首創導入AI自動審核檢測機制
以稻穀生產成本調查為例

單位：行政院農業委員會農糧署統計室

報告人：蔡永輝

大綱

稻穀生產成本調查自民國36年起由前省政府糧食局負責辦理本調查，期間經歷多次之變革及組織調整；87年精省後，調查移由行政院農業委員會辦理，93年農糧署成立，則由本署承辦該項調查工作至今。

1. 研究動機
2. 調查審核作業面臨之問題
3. 偵測模型之建立與導入
4. 上線後系統實作
5. 總結與未來展望

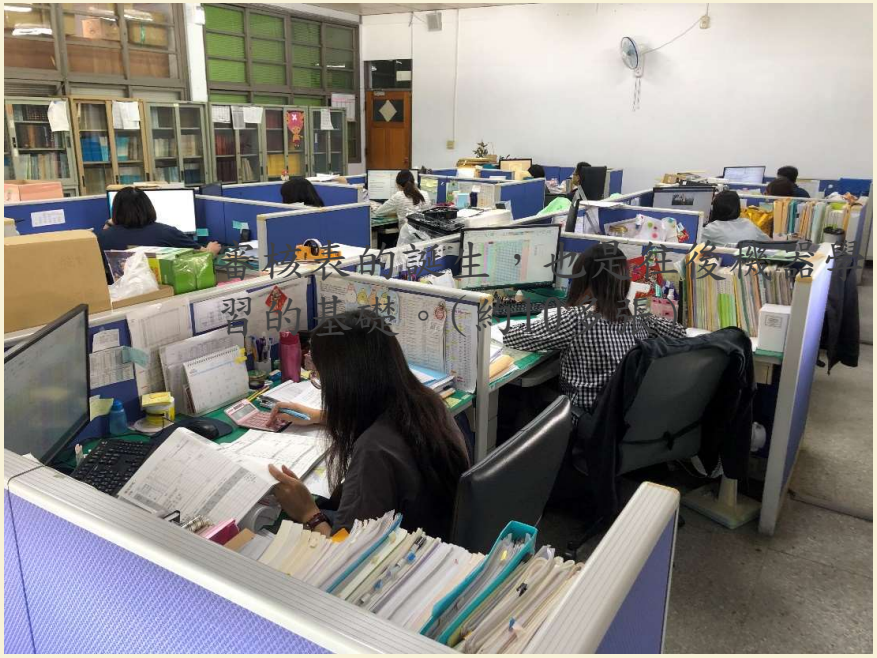
研究動機(1/2)

政府統計調查工作內容是什麼?

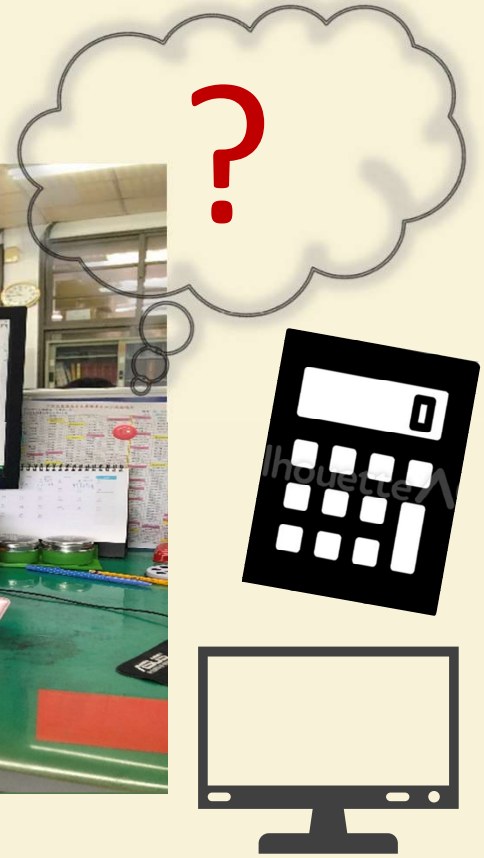
研究動機(2/2)

- 審核員在調查作業流程上扮演之角色?
- 審核確認作業成本?
- 但是，如果....

調查審核作業面臨之問題(1/2)



審核表的誕生，也是往後機器學習的基礎。(編輯中)



調查審核作業面臨之問題(2/2)

過往的嘗試

1. 傳統統計方法：
   - A. 單變數：計算Z值
   - B. 雙變數：計算模型槓桿值、student化殘差等
2. 神經網絡(生成模型)：自動編碼器

失敗的原因

1. 沒有解決非線性可分的情況
2. 自動編碼器模型預測較不穩定
3. 沒有標記資料難以評估

偵測模型之建立與導入(1/10)

目標與需求規劃

- 要能解決非線性可分的情況
- 要能穩定預測
- 可實際上線使用
- 可重複訓練
- 偽陰性與偽陽性誰重要?
  - (Def：資料錯誤稱其呈現陽性)
- 誰使用?

偵測模型之建立與導入(2/10)

- 運用2016年至2020年稻穀生產成本資料，其中80%為訓練資料；20%為測試資料。

```
{
    "DATASET": "每公頃插秧",
    "DATAS":[
        {
            "年度": "2016",
            "序號": "1",
            "農民": "RCIHH_6P688QDK8",
            "原始or審核": "1",
            "縣市": "宜蘭縣",
            "鄉鎮": "壯圍鄉",
            "稻種": "1",
            "插植人工合計時數": "11",
            "插植人工合計費用": "2727",
            "插植機工時數": "5",
            "插植機工費用": "7000",
            "耕犁整地人工合計時數": "0",
            "耕犁整地人工合計費用": "0",
            "耕犁整地機工時數": "5",
            "耕犁整地機工費用": "14000",
            "種苗費": "7000",
            "種籽費": "0"
        },
        {
            "年度": "2016",
            "序號": "1",
```

偵測模型之建立與導入(3/10)

- 加入鄉鎮變數及使用集群分析方法，同時解決非線性可分與避免維度過度擴張。

第1群

桃園市	八德區
桃園市	中壢區
桃園市	平鎮區
桃園市	新屋區
桃園市	觀音區
桃園市	楊梅區
新竹縣	竹北市
新竹縣	湖口鄉
新竹縣	竹東鎮
新竹縣	芎林鄉
新竹市	北區
苗栗縣	竹南鎮
苗栗縣	苑裡鎮
苗栗縣	通霄鎮
苗栗縣	南庄鄉
苗栗縣	造橋鄉
台中市	后里區
嘉義縣	義竹鄉

第2群

新北市	鶯歌區
新北市	金山區
新北市	淡水區
桃園市	大園區
桃園市	大溪區
桃園市	龍潭區
桃園市	蘆竹區
新竹縣	關西鎮
新竹縣	新埔鎮
新竹市	香山區
苗栗縣	後龍鎮
苗栗縣	頭份市
苗栗縣	公館鄉
苗栗縣	銅鑼鄉
苗栗縣	西湖鄉
苗栗縣	苗栗市

宜蘭縣　壯圍鄉

第3群

新竹縣	新豐鄉
台中市	外埔區
彰化縣	大城鄉
彰化縣	和美鎮
彰化縣	埤頭鄉
彰化縣	大村鄉
彰化縣	福興鄉
彰化縣	溪湖鎮
彰化縣	竹塘鄉
彰化縣	埔鹽鄉
彰化縣	北斗鎮
彰化縣	伸港鄉
彰化縣	彰化市
彰化縣	二林鎮
彰化縣	鹿港鎮
彰化縣	線西鄉
彰化縣	溪州鄉
彰化縣	芳苑鄉
彰化縣	花壇鄉
彰化縣	秀水鄉

台南市	西港區
台南市	新市區
台南市	東山區
台南市	佳里區
台南市	下營區
台南市	六甲區
台南市	善化區
台南市	麻豆區
台南市	白河區
台南市	新營區
台南市	後壁區
台南市	柳營區
台南市	官田區
台南市	新化區

嘉義縣	太保市
嘉義縣	溪口鄉
嘉義縣	大林鎮
嘉義縣	東石鄉
嘉義縣	布袋鎮
嘉義縣	新港鄉
嘉義縣	朴子市
嘉義縣	鹿草鄉
嘉義縣	中埔鄉
嘉義縣	水上鄉
嘉義縣	六腳鄉
嘉義縣	竹崎鄉
嘉義縣	民雄鄉
嘉義市	西區
嘉義市	東區

雲林縣	斗南鎮
雲林縣	二崙鄉
雲林縣	北港鎮
雲林縣	土庫鎮
雲林縣	東勢鄉
雲林縣	斗六市
雲林縣	台西鄉
雲林縣	元長鄉
雲林縣	水林鄉
雲林縣	四湖鄉
雲林縣	大埤鄉

屏東縣	南州鄉
屏東縣	萬丹鄉
屏東縣	新園鄉
花蓮縣	富里鄉
花蓮縣	玉里鎮
花蓮縣	壽豐鄉
花蓮縣	光復鄉

偵測模型之建立與導入(4/10)

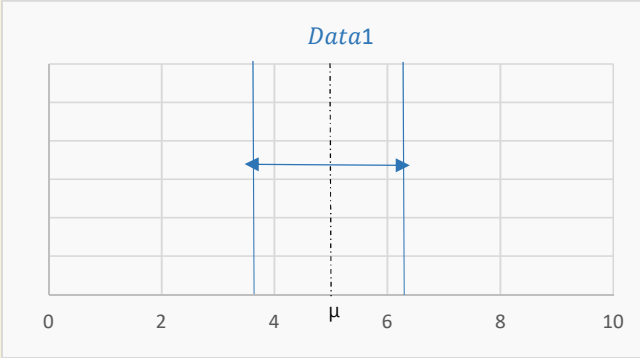
- 第一階段：統計方法判定
- 第二階段：以Adaboost、XGBoost、Decision Tree、Random Forest等七種模型來進行測試評估。

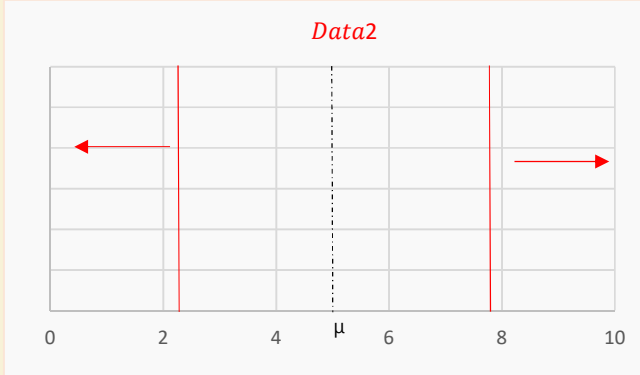
偵測模型之建立與導入(5/10)

統計方法：

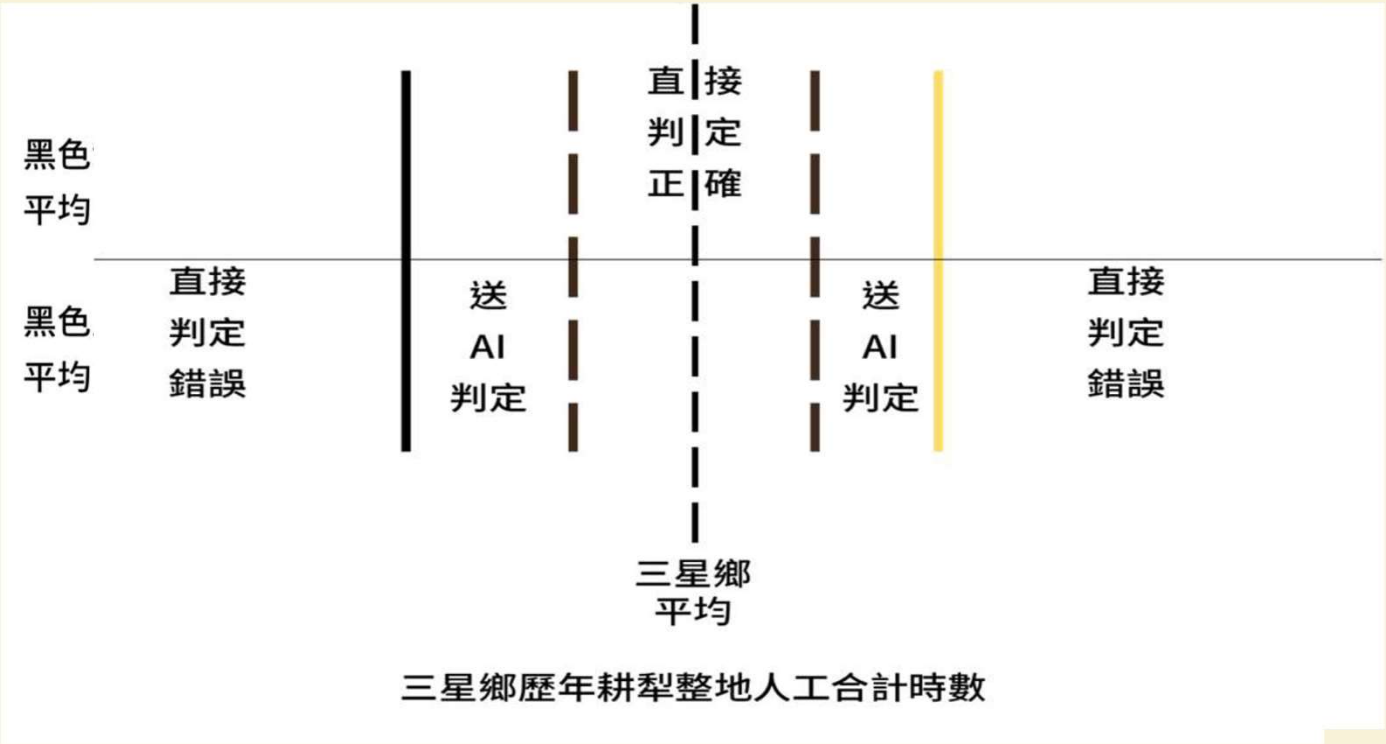
利用暴力法，逐一搜尋從0.01 到0.59 間的各係數，何係數具有最佳錯誤/正確率，即選定為下面二式a、b 值。

- 直接判定正確
  - $Data1 \in [\, x \mid x > Mean - a * StDev \; or \; x < Mean + a * StDev]$
  - a ∈ [0.01,0.59]
  - Max(正確率| a , $Data1$)
- 直接判定錯誤
  - $Data2 \in [\, x \mid x < Mean - b * StDev \; or \; x > Mean + b * StDev]$
  - b ∈ [0.01,0.59]
  - Max(錯誤率| b, $Data2$)
  - Max[Min(x | x ∈ $Data$ for some town),-b]和Min [Max(x | x ∈ $Data$ for some town),b]





偵測模型之建立與導入(6/10)

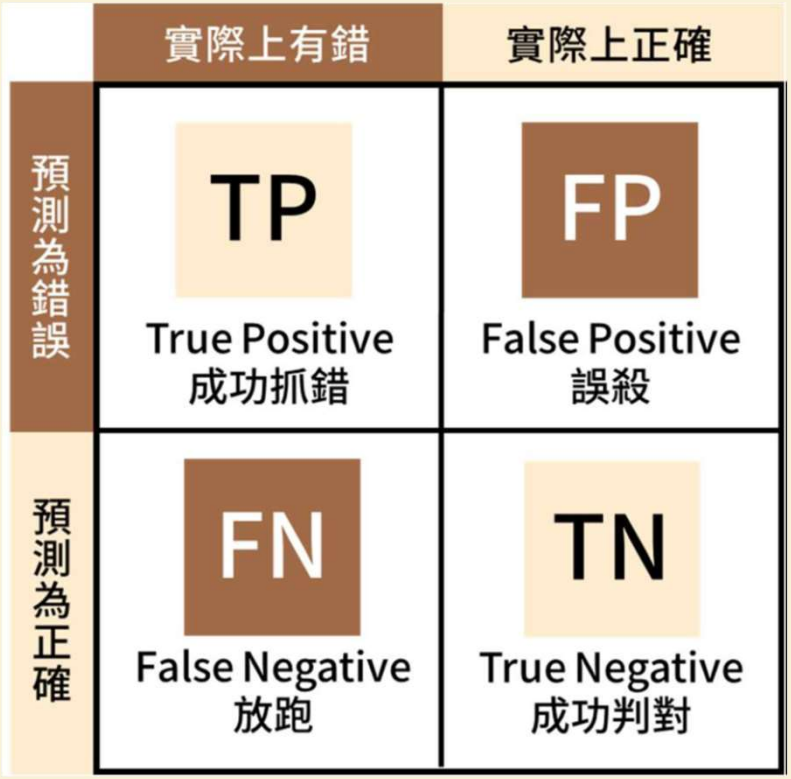


三星鄉歷年耕犁整地人工合計時數

偵測模型之建立與導入(7/10)

混淆矩陣

- 準確性 Accuracy:
  - (TN+TP)/(TN+FN+FP+TP)
- 精度Precision:
  - T*P*/(TP+F*P*)
- 召回率Recall:
  - T*P*/(TP+FN)
- F1 Score:
  - 2 / [(1/Precision) + (1/Recall)]



偵測模型之建立與導入(8/10)

模型選擇- XGBoost 模型

- 以審核表一表現而言，總和各項指標，表現最佳者是 Decision Tree 模型。在其餘各表最佳的模型不一，但值得注意的是，不論在任何表，XGBoost 模型即便不是表現最佳者，其Recall 值皆為最高，意即其抓出錯誤的能力為最佳。

表 2-3 表一之審核 Decision Tree 對比 XGBoost

不包含統計審核模型		
	Decision Tree	XGBoost
TP	41	53
TN	1096	1054
FP	5	47
FN	71	59
Accuracy	0.937	0.913
Precision	0.891	0.53
Recall	0.366	0.4732
F1-Score	0.519	0.5

偵測模型之建立與導入(9/10)

表 2-2 機器學習模型 對比 機器學習加入統計審核模型

左側：僅使用機器學習　右側：加入統計審核										
	表一		表二		表三		表四		表五	
TP	52	88	28	97	73	110	31	111	28	111
TN	1058	1094	1062	1092	1047	1071	1030	1072	1021	1060
FP	43	7	34	4	31	7	49	7	47	8
FN	60	24	89	20	61	25	103	23	117	34
Accuracy	0.915	0.974	0.899	0.980	0.924	0.974	0.875	0.975	0.865	0.965
Precision	0.547	0.926	0.452	0.960	0.702	0.940	0.388	0.941	0.373	0.933
Recall	0.464	0.786	0.239	0.829	0.545	0.815	0.231	0.828	0.193	0.766
F1 Score	0.502	0.850	0.313	0.890	0.613	0.873	0.290	0.881	0.255	0.841
	表六		表七		表八		表九		表十	
TP	20	47	74	105	11	24	35	78	9	14
TN	1109	1141	1072	1086	1174	1179	1090	1113	1185	1190
FP	35	3	26	13	8	3	28	5	7	2
FN	49	22	40	9	20	7	59	17	12	7
Accuracy	0.931	0.979	0.946	0.982	0.977	0.992	0.928	0.982	0.984	0.993
Precision	0.364	0.940	0.740	0.890	0.579	0.889	0.556	0.940	0.563	0.875
Recall	0.290	0.681	0.649	0.921	0.355	0.774	0.372	0.821	0.429	0.667
F1 Score	0.323	0.790	0.692	0.905	0.440	0.828	0.446	0.876	0.486	0.757

偵測模型之建立與導入(10/10)

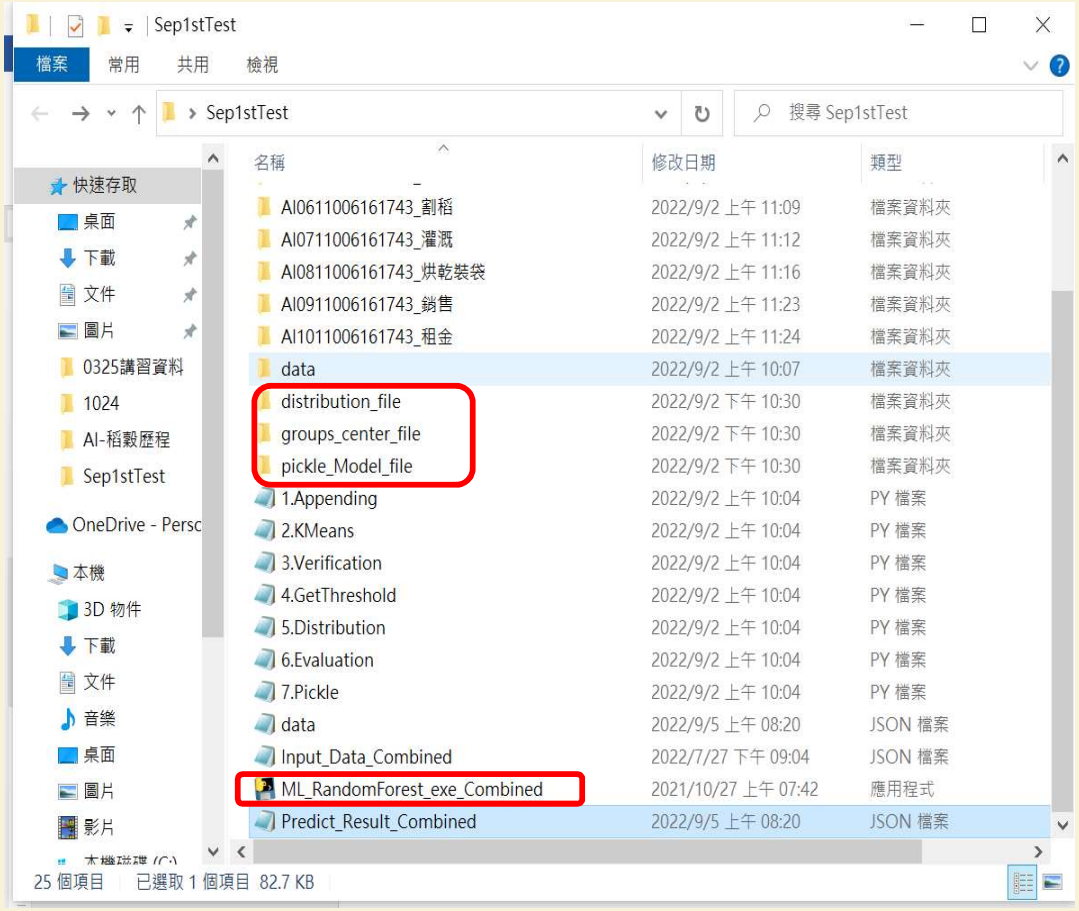
總結：

模型判斷為錯誤的資料，真正錯誤的占八成。而真正錯誤的資料中，有六成可被此模型找出。

表 3-2 2021-1 期資料審核結果

	表一	表二	表三	表四	表五	表六	表七	表八	表九	表十	Total
分群數	8	8	7	5	3	4	6	3	8	3	
資料總數	629	629	629	629	629	629	629	629	629	629	6290
正確資料	263	243	191	167	211	429	274	511	392	583	3264
錯誤資料	366	386	438	462	418	200	355	118	237	46	3026
TP	225	200	313	217	203	86	278	93	177	29	1821
TN	216	184	156	119	167	392	204	496	292	567	2793
FP	47	59	35	48	44	37	70	15	100	16	471
FN	141	186	125	245	215	114	77	25	60	17	1205
Accuracy	0.701	0.610	0.746	0.534	0.588	0.760	0.766	0.936	0.746	0.948	0.734
Precision	0.827	0.772	0.899	0.819	0.822	0.699	0.799	0.861	0.639	0.644	0.795
Recall	0.615	0.518	0.715	0.470	0.486	0.430	0.783	0.788	0.747	0.630	0.602
F1-Score	0.705	0.620	0.796	0.597	0.611	0.533	0.791	0.823	0.689	0.637	0.685

上線後系統實作(1/3)



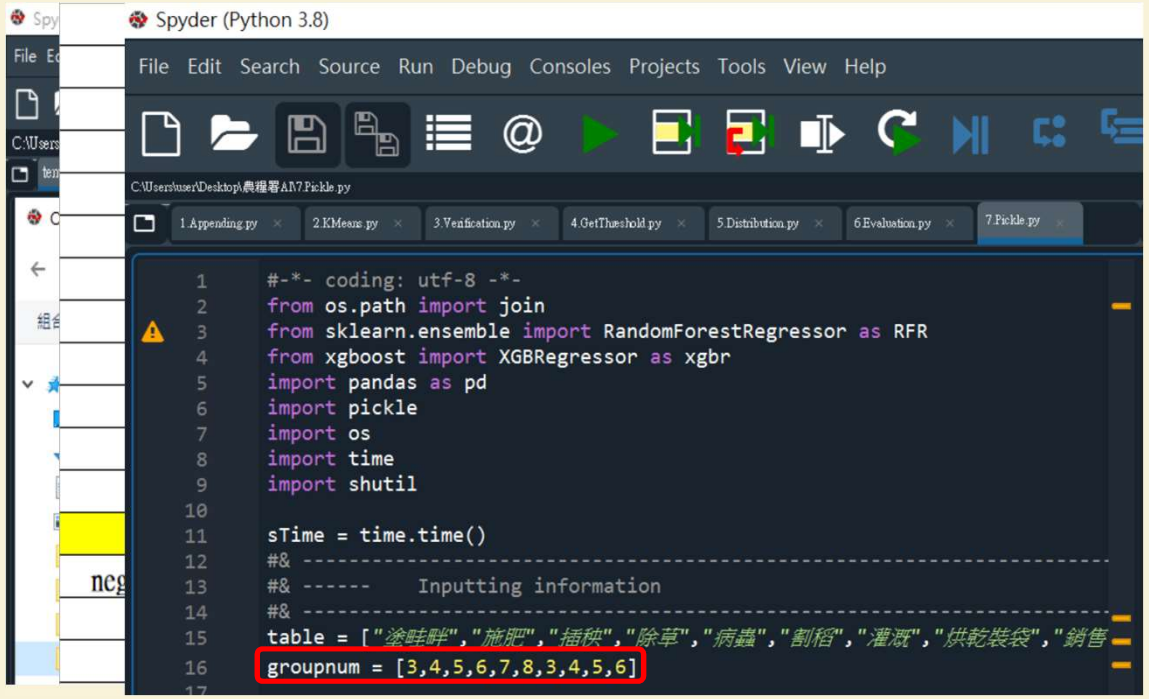
廠商：原系統藉由呼叫執行檔後產生新的Predict Result_Combined

更新模型：由系統中匯出訓練資料，再由重新訓練後之分群資訊、鄉鎮閥值、pickle檔送交系統商

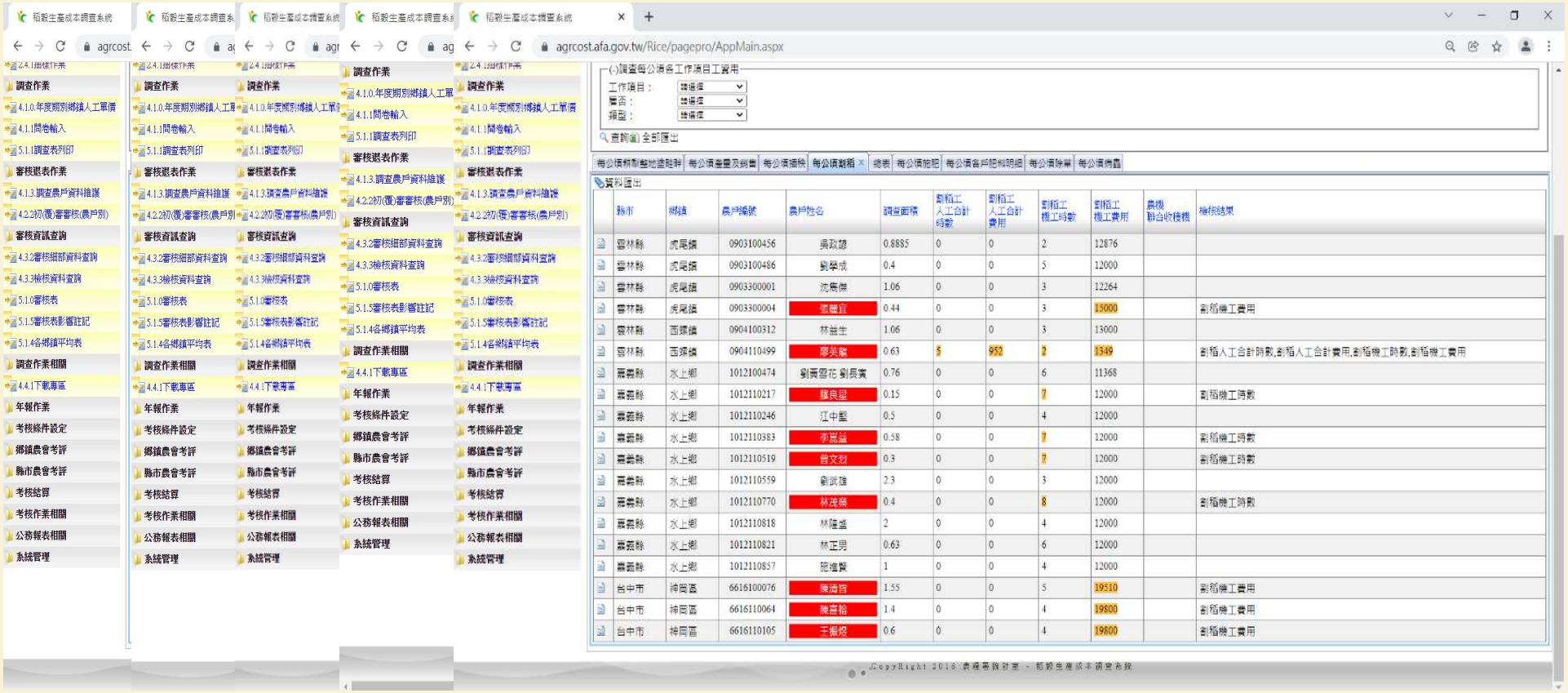
上線後系統實作(2/3)

訓練時畫面

- 步驟一：由spyder執行1~6
- 步驟二：選擇較好的群數
- 步驟三：更新程式(7.Pickle)檔中群數後執行



上線後系統實作(3/3)



調查員、審核員系統畫面

總結與未來展望

- **調查源頭控管**：調查誤差有許多內涵，審核作業事實上可處理的有限，若要精進資料品質，除審核作業要做好外，最好是強化調查源頭會比較有效率。

- **促進產官學合作**：將產業界、學界與使用單位整合，可提升專案功能，強化研究的實用性。

- **鏈結資通訊技術**：未來或許科技再進步，可考慮第一線的輔助系統。而本系統的預測精準度也期待能有所精進。

謝謝聆聽