

BAB I

PENDAHULUAN

Bab I berisi tentang latar belakang, rumusan masalah yang meliputi permasalahan, metode usulan dan pertanyaan penelitian, batasan masalah, tujuan penelitian, manfaat penelitian, sistematika penulisan.

1.1. Latar Belakang

Setiap tahun perkembangan zaman semakin canggih, mulai dari teknologi hingga makanan dan minuman yang beraneka ragam. Di era teknologi ini banyak makanan cepat saji yang bisa kita beli menggunakan aplikasi, sehingga tidak perlu membeli ke pasar atau toko. Dimana minat masyarakat akan makanan cepat saji sangat tinggi, seperti *frozen*, *junk food* dan jenis makanan cepat saji yang lainnya [1]. Namun tidak bisa dipungkiri, masih banyak makanan dengan kandungan gizi dan serat yang masih rendah, sedangkan kandungan lemaknya banyak, sehingga tidak baik bagi kesehatan [1], seperti dapat memicu adanya penyakit yang menyerang tubuh kita. Jika gaya hidup kita kurang sehat, maka berbagai macam penyakit dapat menyerang tubuh kita, baik itu penyakit menular ataupun tidak.

Menurut data WHO pada tahun 2018, persentase penyakit tidak menular lebih tinggi daripada penyakit menular, yaitu 73% dan setiap tahun angka kematian semakin bertambah [2]. Salah satu penyakit tidak menular dengan persentase tinggi penyebab kematian yaitu penyakit jantung, dimana data WHO pada tahun 2018 menunjukkan 35% dan hampir 10 juta jiwa yang meninggal akibat penyakit jantung pada 2020. Pada tahun 2030 mendatang, pihak *World Health Organization* (WHO) telah memprediksi bahwa angka kematian akibat penyakit jantung akan bertambah sebanyak 50 juta setiap tahunnya [3]. Penyakit jantung (*kardiovaskuler*) yaitu penyakit menakutkan yang menimbulkan kekhawatiran masyarakat, dikarenakan penyakit ini dapat menyerang tubuh secara tiba-tiba dan potensi meninggal seketika cukup tinggi. Penyakit jantung koroner dapat menyerang tubuh karena terdapat penyempitan pada aliran darah menuju jantung yang disebabkan adanya penumpukan plak pada arteri darah, serta adanya pola hidup yang kurang sehat, merokok, tingginya angka hipertensi, kolesterol, diabetes dan lainnya [4].

Salah satu teknologi yang sedang populer yaitu *machine learning*, pada *machine learning* terdapat metode klasifikasi yang berperan di berbagai bidang, salah satunya bidang kesehatan. Pada bidang ini, *machine learning* mampu mengklasifikasikan dan memprediksi penyakit hingga menyajikan data hasil diagnosis medis. Terdapat banyak metode *machine learning* yang digunakan untuk klasifikasi dan menganalisis penyakit, salah satunya yaitu *Random forest*. Terdapat penelitian yang mengatakan bahwa metode *Support Vector Machines*, *K-Nearest Neighbors*, *Neural Network*, *Logistic Regression*, *Random Forest*, dan *XGBoost* merupakan algoritma klasifikasi yang banyak digunakan pada bidang medis [5]. Pada penelitian lainnya juga dijelaskan bahwa *KNN*, *Logistic Regression*, *Naïve Bayes*, *Decision Tree*, dan *Random Forest* merupakan metode yang banyak dibahas dan dibandingkan untuk mengidentifikasi pengklasifikasian terbaik untuk prediksi penyakit jantung [6]. *Random forest* yakni salah satu algoritma *supervised learning* yang digunakan untuk klasifikasi data berdasarkan sampel dan atribut data *training* dengan metode *ensemble* [7]. Namun dalam mengklasifikasikan sebuah data, terkadang terdapat beberapa fitur/atribut yang tidak relevan [8], maka dari itu dapat digunakan teknik seleksi fitur guna mendapatkan atribut terbaik untuk proses klasifikasi menggunakan teknik *information gain*.

Pada penelitian terdahulu [9], menggunakan dua *dataset* (*heart disease dataset collected from UCI machine learning repository*, *heart disease dataset (Comprehensive) collected*) untuk objek penelitiannya dengan membandingkan 8 metode (*Naïve Bayes*, *Logistic Regression*, *Support Vector Machine*, *K-Nearest Neighbor*, *Random Forest*, *XGBoost*). Dimana metode *Random Forest* merupakan pemodelan terbaik dengan akurasi 94.12% dan *Naïve Bayes* menghasilkan akurasi terendah sebesar 79.83%.

Pada penelitian lainnya [10], menggunakan 4 model metode klasifikasi (*Adaboost*, *Bagging*, *Voting Ensemble*, *Random Forest*), datanya diperoleh dari *UCI machine learning repository* sebanyak 270 data dan 13 fitur dengan dua kelas untuk pengklasifikasian dan metode *random forest* merupakan pemodelan terbaik dengan akurasi 87%.

Pada penelitian lainnya [7], membandingkan dua metode dalam memprediksi *customer churn* (*Random Forest* dan *XGBoost*) menggunakan *dataset*

Telecommunication Customer Churn yang diperoleh dari *website Kaggle* dan metode *Random Forest* mencapai akurasi 93.5% dan AUC 0.799.

Pada penelitian lainnya [11], menggunakan dua metode yaitu *backpropagation* dan *K-Nearest Neighbor* dalam *Heart Disease (Cleveland) dataset* dengan 304 data, 14 fitur yang diperoleh dari *UCI machine learning*. Metode *Backpropagation* menghasilkan akurasi 64% dan *K-Nearest Neighbor* mencapai akurasi 66%.

Berdasarkan penelitian terdahulu yang sudah diuraikan diatas, maka penulis berinisiatif untuk menganalisis dan mengklasifikasikan adanya penyakit jantung menggunakan metode *Random Forest* guna membantu penderita atau seseorang yang mempunyai gejala mengurangi rasa khawatirnya, meminimalisir angka kematian akibat penyakit jantung dan membantu ahli medis dalam mengatasi kesalahan diagnosis [12]. Hal ini dikarenakan pada penelitian terdahulu [9][10], *Random Forest* memiliki akurasi yang tinggi daripada algoritma lainnya. Pada penelitian terdahulu [11], jika menggunakan metode berbeda (*Backpropagation* dan *K-Nearest Neighbor*) pada data penyakit jantung, tingkat akurasinya lebih rendah dari *Random Forest* yaitu *Backpropagation* menghasilkan akurasi 64% dan *K-Nearest Neighbor* mencapai akurasi 66%. Dan pada penelitian terdahulu [9], terdapat percobaan menggunakan metode *Naive Bayes* yang menghasilkan akurasi lebih rendah dari *Random Forest* yaitu 79.83%. Penelitian yang akan dikerjakan penulis memiliki persamaan dengan 2 penelitian terdahulu yang telah diuraikan diatas [10][7], yaitu sama-sama menggunakan metode *Random Forest* dan salah satu *dataset* yang digunakan pada penelitian lainnya [9] dalam penyelesaiannya. Namun perbedaanya yaitu *Heart Disease Dataset (Comprehensive)* yang digunakan pada penelitian ini diperoleh dari *website IEEE* dengan jumlah data 1190 dan 11 fitur menggunakan teknik seleksi fitur (perbandingan keakuratan model menggunakan seleksi fitur dan tanpa seleksi fitur).

1.2. Rumusan Masalah

1.2.1. Permasalahan

Berdasarkan latar belakang pada uraian diatas, permasalahan dari penelitian ini yaitu tingginya angka kematian penyakit jantung untuk setiap tahunnya

dikarenakan banyak dari masyarakat yang kurang menyadari adanya potensi penyakit jantung pada tubuhnya dan kurangnya ahli medis dalam bidang tersebut. Sehingga mengakibatkan kurang optimalnya dalam proses pengklasifikasian penyakit jantung.

1.2.2. Metode Usulan

Metode usulan untuk mengatasi permasalahan diatas yaitu menggunakan algoritma klasifikasi *Random Forest*

1.2.3. Pertanyaan Penelitian

Bagaimana perbandingan hasil ketepatan tingkat akurasi antara *dataset* dengan teknik seleksi fitur dan *dataset* tanpa seleksi fitur dari pemodelan yang dibuat menggunakan metode *Random Forest* untuk klasifikasi penyakit jantung?

1.3. Batasan Masalah

Berikut batasan masalah untuk penelitian:

1. Data yang digunakan hanya berfokus pada *dataset* penyakit jantung (*Heart Disease Dataset (Comprehensive)*) yang diperoleh dari website online IEEE <https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive>
2. Data yang digunakan sebanyak 1190, 11 fitur (*age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, restingECG, max heart rate, exercise angina*) dan kelas target dari penelitian ini ada 2 yaitu penyakit jantung (*heart disease*) berlabel 1 dan normal (tidak memiliki penyakit jantung) berlabel 0
3. Menggunakan metode *Random Forest*
4. Menggunakan teknik seleksi fitur *information gain*

1.4. Tujuan Penelitian

Berdasarkan pertanyaan penelitian pada uraian rumusan masalah diatas, tujuan dari penelitian ini yaitu menghitung dan mengetahui hasil ketepatan tingkat akurasi antara *dataset* dengan teknik seleksi fitur dan *dataset* tanpa seleksi fitur dari pemodelan yang dibuat menggunakan metode *Random Forest* untuk klasifikasi penyakit jantung.

1.5. Manfaat Penelitian

Adapun manfaat yang diperoleh dari penulis dan pembaca sebagai berikut:

1. Manfaat Bagi Penulis

Berikut manfaat yang diperoleh penulis,

- a. Mengasah kemampuan dan menerapkan ilmu yang diperoleh selama menempuh kuliah
- b. Menambah ilmu dan pengetahuan yang lebih luas lagi akan *data mining*, *machine learning* dan metode yang ada didalamnya (khususnya metode untuk klasifikasi suatu studi kasus tertentu)

2. Manfaat Bagi Pembaca

Berikut manfaat yang diperoleh pembaca,

- a. Manfaat bagi pihak medis membantu dalam pengklasifikasian penyakit jantung dan mengatasi kesalahan diagnosis
- b. Bisa mengetahui cara klasifikasi adanya potensi penyakit jantung agar bisa meminimalisir angka kematian akibat penyakit jantung
- c. Bisa dijadikan referensi dalam mengembangkan suatu sistem klasifikasi atau prediksi penyakit jantung
- d. Bisa dijadikan sebagai bahan pembelajaran untuk mengetahui tingkat akurasi dari metode *Random Forest* dalam klasifikasi penyakit jantung

1.6. Sistematika Penulisan

Adapun sistematika penulisan dari penelitian ini yaitu:

1. Bab I Pendahuluan

Pada bab I berisi tentang latar belakang, rumusan masalah yang meliputi permasalahan, metode usulan dan pertanyaan penelitian, batasan masalah, tujuan penelitian, manfaat penelitian dan sistematika penulisan

2. Bab II Kajian Pustaka

Pada bab II berisi tentang penelitian-penelitian terdahulu dan teori-teori pendukung untuk penelitian

3. Bab III Metode Usulan

Pada bab III berisi tentang kebutuhan sistem kebutuhan sistem, tahapan penelitian, data *collection*, pengelolaan data awal, tahapan metode klasifikasi *Random Forest* untuk klasifikasi penyakit jantung, evaluasi model dan skenario uji coba

4. Bab IV Hasil dan Pembahasan

Pada bab IV berisi tentang hasil proses pengelolaan data awal, hasil proses seleksi fitur, hasil proses tuning parameter menggunakan *Random SearchCV*, hasil proses pemodelan metode *Random Forest* dan hasil proses skenario uji coba pada penelitian ini

5. Bab V Penutup

Pada bab V berisi tentang kesimpulan dan saran untuk pengembangan penelitian selanjutnya.