## **Project: Predictive Modeling of Heart Disease**

### **1. Introduction and Project Goal**

Cardiovascular diseases (CVDs) are the leading cause of death globally, making their early detection a critical public health challenge. Machine learning and deep learning models offer promising avenues for developing accurate predictive tools to assist clinicians in diagnosing heart disease.

This project aims to leverage the UCI Heart Disease dataset to develop and evaluate robust machine learning and deep learning models for predicting the presence of heart disease in patients. Through this hands-on project, students will gain a comprehensive understanding of the entire machine learning pipeline, from data exploration and preprocessing to model training, evaluation, and interpretation.

The primary objective is to build a classifier that can accurately predict the presence of heart disease based on a set of patient attributes. This will involve a binary classification task where the model will predict either the absence (0) or presence (1) of heart disease.

### **2. Dataset Description**

This project utilizes the "Heart Disease" dataset from the UCI Machine Learning Repository. While the full dataset contains 76 attributes, we will focus on a widely used subset of 14 attributes from the Cleveland Clinic Foundation database, which is a common benchmark for this task.

**Dataset Link:**<https://archive.ics.uci.edu/dataset/45/heart+disease>

### **3. Task Outline**

This project is divided into four main tasks designed to guide students through a complete machine learning workflow.

#### **Task 1: Data Understanding and Exploration (EDA)**

The initial and one of the most crucial steps is to thoroughly understand the data you are working with.

* **Objective:** To gain insights into the dataset's structure, identify potential data quality issues, and understand the relationships between different variables.
* **Activities:**
  + Load the dataset using a library like Pandas.
  + Generate descriptive statistics for all numerical and categorical features.
  + Create visualizations to understand the distribution of each feature (e.g., histograms for numerical features, bar charts for categorical features).
  + Explore the relationships between pairs of features using scatter plots and correlation matrices.
  + Analyze the balance of the target variable. Is the dataset imbalanced?
  + Create visualizations to understand how different features relate to the presence or absence of heart disease (e.g., box plots, violin plots).

#### **Task 2: Data Preprocessing and Feature Engineering**

Raw data is rarely ready for direct use in machine learning models. This task focuses on cleaning and preparing the data for modeling.

* **Objective:** To handle missing data, encode categorical variables, and scale numerical features to make them suitable for machine learning algorithms.

#### **Task 3: Model Training and Development**

In this phase, you will train a variety of machine learning and deep learning models to predict heart disease.

* **Objective:** To build, train, and tune different classification models.
* **Models to Implement:**
  + **Traditional Machine Learning Models:**
    - Logistic Regression
    - K-Nearest Neighbors (KNN)
    - Support Vector Machines (SVM)
    - Gradient Boosting Machines (e.g., XGBoost, LightGBM)
  + **Deep Learning Model:**
    - Build a simple feedforward neural network (also known as a multilayer perceptron - MLP) using a framework like PyTorch. You can experiment with the number of layers, neurons per layer, activation functions, and optimizers.
* **Activities:**
  + Train each of the specified models on the training dataset.
  + For models with important hyperparameters (e.g., the number of neighbors in KNN, C and gamma in SVM, tree depth in Random Forest), perform hyperparameter tuning using techniques like GridSearchCV or RandomizedSearchCV to find the optimal settings.
  + Document the architecture and parameters of your deep learning model.

#### **Task 4: Model Evaluation and Comparison**

The final step is to rigorously evaluate the performance of your trained models and select the best one for the task.

* **Objective:** To assess and compare the predictive performance of the different models using appropriate evaluation metrics.
* **Evaluation Metrics:**
  + **Accuracy:** The proportion of correctly classified instances.
  + **Precision:** The proportion of positive identifications that were actually correct.
  + **Recall (Sensitivity):** The proportion of actual positives that were identified correctly.
  + **F1-Score:** The harmonic mean of precision and recall.
  + **Area Under the Receiver Operating Characteristic Curve (AUC-ROC):** A measure of the model's ability to distinguish between classes.
* **Activities:**
  + Evaluate each trained model on the held-out testing set using the specified metrics.
  + Generate a confusion matrix for each model to visualize its performance in detail.
  + Create a summary table comparing the performance of all the models across the different evaluation metrics.
  + Based on your evaluation, identify the best-performing model for this task and justify your choice.
  + For the best-performing traditional model and the deep learning model, analyze the feature importance to understand which patient attributes are most influential in predicting heart disease.

### **5. Deliverables**

Students are expected to submit a comprehensive report and a well-documented Jupyter Notebook that includes:

* A clear explanation of each step of the project.
* The code used for data exploration, preprocessing, model training, and evaluation.
* Visualizations and tables that support your findings.
* A detailed report of the entire project in .pdf format. (whatever you have done)