INTRODUCTION TO DATA-SCIENCE

Assignment:4

Registration number: SP20-BCS-077/B

Submitted to: Dr. Muhammad Sharjeel

Submission Date: 18/12/2022

(Group-2)

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

1. How many instances does the dataset contain?

Ans: Dataset heave 80 instances.

2. How many input attributes does the dataset contain?

Ans: Dataset contains7 input attributes:

- Height
- Weight
- Beard
- Hair Length
- Shoe size
- Scarf
- Eye Color

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

Ans: Output attribute: Gender

Possible Values which can be output = 2

One is Male and the other one is Female.

4. How many input attributes are categorical?

Ans: 4 attributes are categorical:

- Beard
- Hair Length
- Scarf
- Eye Color

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

Ans: Class ratio: 26: 13

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?

Model	Random Forest	Support Vector Machine	Multilayer Perceptron
Instance Incorrectly Classified	0	6	10

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

Ans: Random forest and Multilayer Perceptron accuracy stayed the same, where as the accuracy of SVM increased. The increase in accuracy could be because the model had more instances to train in 80/20 split than in 66/33 split. Therefore, the model was trained better with 80/20 split.

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

Ans: Beard and Scarf are 2 attributes that are believed to be most powerful in this task, because a female is very unlikely to have a beard, and it is also very unlikely for a male to wear a scarf.

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.

Ans: Accuracy of all three models decreased, as the attributes that were the most deterministic and since these are excluded, the accuracy is decreased.

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.
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.

Ans:

Monte Carlo Random Split: 66/33
Monte Carlo Number of Iterations: 5

```
monte_carlo_acc = cross_val_score(decision_tree_model,x_encoded,y_encoded,cv=monte_carlo).mean() * 100
monte_carlo_f1 = cross_val_score(decision_tree_model,x_encoded,y_encoded, scoring="f1", cv=monte_carlo).mean() * 100
print("Monte Carlo cross-validation accuracy", str(round(monte_carlo_acc)), "%")
print("Monte Carlo cross-validation F1 score", str(round(monte_carlo_f1)), "%")
```

```
Monte Carlo cross-validation accuracy 91 %
Monte Carlo cross-validation F1 score 95 %
```

Value for p in Leave P-Out: 5

```
leave_pout_acc = cross_val_score(decision_tree_model,x_encoded,y_encoded,cv=lpo).mean() *100
leave_pout_f1 = cross_val_score(decision_tree_model,x_encoded,y_encoded,cv=lpo, scoring="f1_weighted").mean() * 100
print("Leave P-Out cross-validation accuracy", str(round(leave_pout_acc)), "%")
print("Leave P-Out cross-validation F1 score", str(round(leave_pout_f1)), "%")
```

```
Leave P-Out cross-validation accuracy 95 %
Leave P-Out cross-validation F1 score 95 %
```

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.

Test Instances

Height	Weight	Beard	Hair Length	Shoe Size	Scarf	Eye Color	Gender	
70	160	no	medium	42	no	black	male	
60	130	no	long	36	no	brown	female	
70	170	no	short	41	no	black	male	
68	138	no	medium	39	yes	brown	female	
65	120	yes	medium	40	no	blue	male	