# HEART DISEASE PREDICTION USING MACHINE LEARNING

* Prepared by: Akshat Jain
* B.Tech, 4th Year
* Teerthanker Mahaveer University

## 1. Introduction

This project focuses on predicting the likelihood of a patient having heart disease using data analytics and machine learning techniques.   
The dataset used is the Heart Disease UCI dataset from Kaggle, which includes several medical and demographic features of patients.   
The aim is to build a model that can analyze patient data and predict the probability of heart disease, helping in early detection and prevention.

## 2. Dataset Description

The dataset consists of the following key features:  
- Age — Patient’s age in years  
- Sex — (1 = male, 0 = female)  
- Chest Pain Type (cp) — 0–3  
- Resting Blood Pressure (trestbps)  
- Cholesterol (chol)  
- Fasting Blood Sugar (fbs)  
- Resting ECG Results (restecg)  
- Maximum Heart Rate Achieved (thalach)  
- Exercise-Induced Angina (exang)  
- ST Depression (oldpeak)  
- Slope, ca, thal — Clinical indicators  
- Target — 1 = heart disease, 0 = no disease

## 3. Models Used

Three machine learning models were implemented to analyze and predict heart disease:  
1. Logistic Regression — A statistical model that estimates the probability of heart disease.  
2. Random Forest — An ensemble of multiple decision trees for improved accuracy.  
3. XGBoost — A boosting-based algorithm for optimized performance (optional).

## 4. How Prediction Works

The model takes input features such as age, cholesterol, and blood pressure to estimate the probability of heart disease.  
For Logistic Regression, the equation is:

* z = b0 + b1\*x1 + b2\*x2 + ... + bn\*xn
* P(heart disease) = 1 / (1 + e^(-z))

If P ≥ 0.5, the prediction is 'Heart Disease'; otherwise, 'No Heart Disease'.  
Random Forest and XGBoost, on the other hand, use tree-based decision structures to calculate probabilities.

## 5. Example Calculation

Suppose the Logistic Regression model learned the equation:  
z = -5 + 0.03(age) + 0.02(chol) + 1.5(cp) - 1.2(thalach)  
  
For a patient with: age = 52, chol = 250, cp = 1, thalach = 170  
z = -5 + 0.03(52) + 0.02(250) + 1.5(1) - 1.2(170) = -200.94  
P = 1 / (1 + e^(200.94)) ≈ 0.0 → Prediction: No Heart Disease

## 6. Model Evaluation Metrics

To measure model performance, the following metrics are used:  
- Accuracy — The percentage of correct predictions.  
- Precision — How many predicted positives were actually positive.  
- Recall — How many actual positives were correctly predicted.  
- F1 Score — A balance between precision and recall.  
- ROC-AUC — The model’s ability to distinguish between classes.

## 7. Predicting for a New Patient

After training, new patient data can be used for prediction. Example:

new\_patient = {'age':52, 'sex':1, 'cp':0, 'trestbps':130, 'chol':250,  
 'fbs':0, 'restecg':1, 'thalach':180, 'exang':0,  
 'oldpeak':1.0, 'slope':2, 'ca':0, 'thal':2}

The data is encoded, scaled, and passed into the model.   
If the probability ≥ 0.5, the model predicts Heart Disease, otherwise No Heart Disease.

## 8. Insights and Applications

- Logistic Regression provides interpretability by showing which features increase risk.  
- Random Forest and XGBoost improve accuracy and handle complex data interactions.  
- These models can help hospitals identify high-risk patients and assist doctors in prioritizing further tests.

## 9. Conclusion

This project demonstrates how machine learning can assist healthcare professionals in predicting heart disease based on clinical data.   
The model helps identify at-risk patients, enabling early diagnosis and treatment.   
Combining data science with healthcare enhances decision-making and improves patient outcomes.

Prepared by: Akshat Jain

B.Tech, 4th Year — Teerthanker Mahaveer University