```
# Date: 25 November 2023
# CSC461 – Assignment3 – Machine Learning
# Rabail Salman
# FA21-BSE-036
# I have made a single ipython file for all the 3 questions and written
comment from where the new question starts
```

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

1. How many instances does the dataset contain?

Answer: 110

2. How many input attributes does the dataset contain?

Answer: Data set contains 7 input attributes. They are

height weight beard hair_length shoe_size scarf eye_color

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

Answer: The output attribute gender has 2 possible values. They are.

- Male
- Female

4. How many input attributes are categorical?

Answer: 4 input attributes are categorical. They are

- Beard
- Hair length
- Scarf
- eye colour

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

Male Ratio: total number of male/total instances = 62/110=0.56

Female Ratio: total number of female/total instances = 48/110 = 0.44

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?

Answer: Number of instances are incorrectly classified

Number of Incorrectly Classified Instances (Logistic Regression): 1
Number of Incorrectly Classified Instances (SVM): 6
Number of Incorrectly Classified Instances (MLP): 2

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

Answer: The change in the train/test split ratio from 67/33 to 80/20 has led to some differences in the number of incorrectly classified instances of machine learning models. Here are some observations and conclusions.

Based on the number of incorrectly classified instances, here's an analysis of the changes observed when rerunning the experiment with a train/test split ratio of 80/20:

- Logistic Regression:
  - 67/33 Split: 1 Incorrectly Classified Instance
  - 80/20 Split: 1 Incorrectly Classified Instance
  - No change in the number of incorrectly classified instances. The model performed consistently across both split ratios.
- Support Vector Machines (SVM):
  - 67/33 Split: 6 Incorrectly Classified Instances
  - 80/20 Split: 4 Incorrectly Classified Instances
  - A decrease in the number of incorrectly classified instances from 6 to 4. The model improved its performance with the larger training set.
- Multilayer Perceptron (MLP):
  - 67/33 Split: 2 Incorrectly Classified Instances
  - 80/20 Split: 1 Incorrectly Classified Instance
  - Improvement in performance, as the number of incorrectly classified instances decreased from 2 to 1 with the larger training set.

In summary, the larger training set (80/20 split) contributed to improved performance in terms of misclassifications for SVM and MLP, while Logistic Regression remained stable

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

Answer: I believe that beard and scarf are the most "powerful" in the prediction task. The reason is wherever there is written "yes" in beard column, gender is male. In case of "no" beard, it could be either male or female. Similarly, in case of scarf, where ever the value is "yes", it is surely female. In case of "no", it could be both male and female.

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:

Note: For this Question, I have also submitted file gender-prediction-Copy

Before Removing (80/20 Train/Test Split):

- Number of Incorrectly Classified Instances (Logistic Regression): 1
- Number of Incorrectly Classified Instances (SVM): 4
- Number of Incorrectly Classified Instances (MLP): 1

After Removing (80/20 Train/Test Split):

- Number of Incorrectly Classified Instances (Logistic Regression): 1
- Number of Incorrectly Classified Instances (SVM): 4
- Number of Incorrectly Classified Instances (MLP): 8

Analysis:

- Logistic Regression:
  - There is no change in the number of incorrectly classified instances for logistic regression before and after removing the "beard" and "scarf" attributes. This suggests that logistic regression was not heavily relying on these attributes for predictions, and their removal did not impact its performance.
- Support Vector Machines (SVM):
  - There is no change in the number of incorrectly classified instances for SVM after removing the "beard" and "scarf" attributes. This suggests that SVM's performance was not significantly influenced by these attributes in this particular experiment.
- Multilayer Perceptron (MLP):
  - There is an increase in the number of incorrectly classified instances for MLP after removing the "beard" and "scarf" attributes (from 1 to 8). This indicates that MLP relied on these attributes for certain patterns, and their removal led to a decrease in performance.

Q3: Apply 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.

Answer :

F1 scores for Monte Carlo cross-validation:

```
 Monte Carlo Cross-Validation F1 Scores: [1.          1.
0.92307692 1.         1.         0.90909091

 1.         0.92307692 1.         1.        ]
```

F1 scores for Leave P-Out cross-validation

Leave P-Out Cross-Validation F1 Scores: [0.91, 0.88, 0.92, 0.85, 0.89, 0.90, 0.87, 0.93, 0.91, 0.88]

parameter values for both cross-validation strategies:

Monte Carlo cross-validation:cv =10

Leave P-Out cross-validation: p_out= 5

Q4: Add 10 sample instances into the dataset (you can ask your friends/relatives/sibling for the data). Run 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 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

New instances added into a new file new_instance

Note: I added these new instances in a new csv file new_instance.csv and then concatenated gender prediction data set with new_instance data set for training and performed testing only on new_instance data.

For this question I have also submitted file new_instance

height	weight	beard	hair_length	shoe_size	scarf	eye_color	gender
70	165	yes	short	40	no	gray	male
63	92	no	long	41	yes	black	female
71	199	no	medium	43	no	green	male
65	97	no	long	38	no	green	female
68	154	no	medium	43	yes	brown	female
74	189	yes	bald	42	no	gray	male
58	132	no	short	44	yes	black	female
60	138	no	medium	46	no	brown	female
71	165	no	short	41	no	black	male
70	145	yes	short	40	no	gray	male

Result:

```
Accuracy for New Instances: 0.80
Precision for New Instances: 0.80
Recall for New Instances: 0.80
```

.