Name: Aleena Khan
Roll no: SP20-BCS-112
Group: 4

IDS-FA22-Assignment Due Date: 16-12-2022

Submission: Please upload the PDF report and Python code (preferably iPython notebook) to GitHub.

Download the gender prediction dataset from the following link:
https://drive.google.com/file/d/1EKpArZit1OdkfhaKVC3Beku6tkmASu3M/view?usp=share_link

Q1: 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', 'gender')
3. **How many possible values does the output attribute have?**
   2(male or female)
4. **How many input attributes are categorical**?
   5
   ('beard', 'hair_length', 'shoe_size', 'scarf', 'eye_color')
5. **What is the class ratio (male vs female) in the dataset?**
   Ratio of males: 46 / 80 (57.5%)
   Ratio of females: 34 / 80 (42.5%)

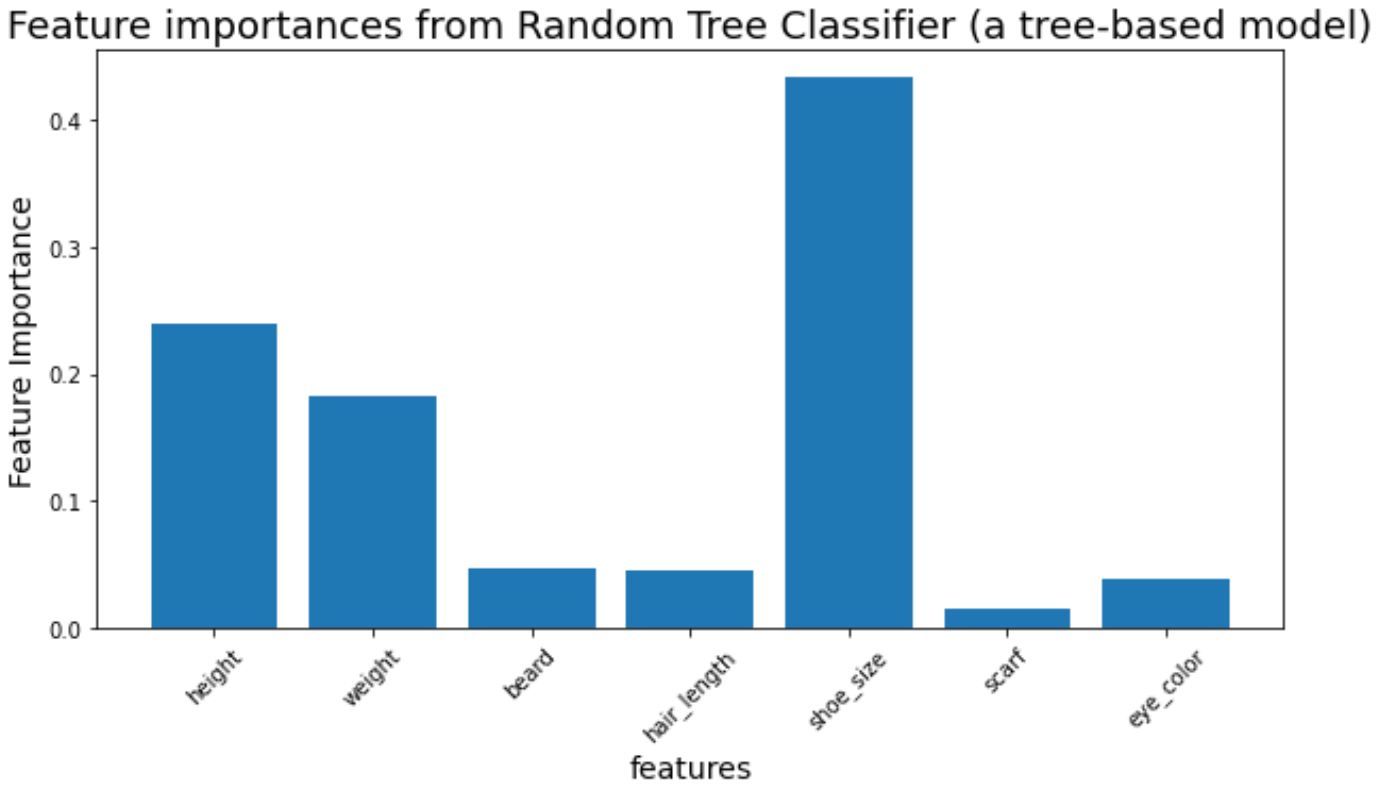
Q2: Apply Random Forest, Support Vector Machines, and Multilayer Perceptron classification algorithms (using Python) on the gender prediction dataset with standard train/test split ratio and answer the following questions.

1. **How many instances are incorrectly classified?**
   SVM: 0
   Random Forrest:0
   Multi-layer Perceptron:2
2. **Rerun the experiment using train/test split ratio of 80/20. Do you see any change in the results? Explain**.

	SVM	Random Forrest	Mutli Layer Perceptron
67%-33%: Accuracy: :	100 %	100%	92.59%
80%-20%: Accuracy:	100 %	93.75%	87.5%

There is some change in the accuracy with false predictions in 80%-20% split than the 67%-33% split because of change in the instances for training and testing when we applied the split again so, it randomly selected instances for testing again. Our dataset is extremely small. The dataset's visual representation leads us to the conclusion that the majority of attributes contain values that do not properly distinguish males from females.

3. **Name 2 attributes that you believe are the most "powerful" in the prediction task. Explain why?**
height and shoe size are important attributes in a way that variation in females and males is shown



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.**

Accuracy has decreased in SVM and random Forrest if we drop the 2 attributes of height and shoe size.

Accuracy:
SVM: 87.75%
Random Forrest:87.5%
Multi-layer Perceptron:93.75%

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 F_1 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:
Average Cross Validation score :0.9627224627224628
n_splits=5

Leave P-Out cross-validation:
Average Cross Validation score for leave p out :0.9424050632911393
p=2

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.

Newly add 5 instances are below:

height	weight	beard	hair_length	shoe_size	scarf	eye_color	gender
74	180	yes	short	43	no	black	male
78	175	no	short	42	no	blue	male
67	150	no	medium	38	no	black	female
75	188	yes	short	43	no	gray	male
64	145	no	long	35	yes	brown	female

After adding the new instances and using Gaussian Naibe Bayes over 80%-20% split:

Accuracy	Precision	Recall
93.75	90	100