BAB I
PENDAHULUAN

1.1. Latar Belakang

Stroke merupakan kondisi darurat medis yang disebabkan adanya gangguan aliran darah ke otak. Stroke dapat menyebabkan kerusakan pada otak jangka pendek, jangka panjang bahkan dapat menyebabkan kematian. Menurut *World Stroke Organization* (WSO) stroke menjadi penyebab kematian kedua dan penyebab kecacatan ketiga di dunia. Di negara maju seperti Amerika Serikat, stroke menjadi penyebab utama terjadinya kecacatan. Menurut *American Heart Association* diperkirakan tiap tahunnya lebih dari 795.000 masyarakat Amerika Serikat mengalami stroke (Benjamin dkk., 2017). Berdasarkan hasil Riset Kesehatan Dasar (Riskesdar) tahun 2018 prevalensi stroke di Indonesia sebanyak 10,9 % dengan 713,783 orang yang menderita stroke setiap tahunnya (RISKESDAS, 2018). Hal ini dapat disimpulkan bahwa penyakit stroke harus diwaspadai.

Faktor utama penyebab stroke adalah tekanan darah tinggi, jantung koroner, obesitas, dan beberapa faktor lainnya. Kemudian ada faktor lain yang tidak dapat diubah seperti jenis kelamin, usia, dan genetika. Berdasarkan faktor tersebut nantinya akan menjadi bahan pengambilan keputusan bagi dokter untuk mengidentifikasi penyakit stroke. Dalam proses identifikasi stroke tentunya diperlukan proses dan waktu yang cukup lama. Namun, mengingat tiap menit ada sel yang mati karena penyumbatan aliran pada otak. Perlu dilakukan diagnosis sedini mungkin untuk mengurangi resiko dari pasien.

Seiring perkembangan zaman, ilmu pengetahuan di bidang teknologi berkembang pesat, salah satu ilmunya adalah *data mining*. *Data mining* merupakan suatu teknik untuk menemukan pola dalam data, pola yang dapat menghasilkan pengetahuan atau memungkinkan pengambilan keputusan yang cepat dan akurat (Witten dkk., 2011). Penerapan *data mining* dapat menjadi solusi untuk melakukan prediksi penyakit stroke. Gagasan untuk mempercepat pasien dalam mengidentifikasi penyakit stroke dapat dilakukan dengan cara pengembangan aplikasi untuk mengklasifikasikan penyakit stroke.

Namun, dalam penerapannya terdapat beberapa permasalahan. Salah satu permasalahan yang dihadapi pada klasifikasi dengan menggunakan *machine learning* yaitu apabila *dataset* yang digunakan merupakan *dataset* dengan jumlah data antar kelas yang tidak seimbang. Jumlah data kelas yang tidak seimbang akan menyebabkan model yang dibentuk bias pada prediksi yang dihasilkan oleh model *machine learning* yang digunakan. Hal ini disebabkan banyaknya algoritma *machine learning* mengandalkan distribusi kelas dalam sekumpulan data latih untuk mengukur kemungkinan dalam mengamati pada setiap kelas saat model nantinya digunakan untuk melakukan prediksi. Salah satu algoritma tersebut adalah *K-Nearest Neighbor* (K-NN), dan *neural networks*. Dengan mengabaikan masalah *imbalanced*, algoritma *machine learning* akan lebih memperhatikan kelas mayoritas dan mengabaikan kelas minoritas. Dengan demikian prediksi model yang dihasilkan akan lebih baik pada kelas mayoritas dibandingkan dengan kelas minoritas (Brownlee, 2020).

Dalam membangun model *machine learning* dengan *dataset* yang *imbalanced* perlu dilakukan penanganan yang khusus. Terdapat empat kategori pendekatan sebagai solusi untuk mengatasi *imbalanced data*, di antaranya *algorithmic level*, *level data*, *cost sensitive*, *ensemble of classifiers* (Beyan & Fisher, 2015). Salah satu teknik level data yang sangat populer untuk menangani *imbalanced data* adalah *resampling* (He & Ma, 2013). *Resampling* merupakan suatu teknik yang dirancang dengan tujuan untuk mengubah distribusi kelas pada *dataset* latih. Secara garis besar teknik *resampling* dikelompokkan menjadi *oversampling*, *undersampling*, dan kombinasi metode. *Oversampling* merupakan metode yang melakukan penggandaan pada contoh data di kelas minoritas atau mensintesiskan contoh baru dari contoh data di kelas minoritas. *Undersampling* merupakan metode yang melakukan penghapusan atau memilih subset suatu contoh data dari kelas mayoritas. Kombinasi metode merupakan kombinasi dari *oversampling* dan *undersampling* (Brownlee, 2020).

Banyak metode *resampling* yang telah dijadikan gagasan baik *oversampling* maupun *undersampling*, beberapa metode *oversampling* yang banyak digunakan dan diimplementasikan diantara lain, *Random*

Oversampling, *Sythetic Minority Oversampling Technique* (SMOTE), Borderline-SMOTE, *Adaptive Synthetic Sampling* (ADASYN). Selain itu metode *undersampling* diantara lain, *random undersampling*, *Near Miss Undersampling*, *Tomek Links*s *Undersampling*, *Edited Nearest Neighboar Rule* (ENN), *One-Sided Selection* (OSS) (Brownlee, 2020).

Masalah yang dihadapi dalam pengaplikasian metode *oversampling* adalah *overfitting*. Sedangkan kelemahan utama dari *undersampling* adalah kehilangan informasi dari kelas mayoritas (Vluymans, 2019). Untuk mengatasi masalah tersebut telah dilakukan percobaan yang menunjukkan bahwa kombinasi *oversampling* dan *undersampling* secara bersamaan dapat meningkatkan kinerja dari model yang lebih baik (Brownlee, 2020). Salah satu percobaan telah dilakukan oleh Indrawati Ariani yang diterbitkan pada 12 Desember 2020. Pada penelitian tersebut metode yang digunakan adalah kombinasi teknik *oversampling* SMOTE dengan teknik *undersampling Tomek Links* dan ENN. Pada penelitian tersebut terjadi kenaikan nilai *f-measure* sebesar 0.23 dan 0.11 pada salah satu *dataset* yang diujikan (Indrawati, 2021).

K-Nearest Neighbor (K-NN) merupakan salah satu dari banyak metode klasifikasi yang dikenal sebagai metode yang efektif dan mampu bekerja dengan baik. Namun, terdapat kekurangan pada metode K-NN yakni hasil akurasi yang lebih rendah dibanding dengan metode lainnya. Hal ini disebabkan, pada proses klasifikasi setiap atribut memiliki pengaruh yang sama (Ginting dkk., 2021). Untuk mengatasi masalah tersebut akan dilakukan atribut yang berpengaruh menggunakan seleksi fitur. Maka berdasarkan latar belakang tersebut diusulkan sebuah penelitian yang berjudul "Implementasi Kombinasi Metode *Resampling* Pada Klasifikasi Penyakit Stroke Dengan Algoritma K-Nearest Neighbor dan Seleksi Fitur *Information gain*".

1.2. Rumusan Masalah

Berdasarkan latar belakang yang telah diuraikan sebelumnya, maka rumusan masalah dalam penelitian ini sebagai berikut:

1. Bagaimana performa dari algoritma K-NN untuk memprediksi stroke dengan menggunakan kombinasi *oversampling* dan *undersampling*.

2. Apakah dengan dilakukan seleksi fitur dengan menggunakan metode *information gain* dapat berpengaruh terhadap performa model.
3. Bagaimana merancang sebuah sistem prediksi untuk memprediksi penyakit stroke berdasarkan data yang telah tersedia.

1.3. Tujuan Penelitian

Berdasarkan latar belakang dan rumusan masalah, maka tujuan penelitian ini adalah sebagai berikut:

1. Menguji performa dari algoritma K-NN untuk memprediksi stroke dengan menggunakan metode *oversampling* dan kombinasi *oversampling* dengan *undersampling*.
2. Mengetahui teknik resampling yang terbaik untuk mengatasi *imbalanced data* pada *dataset* penyakit stroke.
3. Mengetahui pengaruh seleksi fitur *information gain* terhadap performa model.
4. Membangun sistem untuk memprediksi penyakit stroke.

1.4. Ruang Lingkup

Agar pembahasan penelitian ini tidak terlalu luas sehingga dibutuhkan ruang lingkup penelitian. Di bawah ini merupakan ruang lingkup untuk penelitian ini, diantaranya yaitu:

1. Fokus utama penelitian ini adalah untuk mengetahui pengaruh teknik *resampling* dan *information gain* sebagai seleksi fitur terhadap performa algoritma K-NN dalam memprediksi penyakit stroke.
2. Teknik mengatasi *imbalanced data* menggunakan teknik SMOTE, SMOTETomek, dan SMOTE-ENN.
3. Seleksi fitur dilakukan menggunakan algoritma *information gain*.

1.5. Manfaat Penelitian

Manfaat dari penelitian ini sebagai berikut:

1. Hasil dari penelitian ini memberikan pertimbangan dan sebagai bahan referensi dalam penelitian selanjutnya dalam menerapkan kombinasi

teknik *oversampling* dan *undersampling*, metode klasifikasi K-NN, dan seleksi fitur.

2. Dapat dijadikan pembanding dalam mengklasifikasikan menggunakan metode yang berbeda dengan *dataset* yang serupa.

1.6. Luaran yang Diharapkan

Adapun luaran dari penelitian ini berupa sistem pengklasifikasian penyakit stroke oleh algoritma K-NN yang telah dilakukan evaluasi performanya setelah dilakukan kombinasi teknik *resampling* dan seleksi fitur *information gain*.

1.7. Sistematika Penulisan

Penelitian ini disusun dalam 5 bab, sistematika penulisannya sebagai berikut:

BAB I PENDAHULUAN

Bagian ini menjelaskan tentang latar belakang penulisan, rumusan masalah, batasan penelitian, tujuan penulisan, manfaat penelitian, luaran yang diharapkan dan sistematika penulisan.

BAB II LANDASAN TEORI

Bagian ini membahas beberapa teori yang mendukung atau menunjang dengan pokok pembahasan yang mendasari penulisan pada penelitian ini.

BAB III METODOLOGI PENELITIAN

Bagian ini menjelaskan tentang metode yang digunakan dalam penelitian ini. Yang terdiri dari kerangka pikir untuk menggambarkan tahapan pada penelitian ini, serta tahapan perancangan perangkat lunak, alat bantu yang digunakan pada penelitian serta jadwal penelitian.

BAB IV HASIL DAN PEMBAHASAN

Bagian ini menjelaskan perancangan dan hasil yang sudah direncanakan pada bab sebelumnya, Serta menganalisis hasil dari proses tersebut.

BAB V KESIMPULAN DAN SARAN

Bagian ini menjelaskan hasil yang dapat disimpulkan dalam penelitian ini, serta saran untuk proses pengembangan selanjutnya.