**About the dataset:**

The dataset is publically available on the Kaggle website, and it is from an ongoing ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. The classification goal is to predict whether the patient has 10-year risk of future coronary heart disease (CHD).The dataset provides the patients’ information. It includes over 4,240 records and 15 attributes.

**Attributes:**

1. sex: male(0) or female(1);(Nominal)
2. age: age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)
3. currentSmoker: whether or not the patient is a current smoker (Nominal)
4. cigsPerDay: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarretts, even half a cigarette.)
5. BPMeds: whether or not the patient was on blood pressure medication (Nominal)
6. prevalentStroke: whether or not the patient had previously had a stroke (Nominal)
7. prevalentHyp: whether or not the patient was hypertensive (Nominal)
8. diabetes: whether or not the patient had diabetes (Nominal)
9. totChol: total cholesterol level (Continuous)
10. sysBP: systolic blood pressure (Continuous)
11. diaBP: diastolic blood pressure (Continuous)
12. BMI: Body Mass Index (Continuous)
13. heartRate: heart rate (Continuous - In medical research, variables such as heart rate though in fact discrete, yet are considered continuous because of large number of possible values.)
14. glucose: glucose level (Continuous)
15. 10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”) - Target Variable

**Objective: Build a classification model that predicts heart disease in a subject. (note the target column to predict is 'TenYearCHD' where CHD = Coronary heart disease)**

Please do the following steps:

1. Read the file and display columns.
2. Handle missing values, Outliers and Duplicate Data
3. Calculate basic statistics of the data (count, mean, std, etc) and exploratory analysts and describe your observations.
4. Select columns that will be probably important to predict heart disease.
5. If you remove columns explain why you removed those.
6. Create training and testing sets (use 60% of the data for the training and reminder for testing).
7. Build a machine learning model to predict TenYearCHD
8. Evaluate the model (f1 score, Acuuracy, Precision ,Recall and Confusion Matrix)
9. Conclude your findings (Model which is giving best f1 score and why)

**Note :**This coding exercise should be performed in python (which is the programming language used by the team). You are free to use the internet and any other libraries. Please save your work in a Jupyter notebook and email