Report of Assignment-4

IDS-FA22-Assignment **Due Date: 16-12-2022**

Name: Azeem Manzoor **Group: 4**

Reg#: Sp20-BCS-046 **Section: B**

Submitted To.

Dr. Muhammad Sharjeel

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

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

Answer: Total 80 instances of the given dataset. (80 rows and 8 columns)

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

Answer: Total 7 input attributes contain the given dataset. (height, weight, beard, scarf, hair length, shoe size and eye color).

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

Answer: Output attribute only have 2 possible values Male and Female but in numeric form male represents 1 and female represents 0.

4. **How many input attributes are categorical?**

Answer: 4 input attributes are categorical (beard, hair length, scarf, eye color).

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

Answer: 46 males and 34 females.

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?**

Answer:

Multilayer Perceptron gives 12 incorrect instances with (33% test size).

```
[[ 0 12]
 [ 0 15]]
              precision    recall  f1-score   support

           0       0.00      0.00      0.00        12
           1       0.56      1.00      0.71        15

    accuracy                           0.56        27
   macro avg       0.28      0.50      0.36        27
weighted avg       0.31      0.56      0.40        27
```

Random Forest predict 1 instance incorrectly with (33% test size).

```
[[11  1]
 [ 0 15]]
              precision    recall  f1-score   support

           0       1.00      0.92      0.96        12
           1       0.94      1.00      0.97        15

    accuracy                           0.96        27
   macro avg       0.97      0.96      0.96        27
weighted avg       0.97      0.96      0.96        27
```

Support Vector Machine predicts all instances correctly with 33% test size.

```
[[12  0]
 [ 0 15]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00        12
           1       1.00      1.00      1.00        15

    accuracy                           1.00        27
   macro avg       1.00      1.00      1.00        27
weighted avg       1.00      1.00      1.00        27
```

---

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

Answer:

Multilayer Perceptron:

```
[[0 7]
 [0 9]]
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         7
           1       0.56      1.00      0.72         9

    accuracy                           0.56        16
   macro avg       0.28      0.50      0.36        16
weighted avg       0.32      0.56      0.40        16
```

7 instances predicts incorrectly out of 16 while with 67% 33% split 12 were incorrect out of 27 but there is no change in Accuracy with both splits which is 56% .

Random Forest:

```
[[7 0]
 [0 9]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         7
           1       1.00      1.00      1.00         9

    accuracy                           1.00        16
   macro avg       1.00      1.00      1.00        16
weighted avg       1.00      1.00      1.00        16
```

20%, 80% split give better results of accuracy, precision, recall and f1 score which is 100% as compared

to 33% test size.

Support vector Machine:

```
[[7 0]
 [0 9]]
              precision    recall  f1-score   support

           0       1.00      1.00      1.00         7
           1       1.00      1.00      1.00         9

    accuracy                           1.00        16
   macro avg       1.00      1.00      1.00        16
weighted avg       1.00      1.00      1.00        16
```

Same values of accuracy, recall, precision and f1 score gave with out any difference when splits were 33% and 67%.

Table contain values with 33% and 67% splits

Classifier	**Accuracy (33%,67%)**	**Accuracy (20%,80%)**
Random Forest	96%	100%
Support Vector Machine (SVM)	100%	100%
Multilayer Perceptron	56%	56%

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

Answer:

I believe that the attribute "weight" and "beard" are most powerful in this prediction task. Because all beard with 0 are females but few ones are beard 0 output is male and the attribute "weight" shows big difference in males and females.

These two attributes are make clear difference for male and females.

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.

Answer:

By excluding attributes "weight" and "beard" I again applied above classifiers again with 80/20 train/test split.

MLP:

```
                  [[0 7]
 [0 9]]
              precision    recall  f1-score   support

           0       0.00      0.00      0.00         7
           1       0.56      1.00      0.72         9

    accuracy                           0.56        16
   macro avg       0.28      0.50      0.36        16
weighted avg       0.32      0.56      0.40        16
```

ID3:

```
[[6 1]
 [0 9]]
              precision    recall  f1-score   support
```

```
              precision    recall  f1-score   support

           0       1.00      0.86      0.92         7
           1       0.90      1.00      0.95         9

    accuracy                           0.94        16
   macro avg       0.95      0.93      0.94        16
weighted avg       0.94      0.94      0.94        16
```

SVM:

```
[[6 1]
 [0 9]]
              precision    recall  f1-score   support

           0       1.00      0.86      0.92         7
           1       0.90      1.00      0.95         9

    accuracy                           0.94        16
   macro avg       0.95      0.93      0.94        16
weighted avg       0.94      0.94      0.94        16
```

Comparison Table with and without attributes "Weight" and "Beard"

Classifier	**Accuracy (excluded "Weight" and "Beard")**	**Accuracy (included "Weight" and "Beard")**
Random Forest	94%	100%
Support Vector Machine (SVM)	94%	100%
Multilayer Perceptron	56%	56%

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-validationstrategies.

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.

Answer:

Monte Carlo cross-validation:

Cv= 10

```
F1-score = 0.9565217391304348
```

Leave_Pout:

Cv = 5

```
F1-score = 0.9462365591397849
```

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 testinstances. Report accuracy, precision, and recall scores.

Note: You have to add the test instances in your assignment submission document.

Answer:

These 5 instances with values I added in dataset.

new_row=[{'height':67,'weight':126,'beard':1,'hair_length':3,'shoe_size':42,"scarf":0,'eye_color':1,'gender':1},#Hasnain

Ali

{'height':66,'weight':132,'beard':1,'hair_length':0,'shoe_size':42,"scarf":0,'eye_color':2,'gender':1},#Abdul_Baseer
{'height':68,'weight':140,'beard':0,'hair_length':1,'shoe_size':40,"scarf":0,'eye_color':1,'gender':1},#Sudais
{'height':65,'weight':200,'beard':0,'hair_length':3,'shoe_size':42,"scarf":0,'eye_color':1,'gender':1},#Huzaifa
{'height':68,'weight':130,'beard':1,'hair_length':2,'shoe_size':40,"scarf":0,'eye_color':3,'gender':1}]#Azeem

Then apply the Naïve Bayes classification Algorithm in this data-frame.

Here below is the classification report that describes accuracy, precision, recall and f1-score.

Note: Total support is 17 it means 85 instances gives 17 instances for testing with 20% testing and 80% training split.

```
[[ 6  1]
 [ 0 10]]
              precision    recall  f1-score   support

           0       1.00      0.86      0.92         7
           1       0.91      1.00      0.95        10

    accuracy                           0.94        17
   macro avg       0.95      0.93      0.94        17
weighted avg       0.95      0.94      0.94        17
```

ThankYou...