Heart Attack Prediction: A Machine Learning Approach

This project applies machine learning to predict heart attack occurrences using clinical and health-related features. By combining exploratory data analysis, feature engineering, and multiple classification models, we build a predictive system to identify patients at risk of heart disease. The workflow demonstrates practical data science techniques essential for healthcare analytics and demonstrates how structured data preparation leads to effective predictive models.

Dataset Features & Project Workflow

Our analysis focuses on 14 key clinical features that indicate heart disease risk. These span demographic information, diagnostic measurements, and physiological responses to stress testing.

Patient Demographics

- Age and Sex
- Baseline health indicators

Cardiac Measurements

- Resting blood pressure
- Cholesterol levels
- Maximum heart rate

Diagnostic Tests

- Chest pain classification
- Exercise-induced angina
- ST segment analysis

Stress Test Results

- Coronary vessel count
- Thalium test outcome
- ST depression metrics

Our four-phase methodology begins with exploratory data analysis to understand feature distributions and identify data quality issues. We then engineer features through encoding, scaling, and outlier handling. Finally, we train and compare Logistic Regression, K-Nearest Neighbors, and Decision Tree models to select the best performer for heart attack prediction.

Data Pipeline & Model Implementation

Data Preparation Phase

Exploratory Analysis: Inspect distributions, identify missing values, and detect outliers. Correlation analysis reveals feature relationships with target variable.

Feature Engineering: Handle missing data, remove or treat outliers, encode categorical variables (ChestPainType, Sex), and scale numerical features for model compatibility.

Modeling & Evaluation

Model Training: Implement three classification approaches—Logistic Regression for baseline performance, KNN for local pattern detection, and Decision Trees for interpretable rules.

Performance Metrics: Compare accuracy, precision, recall, and F1-score to identify the strongest predictor for clinical deployment.

1

Current Capabilities

Baseline models provide solid foundation for heart disease prediction without hyperparameter optimization.

2

Future Enhancements

Grid search optimization, ensemble methods like Random Forest and XGBoost, plus advanced interpretation using SHAP values.

3

Clinical Impact

Interpretable predictions support clinical decision-making by identifying high-risk patients for preventive intervention.