2022

Assignment-4

INTRODUCTION TO DATA SCIENCE

Course: Introduction to Data Science
Submitted To: Muhammad Sharjeel
Submitted By: Muhammad Faizan
Student Id: CIIT/SP20-BCS-101/LHR
Section: B

Assignment Report

Machine Learning
(IDS)

Question: 1

1. How many instances does the dataset contain?

80 instances

2. How many input attributes does the dataset contain?

7 input attributes:

- height
- weight
- beard
- hair_length
- shoe_size
- scarf
- eye_color

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

2 possible values:

- Male
- Female

4. How many input attributes are categorical?

4 attributes are categorical:

- hair_length
- scarf
- eye_color

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

- Male instances: 46
- Female instances: 34

Class ratio: 23: 17

Question: 2

1. **How many instances are incorrectly classified?**
   - Random Forest: 0
   - Support Vector Machine: 6
   - Multilayer Perceptron: 10

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

   After using train/test split ratio of 80/20, the Accuracy of the Support Vector Machine and Multilayer Perceptron have gone up. Accuracy of Random Forest stays the same (i.e., 100%)

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

   2 most powerful attributes are believed to be "beard" and "scarf" as these can easily distinguish between a male and a female. Only males have beard, and females wear a scarf, so these 2 attributes are the most discriminating attributes.

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.**

   Random forest and Multilayer Perceptron accuracy stayed the same, where as the accuracy of SVM increased. The increase in accuracy could be because the model had more instances to train in 80/20 split than in 66/33 split. Therefore, the model was trained better with 80/20 split..

Question: 3

Monte Carlo Cross Validation

- Parameters: n_splits=4, test_size=0.33, random_state=2
- F1 Score = 96.01%

Leave p-out Cross Validation

- Parameter: LeavePOut(2) //Leave 2 out
- F1 Score = 95.04%

Question: 4

After adding 5 sample instances into the dataset, rerunning the ML experiment 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 test instances, the accuracy, precision, and recall scores are as follows:

- Accuracy = 100%
- Precision = 100%
- Recall = 100%

New Instances Added / Test Instances:

height	**weight**	**beard**	**hair_length**	**shoe_size**	**scarf**	**eye_color**	**gender**
71	157	yes	long	42	no	brown	male
69	158	yes	medium	43	no	gray	male
67	152	yes	short	43	no	black	male
63	100	no	long	38	no	brown	female
65	130	no	medium	40	yes	black	female