



Computer Based Information System Journal

ISSN (Print): 2337-8794 | E- ISSN : 2621-5292
web jurnal : <http://ejournal.upbatam.ac.id/index.php/cbis>



ANALISIS DATA MINING DENGAN METODE LOGISTIC REGRESSION DAN RANDOM FOREST UNTUK PREDIKSI PELUNASAN DAN RISIKO GAGAL BAYAR DEBITUR

Elvis¹, Erlin Elisa²

Universitas Putera Batam, Indonesia ^{1,2}

INFORMASI ARTIKEL

Diterima Redaksi: 2025-10-24
Diterbitkan Online: 2026-09-30

KATA KUNCI

Credit Scoring; Data Mining; Default Risk; Logistic Regression; Random Forest

KORESPONDENSI

E-mail:
pb221510001@upbatam.ac.id

A B S T R A C T

The financing sector in Indonesia faces increasing bad debt risks, particularly in motor vehicle financing. PT XYZ grapples with an 18.2% default rate due to subjective manual credit assessments. This study aims to compare Logistic Regression and Random Forest algorithms to predict creditworthiness using historical data from 2024-2025. The methodology involves data cleaning, transformation, and modeling, with performance evaluated via accuracy and Area Under Curve (AUC) metrics. Results indicate that Random Forest outperforms Logistic Regression, achieving an accuracy of 85% and an AUC of 0.914. It effectively distinguishes high-risk customers and minimizes false negatives, reducing the risk of rejecting eligible applicants. This research demonstrates that data mining, specifically Random Forest, offers a more objective and accurate solution for managing default risk in financing companies.

I. Latar Belakang

Risiko kredit merupakan salah satu tantangan utama yang dihadapi oleh perusahaan pembiayaan, khususnya dalam pembiayaan kendaraan bermotor. Tingginya tingkat gagal bayar debitur dapat berdampak langsung pada kesehatan keuangan perusahaan. Proses analisis kredit secara konvensional sering kali bergantung pada penilaian subjektif dan keterbatasan pengolahan data historis yang besar.

Seiring dengan perkembangan teknologi informasi, pemanfaatan data mining menjadi solusi strategis dalam pengelolaan risiko kredit. Data mining memungkinkan perusahaan untuk menggali pola tersembunyi dari data historis debitur sehingga dapat digunakan sebagai dasar pengambilan keputusan yang lebih objektif dan akurat.

Penelitian ini bertujuan untuk membandingkan kinerja metode Logistic Regression dan Random Forest dalam memprediksi pelunasan dan risiko gagal bayar debitur. Hasil penelitian

diharapkan dapat memberikan kontribusi praktis bagi perusahaan pembiayaan serta kontribusi akademis dalam pengembangan penerapan data mining pada bidang manajemen risiko kredit.

II. Kajian Literatur

Kredit

Kredit didefinisikan sebagai penyediaan uang atau tagihan yang dapat dipersamakan dengan itu, berdasarkan persetujuan atau kesepakatan pinjam-meminjam antara bank dengan pihak lain (Naufal et al., 2024). Dalam kesepakatan ini, pihak peminjam berkewajiban melunasi utangnya setelah jangka waktu tertentu dengan pemberian bunga. Secara umum, kredit berfungsi sebagai layanan peminjaman di mana nasabah harus mengembalikan pokok pinjaman beserta imbalan sesuai persyaratan yang disetujui (Ngurawan et al., 2021).

Risiko Kredit

Risiko kredit adalah potensi kerugian yang timbul akibat kegagalan debitur dalam memenuhi kewajibannya sesuai dengan perjanjian yang telah disepakati (Dwi et al., 2024). Fenomena ini dapat terjadi dalam bentuk keterlambatan pembayaran atau gagal bayar total, yang berpotensi menggerus modal dan keuntungan lembaga keuangan (Kristina Situmorang et al., 2025). Risiko ini meningkat apabila perusahaan tidak memiliki mekanisme pengawasan dan prediksi yang memadai terhadap kemampuan bayar nasabah. (Fitriani & Kristianti Maharani, 2024).

Knowledge Discovery in Database (KDD)

Knowledge Discovery in Database (KDD) merupakan rangkaian proses untuk menggali informasi dan pola penting dari data yang bervolume besar (Elisa et al., 2020). KDD sering diidentikkan dengan data mining, meskipun data mining sebenarnya adalah salah satu langkah dalam proses KDD. Proses ini dimulai dari pengumpulan data mentah, pembersihan, hingga interpretasi pola untuk menghasilkan pengetahuan yang berguna bagi pengambilan keputusan (Widaningsih, 2019).

Data Mining

Data mining merupakan proses untuk menemukan pola, hubungan, dan informasi bermakna dari sekumpulan data berukuran besar dengan menggunakan teknik statistik, matematika, dan pembelajaran mesin (Fitriani & Elisa, 2021). Dalam konteks keuangan, data mining banyak dimanfaatkan untuk credit scoring dan prediksi risiko gagal bayar.

Klasifikasi

Klasifikasi adalah teknik dalam data mining yang bertujuan untuk mengelompokkan data ke dalam kelas tertentu berdasarkan atribut yang dimiliki (Srirahayu & Setya Pribadie, 2023). Metode klasifikasi banyak digunakan untuk memprediksi kondisi debitur, seperti lancar atau gagal bayar (Oktanisa & Supianto, 2018).

Logistic Regression

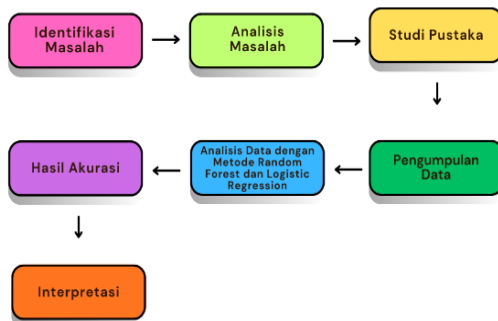
Logistic Regression merupakan metode statistik yang digunakan untuk memodelkan hubungan antara variabel independen dan variabel dependen biner (Gunawan & Gusti Putu Purnaba, 2022). Metode ini banyak digunakan dalam analisis kredit karena hasil modelnya mudah diinterpretasikan dan mampu menunjukkan pengaruh masing-masing variable (Kristiawan & Widjaja, 2021).

Random Forest

Random Forest adalah algoritma ensemble learning yang membangun banyak decision tree dan menggabungkan hasil prediksinya melalui mekanisme voting (Ananda Surya dkk., 2025). Metode ini memiliki keunggulan dalam menangani data berukuran besar dan hubungan non-linear antar variabel.

III. Metodologi

Penelitian ini menerapkan desain penelitian kuantitatif dengan pendekatan data mining atau predictive analytics untuk memecahkan permasalahan risiko kredit. Tahapan pelaksanaan penelitian disusun secara sistematis



Gambar 3.1 Desain Penelitian

Penelitian ini menggunakan desain penelitian kuantitatif dengan pendekatan predictive analytics yang diawali dengan tahap Identifikasi Masalah untuk menyoroti tingginya rasio gagal bayar sekitar 18% serta perlunya metode penilaian yang lebih objektif dibandingkan analisis manual. Selanjutnya, dilakukan Analisis Masalah untuk merumuskan batasan penelitian yang berfokus pada perbandingan algoritma Random Forest dan Logistic Regression, yang dijalankan beriringan dengan Studi Pustaka guna memperkuat landasan teori dan metodologi penelitian.

Tahapan berikutnya adalah Pengumpulan Data yang memanfaatkan data sekunder berupa catatan historis nasabah internal periode tahun 2024 hingga 2025. Setelah data terkumpul, proses berlanjut ke Analisis Data dan Pemodelan Algoritma, di mana dataset dibagi menjadi data latih (training) dan data uji (testing) untuk membangun model klasifikasi Logistic Regression dan Random Forest secara terpisah guna mensimulasikan prediksi pada data baru.

Setelah pemodelan selesai, dilakukan tahap evaluasi Hasil Akurasi untuk mengukur performa model secara ketat menggunakan metrik standar seperti Akurasi, Confusion Matrix, dan Area Under Curve (AUC). Rangkaian penelitian ini diakhiri dengan tahap Interpretasi, yaitu menentukan model mana yang memiliki kinerja terbaik untuk kemudian direkomendasikan sebagai solusi praktis dalam mendukung pengambilan keputusan kredit yang lebih efisien dan objektif bagi manajemen

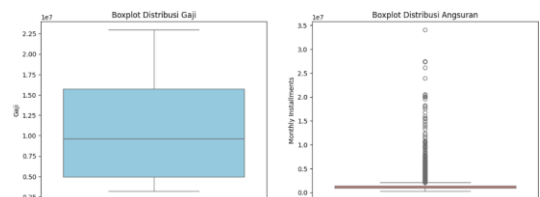
IV. Pembahasan

Penelitian diawali dengan tahapan seleksi data dari total 39.269 data mentah nasabah. Setelah dilakukan penyaringan berdasarkan periode jatuh tempo tahun 2024–2025, didapatkan 19.182 data yang relevan untuk dianalisis.

Total Baris (Observasi)	39269	19182
Total Kolom (Variabel)	62	10
Rentang Tahun	Semua Tahun	2024 - 2025

Gambar 1.1 Data Selection

Proses data cleaning selanjutnya dilakukan untuk menjamin kualitas data, di mana ditemukan dan dihapus 6 baris data yang memiliki nilai kosong (missing values) pada kolom OTR. Selain itu, metode Interquartile Range (IQR) diterapkan untuk mendeteksi outlier, yang menghasilkan penghapusan 1.963 baris data ekstrem yang berpotensi membiaskan model



Gambar 1.2 Boxplot

Setelah pembersihan, dilakukan feature engineering dengan menambahkan variabel baru yaitu Installment to Income Ratio (rasio beban angsuran terhadap pendapatan) untuk memperkuat analisis kemampuan bayar. Data akhir kemudian dibagi dengan rasio 80:20, menghasilkan 15.340 data latih untuk pembangunan model dan 3.836 data uji untuk evaluasi kinerja.

Hasil Pembagian Data:

- Jumlah Data Latih (Training Data 80%): 15340 baris
- Jumlah Data Uji (Testing Data 20%) : 3836 baris

Gambar 1.3 Data Splitting

Kedua algoritma, Logistic Regression dan Random Forest, dievaluasi menggunakan metrik Akurasi, Precision, Recall, F1-Score, dan ROC-AUC. Perbandingan kinerja kedua model disajikan dalam Tabel 1.1

Tabel 1.1 Perbandingan Performa Model

Model	Akurasi	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	81.73%	88.52%	80.85%	84.51%	903
Random Forest	85.22%	89.33%	86.34%	87.81%	914

Berdasarkan Tabel 1.1 algoritma Random Forest menunjukkan performa yang lebih unggul di seluruh metrik pengujian dibandingkan Logistic Regression. Random Forest mencatatkan tingkat akurasi sebesar 85,22% dan nilai AUC 0,914, sedangkan Logistic Regression memperoleh akurasi 81,73% dengan AUC 0,903. Keunggulan Random Forest juga terlihat pada nilai Recall sebesar 86,34%, yang mengindikasikan kemampuan model ini lebih baik dalam mendeteksi nasabah yang sebenarnya berisiko gagal bayar, sehingga dapat meminimalkan risiko False Negative bagi perusahaan.

Hasil penelitian menunjukkan bahwa Random Forest merupakan metode yang paling optimal untuk prediksi risiko kredit di PT XYZ. Keunggulan performa Random Forest disebabkan oleh karakteristik algoritma ensemble yang mampu menangkap hubungan non-linear dan interaksi kompleks antarvariabel yang sering ditemukan dalam data keuangan, berbeda dengan Logistic Regression yang cenderung bekerja lebih baik pada hubungan linear. Model Random Forest juga terbukti lebih tangguh (robust) terhadap variasi data dan tidak memerlukan asumsi distribusi data yang ketat.

Analisis lebih lanjut mengenai Feature Importance pada model Random Forest mengungkapkan faktor-faktor dominan yang memengaruhi prediksi risiko. Temuan menunjukkan bahwa variabel Gaji dan Installment to Income Ratio adalah dua prediktor paling kritis. Hal ini mengonfirmasi bahwa kapasitas finansial nasabah dan proporsi beban cicilan terhadap pendapatan memiliki pengaruh yang jauh lebih signifikan dalam menentukan

kelayakan kredit dibandingkan variabel demografis seperti usia atau jenis pekerjaan.

Implikasi dari temuan ini bagi PT XYZ adalah perlunya pergeseran dari penilaian manual yang subjektif ke penilaian berbasis data. Dengan akurasi mencapai 85%, penerapan model Random Forest dapat membantu manajemen mengambil keputusan kredit yang lebih objektif, menekan angka Non-Performing Loan (NPL), dan menetapkan ambang batas risiko yang lebih presisi sesuai dengan profil keuangan calon debitur

V. Kesimpulan

Pemanfaatan algoritma Logistic Regression dan Random Forest mampu menghasilkan model prediksi risiko gagal bayar dengan tingkat akurasi yang baik. Model Logistic Regression mencapai akurasi sebesar 82% dengan nilai ROC-AUC 0,9028, sedangkan model Random Forest menunjukkan kinerja yang lebih tinggi dengan akurasi 85% dan ROC-AUC 0,914. Hasil ini menunjukkan bahwa penerapan algoritma data mining secara signifikan meningkatkan ketepatan analisis risiko kredit dibandingkan pendekatan konvensional, sehingga mampu mendukung proses penilaian kredit yang lebih objektif dan terukur.

Penelitian ini membuktikan bahwa algoritma Random Forest merupakan model yang paling optimal dibandingkan Regresi Logistik dalam membangun sistem penilaian kredit berbasis data mining. Keunggulan Random Forest ditunjukkan melalui akurasi dan kemampuan diskriminatif yang lebih tinggi, serta kemampuannya menangani hubungan non-linear antarvariabel finansial debitur. Selain itu, analisis feature importance mengungkapkan bahwa variabel gaji dan rasio angsuran terhadap pendapatan merupakan faktor dominan dalam menentukan

kelayakan kredit. Dengan demikian, Random Forest lebih sesuai diterapkan sebagai model pendukung pengambilan keputusan kredit

Ucapan Terima Kasih

Penulis mengucapkan terima kasih kepada seluruh pihak yang telah membantu proses penelitian ini.

Daftar Pustaka

- Ananda Surya, A., Rizki Darmawan, D., & Solichin, A. (2025). Prediksi Kapabilitas Calon Debitur Menggunakan Analisis Data Machine Learning Dengan Metode Random Forest. *Jurnal Algoritma*, 22(1), 777–788. <https://doi.org/10.33364/algoritma/v.22-1.1929>
- Dwi, C., Putri, A., & Defung, F. (2024). Pengaruh Risiko Likuiditas dan Risiko Kredit serta Pertumbuhan Ekonomi terhadap Stabilitas Bank. *KINERJA: Jurnal Ekonomi Dan Manajemen*, 21(1), 2024–2071. <https://doi.org/10.30872/jkin.v21i1.14633>
- Elisa, E., Fajrin, A. A., & Kunci, K. (2020). Computer Based Information System Journal Algoritma Classifier Untuk Menganalisis Faktor Pemilihan Hunian Tempat Tinggal. *CBIS JOURNAL*, 08(02). <http://ejournal.upbatam.ac.id/index.php/cbishttp://ejournal.upbatam.ac.id/index.php/cbis>
- Fitriani, N., & Kristianti Maharani, N. (2024). Pengaruh Risiko Kredit, Risiko Likuiditas, Modal Bank Dan Rentabilitas Terhadap Profitabilitas Bank. 8(2).
- Fitrini, & Elisa, E. (2021). Pemanfaatan Data Mining Clustering Dalam Penentuan Media Promosi Umkm Di Kota Batam Fitrini 1 Erlin Elisap 2 P.
- Gunawan, K., & Gusti Putu Purnaba, I. (2022). Penerapan Analisis Regresi Logistik Ordinal pada Asuransi Kredit Perdagangan Domestik. *Jurnal Statistika Dan Aplikasinya*, 6(2), 366.
- Kristiawan, K., & Widjaja, A. (2021). Perbandingan Algoritma Machine Learning dalam Menilai Sebuah Lokasi Toko Ritel. *Jurnal Teknik Informatika Dan Sistem Informasi*, 7(1). <https://doi.org/10.28932/jutisi.v7i1.3182>
- Kristina Situmorang, T., Cahya Sagala, I., Dewi, S., & Rahman, F. (2025). Balance: Jurnal Akuntansi dan Manajemen. In *Balance: Jurnal Akuntansi dan Manajemen* (Vol. 4, Issue 2).
- Naufal, M. A., Ja'faruddin, J., & Ihsan, H. (2024). Ethnomathematics Exploration of Maggurecceng, Engklek, and Sayyang Pattudu in the Concept of Number. *Sainsmat: Jurnal Ilmiah Ilmu Pengetahuan Alam*, 13(2), 126. <https://doi.org/10.35580/sainsmat132515432024>
- Ngurawan, Y., Morasa, J., & Kapojos, P. M. (2021). *iogi2018,+06.++Yusril+Indra+Ngurawan.+OK*.
- Oktanisa, I., & Supianto, A. A. (2018). Perbandingan Teknik Klasifikasi Dalam Data Mining Untuk Bank Direct Marketing. 5(5), 567–576. <https://doi.org/10.25126/jtiik20185958>
- Widaningsih, S. (2019). PERBANDINGAN METODE DATA MINING UNTUK PREDIKSI NILAI DAN WAKTU KELULUSAN MAHASISWA PRODI TEKNIK INFORMATIKA DENGAN ALGORITMA C4,5, NAÏVE BAYES, KNN DAN SVM. *Jurnal Tekno Insentif*, 13(1), 16–25. <https://doi.org/10.36787/jti.v13i1.78>