BAB 2
LANDASAN TEORI

2.1 Penyakit jantung

Penyakit jantung merupakan kelompok gangguan yang memengaruhi fungsi dan struktur jantung serta sistem pembuluh darah [1]. Penyakit ini sering disebut sebagai penyakit kardiovaskular dan mencakup berbagai kondisi, seperti penyakit jantung koroner, gagal jantung, aritmia, dan penyakit katup jantung. Penyakit jantung umumnya berkembang secara perlahan dan sering kali tidak menunjukkan gejala yang jelas pada tahap awal, sehingga deteksi dini menjadi sangat penting.

Penyakit jantung koroner merupakan salah satu jenis penyakit jantung yang paling umum terjadi. Kondisi ini disebabkan oleh penyempitan atau penyumbatan pembuluh darah koroner akibat penumpukan plak lemak (aterosklerosis). Penyumbatan tersebut dapat menghambat aliran darah ke otot jantung dan berpotensi menyebabkan serangan jantung.

2.1.1 Faktor Risiko Penyakit Jantung

Faktor risiko penyakit jantung dapat dibedakan menjadi faktor yang tidak dapat dimodifikasi [1] dan faktor yang dapat dimodifikasi. Faktor yang tidak dapat dimodifikasi meliputi usia dan jenis kelamin. Risiko penyakit jantung cenderung meningkat seiring bertambahnya usia, serta lebih tinggi pada laki-laki dibandingkan perempuan [10].

Sementara itu, faktor yang dapat dimodifikasi meliputi tekanan darah tinggi, kadar kolesterol yang tidak normal, diabetes, obesitas, kurangnya aktivitas fisik, serta kebiasaan merokok [1]. Faktor-faktor ini memiliki hubungan erat dengan kondisi kesehatan jantung dan sering digunakan sebagai variabel dalam penelitian prediksi risiko penyakit jantung.

2.2 Data Mining dan Machine Learning

Data mining merupakan proses penggalian pola, hubungan, atau informasi yang bernilai dari kumpulan data berukuran besar [7]. Proses ini melibatkan beberapa tahapan, mulai dari pengumpulan data, pembersihan data, pemilihan fitur, hingga pembangunan model.

Machine learning adalah bagian dari data mining yang berfokus pada pengembangan algoritma yang memungkinkan sistem komputer untuk belajar dari data tanpa diprogram secara eksplisit [11]. Dalam konteks kesehatan, machine learning banyak digunakan untuk mendukung pengambilan keputusan klinis, seperti diagnosis penyakit, prediksi risiko, dan analisis prognosis.

Berdasarkan cara pembelajarannya, machine learning dibagi menjadi tiga jenis utama, yaitu supervised learning, unsupervised learning, dan reinforcement learning. Penelitian ini menggunakan pendekatan supervised learning karena dataset yang digunakan memiliki label kelas yang jelas, yaitu kondisi penyakit jantung.

2.3 Klasifikasi dalam Machine Learning

Klasifikasi merupakan salah satu teknik dalam supervised learning yang bertujuan untuk memetakan data ke dalam kelas tertentu berdasarkan fitur yang dimiliki. Dalam penelitian ini, klasifikasi digunakan untuk menentukan apakah seseorang termasuk ke dalam kelompok berisiko penyakit jantung atau tidak.

Algoritma klasifikasi bekerja dengan mempelajari pola dari data latih (training data) dan kemudian digunakan untuk memprediksi kelas data uji (testing data). Beberapa algoritma klasifikasi yang umum digunakan antara lain Logistic Regression, Support Vector Machine, Decision Tree, dan Random Forest.

2.4 Algoritma Random Forest

Random Forest merupakan algoritma machine learning berbasis ensemble learning yang dikembangkan oleh Leo Breiman. Algoritma ini membangun banyak pohon keputusan (decision tree) selama proses pelatihan dan menghasilkan prediksi berdasarkan hasil agregasi dari seluruh pohon tersebut, umumnya melalui metode voting untuk klasifikasi.

Keunggulan utama Random Forest terletak pada kemampuannya untuk mengurangi overfitting yang sering terjadi pada model decision tree tunggal. Hal ini dicapai dengan menerapkan dua konsep utama, yaitu *bootstrap sampling* dan pemilihan fitur secara acak pada setiap node pohon.

2.4.1 Cara Kerja Random Forest

Proses kerja Random Forest dapat dijelaskan melalui tahapan berikut:

A. Melakukan pengambilan sampel data secara acak dengan pengembalian (*bootstrap sampling*) dari dataset pelatihan.

B. Membangun sejumlah *decision tree* berdasarkan sampel data tersebut.

C. Pada setiap node pohon, hanya sebagian fitur yang dipilih secara acak untuk menentukan pemisahan terbaik.

D. Melakukan prediksi dengan menggabungkan hasil prediksi dari seluruh pohon menggunakan mekanisme voting mayoritas.

2.4.2 Parameter Random Forest

Beberapa parameter penting dalam Random Forest antara lain:

A. *n_estimators*, yaitu jumlah pohon keputusan yang dibangun.

B. *max_depth*, yaitu kedalaman maksimum pohon.

C. *min_samples_split*, yaitu jumlah minimum sampel untuk melakukan pemisahan node.

D. *random_state*, digunakan untuk memastikan hasil yang konsisten.

2.5 Ketidakseimbangan Kelas

Ketidakseimbangan kelas merupakan kondisi di mana distribusi data antar kelas dalam sebuah dataset tidak proporsional [12]. Dalam konteks klasifikasi biner, ketidakseimbangan kelas terjadi apabila jumlah sampel pada satu kelas (kelas mayoritas) jauh lebih banyak dibandingkan kelas lainnya (kelas minoritas). Kondisi ini sangat umum ditemui pada dataset medis, termasuk dataset penyakit jantung. Ketidakseimbangan kelas dapat berdampak negatif terhadap performa model *machine learning*. Model yang dilatih pada dataset tidak seimbang cenderung bias terhadap kelas mayoritas, menghasilkan akurasi keseluruhan yang tampak tinggi namun recall pada kelas minoritas yang sangat rendah. Dalam konteks medis, kelas minoritas umumnya merepresentasikan kasus penyakit (positif), sehingga model yang gagal mendeteksi kasus positif (false negative) dapat membawa konsekuensi klinis yang serius. Terdapat beberapa pendekatan untuk mengatasi ketidakseimbangan kelas, yaitu pendekatan pada level data (resampling), pendekatan pada level algoritma (cost-sensitive learning), dan pendekatan hybrid.

Penelitian ini menggunakan pendekatan resampling karena fleksibilitasnya yang tinggi dan kemampuannya untuk diterapkan bersama algoritma klasifikasi apapun tanpa memodifikasi algoritma itu sendiri.

2.6 Teknik Resampling

Teknik resampling merupakan pendekatan pada level data yang bertujuan untuk menyeimbangkan distribusi kelas dalam dataset sebelum proses pelatihan model dilakukan [5]. Terdapat dua strategi utama dalam teknik resampling, yaitu *oversampling* (menambah jumlah sampel pada kelas minoritas) dan *undersampling* (mengurangi jumlah sampel pada kelas mayoritas). Pendekatan *hybrid* menggabungkan keduanya.

2.6.1 SMOTE (Synthetic Minority Over-sampling Technique)

SMOTE adalah teknik oversampling yang dikembangkan oleh Chawla et al. untuk mengatasi ketidakseimbangan kelas dengan cara membuat sampel sintetis baru pada kelas minoritas, bukan hanya menduplikasi sampel yang ada [5]. Sampel sintetis dibuat dengan cara menginterpolasi antara sampel minoritas yang ada dengan tetangga terdekatnya (k-nearest neighbors).

Langkah kerja SMOTE adalah sebagai berikut:

A. Untuk setiap sampel pada kelas minoritas, identifikasi k tetangga terdekat dari kelas yang sama.

B. Pilih secara acak satu dari k tetangga tersebut.

C. Buat sampel sintetis baru pada titik di antara sampel asli dan tetangga yang dipilih menggunakan interpolasi linier.

D. Ulangi proses hingga jumlah kelas minoritas mencapai proporsi yang diinginkan.

2.6.2 SMOTE-ENN (SMOTE combined with Edited Nearest Neighbor)

SMOTE-ENN merupakan teknik resampling hybrid yang menggabungkan oversampling menggunakan SMOTE dengan pembersihan data menggunakan metode Edited Nearest Neighbor (ENN) [5]. ENN bekerja dengan menghapus

sampel yang misklasifikasi oleh k tetangga terdekatnya, baik dari kelas mayoritas maupun kelas minoritas.

Proses kerja SMOTE-ENN terdiri dari dua tahap utama:

A. SMOTE diterapkan untuk melakukan oversampling pada kelas minoritas, sehingga jumlah sampel kelas minoritas meningkat.

B. ENN diterapkan pada seluruh dataset hasil SMOTE. Setiap sampel yang berbeda kelas dengan mayoritas tetangga terdekatnya akan dihapus.

Keunggulan SMOTE-ENN dibandingkan SMOTE adalah kemampuannya membersihkan sampel ambigu dan noise di perbatasan antar kelas, sehingga menghasilkan batas keputusan (decision boundary) yang lebih bersih. Hal ini umumnya menghasilkan performa klasifikasi yang lebih baik pada data uji.

2.6.3 SMOTE-Tomek (SMOTE combined with Tomek Links)

SMOTE-Tomek adalah teknik resampling hybrid lain yang menggabungkan oversampling SMOTE dengan metode undersampling Tomek Links [5]. Tomek Links adalah pasangan sampel dari dua kelas berbeda yang saling menjadi tetangga terdekat satu sama lain. Sepasang sampel (*xi*, *xj*) disebut Tomek Link apabila tidak ada sampel lain *xk* yang lebih dekat ke *xi* daripada *xj*, atau lebih dekat ke *xj* daripada *xi*.

Proses kerja SMOTE-Tomek adalah sebagai berikut:

A. SMOTE diterapkan terlebih dahulu untuk melakukan oversampling pada kelas minoritas.

B. Tomek Links diidentifikasi pada seluruh dataset hasil SMOTE.

C. Sampel yang termasuk dalam Tomek Links dihapus, biasanya sampel dari kelas mayoritas atau keduanya tergantung konfigurasi.

Perbedaan antara SMOTE-ENN dan SMOTE-Tomek terletak pada agresivitas pembersihan datanya. ENN cenderung lebih agresif dalam menghapus sampel, sedangkan Tomek Links hanya menghapus pasangan sampel yang berada di perbatasan kelas. Akibatnya, SMOTE-ENN biasanya menghasilkan dataset yang lebih kecil namun lebih bersih, sementara SMOTE-Tomek menghasilkan batas keputusan yang lebih halus.

2.7 Optimasi Hyperparameter dengan GridSearchCV

Hyperparameter adalah parameter model yang nilainya ditentukan sebelum proses pelatihan dan tidak dipelajari dari data secara langsung. Pemilihan nilai hyperparameter yang tepat sangat berpengaruh terhadap performa model. Proses pencarian nilai hyperparameter optimal dikenal sebagai hyperparameter tuning atau optimasi hyperparameter. GridSearchCV (Grid Search Cross-Validation) adalah metode exhaustive search yang mencoba seluruh kombinasi nilai hyperparameter yang telah ditentukan dalam sebuah grid, dan mengevaluasi setiap kombinasi menggunakan cross-validation [11]. Metode ini menjamin bahwa kombinasi hyperparameter terbaik dalam ruang pencarian yang didefinisikan akan ditemukan.

Proses kerja GridSearchCV adalah sebagai berikut:

A. Pengguna mendefinisikan grid pencarian, yaitu kumpulan nilai yang akan dicoba untuk setiap hyperparameter.

B. GridSearchCV mencoba seluruh kombinasi nilai hyperparameter dalam grid tersebut.

C. Setiap kombinasi dievaluasi menggunakan k-fold cross-validation pada data latih.

D. Kombinasi hyperparameter yang menghasilkan skor evaluasi terbaik dipilih sebagai konfigurasi optimal.

Keunggulan GridSearchCV adalah sistematisitas dan reprodusibilitasnya. Namun, kelemahannya adalah kompleksitas komputasi yang tinggi apabila ruang pencarian sangat besar. Pada penelitian ini, GridSearchCV digunakan untuk mengoptimalkan parameter n_estimators, max_depth, dan min_samples_split pada model Random Forest.

2.8 Explainable AI (XAI)

Explainable Artificial Intelligence (XAI) merupakan bidang dalam kecerdasan buatan yang berfokus pada pengembangan metode dan teknik untuk membuat keputusan model AI dapat dipahami dan dijelaskan kepada manusia [9]. XAI muncul sebagai respons terhadap meluasnya penggunaan model machine learning yang bersifat black-box, di mana mekanisme internal model sulit dipahami bahkan oleh para ahli. Dalam konteks medis, XAI memiliki peran yang sangat

krusial. Tenaga medis memerlukan transparansi dan penjelasan yang logis sebelum dapat mempercayai dan menggunakan rekomendasi dari sistem berbasis AI dalam praktik klinis. Tanpa interpretabilitas yang memadai, model prediksi yang akurat sekalipun sulit diterima sebagai alat bantu keputusan klinis yang sah secara profesional dan etis. Terdapat dua kategori utama dalam XAI berdasarkan cakupan penjelasannya, yaitu penjelasan global yang menggambarkan perilaku model secara keseluruhan pada seluruh dataset, dan penjelasan lokal yang menjelaskan prediksi model untuk satu sampel atau individu tertentu. Penelitian ini memanfaatkan keduanya melalui metode SHAP.

2.9 SHAP (SHapley Additive Explanations)

SHAP (SHapley Additive Explanations) adalah metode interpretasi model berbasis teori permainan (game theory) yang dikembangkan oleh Lundberg dan Lee [7]. SHAP mengukur kontribusi setiap fitur terhadap prediksi model dengan cara yang konsisten dan memiliki dasar matematis yang kuat, yaitu nilai Shapley dari teori permainan kooperatif. Nilai SHAP untuk fitur ke-i pada prediksi ke-j didefinisikan sebagai kontribusi marginal rata-rata fitur tersebut terhadap output model, yang dihitung berdasarkan seluruh kemungkinan kombinasi urutan fitur. Secara formal, nilai Shapley ϕ_i dihitung menggunakan Persamaan 2.1.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f(S \cup \{i\}) - f(S)] \tag{2.1}$$

di mana F merupakan himpunan seluruh fitur, S adalah subset fitur yang tidak mengandung fitur i, $f(S)$ merupakan prediksi model yang dihasilkan menggunakan fitur-fitur dalam subset S, dan $f(S \cup \{i\})$ merupakan prediksi model setelah fitur i ditambahkan ke dalam subset S.

2.9.1 Jenis Output SHAP

SHAP menghasilkan beberapa jenis visualisasi yang berguna untuk interpretasi model:

A. Summary Plot (Global): Menampilkan distribusi nilai SHAP seluruh fitur untuk semua sampel, memberikan gambaran fitur mana yang paling berpengaruh secara keseluruhan.

B. Bar Plot (Global): Menampilkan rata-rata nilai absolut SHAP setiap fitur, diurutkan dari yang paling berpengaruh.

C. Waterfall Plot (Lokal): Menampilkan kontribusi setiap fitur terhadap prediksi untuk satu sampel individu tertentu, sehingga dapat diketahui faktor apa yang mendorong model ke arah prediksi tertentu.

D. Dependence Plot: Menampilkan hubungan antara nilai suatu fitur dengan nilai SHAP-nya, sering disertai warna dari fitur lain untuk menangkap efek interaksi.

2.9.2 TreeSHAP untuk Model Berbasis Pohon

Untuk model berbasis pohon seperti Random Forest, SHAP menggunakan algoritma TreeSHAP yang dikembangkan khusus untuk mengeksploitasi struktur pohon keputusan [8]. TreeSHAP mampu menghitung nilai SHAP yang tepat (exact) dalam waktu polinomial, sehingga jauh lebih efisien dibandingkan pendekatan SHAP umum yang memerlukan waktu eksponensial. Hal ini menjadikan SHAP sangat praktis untuk diaplikasikan pada model Random Forest yang terdiri dari ratusan pohon.

2.10 Evaluasi Performa Model Klasifikasi

Evaluasi performa model klasifikasi bertujuan untuk mengukur seberapa baik model dalam melakukan prediksi pada data yang belum pernah dilihat sebelumnya. Beberapa metrik evaluasi yang digunakan dalam penelitian ini adalah:

A. Akurasi (Accuracy): perbandingan antara jumlah prediksi yang benar dengan total jumlah data. Metrik ini dapat menyesatkan pada dataset tidak seimbang.

B. Precision: kemampuan model dalam memprediksi kelas positif secara tepat, yaitu proporsi prediksi positif yang benar-benar positif.

C. Recall (Sensitivity): kemampuan model dalam mendeteksi seluruh data yang benar-benar positif. Metrik ini sangat penting dalam konteks medis.

D. F1-Score: rata-rata harmonis antara precision dan recall, berguna ketika terdapat ketidakseimbangan antar kelas.

E. AUC-ROC: Area Under the Receiver Operating Characteristic Curve, mengukur kemampuan model dalam membedakan antara kelas positif dan negatif pada berbagai threshold.