BAB I
PENDAHULUAN

1.1. Latar Belakang

Perkembangan teknologi informasi yang sangat pesat turut mempengaruhi pertumbuhan jumlah data. pelbagai bidang misalnya bidang kesehatan turut menghasilkan data dengan jumlah yang banyak sehingga mempengaruhi laju pertumbuhan data. Selain itu, berkat adanya teknologi internet, transfer data dilakukan sangat cepat dan masif baik itu melalui aplikasi berbasis mobile, desktop maupun website.

Dengan adanya ketersediaan data yang banyak ini, dapat dimanfaatkan untuk mencari informasi, pola-pola serta pengetahuan yang bermanfaat. untuk melakukan hal tersebut, maka kita bisa menggunakan alat atau *tool* yaitu data mining. klasifikasi data merupakan salah satu teknik data mining yang sering digunakan.klasifikasi merupakan proses pengelompokkan suatu data atau objek ke dalam kelompok kelas tertentu berdasarkan kemiripan dan kesamaan properti. banyak algoritma klasifikasi yang sering digunakan seperti *decision tree, naive bayes,* SVM, knn dan lain-lain. pada proses klasifikasi terdapat masalah yang sering terjadi yaitu *class imbalance*.

Pada dasarnya *class imbalance* merupakan salah satu masalah pada klasifikasi data mining. *Class imbalance* merupakan kondisi dimana terjadi ketidakseimbangan dalam jumlah distribusi data di kelas-kelas sehingga menghasilkan kelas yang memiliki umlah data yang banyak(kelas mayoritas) dan kelas yang memiliki jumlah data yang sedikit(kelas minoritas)[3].Data yang diambil langsung dari database pada dasarnya merupakan data yang tidak seimbang. Dengan adanya ketidakseimbangan ini membuat algoritma klasifikasi seperti *naïve bayes, k-nn,* C4.5, SVM dan lain-lain menghasilkan performa yang buruk dalam mengklasifikasikan data ke dalam kelas yang benar[1]. Sehingga klasifikasi pada kelas mayoritas menghasilkan akurasi yang sangat tinggi sedangkan klasifikasi pada kelas minoritas menghasilkan akurasi yang sangat

rendah[2]. Tentu saja hal ini tidak bagus, terlebih lagi apabila hasil klasifikasi kelas minoritas itu sangat diperlukan informasinya. Karena pada kasus tertentu, kelas minoritas perlu untuk diidentifikasi secara tepat[1].

Untuk mengatasi masalah tersebut, maka dilakukan proses pendekatan level data dengan cara melakukan *resampling* pada data. terdapat dua jenis teknik resampling data yaitu *undersampling* dan *oversampling*. *Oversampling* dipilih karena jumlah data yang digunakan tidak terlalu banyak[5]. Terdapat algoritma-algoritma *oversampling* seperti *Random sampling* dan Synthetic *Minority Over Sampling Method*.

SMOTE merupakan algoritma *oversampling* yang ditawarkan oleh chawla[6] dan digunakan untuk menangani masalah *imbalanced class*. SMOTE bekerja dengan cara mensintesis data buatan berdasarkan sampel yang diambil dari data kelas minoritas. sintesis dilakukan sepanjang garis segmen dimana melibatkan sebagian atau seluruh *k-neighbor*(sampel yang terdekat)[5]. Data asli dan data yang telah di-resampling akan dilakukan proses klasifikasi.Algoritma klasifikasi yang akan digunakan adalah Algoritma Naive bayes. Algoritma ini dipilih karena bagus untuk klasifikasi data yang sedikit[4].

Dengan dilakukannya penelitian ini diharapkan dapat mengetahui dampak dari penggunaan teknik *resampling* menggunakan algoritma SMOTE terhadap tingkat akurasi klasifikasi algoritma *naive bayes*. Dan Implementasi algoritma SMOTE atau *synthetic minority oversampling technique* pada penelitian ini diharapkan dapat meningkatkan performa pada klasifikasi algoritma *naive bayes*.

1.2. Rumusan Masalah

Berdasarkan latar belakang diatas maka, rumusan masalah dari penelitian ini adalah apakah algoritma SMOTE dapat memperbaiki klasifikasi data pada kelas minoritas dan meningkatkan performa algoritma klasifikasi *naive bayes* pada dataset yang mengalami *imbalance class*?

1.3. Batasan masalah

Batasan masalah dari penelitian ini adalah sebagai berikut :

1. Menggunakan algoritma SMOTE dan Naive bayes.
2. Menguji klasifikasi menggunakan data tidak seimbang dan yang seimbang.
3. Data set yang digunakan adalah data set hasil pengobatan penyakit kutil menggunakan metode imunoterapi dari situs UCI machine learning.

1.4. Maksud dan Tujuan Penelitian

1. Menangani imbalance class menggunakan Algoritma SMOTE.
2. Mengetahui performa algoritma klasifikasi pada data tidak seimbang dan seimbang.
3. Membandingkan dan mengevaluasi dampak dari penggunaan algoritma SMOTE terhadap pada performa algoritma *naive bayes* dalam penanganan masalah *imbalance class* .

1.5. Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut :

1. Dapat mengetahui tingkat performa algoritma Naive bayes pada klasifikasi data yang telah di- *oversampling* menggunakan algoritma SMOTE.
2. Dapat menjadi referensi untuk penelitian selanjutnya yang sejenis.

1.6. Metode Penelitian

1.6.1. Metode Pengumpulan Data

Dataset yang akan digunakan didapatkan dari website UCI *machine learning.*

1.6.2. Preprocessing

Pada tahap ini akan dilakukan *oversampling* data dengan menggunakan algoritma SMOTE dan melakukan diskritisasi pada beberapa *field* atau variable.

1.6.3. Klasifikasi

Dataset yang telah di-*oversampling* maupun dataset asli akan dilakukan pengujian klasifikasi menggunakan algoritma Naive Bayes.

1.6.4. Evaluasi

setiap kali proses klasifikasi akan diukur performanya dan dibandingkan hasilnya antara klasifikasi pada dataset asli dan dataset yang telah di-*oversampling*. hal ini akan dilakukan untuk melihat pengaruh implementasi algoritma SMOTE pada performa klasifikasi algoritma Naive Bayes.

1.7. Sistematika Penulisan

1.7.1. BAB I Pendahuluan

Bab I berisi pemaparan pendahuluan, pengenalan latar belakang, rumusan masalah, batasan masalah, tujuan dan manfaat, dn sistematika penulisan.

1.7.2. BAB II Landasan Teori

Bab ini menjelaskan seluruh teori yang digunakan dalam penelitian ini meliputi klasifikasi, algoritma *Naive Bayes, imbalance class*, algoritma SMOTE dan *confusion matrix*.

1.7.3. BAB III Metodologi Penelitian

Bab ini menjelaskan proses penelitian secara garis besar yang dimulai dari pencarian dataset yang akan digunakan, melakukan proses resampling/*oversampling*, melakukan proses klasifikasi, dan mengevaluasi performa dari algoritma klasifikasi.

1.7.4. BAB IV Hasil Implementasi dan Pembahasan

Berisi pembahasan dari hasil implementasi yang telah dilakukan pada bab sebelumnya serta menganalisis hasilnya.

1.7.5. BAB V Penutup

Pada bab ini akan memaparkan kesimpulan dari hasil penelitian serta menjawab rumusan masalah dan memberikan saran untuk penelitian selanjutnya.

1.7.6. Daftar Pustaka

Mencantumkan semua referensi literatur yang mendukung proses penelitian ini.

1.7.7. Lampiran

Dokumen tambahan atau pendukung lainnya yang dilampirkan.