

BAB I

PENDAHULUAN

1.1 Latar Belakang

Stroke merupakan penyebab kematian kedua dan penyebab kecacatan ketiga di dunia (WHO, 2016). Stroke biasanya menyerang orang tua. Namun, orang muda juga dapat terkena stroke. Penyakit ini terjadi ketika aliran darah dan oksigen ke otak tersumbat karena penyumbatan pembuluh darah atau pecahnya pembuluh darah. Kurangnya suplai darah dan oksigen di otak dapat menyebabkan bagian tubuh tertentu tidak berfungsi dengan baik.

Menurut data *World Stroke Organization*, terdapat 13,7 juta kasus stroke baru setiap tahunnya, dan sekitar 5,5 juta orang meninggal karena stroke. Sekitar 70% penyakit stroke dan 87% kematian dan kecacatan terkait stroke terjadi di negara berpenghasilan rendah dan menengah. Dalam 15 tahun terakhir, dibandingkan dengan negara-negara berpenghasilan tinggi, negara-negara berpenghasilan rendah dan menengah memiliki insiden dan tingkat kematian stroke yang lebih tinggi. Di Indonesia, menurut hasil Riskesdas, prevalensi stroke pada tahun 2018 mengalami peningkatan dibandingkan tahun 2013, dari 7% menjadi 10,9%, dengan perkiraan 2.120.362 orang. Kalimantan Timur merupakan provinsi dengan prevalensi stroke tertinggi (14,7%). Prevalensi stroke terendah di Maluku Utara (4,6%).

Gejala stroke mudah dikenali, antara lain ukuran wajah mengecil, kesulitan tersenyum, ketidakmampuan menggerakkan bagian tubuh tertentu, serta kesulitan

berbicara dan melihat. Gejala lain mungkin termasuk mual, sakit kepala, kehilangan kesadaran, dan kehilangan keseimbangan dan koordinasi antar bagian tubuh. Jika pasien memiliki riwayat kesehatan dan gaya hidup yang buruk seperti tekanan darah tinggi, diabetes, obesitas, kolesterol tinggi, penyakit jantung, kurang olahraga, merokok dan alkoholisme, risiko terkena penyakit ini akan meningkat.

Deteksi dini penyakit stroke sangat penting karena menunda pengobatan penyakit ini dapat menyebabkan sel-sel otak mati dalam beberapa menit. Sehingga pengobatan yang tepat dapat dilakukan untuk mengurangi risiko kerusakan otak dan komplikasi yang dapat mempengaruhi kehidupan seseorang. Menggunakan analisis riwayat kesehatan dan laporan pemeriksaan fisik pasien untuk mendiagnosis penyakit stroke sering kali membutuhkan waktu yang relatif lama dan tidak terlalu akurat. Oleh karena itu, perlu diterapkan sistem pendukung prediksi stroke melalui model klasifikasi *machine learning* (Kunder Akash, Shashank, S, & A. M, 2020).

Klasifikasi merupakan suatu pemodelan prediktif yang memprediksi label kelas pada tiap pengamatan. Salah satu masalah dalam klasifikasi yaitu klasifikasi data tidak seimbang dimana distribusi contoh di seluruh kelas pengamatan yang diketahui tidak memiliki proporsi yang sama. Hal ini mengakibatkan model klasifikasi memiliki kinerja yang buruk dalam memprediksi kelas minoritas.

Metode *Syntetic Minority Oversampling Technique* (SMOTE) digunakan untuk menangani data kelas tidak seimbang. Metode SMOTE menghasilkan data sintetik dari kelas minor dengan mencari nilai tetangga terdekat, sehingga menghasilkan akurasi terbaik ketika berhadapan dengan data yang tidak seimbang.

Metode ini efektif karena data sintetik dari kelas minoritas yang dibuat ini masuk akal, yaitu relatif dekat dengan contoh yang ada di fitur kelas minoritas.

Masalah lain yang sering ditemui dalam klasifikasi *machine learning* pada data medis yaitu adanya data pencilan. Pada kelas data yang sama terdapat sampel data yang memiliki karakteristik berbeda dengan yang lainnya. Misalnya, dalam tugas klasifikasi status penyakit stroke yang melibatkan kelas data berstatus penyakit stroke dan sehat, pasien yang sehat mungkin menunjukkan karakteristik yang umumnya terdapat pada kelas data berstatus penyakit stroke seperti level glukosa dan berat badan. Hal ini dapat mengganggu mekanisme pembelajaran algoritma sehingga mengarah pada kesalahan klasifikasi.

XGBoost atau *eXtreme Gradient Boosting* adalah metode pembelajaran terintegrasi terbaik yang memberikan hasil akurasi tinggi dan biasa digunakan dalam kompetisi Kaggle. Kelebihan XGBoost adalah lebih tahan terhadap outlier, waktu perhitungan yang singkat, dan hasil yang akurat (Syukron, Santoso, & Widiyarih, 2020). XGBoost merupakan metode *machine learning* yang efektif digunakan pada dataset berdistribusi tidak seimbang (Brownlee, 2020) serta mampu menangani data klasifikasi berukuran besar dengan jumlah *feature* lebih kecil daripada jumlah data pengamatan pada data training (Chen & Guestrin, 2016). Penelitian yang dilakukan oleh Mardiansyah (2019) membuktikan bahwa metode SMOTE-XGBoost memberikan akurasi lebih baik(99,40%) dibandingkan metode Logistic Regression, SMOTE logistic regression, random Forest, SMOTE random forest, SVM, SMOTE SVM, dan XGBoost dalam mengklasifikasi *credit scoring*.

Berbagai penelitian pemanfaatan *machine learning* untuk memprediksi penyakit stroke telah dilakukan. *Machine Learning* adalah pengetahuan dan seni pemrograman komputer yang belajar dari data (Geron, 2017). Penelitian terdahulu tentang klasifikasi penyakit stroke telah dilakukan oleh Rado, Fanah, & Taktek (2019) menggunakan metode klasifikasi SVM, KNN dan Decision Tree sebagai klasifikasi tunggal dan Random Forest, Adaboost, Stacking sebagai Ensemble Learning. Hasil menunjukkan, bahwa metode Stacking memiliki nilai akurasi lebih baik dibandingkan dengan beberapa metode lainnya. Selain itu, Govindarajan, et al., (2019) membandingkan klasifikasi penyakit stroke menggunakan metode ANN, SVM, Boosting dan Bagging. Diperoleh hasil bahwa metode ANN memiliki akurasi terbaik sebesar 95% dalam mengklasifikasikan penyakit stroke.

Dalam kasus data kelas tidak seimbang, hasil klasifikasi akan mencapai tingkat akurasi yang tinggi karena hanya melihat kelas mayoritas saja. Oleh karena itu, untuk data kelas tidak seimbang, nilai akurasi klasifikasi tidak cukup sebagai ukuran standar, perlu menggunakan nilai *Area Under Curve* (AUC) dalam memahami kinerja metode klasifikasi data tidak seimbang.

Berdasarkan uraian latar belakang tersebut maka akan dilakukan penelitian berjudul “**KLASIFIKASI PENYAKIT STROKE MENGGUNAKAN SMOTE-XGBOOST**” yang bertujuan membandingkan nilai AUC metode XGBoost dan SMOTE-XGBoost dalam mengklasifikasikan apakah pasien cenderung mengalami penyakit stroke.

1.2 Rumusan Masalah

Berdasarkan latar belakang di atas, maka permasalahan yang dirumuskan adalah sebagai berikut:

1. Bagaimana hasil *Area Under Curve* klasifikasi penyakit stroke menggunakan metode XGBoost?
2. Bagaimana hasil *Area Under Curve* klasifikasi penyakit stroke menggunakan metode SMOTE-XGBoost?
3. Bagaimana perbandingan kinerja metode XGBoost dan SMOTE-XGBoost dalam mengklasifikasi penyakit stroke?

1.3 Tujuan Penelitian

Berdasarkan permasalahan yang telah diuraikan di atas, maka tujuan yang ingin dicapai dalam penelitian ini adalah sebagai berikut:

1. Mengetahui *Area Under Curve* klasifikasi penyakit stroke menggunakan metode XGBoost.
2. Mengetahui *Area Under Curve* klasifikasi penyakit stroke menggunakan metode SMOTE-XGBoost.
3. Metode terbaik dalam mengklasifikasi status penyakit stroke.

1.4 Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut:

1. Menambah pengetahuan tentang algoritma klasifikasi penyakit stroke menggunakan XGBoost dan SMOTE-XGBoost

2. Dapat menjadi bahan acuan untuk penelitian dengan masalah yang sama secara lebih mendalam.
3. Menambah keilmuan statistika di bidang klasifikasi *machine learning* pada data tidak seimbang.

1.5 Batasan Penelitian

Batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Data yang digunakan adalah data penyakit stroke yang diambil dari situs Kaggle.
2. Masalah klasifikasi dalam penelitian ini adalah masalah klasifikasi biner, yaitu ada dua jenis klasifikasi (0 dan 1).
3. Proses analisis data menggunakan metode XGBoost dan SMOTE-XGBoost