

BAB 1

PENDAHULUAN

1.1 Latar Belakang Penelitian

Stroke merupakan kondisi patologis pada sistem vaskular serebral yang ditandai oleh terganggunya atau menurunnya aliran darah serta suplai oksigen ke otak, yang dapat menyebabkan kerusakan atau kematian jaringan otak, disertai gangguan fungsi neurologis[1]. Menurut laporan Organisasi Kesehatan Dunia (WHO), setiap tahunnya terdapat sekitar 15 juta kasus stroke di seluruh dunia, di mana 5 juta di antaranya berakhir dengan kematian dan 5 juta lainnya mengalami disabilitas permanen[2]. Stroke tercatat sebagai penyebab kematian kedua dan penyebab kecacatan ketiga secara global, dengan sekitar 70% kasus terjadi di negara-negara dengan ekonomi menengah dan rendah. Dari keseluruhan beban mortalitas dan morbiditas tersebut, sekitar 87% disebabkan oleh stroke[3]. Oleh karena itu, penyakit ini menjadi salah satu isu kesehatan global yang mendesak dan menimbulkan dampak signifikan secara luas[4][5].

Penelitian ini memanfaatkan berbagai pendekatan machine learning, seperti algoritma klasifikasi XGBoost, Random Forest, dan Logistic Regression, dengan mempertimbangkan sejumlah komponen yang diduga kontribusi pada terhadap risiko stroke. Dengan menghimpun dan mengolah data dari studi-studi sebelumnya serta menganalisis faktor-faktor yang berpotensi memengaruhi kejadian stroke, penelitian ini bertujuan untuk mengidentifikasi pola atau hubungan tersembunyi yang mungkin tidak terdeteksi apabila setiap variabel dianalisis secara terpisah [6]. Penggunaan algoritma machine learning dalam proses klasifikasi penyakit telah menunjukkan hasil yang menjanjikan. Di antara berbagai metode yang tersedia, XGBoost, Random Forest, dan Logistic Regression merupakan tiga algoritma yang populer dan sering diterapkan dalam analisis data medis. XGBoost (Extreme Gradient Boosting) dikenal karena kemampuan untuk mengelola data yang tidak seimbang dan menghasilkan model prediktif yang relevan[7].

Namun, Random Forest adalah teknik kelompok yang meningkatkan ketepatan klasifikasi dengan menggunakan sekumpulan pohon keputusan dan mengurangi risiko overfitting. Adapun Logistic Regression bekerja dengan menghitung probabilitas terjadinya suatu peristiwa berdasarkan data masukan, sehingga sering digunakan dalam analisis yang membutuhkan interpretasi sederhana namun tetap menghasilkan prediksi yang andal

Stroke merupakan salah satu permasalahan kesehatan utama yang menempati peringkat

kedua sebagai penyebab kematian tertinggi secara global. Penyakit ini tidak hanya berdampak pada tingginya angka mortalitas, tetapi juga berkontribusi signifikan terhadap terjadinya disabilitas yang dapat menurunkan kemampuan individu dalam menjalankan aktivitas sehari-hari. Melalui pemahaman yang lebih komprehensif mengenai stroke serta pemanfaatan teknologi dalam proses klasifikasi, tujuan penelitian ini adalah untuk membuat model.

klasifikasi terhadap faktor-faktor penyebab stroke pada manusia. Dalam studi ini, dilakukan perbandingan antara algoritma XGBoost, Random Forest, dan Logistic Regression guna mengevaluasi efektivitas masing-masing metode dalam mengidentifikasi faktor-faktor risiko secara optimal.

1.2 Rumusan Masalah

Rumusan masalah penelitian adalah bagaimana cara agar dapat menggunakan algoritma untuk mengklasifikasikan penyakit stroke, berdasarkan uraian latar belakang sebelumnya. XGBoost, Random Forest, dan Logistic Regression, serta bagaimana membandingkan kinerja masing-masing algoritma.

1.3 Tujuan

Penelitian ini dilakukan untuk mengidentifikasi serta mengklasifikasikan faktor-faktor penyebab stroke dengan menggunakan tiga model klasifikasi: XGBoost, Random Forest, dan Logistic Regression untuk menentukan model yang paling efektif.

1.4 Manfaat Penelitian

Adapun manfaat yang dapat dihasilkan dari penelitian ini adalah:

1. Memberikan sumber referensi dan pendidikan bagi pihak yang tertarik dalam penerapan model klasifikasi XGBoost, Random Forest, dan Logistic Regression untuk mengidentifikasi faktor-faktor penyebab penyakit stroke.
2. Membantu memudahkan bagi praktisi medis dalam pengambilan keputusan yang lebih tepat terkait proses diagnosis dan penanganan pasien yang berisiko mengalami stroke, melalui pemanfaatan teknologi berbasis machine learning

1.5 Batasan Penelitian

Batasan penelitian ini adalah sebagai berikut.

1. Metode untuk mengklasifikasikan diterapkan dalam penelitian ini meliputi algoritma XGBoost, Random Forest, dan Logistic Regression, yang digunakan untuk membandingkan efektivitas masing-masing dalam klasifikasi penyakit stroke.
2. Dataset yang digunakan dalam studi ini diperoleh dari [DATA PASIEN PENYAKIT STROKE](#), dengan fokus pada data medis pasien yang terkait dengan penyakit stroke.
3. Dataset tersebut terdiri atas 5.110 data dengan 12 atribut yang merepresentasikan berbagai faktor penyebab stroke, termasuk usia, jenis kelamin, hipertensi, penyakit jantung, status merokok, serta indeks massa tubuh (BMI).
4. Proses pengujian dan pelatihan model dilakukan menggunakan bahasa pemrograman Python pada lingkungan Jupyter Notebook, dengan pemanfaatan GPU guna mempercepat komputasi.

1.6 Keterbaruan

Dalam hal klasifikasi penyakit stroke, penelitian ini memperbarui penelitian sebelumnya.

1. *Kenny Riva et al* [8]. Uji algoritma SVM untuk klasifikasi data stroke. Penelitian ini menunjukkan bahwa dalam informasi yang (unbalanced), kernel linear menghasilkan kernel polynomial memiliki 80% akurasi. Sementara itu, pada data yang telah diseimbangkan (balanced), akurasi yang diperoleh adalah 77% untuk kernel linear dan 76% untuk kernel polynomial.
2. *Aulia et al.* [8]. Algoritma Decision Tree merupakan metode klasifikasi terbaik dalam studi mereka, dengan akurasi tertinggi sebesar 95,13%, mengungguli algoritma Random Forest dan Naïve Bayes. Nilai True False (TF) tercatat sebanyak 4.861, True True (TT) sebesar 0, False False (FF) sebanyak 249, dan False True (FT) sebesar 0.
3. *Akmal et al.* [9]. Pemanfaatan algoritma Naïve Bayes dan K-Nearest Neighbors (K-NN) sebagai pendekatan evaluatif dalam pengembangan sistem cerdas yang dapat mendukung tenaga medis dalam pengambilan keputusan, baik dalam aspek preventif maupun klinis, untuk mempercepat proses diagnosis pasien stroke. Menurut hasil penelitian, algoritma K-NN bisa mencapai tingkat akurasi 94,36%.