BAB I

PENDAHULUAN

1.1 Latar Belakang Masalah

Secara bertahap Pembelajaran Mesin (*Machine Learning*) menjadi sebuah bidang studi yang sangat diharapkan perkembangannya, karena *Machine Learning* sendiri merupakan bidang studi ilmiah yang mampu memungkinkan sebuah program untuk melakukan tugas tertentu tanpa harus diberikan instruksi secara eksplisit, dengan mengandalkan pola dan inferensi. Tingkat kemampuan belajar suatu program ditentukan oleh algoritmanya. *Machine learning* mampu beradaptasi dengan keadaan yang baru, serta dapat memprediksi dan memperkirakan suatu pola. Implementasi *machine learning* dapat dicapai menggunakan kaidah, pendekatan statistik dan pendekatan fisiologis. Algoritme *machine learning* secara otomatis membangun sebuah model matematika menggunakan data sampel juga lebih dikenal sebagai "*data training*" untuk membuat keputusan-keputusan tanpa diprogram secara khusus [1].

Machine Learning dapat membuat komputer belajar (atau mengembangkan kinerjanya) dengan sendirinya berdasarkan data yang ada didalamnya yaitu menggunakan pendekatan *supervised learning*. *Supervised Learning* merupakan supervisi pembelajaran dari pembelajaran mesin dengan memanfaatkan target label yang merupakan hasil dari dataset yang telah di training [2]. Kemudian algoritma mempelajari dari data latih yang biasanya berupa *array* atau vektor, kadang kadang disebut dengan vektor fitur dan menghasilkan sebuah model statistik yang mampu memetakan masukan yang baru menjadi keluaran yang tepat [3]. Dapat disimpulkan bahwa *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 menyusun data secara sistematis atau menurut aturan atau kaidah yang telah ditetapkan. Sebuah kaidah atau aturan diperoleh dari pembelajaran sebuah himpunan data (*dataset*). Namun pada 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 (kelas positif) sangat jauh lebih kecil dengan kelas mayoritas (kelas negatif) atau sangat kurang memadai sehingga sulit untuk mendapatkan sebuah model klasifikasi yang kuat. Sebagai contoh dalam dunia medis, pasien yang dideteksi menderita kanker ganas (kelas minoritas) lebih sedikit daripada pasien menderita kanker jinak, hal ini berpotensi kesulitan mendapatkan hasil klasifikasi dengan tepat [4][5].

Melihat pentingnya permasalahan kelas *imbalance*, berbagai macam teknik telah dikembangkan untuk mengatasinya. 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 pendekatan pada level algoritma 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. *Bagging* dan *Boosting* adalah metode yang paling popular digunakan. *Bagging* dan *Boosting* adalah salah satu metode *ensemble* yang berbasis variasi data *ensemble*, yang terdiri dari

memanipulasi data training sedemikian rupa sehingga masing-masing *classifier* dilatih dengan data training yang berbeda [5]. *Bagging* terdiri dari beberapa *training classifiers* yang berbeda dengan replika *bootstrap* dari kumpulan data pelatihan yang asli. Artinya, kumpulan data baru dibentuk untuk melatih setiap *classifier* dengan melakukan penggambaran secara acak (*with replacement*) contoh dari kumpulan data asli (biasanya, untuk mempertahankan ukuran kumpulan data asli). Fokus utama *bagging* adalah untuk mencapai varian (*noise*) yang lebih sedikit daripada model manapun secara individual [6][7]. Metode *bagging* akan mencapai kinerja terbaiknya saat bertemu dengan data yang tidak stabil dan cenderung *overfitting*, maksudnya perubahan kecil pada data latih akan mengarahkan pada perubahan prediksi output yang lebih besar (utama). Metode bagging bekerja secara efektif dalam mengurangi varian (*noise*) data dengan melakukan agregasi pada data individu yang sebenarnya memiliki data statistik yang berbeda. Kelebihan metode *bagging* daripada teknik *ensemble* yang lain adalah *bagging* merupakan algoritme *data-specific*, teknik bagging mampu mengurangi model data yang mengalami *overfitting*. Dia juga mampu bekerja dengan baik pada data berdimensi tinggi, terlebih lagi nilai yang hilang dari dataset tidak akan berpengaruh pada kinerja dari algoritmenya. Pada sebuah kasus dengan varian (*noise*) model data yang tinggi seperti *Decision Tree* dapat dieksekusi dengan baik oleh metode bagging. Jika digunakan pada kasus dengan varian (*noise*) model data yang rendah seperti Linear Regresi, itu tidak terlalu mempengaruhi kinerja dari proses belajarnya [6][8]. Pemilihan metode bagging (*Bootstrap Aggregating*) sangat tepat untuk membantu meningkatkan hasil akurasi klasifikasi pada dataset [9]. Teknik *Bagging* merupakan metode yang dapat memperbaiki hasil dari algoritma klasifikasi *machine learning*, oleh karena itu pada penelitian ini peneliti mengusulkan penggunaan *Bagging Method* sebagai algoritma guna menangani ketidakseimbangan distribusi kelas [10]. Di bidang klasifikasi, ada banyak cabang yang berkembang yaitu pohon keputusan (*decision tree*), klasifikasi Bayesian, jaringan syaraf tiruan dan algoritma genetika. Dalam penelitiannya menyatakan bahwa decision tree memang populer dan sering digunakan dalam klasifikasi karena memiliki hasil yang cukup baik jika

dibanding algoritma lainnya [11], algoritma pohon keputusan hanya digunakan sebagai penguji untuk melakukan evaluasi matriks saat training dan testing. Sehingga diharapkan dengan melakukan penanganan terhadap distribusi data dapat meningkatkan kinerja dari algoritma klasifikasi menurut hasil evaluasi matriks akurasi dan nilai geometric mean.

Namun dalam beberapa kasus ditemukan bahwa bagging mengalami penurunan kinerja disebabkan oleh pengaruh dari dataset yang digunakan, Adaboost(*boosting*) yang juga merupakan algoritma dari *ensemble method* tampaknya lebih efektif daripada *bagging* bila diterapkan ke C4.5 (*Decision Tree*), meskipun kinerja *C4.5 bagging* memiliki lebih sedikit variabelnya dibandingkan dengan variabel milik *boosting*. Jika bobot *voting* yang digunakan untuk melakukan *aggregate component classifier* (*bagging*) diubah menjadi *boosting classifier* untuk merefleksikan keyakinan dengan masing-masing *instance* yang diklasifikasikan, hasil yang lebih baik diperoleh di hampir semua kumpulan data yang diselidiki [12]. Oleh karena itu penelitian ini juga akan melakukan perbandingan antara bagging dan boosting untuk melihat batas dari kemampuan bagging, dan melakukan peningkatan pada bagging dengan kombinasi antara metode *sampling* dan *ensemble* atau biasa disebut *balanced-bagging* (*undersampling* dan *bagging*). *Undersampling* sendiri adalah metode yang efisien untuk pembelajaran keseimbangan kelas. Metode ini menggunakan subset dari kelas mayoritas untuk melatih pengklasifikasi. Karena banyak contoh kelas mayoritas diabaikan, set pelatihan menjadi lebih seimbang dan proses pelatihan menjadi lebih cepat. Namun, kelemahan utama dari undersampling adalah bahwa informasi yang berpotensi berguna yang terkandung dalam contoh yang diabaikan ini akan diabaikan. Untuk mengatasinya, dapat dilakukan dengan memodifikasi algoritma dengan mengidentifikasi data penting dari kelas mayoritas dan menggunakan teknik undersampling untuk data sisa [13]. Kombinasi antara *under-sampling* dan *bagging* saat ini menjadi teknik ensembel paling akurat yang dikhususkan untuk data kelas yang tidak seimbang. Ide dasar dibalik kombinasi *bagging* untuk data yang tidak seimbang

adalah untuk memodifikasi distribusi contoh dari kelas minoritas dan mayoritas di bootstraps. Ini dapat dicapai dengan banyak cara, salah satunya adalah menggunakan metode yang biasanya menyeimbangkan jumlah contoh dari kedua kelas. Studi eksperimental menunjukkan bahwa *under-sampling* (yaitu pengurangan contoh dari kelas mayoritas) berkinerja lebih baik daripada *over-sampling* (yaitu, perkalian contoh dari kelas minoritas) [7][14].

1.2 Rumusan Masalah

Berdasarkan dengan hal-hal yang telah dijabarkan pada latar belakang, maka yang akan dibahas dalam penelitian ini adalah sebagai berikut:

1. Bagaimana kemampuan algoritma klasifikasi menghadapi dataset dengan distribusi dari kelas yang tidak seimbang ?
2. Apakah implementasi Teknik Stratified K-Fold Cross Validation berpengaruh terhadap hasil kinerja metode ensemble ?
3. Bagaimana dampak kinerja sebuah model klasifikasi setelah mengalami proses bootstrap aggregating dengan *ensemble technique Bagging Method*?
4. Bagaimana dampak kinerja dari metode bagging setelah dilakukan penyeimbangan kelas data dengan metode *under-sampling* (*Balanced-Bagging*)?

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 metode *ensemble* dan difokuskan sebatas untuk mengetahui dampak dari penanganan distribusi kelas tersebut.

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. Pada penelitian ini dibatasi tiga skenario yang dibandingkan. Pertama, diklasifikasikan dengan pengklasifikasian murni tanpa bagging. Kedua, diklasifikan dengan menggunakan *Base Classifier* yang sudah diimplemntasikan menggunakan metode *Bagging*. Ketiga, melakukan klasifikasi ulang menggunakan metode *Balanced-Bagging*.
5. Dalam penelitian ini menggunakan pendekatan terhadap algoritma, sehingga algoritma klasifikasi akan melalui proses yang didalamnya telah diimplementasikan *ensemble method*.
6. Algoritma klasifikasi yang digunakan adalah *Decision Tree* dan *Logistic Regression*.
7. Matriks Evaluasi yang digunakan berfokus pada kasus *imbalance* dataset menggunakan akurasi (*balanced accuracy*) dan *geometric mean*.
8. Implementasi dituliskan dalam bahasa pemrograman python dan menggunakan *jupyter notebook*.
9. Penelitian ini tidak sampai membuat sistem informasi.

1.4 Tujuan Penelitian

Adapun tujuan dari penelitian ini dilaksanakan adalah sebagai berikut :

1. Menguji pengaruh perbedaan distribusi kelas pada dataset terhadap kemampuan klasifikasi.
2. Menguji pengaruh bootstraping data menggunakan *Bagging Method* (*Bootstrap Aggregating*) pada algoritma klasifikasi sehingga mampu

menghasilkan data dengan varian(*noise*) yang lebih sedikit daripada model mana pun secara individual.

3. Menguji pengaruh kemampuan algoritma klasifikasi yang telah diimplementasikan metode Balanced-Bagging terhadap *imbalanced dataset*.

1.5 Manfaat Penelitian

Dalam Penelitian ini diharapkan metode yang diusulkan mampu menangani masalah distribusi kelas yang tidak seimbang, dengan kemampuan algoritma Bagging (Bootstrap Aggregating) membuat beberapa sampel data baru dari data latih asli. Sampel data dibuat dengan cara *sampling with replacement*, sampel himpunan data baru yang dihasilkan disebut dengan bootstrap sample. Masing-masing bootstrap sample yang dihasilkan kemudian dilatih untuk menghasilkan model klasifikasi yang memiliki *low variance*. Diharapkan dengan mengurangi varian (*noise*) data dari pada model data yang lain dapat membuat data yang digunakan pada algoritma klasifikasi menjadi lebih stabil. Balanced-Bagging merupakan hasil improvement dari metode bagging yang memanfaatkan kemampuan dari *undersampling*, dimana pada subset bootstrap yang didapatkan akan dilakukan penyeimbangan kembali dataset, untuk menghasilkan nilai variance (noise) yang lebih sedikit. Diharapkan implementasi dari balanced-bagging mampu meningkatkan hasil akurasi dari algoritma klasifikasi.

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.

Algoritma Klasifikasi yang digunakan pada penelitian ini dimaksudkan untuk menguji *dataset* dalam tiga skenario berbeda, algoritma klasifikasi yang digunakan untuk pengujian adalah Decision Tree.

Metode untuk mengatasi ketidakseimbangan data menggunakan metode *ensemble*. Pada penelitian ini menggunakan metode *ensemble* yang diimplementasikan pada algoritma klasifikasi untuk mengatasi ketidakseimbangan kelas pada *minority dataset class*. Algoritma *ensemble* yang digunakan yaitu Bagging (Bootstrap Aggregating) dan Balanced-Bagging.

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* dan *geometric mean*. Balanced akurasi difungsikan untuk menghitung kelas hasil positif dan negatif agar tidak memberi hasil yang salah pada data yang tidak seimbang, sementara *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 TINJAUAN PUSTAKA DAN 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 *dataset imbalance*.

BAB V KESIMPULAN DAN SARAN

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