HW8_Programming: The Experiment of K-anonymity

Group: yayaya
108753109 資科碩一 陳庭軒
108753134 資科碩一 高士傑
108753124 資科碩一 楊芝辰
108753102 資科碩一 吳映函

1. Dataset: Adult (test size $= 0.3 \times \text{dataset}$)
2. Programming Language: Python
3. Simple report:
   (1) How do you perturb the data?
   A: By L-diversity.
   Our function support both numeric values and categoric values. For numeric values, each iterator is a mean split. For categoric values, each iterator is a split on GH. The final result is returned in 2-dimensional list.
   (2) What is your machine learning model used in this experiment?
   Besides, provide some details/parameters of this model.
   A: 機器學習模型使用 Python sklearn 底下的 svm.SVC。
   SVC 有很多參數，例如：gamma 值、probability 等，在這次實驗中所有參數皆使用預設值。
   (3) Effectiveness measure: Misclassification Error, Accuracy, Precision, Recall, and AUC
   A: 預測結果為 Occupation

Occupation	One hot encode number
Tech-support	0
Craft-repair	1
Other-service	2
Sales	3
Exec-managerial	4
Prof-specialty	5
Handlers-cleaners	6
Machine-op-inspct	7
Adm-clerical	8
Farming-fishing	9
Transport-moving	10
Priv-house-serv	11
Protective-serv	12
Armed-Forces	13

Misclassification Error: 0.7512998266897747

Accuracy: 0.19493866725605039

Confusion Matrix:

```
        0  1    2   3   4    5   6   7   8   9  10  11  12  13
0   [[  0  19   4   0 150  92   0   0   9   0   0   0   0   0]
1   [  0 287  13   3 823  79   0   0   4   0   0   0   0   0]
2   [  0 103  60  61 532 139   0   0  68   0   0   0   0   0]
3   [  0 109  21  45 629 230   0   0  41   0   0   0   0   0]
4   [  0  69   1   3 884 219   0   0  22   0   0   0   0   0]
5   [  0  25   4   8 758 408   0   0   8   0   0   0   0   0]
6   [  0 130  39   4 201  24   0   0   7   0   0   0   0   0]
7   [  0 133   9   7 379  39   0   0  23   0   0   0   0   0]
8   [  0  70  11  37 685 233   0   0  80   0   0   0   0   0]
9   [  0  65  10   1 197  21   0   0   3   0   0   0   0   0]
10  [  0 115   7   2 328  17   0   0   3   0   0   0   0   0]
11  [  0   0   1   5  27   5   0   0   5   0   0   0   0   0]
12  [  0  28   2   2 124  31   0   0   6   0   0   0   0   0]
13  [  0   1   0   0   2   0   0   0   0   0   0   0   0   0]]
```

Classification Report (Precision and Recall):

	precision	recall	f1-score	support
0	0.00	0.00	0.00	274
1	0.25	0.24	0.24	1209
2	0.33	0.06	0.10	963
3	0.25	0.04	0.07	1075
4	0.15	0.74	0.26	1198
5	0.27	0.34	0.30	1211
6	0.00	0.00	0.00	405
7	0.00	0.00	0.00	590
8	0.29	0.07	0.11	1116
9	0.00	0.00	0.00	297
10	0.00	0.00	0.00	472
11	0.00	0.00	0.00	43
12	0.00	0.00	0.00	193
13	0.00	0.00	0.00	3
accuracy			0.19	9049
macro avg	0.11	0.11	0.08	9049
weighted avg	0.19	0.19	0.14	9049

Area Under the Curve (AUC): 0.39271424140375744

(4) Privacy level: $L = 5$

4. 執行流程：

（因每部電腦效能差異，執行時間大約 2 分鐘以上）

(1) 先執行”anonymizer.py”

會在”data”底下產生”anonymized.data”

(2) 將”data\anonymized.data”轉為”anonymized.csv”（已附在壓縮檔中）

(3) 執行”SVM.py”，即可看到上述的截圖結果