BAB I
PENDAHULUAN

1.1. Latar Belakang

Risiko kredit adalah kegagalan peminjam dalam melunasi pinjaman atau bunga yang terkait sehingga menyebabkan suatu kredit bermasalah (*Non-Performing Loan*). Kredit bermasalah diartikan sebagai pinjaman yang telah jatuh tempo atau berada dalam keterlambatan pembayaran selama periode tertentu, biasanya 90 hari atau lebih. Dengan adanya kredit bermasalah ini membuat nilai *Non-Performing Loan* membesar yang berdampak pada operasional bisnis dan pendapatan bank atau lembaga keuangan. Salah satu keberhasilan operasional bisnis di bank atau lembaga keuangan dapat dilihat dari kemampuannya dalam memberikan kredit dengan meminimalisir risiko yang dihadapi [1].

Salah satu cara untuk mengurangi risiko kredit adalah dengan memanfaatkan pengetahuan dari data histori peminjaman yang sudah ada sebelumnya menggunakan teknik *data mining* [2]. *Data mining* secara luas dibagi menjadi dua kategori yaitu prediktif dan deskriptif [2]. Pada metode prediktif dapat dilakukan dengan pembuatan model klasifikasi. Klasifikasi sendiri merupakan proses pengelompokan objek yang mempunyai kesamaan karakteristik atau ciri kedalam beberapa kelas [3]. Untuk membuat model klasifikasi, diperlukan sebuah teknik *machine learning*.

Beberapa algoritma *machine learning* yang dapat digunakan untuk mengklasifikasikan risiko kredit diantaranya Naïve Bayes, SVM, *K-Nearest Neighbors*, dan *Decision Tree*. Naïve Bayes memiliki kelebihan dapat diimplementasikan secara cepat dan tidak membutuhkan banyak data pelatihan atau *training*, namun algoritma tersebut memiliki kekurangan saat datanya tidak saling independen dan mengabaikan hubungan antar fitur [4]. *Support Vector Machine* (SVM) memiliki kelebihan yang dapat diandalkan saat digunakan karena berfungsi dengan mengoptimalkan batas pemisah yang optimal (*hyperplane*) dan kuat akan data yang berdimensi tinggi [5]. Namun, algoritma ini kurang tepat diterapkan saat data yang digunakan berskala besar yang membutuhkan waktu *training* yang lama.

Selanjutnya terdapat algoritma *K-Nearest Neighbors* yang mudah diimplementasikan dengan menghitung jarak antar kelasnya, namun algoritma ini memiliki kekurangan yaitu sensitif terhadap data pencilan (*outliers*) [6]. Selain itu terdapat algoritma *Decision Tree* yang menggunakan konsep pohon dan mudah diimplementasikan serta divisualkan, namun algoritma ini memiliki kekurangan saat data yang digunakan berskala besar maka rentan terjadi *overfitting*.

Pemilihan algoritma *machine learning* harus disesuaikan dengan karakteristik data yang ada. Data kredit mencakup beberapa fitur seperti pendapatan, pekerjaan, jumlah pinjaman, suku bunga, kepemilikan rumah, dan faktor lainnya. Selain itu, data kredit juga memiliki berbagai jenis tipe data, termasuk kategorikal, diskrit, dan kontinu. Adanya keberagaman tipe fitur pada data, ukuran data yang besar, dan terdapat banyak data *outliers*, maka algoritma *Random Forest* digunakan sebagai metode prediksi yang sesuai untuk kasus ini karena keunggulannya dalam menangani kumpulan data kompleks dengan berbagai jenis fitur. *Random Forest* merupakan algoritma *ensemble* yang menggunakan kumpulan pohon keputusan acak untuk memperoleh prediksi yang akurat dan meminimalkan *overfitting* [7]. Selain kuat terhadap *overfitting*, *Random Forest* memiliki kinerja yang lebih baik dibandingkan dengan algoritma – algoritma lain seperti *Support Vector Machine* (SVM) dan *Discriminant Analysis* [8].

Algoritma *Random Forest* memiliki beberapa parameter yang mempengaruhi kinerjanya. Penggunaan pohon yang terlalu sedikit membuat model mengalami kekurangan, begitupun pohon yang terlalu banyak [9]. Penelitian sebelumnya pada penggunaan algoritma *Random Forest* belum menerapkan *hyperparameter tuning* sehingga kurang menjelajahi parameter – parameter terbaik yang dapat digunakan [10]. Tidak adanya analisis mengenai parameter – parameter terbaik pada model *Random Forest*, maka kualitas model yang dihasilkan kurang optimal dan mengurangi kemampuannya dalam mengidentifikasi risiko kredit secara akurat. Penelitian [11] menjelaskan pentingnya penggunaan algoritma optimasi seperti *Grid Search* dan *Random Search* dalam meningkatkan akurasi model.

Pada data yang tidak seimbang atau *imbalanced* perlu dilakukan *feature engineering* untuk memperbaiki kualitas data. Adanya kelas data yang tidak

seimbang dapat meningkatkan risiko *overfitting* dan representasi terbatas pada kelas minoritasnya. Penelitian [12] menunjukkan bahwa penerapan *feature engineering* khususnya *feature selection* dapat meningkatkan akurasi model *machine learning* pada klasifikasi risiko kredit. Oleh karena itu, penelitian ini bertujuan untuk mengatasi kekurangan – kekurangan tersebut dengan menerapkan *feature engineering* dan *hyperparameter tuning* yang lebih cermat guna menciptakan model klasifikasi yang lebih baik.

Pada penelitian ini akan menerapkan *feature engineering* untuk memperbaiki kualitas data yang digunakan dalam membuat model dengan melibatkan pemilihan fitur yang relevan, pengelolaan nilai yang hilang, dan penanganan *imbalanced* data yang diperlukan. Penelitian ini juga akan melakukan *tuning* parameter untuk mengoptimalkan performa algoritma *Random Forest* dengan melakukan analisis berbagai kombinasi parameter seperti jumlah pohon (*n_estimators*), kriteria (*criterion*), maksimal kedalaman pohon (*max_depth*), jumlah fitur yang dipertimbangkan *(max features*), jumlah sampel minimum untuk pembentukan node baru (*min_samples_split*), dan jumlah sampel minimum pada daun (*min_samples_leaf*).

Dengan mengintegrasikan *feature engineering* dan *hyperparameter tuning* secara komprehensif, penelitian ini diharapkan dapat menghasilkan model klasifikasi yang lebih akurat dalam mengidentifikasi risiko kredit.

1.2. Perumusan Masalah

Penerapan *Random Forest* pada klasifikasi risiko kredit memiliki akurasi yang rendah, sedangkan kondisi data (*credit risk dataset*) yang digunakan memiliki tipe data yang bervariasi, *imbalanced*, *outliers* dan *missing value* sehingga diperlukan *feature engineering* dan *hyperparameter tuning* untuk meningkatkan akurasi model.

1.3. Pertanyaan Penelitian

1. Apakah penggunaan teknik *feature engineering* dan *hyperparameter tuning* secara bersamaan dapat memberikan peningkatan signifikan terhadap akurasi model *Random Forest* dalam klasifikasi risiko kredit?
2. Jenis *feature engineering* dan *hyperparameter tuning* apa saja yang dapat

meningkatkan akurasi model *Random Forest*?

1.4. Tujuan Penelitian

1. Meningkatkan akurasi model *Random Forest* dalam klasifikasi risiko kredit dengan menggunakan *feature engineering* dan *hyperparameter tuning.*
2. Menentukan kombinasi optimal dari *feature engineering* dan *hyperparameter tuning* yang dapat menghasilkan model dengan akurasi yang lebih tinggi dalam klasifikasi risiko kredit.

1.5. Batasan Masalah

1. Data bersumber dari situs *kaggle* yang mempunyai kesamaan variabel dengan data pengajuan kredit di salah satu lembaga keuangan.
2. Pengambilan data risiko kredit di *kaggle* dilakukan pada 04 September 2023.
3. Model yang dibuat tidak akan diuji pada lembaga keuangan kredit karena kendala perizinan yang kompleks dan keterbatasan penerapan metode terbaru di lingkungan operasional bisnisnya.

1.6. Manfaat Penelitian

1. Mengurangi adanya *overfitting* yang terjadi pada data yang tidak seimbang dan mendapatkan parameter terbaik pada model *Random Forest.*
2. Memberikan rekomendasi penerapan *feature engineering* dan *hyperparameter tuning* pada data yang besar, *outliers,* dan *imbalanced*.