

## ABSTRACT

Cardiovascular disease will continue to be one of the most common causes of death globally hence there is an associated need with risk prediction tools to be accurate and interpretable. Here, we build an explainable machine learning model (using Extreme Gradient Boosting: XGBoost) to prospectively estimate the probability of heart attack based on clinical measurements, lifestyle behaviors, and socioeconomic factors. We performed very careful data preprocessing and class-balancing techniques to mitigate class imbalance in medical datasets. The model was fitted using early stopping and assessed with different evaluation metrics. It had ROC-AUC=0.73 and PR-AUC=0.80 on validation set (indicating moderate discriminative ability with strong sensitivity to minority cases). For the separate test set (downstream), we observe better generalization performance with ROC-AUC = 0.79, recall = 0.84, and precision = 0.63 for the model. Such results show that the model can accurately identify high-risk individuals with an acceptable false positive rate, making it feasible to use as a screening tool in healthcare practice. To increase interpretability, quantifying feature contributions were performed using SHapley Additive exPlanations (SHAP). These include factors such as blood pressure, smoking habits, physical activity, sleep patterns, alcohol consumption and other modifiable lifestyles that contribute to the risk prediction analysis. Moreover, it addresses the dependence of cardiovascular risk on a broader socioeconomic context. This study provides a transparent and interpretable predictive model that achieves an acceptable balance between overall performance and clinical relevance, aiding in early identification of patients at high risk for cardiovascular disease enabling targeted efforts to prevent the progression of disease through medical decision-making.

## ABSTRAK

Penyakit kardiovaskular akan terus menjadi salah satu penyebab kematian paling umum di dunia, oleh karena itu diperlukan alat prediksi risiko yang akurat dan mudah diinterpretasikan. Di sini, kami membangun model pembelajaran mesin yang dapat dijelaskan (menggunakan Extreme Gradient Boosting: XGBoost) untuk memperkirakan probabilitas serangan jantung secara prospektif berdasarkan pengukuran klinis, perilaku gaya hidup, dan faktor sosioekonomi. Kami melakukan pra-pemrosesan data dan teknik penyeimbangan kelas yang sangat hati-hati untuk mengurangi ketidakseimbangan kelas dalam dataset medis. Model tersebut diuji menggunakan penghentian dini dan dinilai dengan berbagai metrik evaluasi. Model tersebut memiliki ROC-AUC=0,73 dan PR-AUC=0,80 pada set validasi (menunjukkan kemampuan diskriminatif moderat dengan sensitivitas yang kuat terhadap kasus minoritas). Untuk set pengujian terpisah (downstream), kami mengamati kinerja generalisasi yang lebih baik dengan ROC-AUC = 0,79, recall = 0,84, dan presisi = 0,63 untuk model tersebut. Hasil tersebut menunjukkan bahwa model dapat secara akurat mengidentifikasi individu berisiko tinggi dengan tingkat false positive yang dapat diterima, sehingga layak digunakan sebagai alat skrining dalam praktik perawatan kesehatan. Untuk meningkatkan interpretasi, kuantifikasi kontribusi fitur dilakukan menggunakan SHapley Additive exPlanations (SHAP). Ini termasuk faktor-faktor seperti tekanan darah, kebiasaan merokok, aktivitas fisik, pola tidur, konsumsi alkohol, dan gaya hidup lain yang dapat dimodifikasi yang berkontribusi pada analisis prediksi risiko. Selain itu, hal ini membahas ketergantungan risiko kardiovaskular pada konteks sosioekonomi yang lebih luas. Studi ini menyediakan model prediktif yang transparan dan dapat diinterpretasikan yang mencapai keseimbangan yang dapat diterima antara kinerja keseluruhan dan relevansi klinis, membantu identifikasi dini pasien berisiko tinggi terhadap penyakit kardiovaskular sehingga memungkinkan upaya yang ditargetkan untuk mencegah perkembangan penyakit melalui pengambilan keputusan medis.