1. The Dataset and the Source

The dataset is publicly available on the Kaggle website, and it is from an 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 4,000 records and 15 attributes.

Source: <https://www.kaggle.com/dileep070/heart-disease-prediction-using-logistic-regression> or <https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset>

2. Description of Dataset

**Variables of dataset:**   
Each **input** attribute is a potential risk factor. There are both demographic, behavioral and medical risk factors.

**Demographic:**  
• Sex: male or female(Nominal)  
• Age: Age of the patient;(Continuous - Although the recorded ages have been truncated to whole numbers, the concept of age is continuous)

• Education: 1 = Some High School; 2 = High School or GED; 3 = Some College or Vocational School; 4 = college(Nominal)  
**Behavioral:**  
• Current Smoker: whether or not the patient is a current smoker (Nominal)  
• Cigs Per Day: the number of cigarettes that the person smoked on average in one day.(can be considered continuous as one can have any number of cigarettes, even half a cigarette.)  
**Medical( history):**  
• BP Meds: whether or not the patient was on blood pressure medication (Nominal)  
• Prevalent Stroke: whether or not the patient had previously had a stroke (Nominal)  
• Prevalent Hyp: whether or not the patient was hypertensive (Nominal)  
• Diabetes: whether or not the patient had diabetes (Nominal)  
**Medical(current):**  
• Tot Chol: total cholesterol level (Continuous)  
• Sys BP: systolic blood pressure (Continuous)  
• Dia BP: diastolic blood pressure (Continuous)  
• BMI: Body Mass Index (Continuous)  
• Heart Rate: 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.)  
• Glucose: glucose level (Continuous)  
**Predict variable (desired target):**• 10 year risk of coronary heart disease CHD (binary: “1”, means “Yes”, “0” means “No”)

3. Question

Binary Classification for Heart Disease Prediction:

In this assignment, I will test logistic regression algorithm and make prediction (whether the patient has 10-year risk of future CHD).

4. Data Processing

After using *df.isnull().sum()*, I found columns with mising values are:

'education'(2.5%), 'cigsPerDay'(0.7%), 'BPMeds'(1.3%), 'totChol'(1.2%) .

Except the feature glucose all other missing values are less than 2% of data. So I drop all other missing values.

It is divided into:

a). 70% of the entire Dataset for training (Training data)

b). 15% of the entire Dataset for validation (Validation data)

Logistic regression is used to build a model. BinaryClassificationEvaluator is used to test the result of classification.

For the training dataset, the accuracy score is 0.7073265608706987.

For the test dataset, the accuracy score is 0.6722228710358958