BAB I
PENDAHULUAN

1.1 Latar Belakang Masalah

Dalam beberapa dekade terakhir, teknologi informasi dan basis data telah berkembang dari sistem pemrosesan file primitif menjadi sistem basis data yang canggih. Namun, seiring dengan berjalannya waktu, jumlah data yang sangat besar yang dikumpulkan dan disimpan dalam gudang penyimpanan data sering kali tidak dipakai oleh para *analist* dalam membuat karena adanya kesulitan dalam mengekstrak informasi dari data yang jumlahnya sangat besar. Akibatnya, yang dibuat hanya berdasarkan intuisi bukan berdasarkan informasi dari data yang ada. Oleh karena itu, pembuat membutuhkan *tool* untuk mengekstrak pengetahuan berharga dari data yang sangat besar. Salah satu *tool* tersebut yaitu data mining yang dapat menganalisis data dan menemukan pola data yang penting untuk pengambilan .

Beberapa *task data mining* yang ada yaitu klasifikasi, regresi, asosiasi, klasterisasi, dan *anomaly detection*. Dalam tugas akhir ini dilakukan penelitian tentang klasifikasi. Klasifikasi merupakan proses data mining bersifat prediksi. Klasifikasi memiliki tujuan akhir membentuk pola sederhana/model berupa kelas dari distribusi data input. Model yang ditemukan dapat berupa aturan *"if-then" decision tree*, formula matematis atau *neural network, genetic algorithm, fuzzy, case-based reasoning, k-nearest neighbor*, dan *bayesian*.

Teknik klasifikasi yang digunakan pada pengerjaan tugas akhir ini adalah teknik *decision tree* algoritma C4.5. Yang menjadi perhatian yaitu akurasi *decision tree* pada algoritma C4.5 yang dihasilkan dari data input. Untuk jumlah data yang sangat banyak, pada algoritma C4.4 sering terjadi overlap, dan terdapat kesulitan dalam mendesain pohon keputusan yang optimal. Proses untuk membentuk *decision tree* pada jumlah *record* data yang banyak dapat diminimalkan dengan pengoptimalan algoritma *Tree Augmented Naive Bayes (TAN)*, sebab dalam pembentukan graf TAN hanya mengeksekusi atribut sebanyak satu kali dan menghasilkan satu level *tree*, berbeda dengan C4.5 yang dapat mengakses atribut beberapa kali dalam pembentukan *tree*. Selain itu, *Tree Augmented Naive Bayes (TAN) classifier* merupakan salah satu tipe *Bayesian Belief Network* (BBN) dan merupakan pengembangan dari Naive *Bayes classifier* yang memiliki node-node yang dapat memiliki keterkaitan satu sama lain, sehingga proses pembentukan graf semakin sederhana.

Tujuan dari tugas akhir ini adalah membandingkan proses pembentukan *decision tree* dengan menggunakan algoritma C4.5 dengan pembentukan *decision tree* menggunakan

gabungan algoritma TAN dan C4.5, dari segi waktu proses pembentukan graf yang dibutuhkan (optimalisasi) dan bentuk graf serta nilai akurasi (ketepatan rule yang didapat) untuk data uji.

1.2 Tujuan Penulisan

Adapun tujuan tugas akhir ini adalah :

a. Merancang dan membangun aplikasi pembentukan graf dengan menggunakan kombinasi antara algpritma C4.5 dengan TAN.
b. Menganalisa tingkat optimalisasi pembentukan graf berdasarkan waktu yang dibutuhkan, bentuk graf yang dihasilkan dan akurasi graf (ketepatan pembentukan *rule*) dengan menggunakan algoritma C4.5 dan kombinasi C4.5 dan TAN.
c. Mengevaluasi graf akhir yang terbentuk dari kedua algoritma yang digunakan.

1.3 Perumusan Masalah

Adapun rumusan masalah dalam tugas akhir ini adalah :

a. Bagaimana waktu yang dibutuhkan untuk pembentukan graf serta graf akhir yang dibentuk oleh algoritma C4.5.
b. Bagaimana waktu yang dibutuhkan untuk pembentukan graf serta graf akhir yang terbentuk dari gabungan algoritma C4.5 dan TAN.
c. Bagaimana pengaruh penggunaan ukuran pengambilan sampel data untuk proses pembentukan graf pada gabungan algoritma TAN dan C4.5 dalam akurasi dan waktu pembentukan graf.

1.4 Batasan Masalah

Dalam penulisan tugas akhir ini, ruang lingkup pembahasan masalah hanya dibatasi pada :

a. Tidak menangani proses pre-processing data.
b. Data latih dan data uji yang akan digunakan adalah data sintetik yang dihasilkan oleh generator atau yang berasal dari UCI *Machine Learning Repository*.
c. Data yang digunakan merupakan data kategorikal.

1.5 Metodologi Penyelesaian Masalah

Metode yang digunakan dalam menyelesaiakan tugas akhir ini adalah :

a. Studi literature
   - Pencarian referensi

Mencari referensi dan sumber-sumber lain yang berhubungan dengan masalah *data mining*, klasifikasi, *bayesian network*, *C4.5, TAN*, dan *CI Test Based Algorithms.*

- Pendalaman materi

  Mempelajari dan memahami materi yang berhubungan dengan tugas akhir ini.

b. Mempelajari konsep tentang *data mining*, klasifikasi C4.5 dan TAN.
c. Melakukan implementasi perancangan perangkat lunak sesuai dengan tujuan yang telah disampaikan.
d. Melakukan pengujian perangkat lunak.
e. Menganalisis hasil klasifikasi berdasarkan akurasi dan waktu pembentukan graf, serta bentuk akhir dari graf klasifikasi.
f. Penyusunan Laporan Tugas Akhir dan pengambilan kesimpulan akhir.