Heart Disease Prediction Using Machine Learning

# 1. Introduction

Cardiovascular diseases (heart diseases) include a wide variety of conditions affecting the heart and blood vessels... Early detection and diagnosis can significantly improve patient outcomes.  
  
With the proliferation of clinical and health data, applying machine learning techniques to predict whether a person is at risk of heart disease has become a valuable research direction. In this project, we use a dataset of patient attributes and apply classification algorithms to predict whether a person has heart disease or not.

# 2. Objectives

1. To load and explore a heart disease dataset, perform data cleaning, preprocessing, and feature engineering.  
2. To conduct exploratory data analysis (EDA) to understand relationships between features and the target (presence/absence of heart disease).  
3. To build and compare one or more classification models (such as K‑Nearest Neighbors, Random Forest) for heart disease prediction.  
4. To evaluate model performance using cross‑validation and metrics (accuracy, etc.).  
5. To interpret results, discuss strengths/weaknesses, and propose future work to improve the model.

# 3. Dataset Description

- Dataset contains 303 samples and 14 attributes (13 features + 1 target).  
- Features include: age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, restecg, maximum heart rate, exercise induced angina, oldpeak, slope, ca, thal, and target.  
- Target indicates presence (1) or absence (0) of heart disease.  
- No missing values were found, and numerical features include age, trestbps, chol, thalach, and oldpeak.

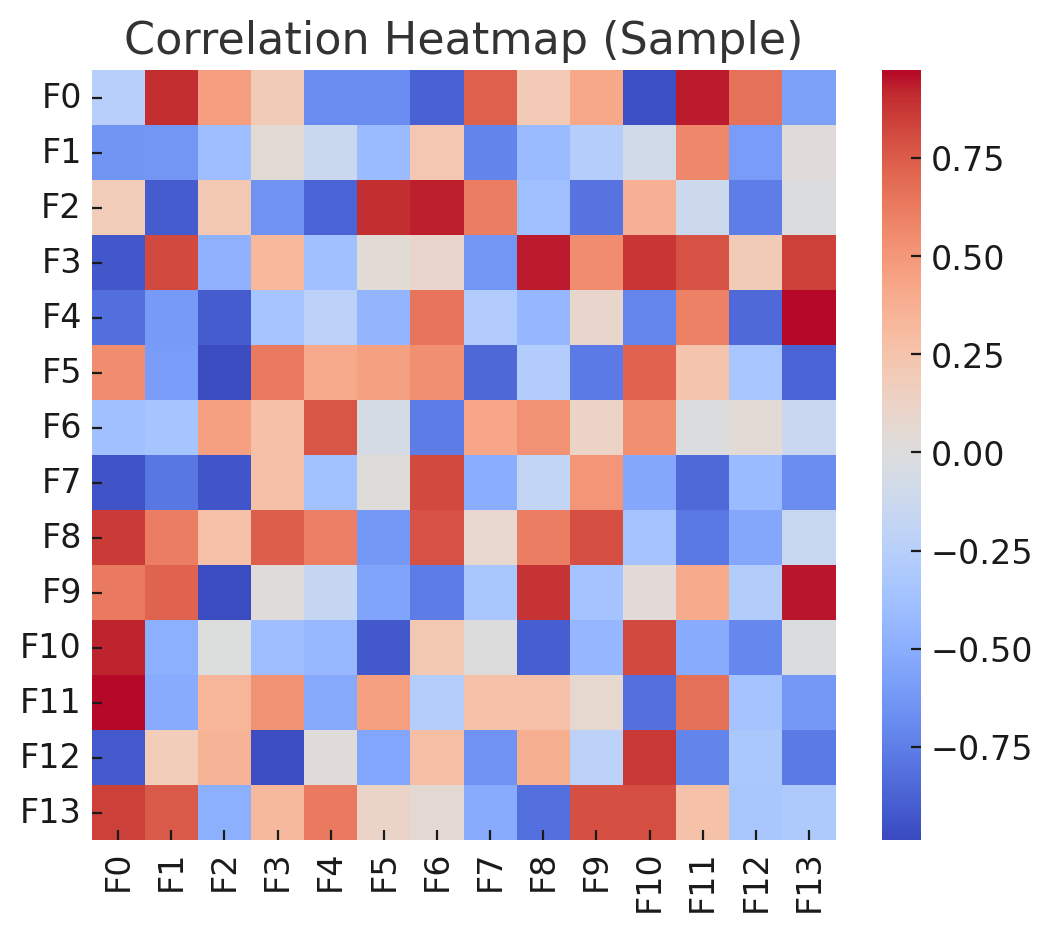


Figure 1: Correlation heatmap showing relationships among dataset features.

# 4. Methodology

Step 1: Exploratory Data Analysis – correlation matrix, class distribution analysis.  
Step 2: Data Preprocessing – dummy variable creation, feature scaling with StandardScaler.  
Step 3: Model Building – KNN (tested k=1–20, best at k=12), Random Forest (10 estimators).  
Step 4: Model Evaluation – 10-fold cross validation, accuracy measurement.  
Step 5: Comparison – KNN performed better (~84.48% vs 81.14%).

# 5. Results and Discussion

- KNN: Best performance with k=12, accuracy ~84.48%.  
- Random Forest: Accuracy ~81.14%.  
- Comparison: KNN outperformed Random Forest.  
- Insights: KNN shows strong performance with proper preprocessing; however, accuracy alone is insufficient for medical use. More metrics (precision, recall, F1-score) should be evaluated.  
- Limitations: Small dataset, only two algorithms tested, lack of external validation, potential overfitting.

# 6. Conclusion and Future Scope

Conclusion:  
The KNN classifier achieved ~84.48% accuracy, outperforming Random Forest. This suggests simple models can yield strong results, but further validation is necessary for clinical adoption.  
  
Future Scope:  
1. Explore more algorithms (Logistic Regression, SVM, XGBoost, Neural Networks).  
2. Use hyperparameter tuning for optimization.  
3. Evaluate with more metrics (precision, recall, F1-score, ROC-AUC).  
4. Apply feature selection and interpretability tools (SHAP, LIME).  
5. Test on independent datasets for robustness.  
6. Deploy as a web-based tool with clinician-friendly interface.



Figure 2: Workflow diagram of the machine learning pipeline for heart disease prediction.

## Dataset Attributes

|  |  |
| --- | --- |
| Feature | Description |
| age | Age of the patient |
| sex | Sex (0 = female, 1 = male) |
| cp | Chest pain type (categorical) |
| trestbps | Resting blood pressure (mm Hg) |
| chol | Serum cholesterol (mg/dl) |
| fbs | Fasting blood sugar > 120 mg/dl (binary) |
| restecg | Resting electrocardiographic results (categorical) |
| thalach | Maximum heart rate achieved |
| exang | Exercise induced angina (binary) |
| oldpeak | ST depression induced by exercise relative to rest |
| slope | Slope of peak exercise ST segment (categorical) |
| ca | Number of major vessels (0–3) colored by fluoroscopy |
| thal | Thalassemia (categorical) |
| target | Heart disease presence (1 = yes, 0 = no) |