BAB I
PENDAHULUAN

1.1. Latar Belakang

Kesehatan merupakan salah satu hal yang sangat penting dalam kehidupan manusia sehingga perlu dijaga dengan baik karena dengan tubuh yang sehat maka aktivitas dapat berjalan dengan baik [1]. Berbagai upaya yang dilakukan manusia dalam menjaga kesehatan seperti mengatur pola makan, rutin berolahraga . Namun, berbagai penyakit dapat muncul sehingga menghambat aktivitas manusia. Salah satu penyakit yang banyak diderita adalah batu ginjal yaitu sebanyak 1.499.400 manusia di Indonesia yang menderita penyakit batu ginjal [2].

Batu ginjal merupakan materi keras seperti batu yang terbentuk dari urin yang berkonsentrasi [3] dengan garam dan mineral [2]. Batu ginjal dapat menyebabkan infeksi berulang, gangguan ginjal dan juga hematuria [4] yang apabila tidak ditangani sedini mungkin dapat menyebabkan kerusakan ginjal seperti gagal ginjal bahkan dapat menyebabkan kanker ginjal [5]. Gejala yang dialami oleh penderita batu ginjal adalah rasa sangat nyeri pada bagian pinggul, muntah secara terus menerus dan berdarah saat berkemih yang disebabkan oleh batu ginjal yang bergerak di dalam saluran ureter melalui saluran urin [3]. Penyakit ini dapat diatasi dengan baik apabila dapat dideteksi sedini mungkin untuk mencegah komplikasi seperti infeksi dan hematuria. Deteksi dini dapat dilakukan dengan memanfaatkan teknik *data mining* untuk mengolah data klinis dengan menemukan pola tersembunyi dan informasi yang relevan dengan diagnosis penyakit batu ginjal [6].

Data mining atau penambangan data merupakan suatu teknik yang digunakan untuk mencari informasi dari suatu kumpulan data yang besar [6]. Data mining akan melakukan sebuah proses ekstrasi dengan memanfaatkan teknik kecerdasan buatan, statistika dan juga teknik matematika dan *machine learning*. Setelah proses ekstrasi data dilakukan, informasi akan didapatkan dari proses data mining tersebut, informasi ini akan membantu dalam prediksi dalam banyak bidang. *Data mining* memiliki fungsi sebagai *clustering, association,* dan *classification* [7].

Classification atau klasifikasi merupakan salah satu teknik dalam *data mining* yang merupakan proses analisis data input dengan model klasifikasi berdasarkan fitur dalam dataset. Klasifikasi memungkinkan memberikan label pada data baru yang belum diketahui labelnya dengan memahami karakteristik kelas [1][8]. Beberapa model klasifikasi pada *data mining* diantaranya adalah *support vector machine, decision tree, random forest, naïve bayes, neural network* [9].

Metode klasifikasi yang umum digunakan untuk klasifikasi penyakit adalah *decision tree* atau pohon keputusan yang merupakan salah satu metode untuk membuat pohon keputusan berdasarkan data pelatihan untuk pengklasifikasian dan prediksi yang mudah dipahami aturannya [1][8]. Dalam pohon keputusan, pembagian himpunan data besar akan dilakukan menjadi himpunan terkecil dalam proses klasifikasi atau prediksi [8]. Pembagian himpunan data besar didasarkan pada nilai *gain* terbesar yang dihasilkan variabel. Nilai *gain* yang terbesar akan digunakan model pohon keputusan sebagai akar pohon [7].

Salah satu algoritma dalam pohon keputusan yang sangat popular digunakan adalah algoritma C4.5 [8]. Algoritma C4.5 memiliki sampel pelatihan yang akan digunakan untuk membangun sebuah pohon dan kebenarannya telah diuji [9]. Untuk membangun sebuah algoritma C4.5 diperlukan atribut sebagai akar. Beberapa kelebihan algoritma ini adalah pohon keputusan yang mudah untuk diinterpretasikan, efisien dalam penanganan data yang bertipe numerik dan diskrit. Kekurangan algoritma pohon keputusan adalah rentan terhadap *overfitting* apabila data memiliki fitur yang terlalu banyak [10] dapat ditangani dengan melakukan pemangkasan pohon keputusan atau disebut dengan *pruning* di algoritma C4.5.

Pada proses klasifikasi, sebuah model yang dilatih menggunakan data *training* tertentu dapat menunjukkan performa yang sangat baik secara kebetulan, namuna model tersebut belum tentu mampu bekerja dengan baik pada data yang belum pernah dilihat sebelumnya. Oleh karena itu, diperlukan proses validasi model untuk memastikan bahwa performa yang dihasilkan model benar-benar menunjukkan kemampuan model secara umum. Beberapa teknik validasi model yang umum digunakan dalam pembelajaran mesin adalah *cross-validation, hold-out validation, dan bootstrapping*. Salah satu teknik yang paling populer adalah

cross-validation yang bekerja dengan membagi data ke dalam beberapa subset untuk kemudian digunakan secara bergiliran sebagai data latih dan data uji. Teknik ini membantu memastikan bahwa hasil yang diperoleh benar-benar mewakili pola dari data yang sesungguhnya, serta dapat meminimalkan risiko kesalahan interpretasi atau pengambilan kesimpulan yang tidak akurat terhadap data.

Penelitian sebelumnya [1] menggunakan algoritma *Decision Tree* dan menerapkan metode C4.5 dengan membandingkan hasil akurasi, namun belum menerapkan teknik evaluasi yang bertujuan untuk mengidentifikasi adanya *overfitting* atau *underfitting*. Teknik seperti *cross-validation* belum digunakan untuk memastikan bahwa model yang dihasilkan memiliki kemampuan generalisasi yang baik terhadap data baru. Selain itu, penelitian tersebut menggunakan dataset dari RSUD Tangerang dengan 18 variabel, yang memiliki karakteristik dan struktur data berbeda dibandingkan dengan dataset yang digunakan dalam penelitian ini. Perbedaan tersebut mencakup jumlah variabel, sumber data, serta definisi dari masing-masing atribut yang digunakan sebagai basis klasifikasi. Penelitian ini menggunakan data dari RSUD Cideres dengan 7 variabel utama.

1.2. Rumusan Masalah

Penerapan algoritma *Decision Tree* C4.5 pada klasifikasi penyakit batu ginjal memiliki akurasi yang tinggi, namun penerapan metode ini pada sumber data yang berbeda dapat menimbulkan sebuah perbedaan penanganan data karena adanya karakteristik yang berbeda antara dataset dan perbedaan distribusi variabel. Penelitian ini berupaya untuk menganalisis efektivitas metode *Decision Tree* C4.5 pada dataset yang berasal dari sumber yang berbeda.

1.3. Tujuan Penelitian

Tujuan penelitian ini adalah:

1. Menerapkan algoritma *Decision Tree* C4.5 untuk mengklasifikasikan penyakit batu ginjal berdasarkan data klinis.
2. Mengukur akurasi algoritma *Decision tree* C4.5 dalam melakukan prediksi pada penyakit batu ginjal.
3. Menemukan variabel yang memiliki pengaruh paling signifikan.

1.4. Manfaat Penelitian

Menerapkan algoritma *Decision Tree* C4.5 dalam klasifikasi penyakit batu ginjal diharapkan mampu menghasilkan sebuah diagnosis dengan melakukan prediksi berdasarkan data prediktor yang ada dalam dataset. Hal ini diharapkan mampu mempermudah dalam mendeteksi dini penyakit batu ginjal dalam mengambil diagnosis.