BAB I

PENDAHULUAN

1.1 Latar Belakang

Bank adalah salah satu bagian penting pada roda perekonomian. Berbagai layanan keuangan biasanya disediakan oleh bank, seperti pengiriman dan penerimaan uang, penyimpanan uang, deposito dan sebagainya. Untuk dapat mempromosikan layanan tersebut maka bank biasanya akan menerapkan suatu teknik pemasaran untuk mendapatkan nasabah. Salah satu teknik yang digunakan adalah *telemarketing*.

Telemarketing merupakan suatu teknik yang diterapkan dengan cara menghubungi calon pelanggan atau nasabah melalui telephone. Perusahaan menggunakan *Telemarketing* dengan tujuan untuk mencari peluang dan calon klien bagi produk atau layanan tertentu, dimana dengan menggunakan komunikasi jarak jauh, perusahaan atau organisasi akan memperoleh keuntungan lebih melalui pemasaran yang efektif. Untuk dapat mengetahui keberhasilan dari telemarketing dapat digunakan suatu teknik *machine learning*. *Machine Learning* (ML) merupakan komponen dari kecerdasan buatan (AI) yang memberikan sistem kemampuan untuk belajar dan meningkatkan kinerjanya secara otomatis dari pengalaman, tanpa perlu diprogram secara eksplisit (Tahyudin, 2020). *Datasets* yang dapat digunakan sebagai data latih adalah *datasets* Bank Marketing yang diambil dari *website* UCI Machine Learning yaitu, *dataset* kampanye pemasaran langsung (*telemarketing*) yang berasal dari lembaga perbankan Portugis. Namun,

terdapat ketidakseimbangan kelas atau *class imbalance* pada dataset tersebut, dimana terdapat kelas yang jauh lebih dominan secara signifikan dari kelas lainnya.

Ketidakseimbangan kelas atau *class imbalanced* yang terjadi pada data tersebut adalah jumlah kelas yang tidak berlangganan deposito jauh lebih banyak dibandingkan dengan kelas yang berlanggan deposito. Kelas yang memiliki distribusi persentase lebih kecil akan disebut dengan kelas minoritas sebaliknya kelas yang memiliki persentase lebih besar akan disebut dengan kelas mayoritas (Ustyannie & Suprapto, 2020). Ketidakseimbangan data antara kelas akan mengakibatkan ketergantungan pada kelas mayoritas dalam proses pengklasifikasian (Mutmainah, 2021). Ketergantungsn itu akan menyebabkan model *machine learning* baik dalam meprediksi kelas mayoritas tetapi sulit untuk memprediksi kelas minoritas (Kaope & Pristyanto, 2023; Wongvorachan dkk., 2023). Selain itu, *class imbalanced* juga dapat menyebabkan terjadi *overfitting* pada model.

Berdasarkan permasalahan itu, maka ketidakseimbangan kelas memerlukan suatu penanganan sehingga model yang dihasilkan dapat bekerja dengan baik dan lebih andal dalam melakukan prediksi terhadap data yang diberikan. Teknik yang dapat digunakan untuk menanganani masalah tersebut diantarnya adalah teknik *oversampling*, *undersampling* dan kombinasi dari kedua teknik tersebut. Teknik ini dilakukan dengan menaikan jumlah sampel pada kelas minoritas (*oversampling*) dan menurunkan jumlah sampel pada kelas mayoritas (*undersampling*). Melalui cara ini, jumlah sampel pada kedua kelas menjadi lebih seimbang dan model dapat mempelajari pola dari kedua kelas secara merata dan akurat.

Penerapan metode *oversampling* yang terlalu banyak dapat menyebabkan terjadi *overfitting* (Kaur & Gosain, 2018) sedangkan penerapan *undersampling* yang berlebihan dapat menyebabkan hilangnya informasi penting dari *datasets* (Guo dkk., 2018; Guzmán-Ponce dkk., 2020). *Synthetic Minority Oversampling Technique* (SMOTE) adalah salah satu teknik *oversampling* yang dapat mengurangi *overfitting* (Cai dkk., 2019) jika diterapkan karena dalam melakukan *oversampling* SMOTE akan membangkitkan data sintetis. Hal itu dapat dilihat dari penelitian oleh (Suparyati dkk., 2022) dimana semua matriks performa menunjukan *score* 1-2% lebih tinggi ketika menggunakan SMOTE saat memprediksi *lumpy skin disease*. Namun, SMOTE memiliki suatu kelemahan dimana data sintetis yang dihasilkan dapat menyebabkan terjadinya *class overlapping* atau tumpang tindih kelas berdasarkan penelitian (Wijayanti dkk., 2021), dimana *class overlapping* juga dapat memengaruhi performa dari model yang dihasilkan.

Oleh karena itu, demi menangani permasalahan itu maka peneliti mengusulkan untuk mengkombinasikan teknik SMOTE dengan teknik *undersampling*. SMOTE akan melakukan *oversampling* terhadap kelas minoritas dan kemudian teknik undersampling akan melakukan *undersampling* terhadap kelas mayoritas. Penerapan teknik *undersampling* ini diharapkan mampu membuat data yang dihasilkan lebih bersih dan terhindar dari data *noise* dan tumpang tindih kelas atau *class overlapping* yang diakibatkan oleh data sintetis yang dihasilkan dari metode SMOTE. Teknik *undersampling* yang akan coba dikombinasikan dengan teknik SMOTE pada penelitian ini diantaranya, *Edited Nearest Neighbor* (ENN), *Neighborhood Cleaning Rule* (NCL) dan *TomekLink*. Lalu, untuk mengukur

performansi dari model yang dilatih nantinya akan digunakan *confusion matrix* sehingga dapat diketahui *sensitivity*, *specificity*, dan *G-means* dari model yang dilatih. Selain itu, akan digunakan digunakan 2 dataset berbeda untuk menguji kembali metode yang digunakan dalam penelitian ini yaitu, dataset *credit card fraud* dan dataset *cerebral stroke* yang diambil dari website Kaggle.

Diharapkan dengan adanya penelitian ini dapat diketahui perbandingan performa dari model yang dilatih menggunakan data tanpa *resampling*, di-*resampling* dengan SMOTE serta yang dilatih dengan data *resampling* kombinasi yaitu dengan SMOTE-ENN, SMOTE-NCL, dan SMOTE-TomekLink. Selain itu, peneliti memiliki harapan agar penelitian ini dapat memberikan informasi yang berguna dalam pengembangan model klasifikasi pada *datasets bank marketing* yang mengalami ketidakseimbangan kelas serta dapat menjadi referensi bagi peneliti selanjutnya yang tertarik untuk mengangkat topik serupa.

1.2 Identifikasi Masalah

Berdasarkan latar belakang yang telah dipaparkan sebelumnya, adapun beberapa permasalahan yang dapat diidentifikasi adalah sebagai berikut:

1. Ketidakseimbangan kelas pada *datasets bank marketing* dapat mempengaruhi performa model *machine learning* yang dihasilkan.
2. Terdapat potensi *class overlapping* atau tumpang tindih kelas akibat dari data sintetik yang dihasilkan melalui penerapan metode SMOTE.
3. Belum diketahui peningkatan dan perbandingan performansi antara kombinasi SMOTE dengan teknik undersampling ENN, NCL, dan Tomek Link dalam

menangani *imbalanced class* ataupun potensi *class overlapping* yang bisa terjadi akibat penerapan SMOTE.

1.3 Batasan Masalah

Adapun batas permasalahan dalam penelitian yang peneliti lakukan adalah sebagai berikut.

1. Aplikasi yang akan digunakan untuk melakukan pengolahan data adalah Jupyter Notebook dengan bahasa pemrograman Python.
2. Metode *oversampling* yang digunakan adalah metode SMOTE.
3. Metode *undersampling* yang akan digunakan diantaranya, ENN, NCL dan Tomek Link.
4. *Classifier* yang digunakan adalah *Logistic Regression*.
5. Dataset yang digunakan adalah *bank marketing* dataset yang berasal dari *website* UCI Machine Learning serta 2 dataset yang diambil dari Kaggle yaitu *credit card fraud datasets* dan *cerebral stroke datasets*.
6. Performansi dari model akan diukur menggunakan *sensitivity, specificity,* dan *g-means*.

1.4 Rumusan Masalah

Adapun rumusan masalah berdasarkan permasalahan yang sudah diidentifikasi sebelumnya, antara lain:

1. Bagaimana proses untuk menyeimbangkan distribusi kelas pada *bank marketing datasets*?

2. Bagaimana proses untuk menangani potensi *class overlapping* yang dapat terjadi akibat penerapan metode *oversampling* SMOTE pada *bank marketing datasets*?
3. Bagaimana peningkatan dan perbandingan performansi kombinasi SMOTE dengan beberapa teknik undersampling dalam menangani *imbalanced class* ataupun potensi *class overlapping* yang bisa terjadi akibat penerapan SMOTE?

1.5 Tujuan Penelitian

Berdasarkan rumusan masalah yang diangkat, maka tujuan dari penelitian ini adalah sebagai berikut:

1. Untuk menyeimbangkan distribusi kelas pada *bank marketing datasets.*
2. Untuk menangani potensi *class overlapping* yang dapat terjadi akibat penerapan metode SMOTE pada *bank marketing dataset*.
3. Untuk mengetahui peningkatan dan perbandingan performansi kombinasi SMOTE dengan beberapa teknik undersampling dalam menangani *imbalanced class* ataupun potensi *class overlapping* yang bisa terjadi akibat penerapan SMOTE.

1.6 Manfaat Penelitian

1.6.1 Manfaat Teoritik

Penelitian ini dimaksudkan untuk dapat membantu dalam pengembangan ilmu tentang proses penerapan kombinasi metode *oversampling* dalam menangani ketidakseimbangan kelas ataupun potensi *class overlapping* yang diakibatkan

metode SMOTE dalam melakukan klasifikasi apakah seseorang bersedia untuk berlangganan deposito berjangka atau tidak, sehingga diharapkan dapat dihasilkan performa model yang lebih baik.

1.6.2 Manfaat Praktis

Peneliti juga berharap dengan penelitian ini manfaat praktis bisa dicapai, diantarnya:

1. Model *machine learning* yang dihasilkan dapat digunakan untuk mengembangkan aplikasi berbasiskan *machine learning* yang dapat digunakan untuk memprediksi seseorang yang berpotensi untuk berlangganan deposito berjangka pada bank.
2. Dapat membantu untuk memilih metode terbaik diantara SMOTE dengan kombinasi antara SMOTE dengan metode *undersampling*, yaitu SMOTE-ENN, SMOTE-NCL, serta SMOTE-TomekLink dalam menangani *class imbalanced* ataupun potensi *class overlapping* yang mungkin terjadi akibat penerapan SMOTE.