BAB I

PENDAHULUAN

1.1 LATAR BELAKANG

Machine learning, atau pembelajaran mesin, merupakan teknologi yang sangat membantu dalam menyelesaikan berbagai masalah dan mempermudah pekerjaan di berbagai bidang, termasuk di sektor kesehatan. Dalam konteks rumah sakit, *Machine learning* memungkinkan dokter untuk mendiagnosis penyakit jantung dengan lebih cepat dan efisien, mengurangi waktu yang diperlukan dalam proses diagnostik [1]. *Machine learning* dapat dimanfaatkan pada analisis stadium penyakit dengan menggunakan algoritma *Random Forest, Decision Tree, Naive Bayes, Logistic Regression, Support Vector Machine, KNN, Gradient Boosting dan Artificial Neural Network* [2].

Penyakit jantung merupakan salah satu penyakit yang berbahaya. Penyakit jantung dapat membahayakan nyawa penderitanya jika ada keterlambatan dalam penanganannya. Sistem prediksi merupakan salah satu opsi yang dapat digunakan untuk melakukan deteksi dini pada penderita penyakit jantung dengan biaya yang lebih murah dalam penggunaanya[3]. Jumlah kematian oleh penyakit jantung di seluruh dunia saat ini menurut data *World Health Organization (WHO)* adalah lebih dari 17 juta jiwa (WHO, 2024). Kematian yag disebabkan oleh penyakit jantung di Indonesia sendiri menurut data Kementerian Kesehatan pada tahun 2023 sebanyak 651.481 jiwa [2].

Dalam prosesnya, klasifikasi dapat dilakukan dengan banyak cara baik secara manual ataupun dengan bantuan teknologi. Klasifikasi yang dilakukan secara manual adalah klasifikasi yang dilakukan oleh manusia tanpa adanya bantuan dari Algortima cerdas komputer. Sedangkan klasifikasi yang dilakukan dengan bantuan teknologi, memiliki beberapa Algoritma, diantaranya *Naïve Bayes, Support Vector Machine, Decission Tree* [4]. Teknik klasifikasi merupakan salah satu dari teknik pengolahan data, Rata- rata nilai akurasi penyakit jantung adalah sebesar 81%. Namun, nilai akurasi tersebut masih belum memuaskan, karena nilai akurasi tersebut masih di bawah rata-rata sebesar 85%. Masalah baru muncul karena adanya akurasi yang rendah, sehingga dapat menimbulkan resiko kesalahan yang signifikan dalam mendeteksi penyakit jantung [5]. Pada penelitian ini, Penulis menggunakan Algoritma *Naive bayes* dan *Decision tree* (C4.5).

Algoritma *Naive Bayes* merupakan metode yang dapat digunakan untuk mengklasifikasikan sekumpulan data. Algoritma ini memanfaatkan metode probabilitas dan Statistik yang dikemukan oleh ilmuwan Inggris Thomas Bayes, yaitu memprediksi probabilitas di masa depan berdasarkan pengalaman di masa sebelumnya [6]. Metode Pengklasifikasi *Naive Bayes* adalah pengklasifikasi paling sederhana dan paling umum digunakan . *Naive Bayes* merupakan teknik prediksi berbasis probabilistik sederhana berdasar pada penerapan teorema Bayes (aturan bayes) dengan asumsi independensi (ketidak tergantungan) yang kuat (naif). Penelitian ini akan melakukan perbandingan Algoritma C4.5 dan *Naive Bayes* untuk mengetahui Algoritma yang memiliki akurasi yang lebih tinggi dalam

mendeteksi penyakit jantung [7]. Berdasarkan penelitian yang dilakukan oleh Khoirudin (2024) [5] menunjukkan bahwa Algoritma *Naïve Bayes* mencapai tingkat ketepatan yang tinggi sebesar 88.52%, Algoritma *Naïve Bayes* menghasilkan kinerja lebih unggul dalam klasifikasi penyakit jantung pada dataset yang digunakan dalam penelitian ini. Hasil penelitian yang di lakukan Ardea Bagas Wibisono dkk (2019) [8] Algoritma *Naive Bayes* menghasilkan akurasi 80,33 persen, dengan recall kelas '1' 84,8 persen, recall kelas '0' 78,4 persen, presisi kelas '1' 80,8 persen, dan presisi kelas '0' 81,2 persen. Hasil penelitian yang dilakukan oleh Dea Haganta Depari et al (2022) [9] dapat diperoleh nilai *accuracy* model *Naive Bayes* sebesar 71%.

Metode Algoritma C4.5 merupakan Algoritma yang digunakan untuk membuat *Decision Tree* atau pohon keputusan. *Decision tree* merupakan salah satu proses klasifikasi dan prediksi yang sangat kuat dan terkenal. Algoritma C4.5 sebagai klasifikasi yang paling sederhana dan mudah diimplementasikan, akan tetapi Algoritma C4.5 masih memiliki kelemahan untuk menangani data berdimensi tinggi. Dalam menggunakan C4.5 harus dilengkapi dengan variabel atau atribut pada data yang digunakan [7]. Berdasarkan penelitian yang dilakukan oleh Arni Sepharni et al. (2022) [3] Algoritma C4.5 dapat digunakan untuk memprediksi penyakit jantung dengan menggunakan beberapa teknik yang dapat digunakan untuk melihat hasil prediksi seperti *cross validation dan confussion matrix*. Tujuan dari penelitian ini adalah untuk membuat sistem prediksi yang menggunakan algoritma C4.5 untuk membuat prediksi berdasarkan data historis pasien yang akan diperiksa. Hasil yang dihasilkan dari penggunaan Algoritma C4.5 untuk membuat

prediksi menunjukkan akurasi 79%, sehingga diharapkan hasil ini akan menjadi sumber informasi untuk penelitian selanjutnya tentang prediksi dengan Algoritma C4.5.

SMOTE merupakan teknik menyeimbangkan jumlah distribusi data sampel pada kelas minoritas dengan cara menyeleksi data sampel tersebut hingga jumlah data sampel menjadi seimbang dengan jumlah sampel pada kelas mayoritas[10]. Pendekatan ini bekerja dengan membuat replikasi dari data minoritas. Replikasi tersebut dikenal dengan data sintetis (syntetic data). Metode SMOTE bekerja dengan mencari *k nearest neighbors* (yaitu ketetanggaan terdekat data sebanyak k) untuk setiap data di kelas minoritas, setelah itu dibuat data sintetis sebanyak prosentase duplikasi yang diinginkan antara data minor dan *knearest neighbors* yang dipilih secara acak.[11].

Data yang di gunakan dalam penelitian ini adalah dataset *Heart Disease Dataset* Dari situs https://www.kaggle.com/datasets/mirzahasnine/heart-disease-dataset dengan format csv. Dengan jumlah 4238 data, Data ini memiliki 16 kolom, yang terdiri dari 1 kolom label dan 15 fitur.

Berdasarkan uraian permasalahan diatas, penulis tertarik untuk melakukan penelitian dalam bentuk Tugas Akhir dengan judul **"Perbandingan Klasifikasi Penyakit Jantung Menggunakan Algoritma Naive Bayes dan C4.5 Dengan Teknik Smote"**

1.2 RUMUSAN MASALAH

Berdasarkan latar belakang diatas, maka dapat dirumuskan bahwa pokok permasalahan yang akan diteliti yaitu :

1. Bagaimana kinerja Algoritma Naive Bayes dan Algoritma C4.5 dalam mengklasifikasikan penyakit jantung berdasarkan dataset *Heart Disease Dataset*?
2. Algoritma mana yang memiliki tingkat akurasi lebih tinggi dalam melakukan klasifikasi penyakit jantung?
3. Apa saja faktor-faktor yang memengaruhi performa Algoritma *Naive Bayes* dan Algoritma C4.5 dalam proses klasifikasi penyakit jantung?

1.3 BATASAN MASALAH

Untuk menghindari terjadinya pembahasan diluar topik dan judul penelitian, maka penulis melakukan pembatasan pada batasan masalah. Adapun batasan masalah dalam penelitian ini adalah sebagai berikut :

1. Data yang digunakan adalah dataset *Heart Disease Dataset* yang diambil dari situs Kaggle, dengan jumlah 4238 data dan 16 kolom (1 kolom label dan 15 fitur).
2. Hanya dua Algoritma yang akan dibandingkan, yaitu Algoritma *Naïve Bayes* dan Algoritma C4.5.
3. Penelitian ini hanya membandingkan akurasi kedua Algoritma dalam klasifikasi penyakit jantung.

4. Penelitian ini akan mengevaluasi performa model klasifikasi menggunakan *confusion matrix* dan kurva ROC/AUC (*Receiver Operating Characterictic/Area Under The Curve).*

1.4 TUJUAN DAN MANFAAT PENELITIAN

1.4.1 Tujuan Penelitian

Tujuan dari penelitian ini yaitu :

1. Menganalisis dan membandingkan performa algoritma *Naïve Bayes* dan Algoritma C4.5 dalam klasifikasi penyakit jantung menggunakan dataset *Heart Disease Dataset.*.
2. Mengetahui tingkat akurasi masing-masing algoritma dalam melakukan prediksi terhadap penyakit jantung.
3. Mengidentifikasi faktor-faktor yang mempengaruhi tingkat akurasi klasifikasi pada Algoritma *Naïve Bayes* dan Algoritma C4.5.

1.4.2 Manfaat Penelitian

Manfaat yang didapatkan dari penelitian ini adalah :

1. Penelitian ini dapat membantu dalam mengidentifikasi tingkat risiko seseorang terkena penyakit jantung berdasarkan data yang tersedia.
2. Menentukan status penyakit jantung dengan akurasi tinggi.
3. Memberikan wawasan yang lebih dalam mengenai efektivitas dan karakteristik masing-masing Algoritma *Naïve Bayes* dan Algoritma C4.5 dalam klasifikasi penyakit jantung.

1.5 SISTEMATIKA PENULISAN

Gambaran mengenai hal-hal yang akan dibahas penelitian ini terdiri dari beberapa bab, yaitu :

BAB I : **PENDAHULUAN**

Pada bab ini menjelaskan secara umum mengenai latar belakang masalah, rumusan masalah, batasan masalah, tujuan dan manfaat penelitian, serta sistematika penulisan.

BAB II : **LANDASAN TEORI**

Dalam bab ini berisi tentang teori-teori dan pendapat para ahli sebagai penunjang yang digunakan oleh penulis yang dikutip dari jurnal, buku, dan lain-lain yang berhubungan dengan permasalahan yang akan di analisis sebagai pedoman penelitian.

BAB III : **METODOLOGI PENELITIAN**

Pembahasan bab ini menjelaskan tentang kerangka penelitian, metode yang akan di gunakan, serta alat bantu yang digunakan dalam penelitian ini.

BAB IV : **ANALISIS DAN HASIL**

Pada bab ini dilakukan perhitungan dan visualisasi dari tools *Google Colabboration* untuk menjelaskan hasil yang didapat dengan menggunakan metode *Naive Bayes* dan C4.5.

BAB V : **PENUTUP**

Pada bab ini berisi tentang penutup dari penelitian ilmiah ini yang berisi dari kesimpulan dari pembahasan pada bab-bab sebelumnya dan saran-saran yang dapat berguna pada semua pihak yang terlibat dalam penelitian ini.