# Problem Statement

Heart disease is broad term used for diseases and conditions affecting the heart and circulatory system. Since heart is amongst the most vital organs of the body, its diseases affect other organs and part of the body as well. There are several different types and forms of heart diseases. The most common ones cause blockage of the coronary arteries known as coronary artery disease, malfunctioning in the valves of the heart, enlargement in the size of heart and several others leading to heart failure and heart attack. It’s the heart disease dataset from this dataset we can derive various insights that help us know the weightage of each feature and how they are interrelated to each other but this time our sole aim is to detect the probability of person that will be affected by a savior heart problem or not.

# Dataset Source

[Kaggle](https://www.kaggle.com/datasets/colewelkins/cardiovascular-disease)

# Description about the dataset

This dataset consolidates information from two primary sources:

1. UCI Machine Learning Repository - Heart Disease Dataset Kaggle
2. Heart Disease Dataset by YasserH

The primary aim is to predict the presence or absence of cardiovascular disease based on various patient metrics.

# Description of the attributes/variables/columns of the dataset

1. **ID**: Unique identifier for each patient.
2. **age**: Age of the patient in days.
3. **age\_years**: Age of the patient in years (derived from age).
4. **gender**: Gender of the patient. Categorical variable (1: Female, 2: Male).
5. **height**: Height of the patient in centimeters.
6. **weight**: Weight of the patient in kilograms.
7. **ap\_hi**: Systolic blood pressure.
8. **ap\_lo**: Diastolic blood pressure.
9. **cholesterol**: Cholesterol levels. Categorical variable (1: Normal, 2: Above Normal, 3: Well Above Normal).
10. **gluc**: Glucose levels. Categorical variable (1: Normal, 2: Above Normal, 3: Well Above Normal).
11. **smoke**: Smoking status. Binary variable (0: Non-smoker, 1: Smoker).
12. **alco**: Alcohol intake. Binary variable (0: Does not consume alcohol, 1: Consumes alcohol).
13. **active**: Physical activity. Binary variable (0: Not physically active, 1: Physically active).
14. **cardio**: Presence or absence of cardiovascular disease. Target variable. Binary (0: Absence, 1: Presence).
15. **bmi**: Body Mass Index, derived from weight and height. Calculated as ( \text{BMI} = \frac{\text{weight (kg)}}{\text{height (m)}^2} ).
16. **bp\_category**: Blood pressure category based on ap\_hi and ap\_lo. Categories include "Normal", "Elevated", "Hypertension Stage 1", "Hypertension Stage 2", and "Hypertensive Crisis".
17. **bp\_category\_encoded**: Encoded form of bp\_category for machine learning purposes.