Feature Engineering Summary

This document outlines the newly engineered features in the dataset, explaining how they were created, what they represent, and why they are important for subsequent analysis and modeling.

1. age years

- **How it was created:**
  The raw “age” was provided in days. To make it more interpretable, the age in days was converted into years by dividing by 365 and rounding to the nearest whole number.
- **Why it helps:**
  Expressing age in years is far more intuitive than dealing with days (e.g., "45 years" vs. "16,425 days"). This transformation simplifies the interpretation of data, plots, model coefficients, and clinical insights.

2. Pulse pressure

- **How it was created:**
  Pulse pressure was computed as the difference between systolic blood pressure (**ap_hi**) and diastolic blood pressure (**ap_lo**):
- **Why it helps:**
  Pulse pressure is a key indicator of arterial strain. Elevated pulse pressure is associated with a higher risk of cardiovascular events. This feature provides valuable insight into the cardiovascular health of patients.

3. BMI (Body Mass Index) & is_obese flag

- **How BMI was computed:**
  Body Mass Index (BMI) is a well-known metric that quantifies an individual’s body weight relative to height:
- **How is_obese was derived:**
  A binary indicator was created based on the BMI value. If the BMI is greater than or equal to 30, the individual is considered obese:
- **Why it helps:**
  BMI is widely used in clinical practice to assess obesity, which is a significant risk factor for various health issues. The **is_obese** flag simplifies the categorization of patients, providing a quick indicator of obesity status.

4. bmi_category

- **How it was created:**
  The continuous BMI values were binned into four distinct categories using the following thresholds:
    - **Underweight:** BMI < 18.5
    - **Normal:** 18.5 ≤ BMI < 25
    - **Overweight:** 25 ≤ BMI < 30
    - **Obese:** BMI ≥ 30

This creates the **bmi_category** feature, with four possible labels: "Underweight," "Normal," "Overweight," and "Obese."

- **Why it helps:**
  This categorization simplifies risk assessment and allows for better interpretation of patterns across different BMI groups. It can help reveal non-linear relationships between BMI and health outcomes that might be missed when using raw BMI values.

5. lifestyle_score

- **How it was created:**
  A composite score was generated by summing three binary indicators of unhealthy behaviors:
    - Smoking (**smoke** = 1 if smoking, else 0)
    - Alcohol intake (**alcohol** = 1 if consuming alcohol, else 0)
    - Physical inactivity (**active** = 1 if physically active, else 0, and inactive if **1 - active**)
- **Why it helps:**
  The **lifestyle_score** provides a single, easy-to-understand metric for evaluating the accumulation of risky behaviors. A higher score indicates a greater number of unhealthy habits (ranging from 0 = healthiest to 3 = most unhealthy), streamlining multivariate risk assessment.

Why These Features Matter

By transforming raw variables into meaningful composites, such as pulse pressure and BMI categories, and adding related behaviors into a single lifestyle index, we improve both **interpretability** and **predictive power**. These engineered features are likely to offer stronger insights than their raw counterparts, making the model more interpretable and actionable in a healthcare context. Additionally, they allow clearer storytelling, helping stakeholders understand the relationships between lifestyle factors, body measurements, and cardiovascular health risks.