DATA-445 In-Class Assignment October 23, 2023

Consider the `framingham.csv` data file. The dataset is publically 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 observations with missing values.
2. (3 points) Split the data into `train` (80%) and `test` (20%) (taking into account the proportions of 0s and 1s).
3. (10 points) Build a `RandomForestClassifier` as follow:
   (i) Using the `train` (from part 2), split the data in `train` (80%) and `test` (20%) (taking into account the proportions of 0s and 1s).

(ii) Build a `RandomForestClassifier` model with `n_estimators = 500` and `max_depth = 5`.

(iii) Store the feature importances.

Repeat (i)-(iii) 100 times. Compute the average importance score of each of the features. Using the top five features; that is, the five features with the highest average importance scores, build a `RandomForestClassifier` with `n_estimators = 500` and `max_depth = 5` on the `train` data-frame from part (2). Using this model, predict the risk of coronary disease of the patients in the `test` data-frame from part (2). Using 10% as cutoff value, report the recall.

4. (10 points) Build a `ExtraTreesClassifier` as follow:

   (i) Using the `train` (from part 2), split the data in `train` (80%) and `test` (20%) (taking into account the proportions of 0s and 1s).

   (ii) Build a `ExtraTreesClassifier` model with `n_estimators = 500` and `max_depth = 5`.

   (iii) Store the feature importances.

   Repeat (i)-(iii) 100 times. Compute the average importance score of each of the features. Using the top five features; that is, the five features with the highest average importance scores, build a `ExtraTreesClassifier` with `n_estimators = 500` and `max_depth = 5` on the `train` data-frame from part (2). Using this model, predict the risk of coronary disease of the patients in the `test` data-frame from part (2). Using 10% as cutoff value, report the recall.

5. (3 points) Using the results from parts (3) and (4), what model would you use to predict `TenYearCHD` based on the recall?

The End.