

BAB I PENDAHULUAN

1.1. Latar Belakang

Stroke merupakan suatu penyakit neurologis yang terjadi ketika sel-sel otak mati karena kurangnya oksigen dan nutrisi [1]. Hal ini disebabkan oleh tersumbatnya pembuluh darah di otak oleh gumpalan ataupun pembuluh darah pecah. Stroke menjadi salah satu penyakit penyebab kematian nomor dua didunia yaitu sebanyak 11,13% yang mana terdapat 33 juta orang terkena dampaknya [2]. Data ini berdasarkan pembaruan statistik penyakit jantung dan stroke pada tahun 2015. Faktor utama stroke adalah faktor resiko yang tidak dapat diubah yaitu usia, jenis kelamin dan riwayat keluarga [3]. Faktor kedua adalah faktor resiko yang dapat diubah, seperti hipertensi, diabetes, merokok dan aktifitas olahraga.

Menurut data dari *South East Asian Medical Information Centre(SEAMIC)* Stroke merupakan salah satu dari 4 penyebab utama kematian di negara Asia sejak 1992 [4]. Indonesia menempati urutan pertama dalam hal kematian akibat stroke, ketiga di Singapura dan Filipina, dan keempat yaitu di Malaysia, Brunei dan Thailand. Oleh karena itu, penyakit ini harus diwaspadai dengan cara pencegahan dini yaitu pemeriksaan oleh dokter, mengingat tingginya angka resiko akan penyakit stroke [5].

Pendeteksian awal penyakit stroke biasanya memakan banyak waktu [6], padahal pendeteksian penyakit stroke dalam beberapa jam pertama akan meningkatkan peluang untuk pencegahan komplikasi, meningkatkan perawatan kesehatan dan manajemen pasien [1]. Terlebih untuk efek obat yang signifikan yang digunakan untuk pengobatan penyakit stroke hanya akan terlihat jika diberikan dalam tiga jam pertama saat pertama stroke terjadi. Dalam dunia kedokteran terkadang banyak pasien yang perlu dideteksi mengenai penyakit stroke yang diderita. Proses pendeteksian yang cepat dan tepat sangat diperlukan, dan hal tersebut dapat dilakukan dengan cara klasifikasi penyakit stroke yang diderita pasien.

Banyak metode dalam data mining yang dapat dipakai dalam mengklasifikasikan data, seperti *Naïve Bayes* [7], *K-Nearest Neighbor(KNN)* [7] [8], *Model Support Vector Machine(SVM)*, *Backpropagation* dan lain sebagainya. Namun untuk mengetahui metode yang paling cocok untuk

penelitian ini, dapat dilakukan perbandingan antara beberapa metode. Penelitian sebelumnya oleh Hedawoo dkk (2014) [7] yang membandingkan metode *Naïve Bayes* dan *KNN* yang digunakan untuk Mammogram(alat skrining untuk mendeteksi dini penyakit kanker payudara pada wanita yang tidak mengalami keputihan). Hasil akurasi yang didapatkan dari metode *Naïve Bayes* sebesar 64.2857% dan *KNN* sebesar 78.5714%. Pada penelitian tersebut dapat disimpulkan bahwa metode *KNN* lebih akurat dari *Naïve Bayes*.

Penelitian kedua dilakukan oleh Athoillah dkk (2015) [8] yang membandingkan metode *SVM*, *KNN*, dan *Backpropagation* dalam hal penelusuran gambar dari lima macam data set gambar. Akurasi metode *KNN* merupakan yang terbaik dari pada dua metode yang lain dengan menggunakan $k=10$, yaitu rata-rata sebesar 82% dan waktu komputasi rata-rata selama 17.99 detik. Waktu rata-rata komputasi metode *SVM* yaitu selama 25,29 detik dan *Backpropagation* 41.70 detik. Dalam penelitian ini metode *Backpropagation* merupakan yang terlama dalam komputasi. Berdasarkan perbandingan metode dari penelitian sebelumnya yang mana didapatkan metode terbaik yaitu *KNN*, maka pada penelitian ini akan menggunakan metode *KNN* untuk klasifikasi.

KNN merupakan suatu metode yang menggunakan algoritma supervised, yang mana akan mengklasifikasikan hasil data uji baru berdasarkan mayoritas dari kategori yang bertujuan untuk mengklasifikasi objek baru berdasarkan fitur dan data latih [9]. *KNN* dilakukan dengan cara mencari kelompok k objek dalam data latih yang paling dekat dengan objek pada data uji.

Dalam klasifikasi data, jumlah fitur yang banyak menyebabkan kurang akuratnya hasil klasifikasi [10]. Terlebih pada penelitian ini, dataset penyakit stroke yang didapatkan dari *Kaggle* dengan jumlah fitur sebanyak 10 yang digunakan untuk diagnosis penyakit stroke, yaitu *gender*(jenis kelamin), *age*(umur), *hypertension*(hipertensi), *heart disease*(penyakit hati), *ever married*(pernah menikah), *work type*(jenis pekerjaan), *residence type*(jenis tempat tinggal), *bmi*(body mass index), *avg glucose level*(kadar glukosa rata-rata), *smoking status*(status merokok), akan membuat keakuratan *KNN* menjadi menurun. Maka perlu digunakan suatu teknik seleksi fitur untuk meningkatkan akurasi dan efektifitas dari proses klasifikasi.

Seleksi fitur merupakan suatu teknik yang digunakan untuk mengurangi dimensi dari fitur [10]. Hal ini dilakukan untuk mendapatkan fitur-fitur yang relevan dan tidak berlebihan sehingga membuat proses klasifikasi semakin cepat dan meningkatkan akurasi dari metode yang dipakai. Pada penelitian ini akan digunakan seleksi fitur *Gain Ratio*. Seluruh fitur akan dipilih fitur terbaik pada metode prediksi dengan diterapkannya seleksi fitur *Gain Ratio*. Fitur *Gain Ratio* merupakan hasil bagi Mutual Information dan Entropy [11].

Mohammad (2018) [12] melakukan perbandingan metode seleksi fitur dalam klasifikasi teks berbahasa arab. Pada penelitian tersebut membandingkan metode seleksi fitur *Information Gain* dan *Gain Ratio* dengan beberapa metode klasifikasi, yaitu *Decesion Tree(C4.5)*, *Support Vector Machine(SVM)* dan *Naïve Bayes*. Hasil rata-rata *Precision*, *Recall* dan *F1 Measure* yang didapatkan untuk *Information gain* pada metode *C4.5* sebesar (0.616, 0.798 dan 0.692) , *SVM* sebesar (0.8536, 0.831 dan 0.840) dan *Naïve Bayes* sebesar (0.832, 0.8186 dan 0.8205). Sedangkan pada *Gain Ratio* hasil rata-rata *Precision*, *Recall* dan *F1 Measure* yang didapatkan pada metode *C4.5* sebesar (0.7156, 0.773 dan 0.740), *SVM* sebesar (0.898, 0.904 dan 0.901) dan *Naïve Bayes* sebesar (0.883, 0.899 dan 0.889). Sehingga dapat dapat disimpulkan metode seleksi fitur terbaik pada penelitian tersebut yaitu *Gain Ratio*.

Berdasarkan uraian yang telah dipaparkan dan penelitian terkait, maka dilakukan penelitian dengan judul “Seleksi Fitur *Gain Ratio* dan Model Klasifikasi *K-Nearest Neighbor* Pada Diagnosis Penyakit Stroke”. Penelitian ini dilakukan untuk mengetahui nilai akurasi dan waktu komputasi model klasifikasi *K-Nearest Neighbor* tanpa seleksi fitur, dan model klasifikasi *Nearest Neighbor* dengan seleksi fitur *Gain Ratio*.

1.2. Perumusan Masalah

1.2.1 Permasalahan (*Problem Statement*)

1. Jumlah pasien yang harus di diagnosis banyak, maka hasil diagnosis yang cepat dan tepat sangat dibutuhkan
2. Pengaruh seleksi fitur gain ratio terhadap akurasi metode *K-Nearest Neighbor*
3. Pengaruh seleksi fitur gain ratio terhadap kecepatan komputasi metode *K-Nearest Neighbor*

1.2.2 Metode Usulan

Metode yang dipakai dalam menyelesaikan permasalahan diagnosis penyakit stroke yaitu *K-Nearest Neighbor* dan seleksi fitur *Gain Ratio*.

1.2.3 Pertanyaan Penelitian

1. Seberapa akurat metode *K-Nearest Neighbor* dalam klasifikasi penyakit stroke setelah menggunakan seleksi fitur *Gain Ratio* ?
2. Seberapa cepat komputasi metode *K-Nearest Neighbor* dalam klasifikasi penyakit stroke setelah menggunakan seleksi fitur *Gain Ratio* ?

1.3. Tujuan & Manfaat

1.3.1 Tujuan Penelitian:

Mengacu pada latar belakang serta perumusan masalah yang telah diuraikan, maka tujuan dari penelitian ini yaitu :

1. Mengetahui pengaruh seleksi fitur *Gain Ratio* terhadap akurasi dari metode *K-Nearest Neighbor* dalam klasifikasi penyakit stroke.
2. Mengetahui pengaruh seleksi fitur *Gain Ratio* terhadap kecepatan komputasi dari metode *K-Nearest Neighbor* dalam klasifikasi penyakit stroke

1.3.2 Manfaat

Hasil penelitian ini dapat menjadi alternatif atau rujukan dalam membangun suatu sistem untuk diagnosis penyakit stroke dengan cepat dan akurat.

1.4. Batasan-batasan

Batasan masalah pada penelitian ini adalah dataset yang digunakan dalam penelitian ini yaitu dataset dari situs <https://www.kaggle.com/asaumya/healthcare-dataset-stroke-data> yang berjumlah 43400 data [13].

1.5 Sistematika Proposal

Sistematika penulisan laporan proposal ini terdiri dari beberapa bab, yaitu :

BAB I PENDAHULUAN

Pada bab ini membahas tentang latar belakang, perumusan masalah, tujuan, manfaat, serta sistematika penulisan laporan.

BAB II KAJIAN PUSTAKA

Pada bab ini membahas tentang kajian pustaka, penelitian sebelumnya, serta contoh penggunaan metode (perhitungan manual).

BAB III METODE USULAN

Pada bab ini membahas tentang tahapan penelitian, rancangan solusi, serta rencana pengujian.

REFERENSI

Berisi tentang jurnal-jurnal atau referensi yang dijadikan sebagai acuan dalam penelitian ini.