INTRO TO DATA SCIENCE

Assignment

NAME: ANAS KHALID

ROLL NO: SP20-BCS-102

SECTION: C

ASSIGNMENT: IDS

[School]
[Course title]

QUESTION NO:1

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

Ans:

Dataset contains 80 instances

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

Ans:

Dataset contains 7 input attributes

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

Ans:

The output attribute has 2 possible values

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

Ans:

4 input attributes are categorical

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

Ans:

male ratio is 57.5% and female ratio is 42.5%

QUESTION NO:2

1. **How many instances are incorrectly classified?**

   Ans:

   In random forest **zero** instances are incorrectly classified.
   In Multilayer perceptron **eight** instances are incorrectly classified.
   In Support vector machine **Seven** instances are incorrectly classified.

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

   Yes, there is a change in the result of some classifiers. On train/test split ratio of 67/33 random forest shows 100% accuracy while MLP shows 70.37% accuracy and Support vector machine shows 74% accuracy. But when we split in 80/20 ratio the random forest shows the same accuracy but there is major change in the accuracy of MLP as it increases to 87.5% on this split while the accuracy of support vector machine increases 1% in this case.

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

   Height and Beard are the most powerful attributes in prediction task. In the following dataset height of male is mostly equal or above 70 while female height is below 70 and beard of every female is no while mostly male have beard which create a major difference to recognize male and 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.**

   After excluding two attributes the random forest and SVM remains the same while the accuracy of MLP decreases to 68%.

QUESTION NO:4

Test instances in your assignment submission document.

61	103	no	long	36	yes	brown	female
72	176	yes	short	40	no	brown	male
65	134	no	medium	42	no	black	female
61	122	no	long	32	yes	black	female
68	150	yes	short	32	no	black	male

Evaluate the trained model using the newly added test instances. Report accuracy, precision, and recall scores.

Accuracy is: 81.81818181818181%

Precision is: 69.23076923076923%

Recall is: 100.0%

QUESTION NO: 3

Parameter values for both cross-validation strategies.

FOR LEAVE P_OUT

LPO =LeavePOut(p=2) use this variable in cv.

FOR Monte Carlo cross-validation

train_size=0.6,

test_size=0.3,

n_splits = 5