1. Pendahuluan

1.1. Latar Belakang

Cepatnya kemajuan dalam pengumpulan data dan teknologi penyimpanan data memungkinkan suatu organisasi untuk menghimpun data dalam jumlah yang sangat besar tanpa dapat dimanfaatkan dengan baik. Akibatnya pengambilan keputusan seringkali berdasarkan intuisi saja, sebab para pengambil keputusan tidak mempunyai alat untuk mengekstrak informasi barharga dari database yang sangat besar. Hal inilah yang melatarbelakangi lahirnya sebuah bidang ilmu yang berupaya menemukan pola ataupun informasi yang belum diketahui sebelumnya dari sekumpulan besar data, yaitu data mining.

Dalam data mining juga terdapat berbagai macam permasalahan, salah satunya yaitu permasalahan *imbalance class* atau kelas *imbalance.* Permasalahan ini diakibatkan karena adanya jumlah data yang tidak seimbang pada kelas yang berbeda, dimana kelas yang satu memiliki jumlah data yang sangat banyak (mayoritas), sedangkan kelas lainnya memiliki jumlah data yang sangat sedikit (minoritas). Pada pengklasifikasian biasa, kelas minoritas tersebut tidak dapat terprediksi. Hal ini disebabkan karena bila jumlah data pada satu kelas sangat kecil, kelas tersebut akan salah diprediksi sebagai kelas mayoritas. Salah satu metode klasifikasi yang biasa digunakan adalah metode *decision tree.* Metode *decision tree* ini akan membentuk model seperti *flow chart* dengan struktur *tree*. *Node* internal menandakan *test* terhadap sebuah atribut. Cabang merepresentasikan keluaran dari *test. Node leaf* merepresentasikan label kelas atau distribusi kelas.

Pada Tugas Akhir ini diperbandingkan tiga algoritma pada *decision tree* yaitu C4.5, CART, dan HDDT. Setiap algoritma tersebut memiliki *splitting criterion* atau *attribute selection measure* yang berbeda dengan tingkat *sensitivitas* yang berbeda pula terhadap kecondongan data. Algoritma C4.5 merupakan algoritma yang paling populer digunakan. Algoritma ini menggunakan *information gain* sebagai *splitting criterion-*

nya. *Information gain* adalah salah satu a*ttribute selection measure* yang digunakan untuk memilih *test* atribut tiap *node* pada *tree*. Atribut dengan *information gain* tertinggi dipilih sebagai *test* atribut dari suatu *node*. CART merupakan kepanjangan dari *classification and regression tree*. Algoritma CART menggunakan *gini index* sebagai *splitting criterion*-nya. Dan algoritma yang ketiga yaitu algoritma *hellinger distance decision tree* (HDDT). Algoritma ini menggunakan suatu *measure* yaitu *hellinger distance* sebagai *splitting criterion*-nya.

Pada Tugas Akhir ini, dianalisis bagaimana performansi dari ketiga algoritma *decision tree* tersebut terhadap data yang bersifat *imbalance* serta akan diperbandingkan algoritma mana yang memiliki performansi yang paling baik. Performansi yang diukur adalah nilai *precession, recall,* dan *f-measure*.

1.2. Perumusan Masalah

Permasalahan yang dijadikan objek penelitian dalam Tugas Akhir ini antara lain :

1. Bagaimana performansi dari algoritma C4.5, CART, dan HDDT yang memiliki *splitting criterion* yang berbeda pada permasalahan kelas *imbalance*.
2. Bagaimana perbandingan performansi dari ketiga algoritma *decision tree* tersebut berdasarkan nilai *precision, recall,* dan *f-measure*.

1.3. Batasan Masalah

Untuk menghindari meluasnya materi pembahasan Tugas Akhir ini, maka permasalahan yang dibahas dalam Tugas Akhir ini mencakup hal-hal berikut :

1. Dataset yang digunakan adalah dataset dari UCI Machine Learning Repository dan dataset buatan yang dihasilkan dari data generator.
2. Dataset yang digunakan merupakan data dengan kelas *binary* tanpa *missing value*.
3. Implementasi algoritma HDDT menggunakan NetBeans IDE 6.0 dengan JDK 1.6.

4. Klasifikasi algoritma C4.5 dan CART menggunakan *classifier* yang ada di WEKA, yaitu J48 yang menerapkan algoritma C4.5 dan SimpleCart yang menerapkan algoritma CART.

1.4. Tujuan Pembahasan

Dalam Tugas Akhir ini, hal-hal yang diharapkan untuk dicapai adalah sebagai berikut :

1. Menganalisis performansi dari algoritma C4.5, CART, dan HDDT terhadap permasalahan kelas *imbalance* yang diterapkan pada beberapa dataset dengan tingkat *imbalance* yang berbeda.
2. Menganalisis dan mengetahui performansi mana yang lebih baik dari ketiga algoritma *decision tree* tersebut pada dataset yang bersifat *imbalance* dilihat dari nilai *precision, recall,* dan *f-measure*. Semakin besar nilai *precision, recall,* dan *f-measure* dari suatu algoritma, semakin baik performansi algoritma tersebut.

1.5. Metodologi Penyelesaian Masalah

Metode yang akan digunakan untuk menyelesaikan Tugas Akhir ini adalah :

1. Studi literatur

   Berupa pencarian sumber-sumber bacaan yang dapat menunjang topik Tugas Akhir ini. Yaitu dengan mencari referensi dan mendalami seluruh materi yang berhubungan dengan data mining, klasifikasi dengan data *imbalance*, algoritma C4.5, CART, HDDT, *information gain, gini index,* dan *hellinger distance measure*.

2. Pengumpulan data-data penunjang Tugas Akhir

   Berupa pengumpulan data penunjang yang dapat membantu perancangan sistem. Data penunjang tersebut akan digunakan untuk pengujian dan analisis, maupun data lain yang membantu terselesainya Tugas Akhir ini.

3. Analisis dan perancangan sistem

   Menganalisis masalah dan perancangan perangkat dengan menggunakan metode yang telah dipilih sebagai batasan masalah. Output dari perangkat lunak yang akan dianalisis, di-*training* dan di-*testing* kemudian dibandingkan satu sama lain sesuai variasi penggunaan algoritmanya.

4. Implementasi

   Berupa realisasi sistem dari rancangan yang dikembangkan. Sistem direalisasikan dengan menggunakan program aplikasi NetBeans IDE 6.0.

5. Evaluasi unjuk kerja sistem

   Melakukan pengujian perangkat lunak dengan memperbandingkan hasil performansi dari setiap algoritma.

6. Penulisan dokumentasi dan laporan

   Mengambil kesimpulan dari analisis hasil yang telah dilakukan, kemudian menyusun laporan dari analisis hasil yang telah dilakukan.