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

1. How many instances does the dataset contain?

It contains 80 instances

2. How many input attributes does the dataset contain?

Dataset have 7 input attributes.

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

It has 2 possible vales. Male or Female.

4. How many input attributes are categorical?

The following four input attributes are categorical

- beard
- hair_length
- eye_color
- scarf

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

- Females are 42.5%
- Male are 57.5%

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?

In Random Forest --> 1 instance

In SVM --> 10 instances

In Multilayer Preceptron --> 4 instances

Rerun the experiment using train/test split ratio of 80/20. Do you see any change in

The results? Explain.

A minor increase in accuracy is observed in case of random forest and multilayer perceptron. This is because the model is now trained on more instances and tested on less instances.

Name 2 attributes that you believe are the most "powerful" in the prediction task. Explain why?
A quantity called mutual information measures the amount of information one can obtain from one random variable given another.

So depending upon the Mutual Information Gain of each attribute, 'Shoe Size' and 'Height' are the most powerful attributes.

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.

If these attributes are excluded, the model gives more false positives and false negatives and a decrease in accuracy is observed.

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.

Monte Carlo Cross Validation:

F1 score = 0.97

Leave P-Out cross-validation

F1 score = 0.77

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

New Test Instances:

[69, 167, 'yes', 'short', 43, 'no', 'brown', 'male'],

[62, 125, 'no', 'long', 37, 'yes', 'black', 'female'],

[73, 181, 'yes', 'medium', 42, 'no', 'black', 'male'],

[65, 102, 'no', 'long', 36, 'no', 'grey', 'female'],

[77, 180, 'yes', 'short', 44, 'no', 'blue', 'male']

The model predicted all 5 instances correctly.

Accuracy = 1.0

Precision = 1.0

Recall = 1.0