# 2. Data Description

## 2.1 Dataset Overview

The dataset consists of 70,000 entries and 13 columns, sourced from routine medical checkups. It is used for predicting the presence of cardiovascular disease (binary classification). Each row represents data for one patient. The features include demographic information, physiological measurements, lifestyle choices, and the target variable 'cardio'.

## 2.2 Features Description

|  |  |  |
| --- | --- | --- |
| Column Name | Type | Description |
| id | int | Unique identifier for each record |
| age | int | Age in days (e.g., 19,468 days ≈ 53.3 years) |
| gender | int | 1 = Female, 2 = Male |
| height | int | Height in centimeters |
| weight | float | Weight in kilograms |
| ap\_hi | int | Systolic blood pressure |
| ap\_lo | int | Diastolic blood pressure |
| cholesterol | int | 1 = Normal, 2 = Above Normal, 3 = Well Above Normal |
| gluc | int | 1 = Normal, 2 = Above Normal, 3 = Well Above Normal |
| smoke | int | 0 = Non-smoker, 1 = Smoker |
| alco | int | 0 = Non-drinker, 1 = Alcohol consumer |
| active | int | 0 = Physically inactive, 1 = Physically active |
| cardio | int | Target variable: 0 = No disease, 1 = Has cardiovascular disease |

## 2.3 Notable Characteristics

- Age is stored in days and should be converted to years for interpretability.  
- Gender is binary encoded: 1 = Female, 2 = Male.  
- Blood pressure values (ap\_hi, ap\_lo) may include outliers or entry errors and should be cleaned.  
- Cholesterol and gluc levels are ordinal (1 to 3).  
- The dataset is balanced, with no missing values detected.