BAB I
PENDAHULUAN

1.1 Latar Belakang

Kemajuan terbaru dalam sains dan teknologi telah membantu mendorong ketersediaan data mentah secara luas, yang kemudian diolah menjadi data yang dapat digunakan, hal ini telah menciptakan peluang penemuan baru dan pengetahuan baru dalam Machine Learning yang memainkan peran penting dalam penerapan masalah nyata di berbagai bidang pada tingkat mikro hingga makro. Salah satu metode Machine Learning yang sering digunakan adalah Decision Tree, Decision Tree adalah model yang sering digunakan sebagai algoritme klasifikasi yang tergolong Machine Learning tradisional. Machine Learning konvensional menawarkan efisiensi dalam menemukan informasi tersembunyi dalam data ketika jumlah kelasnya seimbang. Namun pada kenyataannya, data yang dihasilkan oleh perkembangan ilmu pengetahuan dan teknologi memiliki kelas yang tidak seimbang dan seringkali terkonsentrasi pada satu kelas, sehingga menjadikan kinerja dari algoritma Machine Learning menurun [1], [2].

Masalah di atas telah dipelajari oleh peneliti lain, pada jurnal "ANALYZING THE IMPACT OF RESAMPLING METHOD FOR IMBALANCED DATA TEXT IN INDONESIAN SCIENTIFIC ARTICLES CATEGORIZATION" Jurnal tersebut membahas tentang kelas imbalance pada dataset Indonesian Scientific Journal Database yang didapat, dan metode yang digunakan penulis adalah metode eksperimen untuk menganalisis pengaruh teknik resampling, setelah dilakukan teknik resampling penulis menyimpulkan bahwa teknik resampling dapat meningkatkan performa dari model Machine Learning [3].

Pada jurnal lain yang berjudul "The impact of class imbalance in classification performance metrics based on the binary confusion metrik" dibahas bahwa kelas dataset yang tidak seimbang dapat mempengaruhi kinerja metrik klasifikasi, penulis menggunakan metode pengujian empiris dengan menggunakan beberapa dataset yang tersedia secara publik, dan penulis menyimpulkan kelas yang

tidak seimbang dapat mempengaruhi kinerja metrik klasifikasi, dan metrik kinerja yang umum digunakan seperti akurasi, presisi, dan recall dapat memberikan hasil yang menyesatkan [4].

Namun dari jurnal di atas tidak di sebutkan ciri-ciri dari feature pada dataset yang digunakan sebagai objek eksperimen, sehingga belum diketahui apakah semua metode resampling bisa diterapkan pada semua jenis dataset, maka dari itu penulis mengusulkan untuk menguji teknik resampling terhadap beberapa tipe dataset, untuk mengetahui kekurangan dan kelebihan serta kecocokan metode resampling terhadap tipe dataset tertentu.

1.2 Rumusan Masalah

Rumusan masalah dari latar belakang di atas yang akan di teliti oleh penulis adalah:

1. Bagaimana penanganan ketidakseimbangan data (imbalance data) menggunakan metode resampling.
2. Apakah semua metode resampling bisa diterapkan ke semua jenis dataset?.
3. Bagimana hasil kinerja metode resampling terhadap model Machine Learning.

1.3 Batasan Masalah

Untuk mempersempit pembahasan pada skripsi ini, maka dibuat batasan-batasan sebagai berikut:

a. Dataset yang digunakan dalam penelitian ini diambil dari Kaggle Repository.
b. Menggunakan Decision Tree sebagai model klasifikasi.
c. Penelitian ini belum sampai pada tahap pengembangan aplikasi.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah menerapkan berbagai algoritma resampling untuk mengetahui algoritma yang cocok untuk diterapkan terhadap suatu tipe dataset tertentu.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat untuk peneliti tentang pengguaan library resampling terhadap dataset, diantaranya:

a. Penelitian ini diharapkan bisa membantu Data Analyst dalam menentukan model resampling terhadap suatu tipe dataset tertentu.
b. Penelitian ini diharapkan dapat digunakan sebagai referensi bagi para peneliti dalam mengembangkan model resampling dalam mengolah dataset.