

BAB I PENDAHULUAN

1.1 Latar Belakang

Penyakit jantung merupakan salah satu penyebab kematian tertinggi di dunia, termasuk di Indonesia. Menurut data dari *World Health Organization*, penyakit dengan nama lain kardiovaskular menyebabkan lebih dari 17,9 juta kematian setiap tahunnya, dan sekitar 85% di antaranya disebabkan oleh serangan jantung dan stroke [1]. Peningkatan jumlah penderita penyakit jantung ini sering kali disebabkan oleh berbagai faktor seperti tekanan darah tinggi, kolesterol, obesitas, gaya hidup tidak sehat, serta kurangnya deteksi dini terhadap kondisi kardiovaskular. Oleh karena itu, diperlukan pendekatan berbasis teknologi agar membantu mendeteksi risiko penyakit jantung secara akurat dan efisien.

Perkembangan teknologi *Machine Learning* telah membuka peluang baru dalam dunia medis khususnya dalam memprediksi penyakit jantung [2][3]. Beberapa algoritma *Machine Learning* seperti *Logistic Regression*, *Random Forest*, *Decision Tree*, dan *XGBoost* telah digunakan untuk melakukan klasifikasi terhadap pasien berdasarkan risiko penyakit jantung [4][5][6]. Namun, model tunggal sering kali memiliki keterbatasan, seperti *overfitting* pada jenis data tertentu atau kurang optimal dalam menangani kompleksitas fitur medis yang saling berkorelasi [7].

Selain itu, model tunggal pada umumnya tidak mampu menangkap hubungan non-linear dan pola interaksi yang rumit antar variabel klinis [8]. Misalnya, *Logistic Regression* hanya mampu mempelajari hubungan linear sehingga kurang efektif untuk pola data medis yang lebih kompleks [9]. Sementara itu, model seperti *Random Forest* dan *XGBoost* memang lebih kuat, tetapi tetap memiliki keterbatasan ketika harus berhadapan dengan *noise*, *multikolinearitas*, serta variabilitas data biologis yang tinggi [6]. Karakteristik khas pada data medis semakin memperjelas keterbatasan pendekatan model Tunggal [10].

Data klinis biasanya memiliki fitur yang saling berkaitan, distribusi kelas yang tidak seimbang antara pasien sehat dan pasien dengan penyakit jantung, serta

adanya variabilitas fisiologis antarindividu [11][12]. Selain itu, tidak jarang dataset medis mengandung *noise* atau ketidakkonsistenan yang dapat mengganggu proses pembelajaran model [10]. Kondisi-kondisi ini membuat model tunggal sulit mencapai performa maksimal [13]. Pendekatan *ensemble learning*, khususnya kombinasi beberapa model *Gradient Boosting*, muncul sebagai solusi yang lebih unggul [14]. *Ensemble learning* memungkinkan penggabungan kekuatan dari beberapa algoritma sekaligus sehingga dapat mengurangi bias dan varians yang terdapat pada model Tunggal [15].

Penelitian Menggunakan *Cardiovascular Disease Dataset* dari kaggle dengan total 70,000 data dan 11 fitur. Model yang digunakan meliputi *Random Forest*, *Decision Tree*, *Multilayer Perceptron*, dan *XGBoost* digunakan untuk memprediksi penyakit jantung. Hasil penelitian tersebut menunjukkan bahwa model terbaik yaitu *Multilayer Perceptron*, hanya mencapai akurasi sebesar 87,28% dengan *AUC* 0.95. Berdasarkan hasil tersebut, penelitian ini mengembangkan pendekatan baru dengan menggabungkan dua model *Gradient Boosting*, yaitu *XGBoost* dan *LightGBM* menjadi satu model *Hybrid Gradient Boosting Ensemble*, untuk meningkatkan performa klasifikasi penyakit jantung. dataset yang digunakan memiliki variable utama seperti *age*, *gender*, *height*, *weight*, *ap_hi*, *ap_lo*, *cholesterol*, *gluc*, *smoke*, *alco*, *active* dan *cardio* sebagai label target [16].

Selain itu, berbagai penelitian lain turut menguatkan pentingnya pengembangan model *ensemble* yang lebih akurat, interpretatif, dan stabil. Penerapan *Explainable Artificial Intelligence* (XAI) untuk meningkatkan transparansi model prediksi penyakit jantung. Pendekatan ini menggaabungkan Teknik interpretabilitas seperti *SHAP* dan *LIME*, sehingga memberikan pemahaman yang lebih mendalam mengenai kontribusi setiap fitur terhadap hasil prediksi [17]. Pengembangan *hybrid ensemble*, dengan menggabungkan alogaritma seperti *Random Forest*, *Gradeint Boosting*, dan *XGBoost*. Hasil penelitian tersebut menunjukkan peningkatan akurasi dan *F1-Score* yang signifikan dibandingkan model tunggal, sehingga membuktikan bahwa *hybrid ensemble* lebih efektif dalam menangkap pola kompleks pada data medis [18].

Pendekatan *Stacking Ensemble*, yang menggabungkan beberapa model dasar secara berlapis, dengan meta-model *XGBoost* sebagai Keputusan akhir. Hasil penelitian membuktikan bahwa *stacking ensemble* mampu menghasilkan kaurasi yang lebih tinggi dibandingkan metode *ensemble* lain seperti *voting* atau *bagging*[19]. Kemudian optimasi *hyperparameter* melalui *Grid Search* dan *Random Search* sangat berpengaruh terhadap peningkatan kinerja model *ensemble* seperti *AdaBoost*, *Random Forest*, dan *XGBoost*. Parameter seperti jumlah *estimator* dan *learning rate* terbukti memiliki dampak besar terhadap performa dan kemampuan generalisasi model [20].

Framework Ensemble fleksibel untuk memprediksi penyakit jantung menggunakan alogaritma dan metrik evaluasi, serta menegaskan pentingnya penggunaan validasi silang (*cross-validation*) dalam membandingkan performa model [21]. Permasalahan *Imbalanced Dataset*, yaitu ketimpangan jumlah pasien sehat dan pasien dengan penyakit jantung. Melalui penerapan Teknik penyeimbangan data seperti *SMOTE* dalam model *stacked ensemble*, penelitian ini berhasil meningkatkan *recall* dan *F1-Score* pada kelas minoritas, sehingga memperkuat pentingnya data *balancing* dalam memprediksi medis [22].

Berdasarkan berbagai temua pada penelitian sebelumnya, penelitian ini mengembangkan model Hybrid Gradient Boosting berbasis kombinasi alogaritma *XGBoost* dan *LightGBM*. Melalui proses *preprocessing*, *normalisasi*, serta optimisasi *hyperparameter*, dan evaluasi model *hybrid* diharapkan model *Hybrid Gradient Boosting* yang dikembangkan diharapkan mampu menghasilkan performa klasifikasi yang lebih akurat dan stabil dalam membedakan pasien beresiko dan tidak beresiko mengalami penyakit jantung. Dengan demikian, penelitian ini berupaya memberikan kontribusi signifikan dalam pengembangan model prediksi penyakit jantung.

1.2 Rumusan Masalah

1. Bagaimana proses kerja model *Hybrid Gradient Boosting* yang menggabungkan *XGBoost* dan *LightGBM* dalam membangun system prediksi penyakit jantung, yang ditinjau dari metrik evaluasi seperti akurasi, *AUC*, dan *F1-Score* ?
2. Bagaimana penerapan tahapan data *preprocessing*, *model hybrid* dan *hyperparameter optimization* dalam meningkatkan kinerja model *Gradient Boosting* tersebut ?

1.3 Batasan Masalah

1. Fokus penelitian ini pada prediksi risiko penyakit jantung menggunakan pendekatan *Machine Learning*, dengan metode *Hybrid Gradient Boosting* yang menggabungkan algoritma *XGBoost* dan *LightGBM*.
2. dataset berasal dari platform Kaggle dengan nama *Cardiovascular Disease Dataset* dengan total 70,000 data dan 11 fitur utama yaitu *age*, *gender*, *height*, *weight*, *ap_hi*, *ap_lo*, *cholesterol*, *gluc*, *smoke*, *alco*, *active* serta *cardio* sebagai label target.
3. Penelitian ini hanya difokuskan pada tahap klasifikasi untuk membedakan antara pasien berisiko dan tidak berisiko terhadap penyakit jantung.
4. Tahapan penelitian meliputi identifikasi masalah, studi literatur pengumpulan data, *EDA*, *preprocessing*, optimisasi *hyperparameter* dan evaluasi pada model *Hybrid Gradient Boosting*.
5. Evaluasi performa model dibatasi pada beberapa metrik utama evaluasi model utama yaitu *Accuracy*, *AUC Score*, dan *F1-Score*.
6. Implementasi model dilakukan menggunakan Bahasa pemrograman Python di lingkungan Google Colab.

1.4 Tujuan Penelitian

Penelitian ini bertujuan untuk merancang dan mengimplementasikan model *Hybrid Gradient Boosting* berbasis kombinasi algoritma *XGBoost* dan *LightGBM* guna mengoptimalkan performa prediksi risiko penyakit jantung.

1.5 Manfaat Penelitian

1. Secara akademis, penelitian ini berkontribusi dalam pengembangan ilmu di bidang Machine Learning untuk para peneliti, khususnya pada penerapan model *Hybrid Gradient Boosting* untuk klasifikasi penyakit jantung.
2. Secara teknologis, penelitian ini dapat menjadi dasar untuk peneliti dalam pengembangan system prediksi berbasis *Machine Learning* yang dapat diintegrasikan ke dalam aplikasi Kesehatan untuk memantau risiko terkena penyakit jantung.

1.6 Sistematika Penulisan

BAB I PENDAHULUAN

Bab ini menjelaskan mengenai latar belakang, rumusan masalah, Batasan masalah, tujuan penelitian, serta manfaat penelitian. Bagian ini memberikan Gambaran umum tentang alasan dalam melakukan penelitian.

BAB II TINJAUAN PUSTAKA

Bab ini menyajikan studi literatur mencakup hasil hasil penelitian terdahulu yang relevan, sedangkan dasar teori berisi konsep, model, dan teori yang mendukung serta menjadi acuan dalam pelaksanaan penelitian.

BAB III METODE PENELITIAN

Bab ini menjelaskan mengenai objek penelitian, alur atau tahapan penelitian, serta alat dan bahan yang digunakan, serta menguraikan metode pengumpulan data, Teknik analisis data, dan prosedur eksperimen yang dilakukan mencapai tujuan penelitian

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil penelitian yang diperoleh dari proses pengolahan data dan eksperimen, serta pembahasan terhadap hasil tersebut, analisis dilakukan untuk membandingkan hasil penelitian dengan teori atau penelitian sebelumnya.

BAB V PENUTUP

Bab ini berisi Kesimpulan yang diperoleh dari hasil penelitian serta saran yang dapat diberikan untuk pengembangan penelitian selajutnya.

