

BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Dalam beberapa tahun terakhir, integrasi teknik Statistik dan *Machine Learning* terhadap bidang keuangan semakin mendalam sehingga telah menyebabkan banyak perubahan terhadap permasalahan industri tersebut khususnya dalam hal pendugaan (Yuelin 2020). Model *Machine Learning* dan statistika berfondasi pada ilmu probabilitas yang merupakan bahasa matematika untuk mengukur derajat ketidakpastian, yang membuat pengembangan kedua teknik ini dijadikan sebagai alat untuk memprediksi peristiwa dimasa yang akan datang dengan menggunakan sejumlah data dari masa lampau. *Machine Learning* sendiri adalah cabang ilmu komputer yang memanfaatkan data masa lalu untuk dipelajari dan menggunakan pengetahuannya tersebut untuk membuat keputusan dimasa depan (Pratap 2017).

Ensemble Learning adalah teknik kombinasi algoritma *Machine Learning* yang bertujuan untuk memperoleh hasil akurasi prediksi yang lebih tinggi dibandingkan dengan algoritma *Machine Learning* yang digunakan secara tunggal (Oliver dan Ziad 2015). Selain meningkatkan akurasi dari sistem pengambilan keputusan, *Ensemble Learning* juga telah berhasil digunakan untuk mengatasi berbagai masalah *Machine Learning* seperti pemilihan fitur, estimasi kepercayaan, fitur yang hilang, pembelajaran tambahan, koreksi kesalahan, kelas dengan data yang tidak seimbang, dan lain-lain (Robi 2012). Dengan demikian *Ensemble Learning* telah terbukti sangat efektif dan sangat serbaguna dalam spektrum yang luas dari domain masalah dan aplikasi dunia nyata.

Klasifikasi yang merupakan salah satu permasalahan dalam *Ensemble Learning* diartikan oleh (Suganya dan Vijayarani 2017) sebagai teknik *data mining* untuk menganalisis data dan mengklasifikasikan data ke dalam sekumpulan kelas yang telah ditentukan sebelumnya. *Ensemble Learning* memiliki beberapa algoritma dalam masalah pengklasifikasian yang dapat digunakan dan disesuaikan dengan permasalahan yang dihadapi. Penelitian perbandingan algoritma yang

diimplementasikan pada klasifikasi keputusan kredit ini, akan menggunakan algoritma *XGBoost* dan *Random Forest*. Hal ini dikarenakan dalam penelitian yang dilakukan oleh (Yiheng dan Weidong 2020) yang berjudul "A comparative performance assessment of ensemble learning for credit scoring", *XGBoost* dan *Random Forest* telah diakui sebagai sebuah estimator canggih dengan kinerja yang sangat tinggi baik dalam klasifikasi maupun regresi. Dalam proses pengklasifikasian diharapkan data yang di-*input* adalah data yang seimbang karena data yang tidak seimbang dapat menyebabkan bias dan *overfitting* terhadap kelas mayoritas, bahkan memungkinkan kesulitan dalam pembuatan keputusan bisnis (Haibo dan Edwardo 2009). Meskipun demikian, algoritma *XGBoost* dan *Random Forest* mampu mencegah *overfitting* serta mampu menghasilkan prediksi yang relatif tinggi terhadap data yang hilang dan data yang tidak seimbang (Yiheng dan Weidong 2020).

Walaupun memiliki beberapa perbedaan seperti *XGBoost* mampu menangani data yang lebih besar dan kompleks dengan baik sedangkan *Random Forest* cenderung lambat, kemudian mengenai multikolinieritas *Random Forest* tidak sensitif sedangkan *XGBoost* bersifat sensitif. Namun, kedua algoritma ini mampu memberikan analisis keputusan kredit dengan akurasi yang tinggi seperti penelitian sebelumnya selain yang dilakukan oleh Yiheng Li dan Weidong Chen. Penelitian yang lainnya adalah milik Kui Wang, Meixuan Li, Jingyi Cheng, Xiaomeng Zhou dan Gang Li pada tahun 2021 tentang evaluasi risiko kredit pribadi berdasarkan *XGBoost* dengan 10.000 data kredit bank X serta membandingkan indeks evaluasi kinerja *XGBoost*, *Decision Tree* dan *K-Nearest Neighbor*. Penelitian tersebut menunjukkan *XGBoost* menghasilkan nilai akurasi yang tinggi kemudian disusul oleh *Decision Tree* dan terakhir oleh KNN (Kui 2021). Setelah itu, pada tahun 2022 *XGBoost* kembali diteliti oleh Sri Elina Herni Yulianti, Oni Soesanto dan Yuana Sukmawaty dengan judul Penerapan Metode *Extreme Gradient Boosting (XGBoost)* pada Klasifikasi Nasabah Kartu Kredit. Sebanyak 30.000 data nasabah dengan 24 variabel dan 2 kelas keputusan, *XGBoost* menghasilkan akurasi sebesar 80,039% sehingga dinyatakan cukup baik dalam klasifikasi nasabah kartu kredit (Sri 2022).

Algoritma *Random Forest* sendiri juga telah sering digunakan sebelumnya dalam berbagai masalah klasifikasi kredit diantaranya penelitian yang dilakukan oleh Lingxiao Tang, Fei Cai dan Yao Ouyang pada tahun 2018 dengan judul *Applying a nonparametric random forest algorithm to assess the credit risk of the energy industry in China*. Penelitian ini mengukur risiko kredit secara

ilmiah dari kartu kredit yang digunakan dalam industri energi China berdasarkan analisis beberapa faktor yang memengaruhi risiko kredit tersebut. Penerapan algoritma *Random Forest* dalam penelitian tersebut dengan data bulanan kartu kredit pelanggan industri energi di cabang Postal Savings Bank of China dari bulan April 2014 hingga Juni 2017, menghasilkan akurasi prediksi sebesar 91,5% dan juga stabilitasnya memuaskan (Lingxiao 2018). Penelitian selanjutnya pada tahun 2020 yang dilakukan oleh Yuelin Wang, Yihan Zhang, Yan Lu dan Xinran Yu, *Random Forest* menghasilkan nilai akurasi sebesar 94,57% dibandingkan algoritma KNN, *Decision Tree*, *Nave Bayes* dan Regresi Logistik (Yuelin 2020). Akan tetapi, pada penelitian yang dilakukan oleh (Lkhagvadorj 2019) yang berjudul "An Empirical Comparison of Machine-Learning Methods on Bank Client Credit Assessments", algoritma *XGBoost* menghasilkan nilai akurasi yang lebih tinggi dibandingkan algoritma *Random Forest*.

Algoritma *XGBoost* adalah algoritma *Ensemble Learning* yang dikembangkan pada tahun 2014 oleh Tianqi Chen dan merupakan model yang terintegrasi *Gradient Lifting* berbasis pohon dengan efisiensi dan akurasi prediksi yang tinggi (Kui 2021). *XGBoost* terdiri dari beberapa pohon keputusan yang kemudian akurasinya ditingkatkan oleh *XGBoost* dengan mempelajari kesalahannya dan melatih kembali pengamatan yang diprediksi salah dalam fase pelatihan. Sedangkan *Random Forest* adalah pengembangan dari *Decision Tree*, dimana sampel diambil dari *data training* secara *Bootstrapping* (pengambilan sampel secara berulang dengan penggantian) (Pratap 2017).

Mengenai keputusan kredit, tentunya kehidupan manusia tidak pernah lepas dalam memberikan sebuah keputusan mulai dari permasalahan kecil hingga permasalahan yang lebih besar. Demikian halnya dalam dunia keuangan dan perbankan yang selalu dihadapkan pada pilihan untuk menerima permohonan kredit nasabah atau justru sebaliknya menolak permohonan tersebut. Pengertian bank dalam undang-undang Nomor 10 Tahun 1998 tentang Perbankan, adalah salah satu badan usaha yang menghimpun dana dari masyarakat dalam bentuk simpanan dan menyalurkannya kepada masyarakat dalam bentuk pinjaman atau kredit.

Kata kredit dalam artikel (Harun 2022) dinyatakan berasal dari kata *Credere* yang berarti percaya (*to trust*), maknanya bank memiliki kepercayaan terhadap nasabah untuk menggunakan kredit sebaik mungkin. Dengan demikian pengambilan keputusan kredit sederhananya dapat didefinisikan sebagai pemeriksaan,

penelitian dan analisis terhadap kelengkapan serta kelayakan berkas/surat/data-data permohonan kredit calon debitur guna menghindari timbulnya risiko kredit saat setelah pihak bank memberikan kepercayaan kepada pihak debitur. Namun, pada praktiknya, kredit yang diberikan kepada debitur selalu memiliki sebuah risiko seperti kredit macet (Nisha dan Pankaj 2020). Kredit macet dapat terjadi salah satunya dikarenakan kekurangmampuan pihak bank dalam menilai risiko calon debitur secara tepat. Selain pada dasarnya dalam klasifikasi keputusan kredit jumlah kelas pengguna (debitur) yang 'berisiko' dan pengguna (debitur) yang 'tidak berisiko' sering tidak seimbang, kekurangmampuan pihak bank dalam menilai risiko calon debitur juga dipengaruhi oleh faktor-faktor lainnya seperti data yang tidak lengkap dan data yang salah atau tidak akurat. Adapun juga kendala lainnya terkait algoritma yang digunakan dalam mengklasifikasi keputusan kredit yaitu terjadinya *overfitting* pada model klasifikasi dan penggunaan evaluasi model yang tidak tepat. Untuk itu, kreditor dalam hal ini adalah pihak bank dituntut harus mampu mengevaluasi permohonan kredit secara objektif dan lebih akurat lagi agar tidak salah dalam memberikan keputusan pada setiap permohonan kredit nasabah. Selain itu, sebagian besar kekayaan bank adalah dalam bentuk kredit yang merupakan sumber pendapatan bank (Lean 2018) sehingga bila kredit macet tidak dihindari dengan baik, bank akan mengalami kerugian bahkan mungkin kebangkrutan seperti penelitian yang dilakukan oleh bank dunia pada tahun 1992.

Berdasarkan uraian latar belakang masalah dan alasan pemilihan algoritma diatas, penulis akan melakukan penelitian yang berjudul "Analisis Perbandingan Algoritma *XGBoost* dan *Random Forest Ensemble Learning* pada Klasifikasi Keputusan Kredit" untuk membandingkan dua algoritma tersebut agar pihak bank dapat mengevaluasi permohonan kredit secara objektif dan lebih akurat lagi.

1.2 Identifikasi Masalah

Berdasarkan uraian latar belakang diatas, dapat disimpulkan identifikasi masalah penelitian ini adalah sebagai berikut:

1. Dunia keuangan dan perbankan yang sering dihadapkan pada pilihan untuk menerima atau menolak permohonan kredit oleh calon debitur dan risiko seperti kredit macet yang dikarenakan kekurangmampuan pihak bank dalam menilai risiko calon debitur secara akurat.

2. Data permohonan kredit bank umumnya seringkali tidak lengkap, kurang akurat serta tidak sesuai dengan kebutuhan algoritma sehingga perlu penanganan seperti *cleaning data*, *label encoding* dan penskalaan data.
3. Sebagian besar algoritma klasifikasi membutuhkan jumlah kelas yang seimbang untuk mendapatkan kinerja yang baik. Namun, pada klasifikasi keputusan kredit, jumlah kelas pengguna yang 'berisiko' dan pengguna yang 'tidak berisiko' sering tidak seimbang. Hal ini dapat menyebabkan model menghasilkan klasifikasi yang tidak seimbang dan tidak akurat.
4. *Overfitting* atau *underfitting* dapat terjadi pada model klasifikasi keputusan kredit jika model terlalu rumit atau terlalu sederhana. Hal ini dapat menyebabkan model tidak dapat menggeneralisasi data dengan baik dan menghasilkan prediksi yang tidak akurat.
5. Adanya kesalahan penilaian atau evaluasi pada model yang terjadi karena penggunaan metode evaluasi yang tidak tepat untuk mengukur kinerja model. Sehingga dapat menyebabkan kesalahan dalam pengambilan keputusan terkait kredit, yang dapat berdampak pada pelanggan dan perusahaan (bank).
6. Terkait tuntutan kepada pihak bank agar dapat mengevaluasi permohonan kredit secara objektif dan lebih akurat, perlu ditentukan algoritma mana yang lebih baik dan akurat antara *XGBoost* dan *Random Forest* dalam mengklasifikasikan keputusan kredit bank.

1.3 Ruang Lingkup

Penelitian ini terfokus pada:

1. Penerapan dua algoritma *Ensemble Learning* yakni *XGBoost* dan *Random Forest* pada dataset klasifikasi kartu kredit bank.
2. Analisis perbandingan algoritma *XGBoost* dan *Random Forest* pada klasifikasi keputusan kredit dengan melihat hasil kinerja setiap algoritma berdasarkan nilai metrik evaluasi model yang diperoleh dalam penelitian ini.

1.4 Batasan Masalah

Mengacu pada ruang lingkup yang telah ditetapkan, maka yang menjadi batasan masalah dalam penelitian ini adalah:

1. Data yang digunakan adalah data sekunder yakni dataset kartu kredit bank yang berjudul "*Credit Card Approval- With Target*" yang diperoleh dari www.kaggle.com merupakan platform kompetisi *data science* dan sumber data terbuka yang menyediakan berbagai dataset yang dapat diakses oleh publik.
2. Algoritma yang digunakan adalah algoritma *XGBoost* dan *Random Forest* yang merupakan dua algoritma terbaik dari metode *Ensemble Learning*.
3. Penggunaan *cleaning data*, *label encoding*, dan penskalaan data untuk mengatasi data yang tidak lengkap, kurang akurat, dan tidak sesuai dengan kebutuhan algoritma.
4. Komputasi yang digunakan untuk penelitian ini adalah *software Python* yaitu bahasa pemrograman yang memiliki *library* yang menyediakan berbagai algoritma.

1.5 Rumusan Masalah

Masalah dalam penelitian ini dapat dirumuskan sebagai berikut:

1. Bagaimana hasil kinerja algoritma *XGBoost* dan *Random Forest* pada klasifikasi keputusan kredit berdasarkan nilai metrik evaluasi model yang diperoleh dalam penelitian ini?
2. Apakah data dengan kelas yang tidak seimbang dan ukuran data memiliki pengaruh pada kinerja setiap algoritma dalam penelitian ini?
3. Manakah algoritma diantara *XGBoost* dan *Random Forest* yang memiliki performa model yang paling baik dalam pemberian keputusan kredit?

1.6 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, maka yang menjadi tujuan penelitian ini adalah:

1. Mengetahui hasil kinerja algoritma *XGBoost* dan *Random Forest* pada klasifikasi keputusan kredit berdasarkan nilai metrik evaluasi model yang diperoleh dalam penelitian ini.
2. Mengetahui pengaruh data dengan kelas yang tidak seimbang dan ukuran data pada kinerja setiap algoritma dalam penelitian ini.
3. Mengetahui algoritma mana diantara *XGBoost* dan *Random Forest* yang memiliki performa model yang paling baik dalam mengklasifikasikan debitur 'berisiko' dan 'tidak berisiko' guna pemberian keputusan kredit.

1.7 Manfaat Penelitian

Selain dapat menambah wawasan penulis, penelitian ini juga dapat digunakan sebagai referensi bagi pihak kreditur (bank) dalam mengklasifikasikan debitur yang 'berisiko' dan tidak 'berisiko'. Dengan demikian, pihak kreditur (bank) dapat mengevaluasi permohonan kredit secara cepat, objektif dan lebih akurat dalam rangka menghindari kesalahan pada saat memberikan keputusan kredit yang mengakibatkan adanya kredit macet.