

2024 教育大數據分析競賽

初賽提案構想書

團隊名稱：什麼啊！不揪！

議題：Beacon

By Education Data Analysis, we could observe students in need.

一、提案議題

在教育現場往往都要等到期中期末考才能得到學生的一次學習成效，因此我們藉由學生在平台上的練習成果與操作紀錄，經過資料分析、篩選、量化表現成果，並進一步將三個平台的量化結果經由模型訓練，設計出預警模型以量化學生目前的學習狀況，另外也設立預警門檻，若是低於此標準則標記此學生以讓老師及早給予協助，讓最需要幫助的學生得以即時、準確使用教育資源。

二、動機與目的

老師藉由平日與學生互動及期考成績可以在心中推斷出學生的表現成果，然而現實層面上並不是所有學生都與老師頻繁互動，且期考成績雖可以量化學習表現卻也為時已晚，學生在此範圍的學習狀況木已成舟，且僅能得到檢核而不能及時解決，另外，老師若檢視平台資料，可能會迷失在幾千幾萬筆的表格欄位，而難以快速、方便的鎖定最容易影響學生期考成績的項目，更是難以判斷究竟是題目缺乏鑑別度還是學生的知識吸收不良，因此我們希望藉由程序化的處理學生在學習平台上的使用資料，以設計出預警系統，方便老師得以在期考前即可以隨時標記出需要幫助的學生，並提供相應的協助。

三、數據準備與處理

各平台數據處理目的：

明確判斷測驗答對率與真實段考分數的相關性，確立預警特徵的依據。避免因測驗題目設計缺陷而誤導預警模型，確保預警特徵的準確性。提供學生綜合表現分數，為教師或學習平台提供實質的參考，提升學習輔助效果。

(一) dp001_prac

1. 數據初步檢視與預處理

(1) 篩選科目

我們首先對 dp001_prac 數據集進行了初步檢視。在數據處理中，為了有可以對應到 user_data 的段考成績，篩選出了科目包含「國文」、「英文」和「數學」，而其他科目（「自然」）則被排除，作為未來處理的展望。

(2) 篩除過少人作答的題目

篩除了僅有一名學生作答的題目，以確保題目具有代表性並提高分析的有效性。篩選結果得到了包含多位學生回答的題目，並將這些題目用於後續分析處理。

2. 預警特徵選擇：總得分/單一得分關聯性與作答時間

(1) 總得分/單一得分關聯性

以數學科為例，在篩選後的數學題目上，我們使用皮爾森相關係數計算判斷學生在平台作答「數學練習題目的總得分」與「單一數學題目得分」之間的作答相關性。此相關性被用作題目的加權參數，用於在得分矩陣中將每道題目的得分乘以總得分/單一得分關聯性權重，以提升模型對題目重要性差異的理解。

(2) 作答時間

在選擇預警特徵中，我們額外納入了作答時間來當作模型訓練的其中一個特徵，使我們更能理解學生的學習行為。

3. 數據轉換與標準化

(1) 建構得分矩陣

我們得到了按題目加權的學生得分矩陣，每一行對應一個學生，每一列對應練習題目，作為主要在後續模型構建時被加權的特徵集，有助於強化那些對學生學習成效有較大貢獻的題目。

(2) 使用 MinMaxScaler 將得分標準化

在模型建構之前，我們使用 MinMaxScaler 對學生的得分矩陣和作答時間矩陣進行縮放，將得分標準化到相同的數值範圍，以確保不同特徵在模型中具備均衡的影響力。這些經過縮放的特徵向量組合為特徵矩陣 X，最終用於訓練預測模型。

4. 模型構建與驗證

(1) 使用 Ridge 回歸模型

在這個情境中，我們選擇 Ridge 回歸作為模型，主要是為了在面對多道題目數據時，穩定各題目的影響力。Ridge 回歸在高維度和特徵具相關性的情況下，能有效控制模型中的權重，避免單個題目過度影響結果，這樣能讓每道題目根據其實際貢獻更平衡地參與成績預測。

(2) 模型訓練

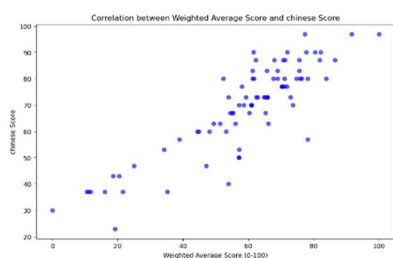
以數學段考為例，將前述特徵矩陣 X 作為輸入，而以學生的「數學段考成績」作為目標變量。模型訓練後，我們分析了模型係數，以了解每個特徵（題目）對最終成績的影響權重。接著，將訓練得到的權重應用於每位學生的得分矩陣上，計算出每位學生的「作答行為分數」，並把資料從原本的數值映射到一般考試百分比（1~100%）。

5. 不同科目實際分數與作答行為分數相關性分析（國文、英文、數學）

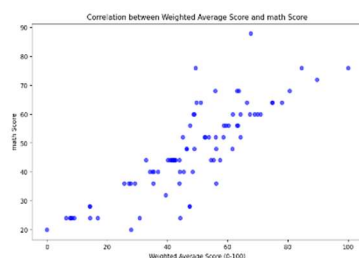
(1) 國文作答行為分數與實際國文分數的相關係數：0.877

(2) 英文作答行為分數與實際英文分數的相關係數：0.886

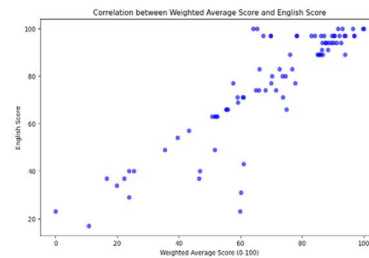
(3) 數學作答行為分數與實際數學分數的相關係數：0.845



圖一、國文作答行為分數與實際成績



圖二、數學作答行為分數與實際成績

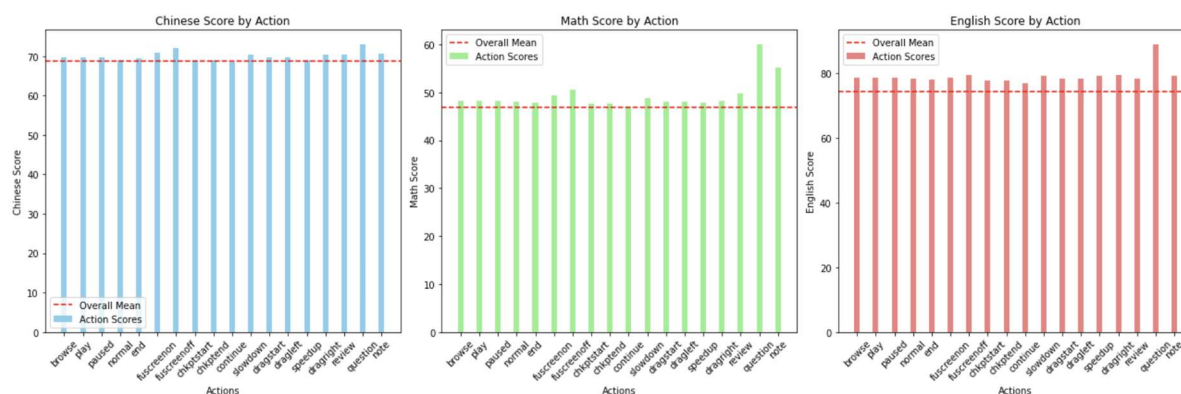


圖三、英文作答行為分數與實際成績

(二) dp001_review_plus

1. 數據初步檢視

我們對常見的影片觀看行為進行分析，並思考這些行為可能代表的意涵；接著進一步繪製圖表觀察各影片行為與各科在校成績表現的關聯，並與全體 `user_data` 中的平均成績做比較：



圖四、各影片行為與各科在校成績表現的關聯

2. 數據分析與比較

為了驗證這些行為特徵的有效性，我們將有影片觀看行為紀錄的使用者與 user_data 中所有使用者進行成績比較。結果顯示，進行 **note**、**question**、**review** 等行為的使用者，其在校成績（包含國文、數學、英文）普遍較高。根據此結果，我們提取了數個可能對學習成效有影響的特徵，作為下一步分析的依據。

3. 特徵提取與行為分數計算

基於數據探勘結果與過往相關研究，我們選定了以下數個主要特徵，並為每個特徵賦予初步的權重：

```
# 理論基礎：他人研究&數據探勘
action_weights = {
    'slowdown': 0.1, #行為科學
    'speedup': 0.1, #行為科學
    'review': 0.2, #自我回饋
    'dragleft': 0.05, #學習深度
    'dragright': 0.05, #學習深度
    'dragstart': 0.1, #學習深度
    'note': 0.2, #行為科學
    'question': 0.15, #行為科學
}
```

圖五、具影響力之影片行為及其權重

為進一步量化這些特徵對學習表現的影響，我們設計了一個**行為分數**，計算公式如下：

$$\text{行為分數} = \sum_{i=1}^n w_i \times x_i$$

其中，w 是行為特徵的權重，x 是行為次數，兩者皆為多維向量。

4. 模型建構與性能評估

我們選用了隨機森林模型作為主要的機器學習模型，並設計輸入特徵為：

- (1) 行為次數向量：使用者各行為的次數。
- (2) 行為分數：根據特徵權重計算出的分數。

```
# 3.2 建立模型
model = RandomForestRegressor(n_estimators=100, random_state=42)
```

模型輸出為平台上的測驗答對率（answer_rate）預測。

考量到數據量較小，我們採用 **8:2 的 train-test split**，以確保模型能在有限數據下進行有效的學習與評估。在模型訓練過程中，我們根據模型表現調整特徵權重與參數，最終獲得最佳化的模型。最後，根據模型最佳化結果，重新計算並

調整了每個行為特徵的權重，計算出 **最終行為分數**，作為後續分析預測學生學習成效的重要指標。

5. 資料處理與補值策略

- (1) user_data 中並非所有使用者都有使用平台一的行為紀錄，若直接刪除這部分資料，將大幅減少可用樣本數。
- (2) 為了保留更多數據，我們採用了 **中位數補值法**，以避免受離群值影響，並能較好地代表典型使用者的行為特徵。

(三) dp002_exam

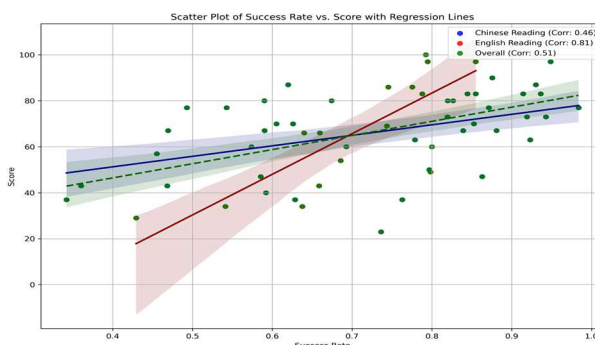
1. 散佈圖與相關係數分析

(1) 建立散佈圖：

使用散佈圖觀察各個測驗變數（如：中文閱讀測驗、英文閱讀測驗等）對真實段考分數的影響，以視覺化方式呈現兩者之間的關係。

(2) 計算相關係數：

- 計算測驗答對率與真實段考分數的皮爾森相關係數，以判斷變數之間的關聯性。
- 根據分析結果：
 - 中文閱讀測驗答對率與真實段考分數呈現**中度正相關**。
 - 英文閱讀測驗答對率與真實段考分數呈現**高度正相關**。
 - 總體答對率與真實段考分數呈現**高度正相關**。
- 結論：測驗答對率具備成為預警特徵的潛力。



圖六、答對率與段考分數散佈圖

2. 測驗題目鑑別度計算

為了確認測驗題目是否具備辨別學生學習成效的能力，我們進一步分析題目的鑑別度：

(1) 篩選題目：

- (a) 篩選出被作答次數大於 23 次的題目，以剔除無代表性且作答人數過少的題目（依據台灣平均班級人數 23 人作為基準）。

(2) 分組學生：

- (a) 將作答此題目的學生根據真實段考成績進行排序，並分為：
 - (i) 高分組 (H)：前 25% 的學生。
 - (ii) 低分組 (L)：後 25% 的學生。

(3) 計算鑑別度指數 (D)：

- (a) 計算高分組與低分組在該題目的平均答對率。
- (b) 利用公式 $D = H - L$ 計算鑑別度指數。
 - (i) 若鑑別度 D 高於 0.25，則視為具備鑑別度的題目。

(4) 篩選具備鑑別度的題目：

- (a) 剔除鑑別度低於 0.25 的題目，因為這些題目無法有效區分高分組和低分組的學生表現，可能是題目本身設計有缺陷。

3. 用戶表現分數計算與標準化

為進一步分析學生在具備鑑別度題目的表現，我們對每位用戶進行以下處理：

(1) 計算高鑑別度題目的答對率：

- (a) 計算每位用戶在高鑑別度題目上的平均答對率，作為其核心表現指標。

(2) 分數標準化：

- (a) 使用均值最大值縮放法 (Mean-Max Scaling) 將每位用戶的分數標準化至 0-100 之間這樣可以有效比較不同用戶之間的表現，避免因分數範圍不同而產生的偏差。

(3) 缺失值填補：

- (a) 若用戶未作答某些測驗題目，則以該題目的中位數作為填補，避免因極端值而影響用戶表現的估計。
- (b) 使用中位數的原因是能夠降低答對率估計值的偏誤，提升分析的穩健性。

(4) 計算最終表現分數：

- (a) 經過標準化和缺失值填補後，所得分數即為用戶在此平台上的綜合表現分數，可作為後續預警分析的依據。

(四) dp003_word

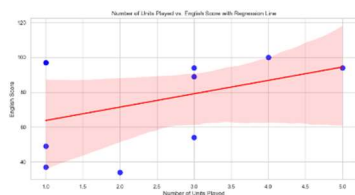
1. 資料初步檢視

在進行分析之前，首先對數據集進行基本檢視。根據 dp003_word 檔案的資料，共有 12 名學生使用平台三的英文單字學習，其中 10 名來自同一所學校。因此，為排除學校差異對結果的影響，將分析範圍集中於這 10 名學生。

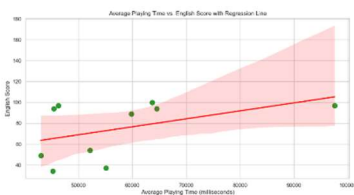
2. 各因素相關性分析

為了找出哪些因素與學生的英語成績 (english_score) 具有顯著相關性，我們進行了多項不同因素與英語成績之相關係數分析，包括學生的學習活動參與度、學習時間、準確率等，並找出最能解釋英語成績變化的因素。最終，鎖定了三個關鍵因素作為後續模型的核心參數：

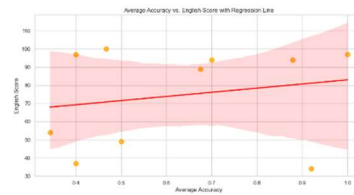
- (1) **num_units_played**：學生在學習過程中參與的單元數量，與英文成績之相關係數為 0.40。
- (2) **avg_playing_time**：學生的平均學習時間，與英文成績之相關係數為 0.45。
- (3) **average_accuracy**：學生每單元的平均正確率，與英文成績之相關係數為 0.20。



圖七、學生參與單元數量與英文成績關聯



圖八、平均學習時間與英文成績關聯



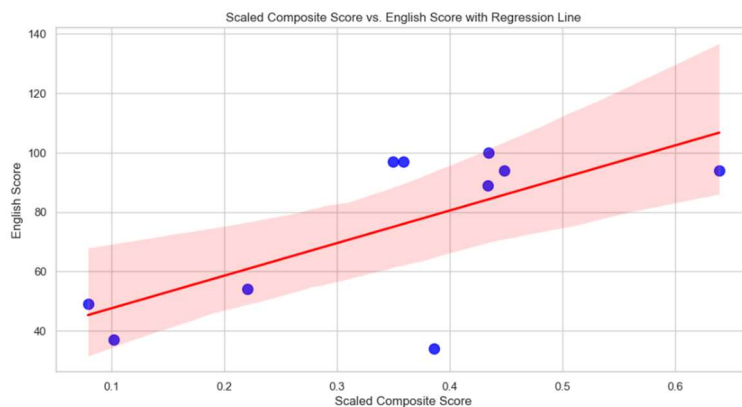
圖九、平均正確率與英文成績關聯

3. 綜合分數計算及權重調整

為了將多個影響因子融合成一個綜合的學習效率分數，給定每個影響因子不同的權重。在參考各個因子的相關係數後，先定下原始權重，再對這些權重進行調整與測試，以得到最能有效預測學生學習成效的權重分配。

4. 預測學習效益

在多種預測方法的比較中，**MinMaxScaler** 顯示出最佳的預測性能。基於 MinMaxScaler 和選定的三個影響因素，構立學習效率分數模型，並為有使用此平台的用戶，計算出一個綜合的學習成效分數，以用來衡量每位學生的學習成效。得到學習成效分數後，再與學生真實英文成績進行迴歸分析，確認其具有相關性。



圖十、綜合分數與實際英文成績關聯

5. 預測普適化

為了避免沒有使用該平台的學生對此組學習效益預測造成缺失值影響，給定他們該平台使用者學習效益分數之中位數，以利後續分析模型訓練。

(五) dp003_math

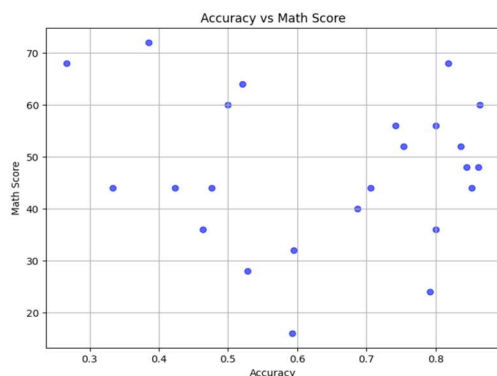
1. 數據檢視與預處理

(1) 數據檢視

- (a) 資料數量：7835
- (b) 使用學生數量：24
- (c) 單元數量：51
- (d) 題目所屬年級數量：6

(2) 有些學生在回答題目時的回答時間小於 1 秒鐘，推測為學生利用猜測方法作答，未真正看清題目並計算答案，因此刪除作答時間小於 1 秒鐘的資料。

(3) 繪製每位用戶的準確度與實際分數的關係圖（相關係數 -0.08）



圖十一、準確率與實際數學成績關聯

2. 各因素相關性分析

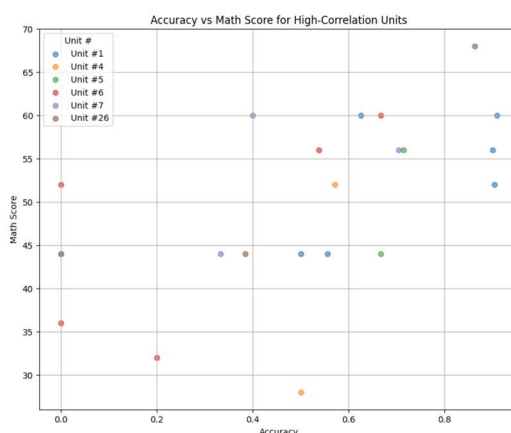
(1) 分析各題目表現

- (a) 個別題目作答人數多為一、二人，由於作答人數過少，較難看出個別題目之特徵。

- (2) 分析不同單元題目的作答情況與實際成績之關聯
 - (a) 將單元名稱分別對應 1-51
 - (b) 計算 1-51 單元的作答準確度與實際分數之關係
 - (c) 相關係數較高之單元
- (3) 分析不同年級題目的作答情況與實際成績之關聯
 - (a) 計算 1-6 年級的題目之作答準確度與實際分數之關聯
 - (b) 與實際成績關聯小

3. 綜合分數計算

篩選出與實際分數之相關係數高於 0.5 的單元之題目作答資料，並繪製準確度與實際分數之關係（相關係數：0.57）



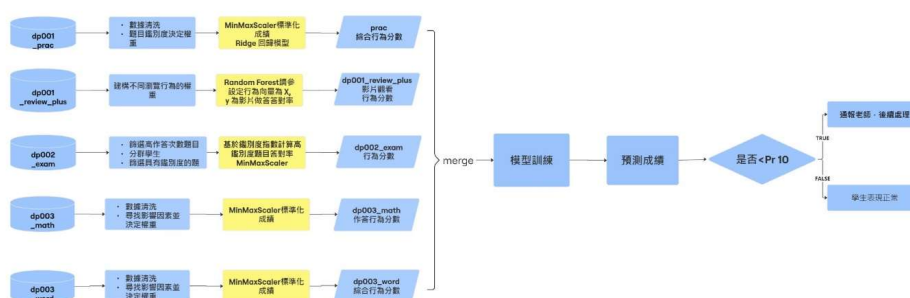
圖十二、高相關性之單元準確率與實際數學成績關聯

- (1) 使用篩選出的資料計算個別學生在這些資料中準確度的平均，作為評斷學生在此平台表現的分數

4. 計算學生在此平台的表現

- (1) 分數標準化：使用 MinMaxScaler 將分數轉換為百分制
- (2) 對沒有資料之學生進行處理
 - (a) 計算轉為百分制後的分數之中位數
 - (b) 沒有資料之學生的分數填補此中位數

四、分析方法與流程



(一) 整合各平台表現分數

以前一部分數據前處理針對各平台上使用者行為分別給予的行為分數作為模型的輸入特徵，共有 dp001_prac 綜合行為分數、dp001_review_plus 影片觀看行為分數、dp002_exam 行為分數、dp003_math 作答行為分數、dp003_word 綜合行為分數等特徵（詳細處理過程可參考前章）。

(二) 模型選擇

我們選擇**線性回歸（Linear Regression）**作為預測模型，主要基於以下幾個理由：

1. **數據的線性關聯**：在前期的數據分析中，我們發現學習時間、單元參與度和準確率等指標與學生的成績之間呈現線性關聯，因此，線性回歸成為了一個合適的選擇。
2. **模型的可解釋性**：線性回歸能夠清晰地捕捉變數與學生成績之間的線性關聯，並提供簡潔、直觀的預測結果，有助於理解不同因素對學生表現的影響。
3. **計算效率**：相比於其他較為複雜的機器學習模型，如決策樹或神經網絡，線性回歸的訓練和預測過程更加高效，能夠在短時間內處理大規模的數據，節省計算資源和時間成本，可能對複賽中更大量的資料集起到幫助。

(三) 模型訓練

1. 資料準備

- (1) 學生在各平台之表現作為（行為分數）特徵
- (2) 預測目標為學生下次考試之成績

2. 數據分割

- (1) 70%訓練集：模型有足夠的數據來學習數據中的潛在模式
- (2) 30%測試集：考量到資料量，百分之三十約為 28 人，符合一班級之人數，方便後續資料處理

3. 模型訓練

- (1) 使用線性回歸

五、初步結果與預期效益

針對個平台的數據分析結果，與期考成績最相關的特徵可以總結如下：

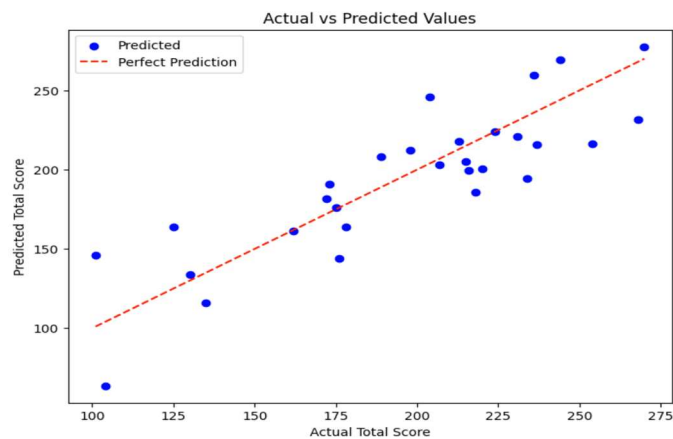
1. dp001_prac: 總得分/單一得分關聯性與作答時間
2. dp001_review_plus: 特定影片觀看行為，即 slowdown、speedup、review、dragleft、dragright、dragstart、note 及 question

3. dp002_exam: 中文閱讀測驗答對率、英文閱讀測驗答對率
4. dp003_word: 學生在學習過程中參與的單元數量、學生的平均學習時間、學生的平均作答正確率
5. dp003_math: 學生平均作答相關係數高於 0.5 的單元之正確率

(一) 模型結果

1. 預測結果與實際結果比較

R-squared : 0.7064305140401216



圖十四、test 資料集實際在校總成績與模型預測在校成績之關聯

2. 鎖定欲預警學生

- (1) 經過模型預測，根據預測之學生下次成績選取分數在後百分之十的學生
- (2) 針對預測分數在後百分之十的學生進行下一步的預警

user_sn	Actual	Predicted	alert
232372	135	115.936568	True
42415	178	163.666256	False
250399	176	144.090435	False
255091	213	217.649738	False
2697	218	185.542362	False
54952	198	212.350339	False
108886	175	176.063889	False

表四、欲預警學生

(二) 預期效益

本提案基於學生在三個平台上的學習表現數據，通過資料分析、量化學生的學習成效，並運用預測模型對學生的學習狀況進行監控和預警。我們預期能夠達成以下幾項效益：

1. 提高學生學習成效的即時預測

傳統的教育評估通常依賴期中期末考試結果來判斷學生的學習狀況，這樣的評估方式往往無法及時反映學生的學習過程，且一旦發現問題，學生的學習已經偏離理想軌道。透過本模型，能夠在學期中及時了解學生在平台上的學習表現，及早發現學習狀況不理想的學生。

2. 精準預警及資源配置

及早給予輔導與支持。預警系統能夠幫助教師根據學生的學習數據和表現狀況，依預警門檻標記可能在下次段考中表現最差的學生，精準識別需要更多關注與幫助的學生，讓教師能及時發現學習進度較慢或學習效果較差的學生，進而調配教學資源，實現更加高效的資源配置。

3. 優化學習平台的效果評估

本提案不僅針對學生進行預測與預警，同時也對學習平台本身的效果進行量化評估。通過分析學生在平台上的學習數據，可以了解平台的學習效果，找出平台內部哪些部分可能對學習成效較無影響，進而為平台的優化提供數據支持。這不僅有助於提升平台本身的設計與功能，還可以提高學生的學習體驗。

4. 促進教師的數據駕駛能力提升

透過此系統，教師能夠獲得更加系統化、可操作的數據，減少繁瑣的數據處理過程，提升教師在數據分析與決策中的駕馭能力。教師不再僅依賴直覺和單一的期末考試結果來判斷學生的學習狀況，而是能夠利用平台數據，基於客觀數據進行更加精準的教育干預與決策。

5. 開創具發展潛力的跨平台學習成效分析系統

本提案開發的預警系統極具發展潛力，未來可以根據學生學習數據的多樣化，逐步擴展到更多學科領域和學習平台。從初步的數學、英文學科到其他如科學、語言、藝術等領域，該系統將能夠整合來自不同學科的學習數據，預測學生在各個學科的學習進展與表現，提供學校更加全面的學生成效分析，協助教師在不同學科領域中制定更符合學生需求的個性化教學策略，從而增強學生的學科綜合能力。