BAB I
PENDAHULUAN

1.1 Latar Belakang

Tantangan dalam membuat model klasifikasi menggunakan *machine learning* adalah menemukan model yang akurat. Terdapat beberapa hambatan yang dihadapi dalam membuat model klasifikasi seperti masalah pada kondisi data yang tidak seimbang. Data dikatakan tidak seimbang jika distribusi dari setiap kelas tidak sama. Ketidakseimbangan distribusi data antara kelas merupakan fenomena yang umum terjadi dalam konteks data [1]. Ketidakseimbangan data antara kelas mayoritas dan minoritas dapat menyebabkan kesalahan dalam proses klasifikasi. Ukuran data yang tidak seimbang dapat membuat model klasifikasi cenderung mengandalkan kelas mayoritas dibandingkan kelas minoritasnya [2].

Penanganan masalah pada dataset yang tidak seimbang memerlukan suatu pendekatan khusus pada data, seperti *oversampling*, *undersampling*, dan gabungan keduanya. Pendekatan *oversampling* seringkali lebih umum digunakan dalam penanganan dataset yang tidak seimbang [3]. Namun, pendekatan *undersampling* berpotensi menghapus data yang pada akhirnya bisa menyebabkan kekurangan data, yang menyebabkan meningkatkan risiko kehilangan informasi penting dalam data [4]. Secara umum, metode *oversampling* memberikan hasil yang lebih baik daripada metode *undersampling*. Terdapat beberapa metode *oversampling*, diantaranya *Adaptive Synthetic Sampling* (*ADASYN*), *Random Oversampling* (*ROS*), dan *Synthetic Minority Over Sampling Technique* (*SMOTE*) [3].

ADASYN merupakan suatu metode yang menggunakan bobot distribusi untuk menghasilkan data sintesis pada kelas minoritas [4]. Kelemahan *ADASYN* adalah dapat menghasilkan data sintetis yang tidak realistis karena perhitungannya menggunakan sampel dari kelas minoritas yang dekat dengan kelas mayoritas [5]. Metode *oversampling* lainya yaitu *Random Oversampling* (*ROS*) yang merupakan teknik menambahkan data ke kelas minoritas secara acak tanpa menambahkan variasi pada data kelas tersebut [6]. Kelemahan metode ini terletak pada produksi duplikasi data yang berlebihan, yang pada akhirnya dapat menyebabkan permasalahan *overfitting* [7].

Metode *oversampling* lainya yaitu *Random Oversampling* (*ROS*) yang merupakan teknik menambahkan data ke kelas minoritas secara acak tanpa menambahkan variasi pada data kelas tersebut [6]. Kelemahan metode ini terletak pada produksi duplikasi data yang berlebihan, yang pada akhirnya dapat menyebabkan permasalahan *overfitting* [7]. Sementara itu, SMOTE dapat menghasilkan pengembangan wilayah keputusan yang lebih luas dan kurang spesifik pada kelas minoritas. Metode ini menciptakan sampel yang lebih terkait dengan kelas minoritas, memungkinkan pengklasifikasi untuk memiliki cakupan yang lebih besar dalam memahami kelas minoritas tersebut. Oleh karena itu, pendekatan *SMOTE* dapat meningkatkan akurasi pengklasifikasi pada kelas minoritas jika dibandingkan dengan metode *oversampling* yang menggunakan duplikasi [5].

Penelitian [8] menggunakan SMOTE dalam mengatasi *imbalanced dataset* dan berpengaruh baik pada hasil akurasi *Fuzzy C-Means*. Penelitian [9] mengenai *SMOTE* menunjukkan performa yang lebih baik dibandingkan dengan Adaptive Synthetic Sampling (*ADASYN*) dalam konteks deteksi penipuan kartu kredit. Pada penelitian ini, akan dilakukan pengujian kinerja SMOTE dalam mengatasi masalah *oversampling* pada beberapa data yang tidak seimbang, menggunakan algoritma *machine learning* seperti *Random Forest* dan *C4.5*. *Random Forest* merupakan algoritma *ensemble learning* yang berdasarkan konsep *Decision Tree*. *Ensemble learning* adalah teknik di mana beberapa model digabungkan untuk membentuk model yang lebih kuat dan handal. Sementara itu, *C4.5* adalah algoritma yang digunakan untuk membentuk pohon keputusan dengan cara pemilihan fitur terbaik.

1.2 Rumusan Masalah

Permasalahan ketidakseimbangan pada dataset mengakibatkan model klasifikasi cenderung bergantung pada kelas mayoritas dan berdampak pada kinerja akurasi dari algoritma yang digunakan sehingga perlu adanya solusi yang tepat untuk menangani masalah tersebut.

1.3 Pertanyaan Penelitian

Pertanyaan pada penelitian ini adalah sebagai berikut berikut:

1. Apa pengaruh penerapan SMOTE pada hasil akurasi kinerja model pada data yang tidak seimbang?

2. Berapa perbandingan akurasi Random Forest dan C4.5 pada data yang dilakukan oversampling dengan SMOTE dan tidak dikenakan dengan SMOTE?

1.4 Batasan Masalah

Batasan masalah yang ada pada penelitian ini adalah sebagai berikut:

1. Penelitian ini memfokuskan pada masalah ketidakseimbangan distribusi data antara kelas dalam konteks data.
2. Penelitian ini tidak mempertimbangkan jenis data yang spesifik, tetapi lebih berfokus pada masalah ketidakseimbangan data secara umum.
3. Variabel yang diteliti adalah efektivitas antara algoritma *Random Forest* dan C4.5 dalam menangani masalah ketidakseimbangan data.

1.5 Tujuan Penelitian

Tujuan dibuatnya penelitian ini adalah sebagai berikut :

1. Menganalisis pengaruh penerapan SMOTE pada hasil akurasi kinerja model pada data yang tidak seimbang
2. Menganalisis performa akurasi Random Forest dan C4.5 pada data yang dilakukan *oversampling* dengan SMOTE dan tidak dikenakan *oversampiling* dengan SMOTE.

1.6 Manfaat Penelitian

Manfaat dari penelitian ini berdasarkan tujuan penelitian yang telah ditetapkan adalah:

1. Memberikan solusi untuk masalah ketidakseimbangan data.
2. Meningkatkan pemahaman tentang perbedaan kinerja antara Random Forest dan C4.5 pada data yang tidak seimbang.
3. Memberikan panduan dalam pemilihan algoritma yang sesuai untuk meningkatkan hasil klasifikasi pada data yang tidak seimbang.