COMSATS University Islamabad, Lahore Campus

Course Title:	Introduction to Data Science				Course Code:	CSC461	Credit Hours:	3(3,0)
Resource Person:	Dr. Muhammad Sharjeel				Programme Name:	BSSE		
Semester:	5th	Batch:	FA21	Section:	C	Max Marks:	10	

Assignment 3 Due Date: 25-11-2023

Submission: Upload the assignment solution (PDF report and Python code, preferably iPython notebook) to your GitHub account (private repository).

Important instructions: Please write the following information at the start of your ipython file.
# Date
# CSC461 – Assignment3 – Machine Learning
# Your Full Name
# You Complete Registration Number
# A brief description of the task

Use the dataset file "gender-prediction.csv" available on shared Google Drive folder for this assignment.

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

1. How many instances does the dataset contain?
2. How many input attributes does the dataset contain?
3. How many possible values does the output attribute have?
4. How many input attributes are categorical?
5. What is the class ratio (male vs female) in the dataset?

Q2: Apply Logistic Regression, Support Vector Machines, and Multilayer Perceptron classification algorithms (using Python) on the gender prediction dataset with 2/3 train and 1/3 test split ratio and answer the following questions.

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

Q3: Apply Random Forest classification algorithm (using Python) on the gender prediction dataset with Monte Carlo cross-validation and Leave P-Out cross-validation. Report F_1 scores 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 10 sample instances into the dataset (you can ask your friends/relatives/sibling for the data). Run 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 10 test instances. Report accuracy, precision, and recall scores.
Note: You must use all the instances in the gender precision dataset for training and only 10 new instances for testing. You must include all the 10 test instances in your assignment submission document.