FA19-BCS-001 Group-2
Muhammad Nouman Tahir

IDS ASSIGN 4

Question No. 1

Part 1:

Answer:

The given dataset contains 80 instances as there are 80 rows in dataset each row represents as a single instance so there are 80 rows means 80 instances.

Part 2:

Answer:

The data set contains 7 input attributes which are as follows:

I. Height
II. Weight
III. Beard
IV. Hair_length
V. Shoe_size
VI. Scarf
VII. Eye_color

Part 3:

Answer:

The data set contains 2 possible values in the output attribute “gender” which are male and female.

Part 4:

Answer:

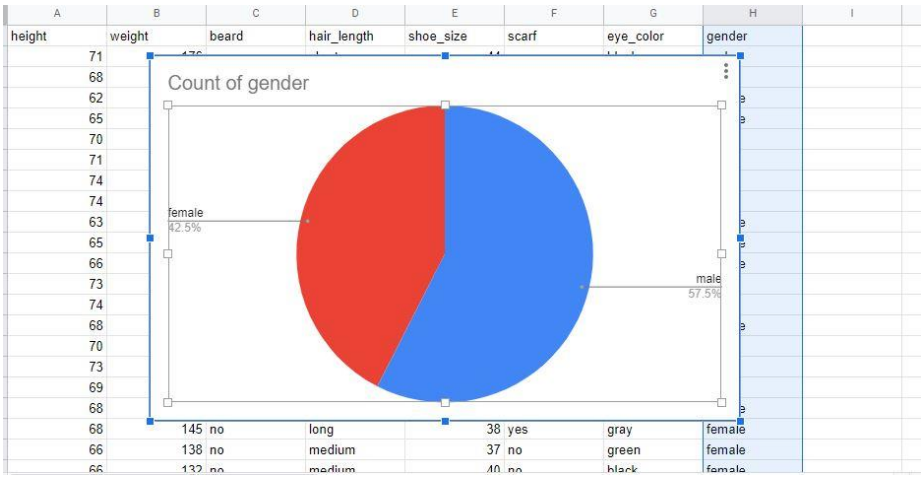
There are four categorical input attributes which are as follows:

I. Beard
II. Hair_length
III. Scarf
IV. Eye_color

Part 5:

Answer:

The class ratio of (male vs female) in the data set is 46:34. The male class is 57.5% of total which are 46 instances and female class is 42.5% of total data set which are 34 instances.



height	weight	beard	hair_length	shoe_size	scarf	eye_color	gender
71							
68							
62							
65							
70							
71							
74							
74							
63							
65							
66							
73							
74							
68							
70							
73							
69							
68							
68	145	no	long	38	yes	gray	female
66	138	no	medium	37	no	green	female
66	132	no	medium	40	no	black	female

Question No.2

Part 1

Answer:

To find incorrectly classified instances for the following I used confusion matrix. It is a matrix with dimension 2x2. To find the result from confusion matrix the non-diagonal values are added with each other which gives us incorrectly classified instances.

For Random Forest Classifier:

```
[[10  0]
 [ 0 17]]
```

The incorrectly classified instances are 0.

The accuracy is 100%.

For SVC_Classifier:

```
[[ 7  3]
 [ 3 14]]
```

The incorrectly classified instances are 6.

The accuracy is 77.778%.

For linearSVC_Classifier:

```
[[ 9  1]
 [ 2 15]]
```

The incorrectly classified instances are 3.

The accuracy is 88.889%.

For MLP_Classifier:

```
[[ 9  1]
 [ 1 16]]
```

The incorrectly classified instances are 2.

The accuracy is 92.5926%.

Part 2 (80/20 split)

When we change the train test split to 80/20 ratio with 80% for training and 20% for testing. The accuracy of Random Forest remains same as above but for SVC its accuracy has increased up to 81.25% from previous accuracy which was 77.778%, for LinearSVC its accuracy has increased up to 93.75% from previous accuracy which was 88.889% and for MLP its accuracy has increased up to 100% from previous accuracy which was 92.5926%. In case for incorrectly classified instances the of Random Forest remains same as above but for SVC the incorrectly classified instances value reduces to 3 from previous which was 6, for LinearSVC the incorrectly classified instances value reduces to 1 from previous which was 3, for MLP the incorrectly classified instances value reduces to 0 from previous which was 2. Hence Overall after the 80/20 split ratio the overall result has improved for prediction of models used.

For Random Forest Classifier:

```
[[ 6  0]
 [ 0 10]]
```

The incorrectly classified instances are 0.

The accuracy is 100%.

For SVC_Classifier:

```
[[4 2]
 [1 9]]
```

The incorrectly classified instances are 3.

The accuracy is 81.25%.

For linearSVC_Classifier:

```
[[6 0]
 [1 9]]
```

The incorrectly classified instances are 1.

The accuracy is 93.75%.

For MLP_Classifier:

```
[[ 6  0]
 [ 0 10]]
```

The incorrectly classified instances are 0.

The accuracy is 100%.

Part 3:

The two attributes that are most powerful in prediction task are “beard” and “scarf”. The reason for choosing the above two are as other features such as height, weight, hair_length, shoe_size and eye color are the attributes which are common among the both genders and any of them can have it but beard is a thing which only male wear and scarf is a thing which only female wear there is possibility that male not wear beard and but in any case male cannot wear scarf on the other hand there is possibility that female not wear scarf but female cannot wear beard in any case.

Part 4 (80/20 split)

Answer:

When we change the train test split to 80/20 ratio and excluding two attributes “beard” and “scarf” The accuracy of Random Forest and remains same in both cases and for SVC its accuracy has remained same, for LinearSVC its accuracy has decreased up to 68.75% from previous accuracy which was 93.75% and for MLP its accuracy has decreased up to 75% from previous accuracy which was 100%. In case for incorrectly classified instances the of Random Forest remains same as above but for SVC the incorrectly classified instances value remains same, for LinearSVC the incorrectly classified instances value increases to 5 from previous which was 1, for MLP the incorrectly classified instances value increases to 4 from previous which was 0.

For Random Forest Classifier:

The incorrectly classified instances are 0.

The accuracy is 100%.

For SVC_Classifier:

The incorrectly classified instances are 3.

The accuracy is 81.25%.

For linearSVC_Classifier:

The incorrectly classified instances are 5.

The accuracy is 68.75%.

For MLP_Classifier:

The incorrectly classified instances are 4.

The accuracy is 75.0%.

Question No. 3:

Cross Validation Strategies:

Applying Decision Tree Classifier Algorithm:

Parameters Used in Monte Carlo Cross Validation:

1. n_splits= 10

2. test_size=0.33

3. random_state=2

4. F1-Score= 0.94

Parameters Used in Leave p-out Cross Validation:

1. P=2

2. F1-Score= 0.7741

Question No. 4

Height	Weight	Beard	Hair_length	Shoe_size	Scarf	Eye_color	gender
73	96	yes	short	41	no	green	male
63	120	no	long	33	no	brown	female
59	100	no	medium	42	no	black	male
75	130	yes	short	33	no	blue	male
60	200	no	medium	40	yes	green	female

In this we have applied gaussian naïve bayes algorithm is used. In this for training we have used 80 instances and for testing we used newly added 5 instances.

I. Training Instances= 80
II. Testing Instances= 5
III. Accuracy= 100%
IV. Precision=1.0
V. Recall=1.0