

## BAB 1. PENDAHULUAN

### 1.1 Latar Belakang

*Decision Tree* merupakan salah satu metode klasifikasi yang paling sering digunakan dalam penelitian seperti kedokteran, astronomi dan biologi molekuler (Sari & Mahmudy, 2019). *Decision tree* merupakan representasi sederhana dari teknik klasifikasi. Dalam teknik klasifikasi biasanya dataset yang diberikan dibagi menjadi data *training* dan data *testing*. Data *training* digunakan untuk membangun model klasifikasi dan data *testing* digunakan untuk memvalidasi hasil klasifikasi dari data *training*. Dalam *decision tree* terdapat beberapa jenis algoritma diantaranya adalah ID3, C4.5 dan J48. *Decision tree* C4.5 memiliki keunggulan seperti mampu menangani *continues attribute*, *missing data*, membangkitkan *rule* dari sebuah pohon, dan pemangkasan pada saat pembangunan pohon (*tree*) atau pada saat proses pembangunan pohon selesai (Muzakir & Wulandari, 2016). Pada proses klasifikasi permasalahan yang terjadi dalam sebuah penelitian adalah hilangnya nilai dari suatu atribut disebut *missing value*. *Missing value* merupakan hilangnya nilai pada dataset atau ketiadaan nilai pada atribut tertentu. Penyebab terjadinya *missing value* adalah kesalahan manusia seperti kelalain dalam pengumpulan data, kesalahan pada saat entri data, dan ketidakmampuan responden dalam memberikan jawaban yang akurat (Mukarromah dkk., 2015). *Missing value* dapat menyebabkan tingkat keakuratan suatu data kurang valid dan cenderung bias. Salah satu solusi yang dilakukan pada kasus ini ialah dengan menggunakan metode yang tepat untuk melengkapi data dengan nilai lain pada kelompok data tersebut (Rizaldi dkk., 2019).

Dalam sepuluh tahun terakhir, penelitian dengan topik menangani *missing value* di bidang kesehatan dalam mendiagnosa penyakit dan pertanian cukup populer. Salah satunya sistem penanganan data *missing value* pada kualitas produksi jagung dengan metode *K-NN Imputation* pada algoritma C4.5. Sistem ini dapat membantu para petani dalam mengetahui kualitas produksi jagung. Hasil akurasi dari penanganan data *missing value* sebesar 94,50% dan akurasi tanpa penanganan *missing value* sebesar 92,90% (Lutfi & Hasyim, 2019). Penelitian

selanjutnya ialah teknik imputasi *missing values* pada data mining pada data hepatitis. Penelitian ini dilakukan untuk mengetahui metode yang paling cocok dalam melakukan imputasi. Tingkat akurasi metode mean dan modus sebesar 84,615%, metode *K-Nearest Neighbor Imputation* sebesar 82,051% dan metode *Singular Value Decomposition Imputation* sebesar 79,487% (Apriliawan, 2015). Pada penelitian prediksi penyakit ginjal kronis menggunakan algoritma *Naive Bayes Classifier* berbasis *Particle Swarm Optimization*. Hasil dari penelitian ini didapatkan hasil akurasi *confusion matrix* sebesar 97.00% dan AUC sebesar 99.8% setelah ditambahkan *Particle Swarm Optimization* hasil akurasi dari *Naive Bayes Classification* meningkat menjadi 98,75% dan AUC sebesar 99% (Arifin & Ariesta, 2019).

Berdasarkan penjelasan diatas penelitian ini melakukan uji coba pengembangan *rule decision tree C4.5* pada dataset yang tidak lengkap dan dataset yang telah diprediksi *missing imputation*. Dengan dilakukan penelitian ini, maka akan diketahui perbedaan *rule* yang dihasilkan dari data yang lengkap dan data yang sudah diprediksi *missing imputation* serta nilai akurasi yang tertinggi pada data input yang tidak lengkap dan data yang sudah diprediksi *missing imputation*.

## 1.2 Rumusan Masalah

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

1. Bagaimana prediksi *missing value* pada data masukan yang tidak lengkap?
2. Bagaimana pengembangan *rule decision tree C4.5* pada data masukan yang tidak lengkap dan data yang telah diprediksi?
3. Bagaimana hasil perbandingan akurasi berdasarkan metode *Decision Tree C4.5* dengan data masukan yang tidak lengkap dan data yang telah diprediksi?

## 1.3 Tujuan Penelitian

Tujuan dari penelitian ini sebagai berikut :

1. Membuat prediksi *missing value* pada data masukan yang tidak lengkap.

2. Mengetahui pengembangan *rule* metode *Decision Tree C4.5* pada data masukan yang tidak lengkap dan data yang telah diprediksi.
3. Mengetahui hasil perbandingan akurasi berdasarkan metode *Decision Tree C4.5* dengan data masukan yang tidak lengkap dan data yang telah diprediksi.

#### **1.4 Manfaat Penelitian**

Manfaat yang diperoleh dalam melakukan penelitian ini sebagai berikut:

1. Mengetahui prediksi *missing value* pada data masukan yang tidak lengkap.
2. Memberi kemudahan bagi pengguna dalam mendiagnosa penyakit ginjal kronis.
3. Mengetahui performa dari metode *Decision Tree C4.5* dengan data masukan yang tidak lengkap dan data yang telah diprediksi

#### **1.5 Batasan Masalah**

Adapun batasan masalah dari penelitian ini adalah :

1. Data yang digunakan dalam penelitian ini adalah data penyakit ginjal kronis.
2. Pembangunan pohon keputusan tidak dilakukan proses *prunning*.
3. Pemilihan atribut dalam penelitian ini tidak menggunakan proses seleksi fitur.