COMSATS University Islamabad (Lahore Campus)

Assignment <4> FALL 2022

Course Title:	Introduction to Data Science			Course Code:		CSC461	Credit Hours: 3
Course Instructor:	Muhammad Sharjeel			Program Name:		BS Computer Science	
Semester:	6th	Batch:	SP20	Section:	C	Date:	11-12-2022
Due Date:	**16-12-2022**			**Maximum Marks:**			**10**
Student Name:	**Ahmad Mujtaba**			**Registration No.**			**SP20-BCS-072**

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

1. How many instances does the dataset contain?

→ ***80 Instances***

height 80 weight 80 beard 80 hair_length 80 shoe_size 80 scarf 80 eye_color 80 gender 80
dtype: int64

2. How many input attributes does the dataset contain?

→ ***7 Attributes***

Index(['height', 'weight', 'beard', 'hair_length', 'shoe_size', 'scarf', 'eye_color'], dtype='object')

3. How many possible values does the output attribute have?

→ ***2 Possible values***

array(['male', 'female'], dtype=object)

4. How many input attributes are categorical?

→ ***4 Attributes***

beard: ['yes' 'no'] hair_length: ['short' 'bald' 'medium' 'long'] scarf: ['no' 'yes'] eye_color:
['black' 'blue' 'gray' 'brown' 'green']

5. What is the class ratio (male vs female) in the dataset?

gender	values
female	34.0
male	46.0
FemaleRatio	42.50
MaleRatio	57.50

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?**

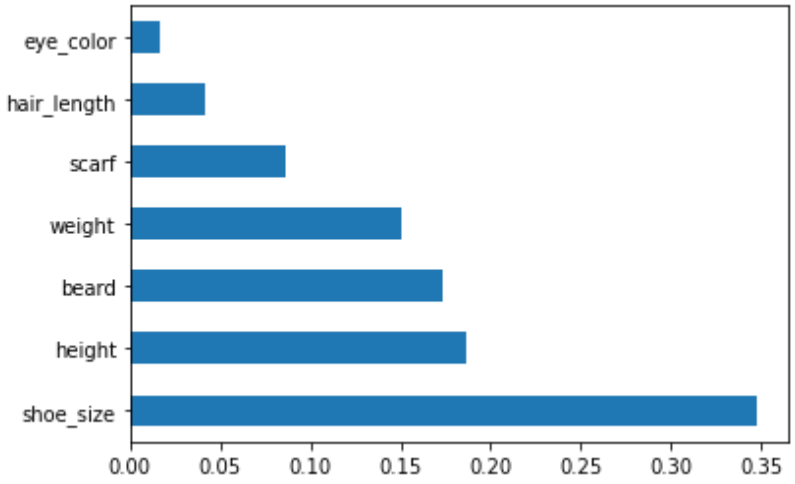
	Random Forest	Support Vector Machines	Multilayer Perceptron
Incorrect Instances	0	0	4

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

	Random Forest	Support Vector Machines	Multilayer Perceptron
Accuracy	100	100	93.75
Incorrect Instances	0	0	1

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

→ height and shoe_size are the most powerful attributes



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.

	Random Forest	Support Vector Machines	Multilayer Perceptron
Accuracy	94.44	77.77	77.77
Incorrect Instances	1	4	4

Q3: Apply Decision Tree Classifier classification algorithm (using Python) on the gender prediction dataset with Monte Carlo cross-validation and Leave P-Out crossvalidation. Report F1 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.

	Monte Carlo cross-validation	Leave P-Out crossvalidation
Time Taken	0.1925s	18.2464s
F1-Score	0.8701	0.7204

Parameters	test_size=0.33, n_splits=10	p=2

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.

EVALUATION	Gaussian Naïve Bayes
Accuracy	0.9655172413793104
Precision	0.9375
Recall	1.0
F1	0.967741935483871

Test Instances:

height	weight	beard	hair_length	shoe_size	scarf	eye_color	gender
70	110	yes	medium	43	no	brown	male
63	109	no	long	40	yes	black	female
68	115	no	short	43	no	black	male
60	94	no	medium	39	no	black	female
69	170	no	short	42	no	brown	male