الجمهورية اليمنية

جامعة إب

كلية العلوم

قسم علوم الحاسوب وتقنية المعلومات

تكليف مقرر

تنقيب بيانات - عملي

Data Mining

المحاضرة السابعة

عمل الطالب :

أسامة سعيد محمد حمود سعيد - مجموعة A

إشراف :

أ. مالك المصنف

2024 - 2025

مقارنة بين أنواع التشفير Encoding

وجه المقارنة / نوع التشفير	المفهوم	الفوائد	العيوب	الاستخدامات	متى تكون مفيدة
Ordinal Encoding **التشفير الترتيبي**	استبدال الفئات بأرقام مرتبة بناءً على تسلسل معين	بسيط وسهل التنفيذ، مناسب للبيانات المرتبة	قد يضيف دلالة رقمية خاطئة إذا لم تكن القيم مرتبة بشكل طبيعي	الميزات الترتيبية مثل تقييمات العملاء (من 1 إلى 5 نجوم)	إذا كانت البيانات مرتبة
Frequency Encoding **التشفير التكراري**	استبدال كل فئة بعدد مرات تكرارها في البيانات	يقلل من عدد القيم الفريدة، يساعد في التنبؤ بناءً على انتشار الفئة	قد يكون مضللًا إذا لم يكن التوزيع طبيعيًا	عندما يكون تكرار الفئة مؤشرًا على الأهمية (مثل تصنيف المنتجات الأكثر مبيعًا)	إذا كان تكرار الفئة مهمًا
Target Encoding **التشفير المستهدف**	استبدال كل فئة بقيمة إحصائية مشتقة من الهدف (مثل المتوسط)	يحافظ على العلاقة بين الفئات والهدف، فعال مع النماذج الخطية	عرضة للتسريب إذا لم يُطبق بطريقة صحيحة	مفيد في المسابقات التنبؤية وفي مشاكل التصنيف ذات الفئات العديدة	إذا كانت الفئات مرتبطة مباشرة بالهدف
Binary Encoding **التشفير الثنائي**	تحويل القيم الفئوية إلى تمثيل ثنائي يقلل من الأبعاد	يحسن الأداء مقارنة بـ One-Hot Encoding، مناسب للبيانات الكبيرة	صعب التفسير مقارنة بأساليب التشفير الأخرى	البيانات الفئوية ذات العدد الكبير من الفئات	إذا كنت بحاجة إلى تقليل الأبعاد مع الحفاظ على الفعالية
Embedded Encoding **التشفير المضمن**	استخدام نماذج تعلم عميق لاستخراج تمثيلات عديدة للفئات	يقدم تمثيلات مضغوطة وعالية الدقة	يحتاج إلى تدريب إضافي ويعتمد على جودة النموذج	معالجة البيانات الفئوية في الشبكات العصبية	إذا كنت تستخدم الشبكات العصبية وتريد تمثيلات أكثر تقدمًا

Read Data

```
import pandas as pd

dataset = pd.read_csv('Employee-Attrition.csv')

dataset
```

```
      Age Attrition     BusinessTravel  DailyRate
Department  \
0      41       Yes      Travel_Rarely       1102
Sales
1      49        No  Travel_Frequently        279  Research &
Development
2      37       Yes      Travel_Rarely       1373  Research &
Development
3      33        No  Travel_Frequently       1392  Research &
Development
4      27        No      Travel_Rarely        591  Research &
Development
...   ...       ...                ...        ...
...
1465   36        No  Travel_Frequently        884  Research &
Development
1466   39        No      Travel_Rarely        613  Research &
Development
1467   27        No      Travel_Rarely        155  Research &
Development
1468   49        No  Travel_Frequently       1023
Sales
1469   34        No      Travel_Rarely        628  Research &
Development

      DistanceFromHome  Education EducationField  EmployeeCount  \
0                    1          2  Life Sciences              1
1                    8          1  Life Sciences              1
2                    2          2          Other              1
3                    3          4  Life Sciences              1
4                    2          1        Medical              1
...                ...        ...            ...            ...
1465                23          2        Medical              1
1466                 6          1        Medical              1
1467                 4          3  Life Sciences              1
1468                 2          3        Medical              1
1469                 8          3        Medical              1

      EmployeeNumber  ...  RelationshipSatisfaction StandardHours  \
0                  1  ...                         1            80
```

```
1                  2  ...                           4            80
2                  4  ...                           2            80
3                  5  ...                           3            80
4                  7  ...                           4            80
...              ...  ...                         ...           ...
1465            2061  ...                           3            80
1466            2062  ...                           1            80
1467            2064  ...                           2            80
1468            2065  ...                           4            80
1469            2068  ...                           1            80

      StockOptionLevel  TotalWorkingYears  TrainingTimesLastYear  \
0                    0                  8                      0
1                    1                 10                      3
2                    0                  7                      3
3                    0                  8                      3
4                    1                  6                      3
...                ...                ...                    ...
1465                 1                 17                      3
1466                 1                  9                      5
1467                 1                  6                      0
1468                 0                 17                      3
1469                 0                  6                      3

     WorkLifeBalance  YearsAtCompany YearsInCurrentRole  \
0                  1               6                  4
1                  3              10                  7
2                  3               0                  0
3                  3               8                  7
4                  3               2                  2
...              ...             ...                ...
1465               3               5                  2
1466               3               7                  7
1467               3               6                  2
1468               2               9                  6
1469               4               4                  3

      YearsSinceLastPromotion  YearsWithCurrManager
0                           0                     5
1                           1                     7
2                           0                     0
3                           3                     0
4                           2                     2
...                       ...                   ...
1465                        0                     3
1466                        1                     7
1467                        0                     3
1468                        0                     8
1469                        1                     2
```

```
[1470 rows x 35 columns]
```

Information about my Data

```
dataset.shape
```

```
(1470, 35)
```

```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 35 columns):
 #   Column                    Non-Null Count  Dtype 
---  ------                    --------------  ----- 
 0   Age                       1470 non-null   int64 
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64 
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64 
 6   Education                 1470 non-null   int64 
 7   EducationField            1470 non-null   object
 8   EmployeeCount             1470 non-null   int64 
 9   EmployeeNumber            1470 non-null   int64 
 10  EnvironmentSatisfaction   1470 non-null   int64 
 11  Gender                    1470 non-null   object
 12  HourlyRate                1470 non-null   int64 
 13  JobInvolvement            1470 non-null   int64 
 14  JobLevel                  1470 non-null   int64 
 15  JobRole                   1470 non-null   object
 16  JobSatisfaction           1470 non-null   int64 
 17  MaritalStatus             1470 non-null   object
 18  MonthlyIncome             1470 non-null   int64 
 19  MonthlyRate               1470 non-null   int64 
 20  NumCompaniesWorked        1470 non-null   int64 
 21  Over18                    1470 non-null   object
 22  OverTime                  1470 non-null   object
 23  PercentSalaryHike         1470 non-null   int64 
 24  PerformanceRating         1470 non-null   int64 
 25  RelationshipSatisfaction  1470 non-null   int64 
 26  StandardHours             1470 non-null   int64 
 27  StockOptionLevel          1470 non-null   int64 
 28  TotalWorkingYears         1470 non-null   int64 
 29  TrainingTimesLastYear     1470 non-null   int64 
 30  WorkLifeBalance           1470 non-null   int64 
 31  YearsAtCompany            1470 non-null   int64 
```

```
 32  YearsInCurrentRole        1470 non-null   int64
 33  YearsSinceLastPromotion   1470 non-null   int64
 34  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(26), object(9)
memory usage: 402.1+ KB
```

```
dataset.describe()
```

```
               Age    DailyRate  DistanceFromHome    Education
EmployeeCount  \
count  1470.000000  1470.000000       1470.000000  1470.000000
1470.0
mean     36.923810   802.485714          9.192517     2.912925
1.0
std       9.135373   403.509100          8.106864     1.024165
0.0
min      18.000000   102.000000          1.000000     1.000000
1.0
25%      30.000000   465.000000          2.000000     2.000000
1.0
50%      36.000000   802.000000          7.000000     3.000000
1.0
75%      43.000000  1157.000000         14.000000     4.000000
1.0
max      60.000000  1499.000000         29.000000     5.000000
1.0

       EmployeeNumber  EnvironmentSatisfaction   HourlyRate
JobInvolvement  \
count     1470.000000              1470.000000  1470.000000
1470.000000
mean      1024.865306                 2.721769    65.891156
2.729932
std        602.024335                 1.093082    20.329428
0.711561
min          1.000000                 1.000000    30.000000
1.000000
25%        491.250000                 2.000000    48.000000
2.000000
50%       1020.500000                 3.000000    66.000000
3.000000
75%       1555.750000                 4.000000    83.750000
3.000000
max       2068.000000                 4.000000   100.000000
4.000000

          JobLevel  ...  RelationshipSatisfaction  StandardHours  \
count  1470.000000  ...               1470.000000         1470.0
mean      2.063946  ...                  2.712245           80.0
std       1.106940  ...                  1.081209            0.0
```

```
min       1.000000  ...                  1.000000          80.0
25%       1.000000  ...                  2.000000          80.0
50%       2.000000  ...                  3.000000          80.0
75%       3.000000  ...                  4.000000          80.0
max       5.000000  ...                  4.000000          80.0

       StockOptionLevel  TotalWorkingYears  TrainingTimesLastYear  \
count       1470.000000        1470.000000            1470.000000
mean           0.793878          11.279592               2.799320
std            0.852077           7.780782               1.289271
min            0.000000           0.000000               0.000000
25%            0.000000           6.000000               2.000000
50%            1.000000          10.000000               3.000000
75%            1.000000          15.000000               3.000000
max            3.000000          40.000000               6.000000

       WorkLifeBalance  YearsAtCompany  YearsInCurrentRole  \
count      1470.000000     1470.000000         1470.000000
mean          2.761224        7.008163            4.229252
std           0.706476        6.126525            3.623137
min           1.000000        0.000000            0.000000
25%           2.000000        3.000000            2.000000
50%           3.000000        5.000000            3.000000
75%           3.000000        9.000000            7.000000
max           4.000000       40.000000           18.000000

       YearsSinceLastPromotion  YearsWithCurrManager
count              1470.000000           1470.000000
mean                  2.187755              4.123129
std                   3.222430              3.568136
min                   0.000000              0.000000
25%                   0.000000              2.000000
50%                   1.000000              3.000000
75%                   3.000000              7.000000
max                  15.000000             17.000000

[8 rows x 26 columns]
```

```
dataset.head()
```

```
   Age Attrition     BusinessTravel  DailyRate              Department
\
0   41       Yes      Travel_Rarely       1102                   Sales

1   49        No  Travel_Frequently        279  Research & Development

2   37       Yes      Travel_Rarely       1373  Research & Development

3   33        No  Travel_Frequently       1392  Research & Development
```

```
4   27        No      Travel_Rarely       591  Research & Development


   DistanceFromHome  Education EducationField  EmployeeCount
EmployeeNumber  \
0                 1          2  Life Sciences              1
1
1                 8          1  Life Sciences              1
2
2                 2          2          Other              1
4
3                 3          4  Life Sciences              1
5
4                 2          1        Medical              1
7

   ...  RelationshipSatisfaction StandardHours  StockOptionLevel  \
0  ...                         1            80                 0
1  ...                         4            80                 1
2  ...                         2            80                 0
3  ...                         3            80                 0
4  ...                         4            80                 1

   TotalWorkingYears  TrainingTimesLastYear WorkLifeBalance
YearsAtCompany  \
0                  8                      0               1
6
1                 10                      3               3
10
2                  7                      3               3
0
3                  8                      3               3
8
4                  6                      3               3
2

  YearsInCurrentRole  YearsSinceLastPromotion  YearsWithCurrManager
0                  4                        0                     5
1                  7                        1                     7
2                  0                        0                     0
3                  7                        3                     0
4                  2                        2                     2

[5 rows x 35 columns]
```

```
dataset.tail()
```

```
      Age Attrition     BusinessTravel  DailyRate
Department  \
1465   36        No  Travel_Frequently       884  Research &
```

```
Development
1466   39        No      Travel_Rarely        613  Research &
Development
1467   27        No      Travel_Rarely        155  Research &
Development
1468   49        No  Travel_Frequently       1023
Sales
1469   34        No      Travel_Rarely        628  Research &
Development

      DistanceFromHome  Education EducationField  EmployeeCount  \
1465                23          2        Medical              1
1466                 6          1        Medical              1
1467                 4          3  Life Sciences              1
1468                 2          3        Medical              1
1469                 8          3        Medical              1

      EmployeeNumber  ...  RelationshipSatisfaction StandardHours  \
1465            2061  ...                         3            80
1466            2062  ...                         1            80
1467            2064  ...                         2            80
1468            2065  ...                         4            80
1469            2068  ...                         1            80

      StockOptionLevel  TotalWorkingYears  TrainingTimesLastYear  \
1465                 1                 17                      3
1466                 1                  9                      5
1467                 1                  6                      0
1468                 0                 17                      3
1469                 0                  6                      3

     WorkLifeBalance  YearsAtCompany YearsInCurrentRole  \
1465               3               5                  2
1466               3               7                  7
1467               3               6                  2
1468               2               9                  6
1469               4               4                  3

      YearsSinceLastPromotion  YearsWithCurrManager
1465                        0                     3
1466                        1                     7
1467                        0                     3
1468                        0                     8
1469                        1                     2

[5 rows x 35 columns]
```

```
dataset.columns
```

```
Index(['Age', 'Attrition', 'BusinessTravel', 'DailyRate',
'Department',
       'DistanceFromHome', 'Education', 'EducationField',
'EmployeeCount',
       'EmployeeNumber', 'EnvironmentSatisfaction', 'Gender',
'HourlyRate',
       'JobInvolvement', 'JobLevel', 'JobRole', 'JobSatisfaction',
       'MaritalStatus', 'MonthlyIncome', 'MonthlyRate',
'NumCompaniesWorked',
       'Over18', 'OverTime', 'PercentSalaryHike', 'PerformanceRating',
       'RelationshipSatisfaction', 'StandardHours',
'StockOptionLevel',
       'TotalWorkingYears', 'TrainingTimesLastYear',
'WorkLifeBalance',
       'YearsAtCompany', 'YearsInCurrentRole',
'YearsSinceLastPromotion',
       'YearsWithCurrManager'],
      dtype='object')
```

```
dataset.EmployeeCount.unique()
```

```
array([1], dtype=int64)
```

```
dataset.Over18.unique()
```

```
array(['Y'], dtype=object)
```

```
dataset.StandardHours.unique()
```

```
array([80], dtype=int64)
```

```
dataset.Age.unique()
```

```
array([41, 49, 37, 33, 27, 32, 59, 30, 38, 36, 35, 29, 31, 34, 28, 22,
53,
       24, 21, 42, 44, 46, 39, 43, 50, 26, 48, 55, 45, 56, 23, 51, 40,
54,
       58, 20, 25, 19, 57, 52, 47, 18, 60], dtype=int64)
```

```
dataset.Attrition.unique()
```

```
array(['Yes', 'No'], dtype=object)
```

```
dataset.BusinessTravel.unique()
```

```
array(['Travel_Rarely', 'Travel_Frequently', 'Non-Travel'],
dtype=object)
```

```
dataset.DailyRate.unique()
```

```
array([1102,  279, 1373, 1392,  591, 1005, 1324, 1358,  216, 1299,
809,
```

```
        153,  670, 1346,  103, 1389,  334, 1123, 1219,  371,  673,
1218,
        419,  391,  699, 1282, 1125,  691,  477,  705,  924, 1459,
125,
        895,  813, 1273,  869,  890,  852, 1141,  464, 1240, 1357,
994,
        721, 1360, 1065,  408, 1211, 1229,  626, 1434, 1488, 1097,
1443,
        515,  853, 1142,  655, 1115,  427,  653,  989, 1435, 1223,
836,
       1195, 1339,  664,  318, 1225, 1328, 1082,  548,  132,  746,
776,
        193,  397,  945, 1214,  111,  573, 1153, 1400,  541,  432,
288,
        669,  530,  632, 1334,  638, 1093, 1217, 1353,  120,  682,
489,
        807,  827,  871,  665, 1040, 1420,  240, 1280,  534, 1456,
658,
        142, 1127, 1031, 1189, 1354, 1467,  922,  394, 1312,  750,
441,
        684,  249,  841,  147,  528,  594,  470,  957,  542,  802,
1355,
       1150, 1329,  959, 1033, 1316,  364,  438,  689,  201, 1427,
857,
        933, 1181, 1395,  662, 1436,  194,  967, 1496, 1169, 1145,
630,
        303, 1256,  440, 1450, 1452,  465,  702, 1157,  602, 1480,
1268,
        713,  134,  526, 1380,  140,  629, 1356,  328, 1084,  931,
692,
       1069,  313,  894,  556, 1344,  290,  138,  926, 1261,  472,
1002,
        878,  905, 1180,  121, 1136,  635, 1151,  644, 1045,  829,
1242,
       1469,  896,  992, 1052, 1147, 1396,  663,  119,  979,  319,
1413,
        944, 1323,  532,  818,  854, 1034,  771, 1401, 1431,  976,
1411,
       1300,  252, 1327,  832, 1017, 1199,  504,  505,  916, 1247,
685,
        269, 1416,  833,  307, 1311,  128,  488,  529, 1210, 1463,
675,
       1385, 1403,  452,  666, 1158,  228,  996,  728, 1315,  322,
1479,
        797, 1070,  442,  496, 1372,  920,  688, 1449, 1117,  636,
506,
        444,  950,  889,  555,  230, 1232,  566, 1302,  812, 1476,
218,
       1132, 1105,  906,  849,  390,  106, 1249,  192,  553,  117,
```

```
185,
        1091,  723, 1220,  588, 1377, 1018, 1275,  798,  672, 1162,
508,
        1482,  559,  210,  928, 1001,  549, 1124,  738,  570, 1130,
1192,
         343,  144, 1296, 1309,  483,  810,  544, 1062, 1319,  641,
1332,
         756,  845,  593, 1171,  350,  921, 1144,  143, 1046,  575,
156,
        1283,  755,  304, 1178,  329, 1362, 1371,  202,  253,  164,
1107,
         759, 1305,  982,  821, 1381,  480, 1473,  891, 1063,  645,
1490,
         317,  422, 1485, 1368, 1448,  296, 1398, 1349,  986, 1099,
1116,
        1499,  983, 1009, 1303, 1274, 1277,  587,  413, 1276,  988,
1474,
         163,  267,  619,  302,  443,  828,  561,  426,  232, 1306,
1094,
         509,  775,  195,  258,  471,  799,  956,  535, 1495,  446,
1245,
         703,  823, 1246,  622, 1287,  448,  254, 1365,  538,  525,
558,
         782,  362, 1236, 1112,  204, 1343,  604, 1216,  646,  160,
238,
        1397,  306,  991,  482, 1176,  913, 1076,  727,  885,  243,
806,
         817, 1410, 1207, 1442,  693,  929,  562,  608,  580,  970,
1179,
         294,  314,  316,  654,  168,  381,  217,  501,  650,  141,
804,
         975, 1090,  346,  430,  268,  167,  621,  527,  883,  954,
310,
         719,  725,  715,  657, 1146,  182,  376,  571,  384,  791,
1111,
        1243, 1092, 1325,  805,  213,  118,  676, 1252,  286, 1258,
932,
        1041,  859,  720,  946, 1184,  436,  589,  760,  887, 1318,
625,
         180,  586, 1012,  661,  930,  342, 1230, 1271, 1278,  607,
130,
         300,  583, 1418, 1269,  379,  395, 1265, 1222,  341,  868,
1231,
         102,  881, 1383, 1075,  374, 1086,  781,  177,  500, 1425,
1454,
         617, 1085,  995, 1122,  618,  546,  462, 1198, 1272,  154,
1137,
        1188,  188, 1333,  867,  263,  938,  129,  616,  498, 1404,
1053,
```

```
       289, 1376,  231,  152,  882,  903, 1379,  335,  722,  461,
974,
      1126,  840, 1134,  248,  955,  939, 1391, 1206,  287, 1441,
109,
      1066,  277,  466, 1055,  265,  135,  247, 1035,  266,  145,
1038,
      1234, 1109, 1089,  788,  124,  660, 1186, 1464,  796,  415,
769,
      1003, 1366,  330, 1492, 1204,  309, 1330,  469,  697, 1262,
1050,
       770,  406,  203, 1308,  984,  439,  793, 1451, 1182,  174,
490,
       718,  433,  773,  603,  874,  367,  199,  481,  647, 1384,
902,
       819,  862, 1457,  977,  942, 1402, 1421, 1361,  917,  200,
150,
       179,  696,  116,  363,  107, 1465,  458, 1212, 1103,  966,
1010,
       326, 1098,  969, 1167,  694, 1320,  536,  373,  599,  251,
131,
       237, 1429,  648,  735,  531,  429,  968,  879,  640,  412,
848,
       360, 1138,  325, 1322,  299, 1030,  634,  524,  256, 1060,
935,
       495,  282,  206,  943,  523,  507,  601,  855, 1291, 1405,
1369,
       999, 1202,  285,  404,  736, 1498, 1200, 1439,  499,  205,
683,
      1462,  949,  652,  332, 1475,  337,  971, 1174,  667,  560,
172,
       383, 1255,  359,  401,  377,  592, 1445, 1221,  866,  981,
447,
      1326,  748,  990,  405,  115,  790,  830, 1193, 1423,  467,
271,
       410, 1083,  516,  224,  136, 1029,  333, 1440,  674, 1342,
898,
       824,  492,  598,  740,  888, 1288,  104, 1108,  479, 1351,
474,
       437,  884, 1370,  264, 1059,  563,  457, 1313,  241, 1015,
336,
      1387,  170,  208,  671,  711,  737, 1470,  365,  763,  567,
486,
       772,  301,  311,  584,  880,  392,  148,  708, 1259,  786,
370,
       678,  146,  581,  918, 1238,  585,  741,  552,  369,  717,
543,
       964,  792,  611,  176,  897,  600, 1054,  428,  181,  211,
1079,
       590,  305,  953,  478, 1375,  244,  511, 1294,  196,  734,
```

```
1239,
        1253, 1128, 1336,  234,  766,  261, 1194,  431,  572, 1422,
1297,
         574,  355,  207,  706,  280,  726,  414,  352, 1224,  459,
1254,
        1131,  835, 1172, 1266,  783,  219, 1213, 1096, 1251, 1394,
605,
        1064, 1337,  937,  157,  754, 1168,  155, 1444,  189,  911,
1321,
        1154,  557,  642,  801,  161, 1382, 1037,  105,  582,  704,
345,
        1120, 1378,  468,  613, 1023,  628], dtype=int64)
```

```
dataset.Department.unique()
```

```
array(['Sales', 'Research & Development', 'Human Resources'],
dtype=object)
```

```
dataset.DistanceFromHome.unique()
```

```
array([ 1,  8,  2,  3, 24, 23, 27, 16, 15, 26, 19, 21,  5, 11,  9,  7,
6,
        10,  4, 25, 12, 18, 29, 22, 14, 20, 28, 17, 13], dtype=int64)
```

```
dataset.Education.unique()
```

```
array([2, 1, 4, 3, 5], dtype=int64)
```

```
dataset.EducationField.unique()
```

```
array(['Life Sciences', 'Other', 'Medical', 'Marketing',
       'Technical Degree', 'Human Resources'], dtype=object)
```

```
dataset.EnvironmentSatisfaction.unique()
```

```
array([2, 3, 4, 1], dtype=int64)
```

```
dataset.Gender.unique()
```

```
array(['Female', 'Male'], dtype=object)
```

```
dataset.HourlyRate.unique()
```

```
array([ 94,  61,  92,  56,  40,  79,  81,  67,  44,  84,  49,  31,
93,
         50,  51,  80,  96,  78,  45,  82,  53,  83,  58,  72,  48,
42,
         41,  86,  97,  75,  33,  37,  73,  98,  36,  47,  71,  30,
43,
         99,  59,  95,  57,  76,  87,  66,  55,  32,  52,  70,  62,
64,
         63,  60, 100,  46,  39,  77,  35,  91,  54,  34,  90,  65,
```

```
88,
        85,  89,  68,  69,  74,  38], dtype=int64)

dataset.JobInvolvement.unique()

array([3, 2, 4, 1], dtype=int64)

dataset.JobLevel.unique()

array([2, 1, 3, 4, 5], dtype=int64)

dataset.JobRole.unique()

array(['Sales Executive', 'Research Scientist', 'Laboratory
Technician',
       'Manufacturing Director', 'Healthcare Representative',
'Manager',
       'Sales Representative', 'Research Director', 'Human
Resources'],
      dtype=object)

dataset.JobSatisfaction.unique()

array([4, 2, 3, 1], dtype=int64)

dataset.MaritalStatus.unique()

array(['Single', 'Married', 'Divorced'], dtype=object)

dataset.MonthlyIncome.unique()

array([5993, 5130, 2090, ..., 9991, 5390, 4404], dtype=int64)

dataset.MonthlyRate.unique()

array([19479, 24907,  2396, ...,  5174, 13243, 10228], dtype=int64)

dataset.NumCompaniesWorked.unique()

array([8, 1, 6, 9, 0, 4, 5, 2, 7, 3], dtype=int64)

dataset.OverTime.unique()

array(['Yes', 'No'], dtype=object)

dataset.PercentSalaryHike.unique()

array([11, 23, 15, 12, 13, 20, 22, 21, 17, 14, 16, 18, 19, 24, 25],
      dtype=int64)

dataset.PerformanceRating.unique()

array([3, 4], dtype=int64)
```

```
dataset.RelationshipSatisfaction.unique()

array([1, 4, 2, 3], dtype=int64)

dataset.StockOptionLevel.unique()

array([0, 1, 3, 2], dtype=int64)

dataset.TotalWorkingYears.unique()

array([ 8, 10,  7,  6, 12,  1, 17,  5,  3, 31, 13,  0, 26, 24, 22,  9,
19,
        2, 23, 14, 15,  4, 29, 28, 21, 25, 20, 11, 16, 37, 38, 30, 40,
18,
       36, 34, 32, 33, 35, 27], dtype=int64)

dataset.TrainingTimesLastYear.unique()

array([0, 3, 2, 5, 1, 4, 6], dtype=int64)

dataset.WorkLifeBalance.unique()

array([1, 3, 2, 4], dtype=int64)

dataset.YearsAtCompany.unique()

array([ 6, 10,  0,  8,  2,  7,  1,  9,  5,  4, 25,  3, 12, 14, 22, 15,
27,
       21, 17, 11, 13, 37, 16, 20, 40, 24, 33, 19, 36, 18, 29, 31, 32,
34,
       26, 30, 23], dtype=int64)

dataset.YearsInCurrentRole.unique()

array([ 4,  7,  0,  2,  5,  9,  8,  3,  6, 13,  1, 15, 14, 16, 11, 10,
12,
       18, 17], dtype=int64)

dataset.YearsSinceLastPromotion.unique()

array([ 0,  1,  3,  2,  7,  4,  8,  6,  5, 15,  9, 13, 12, 10, 11,
14],
      dtype=int64)

dataset.YearsWithCurrManager.unique()

array([ 5,  7,  0,  2,  6,  8,  3, 11, 17,  1,  4, 12,  9, 10, 15, 13,
16,
       14], dtype=int64)
```

Remove Columns

EmployeeCount: دائمًا قيمته 1، غير مفيد.

EmployeeNumber: رقم تعريفي للموظف، لا علاقة له بالتحليل.

Over18: جميع القيم Yes، غير مفيد.

StandardHours: جميع القيم 40، غير مفيد.

```
dataset =
dataset.drop(['EmployeeCount','EmployeeNumber','Over18','StandardHours
'],axis=1)

dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 31 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   Attrition                 1470 non-null   object
 2   BusinessTravel            1470 non-null   object
 3   DailyRate                 1470 non-null   int64
 4   Department                1470 non-null   object
 5   DistanceFromHome          1470 non-null   int64
 6   Education                 1470 non-null   int64
 7   EducationField            1470 non-null   object
 8   EnvironmentSatisfaction   1470 non-null   int64
 9   Gender                    1470 non-null   object
 10  HourlyRate                1470 non-null   int64
 11  JobInvolvement            1470 non-null   int64
 12  JobLevel                  1470 non-null   int64
 13  JobRole                   1470 non-null   object
 14  JobSatisfaction           1470 non-null   int64
 15  MaritalStatus             1470 non-null   object
 16  MonthlyIncome             1470 non-null   int64
 17  MonthlyRate               1470 non-null   int64
 18  NumCompaniesWorked        1470 non-null   int64
 19  OverTime                  1470 non-null   object
 20  PercentSalaryHike         1470 non-null   int64
 21  PerformanceRating         1470 non-null   int64
 22  RelationshipSatisfaction  1470 non-null   int64
 23  StockOptionLevel          1470 non-null   int64
 24  TotalWorkingYears         1470 non-null   int64
 25  TrainingTimesLastYear     1470 non-null   int64
 26  WorkLifeBalance           1470 non-null   int64
 27  YearsAtCompany            1470 non-null   int64
```

```
 28  YearsInCurrentRole        1470 non-null   int64
 29  YearsSinceLastPromotion   1470 non-null   int64
 30  YearsWithCurrManager      1470 non-null   int64
dtypes: int64(23), object(8)
memory usage: 356.1+ KB
```

Move Target Column in the Last Table

```
dataset = dataset[[col for col in dataset.columns if col !=
'Attrition'] + ['Attrition']]

dataset.head()
```

```
   Age     BusinessTravel  DailyRate              Department  \
0   41      Travel_Rarely       1102                   Sales
1   49  Travel_Frequently        279  Research & Development
2   37      Travel_Rarely       1373  Research & Development
3   33  Travel_Frequently       1392  Research & Development
4   27      Travel_Rarely        591  Research & Development

   DistanceFromHome  Education EducationField  EnvironmentSatisfaction
\
0                 1          2  Life Sciences                        2

1                 8          1  Life Sciences                        3

2                 2          2          Other                        4

3                 3          4  Life Sciences                        4

4                 2          1        Medical                        1


   Gender  HourlyRate  ...  RelationshipSatisfaction  StockOptionLevel
\
0  Female          94  ...                         1                 0

1    Male          61  ...                         4                 1

2    Male          92  ...                         2                 0

3  Female          56  ...                         3                 0

4    Male          40  ...                         4                 1


  TotalWorkingYears  TrainingTimesLastYear WorkLifeBalance
YearsAtCompany  \
0                 8                      0               1
```

```
6
1                10                       3               3
10
2                 7                       3               3
0
3                 8                       3               3
8
4                 6                       3               3
2

   YearsInCurrentRole  YearsSinceLastPromotion YearsWithCurrManager
Attrition
0                   4                        0                    5
Yes
1                   7                        1                    7
No
2                   0                        0                    0
Yes
3                   7                        3                    0
No
4                   2                        2                    2
No

[5 rows x 31 columns]
```

```python
from sklearn.preprocessing import LabelEncoder

encoder = LabelEncoder()

categorical_columns = ["BusinessTravel", "Department", 
"EducationField", "Gender", "JobRole", "MaritalStatus", "OverTime"]

for col in categorical_columns:
    dataset[col] = encoder.fit_transform(dataset[col])

dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1470 entries, 0 to 1469
Data columns (total 31 columns):
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Age                       1470 non-null   int64
 1   BusinessTravel            1470 non-null   int32
 2   DailyRate                 1470 non-null   int64
 3   Department                1470 non-null   int32
 4   DistanceFromHome          1470 non-null   int64
 5   Education                 1470 non-null   int64
 6   EducationField            1470 non-null   int32
 7   EnvironmentSatisfaction   1470 non-null   int64
```

```
 8   Gender                    1470 non-null   int32
 9   HourlyRate                1470 non-null   int64
 10  JobInvolvement            1470 non-null   int64
 11  JobLevel                  1470 non-null   int64
 12  JobRole                   1470 non-null   int32
 13  JobSatisfaction           1470 non-null   int64
 14  MaritalStatus             1470 non-null   int32
 15  MonthlyIncome             1470 non-null   int64
 16  MonthlyRate               1470 non-null   int64
 17  NumCompaniesWorked        1470 non-null   int64
 18  OverTime                  1470 non-null   int32
 19  PercentSalaryHike         1470 non-null   int64
 20  PerformanceRating         1470 non-null   int64
 21  RelationshipSatisfaction  1470 non-null   int64
 22  StockOptionLevel          1470 non-null   int64
 23  TotalWorkingYears         1470 non-null   int64
 24  TrainingTimesLastYear     1470 non-null   int64
 25  WorkLifeBalance           1470 non-null   int64
 26  YearsAtCompany            1470 non-null   int64
 27  YearsInCurrentRole        1470 non-null   int64
 28  YearsSinceLastPromotion   1470 non-null   int64
 29  YearsWithCurrManager      1470 non-null   int64
 30  Attrition                 1470 non-null   object
dtypes: int32(7), int64(23), object(1)
memory usage: 315.9+ KB
```

Split Data => Features and Target

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

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

Feature Selection

```
from sklearn.feature_selection import SelectKBest , f_classif

FeatureSelection = SelectKBest(score_func = f_classif , k = 6)

X.shape
```

```
(1470, 30)
```

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

```
array([False, False, False, False, False, False, False, False, False,
       False, False,  True, False, False,  True,  True, False, False,
```

```
        True, False, False, False, False,  True, False, False, False,
        True, False, False])

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

X_best.shape

(1470, 6)

X_best.head()

   JobLevel  MaritalStatus  MonthlyIncome  OverTime  TotalWorkingYears  \
0         2              2           5993         1                  8

1         2              1           5130         0                 10

2         1              2           2090         1                  7

3         1              1           2909         1                  8

4         1              1           3468         0                  6


   YearsInCurrentRole
0                   4
1                   7
2                   0
3                   7
4                   2

Y.shape

(1470,)

Y.info()

<class 'pandas.core.series.Series'>
RangeIndex: 1470 entries, 0 to 1469
Series name: Attrition
Non-Null Count  Dtype
--------------  -----
1470 non-null   object
dtypes: object(1)
memory usage: 11.6+ KB

Y = encoder.fit_transform(Y)
Y = pd.Series(Y, name="Attrition")

Y.info()
```

```
<class 'pandas.core.series.Series'>
RangeIndex: 1470 entries, 0 to 1469
Series name: Attrition
Non-Null Count  Dtype
--------------  -----
1470 non-null   int32
dtypes: int32(1)
memory usage: 5.9 KB
```

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=0.20,
random_state=20, shuffle =True)
X_train
```

```
      JobLevel  MaritalStatus  MonthlyIncome  OverTime
TotalWorkingYears  \
784          3              1           8823         0
20
1032         1              2           3646         1
11
623          1              0           3761         0
10
1347         2              2           3886         0
10
585          1              1           1601         1
1
...        ...            ...            ...       ...
...
924          1              1           3506         1
4
1247         2              1           8346         0
6
271          3              1          11849         1
10
474          1              1           2725         1
6
1379         1              1           2863         0
1

      YearsInCurrentRole
784                    9
1032                   0
623                    4
1347                   1
585                    0
```

```
...                     ...
924                       2
1247                      2
271                       7
474                       5
1379                      0

[1176 rows x 6 columns]
```

X_test

```
      JobLevel  MaritalStatus  MonthlyIncome  OverTime  
TotalWorkingYears  \
1261         2              1           5811         1  
15   
434          3              0          10648         0  
13   
313          3              1          11878         0  
12   
1182         2              1           4374         0  
4    
446          2              2           6230         0  
16   
...        ...            ...            ...       ...  
...   
1059         1              1           2404         0  
1    
1374         4              1          17875         1  
29   
326          5              1          19272         0  
21   
283          2              1           5415         1  
12   
855          2              1           6474         0  
14   

      YearsInCurrentRole  
1261                   0  
434                    8  
313                    6  
1182                   2  
446                    3  
...                  ...  
1059                   0  
1374                   0  
326                    9  
283                    7  
855                    8  

[294 rows x 6 columns]
```

Data Scaling

MinMax Scaling

```
from sklearn.preprocessing import MinMaxScaler

Scaler=MinMaxScaler(feature_range = (2,6))

X_train, X_test, y_train, y_test = tts(X_best, Y, test_size=0.20, 
random_state=20, 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
```

```
      JobLevel  MaritalStatus  MonthlyIncome  OverTime  
TotalWorkingYears  \
0          4.0            4.0       3.645919       2.0  
4.0
1          2.0            6.0       2.555450       6.0  
3.1
2          2.0            2.0       2.579674       2.0  
3.0
3          3.0            6.0       2.606003       2.0  
3.0
4          2.0            4.0       2.124697       6.0  
2.1
...        ...            ...            ...       ...  
...
1171       2.0            4.0       2.525961       6.0  
2.4
1172       3.0            4.0       3.545445       2.0  
2.6
1173       4.0            4.0       4.283307       6.0  
3.0
1174       2.0            4.0       2.361453       6.0  
2.6
1175       2.0            4.0       2.390521       2.0  
2.1

      YearsInCurrentRole  
0               4.000000  
1               2.000000  
2               2.888889  
3               2.222222  
4               2.000000  
...                  ...  
1171            2.444444  
```

```
1172            2.444444
1173            3.555556
1174            3.111111
1175            2.000000

[1176 rows x 6 columns]
```

X_test_Scaled

```
     JobLevel  MaritalStatus  MonthlyIncome  OverTime
TotalWorkingYears  \
0         3.0            4.0       3.011480       6.0
3.5
1         4.0            2.0       4.030332       2.0
3.3
2         4.0            4.0       4.289415       2.0
3.2
3         3.0            4.0       2.708794       2.0
2.4
4         3.0            6.0       3.099737       2.0
3.6
..        ...            ...            ...       ...                .
..
289       2.0            4.0       2.293839       2.0
2.1
290       5.0            4.0       5.552607       6.0
4.9
291       6.0            4.0       5.846867       2.0
4.1
292       3.0            4.0       2.928067       6.0
3.2
293       3.0            4.0       3.151132       2.0
3.4

     YearsInCurrentRole
0              2.000000
1              3.777778
2              3.333333
3              2.444444
4              2.666667
..                  ...
289            2.000000
290            2.000000
291            4.000000
292            3.555556
293            3.777778

[294 rows x 6 columns]
```