教育大數據專題

主題:利用不同變因預測學生成績

4110053140 應數 4 李政峰

(一)研究方法

1. 學生的學習成績資料集:這是由 kaggle 上所找到的資料庫

2. 詳細變數討論，與資料分布

gender : sex of students -> (Male/female)

race/ethnicity : ethnicity of students -> (Group A, B,C, D,E)

parental level of education : parents' final education ->(bachelor's degree,somecollege,master's degree,associate's degree,- high school)

lunch : having lunch before test (standard or free/reduced)

test preparation course : complete or not complete before test

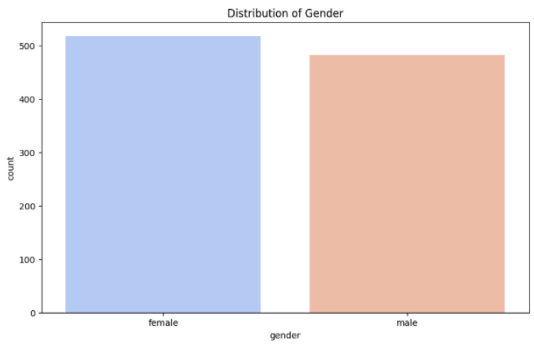
math score

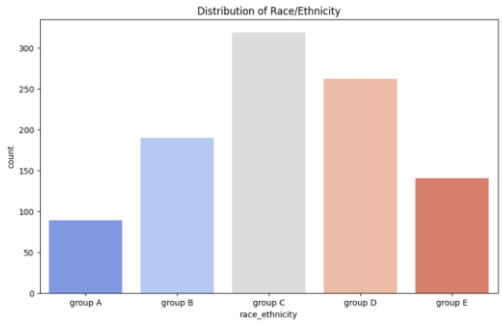
reading score

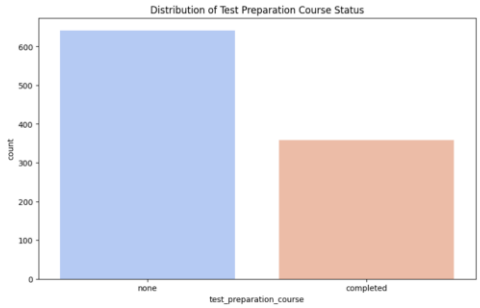
writing score

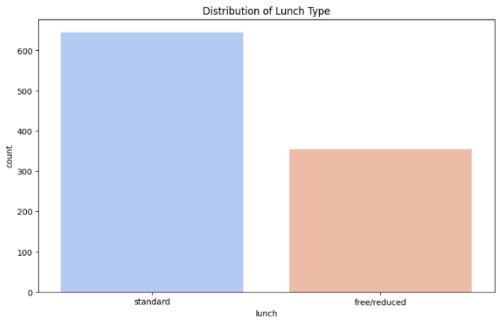
有:性別，種族，雙親教育程度，是否在考前吃午餐，有無準備考試，數學分數，閱讀分數，寫作分數

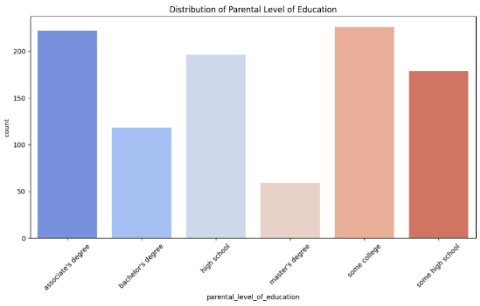
應變數:數學成績　　輸入變數:其他











資料分配並不均勻

是否有 missing data-無

利用不同的機器學習方法用以預測學生學習成績(使用 rf,svm,dt,knn)

(1).資料預處理:

把不同地區/性別/有無吃午餐/數學有無及格等利用 one-hot- encoding 重新編碼，在把有高低排序的學歷由高到低排成 5,4,3,2,1,然後利用 python 內建函式將資料標準化

(2).開始機器學習

在資料預處理後，我們就能開始機器學習。利用 python 內建函式我們可以得到下圖，然而，就算我們做出了還算不錯的準確率，我們仍要注意過擬合的問題，因此我們還要繪製學習曲線

最佳隨機森林準確率: 0.885
最佳隨機森林分類報告:

	precision	recall	f1-score	support
0	0.89	0.74	0.81	65
1	0.88	0.96	0.92	135
accuracy			0.89	200
macro avg	0.89	0.85	0.86	200
weighted avg	0.89	0.89	0.88	200

多項式核SVM準確率: 0.9
多項式核SVM分類報告:

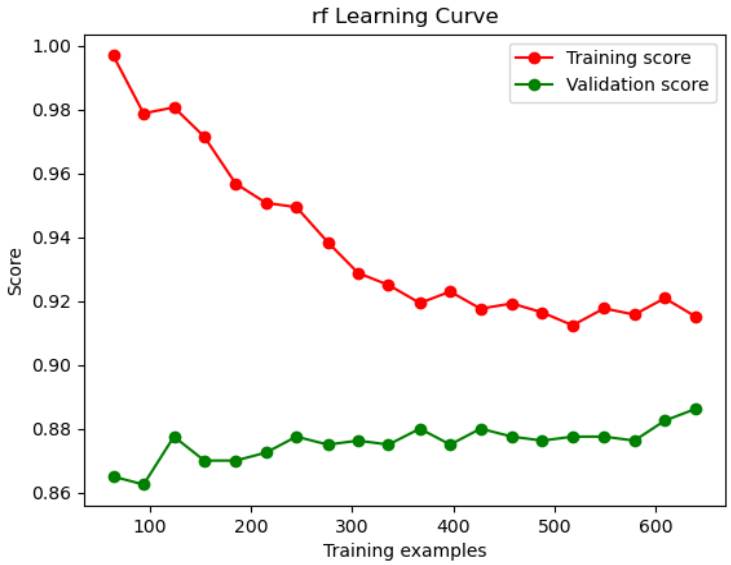
	precision	recall	f1-score	support
0	0.88	0.80	0.84	65
1	0.91	0.95	0.93	135
accuracy			0.90	200
macro avg	0.89	0.87	0.88	200
weighted avg	0.90	0.90	0.90	200

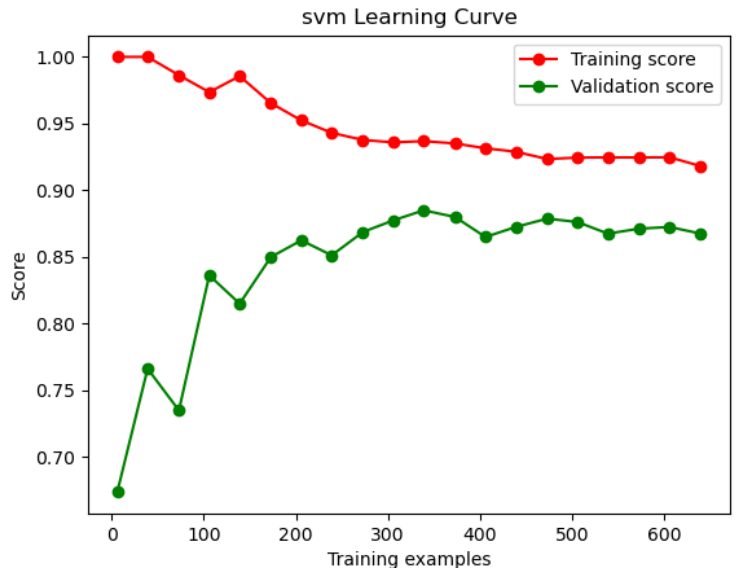
KNN率: 0.86
KNN分告:

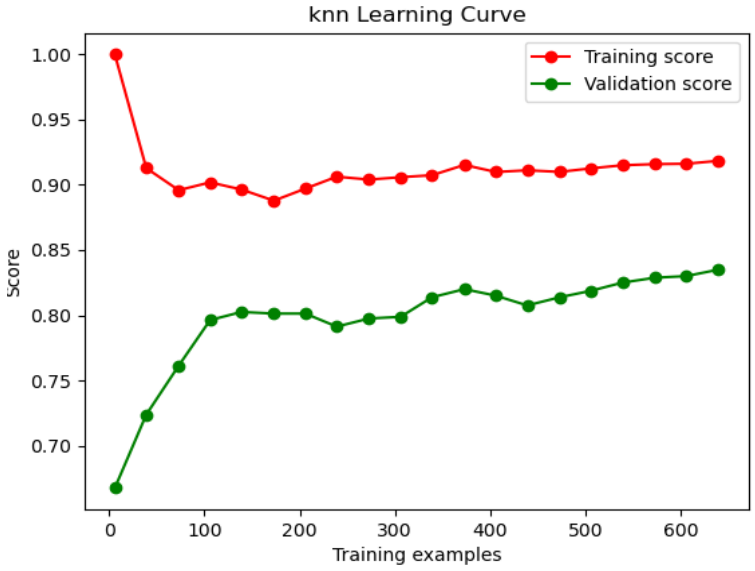
	precision	recall	f1-score	support
0	0.86	0.68	0.76	65
1	0.86	0.95	0.90	135
accuracy			0.86	200
macro avg	0.86	0.81	0.83	200
weighted avg	0.86	0.86	0.86	200

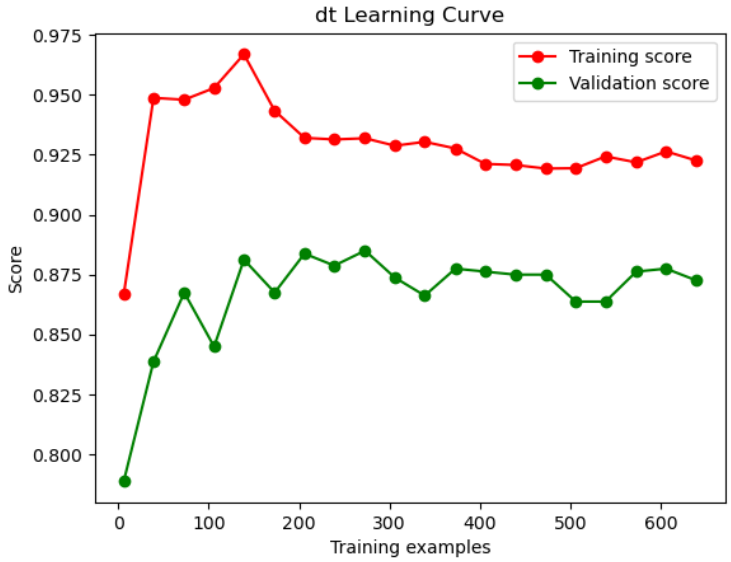
最佳決策樹準確率: 0.855
最佳決策分類報告:

	precision	recall	f1-score	support
0	0.77	0.78	0.78	65
1	0.90	0.89	0.89	135
accuracy			0.85	200
macro avg	0.83	0.84	0.84	200
weighted avg	0.86	0.85	0.86	200


rf Learning Curve
Training score
Validation score
Score
Training examples
1.00
0.98
0.96
0.94
0.92
0.90
0.88
0.86
100
200
300
400
500
600


svm Learning Curve
Training score
Validation score
Score
Training examples
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0
100
200
300
400
500
600


knn Learning Curve
Training score
Validation score
Score
Training examples
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0
100
200
300
400
500
600



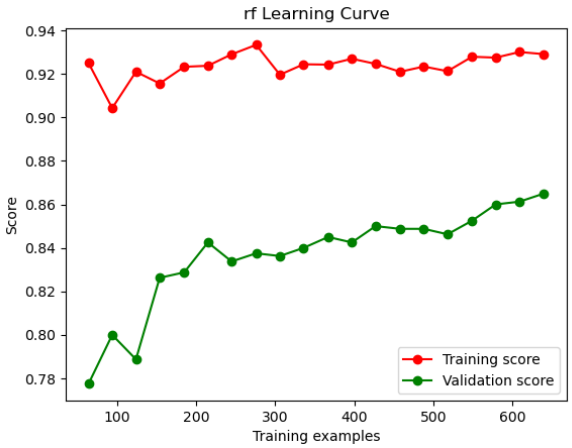
其實我們可以預想到，受限於資料量(1000 筆,11 維)，我們機器學習很容易發生過擬合，不過我們可以利用調整內部參數，並用 grid_search_cv 保證在擬合不錯的前提下，模型仍有一定的精確度。

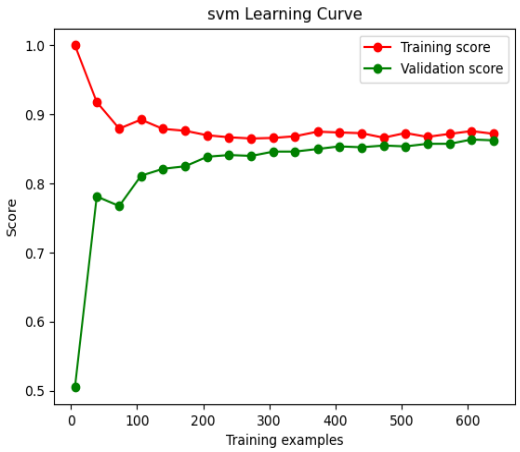
然而，在實際應用上，資料可能更多更複雜，如果還是用以上方法的話，很有可能導致運算資源不夠使用，導致不可預期的錯誤，因此我們需要使用 pca 降維讓我們的資料複雜度減少，從而減少計算資源。

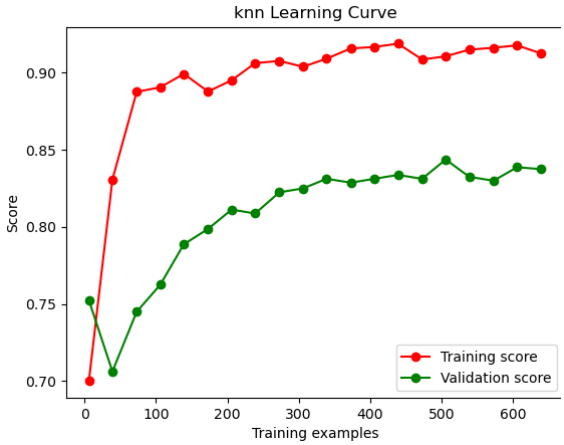
若我們加入 pca 降到三維，在調整 grid_search_cv 參數，我們可以獲得

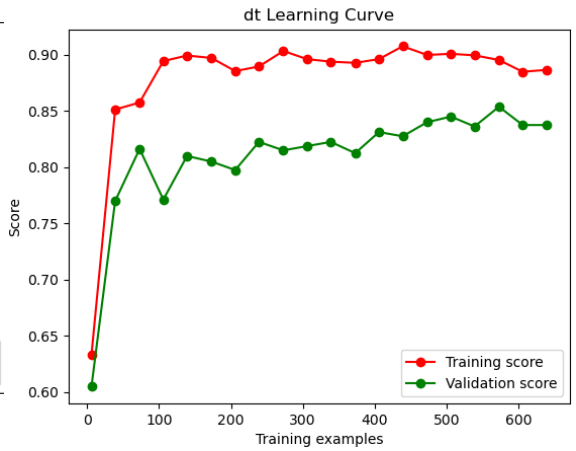
```
最佳隨機森林準確率: 0.85
最佳隨機森林分類報告:
              precision    recall  f1-score   support

           0       0.82      0.69      0.75        65
           1       0.86      0.93      0.89       135

    accuracy                           0.85       200
   macro avg       0.84      0.81      0.82       200
weighted avg       0.85      0.85      0.85       200

多項式核SVM準確率: 0.85
多項式核SVM分類報告:
              precision    recall  f1-score   support

           0       0.81      0.71      0.75        65
           1       0.87      0.92      0.89       135

    accuracy                           0.85       200
   macro avg       0.84      0.81      0.82       200
weighted avg       0.85      0.85      0.85       200
```

```
KNN準確率: 0.83
KNN分類報告:
              precision    recall  f1-score   support

           0       0.82      0.62      0.70        65
           1       0.83      0.93      0.88       135

    accuracy                           0.83       200
   macro avg       0.83      0.77      0.79       200
weighted avg       0.83      0.83      0.82       200

Cross-validation scores: [0.86875 0.84375 0.85625 0.8
0.81875]
Average cross-validation score: 0.8374999999999998
最佳決策樹準確率: 0.83
最佳決策樹分類報告:
              precision    recall  f1-score   support

           0       0.79      0.65      0.71        65
           1       0.84      0.92      0.88       135

    accuracy                           0.83       200
   macro avg       0.82      0.78      0.80       200
weighted avg       0.83      0.83      0.82       200
```





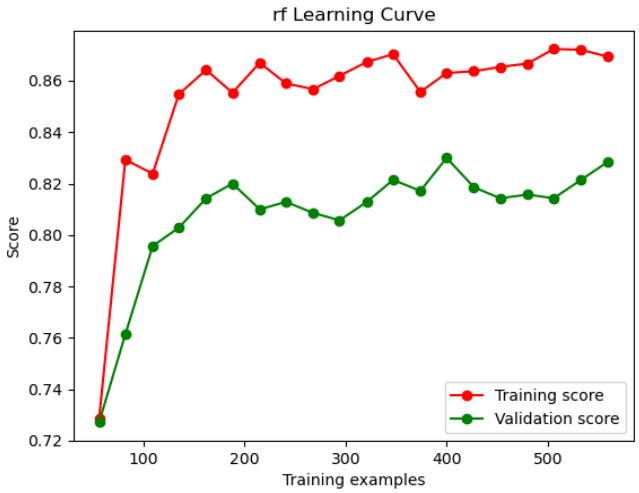


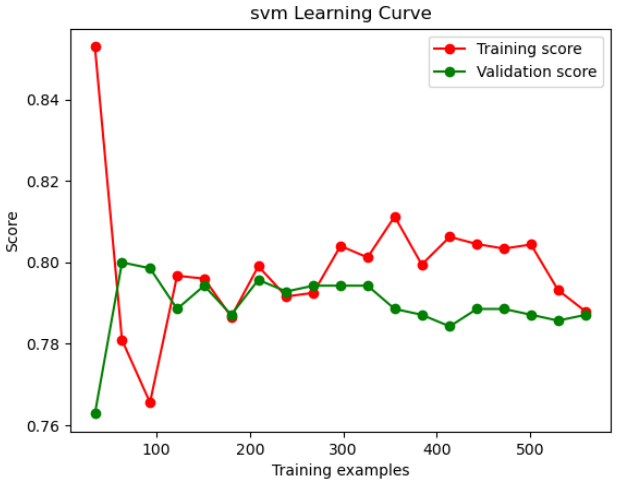


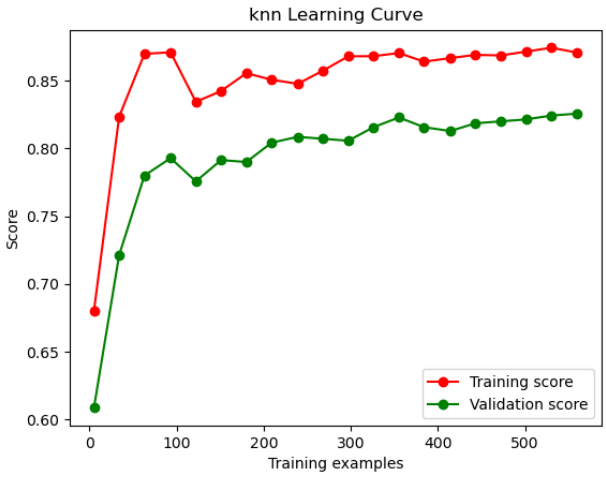
由上圖我們可以看出 svm,dt 擬合得更好了，在降到三維損失了一點精度的前提下。然而資料並不是總是有好的線性關係，這時我們就要用 kener pca(降至 2 維)做非線性主成分分析，再利用 grid_search cv 利用 accuracy score 找出最佳參數組合

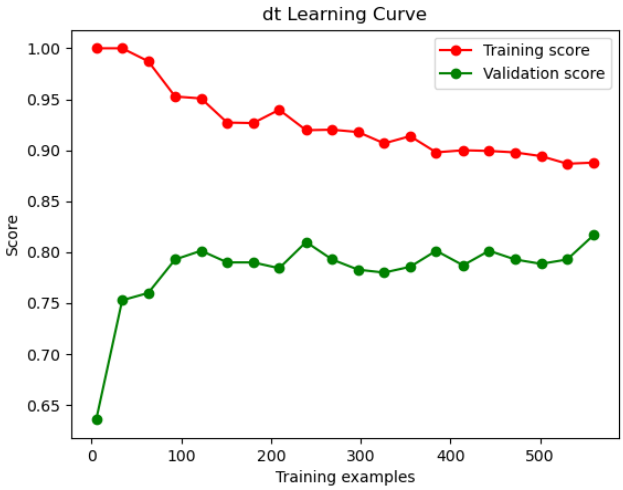
```
最佳隨機森林準確率: 0.8133333333333334
最佳隨機森林分類報告:
              precision    recall  f1-score   support

           0       0.77      0.61      0.68        97
           1       0.83      0.91      0.87       203

    accuracy                           0.81       300
   macro avg       0.80      0.76      0.77       300
weighted avg       0.81      0.81      0.81       300

多項式核SVM準確率: 0.79
多項式核SVM分類報告:
              precision    recall  f1-score   support

           0       0.72      0.57      0.64        97
           1       0.81      0.90      0.85       203

    accuracy                           0.79       300
   macro avg       0.77      0.73      0.74       300
weighted avg       0.78      0.79      0.78       300
```

```
KNN準確率: 0.85
KNN分類報告:
              precision    recall  f1-score   support

           0       0.79      0.73      0.76        97
           1       0.88      0.91      0.89       203

    accuracy                           0.85       300
   macro avg       0.83      0.82      0.83       300
weighted avg       0.85      0.85      0.85       300

Cross-validation scores: [0.84285714 0.80714286 0.82857143
0.81428571 0.83571429]
Average cross-validation score: 0.8257142857142856
最佳決策樹準確率: 0.8066666666666666
最佳決策樹分類報告:
              precision    recall  f1-score   support

           0       0.74      0.62      0.67        97
           1       0.83      0.90      0.86       203

    accuracy                           0.81       300
   macro avg       0.79      0.76      0.77       300
weighted avg       0.80      0.81      0.80       300
```







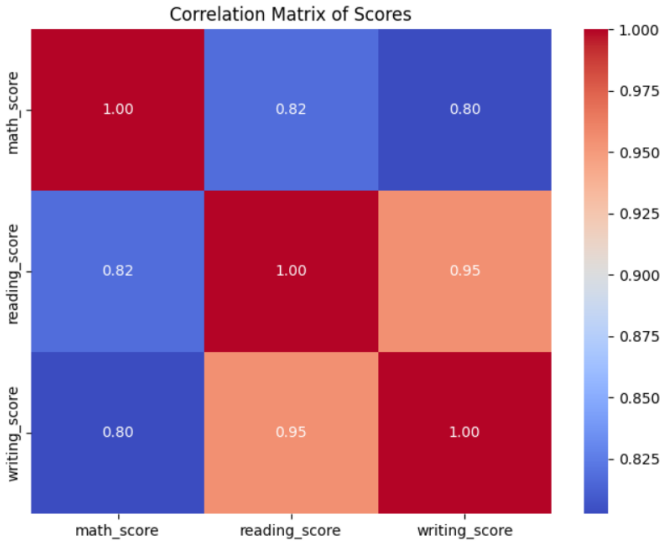


由上，我們得到了一個不錯的 knn 模型，以及勉強能看的 rf，至於如果調參數也沒有效的話，我們也可以利用正則化，早停等方法來避免過擬合

(二)結論:

利用上述方法，我們得到了一些不錯的機器學習模型，雖然只有大概 8 成多的精確度，但在降到低維並且避開過擬合後這樣的表現已經非常不錯，何況他的頻率分布並不均勻，如果有辦法剔除雜訊的

話精度可以再進一步上升，像是移除偏科的極端值，或者利用深度學習調整顯著影響的權重，這些都是可行的方法



並且由相關矩陣可以得知也是有一部份的人偏科