Deepam Sarmah
2020050
deepam20050@iiitd.ac.in

AI Assignment 4

Steps for data preparation/modification:

- Implemented reclassification to reduce target variables to 7 classifications tasks as per the question. These 7 classification tasks are:
    - Architect
    - Engineer
    - SysAdmin
    - Analyst
    - Technician
    - QoS-Engineer
    - SDE
- I used label encoding to encode the various labels for columns such as ‘worked in teams ever?’, certifications’, ‘workshops’ etc

Experiments pedagogy:

- Used various train-test splits as mentioned in the report
- Also shuffled the rows of the data frame to check if the accuracy improved. It improved in the case of 90-10 train-test split

Train-Test Splits:

- 90-10:
    - Accuracy: 15.15
    - Confusion Matrix:

```
[[ 35   0   0   0   0   0   0 359   0   0   0   0   0]
 [  6   0   0   0   0   0   0  50   0   0   0   0   0]
 [  1   0   0   0   0   0   0  63   0   0   0   0   0]
 [  6   0   0   0   0   0   0  61   0   0   0   0   0]
 [  5   0   0   0   0   0   0  46   0   0   0   0   0]
 [  4   0   0   0   0   0   0  50   0   0   0   0   0]
 [  3   0   0   0   0   0   0  50   0   0   0   0   0]
 [ 29   0   0   0   0   0   0 268   0   0   0   0   0]
 [ 16   0   0   0   0   0   0 211   0   0   0   0   0]
 [ 24   0   0   0   0   0   0 255   0   0   0   0   0]
 [  7   0   0   0   0   0   0 113   0   0   0   0   0]
 [ 21   0   0   0   0   0   0 210   0   0   0   0   0]
 [ 10   0   0   0   0   0   0  97   0   0   0   0   0]]
```

    - Class-wise accuracies:

```
[0.08883249 0.         0.         0.         0.         0.
 0.         0.9023569  0.         0.         0.         0.
 0.        ]
```

- 70-30:
  - Accuracy: 19.616666666666667
  - Confusion Matrix:

```
[[1177    0    0    0    0    0    0    0    0    0    0    0    0]
 [ 158    0    0    0    0    0    0    0    0    0    0    0    0]
 [ 173    0    0    0    0    0    0    0    0    0    0    0    0]
 [ 192    0    0    0    0    0    0    0    0    0    0    0    0]
 [ 172    0    0    0    0    0    0    0    0    0    0    0    0]
 [ 164    0    0    0    0    0    0    0    0    0    0    0    0]
 [ 163    0    0    0    0    0    0    0    0    0    0    0    0]
 [ 868    0    0    0    0    0    0    0    0    0    0    0    0]
 [ 683    0    0    0    0    0    0    0    0    0    0    0    0]
 [ 864    0    0    0    0    0    0    0    0    0    0    0    0]
 [ 344    0    0    0    0    0    0    0    0    0    0    0    0]
 [ 706    0    0    0    0    0    0    0    0    0    0    0    0]
 [ 336    0    0    0    0    0    0    0    0    0    0    0    0]]
```

  - Class-wise accuracies:

```
[1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]
```

- 60-40:
  - Accuracy: 11.387500000000001
  - Confusion Matrix:

```
[[  46    0    0    0    0    0    0    0 1527    0    0    0    0]
 [   4    0    0    0    0    0    0    0  200    0    0    0    0]
 [   6    0    0    0    0    0    0    0  226    0    0    0    0]
 [   6    0    0    0    0    0    0    0  241    0    0    0    0]
 [   2    0    0    0    0    0    0    0  224    0    0    0    0]
 [   9    0    0    0    0    0    0    0  222    0    0    0    0]
 [   6    0    0    0    0    0    0    0  216    0    0    0    0]
 [  25    0    0    0    0    0    0    0 1147    0    0    0    0]
 [  28    0    0    0    0    0    0    0  865    0    0    0    0]
 [  34    0    0    0    0    0    0    0 1126    0    0    0    0]
 [  16    0    0    0    0    0    0    0  437    0    0    0    0]
 [  35    0    0    0    0    0    0    0  914    0    0    0    0]
 [  15    0    0    0    0    0    0    0  423    0    0    0    0]]
```

  - Class-wise accuracies:

```
[0.02924348 0.         0.         0.         0.         0.
 0.         0.         0.96864502 0.         0.         0.
 0.        ]
```

- 50-50:
  - Accuracy: 19.64
  - Confusion Matrix:

```
[[1964    0    0    0    0    0    0    0    0    0    0    0    0]
 [ 271    0    0    0    0    0    0    0    0    0    0    0    0]
 [ 285    0    0    0    0    0    0    0    0    0    0    0    0]
```

[309 0 0 0 0 0 0 0 0 0 0 0 0]
[287 0 0 0 0 0 0 0 0 0 0 0 0]
[288 0 0 0 0 0 0 0 0 0 0 0 0]
[295 0 0 0 0 0 0 0 0 0 0 0 0]
[1488 0 0 0 0 0 0 0 0 0 0 0 0]
[1118 0 0 0 0 0 0 0 0 0 0 0 0]
[1418 0 0 0 0 0 0 0 0 0 0 0 0]
[580 0 0 0 0 0 0 0 0 0 0 0 0]
[1149 0 0 0 0 0 0 0 0 0 0 0 0]
[548 0 0 0 0 0 0 0 0 0 0 0 0]]

- Class-wise accuracies:
  [1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

- 40-60:
  - Accuracy: 19.541666666666664
  - Confusion Matrix:
    [[2345 0 0 0 0 0 0 0 0 0 0 0 0]
    [339 0 0 0 0 0 0 0 0 0 0 0 0]
    [332 0 0 0 0 0 0 0 0 0 0 0 0]
    [364 0 0 0 0 0 0 0 0 0 0 0 0]
    [357 0 0 0 0 0 0 0 0 0 0 0 0]
    [341 0 0 0 0 0 0 0 0 0 0 0 0]
    [345 0 0 0 0 0 0 0 0 0 0 0 0]
    [1774 0 0 0 0 0 0 0 0 0 0 0 0]
    [1362 0 0 0 0 0 0 0 0 0 0 0 0]
    [1692 0 0 0 0 0 0 0 0 0 0 0 0]
    [704 0 0 0 0 0 0 0 0 0 0 0 0]
    [1369 0 0 0 0 0 0 0 0 0 0 0 0]
    [676 0 0 0 0 0 0 0 0 0 0 0 0]]
  - Class-wise accuracies:
    [1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

- 90-10 (with shuffled data frame rows):
  - Accuracy: 19.0
  - Confusion Matrix:
    [[380 0 0 0 0 0 0 0 0 1 0 0 0]
    [62 0 0 0 0 0 0 0 0 0 0 0 0]
    [71 0 0 0 0 0 0 0 0 0 0 0 0]
    [57 0 0 0 0 0 0 0 0 0 0 0 0]
    [60 0 0 0 0 0 0 0 0 0 0 0 0]
    [43 0 0 0 0 0 0 0 0 0 0 0 0]
    [66 0 0 0 0 0 0 0 0 0 0 0 0]
    [252 0 0 0 0 0 0 0 0 0 0 0 0]
    [242 0 0 0 0 0 0 0 0 0 0 0 0]
    [296 0 0 0 0 0 0 0 0 0 0 0 0]
    [132 0 0 0 0 0 0 0 0 0 0 0 0]

[232 0 0 0 0 0 0 0 0 0 0 0 0]
[106 0 0 0 0 0 0 0 0 0 0 0 0]]

- Class-wise accuracies:
  [0.99737533 0. 0. 0. 0. 0.
  0. 0. 0. 0. 0. 0.
  0.]

- 50-50 (with shuffled data frame rows):
  - Accuracy: 19.28
  - Confusion Matrix:
    [[1928 0 0 0 0 0 0 0 0 0 0 0 0]
    [290 0 0 0 0 0 0 0 0 0 0 0 0]
    [288 0 0 0 0 0 0 0 0 0 0 0 0]
    [312 0 0 0 0 0 0 0 0 0 0 0 0]
    [292 0 0 0 0 0 0 0 0 0 0 0 0]
    [266 0 0 0 0 0 0 0 0 0 0 0 0]
    [301 0 0 0 0 0 0 0 0 0 0 0 0]
    [1453 0 0 0 0 0 0 0 0 0 0 0 0]
    [1161 0 0 0 0 0 0 0 0 0 0 0 0]
    [1409 0 0 0 0 0 0 0 0 0 0 0 0]
    [573 0 0 0 0 0 0 0 0 0 0 0 0]
    [1139 0 0 0 0 0 0 0 0 0 0 0 0]
    [588 0 0 0 0 0 0 0 0 0 0 0 0]]
  - Class-wise accuracies:
    [1. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0. 0.]

Analysis & Results: When the train-test split is 50-50 the accuracy is the maximum possible