# PEMBENTUKAN POHON KLASIFIKASI BINER DENGAN ALGORITMA CART ( *CLASSIFICATION AND REGRESSION TREES* ) ( STUDI KASUS PENYAKIT DIABETES SUKU PIMA INDIAN )

**SKRIPSI**

**Disusun Oleh :**

**KRISAN APRIAN WIDAGDO**

**J2E 005 233**

**PROGRAM STUDI STATISTIKA**

**JURUSAN MATEMATIKA**

**FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM**

**UNIVERSITAS DIPONEGORO**

**2010**

**DAFTAR ISI**

Halaman

## ABSTRAK

Metode klasifikasi CART (*Classification And Regression Trees*) merupakan metode nonparametrik yang berguna untuk mendapatkan suatu kelompok data yang akurat sebagai penciri dari suatu pengklasifikasian. Metode klasifikasi CART terdiri dari dua metode yaitu metode pohon regresi dan pohon klasifikasi. Jika variabel dependen yang dimiliki bertipe kategorik maka CART menghasilkan pohon klasifikasi (*classification trees*). Sedangkan jika variabel dependen yang dimiliki bertipe kontinu atau numerik maka CART menghasilkan pohon regresi (*regression trees*). Proses pembentukan pohon klasifikasi terbagi menjadi 4 tahapan yaitu pembentukan pohon, pelabelan kelas, proses pemangkasan pohon klasifikasi dan pemilihan pohon klasifikasi optimal. Contoh penerapan metode pohon klasifikasi dipergunakan data penyakit diabetes mellitus suku Pima Indian. Data tersebut diperlakukan dalam tiga kondisi berbeda yaitu proporsi jumlah pembagian data *learning* lebih kecil dari data *testing*, proporsi jumlah pembagian data *learning* sama dengan data *testing* dan proporsi jumlah pembagian data *learning* lebih besar *testing*. Metode klasifikasi menghasilkan ketepatan klasifikasi terbaik pada proporsi jumlah pembagian data *learning* sama dengan data *testing* yaitu sebesar 84.83 %. Kedua kondisi lainnya menghasilkan nilai ketepatan klasifikasi sebesar 81.42% dan 81.75%.



**Kata Kunci** : pohon klasifikasi, CART, diabetes mellitus

# BAB I

# PENDAHULUAN

## 1.1 Latar Belakang

Masalah klasifikasi (pengelompokkan) sering dijumpai pada kehidupan sehari-hari, baik mengenai data sosial, data industri, data kesehatan maupun data perbankan. Masalah tersebut dapat diselesaikan dengan metode klasifikasi. Namun, pada penyelesaian masalah klasifikasi perlu diperhatikan dalam memilih metode klasifikasi yang tepat. Sebagai contoh dalam masalah kesehatan, apabila ingin mengelompokkan pasien yang terkena penyakit diabetes dan tidak terkena diabetes. Jika mengelompokkan pasien yang terkena penyakit diabetes ke dalam kelompok pasien yang tidak terkena penyakit diabetes merupakan kesalahan yang dapat berakibat cukup fatal.

Metode klasifikasi dapat dilakukan dengan pendekatan parametrik dan pendekatan nonparametrik. Dalam pendekatan parametrik terdapat beberapa metode klasifikasi yang sering digunakan antara lain : Analisis Regresi Logistik, Analisis Diskriminan dan Analisis Regresi Probit. Regresi logistik dan regresi probit memiliki kelemahan, yaitu nilai yang dihasilkan model regresi logistik dan probit berupa nilai probabilitas yang dirasa kurang praktis (Webb dan Yohannes, 1999). Pada analisis diskriminan, data diharuskan memenuhi beberapa asumsi yaitu data harus berdistribusi normal multivariat dan matrik kovarian yang sama untuk setiap populasi (Breiman et al, 1984).

Dengan adanya keterbatasan metode klasifikasi parametrik, maka digunakan pendekatan nonparametrik. Karena pendekatan tidak bergantung pada asumsi tertentu

sehingga memberikan fleksibilitas yang lebih besar dalam menganalisa data tetapi tetap mempunyai tingkat akurasi yang tinggi dan mudah dalam penggunaannya. Ada beberapa metode klasifikasi dengan pendekatan nonparametrik yang sering digunakan, salah satunya metode klasifikasi berstruktur pohon yang diperkenalkan oleh Leo Breiman, Jerome H. Friedman, Richard A. Olshen, dan Charles J. Stone. Pada tahun 1984, keempat ilmuwan memperkenalkan metode klasifikasi CART (*Classification And Regression Trees*) yaitu metode pohon regresi dan pohon klasifikasi. Jika variabel dependen yang dimiliki bertipe kategorik maka CART menghasilkan pohon klasifikasi (*classification trees*), sedangkan jika variabel dependen yang dimiliki bertipe kontinu atau numerik maka CART menghasilkan pohon regresi (*regression trees*).

Proses pembentukan pohon klasifikasi (CART) dikenal dengan istilah *binary recursive partition*. Proses disebut *binary* karena setiap *parent node* akan selalu mengalami pemecahan kedalam tepat dua *child node*. Sedangkan *recursive* berarti bahwa proses pemecahan tersebut akan diulang kembali pada setiap *child nodes* sebagai hasil pemecahan terdahulu, sehingga *child nodes* tersebut sekarang menjadi *parent nodes*. Proses pemecahan ini akan terus dilakukan sampai tidak ada kesempatan lagi untuk melakukan pemecahan berikutnya. Istilah *partitioning* berarti bahwa *learning sample* yang dimiliki dipecah kedalam bagian-bagian atau partisi-partisi yang lebih kecil (Lewis, 2000). Beberapa kelebihan metode pohon regresi dan pohon klasifikasi antara lain struktur datanya dapat dilihat secara visual, proses pengklasifikasian lebih mudah dilakukan dengan menelusuri pohon klasifikasi yang dihasilkan, dapat mengeksplorasi struktur data yang kompleks serta bersifat nonparametrik sehingga tidak memerlukan asumsi

tertentu yang sering tidak terpenuhi oleh data.

## 1.2 Perumusan Masalah

Berdasarkan latar belakang tersebut maka digunakan klasifikasi dengan pendekatan nonparametrik yaitu pohon klasifikasi (CART). Permasalahan yang muncul adalah bagaimana cara pembentukan pohon klasifikasi biner dengan algoritma CART.

## 1.3 Pembatasan Masalah

CART ( *Classification And Regression Trees* ) terdiri dari dua metode yang berbeda yaitu pohon klasifikasi dan pohon regresi. Dalam tugas akhir ini pembahasan hanya dilakukan pada pembentukan pohon klasifikasi .

## 1.4 Tujuan

Tujuan penulisan tugas akhir ini adalah :

1. Membentuk pohon klasifikasi biner terbaik untuk melakukan sebuah prediksi.
2. Mengetahui interpretasi dari pohon klasifikasi biner yang terbentuk.
3. Mengaplikasikan pohon klasifikasi biner pada masalah kesehatan (penyakit diabetes).

## 1.5 Sistematika Penulisan

Sistematika penulisan tugas akhir ini adalah sebagai berikut : Bab I merupakan bab pendahuluan yang berisi garis besar permasalahan yang akan dibahas dan diselesaikan sesuai dengan tujuan yang telah dirumuskan. Bab II berisi teori-teori yang mendukung dan mendasari penulisan ini yaitu mengenai konsep atau landasan teori mengenai teknik pohon klasifikasi dengan

menggunakan algoritma CART dan diabetes mellitus. Bab III merupakan bagian utama dari penulisan tugas akhir ini, mengenai aplikasi teknik pohon klasifikasi dengan menggunakan algoritma CART. Pembahasan difokuskan pada prinsip-prinsip kerja dari algoritma CART. Beberapa bagian disertakan dengan contoh dan penerapan terhadap data penyakit diabetes beserta analisisnya. Bab IV berisi kesimpulan secara umum dari keseluruhan penelitian dan saran untuk pengembangan selanjutnya.