# DAYANANDA SAGAR COLLEGE OF ENGINEERING

(An Autonomous Institute affiliated to VTU, Belagavi, Approved by AICTE & ISO 9001:2008 Certified)
Accredited by National Assessment & Accreditation Council (NAAC) with ‘A’ grade,
Shavige Malleshwara Hills, Kumaraswamy Layout, Bengaluru-111

**22CS53 - ARTIFICIAL INTELLIGENCE & MACHINE LEARNING**

**ASSIGNMENT**

**On**

# “PREDICTING HEART DISEASE RISK”

Submitted by

**AMULYA M G- 1DS22CS030**

**ANKITA A Y- 1DS22CS035**

**Fifth Semester B.E (CSE)**

**2024-2025**

Under the guidance of

**Dr. Anupama Girish**
**Assistant Professor**
**Dept. of CSE**
**DSCE, Bangalore**

**Department of Computer Science and Engineering**
**Dayananda Sagar College of Engineering**
**Bangalore-560078**

# Table Of Contents

## Problem Description:

The "Heart Disease Prediction" task aims to create a model that predicts whether a patient has heart disease based on medical attributes like age, gender, cholesterol levels, and more. This classification problem involves handling numerical and categorical data, requiring careful preprocessing and feature engineering. Key challenges include managing missing data, multicollinearity, and feature interactions. Advanced techniques like logistic regression, Random Forest, Gradient Boosting, and Support Vector Machines are utilized to build robust models. Emphasizing exploratory data analysis (EDA), domain understanding, and hyper parameter optimization is crucial for refining predictions and ensuring model interpretability.

## Identified Dataset- (link to theidentified data set)

LINK

Here is an example of what the first few rows of a heart disease prediction dataset might look like:

| Age | Gender | Cholesterol | Blood Pressure | Heart Disease |
|-------|----------|---------------|--------------------|------------------|
| 55 | Male | 240 | 130 | Yes |
| 60 | Female | 210 | 120 | No |
| 45 | Male | 190 | 140 | Yes |
| 50 | Female | 230 | 150 | No |
| 65 | Male | 260 | 160 | Yes |

## Identification of Predictors and Response

The Heart Disease Prediction case study focuses on developing predictive models to identify the presence of heart disease in patients. The first step in solving this problem is to identify the response variable (target) and the predictors (features) that influence the target. These variables are derived from the provided dataset.

### Response Variable:

**Target:** The presence or absence of heart disease. It is a categorical variable with values (1) indicating the presence and (0) indicating the absence of heart disease. This is the dependent variable, influenced by various medical attributes related to the patient's health. The objective is to build a model capable of accurately predicting the presence of heart disease.

## Predictor Variables :

The predictor variables include both quantitative and qualitative medical attributes of the patients. These can be broadly categorized into numerical and categorical variables.

### 1. Numerical Predictors:

**Age:** Age of the patient in years. Older patients may have a higher risk of heart disease.

Resting Blood Pressure (trestbps): Resting blood pressure in mm Hg. Higher blood pressure levels are often associated with cardiovascular issues.

**Serum Cholesterol (chol):** Serum cholesterol in mg/dl. High cholesterol levels can indicate a higher risk of heart disease.

**Maximum Heart Rate Achieved (thalach):** The highest heart rate achieved during exercise. Lower maximum heart rates can indicate cardiac problems.

**Oldpeak :** ST depression induced by exercise relative to rest, which measures the effectiveness of the heart under stress.

### 2. Categorical Predictors:

**Sex :** Gender of the patient (1 = male, 0 = female). Gender differences can affect the risk and manifestation of heart disease.

**Chest Pain Type (cp) :** Type of chest pain experienced (0 to 3). Different types of chest pain can indicate varying levels of heart disease risk.

**Fasting Blood Sugar (fbs) :** Fasting blood sugar > 120 mg/dl (1 = true, 0 = false). High fasting blood sugar levels can be a risk factor for heart disease.

**Resting Electrocardiographic Results (restecg) :** Resting ECG results (0 to 2). Abnormal ECG results can indicate cardiac problems.

**Exercise Induced Angina (exang) :** Exercise-induced angina (1 = yes, 0 = no). The presence of angina during exercise is a significant predictor of heart disease.

**Slope :** The slope of the peak exercise ST segment (0 to 2). The slope of ST segments during peak exercise can indicate heart health.

**Number of Major Vessels (ca) :** Number of major vessels colored by fluoroscopy (0 to 3). More colored vessels can indicate a higher risk of heart disease.

**Thalassemia (thal) :** Blood disorder status (1 to 3). Different types of thalassemia can impact heart health.

### Importance of Predictors:

The combination of numerical and categorical predictors provides a comprehensive understanding of factors influencing the risk of heart disease. Numerical predictors offer measurable health attributes, while categorical variables add context, such as gender and types of symptoms. For example:

Age : Older age typically correlates with higher heart disease risk.

Cholesterol Levels : Higher cholesterol levels indicate a higher risk.

Chest Pain Type : Certain types of chest pain are more indicative of heart disease.

## Design of the Data Analytics Process

It requires a systematic approach to data analytics, focusing on identifying patterns, addressing key challenges, and building predictive models. The design of the data analytics process revolves around answering specific questions that guide the analysis and decision-making process.

### 1. Data Exploration Questions

What is the distribution of the target variable (heart disease)?

Understanding the distribution helps determine if there are class imbalances that need to be addressed.

Are there missing values in the dataset, and how should they be handled?

Missing values can impact model performance. Questions about the proportion and nature of missing data guide the choice of imputation techniques or exclusion of irrelevant features.

What are the relationships between predictors and the target variable?

Exploring correlations or visualizing relationships (e.g., scatter plots for numerical variables and bar plots for categorical variables) identifies key drivers of heart disease.

### 2. Feature Engineering Questions

Which variables have the strongest predictive power?

Techniques like correlation analysis, feature importance ranking, or variance inflation factors help isolate the most influential predictors for heart disease.

Are there opportunities for creating new features?

Questions about potential feature combinations, such as creating a risk score from multiple predictors or combining age and cholesterol levels, enrich the dataset.

Are categorical variables represented appropriately?

Determining whether to use one-hot encoding, ordinal encoding, or another transformation depends on the variable type and its relationship with the target.

### 3. Model Selection and Validation Questions

What classification techniques are most suitable for the dataset?

Depending on the complexity of the data, models like logistic regression, decision trees, Random Forest, or ensemble methods (e.g., XGBoost) can be evaluated.

How should the dataset be split for training and testing?

Addressing the size and proportion of training and testing datasets ensures robust model validation. Cross-validation strategies can also be explored.

Which evaluation metrics best measure model performance?

Metrics like accuracy, precision, recall, F1-score, and ROC-AUC quantify the effectiveness of the model in predicting heart disease.

### 4. Model Optimization Questions

What hyper parameters should be tuned for optimal performance?

Exploring grid search or random search approaches answers questions about parameter tuning, such as the depth of decision trees in Random Forest or the learning rate in gradient boosting models.

How can overfitting or under fitting be addressed?

Techniques like regularization, cross-validation, and feature selection prevent overfitting, ensuring the model generalizes well to unseen data.

### 5. Deployment and Interpretation Questions

How can the model's predictions be interpreted for decision-making?

Understanding feature importance or using techniques like SHAP (SHapley Additive exPlanations) enables stakeholders to make informed decisions.

What insights can be derived for healthcare trends?

Identifying patterns, such as the influence of age or cholesterol levels on heart disease risk, provides actionable insights for medical professionals and policymakers.

## Machine Learning Model to Be Used: Random Forest Classifier

The Random Forest Classifier is a robust choice due to its ability to handle complex relationships and interactions between features while maintaining high prediction accuracy.

**Description of Random Forest Classifier**: Random Forest is an ensemble machine learning technique that creates multiple decision trees and combines their outputs to make a final prediction.

**Features**:

1. **Ensemble Method**: Combines multiple decision trees to improve performance.

2. **Bootstrapping**: Uses bootstrapped datasets to create individual trees.
3. **Feature Importance**: Provides insights into the importance of each feature.
4. **Robustness**: Reduces overfitting compared to single decision trees.

## Implementation Workflow:

1. **Preprocessing**: Handle missing values, encode categorical features, and scale numerical values.
2. **Model Initialization**: Define hyperparameters like the number of trees, maximum depth, and min_samples_split.
3. **Training**: Fit the model using the training dataset.
4. **Hyperparameter Tuning**: Use grid search or random search for optimal hyperparameter selection.
5. **Evaluation**: Validate model performance using metrics like accuracy, precision, recall, and ROC-AUC.
6. **Feature Importance Analysis**: Assess the contribution of each predictor to the final model.

**Tools and Libraries**:

1. **Scikit-Learn**: Provides an easy-to-use implementation of Random Forest Classifier.
2. **XGBoost**: An advanced gradient boosting library for comparison.

**Random Forest Classifier's adaptability and robustness make it an excellent choice for predicting heart disease risk with high accuracy, while its feature interpretability provides valuable insights into the drivers of heart disease.**

## Expressions:

Random Forest involves several key concepts that govern its working.

1. **Bootstrap Sampling**:
   - Randomly sample with replacement from the training data to create multiple subsets.
2. **Building Individual Trees**:
   - Each tree is trained on a bootstrap sample with a random subset of features.
3. **Combining Predictions**:
   - Final prediction is made by averaging the predictions of individual trees (regression) or taking the majority vote (classification).
4. **Feature Importance**:
   - Calculated by measuring the decrease in model accuracy when a feature is permuted.

# Expressions:

Gradient Boosting Regressor (GBR) involves several key formulas that govern it's working.

**1. Initial Model**

The initial model F0(x)F_0(x)F0(x) is often chosen as the constant value that minimizes the loss function. For Mean Squared Error (MSE), it is the mean of the target values:

$$F_0(x) = \arg\min_c \sum_{i=1}^{n} \text{Loss}(y_i, c)$$

For MSE:

$$F_0(x) = \frac{1}{n}\sum_{i=1}^{n} y_i$$

**2. Negative Gradient (Pseudo-Residuals)**

At each iteration mmm, calculate the negative gradient of the loss function with respect to the model's prediction

$$r_{im} = -\frac{\partial \mathrm{Loss}\big(y_i, F_{m-1}(x_i)\big)}{\partial F_{m-1}(x_i)}$$

For MSE:

$$r_{im} = y_i - F_{m-1}(x_i)$$

**3. Fit Weak Learner**

Fit a weak learner (e.g., a decision tree) to the pseudo-residuals. The weak learner minimizes the squared error:

$$h_m(x) = \arg\min_h \sum_{i=1}^{n} \big(r_{im} - h(x_i)\big)^2$$

**4. Update Model**

Update the model by adding the scaled predictions of the weak learner to the previous model:

$$F_m(x) = F_{m-1}(x) + \mathrm{v} \cdot h_m(x)$$

Where:

- ν: Learning rate, 0<ν≤1

**5. Final Prediction**

After MMM iterations, the final model is:

$$F_M(x) = F_0(x) + \mathrm{v} \sum_{m=1}^{M} h_m(x)$$

**Loss Function (Examples)**

**Mean Squared Error (MSE):**

$$\mathrm{Loss}\big(y, F(x)\big) = \frac{1}{n}\sum_{i=1}^{n} \big(y_i - F(x_i)\big)^2$$

**Mean Absolute Error (MAE):**

$$\mathrm{Loss}\big(y, F(x)\big) = \frac{1}{n}\sum_{i=1}^{n} |y_i - F(x_i)|$$