Heart Disease Prediction Using Machine Learning

1. Executive Summary

Objective
To develop a reliable, data-driven system for predicting the presence of heart disease in patients, enabling earlier detection and supporting clinical decision-making.

Approach
A supervised machine learning framework was implemented using a structured medical dataset of patient health metrics. Multiple classification algorithms were benchmarked to identify the most accurate and clinically reliable model for heart disease prediction.

Key Results

- **Best Model:** Random Forest Classifier
- **Prediction Accuracy:** 86.89% on the test set
- **Diagnostic Strength:** Achieved a **ROC AUC score of 0.9427**, indicating excellent discriminative ability between diseased and non-diseased patients
- **Clinical Insight:** The model demonstrated high recall (96.97%), minimizing false negatives—critical for medical screening applications

Business & Clinical Impact
This system can be used as a **clinical decision-support tool** to identify high-risk patients, prioritize diagnostic testing, and improve preventive healthcare outcomes.

2. Data Engineering & Preparation

Data Source

- Dataset: heart.csv
- Records: 303 patient observations
- Features: 13 medical attributes + 1 target variable

Data Quality Assessment

- **Missing Values:** None detected (100% complete dataset)
- **Consistency:** All features were within expected clinical ranges

Preprocessing Pipeline

- **Categorical Encoding:** Binary and ordinal medical variables were encoded numerically
- **Feature Scaling:** StandardScaler applied to numerical features to ensure fair model comparison
- **Train-Test Split:** 80/20 split to ensure unbiased model evaluation

This preprocessing ensured stable convergence of models and robust performance metrics.

3. Feature Overview & Medical Relevance

The model utilized clinically meaningful patient attributes, including:

- **Demographics:** Age, Sex
- **Cardiac Indicators:** Chest pain type (cp), Maximum heart rate (thalach), ST depression (oldpeak), Exercise-induced angina (exang)
- **Physiological Metrics:** Resting blood pressure (trestbps), Serum cholesterol (chol), Fasting blood sugar (fbs)
- **Diagnostic Results:** ECG readings (restecg), Fluoroscopy vessels (ca), Thallium stress test (thal)

Target Variable

- target = 1: Presence of heart disease
- target = 0: Absence of heart disease

4. Supervised Learning: Disease Prediction

Goal

To build a high-performance classification model capable of predicting heart disease presence using patient health data.

Models Evaluated

- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest Classifier
- XGBoost Classifier

Evaluation Metrics

- Accuracy
- Precision
- Recall
- F1-Score
- ROC AUC Score

Model Performance Summary (Test Set)

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
Logistic Regression	0.8033	0.7692	0.9091	0.8333	0.8690
SVM	0.8361	0.7949	0.9394	0.8611	0.8864
Random Forest	**0.8689**	**0.8421**	**0.9697**	**0.9014**	**0.9427**
XGBoost	0.8033	0.7561	0.9394	0.8378	0.8561

5. Why Random Forest?

The **Random Forest Classifier** was selected as the optimal model due to:

- **Highest Overall Accuracy (86.89%)**
- **Exceptional Recall (96.97%)**, critical for minimizing missed diagnoses
- **Strong ROC AUC (0.94)**, demonstrating excellent class separation
- **Robustness to Feature Interactions**, capturing complex non-linear medical relationships

This makes Random Forest particularly well-suited for **healthcare risk prediction**, where sensitivity is paramount.

6. Exploratory Insights & Key Drivers

Analysis and visualizations revealed several important medical insights:

1. **Exercise Capacity Matters**
    - Maximum heart rate achieved (thalach) showed strong differentiation between healthy and diseased patients.
2. **Chest Pain Type is Highly Predictive**
    - Certain chest pain categories were strongly associated with positive heart disease diagnoses.
3. **Cardiac Stress Indicators**
    - ST depression (oldpeak) and exercise-induced angina (exang) were significant indicators of heart disease risk.

These insights align well with established clinical knowledge, reinforcing the model's credibility.

8. Recommendations for Stakeholders

For Healthcare Providers

- Integrate the Random Forest model as a **screening support tool** to flag high-risk patients early.
- Use predictions to prioritize advanced diagnostic tests such as angiography.

For Hospitals & Clinics

- Deploy the model in triage systems to optimize patient flow and reduce diagnostic delays.

For Data Science & Research Teams

- Expand the dataset to improve generalizability across demographics.
- Explore model explainability tools (e.g., SHAP) to enhance clinician trust.

9. Future Enhancements

- **Hyperparameter Optimization:** GridSearchCV and RandomizedSearchCV for further accuracy gains
- **Feature Engineering:** Deriving composite cardiac risk indicators
- **Ensemble Learning:** Blending Random Forest and Gradient Boosting models
- **Deep Learning Models:** Neural networks for large-scale clinical datasets