

BAB 1

PENDAHULUAN

1.1 Latar Belakang

Stroke (*Cerebro-Vascular Accident* atau CVA) adalah suatu kondisi di mana terjadi gangguan mendadak dalam aliran darah menuju otak. Hal ini disebabkan oleh penyumbatan (stroke iskemik) atau pecahnya (hemoragik stroke) pembuluh darah yang bertugas mengantarkan oksigen dan nutrisi ke otak[1]. Saat ini, stroke menjadi salah satu penyakit serius yang tersebar di hampir seluruh dunia. Kejadian stroke yang tiba-tiba dapat berakibat fatal dan menyebabkan kecacatan fisik dan mental, baik pada individu yang sehat maupun pada usia lanjut.[2]

Pada penelitian ini Deteksi dan prediksi dini stroke menjadi sangat penting dalam dunia medis untuk mencegah dampak yang merugikan pada kesehatan pasien. Dalam upaya meningkatkan kualitas prediksi dan deteksi stroke, analisis dan pemodelan prediktif menggunakan teknik *Exploratory Data Analysis* (EDA) dan algoritma klasifikasi telah menjadi fokus penelitian yang menjanjikan.

EDA merupakan sebuah pendekatan yang memungkinkan kita untuk menggali wawasan yang berharga dari data kesehatan yang tersedia. Melalui EDA, kita dapat mengidentifikasi pola, hubungan, dan karakteristik yang relevan dengan deteksi stroke. Dengan demikian, Melalui EDA (*Exploratory Data Analysis*), kita bisa mendapatkan pemahaman yang lebih mendalam tentang faktor risiko dan indikator potensial yang terkait dengan stroke. Selain EDA, pemodelan prediktif juga merupakan langkah penting dalam penelitian ini. Algoritma klasifikasi telah dikembangkan untuk memprediksi kemungkinan terjadinya stroke berdasarkan data pasien. Dalam penelitian ini, kami akan memfokuskan pada perbandingan performa algoritma klasifikasi yang umum digunakan, yaitu *Support Vector Machine* (SVM) dan *Random Forest Classifier* (RFC).

Pada penelitian dari “**Project Heart Diseases**”[3] yang memperbandingkan akurasi dari beberapa model Pada metode *cross-validation*, Hasil tertinggi dari perbandingan beberapa algoritma klasifikasi yang digunakan ialah *Random Forest* dengan skor akurasi terbaik yaitu (99,7%) dibandingkan dengan *Decision Tree* (99,4%). Untuk Hasil dari *hyperparameter tuning* menunjukkan bahwa *Random Forest* memiliki skor terbaik (98,18%) dibandingkan dengan *Decision Tree*

(98,04%). Pada *learning curves*, kedua model memiliki grafik *cross-validation score* yang mirip, namun ketika nilai training semakin besar, nilai *cross-validation score* pada *Random Forest* mendekati *training score* yaitu di sekitar 98%. Pada ROC/AUC *curves*, *Random Forest* memiliki nilai area 1.00 yang menunjukkan bahwa model memiliki kemampuan yang sangat baik dalam membedakan antara kelas positif dan negatif.

Penelitian lain yang menggunakan Citra CT *Scan* pada analisis kinerja algoritma untuk klasifikasi penyakit stroke yaitu Hasil eksperimen menunjukkan bahwa algoritma *Deep Learning* mencapai tingkat akurasi tertinggi sebesar 96.79%. Pencapaian ini didapatkan melalui pengaturan berbagai parameter seperti jumlah *hidden layer*, jumlah *neuron*, *epoch*, *batch size*, *dropout*, *learning rate*, dan fungsi aktivasi. [4]

Pada Penelitian lainnya[1] Untuk meningkatkan model prediksi stroke, diusulkan arsitektur DNN dengan *pre-processing* menggunakan metode *SMOTE* dan dikombinasikan dengan *Tomek links*. Hasilnya, model ini mencapai akurasi, *f1-score*, dan AUC sebesar 0.96, 0.9611, dan 0.981. Performa tersebut lebih unggul dibandingkan dengan model hasil penelitian sebelumnya yang menggunakan klasifikasi *Extreme Gradient Boosting*.

Dalam penelitian lain, dilakukan kombinasi teknik *boosting* menggunakan *AdaBoost* untuk mengoptimasi tingkat algoritma pada algoritma *neural network* untuk memprediksi harapan hidup pasien kanker paru-paru setelah operasi bedah thoraks. Berdasarkan hasil eksperimen, disimpulkan bahwa penggunaan algoritma *AdaBoost* berhasil meningkatkan kinerja algoritma *k-nearest neighbor* dalam memprediksi harapan hidup pasien pasca operasi bedah toraks sebesar 7.45%, dari 77.66% menjadi 85.11%. [5]

Dalam penelitian lain, kesimpulan yang dihasilkan dari prediksi penyakit Stroke menggunakan Algoritma SVM adalah bahwa model prediksi yang dikembangkan dengan SVM menggunakan *Relief-f* menghasilkan akurasi data sebesar 100%. Pengujian dilakukan menggunakan 2.398 data *training* dan 1.028 data *testing* dengan total data sebanyak 3.426 baris dan lima kolom menggunakan metode *Ratio* dan *Confusion Matrix*. Penggunaan Algoritma SVM dengan *Relief-f* dan *Kernel Linear* memberikan hasil terbaik dalam penelitian tersebut. [6]

Penelitian sebelumnya telah menunjukkan keberhasilan algoritma klasifikasi dalam deteksi dan prediksi stroke. Namun, perbandingan langsung antara berbagai algoritma klasifikasi tersebut masih sangat diperlukan untuk menentukan algoritma yang paling efektif dalam hal akurasi, presisi, dan *recall*. Maka pada penelitian ini, kami bertujuan untuk menerapkan EDA pada data kesehatan yang relevan dan membandingkan performa algoritma klasifikasi yang disebutkan sebelumnya dalam deteksi dan prediksi stroke. Melalui analisis komprehensif ini, diharapkan kami dapat memberikan kontribusi yang signifikan dalam pengembangan metode deteksi dini stroke yang lebih akurat dan efisien. Oleh karena itu, penulis melakukan penelitian yang berjudul "Analisis Prediksi Deteksi Stroke dengan Pendekatan EDA dan Perbandingan Algoritma Machine Learning".

1.2 Rumusan Masalah

Adapun rumusan masalah yang menjadi fokus dalam penelitian ini antara lain:

1. Bagaimana pendekatan *Exploratory Data Analysis* (EDA) dapat digunakan untuk menganalisis data terkait stroke?
2. Bagaimana performa dari algoritma machine learning (*Support Vector Machine* (SVM) dan *Random Forest Classifier* (RFC)) dalam memprediksi deteksi stroke?
3. Bagaimana tingkat akurasi dari model prediksi tersebut?

1.3 Batasan Masalah

Beberapa batasan masalah yang diterapkan dalam penelitian ini antara lain:

1. Dalam penelitian ini, data yang digunakan telah melalui proses pengolahan dan persiapan sebelumnya.
2. Data yang peneliti gunakan terkait dengan faktor-faktor yang berpotensi mempengaruhi risiko terjadinya stroke.
3. Penelitian ini tidak mempertimbangkan aspek pengobatan atau perawatan pasien stroke.
4. Fokus penelitian ini adalah pada perbandingan performa algoritma *machine learning* dalam memprediksi deteksi stroke.

1.4 Tujuan Penelitian

Tujuan dari penelitian ini adalah:

1. Melakukan analisis terhadap hubungan antara faktor risiko tertentu (umur, indeks massa tubuh, kadar glukosa rata-rata, tekanan darah tinggi, penyakit jantung.) dengan kejadian stroke.
2. Menganalisis performa algoritma *machine learning* dalam memprediksi deteksi stroke.
3. Menentukan algoritma *machine learning* dengan performa terbaik dalam memprediksi deteksi stroke berdasarkan metrik evaluasi yang digunakan.
4. Memberikan pemahaman yang lebih baik tentang algoritma *machine learning* yang dapat digunakan dalam mendukung deteksi dini dan pencegahan stroke.

1.5 Manfaat Penelitian

Berikut adalah beberapa manfaat yang dapat diperoleh dari penelitian ini:

1. Menyediakan pemahaman yang lebih baik tentang hubungan antara faktor risiko tertentu dan kejadian stroke, yang dapat digunakan sebagai dasar untuk upaya pencegahan dan penanganan stroke.
2. Menyediakan kontribusi pada bidang penelitian kesehatan dengan memperluas pemahaman tentang penggunaan *Exploratory Data Analysis* EDA.
3. Memberikan informasi tentang performa beberapa algoritma *machine learning* dalam memprediksi deteksi stroke, yang dapat membantu para peneliti atau praktisi kesehatan dalam memilih algoritma yang paling cocok untuk digunakan.
4. Meningkatkan pemahaman dan kesadaran masyarakat mengenai pentingnya penggunaan teknologi *machine learning* dalam mendukung upaya pencegahan dan penanganan stroke secara lebih efektif.