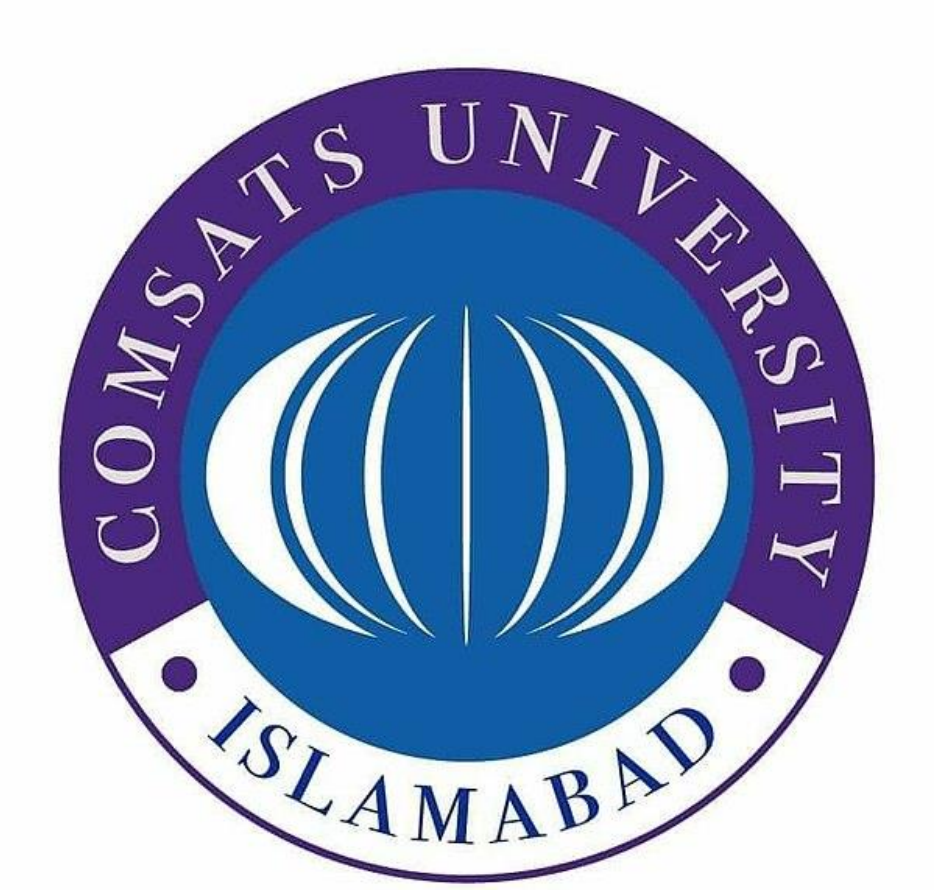

INTRODUCTION TO DATA SCIENCE

-ASSIGNMENT # 4-

Registration Number: SP20-BCS-095

Student Name: Rohaan Farooq

Course: Introduction to Data Science (CSC461)

Group: BCS 6 - G2-A

Assignment #: 4

Submitted to: Prof. Muhammad Sharjeel

Date: 16th December 2022

Question 1

1. How many instances does the dataset contain?
80 instances

2. How many input attributes does the dataset contain?
7 input attributes

3. How many possible values does the output attribute have?
2 possible values (male and female)

4. How many input attributes are categorical?
4 attributes are categorical (beard, hair_length, scarf, eye_color)

5. What is the class ratio (male vs female) in the dataset?
46 Males and 34 Females
Ratio (Male : Female) = 23 : 17

Question 2

1. How many instances are incorrectly classified?
Random Forest: **0** incorrect classifications
Support Vector Machine: **6** incorrect classifications
Multilayer Perceptron: **10** incorrect classifications

2. Rerun the experiment using train/test split ratio of 80/20. Do you see any change in the results? Explain.
After using train/test split ratio of 80/20, the Accuracy of the Support Vector Machine and Multilayer Perceptron have gone up. Accuracy of Random Forest stays the same (i.e., 100%). This is likely because the model has more instances to be trained on as compared to before. The model was hence trained better on 80/20 split and performed better.

3. Name 2 attributes that you believe are the most "powerful" in the prediction task. Explain why?
2 most powerful attributes are believed to be "beard" and "scarf" as these can easily distinguish between a male and a female. Only males have beard, and females wear a scarf so these 2 attributes are the most discriminating 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.
After rerunning the experiment using 80/20 after removing beard and scarf, the evaluation metrics (accuracy, precision, recall, f1) went up.

Question 3

Monte Carlo Cross Validation
Parameters: n_splits=4, test_size=0.33, random_state=1
F1 Score = 97.01%

Leave p-out Cross Validation
Parameter: LeavePOut(2) //Leave 2 out
F1 Score = 94.02%

Question 4

After adding 5 sample instances into the dataset, rerunning the ML experiment 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, the accuracy, precision, and recall scores are as follows:

Accuracy = 100%
Precision = 100%
Recall = 100%

The End