L4. Advanced Topics

Please work with the **framingham** dataset [1,2] (note, it is CSV => `sep=",`)

World Health Organization has estimated 12 million deaths occur worldwide, every year due to Heart diseases. Half the deaths in the United States and other developed countries are due to cardio vascular diseases. The early prognosis of cardiovascular diseases can aid in making decisions on lifestyle changes in high risk patients and in turn reduce the complications. This research intends to pinpoint the most relevant/risk factors of heart disease as well as predict the overall risk using logistic regression.

The dataset includes 4240 records, 16 columns. The goal of the dataset is to predict whether the patient has 10-year risk of future coronary heart disease (TenYearCHD column). Each attribute is a potential risk factor. There are demographic, behavioral and medical risk factors.

Demographic:

- male: 0-female, 1-male
- age: age of the patient

Behavioural:

- currentSmoker: whether (1) or not (0) the patient is a current smoker
- cigsPerDay: the number of cigarettes that the person smoked on average in one day

Medical (history):

- BPMeds: whether or not the patient was on blood pressure medication
- prevalentStroke: whether or not the patient had previously had a stroke
- prevalentHyp: whether or not the patient was hypertensive
- diabetes: whether or not the patient had diabetes

Medical (current):

- totChol: total cholesterol level
- sysBP: systolic blood pressure
- diaBP: diastolic blood pressure
- BMI: Body Mass Index
- heartRate: heart rate
- glucose: glucose level

Predict variable (desired target):

- TenYearCHD: 10 year risk of coronary heart disease CHD ("Yes"= 1, "No"= 0)

Please predict TenYearCHD. As it is binary, so use logistic regression. Check which variables influence the survival and choose the optimal set of predictors. For simplicity, you can exclude patients with *NA* observations.

[1] https://www.kaggle.com/datasets/aasheesh200/framingham-heart-study-dataset

[2] https://colab.research.google.com/github/dphi-official/Data_Science_Bootcamp/blob/master/Week3/Logistic_Regression/Logistic_Regression_Heart_Disease.ipynb#scrollTo=z7rSOAkerwxn