

資料探勘書面報告

資工系 41071102H 徐敏皓

目錄

頁次

專題實作資料集.....	1
採用 data mining 方法.....	5
程式/環境設定，執行方式說明.....	6
改變控制參數/技術說明.....	7
評估方法.....	7
結果及討論.....	7
繳交於 Kaggle 上的結果.....	10

專題實作資料集

採用的資料集為 [Exploring Mental Health Data](#)。

我有先觀察一些資料的資訊：

1. 網站提供的 train data 和 test data 資料和欄位的數目（圖一）。
2. train data 和 test data 前五筆的資料（圖二、圖三）確認了 test data 確實只有少”Depression”（資料集中的 class label）。
3. train data 每個欄位的 missing value 有多少（圖四）。主要鎖定在 Numerical 欄位有 missing value 的部份去觀察分布情形，後續選擇適合的填補策略。
4. 每個欄位的資料型態（圖五）。配合第三點說明的，要去查看有 missing value 的 Numerical 欄位的分布。
5. 印出 5 個含有 missing value 的 Numerical 欄位（圖六～圖十）。其中可發現圖八分布的情形特別不均勻，因此決定在後續前處理時用”median”的方式填補，原因是中位數對極端值不敏感，更能應對那個不均勻的欄位，且對於其他均勻分布的欄位，中位數也能提供合理的替代值，不會造成明顯偏差。
6. 觀察 class label 分布的情形（圖十一）。從圖可發現 class label 分布不均，後續模型用的評估方式就是根據此觀察而決定的。

```
Shape of train data: (140700, 20)
Shape of test data: (93800, 19)
```

圖一 資料和欄位的數目

	id	Name	Gender	Age	City	Working Professional or Student	Profession	Academic Pressure	Work Pressure	CGPA
0	0	Aaradhya	Female	49.0	Ludhiana	Working Professional	Chef	NaN	5.0	NaN
1	1	Vivan	Male	26.0	Varanasi	Working Professional	Teacher	NaN	4.0	NaN
2	2	Yuvraj	Male	33.0	Visakhapatnam	Student	NaN	5.0	NaN	8.97
3	3	Yuvraj	Male	22.0	Mumbai	Working Professional	Teacher	NaN	5.0	NaN
4	4	Rhea	Female	30.0	Kanpur	Working Professional	Business Analyst	NaN	1.0	NaN

	Study Satisfaction	Job Satisfaction	Sleep Duration	Dietary Habits	Degree	Have you ever had suicidal thoughts ?	Work/Study Hours	Financial Stress	Family History of Mental Illness	Depression
	NaN	2.0	More than 8 hours	Healthy	BHM	No	1.0	2.0	No	0
	NaN	3.0	Less than 5 hours	Unhealthy	LLB	Yes	7.0	3.0	No	1
	2.0	NaN	5-6 hours	Healthy	B.Pharm	Yes	3.0	1.0	No	1
	NaN	1.0	Less than 5 hours	Moderate	BBA	Yes	10.0	1.0	Yes	1
	NaN	1.0	5-6 hours	Unhealthy	BBA	Yes	9.0	4.0	Yes	0

圖二 train data 前五筆的資料

	id	Name	Gender	Age	City	Working Professional or Student	Profession	Academic Pressure	Work Pressure	CGPA
0	140700	Shivam	Male	53.0	Visakhapatnam	Working Professional	Judge	NaN	2.0	NaN
1	140701	Sanya	Female	58.0	Kolkata	Working Professional	Educational Consultant	NaN	2.0	NaN
2	140702	Yash	Male	53.0	Jaipur	Working Professional	Teacher	NaN	4.0	NaN
3	140703	Nalini	Female	23.0	Rajkot	Student	NaN	5.0	NaN	6.84
4	140704	Shaurya	Male	47.0	Kalyan	Working Professional	Teacher	NaN	5.0	NaN

	Study Satisfaction	Job Satisfaction	Sleep Duration	Dietary Habits	Degree	Have you ever had suicidal thoughts ?	Work/Study Hours	Financial Stress	Family History of Mental Illness
	NaN	5.0	Less than 5 hours	Moderate	LLB	No	9.0	3.0	Yes
	NaN	4.0	Less than 5 hours	Moderate	B.Ed	No	6.0	4.0	No
	NaN	1.0	7-8 hours	Moderate	B.Arch	Yes	12.0	4.0	No
	1.0	NaN	More than 8 hours	Moderate	BSc	Yes	10.0	4.0	No
	NaN	5.0	7-8 hours	Moderate	BCA	Yes	3.0	4.0	No

圖三 test data 前五筆的資料

```

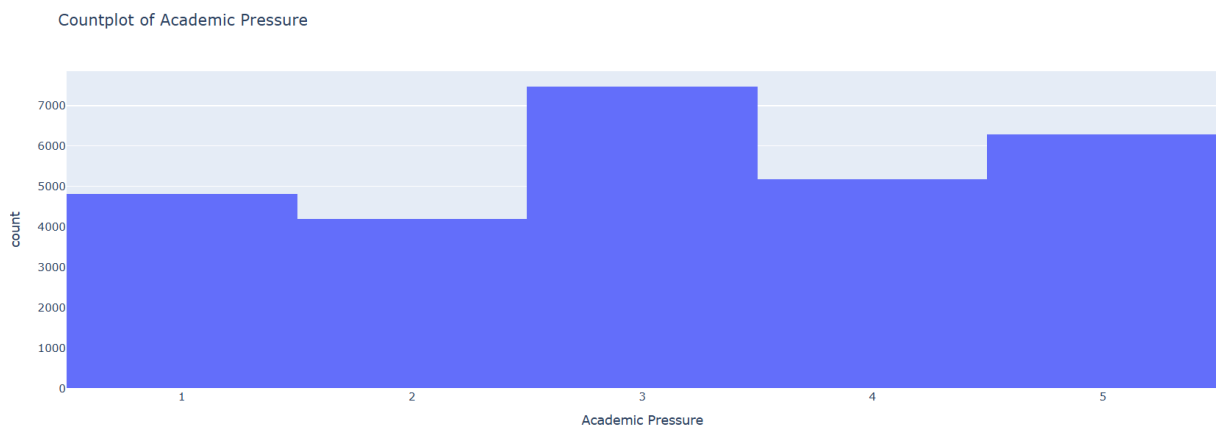
The number of missing values before data processing:
id                                0
Name                              0
Gender                            0
Age                               0
City                              0
Working Professional or Student   0
Profession                        36630
Academic Pressure                 112803
Work Pressure                     27918
CGPA                             112802
Study Satisfaction                112803
Job Satisfaction                  27910
Sleep Duration                    0
Dietary Habits                    4
Degree                            2
Have you ever had suicidal thoughts ? 0
Work/Study Hours                  0
Financial Stress                   4
Family History of Mental Illness  0
Depression                        0
dtype: int64

```

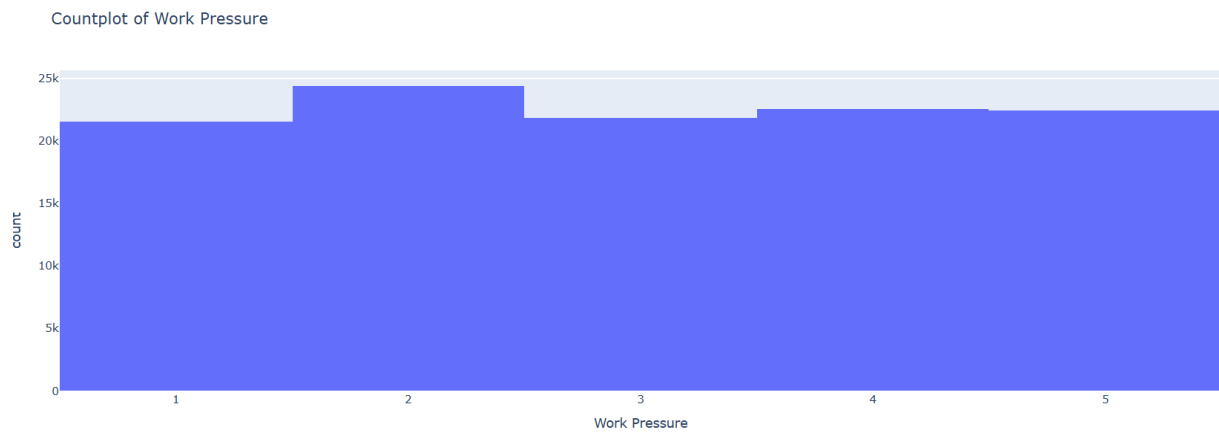
圖四 train data 每個欄位的 missing value 數目

	0
id	int64
Name	object
Gender	object
Age	float64
City	object
Working Professional or Student	object
Profession	object
Academic Pressure	float64
Work Pressure	float64
CGPA	float64
Study Satisfaction	float64
Job Satisfaction	float64
Sleep Duration	object
Dietary Habits	object
Degree	object
Have you ever had suicidal thoughts ?	object
Work/Study Hours	float64
Financial Stress	float64
Family History of Mental Illness	object
Depression	int64

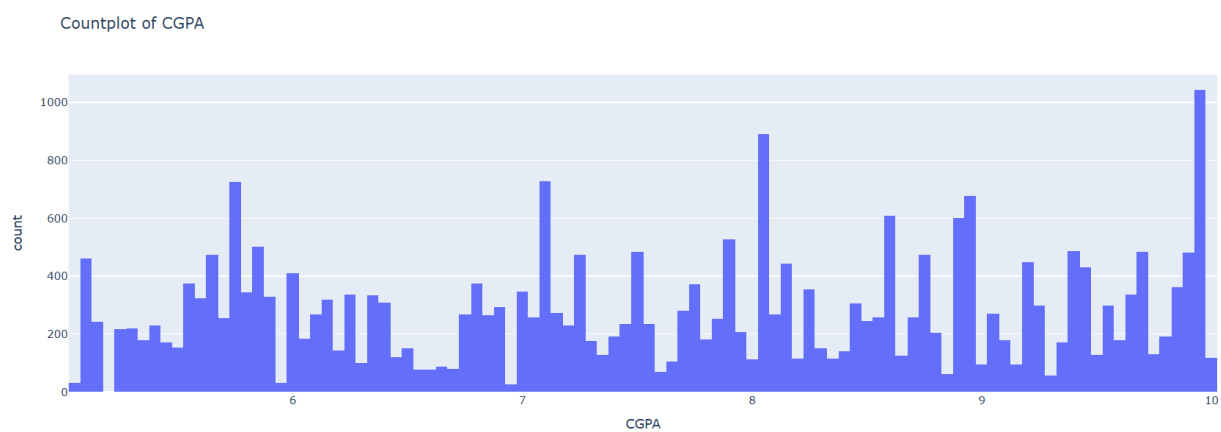
圖五 每個欄位的資料型態



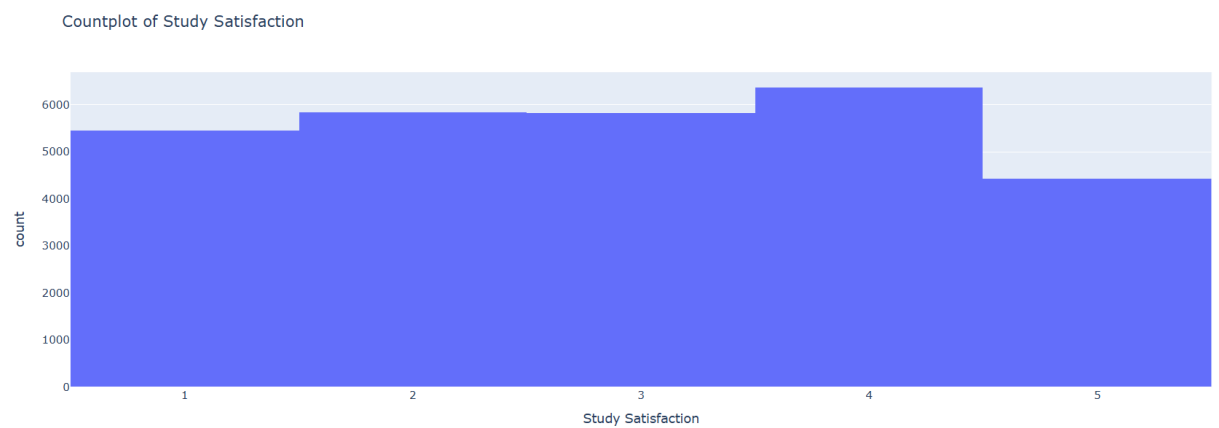
圖六 Academic Pressure



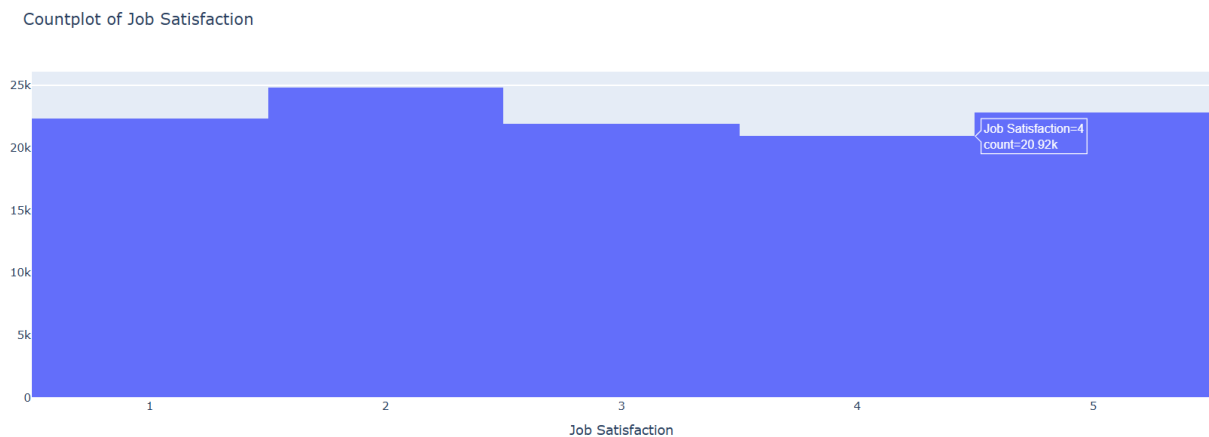
圖七 Work Pressure



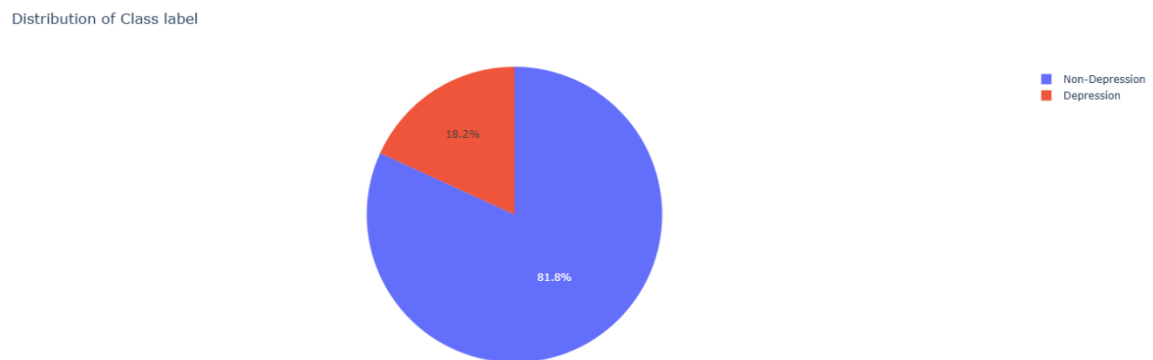
圖八 CGPA



圖九 Study Satisfaction



圖十 Job Satisfaction



圖十一 class label 比例圖

採用 data mining 方法

1. 前處理（圖十二）：

- 將”id”和”Name”從 train data 和 test data 中移除。
- 將 train data 中的 class label 和特徵分離。
- 將欄位分成 Numerical 欄位和 Categorical 欄位分開處理：
 - Numerical 欄位根據上面「專題欄位資料第五點」的結論把 missing value 用”median”填補。另外，這些數值不須做標準化，因為我使用的演算法是 Decision tree 與 Random forest，兩者不會計算特徵之間的距離。
 - Categorical 欄位用”missing”這個字串填補 missing value，接著再做 ordinal encoded 把列舉值變成數字。
- 將處理後的 Numerical 欄位和 Categorical 欄位合併。

2. 實作演算法：由於資料集的 class label 只存在 2 種值，0 和 1，因此用 classifier 來實作比較適合，因為 classifier 主要處理的就是列舉類型的 class label。原本我只有使用 Decision tree。後來 Demo 時老師有推薦我使用 Random forest 的方法用來與 Decision tree 做比較。因此，最後使用 Decision tree 和 Random forest。

	Age	Academic Pressure	Work Pressure	CGPA	Study Satisfaction	Job Satisfaction	Work/Study Hours	Financial Stress
0	49.0	3.0	5.0	7.77	3.0	2.0	1.0	2.0
1	26.0	3.0	4.0	7.77	3.0	3.0	7.0	3.0
2	33.0	5.0	3.0	8.97	2.0	3.0	3.0	1.0
3	22.0	3.0	5.0	7.77	3.0	1.0	10.0	1.0
4	30.0	3.0	1.0	7.77	3.0	1.0	9.0	4.0

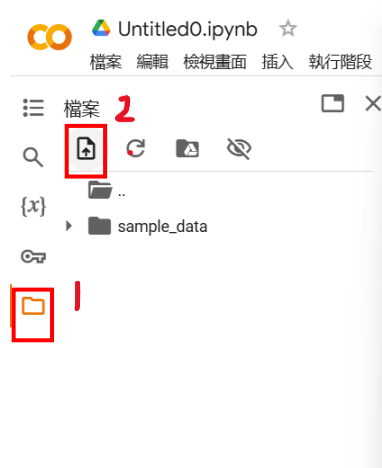
	Gender	City	Working Professional or Student	Profession	Sleep Duration	Dietary Habits	Degree	Have you ever had suicidal thoughts ?	Family History of Mental illness
	0.0	50.0	1.0	10.0	29.0	7.0	33.0	0.0	0.0
	1.0	93.0	1.0	55.0	27.0	20.0	63.0	1.0	0.0
	1.0	97.0	0.0	64.0	15.0	7.0	21.0	1.0	0.0
	1.0	64.0	1.0	55.0	27.0	15.0	28.0	1.0	1.0
	0.0	37.0	1.0	9.0	15.0	20.0	28.0	1.0	1.0

圖十二 前處理後的前 5 筆資料

程式/環境設定，執行方式說明

因為我程式執行是在 colab 上，colab 已經有安裝過這個專題需要用的 package 以及 module，因此不須再安裝。

程式執行的部分需要把 train.cav 和 test.csv 上傳到雲端，操作步驟如圖十三。之後從頭一行一行執行即可。



圖十三 先找到畫面最左側標示為 1 的按鈕，按下去後再按標示為 2 的按鈕上傳檔案

改變控制參數/技術說明

控制參數 1：“max_depth”。樹的深度會影響模型的擬合能力：過淺的深度容易導致 underfitting，而過深的深度可能導致 overfitting。

技術說明 1：用 cross_val_score 這個 function 做 10 次交叉驗證，從 1 到 30 這個範圍一個一個嘗試。

控制參數 2：“criterion”，想確認用“gini”或是用“entropy”的方式計算 impurity 會不會有影響。其中 Gini 的計算偏向節點中錯誤分類的概率，而 Entropy 更強調信息增益的最大化。

技術說明 2：調整參數“criterion”，一個是使用“gini”，另一個是使用“entropy”，其餘不變。

控制參數 3：將單棵樹擴展為多棵樹（Random forest），想比較多棵樹是否會比單棵樹的準確性來的更好。其中 Random Forest 是基於多棵樹的投票結果，通過隨機特徵選擇和 Bagging 技術，能降低單棵樹的 overfitting 問題，提高模型的穩健性。

技術說明 3：「技術說明 1」原本是用 Decision tree 來做，現在改成為 Random forest 來做。

評估方法

評估方法 1：

- cross_val_score 這個 function 中我選擇的評估方式是用“f1_weighted”，因為 F1 weighted score 能綜合考量每個類別的表現，避免因類別分布不均導致評估偏向大類別，對於此資料集非常合適。
- 樹的最佳深度會跟前一次交叉驗證最好的相比，如果相差小於 0.001 就不會更新最佳深度，因為差距已經夠小了。
- 把整個 train data 分成 training set 和 validation set 用以評估模型的效果。

評估方法 2：與「評估方法 1」相同。

評估方法 3：與「評估方法 1」相同。

結果及討論

結果 1：

- cross_val_score 執行後如圖十四，可看出剛開始樹的深度很淺時有 underfitting 的情形，大約超過 10 以後 f1 weighted score 開始下降，這表示訓練集表現提升，但測試集表現下降，反映出模型有 overfitting 的情形。
- 最佳的深度為 9。
- 模型的結果如圖十五，可看出 Non-Depression 的 precision, recall, f1 score 都比 Depression 高，這是因為 class label 分布不均的關係。而整體模型的 weighted avg f1 score 為 0.93，看起來是相當不錯的。

結果 2：

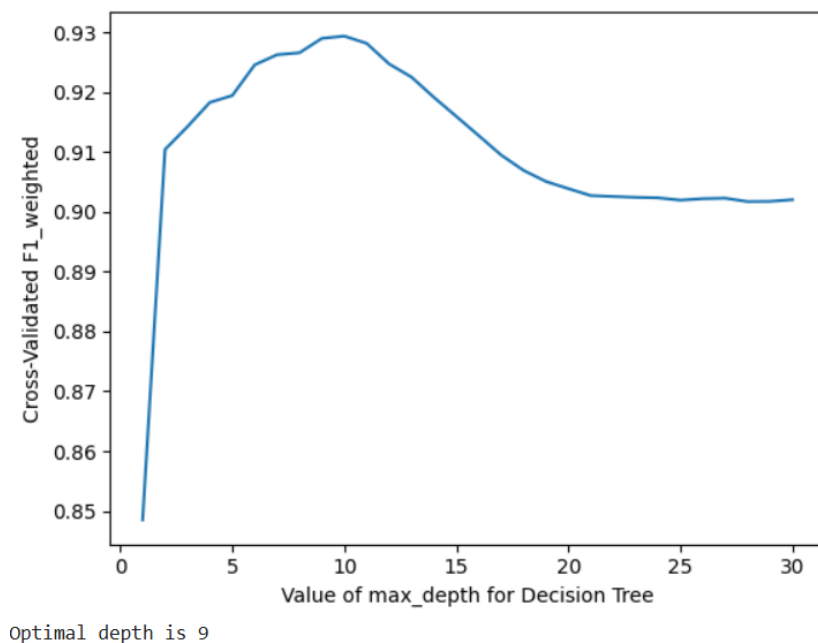
- `cross_val_score` 執行後如圖十六，整個圖形的趨勢與圖十四相近。
- 最佳深度為 10。
- 模型的結果如圖十七，發現 weighted avg f1 score 一樣為 0.93，由於資料特徵和分布的特性，"gini" 和 "entropy" 的計算差異對於分裂結果的影響不明顯，因此最終 F1 weighted score 相近。。

結果 3：

- `cross_val_score` 執行後如圖十八，可看出剛開始樹的深度很淺時有 underfitting 的情形，大約超過 9 以後 f1 score 有持平的現象。
- 最佳的深度為 9。
- 模型的結果如圖十九，發現單看 weighted avg f1 score 一樣為 0.93，但 accuracy 的部分和 Decision tree 比有上升 0.01，這是因為 Random Forest 有 ensemble learning 的特性，使模型在多次隨機特徵選擇和投票中降低了對訓練數據的 overfitting，因而 accuracy 上升 0.01。F1 weighted score 不變可能是因為小類別的表現未明顯改善。

總結：

調整了 Decision tree 的深度與分裂標準，並比較了單棵樹與隨機森林的性能。結果顯示，最佳深度為 9-10，分裂標準的選擇對模型影響甚微；Random forest 雖在 F1 weighted score 上與 Decision tree 持平，但 accuracy 提升了 0.01，顯示多棵樹在大類別預測上的穩定性更高。然而，小類別的表現仍受類別不均影響，未來可考慮資料平衡技術或其他集成學習方法來進一步改善。

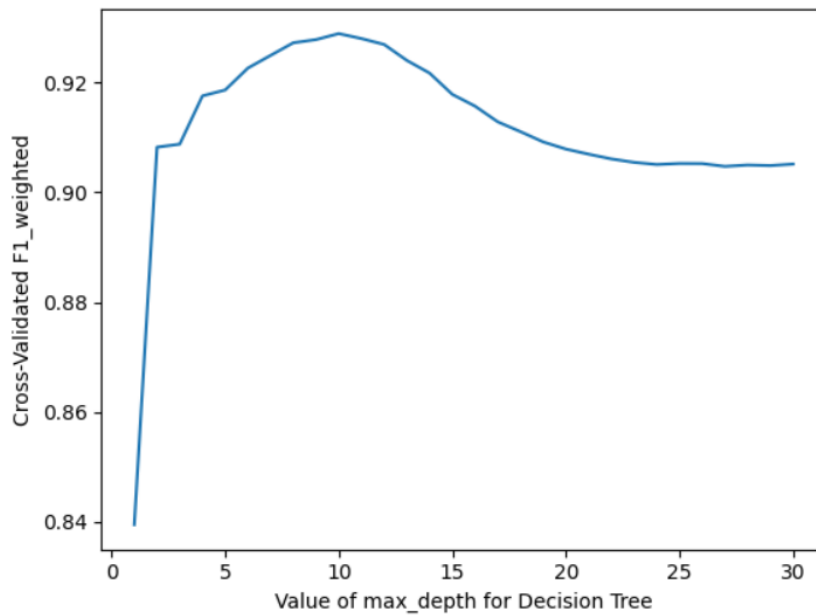


圖十四 用”gini”計算 impurity，找出最佳深度

Validation Set Classification Report:

	precision	recall	f1-score	support
Non-Depression	0.96	0.96	0.96	22986
Depression	0.81	0.80	0.81	5154
accuracy			0.93	28140
macro avg	0.88	0.88	0.88	28140
weighted avg	0.93	0.93	0.93	28140

圖十五 用”gini”計算 impurity，模型訓練的結果



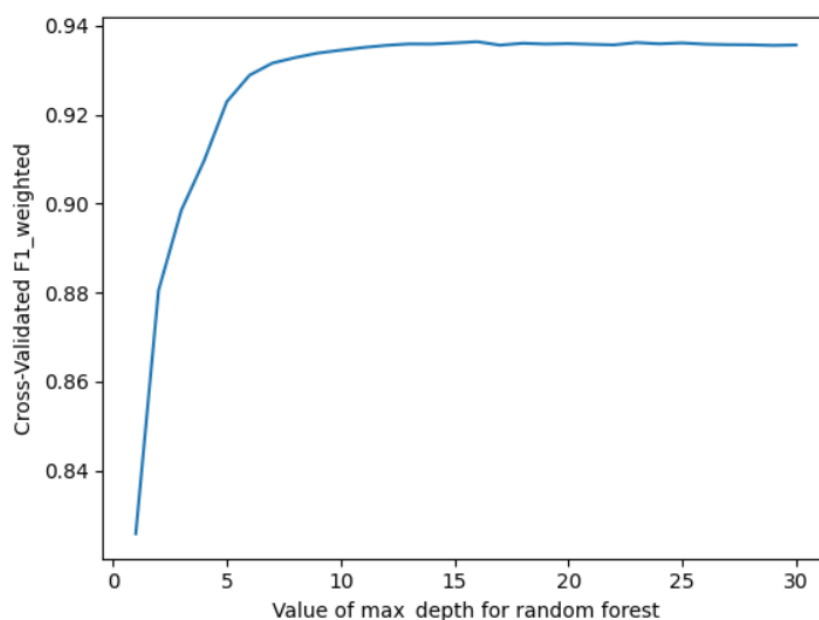
Optimal depth is 10

圖十六 用”entropy”計算 impurity，找出最佳深度

Validation Set Classification Report:

	precision	recall	f1-score	support
Non-Depression	0.95	0.96	0.96	22986
Depression	0.83	0.79	0.80	5154
accuracy			0.93	28140
macro avg	0.89	0.87	0.88	28140
weighted avg	0.93	0.93	0.93	28140

圖十七 用”entropy”計算 impurity，模型訓練的結果



Optimal depth is 9

圖十八 使用 Random forest 找出最佳深度




Validation Set Classification Report:

	precision	recall	f1-score	support
Non-Depression	0.95	0.97	0.96	22986
Depression	0.84	0.79	0.82	5154
accuracy			0.94	28140
macro avg	0.90	0.88	0.89	28140
weighted avg	0.93	0.94	0.93	28140

圖十七 使用 Random forest 模型訓練的結果

繳交於 Kaggle 上的結果

從圖十九中可發現預測效果最好的是 Random forest。而使用”gini”或”entropy”的效果相近，且都略低於 Random forest。

Submission and Description	Private Score ^①	Public Score ^①
 random forest.csv Complete (after deadline) · 15s ago	0.93618	0.93742
 Decision tree entropy ver.csv Complete (after deadline) · 41s ago	0.93043	0.93118
 Decision tree gini ver.csv Complete (after deadline) · 1m ago	0.93086	0.93278

圖十九 在 Kaggle 上繳交模型預測結果