

# 員工離職預測

許雅淑

# 目錄content

資料清洗

特徵選擇

模型建立

離職預測

# 資料清洗

# ● 描述性分析

查看前 5 行資料

	yyyy	PerNo	PerStatus	sex	工作分類	職等	廠區代碼	管理層級	工作資歷1	工作資歷2	...	年資層級A	年資層級B	年資層級C	任職前工作平均年數	最高學歷	畢業學校類別	畢業科系類別	眷屬量	通勤成本	歸屬部門
0	2014	1	0	1.0	1.0	3.0	19.0	4.0	0.0	1.0	...	2.0	1.0	1.0	2.0	NaN	NaN	5.0	0.0	8.0	19138.0
1	2015	1	0	1.0	1.0	3.0	19.0	6.0	0.0	1.0	...	2.0	2.0	1.0	2.0	NaN	NaN	5.0	2.0	8.0	19138.0
2	2016	1	0	1.0	1.0	3.0	19.0	6.0	0.0	1.0	...	2.0	2.0	1.0	2.0	NaN	NaN	5.0	2.0	8.0	19138.0
3	2017	1	0	1.0	1.0	3.0	19.0	6.0	0.0	1.0	...	2.0	2.0	1.0	2.0	NaN	NaN	5.0	2.0	8.0	19138.0
4	2014	3	0	0.0	1.0	4.0	8.0	1.0	0.0	0.0	...	5.0	5.0	0.0	0.0	2.0	4.0	1.0	2.0	8.0	8181.0

5 rows × 47 columns

資料描述性統計

	yyyy	PerNo	PerStatus	sex	工作分類	職等	廠區代碼	管理層級	工作資歷1	工作資歷2	...	年資層
count	14392.000000	14392.000000	14392.000000	14319.000000	14319.000000	14319.000000	14319.000000	14319.000000	14319.000000	14319.000000	14319.000000	14319.000000
mean	2015.520220	4410.755489	0.055309	0.701306	1.109365	4.375655	12.833927	1.791885	0.029122	0.051749	...	3.619
std	1.116459	2530.042411	0.228589	0.457701	0.312108	1.736769	5.634432	1.543562	0.168155	0.221528	...	1.761
min	2014.000000	1.000000	0.000000	0.000000	1.000000	1.000000	2.000000	1.000000	0.000000	0.000000	...	1.000
25%	2015.000000	2237.000000	0.000000	0.000000	1.000000	3.000000	8.000000	1.000000	0.000000	0.000000	...	2.000
50%	2016.000000	4387.000000	0.000000	1.000000	1.000000	4.000000	14.000000	1.000000	0.000000	0.000000	...	3.000
75%	2017.000000	6613.250000	0.000000	1.000000	1.000000	7.000000	18.000000	1.000000	0.000000	0.000000	...	5.000
max	2017.000000	8775.000000	1.000000	1.000000	2.000000	8.000000	20.000000	6.000000	1.000000	1.000000	...	9.000

8 rows × 47 columns

# ● 資料預處理

yyyy	0
PerNo	0
PerStatus	0
sex	73
工作分類	73
職等	73
廠區代碼	73
管理層級	73
工作資歷1	73
工作資歷2	73
工作資歷3	73
工作資歷4	73
工作資歷5	73
專案時數	73
專案總數	73
當前專案角色	73
特殊專案佔比	73
工作地點	73
訓練時數A	73
訓練時數B	73
訓練時數C	73
生產總額	73
榮譽數	73
是否升遷	73
升遷速度	73

近三月請假數A	73
近一年請假數A	73
近三月請假數B	73
近一年請假數B	73
出差數A	73
出差數B	73
出差集中度	73
年度績效等級A	73
年度績效等級B	73
年度績效等級C	73
年齡層級	73
婚姻狀況	73
年資層級A	73
年資層級B	73
年資層級C	73
任職前工作平均年數	73
最高學歷	5326
畢業學校類別	3841
畢業科系類別	73
眷屬量	73
通勤成本	73
歸屬部門	73
dtype: int64	

# ● 資料預處理

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 14392 entries, 0 to 14391
Data columns (total 45 columns):
 #   Column      Non-Null Count  Dtype  
0   yyyy         14392 non-null   int64  
1   PerNo        14392 non-null   int64  
2   PerStatus    14392 non-null   int64  
3   Sex          14392 non-null   float64
4   工作分類     14319 non-null   float64
5   職等         14319 non-null   float64
6   廠區代碼     14319 non-null   float64
7   管理層級     14319 non-null   float64
8   工作資歷1    14319 non-null   float64
9   工作資歷2    14319 non-null   float64
10  工作資歷3   14319 non-null   float64
11  工作資歷4   14319 non-null   float64
12  工作資歷5   14319 non-null   float64
13  專案時數    14319 non-null   float64
14  專案總數    14319 non-null   float64
15  當前專案角色 14319 non-null   float64
16  特殊專案佔比 14319 non-null   float64
17  工作地點     14319 non-null   float64
18  訓練時數A   14319 non-null   float64
19  訓練時數B   14319 non-null   float64
20  訓練時數C   14319 non-null   float64
21  生產總額     14319 non-null   float64
22  榮譽數       14319 non-null   float64
23  是否升遷     14319 non-null   float64
24  升遷速度     14319 non-null   float64
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 14319 entries, 0 to 14391
Data columns (total 45 columns):
 #   Column      Non-Null Count  Dtype  
0   yyyy         14319 non-null   int64  
1   PerNo        14319 non-null   int64  
2   PerStatus    14319 non-null   int64  
3   Sex          14319 non-null   float64
4   工作分類     14319 non-null   float64
5   職等         14319 non-null   float64
6   廠區代碼     14319 non-null   float64
7   管理層級     14319 non-null   float64
8   工作資歷1    14319 non-null   float64
9   工作資歷2    14319 non-null   float64
10  工作資歷3   14319 non-null   float64
11  工作資歷4   14319 non-null   float64
12  工作資歷5   14319 non-null   float64
13  專案時數    14319 non-null   float64
14  專案總數    14319 non-null   float64
15  當前專案角色 14319 non-null   float64
16  特殊專案佔比 14319 non-null   float64
17  工作地點     14319 non-null   float64
18  訓練時數A   14319 non-null   float64
19  訓練時數B   14319 non-null   float64
20  訓練時數C   14319 non-null   float64
21  生產總額     14319 non-null   float64
22  榮譽數       14319 non-null   float64
23  是否升遷     14319 non-null   float64
24  升遷速度     14319 non-null   float64
```



## 可視化分析

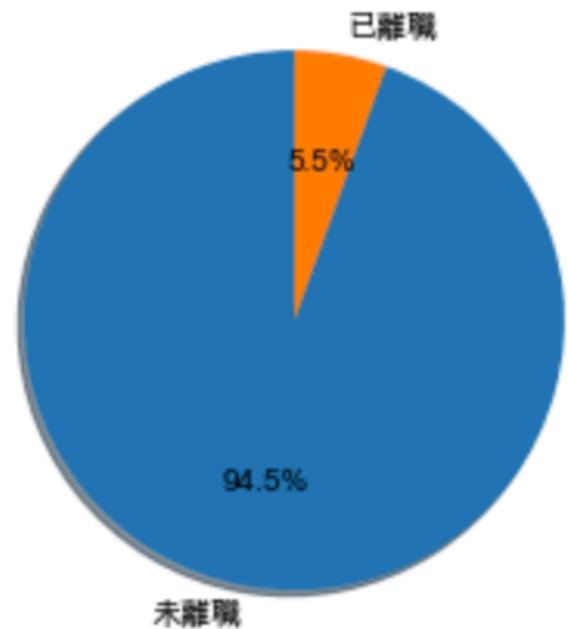
離職人數占比

總人數: 14319

0 13526

1 793

Name: PerStatus, dtype: int64



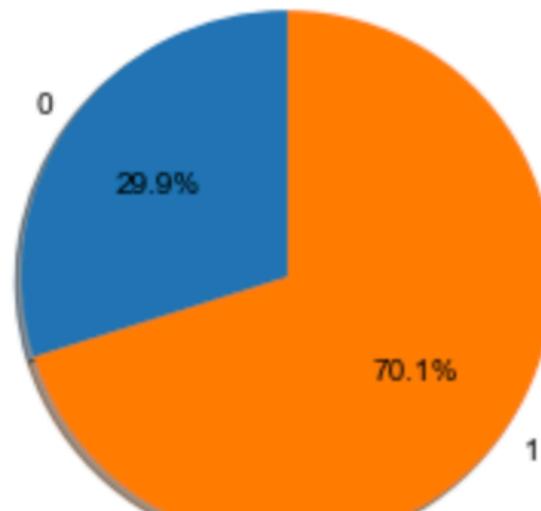
Sex占比

總人數: 14319

1.0 10042

0.0 4277

Name: sex, dtype: int64

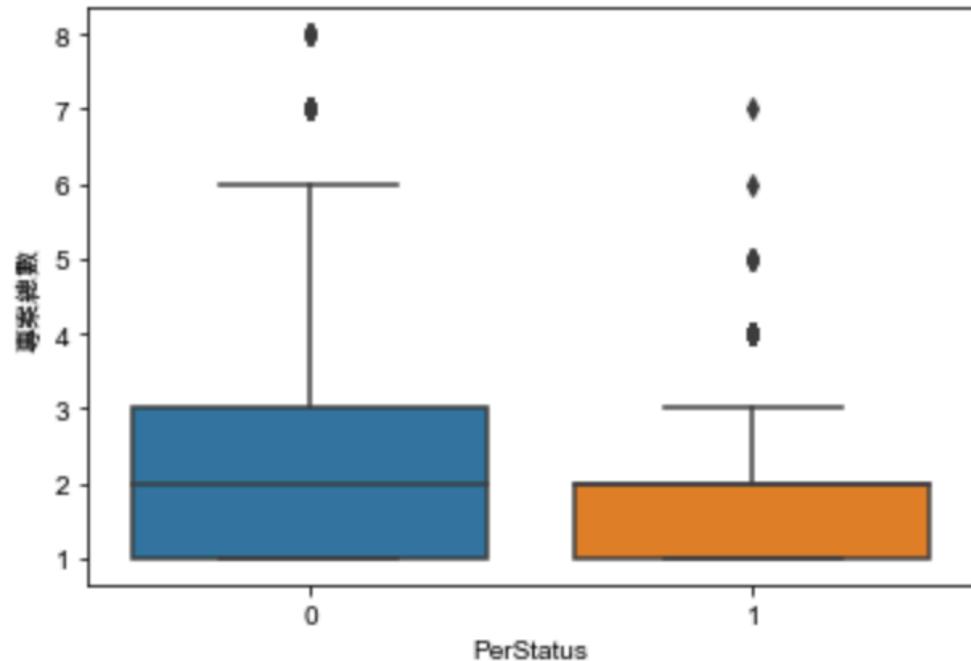




# 可視化分析

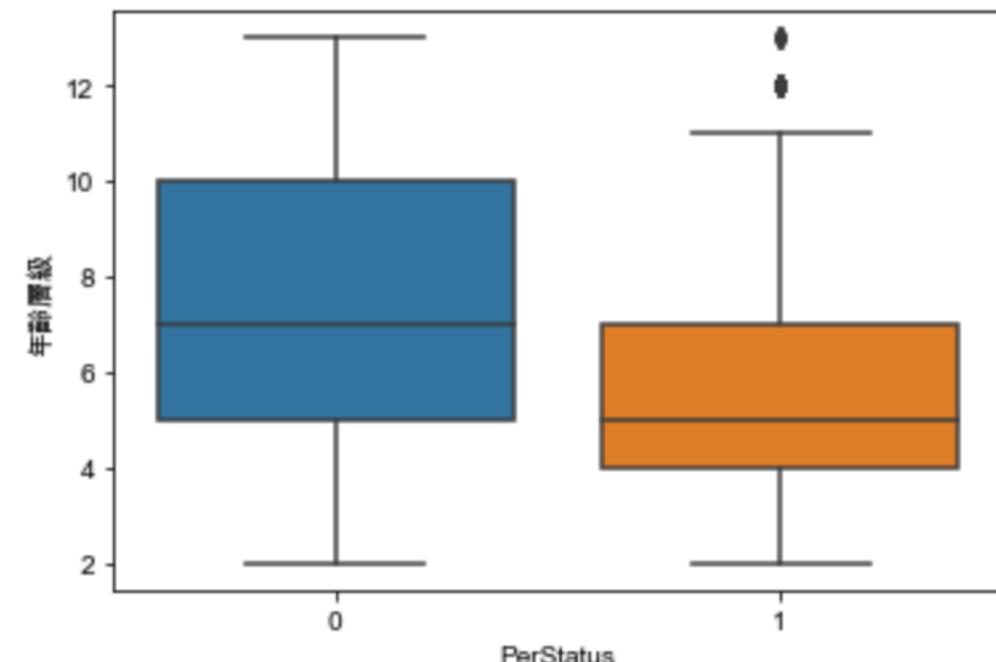
專案總數對離職的影響

```
<AxesSubplot:xlabel='PerStatus', ylabel='專案總數'>
```



年齡層級對離職的影響

```
<AxesSubplot:xlabel='PerStatus', ylabel='年齡層級'>
```

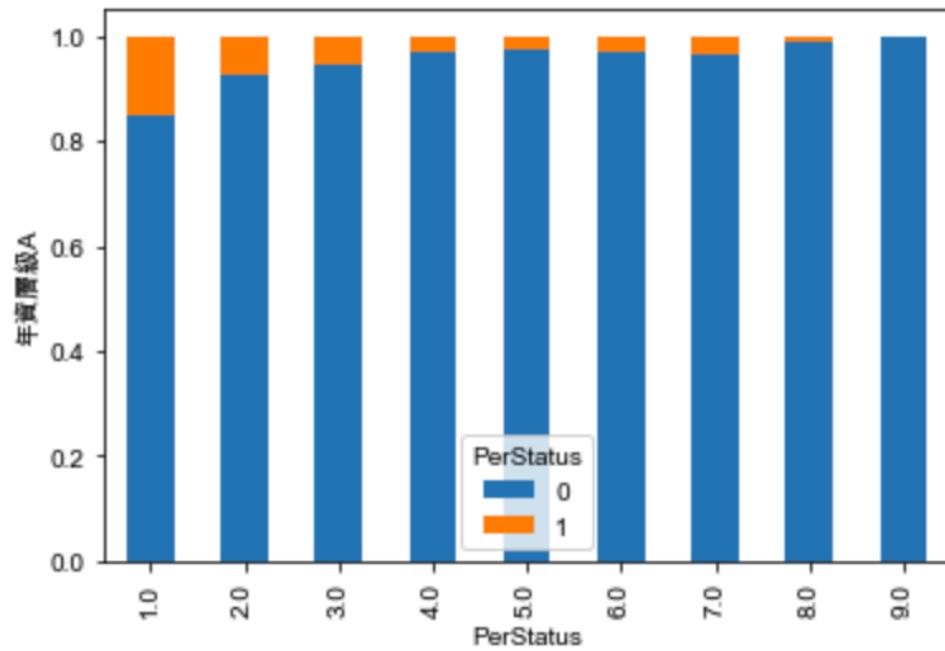




# 可視化分析

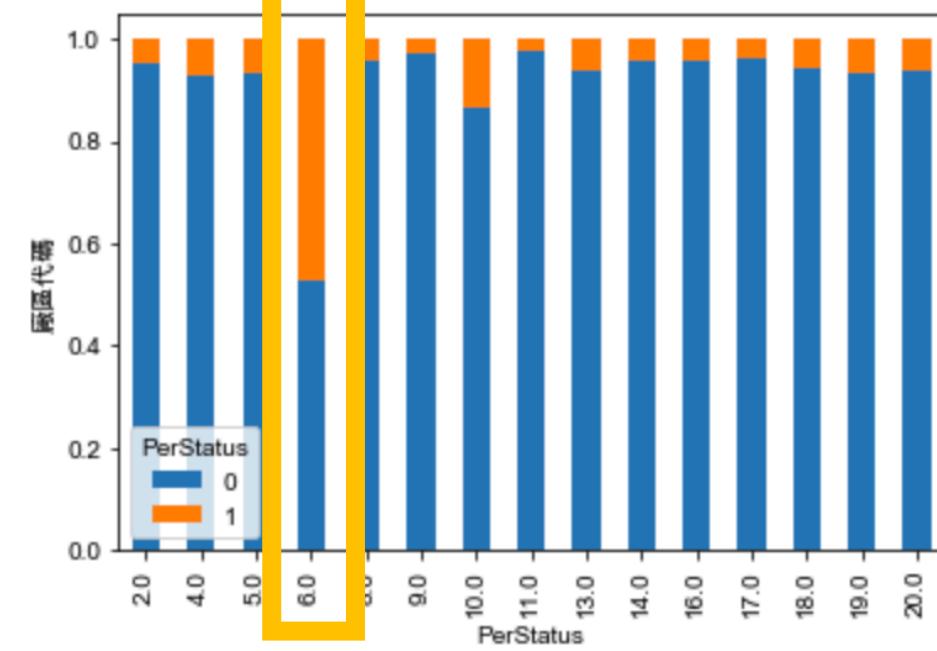
年資層級A對離職的影響

Text(0, 0.5, '年資層級A')

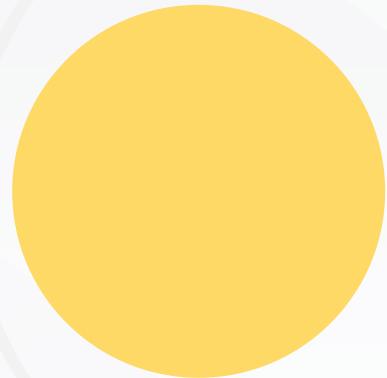


廠區代碼對離職的影響

Text(0, 0.5, '廠區代碼')



# 特徵選擇 模型建立



# ● 描述統計

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 14319 entries, 0 to 14391
Data columns (total 11 columns):
 #   Column      Non-Null Count  Dtype  
--- 
 0   PerStatus    14319 non-null   int64  
 1   管理層級     14319 non-null   float64 
 2   專案時數     14319 non-null   float64 
 3   專案總數     14319 non-null   float64 
 4   生產總額     14319 non-null   float64 
 5   榮譽數       14319 non-null   float64 
 6   是否升遷     14319 non-null   float64 
 7   升遷速度     14319 non-null   float64 
 8   年齡層級     14319 non-null   float64 
 9   年資層級A    14319 non-null   float64 
 10  廠區代碼    14319 non-null   float64 
dtypes: float64(10), int64(1)
memory usage: 1.3 MB
: PerStatus    0
管理層級     0
專案時數     0
專案總數     0
生產總額     0
榮譽數       0
是否升遷     0
升遷速度     0
年齡層級     0
年資層級A    0
廠區代碼    0
dtype: int64
```

訓練集準確率: 0.9172413793103448

測試集準確率: 0.9221368715083799

	precision	recall	f1-score	support
0	0.95	0.97	0.96	2716
1	0.12	0.08	0.10	148
accuracy			0.92	2864
macro avg	0.54	0.52	0.53	2864
weighted avg	0.91	0.92	0.91	2864

F beta score: 0.09027777777777779

# ● 相關係數

```
[ 1.0000000e+00  7.25757834e-02  7.09252676e-02  6.83013504e-02  
 6.35335631e-02  5.52814638e-02  4.31903714e-02  4.09800541e-02  
 3.53570759e-02  3.06993487e-02  2.72988782e-02  2.39931653e-02  
 1.71773775e-02  1.62912389e-02  1.62423741e-02  8.66829016e-03  
 5.46205851e-03  4.85523719e-03  3.40801127e-03  3.39768452e-03  
 1.13716000e-03  9.63249876e-04  -5.33415968e-03 -5.45354392e-03  
 -6.04050783e-03 -7.87833494e-03 -1.15416616e-02 -1.37901457e-02  
 -1.49287901e-02 -1.80034124e-02 -1.93464705e-02 -2.60898955e-02  
 -2.80129118e-02 -3.32231247e-02 -3.58826365e-02 -3.67218623e-02  
 -3.73242258e-02 -4.20897130e-02 -4.20960656e-02 -4.66927792e-02  
 -7.42775216e-02 -7.47824832e-02 -1.19660619e-01 -1.21275686e-01  
 -1.23798976e-01]
```

```
Index(['PerStatus', '婚姻狀況', '工作資歷5',  
       '訓練時數C', '職等', '工作資歷1', '出差',  
       '工作地點', '工作資歷2', '通勤成本', '歸  
       dtype='object')
```

訓練集準確率: 0.8948930597992143

測試集準確率: 0.8938547486033519

	precision	recall	f1-score	support
0	0.95	0.94	0.94	2716
1	0.10	0.14	0.12	148
accuracy			0.89	2864
macro avg	0.53	0.54	0.53	2864
weighted avg	0.91	0.89	0.90	2864

F beta score: 0.12287334593572777



# 過濾法 - 方差選擇

## KNN

KNN  
訓練集準確率: 0.9452640768223484  
測試集準確率: 0.94518156424581

	precision	recall	f1-score	support
0	0.95	1.00	0.97	2716
1	0.15	0.01	0.02	148
accuracy			0.95	2864
macro avg	0.55	0.50	0.50	2864
weighted avg	0.91	0.95	0.92	2864

F beta score: 0.018786127167630062

## GNB

GaussianNB  
訓練集準確率: 0.8945438673068529  
測試集準確率: 0.8942039106145251  
測試集召回率:

	precision	recall	f1-score	support
0	0.96	0.93	0.94	2716
1	0.15	0.22	0.18	148
accuracy			0.89	2864
macro avg	0.55	0.58	0.56	2864
weighted avg	0.91	0.89	0.90	2864

F beta score: 0.19359205776173286

## Random Forest

Random Forest  
訓練集準確率: 0.9999127018769096  
測試集準確率: 0.9479748603351955

	precision	recall	f1-score	support
0	0.95	0.94	0.95	2716
1	0.10	0.13	0.11	148
accuracy			0.90	2864
macro avg	0.53	0.53	0.53	2864
weighted avg	0.91	0.90	0.90	2864

F beta score: 0.1196705426356589

## Decision Tree

(Max depth : 35,Criterion : entropy)

DecisionTree  
訓練集準確率: 1.0  
測試集準確率: 0.897695530726257  
Accuracy: 0.897695530726257

	precision	recall	f1-score	support
0	0.95	0.94	0.95	2716
1	0.10	0.13	0.11	148
accuracy			0.90	2864
macro avg	0.53	0.53	0.53	2864
weighted avg	0.91	0.90	0.90	2864

F beta score: 0.1196705426356589



# 無特徵選擇

## KNN

KNN				
訓練集準確率: 0.9459624618070711				
測試集準確率: 0.946927374301676				
	precision	recall	f1-score	support
0	0.95	1.00	0.97	2716
1	0.39	0.05	0.08	148
accuracy			0.95	2864
macro avg	0.67	0.52	0.53	2864
weighted avg	0.92	0.95	0.93	2864

F beta score: 0.06481481481481483

## GNB

GaussianNB				
訓練集準確率: 0.8762112614578786				
測試集準確率: 0.8736033519553073				
測試集召回率:	precision	recall	f1-score	support
0	0.96	0.91	0.93	2716
1	0.14	0.27	0.18	148
accuracy			0.87	2864
macro avg	0.55	0.59	0.56	2864
weighted avg	0.92	0.87	0.89	2864

F beta score: 0.20733652312599685

## Random Forest

Random Forest				
訓練集準確率: 0.9999127018769096				
測試集準確率: 0.9476256983240223				
	precision	recall	f1-score	support
0	0.95	1.00	0.97	2716
1	0.33	0.01	0.03	148
accuracy			0.95	2864
macro avg	0.64	0.51	0.50	2864
weighted avg	0.92	0.95	0.92	2864

F beta score: 0.019174041297935103

## Decision Tree

(Max depth : 35,Criterion : entropy)

DecisionTree				
訓練集準確率: 1.0				
測試集準確率: 0.8983938547486033				
	precision	recall	f1-score	support
0	0.95	0.94	0.95	2716
1	0.12	0.15	0.13	148
accuracy			0.90	2864
macro avg	0.54	0.54	0.54	2864
weighted avg	0.91	0.90	0.90	2864

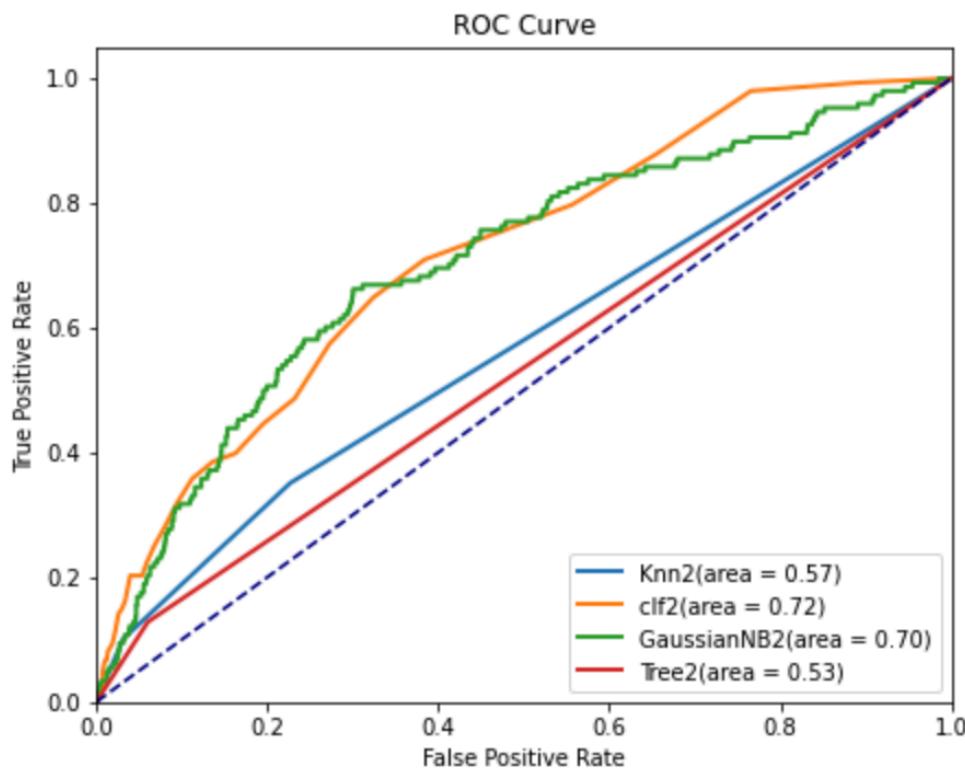
F beta score: 0.1375



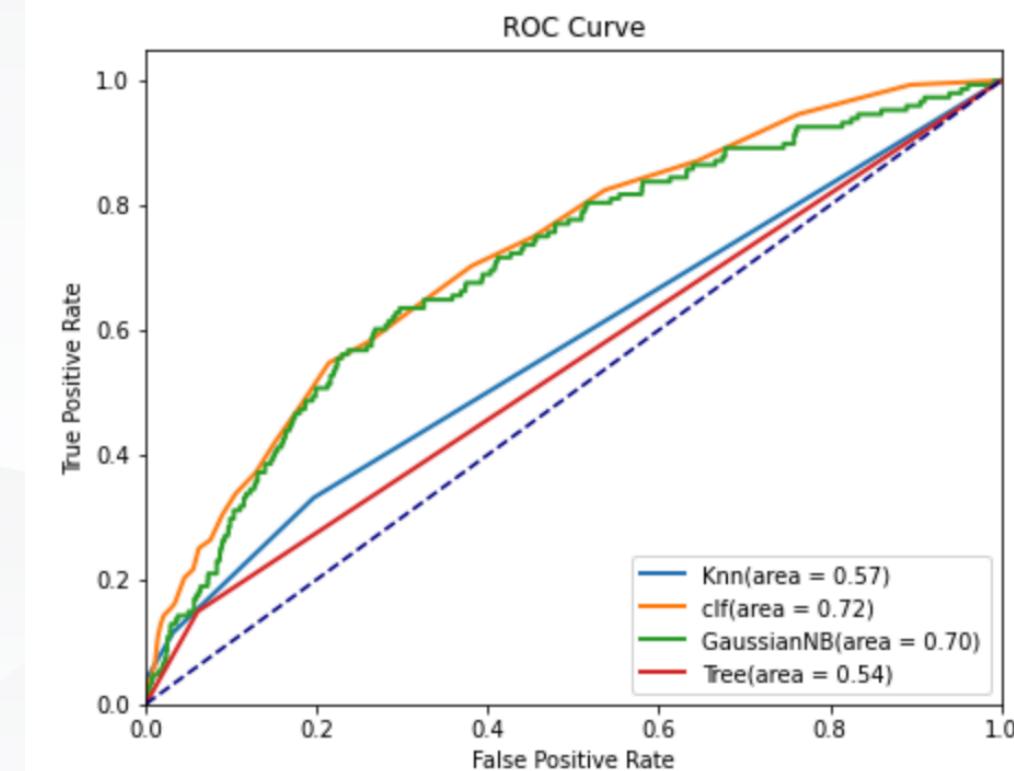


# 模型評估 - ROC

Filter-方差選擇



無特徵選擇



# 離職預測

# 離職預測

```
# prediction
test = pd.read_csv('test.csv')
#test.head()

test.describe()

print(test.isnull().sum())
print(test.info())

test.drop(['PerStatus', '最高學歷', '畢業學校類別'], axis = 1, inplace = True)
# test

col = test.columns
print(col)
for i in col:
    test.loc[test[i].isna(), i] = test[i].mean()
test.info()

test_id = pd.DataFrame(test['PerNo'], columns=['PerNo'])
test_id.head()
```

yyyy	0	近三月請假數A	18
PerNo	0	近一年請假數A	18
PerStatus	3739	近三月請假數B	18
sex	18	近一年請假數B	18
工作分類	18	出差數A	18
職等	18	出差數B	18
廠區代碼	18	出差集中度	18
管理層級	18	年度績效等級A	18
工作資歷1	18	年度績效等級B	18
工作資歷2	18	年度績效等級C	18
工作資歷3	18	年齡層級	18
工作資歷4	18	婚姻狀況	18
工作資歷5	18	年資層級A	18
專案時數	18	年資層級B	18
專案總數	18	年資層級C	18
當前專案角色	18	任職前工作平均年數	18
特殊專案佔比	18	最高學歷	1384
工作地點	18	畢業學校類別	1067
訓練時數A	18	畢業科系類別	18
訓練時數B	18	眷屬量	18
訓練時數C	18	通勤成本	18
生產總額	18	歸屬部門	18
榮譽數	18		
是否升遷	18		
升遷速度	18		

# 離職預測

```
# 無特徵選擇
## GNB
pred = GNB.predict(test)
print(pred)

pred = pd.DataFrame(pred, columns=['PerStatus'])
print(pred.value_counts())
# pred.info()
# test_id.info()
submission = pd.concat([test_id,pred],axis =1)
# submission.to_csv('submission_GNB.csv',index = 0)

# 過濾法(filter)
## 方差選擇法
test_var = selector1.transform(test)
test_var.shape

## GNB2
pred = GNB2.predict(test_var)
print(pred)

pred = pd.DataFrame(pred, columns=['PerStatus'])
print(pred.value_counts())
# pred.info()
# test_id.info()
submission = pd.concat([test_id,pred],axis =1)
# submission.to_csv('submission.csv',index = 0)
```

## Private Leaderboard

評估結果 排名

0.1413043 10/41

0.1282894

0.0981132