Student Grade Prediction

Dataset

- 資料集名稱：Student Grades Prediction
- 資料數量：395筆
- 資料特徵：32個
- 特徵：
  Sex、Address、famsize、guardian、studytime、G1、G2...

- 透過學生的特徵→預測學生的G3考試成績

Project Introduction

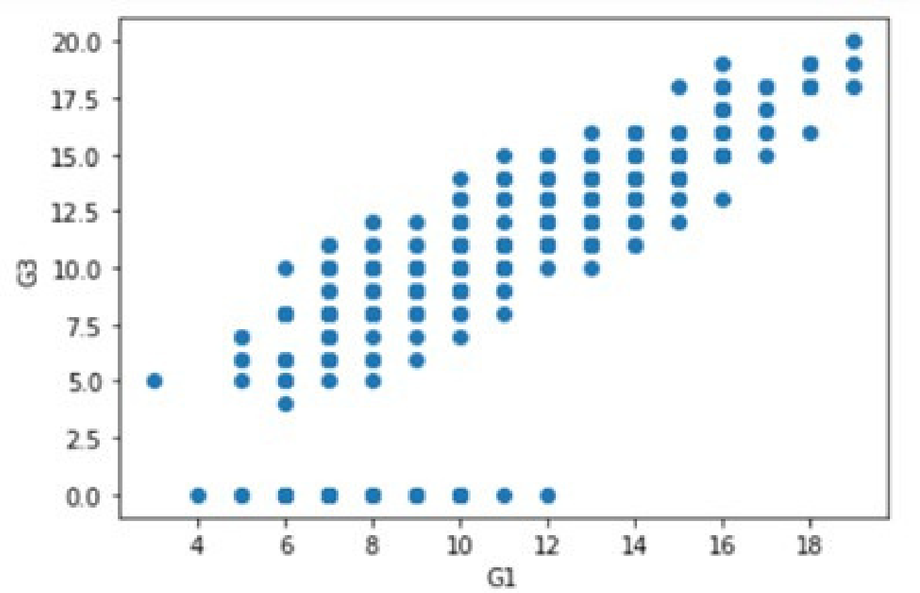
- 目的：希望分析出影響學生成績的最主要因素，而有助於教師針對重點因素幫助學生提升成績；也能對於可能有較差成績的學生產生預警，使教師能盡早發現問題並關心學生學習狀況。

- 實驗方法：
  將其中30個特徵分成四類因素，分析每類與成績的相關性：

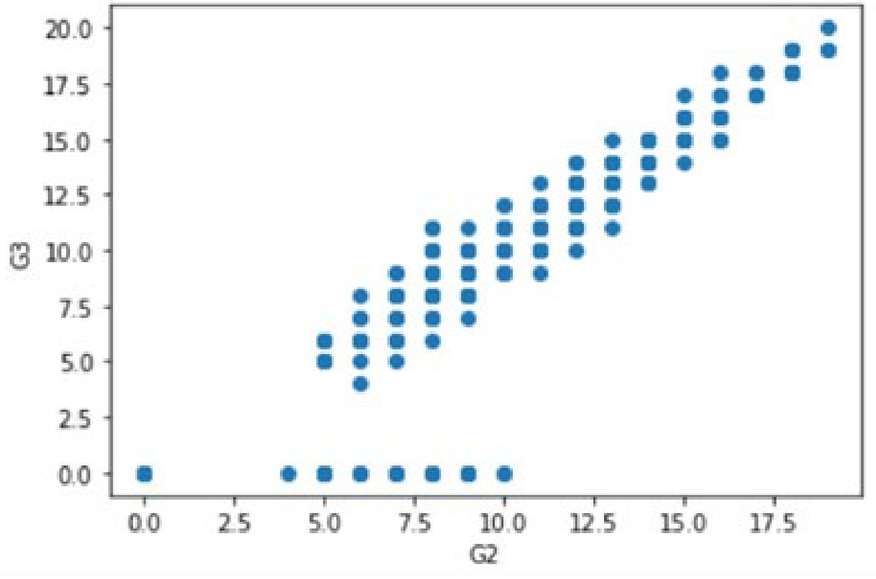
  1. 個人資料(6個特徵)
  2. 家庭因素(9個特徵)
  3. 自我管理(11個特徵)
  4. 額外教育(4個特徵)

Project Introduction

- 因為G1、G2與預測結果高度相關，因此除了各自使用的各類特徵外，我們也共用此兩特徵。



G1與G3的特徵分布



G2與G3的特徵分布

Project Introduction

- 我們根據葡萄牙學校的評分標準，將G3(0-20)的成績區分為及格(>=10)以及不及格(<10)

- 新增一個特徵為''Pass''方便後續分析各類特徵
  (及格：Pass = 1，不及格：Pass = 0)

Student Personal Data

Student Personal Data

- Sex('F' - Female, 'M' - Male)

  前處理：

  將'F'轉為0，'M'轉為1

  學生性別為男生→有較高的及格率

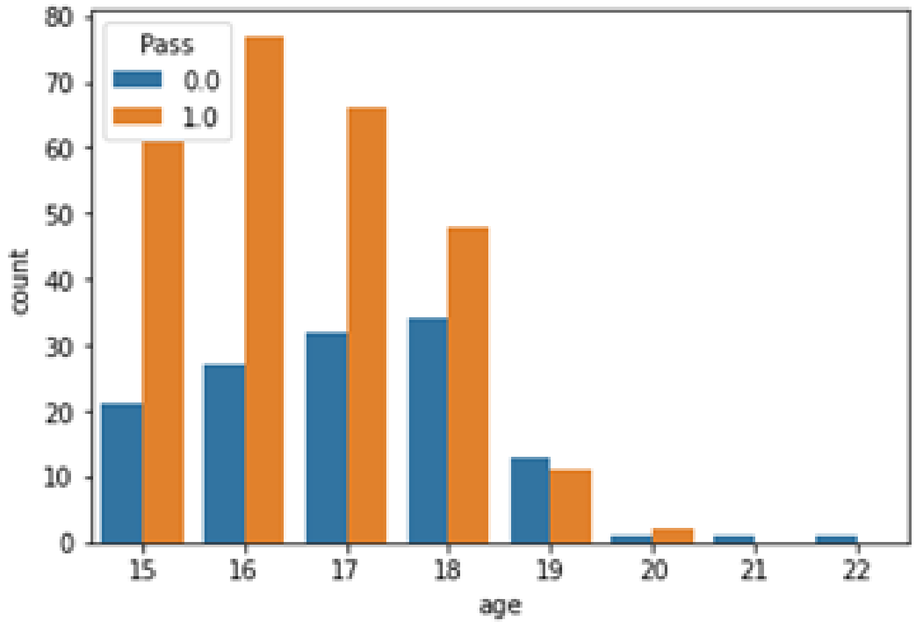
	sex	Pass
0	F	0.639
1	M	0.706

及格率

Student Personal Data

- Age(15 - 22 years old)

  19、21及22的及格率較低，年齡在20歲的學生有最高的平均成績。



	age	Pass
0	15	0.744
1	16	0.740
2	17	0.673
3	18	0.585
4	19	0.458
5	20	0.667
6	21	0.000
7	22	0.000

及格率

	age	G3
0	15	11.256
1	16	11.029
2	17	10.276
3	18	9.549
4	19	8.208
5	20	14.000
6	21	7.000
7	22	8.000

平均分數

Student Personal Data

- traveltime(1-4 hrs)

  通勤時間越長

  →成績有下降的趨勢

	traveltime	Pass
0	1	0.689
1	2	0.636
2	3	0.652
3	4	0.625

及格率

	traveltime	G3
0	1	10.782
1	2	9.907
2	3	9.261
3	4	8.750

平均分數

Student Personal Data

- health(1- very bad ~ 5 - very good)

  身體健康狀態最差的學生

  →有最高的及格率和平均成績

	health	Pass
0	1	0.787
1	2	0.667
2	3	0.659
3	4	0.636
4	5	0.658

及格率

	health	G3
0	1	11.872
1	2	10.222
2	3	10.011
3	4	10.106
4	5	10.397

平均分數

Student Personal Data

- school('GP' - Gabriel Pereira, 'MS' - Mousinho da Silveira)

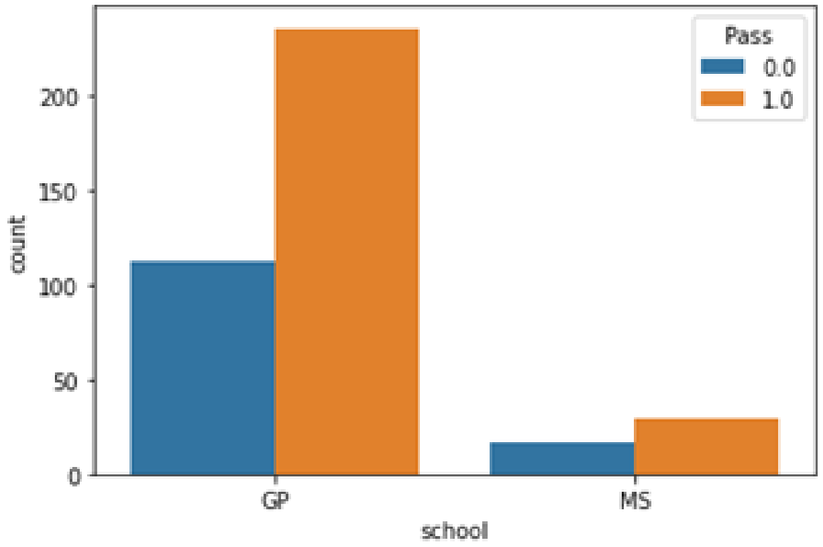
  前處理：

  GP轉為0，MS轉為1

學校為GP的學生→ 有較高的平均分數

school	G3
0 GP	10.490
1 MS	9.848

平均分數



Student Personal Data

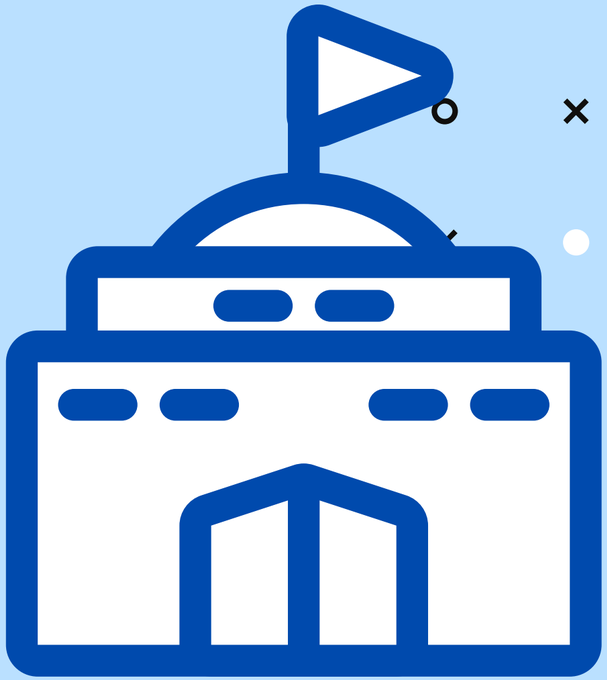
- 最後使用特徵：
  Sex、Age、Traveltime、Health、G1、G2

LinearRegression test error is：6.43

RandomForestRegressor test error is：6.56

KNeighborRegressor test error is：5.83

Student Extra Education



Student Extra Education

- nursery(yes/no)

  前處理：

  yes轉為1，no轉為0

  →成績上影響差異不大

- paid(yes/no)

  前處理：

  yes轉為1，no轉為0

  有參與付費課程的學生

  →成績表現較好

	nursery	Pass
0	no	0.679
1	yes	0.669

及格率

	nursery	G3
0	no	9.951
1	yes	10.535

平均分數

	paid	Pass
0	no	0.631
1	yes	0.718

及格率

	paid	G3
0	no	9.986
1	yes	10.923

平均分數

Student Extra Education

- 新增特徵：extra_course

  前處理：

  nursery + paid

  extra_course = 0 - 沒有參加過額外課程

  extra_course = 1 - 參加過一項額外課程

  extra_course = 2 - 兩項額外課程都參加

  兩項皆參加→成績表現較好

	extra_course	G3
0	0	9.404
1	1	10.288
2	2	10.921

平均分數

	extra_course	Pass
0	0	0.596
1	1	0.670
2	2	0.697

及格率

Student Extra Education

- schoolsup(yes/no)

  前處理：

  yes轉為1，no轉為0

  未參加學校額外課程→成績表現較好

	schoolsup	Pass
0	no	0.689
1	yes	0.549

及格率

Student Extra Education

- 新增特徵：difschool_sup

  前處理：

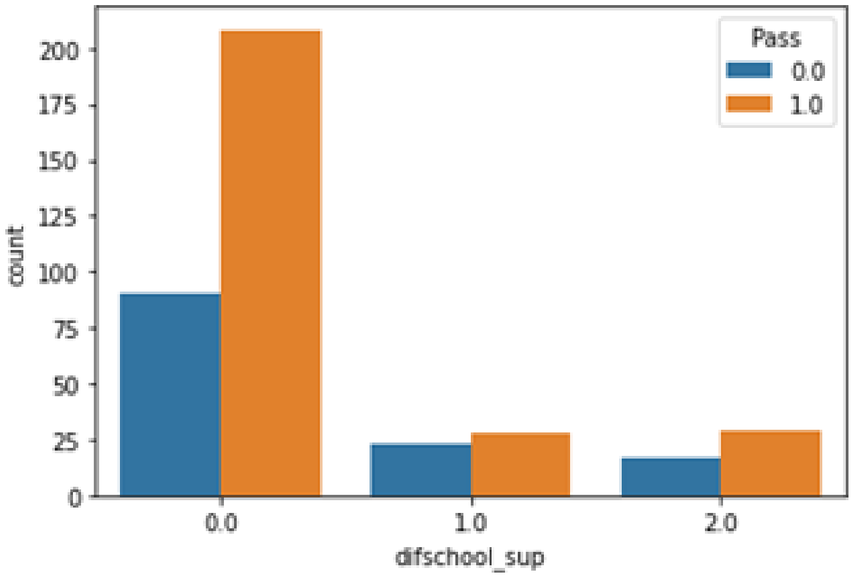
  結合school及schoolsup

  學校為GP者，若無參加schoolsup為0

  學校為GP者，若有參加schoolsup為1

  學校為MS者，若無參加schoolsup為2

  學校為MS者，若有參加schoolsup為3



Student Extra Education

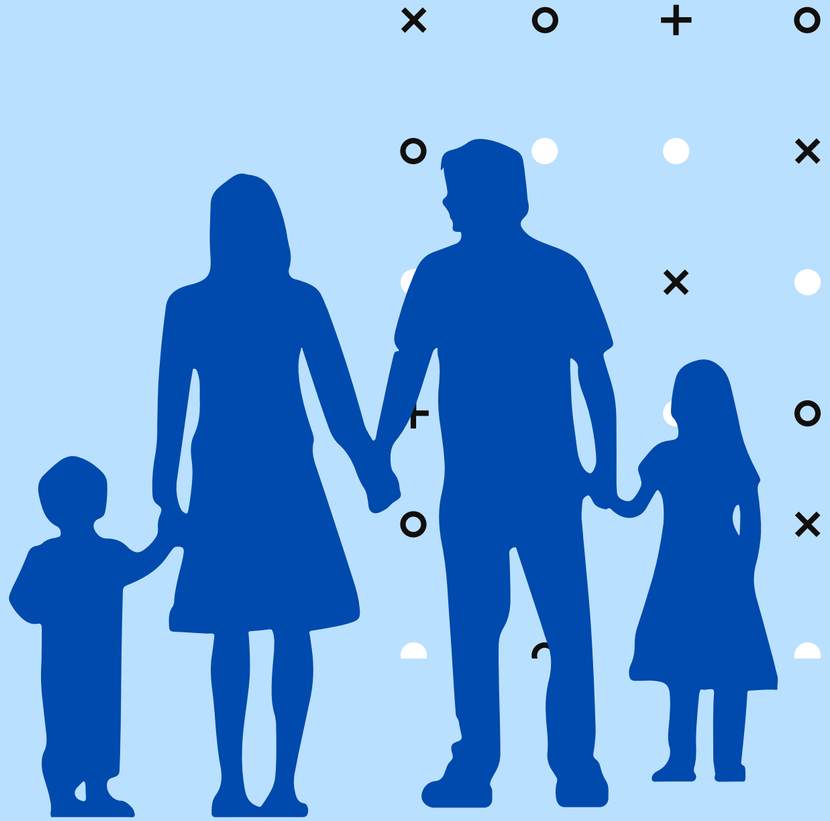
- 最後使用特徵：
  extra_course、difschool_sup、G1、G2

LinearRegression test error is：6.39

RandomForestRegressor test error is：7.42

KNeighborRegressor test error is：6.96

Student Family Factors



Student Family Factors

- Pstatus(父母是否同居：'T'同居 ，'A'沒有同居)

父母沒有同居的學生

→平均分數和及格百分比都比較高

前處理：T轉為1，A轉為0

	Pstatus	Pass
0	A	0.732
1	T	0.664

及格百分比

	Pstatus	G3
0	A	11.195
1	T	10.325

平均分數

Student Family Factors

- 新增特徵：g_edu(監護人教育程度)
  前處理：
  guardian(監護人)結合Medu/Fedu

  0~4 :監護人的教育程度，5監護人為other :
  (1) 0的資料數很少→誤差大
  (2) 1~4 監護人教育程度越高
  →平均分數越高

	g_edu	G3
0	0.0	15.000
1	1.0	9.020
2	2.0	9.522
3	3.0	10.489
4	4.0	11.873
5	5.0	9.062

平均分數

Student Family Factors

- famrel(家庭關係)

  (1) famrel=0 資料很少筆→誤差大

  (2) 家庭關係越好

  →平均分數越高

	famrel	Pass
0	1	0.750
1	2	0.611
2	3	0.618
3	4	0.672
4	5	0.708

及格百分比

	famrel	G3
0	1	10.625
1	2	9.889
2	3	10.044
3	4	10.359
4	5	10.830

平均分數

Student Family Factors

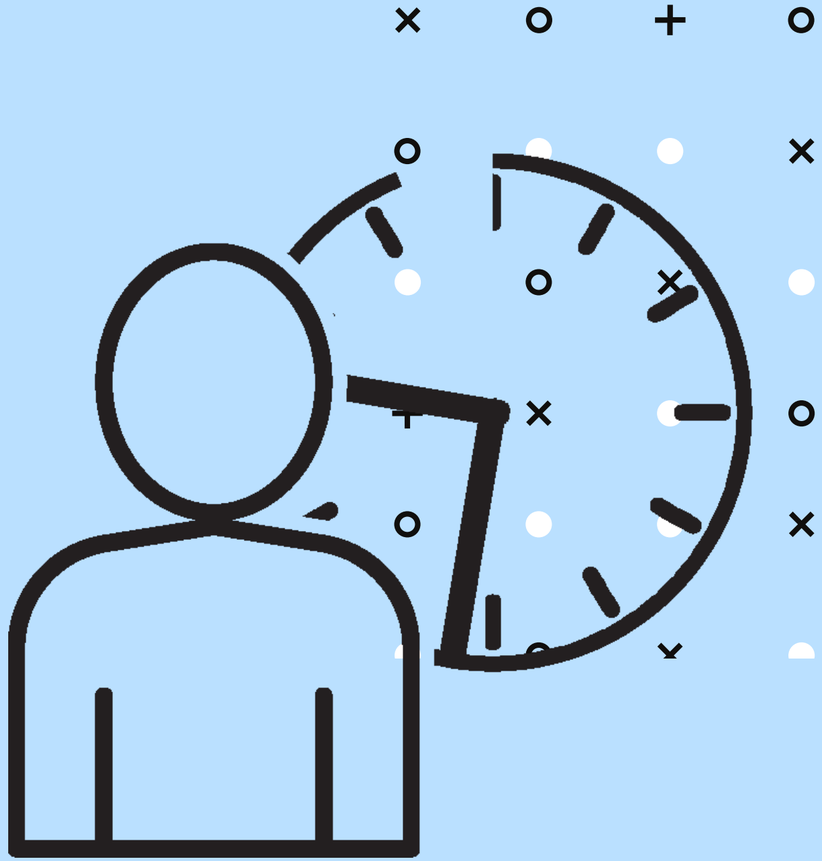
- 最後使用特徵：
  G1、G2、g_edu、Pstatus、famrel

LinearRegression test error is：6.46

RandomForestRegressor test error is：7.00

KNeighborRegressor test error is：7.05

Student Self Management



Student Self Management

- 時間運用方面(freetime、studytime、goout)

	freetime	pass
0	1	0.632
1	2	0.766
2	3	0.643
3	4	0.643
4	5	0.725

	studytime	pass
0	1	0.648
1	2	0.646
2	3	0.754
3	4	0.741

	goout	pass
0	1	0.739
1	2	0.767
2	3	0.715
3	4	0.558
4	5	0.528

Student Self Management

- 時間運用方面(freetime、studytime、goout)

  新建立特徵：spendout(freetime-goout)

  學生花越多課後時間和朋友出去

  →通過率越低

	spendout	pass
0	-4	0.000
1	-3	0.400
2	-2	0.581
3	-1	0.647
4	0	0.646
5	1	0.724
6	2	0.741
7	3	0.929
8	4	0.750

Student Self Management

- failures(0-3)

  學生過去不及格越多次→通過率越低

	failures	pass
0	0	0.750
1	1	0.480
2	2	0.176
3	3	0.250

Student Self Management

- absences(0-93)
  前處理：
  KBinsDiscretizer
  (uniform, n_bins=4)

  出席率越低的學生→通過率越低

	kbd_absences	pass
0	0.0	0.677
1	1.0	0.600
2	2.0	0.500
3	3.0	0.000

Student Self Management

- higher(yes/no)

  前處理：LabelEncoder

  想接受高等教育的學生→通過率越高

- romantic(yes/no)

  前處理：LabelEncoder

  具有戀愛關係的學生→通過率較低

	higher	pass
0	no	0.350
1	yes	0.688

	romantic	pass
0	no	0.703
1	yes	0.606

Student Self Management

- 最後使用特徵：
  G1、G2、failures、absences、spendout、higher、romantic

LinearRegression test error is：6.17

RandomForestRegressor test error is：4.96

KNeighborRegressor test error is：4.97

Result and Conclusion

同模型不同類別特徵比較

3>1>2>4

	3. 自我管理	1. 基本資料	2. 家庭因素	4. 額外教育
Linear	6.17	6.43	6.46	6.39
Random Forest	4.96	6.56	7.00	7.42
KNeighbor	4.97	5.83	7.05	6.96

同類別特徵不同模型比較

- Regressor

  1. Linear：容易過擬合，將連續資料轉為區間可提高準確率。

  2. RandomForest：抗過擬合能力高，特徵幾乎不需要前處理。

  3. KNeighbor：將特徵歸一化後有助於提高準確率。

RandomForest參數變化(n_estimators)

n_estimators	Error
100(default)	4.96
200	4.90
500	4.86

- n_estimators越大
  →由越多決策樹組成
  →越不容易過擬合
  →error越低

KNeighbor參數變化(n_neighbors)

n_neighbors	Error
1	8.55
3	6.46
5(default)	4.97
10	5.45
15	5.85

- n_neighbors 越小
  →採用越少鄰資料
  →使模型過擬合
  →error越高
- n_neighbors 越大
  →採用越多鄰資料
  →參考到不類似的資料
  →error越高


Thank
You