

BAB I

PENDAHULUAN

1.1 Latar Belakang

Otak adalah salah satu organ terpenting pada tubuh manusia. Otak manusia dewasa mengandung hampir 97% jaringan saraf tubuh (Martini et al., 2015). Salah satu penyakit mematikan yang menyerang pembuluh darah pada otak adalah *stroke*. *Stroke* adalah tanda-tanda klinis yang berkembang pesat dari gangguan fokal pada fungsi otak, dengan gejala yang berlangsung selama 24 jam atau lebih, tanpa penyebab yang jelas selain yang berasal dari *vascular* (Truelsen et al., 2000).

Terdapat tiga macam *stroke*, yang pertama, *Ischemic stroke* adalah *stroke* yang terjadi karena adanya penyumbatan saluran darah pada otak, berakibat adanya darah yang membeku. Kedua, *hemorrhagic stroke* yang terjadi karena terdapat kebocoran pada pembuluh darah otak dikarenakan pembuluh darah otak yang melemah. Ketiga, *Subarachnoid hemorrhagic* terjadi akibat pendarahan pada daerah *Subarachnoid* otak (Truelsen et al., 2000). Pada tahun 2018, seseorang meninggal dikarenakan *stroke* setiap 3 menit 33 detik. *Stroke* juga merupakan penyebab kematian terbanyak ke-5 setelah penyakit jantung, kanker, penyakit pernapasan bawah kronis, dan cedera/kecelakaan yang tidak disengaja (Aparicio et al., 2021).

Menurut data statistik *American Heart Association* (AHA) (Aparicio et al., 2021), kesadaran masyarakat terhadap faktor risiko penyebab terjadinya penyakit *stroke* belum optimal. Pada tahun 2014, hanya 68,4% dari total responden hasil *survey* yang dapat mengenali lima gejala umum *stroke*. Lima gejala umum tersebut yaitu, pertama, mati rasa tiba-tiba pada bagian wajah, lengan, kaki, pada satu sisi bagian tubuh. Kedua, kebingungan tiba-tiba atau kesulitan berbicara. Ketiga, memiliki gangguan pengelihatn pada satu atau

kedua mata secara tiba-tiba. Keempat, kesulitan berjalan tiba-tiba yang disertai dengan rasa pusing serta kehilangan keseimbangan. Kelima, sakit kepala parah yang terjadi secara tiba-tiba, tanpa penyebab yang diketahui. Data statistik (Aparicio et al., 2021), juga disebutkan bahwa dalam sebuah penelitian terhadap pasien yang mengidap *atrial fibrillation* (AF), pengetahuan tentang macam-macam *stroke*, gejala umum *stroke*, dan peningkatan risiko *stroke* yang terkait dengan AF sangatlah kurang. Klasifikasi penyakit *stroke* menggunakan *machine learning* dapat memberikan *output* berupa *feature importance*, dimana atribut masukan akan diurutkan, berdasarkan seberapa pentingnya atribut tersebut berperan dalam klasifikasi penyakit *stroke*. *Feature importance* ini dapat digunakan untuk melakukan edukasi terkait faktor risiko dan gejala umum penyebab terjadinya penyakit *stroke*.

Machine learning merupakan percabangan dari *Artificial Intelligence* (AI). *Machine learning* dapat dibagi menjadi dua, *supervised learning* dan *unsupervised learning*. *Supervised learning* terdiri dari klasifikasi dan regresi. Klasifikasi adalah proses untuk mengelompokkan data set menjadi kelompok-kelompok tertentu yang memiliki kesamaan pada atribut tertentu (S. B. Kotsiantis, 2007). *Random Forest* (RF), *Extreme Gradient Boosting* (XGBoost) dan *Support Vector Machine* (SVM) adalah beberapa contoh metode klasifikasi. XGBoost adalah *classifier* yang sudah disempurnakan menggunakan *gradient boosting decision tree algorithm*. XGBoost menggabungkan prediksi dari *weak classifier* (*decision tree*), yang ditambahkan secara berurutan untuk memaksimalkan kinerja prediktif dan meminimalkan kompleksitas model (Xu et al., 2019). RF juga merupakan *classifier* yang memiliki kemampuan *ensemble*, yaitu menggabungkan *weak learner* menjadi *strong learner*. Masing-masing hasil prediksi dari *decision tree* akan dihitung secara terpisah yang kemudian digabungkan untuk memberikan prediksi akhir (Sindhu et al., 2020). Berbeda dengan SVM, yang memanfaatkan pendekatan *hyperplane* berupa garis pembatas antar kelas. SVM memaksimalkan margin dari *hyperplane* dalam

rangka memisahkan antar *target class* (S. B. Kotsiantis, 2007). Terdapat kelemahan yang dimiliki oleh metode klasifikasi yang menggunakan metode *ensemble*, dimana metode klasifikasi memerlukan waktu komputasi relatif lebih lama daripada metode klasifikasi lainnya. Hal ini dikarenakan metode *ensemble* akan melakukan proses *training* dan evaluasi terlebih dahulu sebelum melakukan *pruning* pada setiap *weak learner* yang dihasilkan (Haixiang et al., 2017).

Salah satu tantangan melakukan klasifikasi menggunakan *machine learning* adalah *imbalanced data*. *Imbalanced data* adalah tidak seimbang data, dimana satu atau beberapa kelas memiliki lebih banyak contoh (data) daripada kelas lain (Haixiang et al., 2017a). Beberapa metode untuk menangani *imbalanced data* adalah *Random Undersampling* (RUS), *Random Oversampling* (ROS), dan *Synthetic Minority Oversampling Technique* (SMOTE).

RUS adalah metode *resampling* yang secara acak menghapus data dari kelas mayoritas sampai kedua kelas memiliki jumlah objek data yang sama. ROS adalah metode *resampling*, yang secara acak membuat objek baru pada kelas minoritas sampai kedua kelas memiliki jumlah objek data yang sama. Sedangkan SMOTE adalah metode *resampling* yang akan menghasilkan objek sintesis untuk kelas minoritas berdasarkan *k-nearest neighbors* dari objek tersebut (Loyola-González et al., 2016).

Terdapat beberapa penelitian yang telah melakukan percobaan klasifikasi pada penyakit *stroke*. Penelitian yang dilakukan oleh (R. S. & Kumar Sukesh, 2016) berhasil membuktikan bahwa SVM dapat digunakan untuk melakukan diagnosis awal penyakit *stroke* menggunakan *machine learning*. Hasil tertinggi didapatkan dengan menggunakan *kernel* linear, dengan nilai akurasi 91% dan sensitivitas 100%. Sedangkan penelitian yang dilakukan oleh (Liu et al., 2019) dengan tujuan mengembangkan pendekatan *hybrid machine learning* untuk memprediksi penyakit *stroke* dengan ketidaklengkapan dan ketidakseimbangan

kelas (*imbalanced dataset*). Hasil menunjukkan bahwa hasil akurasi tertinggi didapatkan oleh metode XGBoost ($74,1 \pm 0,4 \%$). Model *classifier* pada penelitian ini juga menghasilkan output berupa *feature importance*, dan tiga atribut yang paling berpengaruh dalam proses klasifikasi adalah umur, nilai glukosa dalam darah, dan *body mass index* (BMI).

Hasil akhir yang berbeda ditemukan pada penelitian yang dilakukan oleh (Alhakami et al., 2020), dengan tujuan melakukan prediksi penyakit *stroke* yang dilakukan dengan menggabungkan dua metode *machine learning*, yaitu *clustering* dan *classification*. Penelitian ini dilakukan dengan menggunakan lima metode, *Logistic Regression*, RF, SVM, *Neural Network* dan XGBoost. Hasil akhir menunjukkan bahwa Random Forest memiliki keseluruhan nilai yang paling baik, dengan nilai akurasi 97,616 % dan F1 Score 0,806. Semua penelitian yang sudah disebutkan sebelumnya mengalami masalah *imbalanced data*, dan data masukan yang digunakan berupa faktor risiko penyebab terjadi nya penyakit *stroke*.

Berdasarkan latar belakang dan hasil dari penelitian yang sudah dilakukan sebelumnya, penelitian ini bertujuan untuk melakukan komparasi *classifier* yang dinyatakan memiliki nilai tertinggi pada tiga penelitian sebelumnya, yaitu XGboost, SVM, dan RF. Selain itu, beberapa metode untuk menangani *imbalanced data* seperti ROS, RUS, dan SMOTE juga akan diterapkan. Hasil dari penelitian ini akan membandingkan ketiga *classifier* dengan ketiga *imbalanced data handling* dalam melakukan melakukan klasifikasi penyakit *stroke*. Data yang digunakan pada penelitian ini adalah data “*The International Stroke Trial database*” yang dipublikasikan oleh *The National Center for Biotechnology Information* (NCBI). Salah satu *output* yang dihasilkan dari proses klasifikasi pada penelitian ini adalah *feature importance*, yang dapat digunakan untuk melakukan edukasi terkait faktor risiko penyebab terjadi nya penyakit *stroke*

1.2 Rumusan Masalah

Berdasarkan latar belakang penelitian, dapat dirumuskan masalah sebagai berikut:

1. Bagaimana penerapan klasifikasi penyakit stroke dengan menggunakan metode SVM, XGBoost dan RF?
2. Manakah diantara SVM, XGBoost dan RF yang lebih baik dalam melakukan klasifikasi penyakit *stroke* ?
3. Manakah diantara ROS, RUS, dan SMOTE yang lebih baik dalam menangani *imbalanced dataset* ?
4. Apa saja atribut yang paling berpengaruh dalam penentuan penyakit *stroke* menggunakan *feature importance* dari hasil klasifikasi?

1.3 Tujuan Penelitian

Berdasarkan rumusan masalah di atas, tujuan dari penelitian ini adalah untuk mengetahui bagaimana penerapan dan manfaat *machine learning* menggunakan algoritma SVM, XGBoost dan RF dalam melakukan klasifikasi penyakit *stroke* dengan data yang tidak seimbang (*imbalanced dataset*).

1.4 Manfaat Penelitian

Adapun manfaat yang didapatkan dari penelitian ini untuk pengembangan ilmu pengetahuan, diantaranya :

1. Mengetahui antara algoritma SVM, XGBoost dan RF, mana yang memberikan nilai keseluruhan tertinggi dalam melakukan klasifikasi penyakit stroke.
2. Mengetahui antara algoritma RUS, ROS dan SMOTE, mana yang dapat menangani *imbalanced dataset* paling tepat.

3. Mempermudah melakukan klasifikasi penyakit *stroke* melalui *Electronic Health Record* (EHR).

Sedangkan manfaat yang didapatkan dari penelitian ini untuk dokter dan pengamat kesehatan adalah :

1. Mengetahui faktor penting penyebab terjadinya penyakit *stroke*, sehingga dapat menjadi salah satu pertimbangan dalam melakukan diagnosis pasien.
2. Dapat melakukan deteksi dini penyakit *stroke* menggunakan *Electronic Health Record* (EHR).

1.5 Batasan Penelitian

Batasan masalah dari penelitian ini ditujukan agar ruang lingkup pembahasan tidak terlalu luas. Batasan masalah yang dimaksud adalah sebagai berikut :

1. Data yang digunakan adalah data sekunder yang didapat melalui website resmi *National Center for Biotechnology Information* (NCBI).
2. Data berisi informasi pasien yang didiagnosa terkena penyakit *ischaemic* dan *hemorrhagic stroke* selama tahun 1992-1996.
3. Pengolahan data dilakukan menggunakan *Jupyter Notebook* dengan bahasa pemrograman *Python*.
4. Tidak semua *attribute* pada data akan digunakan. *Prediction attribute* yang digunakan diambil berdasarkan penelitian (R. S. & Kumar Suresh, 2016), yang mengacu pada faktor risiko dan gejala umum terjadinya *stroke*.