Journal Sensi:Strategic Of Education in Information System
Accreditation rank 5
SK Dirjen DIKTI, RisTek No. 10/C/C3/DT.05.00/2025; E-ISSN: 2655-5298 P-ISSN: 2461-1409

PENERAPAN DATA MINING UNTUK PREDIKSI KELAYAKAN PEMBERIAN KREDIT NASABAH BANK MENGGUNAKAN ALGORITMA K-NEAREST NEIGHBOR (KNN)

Nur Azizah[1], Zulfan Al-Fauzan[2], Rizki Andriyan[3], Ahmad Faqih Mubarok[4]
[1]Program Studi Magister Teknik Informatika, Universitas Raharja;
[234]Program Studi Manajemen Informatika, Politeknik LP3I Jakarta
Email: [1]nur.azizah@raharja.info, [2]zlfnlfzn@gmail.com, [3]andriyanrikzi8@gmail.com, [4]faqih7687@gmail.com

Abstrak

Pemberian kredit merupakan salah satu layanan utama pada sektor perbankan yang memiliki risiko tinggi terhadap terjadinya kredit bermasalah apabila proses penilaian kelayakan nasabah tidak dilakukan secara akurat. Proses evaluasi yang masih dilakukan secara manual cenderung membutuhkan waktu yang lama serta berpotensi dipengaruhi oleh faktor subjektivitas analis kredit. Oleh karena itu, diperlukan pendekatan berbasis data yang mampu mendukung pengambilan keputusan secara lebih objektif, cepat, dan konsisten. Penelitian ini bertujuan menerapkan teknik *data mining* menggunakan algoritma K-Nearest Neighbor (KNN) untuk memprediksi kelayakan pemberian kredit nasabah berdasarkan data historis pengajuan kredit. Metode penelitian menggunakan tahapan Knowledge Discovery in Databases (KDD) yang meliputi *data selection, data preprocessing, data transformation, data mining,* dan *evaluation*. Dataset yang digunakan adalah Loan Prediction Dataset dari Kaggle yang terdiri atas 614 data dan 12 atribut, kemudian diolah menggunakan RapidMiner Studio. Tahap transformasi dilakukan melalui *label encoding* pada atribut kategorikal dan normalisasi atribut numerik sebelum dataset dibagi menjadi 80% data pelatihan dan 20% data pengujian. Model KNN dibangun menggunakan pengukuran Euclidean Distance dan dievaluasi menggunakan Confusion Matrix dengan metrik *accuracy, precision,* dan *recall*. Hasil penelitian menunjukkan bahwa model memperoleh accuracy sebesar 74,80%**,** precision sebesar 75,31%**,** dan recall sebesar 74,80%**.** Hasil tersebut menunjukkan bahwa algoritma K-Nearest Neighbor mampu memberikan performa klasifikasi yang cukup baik sehingga berpotensi digunakan sebagai alternatif pendukung keputusan dalam proses penilaian kelayakan pemberian kredit secara lebih objektif dan berbasis data.

Kata Kunci: *Data Mining*, *Knowledge Discovery in Database (KDD)*, *K-Nearest Neighbor (KNN)*, Kelayakan Kredit, RapidMiner.

Abstract

Credit granting is a core banking service that carries a high risk of non-performing loans if the customer eligibility assessment process is not conducted accurately. Manual evaluation processes tend to be time-consuming and are potentially influenced by the subjectivity of credit analysts. Therefore, a data-driven approach is needed to support decision-making that is more objective, rapid, and consistent. This study aims to apply data mining techniques using the K-Nearest Neighbor (KNN) algorithm to predict customer credit eligibility based on historical loan application data. The research methodology follows the Knowledge Discovery in Databases (KDD) process, comprising data selection, data preprocessing, data transformation, data mining, and evaluation. The dataset used is the "Loan Prediction Dataset" from Kaggle, consisting of 614 records and 12 attributes, processed using RapidMiner Studio. The transformation stage involved label encoding for categorical attributes and normalization for numerical attributes before splitting the dataset into 80% training data and 20% testing data. The KNN model was constructed using the Euclidean Distance metric and evaluated using a confusion matrix based on accuracy, precision, and recall. The results indicate that the model achieved an accuracy of 74.80%, a precision of 75.31%, and a recall of 74.80%. These findings demonstrate that the K-Nearest Neighbor algorithm delivers satisfactory classification performance, making it a viable alternative for supporting objective, data-driven decision-making in the credit eligibility assessment process.

Keywords: *Data Mining, Knowledge Discovery in Databases (KDD), K-Nearest Neighbor (KNN), Credit Eligibility, RapidMiner.*

1. PENDAHULUAN

Perkembangan teknologi informasi telah mendorong peningkatan volume data secara signifikan pada berbagai sektor, termasuk sektor perbankan. Digitalisasi layanan perbankan menyebabkan setiap aktivitas nasabah, mulai dari pembukaan rekening, transaksi keuangan, hingga pengajuan kredit, menghasilkan data dalam jumlah besar yang terus bertambah setiap hari. Kondisi tersebut menjadikan data sebagai salah satu aset strategis yang tidak hanya berfungsi sebagai media penyimpanan informasi, tetapi juga sebagai sumber pengetahuan yang dapat dimanfaatkan untuk mendukung proses pengambilan keputusan secara lebih efektif dan berbasis data [1–3].

Pemanfaatan data dalam jumlah besar memerlukan pendekatan analitis yang mampu mengekstraksi informasi bernilai dari kumpulan data yang kompleks. Salah satu pendekatan yang banyak digunakan adalah data mining, yaitu proses menemukan pola, hubungan, maupun informasi tersembunyi dari sekumpulan data menggunakan teknik statistik, pembelajaran mesin (machine learning), dan basis data. Informasi yang dihasilkan melalui proses tersebut dapat digunakan untuk mendukung pengambilan keputusan pada berbagai bidang, seperti pemasaran, kesehatan, pendidikan, manufaktur, hingga sektor jasa keuangan [1, 3, 7].

Dalam industri perbankan, penerapan data mining semakin berkembang seiring meningkatnya kebutuhan akan sistem analisis yang mampu memberikan keputusan secara cepat, konsisten, dan objektif. Salah satu penerapan yang paling banyak dikembangkan adalah analisis kelayakan pemberian kredit (credit scoring). Proses ini bertujuan untuk mengidentifikasi tingkat risiko calon nasabah berdasarkan karakteristik finansial maupun nonfinansial sebelum keputusan pemberian kredit dilakukan. Dengan demikian, lembaga keuangan dapat meminimalkan risiko kredit bermasalah (non-performing loan) sekaligus meningkatkan kualitas proses evaluasi kredit [4–6].

Secara umum, credit scoring merupakan proses klasifikasi yang memanfaatkan data historis untuk memprediksi apakah seorang calon nasabah layak atau tidak layak menerima fasilitas kredit. Model credit scoring dibangun berdasarkan berbagai atribut yang dianggap memengaruhi kemampuan pengembalian pinjaman, seperti tingkat pendapatan, status pekerjaan, jumlah pinjaman, jangka waktu kredit, riwayat pembayaran, jumlah tanggungan, tingkat pendidikan, serta lokasi properti. Hubungan antaratribut tersebut kemudian dipelajari untuk menghasilkan model yang mampu memberikan prediksi terhadap pengajuan kredit baru secara lebih objektif dibandingkan proses evaluasi manual [4, 5, 8].

Meskipun demikian, proses penilaian kelayakan kredit pada banyak lembaga keuangan masih melibatkan evaluasi manual yang memerlukan analisis terhadap berbagai dokumen pendukung. Pendekatan tersebut membutuhkan waktu yang relatif lama dan berpotensi menghasilkan keputusan yang tidak konsisten karena dipengaruhi oleh pengalaman maupun subjektivitas analis kredit. Selain itu, meningkatnya jumlah pengajuan kredit menyebabkan proses evaluasi secara manual menjadi semakin sulit dilakukan secara efisien. Oleh karena itu, penerapan metode klasifikasi berbasis machine learning menjadi salah satu alternatif yang banyak dikembangkan untuk meningkatkan kecepatan, konsistensi, dan akurasi proses penilaian kelayakan kredit [5, 6, 8].

Berbagai algoritma klasifikasi telah diterapkan dalam penelitian mengenai credit scoring. Algoritma Decision Tree, Naïve Bayes, Support Vector Machine (SVM), Random Forest, hingga K-Nearest Neighbor (KNN) merupakan beberapa metode yang sering digunakan karena memiliki karakteristik dan tingkat performa yang berbeda pada setiap jenis dataset. Pemilihan algoritma yang tepat menjadi salah satu faktor penting dalam menghasilkan model klasifikasi yang memiliki kemampuan generalisasi yang baik terhadap data baru [9–13].

Algoritma K-Nearest Neighbor (KNN) merupakan salah satu metode klasifikasi berbasis supervised learning yang bekerja dengan menentukan kelas suatu data berdasarkan kedekatan terhadap sejumlah data pelatihan. Algoritma ini memiliki mekanisme yang sederhana, tidak memerlukan proses pembentukan model yang kompleks (lazy learning), serta mampu menghasilkan performa klasifikasi yang kompetitif pada berbagai permasalahan prediksi. Selain itu, KNN relatif mudah diimplementasikan dan memiliki kemampuan yang baik dalam menangani data dengan pola distribusi yang beragam apabila didukung oleh proses prapemrosesan data yang tepat [1, 9, 10].

Beberapa penelitian terdahulu menunjukkan bahwa algoritma KNN memiliki potensi yang baik dalam menyelesaikan permasalahan klasifikasi pada bidang keuangan. Harlina menerapkan algoritma KNN yang dikombinasikan dengan forward selection untuk memprediksi kelayakan kredit dan menunjukkan adanya peningkatan performa klasifikasi dibandingkan penggunaan KNN tanpa seleksi fitur [14]. Penelitian yang dilakukan oleh Faqih et al. juga menunjukkan bahwa algoritma KNN mampu mengklasifikasikan pengajuan kartu kredit berdasarkan karakteristik data historis dengan tingkat akurasi yang memadai [15]. Selain itu, Nawawi dan Suhendar berhasil menerapkan algoritma KNN sebagai metode pendukung keputusan dalam penentuan kelayakan kredit UMKM [16]. Penelitian lain oleh Santoso, Kusrini, dan Hartanto membandingkan algoritma Random Forest dan K-Nearest Neighbor pada prediksi kelayakan kredit, sedangkan Sari dan Simamora mengembangkan metode Bootstrap Aggregating K-Nearest Neighbor untuk meningkatkan stabilitas model klasifikasi pada analisis credit scoring [13, 17].

Meskipun berbagai penelitian telah berhasil menerapkan algoritma klasifikasi pada permasalahan *credit scoring*, masih terdapat beberapa aspek yang memerlukan perhatian lebih lanjut. Sebagian besar penelitian berfokus pada perbandingan tingkat akurasi antaralgoritma, seperti Decision Tree, Random Forest, Support Vector Machine, maupun K-Nearest Neighbor, tanpa memberikan penjelasan yang komprehensif mengenai tahapan pengolahan data sebelum proses klasifikasi dilakukan [6, 8, 13]. Padahal, kualitas data merupakan salah satu faktor yang sangat memengaruhi performa model klasifikasi. Dataset yang masih mengandung atribut

kategorikal, memiliki skala data yang berbeda, atau belum melalui proses transformasi yang tepat dapat menghasilkan model dengan performa yang kurang optimal meskipun menggunakan algoritma yang memiliki kemampuan klasifikasi yang baik [1, 2, 3].

Selain itu, beberapa penelitian lebih menitikberatkan pada hasil evaluasi model dibandingkan dokumentasi proses pembangunan model secara menyeluruh. Tahapan seperti pemilihan atribut, prapemrosesan data, transformasi data, hingga evaluasi model sering kali hanya dijelaskan secara singkat sehingga proses penelitian menjadi kurang mudah direplikasi. Padahal, dokumentasi setiap tahapan pengolahan data merupakan bagian penting dalam penelitian *data mining* karena berpengaruh terhadap validitas dan konsistensi hasil yang diperoleh [2, 7, 9].

Permasalahan lain yang ditemukan pada penelitian terdahulu berkaitan dengan penggunaan dataset. Beberapa penelitian menggunakan data yang berasal dari institusi tertentu sehingga karakteristik dataset cenderung spesifik terhadap lingkungan penelitian masing-masing. Kondisi tersebut menyebabkan hasil yang diperoleh belum tentu dapat digeneralisasikan pada dataset lain yang memiliki distribusi atribut berbeda. Oleh karena itu, penggunaan dataset yang tersedia secara publik menjadi salah satu alternatif untuk meningkatkan reproduktifitas penelitian sekaligus mempermudah proses perbandingan hasil dengan penelitian sebelumnya [6, 8].

Berdasarkan kondisi tersebut, penelitian ini menggunakan Loan Prediction Dataset yang diperoleh dari platform Kaggle sebagai sumber data penelitian [18]. Dataset tersebut dipilih karena telah banyak digunakan dalam penelitian prediksi kelayakan kredit dan memiliki atribut yang mewakili karakteristik umum calon nasabah, seperti jenis kelamin, status pernikahan, jumlah tanggungan, tingkat pendidikan, status pekerjaan, pendapatan pemohon, pendapatan pasangan, jumlah pinjaman, jangka waktu pinjaman, riwayat kredit, lokasi properti, serta status kelayakan kredit sebagai variabel target. Penggunaan dataset publik diharapkan dapat meningkatkan transparansi penelitian sekaligus memudahkan proses validasi maupun pengembangan penelitian pada masa mendatang.

Dalam penelitian ini dipilih algoritma K-Nearest Neighbor (KNN) sebagai metode klasifikasi karena memiliki mekanisme yang sederhana, mudah diimplementasikan, serta mampu memberikan performa yang kompetitif pada berbagai permasalahan klasifikasi. Berbeda dengan beberapa algoritma lain yang memerlukan proses pembentukan model (*model training*) yang kompleks, KNN bekerja dengan menghitung tingkat kedekatan antara data baru dan data historis menggunakan metode pengukuran jarak. Pendekatan tersebut memungkinkan proses klasifikasi dilakukan berdasarkan karakteristik data yang memiliki tingkat kemiripan tertinggi sehingga sesuai diterapkan pada permasalahan prediksi kelayakan kredit [10, 14, 16].

Meskipun demikian, performa algoritma KNN sangat dipengaruhi oleh kualitas data yang digunakan. Atribut dengan skala nilai yang berbeda dapat menyebabkan proses perhitungan Euclidean Distance menjadi kurang representatif apabila tidak dilakukan normalisasi terlebih dahulu. Selain itu, atribut kategorikal tidak dapat diproses secara langsung oleh algoritma KNN sehingga memerlukan proses transformasi ke dalam bentuk numerik sebelum dilakukan klasifikasi. Oleh karena itu, penelitian ini tidak hanya berfokus pada implementasi algoritma KNN, tetapi juga menekankan pentingnya tahapan pengolahan data sebagai bagian dari proses pembangunan model klasifikasi [1, 3, 9].

Untuk mendukung proses tersebut, penelitian ini menerapkan metode Knowledge Discovery in Databases (KDD) sebagai kerangka kerja penelitian. Metode KDD dipilih karena menyediakan tahapan yang sistematis dalam proses pengolahan data, mulai dari data selection**,** data preprocessing**,** data transformation**,** data mining, hingga evaluation [2, 7, 9]. Melalui penerapan metode KDD, setiap tahapan pengolahan data dapat dilakukan secara terstruktur sehingga kualitas data yang digunakan dalam proses klasifikasi dapat dipertahankan. Selain itu, penggunaan kerangka kerja KDD juga mempermudah proses dokumentasi penelitian sehingga setiap tahapan dapat direplikasi pada penelitian selanjutnya.

Seluruh tahapan penelitian diimplementasikan menggunakan perangkat lunak RapidMiner Studio [19]. Perangkat lunak ini dipilih karena menyediakan berbagai operator yang mendukung implementasi metode KDD secara terintegrasi, mulai dari pembacaan dataset, transformasi atribut, normalisasi data, pembagian data pelatihan dan data pengujian, pembangunan model klasifikasi menggunakan algoritma K-Nearest Neighbor, hingga evaluasi model menggunakan Confusion Matrix. Penggunaan RapidMiner juga memungkinkan proses eksperimen dilakukan secara visual sehingga memudahkan analisis maupun dokumentasi penelitian.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk menerapkan algoritma K-Nearest Neighbor (KNN) dalam memprediksi kelayakan pemberian kredit nasabah berdasarkan data historis pengajuan kredit menggunakan pendekatan Knowledge Discovery in Databases (KDD). Selain membangun model klasifikasi, penelitian ini juga mengevaluasi performa model menggunakan metrik Accuracy, Precision, dan Recall untuk mengetahui tingkat kemampuan algoritma dalam mengklasifikasikan data kelayakan kredit secara objektif.

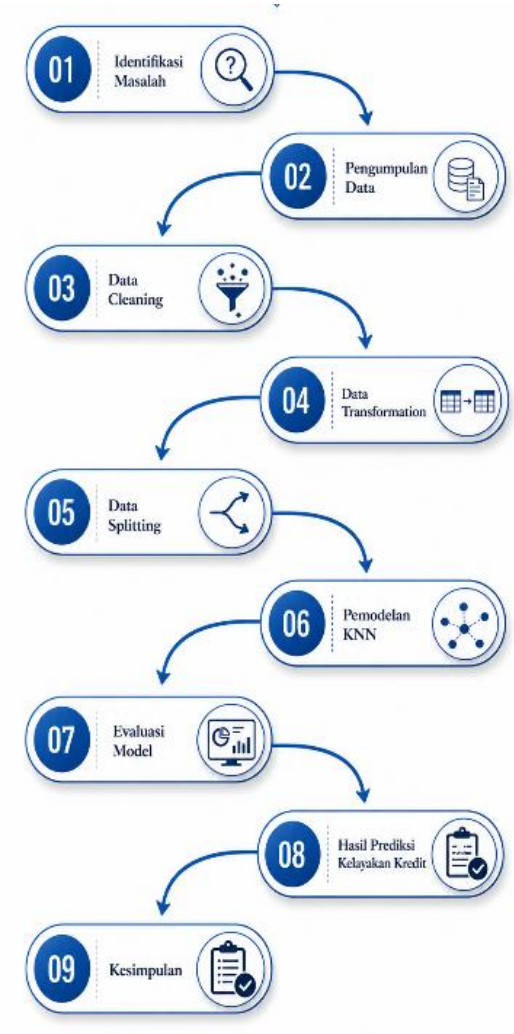
Kontribusi penelitian ini terletak pada penerapan tahapan Knowledge Discovery in Databases (KDD) secara menyeluruh dalam proses pembangunan model klasifikasi menggunakan algoritma K-Nearest Neighbor. Berbeda dengan beberapa penelitian sebelumnya yang lebih banyak berfokus pada hasil evaluasi model, penelitian ini mendokumentasikan setiap tahapan pengolahan data mulai dari pemilihan dataset, prapemrosesan, transformasi, pembangunan model, hingga evaluasi menggunakan Confusion Matrix. Dengan demikian, penelitian ini tidak hanya menghasilkan model prediksi kelayakan kredit, tetapi juga menyediakan alur implementasi yang sistematis dan mudah direplikasi sebagai referensi bagi penelitian selanjutnya.

Selanjutnya, untuk menjelaskan tahapan pembangunan model secara rinci, bagian berikutnya memaparkan metode penelitian yang digunakan. Pembahasan meliputi karakteristik dataset, tahapan Knowledge Discovery in Databases (KDD), proses prapemrosesan dan transformasi data, implementasi algoritma K-Nearest Neighbor menggunakan RapidMiner Studio, serta metode evaluasi yang digunakan untuk mengukur performa model klasifikasi.

2. METODE PENELITIAN

Penelitian ini menggunakan metode Knowledge Discovery in Databases (KDD) sebagai kerangka kerja dalam membangun model prediksi kelayakan pemberian kredit menggunakan algoritma K-Nearest Neighbor (KNN). Metode KDD dipilih karena menyediakan tahapan pengolahan data yang sistematis, mulai dari pemilihan data, prapemrosesan, transformasi data, pembangunan model, hingga evaluasi hasil klasifikasi [2, 7, 9]. Seluruh proses implementasi dilakukan menggunakan perangkat lunak RapidMiner Studio [19].

Secara umum penelitian terdiri atas lima tahapan utama, yaitu data selection, data preprocessing, data transformation, data mining, dan evaluation. Setiap tahapan saling berkaitan untuk menghasilkan model klasifikasi yang mampu memberikan prediksi kelayakan kredit secara objektif berdasarkan karakteristik data historis nasabah.



Gambar 1. Tahapan Knowledge Discovery in Databases (KDD)

A. Dataset Penelitian

Dataset yang digunakan pada penelitian ini merupakan Loan Prediction Dataset yang diperoleh dari platform Kaggle [18]. Dataset dipilih karena telah banyak digunakan sebagai acuan pada penelitian prediksi kelayakan kredit sehingga memungkinkan hasil penelitian dibandingkan dengan penelitian lain yang menggunakan sumber data yang sama. Dataset terdiri atas 614 data dengan 12 atribut, yang meliputi atribut kategorikal maupun numerik. Variabel target yang digunakan adalah Loan_Status, sedangkan atribut lainnya digunakan sebagai variabel prediktor dalam proses klasifikasi.

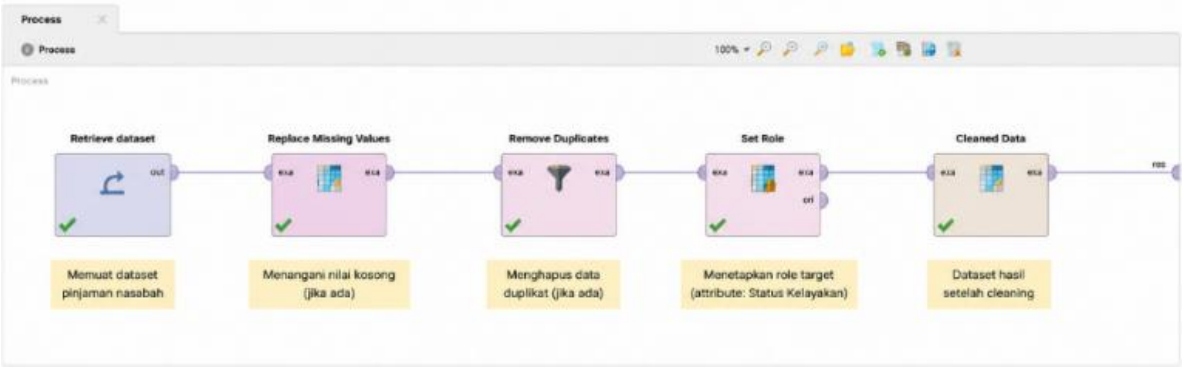
Tabel 1. Karakteristik Atribut Dataset Penelitian

No	Atribut	Jenis Data	Keterangan
1	Gender	Nominal	Jenis kelamin pemohon
2	Married	Nominal	Status pernikahan

No	Atribut	Jenis Data	Keterangan
3	Dependents	Nominal	Jumlah tanggungan
4	Education	Nominal	Tingkat pendidikan
5	Self_Employed	Nominal	Status pekerjaan
6	ApplicantIncome	Numerik	Pendapatan pemohon
7	CoapplicantIncome	Numerik	Pendapatan pasangan
8	LoanAmount	Numerik	Jumlah pinjaman
9	Loan_Amount_Term	Numerik	Jangka waktu pinjaman
10	Credit_History	Nominal	Riwayat kredit
11	Property_Area	Nominal	Lokasi properti
12	Loan_Status	Target	Status kelayakan kredit

B. Data Preprocessing

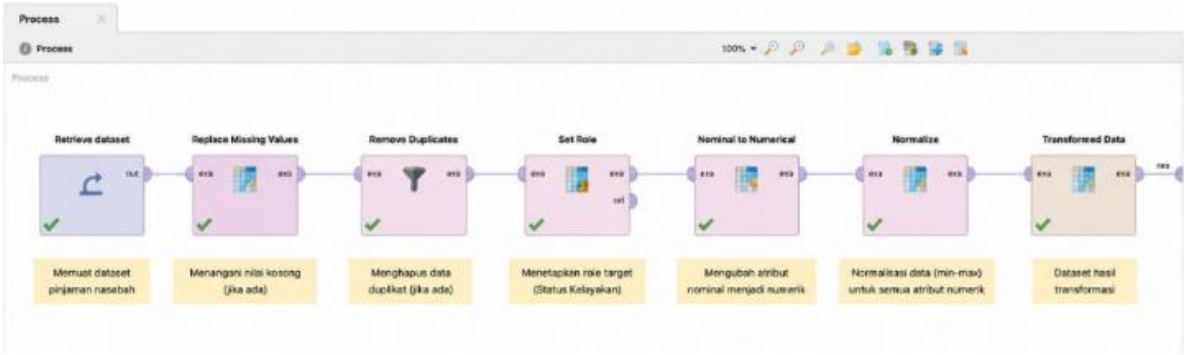
Tahap data preprocessing dilakukan untuk meningkatkan kualitas dataset sebelum proses klasifikasi. Pada tahap ini dilakukan pemeriksaan terhadap data duplikat, *missing value*, serta inkonsistensi nilai pada setiap atribut. Hasil pemeriksaan menunjukkan bahwa dataset tidak mengandung data duplikat sehingga seluruh data dapat digunakan pada tahapan berikutnya. Proses prapemrosesan juga memastikan setiap atribut memiliki representasi data yang konsisten sebelum dilakukan transformasi [1, 2, 7].



Gambar 2. Proses Data Preprocessing

C. Data Transformation

Tahap data transformation dilakukan untuk mengubah seluruh atribut ke dalam format yang dapat diproses oleh algoritma K-Nearest Neighbor. Atribut kategorikal dikonversi menggunakan metode Label Encoding, sedangkan atribut numerik dinormalisasi menggunakan metode Min-Max Normalization agar seluruh atribut berada pada rentang nilai yang seragam. Proses normalisasi diperlukan karena algoritma KNN menggunakan pengukuran jarak Euclidean Distance yang sensitif terhadap perbedaan skala data [1, 3, 9].



Gambar 3. Proses Transformasi Data

D. Pemodelan K-Nearest Neighbor

Model klasifikasi dibangun menggunakan algoritma K-Nearest Neighbor (KNN). Setelah proses transformasi selesai dilakukan, dataset dibagi menjadi 80% data pelatihan dan 20% data pengujian menggunakan operator Split Data pada RapidMiner. Selanjutnya algoritma menghitung tingkat kedekatan antara data uji dengan seluruh data pelatihan menggunakan **metode** Euclidean Distance.

Proses klasifikasi pada algoritma K-Nearest Neighbor dilakukan dengan menghitung tingkat kedekatan antara data uji dan data pelatihan menggunakan metode Euclidean Distance. Data dengan nilai jarak terkecil dianggap sebagai tetangga terdekat (*nearest neighbor*) dan digunakan sebagai dasar penentuan kelas terhadap data baru [1, 10, 16].

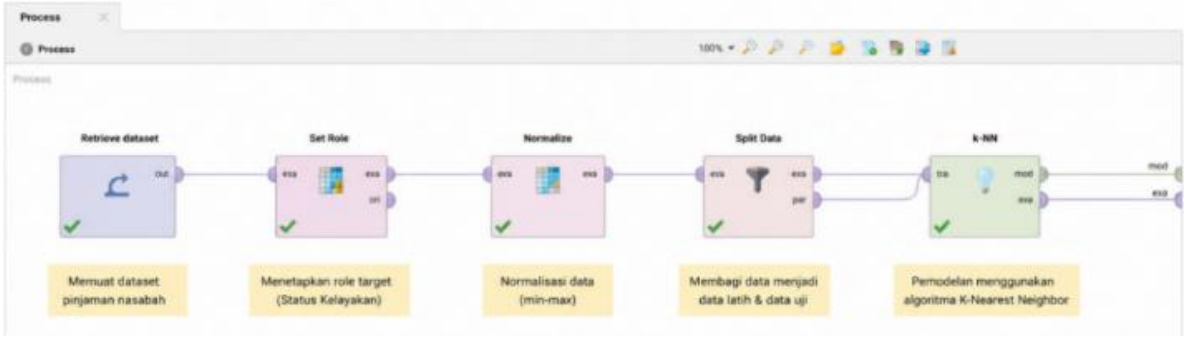
Persamaan Euclidean Distance ditunjukkan pada Persamaan (1).

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

dengan:

- $d\ (x,y)$ = jarak antara dua data,
- x_i = nilai atribut pada data uji,
- y_i = nilai atribut pada data pelatihan,
- n = jumlah atribut yang digunakan.

Setelah seluruh jarak dihitung, algoritma memilih sejumlah tetangga terdekat berdasarkan nilai parameter K = 5. Selanjutnya kelas mayoritas dari tetangga tersebut digunakan sebagai hasil prediksi terhadap data uji.



Gambar 4. Model K-Nearest Neighbor pada RapidMiner

E. Evaluasi Model

Performa model dievaluasi menggunakan Confusion Matrix yang menghasilkan nilai Accuracy, Precision, dan Recall. Accuracy digunakan untuk mengukur tingkat ketepatan model dalam mengklasifikasikan seluruh data uji, sedangkan Precision menunjukkan ketepatan prediksi terhadap kelas tertentu dan Recall menggambarkan kemampuan model dalam mengenali seluruh data pada kelas yang sama [20].

Nilai Accuracy dihitung menggunakan Persamaan (2).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

Nilai **Precision** dihitung menggunakan Persamaan (3).

$$Precision = \frac{TP}{FP + TP} \times 100\%$$

Nilai **Recall** dihitung menggunakan Persamaan (4).

$$Recall = \frac{TP}{FP + FN} \times 100\%$$

dengan:

- TP (True Positive) = jumlah data positif yang diprediksi benar,
- TN (True Negative) = jumlah data negatif yang diprediksi benar,
- FP (False Positive) = jumlah data negatif yang diprediksi positif,
- FN (False Negative) = jumlah data positif yang diprediksi negatif.

Gambar 5. Proses Evaluasi Model

3. HASIL DAN PEMBAHASAN

Bagian ini menyajikan hasil implementasi algoritma K-Nearest Neighbor (KNN) untuk memprediksi kelayakan pemberian kredit berdasarkan tahapan Knowledge Discovery in Databases (KDD). Pembahasan difokuskan pada hasil pengolahan data, pembangunan model klasifikasi, evaluasi performa model menggunakan Confusion Matrix, serta interpretasi hasil berdasarkan teori dan penelitian terdahulu. Pendekatan tersebut dilakukan untuk mengetahui sejauh mana algoritma KNN mampu mengklasifikasikan data kelayakan kredit secara objektif berdasarkan karakteristik data historis [1, 2, 20].

A. Hasil Pengolahan Data

Tahap pengolahan data dilakukan untuk menghasilkan dataset yang siap digunakan pada proses klasifikasi. Seluruh atribut kategorikal dikonversi menjadi nilai numerik menggunakan metode Label Encoding, sedangkan atribut numerik dinormalisasi menggunakan Min-Max Normalization sehingga seluruh atribut memiliki rentang nilai yang seragam. Proses ini diperlukan karena algoritma K-Nearest Neighbor menggunakan pengukuran jarak Euclidean Distance, sehingga perbedaan skala data dapat memengaruhi hasil klasifikasi apabila tidak dilakukan normalisasi terlebih dahulu [1, 3, 9].

Transformasi data menghasilkan seluruh atribut dalam bentuk numerik tanpa mengubah makna informasi yang dikandungnya. Dengan demikian, setiap atribut dapat diproses secara optimal pada tahap klasifikasi menggunakan algoritma KNN.

Tabel 2. Hasil Transformasi Data Menggunakan Label Encoding dan Min-Max Normalization

Atribut	Sebelum Transformasi	Sesudah Transformasi	Metode
Gender	Laki-laki / Perempuan	1 / 0	Label Encoding
Status Menikah	Ya / Tidak	1 / 0	Label Encoding
Pendidikan	Sarjana / Tidak Sarjana	1 / 0	Label Encoding
Pekerjaan	Ya / Tidak	1 / 0	Label Encoding
Lokasi Properti	Pedesaan / Semi Perkotaan / Perkotaan	0 / 1 / 2	Label Encoding
Tenor, Jumlah Pinjaman, Income Pemohon, Income Pasangan	Nilai numerik	0,000–1,000	Min-Max Normalization
Status Kelayakan	Layak / Tidak Layak	1 / 0	Label Encoding

Gambar 6. Hasil Transformasi Data

B. Pembagian Dataset

Setelah proses transformasi selesai dilakukan, dataset dibagi menjadi 491 data pelatihan dan 123 data pengujian menggunakan rasio 80:20. Proporsi tersebut dipilih untuk memberikan jumlah data pelatihan yang memadai sehingga model mampu mempelajari karakteristik data historis, sementara data pengujian digunakan untuk mengevaluasi kemampuan model dalam mengklasifikasikan data yang belum pernah dipelajari sebelumnya. Rasio pembagian ini merupakan salah satu pendekatan yang umum digunakan pada penelitian klasifikasi karena mampu memberikan keseimbangan antara proses pembelajaran model dan evaluasi performa [1, 5].

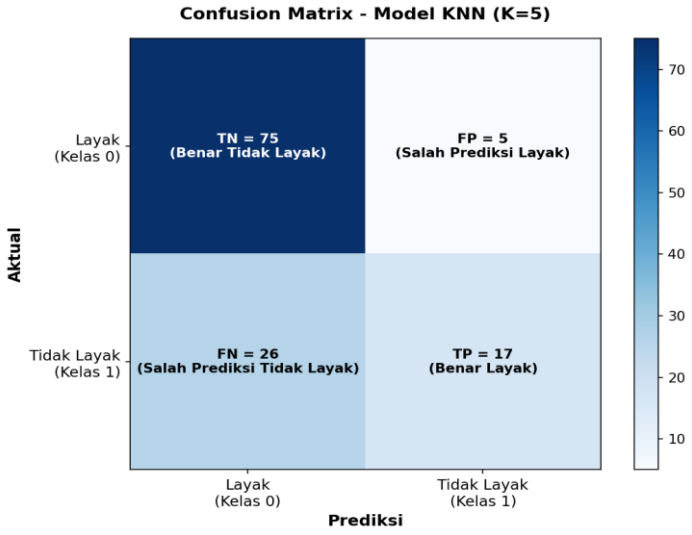
Tabel 3. Pembagian Dataset Penelitian

Jenis Data	Jumlah Data	Persentase	Keterangan
Data Training	491	80%	Digunakan untuk membangun model KNN
Data Testing	123	20%	Digunakan untuk pengujian model
Total	614	100%	Dataset penelitian

Pembagian dataset menggunakan rasio 80:20 dipilih karena mampu memberikan jumlah data pelatihan yang cukup besar untuk proses pembelajaran model, sekaligus menyediakan data pengujian yang memadai untuk mengevaluasi performa klasifikasi secara objektif.

C. Hasil Pemodelan K-Nearest Neighbor

Model klasifikasi dibangun menggunakan algoritma K-Nearest Neighbor dengan parameter K = 5. Setelah proses pelatihan selesai dilakukan, model diterapkan pada data pengujian untuk menghasilkan prediksi terhadap status kelayakan pemberian kredit. Hasil klasifikasi kemudian dievaluasi menggunakan Confusion Matrix guna mengetahui kemampuan model dalam mengidentifikasi setiap kelas secara benar maupun kesalahan klasifikasi yang terjadi [20].



Gambar 7 Confusion Matrix Hasil Klasifikasi KNN

Hasil Confusion Matrix menunjukkan bahwa model berhasil mengklasifikasikan 75 data layak dan 17 data tidak layak secara benar. Di sisi lain masih terdapat 5 data layak yang diprediksi

sebagai tidak layak (False Positive) serta 26 data tidak layak yang diprediksi sebagai layak (False Negative).

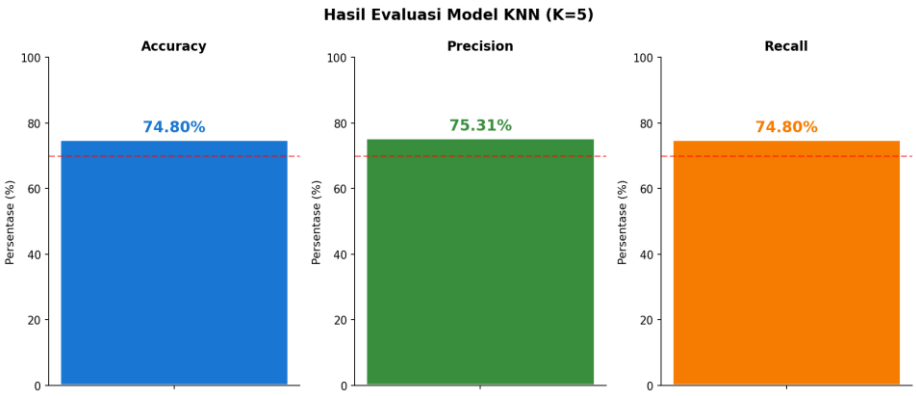
Nilai False Negative yang lebih tinggi dibandingkan False Positive menunjukkan bahwa model masih mengalami kesulitan dalam mengenali sebagian data pada kelas tidak layak. Dalam konteks pemberian kredit, kondisi tersebut perlu menjadi perhatian karena dapat menyebabkan calon nasabah yang memiliki risiko gagal bayar diprediksi sebagai layak menerima kredit. Oleh karena itu, peningkatan kemampuan model dalam mengenali kelas minoritas menjadi salah satu aspek yang dapat dikembangkan pada penelitian selanjutnya [5, 6, 8].

Tabel 4. Hasil Confusion Matrix

Komponen	Nilai	Keterangan
True Positive (TP)	17	Data Tidak Layak diprediksi benar
True Negative (TN)	75	Data Layak diprediksi benar
False Positive (FP)	5	Data Layak diprediksi Tidak Layak
False Negative (FN)	26	Data Tidak Layak diprediksi Layak
Total Data Testing	123	TP + TN + FP + FN

D. Evaluasi Model

Performa model dievaluasi menggunakan metrik Accuracy, Precision, dan Recall yang diperoleh melalui Confusion Matrix. Ketiga metrik tersebut digunakan untuk memberikan gambaran menyeluruh mengenai kemampuan model dalam melakukan klasifikasi terhadap data pengujian [20].



Gambar 8. Hasil Evaluasi Model KNN

Model menghasilkan Accuracy sebesar 74,80%, yang menunjukkan bahwa sebagian besar data pengujian berhasil diklasifikasikan dengan benar. Nilai tersebut mengindikasikan bahwa algoritma K-Nearest Neighbor mampu mempelajari pola hubungan antaratribut pada dataset sehingga dapat digunakan sebagai model prediksi kelayakan kredit dengan tingkat ketepatan yang cukup baik.

Nilai Precision sebesar 75,31% menunjukkan bahwa sebagian besar prediksi yang dihasilkan model sesuai dengan kondisi sebenarnya. Sementara itu, Recall sebesar 74,80% menunjukkan

bahwa model mampu mengenali sebagian besar data pada masing-masing kelas, meskipun masih terdapat sejumlah data yang belum berhasil diklasifikasikan secara benar. Secara keseluruhan, ketiga metrik tersebut menunjukkan bahwa model memiliki performa yang cukup stabil dalam mengklasifikasikan data kelayakan kredit.

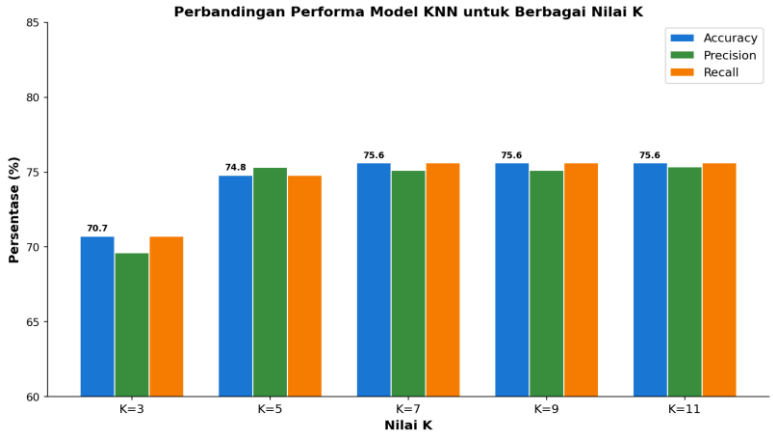
Tabel 5. Rekapitulasi Hasil Evaluasi Model Knn

Metrik	Nilai
Accuracy	**74,80%**
Precision	**75,31%**
Recall	**74,80%**
TP	17
TN	75
FP	5
FN	26

Berdasarkan hasil tersebut, algoritma KNN mampu menghasilkan performa klasifikasi yang cukup baik dalam memprediksi kelayakan pemberian kredit. Meskipun demikian, nilai *False Negative* yang relatif lebih tinggi menunjukkan bahwa masih terdapat data nasabah yang seharusnya dikategorikan tidak layak tetapi diprediksi sebagai layak. Kondisi ini menjadi perhatian penting karena berpotensi meningkatkan risiko pemberian kredit kepada nasabah yang memiliki kemungkinan gagal bayar.

E. Perbandingan Performa pada Berbagai Nilai K

Pengujian juga dilakukan pada beberapa nilai **K** untuk mengetahui pengaruh parameter tersebut terhadap performa algoritma K-Nearest Neighbor.



Gambar 9. Perbandingan Performa Model KNN pada Berbagai Nilai K

Tabel 6. Perbandingan Performa Model Knn

Nilai K	Accuracy	Precision	Recall
3	70,73%	69,63%	70,73%
5	74,80%	75,31%	74,80%
7	75,61%	75,13%	75,61%
9	75,61%	75,13%	75,61%
11	75,61%	75,35%	75,61%

Hasil pengujian menunjukkan bahwa peningkatan nilai K memberikan perubahan terhadap performa model. Nilai K = 3 menghasilkan Accuracy sebesar 70,73%, sedangkan peningkatan nilai K menjadi 7, 9, dan 11 menghasilkan Accuracy sebesar 75,61%. Hasil tersebut menunjukkan bahwa pemilihan parameter K berpengaruh terhadap kemampuan model dalam mengklasifikasikan data.

Meskipun nilai Accuracy tertinggi diperoleh pada K = 7, K = 9, dan K = 11, penelitian ini tetap menggunakan K = 5 sebagai konfigurasi utama karena parameter tersebut digunakan pada proses implementasi model yang dijelaskan pada penelitian ini serta menghasilkan keseimbangan nilai Accuracy, Precision, dan Recall. Oleh karena itu, konfigurasi tersebut dipertahankan agar seluruh tahapan eksperimen tetap konsisten.

F. Pembahasan

Hasil penelitian menunjukkan bahwa penerapan algoritma K-Nearest Neighbor mampu memberikan performa klasifikasi yang cukup baik dalam memprediksi kelayakan pemberian kredit berdasarkan karakteristik data historis. Keberhasilan model tidak hanya dipengaruhi oleh algoritma yang digunakan, tetapi juga oleh kualitas proses pengolahan data melalui tahapan Knowledge Discovery in Databases (KDD). Proses Label Encoding dan Min-Max Normalization memungkinkan seluruh atribut berada pada format yang sesuai untuk proses perhitungan Euclidean Distance, sehingga meningkatkan kualitas proses klasifikasi [1, 2, 7].

Hasil penelitian ini sejalan dengan penelitian Nawawi dan Suhendar [16], Santoso et al. [13], serta Sari dan Simamora [17], yang menunjukkan bahwa algoritma K-Nearest Neighbor mampu diterapkan pada permasalahan prediksi kelayakan kredit dengan tingkat performa yang kompetitif. Kesamaan hasil tersebut menunjukkan bahwa algoritma KNN masih menjadi salah satu metode klasifikasi yang relevan untuk diterapkan pada data kredit, terutama apabila didukung oleh proses prapemrosesan data yang baik.

Meskipun demikian, nilai False Negative yang masih relatif tinggi menunjukkan bahwa model memiliki keterbatasan dalam mengenali sebagian data pada kelas tidak layak. Kondisi tersebut menunjukkan bahwa pengembangan model melalui optimasi parameter K, penerapan teknik seleksi fitur, maupun penggunaan algoritma lain sebagai pembanding masih diperlukan untuk meningkatkan kemampuan model dalam mengenali kelas minoritas.

Secara keseluruhan, hasil penelitian membuktikan bahwa algoritma K-Nearest Neighbor yang diterapkan menggunakan pendekatan Knowledge Discovery in Databases (KDD) mampu memberikan hasil klasifikasi yang cukup baik dan berpotensi diterapkan sebagai sistem pendukung keputusan dalam proses penilaian kelayakan pemberian kredit pada lembaga keuangan.

4. KESIMPULAN

Penelitian ini berhasil menerapkan algoritma K-Nearest Neighbor (KNN) dengan pendekatan Knowledge Discovery in Databases (KDD) untuk memprediksi kelayakan pemberian kredit nasabah berdasarkan data historis pengajuan kredit. Seluruh tahapan penelitian, mulai dari pemilihan dataset, prapemrosesan, transformasi data, pembangunan model, hingga evaluasi, telah diimplementasikan secara sistematis menggunakan perangkat lunak RapidMiner Studio.

Hasil evaluasi menunjukkan bahwa model KNN menghasilkan nilai Accuracy sebesar 74,80%, Precision sebesar 75,31%, dan Recall sebesar 74,80%. Nilai tersebut menunjukkan bahwa algoritma KNN memiliki kemampuan yang cukup baik dalam mengklasifikasikan kelayakan pemberian kredit berdasarkan karakteristik data historis. Penerapan tahapan KDD juga memberikan proses pengolahan data yang terstruktur sehingga mendukung pembentukan model klasifikasi yang konsisten dan mudah direplikasi.

Meskipun demikian, hasil Confusion Matrix menunjukkan bahwa model masih menghasilkan jumlah False Negative yang relatif lebih tinggi dibandingkan False Positive, sehingga masih terdapat data nasabah yang seharusnya dikategorikan tidak layak tetapi diprediksi layak menerima kredit. Oleh karena itu, penelitian selanjutnya dapat mengembangkan model melalui optimasi nilai parameter K, penerapan teknik seleksi fitur, maupun perbandingan dengan algoritma klasifikasi lainnya untuk meningkatkan kemampuan model dalam mengenali kelas minoritas dan menghasilkan prediksi yang lebih akurat.

DAFTAR PUSTAKA

[1] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA, USA: Morgan Kaufmann, 2011.

[2] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From Data Mining to Knowledge Discovery in Databases," AI Magazine, vol. 17, no. 3, pp. 37–54, 1996.

[3] M. J. A. Berry and G. S. Linoff, Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management, 2nd ed. Indianapolis, IN, USA: Wiley Publishing, 2004.

[4] L. C. Thomas, Consumer Credit Models: Pricing, Profit and Portfolios. Oxford, U.K.: Oxford University Press, 2009.

[5] D. J. Hand and W. E. Henley, "Statistical Classification Methods in Consumer Credit Scoring: A Review," Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 160, no. 3, pp. 523–541, 1997.

[6] I. Brown and C. Mues, "An Experimental Comparison of Classification Algorithms for Imbalanced Credit Scoring Data Sets," Expert Systems with Applications, vol. 39, no. 3, pp. 3446–3453, 2012, doi: 10.1016/j.eswa.2011.09.033.

[7] F. Gorunescu, Data Mining: Concepts, Models and Techniques. Berlin, Germany: Springer, 2011.

[8] S. Lessmann, B. Baesens, H. V. Seow, and L. C. Thomas, "Benchmarking State-of-the-Art Classification Algorithms for Credit Scoring: An Update of Research," European Journal of Operational Research, vol. 247, no. 1, pp. 124–136, 2015, doi: 10.1016/j.ejor.2015.05.030.

[9] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.

[10] A. Prastyo et al., "Implementasi Algoritma Naïve Bayes pada Klasifikasi Data," Jurnal Teknologi Informasi, 2021.

[11] A. Santoso, Kusrini, and R. Hartanto, "Perbandingan Algoritma Random Forest dan K-Nearest Neighbor pada Prediksi Kelayakan Kredit," Jurnal Informatika, 2024.

[12] Harlina, "Penerapan K-Nearest Neighbor dengan Forward Selection untuk Prediksi Kelayakan Kredit," Jurnal Teknologi Informasi, 2018.

[13] M. Faqih, R. Prayoga, M. Dzakiir, and D. Lestari, "Klasifikasi Pengajuan Kartu Kredit Menggunakan Algoritma K-Nearest Neighbor (KNN)," Jurnal Informatika, 2024.

[14] M. T. Nawawi and A. Suhendar, "Implementation of MSME Credit Loan Determination Using Machine Learning Technology with KNN (K-Nearest Neighbors) Algorithm," JIKO (Jurnal Informatika dan Komputer), vol. 7, no. 3, pp. 217–221, 2024, doi: 10.33387/jiko.v7i3.9064.

[15] R. A. Sari and E. Simamora, "Penerapan Metode Bootstrap Aggregating K-Nearest Neighbor dengan Analisis Credit Scoring Berdasarkan Status Pembayaran Kredit Motor," EMASAINS: Jurnal Edukasi Matematika dan Sains, vol. 14, no. 2, 2024, doi: 10.59672/emasains.v14i2.5019.

[16] Kaggle, "Loan Prediction Dataset." [Online]. Available: https://www.kaggle.com/.

[17] RapidMiner, "RapidMiner Studio Documentation." [Online]. Available: https://docs.rapidminer.com/.

[18] N. S. Altman, "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," The American Statistician, vol. 46, no. 3, pp. 175–185, 1992.

[19] D. M. W. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness and Correlation," Journal of Machine Learning Technologies, vol. 2, no. 1, pp. 37–63, 2011.

[20] A. Zanuardi and H. Suprayitno, "Penerapan Tahapan Knowledge Discovery in Databases (KDD) pada Data Mining," Jurnal Teknologi Informasi, 2018.