BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Pembelajaran Mesin (*Machine Learning*) merupakan bidang studi ilmiah yang mampu memberikan kemampuan melakukan tugas tertentu tanpa diberikan instruksi secara eksplisit, dengan mengandalkan pola dan inferensi. Kemampuan belajar menjadi dominan ditentukan oleh algoritma yang dapat dicapai baik menggunakan kaidah, pendekatan statistik dan pendekatan fisiologis. Algoritma pembelajaran membangun model matematika berdasarkan data sampel atau lebih dikenal dengan istilah data *training*, untuk membuat prediksi atau keputusan [1]. Aplikasi pembelajaran mesin sangat beraneka ragam, seperti penyaringan spam email, *computer vision*, dan sebagainya, Dimana program yang semacam ini akan sulit dikembangan dengan pemrograman eksplisit.

Supervised Learning merupakan bagian dari *Machine Learning* untuk jenis pembelajaran mesin memetakan data yang disertai dengan anotasi manusia biasa disebut dengan istilah label. Label merupakan target yang diinginkan dari data latih [2]. Kemudian algoritma mempelajari dari data latih 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]. Bisa diasumsikan bahwa *supervised learning* belajar dari sebuah contoh. Salah satu metode supervised learning yang populer ialah klasifikasi.

Klasifikasi adalah metode untuk menyusun data secara sistematis atau menurut aturan atau kaidah 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 *datasets*. Ketidakseimbangan kelas pada dataset merupakan situasi atau kondisi dimana 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, sedikit pasien dideteksi menderita kanker ganas (kelas minoritas) lebih sedikit daripada pasien menderita kanker jinak, hal ini berpotensi kesulitan mendapatkan hasil klasifikasi dengan tepat [4][5].

Selain keberadaan ketidakseimbangan kelas pada *dataset*, dalam penerapan *machine learning* seringkali dijumpai *dataset* yang memiliki dimensi yang tinggi. Hal ini ditandai dengan banyaknya jumlah *feature*. *Dataset* dengan dimensi tinggi tentunya akan memiliki pengaruh terhadap proses penerapan teknik data mining itu sendiri baik klasifikasi, klasterisasi maupun prediksi. Beberapa permasalahan yang sering disebabkan oleh *dataset* berdimensi tinggi antara lain kinerja algoritma klasifikasi baik dari sisi waktu komputasi maupun dari sisi akurasinya, penyebabnya adalah beberapa *attribute feature* yang tidak memiliki relevansi dengan *attribute class*, sehingga secara tidak langsung berpengaruh terhadap algoritma klasifikasi yang digunakan, karena algoritma klasifikasi bekerja dengan mengenali pola pada *attribute feature* untuk memprediksi maupun mengklasifikasi *attribute class* [6].

Penanganan *imbalance dataset* menggunakan teknik sampling *oversampling* ada bermacam-macam, ADASYN (Adaptive Synthetic k-NN) merupakan salah satu teknik sampling *oversampling* yang mana cara kerja untuk mengatasi ketidakseimbangan distribusi kelas pada *dataset*, metode ini melakukan sintesis berdasar pada bobot distribusi untuk kelas minoritas, dan dapat secara adaptif menggeser pada sampel yang sulit dipelajari. Keunggulan metode ADASYN (Adaptive Synthetic k-NN) terhadap metode *oversampling* lain terdapat pada proses penambahan data dilakukan berdasarkan bobot atribut yang tidak hanya dilakukan duplikasi data memicu terjadinya *overfitting* pada model yang dibangun [7][8]. Oleh karena itu pada penelitian ini akan mengusulkan ADASYN (Adaptive Synthetic k-NN) sebagai metode penanganan distribusi ketidaksemimbangan distribusi kelas dan pada proses fitur seleksi menggunakan Information Gain (IG) mampu mendeteksi fitur yang memiliki relevansi terhadap kelas – kelas tertentu dan mengurangi *noise* yang disebabkan oleh fitur tidak relevan [6][27]. Naïve Bayes dan Decision Tree merupakan algoritma yang umum digunakan dalam kasus klasifikasi namun masih jarang diujikan terhadap *imbalance dataset* [6], algoritma ini hanya berperan sebagai penguji untuk melakukan evaluasi matriks saat *training* dan *testing*. Sehingga diharapkan dengan melakukan penanganan terhadap distribusi data dan pengurangan dimensi fitur dapat meningkatkan kinerja algoritma klasifikasi menurut hasil evaluasi matriks akurasi dan nilai *geometric mean*.

1.2 Rumusan Masalah

Berdasarkan yang sudah dijabarkan pada latar belakang, maka yang akan dibahas dalam penelitian ini adalah sebagai berikut:

1. Bagaimana kemampuan algoritma klasifikasi menghadapi *dataset* dengan distribusi yang tidak seimbang?
2. Apakah dominasi dalam sebuah kelas mempengaruhi kemampuan dari sebuah model dalam melakukan klasifikasi?
3. Bagaimana dampak performa sebuah model setelah mengalami penyeimbangan dataset dengan teknik sampling *oversampling* ADASYN (Adaptive Synthetic k-NN)?
4. Apakah *High Dimensional Data* mempengaruhi kinerja sebuah algoritma klasifikasi?
5. Bagaimana hasil evaluasi dari model yang sudah diberikan perlakuan terhadap 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 metode *resampling oversampling* ADASYN (Adaptive Synthetic k-NN) 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 digunakan untuk 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 dataset murni. Kedua, diklasifikasikan dengan menggunakan dataset melalui proses *resampling* ADASYN. Ketiga, diklasifikasikan dengan hasil *resampling* ADASYN dan reduksi fitur Information Gain.
5. Dalam penelitian ini menggunakan pendekatan terhadap data, sehingga algoritma klasifikasi hanya digunakan sebagai penguji untuk melakukan evaluasi matriks saat *training* dan *testing*.
6. Algoritma klasifikasi yang digunakan adalah Decision Tree dan Naïve Bayes.
7. Matriks evaluasi yang digunakan berfokus pada kasus *imbalance dataset* menggunakan akurasi dan *geometric mean*.
8. Implementasi dituliskan dalam bahasa pemrograman python dan menggunakan IDE 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 sintesis data menggunakan ADASYN (Adaptive Synthetic k-NN) yang menghasilkan data minoritas secara adaptif sampel sesuai distribusinya.
3. Menguji pengaruh penggunaan seleksi fitur terhadap kemampuan algoritma klasifikasi dalam membangun sebuah model dengan dimensi fitur yang lebih sederhana.
4. Menguji pengaruh kemampuan klasifikasi algoritma dengan balance dataset dengan metode sampling ADASYN (Adaptive Synthetic k-NN).

1.5 Manfaat Penelitian

Dalam penelitian ini diharapkan metode yang diusulkan mampu menangani masalah distribusi kelas yang tidak seimbang, dengan kemampuan algoritma ADASYN (Adaptive Synthetic k-NN) memberikan suatu distribusi bobot kepada pengamatan kelas minor berdasarkan tingkat kesulitan sampel, untuk dipelajari dan digunakan sebagai suatu acuan dalam memutuskan jumlah sampel sintetis yang akan di bangkitkan oleh setiap pengamatan kelas minor secara otomatis. Sehingga dapat digunakan untuk memperbaiki kemampuan algoritma klasifikasi mengurangi asumsi *noise* pada data minoritas. Information Gain (IG) merupakan salah satu metode seleksi fitur untuk melakukan perangkingan atribut dan digunakan untuk mengevaluasi suatu atribut yang tidak memiliki relevansi terhadap kelasnya sehingga hanya atribut yang memenuhi kriteria (*threshold*) yang akan dipertahankan, dengan menghilangkan fitur – fitur yang tidak sesuai mengakibatkan dimensi data menjadi lebih sederhana diharapkan mampu menciptakan model dengan kemampuan lebih baik.

1.6 Metode Penelitian

1.6.1 Metode Pengumpulan Data.

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

1.6.2 Metode Penanganan Ketidakseimbangan Kelas.

Metode untuk mengatasi ketidak seimbangan data menggunakan Teknik *resampling*. Pada penelitian ini menggunakan teknik *resampling oversampling* untuk mengatasi ketidak seimbangan kelas pada *dataset* yang diimplementasikan terhadap kelas minoritas. Algoritma *oversampling* yaitu ADASYN (Adaptive Synthetic k-NN).

1.6.3 Metode Seleksi Fitur.

Seleksi fitur digunakan untuk menyederhanakan dimensi data dalam *dataset*. Dalam penelitian ini digunakan metode Information Gain (IG) adalah seleksi fitur untuk melakukan perangkingan atribut dan digunakan untuk mengevaluasi suatu atribut yang tidak memiliki relevansi terhadap kelasnya sehingga hanya atribut yang memenuhi kriteria (*threshold*) yang akan dipertahankan.

1.6.4 Metode Klasifikasi.

Algoritma klasifikasi yang digunakan pada penelitian ini dimaksudkan untuk menguji *dataset* dalam tiga skenario berbeda, algoritma klasifikasi yang digunakan untuk pengujian adalah Naive Bayes dan Decision tree.

1.6.5 Metode Evaluasi.

Pada tahap ini dilakukan perbandingan antara kinerja algoritma klasifikasi pada dataset asli, algoritma klasifikasi dengan penambahan *resampling* dan terakhir dengan ditambahkan fitur seleksi. Indikator 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, 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 PEMBAHASAN

Bab ini akan dibahas mengenai hasil dari penelitian yang telah dilakukan yaitu, hasil implementasi teknik *resampling*, dan implementasi *feature selection* pada *dataset imbalance*.

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 selanjutnya.