Interpretable Machine Learning for Cardiovascular Disease Risk Prediction

Byte to Beat Hackathon

B.Durga Devihttps://github.com/durgadeviboomi03-a11y/Heart-disease-risk-prediction/upload/main

Introduction

Heart disease kills more people than anything else around the world, and it often shows no signs until it's bad. Noticing it early is critically important. Hospitals already collect clinical measures like blood pressure, cholesterol, and daily habits, so I figured machine learning could help spot risks using that.

ML works great for predictions in medicine, but doctors need systems they can understand and rely on, not just ones that perform well on paper. In this project, I used easy-to-explain models to predict heart disease from anonymized patient data, testing them like they'd be used for real. The objective of this study is to evaluate whether interpretable machine learning models using routine clinical measurements can reliably predict cardiovascular disease risk while maintaining clinical interpretability

Dataset and Preprocessing

Multiple cardiovascular datasets provided through the Hackathon platform were used. Primary analysis focused on cardio_base.csv (anonymized patient health metrics). Integrated supplementary datasets cardiac_failure_processed.csv, heart_processed.csv, and eg_timeseries.csv. The time-series dataset was explored but

excluded from final modeling due to scope constraints. All data were fully anonymized prior to analysis to maintain strict ethical compliance

Methodology

Models and Evaluation

Two models were evaluated: Logistic Regression and Random Forest. Data was split into 80% training and 20% testing sets. Performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC. For Logistic Regression, various judgements in thresholds were tested to analyze replacements between neglected cases and false positives, which is critical in diagnostic evaluation.

Interpretability Analysis

Logistic regression analysis identified age, systolic blood pressure, and cholesterol as the key factors of risk, co-ordinating with established fundamental research.

Reproducibility

To ensure consistency all preprocessing, training, and validation code along with environment set-up files are filed in the project's GitHub repository.

Results

Both models showed similar performance. Logistic Regression achieved about 71% accuracy with a ROC-AUC of 0.78, while Random Forest performed in the same way but had a marginally lower AUC. Logistic Regression was chosen because its results are easier to interpret.

Adjusting the threshold revealed important trade-offs. Using the default value of 0.5 missed many individuals who should have received a warning. Reducing the threshold to 0.4 improved detection while keeping false positives at a reasonable level.

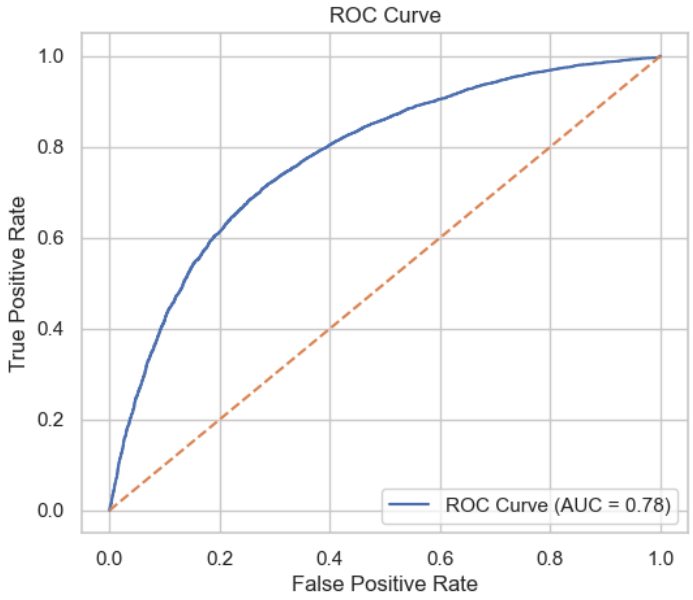

Figure 1: ROC Curve for Logistic Regression model

Limitations

This study used past, anonymized data. Time-series data was not included because of scope limits, and these results may not apply to all populations due to possible sampling or measurement limitations.

Conclusion and Future Work

This project shows that interpretable machine learning models can effectively predict cardiovascular disease risk using routine clinical and lifestyle data. Sorting case tracing through threshold modification better serves public health by reducing missed diagnoses in elevated-risk in populations. The predictive features follow known factors, helping clinicians better understand and trust the result. Future work includes integrating time-series signals, validating on external datasets, and assessing fairness before clinical use.

AI Usage Disclosure

Generative AI tools were used only for coding assistance and debugging. All research ideas, hypotheses, analysis, interpretation of results, and written content were created solely by myself.