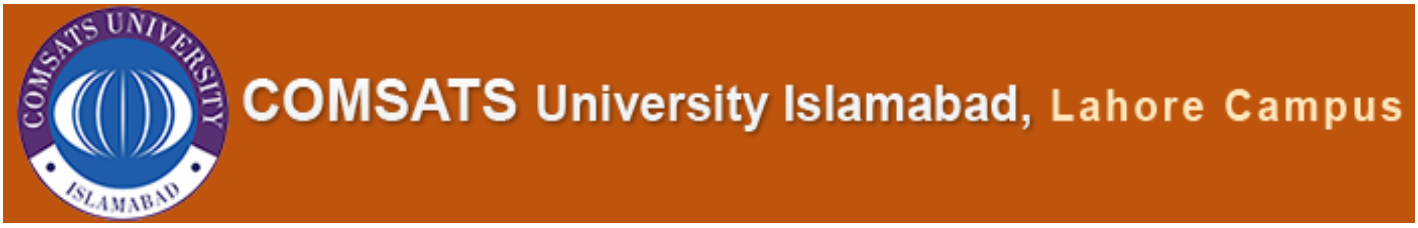

Assignment – 4

(Report)

Submitted By:

Muhammad Aun Raza

SP20-BCS-092

Group-2

Semester 6th

Course Title:

Introduction to Data Science
(CSC-461)

Submitted To:

Sir Muhammad Sharjeel

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

- **How many instances does the dataset contain?**
  There are 80 instances.
- **How many input attributes does the dataset contain?**
  There are 7 input attributes.
- **How many possible values does the output attribute have?**
  There are 2 possible values for the output attribute that are Male or Female.
- **How many input attributes are categorical?**
  4 input attributes are categorical.
- **What is the class ratio (male vs female) in the dataset?**
  There are 46 males and 34 Females in the gender attribute. Ratio of Male: Female would be 23:17.

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.

- **How many instances are incorrectly classified?**

  With random state=42

  In RF >> TP=12, TN=16, FP=0, FN=0

  In SVM >> TP=9, TN=11, FP=3, FN=4

  In MLP >> TP=10, TN=13, FP=2, FN=2

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

  Yes, there is a visible change in the values of confusion matrix that are TP, TN, FP, FN when we change the train/test split ratio to 80/20, having the random state=42.
  The updated values are

  In RF >> TP=7, TN=9, FP=0, FN=0

  In SVM >> TP=5, TN=5, FP=2, FN=4

  In MLP >> TP=5, TN=8, FP=2, FN=1

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

  The attributes **Beard** and **Scarf** are the most powerful in terms of prediction of the gender because, Beard is such a characteristic that only Male gender possesses.
  On the other hand, Scarf is such a characteristic that mostly Female gender possesses.

- **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.**

  After excluding those 2 attributes and after re-running the experiment using same 80/20 train/test split no change was observed in **RF**, **SVM**. and **MLP** classifier. The values are,

  In RF >> TP=7, TN=9, FP=0, FN=0

  In SVM >> TP=5, TN=5, FP=2, FN=4

  In MLP >> TP=5, TN=8, FP=2, FN=1

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 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.

For Monte-Carlo (Shuffle Split) Cross Validation:

test_size=0.33, n_splits=10

F1- Score: 1.00

For Leave P-Out Cross Validation:

test_size=0.33,

F1-Score: 0.93

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.

Newly added 5 instances are

Height	Weight	Beard	Hair Length	Shoe Size	Scarf	Eye Colour	Gender
75	162	Yes	Medium	42	No	Black	Male
85	176	Yes	Short	37	No	Blue	Male
120	175	Yes	Short	36	No	Gray	Male
57	143	No	Long	38	No	Black	No
50	159	No	Long	43	Yes	Black	No

After adding newly added 5 instances and training the model on Gaussian Naïve Bayes classification algorithm and selecting the above-mentioned instances using Loc for the testing purpose, the required values are listed below. After selecting the above-mentioned instances using Loc for the testing purpose,

Accuracy: 100%
Precision: 1.00
Recall: 1.00
