Intro to Data Science

Assignment No. 3

Name: Zohaib Murtaza
Roll No: FA21-BSE-138
Section: C
Submitted to: Dr. Muhammad Sharjeel

Date: Nov 25, 2023

Topic:

Usi the dataset file “gender-prediction.csv” available on shared Google Drive folder and doing the given tasks.

COMSATS University Islamabad, Lahore Campus

Question: 1

1. **How many instances does the dataset contain?**

110 Instances

2. **How many input attributes does the dataset contain?**

7 attributes

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

2 (Male & Female)

4. **How many input attributes are categorical?**

The categorical attributes are Beard, Hair Length, Scarf, Eye Color, and Gender.

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

Class ratio for male and female is 62/48 = 31/24
(62 male and 48 female)

Question: 2

1. **How many instances are incorrectly classified?**

Incorrectly Classified Instances (LR): 1
Incorrectly Classified Instances (SVM): 6
Incorrectly Classified Instances (MLP): 0

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

Yes, results changed.
LR Accuracy: 95.45454545454545
Incorrectly Classified Instances (LR): 1

SVM Accuracy: 81.81818181818183
Incorrectly Classified Instances (SVM): 4

MLP Accuracy: 95.45454545454545
Incorrectly Classified Instances (MLP): 1

The change in results with an 80/20 split indicates sensitivity to the training set size. Logistic Regression and MLP show minor accuracy fluctuations, suggesting robustness, while SVM is more sensitive. The counts of incorrectly classified instances provide insights into each model's performance on the new split. Overall, the variation highlights the importance of selecting an appropriate train/test split ratio for reliable model evaluation.

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

    - **Hair Length:**
        - Hair length might be influential in gender prediction, as it often correlates with gender norms and societal expectations.
        - Longer hair is traditionally associated with females, and shorter hair with males, making it a potentially significant predictor.
    - **Shoe Size:**
        - Shoe size could be indicative of physical characteristics related to gender, such as height.
        - While not a definitive measure, it might capture some aspects of physiological differences between genders.

    Although beard is a prominent difference between male and female, but it cannot be applied anywhere. In some cultures, beards are common among males, while in others, it’s not.

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.**

```
LR Accuracy:  86.36363636363636
Incorrectly Classified Instances (LR):  3

SVM Accuracy:  77.27272727272727
Incorrectly Classified Instances (SVM):  5

MLP Accuracy:  63.63636363636363
Incorrectly Classified Instances (MLP):  8
```

Excluding "Hair Length" and "Shoe Size" led to decreased accuracy in all models, with the most significant impact on MLP. This highlights the importance of these attributes in gender prediction and the need for careful feature selection. The most substantial impact was on the MLP model, indicating that these excluded attributes played a crucial role in capturing patterns and variations in the data.

Question: 3

Monte Carlo Cross-Validation F1 Scores:

```
cv=5, scoring='f1_weighted'
```

Results:

```
[1.        0.95405031 1.        1.        0.95425837]
```

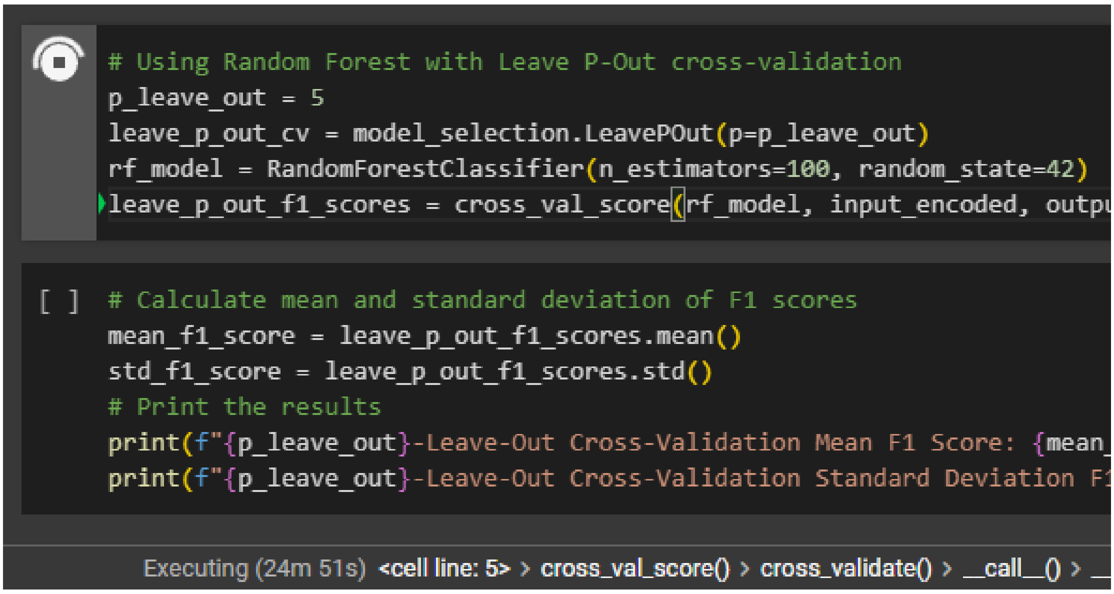
P Leave-Out Cross-Validation F1 Scores:

```
cv=2, scoring='f1_weighted'
```

Results:

2-Leave-Out Cross-Validation Mean F1 Score: 0.9751

2-Leave-Out Cross-Validation Standard Deviation F1 Score: 0.1166



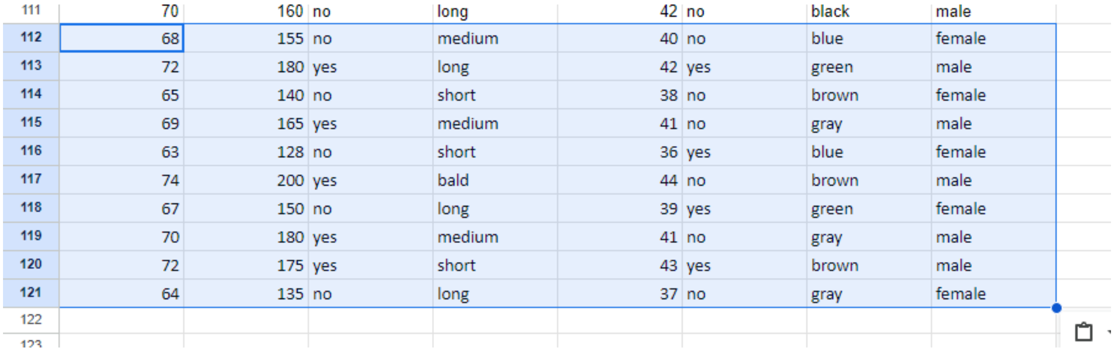
```
# Using Random Forest with Leave P-Out cross-validation
p_leave_out = 5
leave_p_out_cv = model_selection.LeavePOut(p=p_leave_out)
rf_model = RandomForestClassifier(n_estimators=100, random_state=42)
leave_p_out_f1_scores = cross_val_score(rf_model, input_encoded, outpu
```

```
[ ] # Calculate mean and standard deviation of F1 scores
    mean_f1_score = leave_p_out_f1_scores.mean()
    std_f1_score = leave_p_out_f1_scores.std()
    # Print the results
    print(f"{p_leave_out}-Leave-Out Cross-Validation Mean F1 Score: {mean_
    print(f"{p_leave_out}-Leave-Out Cross-Validation Standard Deviation F1
```

```
Executing (24m 51s) <cell line: 5> › cross_val_score() › cross_validate() › __call__() › _
```

(Took approximately 30 minutes to complete)

Question: 4



111	70	160	no	long	42	no	black	male
112	68	155	no	medium	40	no	blue	female
113	72	180	yes	long	42	yes	green	male
114	65	140	no	short	38	no	brown	female
115	69	165	yes	medium	41	no	gray	male
116	63	128	no	short	36	yes	blue	female
117	74	200	yes	bald	44	no	brown	male
118	67	150	no	long	39	yes	green	female
119	70	180	yes	medium	41	no	gray	male
120	72	175	yes	short	43	yes	brown	male
121	64	135	no	long	37	no	gray	female
122								
123								

Gaussian Naive Bayes Accuracy: 97.5
Gaussian Naive Bayes Precision: 0.9761363636363637
Gaussian Naive Bayes Recall: 0.975