IDS-FA22-Assignment **Due Date: 16-12-2022**

Submission: Please upload the PDF report and Python code (preferably python notebook) to GitHub.

Download the gender prediction dataset from the following link:
https://drive.google.com/file/d/1EKpArZit1OdkfhaKVC3Beku6tkmASu3M/view?usp=share_link

Q1: Provide responses to the following questions about the dataset.

1. How many instances does the dataset contain?
   There are 18 instances in the dataset.

2. How many input attributes does the dataset contain?
   There are 7 input instances in the dataset.

3. How many possible values does the output attribute have?
   There are 2 possible output instances in the dataset.
4. How many input attributes are categorical?
   There are 4 categorical input instances in the dataset.

5. What is the class ratio (male vs female) in the dataset?
   Ratio of Male: 0.575
   Ratio of Female: 0.425
   46 MALES
   34 FEMALES

Q2: Apply Random Forest, Support Vector Machines, and Multilayer Perceptron classification algorithms (using Python) on the gender prediction dataset with standard train/test split ratio and answer the following questions.

1. How many instances are incorrectly classified?

   MLP: 12 instances are incorrectly classified.
   Random Forest: 1 instance are incorrectly classified.
   SVM: 0 instances are incorrectly classified.

2. Rerun the experiment using train/test split ratio of 80/20. Do you see any change in the results? Explain.

   MLP: 7 instances are incorrectly classified.
   Random Forest: 0 instances are incorrectly classified.
   SVM: 1 instance are incorrectly classified.

3. Name 2 attributes that you believe are the most "powerful" in the prediction task. Explain why?

BEARD and SACRF are the most powerful attributes in given the dataset. As Scarf represent the female and beard represent the female.

4. Try to exclude these 2 attribute(s) from the dataset. Rerun the experiment (using 80/20 train/test split), did you find any change in the results? Explain.

BY DROPING BEARD AND SCARF

MLP:

15 instances are correctly classified.
2 instances are incorrectly classified.

Random Forest:

10 instances are correctly classified.
3 instances are incorrectly classified.

SVM:

13 instances are correctly classified.
3 instances are incorrectly classified.

Q3: Apply Decision Tree Classifier classification algorithm (using Python) on the gender prediction dataset with Monte Carlo cross-validation and Leave P-Out cross-validation. Report F_1 score for both cross-validation strategies.
Note: You are free to choose any parameter values for both cross-validation strategies, however, you have to provide these values in your submission document.

Q4: Add 5 sample instances into the dataset (you can ask your friends/relatives/sibling for the data). Rerun the ML experiment (using Python) by training the model using Gaussian Naïve Bayes classification algorithm and all the instances from the gender prediction dataset. Evaluate the trained model using the newly added test instances. Report accuracy, precision, and recall scores.
Note: You have to add the test instances in your assignment submission document.