



BAB I

PENDAHULUAN

1.1 Latar Belakang

Penyakit kardiovaskular, terutama penyakit jantung koroner, merupakan penyebab kematian tertinggi di dunia. Data World Health Organization (WHO) menunjukkan bahwa sekitar 17,9 juta jiwa meninggal akibat penyakit kardiovaskular setiap tahunnya, yang mencakup lebih dari 30% dari total kematian global [1]. Di Indonesia, Riset Kesehatan Dasar (Riskesdas) 2018 mencatat prevalensi penyakit jantung mencapai 1,5% dengan kecenderungan peningkatan dari tahun ke tahun [2]. Tingginya angka mortalitas dan morbiditas ini tidak hanya berdampak pada kualitas hidup individu, tetapi juga membebani sistem kesehatan nasional dan ekonomi masyarakat [3]. Oleh karena itu, upaya deteksi dini dan prediksi risiko penyakit jantung menjadi sangat krusial untuk menurunkan angka kesakitan dan kematian.

Metode diagnosis konvensional yang umum digunakan saat ini bergantung pada pemeriksaan klinis oleh dokter, seperti elektrokardiogram (EKG), tes darah, serta prosedur pencitraan invasif seperti angiografi [4]. Pendekatan ini memerlukan biaya tinggi, waktu yang relatif lama, serta akses ke fasilitas kesehatan dan tenaga medis spesialis yang tidak merata, terutama di daerah terpencil. Akibatnya, banyak kasus penyakit jantung terdeteksi pada stadium lanjut ketika penanganan menjadi lebih sulit dan risiko komplikasi semakin tinggi [5]. Keterlambatan diagnosis ini berkontribusi signifikan terhadap tingginya angka kematian dan beban perawatan jangka panjang yang tidak dapat ditangani secara optimal.

Penderita penyakit jantung umumnya memiliki profil klinis yang khas. Berdasarkan studi epidemiologi dan dataset medis, faktor-faktor seperti usia lanjut (di atas 50 tahun), jenis kelamin laki-laki, tekanan darah istirahat yang tinggi (>140 mmHg), kadar kolesterol serum yang melebihi 200 mg/dL, serta kadar gula darah puasa >120 mg/dL sering ditemukan pada pasien yang terdiagnosis penyakit jantung. Selain itu, jenis nyeri dada yang dialami juga menjadi indikator penting; pasien dengan nyeri tumpul (non-anginal pain) atau nyeri tajam (asymptomatic) cenderung memiliki risiko lebih tinggi dibandingkan mereka yang tidak merasakan nyeri sama sekali [6]. Denyut jantung maksimal saat aktivitas fisik pada penderita penyakit jantung biasanya lebih rendah



(rata-rata 140–150 bpm) dibandingkan individu sehat (160–170 bpm) [6]. Dengan memahami karakteristik ini, deteksi dini dapat dilakukan lebih terarah, dan masyarakat awam pun dapat mengenali tanda-tanda awal tanpa harus langsung menjalani pemeriksaan medis invasif. Seiring perkembangan teknologi kecerdasan buatan (artificial intelligence), metode machine learning mulai banyak digunakan untuk membantu prediksi penyakit jantung secara non-invasif dan efisien [6]. Berbagai algoritma klasifikasi seperti Random Forest (RF), Extreme Gradient Boosting (XGBoost), dan Support Vector Machine (SVM) telah diimplementasikan pada dataset medis seperti UCI Heart Disease Dataset [7], [8]. Penelitian oleh Anderies dkk. (2022) melaporkan bahwa SVM mencapai akurasi 85% dalam memprediksi penyakit jantung [7]. Sementara itu, studi lain menunjukkan bahwa algoritma berbasis ensemble learning seperti RF dan XGBoost mampu mencapai akurasi antara 85% hingga 93%, dengan keunggulan masing-masing dalam stabilitas dan kecepatan komputasi [9], [10].

Namun, sebagian besar penelitian yang ada masih menggunakan model tunggal (single classifier) yang memiliki keterbatasan, seperti variance tinggi pada RF atau risiko overfitting pada XGBoost jika tidak dituning dengan baik [9]. Selain itu, penerapan validasi silang (cross-validation) sering diabaikan, sehingga estimasi performa model kurang akurat dan sulit digeneralisasi [11].

Penggunaan model hybrid Random Forest–XGBoost didasarkan pada keterbatasan masing-masing model tunggal. Random Forest yang berbasis bagging stabil dan tahan overfitting, tetapi kurang agresif dalam menurunkan bias. Sebaliknya, XGBoost yang berbasis boosting mampu mengurangi bias secara signifikan dengan akurasi tinggi, namun rentan overfitting. Dengan menggabungkan keduanya melalui soft voting berbobot, model hybrid memanfaatkan stabilitas Random Forest sekaligus kemampuan diskriminasi XGBoost. Hasil penelitian membuktikan hybrid mencapai akurasi yang cukup tinggi. Belum banyak studi yang secara sistematis menggabungkan kedua algoritma tersebut dalam suatu model hibrida, khususnya pada dataset yang besar dan beragam. Kelemahan lainnya adalah minimnya upaya untuk menyederhanakan fitur-fitur klinis sehingga masyarakat awam tidak dapat menggunakan model secara mandiri tanpa harus melakukan pemeriksaan medis yang rumit [6].

Berdasarkan celah tersebut, penelitian ini mengusulkan model hibrida yang menggabungkan Random Forest dan XGBoost melalui teknik soft voting dengan bobot optimal, serta membandingkannya dengan masing-masing model tunggal. Kontribusi penelitian ini meliputi: (a) metodologis – pengembangan model hibrida RF-XGBoost untuk prediksi penyakit jantung; (b) empiris – perbandingan performa kuantitatif (akurasi, presisi, recall, F1-score, AUC) yang menunjukkan peningkatan signifikan dibanding model tunggal; (c) eksperimental – penggunaan dataset gabungan 11.543 rekam medis dari lima sumber (Cleveland, Hungary, Switzerland, Long Beach VA, Statlog) serta skenario uji 5-fold cross-validation dan backward feature selection dari 11 ke 7 fitur yang mudah diakses [12]; (d) implementasi – pembuatan prototipe aplikasi web bernama CardioPredict menggunakan framework Streamlit, yang memungkinkan masyarakat memasukkan 7 parameter klinis (usia, jenis kelamin, nyeri dada, tekanan darah, kolesterol, gula darah puasa, denyut jantung maksimal) tanpa harus ke dokter terlebih dahulu; (e) dataset – penggabungan dan pembersihan data dari berbagai sumber sehingga menghasilkan dataset yang lebih representatif dibandingkan penelitian sebelumnya. Dengan demikian, penelitian ini diharapkan menghasilkan model prediksi yang lebih akurat, stabil, dan aplikatif untuk deteksi dini penyakit jantung..

Berdasarkan uraian di atas, penelitian ini bertujuan untuk mengembangkan model prediksi penyakit jantung yang berfungsi sebagai alat skrining awal (*early screening tool*), bukan sebagai alat diagnosis definitif. Perbedaan mendasar antara skrining dan diagnosis adalah bahwa skrining bertujuan untuk mengidentifikasi individu yang berisiko tinggi di antara populasi yang tampak sehat, sedangkan diagnosis merupakan proses penegakan penyakit yang dilakukan oleh tenaga medis profesional melalui serangkaian pemeriksaan klinis dan laboratorium yang komprehensif [14]. Model yang dibangun dalam penelitian ini dirancang untuk memberikan peringatan dini kepada masyarakat mengenai kemungkinan risiko penyakit jantung berdasarkan parameter klinis yang mudah diakses, sehingga mereka dapat segera melakukan konsultasi lebih lanjut dengan dokter untuk pemeriksaan diagnostik yang lebih akurat

1.2 Rumusan Masalah

1. Bagaimana perbedaan tingkat performa (akurasi, presisi, recall, F1-score, dan AUC) antara algoritma Random Forest, XGBoost, serta model hibrida Random Forest – XGBoost dalam mengklasifikasikan kemungkinan seseorang menderita penyakit jantung?
2. Model algoritma manakah yang menghasilkan hasil evaluasi paling tinggi di antara ketiga pendekatan tersebut untuk digunakan dalam prediksi penyakit jantung berdasarkan metrik evaluasi yang telah ditetapkan?
3. Bagaimana cara agar hasil pemodelan klasifikasi dapat diimplementasikan ke dalam sebuah aplikasi yang dapat digunakan secara mandiri oleh masyarakat tanpa harus terlebih dahulu melakukan pemeriksaan ke dokter?

1.3. Batasan Masalah

1. Penelitian ini hanya berfokus pada perbandingan performa antara klasifikasi tunggal menggunakan algoritma Random Forest (RF) dan XGBoost dengan klasifikasi hibrida Random Forest – XGBoost dalam memprediksi penyakit jantung.
2. Dataset yang digunakan merupakan dataset sekunder "Heart Disease Dataset" yang berasal dari UCI Machine Learning Repository, yang merupakan gabungan dari lima sumber data: Cleveland, Hungarian, Switzerland, Long Beach VA, dan Statlog. Dataset diakses pada tahun 2024/2025.
3. Penelitian ini hanya menggunakan 7 variable pada dataset. Setelah pembersihan data, diperoleh 11.543 baris data dengan 11 atribut awal. Melalui proses backward feature selection, dipilih 7 atribut final yang mudah diakses masyarakat, yaitu: age (usia), gender (jenis kelamin), chestpain (jenis nyeri dada), restingBP (tekanan darah istirahat), serumcholesterol (kolesterol serum), fastingbloodsugar (gula darah puasa), dan maxheartrate (denyut jantung maksimum). Atribut lain seperti restingelectro, exerciseangia, oldpeak, dan noofmajorvessels tidak digunakan dalam model final karena memerlukan pemeriksaan klinis.

1.4 Tujuan dan Manfaat Penelitian

1.4.1 Tujuan Penelitian

1. Menganalisis dan membandingkan performa (akurasi, presisi, recall, F1-score, AUC) dari Random Forest, XGBoost, dan hibrida RF XGBoost dalam prediksi penyakit jantung.
2. Menentukan model terbaik berdasarkan matrix evaluasi di antara ketiganya untuk sistem prediksi penyakit jantung.
3. Mengimplementasikan model terbaik ke dalam prototipe aplikasi berbasis web (Streamlit) yang dapat digunakan masyarakat secara mandiri.

1.4.2 Manfaat Penelitian

1. Memberikan pemahaman tentang perbandingan performa algoritma Random Forest, XGBoost, dan hybrid RF XGBoost dalam prediksi penyakit jantung.
2. Menjadi referensi dan bahan perbandingan bagi peneliti lain yang mengembangkan model klasifikasi hibrida untuk prediksi penyakit jantung.
3. Menyediakan prototipe aplikasi yang mudah digunakan masyarakat untuk deteksi dini risiko penyakit jantung, sehingga dapat meningkatkan kesadaran preventif tanpa perlu ke dokter terlebih dahulu.

1.5 Sistematika Penulisan

BAB 1 PENDAHULUAN

Dalam pendahuluan terdapat latar belakang, rumusan masalah, batasan masalah, tujuan penelitian, manfaat penelitian, serta sistematika penulisan yang dilakukan.

BAB II LANDASAN TEORI

Dalam landasan teori terdapat penelitian-penelitian terdahulu serta kajian teori yang relevan dan berkaitan dengan penelitian yang dilakukan.

BAB III METODOLOGI PENELITIAN

Dalam metodologi penelitian terdapat penjelasan mengenai model yang akan dibuat. Dalam bab ini akan dibahas mengenai sumber data, analisis pengolahan data, serta pembuatan model *machine learning* yang digunakan dalam penelitian.



BAB IV ANALISIS DAN HASIL PENELITIAN

Dalam analisis dan hasil penelitian terdapat pembahasan mengenai hasil pemodelan yang telah dibuat dalam penelitian beserta penjelasan analisisnya.

BAB V SIMPULAN DAN SARAN

Dalam simpulan dan saran terdapat pembahasan mengenai kesimpulan serta saran berdasarkan penelitian yang telah dilakukan.