DATA SCIENCE

ASSIGNMENT :03

Date:**11/26/2023**

CSC461 – Assignment3 – Machine Learning

Your Full Name : **AHMAD AWAIS**

You Complete Registration Number : **FA21-BSE-123**

Question : 01

1. How many instances does the dataset contain?

The dataset contains 110 instances (rows).

2. How many input attributes does the dataset contain?

The dataset contains 8 input attributes: height, weight, beard, hair_length, shoe_size, scarf, eye color, and gender.

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

The output attribute is gender, and it seems to have two possible values: male and female.

4. How many input attributes are categorical?

Categorical attributes are those that represent discrete categories. In this dataset, the categorical input attributes are:

beard

hair_length

scarf

eye_color

gender

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

To determine the class ratio, we can count the occurrences of each gender in the "gender" column. It would be helpful to know the counts for both "male" and "female."

After counting, you can calculate the ratio using the formula:

Ratio = count of Male / count of female

If there are 60 males and 40 females, the ratio would be **60/40 = 1.5** , indicating ***1.5*** males for every female in the dataset.

Question : 02

1. How many instances are incorrectly classified?

```
Logistic Regression Incorrectly Classified Instances: 5
Support Vector Machines Incorrectly Classified Instances: 9
Multilayer Perceptron Incorrectly Classified Instances: 10
```

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

   After changing the test / train split ratio by 80/20 . The accuracy values of three models remains almost same. But , the number of instances incorrectly classified changes , in logistic regression remains same , support vector machines decreases by 2 and multilayer perceptron accuracy increases by 2.

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

   Height and weight because the values of these attributes highly fluctuate and is distinctive.

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 the height and weight attributes , the value of logistic regression accuracy becomes 1.0, support vector machines and multilayer perceptron accuracy remain almost same.
The values of number of instances incorrectly classified for Logistic regression becomes 0 , for support vector machines it decreases and becomes 2.
Lastly, for multilayer perceptron it decreases and becomes 6.

Question : 03

Parameter Values:

The n_estimators is set 100 while random_state is set 42 in random forest classifier.

The p_out is set 15 in p out Cross Validation

Question : 04

Instances for testing:

72,132,no,medium,45,no,brown,male

72,154,yes,medium,42,no,blue,male

70,119,yes,medium,40,no,brown,male

65,160,no,long,41,no,black,male

72,121,yes,medium,45,no,brown,male

61,103,no,long,38,no,green,female

68,135,no,long,37,yes,green,female

66,140,no,long,36,yes,gray,female

66,132,no,medium,37,no,black,female

70,160,no,long,42,no,black,male