BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Pembelajaran Mesin (*Machine Learning*) merupakan salah satu bidang ilmu komputer yang berkembang dengan pesat serta memiliki penerapan yang sangat luas. Pembelajaran Mesin sering direferensikan sebagai sebuah pengenalan pola otomatis pada suatu data. Pada beberapa dekade terakhir, hal tersebut telah digunakan sebagai solusi untuk hampir setiap permasalahan yang membutuhkan ekstraksi informasi dari kumpulan data yang besar. Pembelajaran mesin memungkinkan computer untuk mempelajari atau bahkan mengembangkan kinerja program menggunakan sekumpulan data atau biasa disebut *dataset* [1].

Salah satu jenis pendekatan pembelajaran pada *Machine Learning* adalah *supervised learning*. *Supervised Learning* merupakan pendekatan pembelajaran dengan memanfaatkan label dari *dataset* [2]. Dengan kata lain, *supervised learning* belajar menggunakan contoh yang sudah disediakan agar dapat menemukan *output* yang optimal. Salah satu metode *supervised learning* yang populer adalah klasifikasi.

Klasifikasi adalah metode untuk mengategorisasikan sekumpulan data kedalam kelas-kelas yang berbeda. Kelas-kelas tersebut biasa direferensikan sebagai target atau label yang sudah ditetapkan atau diperoleh dari dataset yang digunakan [2].

Namun pada beberapa penelitian yang telah dilakukan terkait penerapan metode klasifikasi, seringkali para peneliti tidak memperhatikan keseimbangan distribusi kelas pada *dataset*. Ketidakseimbangan kelas pada dataset (*imbalanced dataset*) merupakan situasi atau kondisi dimana nilai dari kelas minoritas sangat jauh lebih kecil dengan kelas mayoritas atau sangat kurang memadai sehingga model lebih mengenali pola pada kelas mayoritas dibanding kelas minoritas [3]. Sebagai contoh dalam dunia medis, jumlah data dari pasien yang mengidap diabetes lebih sedikit dibandingkan dengan pasien yang tidak mengidap diabetes, hal ini berpotensi menghasilkan klasifikasi yang kurang tepat dan dapat berakibat fatal jika diimplementasikan di dunia nyata [4].

Permasalahan *imbalanced dataset* termasuk salah satu tantangan yang sangat penting dalam komunitas penelitian pembelajaran mesin [5]. Faktanya, berbagai jenis metode telah dikembangkan untuk mengatasi permasalahan tersebut, seperti *resampling methods, cost-sensitive approaches, ensemble learning algorithms, kernel-based methods,* dan *active learning methods* [6]. Teknik tersebut bisa dikategorikan ke dalam beberapa pendekatan, berdasarkan bagaimana cara mereka mengatasi masalah *imbalance data*. Pendekatan pada level algoritma (internal) dengan membuat atau memodifikasi sebuah algoritma, untuk memperhitungkan signifikansi dari kelas positif. Pendekatan algoritma mencakup metode *cost-sensitive* dan pendekatan berbasis pengenalan. Pendekatan level data (*external*) menambahkan langkah *preprocessing*, dimana distribusi data diseimbangkan kembali untuk

mengurangi efek distribusi kelas mayoritas dalam proses *learning*. Salah satu alternatif lain dalam meningkatkan akurasi kelas *imbalance* adalah dengan menggunakan metode *ensemble*. Metode *ensemble* pada prinsipnya mengkombinasikan sekumpulan classifier yang dilatih dengan tujuan untuk membuat model klasifikasi (*classifier*) campuran yang terimprovisasi sehingga membuat *classifier ensemble* yang terbentuk lebih akurat dari pada *classifier* asalnya dalam melakukan suatu pengklasifikasian. *Boosting, Stacking, dan Bagging* adalah metode yang paling popular digunakan [7]. *Boosting* adalah metode yang mengonversi *weak learners* atau model yang lemah dalam belajar menjadi *strong learners* atau model yang baik dalam belajar. *Boosting* dimulai dengan melatih *base learner* atau model dasar kemudian menyesuaikan distribusi dari sampel latihan berdasarkan hasil dari model dasar sehingga sampel yang diklasifikasikan dengan tidak benar akan mendapat perhatian lebih oleh model berikutnya. Setelah model dasar pertama dilatih, maka model dasar kedua dilatih dengan sampel latihan yang sudah disesuaikan pada model dasar sebelumnya, kemudian hasilnya digunakan untuk menyesuaikan distribusi sampel latihan selanjutnya. Proses tersebut dilakukan berulang-ulang hingga model dasar mencapai nilai T yang telah ditentukan dan akhirnya model dasar akan ditimbang dan dikombinasikan [2]. Salah satu jenis teknik boosting adalah Gradient Boosting, yang mana merupakan ensemble berbasis pohon yang dapat diaplikasikan pada berbagai *loss function*. Berbeda dengan metode boosting tradisional dimana *weak learners* melakukan *fitting* model pada hasil output dari sampel, pada setiap iterasi dalam algoritma ini, *decision trees* dihasilkan dengan

melakukan *fitting* menggunakan gradien negatif. Gradien negatif biasa disebut juga residual error yang mana merupakan fungsi dari selisih antara nilai prediksi dan nilai asli dari output [8].

Namun, mayoritas metode tersebut hanya terfokus pada *binary dataset*. Jelas, masalah pembelajaran ketidakseimbangan multi-kelas jauh lebih sulit untuk diatasi daripada skenario biner, karena batas keputusan melibatkan pembedaan antara lebih banyak kelas. Sayangnya, secara langsung menerapkan metode yang diusulkan untuk menangani ketidakseimbangan *binary dataset* pada *multiclass dataset* mungkin tidak valid [9]. Ada tiga jenis kesulitan yang terkait dengan kumpulan data tidak seimbang multi-kelas: satu kelas mayoritas dan banyak kelas minoritas, satu minoritas dan banyak kelas mayoritas, dan banyak minoritas dan banyak kelas mayoritas [10]. Selain itu, distribusi instance yang miring di antara kelas bukan satu-satunya sumber kesulitan bagi algoritma klasifikasi untuk menangani kumpulan data yang tidak seimbang multi-kelas. Kesulitan yang tertanam dalam struktur data juga selalu ada, seperti kelas yang tumpang tindih, pemisahan kecil (kelas minoritas dapat terdiri dari beberapa subkonsep) dan ukuran sampel yang kecil (kurangnya contoh minoritas yang representatif) [6]. Seperti dijelaskan di atas, dalam *multiclass imbalance learning*, karakteristik yang berbeda direpresentasikan di wilayah yang berbeda, yang terkait dengan kesulitan klasifikasi yang berbeda. Oleh karena itu, dibutuhkan sebuah algoritma yang dapat menyesuaikan dengan berbagai permasalahan yang berbeda. Salah satu metode yang memiliki berbagai fitur yang dapat membantu dalam

penyesuaian adalah XGBoost (*Extreme Gradient Boosting*). XGBoost merupakan versi Gradient Boosting yang telah dioptimalisasi, yang mana diperkenalkan oleh Chen pada tahun 2016. XGBoost merupakan sebuah sistem pembelajaran mesin yang dapat diskalakan untuk *tree boosting*. XGBoost dapat menambahkan komponen regularisasi pada *loss function* sehingga kompleksitas *ensemble* yang dihasilkan dipertimbangkan bersama dengan prediktabilitas di setiap split. Selain itu, XGBoost memungkinkan penggunanya untuk mengurangi *overfitting* pada model dengan mengatur beberapa *hyper-parameter* seperti kompleksitas *single tree*, kompleksitas *forest*, *learning rate*, *regularization terms*, *column subspaces*, *dropouts*, dan sebagainya. XGBoost juga menyediakan fitur tambahan seperti penanganan data hilang dengan *nodes default directions*, menghitung secara efisien ambang batas pemisahan potensial selama pemisahan node, dan mendukung *platform* terdistribusi seperti Apache Hadoop [11]. Salah satu kelebihan terbesar dari algoritma XGBoost adalah skalabilitas dalam segala skenario. Sistem pada XGBoost berjalan sepuluh kali lebih cepat dibanding metode-metode populer pada single machine dan menskalakan hingga miliaran contoh pada pengaturan terdistribusi maupun pengaturan yang dibatasi oleh penggunaan memori. Skalabilitas XGBoost disebabkan oleh sejumlah sistem penting dan optimisasi algoritma. Inovasi-inovasi tersebut antara lain: sebuah algoritma berbasis pohon baru yang digunakan untuk menangani sparse data, prosedur sketsa kuantil berbobot yang dibenarkan secara teoritis memungkinkan penanganan bobot instans dalam pembelajaran pohon perkiraan. Komputasi paralel dan terdistribusi membuat proses pembelajaran jadi lebih cepat yang mana

memungkinkan eksplorasi model yang lebih cepat. Lebih penting lagi, XGBoost mengeksploitasi komputasi out-of-core dan memungkinkan data scientist untuk memproses ratusan juta contoh [12].

1.2 Rumusan Masalah

Berdasarkan dengan hal-hal yang telah dijabarkan pada latar belakang, maka yang akan dibahas dalam penelitian ini adalah bagaimana kinerja algoritma XGBoost dalam menangani *multiclass imbalanced dataset*?

1.3 Batasan Penelitian

Adapun batasan masalah terkait dengan penelitian ini agar tidak menyimpang dalam pembahasan adalah sebagai berikut:

1. Penelitian ini untuk mengatasi ketidakseimbangan dataset menggunakan algoritma XGBoost.
2. Dataset dalam penelitian ini merupakan jenis data sekunder yang sudah tersedia di internet dan legal untuk digunakan oleh umum bersumber dari *repository KEEL* dan *UCI*.
3. Dataset yang digunakan berupa data numerik.

4. Dalam penelitian ini menggunakan pendekatan terhadap data, sehingga algoritma klasifikasi hanya digunakan sebagai penguji untuk melakukan evaluasi matriks saat *training* dan *testing*.
5. Matriks Evaluasi yang digunakan berfokus pada kasus *imbalanced dataset* menggunakan *balanced accuracy*, *geometric mean*, MAUC, *sensitivity*, *specificity*.
6. Implementasi dituliskan dalam bahasa pemrograman python dan menggunakan IDE Jupyter Notebook.
7. Penelitian ini tidak sampai membuat sistem informasi.

1.4 Tujuan Penelitian

Adapun tujuan dari penelitian ini dilaksanakan adalah sebagai berikut :

1. Menguji kinerja algoritma XGBoost dalam menangani *multiclass imbalanced dataset*.

1.5 Manfaat Penelitian

Dalam penelitian ini diharapkan metode yang diusulkan yaitu XGBoost dapat menangani permasalahan *multiclass imbalanced dataset*.

1.6 Metode Penelitian

1.6.1 Metode Pengumpulan Data.

Dataset yang digunakan pada penelitian ini merupakan data sekunder yang bersumber dari *repository* KEEL da UCI.

1.6.2 Metode Klasifikasi dan Penanganan Ketidakseimbangan Kelas.

Metode kalsifikasi dan metode untuk mengatasi ketidakseimbangan data yang digunakan adalah algoritma XGBoost yang memiliki *built-in classifier* berbasis *tree*.

1.6.3 Metode Evaluasi.

Pada tahap ini dilakukan perbandingan antara kinerja algoritma klasifikasi pada dataset asli, algoritma klasifikasi dengan penambahan teknik *ensemble*. Indikator evaluasi yang digunakan pada penelitian ini adalah *balanced accuracy*, *geometric mean*, MAUC, *sensitivity*, dan *specificity*. *Balanced accuracy* difungsikan untuk menghitung kelas hasil positif dan negatif dan tidak memberi hasil yang salah pada data yang tidak seimbang semntara *geometric mean* atau *G-mean* mengindikasikan keseimbangan antara kinerja

klasifikasi pada kelas mayoritas dan minoritas. Ukuran G-mean diambil berdasarkan *sensitifity* (akurasi dari data positif) dan *specificity* (akurasi data negatif).

1.7 Sistematika Penulisan

Materi - materi dalam Laporan Skripsi meliputi beberapa sub bab dan diuraikan dengan sistematika penulisan sebagai berikut:

BAB I PENDAHULUAN

Berisi tentang latar belakang, rumusan masalah, batasan penelitian, tujuan penelitian, manfaat penelitian, metode penelitian dan sistematika penulisan.

BAB II LANDASAN TEORI

Bab ini berisi tentang penelitian terdahulu yang berkaitan dengan masalah penelitian, dan pada bab ini juga memuat teori - teori dan konsep untuk penyelesaian masalah yang diusulkan.

BAB III METODE PENELITIAN

Bab ini berisi tentang metode penelitian yang akan dilakukan seperti alat dan bahan, dan alur penelitian yang akan dilakukan.

BAB IV HASIL DAN PEBAHASAN

Bab ini akan dibahas mengenai hasil dari penelitian yang telah dilakukan yaitu, hasil implementasi teknik *ensemble* pada *multiclass imbalanced dataset*.

BAB V PENUTUP

Berisi tentang kesimpulan dari penelitian yang sudah dilakukan serta saran yang didasarkan pada hasil penelitian dan diharapkan dapat menjadi tambahan informasi untuk penelitian – penelitian selanjutny