HSBC DNA Deep Dive Batch 3 – Hands-on Assessment

Duration: 1 hour Max marks: 20

General Instructions:

1. Read the problem statement carefully before answering. In case of any ambiguity please contact the invigilator
2. Write the solution in a python notebook named **DNADD_B3_<<your Infosys id>>**; for example DNADD_B3_HSBC2019_Alex and upload the notebook at http://192.168.1.62/uploadfile/ upon completion of the exam.
3. Before you submit the notebook, please ensure that you have saved the last modified code. Please use **File → Save and Checkpoint** option to do this.
4. Access to your previous notebooks, instructor notes, and internet is permitted. This is an open book exam.
5. Discussion among peers and any form of malpractice is prohibited.

---- All the best ----

Problem Statement(s)

You are given 2 datasets – abalone.csv and adult.csv. The abalone.csv file contains several columns that describe features of abalones (marine snails) such as gender, length, diameter, etc. The adult.csv file contains features of various individuals such as age, employment status, marital status, etc.

Task 1: In the abalone data, the task is to predict the value of the **Rings** variable using other features.

Task 2: In the adult data, the task is to predict the value of **Income Group** variable using other features.

1. Determine the type of supervised machine learning algorithm that should be used for task1 and task2 respectively **[1M]**
2. Choose any one of the above datasets to work with and import only that dataset into Python as a pandas data frame. **[1M].**
3. For the chosen dataset, do the following:

a. Determine the number of missing values in each column. The missing values are indicated using the character **'?' [1M].**
   Note: You are not required to take any further action. Ensure that you retain the missing values as '?' only.
b. Determine the data type of each column in the data and list out all columns that you think are categorical **[1M]**
c. One hot encode the categorical columns such that k-1 dummy variables are created for k categories **[1M]**
d. Scale all numerical columns such that after scaling the minimum value in the column is 0 and the maximum is 1. **[1M]**
e. Comment on whether the data is adequately balanced **[1M]**
    i. For the abalone data check only the 'Sex' column.
    ii. For the adult data check only the 'income_group' column
f. List out any 2 algorithms (or methods in sklearn) that can be used for making predictions on the chosen dataset. **[1M]**
g. Split the data into train and test such that 20% of the total data is considered as test data **[1M]**
h. Build a predictive machine learning model on the train data to predict the outcome (rings for abalone data, income_group for adult data) using an appropriate algorithm **[2M]**
i. Evaluate the model's performance on the train data using an appropriate metric. Note: choose among RMSE or Accuracy. **[2M]**
j. Evaluate the model's performance on the test data using an appropriate metric. Note: choose among RMSE or Accuracy. **[2M]**
k. Comment on the usefulness of the model based on its performance in train and test data **[2M]**
l. Finally build a model using an ensemble learner (either RandomForest or GradientBoosting) for your chosen data. Display the model's score (RMSE or Accuracy) on both train and test data. **[3M]**