使用機器學習技術於招生錄取標準之研究分析

指導教授：陳建彰　博士
林承賢　博士
研 究 生：戎書玄

中華民國110年7月19日

淡江大學 Tamkang University

簡報大綱
Outline

PART ONE 緒　論
PART TWO 文獻探討
PART THREE 研究方法
PART FOUR 結果分析
PART FIVE 結論建議

PART ONE
緒論

PART ONE
緒論
研究背景

- **入學考試制度改革**
  - 大學自主、多元評量、多元管道的招生策略學生依照自己的性向興趣自主選擇校系
  - 學校重視學生的「才能」，將最有受高等教育
  - 潛力的學生，選擇出來，讓其進入大學就讀 2002年開始全面實施的「大學多元入學制度」

PART ONE

緒論

研究動機

- 因應大學多元入學
  - 進入12年國教後，「大學學力測驗」成為大多數學生第一個接觸到的重大考試
  - 多元入學三管道中，其中兩者：繁星推薦及個人申請均以「大學學力測驗」成績為標準
- 社會少子化嚴重
  - 如何挑選出合適就讀的學生也成學校的一大課題

PART ONE

緒論

研究目的與問題

在申請入學階段為學校提供一項挑選學生之參考依據

- 研究問題

一、依據學測成績、高中PR值、高中GPA對入學後表現預測的結果

二、探討不同分類方式所獲得的預測表現

三、各特徵值與預測結果的相關性

PART TWO
文獻探討

PART TWO
文獻探討

成績預測

相關研究結果

- 教育資料探勘(Educational data mining，EDM)領域中已有許多成績相關的研究
- Shahiri等人在2015年的研究中指出累計平均積點(CGPA)是影響最大的特徵值，其次較多人使用人口統計和外部評估，最後則是行為模式
- Shahiri也在研究中統計出過往成績預測相關研究中各種方法的預測表現，其中表現最優為神經網路，其次是決策樹，接下來是支援向量機

PART TWO

文獻探討

深度神經網路

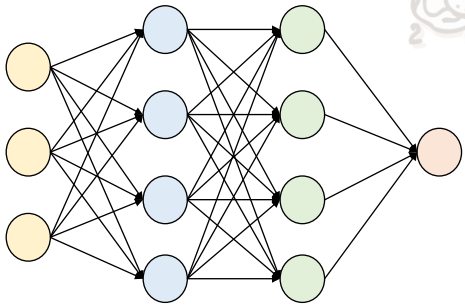
相關研究結果

- Hanan於2020年提出的研究中以三項入學成績進行學生首年的累計平均積點(CGPA)預測
- Hanan在研究中使用ANN模型，結果顯示預測出79.22%的準確率

PART TWO

文獻探討

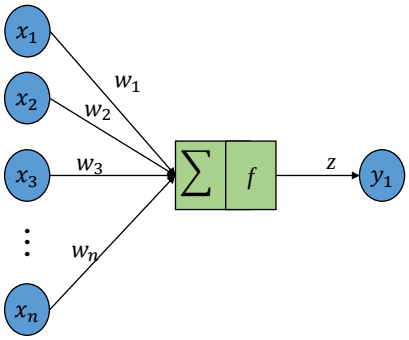
深度神經網路



PART TWO

文獻探討

深度神經網路



- 節點獲取輸入的加權和，激勵函數(Activation function)轉換成非線性，算出結果傳遞給下一層
- $ReLU(z_j) = max(0, z_j)$

  優點：解決梯度爆炸問題、計算數度快、收斂速度快
- $Softmax(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}}$

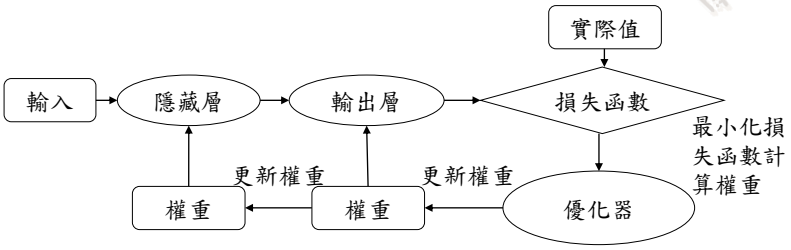
  for $j = 1, \dots, K$

  它將多個神經元的輸出，映射到(0,1)區間內如，可以看成機率來理解，從而來進行多分類

PART TWO

文獻探討

深度神經網路



PART TWO

文獻探討

深度神經網路

- 損失函數（loss function）：用來描述模型的預測值與真實值不一致的程度
- 交叉熵cross-entropy：觀測預測的機率分佈與實際機率分布的誤差範圍

$$H = \sum_{c=1}^{C}\sum_{i=1}^{n} -y_{c,i}\ln(p_{c,i})$$

C：類別數，n：所有資料數，$y_{c,i}$：第i筆資料屬於第c類真實類別，$p_{c,i}$：第i筆資料屬於第c預測出來的機率

- 優化器(optimizer)：決定權重更新方式
- 準確步梯度下降法Stochastic Gradient Decent(SGD)：最單純的梯度下降方法，利用微分的方法找出梯度，往梯度的方向去更新權重

$$W \leftarrow W - \eta\frac{\partial L}{\partial W}$$

W：權重，L：損失函數，η：學習率，$\frac{\partial L}{\partial W}$：損失函數對權重的梯度

PART TWO

文獻探討

隨機森林

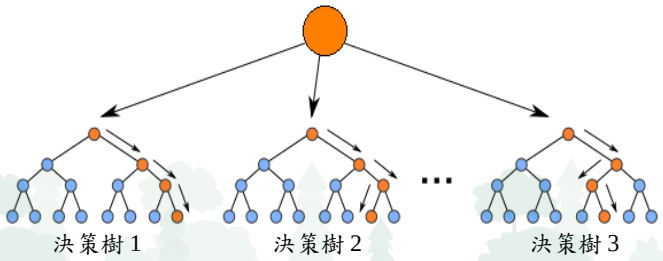
相關研究結果

- Anuradha等人於2015年提出的研究使用學生的人口統計數據及外部評估來預測學生的期末學期標記(ESM)。
- Anuradha在這項研究中使用決策樹模型，結果顯示預測出72.51%的準確率

PART TWO

文獻探討

隨機森林

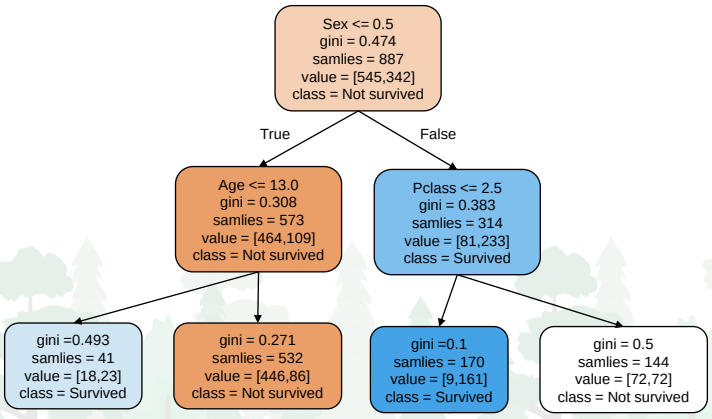


圖片來源：https://medium.com/chung-yi/ml%E5%85%A5%E9%96%80-%E5%8D%81%E4%B8%83-%E9%9A%A8%E6%A9%9F%E6%A3%AE%E6%9E%97-random-forest-6afc24871857

PART TWO

文獻探討

隨機森林



PART TWO

文獻探討

隨機森林

- 為了要在結點上使用最具意義的特徵來做分割，我們定義一個函數：在每個分割處理時，其資訊增益(information gain，IG)必須要最大

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j)$$

f：結點用來作分割的特徵
D_p、D_j：父結點、第j個子結點的數據集
I：不純度(impurity measure)
N_j、N_p：父結點、第j個子結點的樣本個數

- Gini不純度(Gini impurity，I_G)

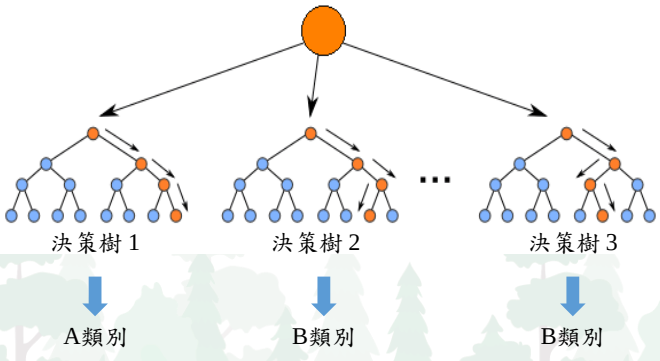
$$I_G(t) = \sum_{i=1}^{c} p(i|t)\big(1 - p(i|t)\big) = 1 - \sum_{i=1}^{c} p(i|t)^2$$

$p(i|t)$：特定結點t中「樣本屬於類別i」的比例
c：分類方式，若為二元分類則c=2

PART TWO

文獻探討

隨機森林



圖片來源：https://medium.com/chung-yi/ml%E5%85%A5%E9%96%80-%E5%8D%81%E4%B8%83-%E9%9A%A8%E6%A9%9F%E6%A3%AE%E6%9E%97-random-forest-6afc24871857

PART TWO

文獻探討

支援向量機

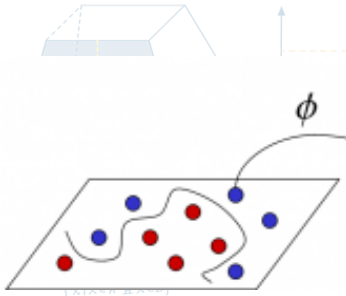
相關研究結果

- Miguéis等人於2015年提出的研究使用大學前兩學期成績與及一些人口統計數據預測學生在學位課程結束時的 AP - 在學術生涯中獲得的成績的加權平均值之間的比率。
- Miguéis在這項研究中使用支援向量機模型，結果顯示預測出92.6%的準確率

PART TWO

文獻探討

支援向量機



輸入所在空間　　特徵分布空間

圖片來源：http://bytesizebio.net/2014/02/05/support-vector-machines-explained-well/

PART TWO

文獻探討

支援向量機

- 核函式：協助將特徵映射至高維度空間。
- 多項式核函式(polynomial kernel function)：表示特徵空間中的向量與原始變量的多項式的相似性，將原不可線性分割之問題投到高維後以線性分割處理。

$$f(x) = sgn\left(\sum_{i=1}^{n} a_i y_i K\left(x_i^t \cdot x\right) + b\right)$$

$$k\left(x_i, x_j\right) = \left(x_i^t \cdot x_j\right)^d \text{，} d \geq 1$$

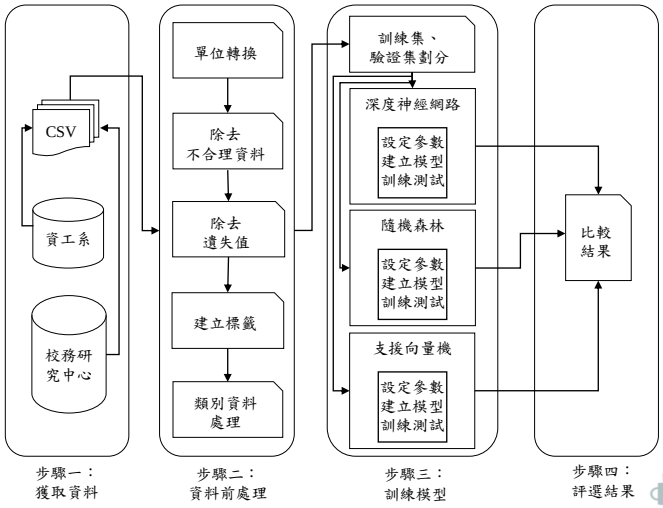
PART THREE

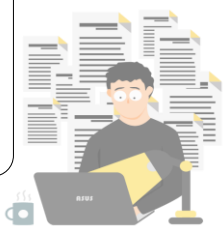
研究方法

PART THREE

研究方法

實驗架構與流程

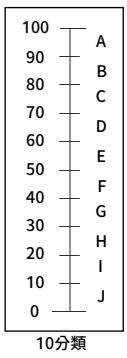




PART THREE

研究方法

資料標籤方式

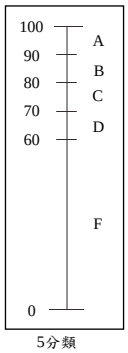


10分類
以10分為間隔區間分為A-J，10個類別

PART THREE

研究方法

資料標籤方式



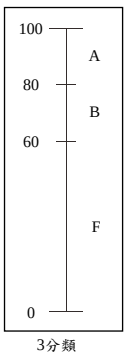
10分類
以10分為間隔區間分為A-J，10個類別

5分類
60分以上以10分為區間分為A-D，以下為F

PART THREE

研究方法

資料標籤方式



10分類
以10分為間隔區間分為A-J，10個類別

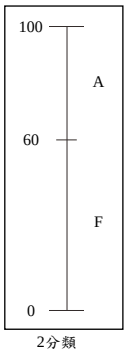
3分類
60分以上以20分為區間分為A-B，以下為F

5分類
60分以上以10分為區間分為A-D，以下為F

PART THREE

研究方法

資料標籤方式



10分類
以10分為間隔區間分為A-J，10個類別

5分類
60分以上以10分為區間分為A-D，以下為F

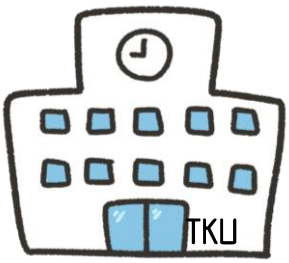
3分類
60分以上以20分為區間分為A-B，以下為F

2分類
60以上為A，以下為F

PART THREE

研究方法

資料來源



校務研究中心

資工系

PART THREE

研究方法

資料前處理

欄位名稱	欄位屬性	是否經過前處理	遺失數量
國文	特徵	○	
英文	特徵	○	
數學	特徵	○	
社會	特徵	○	
自然	特徵	○	
高中PR值	特徵	○	
高中GPA	特徵		
總平均	標籤	○	
程式科	標籤	○	1
數學科	標籤	○	18
基礎科	標籤	○	1

- 學測成績：通過累計人數百分比由遠使集分轉換為百分等級
- 高中PR值：公立學校有公開PR值資訊，私立集獨招學校則從網上搜尋而來，較不可靠
- 總平均：從原始學生修課成績中篩出系必修專業科目做平均，再依照標籤方式分類
- 科目類：依照課程地圖進行類別分類，分別平均後再依照標籤方式分類

PART THREE

研究方法

遺失值處理

欄位名稱	欄位屬性	是否經過前處理	遺失數量
國文	特徵	○	
英文	特徵	○	
數學	特徵	○	
社會	特徵	○	
自然	特徵	○	
高中PR值	特徵	○	
高中GPA	特徵		
總平均	標籤	○	
程式科	標籤	○	1
數學科	標籤	○	18
基礎科	標籤	○	1

- 因學生成績皆為真實存在之資料，為求公平，將遺失值去除

PART THREE

研究方法

模型參數設定

- 深度神經網路

參數名稱	參數值
層數	3
隱藏層節點個數	10
Dropout	0.4
優化器	SGD
評估標準	categorical_accuracy
激勵函數	隱藏層ReLU / 輸出層softmax
損失函數	categorical_crossentropy
批次大小	10
epochs	5

PART THREE

研究方法

選定參數過程

- 深度神經網路

	10分類	5分類	3分類	2分類
層數：3 節點數：10 batch_size：10 epoch：5	36.36%	36.36%	60.61%	84.85%
層數：50 節點數：10 batch_size：10 epoch：5	24.24%	36.36%	60.61%	84.85%
層數：200 節點數：10 batch_size：10 epoch：5	36.36%	24.24%	60.61%	84.85%
層數：3 節點數：200 batch_size：10 epoch：5	36.36%	24.24%	60.61%	84.85%
層數：50 節點數：50 batch_size：100 epoch：100	24.24%	24.24%	60.61%	84.85%

PART THREE

研究方法

模型參數設定

- 隨機森林

參數名稱	參數值
n_estimators	5
Criterion	Gini

選定參數過程

	10分類	5分類	3分類	2分類
n_estimators：1	9.09%	30.30%	54.55%	81.82%
n_estimators：5	27.27%	27.27%	72.73%	87.88%
n_estimators：10	27.27%	21.21%	57.58%	69.70%
n_estimators：100	6.06%	18.18%	69.70%	87.88%
n_estimators：200	6.06%	18.18%	75.76%	87.88%

PART THREE

研究方法

模型參數設定

- 支援向量機

參數名稱	參數值
kernel	poly
decision_function_shape	ovr
C	1

PART THREE

研究方法

選定參數過程

- 支援向量機

	10分類	5分類	3分類	2分類
kernel：poly shape：ovr degree：3 C：1	0.00%	6.06%	63.64%	90.91%
kernel：poly shape：ovr degree：5 C：1	9.09%	18.18%	39.39%	84.85%
kernel：poly shape：ovr degree：3 C：50	3.03%	15.15%	39.39%	84.85%
kernel：poly shape：ovr degree：3 C：100	6.06%	18.18%	39.39%	84.85%
kernel：rbf shape：ovr C：1	0.00%	0.00%	48.48%	84.85%
kernel：sigmoid shape：ovr C：1	0.00%	0.00%	60.61%	84.85%

0.00%	9.09%	51.52%	84.85%
0.00%	0.00%	60.61%	84.85%
0.00%	0.00%	45.45%	84.85%
3.03%	12.12%	42.42%	87.88%

PART FOUR

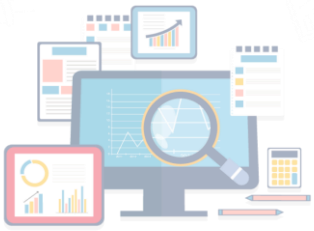
結果分析

PART FOUR

結果分析

評估指標

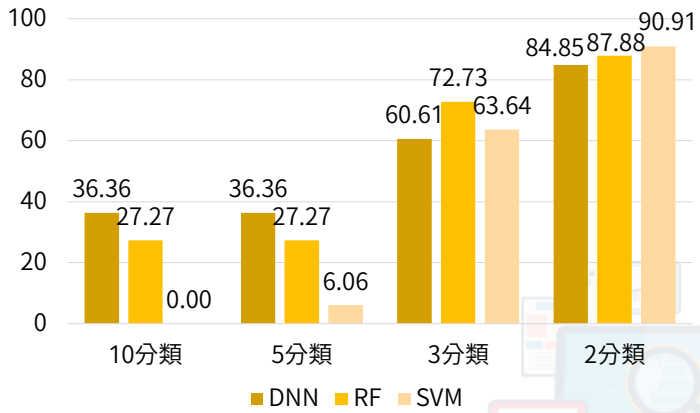
Bingo　　1-Away



PART FOUR

結果分析

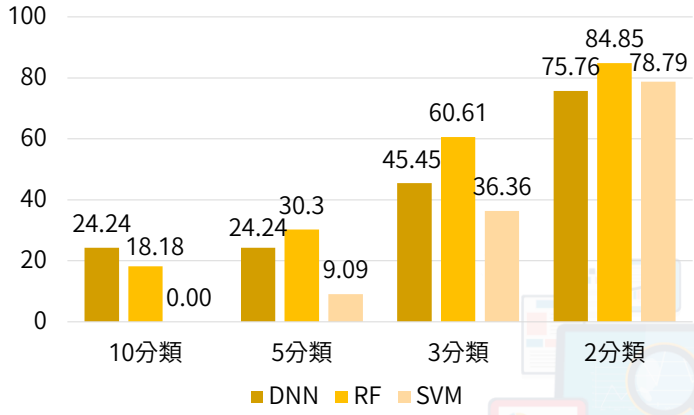
總平均完全正確率預測結果



PART FOUR

結果分析

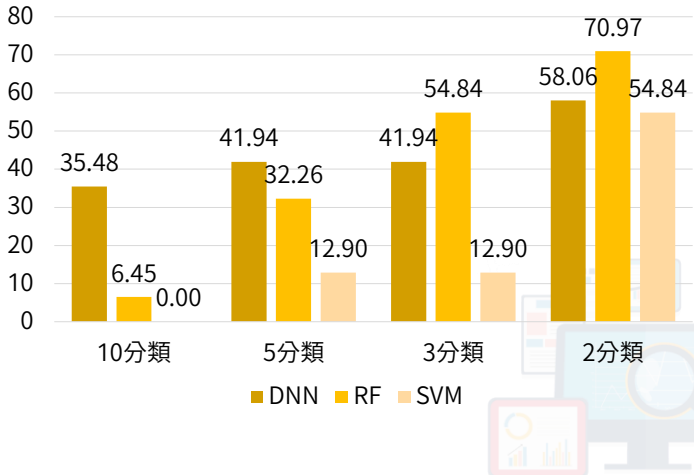
程式科完全正確率預測結果



PART FOUR

結果分析

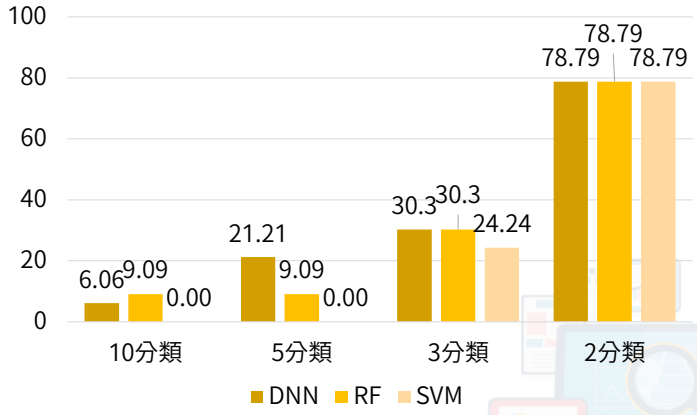
數學科完全正確率預測結果



PART FOUR

結果分析

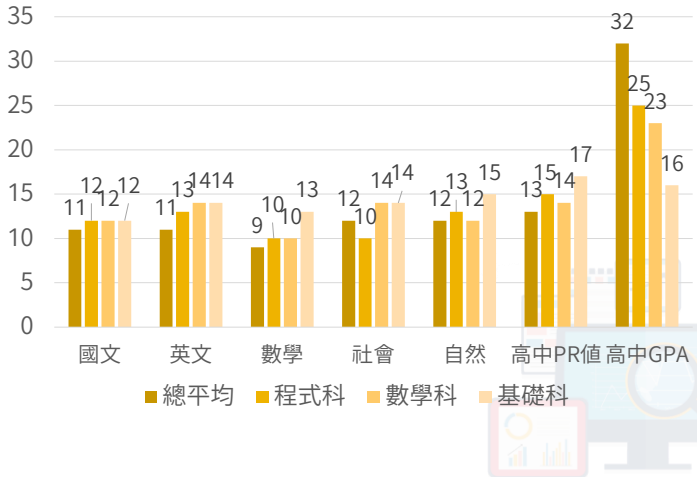
基礎科完全正確率預測結果



PART FOUR

結果分析

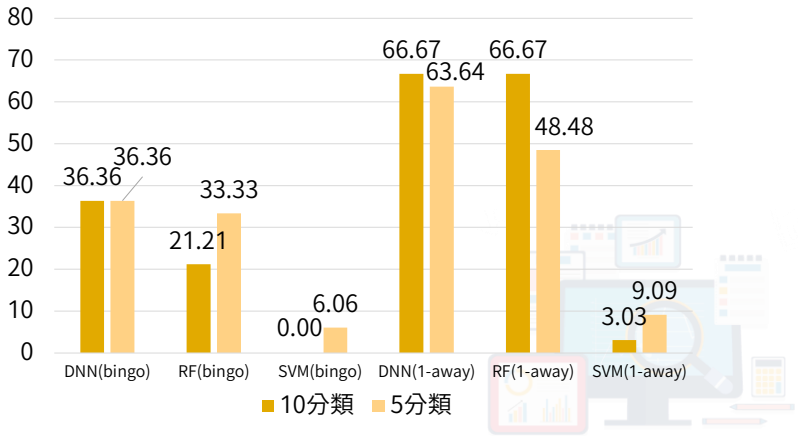
隨機森林－特徵重要性



PART FOUR

結果分析

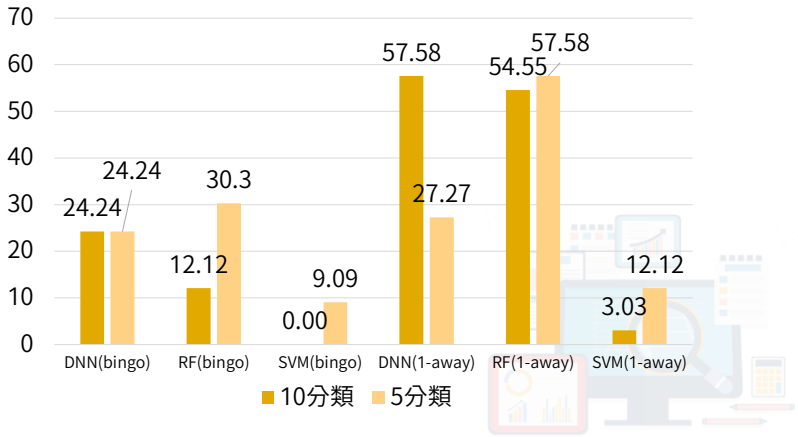
完全正確率與鄰近正確率總平均



PART FOUR

結果分析

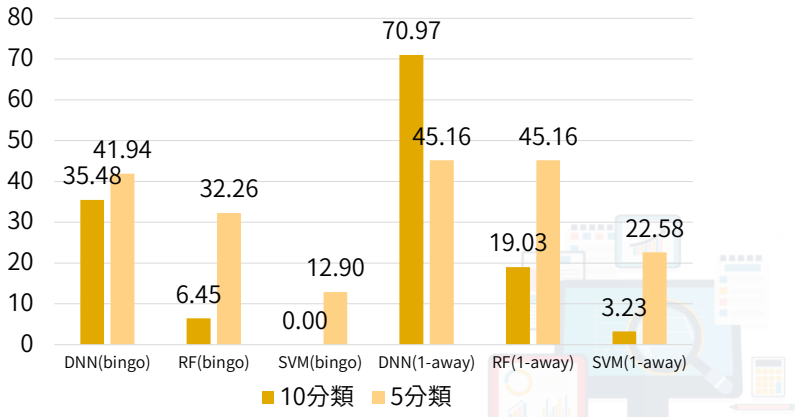
完全正確率與鄰近正確率程式科



PART FOUR

結果分析

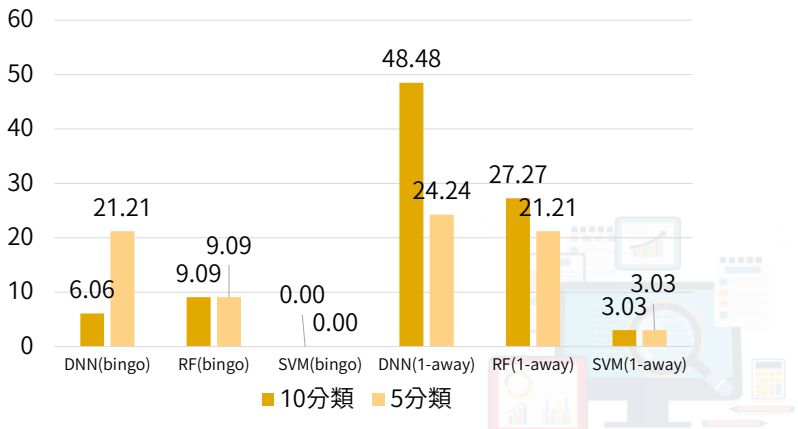
完全正確率與鄰近正確率
數學科



PART FOUR

結果分析

完全正確率與鄰近正確率
基礎科



PART FOUR

結果分析

結果統整

	10分類		5分類		3分類	2分類
	Bingo	1-Away	Bingo	1-Away	Bingo	Bingo
DNN	1	1	1	1	3	3
RF	2	1	2	2	1	2
SVM	3	3	3	3	2	1

- 高中GPA在各科目中皆表現出高度相關性
- 使用鄰近正確率後在不失去分類意義的情況下得到更高的準確率，在10分類尤其明顯

PART FIVE

結論與建議

PART FIVE

結論與建議

結論

- 由於入學前可取得的學生資料相當有限，且少部分特徵欄位不是官方釋出的數據，在資料正確性上有疑慮。爲求資料及正確性，所能匯入的學生資料有限，因此無法增加樣本數，使得訓練成果有限
- 在不同的機器學習模型中，進行不同標籤方式的比較，探討各個方式在不同模型下的預測結果，透過合理運用準確率評估標準後得出了很不錯的預測表現

PART FIVE

結論與建議

未來研究方向

- 取得更多有價值之特徵資料，如高中每學期的學科成績或是人口統計、外部評估、行爲模式等
- 嘗試使用不同租入組合以提升準確率

使用機器學習技術於
招生錄取標準之研究分析

指導教授：陳建彰　博士
研 究 生：戎書玄

中華民國110年7月20日

淡江大學 Tamkang University

使用機器學習技術於
招生錄取標準之研究分析

Q&A時間

淡江大學 Tamkang University