SP20-BCS-087

Sana Imran

Group-4

Machine Learning Models & Techniques

IDS-Assignment-4

**IDS-FA22-Assignment Due Date: 16-12-2022**

Submission: Please upload the PDF report and Python code (preferably iPython notebook) to GitHub. Download the gender prediction dataset from the following link:

https://drive.google.com/file/d/1EKpArZit1OdkfhaKVC3Beku6tkmASu3M/view?usp=share\_link

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

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

Ans: 80

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

Ans: 7

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

Ans: 2

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

Ans: 4(beard, hair\_length, scarf, eye\_color)

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

Ans: From pie chart and calculations, we can conclude that there are **46/80** **male** i-e; **57.5%** and **34/80 female** i-e; **42.5%** instances in the data.

**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?**

**Ans:** 0

* 1. In Random Forest Classifier, **0** instances are incorrectly classified.
  2. In SVM Classifier, **0** instances are incorrectly classified.
  3. In Multilayer Perceptron Classifier, **0** instances are incorrectly classified.

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

**Ans:** Yes.

* 1. In Random Forest Classifier, **0** instances are incorrectly classified.
  2. In SVM Classifier, **0** instances are incorrectly classified.
  3. In Multilayer Perceptron Classifier, **1** instance is incorrectly classified.

There is a change in the accuracy with false predictions in 80%-20% split than the 67%-33% split because of change in the instances for training and testing when we applied the split again so, it randomly selected instances for testing again. Also, the size of our dataset is very small to train our models properly. From visualizing the dataset, it is concluded that, most of the attributes contain values which do not create difference between male and female properly.

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

**Ans:** I tested through computing features’ importance using random forest classifier and got that **“Shoe Size”** and **“Beard”** are the most powerful attributes which play a major role in prediction of classes because they contain more distinctive values for separate classes than other attributes hence, making prediction powerful.

1. **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.**

**Ans:** Yes. The Random Forest Classifier, which was previously giving 100% accuracy, is now reduced in its accuracy. The Multilayer Perceptron has a different accuracy too. This is because of a small dataset and even the most powerful attributes didn’t make much change because there is a very less difference in the importance/powerfulness of “Shoe Size”, “Beard” and “Weight” (3rd number feature on importance scale).

**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.  
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.**

**Ans:**

* 1. For **Monte Carlo cross-validation**, the parameters used in ShuffleSplit are:

train\_size=0.7, test\_size=0.3, n\_splits = 5

The results are:

**Cross Validation F1 Scores:**

[0.89655172, 0.92857143, 0.83333333, 0.96, 0.91666667]

**Average F1 Score:**

0.9445370370370372

**Standard Deviation F1 Score:**

0.034228823400527214

* 1. For **Leave P-Out cross-validation,** the “p” value I gave in the parameter is “2”. The results are:

**Cross Validation F1 Scores:**

[1. 1. 1. ... 1. 0.66666667 0.66666667]

**Average F1 Score:**

0.7821729957805906

**Standard Deviation F1 Score:**

0.3973170265420241

**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.**

**Note: You have to add the test instances in your assignment submission document.**