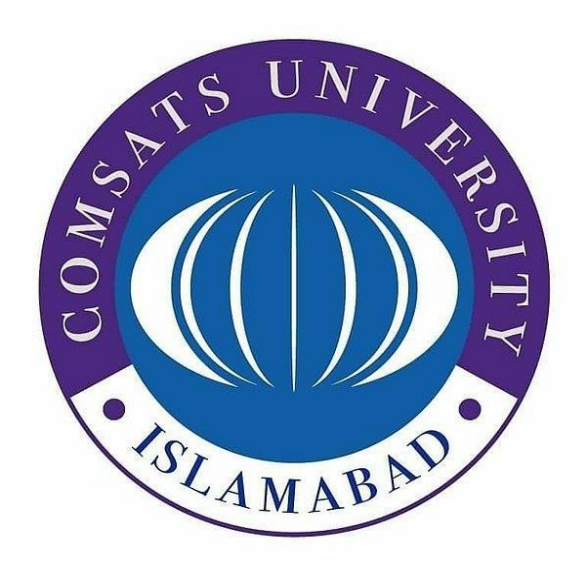

Introduction to Data Science

(CSC461)

Name: Ahsan Aijaz

Reg No: SP20-BCS-044

Section: BCS-B

Assignment no: 4

Section: SP20-BCS-044 (G-II)

Instructor: Dr. Muhammad Sharjeel

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

1. **How many instances does the dataset contain?**
   This dataset contains 80 instances.
2. **How many input attributes does the dataset contain?**
   This dataset contains 7 attributes.
3. **How many possible values does the output attribute have?**
   The output attribute has 2 possible values:
   - Male
   - Female
4. **How many input attributes are categorical?**
   4 input attributes are categorical:
   - Beard
   - Hair_length
   - Scarf
   - Eye_color
5. **What is the class ratio (male vs female) in the dataset?**
   Class ratio is 0.73.

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?**
   Out of 26, incorrectly classified test instances are:
   - Random Forest: 0 (100% accuracy)
   - Support Vector Machine: 6 (77.77% accuracy)
   - Multilayer Perceptron: 10 (62.96% accuracy)

2. **Rerun the experiment using train/test split ratio of 80/20. Do you see any change in the results? Explain.**
   After re-running the experiment using train/test split ratio of 80/20, the changes were as follows:
   - Random Forest: 100% accuracy
   - Support Vector Machine: 81.25% accuracy
   - Multilayer Perceptron: 62.5% accuracy

   As we can tell, the accuracies of Random Forest and Multilayer Perceptron almost stood same. However, the accuracy of support vector machine increased significantly.

3. **Name 2 attributes that you believe are the most "powerful" in the prediction task. Explain why?**
   Most important attributes are:
   - Scarf (since only women wear scarfs)

- Beard (since beard is only limited to men)

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 excluding scarf and beard, when we re-ran the experiment using 80-20 train-test split, the results were:
   - Random Forest: 100% accuracy
   - Support Vector Machine: 77.77% accuracy
   - Multilayer Perceptron: 85.19% accuracy

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.

Monte Carlo Cross Validation Parameters:

n_splits=5, test_size=0.33, random_state=7

Accuracy: 92.59

F1 Score: 95.07%

Leave p-out Cross Validation Parameter:

LeavePOut(2)

Accuracy: 94.17

F1 Score = 94.12%

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.

After performing experiment according to given conditions, the results were:

Accuracy: 95%

Precision: 96%

Recall: 94.11%