Name: Muhammad Umair Tariq **Reg#FA19-BCS-055**

Introduction to Data Science

Assigment#4

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

1. How many instances does the dataset contain?

Answer: There 80 instances in dataset.

2. How many input attributes does the dataset contain?

Answer: Dataset contains 7 input attributes.

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

Answer: There are 2 possible values of output e.g. male and female.

4. How many input attributes are categorical?

Answer: 4 Input attributes are categorical.

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

Answer: Males=46 Females=34

Ratio of male vs female is 46:34=23:17 = 1.35

Q2: Apply Random Forest, Support Vector Machines, and Multilayer Perceptron classification algorithm (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?

Answer:

Random Forest: No instances are incorrectly classified.

SVC: 6 instances are incorrectly classified.

LinearSVC: 2 instances are incorrectly classified but it is different every time we run it due to different data in different iterations.

Multilayer Perceptron: 3 instances are incorrectly classified but it is different every time we run it due to different data in different iterations.

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

Answer: Yes, there are changes in results. Accuracy increased as we are giving more data for training the model.

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

Answer: Beard and scarf are most powerful attributes because they are beard is only for males and scarf is only for males hence it helps model to predict easily that if beard is yes than it is male and if scarf is yes then it is female. Other attributes can have same values for both male and female but these 2 mostly 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.

Answer: Yes there is little decrement in accuracy of SVC and Random Forrest as we have decreases attributes that have important role for gender prediction. So, model training is not very good.

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.

Answer:

F1 Score for Monte Carlo is 0.9622 or 96.22%

F1 score for Leave P-Out cross-validation is 0.777 or 77.7%

Values for Monte Carlo are Test data size = 20% and Train data size = 80% and Iterations=10.

Value of p in leave p-out is 2.

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

Answer:

Test Instances:

Height	Weight	Beard	Hair_length	Shoe_size	Scarf	Eye_color	Gender
57	147	No	Medium	33	Yes	Gray	Female
68	156	No	Short	45	No	Black	Male
75	160	Yes	Long	42	No	Blue	Male
61	135	No	Long	38	No	Brown	Female
66	140	no	medium	43	yes	blue	female

Accuracy: 100%

Precision: 100%

Recall Scores: 100%