# 1. PENDAHULUAN

## 1.1 Latar belakang masalah

Akurasi dalam algoritma klasifikasi adalah seberapa tepat algotima klasifikasi tersebut berhasil mengklasifikasikan data berdasarkan pembelajaran dari data *training*, akurasi sangat penting dalam pengklasifikasian karena jika salah dalam mengklasifikasikan maka akan sangat merugikan berbagai pihak, misalnya jika SPPK rumah sakit salah dalam mendiagnosa penyakit pasien maka akan berakibat *fatal* jika salah dalam pemberian perawatan atau dosis obat, dan dalam SPPK kredit nasabah bank jika salah memberikan pinjaman kredit ke nasabah maka dapat menimbulkan kredit macet yang dapat memberikan dampak buruk bagi bank karena uang yang akan diputar tidak lancar.

*Ensemble method* adalah metode yang digunakan untuk meningkatkan akurasi algoritma klasifikasi dengan membangun beberapa *classifier* dari data *training* kemudian pada saat klasifikasi metode ini menggunakan *voting/aggregating* dari *classifier-classifer* tersebut[10]. Metode ini dapat meningkatkan akurasi karena *ensemble method* akan membangun beberapa *classifier* yang saling independent, jika *classifier-classifier* itu dikombinasikan hasilnya lebih baik daripada *single classifier*. Salah satu contoh dari e*nsemble method* adalah *bagging*.

*Bagging* ditemukan oleh Breiman(1996) yang merupakan kepanjangan dari ” *boostrap agregating*”[11]. *Bagging* adalah salah satu teknik dari *ensemble method* dengan cara memanipulasi data *training*, data *training* di duplikasi sebanyak *d* kali dengan pengembalian (*sampling with replacement*), yang akan menghasilkan sebanyak *d* data *training* yang baru, kemudian dari data *d training* tersebut akan dibangun *classifier-classifier* yang disebut sebagai bag*ged classifier* [4]. Karena data pada tehnik *bagging* dilakukan *sampling with replacement*, ukuran data *bagging* sama dengan data aslinya, tetapi distribusi data dari tiap data *bagging* berbeda, beberapa data dari data *training* bisa saja muncul beberapa kali atau mungkin tidak muncul sama sekali[4]. Hal inilah yang menjadi kunci kenapa *bagging* bisa meningkatkan akurasi karena dengan *sampling with replacement* dapat memperkecil *variance* dari *dataset*[10]

Ilustrasi kinerja *bagging* dapat dilihat jika seorang pasien ingin melakukan *diagnosis* terhadap penyakitnya, jika dia akan datang ke beberapa dokter daripada hanya kepada satu dokter, dan banyak dokter yang menyatakan pernyataan yang sama atas penyakit yang dia derita maka pasien tersebut akan memilihnya sebagai *diagnosis* yang paling banyak atas penyakit yang dialaminya, begitu juga cara kerja *bagging* jika ada sebuah data baru *bagging* akan melakukan *voting* terhadap data baru tersebut sehingga hasil *voting* terbanyak adalah klasifikasi yang terbaik untuk data baru tersebut.

Pada [10] menyebutkan bahwa *bagging*(*ensemble method*) cocok diterapkan untuk algortima yang memiliki sifat *unstable learning algortims/ unstable classifiers* yang dimana *classifier* sangat sensitif terhadap setiap perubahan yang terjadi di data *training*, jika data *training* berubah maka *classfier* pun akan ikut berubah. Algoritma yang termasuk *unstable classifiers* adalah *desicion tree*, *regression tree*, *artificial neural network* dan *rule-based classifiers[*10]. Dalam tugas akhir kali ini peneliti berfokus kepada algoritma CART dan C4.5

karena CART termasuk kedalam kategori *regression tree* sedangkan C4.5 termasuk *desicion tree*. CART dan C4.5 juga termasuk dalam sepuluh besar algoritma dalam *data mining*[14].

Seperti yang sudah dijelaskan diatas *bagging* melakukan duplikasi terhadap data *training*, tetapi untuk berapa banyak *bag* efisien yang harus dibuat oleh *bagging* ada beberapa versi, [9] merekomendasikan sebanyak 25 atau 50 *bag* sedangkan [5] merekomendasikan sebanyak 50 sampai 100 *bag*, [5] juga mengatakan jika *bag* ditingkatkan sebanyak 100 *bag* maka akurasinyapun tidak lebih besar dari *bag* yang dibawah 100, jadi tingkat akurasi *bagging* ditentukan oleh banyak *bag* yang dibuat.

Pada tugas akhir kali ini akan menganalisis pengaruh *bagging* pada algoritma CART dan C4.5, berapa persen tingkat peningkatan akurasi yang dihasilkan oleh *bagging* dan berapa *bag* yang harus dibuat oleh *bagging* untuk tiap data *training* yang akan diuji. Penelitian akan difokuskan kepada penerapan tehnik *bagging* terhadap kedua algoritma tersebut kemudian akan dianalisis berapa peningkatan akurasi yang diakibatkan oleh *bagging*, lalu berapa *bag* yang dibutuhkan agar hasil *bagging* efektif, karena banyaknya *bag* dapat mempengaruhi waktu proses.

## 1.2 Perumusan Masalah

Permasalahan yang akan diangkat pada tugas akhir kali ini adalah:

1. Bagaimana cara mengimplementasikan *bagging* pada algoritma CART dan C4.5.
2. Bagaimana menganalisis *bagging* dalam meningkatkan akurasi dari algoritma CART dan C4.5.

## 1.3 Batasan Masalah

Dalam tugas akhir kali ini dibatasi oleh beberapa hal, sebagai berikut :

1. Data yang digunakan diambil dari *UCI Repository*.
2. Tipe data yang dimasukkan ke dalam sistem terdiri data numerik, data kontinu, dan data campuran antara kontinu dan numerik. Bukan termasuk *dirty data* dan *dependence structure* (seperti *time series*)[11].
3. *Data preprocessing* dilakukan *manual* diluar program.
4. *Dataset* yang akan diuji sudah di *setting* didalam program.

## 1.4 Tujuan

Tujuan dari Tugas Akhir ini adalah :

1. Membangun perangkat lunak untuk klasifikasi dengan menggunakan teknik *bagging* pada CART dan C4.5.
2. Menganalisis dan membandingkan peningkatan akurasi yang diakibatkan *bagging* pada algoritma CART dan C4.5.
3. Menganalisis performansi *bagging* untuk algotima CART dan C4.5 berdasarkan data set yang diuji, dan berapa banyak *bag* yang digunakan supaya hasil *bagging* efisien.

## 1.5 Metode Penyelesaian Masalah

Metode yang dipakai dalam penyelesaian masalah dalam tugas akhir ini adalah :

1. Studi literatur
   a. Pencarian sumber dan refrensi yang berhubungan dengan *data mining* khususnya algoritma klasifikasi CART dan C4.5.
   b. Pencarian sumber dan referensi yang berhubungan dengan *bagging*.

2. Pengumpulan data
   Pada tahap ini data-data yang akan dipakai dalam sistem sebagian diambil secara *online* dari UCI *Repository*.

3. Perancangan dan pemodelan sistem
   a. Merancang sistem untuk proses algoritma CART dan C4.5
   b. Merancang sistem untuk proses *Bagging.*
   c. Merancang *user interface* dari sistem.

4. Implementasi Sistem
   Mengimplementasikan perangkat lunak sesuai dengan analisis dan desain sistem. Perangkat lunak rencananya akan dibangun dalam bahasa pemograman Java dan database menggunakan MySQL.

5. Pengujian dan Analisis hasil
   Pengujian ini dilakukan dengan menganalisis berapa akurasi yang dihasilkan *bagging* pada algoritma CART dan C4.5 untuk tiap data kemudian akan dicari berapa *bag* yang efisien sehingga akurasi yang dihasilkan maksimal.

6. Pengambilan kesimpulan dan pembuatan laporan
   Menyimpulkan hasil dari penelitian dan menyusun laporan sesuai kaidah penulisan yang benar dan sistematika yang telah ditetapkan oleh institusi sebagai salah satu dokumentasi tugas akhir.