Proposal for Data science - project 1

Name: Salma Alarfaj

Purpose

To get exposed to a very large data set that has more than 10 features and more than 10000 records and apply data cleaning, Machine learning Algorithms, and statistical processes to better study the data and practice data analysis and science.

Picked Dataset

Cardiovascular Disease dataset
The dataset consists of 70 000 records of patient's data, 11 features + target.

Target Variable

Presence or absence of cardiovascular disease (binary).

Features

1. Age | age | int (days)
2. Height | height | int (cm)
3. Weight | weight | float (kg)
4. Gender | gender | categorical code
5. Systolic blood pressure | ap_hi | int
6. Diastolic blood pressure | ap_lo | int
7. Cholesterol | cholesterol | 1: normal, 2: above normal, 3: well above normal
8. Glucose | gluc | 1: normal, 2: above normal, 3: well above normal |
9. Smoking | smoke | binary |
10. Alcohol intake | alco | binary |
11. Physical activity | active | binary |

Analysis

The most common three factors that usually accompany **cardiovascular diseases.**
Run statistical description on the data.
The age-groups commonly carry **cardiovascular diseases.**
More analysis will be carried out after exploring the data closely.