# BAB 1. PENDAHULUAN

## 1.1 Latar Belakang

Data mining merupakan bagian dari tahapan proses *Knowledge Discovery in Database* (KDD). Data mining merupakan sebuah proses ekstraksi untuk mendapatkan suatu informasi yang sebelumnya tidak diketahui dari sebuah data. Salah satu cara untuk mendapatkan informasi atau pola dari kumpulan data yang kecil hingga terbesar yaitu menggunakan teknik dalam data mining(Ardiyansyah et al., 2018).Data mining sendiri memiliki fungsi yaitu sebagai deskripsi, prediksi, pengelompokan, assosiasi, peramalan, pengurutan dan klasifikasi. Rangkaian proses data mining dibagi menjadi beberapa tahapan yaitu pembersihan data, integrasi data, seleksi data, transformasi data, penambangan data, evaluasi pola dan presentasi pengetahuan(Harlina, 2018).

Pendekatan komputerisasi dengan menerapkan metode seperti klasifikasi sangat umum dilakukan. Klasifikasi merupakan suatu bentuk dari analisis terhadap data dengan merepresentasikan model dari data yang penting melalui fitur tertentu. Klasifikasi merupakan metode data mining yang dalam pembelajarannya bersifat *supervised learning*. *Supervised learning* merupakan algoritma yang membangkitkan suatu fungsi yang memetakan input ke output yang diinginkan. Dalam klasifikasi dapat dilakukan dengan beberapa algoritma diantaranya yaitu *Decision Tree, Naive Bayes, SVM,K-NN, Random Forest, dan Regresi Logistik*(Hidayah et al., 2019).Dalam klasifikasi suatu metode pembelajaran untuk memprediksi nilai dari sekelompok *attribut* dalam menggambarkan atau membedakan kelas data membutuhkan data *training* untuk menentukan sebuah pola dengan data *training* yang semakin banyak data maka semakin bagus untuk menghasilkan sebuah pola klasifikasi(Ardiyansyah et al., 2018).

*Decision Tree* merupakan salah satu metode klasifikasi pada *data Mining*. *Decision Tree* merupakan salah satu metode klasifikasi yang sering digunakan dalam penelitian seperti kedokteran, astronomi dan biologi molekuler(Sari &

Mahmudy, 2019).Dalam *decision tree* terdapat beberapa jenis algoritma diantaranya adalah ID3, C4.5, dan J48. Algoritma C4.5 merupakan algoritma klasifikasi *decision tree* yang banyak digunakan karena memiliki kelebihan utama dari algoritma lainnya. Dalam *pre-processing* pada *data mining* salah satu yang biasa dilakukan adalah seleksi fitur (*feature selection*). Tujuan seleksi fitur yaitu untuk memilih subset variabel dari masukan yang bisa menggambarkan efisiensi input data dalam mengurangi dampak *noise* atau variabel yang tidak relevan sehingga tetap memberikan hasil prediksi yang baik(Sufarnap, 2018). Seleksi fitur dibagi menjadi tiga teknik yaitu *filtering*, *wrapper* dan *hybrid*(Shailendra, 2017).

Beberapa penelitian terkait dengan topik mengenai seleksi fitur di bidang kesehatan dalam mendiagnosa penyakit dan kesehatan cukup populer. Salah satunya penerapan *decision tree* C4.5 sebagai seleksi fitur. Pada studi kasus ini dapat membuktikan bahwa performansi meningkat ketika dilakukannya seleksi fitur menggunakan algoritma decision tree C4.5 pada dataset diagnosa kanker payudara dari nilai akurasi 97,56% menjadi 99,02%(Riswanto et al., 2019). Penelitian selanjutnya ialah dengan topik analisis *decision tree* dan *chi square* untuk menentukan kriteria siswa dalam memilih program studi. Pada studi kasus ini mengimplementasikan *decision tree* untuk menghasilkan pohon keputusan dan akan disederhanakan melalui uji *independen* dengan menerapkan *chi-square*untuk mendapatkan hasil yang lebih sederhana dan akurat. Pada hasil analisis tersebut terdapat penyederhanaan dari 11 rule atau aturan menjadi 7 rule atau aturan setelah menerapkan *chi-square*. Berdasarkan penjelasan diatas penelitian ini melakukan pembuktian mengenai seleksi fitur menggunakan *decision tree* dan *chi-square* pada dataset(Jollyta et al., 2018). Dengan dilakukan penelitian ini, maka algoritma ini dapat membuktikan bahwa hasil seleksi fitur algoritma C4.5 sesuai dengan*chi-square*.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan di atas maka rumusan masalah pada penelitian ini sebagai berikut :

1. Bagaimana *decision tree* C4.5 mampu mengklasifikasikan dataset yang kecil?
2. Bagaimana perbandingan algoritma C4.5 dengan *Chi Square* dalam seleksi fitur?

## 1.3 Batasan Masalah

Adapun batasan masalah dari penelitian ini sebagai berikut :

1. Penelitian ini berfokus pada *Decision Tree* C4.5 dalam seleksi fitur dengan *Chi square*.
2. Data penderita penyakit kanker *serviks* yang dapat diolah adalah dataset yang didapatkan dari UCI *Machine Learning Repository* (https://archive.ics.uci.edu/ml/datasets/Cervical+Cancer+Behavior+Risk).

## 1.4 Tujuan Penelitian

Tujuan dari penelitian ini sebagai berikut :

1. Membuktikan bahwa *decision tree* merupakan metode klasifikasi yang sekaligus bisa menyeleksi fitur yang tidak berkaitan.
2. Membuktikan bahwa dengan adanya seleksi fitur menggunakan *decision tree* dengan algoritma C4.5 dapat dikomparasikan dengan *Chi Square.*

## 1.5 Manfaat Penelitian

Manfaat yang diperoleh dari penelitian ini sebagai berikut :

1. Mengetahui kemampuan *decision tree* C4.5 untuk klasifikasi data training yang kecil (<100).
2. Mengetahui bahwa seleksi fitur menggunakan *decision tree* C4.5 dapat dikomparasikan dengan *chi square*.