# Cardiovascular disease Prediction

Shahd Omar Aljabarti

## ABSTRACT

The cardiovascular diseases lead the cause of death across the world as the statement of World Health Organization. In 2019, 32% of the global death were due to the heart attacks and strokes. [1] This proposed project intended to predict the heart disease presence based on combination of several features and that could help in prevent life threats.

## PROBLEM STATEMENT

The cardiovascular disease prediction is an interesting problem in the data science field. This project aims to explore the different features that might increase or affect the chance of cardiovascular disease occurrence and then use the extracted knowledge to determine whether a person is at risk of cardiovascular disease.

## DATA DESCRIPTION

Source of the dataset is Kaggle website where the dataset was extracted in excel format then converted to csv format. It has 70,000 records and 12 features. For the features, it has an age column that measured in days, gender which has two values 1 for female and 2 for male, height in cm and weight in kg, systolic blood pressure, diastolic blood pressure, cholesterol levels (1: normal, 2: above normal, 3: well above normal), glucose levels (1: normal, 2: above normal, 3: well above normal), alcohol intake, physical activity which has binary values 1 for active and 0 for inactive, smoking , and the last feature is the cardio disease presence which has 1 for heart disease presence and 0 for no presence.

## DATA CLEANING

By using python packages (numpy, pandas, matplitlib) I started with renaming some columns to understandable names. I filled the null values of the height, systolic blood pressure and diastolic blood pressure columns with their mean values. I added new column named ‘age_in_years’ and filled it with the age in years by applying a function to the age in days column.

Pandas.corr() function was used to discover the correlated features. The dataset has 8 positive correlations.

1. high correlation: cholesterol & glucose.
2. moderate correlations: alcohol intake & smoking, gender & smoking.
3. low correlations: age & cardio disease presence, age & cholesterol, weight & cardio disease presence, weight & glucose, cholesterol & cardio disease presence.

I plotted three graphs in order to capture insights. First one is box-and-whisker plot for the age in years column, it showed that it has two outliers (see figure 1).

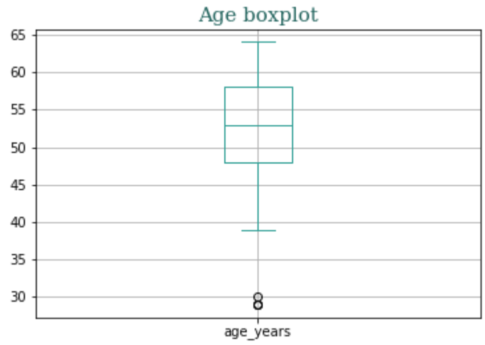

Figure 1: age boxplot

The second graph is box-and-whisker plot for age grouped by cholesterol (see figure 2) and it showed interesting outliers for some cholesterol's levels.

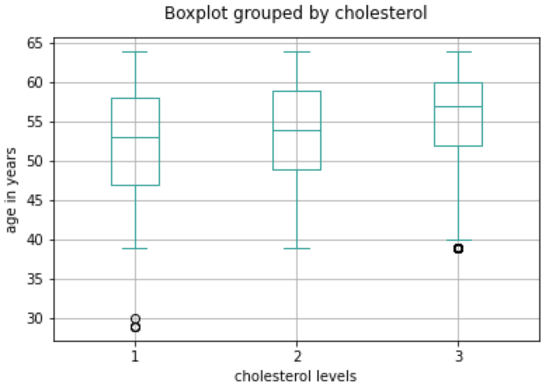

Figure 2: age_years x cholesterol

I also plotted a histogram for age distribution and it showed that the age of 55 is most occurred age in the dataset.

I also created new column named "Blood_pressure" which contains the levels of the blood pressure reading (systolic blood pressures/ diastolic blood pressure) whether it's normal, at risk or high blood pressure based on the Center of Disease Control and Prevention categorization [2].

Reference:

[1] 'Cardiovascular diseases (CVDs)'. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds) (accessed Oct. 03, 2021).

[2] CDC, 'High Blood Pressure Symptoms, Causes, and Problems | cdc.gov', *Centers for Disease Control and Prevention*, May 18, 2021. https://www.cdc.gov/bloodpressure/about.htm (accessed Oct. 04, 2021).