AI Assignment 4

Arjun Temura
2020497

Code Workflow:

1. Data wrangling/preprocessing:
   The original dataset is of dimensions (20000, 59).
   Label: “Suggested Job role” (consists of 34 distinct job roles)

   a) Data is cleansed of all the unnecessary information regarding the predictions such as 'interested in games','Interested Type of Books','In a Relationship?','Gentle or Tuff behaviour?','reading and writing skills','Taken inputs from seniors or elders' etc.
   b) Fields containing data in the form of ‘yes’ or ‘no’ are modified to contain data in ‘1’ and ‘0’.
   c) Fields with data in the form of object (eg. 'Interested subjects',’interested career areas’,’Type of company want to settle in’ etc.) are converting into numeric data values using binary encoding.
   d) Further, fields with marks in different technical fields are merged as a single field as ‘Computer Science Percentage’ to reduce dimensionality of the dataset.
   e) Some fields like ‘workshops’, ‘worked in teams ever?’ and ‘memory capability’ are removed through accuracy testing.
   f) Since there are 34 categories of Job roles, these categories are merged to form 5 broad categories namely,
      - ***'Data Analyst'***
      - ***'Technical Analyst',***
      - ***'Security/Networking Analyst',***
      - ***'Business Analyst',***
      - ***‘Software/UX Analyst'***

      We have 29 columns after data preprocessing

2.Data Training and Testing:
a)The data is split into training and testing data.
b) This data is normalized to scale all the field values to the same scale.
c) tSNE is implemented(perplexity=2) but does not give better results.
d) Data is passed through the MLPClassifier with hidden layers=(40,20,2), max_iter=1000,activation = 'relu',solver='adam',random_state=1.

e) The classification report, accuracy and confusion matrix are calculated.

Analysis of results:

a) We achieve maximum accuracy (i.e. **27 %**) with a split ratio of 10:90 in comparison to 30:70, 40:60 and 10:90.
b) We observe that some unnecessary fields can be detected through accuracy testing.
c) The MLPClassifier with the above arguments gives optimal results for our ANN model.
d) We have successfully created our ANN model for the given dataset.

1) 10:90 split

```
Classification report:
                           precision    recall  f1-score   support

          Business Analyst       0.00      0.00      0.00       301
              Data Analyst       0.05      0.00      0.01       209
Security/Networking Analyst       0.00      0.00      0.00       467
        Software/UX Analyst       0.27      0.92      0.42       534
          Technical Analyst       0.22      0.07      0.11       489

                   accuracy                           0.27      2000
                  macro avg       0.11      0.20      0.11      2000
               weighted avg       0.13      0.27      0.14      2000
```

```
Confusion matrix:
[[  0   4   0 256  41]
 [  0   1   0 196  12]
 [  0   8   0 419  40]
 [  0   5   0 493  36]
 [  0   3   0 450  36]]
The accuracy of model is 26.5 %
```

2) 25:75 split

	precision	recall	f1-score	support
Business Analyst	0.00	0.00	0.00	943
Data Analyst	0.00	0.00	0.00	751
Security/Networking Analyst	0.00	0.00	0.00	1536
Software/UX Analyst	0.25	0.90	0.40	1674
Technical Analyst	0.26	0.10	0.15	1696
accuracy			0.26	6600
macro avg	0.10	0.20	0.11	6600
weighted avg	0.13	0.26	0.14	6600

```
Confusion matrix:
[[   0    0    0  843  100]
 [   0    0    0  672   79]
 [   0    0    0 1374  162]
 [   0    0    0 1507  167]
 [   0    0    0 1519  177]]
The accuracy of model is 25.515151515151512 %
```

3) 30:70 split

	precision	recall	f1-score	support
Business Analyst	0.00	0.00	0.00	853
Data Analyst	0.00	0.00	0.00	684
Security/Networking Analyst	0.00	0.00	0.00	1415
Software/UX Analyst	0.25	0.92	0.40	1531
Technical Analyst	0.21	0.06	0.10	1517
accuracy			0.25	6000
macro avg	0.09	0.20	0.10	6000
weighted avg	0.12	0.25	0.13	6000

```
Confusion matrix:
[[   0    0    0  785   68]
 [   0    0    0  632   52]
 [   0    0    0 1296  119]
 [   0    0    0 1401  130]
 [   0    0    0 1419   98]]
The accuracy of model is 24.98333333333333 %
```

4) 40:60 split

	precision	recall	f1-score	support
Business Analyst	0.00	0.00	0.00	1156
Data Analyst	0.11	0.02	0.04	906
Security/Networking Analyst	0.00	0.00	0.00	1862
Software/UX Analyst	0.25	0.86	0.39	2022
Technical Analyst	0.24	0.11	0.15	2054
accuracy			0.25	8000
macro avg	0.12	0.20	0.12	8000
weighted avg	0.14	0.25	0.14	8000

```
Confusion matrix:
[[   0   32    0  983  141]
 [   0   22    0  787   97]
 [   0   50    0 1599  213]
 [   0   50    0 1738  234]
 [   0   44    0 1788  222]]
The accuracy of model is 24.775 %
```

Assignment 1 Extension

Code Workflow:

1. Data wrangling/preprocessing
    a) We use fields with percentage in various courses, interest of subject, interest in career area, certifications and workshops as the input attributes for our training data.
    b) Certifications and workshops are renamed as 'Elective1' and 'Elective2' fields.
    c) We use these 3 parameters namely, grades, interests and electives as a criteria to decide future job roles.
    d) Binary encoding is used to represent string data in the interests and elective fields.
2. Data Training and Testing
    a) The data is trained using MLPClassifier to build an ANN model.
    b) The data is normalized before training using StandardScaler().
    c) The accuracy, confusion matrix and classification report is evaluated.

Analysis of Results:

a) The average accuracy is 25 % using a 25:75 split using (40,20,2) hidden layers. This accuracy is highest amongst other splits like 30:70, 40:60, 10:90 etc.

1. 10:90 split

	precision	recall	f1-score	support
Business Analyst	0.00	0.00	0.00	301
Data Analyst	0.00	0.00	0.00	209
Security/Networking Analyst	0.23	0.53	0.32	467
Software/UX Analyst	0.26	0.26	0.26	534
Technical Analyst	0.23	0.17	0.20	489
accuracy			0.24	2000
macro avg	0.14	0.19	0.16	2000
weighted avg	0.18	0.24	0.19	2000

```
Confusion matrix:
[[  0   0 176  80  45]
 [  0   0 110  58  41]
 [  0   0 248 127  92]
 [  0   0 290 141 103]
 [  0   0 265 139  85]]
The accuracy of model is 23.7 %
```

2. 25:75 split

	precision	recall	f1-score	support
Business Analyst	0.00	0.00	0.00	711
Data Analyst	0.00	0.00	0.00	579
Security/Networking Analyst	0.24	0.54	0.33	1186
Software/UX Analyst	0.25	0.25	0.25	1276
Technical Analyst	0.27	0.22	0.24	1248
accuracy			0.25	5000
macro avg	0.15	0.20	0.17	5000
weighted avg	0.19	0.25	0.20	5000

```
Confusion matrix:
[[  0   0 384 186 141]
 [  0   0 306 159 114]
 [  0   0 645 300 241]
 [  0   0 684 317 275]
 [  0   0 672 297 279]]
The accuracy of model is 24.82 %
```

3. 30:70 split

	precision	recall	f1-score	support
Business Analyst	0.11	0.00	0.00	853
Data Analyst	0.00	0.00	0.00	684
Security/Networking Analyst	0.23	0.52	0.32	1415
Software/UX Analyst	0.25	0.28	0.26	1531
Technical Analyst	0.25	0.18	0.21	1517
accuracy			0.24	6000
macro avg	0.17	0.20	0.16	6000
weighted avg	0.20	0.24	0.20	6000

```
Confusion matrix:
[[  2   0 451 240 160]
 [  3   0 361 197 123]
 [  3   0 736 411 265]
 [  4   0 823 428 276]
 [  6   0 788 443 280]]
The accuracy of model is 24.099999999999998 %
```

4. 40:60 split

	precision	recall	f1-score	support
Business Analyst	0.00	0.00	0.00	1156
Data Analyst	0.00	0.00	0.00	906
Security/Networking Analyst	0.23	0.54	0.32	1862
Software/UX Analyst	0.26	0.26	0.26	2022
Technical Analyst	0.26	0.20	0.22	2054
accuracy			0.24	8000
macro avg	0.15	0.20	0.16	8000
weighted avg	0.19	0.24	0.20	8000

```
Confusion matrix:
[[   0    0  618  315  223]
 [   0    0  507  214  185]
 [   0    0 1008  478  376]
 [   0    0 1112  535  375]
 [   0    0 1101  548  405]]
The accuracy of model is 24.349999999999998 %
```