

BAB I

PENDAHULUAN

1.1 Latar Belakang

Stroke merupakan salah satu masalah kesehatan utama di dunia, termasuk di Indonesia. Organisasi Kesehatan Dunia (WHO) menjelaskan bahwa kondisi ini terjadi akibat aliran darah ke otak terganggu, baik karena pecahnya pembuluh darah maupun karena adanya sumbatan dari bekuan darah. Gangguan tersebut mengakibatkan jaringan otak mengalami kerusakan karena tidak mendapatkan suplai oksigen serta nutrisi yang dibutuhkan. Data WHO tahun 2018 mencatat sekitar 7,75 juta kematian di seluruh dunia disebabkan oleh stroke. Sementara itu, *Centers for Disease Control* (CDC) melaporkan bahwa di Amerika Serikat pada tahun 2020, rata-rata satu orang meninggal setiap empat menit akibat penyakit ini. Laporan WHO tahun 2017 juga menegaskan bahwa pada tahun 2015, penyakit kardiovaskular, termasuk stroke iskemik, menyumbang 17,7 juta kematian atau sekitar 31% dari seluruh kasus kematian global [1]. Di Indonesia, prevalensi stroke meningkat dari 12,1% pada 2020 menjadi 14,9% pada 2021. Peningkatan kasus ini dipengaruhi oleh bertambahnya jumlah penduduk lanjut usia, perubahan pola hidup, serta meningkatnya prevalensi hipertensi di masyarakat. Dengan kondisi tersebut, stroke menjadi penyebab utama kecacatan dan kematian di Indonesia, dan prevalensinya terus menunjukkan tren peningkatan setiap tahunnya [2].

Klasifikasi merupakan suatu teknik untuk menempatkan objek ke dalam kelompok atau kategori tertentu yang telah ditentukan sebelumnya. Penempatan ini didasarkan pada kesamaan karakteristik yang dimiliki oleh data [3][4]. Beberapa penelitian terdahulu terkait klasifikasi penyakit stroke telah dilakukan, salah satunya memanfaatkan algoritma *Decision Tree C4.5* pada data pasien di Rumah Sakit Santa Maria Pematang. Studi tersebut menunjukkan bahwa teknik data mining, khususnya algoritma C4.5, dapat digunakan untuk mengelompokkan pasien ke dalam kategori stroke maupun non-stroke. Hasil eksperimen dengan metode tersebut menghasilkan tingkat akurasi pada data latih sebesar 82,31% (130 dari 156 data pasien), sedangkan pada data uji diperoleh akurasi 76,92% (26 dari

156 data pasien). Evaluasi performa dilakukan dengan menggunakan confusion matrix. Penelitian tersebut menekankan bahwa algoritma Decision Tree C4.5 dapat diimplementasikan secara efektif dalam mendukung klasifikasi penyakit stroke.

Metode pemungutan suara yang paling umum digunakan adalah pemungutan suara mayoritas, yang berlandaskan pada algoritma *Random Forest* (RF). Klasifikasi menggunakan *Random Forest* dapat mencapai ketepatan tinggi dalam pengelompokan data bila dibandingkan dengan berbagai metode klasifikasi lainnya, dan bisa mengurangi tingkat kesalahan klasifikasi keseluruhan serta mampu menangani data yang tidak seimbang. Namun, ketika rasio ketidakseimbangan meningkat (seperti 15%), kemampuan klasifikasi menurun, disebabkan karena setiap klasifikasi mempunyai bobot yang sama saat digabungkan. Untuk mengatasi permasalahan ini, diperkenalkan metode penggabungan klasifikasi dengan bobot yang berbeda. *Weighted Voting Random Forest* (WRF) merupakan algoritma yang memperhitungkan setiap keputusan klasifikasi dengan bobot yang mencerminkan kepercayaan individu terhadap keputusan tersebut. Penggunaan metode pemungutan suara mayoritas berbobot yang didasarkan pada independensi kondisional kelas dari hasil klasifikasi juga disarankan. Bobot suara endogen disebutkan dalam konteks perwakilan yang dipilih dan pemetaan ulang distrik. Kekuatan dalam Permainan Pemungutan Suara Berbobot dengan Bobot Super-Meningkat dianalisis. Metode *Random Forest* (RF) dengan pemungutan suara berbobot untuk tugas deteksi anomali juga diperkenalkan. Metrik kesalahan out-of-bag (OOB) internal dialokasikan sebagai bobot pohon dalam RF. Namun, pendekatan ini belum memberikan peningkatan yang signifikan pada kemampuan prediktif. Setiap klasifikasi hanya memiliki satu bobot saat mereka digabungkan, yang tidak cukup memisahkan kelas mayoritas dari minoritas; bobot kelas (*class weight*), yang merujuk pada bobot berbeda untuk setiap kelas dalam klasifikasi, tidak ditentukan. Berdasarkan pengamatan tersebut, klasifikasi membutuhkan beberapa bobot, dan meningkatkan *class weight* untuk setiap klasifikasi dapat berpotensi meningkatkan kinerja prediktif secara keseluruhan. Oleh karena itu, *class weight* perlu ditetapkan untuk lebih mencerminkan kemampuan klasifikasi dalam membedakan antara kelas mayoritas dan minoritas.

Random Forest (RF) melibatkan berbagai parameter yang mengatur struktur setiap pohon secara individu (misalnya, ukuran minimum simpul yang diperlukan untuk pemecahan) serta struktur dan jumlah pohon dalam hutan (misalnya, total pohon) dan derajat kerandoman (misalnya, jumlah variabel *mtry* yang dipertimbangkan sebagai kandidat untuk pemecahan dalam setiap pembagian atau metode pengambilan sampel untuk menghasilkan dataset yang menjadi dasar pembangunan pohon-pohon tersebut). Pengaruh dari parameter ini telah diteliti dalam beberapa karya ilmiah. Namun, hasil yang berkaitan dengan pengaruh ini sering kali berfokus pada satu parameter dan hanya disajikan sebagai hasil sampingan dari penelitian yang ditujukan untuk topik lain (contohnya, varian baru RF), sehingga sulit diakses oleh pembaca yang tidak memiliki pengetahuan mendalam terkait literatur tersebut. Panduan yang jelas masih sangat dibutuhkan, dan memilih nilai yang tepat untuk parameter tetap menjadi tantangan dalam praktik.

Oleh sebab itu, saya mengusulkan *class weight* untuk klasifikasi dengan bobot yang berbeda di setiap kelas. Bobot berbeda untuk setiap kelas diambil dari kesalahan empiris dari klasifikasi yang beragam. Algoritma ini memberikan bobot yang berbeda untuk setiap kelas, bukan satu bobot tunggal, dan memfokuskan perhatian pada tantangan identifikasi yang efektif untuk kelas minoritas. Pendekatan ini mampu meningkatkan kinerja pengenalan untuk kelas minoritas sambil tetap menjaga kinerja untuk kelas mayoritas.

1.2 Rumusan Masalah

Beberapa rumusan masalah yang didapat dari latar belakang yang telah dijelaskan adalah sebagai berikut ini.

1.2.1 Permasalahan

Dalam klasifikasi penyakit stroke, salah satu hambatan utama yang dihadapi adalah ketidakseimbangan distribusi data. Jumlah pasien yang tidak mengalami stroke jauh lebih dominan dibandingkan dengan pasien yang terkena stroke. Kondisi ini membuat algoritma klasifikasi, seperti *Random Forest*, cenderung

berpihak pada kelas mayoritas. Karena kinerja *Random Forest* sangat dipengaruhi oleh pengaturan parameter, maka diperlukan *class weight* sebagai strategi agar perlakuan di setiap kelas nya menjadi setara. Dengan pengaturan bobot kelas yang sesuai, model diharapkan mampu meningkatkan ketepatan klasifikasi, terutama dalam mengidentifikasi pasien stroke yang termasuk dalam kelas minoritas.

1.2.2 Solusi Permasalahan

Pada penelitian ini solusi yang diusulkan adalah klasifikasi penyakit stroke menggunakan metode *Class Weight Random Forest*

1.2.3 Pertanyaan Penelitian

Bagaimana perbandingan kinerja akurasi model klasifikasi penyakit stroke antara metode Random Forest yang menerapkan penyesuaian bobot kelas (*Class Weight Random Forest*) dengan metode Random Forest yang menerapkan teknik penyeimbangan data menggunakan *Synthetic Minority Oversampling Technique* (SMOTE)?

1.3 Batasan Masalah

Agar penelitian ini lebih terarah dan tetap fokus pada topik yang dibahas, maka diperlukan adanya batasan masalah. Batasan dalam penelitian ini dijelaskan sebagai berikut:

1. Dataset yang digunakan berasal dari situs kaggle.com, tepatnya dari repository dataset *Stroke Prediction*
2. Jumlah data yang dianalisis dalam penelitian ini sebanyak 5110 entri.
3. Terdapat 11 fitur yang dimanfaatkan untuk proses klasifikasi apakah seorang pasien menderita stroke atau tidak. Fitur-fitur tersebut meliputi: *gender*, *age*, *hypertension*, *heart_disease*, *ever_married*, *work_type*, *residence_type*, *avg_glucose_level*, *bmi*, *smoking_status*, dan *stroke*.
4. Klasifikasi penyakit stroke yang dilakukan hanya berdasarkan pengidap atau tidak penyakit stroke.
5. Metode *Class Weight Random Forest* digunakan untuk memperhatikan hubungan kontekstual dalam data penyakit stroke.

1.4 Tujuan dan Manfaat

Berikut adalah tujuan dan manfaat dari penelitian ini berdasarkan rumusan masalah yang telah dijelaskan.

1.4.1 Tujuan

Untuk mengetahui seberapa besar hasil kinerja metode dari klasifikasi penyakit stroke menggunakan metode *Class Weight Random Forest*.

1.4.2 Manfaat

Dapat mengetahui seberapa besar nilai akurasi yang didapat dalam klasifikasi penyakit stroke menggunakan metode *Class Weight Random Forest* yang pada akhirnya dapat digunakan sebagai rujukan dalam pembangunan sistem diagnosa penyakit stroke.

1.5 Sistematika Laporan

Sistematika laporan tugas akhir ini disusun dalam lima bab sebagai berikut:

BAB I PENDAHULUAN

Bab ini menyajikan uraian umum mengenai penelitian yang akan dilaksanakan. Di dalamnya mencakup penjelasan singkat mengenai latar belakang, rumusan masalah, batasan penelitian, tujuan serta manfaat yang ingin dicapai, dan juga sistematika penyusunan laporan.

BAB II KAJIAN PUSTAKA

Bab ini menjelaskan tentang landasan teori yang menjadi acuan atau pedoman dan landasan dalam melakukan penelitian klasifikasi penyakit stroke menggunakan metode *Class Weight Random Forest*, serta penelitian terkait.

BAB III METODOLOGI PENELITIAN

Bab ini memaparkan rancangan arsitektur sistem, tahapan persiapan data, serta alur dari setiap metode yang diusulkan. Selain itu, juga dijelaskan skenario pengujian yang digunakan.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil implementasi sistem klasifikasi penyakit stroke menggunakan metode *Class Weight Random Forest*. Di dalamnya dibahas hasil pengujian model berdasarkan skenario yang telah ditetapkan, termasuk evaluasi performa menggunakan metrik seperti accuracy, precision, recall, F1-score,

confusion matrix, dan ROC curve. Selain itu, pada bab ini juga dilakukan analisis dan pembahasan terhadap hasil yang diperoleh untuk menilai efektivitas metode yang digunakan serta membandingkan performa antara skenario Class Weight Random Forest dan SMOTE Random Forest.

BAB V KESIMPULAN DAN SARAN

Bab ini berisi kesimpulan yang diperoleh dari seluruh rangkaian penelitian yang telah dilakukan, mencakup pencapaian tujuan penelitian serta temuan utama yang dihasilkan. Selain itu, bab ini juga memuat saran untuk pengembangan penelitian selanjutnya, baik dari sisi metode, data, maupun pengembangan sistem, agar dapat diperoleh hasil yang lebih optimal di masa mendatang.