Predicting Cardiovascular Disease Using Machine Learning

Cardiovascular disease remains the leading cause of death globally, claiming millions of lives annually. This project addresses a critical healthcare challenge**: Can machine learning models identify cardiovascular disease risk early enough to enable timely medical intervention and dramatically improve patient outcomes?**

Our research compares multiple machine learning approaches to predict cardiovascular disease from patient health records. We evaluated Logistic Regression, Random Forest, Gradient Boosting, and XGBoost models on a comprehensive dataset of 70,000 patient records, ultimately identifying XGBoost as the superior performer.

Strong Predictive Performance

Test accuracy of 0.73 with F1 score of 0.713, demonstrating reliable cardiovascular risk classification

GPU-Accelerated Training

Leverages GPU computing to efficiently process 70,000+ patient records with rapid model iteration

Non-Linear Pattern Recognition

Captures complex relationships between risk factors that linear models miss, improving predictive power

Dataset Overview and Preprocessing

Data Source and Characteristics

Our analysis utilizes the Cardiovascular Disease Dataset from Kaggle (cardio_train.csv), comprising 70,000 patient records. After rigorous data cleaning to remove physiologically unrealistic values, we retained 68,591 high-quality records for model training and evaluation. The dataset encompasses 13 comprehensive features spanning demographic information (age, gender), clinical measurements (height, weight, systolic blood pressure, diastolic blood pressure, cholesterol levels, glucose levels), and lifestyle factors (smoking status, alcohol consumption, physical activity levels). Our binary target variable indicates cardiovascular disease presence or absence..

01

Data Cleaning

Removed outliers and physiologically impossible values for blood pressure readings, height, and weight measurements

02

Exploratory Data Analysis

Explored feature distributions, correlations, and class balance to understand data and guide modeling.

03

Feature Engineering

Created age_years from days and calculated BMI (weight/height^2) to capture clinically relevant risk indicators

04

Train-Test Split

Implemented stratified 80/20 split to maintain class balance and ensure representative evaluation sets

Methodology and Model Development

We implemented a comprehensive model comparison strategy, evaluating four distinct machine learning approaches to identify the optimal predictor of cardiovascular disease risk. Our baseline Logistic Regression model provided interpretability, while ensemble methods - Random Forest, Gradient Boosting, and XGBoost - offered increasingly sophisticated pattern recognition capabilities.

Logistic Regression

Baseline linear model providing interpretable coefficients and establishing performance benchmarks for comparison

Random Forest

Ensemble of decision trees capturing feature interactions through parallel tree construction and majority voting

Gradient Boosting

Sequential ensemble method building trees iteratively to correct predecessor errors and minimize loss

XGBoost (CPU & GPU)

Advanced gradient boosting with regularization, parallel processing, and GPU acceleration for optimal performance

Rigorous Validation Strategy

To ensure robust performance estimates and model stability, we employed five-fold stratified cross-validation using XGBoost with GPU acceleration. This approach revealed exceptional consistency with a mean cross-validation F1 score of 0.7211 and remarkably low standard deviation of approximately 0.006.

Our feature engineering focused on clinically meaningful transformations: calculating Body Mass Index (BMI) from weight and height measurements, and converting age from the original day-based format to more interpretable years. These engineered features enhanced model interpretability while improving predictive accuracy.

Model Performance and Key Results

Our comprehensive evaluation on the held-out test set reveals XGBoost's superiority across multiple performance dimensions. While all models demonstrated reasonable predictive capability, XGBoost (GPU) achieving the highest accuracy and F1 score.

0.729	0.714	0.795	68.6K
XGBoost Accuracy	F1 Score	ROC AUC	Training Records
Best overall test set performance	Optimal precision-recall balance	Strong discrimination capability	After quality preprocessing

Confusion Matrix Analysis The XGBoost GPU model's confusion matrix reveals balanced performance: 5,370 true negatives and 4,635 true positives demonstrate strong predictive capability across both classes. With 2,152 false negatives and 1,562 false positives, the model maintains acceptable error rates while prioritizing sensitivity in this clinical context where missed diagnoses carry significant consequences.

Model Comparison Results

Logistic Regression

Accuracy: 0.7205

F1 Score: 0.7001

Baseline performance

Gradient Boosting

Accuracy: 0.7281

F1 Score: 0.7124

Strong results

Random Forest

Accuracy: 0.7077

F1 Score: 0.6992

Moderate performance

XGBoost (GPU)

Accuracy: 0.7299

F1 Score: 0.7135

Best overall

Feature importance analysis reveals that Body Mass Index (BMI), blood pressure measurements (both systolic and diastolic), and patient age emerge as the strongest predictors of cardiovascular disease risk. These findings align with established clinical knowledge, validating our model's ability to identify medically relevant risk factors while suggesting potential focus areas for preventive intervention strategies.

Team Contributions

TEAM 52

This cardiovascular disease prediction project represents a collaborative effort combining expertise in machine learning, data analysis, medical interpretation, and ethical considerations. Each team member contributed specialized skills to deliver a comprehensive and responsible predictive modeling solution.

Achal Nanjundamurthy

Led model architecture development, implemented cross-validation strategies, conducted hyperparameter tuning experiments, and researched advanced machine learning techniques

Khushi Bijkal

Managed data preprocessing pipeline, designed exploratory data analysis visualizations, interpreted clinical findings, and authored technical documentation

Suman Neupane

Provided model development support, conducted detailed error analysis, reviewed academic literature, and validated methodological approaches

Durga Prasad Narsing

Evaluated ethical implications and legal compliance considerations, ensured responsible AI practices, and contributed to project documentation