

BAB I PENDAHULUAN

1.1. Latar Belakang

Klasifikasi adalah kegiatan ataupun proses untuk melakukan ekstraksi data dan kemudian melakukan prediksi dari kelas ataupun label kategori untuk masing-masing data yang digunakan. Dalam istilah lain, klasifikasi yaitu suatu proses dimana dilakukan pencarian model yang bisa membedakan diantara kelas data dan bertujuan agar model itu bisa digunakan untuk melakukan prediksi kelas dari suatu obyek yang dimana masih belum diketahui kelasnya[1]. Terdapat banyak sekali metode-metode yang berbeda untuk melakukan perbandingan dan mendapatkan hasil klasifikasi terbaik pada bidang klasifikasi, terdapat banyak sekali cabang yang masih dilakukan pengembangan diantaranya yaitu algoritma genetika, pohon keputusan (decision tree), jaringan syaraf, klasifikasi Bayesian. Pada penelitian yang telah dilakukan mengatakan bahwa *decision tree* memang sangat populer dan masih banyak digunakan pada klasifikasi dikarenakan dapat memiliki hasil yang cukup memuaskan apabila dibandingkan dengan beberapa algoritma lainnya. Salah satu metode atau algoritma dari decision tree yang populer adalah C4.5. Kemudahan dari algoritma C4.5 ini adalah dapat menangani data diskrit ataupun numerik yang dimana apabila diimplementasikan untuk mencari tingkat keakurasian sistem dengan dataset yang digunakan adalah penyakit Stroke terdapat data numerik nya untuk dilakukan perhitungan sangat cocok. Stroke yaitu kerusakan pada sistem otak yang menyerang secara tiba-tiba, sangat progresif, dan cepat diakibat dari gangguan peredaran darah pada otak non traumatic[2]. Adapun tingkat keakurasian dari algoritma ini dapat diterima. Algoritma C4.5 ini mengaplikasikan konsep *Gain ratio* ataupun *entropy reduction* yang dapat memilih cabang yang paling optimal digunakan pada saat mengolah data yang ada[3].

Untuk meningkatkan tingkat akurasi dari algoritma C4.5 ke lebih tinggi, maka bisa ditambahkan metode metode untuk melakukan pre

prosesing dataset. Dari dataset yang terdapat *missing value* dan juga memiliki atribut kontinyu maka bisa dilakukan imputasi dan diskritisasi. Imputasi adalah cara mengisi *missing value* yang terdapat di dataset diisi dengan nilai yang didapat mungkin berdasarkan pada informasi ataupun data yang sudah ada dan tersedia pada dataset tersebut. Metode imputasi mean sering digunakan untuk menghilangkan missing value, dengan cara mengganti data abnormal tersebut dengan hasil dari rata rata semua nilai yang diketahui pada suatu atribut [4]. Diskritisasi adalah suatu proses dimana adanya perubahan dari data yang awalnya kontinyu menjadi data diskrit dengan interval interval yang telah ditetapkan sebelumnya. Pada studi kasus dataset penyakit stroke dari Kaggle ini dibutuhkan diskritisasi karena terdapat data kontinyu yang bisa dirubah menjadi data diskrit untuk memudahkan penelitian. Metode diskritisasi yang digunakan adalah *equal-width interval discretization*, metode ini adalah metode paling sederhana untuk membagi data kontinyu yang telah dirubah ke data diskrit ke interval interval yang telah ditetapkan berdasarkan nilai atribut [5]. Dari penelitian terkait yaitu dengan adanya imputasi dan diskritisasi berdasarkan entropy dapat membantu meningkatkan tingkat akurasi Algoritma C4.5 dari awalnya 68,61% menjadi 74,87% [6].

Dalam penelitian lain yang mengkaji dan meneliti perbandingan algoritma C4.5 dengan algoritma CART dalam klasifikasi data nilai mahasiswa prodi teknik komputer politeknik negeri padang. Dalam penelitiannya menggunakan dataset nilai semester pertama mahasiswa prodi teknik komputer dari angkatan 2006-2010. Hasil penelitian tersebut menunjukkan bahwa algoritma C4.5 memiliki tingkat keakurasian lebih besar dengan presentase 85,61% daripada algoritma CART yang menghasilkan tingkat keakruasian sebesar 84,95% [3]. Pada penelitian yang menggunakan Algoritma C4.5 lain yaitu Prediksi Potensi kebakaran hutan Studi Kasus Provinsi Kalimantan Barat menghasilkan tingkat ke akurasian sebesar 89,7859% Adapun rule yang dimiliki oleh sistem sebanyak 5 rule untuk mengklasifikasikan data tersebut [7]. Pada penelitian lain juga yang menggunakan Algoritma C4.5 yaitu

Menganalisis Kemungkinan keterlamabatan Pembayaran SPP menyatakan dengan menggunakan 4 macam partisi peembandingan dimana setiap partisi memiliki rasio pengujian data testing dan data training berbeda menghasilkan bahwa partisi ke 2 memiliki tingkat keakurasian tertinggi yaitu 75% dengan rasio 80% data training dan 20% data testing [8]. Perbandingan antara ID3 dan C4.5 pernah dilakukan yaitu A comparative study of decision tree ID3 and C4.5. Menghasilkan yaitu dari tingkat akurasi dengan banyaknya dataset yang digunakan algoritma C4.5 lebih unggul daripada ID3, sedangkan untuk waktu komputasi yang digunakan pada saat eksekusi algoritma tersebut C4.5 memiliki waktu yang sangat cepat berbanding terbalik sangat jauh daripada ID3 yang membutuhkan waktu lebih lama [9]. Pada penelitian selanjutnya digunakan kombinasi K-Means dan Algoritma C4.5 Untuk Prediksi Penderita Diabetes. Di penelitian ini menghasilkan tingkat keakurasian sebesar 72% Adapun beberapa faktor-faktor yang menyebabkan tingkat keakurasian rendah yaitu dibutuhkan data training yang banyak dan bervariasi yang memiliki unsur unsur unik pada setiap data [10]. Metode lain yang juga digunakan untuk klasifikasi adalah diantaranya ada ID3, Naïve bayes, dan TAN (*Tree Augmented Naïve*). Yang melakukan perbandingan menghasilkan bahwa algoritma ID3 yang merupakan versi sebelum Algoritma C4.5 menghasilkan tingkat keakurasian paling tinggi dibanding dengan dua algoritma lainnya dengan menggunakan pembagian dataset yaitu 80% data training dan 20% data testing [11]. Penelitian lain yang mengemukakan bahwa klasifikasi dengan menggunakan metode *Certainty factor* dapat melakukan diagnosa penyakit stroke namun pada penelitian tersebut tidak dilakukan testing tingkat akurasi apakah hasil diagnosa penyakit stroke tersebut akurat atau tidak [12].

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang diatas, maka dapat dibuat rumusan masalah sebagai berikut :

1.2.1. Permasalahan

Bagaimana hasil tingkat akurasi dari Algoritma C4.5 untuk mendiagnosis penyakit stroke dengan mengatasi *missing value* dan atribut kontinyu serta penambahan metode untuk penyeimbangan data.

1.2.2. Usulan Solusi

Solusi yang digunakan untuk dapat menghasilkan tingkat akurasi yang tinggi dibanding menggunakan algoritma C4.5 saja adalah dengan menggunakan metode algoritma C4.5 dengan tambahan metode imputasi mean untuk mengatasi *missing value* dan menggunakan diskritisasi equal-width interval discretization untuk mengatasi atribut kontinyu serta penggunaan algoritma *smote* untuk klasifikasi penyakit stroke dengan akurasi tinggi.

1.2.3. Pertanyaan Penelitian

1. Bagaimana pengaruh metode imputasi dan diskritisasi pada algoritma C4.5 pada hasil klasifikasi penyakit stroke?
2. Bagaimana pengaruh algoritma *smote* untuk mengatasi penyeimbangan data klasifikasi?

1.3 Batasan Masalah

1. Dataset yang digunakan dalam penelitian ini berupa dataset dari Kaggle yaitu *stroke prediction dataset*. Yang dimana terdapat *missing value*. Total data dari dataset ini berjumlah 5110 dan menghasilkan klasifikasi.
2. Atribut yang digunakan untuk menjadi klasifier adalah 0(tidak) dan 1(ya)
3. Pra-proses dilakukan dengan mengatasi *missing value*, diskritisasi, dan algoritma *smote*.
4. Sumber dari data repository Kaggle merahasiakan tempat asal data diambil <https://www.kaggle.com/fedesoriano/stroke-prediction-dataset>.

1.4 Manfaat dan Tujuan

1.4.1 Manfaat

Manfaat yang diperoleh dari penelitian ini bagi peneliti adalah dapat mengukur tingkat ke-akurasian sistem yang menggunakan model metode Algoritma C4.5 dan dikombinasikan dengan diskritisasi untuk mencari

tingkat akurasi paling tinggi dan juga algoritma *smote* untuk menyeimbangkan data.

1.4.2 Tujuan

Mampu mengukur tingkat keakurasian sistem menggunakan metode Algoritma C4.5 dan diskritisasi dengan dataset penyakit stroke dan fitur-fitur yang terdapat pada dataset tersebut.

1.5 Sistematika Proposal

Sistematika penulisan proposal tugas akhir ini adalah sebagai berikut :

- **BAB I PENDAHULUAN**

Bab ini membahas mengenai latar belakang tentang ruang lingkup permasalahan yang dibahas, solusi yang diusulkan agar dapat menyelesaikan permasalahan, manfaat dan tujuan dengan diterapkannya solusi yang diusulkan, serta sistematika penulisan proposal.

- **BAB II KAJIAN PUSTAKA**

Bab ini membahas tentang landasan teori dan juga penelitian-penelitian sebelumnya yang menyangkut dari penelitian ini. Materi-materi tersebut meliputi penyakit Stroke dan gejala-gejalanya, klasifikasi dan tingkat keakurasian dari metode-metode yang digunakan pada penelitian ini.

- **BAB III METODE PENELITIAN**

Bab ini membahas mengenai keseluruhan analisa dan rancangan sistem yang dilakukan, *dataset* yang digunakan, arsitektur sistem, serta arsitektur pelatihan dan pengujian.

- **BAB IV IMPLEMENTASI METODE**

Bab ini membahas tentang implementasi metode-metode yang digunakan dan juga tahapan-tahapan yang dibutuhkan untuk melakukan implementasi tersebut, serta hasil dari pengujian.

- **BAB V PENUTUP**

Bab terakhir membahas mengenai kesimpulan yang didapat dalam penelitian dan juga saran saran untuk penelitian kedepannya.