資料科學導論 Competition 1 說明

A general model for Binary Classification (ABC)

Competition 1 (CP1) 競賽時程

- 10/21 (Mon) CP1 announce
- 10/23 (Wed) Form your team
- 11/23 (Sat) Upload Predictions at Website
- 11/24 (Sun) Release of private leaderboard ranking
- 11/27 (Wed) Upload Report and Code at Moodle
- 12/2 (Mon) CP1 Award Ceremony & Invite top-5 teams for presentations

Outline

- 任務說明
- 資料介紹
- 資料切割
- 評估指標排名方式
- 範例程式說明
- 上傳格式
- 平台使用
- Public / Private leaderboards 介紹

Outline

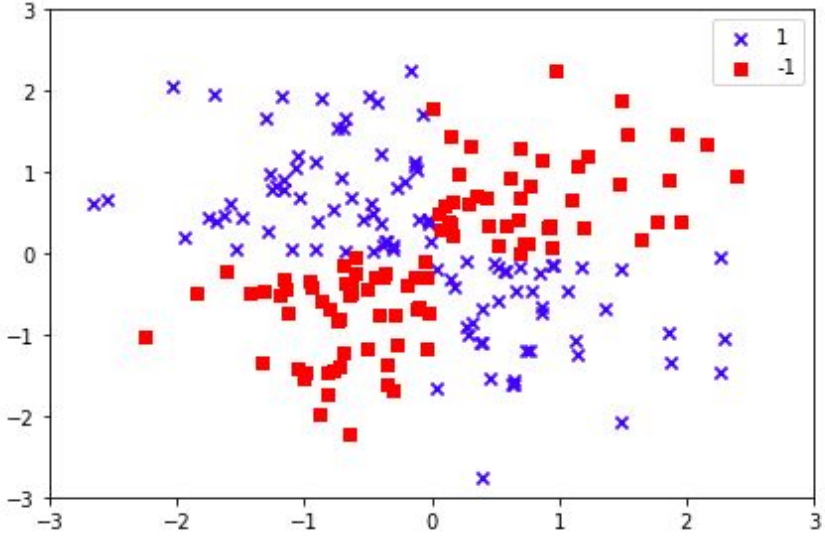
- **任務說明**
- 資料介紹
- 資料切割
- 評估指標排名方式
- 範例程式說明
- 上傳格式
- 平台使用
- Public / Private leaderboards 介紹

任務說明

本次競賽中，我們會提供許多資料集，同學們需要對我們提供的資料集進行分析、建模，並產生對應的預測結果。最終分數會是**模型在每個資料集上的預測表現之加權平均**，因此同學們設計的資料處理流程與模型必須具有**通用性**，能夠在大多數資料集上具有好的表現。

資料集個數: **49**

任務類型: **二分類**

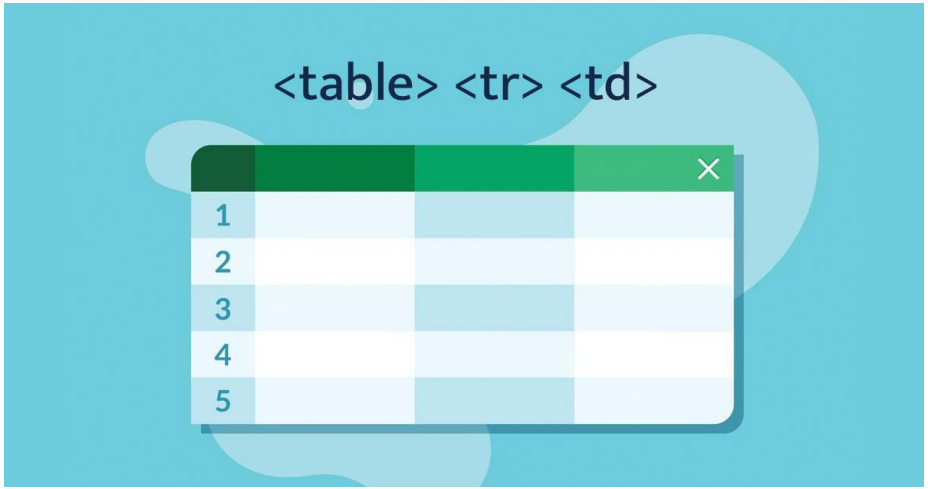


Outline

- 任務說明
- **資料介紹**
- 資料切割
- 評估指標排名方式
- 範例程式說明
- 上傳格式
- 平台使用
- Public / Private leaderboards 介紹

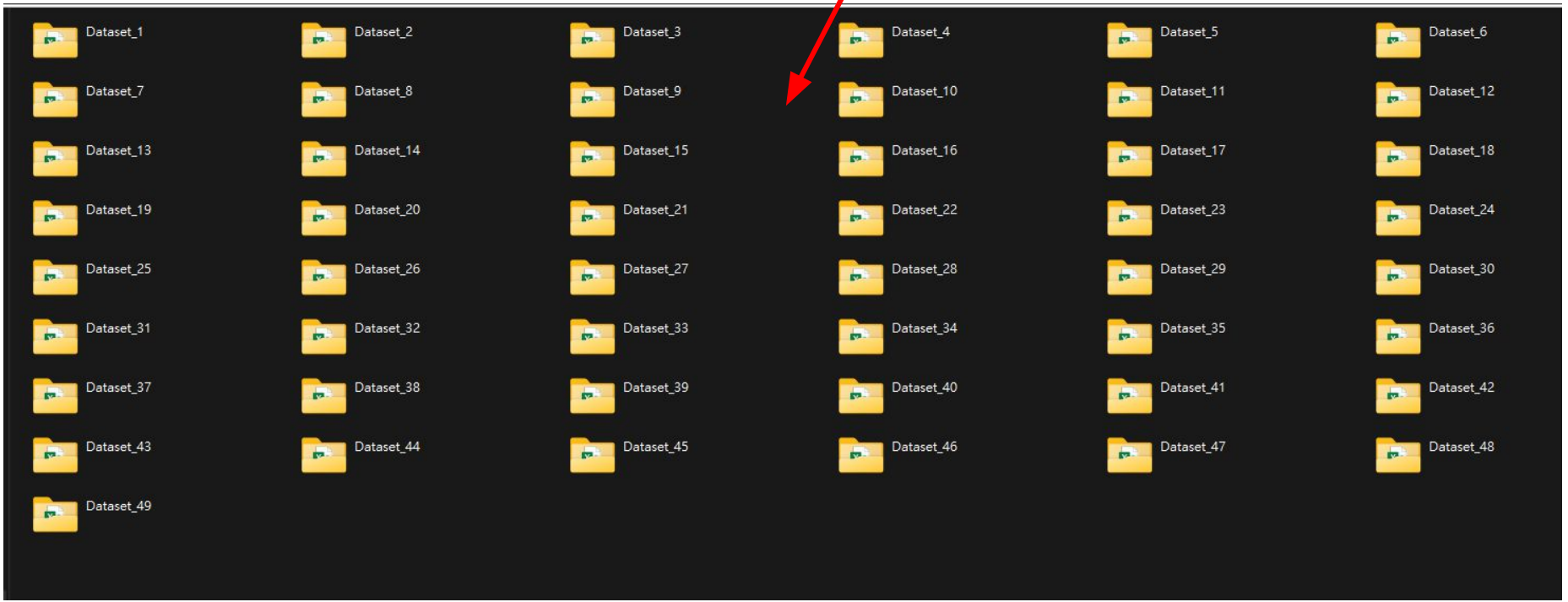
資料介紹

- 本次競賽所使用的資料集為49個二分類任務之小資料集(資料筆數小於10000)



資料介紹

每個資料集對應到一個資料夾



資料介紹

名稱	修改日期	類型	大小
X_test	2024/10/21 上午 12:14	Microsoft Excel ...	48 KB
X_train	2024/10/21 上午 12:14	Microsoft Excel ...	72 KB
y_predict	2024/10/21 上午 12:15	Microsoft Excel ...	1 KB
y_train	2024/10/21 上午 12:14	Microsoft Excel ...	1 KB

資料集內有X_train, X_test, y_train, y_predict四個csv檔案

y_predict為同學們需要透過模型預測產生的，此處原有的y_predict僅為格式參考，不為正確答案

資料介紹

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Feature_1	Feature_2	Feature_3	Feature_4	Feature_5	Feature_6	Feature_7	Feature_8	Feature_9	Feature_10	Feature_11	Feature_12	Feature_13	Feature_14	Feature_15	Feature_16	Feature_17	Feature_18	Feature_19	Feature_20
2	0.791656	-0.4461	-0.89421	-0.79165	-1.06757	-4.45096	-4.14061	1	7	4	2	3	7	0	2	0	0	0	0	4
3	-0.17998	1.067571	0.080683	-0.56589	0.018889	1.067571	0.539817	6	1	4	3	6	8	0	0	2	1	0	1	5
4	-0.31864	2.247148	-0.36603	-0.18002	1.216977	-1.06757	-1.46524	6	6	2	2	6	4	0	0	2	0	1	3	7
5	0.166808	-1.47573	-0.56589	-2.24E-07	-0.72859	0.119611	0.180045	4	11	0	0	0	1	0	0	0	0	0	0	5
6	-1.46526	0.791639	-0.36603	-0.18002	-0.82298	-1.06757	-0.79174	0	9	1	0	4	5	0	0	3	1	0	0	1
7	-0.6364	-0.4461	1.190714	0.348363	1.216977	0.636138	0.927461	3	6	2	2	5	4	0	0	0	1	0	0	3
8	1.464391	0.791639	-0.36603	-0.18002	0.018889	-1.06757	-0.79174	0	1	4	3	4	8	0	0	3	1	0	0	1
9	0.166808	0.565949	0.56597	0.876128	-0.3932	-0.79164	-0.79174	6	4	1	1	2	4	0	0	0	1	0	0	3
10	-1.06756	1.465234	-1.80273	-1.34872	-5.19934	0.216532	-0.11932	6	10	2	0	6	1	0	1	0	0	0	0	7
11	0.166808	-0.4461	-0.89421	-0.79165	0.018889	-4.45096	-4.14061	1	1	3	3	3	8	0	2	0	0	0	0	4
12	-0.36611	-1.06757	0.366114	-0.3661	1.067571	1.291795	-0.11932	6	0	1	1	6	4	0	0	2	1	0	0	7
13	1.464391	1.15672	0.080683	1.067564	-0.83987	-0.13469	-0.11932	6	6	4	2	3	6	0	0	2	1	1	1	7
14	0.565979	0.366106	0.080683	1.067564	0.473478	2.189352	4.342817	6	3	4	3	6	6	0	0	1	1	0	0	0
15	-0.31864	4.48E-07	0.463721	0.876128	-1.06757	1.241868	0.180045	6	7	2	2	6	7	0	3	2	1	0	2	7

X_train

資料介紹

數值型資料　**類別型資料**

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Feature_1	Feature_2	Feature_3	Feature_4	Feature_5	Feature_6	Feature_7	Feature_8	Feature_9	Feature_10	Feature_11	Feature_12	Feature_13	Feature_14	Feature_15	Feature_16	Feature_17	Feature_18	Feature_19	Feature_20
2	0.791656	-0.4461	-0.89421	-0.79165	-1.06757	-4.45096	-4.14061	1	7	4	2	3	7	0	2	0	0	0	0	4
3	-0.17998	1.067571	0.080683	-0.56589	0.018889	1.067571	0.539817	6	1	4	3	6	8	0	0	2	1	0	1	5
4	-0.31864	2.247148	-0.36603	-0.18002	1.216977	-1.06757	-1.46524	6	6	2	2	6	4	0	0	2	0	1	3	7
5	0.166808	-1.47573	-0.56589	-2.24E-07	-0.72859	0.119611	0.180045	4	11	0	0	0	1	0	0	0	0	0	0	5
6	-1.46526	0.791639	-0.36603	-0.18002	-0.82298	-1.06757	-0.79174	0	9	1	0	4	5	0	0	3	1	0	0	1
7	-0.6364	-0.4461	1.190714	0.348363	1.216977	0.636138	0.927461	3	6	2	2	5	4	0	0	0	1	0	0	3
8	1.464391	0.791639	-0.36603	-0.18002	0.018889	-1.06757	-0.79174	0	1	4	3	4	8	0	0	3	1	0	0	1
9	0.166808	0.565949	0.56597	0.876128	-0.3932	-0.79164	-0.79174	6	4	1	1	2	4	0	0	0	1	0	0	3
10	-1.06756	1.465234	-1.80273	-1.34872	-5.19934	0.216532	-0.11932	6	10	2	0	6	1	0	1	0	0	0	0	7
11	0.166808	-0.4461	-0.89421	-0.79165	0.018889	-4.45096	-4.14061	1	1	3	3	3	8	0	2	0	0	0	0	4
12	-0.36611	-1.06757	0.366114	-0.3661	1.067571	1.291795	-0.11932	6	0	1	1	6	4	0	0	2	1	0	0	7
13	1.464391	1.15672	0.080683	1.067564	-0.83987	-0.13469	-0.11932	6	6	4	2	3	6	0	0	2	1	1	1	7
14	0.565979	0.366106	0.080683	1.067564	0.473478	2.189352	4.342817	6	3	4	3	6	6	0	0	1	1	0	0	0
15	-0.31864	4.48E-07	0.463721	0.876128	-1.06757	1.241868	0.180045	6	7	2	2	6	7	0	3	2	1	0	2	7

X_train

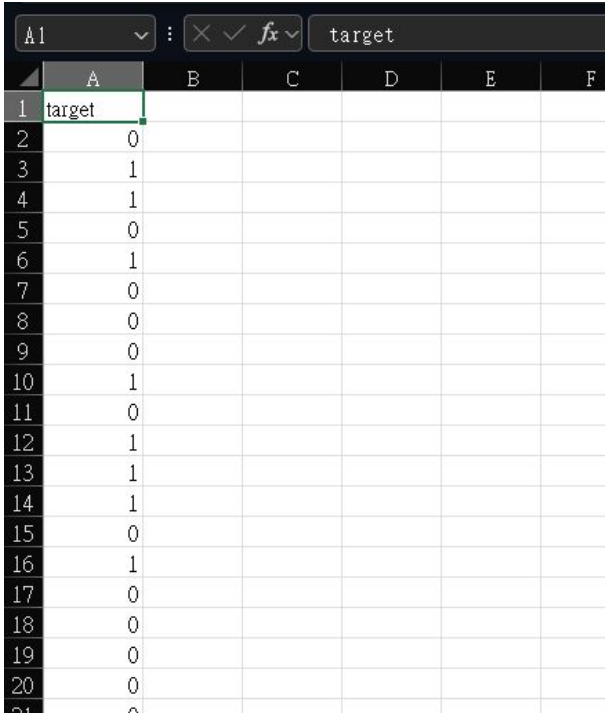
資料介紹

Feature_1

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
1	Feature_1	Feature_2	Feature_3	Feature_4	Feature_5	Feature_6	Feature_7	Feature_8	Feature_9	Feature_10	Feature_11	Feature_12	Feature_13	Feature_14	Feature_15	Feature_16	Feature_17	Feature_18	Feature_19	Feature_20
2	0.166808	4.48E-07	-0.89421	-1.34872	0.75451	-0.51335	-0.51728	2	4	1	1	3	7	0	0	1	0	0	2	2
3	-1.06756	1.988124	-0.17994	-0.71244	0.565949	-0.33438	-0.11932	6	1	4	3	6	0	0	0	4	1	0	0	2
4	1.464391	1.083141	-0.17994	0.791584	-0.18001	-0.24104	-0.11932	6	11	3	0	6	8	0	0	1	1	0	1	4
5	1.464391	1.083141	-0.17994	0.791584	0.018889	-0.24104	-0.11932	6	1	1	3	6	8	0	0	1	1	0	1	4
6	-0.31864	1.067571	0.080683	-0.56589	0.366106	1.067571	0.539817	6	9	0	0	6	6	0	0	2	1	0	1	5
7	-0.31864	-1.47573	-0.56589	-2.24E-07	-0.18912	0.119611	0.180045	4	3	1	3	0	5	0	0	0	0	0	0	5
8	-0.96741	0.791639	-0.36603	-0.18002	-0.82298	-1.06757	-0.79174	0	9	2	0	4	5	0	0	3	1	0	0	1
9	-0.6364	0.565949	0.56597	0.876128	1.216977	-0.79164	-0.79174	6	6	2	2	2	4	0	0	0	1	0	0	3
10	0.166808	4.48E-07	-0.89421	-1.34872	1.216977	-0.51335	-0.51728	2	6	2	2	3	4	0	0	1	0	0	2	2
11	1.464391	4.48E-07	-0.89421	-1.34872	-0.35113	-0.51335	-0.51728	2	1	1	3	3	2	0	0	1	0	0	2	2
12	1.464391	-0.4461	1.190714	0.348363	0.565949	0.636138	0.927461	3	1	2	3	5	0	0	0	0	1	0	0	3
13	-4.21305	-1.47573	1.190714	1.465243	0.018889	1.465235	0.927461	5	1	4	3	1	8	1	0	1	1	0	0	6
14	-1.06756	0.565949	0.56597	0.876128	-0.3932	-0.79164	-0.79174	6	4	4	1	2	4	0	0	0	1	0	0	3
15	0.166808	0.366106	-1.24188	-0.63747	-1.06757	-0.18001	-0.11932	6	7	3	2	6	7	0	2	0	0	0	0	6
16	0.166808	1.067571	0.080683	-0.56589	-0.82298	1.067571	0.539817	6	9	3	0	6	5	0	0	2	1	0	1	5

X_test

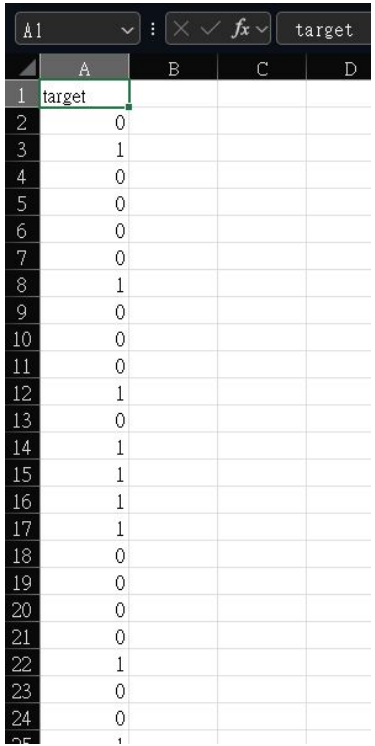
資料介紹



	A
1	target
2	0
3	1
4	1
5	0
6	1
7	0
8	0
9	0
10	1
11	0
12	1
13	1
14	1
15	0
16	1
17	0
18	0
19	0
20	0

y_train

資料介紹



y_predict

此處原有的y_predict僅為格式參考，不為正確答案

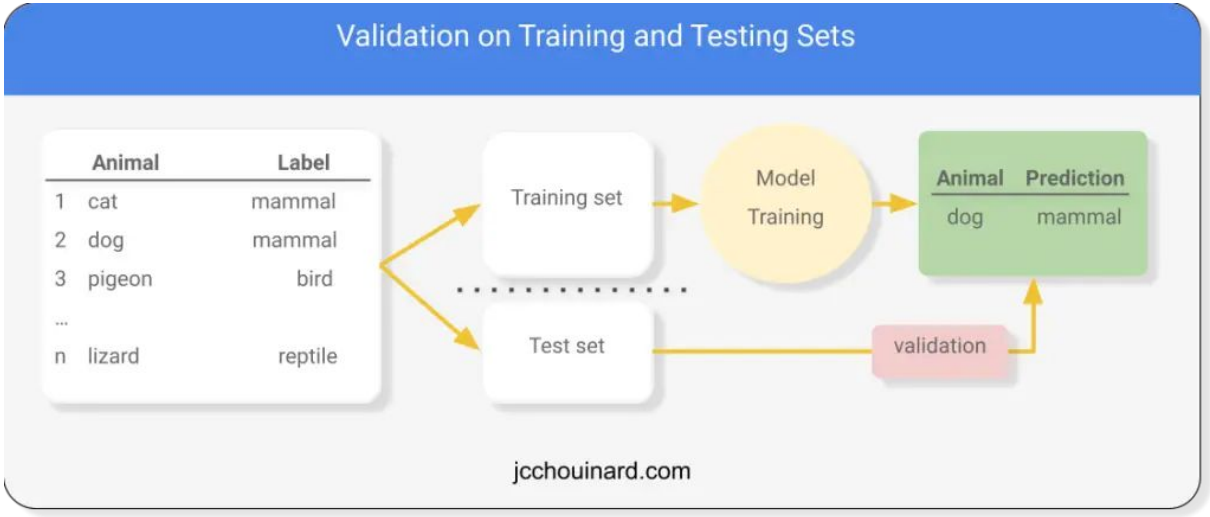
Outline

- 任務說明
- 資料介紹
- **資料切割**
- 評估指標排名方式
- 範例程式說明
- 上傳格式
- 平台使用
- Public / Private leaderboards 介紹

資料切割

train: 60%

test: 40%

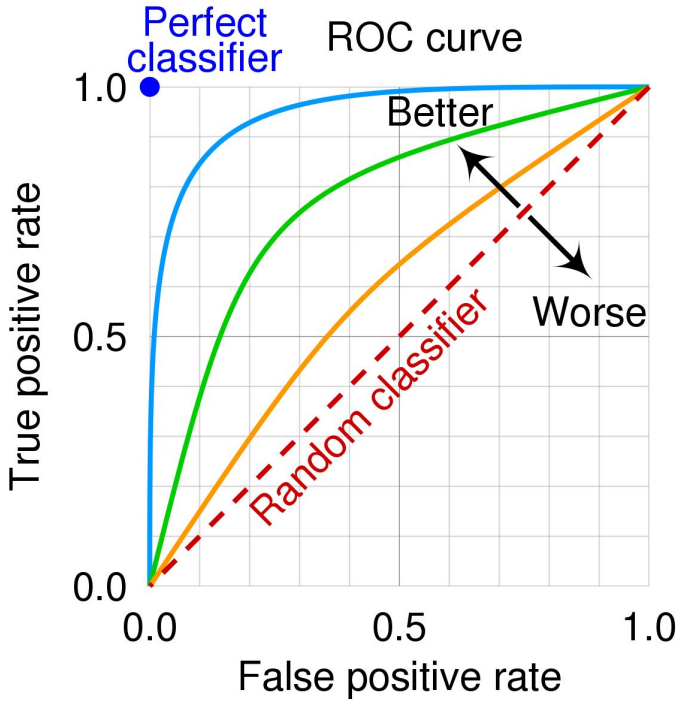


Outline

- 任務說明
- 資料介紹
- 資料切割
- **評估指標排名方式**
- 範例程式說明
- 上傳格式
- 平台使用
- Public / Private leaderboards 介紹

評估指標排名方式

Score = AUC



ROC-AUC 介紹

競賽可實作的方法

- **上課介紹過的方法**
  - 各種監督式學習方法
  - 訓練技巧 (資料前處理、類別不平行處理方式、特徵選擇, etc.)
- **上課沒介紹過的方法**
  - 自行找state-of-the-art的二元分類方法 (進階神經網路方法)
- **自行設計新方法** (須於報告具體說明)

若最終方法有採用或修改他人的方法，請務必於報告中明確引用paper/github

Outline

- 任務說明
- 資料介紹
- 資料切割
- 評估指標排名方式
- **範例程式說明**
- 上傳格式
- 平台使用
- Public / Private leaderboards 介紹

範例程式說明

Read All Dataset CSV

In [39]:
```
import os
import csv
import pandas as pd
import numpy as np
```

In [40]:
```
dataset_names=[]
X_trains=[]
y_trains=[]
X_tests=[]
for folder_name in os.listdir("./Competition_data"):
    # print(folder_name)
    dataset_names.append(folder_name)
    X_trains.append(pd.read_csv(f"./Competition_data/{folder_name}/X_train.csv",header=0))
    y_trains.append(pd.read_csv(f"./Competition_data/{folder_name}/y_train.csv",header=0))
    X_tests.append(pd.read_csv(f"./Competition_data/{folder_name}/X_test.csv",header=0))
```

Data Preprocessing & Feature Engineering

In [41]:
```
## your code here
```

範例程式說明

由於此次競賽的評估指標是以AUC為主，
因此可以改為嘗試AUC來印出模型表現

train test split & build Model

You can select an appropriate model and perform corresponding hyperparameter tuning.

In [42]:
```
from sklearn.model_selection import train_test_split
from sklearn.neighbors import KNeighborsClassifier
from sklearn.metrics import accuracy_score, f1_score, precision_score
```

In [43]:
```
models=[]
for i in range(len(dataset_names)):
    tmp_X_train, tmp_X_test, tmp_y_train, tmp_y_test = train_test_split(X_trains[i], y_trains[i], test_size=0.2, random_sta
    model = KNeighborsClassifier(n_neighbors=3)
    model.fit(tmp_X_train, tmp_y_train.squeeze())
    tmp_y_predict = model.predict(tmp_X_test)
    acc=accuracy_score(tmp_y_test, tmp_y_predict)
    precision=precision_score(tmp_y_test, tmp_y_predict)
    f1=f1_score(tmp_y_test, tmp_y_predict)
    # print(f"{dataset_names[i]}: accuracy={round(acc,3)}, precision={round(precision,3)}, f1={round(f1,3)}\n")
    models.append(model)
```

範例程式說明

Inference Model

In [58]:
```
y_predicts=[]
for i in range(len(dataset_names)):
    y_predict=models[i].predict(X_tests[i])
    df = pd.DataFrame(y_predict, columns=['target'])
    y_predicts.append(df)
```

Save result

In [59]:
```
for idx,dataset_name in enumerate(dataset_names):
    df=y_predicts[idx]
    df.to_csv(f'./Competition_data/{dataset_name}/y_predict.csv', index=False,header=True)
```

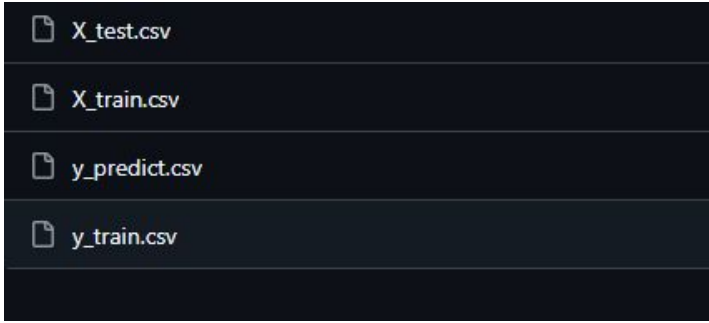
In [ ]:
```

```

Outline

- 任務說明
- 資料介紹
- 資料切割
- 評估指標排名方式
- 範例程式說明
- **上傳格式**
- 平台使用
- Public / Private leaderboards 介紹

上傳格式

將模型對於每個資料集的預測結果之y_predict.csv寫回，也就是如同以下路徑
/Competition_data/對應的資料集名稱資料夾/y_predict.csv



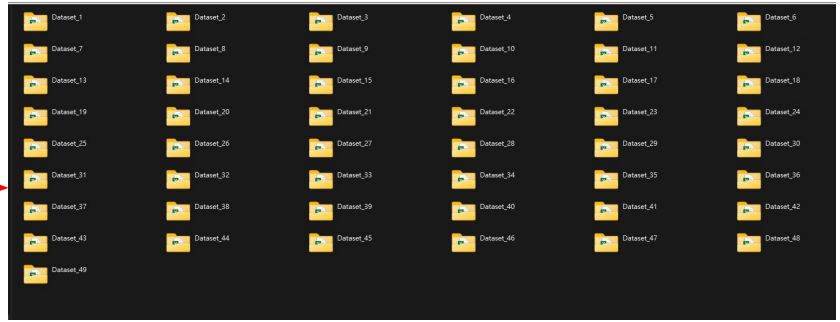
最後將整個Competition_data資料夾壓縮成"Group_組別編號.zip"並上傳

上傳格式

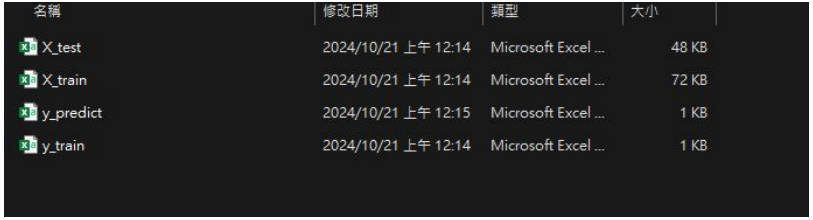
檔案架構如下

壓縮檔包含

資料夾中包含各個資料集對應的子資料夾



每個子資料夾中包含這四個檔案



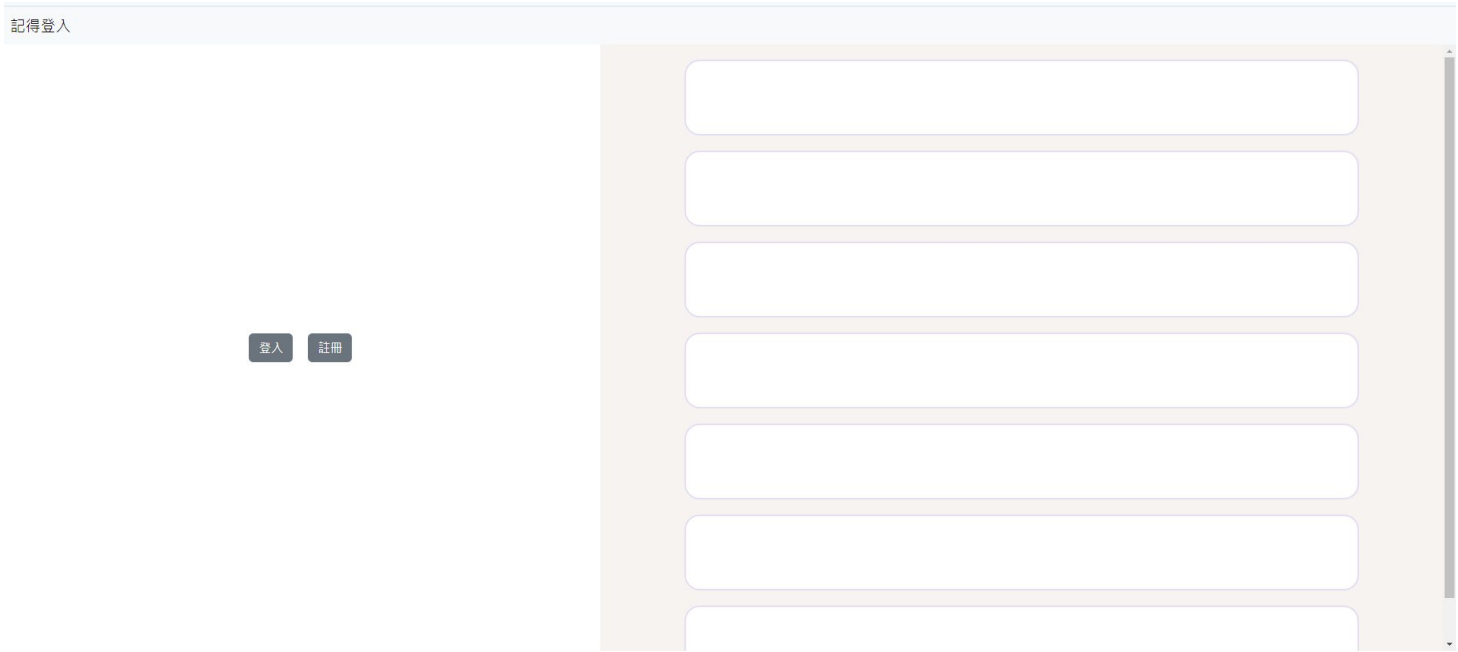
Outline

- 任務說明
- 資料介紹
- 資料切割
- 評估指標排名方式
- 範例程式說明
- 上傳格式
- **平台使用**
- Public / Private leaderboards 介紹

分組資訊

- 請於10/23 23:59前填寫分組資訊 (每組至多3人, 可1人): https://forms.gle/jHDEb1aztQ71HLGT6
- 分組進行兩次競賽與期末專題
- 若需協助找人組隊, 可單人填寫後備註
- 每一組可註冊一個帳號
- 分組完成後, 會請大家提供 "帳號--隊名" 的 mapping

平台使用



http://140.116.246.240/

平台使用


登入
帳號
密碼
登入


註冊新帳號
帳號
密碼
確認密碼
註冊


選擇壓縮檔
Choose Files
No file chosen
上傳

平台使用

到時候你們的結果會在這裡

平台使用

- 頁面會持續改進，陸續提供新功能。
- 一組一個帳號
- 不要亂上傳檔案

Outline

- 任務說明
- 資料介紹
- 資料切割
- 評估指標排名方式
- 範例程式說明
- 上傳格式
- 平台使用
- **Public / Private leaderboards 介紹**

Public / Private Leaderboards 介紹

本次競賽也會有public / private leaderboard的機制，會將test data分成兩部分(public與private)，同學們上傳的y_predict中，一部分的預測結果會當成計算 public score rank的依據，並將分數與排名顯示在 public leaderboard上，而另一部分的預測結果則是當成計算 private score rank的依據，但private leaderboard只有在**競賽結束後才會公布**，競賽進行期間是**隱藏**起來的。且我們不會公布 public與private之切分方式！

Public / Private Leaderboards 介紹

- 大家需要根據自己上傳後，在競賽平台 public leaderboard上顯示的排名與評估指標 (AUC)來得知此次上傳的結果，並以此為依據來對自己的模型與演算法進行優化。
- public leaderboard僅代表同學在測試集內的一部份資料的表現，請大家不要為了追求在 public leaderboard上的高排名 (也就是overfit了public leaderboard)，而用一些不夠具有泛用性的方式來建模與處理資料 (儘管public與private leaderboard的排名是高度相關，但些微的差異就足以造成排名的變化)
- **一個好的資料處理流程與 AI模型應該要能在 public與private測試資料上都有好的表現**
- **最終排名會以競賽結束後公布的 private leaderboard為主**

Leaderboard上面會有一些我們提供的 baselines，也佔據著排名，
請同學們努力打敗這些 baselines😝😝

歡迎協助平台debug，有任何建議或發現問題，請讓助教知道，我們會努力開發讓平台更obust，**被採用有機會酌量bonus!!!**😆

祝大家競賽順利，謝謝大家 !!