BAB I

PENDAHULUAN

1.1 LATAR BELAKANG MASALAH

Data mining atau juga sering disebut *Knowledge Discovery in Database* (KDD), adalah teknik pengambilan keputusan di masa depan berdasarkan informasi yang diperoleh dari masa lalu untuk mendapatkan prediksi, estimasi, pengelompokan, dan deskripsi [1]. Tahapan yang dilakukan pada proses *data mining* yaitu dari seleksi data, tahap *pre-processing*, transformasi, *data mining* serta tahap evaluasi yang menghasilkan pengetahuan baru [2]. Perkembangan zaman pada bidang teknologi dapat membantu banyak hal, salah satu contoh nya dapat membantu bidang Kesehatan [3]. Sehingga *data mining* dapat digunakan untuk membantu bidang Kesehatan, terkhususnya pada penyakit stroke.

Klasifikasi merupakan suatu proses dalam menemukan model yang menjelaskam, membedakan konsep atau kelas data, yang bertujuan memprediksi kelas dari suatu objek yang labelnya sudah diketahui, Dimana dalam klasifikasi, memiliki kategori label dari atribut-atribut yang ada, Algoritma klasifikasi antara lain adalah *Decision Tree, Naïve Bayes, Neural Network, Genetic Algorithm, Fuzzy, Case-Based Reasoning, Random Forest, dan K-Nearest Neighbor* [4], [5]. Dengan mengunakan salah satu algoritma dari metode klasifikasi, penelitan ini akan menghasilkan model yang dapat digunakan untuk memprediksi.

Random Forest merupakan teknik komputasi efisien yang dapat beroperasi dengan cepat pada *Dataset* besar. *Random Forest* menggunakan pengacakan untuk membuat sejumlah besar pohon keputusan. *Output* dari pohon-pohon ini dikumpulkan menjadi *output* tunggal menggunakan *voting* untuk masalah klasifikasi atau rata-rata untuk masalah regresi [6], [7]. Algoritma ini efektif dengan data yang besar sehinga mendukung untuk mengklasifikasi jumlah data yang besar seperti yang digunakan dalam penelitian ini.

Stroke merupakan penyakit yang disebabkan oleh penyumbatan atau pecahnya pembuluh darah di otak sehingga suplai darah ke suatu area otak tiba-tiba terganggu [8], [9]. Pasien stroke otak dapat digolongkan sebagai masalah serius, Hal ini tergambar dari kematian akibat stroke di Indonesia yang mencapai angka sebanyak 131,8 kasus. sehingga prediksi tepat waktu dan akurat dapat meningkatkan peluang pulihnya pasien atau juga dapat mengurangi resiko kematian yang di akibatkan stroke [10], [11], [12]. Maka dari itu penelitian ini akan mengunakan algoritma *Random Forest* untuk mengklasifikasikan *Barain stroke Dataset*.

Dataset penting untuk digunakan dalam klasifikasi *data mining* karena dapat membantu peneliti untuk mengembangkan model yang dapat memprediksi risiko stroke pada pasien. *Dataset* stroke yang tersedia di *repository* Kaggle adalah *Dataset* yang baik untuk digunakan dalam klasifikasi *data mining*. *Dataset* yang dimiliki oleh Jillani Soft Tech dengan judul *Brain stroke dataset* ini mencakup data yang terdiri dari 11 fitur, yaitu *gender*, *age*, *hypertension*, *heart_disease*, *ever_married*, *work_type*, *Residence_type*, *avg_glucose_level*, *bmi*,

smoking_status, *stroke*. *Dataset* ini juga memiliki jumlah data yang cukup besar, yaitu 4981 data.

Terdapat penelitian sebelumnya mengenai klasifikasi penyakit stroke seperti pada penelitian yang dilakukan oleh Agus Byna, dan Muhammad Basit [13] mengunakan algoritma *Naïve bayes* yang di optimalisasi dengan *Adaboost* mendapatkan hasil akurasi sebesar 89.65% menggunakan bahasa pemrograman *Python 3.9*. Selanjutnya penelitian yang di lakukan oleh Taufk Djatna dkk. [14] menggunakan algoritma *Decision Tree* yang berbasis *Fuzzy* dan menunjukkan akurasi sebesar 90,59%. Dan juga penelitian yang dilakukan oleh Muhammad Firdaus Banjar dkk. [15] yang megunakan algoritma *Random Forest* dengan 4 jumlah pohon keputusan yang berbeda yaitu 50 dengan akurasi sebesar 86,49%, 100 sebesar 86,82%, 200 sebesar 86,30%, dan 500 mendapatkan akurasi sebesar 86,49%, sehingga kinerja terbaik pada jumlah pohon keputusan sebanyak 100.

Dengan pemaparan latar belakang diatas, maka penulis melakukan penelitian dengan judul **"Penerapan Algoritma *Random Forest* Untuk Klasifikasi Penyakit Stroke"**

1.2 PERUMUSAN MASALAH

Berdasarkan latar belakang yang telah diuraikan, maka perumusan masalah yang dapat diambil dalam penelitian ini adalah:

1. Bagaimana menerapkan algoritma *Random Forest* dalam mengklasifikasi penyakit stroke pada *Brain Stroke Dataset*?

2. Seberapa akurat algoritma *Random Forest* dalam mengklasifikasi penyakit stroke pada *Brain Stroke Dataset*?

1.3 BATASAN MASALAH

Agar pembahasan tidak keluar dari pembahasan dan tujuan penulis, maka di perlukan batasan masaslah, yaitu:

1. Data yang di gunakan pada penelitian ini merupakan *Dataset* publik yang di ambil dari *repository* kaggle yang diberi judul *Brain Stroke Dataset.*
2. Penelitian ini yang hanya berfokus pada atribut yang akan di gunakan yaitu *gender*, *age, hypertension, heart_disease, ever_married, work_type, Residence_type, avg_glucose_level, bmi, smoking_status,* dan *stroke.*
3. Pengukuran dilakukan hanya pada algoritma *data mining* yaitu *Random Forest* dengan memperhatikan nilai akurasi.
4. Aplikasi yang digunakan pada penelitian ini berupa *software* yaitu *RapidMiner*.

1.4 TUJUAN DAN MANFAAT PENELITIAN

1.4.1 Tujuan Penelitian

Berdasarkan rumusan masalah di atas maka didapatkan tujuan yang ingin dicapai dari penelitian ini yaitu:

1. Untuk menerapkan algoritma *Random Forest* dalam mengklasifikasi penyakit stroke pada *Brain Stroke Dataset*.

2. Untuk mengukur tingkat akurasi algoritma *Random Forest* dalam mengklasifikasi penyakit stroke pada *Brain Stroke Dataset*.

1.4.2 Manfaat Penelitian

Beberapa manfaat yang dapat diperoleh dari penelitian ini yaitu:

1. Penelitian ini dapat menghasilkan model yang berguna untuk pengambilan keputusan atau prediksi.
2. Penelitian ini dapat mengetahui seberapa akurasi dari model yang dihasilkan algoritma *Random Forest.*
3. Penelitian ini dapat memberikan rekomendasi model bagi bidang kesehatan dalam memperediksi penyakit stroke.
4. Dari hasil penelitian ini dapat menembah wawasan dan informasi baru bagi penulis dan pembaca.

1.5 SISTEMATIKA PENULISAN

Gambaran umum mengenai keseluruhan penulisan ilmiah, dapat di lihat melalui sistematika penulisan berikut ini:

BAB I : PENDAHULUAN

Pada bab ini menguraikan tentang latar belakang maslah, Batasan masalah, tujuan penelitian,manfaat penelitian,serta sistematika penulisan.

BAB II : LANDASAN TEORI

Pada baba ini berisikan landasan teori yang akan membahas teori-teori dan pendapat para ahli yang berhubungan dengan pembahasan yang di analisis teori teori yang di gunakan antara lain *Data mining*, *Naïve Bayes*, *Decision Tree*, C4.5, *RapidMiner*, dan Strok otak.

BAB III : METODOLOGI PENELITIAN

Dalam bab ini menjelaskan tentang kerangka kerja penelitian, metode pengumpulan data, proses yang dilakukan, prosedur penelitian, dan alat bantu (tools) yang digunakan untuk mendukung penelitian

BAB IV : ANALISIS

Bab ini berisikan tentang perhitungan,hasil,visualisasi,serta analisis permasalahan penelitian yang berupa pengujian mengenal algoritma yang digunakan untuk klasifikasi penyakit strok otak.

BAB V : PENUTUP

Pada bab penutup ini berisikan uraian-uraian dari seluruh kegiatan penelitian serta saran-saran yang dapat bisa berguna untuk para pembaca.