BAB I

PENDAHULUAN

1.1 Latar Belakang

Stroke merupakan penyakit serebrovaskuler (pembuluh darah otak) yang disebabkan oleh gangguan atau kerusakan otak yang ditandai dengan kematian jaringan otak (*Infark serebral* (Mutmainah, 2021). Secara global penyakit stroke merupakan penyebab kematian tertinggi kedua, serta di indonesia penyakit stroke juga menjadi salah satu penyebab kematian tertinggi (Zuama, Rahmatullah and Yuliani, 2022). Menurut WHO sebanyak 70% kematian disebabkan dari penyakit stroke, serta 87% kematian yang diakibatkan oleh stroke terjadi pada negara-negara berpenghasilan rendah dan menengah (Byna and Basit, 2020). Semakin meningkatnya angka kasus penyakit stroke maka bisa membuat ahli medis mengalami kewalahan untuk melakukan diagnosis.

Dibutuhkan waktu yang cukup lama untuk seorang praktisi medis dalam melakukan identifikasi penyakit stroke, sehingga untuk mengatasi hal tersebut diperlukan sistem otomatis untuk melakukan prediksi stroke dari data demografi pasien. Maka dengan membuat suatu model untuk memprediksi risiko stroke akan memberikan kontribusi untuk ahli medis dalam melakukan prediksi faktor risiko untuk mencegah dan melakukan penangan dini dari bahaya penyakit stroke (Faisal and Subekti, 2021).

Pengetahuan awal terhadap penyakit stroke dapat dilihat dari data pasien terdahulu, data penderita penyakit stroke terdahulu tersimpan dan tersusun dalam suatu *dataset*. Seperti pada penelitian ini digunakan suatu *dataset* stroke publik dari

situs web kaggle.com. Pada penelitian ini dilakukan pengolahan data dengan menerapkan teknik *data mining* dengan menggunakan teknik klasifikasi. Pada penelitian ini membandingkan algoritma klasifikasi *k-nearest neighbor* dan *adaptive boosting* serta menerapkan cara mengatasi ketidakseimbangan data menggunakan teknik *SMOTE*. Pada penelitian ini tidak hanya nilai akurasi yang diperhatikan tetapi juga nilai *recall*, karena nilai *recall* mengukur seberapa baik model mendeteksi semua data positif yang sebenarnya ada (Clara *et al.*, 2021). Dalam kasus kesehatan akan sangat fatal apabila seseorang terkena penyakit diklasifikasikan sebagai sehat, maka akan lebih baik model memiliki nilai *recall* yang tinggi.

Algoritma *KNN (K-Nearest Neighbor)* digunakan karena memiliki kemampan untuk mendeteksi dan menganalisa permasalahan yang bersifat kompleks dan *non-linier*, serta dalam proses perhitungan dilakukan secara pararel sehinga waktu komputasi lebih cepat (Isnain, Supriyanto and Kharisma, 2021). Sedangkan algoritma *adaptive boosting* merupakan algoritma yang dapat dikombinasikan dengan mudah untuk meningkatkan nilai akurasi dari suatu klasifikasi, algoritma *adaboost* memiliki dasar teori yang kuat, prediksi yang akurat, dan sangat sederhana, dimana hasil dari algoritma ini berasal dari *weak learner* atau *basic clasifier*. Dengan menerapkan algoritma *adaboost* akan menghasilkan suatu *strong learner* yang berpotensi meningkatkan kekuatan model prediksi (Gultom, 2020).

Data yang tidak seimbang (*imbalance dataset*) merupakan salah satu masalah yang bisa saja muncul dalam proses klasifikasi, ketidakseimbangan data berarti

terdapat salah satu kelas yang menonjol daripada yang lain (Susan and Kumar, 2021). Munculnya kelas minoritas akan menyebabkan klasifikasi yang dibangun tidak berjalan dengan optimal dan akan menghasilkan prediksi yang kurang akurat (Arifiyanti and Wahyuni, 2020) maka dari itu penting sekali untuk menangan ketidakseimbangan data. Terdapat beberapa teknik yang dapat digunakan untuk menangani ketidakseimbangan data, salah satunya yaitu *teknik synthetic minority oversampling technique* (*SMOTE*). Teknik *SMOTE* dalam menangani ketidakseimbangan data dengan cara membuat data sintetis pada kelas minor sehingga data menjadi seimbang.

Berdasarkan penelitian yang telah dilakukan oleh (Sholekhah *et al.*, 2024) mengenai perbandingan algoritma *naïve bayes* dan *KNN (K-Nearest Neighbor)s* untuk klasifikasi metabolik sindrom. Penelitian ini mendapatkan kesimpulan bahwa hasil yang paling optimal adalah pada algoritma KNN. Dengan hasil pada skema pembagian data 50:50, dan parameter K = 5 mendapatkan *accuracy* mencapai 82%, sedangkan *Naïve Bayes* menunjukkan hasil dengan nilai a*ccuracy* sebesar 79%. Kajian terkait perbandingan algortitma *machine learning* juga telah dilakukan dalam penelitian yang dilakukan oleh (Yustikasari, Mubarok and Rianto, 2022) yaitu mengenai perbandingan algoritma *KNN (K-Nearest Neighbor)* dan *adaptive boosting* . Hasil dari penelitian tersebut algoritma *adaptive boosting* memperoleh akurasi rata-rata yang lebih akurat yaitu 94,5% dibandingkan dengan algoritma *KNN (K-Nearest Neighbor)* yang menghasilkan akurasi rata-rata 93%.

Selain itu telah dilakukan penelitian oleh (Prianggi, Nilogiri and Umilasari, 2021) mengenai pengaruh teknik *SMOTE* terhadap prediksi harapan hidup

penderita penyakit hepatitis dengan menggunakan algoritma *KNN (K-Nearest Neighbor)* mendapatkan kesimpulan bahwa teknik *SMOTE* terbukti dapat mengatasi ketidakseimbangan data dengan menghasilkan nilai akurasi lebih baik yaitu 93,48% sedangkan nilai akurasi tanpa menerapkan teknik *SMOTE* didapatkan nilai akurasi 89,66%.

Berdasarkan kajian yang telah dipaparkan, maka penelitian ini akan membandingkan proses klasifikasi menggunakan algoritma *KNN (K-Nearest Neighbor)* dan algoritma *adaptive boosting* dengan menerapkan teknik *SMOTE* untuk menangani *imbalance* data, dengan mengambil judul **"PENERAPAN ALGORITMA *K-NEAREST NEIGHBOR* DAN *ADAPTIVE BOOSTING* MENGGUNAKAN TEKNIK *SMOTE* PADA PREDIKSI PENYAKIT STROKE."**

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang di atas, maka perumusan masalah dalam penelitian ini dapat dinyatakan sebagai berikut:

1. Bagaimana mengimplementasikan model *data mining* yaitu klasifikasi dengan metode *adaptive boosting* dan algoritma *KNN (K-Nearest Neighbor)* untuk memprediksi penyakit stroke?
2. Bagaimana performa yang dihasilkan oleh algoritma *KNN (K-Nearest Neighbor)* dan *adaptive boosting* setelah menerapkan metode *SMOTE* terhadap data stroke?

1.3 Batasan Masalah

Untuk memperjelas arah dan ruang lingkup penelitian, ditetapkan beberapa batasan masalah agar pembahasan fokus terhadap tujuan yang ingin dicapai dan tidak meluas. Adapun batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Data yang digunakan merupakan *dataset stroke disease public* yang diambil dari situs web *kaggle* 2021 *version* 1.
2. Algoritma yang digunakan pada penelitian ini adalah algoritma klasifikasi.
3. Membahas prediksi serta optimalisasi pada penyakit stroke dengan metode *adaptive boosting* pada algoritma *KNN (K-Nearest Neighbor)*.
4. Menerapkan metode *SMOTE* yang berfungsi untuk menangani ketidakseimbangan *data stroke*.

1.4 Tujuan Penelitian

Setelah dirumuskan permasalahan dalam penelitian ini, maka tujuan yang ingin dicapai dapat dijelaskan sebagai berikut :

1. Mengimplementasikan *data mining* dengan penerapan metode *adaptive boosting* dan algoritma *KNN (K-Nearest Neighbor)* untuk memprediksi penyakit stroke.
2. Mengetahui performa metode algoritma *adaptive boosting* dan *KNN (K-Nearest Neighbor)* dengan menerapkan teknik *SMOTE* yang berfungsi untuk menangani ketidakseimbangan data.

1.5 Manfaat Penelitian

Penelitian ini diharapkan dapat memberikan manfaat bagi berbagai pihak yang terkait, antara lain sebagai berikut :

1. Secara Aplikatif
   a. Diharapkan dari hasil prediksi penyakit stroke ini dapat membantu seluruh ahli medis dalam pengambilan keputusan untuk menentukan seseorang mengidap penyakit stroke atau tidak agar tidak terjadi peningkatan kasus stroke yang semakin tinggi.
2. Secara Akademis
   a. Mengetahui cara pembuatan model *data mining* dengan menerapkan metode *adaptive boosting* dan algoritma *KNN (K-Nearest Neighbor).*
   b. Memberikan wawasan bagaimana performa penggunaan metode *adaptive boosting* dan algoritma *KNN (K-Nearest Neighbor).*
   c. Memberikan wawasan bagaimana performa penggunaan metode *adaptive boosting* dan algoritma *KNN (K-Nearest Neighbor)* dengan menerapkan metode *SMOTE* yang berfungsi untuk menangani ketidakseimbangan data.
   d. Membantu peneliti di masa yang akan datang sebagai bahan referensi.

1.6 Sistematika Penulisan

Adapun aturan dan sistematika penulisan laporan yang digunakan dalam penelitian ini sebagai berikut:

BAB I PENDAHULUAN

Bab pendahuluan memuat uraian secara umum mengenai isi laporan, mencakup latar belakang, perumusan masalah, batasan masalah dalam penelitian, tujuan dan manfaat yang ingin dicapai, metode yang digunakan, serta susunan sistematika penulisan dalam laporan tugas akhir.

BAB II LANDASAN TEORI

Bab ini membahas landasan teori yang berkaitan dengan penelitian, meliputi konsep-konsep dasar, metode, serta algoritma yang digunakan. Selain itu, disajikan pula kajian terhadap penelitian-penelitian sebelumnya yang memiliki relevansi dengan topik yang diangkat dalam penelitian ini.

BAB III METODOLOGI

Bab ini menguraikan metode yang digunakan dalam pelaksanaan penelitian, mencakup waktu dan lokasi penelitian, objek dan variabel yang diteliti, matriks penelitian, serta langkah-langkah atau tahapan proses penelitian yang dilakukan.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil dan pembahasan dari proses perancangan yang telah dijelaskan pada bab sebelumnya. Fokus utamanya adalah alur pengolahan data menggunakan metode *Adaptive Boosting* dan algoritma *KNN (K-Nearest Neighbor)*, baik tanpa penerapan teknik *SMOTE* maupun dengan penerapan teknik SMOTE untuk mengatasi ketidakseimbangan data. Selain itu, dilakukan pengujian model guna memperoleh hasil yang selaras dengan tujuan awal penelitian.

BAB V KESIMPULAN DAN SARAN

Bab ini merupakan bagian penutup yang memuat kesimpulan dan saran berdasarkan hasil penelitian. Kesimpulan merangkum temuan utama sebagai hasil akhir dari proses penelitian, sementara saran berisi rekomendasi yang disesuaikan dengan keterbatasan yang ditemui selama pelaksanaan penelitian. Bab ini juga memberikan gambaran umum mengenai metode yang telah diterapkan.