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

1. **How many instances does the dataset contain?**
   The dataset given for gender prediction task contain 80 instances. There are total 80 rows in dataset means 80 intances.
2. **How many input attributes does the dataset contain?**
   The dataset contain 7 input attributes. There are total 8 attributes in the data. One is output attribute and remaining 7 are input attributes.
3. **How many possible values does the output attribute have?**
   There are two possible values of output attribute. One is male and other one is female.
4. **How many input attributes are categorical?**
   There are four input attributes which are categorical. The name of categorical attributes are as follow.
   Beard.
   Hair_length.
   Scarf.
   Eye_color.
5. **What is the class ratio (male vs female) in the dataset?**
   There are 46 instances for male class which is 57.5% of total dataset and 34 instances for female instances which is 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?**
   The one way to find the count of incorrectly classified instances is iterating over predictions and testing data output. But I am using the confusion matrix to find incorrectly classified instances. This is classification problem so matrix of 2x2 will be generated. The sum of non-diagonal entries is equal to total incorrectly classified instances.

Model	Training instances (67%)	Testing instances (33%)	Accuracy	Incorrectly Classified instances
Random Forest	53	27	100%	0
Support Vector Machine	53	27	77.78%	6
Multilayer Perceptron	53	27	48.19%	14

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

Yes, when we run the experiment with 80% and 20% split. The accuracy of random forest remain same 100% but accuracy for Support vector machine and Multilayer Perceptron changed. Accuracy of SVM improves and accuracy of MLP decreases. Accuracy and number of incorrectly classified instances mentioned in table.

Model	Training instances (80%)	Testing instances (20%)	Accuracy	Incorrectly Classified instances
Random Forest	64	16	100%	0
Support Vector Machine	64	16	87.5%	2
Multilayer Perceptron	64	16	43.75%	9

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

   Beard and scarf are more powerful attribute in predicting task. These are the attributes which mainly distinguish between male and female. Female don't have beard and men do not wear scarf. Male and female may have same value of other attributes. Like both male and female can have short hair length. They may have same value of height, shoe size, eye color and weight.

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.**

   Yes, there is difference between accuracy of multilayer perceptron model trained without two selected attributes vs accuracy of models trained on all 7 attributes. It is 2 times of the accuracy with 7 attributes. Number of incorrectly classified instances by multilayer perceptron reduces when we trained the model without 2 selected attributes.

Model	Accuracy with training on 7 attributes	Accuracy with training the model without 2 selected attributes	Incorrectly Classified instances with training on 7 attributes	Incorrectly Classified instances with training the model without 2 selected attributes
Random Forest	100%	100%	0	0
Support Vector Machine	87.5%	87.5%	2	2
Multilayer Perceptron	43.75%	87.5%	9	2

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.

Cross Validation strategies	Model	Parameters used in cross-validation strategies	F1-score
Monte Carlo	Decision Tree	n_splits=10,test_size=0.20,random_state=1	0.96
Leave p-out	Decision Tree	P=3	0.87

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 added instances in dataset are as follow.

height	weight	beard	hair_length	shoe_size	scarf	eye_color	gender
72	69	no	short	42	no	brown	male
65	71	yes	short	39	no	black	male
57	47	no	long	37	yes	gray	female
68	83	yes	short	38	no	brown	male
62	50	no	medium	36	yes	black	female

In this experiment, gausian naïve bayes algorithm is used. 80 instances are used to train the data and newly added 5 instances are used to test the model.

Model	Number of training example	Number of testing example	Accuracy	Precision	Recall
Gausian Naïve Bayes	80	5	100%	1.00	1.00