



Intro To Data Science

Assignment 4 Report

SUBMITTED BY:

MUHAMMAD ASAD

SP20-BCS-158

SECTION C

Q1:

1. The dataset contains 80 instances.
2. The dataset contains 7 input attributes.
3. The output attribute has 2 possible values (male and female).
4. 4 input attributes are categorical.
5. There are 46 male instances and 36 female instances, so male vs female ratio is 1.35.

Q2:

Incorrect instances found using confusion matrix.

Incorrectly Classified Using Standard Split

- Random Forest = 0
- SVC = 6
- MLP = 3

Incorrectly Classified Using 80/20 Split

- Random Forest = 0
- SVC = 3
- MLP = 0

Excluding Attributes

In my opinion **scarf** and **beard** are the two most distinguishable and defining attributes to classify between males and females. They have consistent values for both males and females, so they are most important.

Incorrectly Classified After Excluding Scarf and Beard

- Random Forest = 1
- SVC = 3
- MLP = 4

Excluding these features reduces the performance of model. Excluding these features makes it a little harder for ML model to predict as these are the most defining features for both output variables.

Q3:

Cross Validation Scores

Monte Carlo F1 Score = 95 .48

No of Splits = 5

Standard Split

Random State = 2

Leave P Out F1 Score = 77.648

P = 2

Q4:

Newly Added Instances:

| | | | | | | | |
|----|-----|-----|--------|----|-----|-------|--------|
| 65 | 169 | no | long | 41 | no | blue | female |
| 72 | 172 | yes | short | 43 | no | brown | male |
| 69 | 168 | no | medium | 42 | yes | brown | female |
| 82 | 179 | no | short | 45 | no | black | male |
| 77 | 175 | yes | medium | 44 | no | brown | male |

New Instances Evaluation Scores:

Accuracy Score = 80

Recall Score = 1.00

Precision Score = 0.75

