DATA-445 IN-CLASS ASSIGNMENT NOVEMBER 3, 2021

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. (3 points) Using the pandas library, read the csv data file and create a data-frame called `heart`.
2. (3 points) Remove observations with missing values.
3. (4 points) Based on the different interactions that we have so far with this data file, it seems that `age`, `totChol`, `sysBP`,`BMI`, `heartRate`, and `glucose` 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): train (80%) and test (20%). Make sure to standardize the input data to 0-1 scale.

4. (7 points) Using the train data-frame, build AdaBoost model (with 500 trees, max tree depth equal 3, and learning rate equal to 0.01) in which: `age`, `totChol`, `sysBP`, `BMI`, `heartRate`, and `glucose` are the predictor variables, and `TenYearCHD` is the target variable. Using this model, predict the risk of coronary disease of the patients in the test data-frame. Using 15% as cutoff value, report the accuracy and recall.

5. (7 points) Using the train data-frame, build an gradient boosting model using `GradientBoostingClassifier` (with 500 trees, max tree depth equal 3, and learning rate equal to 0.01) in which: `age`, `totChol`, `sysBP`, `BMI`, `heartRate`, and `glucose` are the predictor variables, and `TenYearCHD` is the target variable. Using this model, predict the risk of coronary disease of the patients in the test data-frame. Using 15% as cutoff value, report the accuracy and recall.

6. (7 points) Using the train data-frame, build a support vector machine model (with `kernel = 'rbf'`) in which: `age`, `totChol`, `sysBP`, `BMI`, `heartRate`, and `glucose` are the predictor variables, and `TenYearCHD` is the target variable. Using this model, predict the risk of coronary disease of the patients in the test data-frame. Using 15% as cutoff value, report the accuracy and recall.

7. (7 points) Using the estimated likelihoods from part (4), (5) and (6), combine those likelihoods using a random forest model with 500 trees and maximum depth equal to 3. Using 15% as cutoff value, report the accuracy and recall.

The End.