INTRODUCTION TO DATA SCIENCE

ASSIGNMENT REPORT

DECEMBER 16, 2022

NAME: MINAHIL SADIQ

Reg # sp20-bcs-023

SUBMITTED TO:

DR. MUHAMMAD SHARJEEL

QUESTION 1

1) How many instances does the dataset contain?

Answer:

80 instances are in the dataset.

2) How many input attributes does the dataset contain?

Answer:

7 input attributes.

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

Answer:

2 possible values. (**Male and Female**).

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 male and 34 female)

Ratio of **male** in dataset: **57.5**

Ratio of **female**: **42.5**

QUESTION 2

1) How many instances are incorrectly classified?

Answer: (67/33 ratio)

- ➢ Random Forest Classifier:

  No instance classified incorrectly.
- ➢ Support Vector Machine:

  6 instances classified incorrectly.
- ➢ Multilayer Perceptron:

  14 instances classified incorrectly.

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

Answer:

- Random Forest Classifier:
  No instance classified incorrectly.
- Support Vector Machine:
  2 instances classified incorrectly.
- Multilayer Perceptron:
  1 instance classified incorrectly.

- ✓ There is no change in results of **Random Forest classifier**, with both split ratios it gives us 100% accuracy results.
- ✓ **Support Vector Machine Classifie**r (**with 80/20 ratio**) gives us good accuracy rate of **87.5%** which was only **77.8%** (**with 67/33 ratio**).
- ✓ With the **Multilayer Perceptron** change in results are amazing as it was giving us only **48.1%** accuracy (**with 67/33 ratio**) and it failed to classify Male category, but it predicts all Female right but did not classify even one Male category rightly. Now (**with 80/20 ratio**), it gives us good results with accuracy of **93.75%,** it predicts all male category correctly and only misclassified one female instance.

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

Answer:

Beard and **scarf** are the two most powerful attributes because Beard would always be false in case of female which is a good attribute to perfectly classify females, but on the other hand male can or can not have beard, therefore the second attribute I considered as powerful is scarf because scarf will never be true in case of male and it can or can not be true in case of female.

Suppose 0 represent false, 1 represent true,

Beard	Scarf	Gender
1	0	Male
0	1	Female

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:

- Random Forest Classifier:

  No change.
- Support Vector Machine:

  Works exactly as before (with 80/20 ratio), gives same accuracy of **87.5%.**
- Multilayer Perceptron:

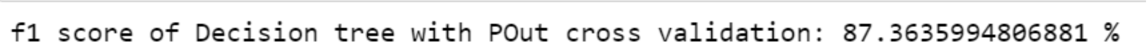
  There is huge change in this as with **80/20 ratio** and all **seven attributes** Multilayer perceptron gives us **93.75%** accuracy, but with the **same ratio** and **five attributes** (**excluding beard and scarf**), it gives us only **43.75%** accuracy and misclassified all the male instances as female.

QUESTION 3

Leave P-out cross validation:

The value of 'p' set to 3.

p=3

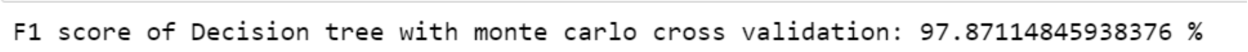


```
f1 score of Decision tree with POut cross validation: 87.3635994806881 %
```

Monte Carlo cross validation:

The value of n-split set to 5.

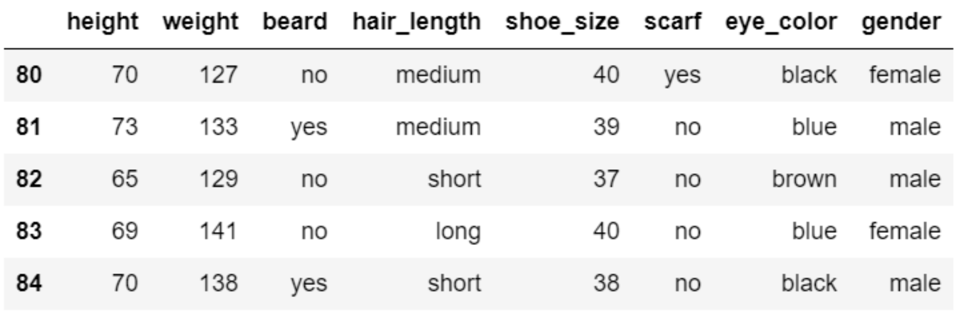
n_splits = 5



```
F1 score of Decision tree with monte carlo cross validation: 97.87114845938376 %
```

QUESTION 4

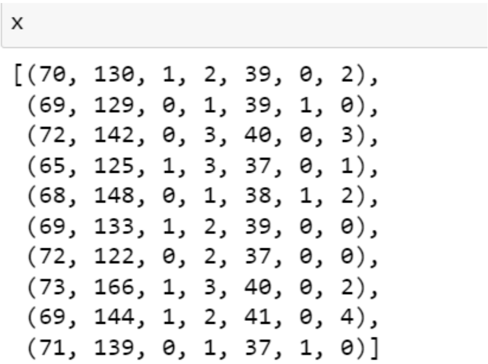
New five training instances:



	height	weight	beard	hair_length	shoe_size	scarf	eye_color	gender
80	70	127	no	medium	40	yes	black	female
81	73	133	yes	medium	39	no	blue	male
82	65	129	no	short	37	no	brown	male
83	69	141	no	long	40	no	blue	female
84	70	138	yes	short	38	no	black	male

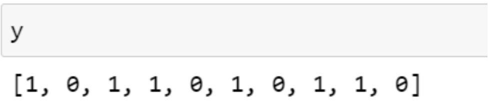
Test instances:

height	weight	beard	hair_length	shoe_size	scarf	eye_color	gender
70	130	yes	medium	39	no	brown	male
69	129	no	long	39	yes	black	female
72	142	no	short	40	no	grey	male
65	125	yes	short	37	no	blue	male
68	148	no	long	38	yes	brown	female
69	133	yes	medium	39	no	black	male
72	122	no	medium	37	no	black	female
73	166	yes	short	40	no	brown	male
69	144	no	medium	41	no	green	male
71	139	yes	long	37	yes	black	female



```
x
```

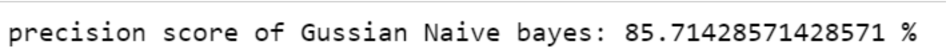
```
[(70, 130, 1, 2, 39, 0, 2),
 (69, 129, 0, 1, 39, 1, 0),
 (72, 142, 0, 3, 40, 0, 3),
 (65, 125, 1, 3, 37, 0, 1),
 (68, 148, 0, 1, 38, 1, 2),
 (69, 133, 1, 2, 39, 0, 0),
 (72, 122, 0, 2, 37, 0, 0),
 (73, 166, 1, 3, 40, 0, 2),
 (69, 144, 1, 2, 41, 0, 4),
 (71, 139, 0, 1, 37, 1, 0)]
```



```
y
```

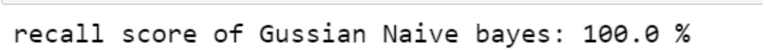
```
[1, 0, 1, 1, 0, 1, 0, 1, 1, 0]
```

PRECISION:



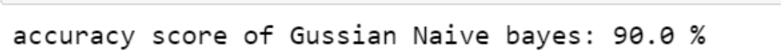
```
precision score of Gussian Naive bayes: 85.71428571428571 %
```

RECALL:



```
recall score of Gussian Naive bayes: 100.0 %
```

ACCURACY:



```
accuracy score of Gussian Naive bayes: 90.0 %
```