Consider the `framingham.csv` data file. The dataset is 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. Each 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)
- 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”)

In **Python**, answer the following:

1. (4 points) Using the `pandas` library, read the csv data file and create a data-frame called `heart`. Remove the missing values.
2. (22 points) Using `age`, `currentSmoker`, `totChol`, `BMI`, and `heartRate` as the predictor variables, and `TenYearCHD` is the target variable, split the data into two data-frames (taking into account the proportion of 0s and 1s), conduct the following:
   - Over 10-folds, train a `LogisticRegression` model and use the area under the ROC-AUC curve as a measure of performance. Make sure to use the `cross_val_score` function.

- Over 10-folds, train a `LogisticRegression` model and use the area under the ROC-AUC curve as a measure of performance. Make sure to use the `cross_val_score` function and put the input features on the same scale using `MinMaxScaler`.
- Over 10-folds, train a `LogisticRegression` model and use the area under the ROC-AUC curve as a measure of performance. Make sure to use the `cross_val_score` function and put the input features on the same scale using `StandardScaler`.
- Over 10-folds, train a `LogisticRegression` model and use the area under the ROC-AUC curve as a measure of performance. Make sure to use the `cross_val_score` function and put the input features on the same scale using `RobustScaler`.
- Over 10-folds, train a `LinearDiscriminantAnalysis` model and use the area under the ROC-AUC curve as a measure of performance. Make sure to use the `cross_val_score` function.
- Over 10-folds, train a `QuadraticDiscriminantAnalysis` model and use the area under the ROC-AUC curve as a measure of performance. Make sure to use the `cross_val_score` function.

3. (4 points) Using the results from part (2), what model would you use to predict `TenYearCHD` based on the area under the ROC-AUC curve?

The End.