Heart Risk Predictor: Integrating Machine Learning for Cardiovascular Risk Assessment

By Vedh Bagare

Abstract:
Cardiovascular disease is the leading cause of death globally. By identifying high-risk patients early, preventative interventions can improve outcomes and reduce healthcare costs. This project uses the UCI Heart Disease dataset to develop a machine learning model that predicts heart disease risk based on clinical biomarkers and patient demographics.

Introduction:
Heart disease risk is influenced by several factors, including age, sex, cholesterol, and blood pressure. Traditional risk calculators often rely on limited variables or linear assumptions. This project applies both logistic regression and random forest classifiers to model relationships between variables, providing a data-driven approach to predicting heart disease risk.

Data and Methodology:
The UCI Heart Disease dataset contains 1,026 patient records with 14 features. The primary outcome variable, target, indicates the presence of heart disease. Input features include age, sex, cp (chest pain type), trestbps (resting blood pressure), chol (cholesterol), fbs (fasting blood sugar), restecg (resting ECG), thalach (maximum heart rate), exang (exercise-induced angina), oldpeak (ST depression), slope (ST slope), ca (number of major vessels), and thal (thalassemia test result). Categorical variables like chest pain type and thalassemia were encoded numerically, while continuous variables were standardized to improve model performance.

Data preprocessing included handling missing values, scaling numeric features, and splitting the dataset into training (80%) and testing (20%) sets. Two machine learning models were trained: a Logistic Regression model to establish baseline performance and a Random Forest Classifier to identify non-linear interactions between features. Model evaluation metrics included accuracy, precision, recall, F1-score, and feature importance, allowing the identification of which clinical indicators most strongly predict heart disease. This framework enabled both predictive performance and interpretability of key risk factors.

Limitations:
The limited dataset (1,026 samples) may restrict generalizability. Model predictions are probabilistic and should not be interpreted as clinical diagnoses. The dataset may contain bias (sex and age distribution skew).

Models and Performance:
Two models were trained: logistic regression and random forest. The logistic regression was used as a baseline, while the random forest was selected in the end, because it could capture non-linear interactions and was better with outliers. The models were evaluated using accuracy, precision, recall, F1-score, and confusion matrices. A simple baseline heuristic was used as well; if the patient was over 55 years old or had a cholesterol level above 240 mg/dL, they were classified as high risk.

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	0.815	0.764	0.924	0.836
Random Forest	0.900	0.880	0.850	0.865
Baseline Heuristic	0.720	0.700	0.600	0.647

Feature Importance and Interpretation:

Feature	**Importance**
cp	0.21
thalach	0.18
exang	0.15
oldpeak	0.12
ca	0.10
thal	0.08
slope	0.05
restecg	0.04
age	0.03
trestbps	0.02

chol	0.01
fbs	0.005
sex	0.005

1. Chest Pain (cp) - Most predictive feature: abnormal chest pain patterns often reflect reduced blood flow (ischemia) to the heart muscle, caused by partial blockage of the coronary arteries. When left untreated, this can progress into a heart attack, which is why there is a strong positive correlation between certain chest pain types and heart disease.
2. Maximum Heart Rate (thalach) - Second most predictive feature: Lower maximum heart rate signals reduced cardiovascular function, which is a precursor to heart disease.
3. Exercise-Induced Angina (exang) - Third most predictive feature: Presence of angina during physical exertion occurs when the heart's oxygen demand exceeds supply due to narrowed arteries.
4. ST Depression (oldpeak) and ST Slope (slope) - Fourth most predictive feature: Reflects ischemic changes (tissue damage from reduced blood flow) in ECG, relevant for cardiac events.
5. Number of Major Vessels (ca) - Fifth most predictive feature: Each blocked vessel reduces oxygen delivery to the heart and increases the likelihood of cardiac failure.

Business Implications and Recommendations:
The Heart Disease Predictor model holds significant value as it provides a quantitative foundation for cost optimization and patient prioritization. Cardiovascular diseases account for nearly $239 billion annually in U.S. healthcare costs, with hospitalizations making up nearly 60% of that figure (CDC 2024). Insurers can use this model to identify high-risk individuals (abnormal cp, thalach, exang, slope, ca, etc) early on and offer interventions that reduce hospitalization risk by about 20% and $2,400-$3,000 saved per patient per year. Across a sample size of 10,000 insured adults, this approach can yield $24-30 million in annual savings while improving health outcomes.

Hospitals spend heavily on redundant or unnecessary cardiac tests. According to the American Heart Association, nearly 25% of stress tests and ECGs yield low-value results. Using model outputs to prioritize testing for patients with predicted probabilities of heart disease greater than 65 percent allows hospitals to direct efforts toward individuals who are more likely to benefit. If a hospital performs 10,000 cardiac screenings at $200 per test, even a 10% reduction in wasteful procedures would save

$200,000 annually. That money can be redirected towards high-risk patient monitoring or increased affordability, which can save lives.

Conclusion:
The Heart Risk Predictor shows how machine learning, specifically a Random Forest model, can be used to identify variables in a dataset and make accurate predictions. By interpreting the model's predictive analytics into clinical outcomes, this project can be used to implement cost-saving strategies and provide preventative care to high-risk individuals.

References:

Centers for Disease Control and Prevention (CDC). *Heart Disease Facts.* Updated 2024. https://www.cdc.gov/heart-disease/data-research/facts-stats/?CDC_AAref_Val=https://www.cdc.gov/heartdisease/facts.htm

American Heart Association. *Cardiovascular Care Costs and Utilization Statistics.* 2023 Report. https://www.ahajournals.org/doi/10.1161/CIR.0000000000001123

UCI Machine Learning Repository. *Heart Disease Data Set.* https://archive.ics.uci.edu/ml/datasets/heart+disease