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

1. How many instances does the dataset contain?
   Ans: 110 instances

2. How many input attributes does the dataset contain?
   Ans: 7 input instances. Height, weight, beard, hair length, shoe size, scarf, and eye color.

3. How many possible values does the output attribute have?
   Ans: There is only one output attribute gender that has two output values male and female.

4. How many input attributes are categorical?
   Ans: Beard, hair length, scarf, and eye color are the input attributes that are categorical.

5. What is the dataset's class ratio (male vs female)?
   Ans: 62 were males and 48 were females.

Q2: Apply Logistic Regression, Support Vector Machines, and Multilayer Perceptron classification algorithms (using Python) on the gender prediction dataset with 2/3 train and 1/3 test split ratio and answer the following questions.

1. How many instances are incorrectly classified?
   Ans: In Logistic Regression, the incorrectly classified are: 1

   In Support Vector Machines, the incorrectly classified are: 10

   In the Multilayer Perceptron classification, the incorrectly classified are: 14

2. Rerun the experiment using a train/test split ratio of 80/20. Do you see any change in the results? Explain.
   Ans: In Logistic Regression, the incorrectly classified are: 2

   In Support Vector Machines, the incorrectly classified are: 6

   In the Multilayer Perceptron classification, the incorrectly classified are: 2

   The reason is that we train the model with 80% of the data, which learns better than 67%. Therefore, the incorrectly classified were lesser.

3. Name 2 attributes you believe are the most "powerful" in the prediction task. Explain why?
   Ans: Beard and Scarf are the two most powerful attributes in the prediction because no female has a beard and no man has a scarf.

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
   Ans: In Logistic Regression, the incorrectly classified are: 2

   In Support Vector Machines, the incorrectly classified are: 6

   In the Multilayer Perceptron classification, the incorrectly classified are: 6

Q3: Apply the Random Forest classification algorithm (using Python) on the gender prediction dataset with Monte Carlo cross-validation and Leave P-Out cross-validation. Report F1 scores 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.

Ans: For Monte Carlo cross-validation Parameters are: (iteration = 11)

```
cross_val_score(randomForest_classifier,x,y,cv=iteration,scoring='f1_weighted')
```

For Leave P-out cross-validation Parameters are:

```
cross_val_score(randomForest_classifier, x, y, cv=2, scoring='f1_weighted')
```

Leave P-out cross-validation F1 score is: 0.9818118688216848

Monte Carlo cross-validation F1 score is: 0.9724275724275725

Add 10 sample instances into the dataset (you can ask your friends/relatives/siblings for the data). Run the ML experiment (using Python) by training the model using the Gaussian Naïve Bayes classification algorithm and all the instances from the gender prediction dataset. Evaluate the trained model using the newly added 10 test instances. Report accuracy, precision, and recall scores.

Note: You must use all the instances in the gender precision dataset for training and only 10 new instances for testing. You must include all the 10 test instances in your assignment submission document.

Ans: Accuracy: 0.7

Precision: 0.6952380952380952

Recall: 0.7