BAB I
PENDAHULUAN

1.1. Latar Belakang

Data terus dihasilkan oleh organisasi, perusahaan, ataupun individu. Data merupakan sesuatu yang bersifat mentah dan dapat berupa angka, tulisan, gambar, suara, huruf, dan dalam bentuk lainnya. Data-data tersebut terus bertambah dan terus menumpuk sehingga membuat jumlah data menjadi sangat besar, jumlah data yang besar membuat data menjadi sulit untuk diolah dan menjadi tidak berguna. Data memiliki nilai yang dapat diolah untuk membentuk suatu informasi yang bermanfaat dan menambah pengetahuan yang tidak diketahui sebelumnya.

Data Mining memiliki kemampuan untuk melakukan analisis data dan dapat menampilkan pola data penting yang tidak terlihat dalam data yang memiliki jumlah sangat besar, berkontribusi terhadap strategi bisnis, dasar pengetahuan, serta penelitian sains dan kedokteran (Han dan Kamber, 2006: 4). Data yang terdapat dalam sebuah kumpulan data diklasifikasikan untuk dapat mendiskripsikan suatu data sehingga menjadi berguna sebagai pengetahuan dasar untuk diterapkan pada data yang baru. Pengelompokan data (klasifikasi) dilakukan untuk membedakan data berdasarkan kelompok (*class*) tertentu (Han dan Kamber, 2006: 287), misalnya terdapat data lulusan mahasisiwa suatu universitas yang terbagi dalam dua kelompok yaitu, mahasiswa yang aktif dan mahasiswa yang DO (*drop out*), data kemudian diklasifikasi untuk mendapatkan pola yang dapat diterapkan pada data mahasiswa angkatan baru.

Data yang beragam dan jumlah data yang besar dapat mempengaruhi sulitnya pengelompokan data kerena data memiliki kecenderungan menjadi tidak seimbang atau disebut *imbalanced data* (Longadge, 2013: 1). Data yang tidak seimbang akan memberikan hasil yang tidak akurat dikarenakan jumlah data minoritas sulit untuk dikenali atau dideskripsikan mengingat jumlah data mayoritas menguasai penyebaran data sehingga menyebabkan data minoritas berada diantara data mayoritas. Contohnya seperti data lulusan mahasiswa Fakultas Teknologi Informasi Universitas Kristen Maranatha, jika kita perhatikan data kelulusan mahasiswa yang diklasifikasi berdasarkan predikatnya terbagi menjadi 3 kategori, yaitu: predikat ‘Dengan Pujian’ (IPK 3.51-4.00) berjumlah

152 data, predikat ‘Sangat Memuaskan’ (IPK 2.76-3.50) berjumlah 272 data, dan predikat ‘Memuaskan’ (IPK 2.01-2.75) berjumlah 13 data. Predikat ‘Sangat Memuaskan’ memiliki jumlah yang sangat besar jika dibandingkan dengan predikat "Memuaskan", besar perbandingan tersebut sebesar 20,92 (hasil dari 272 dibagi dengan 13). Hal ini menyebabkan seluruh data dengan predikat "Memuaskan" menjadi *misclassified* dilihat dari hasil *confussion matrix* pada penelitian yang dilakukan oleh Tjioe Marvin (Tjioe, 2014).

Kasus diatas menunjukkan kesimpulan yang dihasilkan dari penelitian oleh Tjioe Marvin dalam menggunakan metode NBTree sebagai *classifier* memberikan pola atau *pettern* dengan kecenderungan *missclassfied* pada *dataset* yang bersifat *imbalanced.* Untuk itu dalam tugas akhir ini akan digunakan metode klasifikasi *Support Vector Machine* yang bekerja tidak menggunakan struktur pohon tetapi bekerja dengan memisahkan *class data* secara *nonlinear* melalui *kernel.*

Seluruh kategori menjadi penting untuk dikenali dan membentuk pola sehingga dapat digunakan pada data mahasiswa yang baru, permasalahan muncul ketika data yang tidak seimbang membuat kelompok tertentu menjadi *misclassified.* Penelitian ini bertujuan untuk mengenali karakteristik data seimbang berdasarkan besar perbandingan data tidak seimbang dan menyelesaikan permasalahan klasifikasi data yang tidak seimbang atau disebut *imbalanced data* dengan menggunakan beberapa metode yang diterapkan dalam beberapa contoh kasus data yang dimiliki.

1.2. Rumusan Masalah

Rumusan masalah yang muncul dari latar belakang tersebut adalah:

1. Bagaimana menganalisis karakteristik data dalam sebuah *imbalanced dataset*?
2. Sejauh mana *Support Vector Machine* dan kombinasi algoritma dapat diterapkan untuk *imbalanced data*?

1.3. Tujuan

Tujuan dilakukannya penelitian dan pembuatan aplikasi adalah:

1. Melakukan analisis karakteristik data dalam sebuah *imbalanced dataset.*
2. Menunjukkan hasil penerapan *Support Vector Machine* dan kombinasi algoritma untuk *imbalanced dataset.*

1.4. Batasan Masalah

Beberapa batasan masalah yang diberikan adalah:

1. Data *testing* yang digunakan adalah *dataset* mahasiswa Fakultas Teknologi Informasi Universitas Kristen Maranatha.
2. *Dataset* Mahasiswa yang digunakan dipisahkan menjadi dua, yaitu data Mahasiswa Aktif tahun pertama (semester 1 & 2) dan data Mahasiswa Lulusan.
3. *Dataset* Mahasiswa yang digunakan dalam tugas akhir ini berasal dari penelitian yang dilakukan oleh Tjioe Marvin (Tjioe, 2014).
4. *Dataset* Mahasiswa yang digunakan dalam tugas akhir ini memiliki atribut data yang telah diubah menjadi tipe data *nominal*
5. Algoritma *classification* yang digunakan adalah SVM (*Support Vector Machines)*
6. SVM yang terdapat pada aplikasi menggunakan *library* dari LIBSVM.
7. *Kernel* SVM yang digunakan adalah RBF kernel (*Radial Basis Function Kernel)*
8. Metode algoritma yang dikombinasikan dengan SVM adalah algoritma KSMOTE (*Kernel-based Synthetic Minority Oversampling Technique)*, algoritma *AdaBoost* dan algoritma SSO (*Sample Subset Optimization*).
9. Metode klasifikasi dalam aplikasi menggunakan *library* yang disediakan oleh WEKA dan SSO.
10. Karakteristik *dataset* dibagi menjadi *low imbalanced, medium imbalanced,* dan *high imbalanced*
11. Data *training* dibuat berdasarkan *dataset* yang diambil dari repositori KEEL (*Knowledge Extraction based on Evolutionary Learning*) dan repositori UCI (University of California, Irvine).

12. *Dataset training* telah disesuaikan dengan dataset mahasiswa Fakultas Teknologi Informasi Universitas Kristen Maranatha, penyesuaian meliputi tipe data dan kategori *imbalanced dataset.*
13. *Dataset training* memiliki atribut dengan tipe data yang telah diubah dengan tipe data *nominal.*

1.5. Sistematika Pembahasan

Sistematika penulisan untuk tugas akhir ini adalah :

BAB I PENDAHULUAN

Bab ini menjelaskan latar belakang, rumusan masalah, tujuan, batasan masalah dan sistematika pembahasan.

BAB II LANDASAN TEORI

Bab ini menjelaskan teori-teori yang berkaitan dengan pembuatan sistem dan mendukung pembuatan sistem atau aplikasi yang ada.

BAB III ANALISIS DAN DISAIN

Bab ini menjelaskan bagaimana analisis keadaan, kebutuhan dari aplikasi, perancangan aplikasi, UML, dan gambaran arsitektur dari aplikasi yang dibuat.

BAB IV PENGEMBANGAN PERANGKAT LUNAK

Bab ini digunakan untuk menjelaskan mengenai implementasi dari teknik serta perancangan aplikasi.

BAB V TESTING DAN EVALUASI SISTEM

Bab ini menjelaskan tentang pengujian dari teknik atau aplikasi yang telah dibuat.

BAB VI KESIMPULAN DAN SARAN

Dalam bab ini, dikemukakan pengetahuan yang didapat setelah mengerjakan penelitian pengembangan aplikasi dan saran berupa hal-hal baru yang dapat dilakukan untuk pengembangan lebih lanjut.