學生的數學和葡萄牙語成績預測

組別：圖靈觀眾
組員：統計 108 劉臣洋
物理 112 李國成
物理 112 陳昱丞

(0) 報告簡介

本組想探討哪些因素會影響學生學業表現，並由各項特徵中盡可能預測出學生學業表現。以下為幾個想要嘗試的目標：

探討現象

儘管過去幾十年葡萄牙人口的教育水平有所提高，但統計數據由於學生不及格率高葡萄牙一直處於歐洲的尾端。因此我們將探討有哪些因素會對於像是數學和葡萄牙語的核心課程的影響。

預測成績

本組分別用**已知成績**來預測下一次的考試成績，和**未知考試成績**的狀況下來預測成績的表現。

由於每次考試成績多少會有一些程度的波動，不過用已知成績預測下次成績的模型若有一定的準確率，可以幫助我們更容易發現進步或退步不尋常的學生。能幫助判斷並鼓勵進步的同學，或注意到成績退步得不合理的同學。

一個家庭的孩子甚至雙胞胎的成績相差都甚，因此我們認為以不和成績有直接關聯的線索來預測成績的誤差一定非常大。在未知成績的狀況下只能大致預測學生成績是偏好或偏壞，對於其準確率需要進一步探討。

尋找成績極端的學生

回歸問題下，中等水平的學生應該比較沒有什麼研究的價值。我們可以把目標朝向尋找極端差或極端好的學生。不過最大的問題是數據量不夠，也不平衡。雖然難度最高，但可以嘗試從成績極端差或好的學生的數據中提取特徵，甚至嘗試用模型去自動化尋找潛在的學生。

因為資料不平衡的問題，我們會先著重在從分類及格和不及格下手。

(1) 資料描述與前處理

資料描述

有 30 個特徵點以及期末分數

1 school - student's school (binary: 'GP' - Gabriel Pereira or 'MS' - Mousinho da Silveira)
2 sex - student's sex (binary: 'F' - female or 'M' - male)
3 age - student's age (numeric: from 15 to 22)
4 address - student's home address type (binary: 'U' - urban or 'R' - rural)
5 famsize - family size (binary: 'LE3' - less or equal to 3 or 'GT3' - greater than 3)
6 Pstatus - parent's cohabitation status (binary: 'T' - living together or 'A' - apart)
7 Medu - mother's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
8 Fedu - father's education (numeric: 0 - none, 1 - primary education (4th grade), 2 â€“ 5th to 9th grade, 3 â€“ secondary education or 4 â€“ higher education)
9 Mjob - mother's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
10 Fjob - father's job (nominal: 'teacher', 'health' care related, civil 'services' (e.g. administrative or police), 'at_home' or 'other')
11 reason - reason to choose this school (nominal: close to 'home', school 'reputation', 'course' preference or 'other')
12 guardian - student's guardian (nominal: 'mother', 'father' or 'other')
13 traveltime - home to school travel time (numeric: 1 - <15 min., 2 - 15 to 30 min., 3 - 30 min. to 1 hour, or 4 - >1 hour)
14 studytime - weekly study time (numeric: 1 - <2 hours, 2 - 2 to 5 hours, 3 - 5 to 10 hours, or 4 - >10 hours)
15 failures - number of past class failures (numeric: n if 1<=n<3, else 4)
16 schoolsup - extra educational support (binary: yes or no)
17 famsup - family educational support (binary: yes or no)
18 paid - extra paid classes within the course subject (Math or Portuguese) (binary: yes or no)
19 activities - extra-curricular activities (binary: yes or no)
20 nursery - attended nursery school (binary: yes or no)
21 higher - wants to take higher education (binary: yes or no)
22 internet - Internet access at home (binary: yes or no)
23 romantic - with a romantic relationship (binary: yes or no)
24 famrel - quality of family relationships (numeric: from 1 - very bad to 5 - excellent)
25 freetime - free time after school (numeric: from 1 - very low to 5 - very high)
26 goout - going out with friends (numeric: from 1 - very low to 5 - very high)
27 Dalc - workday alcohol consumption (numeric: from 1 - very low to 5 - very high)
28 Walc - weekend alcohol consumption (numeric: from 1 - very low to 5 - very high)
29 health - current health status (numeric: from 1 - very bad to 5 - very good)
30 absences - number of school absences (numeric: from 0 to 93)

these grades are related with the course subject, Math or Portuguese:
31 G1 - first period grade (numeric: from 0 to 20)
31 G2 - second period grade (numeric: from 0 to 20)
32 G3 - final grade (numeric: from 0 to 20, output target)

資料前處理

1. 使用 onehot encoding：將資料中有的類別資料轉換為數值
2. 將成績改變為 failed 和 pass：以三次考試成績加總後的一半做為基準，並將考生的成績判定為過和沒過

檢討：由於考試成績為連續型資料，直接分為過與不過比較不精準，下次可以嘗試分為五個等第(超級爛、爛、中等、好、超級好)

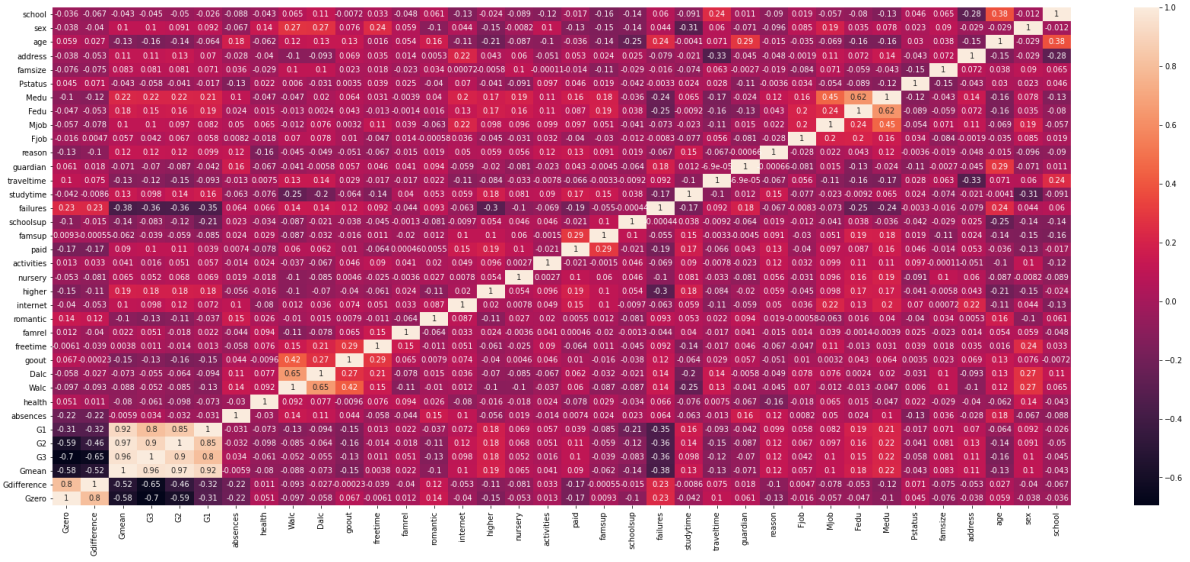
(2) 從數據中發現的見解

1. 觀察數學與葡萄牙語資料是否為相同的一批測試者

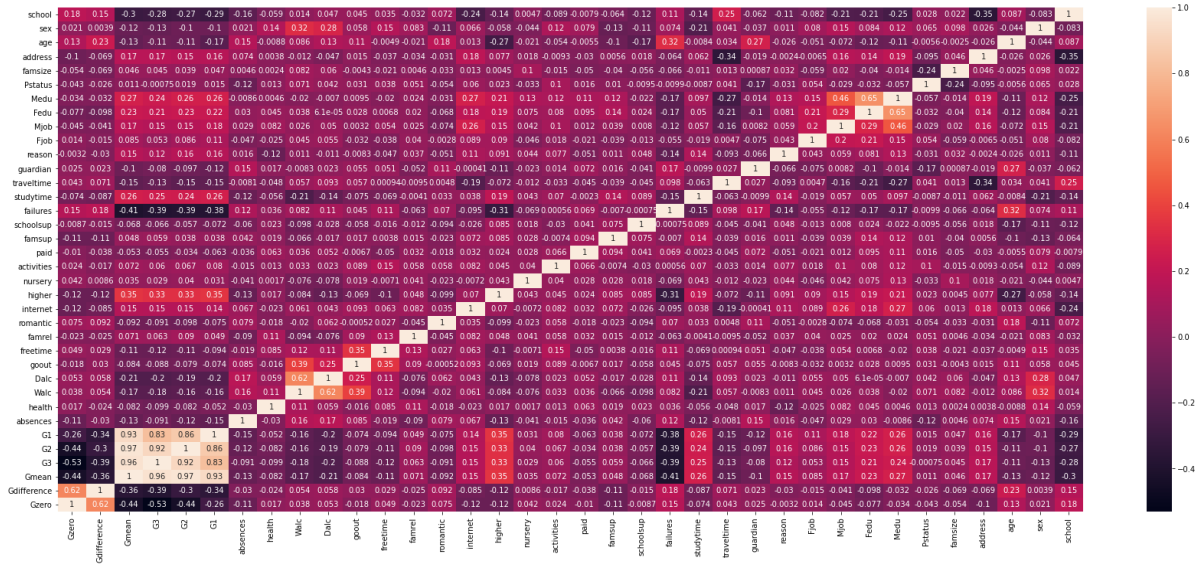
兩筆資料重複的人：數學和葡萄牙資料完全重複的筆數有 39 筆，意味着至少有 39 個學生是同一批人。但因為人數太少不繼續做更深入的探討。

2. 從熱圖判斷一些相關特性

Math correlation heatmap



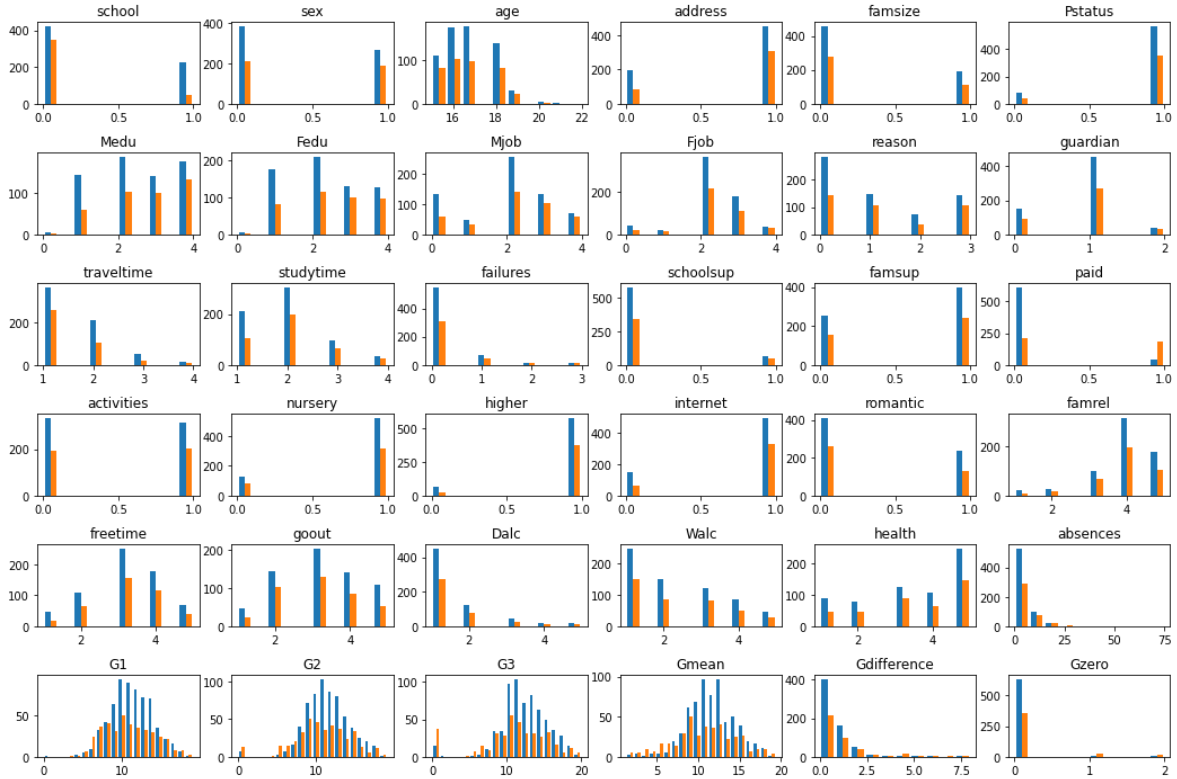
Portuguese correlation heatmap



1. 成績幅度與成績為負相關，看出成績波動越大者其成績越不理想
2. 周末喝酒的和平日喝酒的大致都是同一群人(相關性 0.69)
3. 通常外出的也較常喝酒，特別是假日，推論出由於平日工作，因此假日是適合休息享受的日子(相關性 0.42)
4. 從父母教育的相關程度顯示通常會與相同教育程度者結識(相關性 0.62)

3. 從特徵分配圖觀察相關特性

葡萄牙和數學的特徵分配圖（藍色為葡萄牙語、橙色為數學）



1. 由此兩張圖可看出兩者特徵分布圖除了分數分布有些差異，其餘大致相同。
2. 成績不太符合高斯常態分佈，左右不對稱，可以推論出老師有調整分數的動作。
3. 考 0 分的人在第二次和第三次逐漸增加
4. 分數變動數學較葡萄牙語大

分數中位數及標準差

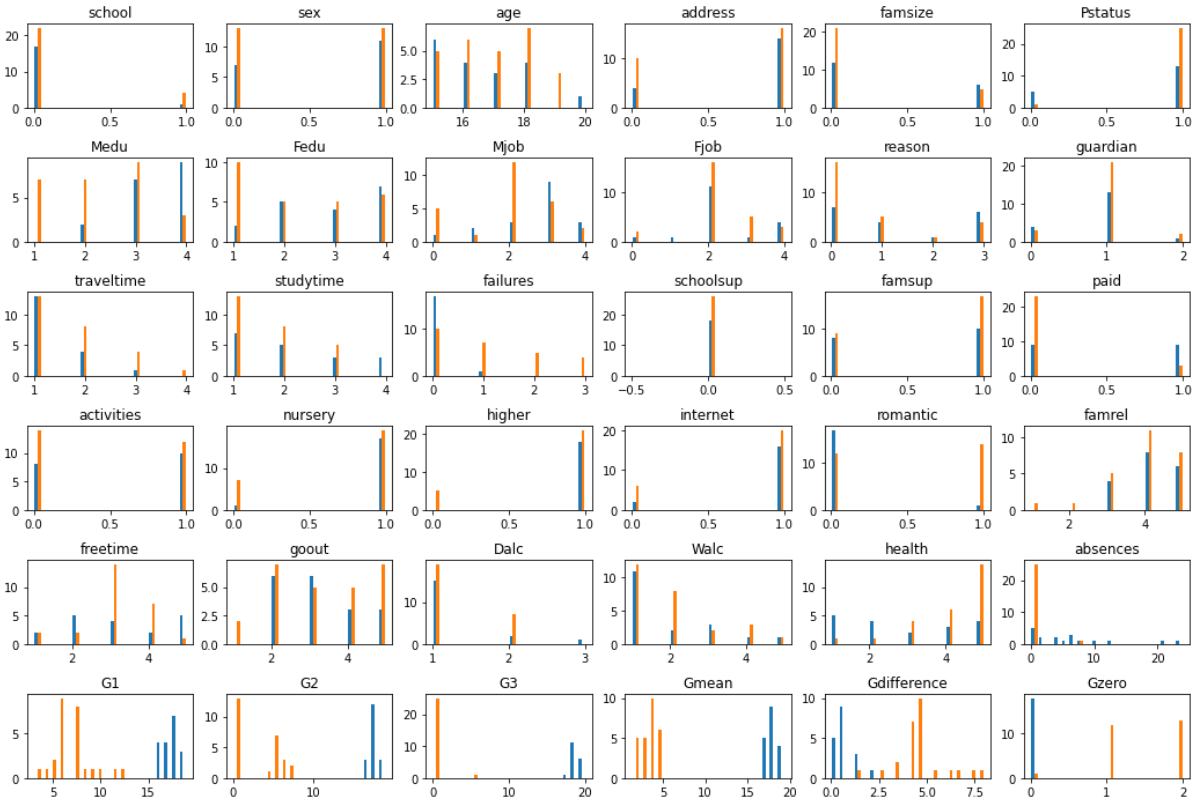
	Portugues		Math	
	Mean	**Standard Dev**	**Mean**	**Standard Dev**
G1	11.399076	2.743149	10.908861	3.31499
G2	11.570108	2.911393	10.713924	3.75674
G3	11.906009	3.228166	10.41519	4.57564
Avg	11.625064	2.831177	10.679325	3.692103

4. 從極端成績分配圖觀察相關特性

數學極端成績分配圖

藍色：成績高於平均 1.7 標準差，約 17 分，26 個樣本

橙色：成績低於平均 1.7 標準差，約 4.4 分，26 個樣本



這些數學的分配圖主要可以觀察到成績表現非常差的同學有以下特徵：

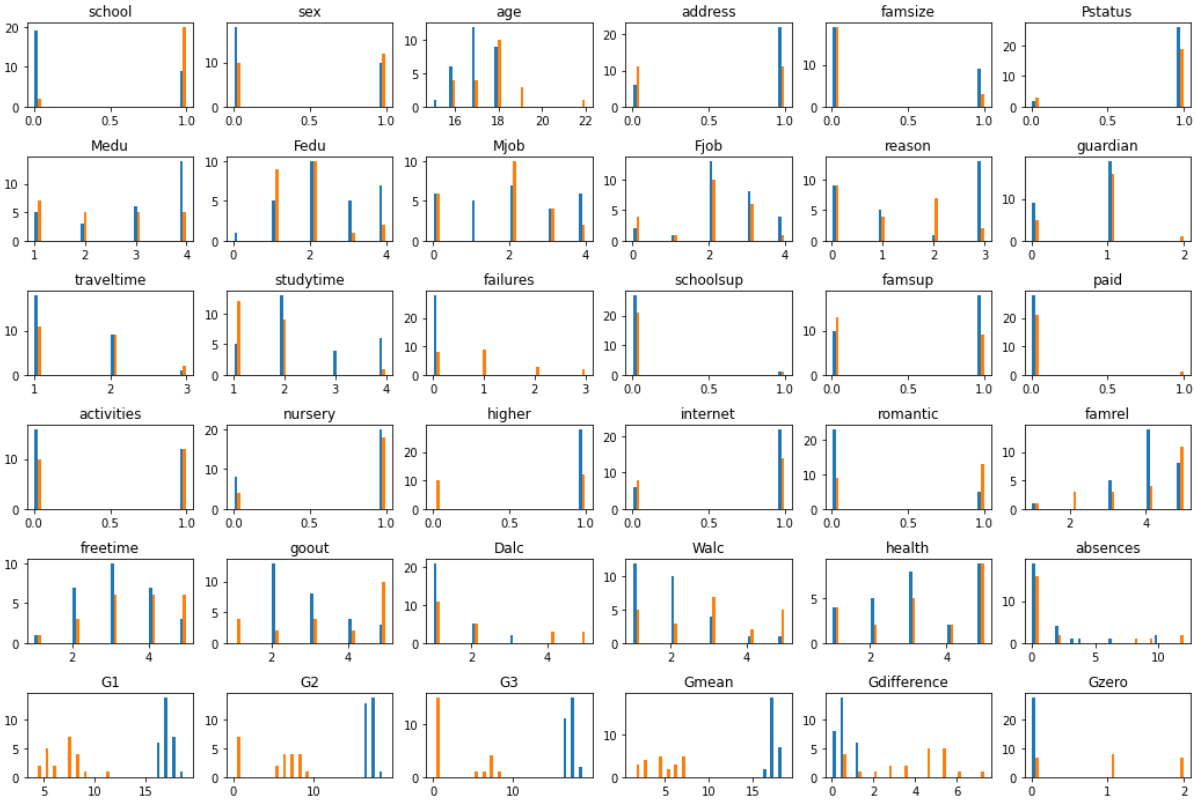
1. 同學父母教育程度都偏低

2. 有談戀愛的比例相對高
3. 年齡也比較大
4. 住家地址比例也會改變，有搬家過
5. 另外也有大量成績非常好的同學缺席次數非常高。

葡萄牙語極端成績分配圖

藍色：成績高於平均 1.7 標準差，約 16.4 分，28 個樣本

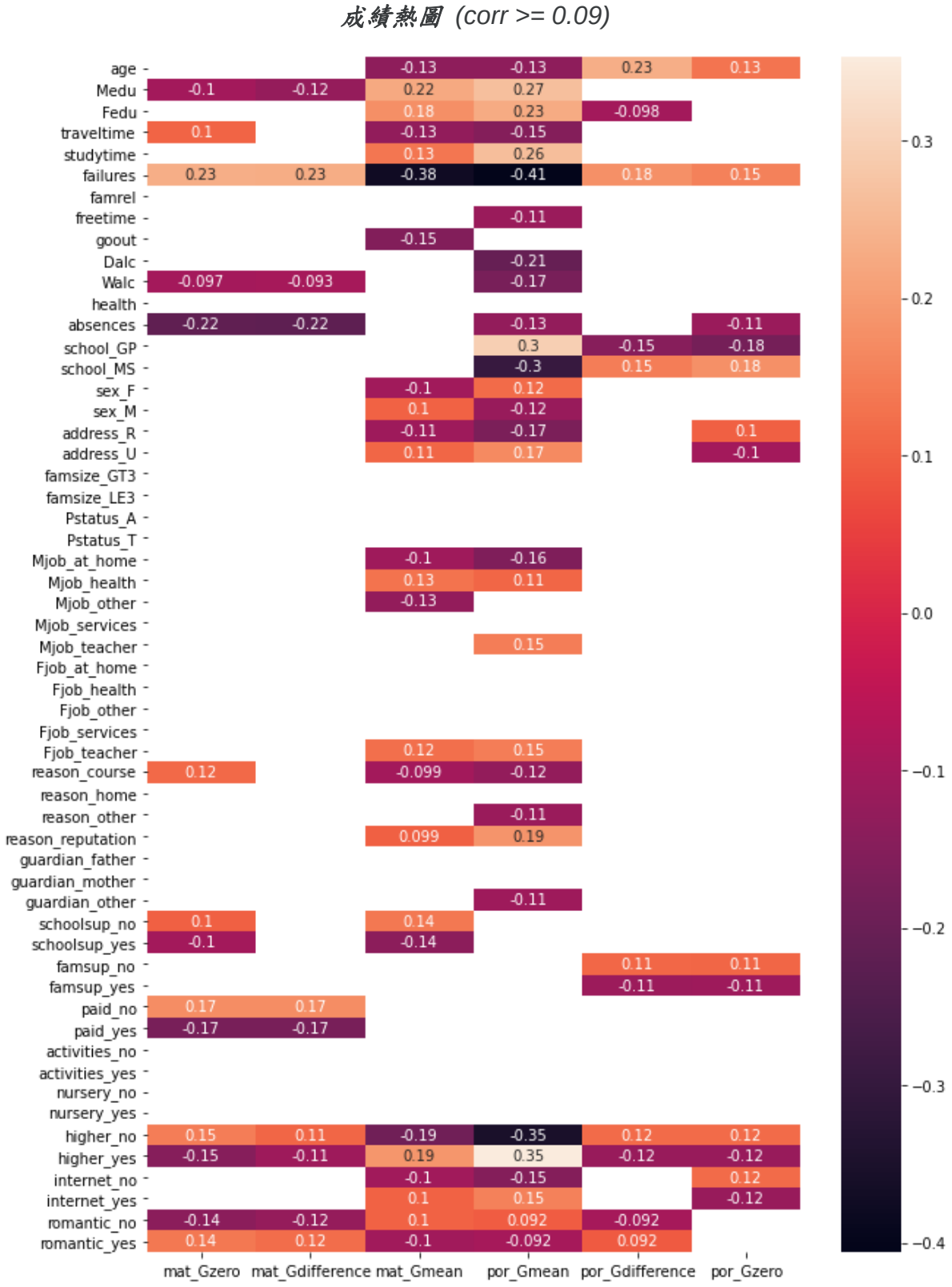
橙色：成績低於平均 1.7 標準差，約 6.8 分，22 個樣本



葡萄牙語也同數學類似，不同的是升學意願、飲酒、自由時間、有更大的關聯性。

而我們也拿閾值為平均分、1、1.7 和 2 個標準差做過比較，發現它們之間的變化並沒有非常明顯，而是漸進式的。這意味着原本我們要尋找成績極端的同學的可行性非常低。並且做其它類似的分群動作也會至少有一個來自數據中個體差異導致的標準差的誤差。這也意味這不靠歷史成績我們很難做準確預測。因為個體差異是無法準確在我們的數據中得到的。並且我們也開始對於之後不參考歷史成績來預測成績的期望並不高。

5. 從分數熱圖觀察相關特性

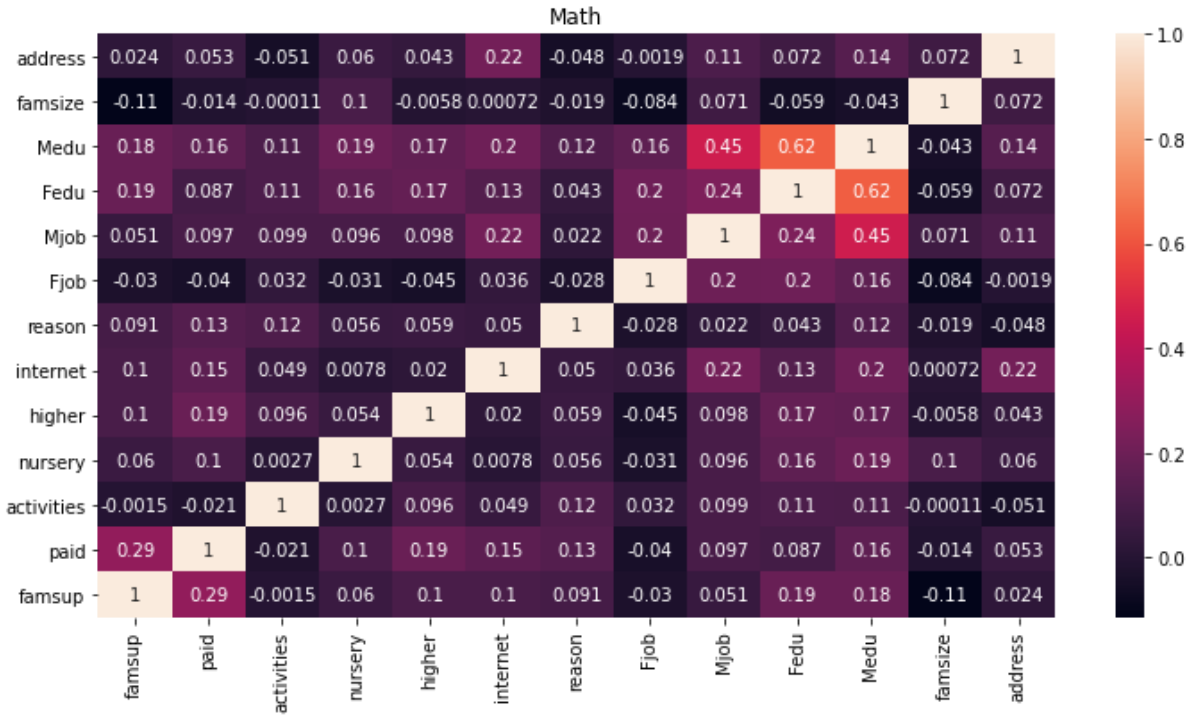


1. 年齡和成績呈現負相關
2. 男生在數學整體表現較佳；女生在葡萄牙語整體表現較佳

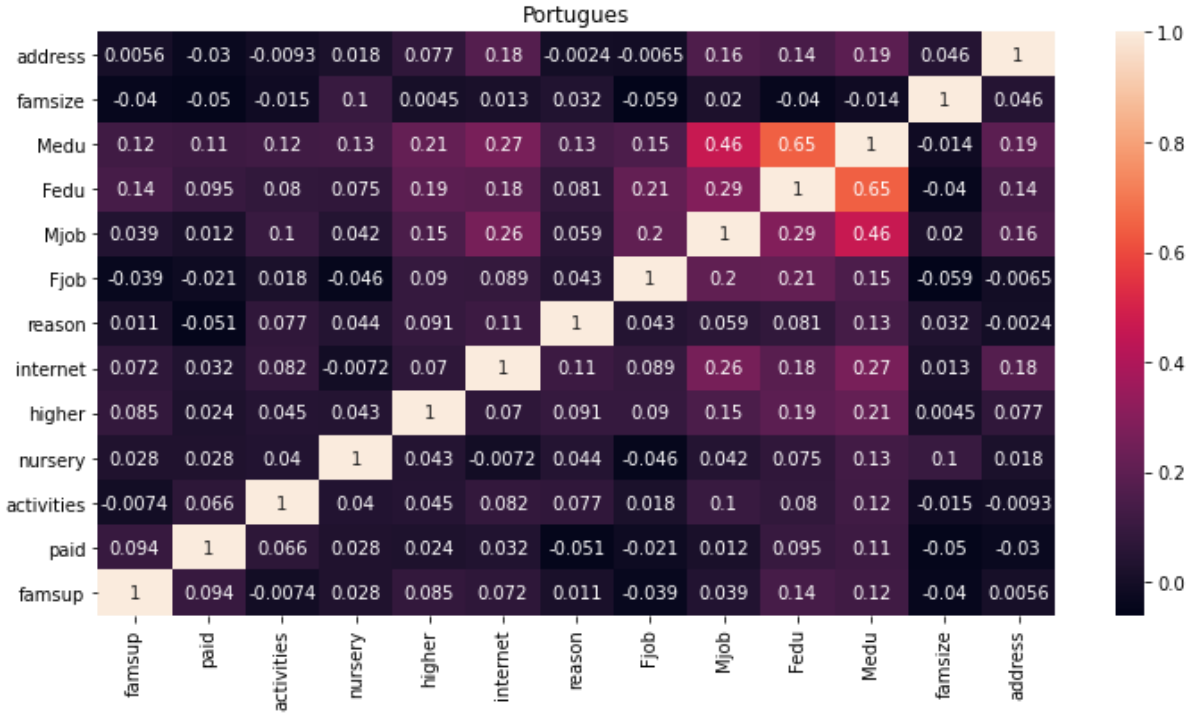
3. 地區、學校對成績有很大關聯性，可看出學校教育對於葡萄牙語呈現一定程度的成績表現
4. 談戀愛關聯性很大，推測會無法專注
5. 缺席造成成績不小的影響
6. 父母的教育程度和孩子有很大的正相關性
7. 除了學校課程額外資源(schoolsup)和數學成績呈負相關，大部分特徵對數學的關聯性比葡萄牙語少很多。

另外我們也從 heatmap 上觀察到一些區塊顯示的有趣的特徵相關性是值得一提的：

數學特徵熱圖

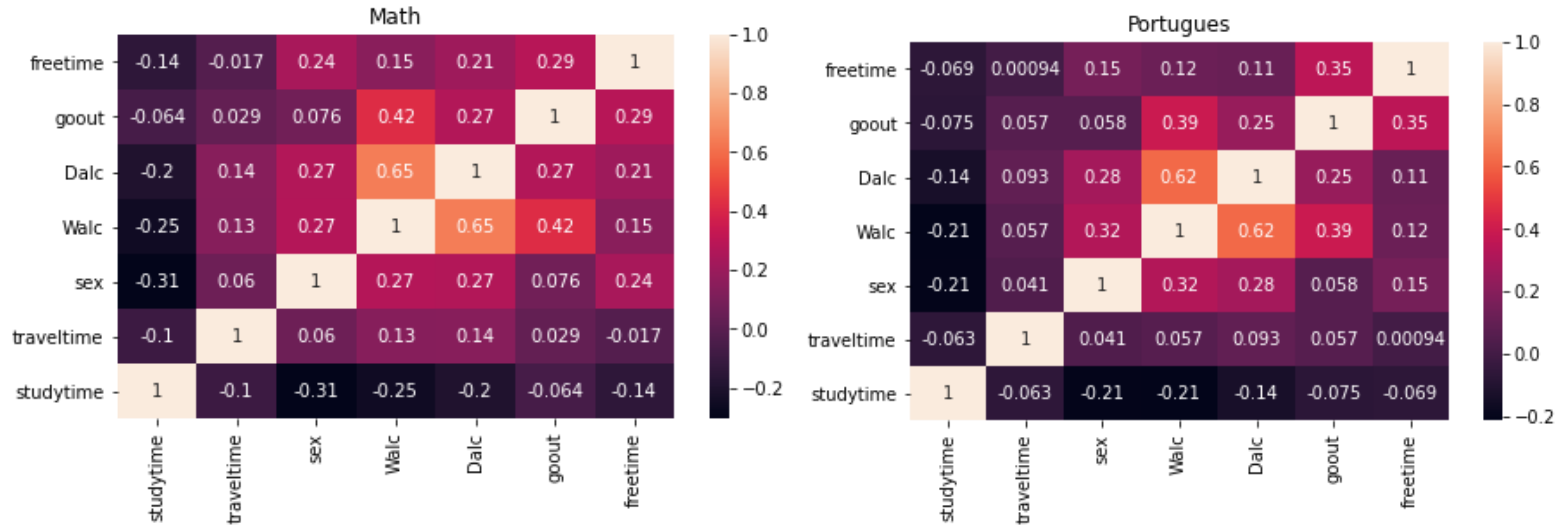


葡萄牙語特徵熱圖



可以觀察出門父母的教育程度和工作有很強烈的(門當戶對)關聯性，並且對於學生的其它背景也有一些影響。但和學習時間、飲酒習慣的相關性不高。

還有另外一個區塊：



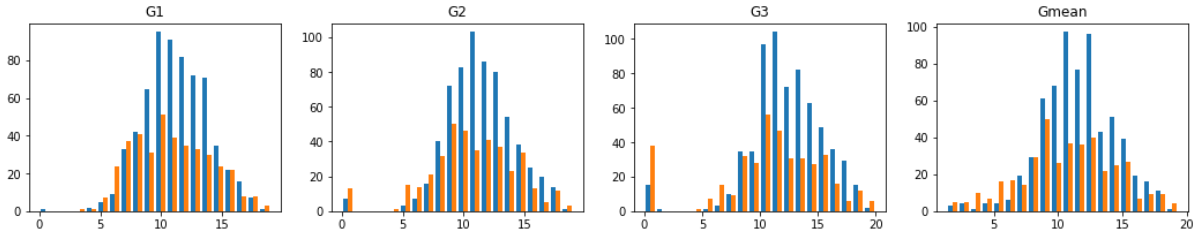
飲酒習慣和自由時間、同朋友出門、性別是有正相關性的。

(3) 執行步驟

資料前處理、分別做幾種分群和回歸預測。

分群 - 判斷及格和不及格：

由於看到了會調分的趨勢(尤其是葡萄牙語)，所以我們把過與不過大略設為總成績的一半，因此模型的預測目標設為三次成績加總的一半當作及格線。後來發現還是有一點小問題，因為三次成績加總的一半，其調分的效應會稍稍往高分偏移。但不影響大致判斷。



1. 分別執行三個模型(RandomForest, Gradient Boosting, XGBoosting)
2. 並以 accuracy，precision， recall 三種評分方式來衡量預測結果的好壞，選擇出最合適的模型。
3. 透過交叉驗證來避免模型過度擬合(overfitting)
4. 透過模型提供的特徵重要性，選擇其中相對重要的特徵(特徵值大於 0.3)
5. 最後，使用這些特徵來預測結果

另外我們大部分的精力都在做這部分的分群。其它分群只是點到為止。

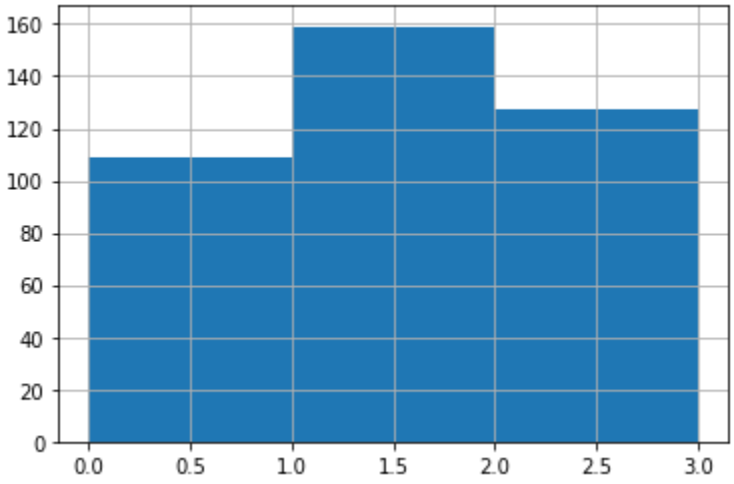
分群 - 分多個群：

除了判斷及格和不及格外，分類學生的發揮價值更大。

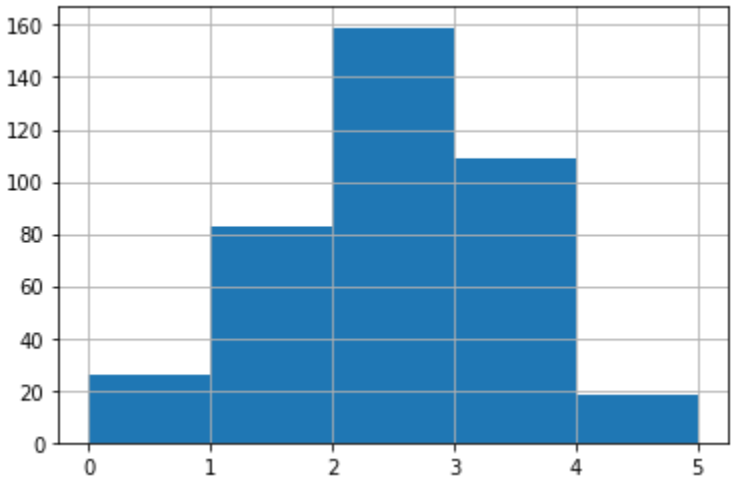
因為時間不夠，分多個群的部分只做初步的 Random Forrest Classifier 便取斷點，也沒有做 resampling 來處理補平衡的問題。並只用 Confusion Matrix 來大致評估其預測可行性。

嘗試兩種組合：

壞、中、好三群(以標準差取成績績切成 3 段)，以下為分配圖



極壞、壞、中、好、極好 5 群（標準差切割 -1.7,-0.5,0.5,1.7），以下為分配圖



可以發現第二個組合有數據不平衡的問題。

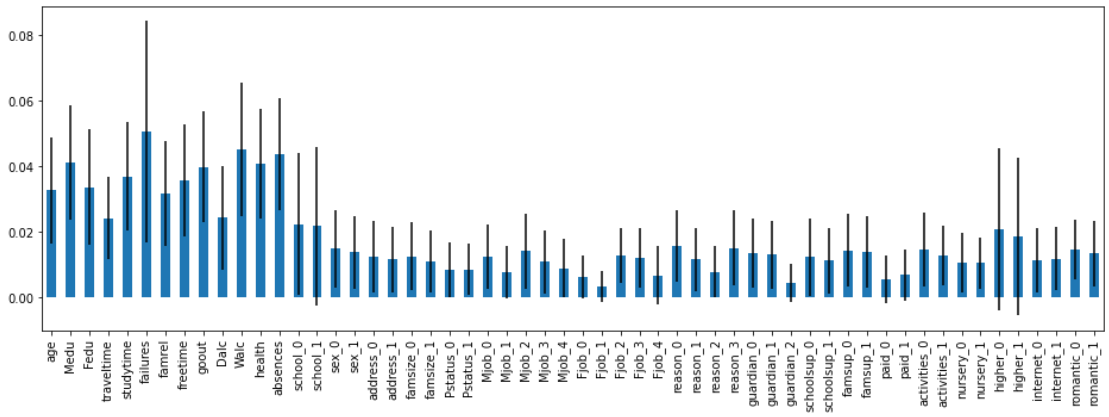
回歸：

嘗試用回歸來預測學生成績。

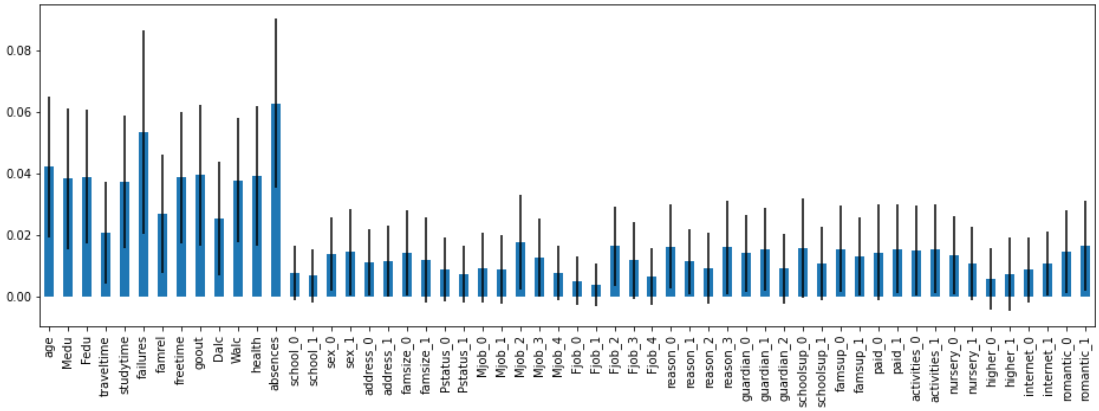
(4) 評估與結果

1. 影響成績的最大因素

葡萄牙語 feature importance:



數學 feature importance：



2. 模型預測成績結果

及格與不及格分群（在訓練集中的結果）：

數學

	accuracy	precision	recall
RandomForest	0.774194	0.744113	0.690330
GradientBoosting	0.661290	0.597222	0.594247
XGBoosting	0.693548	0.636364	0.632191
CatBoost	0.629032	0.496154	0.497552

葡萄牙語

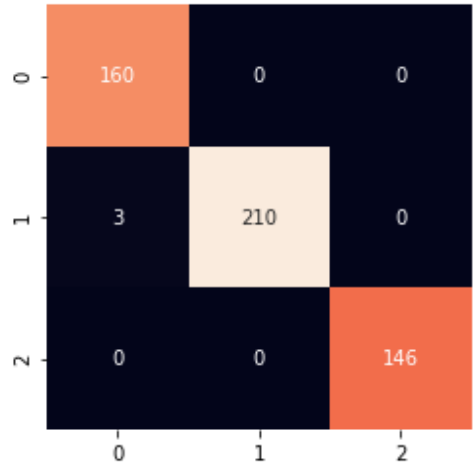
	accuracy	precision	recall
RandomForest	0.84	0.828869	0.729730

GradientBoosting	0.77	0.700000	0.632536
XGBoosting	0.83	0.787523	0.747921
CatBoost	0.79	0.64	0.58

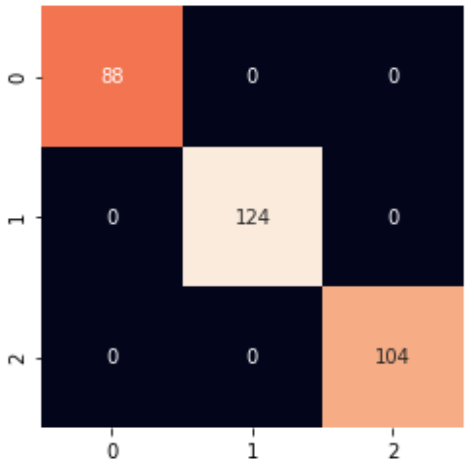
以上兩張圖可以大約看出可以大約判斷及格，不過由於 recall 值不高，因此對於預測不及格不是很理想。

好中壞分群：

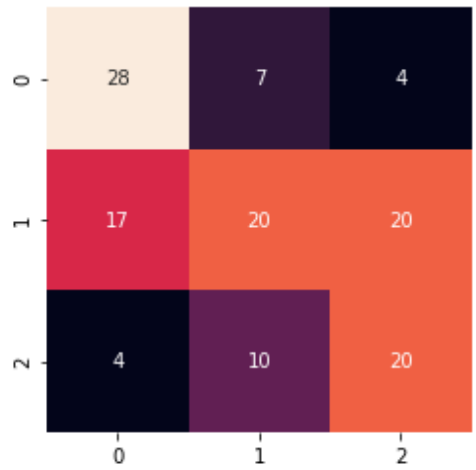
Portuguese
The confusion matrix of tranning set



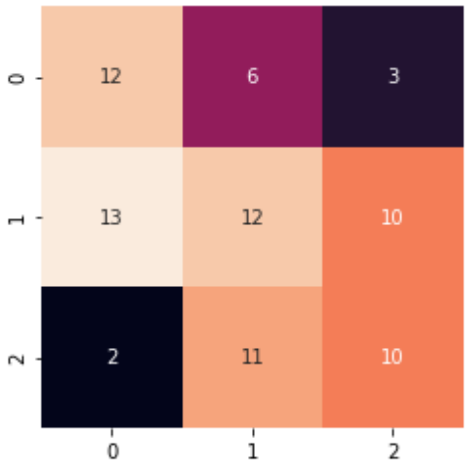
Math
The confusion matrix of tranning set



The confusion matrix of testing set



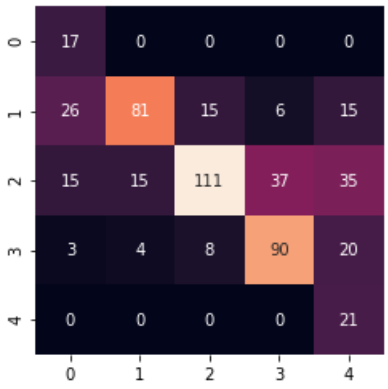
The confusion matrix of testing set



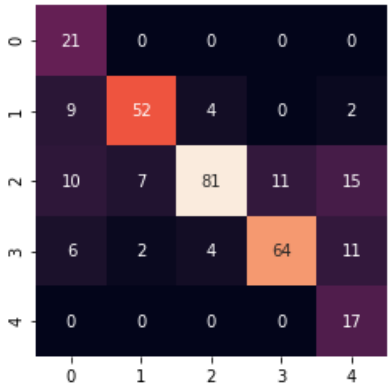
可以發現模型擬合得很不錯，但是預測的部分很容易差到一個類別。也就是說當分類器把學生分類為差時，他有一半概率是中等，但幾乎不可能是優秀。另外，數學相對葡萄牙語更難預測。

極壞、壞、中、好、極好分群：

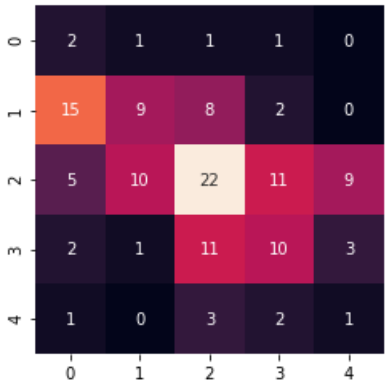
Portuguese
The confusion matrix of tranning set



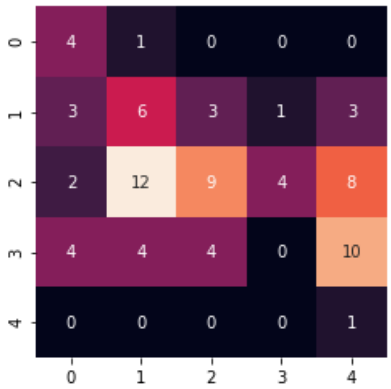
Math
The confusion matrix of tranning set



The confusion matrix of testing set

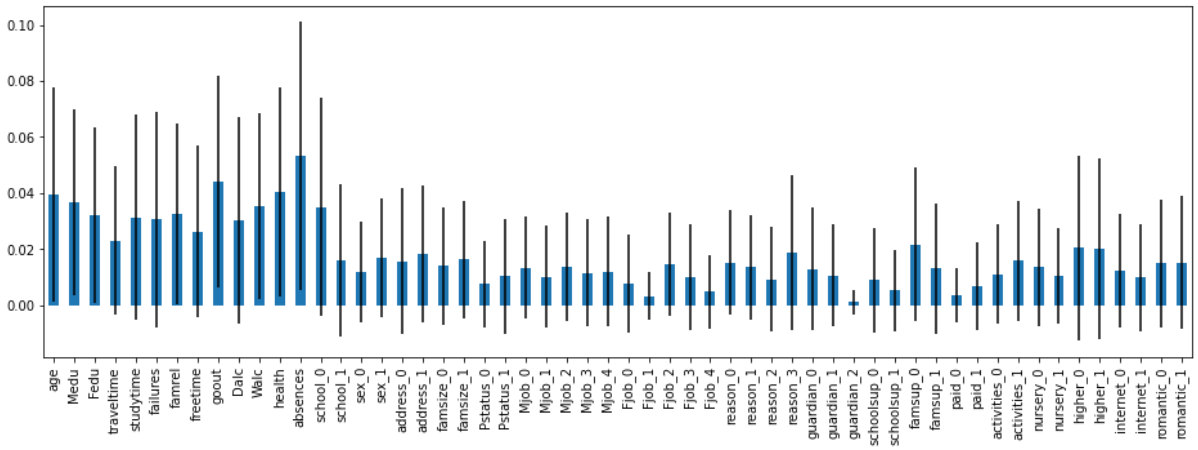


The confusion matrix of testing set

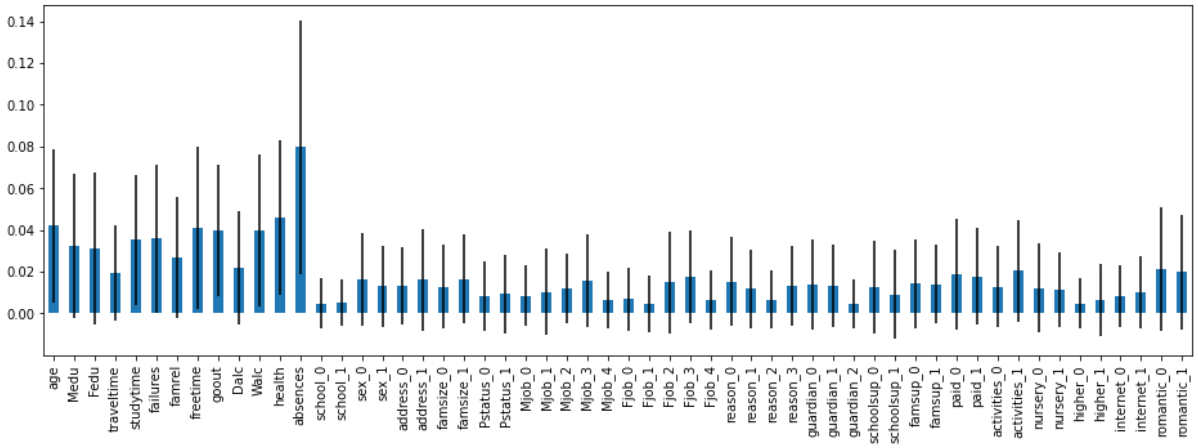


模型趨向預測中等成績。這不意外因為極端成績的數據比較少，是 imbalance data，而我們還沒有嘗試 resampling 的動作。同樣葡萄牙語的準確率比較高。

葡萄牙語 Feature Importance：



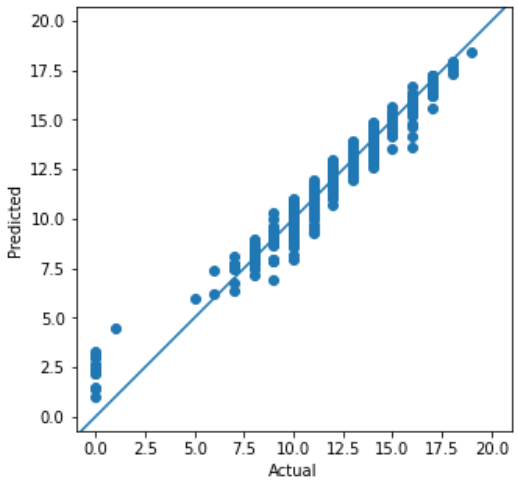
數學 Feature Importance：



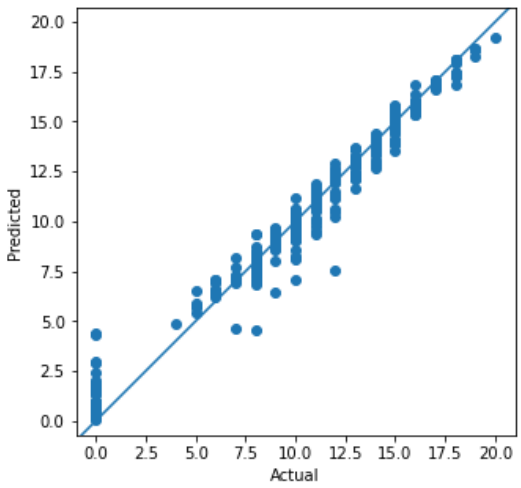
回歸：

RandomForrest Regressor，利用歷史成績預測的結果：

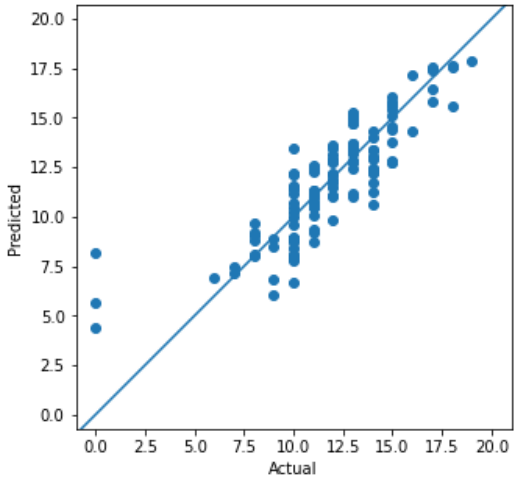
Protuguese
Train Result:



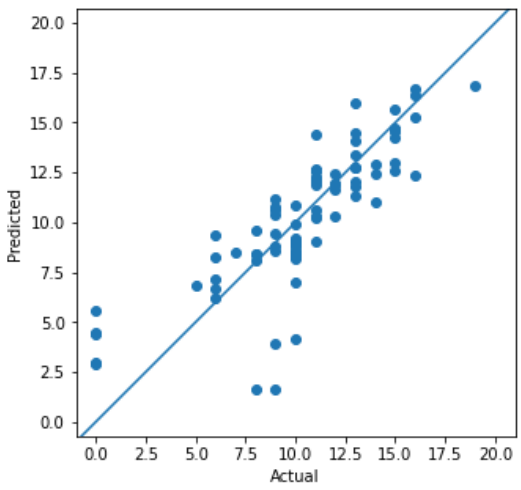
Math
Train Result:



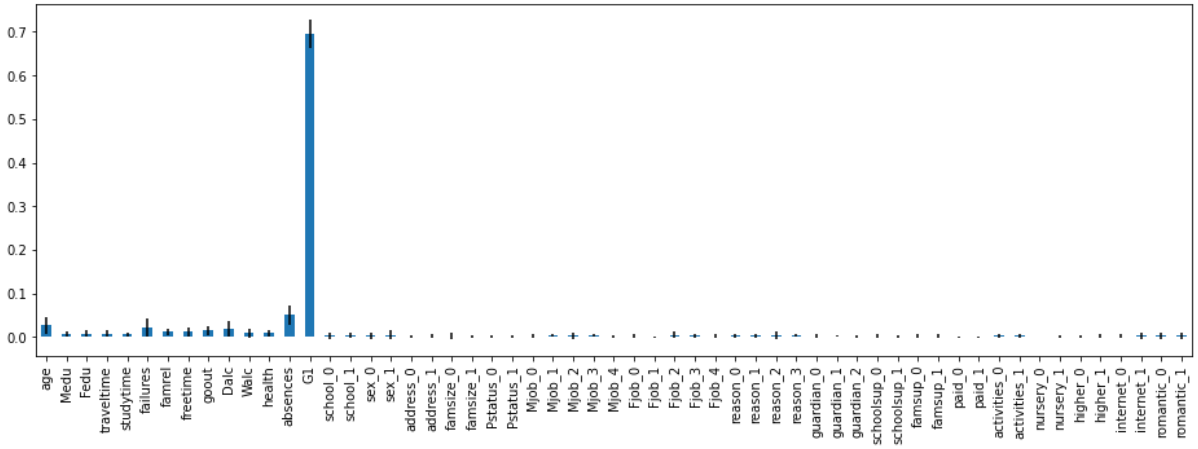
Test Result:



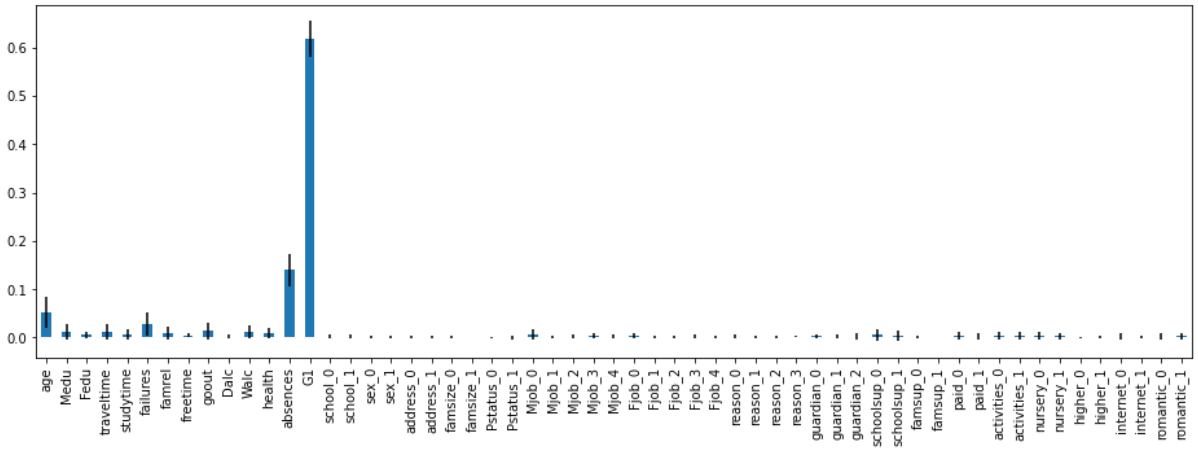
Test Result:



葡萄牙語 Feature Importance:

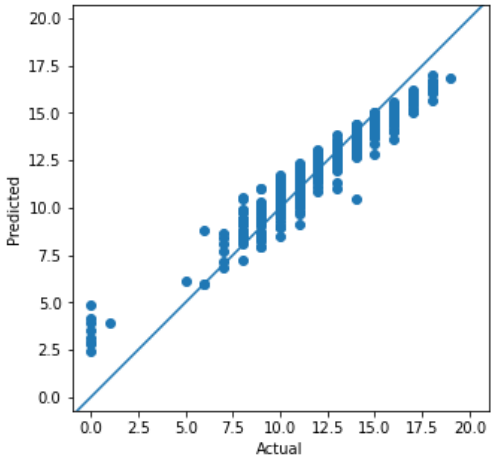


數學 Feature Importance:

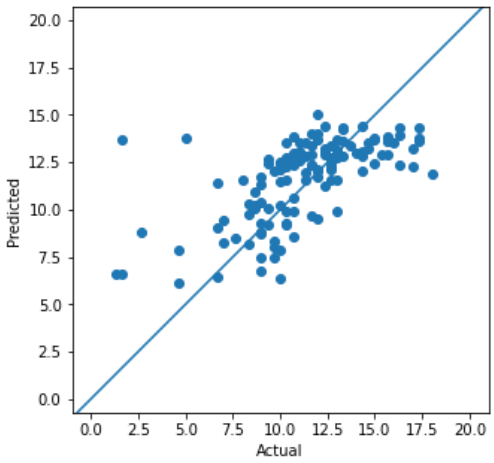


不利用歷史成績預測：

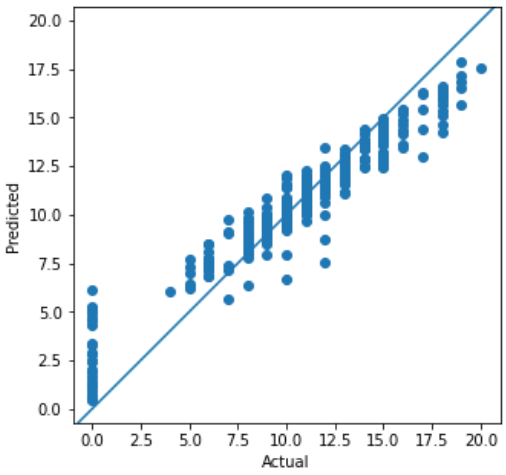
Portuguese
CV nmse: [8.78397404 5.68114423 5.446679
7.011970782300224
Train Result:



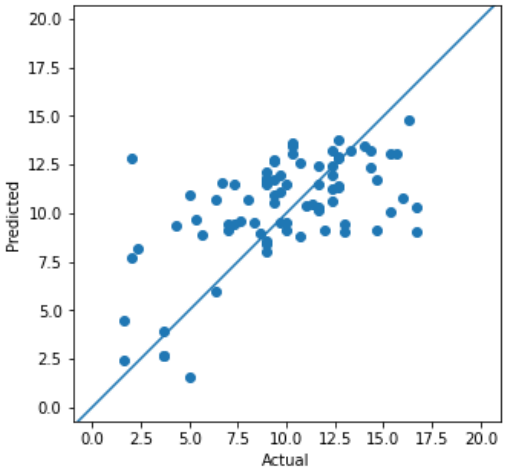
Test Result:



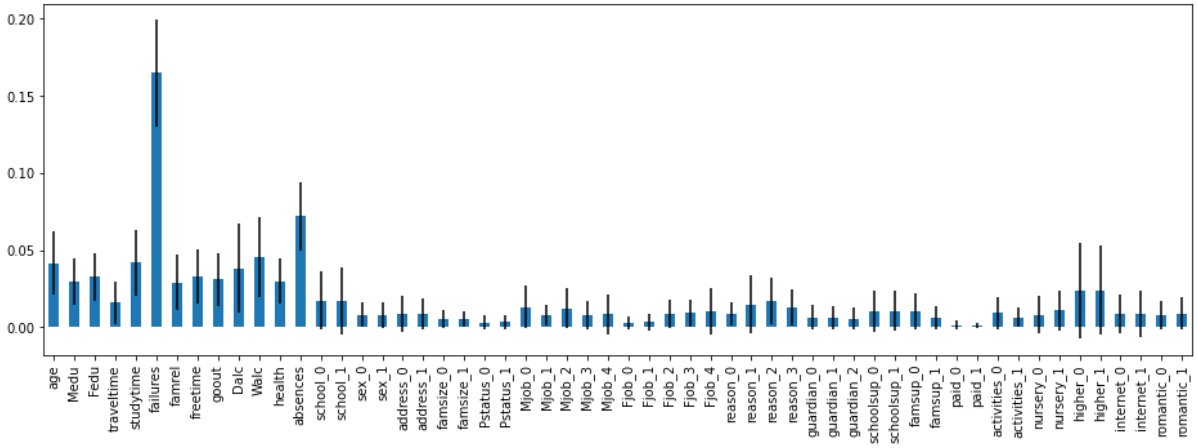
Math
CV nmse: [14.61656562 14.3144873 19.817
15.897880426587301
Train Result:



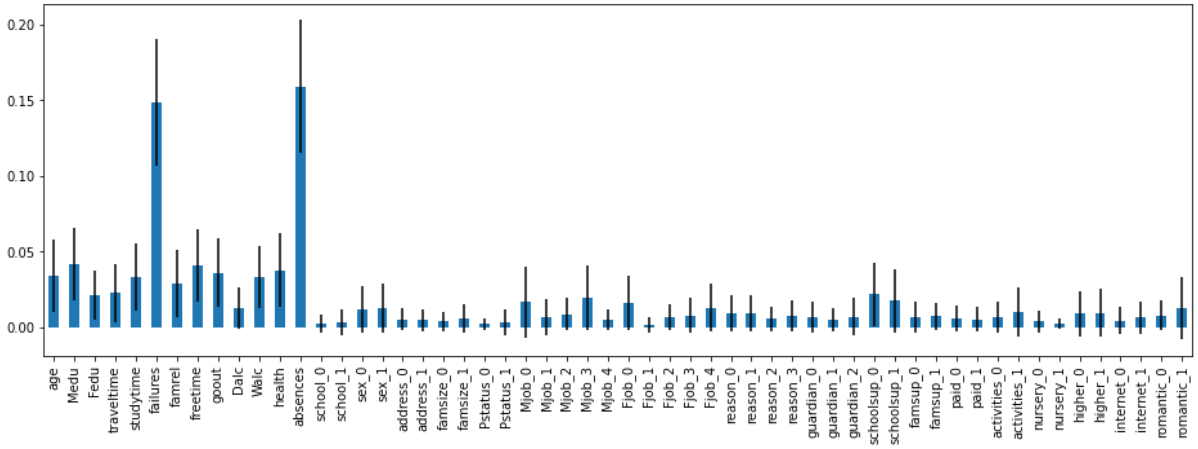
Test Result:



葡萄牙語 Feature Importance



數學 Feature Importance



1. 相較於數學，葡萄牙語的預測結果較好，可以推論出學生背景對於葡萄牙語的影響較大，數學更偏向於額外因素或者是天生的。

2. 數學的 Features Importance 上來看，Absences 是佔很重要的一部分。但是從分配圖來看，多數數學成績極好的缺席次數也很多。也就是說缺席在一定範圍內和成績是正相關的。
3. 不同模型的特徵選取差異有點大，想嘗試其他選取特徵的方法或者在相同的特徵下模型的預測結果特徵或找到 outlier。
4. 經過查詢後發現，未來可以使用非線性的機器學習，像是 NN 和 SVM，他們對於不相關的資料更敏感

(5) 額外嘗試 - Catboost 和 Error Analysis

後來根據老師的建議做了 Catboost 和這個模型下的 Error Analysis。因為時間的關係 Error Analysis 沒來得及對之前的模型做過一邊。

CatBoost 模型：

新增老師建議的模型 CatBoost(Categorical and Boosting)。使用套件中的 Pool function 作爲訓練集合(這 function 有利於模型的後續分析)。提取相對重要的特徵(設定特徵值大於 3)後，再次訓練模型並觀察其結果。其結果與未刪減前相同，故認爲這些特徵都是重要特徵。然而，相較於其他模型，其預測結果並沒有更好。

基於 Catboost 下的 Error analysis：

透過將 CatBoost 的預測結果分類，發現無論在葡萄牙語或是數學中(數學是 schoolsup, 葡萄牙語是 failures)，其最重要的特徵都出現特徵不平衡的現象(某一類特徵特別多，其他特別少)。因此，對特徵較少的結果進行不重複隨機抽樣，並將擴充後的資料丟進 CatBoost 模型裡進行預測。其預測結果大幅優化，成爲宗和預測結果最好的模型。

數學

	accuracy	precision	recall
RandomForest	0.50	0.65	0.61
GradientBoosting	0.56	0.62	0.62
XGBoosting	0.52	0.66	0.62
CatBoost	0.54	0.67	0.64

葡萄牙語

	accuracy	precision	recall
RandomForest	0.82	0.78	0.72
GradientBoosting	0.77	0.70	0.63
XGBoosting	0.83	0.79	0.75
CatBoost	0.83	0.79	0.74

(6) 結論

於已知歷史成績的前提下可以準確預測學生的下次成績。但是不參考歷史成績下模型只能大致預測成績的趨勢，無法準確預測成績，也是我們意料之中的。

且經由這次的報告過程中，發現可以經由這些收集數據的方法，進而了解一些關於社會層面的問題，並進一步將其解決。

(7) 組員主要貢獻

劉臣洋：數據前處理，調試模型
李國成：資料探勘
陳昱丞：號召、分配、整合，整理資料及報告
其餘資料洞見及方向皆由每周開會共同討論