

# **BAB I**

## **PENDAHULUAN**

### **1.1 Latar Belakang Masalah**

Stroke atau cerebrovascular accident (CVA) adalah keadaan dimana fungsi otak hilang akibat gangguan mendadak pada suplai darah ke otak. Gangguan ini terjadi karena masalah sirkulasi darah di otak yang menyebabkan kelumpuhan atau kematian.[1] Sel otak mengalami kematian akibat penyumbatan saluran yang menyediakan nutrisi dan oksigen ke otak. [2] Stroke terdiri dari dua jenis yaitu Stroke iskemik merupakan salah satu jenis penyakit stroke yang biasanya disebabkan oleh penyumbatan pembuluh darah. Sedangkan stroke hemoragik adalah jenis stroke yang disebabkan karena terjadinya pendarahan langsung ke otak atau ke dalam ruang antara membran otak. Pendarahan dapat terjadi karena pecahnya aneurisma otak. Keduanya menyebabkan sebagian otak berhenti bekerja. Tanda-tanda dan efek samping stroke mungkin termasuk kegagalan untuk menggerakkan atau merasakan satu sisi tubuh, ketidaknyamanan dalam memahami atau berbicara, rasa pusing, kehilangan penglihatan pada satu sisi atau mengalami nyeri otak yang parah.

Informasi dari World Wellbeing Association (WHO) menunjukkan ada 10 penyakit yang menjadi penyebab kematian terbesar di Indonesia. Stroke menempati urutan pertama, dengan 131,8 kematian per 100.000 orang. Kedua, penyakit jantung koroner iskemik atau penyebab gagal jantung sebanyak 95,68 kasus. Urutan ketiga dengan gap yang cukup besar adalah penyakit diabetes melitus yaitu sebanyak 40,78 kasus. Keempat, tuberkulosis (TB) bertambah sebanyak 33,24 kasus. Dengan 33,06 kasus, sirosis hati berada tepat di bawah persentil kelima. Urutan terakhir adalah kematian bayi baru lahir (neonatal) sebanyak 16,77 kasus.[3] Angka kejadian stroke di Indonesia menurut data Riskedas tahun 2013 adalah 7 untuk setiap 1.000 orang yang terkena stroke. Angka kejadian ini meningkat pada tahun 2018 dengan 10,9 per 1.000 orang menderita stroke. Banyak faktor yang masuk dalam empat layanan prioritas berdasarkan data Kementerian Kesehatan tahun

2022, termasuk fakta bahwa tiga dari setiap 1.000 orang berisiko terkena stroke setiap tahunnya. 15% korban stroke terancam mati dan 65% korban stroke terancam mengalami kecacatan.[4] Permasalahan stroke di Indonesia memerlukan perhatian yang serius mengingat jumlah kasus yang terus meningkat dan angka kematian yang tinggi.[5]

Oleh karena itu, diperlukan upaya untuk menemukan solusi yang dapat memprediksi awal penyakit stroke guna mendukung langkah-langkah pencegahan dan pengobatan sejak dini. Salah satu cara untuk menangani permasalahan tersebut adalah dengan mengumpulkan informasi dari data penyakit stroke, yang nantinya dapat dimanfaatkan untuk mengembangkan model baru yang dapat mengkarakterisasi pasien dan memutuskan apakah mereka berisiko terkena stroke atau tidak. Metode yang melibatkan pengambilan data penting dari indeks informasi yang sangat besar dikenal dengan istilah data mining.

Salah satu metode pengenalan pola dalam data mining melibatkan penerapan machine learning, di mana komputer dapat memahami pola dari data tertentu dan membentuk suatu model. Dalam machine learning, terdapat dua pendekatan pembelajaran utama, yaitu supervised learning dan unsupervised learning. Supervised learning merupakan metode pembelajaran yang diterapkan ketika hasil yang diinginkan atau yang diperkirakan sudah diketahui. Salah satu subkategori dalam supervised learning adalah klasifikasi. Klasifikasi adalah salah satu metode dalam data mining yang digunakan untuk menggali sekumpulan informasi dari data penting. Teknik klasifikasi dapat secara otomatis memprediksi kelas dari data yang tidak terklasifikasi.[6]

Metode klasifikasi pada data mining yang sering digunakan dalam prediksi penyakit yaitu Decision Tree dan Naïve Bayes. Decision Tree merupakan metode klasifikasi yang menghasilkan keputusan dalam bentuk struktur pohon. Algoritma ini membuat model keputusan dengan menggunakan simpul daun sebagai akar pohon yang merupakan keluaran, simpul internal sebagai akar pohon yang mempunyai masukan dan keluaran, dan simpul akar sebagai akar pohon yang diprioritaskan dan tidak mempunyai input. Setiap simpul menyimpan data yang telah diklasifikasikan dengan mempertimbangkan variabel targetnya.[7] Decision

Tree memiliki kelebihan yaitu sifatnya yang fleksibel sehingga mampu meningkatkan kualitas keputusan yang dihasilkan.[8] Keunggulan lain dari algoritma decision tree yaitu kemampuan menangani nilai atribut yang hilang, meningkatkan efisiensi komputasi, dan menyelesaikan masalah overfitting.[7] Namun, beberapa kelemahan dari algoritma ini adalah kemungkinan terjadi tumpang tindih (*overlap*) jika digunakan pada data yang memiliki banyak kelas dan kriteria.[8]

Metode klasifikasi selanjutnya dalam penelitian ini adalah naïve bayes. Naïve Bayes adalah metode klasifikasi probabilitas sederhana yang menghitung sejumlah probabilitas dengan menambahkan frekuensi dan kombinasi nilai dari dataset yang disediakan. Keunggulan algoritma naïve bayes terletak pada kesederhanaan perhitungannya, sehingga memungkinkan proses yang lebih cepat dan efisien. Metode naïve bayes hanya memerlukan jumlah data pelatihan yang relatif kecil untuk menentukan parameter yang diperlukan dalam proses klasifikasi.[9] Namun, metode naïve bayes hanya mendukung pada atribut yang bertipe data discrete atau discretized, dan tidak mendukung atribut yang bernilai continuous (numerik), sehingga semua atribut menjadi independent. Selain itu, atribut tersebut dapat memberikan kontribusi kepada atribut yang diprediksi.[10] Disisi lain naïve bayes memiliki kelemahan dimana masing-masing variabel bersifat independent, hal ini dapat menurunkan nilai akurasi. Dalam mengatasi permasalahan tersebut, beberapa penelitian telah dilakukan dengan menerapkan teknik ensemble untuk meningkatkan akurasi prediksi pada algoritma decision tree dan naïve bayes. Maka dari itu, penggunaan teknik ensemble dalam penelitian ini bertujuan untuk menghasilkan prediksi yang lebih akurat.

Teknik ensemble adalah metode yang dapat meningkatkan kinerja algoritma klasifikasi tunggal dengan menggabungkan beberapa pengklasifikasi yang dibuat berdasarkan data latih sekaligus mengelompokkan dan menangani ketidakrataan informasi dengan lebih baik dibandingkan dengan pengambilan sampel ulang.[7] Terdapat dua metode yang sering digunakan dalam teknik ensemble, yaitu bagging dan boosting. Bagging memiliki kemampuan untuk mendukung algoritma klasifikasi yang tidak stabil, seperti Decision Tree. Algoritma ini cenderung

mengalami perubahan yang besar dalam prediksi ketika digabungkan dengan teknik bagging. Oleh karena itu, algoritma decision tree dapat mencapai akurasi yang lebih baik daripada pengklasifikasian tunggal.[7] Sedangkan teknik boosting memiliki fungsi utama untuk meningkatkan performa klasifikasi dengan memanfaatkan kekuatan kolektif dari beberapa model lemah yang bekerja secara sinergis.

Beberapa penelitian yang telah dilakukan sebelumnya terkait prediksi penyakit stroke menggunakan algoritma decision tree, naïve bayes dan teknik ensemble berjudul “Komparasi Penerapan Metode Bagging dan Adaboost pada Algoritma C4.5 untuk Prediksi Penyakit Stroke” di teliti oleh [7] menghasilkan nilai akurasi algoritma C4.5 sebesar 92,87%. Namun, setelah menerapkan teknik bagging, akurasi meningkat menjadi 95,02%, sedangkan setelah menerapkan metode Adaboost, akurasi mencapai 94,63%. Perbandingan antara teknik Bagging dan Adaboost pada algoritma C4.5 membuktikan peningkatan dan perbaikan kinerja klasifikasi. Akurasi algoritma C4.5 meningkat sebesar 3% dan 2% setelah digunakan metode bagging dan Adaboost secara berturut-turut. Selanjutnya penelitian yang dilakukan oleh [11] berjudul “*Analyzing the Performance of Stroke Prediction using ML Classification Algorithms*” dengan dataset berjumlah 5110 data. Hasil uji coba pada 6 algoritma klasifikasi yang dipilih, algoritma Naïve Bayes Classification memiliki kinerja terbaik dengan akurasi 82%. Sedangkan algoritma Logistic Regression memiliki akurasi 78%, Decision Tree Classification 66%, Random Forest Classification 73%, K-Nearest Neighbors Classification 80%, dan Support Vector Machine sebesar 80%.

Berdasarkan penelitian – penelitian terdahulu, maka penulis akan membandingkan algoritma klasifikasi Decision Tree dan Naïve Bayes dengan penerapan teknik ensemble yaitu teknik Bagging dan Adaboost untuk memprediksi penyakit stroke dan menentukan algoritma mana yang memberikan nilai akurasi tertinggi dalam prediksi penyakit stroke. Penelitian ini bertujuan untuk menanggulangi ketidakakuratan prediksi dan mengevaluasi kemampuan kinerja algoritma Decision Tree, Naïve Bayes yang dikombinasikan dengan teknik ensemble dalam melakukan prediksi penyakit stroke sehingga didapatkan hasil akurasi yang lebih baik. Berdasarkan pemaparan tersebut maka penulis mengambil

judul “KOMPARASI PENERAPAN TEKNIK BAGGING DAN ADABOOST PADA ALGORITMA DECISION TREE DAN NAIVE BAYES UNTUK PREDIKSI PENYAKIT STROKE”.

## **1.2 Identifikasi Masalah**

1. Meningkatnya jumlah kasus penyakit stroke, yang saat ini merupakan penyebab kematian utama di Indonesia, langkah-langkah prediktif diperlukan untuk mengidentifikasi potensi risiko seseorang terkena stroke. Hal ini bertujuan untuk mencegah dan mengurangi jumlah kasus penyakit stroke di Indonesia.
2. Penerapan Metode Klasifikasi Data Mining masih menunjukkan beberapa kelemahan dan ketidaktepatan dalam melakukan prediksi. Oleh karena itu, diperlukan pemanfaatan teknik ensemble untuk meningkatkan nilai akurasi algoritma klasifikasi data mining.
3. Belum diketahui algoritma mana yang memberikan hasil akurasi terbaik pada prediksi penyakit stroke.

## **1.3 Rumusan Masalah**

Berdasarkan latar belakang masalah yang telah dipaparkan, maka perumusan masalah dalam penelitian ini adalah sebagai berikut.

1. Bagaimana penerapan teknik Bagging dan Adaboost pada algoritma Decision Tree dan Naïve Bayes untuk prediksi penyakit stroke?
2. Bagaimana teknik Bagging dan Adaboost meningkatkan kinerja algoritma Decision Tree dan Naive Bayes dalam memprediksi penyakit stroke?
3. Apa model algoritma terbaik dari hasil komparasi teknik Bagging dan Adaboost pada algoritma Decision Tree dan Naïve Bayes untuk memprediksi penyakit stroke?

#### **1.4 Batasan Masalah**

Beberapa batasan masalah pada penelitian ini adalah:

1. Data yang digunakan pada penelitian ini diunduh melalui website kaggle.com yaitu *dataset stroke prediction*.
2. Data yang digunakan sebanyak 5110 data dengan 11 variabel prediktor dan 1 variabel target/label. Variabel prediktor terdiri dari id, gender, age, hypertension, heart\_disease, ever\_married, work\_type, residence\_type, avg\_glucose\_level, bmi, smoking\_status. Sedangkan variabel target terdapat output stroke yang memiliki 2 kategori yaitu 0 berarti pasien tidak stroke dan 1 berarti pasien stroke.
3. Penelitian ini hanya mencakup prediksi penyakit stroke dengan menggunakan dataset healthcare-dataset-stroke-data.
4. Metode klasifikasi yang digunakan yaitu algoritma Decision Tree dan Naïve Bayes. Sedangkan teknik Bagging dan Adaboost untuk meningkatkan akurasi kinerja dari Decision Tree dan Naïve Bayes.

#### **1.5 Tujuan Penelitian**

Dari permasalahan yang telah dirumuskan, maka peneliti memiliki tujuan penelitian sebagai berikut:

1. Menerapkan teknik Bagging dan Adaboost pada algoritma Decision Tree dan Naïve Bayes untuk meningkatkan akurasi prediksi penyakit stroke.
2. Mendapatkan hasil komparasi teknik Bagging dan Adaboost dalam meningkatkan kinerja algoritma Decision Tree dan Naive Bayes pada prediksi penyakit stroke.
3. Mendapatkan model algoritma dengan hasil terbaik dalam memprediksi penyakit stroke.

#### **1.6 Manfaat Penelitian**

Hasil penelitian ini memberikan beberapa manfaat sebagai berikut:

1. Penelitian ini dapat memberikan wawasan yang lebih dalam tentang bagaimana teknik bagging dan adaboost dapat diterapkan pada algoritma decision tree dan naive bayes untuk meningkatkan prediksi penyakit stroke.
2. Dengan membandingkan dua teknik ensemble (bagging dan adaboost) pada dua algoritma berbeda (decision tree dan naive bayes), penelitian ini dapat membantu mengidentifikasi kombinasi terbaik yang memberikan hasil prediksi penyakit stroke yang lebih akurat.
3. Menambahkan literatur ilmiah mengenai penggunaan teknik ensemble pada algoritma decision tree dan naive bayes untuk prediksi penyakit stroke.
4. Dalam bidang kesehatan, hasil penelitian dapat diimplementasikan dalam sistem kesehatan untuk meningkatkan keakuratan prediksi penyakit stroke.

## **1.7 Sistematika Penulisan**

Berikut adalah deskripsi singkat mengenai struktur penulisan pada setiap bab yang digunakan oleh penulis untuk memudahkan penyusunan tesis ini:

### **BAB I PENDAHULUAN**

Bab I memberikan gambaran mengenai konteks latar belakang masalah, identifikasi masalah, perumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, dan struktur penulisan.

### **BAB II LANDASAN TEORI**

Bagian ini mencakup studi terdahulu terkait topik penelitian serta teori-teori pendukung yang relevan dan berkaitan erat dengan landasan berpikir utama yang terkait dengan topik penelitian.

### **BAB III METODE PENELITIAN**

Pada bagian ini berisi metode penelitian menggunakan Metode Decision Tree dan Naïve Bayes dengan teknik ensemble yaitu

Bagging dan Adaboost serta tahapan yang dilakukan dalam penelitian antara lain: alur penelitian dan metode pengumpulan data.

#### **BAB IV HASIL PENELITIAN DAN PEMBAHASAN**

Dalam bab ini menjelaskan hasil eksperimen meliputi proses pengolahan data, pelatihan model per-skenario pengujian, perhitungan confusion matrix, pengujian model yaitu pengujian algoritma Decision Tree, Naïve Bayes dan penerapan teknik ensemble yaitu Bagging dan Adaboost untuk meningkatkan kinerja algoritma Decision Tree dan Naïve Bayes serta analisis hasil pengujian.

#### **BAB V SIMPULAN DAN SARAN**

Pada bab ini, berisi ringkasan dari pembahasan yang mencakup jawaban terhadap permasalahan yang dirumuskan, tujuan penelitian, dan hipotesis. Selain itu, juga terdapat saran untuk penelitian mendatang, yang merupakan hasil refleksi dari keterbatasan penelitian yang telah dilakukan.

#### **DAFTAR PUSTAKA**

Daftar pustaka memuat referensi literatur yang digunakan atau dirujuk dalam penelitian, mencakup buku, jurnal, baik dalam format fisik maupun elektronik.

#### **LAMPIRAN**

Bagian ini mencakup informasi atau data yang mendukung penelitian, termasuk gambar, tabel, diagram, flowchart dan elemen visual lainnya.