LABS

Data Mining

Eng: Malek Almosanif

```
import pandas as pd

Hr=pd.read_csv('Hr_report.csv')

Hr.shape

(14999, 12)

Hr.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14999 entries, 0 to 14998
Data columns (total 10 columns):
 #   Column                 Non-Null Count  Dtype  
---  ------                 --------------  -----  
 0   satisfaction_level     14999 non-null  float64
 1   last_evaluation        14999 non-null  float64
 2   number_project         14999 non-null  int64  
 3   average_montly_hours   14999 non-null  int64  
 4   time_spend_company     14999 non-null  int64  
 5   Work_accident          14999 non-null  int64  
 6   promotion_last_5years  14999 non-null  int64  
 7   dept                   14999 non-null  int64  
 8   salary                 14999 non-null  int64  
 9   left                   14999 non-null  int64  
dtypes: float64(2), int64(8)
memory usage: 1.1 MB

Hr.describe()

       Unnamed: 0.1    Unnamed: 0  satisfaction_level  last_evaluation
\
count  14999.000000  14999.000000        14999.000000     14999.000000

mean    7499.000000   7499.000000            0.612834         0.716102

std     4329.982679   4329.982679            0.248631         0.171169

min        0.000000      0.000000            0.090000         0.360000

25%     3749.500000   3749.500000            0.440000         0.560000

50%     7499.000000   7499.000000            0.640000         0.720000

75%    11248.500000  11248.500000            0.820000         0.870000

max    14998.000000  14998.000000            1.000000         1.000000


       number_project  average_montly_hours  time_spend_company  \
count    14999.000000          14999.000000        14999.000000
mean         3.803054            201.050337            3.498233
```

```
std          1.232592             49.943099             1.460136
min          2.000000             96.000000             2.000000
25%          3.000000            156.000000             3.000000
50%          4.000000            200.000000             3.000000
75%          5.000000            245.000000             4.000000
max          7.000000            310.000000            10.000000

       Work_accident  promotion_last_5years          dept
salary  \
count   14999.000000           14999.000000  14999.000000
14999.000000
mean        0.144610               0.021268      5.870525
0.594706
std         0.351719               0.144281      2.868786
0.637183
min         0.000000               0.000000      0.000000
0.000000
25%         0.000000               0.000000      4.000000
0.000000
50%         0.000000               0.000000      7.000000
1.000000
75%         0.000000               0.000000      8.000000
1.000000
max         1.000000               1.000000      9.000000
2.000000

               left
count  14999.000000
mean       0.238083
std        0.425924
min        0.000000
25%        0.000000
50%        0.000000
75%        0.000000
max        1.000000
```

```
Hr.head()
```

```
   Unnamed: 0.1  Unnamed: 0  satisfaction_level  last_evaluation  \
0             0           0                0.38             0.53
1             1           1                0.80             0.86
2             2           2                0.11             0.88
3             3           3                0.72             0.87
4             4           4                0.37             0.52

   number_project  average_montly_hours  time_spend_company
Work_accident  \
0               2                   157                   3
0
1               5                   262                   6
```

```
0
2                7                    272                    4
0
3                5                    223                    5
0
4                2                    159                    3
0
```

```
   promotion_last_5years  dept  salary  left
0                      0     7       0     1
1                      0     7       1     1
2                      0     7       1     1
3                      0     7       0     1
4                      0     7       0     1
```

```
Hr.tail()
```

```
       Unnamed: 0.1  Unnamed: 0  satisfaction_level
last_evaluation  \
14994         14994       14994                0.40              0.57

14995         14995       14995                0.37              0.48

14996         14996       14996                0.37              0.53

14997         14997       14997                0.11              0.96

14998         14998       14998                0.37              0.52


       number_project  average_montly_hours  time_spend_company  \
14994               2                   151                   3
14995               2                   160                   3
14996               2                   143                   3
14997               6                   280                   4
14998               2                   158                   3

       Work_accident  promotion_last_5years  dept  salary  left
14994              0                      0     8       0     1
14995              0                      0     8       0     1
14996              0                      0     8       0     1
14997              0                      0     8       0     1
14998              0                      0     8       0     1
```

```
Hr = Hr.drop(['Unnamed: 0','Unnamed: 0.1'], axis=1)
```

Split Data => Features And Target

```
X=Hr.iloc[:,:-1]

Y=Hr.iloc[:,-1]
```

Features Selection

SelectKBest

```
from sklearn.feature_selection import SelectKBest,f_classif

FeatureSelection = SelectKBest(score_func= f_classif ,k=4)

X.shape
```

```
(14999, 9)
```

```
X_best = FeatureSelection.fit_transform(X, Y)

FeatureSelection.get_support()
```

```
array([ True, False, False, False,  True,  True, False, False,  True])
```

```
selected_features = X.columns[FeatureSelection.get_support()]
X_best=pd.DataFrame(X_best,columns=selected_features)

X_best.shape
```

```
(14999, 4)
```

```
X_best.head()
```

	satisfaction_level	time_spend_company	Work_accident	salary
0	0.38	3.0	0.0	0.0
1	0.80	6.0	0.0	1.0
2	0.11	4.0	0.0	1.0
3	0.72	5.0	0.0	0.0
4	0.37	3.0	0.0	0.0

```
Y.shape
```

```
(14999,)
```

Select Percentile

```
from sklearn.feature_selection import SelectPercentile,f_classif

FeatureSelection2 = SelectPercentile(score_func=
f_classif ,percentile=50)

X.shape
```

```
(14999, 9)
```

```
X_best = FeatureSelection2.fit_transform(X, Y)

FeatureSelection2.get_support()
```

```
array([ True, False, False, False,  True,  True, False, False,  True])
```

```
selected_features = X.columns[FeatureSelection2.get_support()]
X_best=pd.DataFrame(X_best,columns=selected_features)

X_best.shape
```

```
(14999, 4)
```

```
X_best.head()
```

```
   satisfaction_level  time_spend_company  Work_accident  salary
0                0.38                 3.0            0.0     0.0
1                0.80                 6.0            0.0     1.0
2                0.11                 4.0            0.0     1.0
3                0.72                 5.0            0.0     0.0
4                0.37                 3.0            0.0     0.0
```

Data Split

```
from sklearn.model_selection import train_test_split as tts

X_train, X_test, y_train, y_test = tts(X_best, Y, test_size=0.30,
random_state=30, shuffle =True)

X_train
```

```
       satisfaction_level  time_spend_company  Work_accident  salary
8436                 0.65                 3.0            0.0     0.0
2160                 0.46                 2.0            0.0     0.0
8516                 0.67                 4.0            0.0     0.0
12495                0.82                 6.0            0.0     1.0
14759                0.85                 5.0            0.0     0.0
...                   ...                 ...            ...     ...
13452                0.62                 3.0            0.0     1.0
500                  0.91                 5.0            0.0     1.0
12077                0.40                 3.0            0.0     1.0
4517                 0.25                 3.0            1.0     0.0
5925                 0.89                 3.0            0.0     0.0

[10499 rows x 4 columns]
```

```
X_test
```

```
       satisfaction_level  time_spend_company  Work_accident  salary
2463                 0.79                 3.0            0.0     0.0
6579                 0.72                 3.0            0.0     0.0
9578                 0.55                 4.0            0.0     1.0
8174                 0.63                 2.0            0.0     0.0
20                   0.11                 4.0            0.0     0.0
...                   ...                 ...            ...     ...
9433                 0.49                 3.0            0.0     0.0
2533                 0.74                 4.0            0.0     1.0
```

```
13375                0.91                   4.0            0.0     1.0
739                  0.11                   4.0            0.0     1.0
1293                 0.72                   2.0            0.0     0.0

[4500 rows x 4 columns]
```

Data Scaling

MinMax Scaling

MinMax:

صورة محلية

```
X

       satisfaction_level  last_evaluation  number_project  \
0                    0.38             0.53               2
1                    0.80             0.86               5
2                    0.11             0.88               7
3                    0.72             0.87               5
4                    0.37             0.52               2
...                   ...              ...             ...
14994                0.40             0.57               2
14995                0.37             0.48               2
14996                0.37             0.53               2
14997                0.11             0.96               6
14998                0.37             0.52               2

       average_montly_hours  time_spend_company  Work_accident  \
0                       157                   3              0
1                       262                   6              0
2                       272                   4              0
3                       223                   5              0
4                       159                   3              0
...                     ...                 ...            ...
14994                   151                   3              0
14995                   160                   3              0
14996                   143                   3              0
14997                   280                   4              0
14998                   158                   3              0

       promotion_last_5years  dept  salary
0                          0     7       0
1                          0     7       1
2                          0     7       1
3                          0     7       0
4                          0     7       0
...                      ...   ...     ...
```

```
14994                       0     8       0
14995                       0     8       0
14996                       0     8       0
14997                       0     8       0
14998                       0     8       0

[14999 rows x 9 columns]
```

```
from sklearn.preprocessing import MinMaxScaler

Scaler=MinMaxScaler(feature_range = (0,1))

from sklearn.model_selection import train_test_split as tts

X_train, X_test, y_train, y_test = tts(X, Y, test_size=0.30,
random_state=30, shuffle =True)

X_train_Scaled=Scaler.fit_transform(X_train)
X_test_Scaled=Scaler.transform(X_test)

X_train_Scaled=pd.DataFrame(X_train_Scaled,columns=X_train.columns)
X_test_Scaled=pd.DataFrame(X_test_Scaled,columns=X_test.columns)

X_train_Scaled
```

```
       satisfaction_level  last_evaluation  number_project  \
0                0.615385         0.437500             0.6
1                0.406593         0.515625             0.0
2                0.637363         0.437500             0.4
3                0.802198         0.984375             0.4
4                0.835165         0.718750             0.4
...                   ...              ...             ...
10494            0.582418         0.671875             0.2
10495            0.901099         0.750000             0.6
10496            0.340659         0.281250             0.0
10497            0.175824         0.359375             0.4
10498            0.879121         0.781250             0.2

       average_montly_hours  time_spend_company  Work_accident  \
0                  0.514019               0.125            0.0
1                  0.294393               0.000            0.0
2                  0.247664               0.250            0.0
3                  0.780374               0.500            0.0
4                  0.813084               0.375            0.0
...                     ...                 ...            ...
10494              0.761682               0.125            0.0
10495              0.789720               0.375            0.0
10496              0.191589               0.125            0.0
10497              0.570093               0.125            1.0
10498              0.523364               0.125            0.0
```

```
       promotion_last_5years      dept  salary
0                        0.0  1.000000     0.0
1                        0.0  0.777778     0.0
2                        0.0  1.000000     0.0
3                        0.0  1.000000     0.5
4                        0.0  0.777778     0.0
...                      ...       ...     ...
10494                    0.0  0.888889     0.5
10495                    0.0  1.000000     0.5
10496                    0.0  0.555556     0.5
10497                    0.0  0.777778     0.0
10498                    0.0  0.888889     0.0

[10499 rows x 9 columns]
```

X_test_Scaled

```
      satisfaction_level  last_evaluation  number_project  \
0               0.769231         0.328125             0.2
1               0.692308         0.421875             0.4
2               0.505495         0.921875             0.4
3               0.593407         0.765625             0.6
4               0.021978         0.734375             0.8
...                  ...              ...             ...
4495            0.439560         0.640625             0.2
4496            0.714286         0.750000             0.2
4497            0.901099         0.500000             0.4
4498            0.021978         0.765625             1.0
4499            0.692308         1.000000             0.0

      average_montly_hours  time_spend_company  Work_accident  \
0                 0.719626               0.125            0.0
1                 0.518692               0.125            0.0
2                 0.714953               0.250            0.0
3                 0.551402               0.000            0.0
4                 0.869159               0.250            0.0
...                    ...                 ...            ...
4495              0.238318               0.125            0.0
4496              0.668224               0.250            0.0
4497              0.168224               0.250            0.0
4498              0.836449               0.250            0.0
4499              0.672897               0.000            0.0

      promotion_last_5years      dept  salary
0                       0.0  0.777778     0.0
1                       0.0  0.777778     0.0
2                       0.0  1.000000     0.5
3                       0.0  0.555556     0.0
4                       0.0  0.777778     0.0
...                     ...       ...     ...
```

```
4495                    0.0  0.888889     0.0
4496                    0.0  0.777778     0.5
4497                    0.0  0.000000     0.5
4498                    0.0  0.888889     0.5
4499                    0.0  0.111111     0.0

[4500 rows x 9 columns]
```

Z-Score

z-Score:

صورة محلية صورة محلية

```
from sklearn.preprocessing import StandardScaler # z-Score
Scaler= StandardScaler()

from sklearn.model_selection import train_test_split as tts
X_train, X_test, y_train, y_test = tts(X, Y, test_size=0.30,
random_state=30, shuffle =True)

X_train_Scaled=Scaler.fit_transform(X_train)
X_test_Scaled=Scaler.transform(X_test)

X_train_Scaled=pd.DataFrame(X_train_Scaled,columns=X_train.columns)
X_test_Scaled=pd.DataFrame(X_test_Scaled,columns=X_test.columns)

X_train_Scaled
```

```
       satisfaction_level  last_evaluation  number_project  \
0                0.147461        -0.443673        0.974097
1               -0.621556        -0.152562       -1.466530
2                0.228410        -0.443673        0.160554
3                0.835529         1.594104        0.160554
4                0.956953         0.604327        0.160554
...                   ...              ...             ...
10494            0.026037         0.429660       -0.652988
10495            1.199800         0.720771        0.974097
10496           -0.864404        -1.025896       -1.466530
10497           -1.471523        -0.734784        0.160554
10498            1.118851         0.837215       -0.652988

       average_montly_hours  time_spend_company  Work_accident  \
0                  0.102223           -0.333841      -0.415074
1                 -0.844424           -1.018812      -0.415074
2                 -1.045838            0.351130      -0.415074
3                  1.250283            1.721072      -0.415074
4                  1.391273            1.036101      -0.415074
...                     ...                 ...            ...
10494              1.169717           -0.333841      -0.415074
```

```
10495              1.290566            1.036101      -0.415074
10496             -1.287534           -0.333841      -0.415074
10497              0.343919           -0.333841       2.409207
10498              0.142505           -0.333841      -0.415074
```

```
       promotion_last_5years      dept    salary
0                  -0.147986  1.097153 -0.930457
1                  -0.147986  0.400804 -0.930457
2                  -0.147986  1.097153 -0.930457
3                  -0.147986  1.097153  0.639344
4                  -0.147986  0.400804 -0.930457
...                      ...       ...       ...
10494              -0.147986  0.748978  0.639344
10495              -0.147986  1.097153  0.639344
10496              -0.147986 -0.295546  0.639344
10497              -0.147986  0.400804 -0.930457
10498              -0.147986  0.748978 -0.930457

[10499 rows x 9 columns]
```

X_test_Scaled

```
      satisfaction_level  last_evaluation  number_project  \
0                0.714105        -0.851229       -0.652988
1                0.430783        -0.501896        0.160554
2               -0.257285         1.361215        0.160554
3                0.066512         0.778993        0.974097
4               -2.038167         0.662549        1.787639
...                   ...              ...             ...
4495            -0.500132         0.313215       -0.652988
4496             0.511732         0.720771       -0.652988
4497             1.199800        -0.210784        0.160554
4498            -2.038167         0.778993        2.601181
4499             0.430783         1.652326       -1.466530

      average_montly_hours  time_spend_company  Work_accident  \
0                 0.988444           -0.333841      -0.415074
1                 0.122364           -0.333841      -0.415074
2                 0.968303            0.351130      -0.415074
3                 0.263354           -1.018812      -0.415074
4                 1.632969            0.351130      -0.415074
...                    ...                 ...            ...
4495             -1.086120           -0.333841      -0.415074
4496              0.766889            0.351130      -0.415074
4497             -1.388242            0.351130      -0.415074
4498              1.491980            0.351130      -0.415074
4499              0.787030           -1.018812      -0.415074

      promotion_last_5years      dept    salary
0                  -0.147986  0.400804 -0.930457
```

```
1                -0.147986  0.400804 -0.930457
2                -0.147986  1.097153  0.639344
3                -0.147986 -0.295546 -0.930457
4                -0.147986  0.400804 -0.930457
...                    ...       ...       ...
4495             -0.147986  0.748978 -0.930457
4496             -0.147986  0.400804  0.639344
4497             -0.147986 -2.036419  0.639344
4498             -0.147986  0.748978  0.639344
4499             -0.147986 -1.688244 -0.930457

[4500 rows x 9 columns]
```