COMSATS UNIVERSITY

Introduction to Data Science

Assignment 4

Submitted To	Sir Muhammad Sharjeel
From	Muhammad Huzaifa Jawad
Roll No.	CIIT/SP20-BCS-144/LHR
Batch	SP20
Department	Computer Science
Section	BCS-B
Group	IV
Course Code	CSC-461

Question 1:

Provide responses to the following questions about the dataset.

1. **How many instances does the dataset contain?**
   80

2. How many input attributes does the dataset contain?

7 (height, weight, beard, hair_length, shoe_size, scarf, eye_color)

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

2 (male, female)

4. How many input attributes are categorical?

4 (beard, hair_length, scarf, eye_color)

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

Total Instances = 80

Number of Male Instances = 46

Number of Female Instances = 34

Ratio of Male vs Female = 46 / 34

Male = 57.5%

Female = 42.5%

Question 2:

Apply Random Forest, Support Vector Machines, and Multilayer Perceptron classification algorithms (using Python) on the gender prediction dataset with standard train/test split ratio and answer the following questions.

1. **How many instances are incorrectly classified?**
   Random Forest = 2,
   Support Vector Machines = 9,
   Multilayer Perceptron = 4

2. **Rerun the experiment using train/test split ratio of 80/20. Do you see any change in the results? Explain.**
   Using 80/20 train/test split ratio:
   Random Forest:
   F1_score increased from 94.11% to 100%.

   Support Vector Machines:
   F1_score increases from 75% to 84.61%.

   Multilayer Perceptron:

F1_score increases from 80.95% to 94.73%.

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

   “beard” and “scarf” are the most powerful attributes in my opinion. Because having a beard can instantly classify someone as male and wearing a scarf is mostly for females. As these 2 attributes can classify the gender easily, so they can be most powerful attributes in the dataset.

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:

   F1_score decreases from 100% to 94.73%.

   Support Vector Machines:

   F1_score decreases from 84.61% to 70.58.

   Multilayer Perceptron:

   F1_score decreases from 94.73% to 69.23.

Question 03:

Apply Decision Tree Classifier classification algorithm (using Python) on the gender redaction dataset with Monte Carlo cross-validation and Leave P-Out cross-validation. Report F_1 score for both cross-validation strategies.

For Monte Carlo cross-validation:

Splits = 3
Train / Test = 67-33 %
F_1 Score = 0.94813

For Leave-P-Out cross-validation:

P = 3
F_1 Score = 0.87281

Question 04:

Add 5 sample instances into the dataset (you can ask your friends/relatives/sibling for the data). Rerun the ML experiment (using Python) 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. Report accuracy, precision, and recall scores.

Accuracy = 96.55172
Precision = 93.75
Recall = 100

Instances:

height	weight	beard	hair_length	shoe_size	scarf	eye_color	gender
84	150	yes	medium	43	no	brown	male
78	142	no	long	42	no	brown	male
63	128	no	long	35	yes	brown	female
73	175	yes	short	44	no	black	male
68	186	no	long	38	yes	brown	female