Assignment-4

Subject: Introduction to Data Science

Submitted to: Muhammad Sharjeel

Submitted By:

Fatima Naqvi

CIIT/SP20-BCS-145/LHR

Section: B

Q1.Provide responses to the following questions about the dataset

1. **How many instances does the dataset contain?**
   It contains 80 instances.
2. **How many input attributes does the dataset contain?**
   It contains 7 attributes.
3. **How many possible values does the output attribute have?**
   2 values are possible (male, female).
4. **How many input attributes are categorical?**
   4 attributes are categorical.
5. **What is the class ratio (male vs female) in the dataset?**
   Total = 80
   Male = 46
   Female = 34
   Ratio:
   Male = 57.5%
   Female = 42.5%

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?**
   Random Forest = 1
   Support Vector Machines = 7
   Multiplayer Perceptron = 3
2) **Rerun the experiment using train/test split ratio of 80/20. Do you see any change in the results? Explain**
   With both ratios, results are almost same with respect to incorrect instances but there is a little difference in accuracy of SVM and perceptron. 80/20 split have a little less accuracy in comparison.
3) **Name 2 attributes that you believe are the most "powerful" in the prediction task. Explain why?**
   I feel beard and scarf are most powerful attributes for prediction because no women have beard and same goes for scarf, there is no man who wear scarfs so there is a clear distinction between genders.
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.**
   SVM and random forest have remained same but the accuracy have become less in multiclass perceptron.

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.

- F1 score for P-out cross validation is 0.96
- F1 score for Monte Carlo cross validation is 0.96

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.

- Accuracy = 92%
- Precision = 83.33%
- Recall = 100%