Slide 1:

討論有關中學生的數學成績以葡萄牙高中生爲例

Slide 2:

本簡報的內容包括以下幾個部分：

1. 研究介紹
2. 資料觀察
3. 降維處理
4. 回歸分析
5. 結論
6. 參考資料

調查人數： 數學：395， 語文（葡萄牙語）：649

總數： 788 學生

Slide 3 :

我們的資料集（Math-student performance Data）取自 Kaggle 平台，由 Shamim (2025) 所提供。

G1、 G2、 G3 、absences 來自學校提供的資料

G1、G2、G3 分別代表不同階段的數學成績，其中 G3 為我們的分析目標。

其餘變數為問卷調查結果，類型以類別變數（categorical）為主。

Slide 4:

本研究的目的是探討哪些變數對學生的最終數學成績（G3）有影響。

Slide 5:

我們將變數依性質分為幾個群組，如下:

Groups	Variables
support	schoolsup, famsup, paid
family	address, famsize, Pstatus, guardian, traveltime, famrel
parents	Medu, Fedu, Mjob, Fjob
performance	failures, studytime, absences
alcohol	Dalc, Walc, health
after_class	activities, freetime, goout
school_choice	reason, nursery, higher
score	G1, G2, G3

Slide 5:

未歸類的變數:

Sex, age, internet(網絡狀態), romantic(是否談戀愛)

Slide 6:
本資料收集自兩所中學分別是:
GP（Gabriel Pereira）
MS （Mousinho da Silveria）
由於來自 MS 的學生數量遠少於 GP 所以我們决定刪除來自 MS 的資料只專注分析來自 GP 的資料

Slide 7:

以下是學生年齡的直方圖，幫助我們了解整體樣本的年齡分布情形。
我們可以發現到多數學生年齡分佈在 15-18 歲（普通中學生），少數為 19-22 歲（可能是留級生）

Slide 8:
由於問卷調查的每周讀書時間時數是是以區段的方式評分 所以我們取每個區段的最少時間來計算(如 2-5 小時 取 5 小時)
這張圖展示不同年齡層學生的最少學習時間。
15-18 歲的學生最少學習時間較長，可能是爲了要升學做準備
至於留級生最少學習時間趨近於零

Slide 9:
這裡比較了不同年齡層的學生在外出（goout）與空閒時間（freetime）上的分布。例如，16–17 歲學生的 goout 與 freetime 時間大致相當，而 19–21 歲的學生則顯示 freetime 顯著多於 goout。
而 20 22 的學生的 gout freretime 有明顯比較多

Slide 10:
此圖顯示學生在不同年齡層的飲酒及健康的平均分數，由於這 3 組變數都是同樣的 5 分作標準 所以我們我們將 Dalc X 4/7 Walc X 3/7(週五六日)算出一組新的分數來與 health 做對比 我可以看到 15-20 歲學生不怎麽飲酒，但 22 歲的學生有酗酒的現象。(葡萄牙法定飲酒年齡：18 歲)
此外能在圖表中看到兩者也成反比的情形

Slide 11:
此圖左側可以看到我們找出了父母的職業與教育程度前 10 多的組合， 並將

每個組合的平均 G3 算出來放在右側 我們可以看到 當父母都是教師且教育程度都為 4 時 平均 G3 是前 10 組合中最高的 但在教育程度與母親工作相同的條件下父親工作改成服務業時的平均 G3 有所下降

Slide 12:

在這張圖我們選取出了前 10 高的平均 G3 的父母職業與教育程度組合 但這裡大部分的組合的樣本都很少

Slide 13

我們利用多維尺度縮減（Multidimensional Scaling, MDS）來觀察變數間的相關性結構。

每個群組以熱力圖（heatmap）的方式展示內部變數間的皮爾森相關係數（correlation coefficient）。

1. 父母教育程度（Medu, Fedu）有中等正相關（0.63），父親與母親職業間也呈現些微相關。

failures 與 studytime 呈現負相關（-0.17），顯示學習時間越多，不及格次數可能越少。

Slide 14:

黑線（Explained Variance）：單一主成分所解釋的變異比例。可看出越往後的主成分，解釋變異量越少。

- **藍線（Remaining Explained Variance）**：剩餘未被解釋的變異量。

- **紅線（Cumulative Explained Variance）**：累積解釋變異，顯示前幾個主成分累積能夠解釋多少總變異。

如圖所示，我們可以看到主成分 PC1 至 PC8 解釋了大部分的變異量，而 PC9 之後的貢獻迅速下降。因此，在後續建模時，我們選擇保留前幾個主成分，以兼顧準確性與簡潔性。本研究也特別比較了「保留全部主成分」與「只保留 80%累積變異」這兩種策略對回歸模型的影響

Slide 15:

我們比較了以下幾種多變量線性回歸模型：

1. 原始模型：使用所有變數
2. MLR 模型：先以 MDS 降維再進行回歸
3. PCA 模型：以 PCA 降維再回歸

4. PCA 80% 模型：僅使用可解釋 80% 變異的主成分

Slide 16:
介紹 Q-Q plot

Slide 17:

我們使用 bidirectional stepwise 之方法挑選重要變數:

Stepwise forward: 變數的選擇慢慢纍加並計算最佳解

Backward: 變數的選擇從所有變數遞減并計算組合最佳解

Bidirectional: 結合兩者之方法

Method	Selected Features
Original (Reduced) + MLR	failures, goout, Mjob_at_home, romantic, schoolsup
PCA (Reduced) + MLR	after_class_dim2, romantic, parents_dim2

Slide 18:
PC 的重要變數
PC1: parents_dim2
PC2: romantic
PC3: parents_dim2

Slide 19:
我們把 0-5 15-20 的低分高分分別用不同顏色標記出來 我們看到右邊這張圖可以看到在 PC1 上 高分的點大多都比較偏向右邊 低分的大多都比較偏左邊

Slide 20:
這張表比較了多個模型的效能指標，包括 R^2、Adj. R^2、MSE、P-value 和 AIC。

我們可以看到，**Original + MLR 模型**在 R^2 與 Adj. R^2 均達到最高值，代表解釋力最好，但它的 AIC 相對偏高，顯示模型複雜度大。

相對而言，**Original (Reduced) + MLR** 模型的 AIC 最低（1995.23），代表它的模型簡潔且擬合效果良好；而且 P-value = 0.0375，顯示統計顯著。

此外，**Paper Proposed + MLR** 模型同樣有不錯的 R^2（0.19）和 MSE（176.54），兼具解釋力與預測穩定性。

因此，在平衡模型解釋力與複雜度時，Original (Reduced) 和 Paper Proposed 這兩個模型都是值得考慮的選項。

Slide 21:

在這張表中，我們比較了各模型中最重要的三個特徵。

例如，在 **Original + MLR** 模型中，higher（是否想上大學）是影響力最大的變數；Fjob_teacher 與 failures 也排在前列。

在我們提出的 **Paper Proposed + MLR** 模型中，higher 依然是最重要的特徵，failures 與 Mjob_health 也是主要影響因子。

這些資訊有助於我們了解，哪些特徵在各種資料處理方法下，對最終數學成績 G3 的影響力最顯著。

Mathematics Performance in Secondary Education

Applied Multivariate Statistical Analysis

National Dong Hwa University, Spring 2025

Yao-Chih Hsu Xuan-Chun Wang Wen-Lee Sin

變數	說明	資料類型	範圍
schoolsup	額外教育支持	二元型	yes; no
famsup	家庭教育支持	二元型	yes; no
paid	額外付費課程	二元型	yes; no
address	學生家庭住址類型	二元型	U - 城市; R - 鄉村
famsize	家庭規模	二元型	LE3 - 小於等於 3; GT3 - 大於 3
Pstatus	父母的同居狀態	二元型	T - 同居; A - 分開
guardian	監護人	名目型	mother - 母親; father - 父親; other - 其他
traveltime	通勤時間	數值型	1 - < 15 min; 2 - 15 to 30 min; 3 - 30 to 60 min; 4 - > 60 min
famrel	家庭關係品質	數值型	1（非常差） 5（非常好）
Medu	母親教育程度	數值型	0 - 無; 1 - 4 年級; 2 - 5 to 9 年級; 3 - 中學教育; 4 - 高等教育
Fedu	父親教育程度	數值型	0 - 無, 1 - 4 年級; 2 - 5 to 9 年級; 3 - 中學教育; 4 - 高等教育
Mjob	母親的工作	名目型	teacher - 教師; health - 醫療相關; services - 公務員; at_home - 在家工作; other - 其他
Fjob	父親的工作	名目型	teacher - 教師; health - 醫療相關; services - 公務員; at_home - 在家工作; other - 其他

failures	課程失敗次數	數值型	0 to 3; 4 表示超過 3 次
studytime	每週學習時間	數值型	1 - < 2 小時, 2 - 2 to 5 小時; 3 - 5 to 10 小時; 4 - > 10 小時
absences	缺席次數	數值型	0 to 93
Dalc	平日酒精消費量	數值型	1（非常少）to 5（非常多）
Walc	週末酒精消費量	數值型	1（非常少）to 5（非常多）
health	健康狀況	數值型	1（非常差）to 5（非常好）
activities	課外活動	二元型	yes; no
freetime	放學後空閒時間	數值型	1（非常少）to 5（非常多）
goout	與朋友外出頻率	數值型	1（非常少）to 5（非常多）
reason	選擇學校的原因	名目型	home - 離家近; reputation - 學校聲望; course - 課程偏好; other - 其他
nursery	是否上過幼兒園	二元型	yes; no
higher	是否想上大學	二元型	yes; no
school	學生學校	二元型	GP - Gabriel Pereira; MS - Mousinho da Silveira
sex	學生的性別	二元型	F - 女性; M - 男性
age	學生年齡	數值型	15 to 22
internet	家中是否有網路	二元型	yes; no
romantic	是否有戀愛關係	二元型	yes; no
G1	第一階段成績	數值型	0 to 20
G2	第二階段成績	數值型	0 to 20
G3	最終成績	數值型	0 to 20