BAB I
PENDAHULUAN

1.1 Latar Belakang

Klasifikasi merupakan pengelompokan data kedalam kelas-kelas yang sebelumnya sudah didefinisikan. Pada klasifikasi sering ditemukan dataset dalam kondisi tidak seimbang. Kondisi dataset yang tidak seimbang merupakan kondisi dimana sebuah dataset memiliki kelas mayoritas dan kelas minoritas.

Kelas mayoritas merupakan kelas yang jumlah datanya jauh lebih besar dibandingkan kelas minoritas[1]. Perbandingan antara kedua kelas disebut dengan *Imbalance Ratio* (IR). Nilai Imbalance Ratio yang semakin besar menandakan perbedaan antar kelas yang semakin besar[2].

Dataset yang tidak seimbang akan mempengaruhi pada hasil klasfikasi. Ketidakseimbangan ini, menjadikan kelas yang seharusnya diklasifikasikan kedalam kelas minoritas tetapi diklasifikasikan kedalam kelas mayoritas. Kesalahan pada klasifikasi akan mempengaruhi kinerja algoritma klasifikasi[3].

Kesalahan klasifikasi perlu ditangani agar terhindar dari dampak yang ditimbulkan. Dataset yang tidak seimbang dapat ditangani dengan 3 pendekatan, yaitu pendekatan pada tingkat data, tingkat algoritma, dan ensemble learning. Pendekatan tingkat data akan menyeimbangkan distribusi data setiap kelas. Pada tingkat algoritma, algoritma akan dimodifikasi untuk mengatasi data yang tidak seimbang. Pada ensemble learning, ketidakseimbangan data diatasi menggunakan penggabungan tingkat data dan algoritma [4].

Pendekatan tingkat data dapat menggunakan teknik *undersampling* , *oversampling* dan kombinasi teknik *undersampling* dan *oversampling*. Teknik *undersampling* akan menghapus data kelas mayoritas hingga seimbang dengan kelas minoritas sedangkan *oversampling* akan menambahkan data kedalam kelas minoritas.

Teknik oversampling banyak digunakan daripada teknik *undersampling*. Hal ini dikarenakan *undersampling* menghilangkan sebagian data sehingga informasi terpenting didalamnya ikut hilang sehingga mempengaruhi kinerja model. Selain itu, menurut penelitian[5] *oversampling* cocok digunakan untuk data berukuran kecil.

Teknik oversampling memiliki bannyak algoritma, ROS dan SMOTE adalah dua dari beberapa algoritma yang ada. Seiring berjalanya waktu, SMOTE telah memiliki berbagai varian yang mengatasi kekurangan SMOTE. Adaptive Synthetic (ADASYN) dan Borderline-SMOTE adalah algoritma dari varian SMOTE. ADASYN menggunakan distribusi kepadatan kelas minoritas untuk menentukan data sintetis yang dihasilkan[6],[7]. Berbeda dengan ADASYN, Borderline-SMOTE menggunakan kelas minoritas di daerah perbatasan untuk menghasilkan data sintetis[8].

Penelitian[9] ADASYN mampu meningkatkan kemampuan algoritma Naïve Bayes, Decision Tree, dan Artificial Neural Network secara signifikan. Hasil evaluasi menunjukan bahwa hampir semua kelas hipertensi dapat diklasifikasikan dengan tepat. Nilai akurasi, precission, dan recall terbaik didapat oleh algoritma ANN dengan akurasi 91%, precission 86% dan recall 99%.

Penelitian[10] memprediksi customer churn menggunakan Borderline-SMOTE untuk menyeimbangkan data. Random forest, Naïve Bayes dan K-Nearest Neighbors digunakan sebagai algoritma classifiernya dan hasil ketiganya dibandingkan. Penelitian tersebut menunjukan bahwa Borderline-SMOTE meningkatkan kinerja Random Forest sebesar 4% dan lebih unggul dari model classifier yang lain.

Berdasarkan penelitian[9], [8] yang menunjukan bahwa ADASYN dan Borderline-SMOTE dapat menangani ketidakseimbangan dataset, maka penelitian ini menggunakan ADASYN dan Borderline-SMOTE untuk menyeimbangkan beberapa dataset.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dijabarkan, maka rumusan masalah dalam penelitian ini adalah sebagai berikut:

1. Apakah teknik balancing bisa meningkatkan kinerja pada Random Forest?
2. Algoritma *oversampling* apa yang menghasilkan kinerja terbaik?

1.3 Batasan Masalah

Agar penelitian tidak meluas dan sesuai tujuan penelitian, maka diterapkan batasan masalah sebagai berikut :

1. Teknik *oversampling* yang digunakan yaitu *Adaptive Synthetic* (ADASYN) dan Borderline-SMOTE.
2. Menggunakan Random Forest untuk klasifikasi dataset.
3. Evaluasi kinerja menggunakan nilai AUC-ROC curve.
4. Dataset yang digunakan dalam penelitian merupakan data binary yang legal digunakan untuk umum dan bersumber dari https://www.kaggle.com/.

1.4 Tujuan Penelitian

Adapun tujuan penelitian yang ingin dicapai sebagai berikut :

1. Mengetahui kinerja Random Forest sebelum dan setelah balancing.
2. Mencari algoritma terbaik dari eksperimen yang telah disusun.

1.5 Manfaat Penelitian

Hasil penelitian ini diharapkan mampu memberikan referensi bagi peneliti berikutnya yang sedang meneliti ketidakseimbangan dataset sehingga penelitian berikutnya dapat jauh lebih baik.

1.6 Sistematika Penulisan

Penulisan skripsi ini tersusun menjadi lima bab dan ditulis secara sistematis. Sistematika penulisan skripsi sebagai berikut :

BAB I PENDAHULUAN

Bab ini berisi latar belakang, rumusan masalah, batasan penelitian , tujuan penelitian, manfaat penelitian dan sistematika penelitian.

BAB II TINJAUAN PUSTAKA

Bab ini memuat penelitian-penelitian terdahulu yang relevan dengan penelitian yang dilakukan, bab ini juga memuat teori-teori yang berkaitan dengan masalah yang diteliti.

BAB III METODE PENELITIAN

Pada bab ini akan mengulas gambaran secara umum penelitian, alat dan bahan, serta langkah-langkah dalam penelitian.

BAB IV HASIL DAN PEMBAHASAN

Bab ini membahas hasil dari penelitian yang dilakukan, berupa hasil balancing menggunakan teknik *oversampling*.

BAB V PENUTUP

Bab ini memuat kesimpulan dari penelitian yang telah dilakukan yang menjawab rumusan masalah yang telah dibuat, serta memberikan saran sesuai dengan penelitin yang telah dilakukan untuk penelitian selanjutnya