BAB I
PENDAHULUAN

1.1 Latar Belakang

Secara pengetahuan (Machine Learning) telah mengubah materi dari sains. Dalam dua dekade terakhir ini ditandai dengan adanya peningkatan secara signifikan di dalam jumlah data yang dihasilkan, serta Machine Learning menyediakan alat penting untuk mengekstrak informasi dari perangkat lunak, sehingga membantu dalam menghasilkan sebuah kesimpulan tentang materi yang tersedia seperti saat sekarang[1]. Akan tetapi dalam pemanfaatannya masih dibutuhkan pemakai sebagai pengatur sistem serta belum memperlihatkan alat dengan peran yang cerdas.

Machine learning dapat diartikan seperti aplikasi komputer serta algoritma yang diadopsi dengan aturan yang bersumber dari data sehingga menghasilkan prediksi, proses yang dimaksud adalah proses yang melalui dua tahapan yaitu latihan (*training*) dan pengujian (*testing*). Dalam penelitian terkini mengatakan bahwa terdapat tiga bagian dalam *machine learning* : Supervised Learning, Unsupervised Learning, Reinforcement Learning.Teknik Supervised Learning merupakan metode klasifikasi yang mana semua kumpulan data diberi label untuk mengklasifikasikan kelas yang tidak dikenal. Sedangkan teknik Unsupervised Learning biasa disebut *cluster* disebabkan tidak terdapat pemberian label di kumpulan data serta hasilnya tidak mengenali kelas yang telah ditentukan. Sedangkan Reinforcement Learning berjalan dalam area yang dinamis dimana rancangannya harus mengatasi tujuan tanpa adanya informasi dari komputer secara jelas jika objek tersebut sudah berhasil[2]. Pembelajaran pada metode ini berdasarkan regulasi yang konsisten dengan memanfaatkan augmentasi dari data yang berpengaruh untuk mempertahankan dugaan yang konsisten pada gambar yang tidak berlabel. Akan tetapi teknik augmentasi yang umum digunakan dalam klasifikasi terbukti kurang efektif[3]. Metode Semi-Supervised Learning dapat digunakan dalam metode lainnya seperti klasifikasi, regresi, dan prediksi. Contoh

penggunaan Semi-Supervised Learning yaitu untuk teknik identifikasi wajah seseorang pada webcam atau kamera smartphone.

Algoritma K – Nearest Neighbor merupakan satu diantara beberapa algoritma lain yang cukup sederhana dalam menyelesaikan kasus dalam klasifikasi, algoritma K – Nearest Neighbor kerap menghasilkan hasil yang signifikan. Selain itu algoritma K – Nearest Neighbor merupakan salah satu algoritma yang paling sederhana dalam menyelesaikan masalah dalam klasifikasi[4].

Algoritma Naïve Bayes merupakan klasifikasi statistik yang digunakan dalam mempridiksi probabilitas keanggotaan dalam suatu kelas. Kelebihan dari algoritma Naïve Bayes adalah memiliki akurasi serta kecepatan yang sangat tinggi ketika diterapkan ke dalam sebuah database dengan jumlah data yang besar. Selain itu algoritma Naïve Bayes merupakan metode yang cukup mudah digunakan, dimana tidak diperlukan estimasi parameter perulangan yang rumit, serta dapat diaplikasikan untuk dataset dengan ukuran besar. Mudah didefinisikan serta dianalisa sehingga pengguna yang tidak memiliki kemampuan dalam bidang teknologi klasifikasi bisa mengerti. Algoritma Naïve Bayes adalah algoritma yang dapat menekan kesalahan dibandingkan dengan *classifier* lainnya, akan tetapi dalam penerapannya tidak selalu terjadi, dikarenakan ketidak akuratan dalam teori yang dibuat untuk penggunaannya kelas yang tidak utuh serta kurangnya data probabilitas yang tersedia[3].

Klasifikasi adalah sebuah metode dalam menemukan sebuah model yang menjelaskan serta membedakan ide maupun kelas data dengan maksud mengestimasi kelas dari suatu objek yang kelasnya tidak diketahui. Klasifikasi mampu diterapkan dalam beragam bagian sehingga seiring berjalannya waktu proses klasifikasi cukup banyak berkembang, tetapi terdapat persoalan yang sering dijumpai dalam klasifikasi yaitu pada masalah ketidakseimbangan data[4]. Ketidakseimbangan data terjadi saat salah satu kelas mempunyai jumlah yang jauh lebih besar dibandingkan kelas lainnya sehingga mengakibatkan menurunnya kemampuan klasifikasi pada kelas minoritas. Kemampuan algoritma machine learning umumnya dievaluasi oleh akurasi hasil prediksi. Metode machine learning

mengarah pada pemberian label berupa kelas mayoritas pada data yang diprediksi serta mengabaikan kelas minoritas sehingga hanya akan menghasilkan hasil akurasi prediksi yang baik pada kelas mayoritas itu sendiri[4].

Sebelum melakukan klasifikasi data, sampel data perlu dijadikan pertimbangan terlebih dahulu. Sehingga harus memastikan dataset terlebih dahulu apakah jumlah dari setiap bagian data seimbang atau *imbalanced*. Jika jumlah dataset *imbalanced*, maka kelas minor (kelas dengan jumlah lebih sedikit) selama klasifikasi data akan diabaikan. Keadaan ini mengakibatkan rata-rata misklasifikasi akan lebih tinggi didalam kelas minor. Proses untuk menangani permasalahan tersebut yaitu dengan menggunakan teknik *sampling*. Teknik tersebut memperbaiki dataset yang tidak seimbang dengan langkah yang berbeda dalam menghasilkan persebaran data yang seimbang[6].

1.2 Rumusan Masalah

Berdasarkan dengan yang telah dijabarkan pada latar belakang tentang implementasi metode oversampling untuk meningkatkan kinerja algoritma klasifikasi pada kasus imbalanced dataset, peneliti menggunakan metode SMOTE (Synthetic Minority Oversampling Technique) dan metode ADASYN (Adaptive Synthetic). Berdasarkan latar belakang yang diuraikan, maka rumusan masalah dari penelitian tersebut adalah bagaimana pengaruh algoritma K – Nearest Neighbor dan Naïve Bayes terhadap hasil akurasi metode SMOTE dan ADASYN. Bagaimana pengaruh hasil akurasi *balance accuracy* dan *geometric – mean* pada metode SMOTE dan ADASYN dengan kedua algoritma dengan jenis dataset yang berbeda.

1.3 Batasan Masalah

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

1. Penelitian ini untuk mengatasi ketidakseimbangan dataset menggunakan metode *resampling oversampling* SMOTE (*Synthetic Minority Oversampling Technique*) dan ADASYN (*Adaptive Synthetic*).

2. Dataset dalam penelitian ini merupakan jenis data sekunder yang sudah tersedia di internet dan legal digunakan untuk umum untuk umum bersumber dari *repository KEEL* dan *UCI*.
3. Dataset yang digunakan berupa data numerik.
4. Pada penelitian ini dibatasi dengan tiga skema yang dibandingkan. Pertama, diklasifikasikan dengan dataset murni tanpa *resampling*. Kedua, diklasifikasikan melalui proses *resampling* SMOTE. Ketiga, diklasifikasikan melalui proses *resampling* denga metode ADASYN.
5. Dalam penelitian ini algoritma klasifikasi hanya digunakan sebagai penguji dalam melakukan evaluasi matriks pada waktu *training* dan *testing*.
6. Penelitian ini menggunakan algoritma K-Nearest Neighbor dan Naïve Bayes sebagai pembanding.
7. Inti dari kasus imbalanced dataset menggunakan evaluasi *Balance Accuracy*, *True Negative Rate* (TNR), *True Positive Rate* (TPR), *Geometric Mean* sebagai evaluasi matriks.
8. Proses implementasi dituliskan dalam bahasa pemrograman python dengan menggunakan *google colaboratory*.
9. Penelitian ini tidak sampai membuat sistem informasi.

1.4 Tujuan Penelitian

Adapun tujuan dari penelitian ini dilaksanakan adalah sebagai berikut :

1. Mengetahui pengaruh perbedaan pada masing- masing dataset terhadap kemampuan algoritma klasifikasi.
2. Mengetahui pengaruh kemampuan metode SMOTE dan ADASYN pada distribusi kelas dataset terhadap kemampuan klasifikasi.
3. Mengetahui hasil evaluasi *balance accuracy* dan *geometric – mean* dari metode yang sudah diberikan perlakuan terhadap dataset.

1.5 Manfaat Penelitian

Pada penelitian yang akan dicapai ini diharapkan dapat memberikan manfaat baik secara langsung maupun tidak langsung serta mampu dalam

menangani ketidakseimbangan kelas terhadap *dataset* dengan menggunakan kinerja algoritma SMOTE dan ADASYN dalam melakukan pemantauan kelas minoritas berdasarkan tingkat kesulitan sampel. Sehingga dapat digunakan dalam memperbaiki kekuatan algoritma klasifikasi dalam menurunkan *noise* terhadap data minoritas.

1.6 Metode Penelitian

1.6.1 Metode Pengumpulan Data

Dataset yang digunakan pada penelitian ini adalah data sekunder yang bersumber dari *repository* KEEL dan UCI.

1.6.2 Metode Klasifikasi

Algoritma yang digunakan adalah K-Nearest Neighbor dan Naïve Bayes sehingga diharapkan dapat menguji *dataset* dalam skenario yang berbeda.

1.6.3 Metode Penanganan Ketidakseimbangan Kelas

Metode untuk mengatasi ketidakseimbangan data menggunakan proses *resampling oversampling* pada *dataset* yang diimplementasikan terhadap *minority dataset class*. Algoritma *oversampling* yaitu SMOTE (*Synthetic Minority Oversampling Technique*) dan ADASYN (*Adaptive Synthetic*).

1.6.4 Metode Evaluasi

Pada bagian ini dilakukan komparasi antara performa kedua algoritma klasifikasi terhadap dataset, algoritma klasifikasi dengan ditambahkan *resampling*. Evaluasi yang digunakan pada penelitian ini adalah akurasi dan *geometric mean*.

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 PEMBAHASAN

Bab ini akan dibahas mengenai hasil dari penelitian yang telah dilakukan yaitu. Hasil implementasi dari teknik *resampling* pada *dataset imbalanced.*

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.