資料探勘

B074011005 王濬瑋　B094020017 郭楨君　B094020046 黃奕瑋

B124020018 劉佳瑜　B124020027 陳闈霆

CONTENT

01 KNN實作

02 相關研究

03 其他演算法

04 總結

KNN實作

KNN流程

<資料前處理>
缺失值填補
正規化資料

→

輸入
DATASET和K值

→

計算歐式距離
存至LIST中

SORT LIST
取前K小的值

→

判斷結果
0與1何者較多

KNN流程

<資料前處理>
缺失值填補
正規化資料

缺失值填補
將**GLUCOSE**、**BLOODPRESSURE**、**SKINTHICKNESS**、**INSULIN**、**BMI**
欄位中為**0**的值替換為平均值。

SORT LIST
取前**K**小的值

判斷結果
0與**1**何者較多

KNN流程

<資料前處理>
缺失值填補
正規化資料

正規化資料
將所有的值投影到**0~1**之間

計算歐式距離
存至LIST中

SORT LIST
取前K小的值

判斷結果
0與1何者較多

KNN流程

<資料前處理>
缺失值填補
正規化資料

→

輸入
DATASET和K值

→

計算歐式距離
存至LIST中

SORT LIST
取前K小的值

→

判斷結果
0與1何者較多

KNN流程

新增 **THRESHOLD** 變數
TRUE 數量 **>** **(K / THRESHOLD)**
就判定為 **TRUE**

T和K值

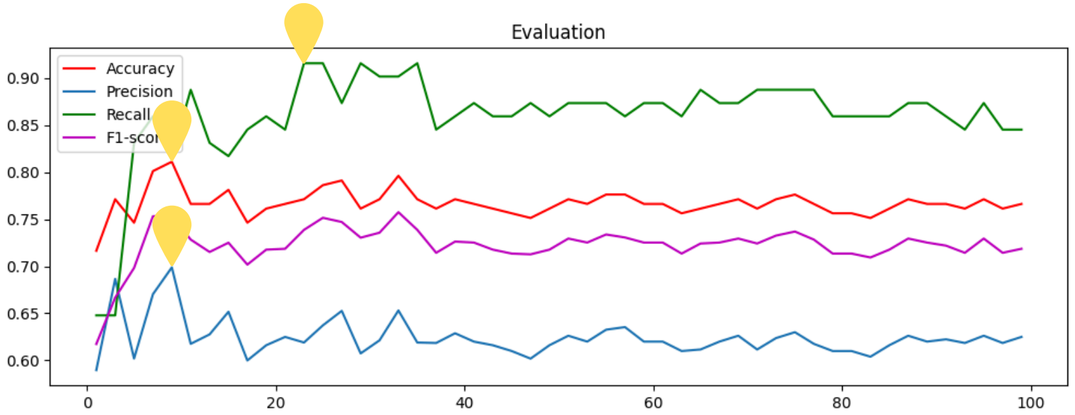
計算歐式距離
存至LIST中

SORT LIST
取前K小的值

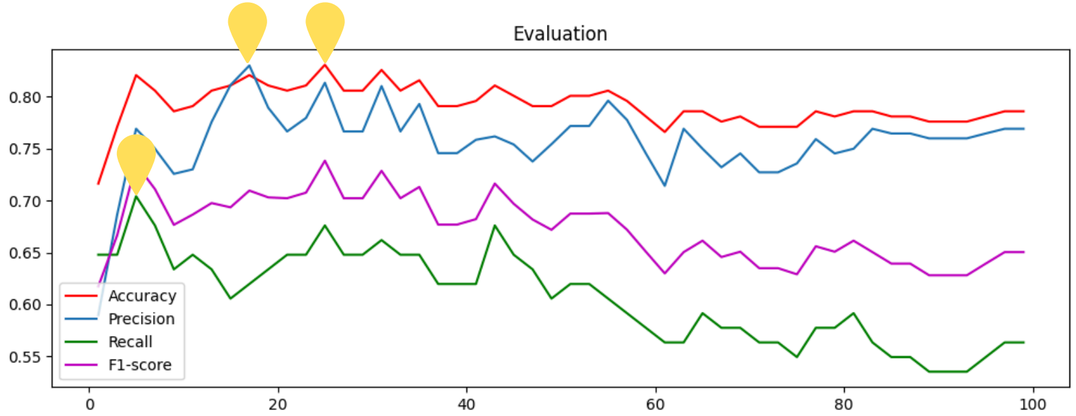
判斷結果
0與1何者較多

找出最佳K值

實驗A	Accuracy	Precision	Recall	F1-score
K = 25 Threshold = 2	0.830845	0.813559	0.676056	0.738461
K = 17 Threshold = 2	0.820895	0.830188	0.619718	0.709677
K = 9 Threshold = 3	0.810945	0.698795	0.816901	0.753246
K = 23 Threshold = 3	0.771144	0.619047	0.915492	0.738636



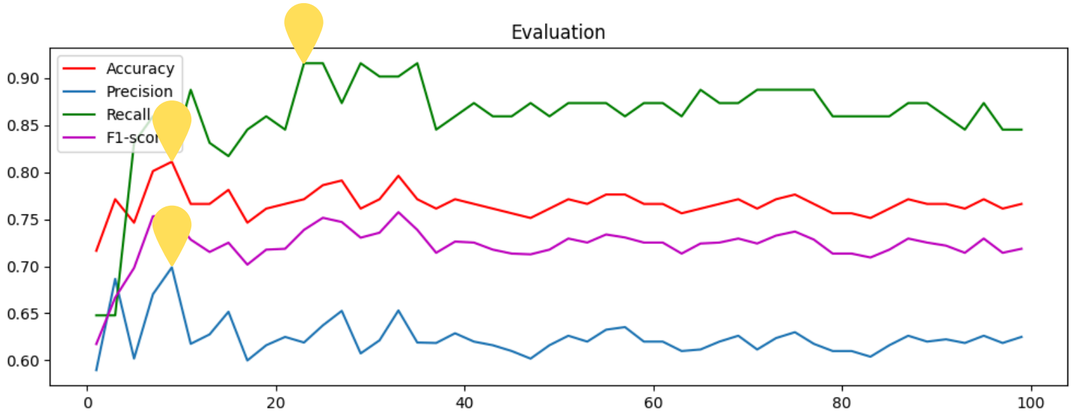
THERSHOLD=3



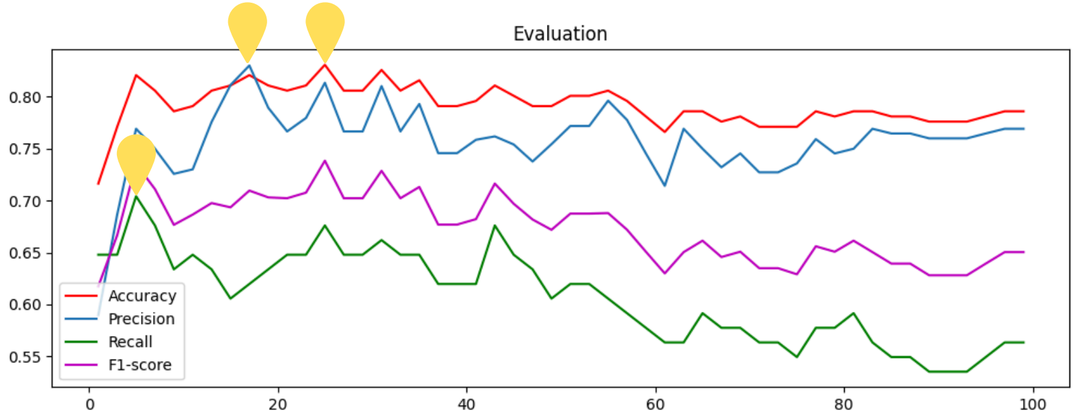
THERSHOLD=2

找出最佳K值

實驗B	Accuracy	Precision	Recall	F1-score
K = 61 Threshold = 2	0.79000	0.90000	0.486486	0.631578
K = 13 Threshold = 2	0.78000	0.727272	0.648648	0.685714
K = 9 Threshold = 3	0.80000	0.697674	0.810810	0.75000
K = 35 Threshold = 3	0.78000	0.647058	0.891891	0.749999



THERSHOLD=3



THERSHOLD=2

相關研究

通過整合PCA和K-MEANS技術改善糖尿病預測的邏輯回歸模型

(2019,CHANGSHENG ZHU A, CHRISTIAN UWA IDEMUDIA A, WENFANG FENG B)

Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques

整合PCA和K-means改善 糖尿病預測的邏輯回歸模型

PCA ✚ K-means ✚ Logistic Regression

其他可參考重點

- **缺失值：血糖濃度、血壓為0，進行處理。**
- **正規化：本篇論文有實行數據正規化**
- **轉換數據：將「懷孕次數」轉換為名義特徵（0或1表示是否懷孕過)**

01 LOGISTIC REGRESSION

02 PCA

03 DECISION TREE

04 RANDOM FOREST

05 MLP

LOGISTIC REGRESSION

未調参數

糖尿病患者誤判為無病的機率高

實驗A

Accuracy： 0.80

Recall ： 0.57

Precision： 0.82

F1-Score ： 0.67

實驗B

Accuracy： 0.83

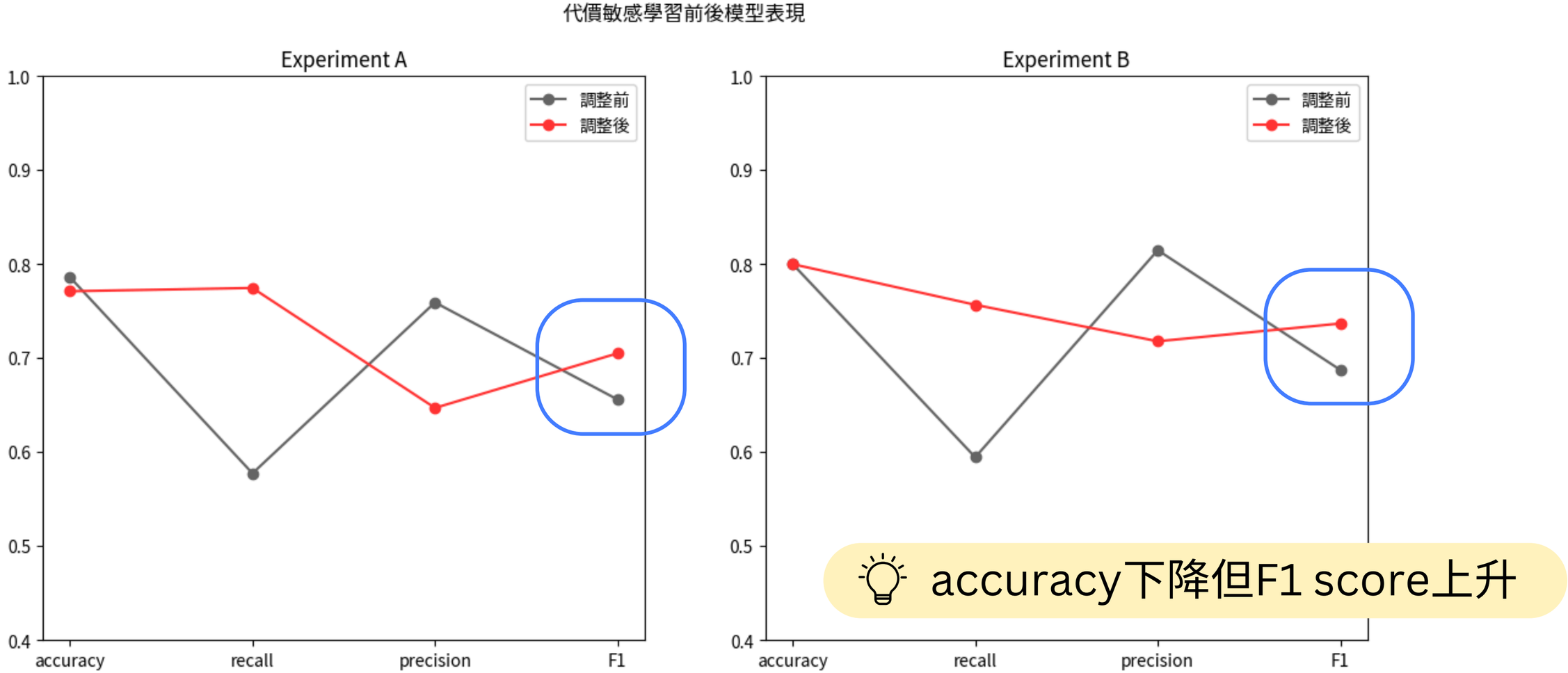
Recall ： 0.59

Precision： 0.91

F1-Score ： 0.72

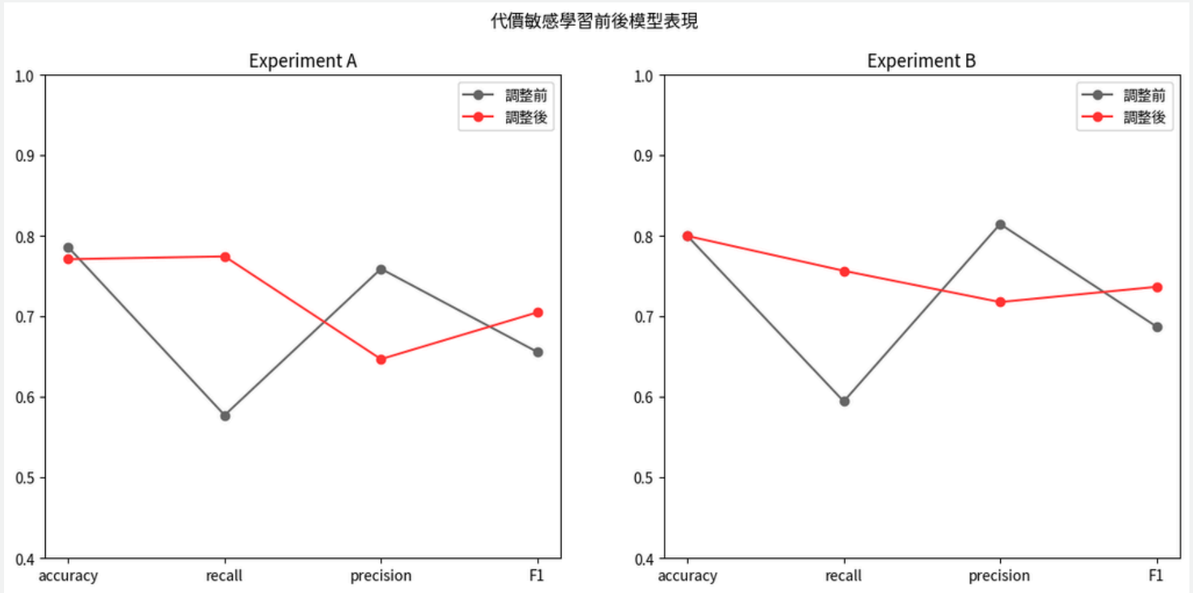
LOGISTIC REGRESSION

正樣本和負樣本的比率調整



accuracy下降但F1 score上升

代價敏感學習



正樣本和負樣本的比率調整
2:1

實驗A

GridSearchCV

最佳參數

'C': 0.042813323987193 94

'penalty': 'l2'

最佳交叉驗證得分

0.77

準確率

accuracy: 0.7711442786069652

recall: 0.5352112676056338

precision: 0.745098039215686 3

F1 score: 0.6229508196721312

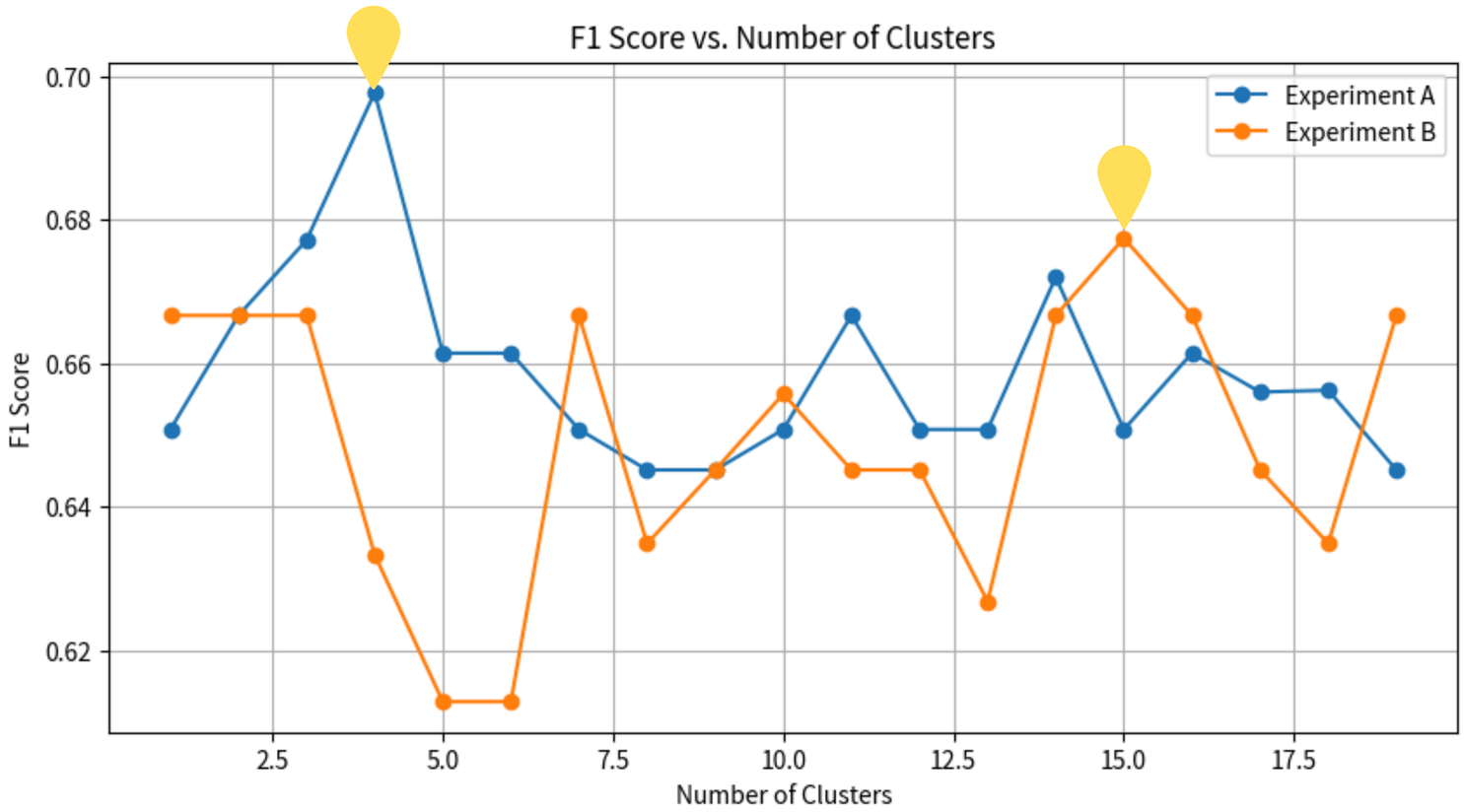
PCA + K-means
Logistic Regression

PCA n_components = 0.95

實驗A k-means cluster = 4
實驗B k-means cluster = 15

實驗A:
'Accuracy': 0.80,
'Recall': 0.63
'Precision': 0.77
'F1 Score': 0.69

實驗B:
'Accuracy': 0.80,
'Recall': 0.56
'Precision': 0.84
'F1 Score': 0.67



DECISION TREE

GridSearchCV

實驗A

最佳參數

'max_depth': 5

'min_samples_leaf': 17

'min_samples_split': 2

最佳交叉驗證得分

0.7583

準確率

accuracy:	0.76
recall:	0.74
precision:	0.73
F1 score:	0.74

實驗B

最佳參數

'max_depth': 7

'min_samples_leaf': 16

'min_samples_split': 2

最佳交叉驗證得分

0.74099

準確率

accuracy:	0.76
recall:	0.74
precision:	0.74
F1 score:	0.74

RANDOM FOREST

RandomizedSearchCV

實驗A

最佳參數

'n_estimators': 79
'min_samples_split': 29
'min_samples_leaf': 30
'max_features': 'sqrt'
'max_depth': 22
'bootstrap': True

準確率

accuracy:	0.76
recall:	0.73
precision:	0.76
F1 score:	0.71

實驗B

最佳參數

'n_estimators': 324
'min_samples_split': 8
'min_samples_leaf': 2
'max_features': 'auto'
'max_depth': 38
'bootstrap': True

準確率

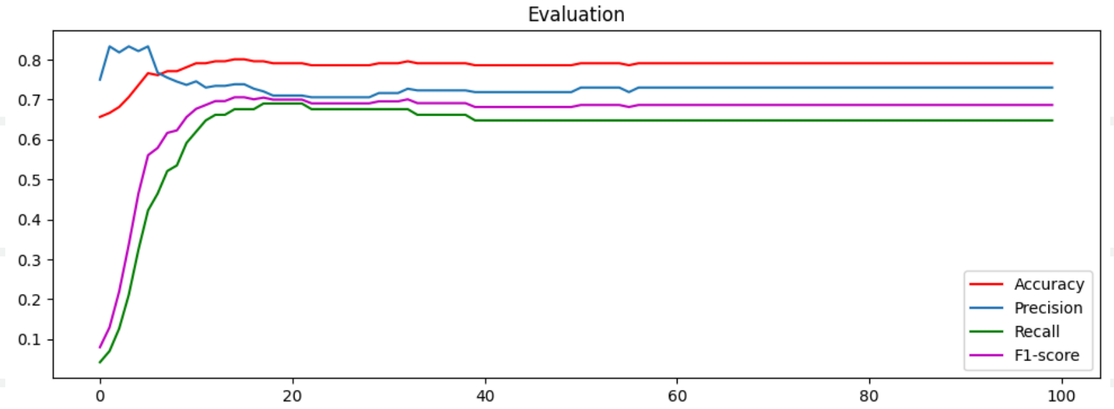
accuracy:	0.79
recall:	0.74
precision:	0.78
F1 score:	0.76

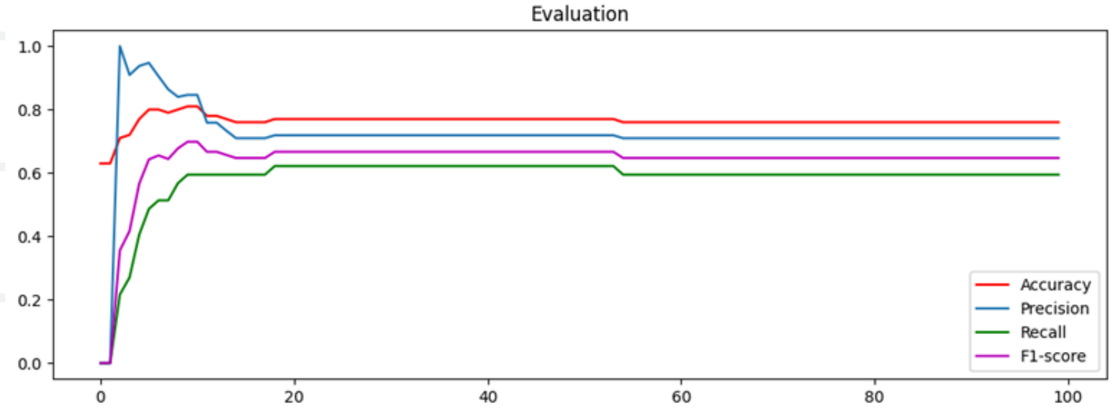
MLP

模型架構 ：

FEATURES_DIM -> 8 -> 1 -> RELU

實驗A	實驗B
accuracy: 0.800995	accuracy: 0.810000
recall: 0.676056	recall: 0.594595
precision: 0.738462	precision: 0.846154
F1 score: 0.705882	F1 score: 0.698413





結論

01 各模型比較:F1 SCORE

02 總結

取各模型最高 F1 分數比較

實驗A	Accuracy	Precision	Recall	F1-score
KNN	0.81	0.69	0.81	0.75
Logistic Regression	0.77	0.64	0.77	0.70
Decision Tree	0.76	0.73	0.74	0.74
Random Forest	0.76	0.76	0.70	0.71
MLP	0.80	0.73	0.67	0.70
PCA+K-means	0.71	0.57	0.78	0.66
PCA+Kmeans+ Logostic Regression	0.8	0.77	0.63	0.69

實驗B	Accuracy	Precision	Recall	F1-score
KNN	0.80	0.69	0.81	0.75
Logistic Regression	0.8	0.71	0.75	0.73
Decision Tree	0.76	0.74	0.74	0.74
Random Forest	0.79	0.74	0.78	0.76
MLP	0.81	0.84	0.59	0.69
PCA+K-means	0.73	0.61	0.72	0.66
PCA+Kmeans+ Logostic Regression	0.8	0.84	0.56	0.67

總結

訓練資料量多寡影響模型表現

- KNN 的 Recall 值皆比較高，因為我們有設 Threshold
- 實驗A訓練資料較少，其他模型較難訓練

訓練資料量少時：KNN

訓練資料量多時：視情況選擇演算法模型

參考文獻

https://pyecontech.com/2020/04/19/knn/https://medium.com/@SCU.Datascientist/python%E5%AD%B8%E7%BF%92%E7%AD%86%E8%A8%98-knn-k-nearest-neighbor-531a95336f71

https://ithelp.ithome.com.tw/m/articles/10269006

https://aws.amazon.com/tw/what-is/logistic-regression/

https://medium.com/@whchang022/%E6%B1%BA%E7%AD%96%E6%A8%B9-decision-tree-E5%AD%B8%E7%BF%92%E7%AD%86%E8%A8%98-e763c5c5b933

https://www.sciencedirect.com/science/article/pii/S2352914819300139

https://ithelp.ithome.com.tw/articles/10272586?sc=hot

分工表

B074011005 王濬瑋	**程式、書面報告**
B094020017 郭楨君	**程式、書面報告**
B094020046 黃奕瑋	**程式、書面報告**
B124020018 劉佳瑜	**投影片**
B124020027 陳闈霆	**上台報告**

THANK YOU