BAB I

PENDAHULUAN

1.1 LATAR BELAKANG MASALAH

Data mining merupakan proses menggali informasi dan data yang signifikan dari sejumlah besar data dengan memanfaatkan ilmu *machine learning* dan ilmu manajemen basis data, yang kemudian dianalisis untuk mengidentifikasi pola dan hubungan sehingga menghasilkan informasi baru yang berharga tersimpan di *database* dalam konteks pengambilan keputusan dengan teknik tertentu[1], [2], [3], [4]. Tujuan dari proses *data mining* sendiri adalah menemukan pola atau model data yang sebelumnya tidak diketahui, sehingga model tersebut dapat digunakan untuk mengambil keputusan[3]. Terdapat beberapa tugas dalam *data mining* yakni deskirpsi, estimasi, prediksi, klasifikasi, pengklasteran, dan asosiasi[5]. Selain itu dalam proses mencari pola atau model yang tersembunyi tersebut *data mining* menggunakan algoritma[6]. Hingga saat ini *data mining* sendiri telah digunakan di berbagai sektor bidang seperti industri keuangan, telekomunikasi, manajemen bisnis, sains serta kedokteran[1]. Melihat potensi dari data mining sendiri yang telah di terapkan di berbagai bidang termasuk kedokteran, sehingga pemanfaatan *data mining* ini patut dicoba dalam mencari metode alternatif untuk memprediksi penyakit stroke otak.

Klasifikasi sendiri merupakan salah satu metode atau teknik *data mining* dengan mengelompokan berdasarkan jumlah dan nama kelompoknya dalam mengambil keputusan[7]. klasifikasi termasuk dalam *supervised learning,* yakni

proses mencari algoritma yang berdasarkan pada contoh-contoh yang diberikan dari luar, untuk membentuk hipotesis yang bersifat umum, yang nantinya dapat digunakan untuk membuat prediksi tentang contoh-contoh di masa mendatang[8]. Dalam klasifikasi sendiri memiliki beberapa algoritma seperti *decision tree, naïve bayes, support vector machine, neural network* dan *K-NN* [7][8]. Dengan menggunakan algoritma tersebut penyebab-penyebab dari stroke otak nantinya akan diklasifikasikan sehingga dapat digunakan untuk memprediksi stroke otak..

Decision tree adalah algoritma umum pada klasifikasi prediksi yang terkenal, sangat kuat, dan ampuh [9], [10], algoritma ini membentuk model keputusan dalam bentuk pohon, terdiri dari root *node* yang merupakan akar pohon utama tanpa *input*, internal *node* yang memiliki *input* dan *output*, dan leaf *node* yang menjadi *output*[7]. Algoritma *Iterative Dichotomizer* 3(ID3), *Classification* 4.5 (C4.5), dan *Classification* 5.0 (C5.0) yang dikembangkan oleh J.Ross Quinlan merupakan algoritma yanng umum digunakan untuk membentuk *decision tree*[3], [11]. C5.0 adalah peningkatan dan modifikasi dari algoritma sebelumnya, yaitu ID3 dan C4.5 selain itu ini juga merupakan versi terakhir dari algoritma pembentuk decision tree yang dikembangkan oleh J.Ross Quinland[3], [11]. C5.0 salah satu algoritma klasifikasi yang efektif dalam mengolah data dengan atribut numerik maupun kategorikal[12]. Proses pembuatan pohon keputusan pada C5.0 memiliki kemiripan dengan C4.5, khususnya dalam menghitung *entropy* dan *information gain*. Tetapi, C5.0 memiliki tahap tambahan yaitu perhitungan *gain ratio*[12]. Pada penelitian ini C5.0 akan digunakan untuk membuat model yang nantinya digunakan untuk memprediksi stroke otak.

Secara umum stroke otak merupakan merupakan penyakit *serebrovaskular* akut, dimana pembuluh darah yang memasuk darah ke suatu bagian otak tersumbat atau pecah sehingga otak kehilangan sumber nutrisi, glukosa, dan oksigen yang dapat mengakibatkan gangguan fungsi otak[13], [14], [15], [16]. Menurut *World Stroke Organization* (WSO) stroke merupakan penyebab kematian kedua terbanyak secara global yang dimana terdapat 17 juta pasien stroke diseluruh dunian dan rata-rata 6.5 juta orang meninggal tiap tahunnya[1]. Hal ini juga sejalan dengan *World Health Organization* (WHO) dimana stroke merupakan penyakit penyebab kematian nomor dua di dunia dan bertanggung jawab sekitar 11% dari seluruh kematian[17]. Selain kematian stroke juga dapat menyebabkan kecacatan jangka panjang[1], [13], [14]. Dari dampak yang ditimbulkan dari stroke otak ini sudah seharusnya kita tidak menganggap remeh penyakit ini, sehingga algoritma *decision tree* C5.0 akan digunakan sebagai metode alternatif dalam memprediksi stroke otak.

Terdapat pula beberapa penelitian yang dilakukan peneliti sebelumnya mengenai penyakit stroke, seperti pada penelitian yang dilakukan oleh Fazrin Meila Azzahra Sofyan dkk [12], membuktikan hasil akurasi yang diperoleh model algoritma *decision tree* C5.0 sebesar 95% dengan melakukan *split data* 80% (*data training*) dan 20% (*data testing*). Adapun penelitian yang dilakukan oleh Iqram Hussain dan Se-Jin Park [18], dengan merekam *Electroencephalography* (EEG) pada empat puluh delapan pasien stroke dan tujuh puluh lima orang dewasa sehat, model C5.0 menunjukkan akurasi 78% untuk keadaan istirahat, akurasi 89% dalam kondisi berjalan, akurasi 84% dalam kondisi kerja, dan akurasi 85% dalam keadaan membaca kognitif. Dari beberapa penelitian diatas dengan menggunakan algoritma

decision tree C5.0 dimana menghasilkan akurasi yang tergolong baik dalam memprediksi stroke.

Dataset yang nantinya digunakan pada penelitian ini untuk membuat model memprediksi stroke otak diambil dari repositori *kaggle* yang dibuat oleh jillani soft tech bernama *Brain Stroke Dataset*. *Dataset* ini berisi data dari 4981 pasien stroke secara umum. Data tersebut dikumpulkan dari berbagai sumber, termasuk rumah sakit, klinik, dan penelitian. Data terdiri dari 11 atribut, dengan atribut *gender, age, hypertension, heart_disease, ever_married, work_type, residence_type, avg_glucose_level, bmi* dan *smoking_status* merupakan *feature* dan *stroke* sebagai label dengan 0 jika tidak stroke dan 1 jika stroke. Selain itu pada *dataset* ini telah ada sedikit proses *preprocessing* sebelumnya.

Dengan demikian berhubungan permasalahan maka akan sangat dibutuhkan informasi dalam upaya penanganan dini penyakit stroke otak, sehingga penulis memutuskan melakukan penelitian yang berjudul **"Penerapan Algoritma *Decision Tree* C5.0 Untuk Memeprediksi Penyakit Stroke Otak"** yang diharapkan menjadi referensi dalam mengambil keputusan untuk memprediksi penyakit stroke otak**.** Dimana penelitian ini stroke otak yang dimaksud merupakan stroke otak secara umum.

1.2 RUMUSAN MASALAH

Dari latar belakang tersebut dapat diuiraikanlah rumusan masalah dalam penelitian ini, yakni:

1. Bagaimana menerapkan algoritma *decision tree* C5.0 dalam *Brain Stroke Dataset* untuk prediksi stroke otak?
2. Seberapa Baik akurasi, *recall,* dan *precision* model pohon keputusan yang dihasilkan dari algoritma *decision tree* C5.0 dengan menggunakan *dataset Brain Stroke Dataset* dalam mengklasifikasikan penyakit stroke otak?

1.3 BATASAN MASALAH

Berikut merupakan batasan masalah dalam penelitian supaya pembahasa tidak meluas serta lebih terarah:

1. Hanya terbatas pada penyakit stroke secara umum
2. Menggunakan *dataset* yang diperoleh dari repositori *kaggle* pada bidang kesehatan mengenai penyakit stroke otak bernama *Brain Stroke Dataset*.
3. Menggunakan 11 atribut: *Gender, Age, Hypertension, Heart Disease, Ever Maried, Work Type, Residence Type, AVG Glucose Level, BMI, Smoking Status* sebagai *feature* dan *stroke* sebagai label dengan nilai 0 jika stroke dan nilai 1 jika tidak stroke.
4. Akurasi, *recall,* dan *precision* dari model yang telah dibuat diukur menggunakan *confusion matrix*.
5. Analisis dilakukan dengan aplikasi *RapidMiner Studio*.

1.4 TUJUAN DAN MANFAAT PENELITIAN

1.4.1 Tujuan Penelitian

Berikut tujuan dari penelitian ini:

1. Menerapkan algoritma *decision tree* C5.0 untuk memprediksi penyakit stroke otak.
2. Untuk mengukur seberapa baik algoritma *decision tree* C5.0 dalam memprediksi penyakit stroke otak.

1.4.2 Manfaat Penelitian

Berikut beberapa manfaat yang diperoleh dari penelitian ini:

1. Dapat memberikan kontribusi pencegahan dini terhadap penyakit stroke otak dengan informasi yang diberikan oleh model berkaitan tentang faktor-faktor penyebab penyakit stroke otak.
2. Dapat mengetahui nilai akurasi, *recall*, dan *precision* dari algoritma *decision tree* C5.0 dalam memprediksi penyakit stroke otak.
3. Dapat membantu dalam mendukung pengambilan keputusan untuk memprediksi penyakit stroke otak.
4. Dapat memberikan kontribusi dalam pengembangan metode untuk memprediksi stroke otak.

1.5 SISTEMATIKA PENULISAN

Untuk memberikan gambaran keseluruhan tentang penulisan ilmiah, dapat dilihat sistematika penulisan yang meliputi:

BAB I : PENDAHULUAN

Pada bab ini berisi tentan uraian dari latar belakang masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan.

BAB II : LANDASAN TEORI

Pada bab ini akan membahas tentang teori-teori serta pendapat dari para ahli yang berhubungan dengan pembahasan yang sedang dianalisis. Untuk teori-teori yang dipergunakan antara lain : *Data mining*, *decision tree, C5.0* dan Stroke otak

BAB III : METODOLOGI PENELITIAN

Pada bab ini akan membahas tentang penjelasan kerangka kerja penelitian, metode pengumpulan data, proses yang dilakukan, prosedur penelitian dan alat bantu(*tools)* yang digunakan sebagai pendukung penelitian.

BAB IV : ANALISIS

Pada bab ini akan membahas perhitungan, hasil, visualisasi, serta analisis permasalahan penelitian yang berupa pengujian mengenai algoritma yang dipergunakan dalam klasifikasi penyakit stroke otak.

BAB V : PENUTUP

Pada bab ini akan menyajikan rangkaian poin-poin dari penelitian serta saran-saran yang nantinya akan ditujukan kepada semua pihak yang bersangkutan.