tantangan Koperasi dalam Pembangunan Ekonomi Masyarakat yang Bebas di Desa Namo Bintang Kecamatan Pancur Batu," *El-Mujtama J. Pengabdi. Masyarakat* , vol. 4, no. 2, pp. 710–719, 2024, doi: 10.47467/elmujtama.v4i2.1001.

[6] Indah Werdiningsih *et al.*, "Analisis Prediksi Stroke Menggunakan Pendekatan Decision Tree dengan Seleksi Fitur dan Neural Network," *J. Sist. Cerdas*, vol. 6, no. 3, pp. 213–221, 2023, doi: 10.37396/jsc.v6i3.310.

[7] H. Suroyo, "Penerapan Machine Learning dengan Aplikasi Orange Data Mining Untuk Menentukan Jenis Buah Mangga," *Semin. Nas. Teknol. Komput. Sains*, vol. 1, no. 1, pp. 343–347, 2019, [Online]. Available: https://prosiding.seminar-id.com/index.php/sainteks/article/view/177

[8] D. Anggraini and T. Sutabri, "IJM: Indonesian Journal of Multidisciplinary Pengembangan Aplikasi Penyaringan Spam e-Mail Menggunakan Teknik Machine Learning dengan Metode Support Vector Machines," *IJM Indones. J. Multidiscip.*, vol. 2, pp. 106–114, 2024, [Online]. Available: https://journal.csspublishing/index.php/ijm

[9] aws.amazon.com, "Apa itu Regresi Logistik?," aws.amazon.com. Accessed: Oct. 20, 2024. [Online]. Available: https://aws.amazon.com/id/what-is/logistic-regression/

[10] G. Biagetti, P. Crippa, L. Falaschetti, G. Tanoni, and C. Turchetti, "A comparative study of machine learning algorithms for physiological signal classification," *Procedia Comput. Sci.*, vol. 126, pp. 1977–1984, 2018, doi: 10.1016/j.procS.2018.07.255.

[11] algorit.ma, "Mengenal Algoritma Random Forest," algorit.ma. Accessed: Oct. 20, 2024. [Online]. Available: https://algorit.ma/blog/cara-kerja-algoritma-random-forest-2022/

[12] S. A. Utiarahman and A. M. M. Pratama, "Analisis Perbandingan KNN, SVM, Decision Tree dan Regresi Logistik Untuk Klasifikasi Obesitas Multi Kelas," *KLIK Kaji. Ilm. Inform. dan Komput.*, vol. 4, no. 6, pp. 3137–3146, 2024, doi: 10.30865/klik.v4i6.1871.

[13] P. Handayani and A. Charis Fauzan, "KLIK: Kajian Ilmiah Informatika dan Komputer Machine Learning Klasifikasi Status Gizi Balita Menggunakan Algoritma Random Forest," *Media Online)*, vol. 4, no. 6, pp. 3064–3072, 2024, doi: 10.30865/klik.v4i6.1909.

[14] Adi Nugroho, Agustinus Bimo Gumelar, Adri Gabriel Sooai, Dyana Sarvasti, and Paul L Tahalele, "Perbandingan Performansi Kinerja Algoritma Pengklasifikasian Terpandu Untuk Kasus Penyakit Kardiovaskular," *J. RESTI (Rekayasa Sist. dan Teknol. Informasi)*, vol. 4, no. 5, pp. 998–1006, 2020, doi: 10.29207/resti.v4i5.2316.

[15] M. I. Akbar, "Perbandingan Teknik Machine Learning Untuk Diagnosisi dan Prediksi Penyakit Jantung," Amikom Yogyakarta, 2024. [Online]. Available: http://eprints.amikom.ac.id//id/eprint/26265/

[16] S. I. Maulina, "Komparasi Algoritma Support Vector Machine dan Random Forest Classifier Untuk Mengklasifikasikan Pembagian Kelas Siswa Kelas VIII," Amikom Yogyakarta, 2024. [Online]. Available: http://eprints.amikom.ac.id//id/eprint/26478/

[17] H. Gustian, "Prediksi Penyakit Diabetes menggunakan Machine Learning," Amikom Yogyakarta, 2023. [Online]. Available: http://eprints.amikom.ac.id//id/eprint/22668/

[18] M. M. Hidayat, "Type Data Mining," *Artic. Min. Massive Datasets*, vol. 2, no. January 2013, pp. 5–20, 2015.

[19] D. Santiago, "Menyeimbangkan Data yang Tidak Seimbang: Teknik Undersampling dan Oversampling dalam Python." Accessed: Nov. 23, 2024. [Online]. Available: https://medium-com.translate.goog/@daniele.santiago/balancing-imbalanced-data-undersampling-and-oversampling-techniques-in-python-7c5378282290?_x_tr_sl=en&_x_tr_tl=id&_x_tr_hl=id&_x_tr_pto=rq&_x_tr_hist=true

[20] A. Mahapatra, "[ML basics][Regression] How to tell if a dataset is linear or not?," medium.com. Accessed: Nov. 15, 2024. [Online]. Available: https://medium.com/@abhinav.mahapatra10/ml-basics-regression-how-to-tell-if-a-dataset-is-linear-or-not-594a4f1e8aaf

[21] R. Zarra, "Machine Learning: Linearity vs Nonlinearity," linkedin.com. Accessed: Nov. 15, 2024. [Online]. Available: https://www.linkedin.com/pulse/machine-learning-linearity-vs-nonlinearity-reday-zarra/

[22] I. O. Muraina, "Ideal Dataset Splitting Ratios in Machine Learning Algorithms: General Concerns for Data Scientists and Data Analysts," *7th Int. Mardin Artuklu Sci. Res. Conf.*, no. February, pp. 496–504, 2022, [Online]. Available: https://www.researchgate.net/publication/358284895_IDEAL_DATASET_SPLITTING_RATIOS_IN_MACHINE_LEARNING_ALGORITHMS_GENERAL_CONCERNS_FOR_DATA_SCIENTISTS_AND_DATA_ANALYSTS

[23] R. Arthana, "Mengenal Accuracy, Precision, Recall dan Specificity serta yang diprioritaskan dalam Machine Learning." Accessed: Nov. 15, 2024. [Online]. Available: https://rey1024.medium.com/mengenal-accuracy-precission-recall-dan-specificity-serta-yang-diprioritaskan-b79ff4d77de8

[24] N. A. Ahmed, "What is A Confusion Matrix in Machine Learning? The Model Evaluation Tool Explained." Accessed: Nov. 15, 2024. [Online]. Available: https://www.datacamp.com/tutorial/what-is-a-confusion-matrix-in-machine-learning

[25] I. Kurniawan and A. Susanto, "Implementasi Metode K-Means dan Naïve Bayes Classifier untuk Analisis Sentimen Pemilihan Presiden (Pilpres) 2019," *Eksplora Inform.*, vol. 9, no. 1, pp. 1–10, 2019, doi: 10.30864/eksplora.v9i1.237.

[26] Siti Alvi Sholikhatin Khairunnisak Nur Isnaini, "Faculty of Sains and Technology, Ibrahimy University," *Ilm. Inform.*, vol. 6, no. 1, pp. 43–49, 2021, [Online]. Available: Siti Alvi Sholikhatin1), Khairunnisak Nur Isnaini2)