BAB I
PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi digital telah mendorong inovasi di sektor keuangan, salah satunya melalui layanan financial technology (fintech) berbasis peer-to-peer lending (P2P lending). Layanan ini mempertemukan pemberi pinjaman dan peminjam secara langsung melalui platform daring tanpa perantara lembaga keuangan konvensional. P2P lending menawarkan kemudahan akses, proses cepat, serta inklusi keuangan bagi masyarakat yang belum terlayani oleh sistem perbankan formal.

Menurut data Otoritas Jasa Keuangan (OJK), outstanding pembiayaan melalui platform P2P *lending* di Indonesia telah mencapai Rp72,03 triliun per Agustus 2024. Sementara itu, penyaluran pembiayaan industri P2P *lending* pada Februari 2024 tercatat sebesar Rp61,09 triliun, meningkat sekitar 21,98% jika dibandingkan dengan periode yang sama pada tahun 2023 [1]. Pertumbuhan masif ini menunjukkan bahwa P2P lending berperan penting dalam memperluas akses pembiayaan.

Namun, pertumbuhan pesat ini juga diikuti oleh meningkatnya potensi risiko gagal bayar (*credit default*). Data OJK mencatat TWP90 (tunggakan lebih dari 90 hari) industri P2P *lending* berada pada level 2,38% per September 2024 [1]. Meskipun angka ini masih di bawah ambang batas toleransi OJK sebesar 5%, tren fluktuatif dan munculnya sejumlah kasus gagal bayar pada beberapa platform menunjukkan bahwa peningkatan volume pinjaman belum diiringi dengan sistem penilaian risiko kredit yang memadai. Untuk itu, dibutuhkan sistem penilaian risiko yang lebih akurat dan adaptif agar platform P2P lending tetap berkelanjutan.

Risiko gagal bayar pada platform P2P lending sangat bergantung pada sistem penilaian kredit yang digunakan dalam menentukan kelayakan calon peminjam. Sejumlah penelitian menunjukkan bahwa penilaian kredit

masih dilakukan dengan pendekatan konvensional umumnya berbasis *rule-based system* atau model statistik klasik seperti *logistic regression* yang telah lama digunakan dalam perbankan, di mana model penilaian masih memiliki keterbatasan karena mengasumsikan hubungan *linear* antarvariabel dan mengabaikan interaksi kompleks yang mungkin terjadi antar faktor risiko [2], [3].

Dalam konteks data calon peminjam, hubungan antara variabel seperti pendapatan, riwayat pinjaman, dan jumlah pinjaman tidak selalu bersifat linear atau independen, Sehingga membentuk pola hubungan yang kompleks dalam penentuan risiko kredit. Kondisi tersebut semakin diperumit ketika data menunjukkan ketidakseimbangan distribusi kelas (*imbalanced data*), di mana proporsi peminjam yang mengalami gagal bayar umumnya jauh lebih kecil dibandingkan peminjam dengan status lancar, yang merupakan karakteristik umum pada data risiko kredit [4].

Ketidakseimbangan ini menyebabkan model klasifikasi cenderung bias terhadap kelas mayoritas dan kurang efektif dalam mendeteksi peminjam berisiko tinggi, sehingga menurunkan kualitas evaluasi risiko kredit secara keseluruhan [4]. Oleh karena itu, diperlukan pendekatan prediksi yang lebih adaptif, mampu menangkap hubungan non-linear antarvariabel, serta lebih robust dalam menangani ketidakseimbangan distribusi kelas, agar akurasi evaluasi risiko kredit dapat ditingkatkan.

Upaya sejumlah penelitian terdahulu mengatasi ketidakseimbangan kelas umumnya dilakukan melalui teknik penyeimbangan data, seperti *undersampling* atau *oversampling* [5], [6]. Meskipun pendekatan tersebut dapat meningkatkan metrik evaluasi tertentu, pengubahan distribusi kelas berpotensi mengaburkan karakteristik risiko yang sebenarnya dihadapi platform P2P lending dan menghasilkan estimasi probabilitas yang tidak merepresentasikan kondisi operasional nyata.

Berbeda dengan pendekatan tersebut, penelitian ini mempertahankan distribusi kelas asli pada dataset Lending Club.

Ketidakseimbangan distribusi kelas dipandang sebagai karakteristik alami data risiko kredit yang mencerminkan kondisi operasional nyata pada layanan peer-to-peer lending. Dengan mempertahankan distribusi asli, penelitian ini bertujuan untuk mengevaluasi kemampuan model dalam mendeteksi kelas minoritas (*default*) secara lebih realistis, tanpa distorsi akibat modifikasi data.

Dalam konteks penelitian ini, dua algoritma machine learning yaitu *Naive Bayes* dan *Random Forest* digunakan untuk memprediksi risiko kredit calon nasabah. Model Naive Bayes yang digunakan merupakan varian *Gaussian Naive Bayes*, yang sesuai untuk menangani fitur numerik kontinu seperti pendapatan, rasio utang, dan jumlah pinjaman yang mendominasi dataset kredit. Naive Bayes dipilih karena sederhana, efisien secara komputasi, dan mampu memberikan hasil cepat pada dataset besar [7]. Model ini digunakan sebagai pembanding (*baseline model*) terhadap Random Forest, yang merupakan metode *ensemble learning* dengan kemampuan menangkap hubungan non-linear antar fitur serta menghasilkan prediksi yang lebih stabil.

Penelitian ini menganalisis dan membandingkan kinerja Gaussian Naive Bayes dan Random Forest dalam memprediksi risiko gagal bayar pada data P2P lending dengan distribusi kelas yang tidak seimbang, dengan fokus pada kemampuan masing-masing model dalam mengidentifikasi kelas minoritas secara realistis.

Pada penelitian menggunakan dataset publik Lending Club, dataset berisi data historis pinjaman yang mencakup informasi demografis, keuangan, serta riwayat pembayaran individu. Dataset ini digunakan sebagai pembuktian konsep (*proof of concept*) untuk menguji efektivitas algoritma *machine learning* dalam memprediksi risiko gagal bayar.

1.2 Rumusan Masalah

Berdasar latar belakang yang telah dijelaskan sebelumnya, maka rumusan

masalah dalam penelitian ini adalah sebagai berikut:

1. Bagaimana penerapan algoritma machine learning dalam memprediksi risiko kredit calon nasabah pada layanan peer-to-peer lending?
2. Bagaimana perbandingan kinerja Gaussian Naive Bayes dan Random Forest dalam memprediksi risiko gagal bayar pada dataset peer-to-peer lending dengan distribusi kelas yang tidak seimbang, khususnya dalam mendeteksi kelas minoritas (*default*)?

1.3 Batasan Masalah

Agar penelitian ini lebih terarah dan tidak meluas, maka ditetapkan beberapa batasan masalah sebagai berikut:

1. Penelitian ini menggunakan dataset publik dari platform Kaggle (Lending Club Dataset) yang berisi data demografi, keuangan, dan riwayat pinjaman nasabah.
2. Model pembelajaran mesin yang digunakan dalam penelitian ini adalah Gaussian Naive Bayes sebagai representasi model sederhana dan Random Forest sebagai representasi model kompleks.
3. Penelitian ini berfokus pada perbandingan kinerja model sederhana dan model kompleks dalam memprediksi risiko gagal bayar pada dataset kredit dengan distribusi kelas yang tidak seimbang.
4. Analisis performa model dilakukan berdasarkan metrik evaluasi Accuracy, Precision, Recall, F1-Score, dan ROC-AUC.
5. Penelitian ini berfokus pada aspek prediksi risiko gagal bayar (default) tanpa membahas faktor sosial, hukum, atau regulasi yang terkait dengan layanan P2P lending.
6. Proses implementasi terbatas pada pembuatan model prediktif dan tidak mencakup pengembangan sistem aplikasi penuh.

1.4 Tujuan Penelitian

Adapun tujuan yang ingin dicapai dalam penelitian ini adalah:

1. Menganalisis penerapan algoritma *machine learning* dalam memprediksi risiko kredit calon nasabah pada layanan *peer-to-peer lending*.
2. Menganalisis perbedaan kinerja antara model machine learning sederhana (Gaussian Naive Bayes) dan model kompleks (Random Forest) dalam memprediksi risiko gagal bayar pada data P2P lending yang tidak seimbang, khususnya dalam mendeteksi kelas *default*.

1.5 Manfaat Penelitian

1. Membantu pengembangan penerapan algoritma *machine learning* dalam klasifikasi risiko kredit pada layanan *peer-to-peer lending*.
2. Menambah referensi ilmiah mengenai penerapan algoritma *Naive Bayes* dan *Random Forest* dalam sistem penilaian kredit otomatis berbasis *machine learning*.

1.6 Sistematika Penulisan

Sistematika penulisan dalam laporan penelitian berdasarkan petunjuk penulisan yang berlaku di Universitas Amikon Yogyakarta sebagai berikut:

BAB I PENDAHULUAN

Berisi latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan sistematika penulisan.

BAB II TINJAUAN PUSTAKA

Berisi studi literatur dan dasar teori yang mendukung penelitian, termasuk konsep machine learning, algoritma Naive Bayes, Random Forest, serta penelitian terdahulu yang relevan.

BAB III METODE PENELITIAN

Menjelaskan objek penelitian, alur penelitian, serta alat dan bahan yang digunakan. Bab ini juga membahas tahapan seperti pengumpulan data, prapemrosesan, pembagian data latih dan uji, pelatihan model, serta metode evaluasi.

BAB IV HASIL DAN PEMBAHASAN

Menyajikan hasil eksperimen dari penerapan model machine learning, analisis perbandingan kinerja antar model, serta pembahasan mengenai hasil evaluasi model terhadap dataset.

BAB V PENUTUP

Berisi kesimpulan dari hasil penelitian dan saran untuk pengembangan penelitian lebih lanjut di masa mendatangrangkum selama proses penelitian,