

ABSTRAK

Penelitian ini bertujuan untuk menganalisis dan memprediksi risiko penyakit jantung menggunakan teknik eksplorasi data (EDA) dan pemodelan *machine learning*, khususnya *Random Forest* dan *Gradient Boosting Machine* (GBM). Data yang digunakan berasal dari dataset Kaggle yang berisi atribut medis seperti usia, jenis kelamin, tekanan darah, kolesterol, dan hasil diagnosis. Melalui analisis statistik deskriptif, dilakukan identifikasi outlier dan duplikat untuk memastikan kualitas data. Model *machine learning* menunjukkan performa tinggi dengan akurasi mencapai 98%, di mana GBM menunjukkan keunggulan dibandingkan *Random Forest*. Fitur penting yang berkontribusi besar meliputi q_25th, ST slope, dan mode, yang membantu dalam interpretasi faktor risiko. Hasil penelitian ini mendukung penggunaan algoritma *machine learning* dalam prediksi risiko penyakit jantung dan pengambilan keputusan klinis yang lebih cepat dan akurat. Untuk meningkatkan akurasi dan generalisasi model, penelitian lebih lanjut menggunakan dataset yang lebih besar dan algoritma eksplorasi alternatif direkomendasikan.

Kata kunci : *Random Forest, Gradient Boosting Machine, Prediksi Risiko Penyakit Jantung, Analisis Perbandingan, Akurasi Model*

ABSTRACT

This study aims to analyze and predict the risk of heart disease using data exploration techniques (EDA) and machine learning modeling, specifically Random Forest and Gradient Boosting Machine (GBM). The data used comes from the Kaggle dataset containing medical attributes such as age, gender, blood pressure, cholesterol, and diagnosis results. Through descriptive statistical analysis, outliers and duplicates were identified to ensure data quality. The machine learning model showed high performance with an accuracy of 98%, whereas GBM showed superiority over Random Forest. Important features that contributed greatly included q_25th, ST slope, and mode, which helped interpret risk factors. The results of this study support the use of machine learning algorithms in predicting the risk of heart disease and making clinical decisions faster and more accurately. To improve the accuracy and generalization of the model, further research using larger datasets and alternative exploration algorithms is recommended.

Keywords: *Random Forest, Gradient Boosting Machine, Heart Disease Risk Prediction, Comparative Analysis, Model Accuracy*