BAB I
PENDAHULUAN

1.1. Latar Belakang

Penyakit jantung, terutama Penyakit Kardiovaskular (PKV), tetap menjadi penyebab utama kematian dan disabilitas di seluruh dunia, menimbulkan beban kesehatan dan ekonomi yang signifikan bagi individu, keluarga, dan sistem pelayanan kesehatan. Menurut laporan WHO, penyakit kardiovaskular adalah penyebab utama kematian dan disabilitas global, merenggut jutaan nyawa tiap tahunnya[1] . Oleh karena itu, deteksi dini dan identifikasi faktor risikonya sangat penting untuk strategi pencegahan dan penanganan yang efektif. Satu dekade terakhir, machine learning (ML) telah menghadirkan potensi besar dalam mengembangkan sistem prediksi akurat untuk risiko penyakit jantung, memungkinkan penanganan yang lebih proaktif dan personal. Dengan menganalisis banyak data medis, algoritma ML mampu menemukan pola dan hubungan kompleks yang meningkatkan ketepatan prediksi risiko kardiovaskular pada individu.

Dalam konteks pengembangan model prediktif ini, ketersediaan data medis yang relevan dan terstruktur menjadi fondasi utama. Dataset Heart Disease dari repositori UCI Machine Learning, khususnya subset yang berasal dari Cleveland Clinic, telah lama menjadi sumber daya yang sangat berharga dan sering dijadikan benchmark dalam riset ML untuk prediksi penyakit jantung. Dataset ini mencakup berbagai fitur klinis dan demografis penting seperti usia, jenis kelamin, tipe nyeri dada, tekanan darah istirahat, kadar kolesterol serum, hasil elektrokardiogram, detak jantung maksimum, dan informasi terkait latihan fisik serta kondisi pembuluh darah. Dataset klasik ini tetap digunakan dalam penelitian terbaru untuk menguji, memvalidasi, dan membandingkan kinerja algoritma ML serta metodologi baru dalam prediksi PKV [2]. Penggunaan dataset standar tidak hanya mempermudah pengembangan model yang efektif, tetapi juga memungkinkan peneliti membandingkan hasil dengan penelitian sebelumnya dan memahami bagaimana berbagai faktor berkontribusi terhadap risiko penyakit kardiovaskular.

Salah satu tantangan inheren dalam bekerja dengan data medis nyata adalah keberadaan nilai yang hilang (missing values). Nilai yang hilang dapat mengurangi kualitas dataset pelatihan dan berpotensi mendistorsi hubungan antar fitur, yang pada akhirnya dapat menurunkan kinerja dan reliabilitas model prediktif yang dihasilkan. Penanganan missing values secara tepat adalah langkah krusial dalam tahap preprocessing

data untuk memastikan integritas data pelatihan. K-Nearest Neighbors (KNN) Imputer adalah metode untuk mengisi nilai yang hilang dalam dataset. Metode ini bekerja dengan memperkirakan nilai yang hilang berdasarkan rata-rata atau modus dari k titik data terdekat dalam ruang fitur [3]. Metode ini memanfaatkan hubungan antar fitur dalam data, menghasilkan estimasi yang lebih akurat daripada metode imputasi sederhana. Dengan demikian, data yang dihasilkan menjadi lebih bersih dan siap untuk membangun model prediksi yang kuat.

Untuk menciptakan model prediksi penyakit jantung yang andal, diperlukan algoritma klasifikasi yang kuat dan fleksibel. XGBoost (Extreme Gradient Boosting) dikenal luas sebagai algoritma ensemble learning berbasis pohon yang unggul dan selalu memberikan performa terbaik dalam berbagai tugas klasifikasi data terstruktur [5]. XGBoost beroperasi dengan membangun serangkaian pohon keputusan secara sekuensial, di mana setiap pohon baru dilatih untuk mengoreksi kesalahan yang dibuat oleh kombinasi pohon-pohon sebelumnya. Di sisi lain, Random Forest (RF), sebagai algoritma ensemble learning berbasis pohon, telah terbukti sangat efektif dalam berbagai tugas klasifikasi, termasuk dalam domain medis. RF bekerja dengan membangun sejumlah besar pohon keputusan secara independen dan menggabungkan prediksi mereka melalui voting mayoritas, sehingga mampu mengurangi overfitting dan meningkatkan akurasi serta stabilitas model [4]. Algoritma ini menggabungkan optimalisasi sistem dan teknik regularisasi yang canggih untuk mengontrol kompleksitas model dan mencegah overfitting, memungkinkannya untuk melakukan generalisasi dengan baik pada data yang belum pernah dilihat sebelumnya. Mengingat kekuatan individu dari Random Forest dan XGBoost, penelitian terkini semakin mengeksplorasi kombinasi atau hibridisasi kedua algoritma ini. Pendekatan XGB-RF bertujuan untuk memanfaatkan kelebihan masing-masing algoritma seperti kemampuan XGBoost dalam mengoptimalkan bias dan meningkatkan akurasi melalui boosting sekuensial, serta kemampuan Random Forest dalam menangani overfitting dan variasi yang tinggi [6]. Kombinasi ini diharapkan dapat menghasilkan model prediksi yang lebih robust dan akurat, mampu menangkap pola data yang kompleks sambil tetap menjaga kemampuan generalisasi yang baik. Karena unggul dalam komputasi dan prediksi, kedua algoritma ini baik secara tunggal maupun digabungkan sangat cocok untuk tugas prediksi dalam penelitian ini.

Meskipun XGB-RF mampu mencapai akurasi prediksi yang tinggi, sifat kompleks dari model ensemble ini seringkali membuatnya tampak seperti "kotak hitam" (black-box), terutama dalam aplikasi kritis seperti diagnosis medis. Penting untuk

memahami alasan di balik prediksi model, bahkan mungkin lebih penting daripada prediksinya sendiri. Kebutuhan akan transparansi dan interpretasi model ini mendorong pengembangan bidang Explainable AI (XAI). SHapley Additive exPlanations (SHAP) adalah kerangka kerja interpretasi model mutakhir yang berakar pada teori permainan kooperatif Shapley values. SHAP mengukur kontribusi unik setiap fitur terhadap prediksi individu dengan membandingkannya dengan nilai dasar [7]. Dengan menerapkan SHAP pada model XGB-RF, kita bisa mengetahui fitur apa saja yang paling mempengaruhi tinggi rendahnya risiko penyakit jantung pada seseorang, serta bagaimana pengaruh fitur-fitur tersebut. Interpretasi berbasis SHAP ini tidak hanya meningkatkan kepercayaan pengguna terhadap hasil prediksi tetapi juga memberikan wawasan klinis yang berharga, mendukung pengambilan keputusan yang lebih terinformasi.

1.2. Rumusan Masalah

1.2.1 Bagaimana model stacking XGBoost - Random Forest (XGB-RF) dapat dibangun dan dioptimalkan untuk memprediksi risiko penyakit jantung menggunakan dataset Heart Disease UCI?

1.2.2 Seberapa efektif model stacking XGB-RF dalam mencapai performa prediksi yang unggul dibandingkan dengan model tunggal atau metode tradisional dalam konteks prediksi penyakit jantung?

1.2.3 Bagaimana penanganan nilai yang hilang dalam dataset Heart Disease UCI dapat dilakukan secara robust menggunakan KNN Imputer untuk memastikan kualitas data yang tinggi dan kinerja model yang optimal?

1.2.4 Bagaimana metode SHapley Additive exPlanations (SHAP) dapat diintegrasikan ke dalam model ensemble stacking yang kompleks untuk menginterpretasi hasil prediksi, secara spesifik mengidentifikasi kontribusi setiap fitur terhadap risiko penyakit jantung pada setiap individu?

1.3. Tujuan Penelitian

1.3.1. Membangun dan mengoptimalkan model stacking XGBoost - Random Forest (XGB-RF) untuk prediksi penyakit jantung menggunakan dataset Heart Disease UCI, setelah melakukan preprocessing data yang teliti, termasuk penanganan missing values menggunakan KNN Imputer.

1.3.2. Menganalisis dan mengevaluasi performa prediksi model stacking XGB-RF dalam mendeteksi risiko penyakit jantung, serta membandingkannya dengan model machine learning tunggalnya untuk menunjukkan keunggulannya.

1.3.3. Mengimplementasikan pipeline preprocessing data yang menyeluruh, termasuk penerapan KNN Imputer untuk penanganan nilai yang hilang secara efektif dalam dataset.

1.3.4. Mengintegrasikan metode SHapley Additive exPlanations (SHAP) ke dalam model ensemble stacking untuk menyediakan interpretasi yang jelas dan dapat ditindaklanjuti dari setiap prediksi individual.

1.4. Batasan dan Asumsi Penelitian.

1.4.1. Penelitian ini terbatas pada penggunaan dataset Heart Disease UCI yang tersedia di Kaggle. Tidak ada penambahan fitur eksternal atau rekayasa fitur yang kompleks di luar fitur yang ada.

1.4.2. Penelitian ini membatasi fokus pada kombinasi XGBoost dan Random Forest. Hal ini dilakukan untuk mendalami sinergi antara kedua algoritma. Model ensemble atau algoritma pembelajaran mesin lainnya tidak akan dieksplorasi secara mendalam sebagai bagian dari penelitian ini, meskipun ada banyak pilihan lain yang tersedia.

1.4.3. Perubahan tren atau pola penyakit jantung dari waktu ke waktu di luar cakupan dataset tidak dipertimbangkan.

1.5. Manfaat Penelitian

1.5.1. Hasil penelitian ini dapat memperkaya pemahaman mengenai efektivitas model hibrida Random Forest-XGBoost dalam aplikasi medis, khususnya prediksi penyakit jantung, yang menggabungkan keunggulan ensemble learning.

1.5.2. Menghasilkan prototipe sistem prediksi risiko penyakit jantung yang tidak hanya memberikan hasil prediksi (risiko) tetapi juga penjelasan yang dapat dipahami pengguna mengenai faktor yang mempengaruhi risiko tersebut.

1.6. Sistematika Penulisan

1.6.1. **Identifikasi Masalah dan Perumusan Tujuan**

Menentukan masalah yang ingin dipecahkan (prediksi penyakit jantung yang interpretable) serta merumuskan pertanyaan dan target spesifik penelitian.

1.6.2. **Studi Literatur dan Tinjauan Pustaka**

Mengkaji teori dan penelitian terdahulu terkait prediksi penyakit jantung,

metode ensemble stacking (XGB-RF), penanganan data (KNN Imputer), dan interpretasi model (SHAP).

1.6.3. **Metodologi Penelitian**

Menjelaskan secara rinci tahapan-tahapan yang akan dilakukan dalam penelitian, meliputi pengumpulan data, pra-pemrosesan data (termasuk penanganan nilai hilang dengan KNN Imputer), pengembangan model stacking XGBoost-Random Forest, evaluasi model, dan interpretasi model menggunakan SHAP.

1.6.4. **Analisis dan Pembahasan**

Menyajikan hasil analisis data, proses *tuning hyperparameter* untuk XGBoost dan Random Forest, serta interpretasi mendalam dari hasil prediksi model menggunakan SHAP, termasuk dampaknya terhadap pemahaman faktor risiko penyakit jantung.

1.6.5. **Perumusan Kesimpulan dan Saran**

Menyimpulkan seluruh temuan utama penelitian berdasarkan analisis hasil dan memberikan rekomendasi untuk pengembangan atau penelitian di masa mendatang.