Introduction to data science

(Assignment 4)

Instructor:

Muhammad Sharjeel

Name: Husnain Ahmed

Registration: SP20-BCS-033

Group: 4

Q1: Provide responses to the following questions about the dataset.

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

   Answer:

   80 instances.

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

   Answer:

   7 input attributes.

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

   Answer:

   Output attribute has 2 possible values.

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

   Answer:

   4 input attributes (beard, hair_length, scarf, eye_color) are categorical.

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

   Answer:

   57.5 : 42.5 (male : female) Ratio

   46 male

   34 female

Q2: 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?**

   Answer:

   We can find incorrectly classified instances by confusion matrix. So the results are as follows.

	Random Forest	**Support Vector Machines**	**Multilayer Perceptron**
Incorrect classified instances	2	1	3

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

   Answer:

   Random Forest:

   After applying the 80/20 train/test split, the accuracy increased from **92.59** to **100.**

   Confusion Matrix:

```
[9  0]
[0  7]
```

   SVM:

   After applying the 80/20 train/test split, the accuracy increased from **96.29** to **100.**

   Confusion Matrix:

```
[9  0]
[0  7]
```

Multilayer Perceptron:

After applying the 80/20 train/test split, the accuracy increased from **88.88** to **93.75.**

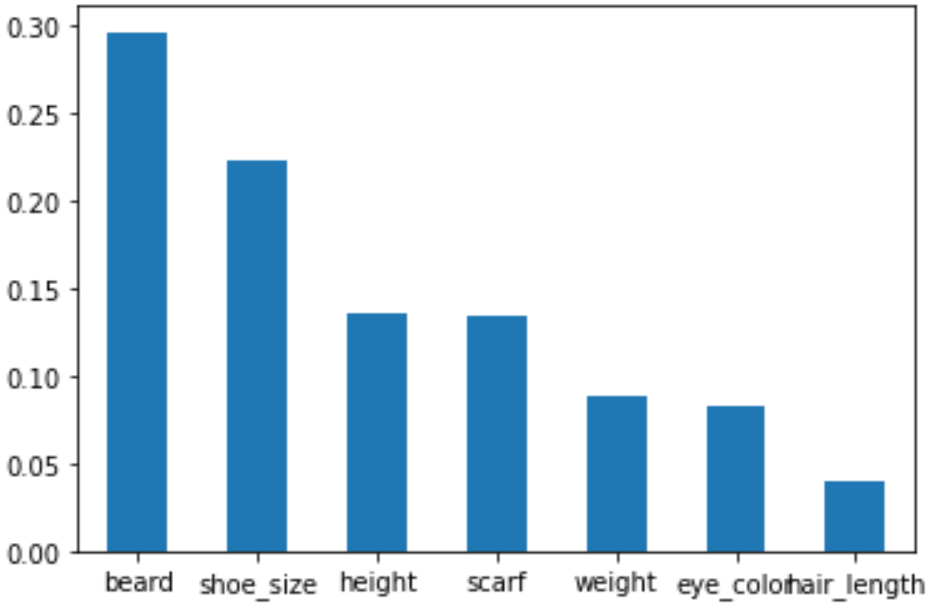
Confusion Matrix:

```
[9  0]
[1  6]
```

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

Answer:

```
Height:          0.135617
Weight:          0.088450
Beard:           0.295695
hair_length:     0.040078
shoe_size:       0.223117
scarf:           0.133774
eye_color:       0.083269
```



Above bar plot shows the feature importance of all the input attributes of the dataset. By visualizing this plot, we can clearly say that **beard** and **shoe_size** are the most powerful 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.

Answer:

	Random Forest	**Support Vector Machines**	**Multilayer Perceptron**
Accuracy	75.0	75.0	75.0
Confusion matrix	[3 1] [3 9]	[3 1] [3 9]	[0 4] [0 12]
Incorrect classified instances	4	4	4

Q3: Apply Decision Tree Classifier classification algorithm (using Python) on the gender prediction dataset with Monte Carlo cross-validation and Leave P-Out cross-validation. Report F1 score for both cross-validation strategies.
Note:

You are free to choose any parameter values for both cross-validation strategies, however, you have to provide these values in your submission document.

Monte Carlo cross-validation:

F1 Score: 0.9566

For Monte Carlo cross-validation, I have used splits = 10 and standard test/train size.

Leave P-out cross-validation:

F1 Score: 0.7777

For Monte Carlo cross-validation, I have used P = 2 and standard test/train size.

Q4: 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.
Note: You have to add the test instances in your assignment submission document.

Answer:

Gaussian Naive Bayes accuracy: 100.0 %

	precision	recall	f1-score	support
0	1.00	1.00	1.00	2
1	1.00	1.00	1.00	3
accuracy			1.00	5
macro avg	1.00	1.00	1.00	5
weighted avg	1.00	1.00	1.00	5