**Assignment 3(Report)**

**Introduction to Data Science**

**A logo of a university

Description automatically generated**

**Name :** Usama Tufail

**Reg. No. :** FA21-BSE-053

**Section :** C

**Submitted to :** Dr. Muhammad Sharjeel

**Comsats University Islamabad, Lahore Campus**

**Question 1:**

1. The dataset contains 110 instances.
2. Dataset contains 7 input attributes.
3. Output has only 2 values which are male and female.
4. 4 input attributes are categorical which are beard, hair\_length, scarf, eye\_color.
5. 62 are males and 48 are females.

**Question 2:**

**In Logistic regression:**

1. When I use 2/3 for training and 1/3 for testing, 1 instance is incorrectly classified. A close-up of numbers

   Description automatically generated
2. When I use 80% for training and 20% for testing, 0 instances are incorrectly classified. A screenshot of a computer

   Description automatically generated
3. shoe\_size and scarf are the most important attributes; I have applied Random Forest Algorithm on data, and I got these 2 attributes most important. A close-up of a number

   Description automatically generated
4. Yes, I found change. When I dropped shoe\_size and scarf then 1 instance is incorrectly classified. But when we have these both attributes in input and our data is split in 80%(training) 20%(testing) ratio then 0 instances are incorrectly classified. A screenshot of a computer

   Description automatically generated

**In Support vector machines:**

1. When I use 2/3 for training and 1/3 for testing, 5 instances are incorrectly classified. A screenshot of a computer

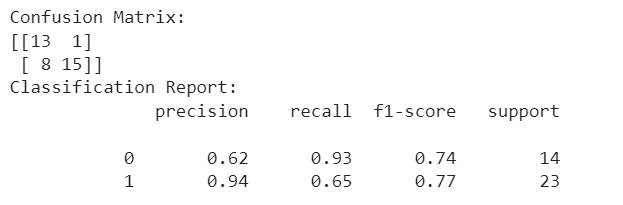
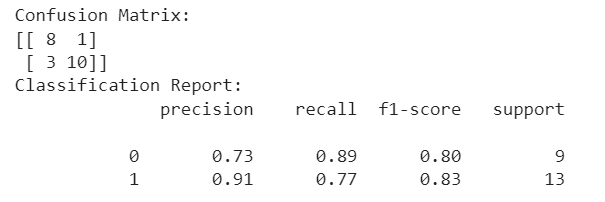
   Description automatically generated
2. When I use 80% for training and 20% for testing, 4 instances are incorrectly classified. A screenshot of a computer

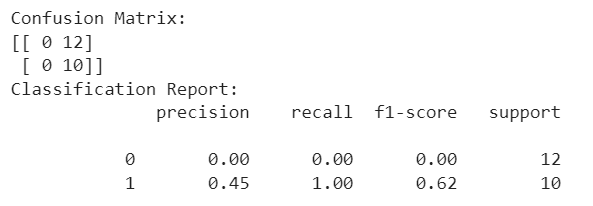
   Description automatically generated
3. shoe\_size and scarf are the most important attributes; I have applied Random Forest Algorithm on data, and I got these 2 attributes most important. A close-up of a number

   Description automatically generated
4. Yes, I found change. When I dropped shoe\_size and scarf then 6 instances are incorrectly classified. But when we have these both attributes in input and our data is split in 80%(training) 20%(testing) ratio then 4 instances are incorrectly classified. A white background with black text

   Description automatically generated

**In Multilayer Perceptron classification:**

1. When I use 2/3 for training and 1/3 for testing, 9 instances are incorrectly classified. 
2. When I use 80% for training and 20% for testing, 4 instances are incorrectly classified. 
3. shoe\_size and scarf are the most important attributes; I have applied Random Forest Algorithm on data, and I got these 2 attributes most important. A close-up of a number

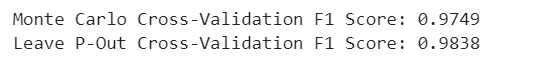
   Description automatically generated
4. Yes, I found change. When I dropped shoe\_size and scarf then 12 instances are incorrectly classified. But when we have these both attributes in input and our data is split in 80%(training) 20%(testing) ratio then 9 instances are incorrectly classified. 

**Question 3:**

**Monte Carlo cross-validation technique:**

I applied Random Forest classification algorithm on the gender prediction dataset with Monte Carlo cross-validation. When I use Monte Carlo cross-validation technique I use number of iterations, input attributes, output attributes and model as a parameter and value of number of iterations is 100 and I get f1 Score for Monte Carlo cross-validation is 0.9749.

**Leave P-Out Cross-Validation technique:**

I applied Random Forest classification algorithm on the gender prediction dataset with Leave P-Out Cross-Validation. When I use Leave P-Out Cross-Validation technique I use number of folds p, input attributes, output attributes and model as a parameter and value of p is 4 and I get f1 Score for Leave P-Out Cross-Validation is 0.9838. 

**Q4:**

10 new instances are:

| **Height** | **Weight** | **Beard** | **Hair Length** | **Shoe Size** | **Scarf** | **Eye Color** | **Gender** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| 68 | 150 | Yes | Medium | 42 | No | Brown | Male |
| 65 | 135 | No | Short | 38 | Yes | Green | Female |
| 70 | 180 | Yes | Short | 44 | No | Blue | Male |
| 72 | 160 | Yes | Bald | 41 | No | Brown | Male |
| 67 | 140 | No | Medium | 39 | Yes | Gray | Female |
| 69 | 155 | No | Medium | 41 | Yes | Blue | Female |
| 63 | 110 | Yes | Short | 38 | No | Black | Male |
| 75 | 190 | Yes | Bald | 46 | No | Brown | Male |
| 66 | 145 | No | Medium | 40 | Yes | Green | Female |
| 68 | 130 | Yes | Medium | 40 | No | Black | Male |

I applied Gaussian Naïve Bayes classification algorithm on all the instances from the gender prediction dataset. I evaluated model on these 10 instances, 0 instances are incorrectly classified.

