BAB I
PENDAHULUAN

1.1 Latar Belakang

World Health Organization (WHO) tahun 2025 penyakit jantung menjadi penyebab utama kematian di seluruh dunia. Pernyataan ini diperkuat data dari *American Hearth Association* (AHA) yang menunjukan bahwa penyakit jantung masih menjadi penyebab kematian tertinggi, mencerminkan konsistensi tren global terkait tingginya beban penyakit jantung (AHA, 2025). Data Survei Kesehatan Indonesia (SKI) tahun 2024 mengenai jumlah pasien jantung berdasarkan kelompok usia menyebutkan bahwa kelompok usia 25-34 tahun mendominasi dengan jumlah 140.206 orang. Angka ini sedikit diatas kelompok usia 15-24 tahun yang mencapai 139.891 orang (Yashilva, 2024). Menurut *World Health Organization (WHO)* dari 17 juta kematian dini usia di bawah 70 tahun akibat penyakit tidak menular, sekitar 38% disebabkan oleh penyakit jantung. Selain itu, lebih dari 75% kematian akibat penyakit ini terjadi di negara-negara berkembang dengan pendapatan rendah dan menengah (Daniela Andrena, 2024). Berdasarkan data BPJS Kesehatan tahun 2021, biaya pengobatan penyakit jantung mencapai Rp. 7,7 triliun menjadikannya pengeluaran tertinggi dalam layanan kesehatan serta keterlambatan dalam diagnosis mendorong kebutuhan akan metode prediksi yang lebih akurat dan efisien (dr.Siti Nadia Tarmizi, 2022). Penderita penyakit jantung seringkali tidak menunjukan gejala pada tahap awal dan banyak diantaranya meninggal karena serangan jantung (Ahmadi et al., 2023). Beberapa pemeriksaan sangat penting untuk mendeteksi penyakit jantung (Adi & Wintarti, 2022).

Kemajuan teknologi, khususnya dalam bidang kecerdasan buatan (AI) memungkinkan pemanfaatan *machine learning* dalam analisis data medis (Ahmadi et al., 2023). Beberapa penelitian telah menggunakan model *Random Forest* dan *XGBoost* dalam prediksi penyakit jantung. (Ansyari et al., 2023) menunjukan bahwa penggunaan *Particle Swarm Optimization (PSO)* untuk seleksi fitur mampu meningkatkan performa *XGBoost* dan *Random Forest*. Akan tetapi, penelitian tersebut belum mengoptimalkan *hyperparameter* model secara langsung yang dapat mempengaruhi stabilitas dan akurasi prediksi. Beberapa penelitian lain juga telah

mengembangkan metode prediksi penyakit jantung dengan pendekatan yang berbeda. (Mohan et al., 2019) mengusulkan *Hybrid Random Forest with a Linear Model (HRFLM)* yang berhasil mencapai tingkat akurasi 88,7% lebih tinggi dibandingkan dengan model individu. (Saputra et al., 2022) membandingkan *Support Vector Machine (SVM)* dengan kombinasi kombinasi SVM-PSO, dimana model SVM-PSO mampu mencapai akurasi 84,81% dan nilai AUC - ROCsebesar 0,898. Penilitian lain oleh (Agus Dendi Rachmatsyah, 2024) menunjukan bahwa *Randomized search* dapat menghasilkan performa yang setara atau lebih baik dengan waktu eksekusi lebih singkat sehingga lebih efisien dibandingkan metode pencarian *hyperparameter* tradisional seperti *Grid Search*. Penelitian lain yang dilakukan oleh (Jain et al., 2022) menunjukan bahwa model evolusioner seperti *Particle Swarm Optimization (PSO)* dapat secara efektif menemukan kombinasi *hyperparameter* terbaik dalam berbagai model pembelajaran mesin. PSO terbukti memiliki tingkat konvergensi yang lebih cepat serta memiliki flekbilitas tinggi dalam menangani optimasi parameter kompleks sehingga ideal untuk tuning model berbasis *ensemble*.

Model *Random Forest* merupakan kumpulan *decision tree* yang dibangun dari sampel acak, bekerja dengan subset fitur untuk menemukan nilai ambang optimal dalam pemisahan data. Kelebihannya adalah ketahanan terhadap overfitting, namun cenderung kurang baik dalam menangani data high-dimensional dan sulit dioptimalkan secara manual (Nugroho & Harini, 2024). Sementara itu, *XGBoost* merupakan model berbasis *gradient boosting* yang memiliki performa tinggi dan dapat menangani *missing values* dengan baik, tetapi *sensitive* terhadap pemilihan *hyperparameter* yang jika tidak dioptimalkan dapat menyebabkan overfitting (Andriansyah & Eka Wulansari Fridayanthie, 2023).

Penelitian ini mengusulkan *Randomized search* untuk optimasi *hyperparameter* kedua model. *Randomized Seach* memilih kombinasi parameter acak dalam ruang pencarian untuk mendapatkan parameter terbaik dengan efisensi lebih tinggi (Agus Dendi Rachmatsyah, 2024). Selain optimasi *hyperparameter*, penelitian ini juga menerapkan metode *stacking ensemble* untuk menggabungkan hasil prediksi *XGBoost* dan *Random Forest* guna meningkatkan akurasi dan

stabilitas model. Metode ini menggunakan meta-model *Logistic Regression* yang berfungsi sebagai pengambil keputusan akhir berdasarkan output dari model dasar. *Logistic Regression* dipilih karena kesederhanaannya dan kemampuannya dalam menggabungkan prediksi dari berbagai model dengan efektif, sehingga meningkatkan generalisasi dan mengurangi kemungkinan kesalahan prediksi (Zian et al., 2021).

Selain metode *stacking*, terdapat pendekatan *ensemble* lain seperti *Voting Ensemble* yang juga sering digunakan untuk meningkatkan akurasi prediksi. *Voting* bekerja dengan cara menggabungkan hasil prediksi dari beberapa model dan memilih prediksi yang paling banyak dipilih *(majority voting)* atau rata-rata skor probabilitas *(soft voting)*. Metode ini sederhana dan cukup efektif, *Voting Ensemble* memiliki keterbatasan dalam mempertimbangkan hubungan antar model, karena semua model diberi bobot yang sama tanpa memperhatikan kinerja masing-masing (Almasri et al., 2025). Metode *stacking ensemble* menawarkan pendekatan yang lebih fleksibel dan kompleks dengan membangun model meta (meta-learner) yang belajar dari prediksi model-model dasar. Hal ini memungkinkan *stacking* untuk menangkap pola kesalahan antar model dasar dan menghasilkan prediksi akhir yang lebih adaptif. Beberapa penelitian terbaru juga mendukung keunggulan *stacking* dalam meningkatkan performa model. Pada penelitian oleh (Majid et al., 2025) menunjukkan bahwa kombinasi *stacking* dengan meta-model *Logistic Regression* mampu mencapai akurasi lebih tinggi dibandingkan *voting* dalam klasifikasi penyakit diabetes. Penelitian (Fonda et al., 2024) membuktikan bahwa *stacking ensemble* memberikan hasil lebih stabil dan akurat pada data fisiologis mahasiswa dibandingkan metode *voting*. Pada penelitian ini, *stacking* dipilih karena kemampuannya dalam menggabungkan keunggulan berbagai model dasar secara lebih optimal dan dinamis dibanding *voting ensemble*.

Masalah utama dalam penelitian ini adalah bagaimana membangun metode prediksi penyakit jantung yang mampu memberikan performa lebih optimal melalui kombinasi dua model dan optimasi *hyperparameter* untuk meningkatkan akurasi. Penelitian ini menggunakan model *XGBoost* karena kemampuannya dalam membentuk model secara bertahap (iteratif) dan mampu menangani data kompleks

dengan akurasi tinggi. *Random Forest* dipilih karena kestabilannya dan kemampuannya menghindari *overfitting* dengan membangun banyak pohon secara acak dan paralel. Kedua model ini kemudian digabungkan menggunakan pendekatan metode *stacking ensemble* agar saling melengkapi kelemahan masing-masing dan menghasilkan prediksi yang lebih kuat. Proses penggabungan prediksi dari kedua model dasar dilakukan melalui meta-model *Logistic Regression*. Pemilihan *Logistic Regression* didasarkan pada kesederhanaannya, waktu pelatihan yang singkat, dinilai lebih stabil tidak mudah *overvitting* pada dataset berukuran kecil. Meningkatkan performa model dilakukan optimasi *hyperparameter* menggunakan *Randomized search* yaitu metode pencarian kombinasi parameter terbaik secara acak.

Penelitian ini merupakan pengembangan dari penelitian (Ansyari et al., 2023) yang mengimplementasikan *XGBoost* dan *Random Forest* dengan seleksi fitur PSO, namun belum mengoptimalkan *hyperparameter* dan belum menerapkan metode *ensemble learning*. Gap penelitian ini yaitu dengan mengimplementasikan *Randomized search* dan metode *stacking ensemble*, diharapkan mampu meningkatkan akurasi prediksi penyakit jantung serta mengurangi risiko *overfitting*. Pendekatan ini memungkinkan pemanfaatan keunggulan masing-masing model dasar dalam menangani kompleksitas data medis, sehingga menghasilkan prediksi yang lebih akurat dan stabil. Selain meningkatkan akurasi, interpretasi terhadap hasil prediksi juga menjadi perhatian penting dalam bidang medis. Salah satu pendekatan yang digunakan adalah analisis *feature importance*, yaitu pengukuran kontribusi setiap fitur terhadap prediksi yang dihasilkan model. Informasi ini sangat bermanfaat dalam konteks klinis, karena dapat membantu dokter memahami faktor risiko utama yang berperan dalam klasifikasi penyakit jantung. Oleh karena itu, penelitian ini berjudul **"Pemodelan Prediksi Penyakit Jantung Menggunakan *Stacking Ensemble*".**

1.2 Rumusan Masalah

Berdasarkan uraian pada latar belakang, dapat dirumuskan masalah pada penelitian ini, yaitu:

1. Bagaimana metode *stacking ensemble* memprediksi penyakit jantung serta fitur-fitur yang mempengaruhi prediksi?
2. Bagaimana performa *Randomized search* dalam mengoptimalkan *hyperparameter* pada model *XGBoost* dan *Random Forest*?

1.3 Tujuan Penelitian

Dari uraian masalah yang sudah dirumuskan, didapat tujuan penelitian sebagai berikut:

1. Menerapkan metode *stacking ensemble* untuk memprediksi penyakit jantung serta mengidentifikasi fitur-fitur yang paling berpengaruh terhadap hasil prediksi.
2. Mengevaluasi performa *Randomized Search* dalam proses optimasi *hyperparameter* pada model *XGBoost* dan *Random Forest*.

1.4 Manfaat Penelitian

Adapun beberapa manfaat yang didapatkan dari penelitian ini, sebagai berikut:

1. Menyediakan pendekatan yang lebih optimal dalam prediksi penyakit jantung.
2. Meningkatkan efisiensi dalam optimasi *hyperparameter*.
3. Menjadi referensi bagi penelitian selanjutnya.

1.5 Batasan Masalah

Berikut merupakan batasan masalah yang terdapat pada penelitian ini:

1. Dataset yang digunakan berasal dari Kaggle Heart Disease Dataset UCI Repository.
2. Model yang digunakan *XGBoost* dan *Random Forest* sebagai base model dengan *Logistic Regression* sebagai meta-model.
3. Evaluasi performa model dilakukan menggunakan AUC-ROC.