# Heart Disease Prediction

## Main Objective

Cardiovascular diseases (CVDs) are the leading cause of death globally, responsible for an estimated 17.9 million deaths each year — about 31% of all deaths worldwide. Four out of five CVD-related deaths are caused by heart attacks and strokes, and one-third occur prematurely in people under the age of 70. Heart failure is a common event triggered by CVDs.

In this project, we aim to predict the likelihood of heart disease using machine learning techniques based on 11 health-related features. Early detection and intervention, especially for individuals with risk factors like hypertension, diabetes, hyperlipidemia, or an established cardiovascular condition, can significantly reduce mortality — and machine learning models can assist greatly in this early identification.

## Dataset Description

The dataset contains 12 columns:

* **Features**: Age, Sex, ChestPainType, RestingBP, Cholesterol, FastingBS, RestingECG, MaxHR, ExerciseAngina, Oldpeak, ST\_Slope
* **Target**: HeartDisease (1: presence of heart disease, 0: absence)

The goal is to build a predictive model using these features to determine the presence of heart disease.

## Data Exploration and Cleaning

* The dataset was reviewed for missing values, inconsistencies, outliers, and duplicates. No major issues were found.
* Five features (Sex, ChestPainType, RestingECG, ExerciseAngina, ST\_Slope) were identified as categorical and were transformed into numerical values using one-hot encoding (pd.get\_dummies with drop\_first=True) to avoid multicollinearity.
* The continuous variables were left unchanged after basic validation.

## Model Training Summary

Six machine learning models were trained:

1. **Logistic Regression** (baseline, with Ridge, and with Lasso penalties)
2. **K-Nearest Neighbors (KNN)**
3. **Support Vector Machine (SVM)**
4. **Decision Tree Classifier**
5. **Random Forest Classifier**
6. **XGBoost Classifier**

All models were evaluated using a **70/30 train-test split**, and random\_state=42 was consistently applied for reproducibility.

## Model Summary

**1. Logistic Regression**

* Performed strongly as a baseline model.
* Ridge and Lasso regularization were applied to prevent overfitting.
* The model was highly interpretable and provided good balance between precision and recall.

**2. K-Nearest Neighbors (KNN)**

* Achieved comparable or better performance compared to more complex models.
* Simple and effective, though sensitive to the choice of 'k' (number of neighbors).
* Easy to explain predictions based on nearest data points.

**3. Support Vector Machine (SVM)**

* Slightly lower performance compared to Logistic Regression and KNN.
* SVM requires careful tuning of hyperparameters like the kernel type and regularization.
* Higher computational cost for larger datasets.

**4. Decision Tree Classifier**

* Easy to interpret and visualize.
* Captured non-linear relationships effectively.
* However, prone to overfitting without pruning or depth control.

**5. Random Forest Classifier**

* Improved performance over single decision trees due to ensemble learning.
* More robust and less overfitting compared to a single decision tree.
* Slightly more complex and less interpretable than Logistic Regression or KNN.

**6. XGBoost Classifier**

* Powerful gradient boosting algorithm.
* Achieved reasonable performance but did not significantly outperform simpler models.
* Requires careful hyperparameter tuning and is computationally heavier.

## Model Evaluation

Models were compared based on four performance metrics:

* **Accuracy**
* **Precision**
* **Recall**
* **F1-score**

Overall, **Logistic Regression** and **K-Nearest Neighbors (KNN)** stood out in balancing performance and interpretability.

## Model Recommendation

For this dataset, **Logistic Regression** and **K-Nearest Neighbors (KNN)** are recommended:

* They performed comparably or better than the more complex models.
* They are computationally efficient.
* They are easy to interpret — a crucial factor for medical machine learning applications where transparency is important.

## Key Findings and Insights

* The strongest predictors for heart disease in the dataset are:
  + **Exercise Angina**
  + **Sex**
  + **Age**
  + **ST\_Slope (Flat)**
* Other important features include:
  + **ST\_Slope (Up)**
  + **Chest Pain Type**
  + **Maximum Heart Rate (MaxHR)**

## Suggestions for Next Steps

* Future analyses could benefit from incorporating additional features such as:
  + Migraine history
  + Nationality
  + Other demographic or lifestyle factors
* Additionally, exploring deep learning models could be valuable for larger and more complex datasets.