從學生入學資訊與在學表現分析與改善招生標準

王顥鈞、何宜融、賴沂璘、許毓玹

研究動機

大學的錄取標準對每位高中的準考生來說，是重要的評判依歸，也是年年追逐的目標。與此同時，考試的錄取標準對學校來說，也深深引響到來年學生的優劣，是個不得不慎重的議題。甚至在入學之後，身邊的同學時常以入學成績自己在未來大學四年的表現。所以希望能透過這次比賽的機會，了解是否能以學生的入學資料，預估或判斷學生的未來表現。

研究目的

我們想藉由分析經由各種管道入學的新生入學時的成績及各項相關資料，以機器學習的方式模擬未來的在校成績，幫助學校調整入學標準，收到可能更有潛力的學生。

研究方法及過程

1. 樣本描述：

將整個資料做前處理，包括去除缺失職、資料整合與轉換之後，選取我們要研究的對象，也就是同時擁有入學資料以及在學表現的台灣籍大學部學生。合併數據後，共有四千多筆樣本(以人為單位)。

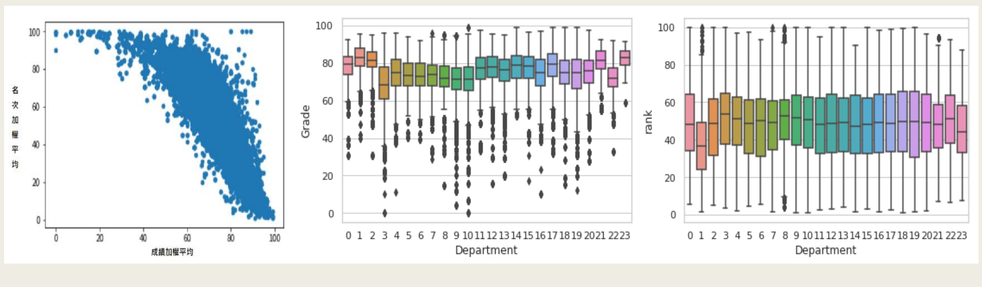
2. 樣本標籤定義：

A. 分數加權：(學生所修習的所有必修課程成績對學分數做加權)

$$\frac{\sum \text{課程成績} \times \text{課程學分數}}{\sum \text{課程學分數}}$$

B. 名次加權：(學生所修習的所有必修課程名次對學分數做加權)

$$\frac{\sum \text{課程名次} \times \text{課程學分數}}{\sum \text{課程學分數}}$$



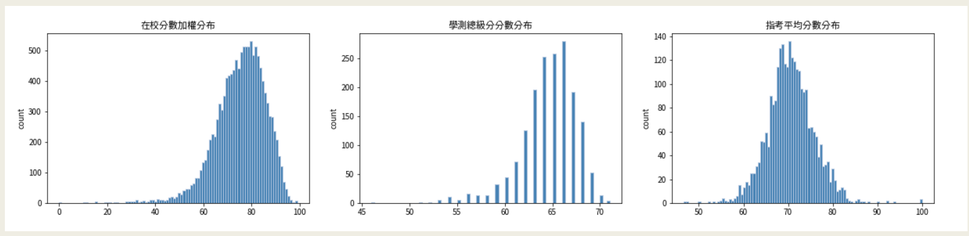
的分數加權與名次加權分數在數據中有高度的相關性，同時也分別具有不同的特性，分數加權可以看出樣本原始的資料分布，名次加權則在跨課程比較的過程中，減少因教授給分的差異而對結果造成的影響。

3. 母體樣本提取:

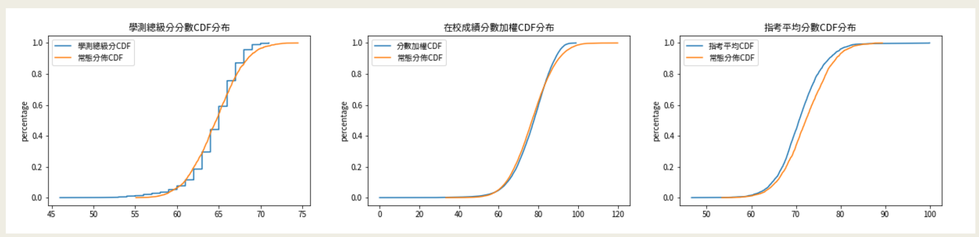
初期，為了保持樣本標籤之間的統一，希望能以系所為單位提取資料，確保學生之間必修課程的重疊性，所以以數學系為最初的母體樣本。但由於個別系所的樣本數過少(僅128筆)，我們以院為單位為提取資料，並以較多共同必修科目的資電學院及管理學院為目標對象。最後，我們選擇以名次加權標準化標籤，消弭系所之間的差異，嘗試以全校學生作為母體樣本進行資料分析研究。

4. 資料勘查：

A. 指考、學測成績與在校成績分布



B. 指考、學測成績及在校成績與常態分佈CDF比較



透過比較可視化資料處理後的樣本分布與常態分布，確認樣本空間沒有因為資料選取、整合或轉換，而大量減少特定型態的樣本，導致樣本空間的分布與原始數據出現差異，進而失去代表性。

5. 資料處理套件工具：

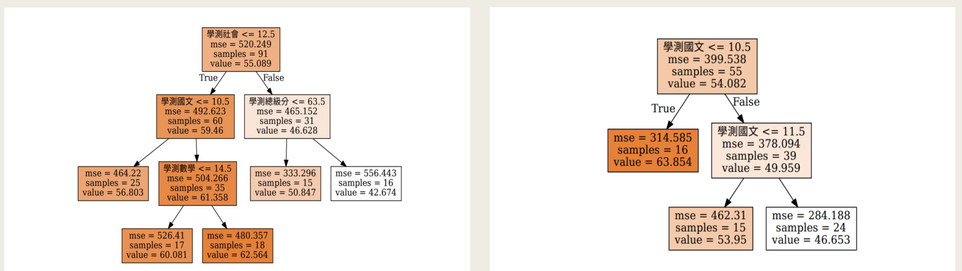
A. numpy（高維度及矩陣運算）
B. pandas（數據處理和分析的主要的函式庫）
C. matplotlib / seaborn（資料可視化工具）
D. sklearn（機器學習函式庫）

整理合併學生入學資料以及在校成績後，以資料可視化工具了解資料分布，再主要利用sklearn中的decision tree及random forest機器學習的算法，嘗試訓練出能以學生入學資料估計該學生未來在校成績的模型。

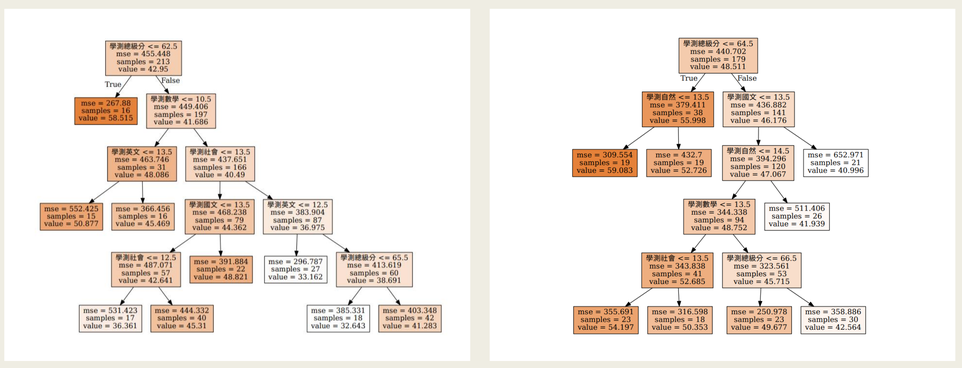
研究結論

1. Decision tree/ Random forest 研究結果：

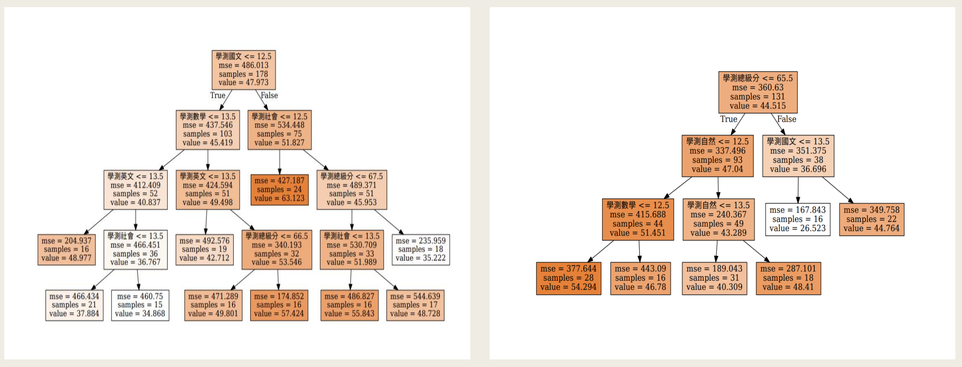
A. 數學系(Decision tree/ Random forest)：



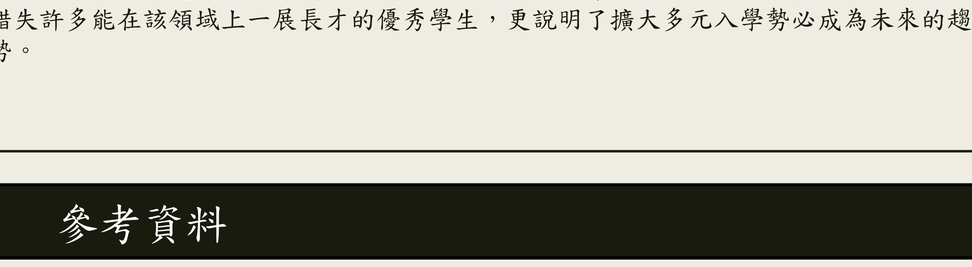
B. 資電學院(Decision tree/ Random forest) ：



C. 管理學院(Decision tree/ Random forest)：



D. 全校學生入學成績包含學測成績/指考成績(Decision tree)：

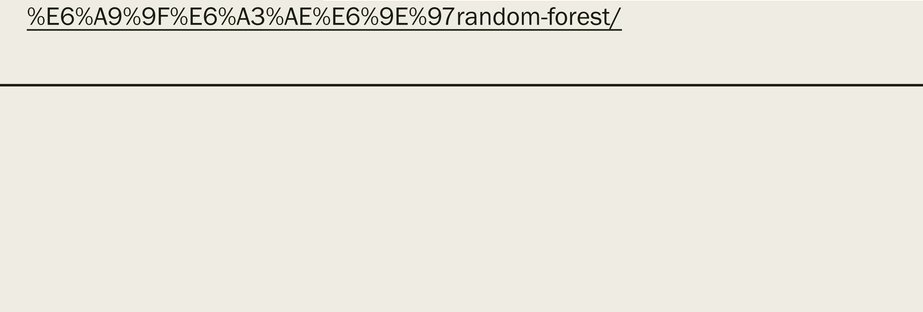


多次對數據做決策樹分類後，數據被分割的方式隨機，決策樹每一層的均方誤差值也沒有遞減的趨勢，進而發現資料的主要特徵(學測及指考成績相關)與標籤的相關性不高的問題。

2. 數據相關性驗證結果：

A. 在校成績與學測關係
（相關係數：0.11505053）

B. 在校成績與指考成績關係
（相關係數：0.20909293）



計算在校成績與學測/指考成的數據相關係數後，發現兩者皆小於0.3，可視為兩者與在校成績呈非常不相關。顯示出在過去升學方式中最主要的評判標準-學測和指考分數，與學生入學後的學業表現並無明顯相關，說明以考試成績為錄取學生的唯一標準將錯失許多能在該領域上一展長才的優秀學生，更說明了擴大多元入學勢必成為未來的趨勢。

參考資料

- Kaggle – https://www.kaggle.com/
- Decision Tree Classification – https://www.saedsayad.com/decision_tree.htm
- 隨機森林Random Forest – https://chtseng.wordpress.com/2017/02/24/%E9%9A%A8%E6%A9%9F%E6%A3%AE%E6%9E%97random-forest/