



# 線上課程對教育的影響

資料探勘第18組期末報告

4109029022 蕭聰宇

4109029033 林祥吉

指導教授：蔡孟勳 教授

# 目錄



## 1. 資料集介紹 與研究動機

- 資料集介紹
- 研究動機
- 使用到的SDGs4
- 魚骨圖

## 2. 文獻探討

- 影響線上學習的因素
- 線上課程對某些學生來說效果更好嗎?

## 3. 資料預處理

- 檢查並處理缺失值
- 類別型資料轉換
- 個別欄位處理
- 資料標準化
- 特徵選取

## 4. 研究方法

- 資料分析
- 資料分割
- 建立AI模型 — Decision Tree、Random Forest、貝式

## 5. 結論

- ROC曲線
- AI模型結果
- 小結論
- 總結論

## 6. 參考資料

- 資料集來源
- 教授、助教及學長姐PPT
- 網路資料來源

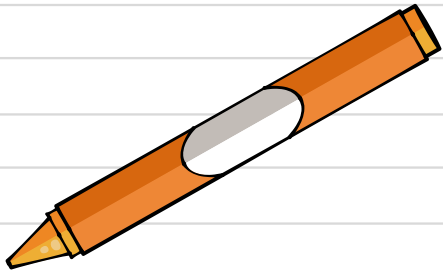





1.

# 資料集介紹與研究動機



- 資料集介紹
  - 研究動機
  - 使用到的SDGs
  - 魚骨圖
- 
- 



# 資料集介紹

原始資料集 :Online Education System – Review (kaggle)

總共 : 1033筆資料 (23個欄位)

連結 : <https://www.kaggle.com/datasets/sujaradha/online-education-system-review>





# 欄位說明

欄位	說明	內容
Gender	性別	Female=0, Male=1
Home Location	住址	Rural=0, Urban=1
Level of Education	教育程度	Post Graduate=0, School=1, Under Graduate=2
Age(Years)	年紀	9~40 歲
Number of Subjects	科目數	1~20 科
Device type used to attend classes	用於上課的設備類型	Desktop=0, Laptop=1, Mobile=2
Economic status	經濟狀況	Middle Class=0, Poor=1, Rich=2
Family size	家庭規模	2~10 人





# 欄位說明

欄位	說明	內容
Internet facility in your locality	所在地區的網路設施	1~5, 數字越大, 網速越快
Are you involved in any sports?	有參加任何運動嗎	No=0, Yes=1
Do elderly people monitor you?	長輩會監視你嗎	No=0, Yes=1
Study time	學習時間	1~10小時
Sleep time	睡眠時間	1~10小時
Time spent on social media	在社交媒體上花費的時間	1~10小時
Interested in Gaming?	對遊戲感興趣嗎	No=0, Yes=1
Have separate room for studying?	有獨立的學習空間嗎	No=0, Yes=1





# 欄位說明

欄位	說明	內容
Engaged in group studies?	有參加小組學習嗎	No=0, Yes=1
Average marks scored before pandemic in traditional classroom	疫情前傳統課堂的平均分數	1~100分，每10分為一間距
Your interaction in online mode	在線上教學下的互動	1~5分
Clearing doubts with faculties in online mode	在線上教學解決問題	1~5分
Interested in?	對...興趣	Both=0, Practical=1, Theory=2
Performance in online	線上表現	2~10分
Your level of satisfaction in Online Education	對線上教育的滿意度	Average=0, Bad=1, Good=2



# 研究目的



使用資料探勘得到的結果，希望提供教育界一些更有效提升線上教學成效的方向，同時達到聯合國SDGs中良質教育，藉由線上教學消除教育的不平等，解決教育城鄉差距的問題，提供國家教育機構或非盈利組職免費且優質的線上教學方案。







## SDGs 4良質教育「確保人人享有優質教育」



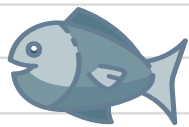
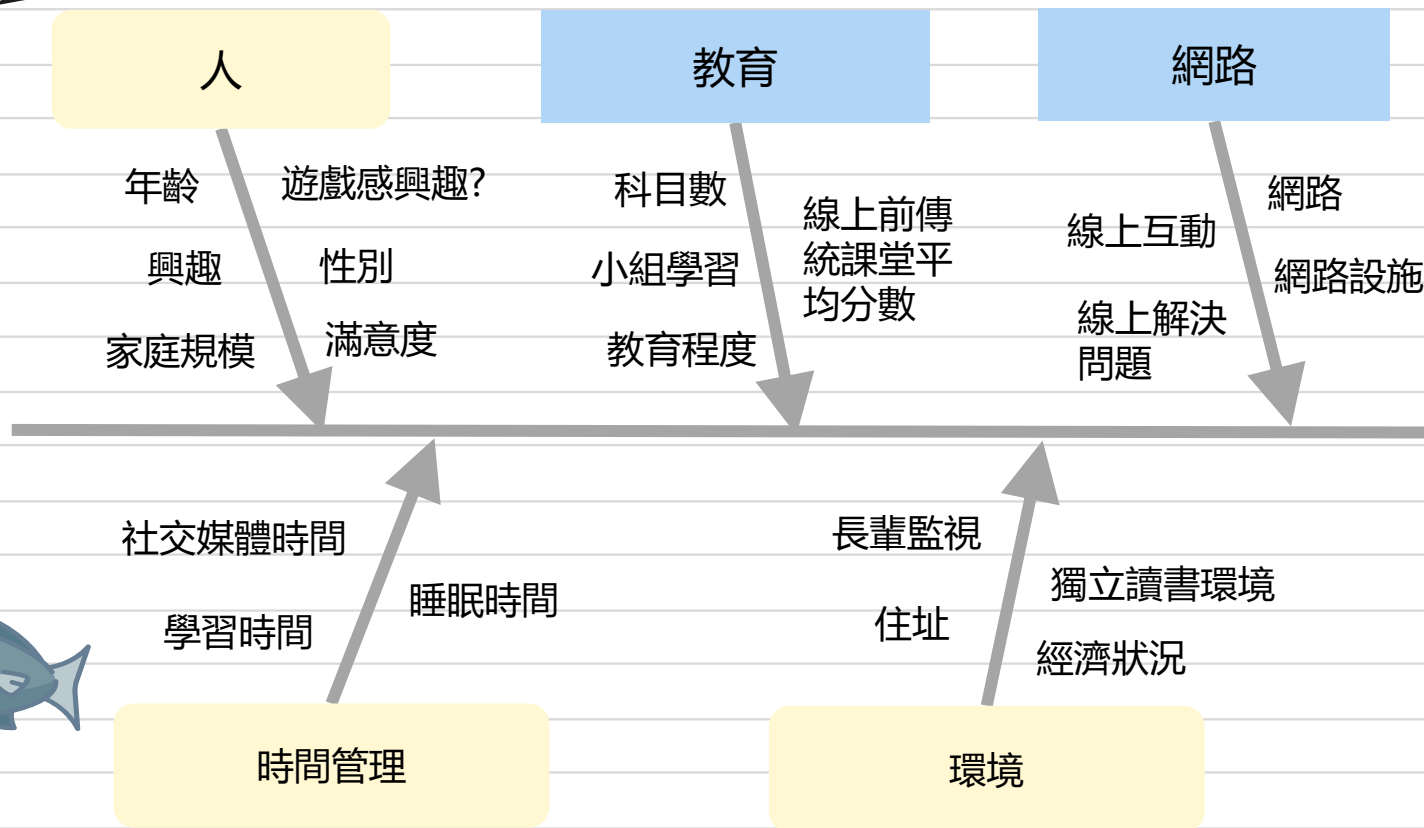
從2019年至今疫情在全球蔓延，導致所有學生的教育受到影響，不只許多學校全面改採遠端教學，不少補習班更直接宣布停課，讓學生在一片混亂中結束一學期。所幸疫情在大家的共同努力下趨緩，學生們終於能回歸正常的學習生活，但與此同時，全球卻仍有超過一半的無法獲得高品質的教育，而線上課程正是疫情期中學習的一個解套，故我們將配合所選用的資料集來檢測線上教學是否有機會促進SDGs 4良質教育。

包含五項目標：提供免費中小學教育、讓孩童接受平等且優質的學前教育、確保人人獲得公平且可負擔的高等教育受教機會、增加具備就業技能的人數、消除教育中所有不平等問題，讓所有孩童都有良好的教育體驗。

4 QUALITY  
EDUCATION




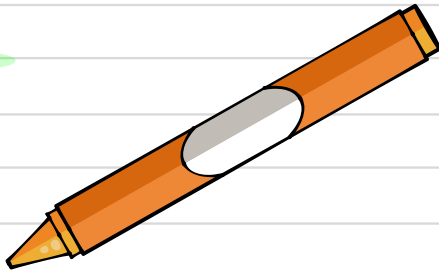

# 魚骨圖





# 2.

## 文獻探討

- 
- 
- 影響線上學習的因素
  - 線上課程對某些學生來說效果更好嗎?
- 



## 影響線上學習的因素

影響線上學習成效的因素有很多，在Al-Ammar及Kirkpatrick研究指出，影響學習成效的因素可分為個人特質、動機、環境因素、等等。

### 個人特質

年齡

教育程度



### 動機

上課態度

上課反映



### 環境因素

工作環境





## 線上課程對某些學生來說效果更好嗎？

線上課程通常成功的學生是那些具有獨立學習取向、受到內在資源高度激勵並且具有強大的時間管理、識字和技術技能的學生。

“


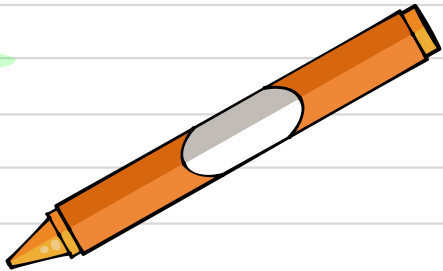

The benefits associated with virtual schooling are expanding educational access, providing high-quality learning opportunities, improving student outcomes and skills, allowing for educational choice, and achieving administrative efficiency. However, the research to support these conjectures is limited at best. The challenges associated with virtual schooling include the conclusion that the only students typically successful in online learning environments are those who have independent orientations towards learning, highly motivated by intrinsic sources, and have strong time management, literacy, and technology skills. These characteristics are typically associated with adult learners. This stems from the fact that research into and practice of distance education has typically been targeted to adult learners.<sup>[17]</sup>





# 3.

## 資料預處理

- 
- 
- 檢查並處理缺失值
  - 類別型資料轉換
  - 個別欄位處理
  - 資料標準化
  - 特徵選取
- 

# 檢查並處理缺失值

```
data.isnull().sum() #show出有缺失的項目
```

```
Gender 0
Home Location 0
Level of Education 0
Age(Years) 0
Number of Subjects 0
Device type used to attend classes 0
Economic status 0
Family size 0
Internet facility in your locality 0
Are you involved in any sports? 0
Do elderly people monitor you? 0
Study time (Hours) 0
Sleep time (Hours) 0
Time spent on social media (Hours) 0
Interested in Gaming? 0
Have separate room for studying? 0
Engaged in group studies? 0
Average marks scored before pandemic in traditional classroom 0
Your interaction in online mode 0
Clearing doubts with faculties in online mode 0
Interested in? 0
Performance in online 0
Your level of satisfaction in Online Education 0
dtype: int64
```

使用data.isnull().sum()  
查看資料是否有缺失值

結果為無缺失值!





# 類別資料轉換

利用LabelEncoder將類別型資料轉換成數值需要轉換的資料如下

'Gender' , 'Home Location',  
'Level of Education' , 'Economic status',  
'Device type used to attend classes',  
'Internet facility in your locality',  
'Are you involved in any sports?',  
'Do elderly people monitor you?',  
'Interested in Gaming?',  
'Have separate room for studying?',  
'Engaged in group studies?',  
'Your interaction in online mode',  
'Clearing doubts with faculties in online mode',  
'Interested in?',  
'Your level of satisfaction in Online Education',  
'Average marks scored before pandemic in traditional classroom'

```
from sklearn.preprocessing import LabelEncoder
le = LabelEncoder()
list_data = [
    'Gender','Home Location',
    'Level of Education','Device type used to attend classes',
    'Economic status','Internet facility in your locality',
    'Are you involved in any sports?',
    'Do elderly people monitor you?',
    'Interested in Gaming?','Have separate room for studying?',
    'Engaged in group studies?','Your interaction in online mode',
    'Clearing doubts with faculties in online mode','Interested in?',
    'Your level of satisfaction in Online Education',
    'Average marks scored before pandemic in traditional
classroom'
]
#先出現的當1，後出現的為0 下面以此類推 用for
for i in range(len(list_data)):
    x = list_data[i]
    data[x] = le.fit_transform(data[x])
```





# 類別資料轉換

Gender	Home Location	Level of Education	Age(Years)	Number of Subjects	Device type used to attend classes	Economic status	Family size	Internet facility in your locality	Are you involved in any sports?	
0	1	1	2	18	11	1	0	4	4	0
1	1	1	2	19	7	1	0	4	0	1
2	1	0	2	18	5	1	0	5	1	0
3	1	1	2	18	5	1	0	4	3	1
4	1	0	2	18	5	1	0	4	2	0

target 無轉換!

Do elderly people monitor you?	Study time (Hours)	Sleep time (Hours)	Time spent on social media (Hours)	Interested in Gaming?	Have separate room for studying?	Engaged in group studies?	Average marks scored before pandemic in traditional classroom	Your interaction in online mode	Clearing doubts with faculties in online mode	Interested in?	Performance in online	Your level of satisfaction in Online Education
1	3	6	1	0	0	0	9	0	0	1	6	0
1	7	5	1	1	1	0	9	0	0	2	3	1
1	6	7	1	0	1	0	7	0	0	0	6	1
1	3	6	2	0	0	1	9	0	1	2	4	1
0	8	7	2	1	1	1	8	2	2	0	6	0





# 個別欄位處理

## Performance in online

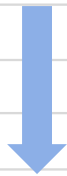
利用replace()將Performance in online  
分割成3個區間

2~4 => 0

5~7 => 1

8~10 => 2

```
data['Performance in online'].unique()  
array([ 6,  3,  4,  2,  9,  7,  5, 10,  8])
```



```
data['Performance in online'] = data['Performance in online'].replace([2, 3, 4], 0)  
data['Performance in online'] = data['Performance in online'].replace([5, 6, 7], 1)  
data['Performance in online'] = data['Performance in online'].replace([8, 9, 10], 2)  
data['Performance in online'].unique()  
array([1, 0, 2])
```





# 資料標準化

將所有欄位做標準化，除了 Performance in online (target)

```
from sklearn.preprocessing import StandardScaler
#要標準化的欄位
data1 = data.drop(['Performance in online'], axis=1)
col_names = data1.columns.values #把欄位名稱指派到col_names
features = data[col_names] #把欄位裡的資料指派到features
scaler = StandardScaler().fit(features.values) #特徵標準化，也就是高斯分佈。使得數據的平均值為0，方差為1
features = scaler.transform(features.values) #transform: 返回完整數據的某一變換版本供我們重組
data[col_names] = features #把標準化完的 features 指派回 dataset
data.head(3)
```



# 資料標準化

	Gender	Home Location	Level of Education	Age(Years)	Number of Subjects	Device type used to attend classes	Economic status	Family size	Internet facility in your locality	Are you involved in any sports?
0	0.826081	0.722049	0.486024	-0.562497	1.411598	-0.581565	-0.270233	-0.334392	1.378125	-0.747040
1	0.826081	0.722049	0.486024	-0.249763	-0.012407	-0.581565	-0.270233	-0.334392	-2.522158	1.338616
2	0.826081	-1.384947	0.486024	-0.562497	-0.724409	-0.581565	-0.270233	0.474571	-1.547087	-0.747040

target 保留

Do elderly people monitor you?	Study time (Hours)	Sleep time (Hours)	Time spent on social media (Hours)	Interested in Gaming?	Have separate room for studying?	Engaged in group studies?	Average marks scored before pandemic in traditional classroom	Your interaction in online mode	Clearing doubts with faculties in online mode	Interested in?	Performance in online	Your level of satisfaction in Online Education
0.946263	-0.621257	-0.716130	-0.8807	-1.088087	-1.196072	-0.822769	1.244915	-1.747111	-1.576433	0.128714	6	-0.867222
0.946263	1.253860	-1.471761	-0.8807	0.919045	0.836070	-0.822769	1.244915	-1.747111	-1.576433	1.458326	3	0.338485
0.946263	0.785080	0.039501	-0.8807	-1.088087	0.836070	-0.822769	-0.165943	-1.747111	-1.576433	-1.200899	6	0.338485



# 特徵選取

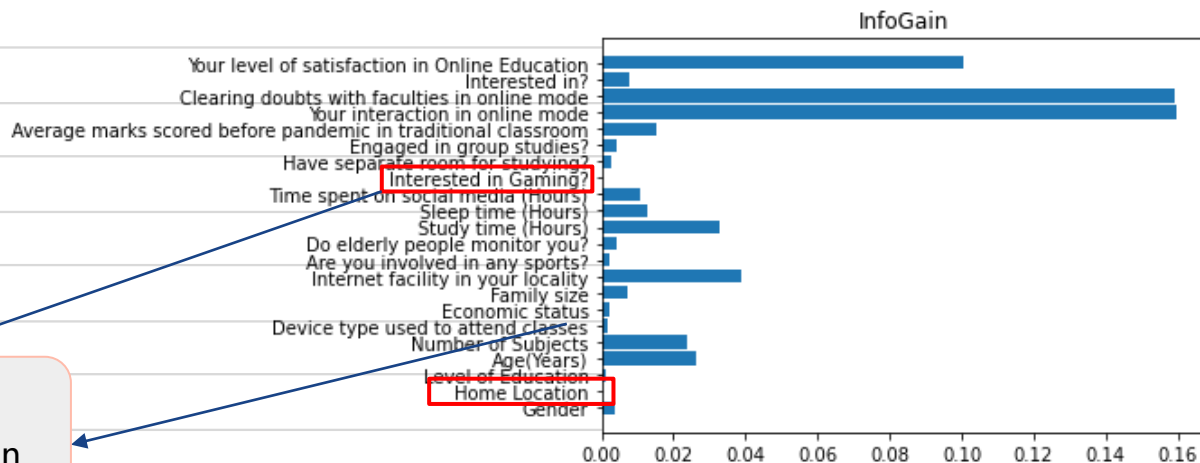
使用PCA(0.97)配合Info Gain與Gain Ratio 做特徵選取，找出需刪除的欄位

```
from sklearn.decomposition import PCA
pca = PCA(n_components=0.97, whiten=True)
features_pca = pca.fit_transform(features)
print("Original number of features(原本有多少欄位):", features.shape[1])
print("Reduce number of features(減少後變幾個欄位):", features_pca.shape[1])
```

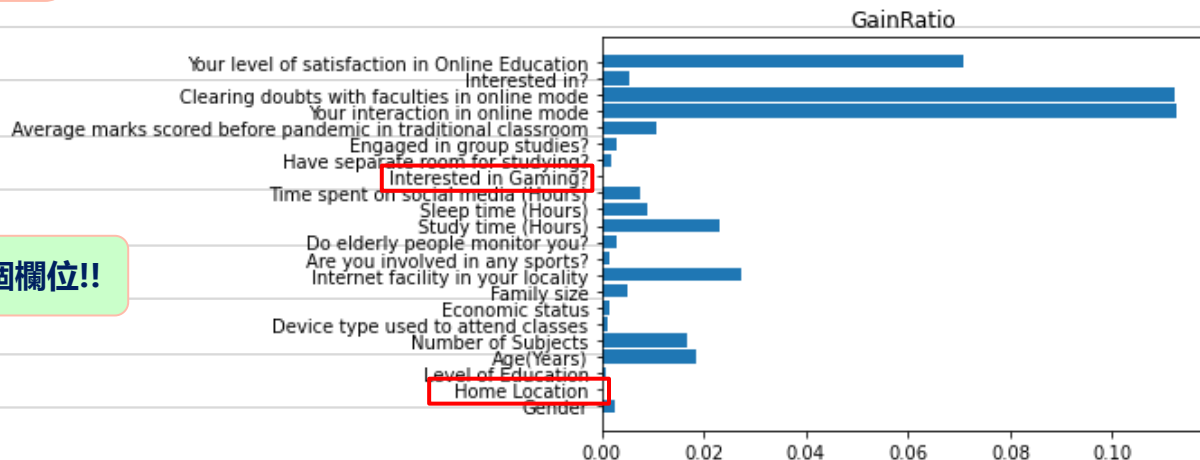


```
Original number of features(原本有多少欄位): 23
Reduce number of features(減少後變幾個欄位): 21
```





```
data =  
data.drop(['Home  
Location','Interested in  
Gaming?'], axis=1)
```




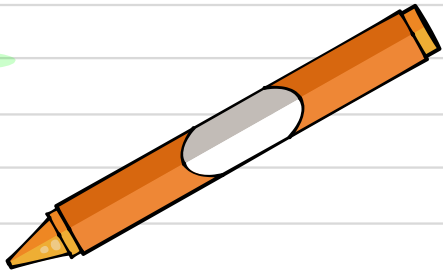

刪除最不重要的兩個欄位!!



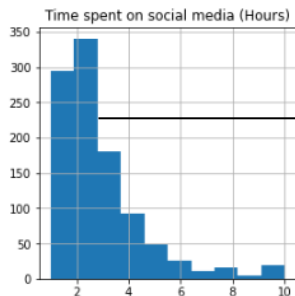


# 4.

## 研究方法

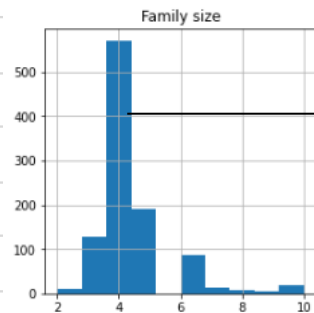
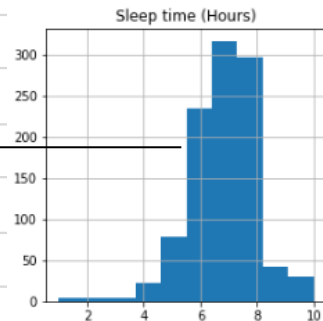
- 
- 
- 資料分析
  - 資料分割
  - 建立AI模型 — Decision Tree、Random Forest、貝式
- 

# 資料分析 直方圖



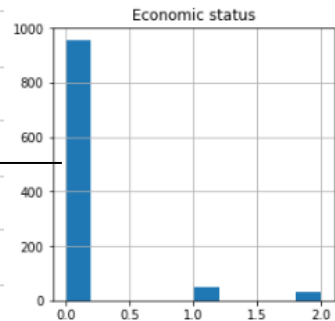
花1-3小時在社交媒體上的人佔大多數

大部分的睡眠時間介於6-8小時



4個人的家庭規模為最多

大部分人的經濟狀況都為中等





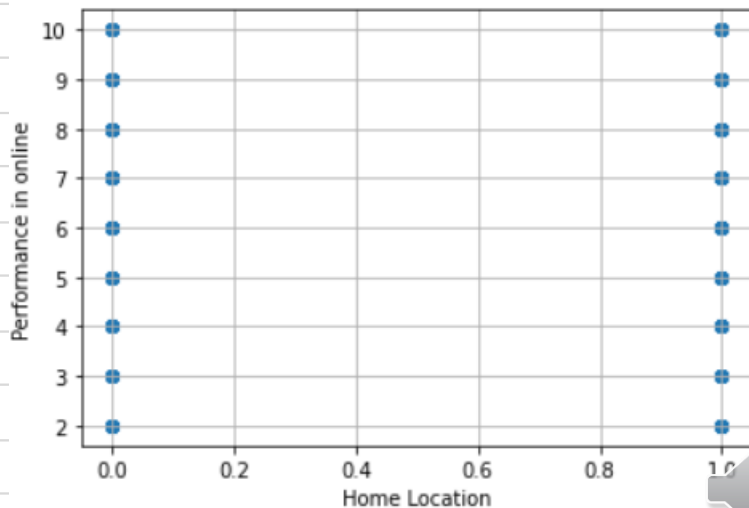
# 資料分析

## 散佈圖

```
plt.scatter(data["Home Location"], data["Performance in online"])
plt.xlabel("Home Location") #X軸標籤
plt.ylabel("Performance in online") #Y軸標籤
plt.grid(True)
```

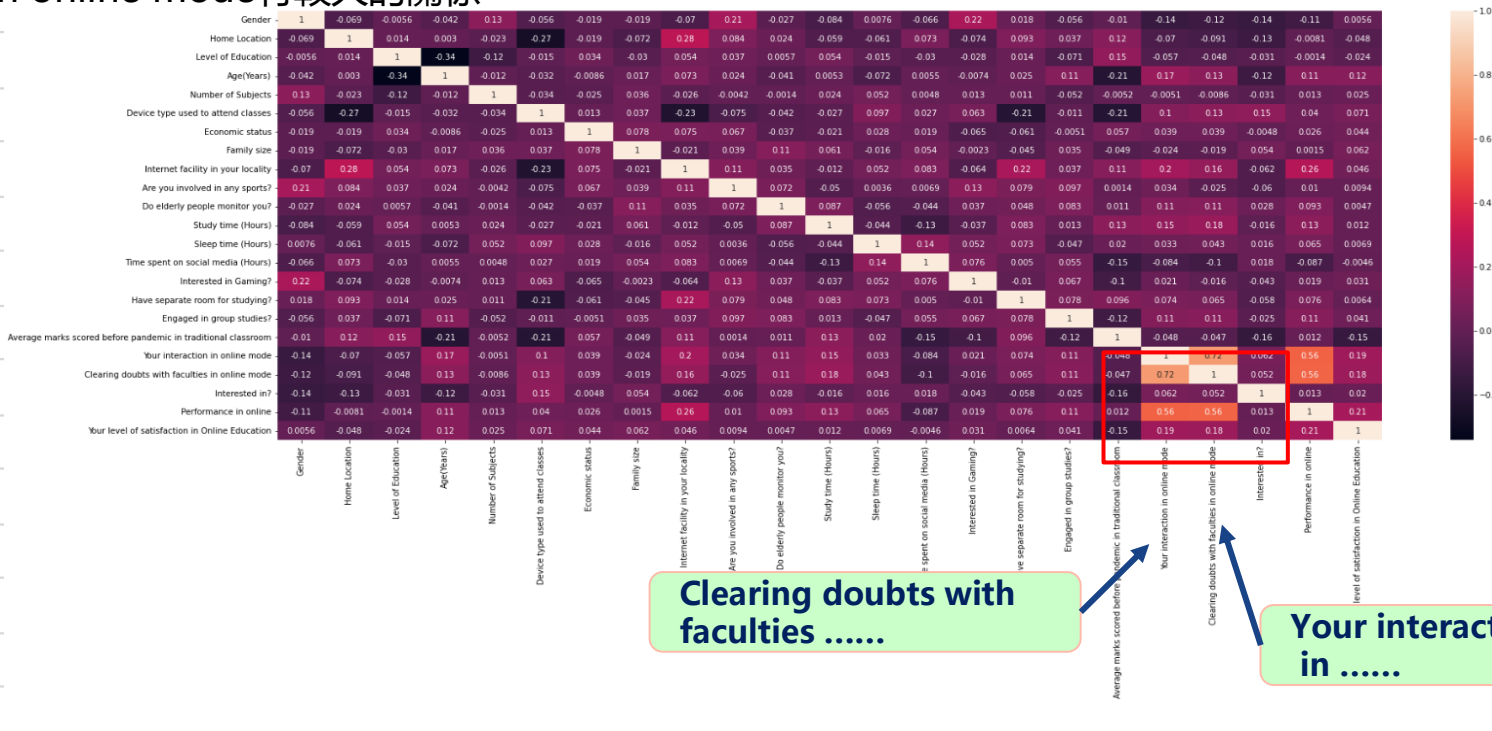
線上教學是否可以消除城鄉差距?

藉由把 Home Location 跟 Performance in online 做成散佈圖，可以發現其實兩者無太大的關係，或是說關係性趨近於0，可以呼應前面 info\_gain的結果，表現出實施線上教育後，城鎮與鄉村對於學習表現無相關，推得線上教學可能可以消除城鄉差距！



# 資料分析 熱圖

Performance in online 跟 Clearing doubts with faculties in online mode 及 Your interaction in online mode 有較大的關係

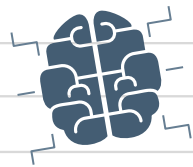


Clearing doubts with faculties .....

Your interaction in .....



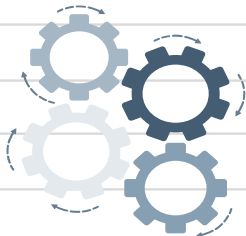
# 資料分割



以Performance in online 為 target將資料分成訓練集(0.8)以及測試集(0.2)

```
from sklearn.model_selection import train_test_split
X = data.drop(['Performance in online'],axis=1) #除了目標以外剩餘的欄位
y = data['Performance in online'] #目標

#train訓練集, test測試集大小
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=100) #大約80%資料給機器去學習
```





# 建立 AI 模型 Decision Tree

```
from sklearn.tree import DecisionTreeClassifier
dtree = DecisionTreeClassifier()
dtree.fit(X_train,y_train)
```

```
predictions = dtree.predict(X_test)
from sklearn.metrics import classification_report, confusion_matrix
dtree_pred=dtree.predict(X_test)
dtree.score1 = dtree.score(X_test,y_test)
print('準確度:', dtree.score1)
print('\n')
print(confusion_matrix(y_test, dtree_pred))
print('\n')
print(classification_report(y_test, predictions))
```

準確度: 0.5265700483091788

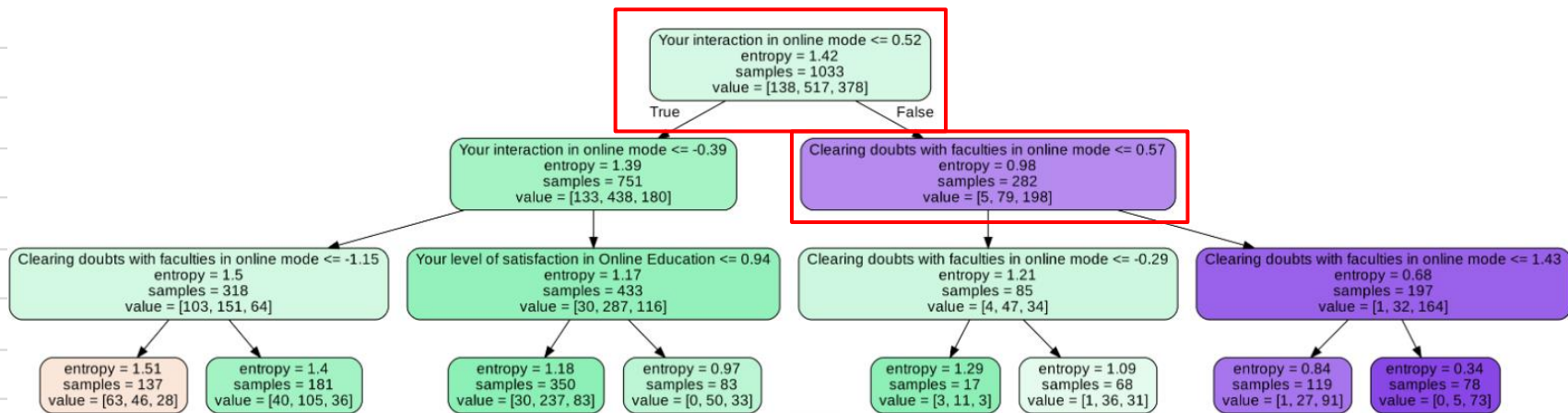
```
[[ 7  8  3]
 [13 65 31]
 [ 8 35 37]]
```

	precision	recall	f1-score	support
0	0.25	0.39	0.30	18
1	0.60	0.60	0.60	109
2	0.52	0.46	0.49	80
accuracy			0.53	207
macro avg	0.46	0.48	0.46	207
weighted avg	0.54	0.53	0.53	207



# Decision Tree 視覺化

Your interaction in online mode 作為決策樹的第一個分支  
可以看出 Your interaction in online mode 是分類 Performance in online 最重要的特徵  
而第二重要的為 Clearing doubts with faculties in online mode 與熱圖結果相符!





# 建立 AI 模型 Random Forest

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import roc_curve, roc_auc_score, auc, accuracy_score, confusion_matrix, classification_report
rfc=RandomForestClassifier(n_estimators=5) #n_estimators代表要用多少CartTree(為使用gini算法的決策樹)
```

```
#從訓練組資料中建立隨機森林模型
rfc.fit(X_train, y_train)
rfc_pred=rfc.predict(X_test)
rfc.score1 = rfc.score(X_test, y_test)
print('準確度:', rfc.score1)
print('\n')
print(confusion_matrix(y_test, rfc_pred))
print(classification_report(y_test, rfc_pred))
```

準確度: 0.6135265700483091

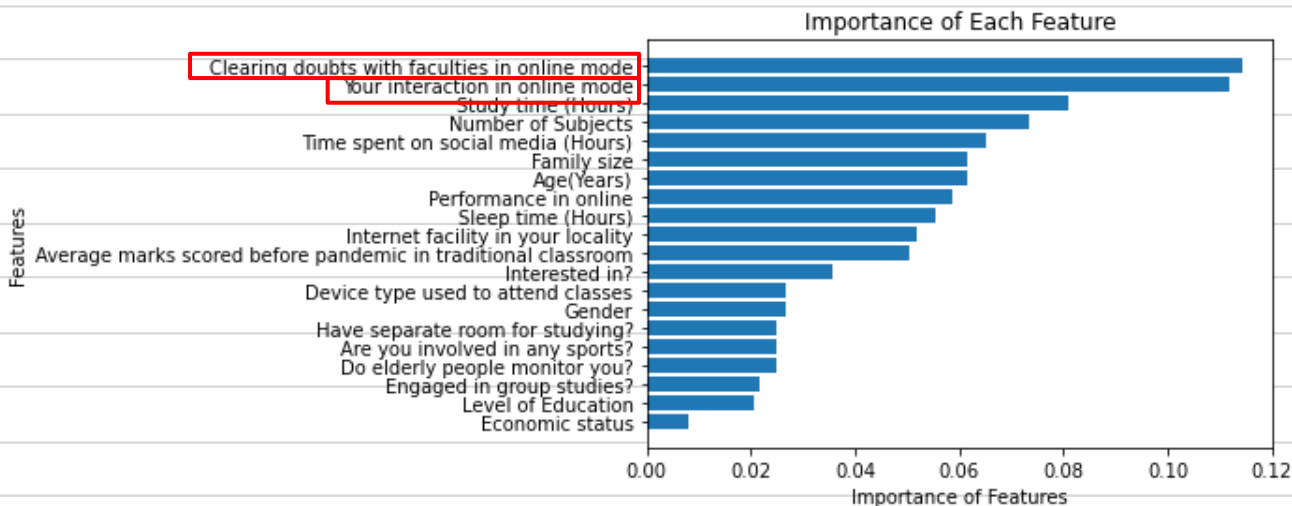
```
[[ 8  9 1]
 [ 9 75 25]
 [ 3 33 44]]
```

	precision	recall	f1-score	support
0	0.40	0.44	0.42	18
1	0.64	0.69	0.66	109
2	0.63	0.55	0.59	80
accuracy			0.61	207
macro avg	0.56	0.56	0.56	207
weighted avg	0.62	0.61	0.61	207



# 用 Random Forest 視覺化重要性

可以看出 Your interaction in online mode , Clearing doubts with faculties in online mode 是最重要的特徵與熱圖結果相符!



# 建立 AI 模型 貝式

```
from sklearn.naive_bayes import GaussianNB
gnb=GaussianNB()
gnb.fit(X_train,y_train)
pred = gnb.predict(X_test)
y_gnb_score = gnb.predict_proba(X_test)
gnb.score1 = gnb.score(X_test,y_test)
print('準確度:', gnb.score1)
print('\n')
print(confusion_matrix(y_test,pred))
print('\n')
print(classification_report(y_test,pred))
```

準確度: 0.6280193236714976

```
[[10  7  1]
 [ 9 79 21]
 [ 9 30 41]]
```

	precision	recall	f1-score	support
0	0.36	0.56	0.43	18
1	0.68	0.72	0.70	109
2	0.65	0.51	0.57	80
accuracy			0.63	207
macro avg	0.56	0.60	0.57	207
weighted avg	0.64	0.63	0.63	207


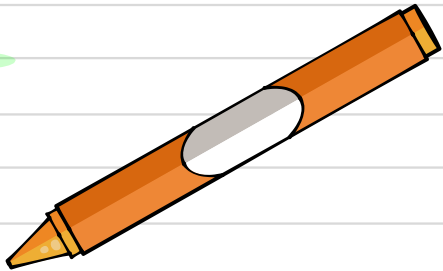







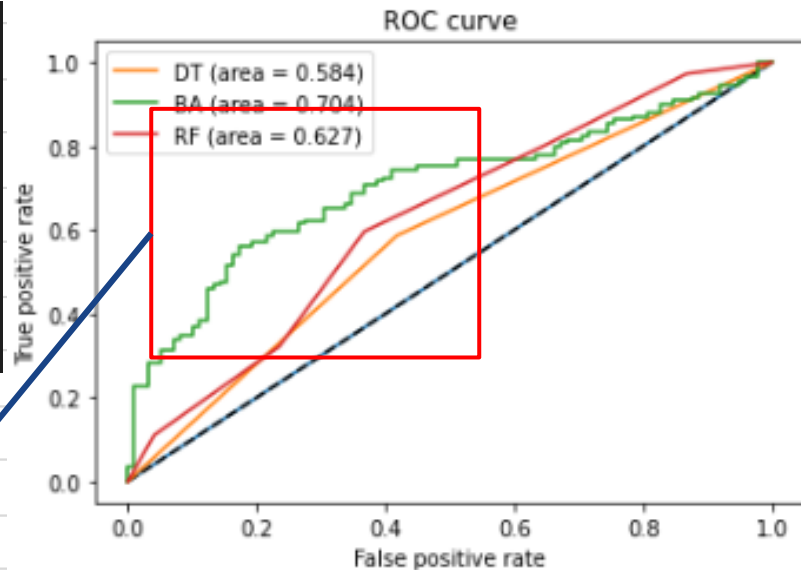
# 5.

## 結論

- 
- 
- ROC曲線
  - AI模型結果
  - 小結論
  - 總結論
- 

# ROC曲線

```
#繪製roc curve
plt.figure(1)
plt.plot([0, 1], [0, 1], 'k--')
plt.plot(fpr_dtc, tpr_dtc, label='DT (area = {:.3f})'.format(auc_dtc))
plt.plot(fpr_gnb, tpr_gnb, label='BA (area = {:.3f})'.format(auc_gnb))
plt.plot(fpr_rf, tpr_rf, label='RF (area = {:.3f})'.format(auc_rf))
# plt.plot(fpr_knn, tpr_knn, label='KNN (area = {:.3f})'.format(auc_knn))
plt.xlabel('False positive rate')
plt.ylabel('True positive rate')
plt.title('ROC curve')
plt.legend(loc='best')
plt.show()
```



由曲線下面積大小可看出貝氏的準確度最高  
再來是隨機森林，最後則決策樹





# AI 模型結果

## 決策樹

準確度約為53%  
在混淆矩陣可以發現  
表現5~7分(中間  
分數)的人預測正確  
率是最好的



## 隨機森林

比決策樹準確度高,  
準確度為62%  
在混淆矩陣可以發現  
表現5~7分(中間分數)  
的人預測正確率是最  
好的

## 貝氏

準確度是最高的為63%  
在混淆矩陣可以發現表  
現5~7分(中間分數)的  
人預測正確率是最好的



## ROC

從ROC曲線圖看出  
貝氏準確度最高,  
因為它曲線下面積  
最大



# 小結論



使用決策樹，隨機森林，貝氏來預測可達5成以上的準確度



由熱圖與決策樹的重要性可看出

- Clearing doubts with faculties in online mode
- Your interaction in online mode

對Performance in online也就是對教育的影響程度最高



由散佈圖跟info\_gain可看出，線上教學中，城鎮跟鄉村對學習表現的重要性較低，足以忽略，可推估線上教學有機會消彌城鄉差距



# 總結論


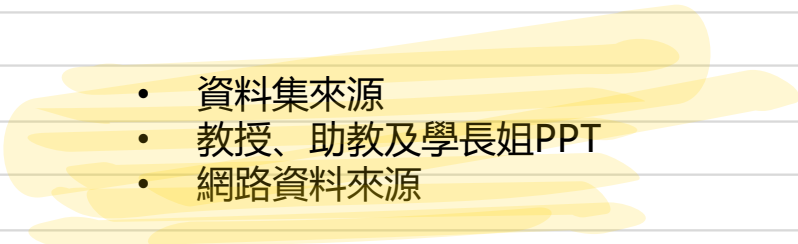
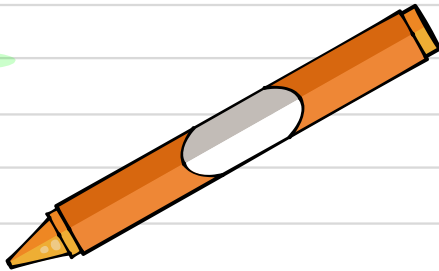
通過前述的分析我們會發現，在線上教學解決問題與在線上教學下的互動對學習表現的影響最高，所以教育者或是業者可以對此兩項結果，發展更完善的教材、互動系統或是解題系統，以利線上教學的學生有更優質的學習環境，達到聯合國SDGS 4 中的「確保人人獲得公平且可負擔的高等教育受教機會」，同時藉由數據的分析，也可以得出線上教學可能有利於減少教育中常面臨到的城鄉差距問題，達成「消除教育中不平等問題」。





6.

參考資料

- 
- 
- 資料集來源
  - 教授、助教及學長姐PPT
  - 網路資料來源
- 





# 參考資料

- 資料集來源：

<https://www.kaggle.com/datasets/sujaradha/online-education-system-review>

- 教授、助教及學長姐PPT

- 網路資料來源：

- 線上學習成效影響因素模式之探討

[https://nccur.lib.nccu.edu.tw/bitstream/140.119/90355/1/79\(%20p1-21%20\).pdf](https://nccur.lib.nccu.edu.tw/bitstream/140.119/90355/1/79(%20p1-21%20).pdf)

- Online Distance Learning: A Literature Review

<https://cirl.etoncollege.com/online-distance-learning-a-literature-review/>





**Thank You**

謝謝大家