BAB 1

PENDAHULUAN

1.1 Latar Belakang

Data mining adalah suatu kegiatan pengumpulan data, pemakaian data historis untuk menemukan pengetahuan dari sebuah data [1]. Salah satu cabang dari data mining adalah klasifikasi. Klasifikasi merupakan teknik pengalian informasi dari data. Salah satu metode klasifikasi adalah supervised classification dimana batasan kelas ditentukan dari awal [2]. Akurasi didalam supervised classification dapat dikontrol dengan memperhatikan kualitas data yang digunakan untuk training terhadap algoritma. Permasalahan terhadap data mining klasifikasi adalah keadaan dataset yang tidak seimbang untuk data training terhadap algoritma, agar menghasilkan akurasi yang optimal. Ketidakseimbangan dataset adalah keadaan dimana distribusi kelas didalam dataset tidak seimbang.

Sebuah kelas dikatakan tidak seimbang apabila ada suatu kelas yang memiliki data yang lebih banyak dibandingkan dengan kelas lainnya [1]. Kelompok kelas dengan jumlah data yang banyak disebut dengan kelas mayoritas, sedangkan kelompok kelas dengan jumlah yang sedikit disebut dengan kelas minoritas. Perbandingan antara kelas minoritas dengan kelas mayoritas disebut dengan Imbalance Ratio (IR) atau rasio ketidakseimbangan. Semakin besar perbedaan antara kelas minoritas dengan kelas mayoritas maka nilai dari Imbalance Ratio (IR) atau rasio ketidakseimbangan semakin besar.

Ketidakseimbangan dataset pada data mining adalah masalah yang serius. Dataset yang tidak seimbang menyebabkan misleading atau kesesatan dalam Hasil klasifikasi dimana data kelas minoritas sering diklasifikasikan sebagai kelas mayoritas [4]. Penerapan algoritma klasifikasi tanpa memperhatikan keseimbangan kelas mengakibatkan prediksi yang baik bagi kelas mayoritas dan kelas minoritas diabaikan.

Apabila algoritma klasifikasi di implementasikan langsung terhadap dataset yang imbalance maka akan mengalami penurunan performa [3]. Pada penelitian ini, peneliti akan melakukan penangganan ketidakseimbangan kelas terhadap kelas minoritas. Pada penelitian ini peneliti menggunakan teknik resampling yaitu oversampling. Teknik oversampling dipilih karena tidak menggurangi dataset akan tetapi menambah dataset yang kurang pada kelas minoritas. Algoritma oversampling yang digunakan adalah Synthetic Minority Over-sampling Technique (SMOTE), algoritma ini dipilih dari beberapa algoritma resampling karena SMOTE menghasilkan akurasi yang baik dan efektif dalam menangani kelas yang tidak seimbang karena mengurangi overfitting [3].

Hal ini bertujuan untuk menyeimbangkan kelas pada dataset sehingga dapat meningkatkan kinerja dari algoritma klasifikasi. Hasil dari penelitian ini akan dibandingkan dengan Hasil klasifikasi tanpa resampling. Uji evaluasi yang digunakan ialah akurasi, Geometric Mean (g-mean), dan Confussion Matrix (CM) [4]. Data pengujian yang digunakan adalah data public yang peneliti dapatkan dari situs KEELS data mining yang menyediakan dataset dengan angka Imbalance

Rasio (IR) yang berbeda – beda. Dataset dari KEELS ini digunakan untuk menguji dan membandingkan usulan peneliti dalam menangani masalah ketidakseimbangan kelas pada sebuah dataset klasifikasi.

1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah disampaikan, maka dapat dirumuskan masalah yang diangkat pada penelitian ini adalah sebagai berikut :

1. Bagaimana implementasi Algoritma Synthetic Minority Over-sampling Technique (SMOTE) terhadap dataset yang imbalance?
2. Bagaimana pengaruh dari algoritma klasifikasi sebelum dan sesudah implementasi algoritma Synthetic Minority Over-sampling Technique (SMOTE) ?
3. Bagaimana tingkat akurasi dari algoritma klasifikasi setelah dimplementasikan Synthetic Minority Over-sampling Technique (SMOTE) pada dataset imbalance dengan tingkat Imbalance Rasio (IR) yang berbeda – beda ?

1.3 Batasan Masalah

Berdasarkan rumusan masalah yang dipaparkan dengan tujuan implementasi algoritma Synthetic Minority Over-sampling Technique (SMOTE) pada dataset imbalance, maka penelitian ini diberi Batasan masalah sebagai berikut :

1. Penelitian ini menggunakan pendekatan terhadap data atau data approach sehingga penelitian ini tidak membahas algoritma klasifikasi,

algoritma klasifikasi yang dipakai pada penelitian ini digunakan untuk pengujian akurasi antara sebelum dilakukan resampling dengan SMOTE dan setelah dilakukan resampling dengan SMOTE.

2. Algoritma klasifikasi yang digunakan untuk uji akurasi adalah C45, Naïve Bayes, K-NN, dan SVM.
3. Implementasi algoritma Synthetic Minority Over-sampling Technique (SMOTE) menggunakan bahasa pemrograman python.
4. Dataset yang digunakan merupakan data public imbalance class dari KEELS dataset.
5. Dataset yang digunakan berupa data numerik.
6. Implementasi dataset terhadap algoritma klasifikasi menggunakan tools Orange 3.
7. Jupyter notebook untuk penulisan kode python.

1.4 Maksud dan Tujuan Penelitian

Maksud dan tujuan dari penelitian ini ialah untuk menguji apakah ada perbedaan antara kinerja algoritma klasifikasi yang menggunakan teknik resampling SMOTE pada dataset klasifikasi dengan kinerja algoritma klasifikasi tanpa dilakukan teknik resampling SMOTE pada beberapa level imbalance rasio (IR).

1.5 Manfaat Penelitian

Adapun manfaat yang dapat diperoleh dalam penelitian dan penyusunan skripsi ini adalah sebagai berikut :

1. Bagi penulis

Menambah pengetahuan penulis dalam mengatasi masalah pada imbalance data menggunakan algotirma Synthetic Minority Over-sampling Technique (SMOTE).

2. Bagi Universitas Amikom Yogyakarta

Menjadi bahan referensi bagi mahasiswa yang tertarik dengan topik penelitian ini, dan atau yang mempunyai permasalahan yang sama dengan penelitian yang diangkat.

1.6 Metode Penelitian

Dataset yang digunakan pada penelitian ini adalah data public imbalance yang didapat dari KEELS dataset dengan berbagai level imbalance rasio (IR).

1.6.1 Metode mengatasi ketidakseimbangan kelas

Metode untuk mengatasi ketidak seimbangan data menggunakan Teknik resampling. Terdapat 2 metode dalam Teknik resampling yaitu oversampling dan undersampling. Pada penelitian ini penulis menggunakan teknik resampling oversampling untuk mengatasi ketidak seimbangan kelas pada dataset yang

diimplementasikan terhadap kelas minoritas. Algoritma oversampling yaitu Synthetic Minority Over-sampling Technique (SMOTE).

1.6.2 Metode Klasifikasi

Algoritma klasifikasi yang digunakan pada penelitian ini dimaksudkan untuk mengetahui akurasi algoritma klasifikasi terhadap dataset sebelum dan sesudah dilakukan resampling. Algoritma klasiikasi yang digunakan untuk penggujian adalah C45, Naïve Bayes, KNN dan SVM.

1.6.3 Metode Evaluasi

Pada tahap ini dilakukan perbandingan antara kinerja keempat algoritma klasifikasi tanpa teknik resampling, keempat algoritma klasifikasi dengan penambahan SMOTE. Indikator evaluasi yang digunakan pada penelitian ini meliputi confusion matrix dan geometric mean (g-mean). Perhitungan akurasi klasifikasi menggunakan confusion matrix sedang untuk mengukur.

1.7 Sistematika Penulisan

Adapun sistematika penulisan yang digunakan dalam penyusunan skripsi ini, terdapat 5 bab, serta daftar pustaka yaitu sebagai berikut : BAB I Pendahuluan, BAB II Landasan Teori, BAB III Metode Penelitian, BAB IV Implementasi Dan Pembahasan, BAB V Penutup, dan Daftar Pustaka. Secara garis besar pembahasan dari setiap bab adalah sebagai berikut :

BAB I PENDAHULUAN

Pada bab I pendahuluan ini, membahas tentang latar belakang permasalahan yang diangkat, rumusan masalah, Batasan masalah, maksud dan tujuan dilakukan

penelitian, manfaat penelitian, metode penelitian yang digunakan, dan sistematika yang digunakan pada penyusunan skripsi ini.

BAB II LANDASAN TEORI

Pada bab II landasan teori ini, membahas tentang tinjauan pustaka, seperti pengertian dan definisi dari setiap komponen yang digunakan pada penelitian. Tinjauan pustaka diambil dari jurnal, buku, dan seminar atau proseding yang sesuai dengan topik penelitian.

BAB III METODE PENELITIAN

Pada bab III metode penelitian ini berisi tentang penjelasan metode penelitian yang digunakan, permodelan data yang digunakan untuk penelitian, dan tahapan penelitian.

BAB IV IMPLEMENTASI DAN PEMBAHASAN

Bab IV implementasi dan pembahasan ini menjelaskan tentang implementasi algoritma yang diusulkan terhadap dataset yang imbalance, penguraian tahapan peneliti dalam melakukan penelitian, dan hasil dari penelitian.

BAB V PENUTUP

Bab V penutup ini berisi kesimpulan dari penelitian yang sudah dilakukan, penarikan hasil dari penelitian, dan berisi saran yang dapat digunakan peneliti lain untuk mengembangkan hasil dari penelitian yang sudah dilakukan.