

# BAB I

## PENDAHULUAN

### 1.1 Latar Belakang

Hingga saat ini, penyakit jantung masih menduduki posisi pertama sebagai penyebab kematian tertinggi secara *global*. *Data World Health Organization* [1] menunjukkan bahwa angka kematian akibat gangguan kardiovaskular mencapai kurang lebih 17,9 jiwa setiap tahunnya, angka yang setara dengan 32% dari total kematian di seluruh dunia. Dampak yang ditimbulkan tidak sebatas pada persoalan kesehatan semata, melainkan turut membebani sistem pembiayaan layanan kesehatan di berbagai negara, khususnya negara-negara berkembang yang masih menghadapi keterbatasan infrastruktur dan sumber daya medis.

Di Indonesia, gambaran epidemiologis penyakit jantung menunjukkan perkembangan yang patut diwaspadai. Berdasarkan hasil riset Kesehatan Dasar (Riskesdas) [2], prevalensi penyakit jantung koroner tercatat sebesar 1,5%, melonjak tiga kali lipat dibandingkan angka 0,5% yang dilaporkan pada tahun 2013. Peningkatan ini tidak terlepas dari pergeseran pola hidup masyarakat [3], bertambahnya proporsi kelompok usia lanjut, serta tingginya prevalensi berbagai faktor risiko. Faktor-faktor risiko yang paling umum di antaranya adalah tekanan darah tinggi sebesar 34,1%, kelebihan berat badan sebesar 21,8%, serta diabetes melitus sebesar 10,9% pada populasi Indonesia. Realitas tersebut menempatkan penyakit jantung sebagai salah satu isu kesehatan yang mendapat perhatian serius dalam kebijakan kesehatan nasional.

Salah satu persoalan mendasar dalam penanganan penyakit jantung di Indonesia adalah keterlambatan deteksi. Sebagian besar pasien baru mengetahui kondisinya ketika penyakit telah berada pada tahap lanjut, mengingat gejala awal yang muncul kerap bersifat tidak khas atau bahkan tidak dirasakan sama sekali. Di samping itu, modalitas pemeriksaan klinis yang lazim digunakan, seperti ekokardiografi dan kateterisasi jantung, menuntut pembiayaan yang tidak sedikit, ketersediaan fasilitas yang memadai, serta kompetensi tenaga spesialis persyaratan yang belum dapat dipenuhi secara merata, terutama di daerah terpencil dan pedesaan.

Kemajuan di bidang kecerdasan buatan, utamanya *machine learning*, membuka dimensi baru guna mengidentifikasi risiko penyakit jantung sejak tahap awal, pendekatan analitis terhadap data digunakan sebagai dasar proses evaluasi. Melalui pendekatan ini, sistem komputer mampu mengenali pola yang tersembunyi dalam data rekam medis pasien dan menghasilkan estimasi risiko secara otomatis tanpa prosedur invasif. Kemampuan tersebut menjadikan *machine learning* potensial untuk difungsikan sebagai mekanisme skrining awal guna mengidentifikasi individu berisiko tinggi sebelum pemeriksaan lanjutan dilakukan.

Sejumlah algoritma *machine learning* telah dikaji dalam konteks prediksi penyakit jantung, diantaranya beberapa teknik klasifikasi yang diimplementasikan meliputi *Decision Tree* [4], *Random Forest* [4], *K-Nearest Neighbor* [5], *Support Vector Machine* [4], *Naïve Bayes* [4], serta *Logistic Regression* [6]. Sreekumari [7] menemukan bahwa *Decision Tree* mengungguli algoritma lainnya dengan akurasi 92% dalam perbandingan lima algoritma, sementara Mohammed [8] melaporkan tujuh algoritma yang kesemuanya mencapai akurasi 97%. Meskipun demikian, sebagian besar studi tersebut bertumpu pada dataset internasional seperti *Cleveland Heart Disease Dataset* dari UCI [9] yang membatasi relevansinya terhadap karakteristik klinis dan demografis populasi Indonesia. Penelitian yang secara spesifik menggunakan data Indonesia dalam skala besar dengan cakupan fitur yang komprehensif masih sangat terbatas.

Meskipun berbagai penelitian menunjukkan tingkat akurasi yang tinggi dalam prediksi penyakit jantung, masih terdapat beberapa keterbatasan yang belum banyak dibahas. Sebagian besar penelitian menggunakan dataset internasional dengan jumlah data yang relatif terbatas sehingga belum tentu merepresentasikan karakteristik populasi Indonesia. Selain itu, fokus penelitian terdahulu umumnya terletak pada peningkatan akurasi model, sementara aspek interpretabilitas model dan kemampuan mendeteksi pasien berisiko tinggi (*recall*) masih belum menjadi perhatian utama. Padahal dalam konteks kesehatan, kegagalan mendeteksi pasien berisiko dapat menimbulkan konsekuensi yang lebih serius dibandingkan kesalahan klasifikasi lainnya.

Dalam penerapan *machine learning* di bidang kesehatan, akurasi prediksi saja tidak cukup menjadi satu-satunya tolok ukur pemilihan algoritma. Aspek lain seperti kemampuan model untuk dijelaskan secara logis, kemudahan integrasi dalam sistem klinis, serta kapasitasnya dalam mendukung proses pengambilan keputusan medis turut perlu dipertimbangkan. Dalam konteks prediksi penyakit jantung, model yang tidak hanya presisi secara statistik tetapi juga mampu memaparkan faktor risiko secara transparan akan lebih mudah diterima dan diimplementasikan oleh tenaga kesehatan di lapangan.

Dari sejumlah alternatif algoritma yang tersedia, *Logistic Regression* dipilih dalam penelitian ini atas beberapa pertimbangan. Algoritma ini menghasilkan output berupa nilai probabilitas yang lebih informatif dibandingkan klasifikasi biner semata, sehingga tenaga medis dapat menilai tingkat risiko pasien secara lebih terukur. Selain itu, model ini juga bersifat interpretabel karena koefisien yang dihasilkan dapat digunakan untuk melihat pengaruh masing-masing variabel terhadap risiko penyakit, sehingga mendukung pengambilan keputusan berbasis bukti.

Dari sisi performa, penelitian oleh Daharwal [5] dan Ayankoya [10] menunjukkan bahwa *Logistic Regression* memiliki hasil yang kompetitif dengan akurasi masing-masing di atas 89%, bahkan mencapai 91,4% pada beberapa studi. Hal ini menunjukkan bahwa meskipun merupakan model yang relatif sederhana,

*Logistic Regression* tetap mampu memberikan performa yang kompetitif serta menawarkan tingkat interpretabilitas yang lebih tinggi dibandingkan beberapa metode klasifikasi lainnya.

Secara teoritis, *Logistic Regression* merupakan model klasifikasi *linear* dengan proses optimisasi *convex* sehingga memiliki kompleksitas komputasi yang relatif lebih rendah dibandingkan metode berbasis *ensemble* seperti *Random Forest* maupun *boosting*. Kondisi ini menjadikan *Logistic Regression* [11] lebih efisien dalam proses pelatihan maupun prediksi, terutama pada sistem yang membutuhkan respon cepat dan sumber daya komputasi terbatas. Kombinasi dari interpretabilitas, performa yang kompetitif berdasarkan penelitian terdahulu, serta efisiensi komputasi tersebut menjadikan *Logistic Regression* sebagai pilihan yang tepat dalam penelitian ini.

Dataset yang digunakan dalam penelitian ini adalah *Heart Attack Prediction Indonesia Dataset* [12] yang dapat diakses secara publik melalui platform Kaggle. Dataset ini dirancang dengan konteks populasi Indonesia dan memuat 158.355 entri data dengan 27 fitur yang mencakup informasi demografis, kondisi klinis, perilaku kesehatan, serta parameter laboratorium. Volume data yang besar dan kelengkapan fitur yang dimiliki dataset ini memberikan landasan yang kokoh untuk pengembangan model prediksi yang representatif bagi populasi Indonesia. Berdasarkan kondisi tersebut, diperlukan penelitian yang memanfaatkan dataset kesehatan Indonesia dalam skala besar, memiliki kemampuan deteksi risiko yang baik, serta tetap mampu memberikan interpretasi yang jelas terhadap faktor-faktor yang memengaruhi risiko penyakit jantung.

Penelitian ini memiliki beberapa pembeda dibandingkan penelitian terdahulu. Pertama, penelitian menggunakan *Heart Attack Prediction Indonesia Dataset* [12] yang terdiri dari 158.355 data dan dirancang berdasarkan konteks populasi Indonesia. Kedua, penelitian tidak hanya menerapkan algoritma *Logistic Regression*, tetapi juga mengintegrasikan teknik *Synthetic Minority Oversampling Technique (SMOTE)* untuk menangani ketidakseimbangan kelas, *Grid Search Cross Validation* untuk memperoleh kombinasi *hyperparameter* terbaik, serta

penyesuaian *threshold* klasifikasi untuk meningkatkan kemampuan model dalam mendeteksi pasien berisiko. Pendekatan tersebut dipilih karena pada kasus prediksi penyakit jantung kemampuan model dalam mendeteksi pasien berisiko (*recall*) menjadi aspek yang sangat penting untuk meminimalkan kemungkinan pasien berisiko tinggi tidak teridentifikasi sejak dini.

Ketiga, performa model tidak hanya dievaluasi berdasarkan *accuracy*, tetapi juga *precision*, *recall*, *F1-score*, dan *ROC-AUC*. Fokus terhadap *recall* dipilih karena tujuan utama penelitian adalah meminimalkan kemungkinan pasien berisiko tinggi tidak terdeteksi sejak dini (*false negative*). Oleh karena itu, model dengan kemampuan deteksi yang lebih tinggi dinilai lebih sesuai untuk mendukung proses skrining awal penyakit jantung dibandingkan model yang hanya berorientasi pada nilai *accuracy* yang tinggi.

Keempat, penelitian tidak hanya mengevaluasi performa model berdasarkan metrik klasifikasi, tetapi juga menganalisis pengaruh setiap faktor risiko menggunakan pendekatan *Odds Ratio* sehingga hasil model dapat diinterpretasikan secara lebih mudah dalam konteks medis. Kombinasi pendekatan tersebut diharapkan dapat menghasilkan model yang tidak hanya memiliki performa prediksi yang baik, tetapi juga lebih sensitif dalam mendeteksi pasien berisiko serta lebih mudah diinterpretasikan dalam konteks pengambilan keputusan klinis. Berdasarkan kesenjangan dan pembeda tersebut, penelitian ini tidak hanya berfokus pada pembangunan model prediksi, tetapi juga pada upaya meningkatkan kemampuan deteksi risiko serta menjelaskan faktor-faktor yang berkontribusi terhadap terjadinya penyakit jantung pada populasi Indonesia.

## 1.2 Rumusan Masalah

Berdasarkan latar belakang yang telah dipaparkan, berikut adalah hasil rumusan masalah dalam penelitian ini.

1. Bagaimana membangun model prediksi risiko penyakit jantung menggunakan algoritma *machine learning* berdasarkan *Heart Attack Prediction Indonesia Dataset*?

2. Bagaimana performa model *Logistic Regression* yang dioptimasi menggunakan *Synthetic Minority Oversampling Technique (SMOTE)*, *Grid Search Cross Validation*, dan penyesuaian *threshold* klasifikasi dalam memprediksi risiko penyakit jantung berdasarkan metrik *akurasi*, *presisi*, *recall*, *F1-score*, dan *ROC-AUC*?
3. Faktor-faktor risiko kesehatan apa yang memiliki pengaruh paling signifikan terhadap prediksi risiko penyakit jantung berdasarkan hasil analisis model *Logistic Regression*?

### 1.3 Batasan Masalah

Agar penelitian ini dapat dilaksanakan secara terarah dan terukur, ditetapkan batasan-batasan masalah sebagai berikut.

1. *Dataset* yang dimanfaatkan dalam penelitian ini berasal dari *Kaggle* dengan nama *Heart Attack Prediction Indonesia Dataset* [12], yang memuat 158.355 data rekaman serta terdiri atas 27 atribut dan satu *variabel target*, yaitu *heart\_attack*.
2. Algoritma utama yang digunakan adalah *Logistic Regression* dengan penerapan *Synthetic Minority Oversampling Technique (SMOTE)*, *optimasi hyperparameter* menggunakan *Grid Search Cross Validation*, serta penyesuaian *threshold* klasifikasi untuk meningkatkan kemampuan deteksi pasien berisiko.
3. Pemrosesan data dan pembangunan model dilakukan menggunakan bahasa pemrograman *Python* dengan *library Scikit-learn*, *Pandas*, *NumPy*, dan *Matplotlib*.
4. Penilaian kinerja model dilakukan menggunakan metrik *akurasi*, *presisi*, *recall*, *F1-score*, *ROC-AUC*, dan *confusion matrix*. *Optimasi hyperparameter* dilakukan menggunakan *Grid Search Cross Validation* dengan skema *5-Fold Cross Validation*.
5. Implementasi sistem pada penelitian ini dibatasi pada *deployment* model dalam bentuk aplikasi prediksi berbasis *web* untuk keperluan demonstrasi

dan pengujian model, serta tidak mencakup integrasi dengan sistem informasi rumah sakit atau penerapan langsung pada lingkungan klinis.

#### **1.4 Tujuan dan Manfaat Penelitian**

Dalam penelitian ini, tujuan dan manfaat dirumuskan sebagai pedoman serta luaran yang diharapkan. Berikut adalah tujuan dan manfaat dari penelitian ini.

##### **1.4.1 Tujuan Penelitian**

Penelitian ini bertujuan untuk mencapai sasaran-sasaran komprehensif sebagai berikut.

1. Membangun model prediksi risiko penyakit jantung menggunakan algoritma *Logistic Regression* berdasarkan *Heart Attack Prediction Indonesia Dataset* [12].
2. Mengevaluasi performa model *Logistic Regression* yang dioptimasi menggunakan *Synthetic Minority Oversampling Technique (SMOTE)*, *Grid Search Cross Validation*, dan penyesuaian *threshold* klasifikasi berdasarkan metrik *akurasi*, *presisi*, *recall*, *F1-score*, dan *ROC-AUC*.
3. Mengidentifikasi faktor-faktor risiko kesehatan yang memiliki pengaruh paling signifikan terhadap risiko penyakit jantung berdasarkan hasil analisis model *Logistic Regression*.

##### **1.4.2 Manfaat Penelitian**

Diharapkan hasil penelitian ini dapat membawa manfaat bagi berbagai pihak, sebagai berikut.

###### **1. Bagi Dunia Akademis**

Penelitian ini dapat menjadi referensi dan bahan perbandingan bagi penelitian selanjutnya yang berkaitan dengan penerapan *machine learning* dalam prediksi penyakit jantung, khususnya dengan konteks data populasi Indonesia.

###### **2. Bagi Tenaga Medis dan Institusi Kesehatan**

Temuan penelitian ini berpotensi menjadi landasan pengembangan sistem pendukung keputusan klinis yang memfasilitasi tenaga medis dalam mendeteksi pasien berisiko tinggi penyakit jantung secara lebih

efisien dan berbasis bukti data.

### 3. Bagi Masyarakat

Melalui adanya penelitian ini, masyarakat diharapkan menjadi lebih sadar mengenai faktor-faktor risiko penyakit jantung, sehingga mendorong gaya hidup yang lebih sehat dan pemeriksaan kesehatan secara berkala.

## 1.5 Sistematika Penulisan

Penyusunan sistematika penulisan ini bertujuan untuk memberikan uraian yang terstruktur terkait cakupan serta isi setiap bab pada studi ini. Adapun susunannya dapat dijabarkan sebagai berikut.

### 1. BAB I PENDAHULUAN

Pada bab ini dibahas mengenai latar belakang, perumusan dan pembatasan masalah, tujuan serta manfaat penelitian, dan juga sistematika penulisan laporan.

### 2. BAB II LANDASAN TEORI

Bab ini memaparkan kajian penelitian terdahulu yang relevan serta teori-teori yang mendasari penelitian, meliputi konsep penyakit jantung, *machine learning*, *algoritma Logistic Regression*, teknik *preprocessing data*, dan metode evaluasi model, beserta *framework* serta perangkat lunak yang dimanfaatkan.

### 3. BAB III METODOLOGI PENELITIAN

Bab ini memberikan gambaran umum mengenai objek yang diteliti, tahapan metode penelitian yang ditempuh mulai dari pengumpulan data, *preprocessing*, pembangunan model, hingga evaluasi, serta variabel-variabel yang digunakan dalam penelitian.

#### **4. BAB IV ANALISIS DAN HASIL PENELITIAN**

Bab ini menyajikan hasil eksplorasi dan *preprocessing data*, proses pembangunan *model Logistic Regression*, penerapan *SMOTE*, optimasi *hyperparameter* menggunakan *Grid Search Cross Validation*, penyesuaian *threshold* klasifikasi, evaluasi performa model, serta analisis faktor-faktor yang berpengaruh terhadap prediksi risiko penyakit jantung melalui pendekatan *Odds Ratio*.

#### **5. BAB V PENUTUP**

Bab ini memuat simpulan yang diperoleh dari keseluruhan proses dan hasil penelitian, disertai saran yang dapat menjadi acuan bagi penelitian selanjutnya.

