Problem Statement

A Healthcare organization maintains separate health risk datasets for male and female patients, containing demographic and behavioural information such as age, weight, height, smoking, alcohol, marital status, and exercise habits. However, the datasets suffer from inconsistencies like missing values, duplicate records, and non-uniform categorical entries (e.g., inconsistent gender or smoking labels).

To ensure accurate analysis and modeling, it is necessary to **clean, merge, and standardize** these datasets. This involves removing duplicates, handling missing values, aligning inconsistent categories, and deriving new calculated metrics like Corrected Body Mass Index (BMI).

The goal is to produce a **high-quality, unified dataset** that accurately reflects patient health patterns and is ready for advanced predictive analytics.