淡江大學資訊工程學系碩士班

碩士論文

指導教授：陳建彰博士
林承賢博士

使用機器學習技術於招生錄取標準之研究分析－以淡江大學資工系為例

Using Machine Learning Techniques to Find the Relationship between University Admission Score and Student Performance －Taking Department of Computer Science and Information Engineering of Tamkang University as an example

研究生：戎書玄撰

日期：中華民國 110 年 6 月

論文名稱：	使用機器學習技術於招生錄取標準之研究分析—以淡江大學資工系為例	頁數：51
校系(所)組別：	淡江大學資訊工程學系碩士班	
畢業時間及提別：	109 學年度第 2 學期碩士學位論文提要	
研究生：戎書玄	指導教授：陳建彰博士 林承賢博士	

論文提要內容：

隨著少子化的影響，所有的高等教育機構都面臨了招生方面的挑戰，如何挑選適合的學生進入校系就讀需要經過各項數據的支持。本研究透過深度神經網路、隨機森林、支援向量機等機器學習技術，分析新生入學資料與大一學業表現的關聯性。實驗結果顯示，在總平均科目的 2 分類預測上，支援向量機法具有最佳的預測結果；在數學與程式類科目中，2 分類的隨機森林法具有最佳的預測結果；然而在基礎類科目中，各方法在二分類有最佳結果且表現同樣優異。在其他的分類上，包括 3 分類、5 分類、及 10 分類，三種方法的預測能力各有優劣。特徵中的高中 GPA 對預測結果影響重大。

關鍵字：機器學習、學生學習表現、成績預測、深度神經網路、隨機森林、支援向量機

Title of Thesis：	Using Machine Learning Techniques to Find the Relationship between University Admission Score and Student Performance －Taking Department of Computer Science and Information Engineering of Tamkang University as an example	Total pages：51
Keyword:	Machine Learning, Student Performance, Performance Prediction, Deep Neural Networks, Random Forest, Support Vector Machine	
Name of Institute:	MASTER'S PROGRAM, DEPARTMENT OF COMPUTER SCIENCE AND INFORMATION ENGINEERING, TAMKANG UNIVERSITY	
Graduate date:	June,2021	Degree conferred: Master
Name of student:	Shu-Shiuan Rong 戎書玄	Advisor: Dr. Chieh-Chang Chen 陳建彰博士 Dr. Cheng-Shian Lin 林承賢博士

Abstract:

With the impact of the declining birthrate, all higher education institutions are facing enrollment challenges. Collecting useful data can find appropriate students for university admission. In this study, we adopt machine learning techniques like Deep Neural Network (DNN), Random Forest (RF), and Support Vector Machine (SVM) to find the relationship between entrance score and freshman academic performance. Experimental results show that the SVM has the best prediction results on average of total courses in 2-category classification. The RF performs best on mathematical courses and programming courses in 2-category classification. However, on fundamental courses, all methods have the best prediction results in 2-category and perform equally well. Moreover, on other category classifications, like 3-category, 5-category, or 10-category classifications, we cannot find a universal best method. Another result shows that High school GPA has a significant impact on results.

目錄

圖目錄

表目錄

第一章　緒論

1.1　研究背景

大學入學制度是決定學生進入大學就讀的方式，也是學校用來挑選合適就讀之學生的一種方式。台灣早年採用聯考制度，在當時深獲社會所肯定，原因在於其具有公平、公正、考科固定、容易準備等特性。但傳統的聯考制度也有其弊病，例如考試制度下的教學使得教育目標遭到扭曲，狹隘的培養考試機器、大學招生無法自主，不利營造多元學習環境、考生依錄取分數做為選題填校系的依據而忽略了大學的辦學績效及自己的性向興趣。

關於大學入學考試制度，各國的情形不同所採用的方式也不同，不過目標卻相同，就是重視學生的「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，然後由大學來培養造就他，使他成為知識份子[1]。

我國的大學聯招制度與當時其他先進國家有著明顯的差異，當時其他先進國家，如美國、英國、日本等，均採取大學自主、多元評量、多元管道的招生策略，此種強調大學自主性、學生自主選擇校系、評量方式多元的招生制度與臺灣教育界長期以來所期待的改革方向不謀而合，對當時推動招生政策改革的學者有著深遠的影響[2]。在此背景下，促成了 1994 年至 2001 年進行試辦，自 2002 年開始全面實施的「大學多元入學制度」。大學多元入學主要有三種管道，茲說明如下：

一、繁星推薦入學

「大學繁星推薦入學」招生管道係延續繁星計畫「高中均質、區域均衡」之理念，由高中向大學校系推薦符合資格的學生，提供各地區學生適性揚才之均等機會，並引導學生就近入學高中。

二、個人申請入學

「大學個人申請入學」招生管道則兼具學生可依個人志趣選擇大學校系及大學校系依其特色適性選才之目的。

三、考試入學

「大學考試入學」招生管道打破傳統類組概念，提供校系彈性自主的招生空間，考生多元選擇及適性發展的機會。

1.2 研究動機

台灣經過一系列的教育改革之後，現已進入 12 年國教，對台灣的學生來說「大學學力測驗」成為求學過程中第一個接觸到的重大考試。而隨著少子化的衝擊與政策改變，「繁星推甄」及「個人申請」所開出的錄取名額逐年增高，許多學生選擇以此兩種方式作為入學的管道，而這兩種方式均是以學科能力測驗為篩選依據，所以學科能力測驗成績就顯得格外重要。學生在進入高中時，經過了國中基本學力測驗或是如今的會考，即便現在的制度漸漸在消除明星高中的迷思但多多少少還是會有些學習環境上的差異，因此畢業學校也可能成為影響大學表現的要素之一。

1.3 研究目的

本研究擬以機器學習之多種方法，建立預測大學入學後表現之模型並進行不同預測建模之比較，以期能找出最適合之預測模型。目的列舉如下：

1. 輔助系上在申請入學階段有更多挑選學生的參考依據。
2. 學生可以通過此預測了解自己是否合適就讀。

1.4 研究問題

本研究將資料集以不同標籤方式作為模型訓練的輸出，應用至不同模型。依據研究目的，列出下列問題作為本研究之研究問題：

1. 依據學測成績、高中 PR 值、高中 GPA 對入學後表現預測的結果。
2. 探討不同分類方式所獲得的預測表現。
3. 各特徵欄位與預測結果的關聯性。

1.5 研究架構

本研究共分為五個章節，後續各章之簡要說明如下：

1. 第二章為文獻探討，主要在探討深度神經網路、隨機森林及支援向量機的相關研究與參考文獻。
2. 第三章主要介紹實驗架構、模型架構、實驗流程與資料處理。
3. 第四章為實驗結果，比較本研究資料集在不同模型上的預測結果。
4. 第五章總結研究的結論，探討研究限制以及未來研究的發展。

第二章　文獻探討

本章整理過往有關於成績預測的重要相關研究，其不同方向的論文有不同的結果預測方式及關注方向。隨著科技日新月異，學者們紛紛採用更有效的預測方法例如 Educational Data Mining(EDM)[3]，這在成績預測相關研究是項非常熱門的方法。EDM 是從巨大的教育資料庫提取有用的信息和模式來更有效預測學生學習表現。在 Shahiri 等人的 2015 年提出的研究[9]中指出累計平均績點(CGPA)是影響最大的特徵欄位，其次較多人使用人口統計(例如：性別、年齡、家庭背景、是否殘疾)和外部評估(例如：特並科目期末考試成績)，最後則是行為模式(例如：課外活動、社交網路)。Shahiri 也在研究中統計了過往預測研究中各種預測方法的準確率，其中準確率最高的為神經網路，其次是決策樹，接下來是支援向量機。

2.1　深度神經網路

深度學習是機器學習的其中一分支，早在 2006 年由 Hinton 等人[10]提出，概念是透過疊加多層的隱藏層，用以模擬大腦進行學習之神經網路。在很多已發表的研究[11]中，都有使用到此方法作為預測學生表現的方法。Hanan 於 2020 年提出的研究中[16]以三項入學成績進行學生首年的累計平均績點(CGPA)預測，其研究中使用 Artificial Neural Networks(ANN)模型，結果顯示預測出 79.22%的準確率。

2.1.1 基本架構

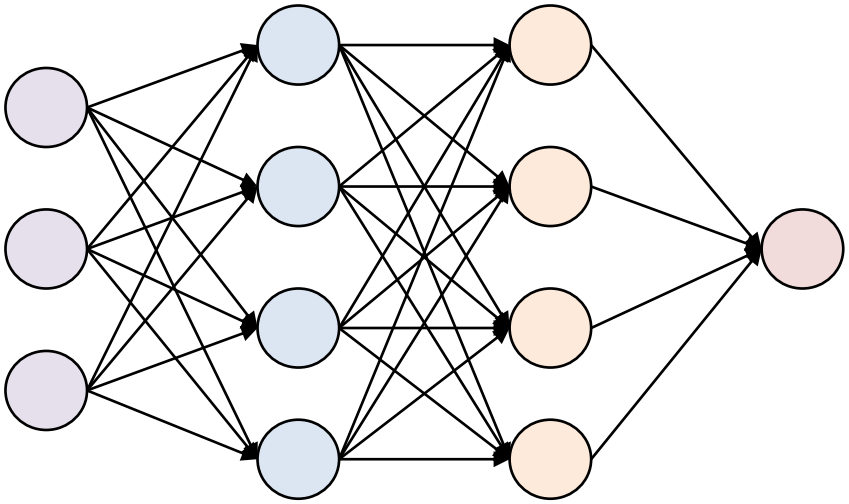


圖 1、深度神經網路架構示意圖

1. 輸入層(Input layer)：接受訊息的神經元，稱為輸入向量。
2. 隱藏層(Hidden layer)：輸入層和輸出層之間眾多神經元和連結組成的各個層面，可以有多層。
3. 輸出層(Output layer)：訊息在神經元連結中傳輸、分析、權衡後所形成的輸出結果(稱為輸出向量)。[17]

2.1.2 運作流程

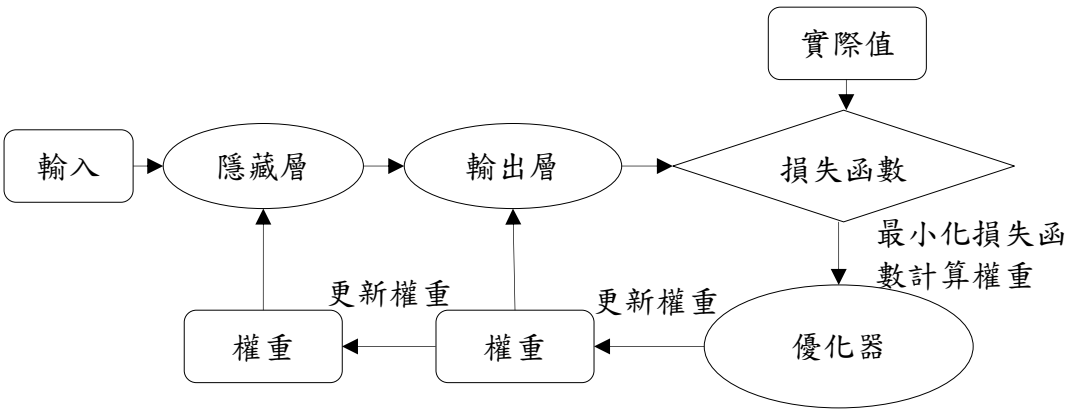


圖 2、深度神經網路流程

建立一個深度神經網路主要可以分為以下六步驟[17]：

1. 決定隱藏的深度(層數)和寬度(神經元數)。
2. 決定每層使用的激勵函數。
3. 決定模型的損失函數。
4. 決定優化器。
5. 編譯模型。
6. 開始訓練。

2.1.3 相關專有名詞介紹

1. 激勵函數(Activation function)：主要作用是將問題轉為非線性。

這裡介紹本實驗使用的 Rectified Linear Unit (ReLU)及輸出層用到的 Softmax。

ReLU 代表的意思就是當事件的值大於 0 的時候，就將這完整事件作為輸出，否則就以 0 作為這個函式的輸出，也就代表這件事不被參考。如下公式(1)及下圖 3 所示，若值為正數，則輸出該值大小，若值為負數，則輸出為 0，有解決梯度爆炸問題、計算數度快、收斂速度快等特性。z_j：第 j 個節點的輸出值。

$$ReLU(z_j) = max(0, z_j) \tag{1}$$

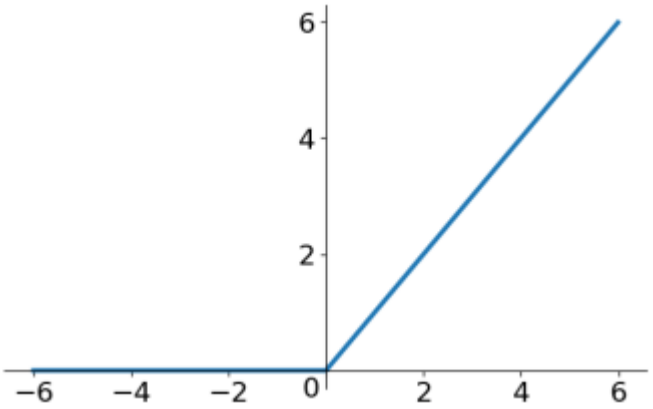


圖 3、ReLU 函數圖[18]

Softmax 用於多分類過程中，它將多個神經元的輸出，映射到(0,1)區間內，可以看成機率來理解，從而來進行多分類。作法如下公式(2)，K：類別數z_j：第j個節點的輸出值。

$$Softmax(z_j) = \frac{e^{z_j}}{\sum_{k=1}^{K} e^{z_k}} \quad \text{for } j = 1, \dots, K \tag{2}$$

2. 損失函數(Loss function)：用來描述模型預測值與真實值不一致的程度。

這裡介紹本研究中使用的 cross-entropy，其用意是在觀測預測的機率分佈與實際機率分布的誤差範圍。作法如下公式(3)，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。

$$H = \sum_{c=1}^{C} \sum_{i=1}^{n} -y_{c,i} \ln(p_{c,i}) \tag{3}$$

3. 優化器(Optimizer)：決定權種更新方式。

本研究使用的是準確步梯度下降法 Stochastic Gradient Decent(SGD)，是最單純的梯度下降方法，再利用微分的方法找出參數的梯度，往梯度的方向去更新參數(weight)。公式如下公式(4)。W：權重，L：損失函數，η：學習率，$\frac{\partial L}{\partial W}$：損失函數對參數的梯度(微分)。

$$W \leftarrow W - \eta \frac{\partial L}{\partial W} \tag{4}$$

2.2 隨機森林

隨機森林最早於 1995 年由貝爾實驗室的 T.K.Ho 所提出[19]，後 L.Breiman 和 Cutler 發展推論出隨機森林的演算法[21]。其基本原理是，結合多顆 Classification and Regression Trees(CART)，使用 GINI 不純度的決策樹，並加入隨機分配的訓練資料，以大幅增進最終的運算結果。目前已被廣泛利用於各種成績預測相關研究[7][22]當中。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(End Semester Mark ，ESM)。其研究中使用決策樹模型，結果顯示預測出 72.51%的準確率

2.2.1 基本架構

介紹隨機森林之前首先須了解決策樹。決策樹是用來處理分類問題的樹狀結構，每個內部節點表示一個特徵欄位，每條分支代表一個可能的欄位輸出結果，而每個葉節點代表不同分類的類別標籤。透過計算資訊增益(Information Gain，IG)挑選特徵作為分割條件，公式如下式(5)。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。

$$IG(D_p, f) = I(D_p) - \sum_{j=1}^{m} \frac{N_j}{N_p} I(D_j) \tag{5}$$

隨機森林使用 CART 樹，使用 Gini 不純度，算法如下公式(6)。$p(i \mid t)$：特定結點t中「樣本屬於類別i」的比例，c：分類數。

$$I_G(t) = \sum_{i=1}^{c} p(i|t)\big(1 - p(i|t)\big) = 1 - \sum_{i=1}^{c} p(i|t)^2 \tag{6}$$

這裡以存歿與否為例展示架構圖如下圖 4。

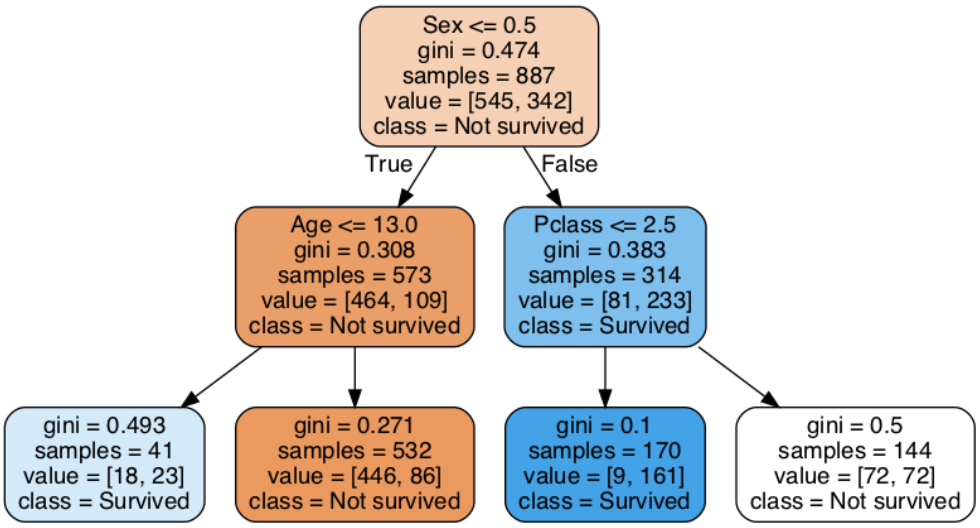


圖 4、決策樹架構示意圖[25]

接著進入隨機森林的部分，隨機森林使用集成學習：結合多個「弱學習器」來建構一個更強的模型：「強學習器」。

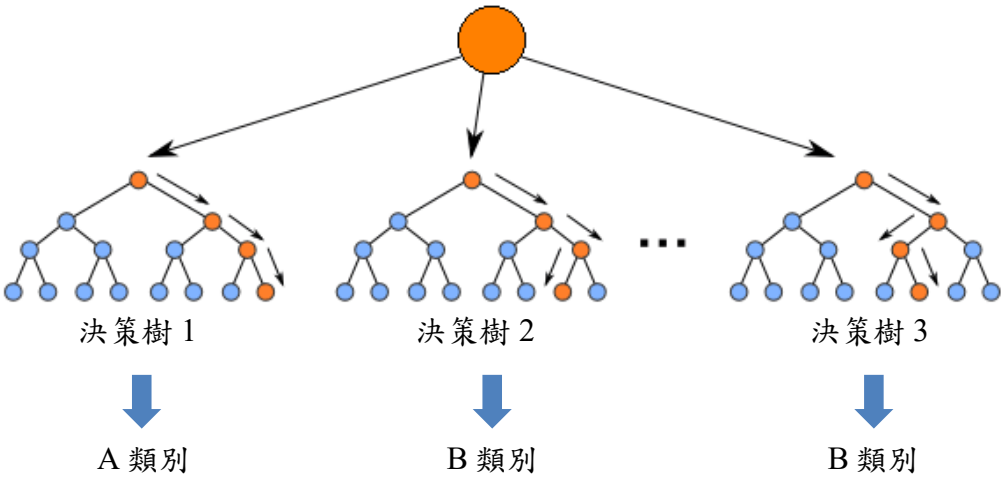


圖 5、隨機森林架構示意圖[26]

2.2.2 運作流程

建立隨機森林分類器的步驟如下步驟 1~4[26]，在這裡簡單介紹步驟 1 中使用的抽樣方法 Bootstrap Aggregation(Bagging)，指的是重新取樣原有 Data 產生新的 Data，取樣的過程是均勻且可以重複取樣的。如此一來就可以從一組 Data 中生出多組 Dataset。

1. 使用 Bagging 的方式從資料集中隨機選取 *n* 個資料，取完後放回。
2. 從選取的 *n* 個資料中，訓練出決策樹。對每一節點：
    a. 隨機選取 *d* 個特徵。
    b. 使用 CART 方式進行特徵分割該節點。
3. 重複 *k* 次步驟 1~步驟 2。
4. 彙總所有決策樹的預測，以多數決方式來決定結果。

2.3 支援向量機

Vapnik 等人於 1960 年代發明 Vapnik–Chervonenkis theory[27]。1992 年，Boser 等人提出通過將核技巧應用於最大間隔超平面來建立非線性分類器的方法[28]，當前標準的前身(軟間隔)由 Cortes 和 Vapnik 於 1993 年提出，並於 1995 年發表[29]。在學習表現預測相關研究上已有被廣泛使用[8][13]。其中 Miguéis 提出的研究[8]，使用大學前兩學期成績與及一些人口統計數據(例如：性別、婚姻狀況等)預測四年後的學業表現。其 SVM 模型的預測結果達到 92.6%。

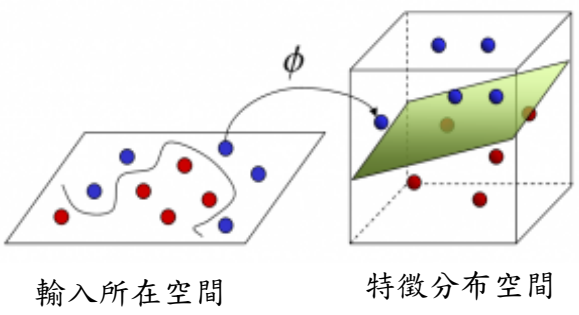


圖 6、SVM 示意圖[30]

SVM 透過核轉換將在低微度空間線性不可分的樣本映射到高維度空間去，找到一個超平面將這些樣本做有效的切割(如上圖 6)，超平面兩邊的樣本與超平面本身的距離會決定模型訓練的好壞。超平面公式如下公式(7)，其中k表核函式。公式(8)為多項式核函式(Polynomial kernel function)，它表示特徵空間中的向量與原始變量的多項式的相似性，從而允許學習非線性模型。

$$f(x) = sgn(\sum_{i=1}^{n} a_i y_i K(x_i^t \cdot x) + b) \tag{7}$$

$$k(x_i, x_j) = (x_i^t \cdot x_j)^d \text{，} d \geq 1 \tag{8}$$

第三章　實驗方法與資料集

本研究之目的為比較不同模型，用以預測成績表現，並從中建立最佳的資料標籤方式及預測模型。因此，何種標籤方式與適當的預測模型為本研究之重點。本研究所收錄 104 年至 108 年申請入學之學生資料，進行資料前處理。本章於 3.1 節首先介紹實驗架構及流程，3.2 節介紹程式編譯環境設定、函數使用，3.3 節介紹資料來源，3.4 節介紹資料前處理與遺失值處裡，最後於 3.5 節介紹訓練模型的建置方式。

3.1　實驗架構及流程

本研究以學生入學成績為輸入集，先將入學考試成績進行預處理，預測結果為入學後第一、二學期的成績表現，學生成績依照系所安排之課綱地圖分成學科總平均、程式科、數學科、基礎科四類，分別預測各類別之表現。本研究主要提出因應特定學生資料集的最佳預測模型。

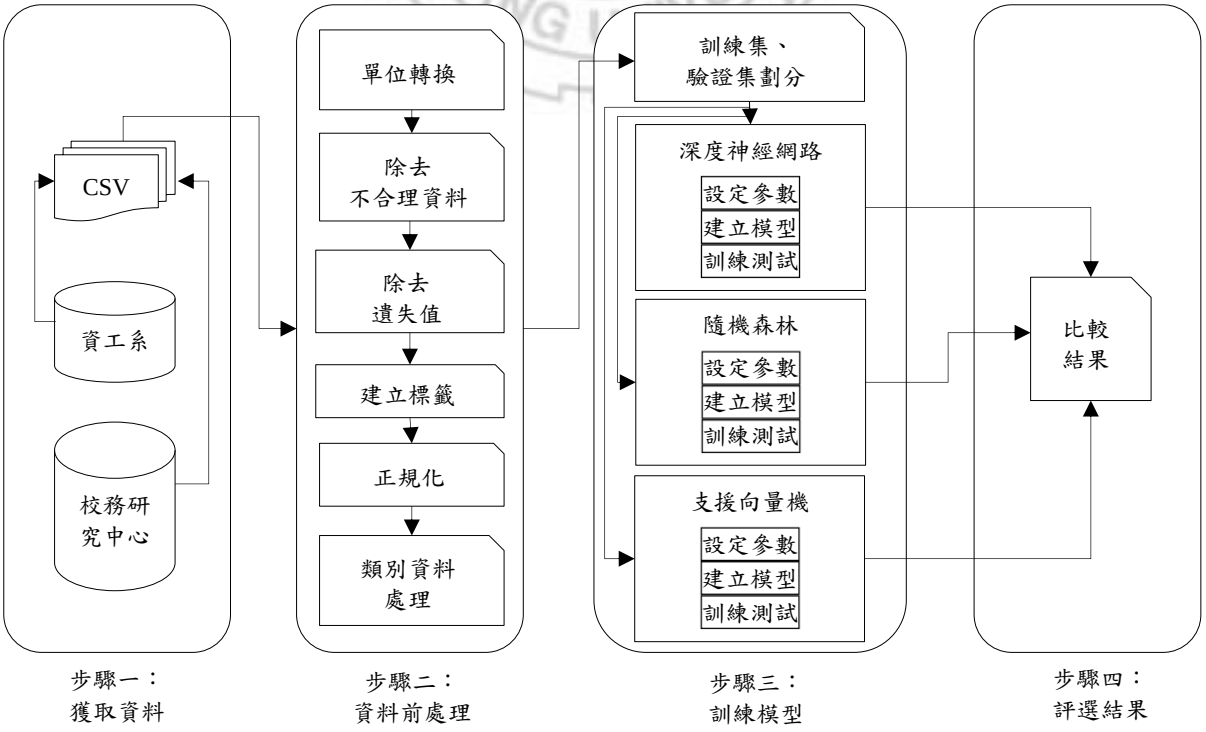
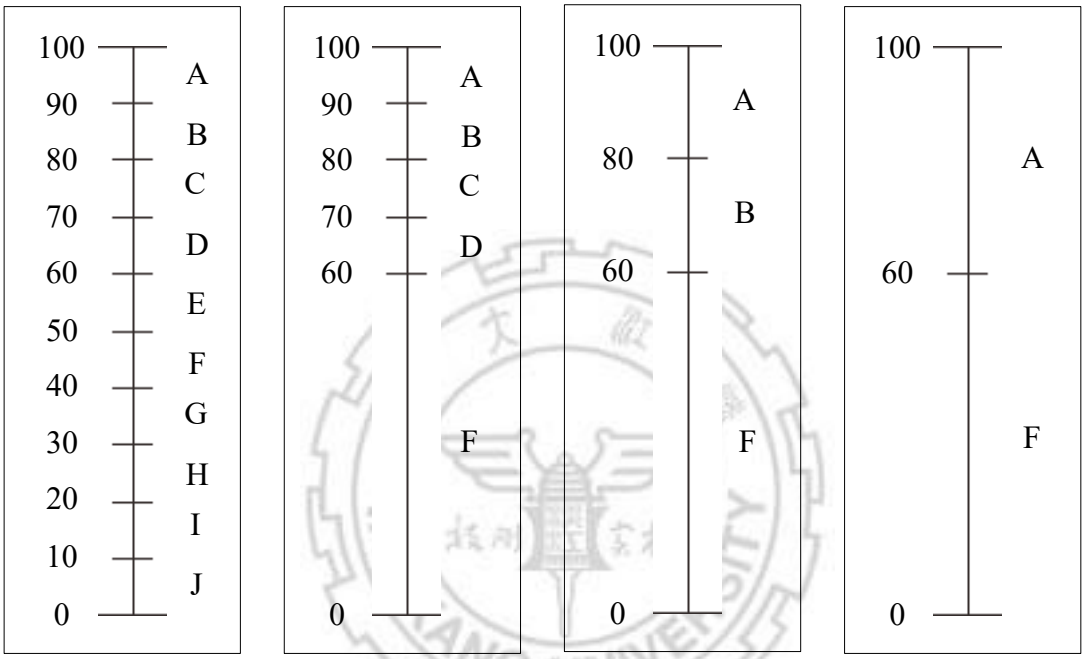


圖 7、實驗架構圖

在本研究架構中，步驟一為獲取資料。步驟二為資料前處理，資料來源及處理方式在 3.3 節作詳細解釋。各個資料輸入欄位依據不同性質做不同處理，再去除所有遺失值，可作為預測的資料集共計有 328 名學生。接下來說明本研究的標籤方式如下圖 8 所示，共分為四種，分別為「10 分類」、「5 分類」、「3 分類」、「2 分類」。



(a)10 分類　(b)5 分類　(c)3 分類　(d)2 分類

圖 8、資料標籤分類

下圖 9 至圖 12 分別為學科總平均、程式科平均、數學科平均、基礎科平均分數依照以上四種標籤方式分組後的人數分布圖。

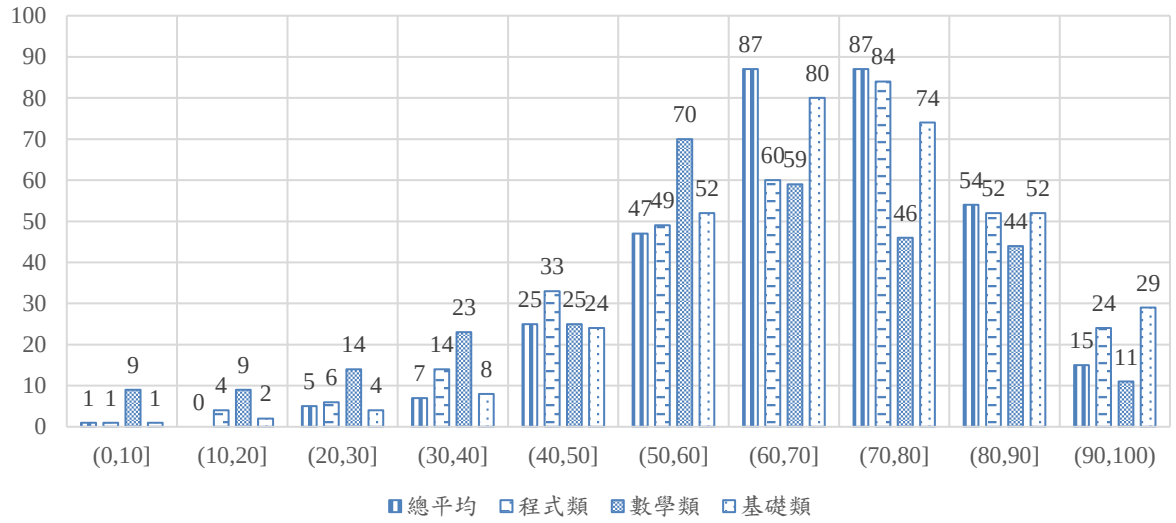


圖 9、「10 分類」成績人數分布圖

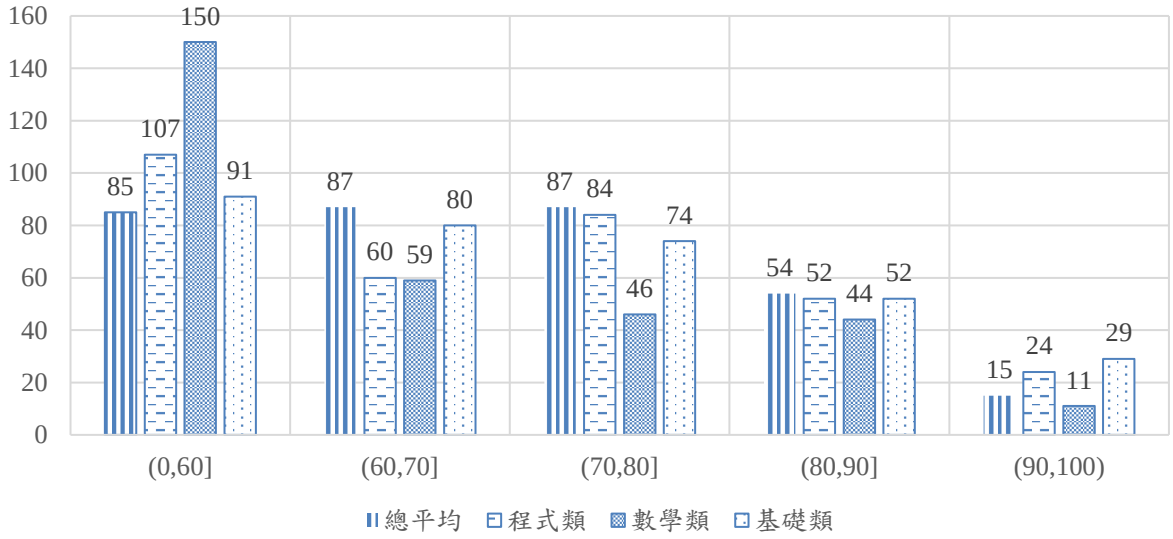


圖 10、「5 分類」成績人數分布圖

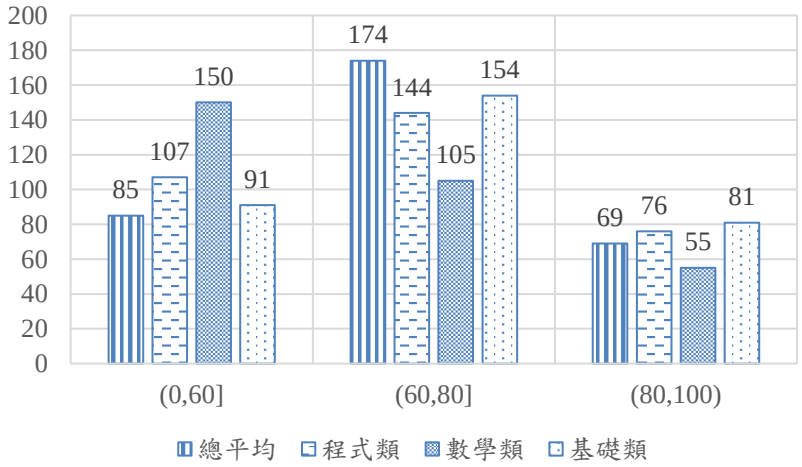


圖 11、「3 分類」成績人數分布圖

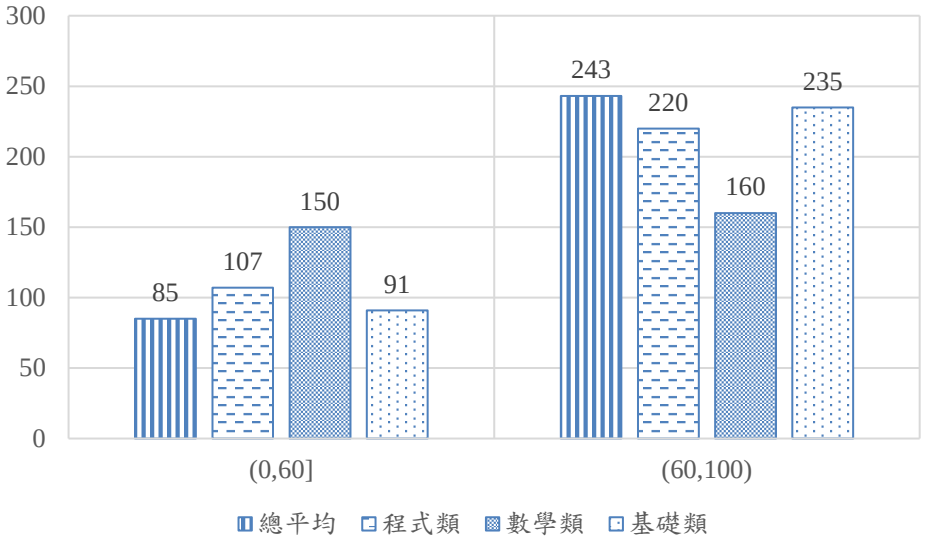


圖 12、「2 分類」成績人數分布圖

在建立標籤之後，進入步驟三：模型訓練。本研究將資料集進行深度神經網路、隨機森林、支援向量機模型的訓練，分別對學科總平均、程式科平均、數學科平均、基礎科平均四種成績依不同的標籤方式分別進行預測。測試集驗證將資料集劃分為訓練集、驗證集、以及最後用來評估結果的測試集，其比例為 8:1:1。其中這三種資料集的資料互不交集。本研究在此次實驗使用的各模型結構詳細參數設定列舉至 3.5。

最後第四步驟為評估結果，主要觀察目標為模型套件的評估函數所產生之正確率。模型評估方式以所有資料集獨立分出來的測試集作為輸入，利用真實答案以及模型預測的答案進行正確率計算。

3.2 實驗之程式編譯環境設定、函數使用

本研究所有實驗過程：資料收集、前處理、模型訓練以及模型評估皆由 python 語言實現，使用 jupyter notebook 工具進行編譯。系統環境為 Window10 x64。資料前處理從 jupyter notebook 連接本地端 MySQL 伺服器，透過 SQL 指

令進行。模型建立、訓練與評估使用 Keras 建立深度學習模型與環境，Scikit-learn 建立隨機森林與支援向量機的模型。下表 1 列舉所各模型使用的函數。

表 1、各模型使用之函數清單與說明

模型	套件名稱	函數名稱	說明
資料前處理	pymysql	connect	建立 MySQL server 連線
資料前處理	pymysql	cursor	建立 cursor 物件
資料前處理	pymysql	execute	執行 SQL 指令
DNN、RF、SVM	os	chdir	指定路徑
DNN、RF、SVM	pandas	read_csv	讀取 csv 檔
DNN、RF、SVM	pandas	get_dummies	將無序資料轉換為讀熱碼
DNN、RF、SVM	numpy	split	分割資料集
DNN	keras	Sequential	定義深度神經網路模型
DNN	keras	layers	定義層內容(包含 input、active function、dropout 等)
DNN	keras	summary_ generator	顯示模型大綱
DNN	keras	compile_ generator	編譯模型(給定優化方式、loss function、metrics 等)
DNN	keras、sklearn.ensemble	fit_generator	配合批次讀取的模型訓練
DNN	keras	history_ generator	訓練結果歷程記錄
DNN	matplotlib.pyplot	plot、title、ylabel、xlabel、legend	給定圖表參數
DNN	matplotlib.pyplot	show_ generator	顯示圖表
DNN	keras	evaluate_genetator	配合批次讀取的模型評估
RF	sklearn.ensemble	RandomForestClassifier	建立隨機森林分類模型
RF	sklearn.ensemble、sklearn.svm	predict_generator	配合批次讀取的模型預測
RF	sklearn.ensemble	feature_importances_ generator	找出特徵重要性
RF、SVM	sklearn.metrics	classification_report	產生結果報告
SVM	sklearn.svm	SVC	建立支援向量機模型
SVM	sklearn.multiclass	OneVsRestClassifier	建立 ovr 的模型
SVM	sklearn.multioutput	MultiOutputClassifier	建立多分類模型

3.3 資料集來源說明

本研究資料集的匯整總共有兩個來源，於 3.3.1 及 3.3.2 節做說明：

1. 淡江大學資訊工程學系辦公室
2. 淡江大學校務研究中心

3.3.1 淡江大學資訊工程系辦公室

系辦是系所招生、系上學生課程管理、系上學生相關業務處裡的主要角色。透過提出申請並提交申請原因後，由行政助理協助至資料庫撈取資料。本研究主要取得的資料為學測成績、前兩學期在校成績。欄位內容及處理方式於 3.4 節說明。

3.3.2 校務研究中心

校務研究中心藉由蒐集教學、研究、學生學習及行政支援等校務資料，並結合自我評鑑結果，適時提供適量具事實依據之分析研究，以為校務發展決策及成效評估之參考依據。透過提出申請並提交申請原因後，由專任助理協助向資訊處提出撈取資料之需求。因應個資法的規範，校務研究中心提供之資料皆經過去識別化處理，故有另外提出請資訊處協助將系辦提供之資料一併去識別化之申請。本研究主要取得的資料為學生基本資料。欄位內容及處理方式於 3.4 節說明。

3.4 資料前處理與遺失值處理

本節主要介紹資料前處裡，分為欄位資料處理及遺失值處理兩部分。下表 2 列出所有實驗中使用的特徵與標籤及其是否需要前處理與遺失數量。

表 2、特徵與標籤清單

欄位名稱	欄位屬性	是否經過前處理	遺失數量
國文	特徵	✓	0
英文	特徵	✓	0
數學	特徵	✓	0
社會	特徵	✓	0
自然	特徵	✓	0
高中 PR 值	特徵	✓	0
高中 GPA	特徵	×	0
總平均	標籤	✓	0
程式科	標籤	✓	1
數學科	標籤	✓	18
基礎科	標籤	✓	1

3.4.1 欄位資料處理

1. 國文、英文、數學、社會、自然：

- 來源：淡江資工系辦提供原始學測級分。
- 處裡方式：級分對應至人數累積百分比。

學測成績皆以級分表示，但因每年的考試難度有差異，同級分於不同年度無法代表相同的原始成績。故這裡使用大考中心提供之成績統計資料－各科級分人數百分比累計表，將級分對應至人數累積百分比做為學測成績資料值。

2. 高中 PR 值：

- 來源：校務研究中心提供學生畢業學校。
- 處裡方式：網上搜尋 PR 值。

PR 值(又稱為百分等級)，是先將該次測驗所有考生的量尺總分排序後，依照人數均分成一百等分，該生大約會落在第幾個等分中。高中的 PR 值則由國中基本學力測驗來決定。但由於只有公立高中會公布 PR 值，且近幾年推行 12 年國教，改為國中會考制度，已無 PR 值可參考，故此欄位值多由網路上搜尋而來，較不可靠。

3. 總平均：

- 來源：淡江資工系辦提供該年度大一班成績。
- 處理方式：篩出系必修科目後做平均，並依實驗需要按四種指定區間做資料標籤分類。

本研究主要目的即希望能預測出學生課業表現，故這裡將原始的成績資料依照系辦提供之必修科目表篩選出系必修課程成績做總平均。

4. 程式科、數學科、基礎科：

- 來源：淡江資工系辦提供該年度大一班成績。
- 處理方式：篩出類別科目後做平均，並依實驗需要按四種指定區間做資料標籤分類。

本研究想更進一步知道學生於不同領域的課業表現，但因每年課綱都有故微調，故將必修科目依系辦提供的課程地圖分成程式科、數學科、基礎科三類，必修科目及其對應分類詳下表 3，並對各類別成績分別求平均值，最後再依實驗需要按 3.1 節提到的四種資料標籤方式分類。

表 3、必修科目分類表

科目名稱	科目分類
計算機程式語言	程式科
高等程式語言	程式科
Linux 作業系統實務	程式科
機率論	數學科
微積分	數學科
離散數學	數學科
數位系統導論	基礎科
邏輯設計實驗	基礎科
計算機概論	基礎科
資訊概論	基礎科

3.4.2 遺失值處裡

因學生有自行選擇修習哪門課程的自由，以及期中加退選制度，若有沒選課或退選情況則會產生遺失值。若該學生於該課程類別僅有部分科目成績遺失，仍可以剩下的科目成績做該科目類的平均，但若該學生於整學年皆無該類別科目成績則成為遺失值。因所有學生成績皆為真實存在，為求公平本實驗將遺失值去除。

3.5 模型參數設定

本實驗使用 7 個輸入特徵：國文、英文、數學、社會、自然、高中 PR 值、高中 GPA，透過 3 種模型：深度神經網路、隨機森林、支援向量機，對 4 種標籤方式：10 分類、5 分類、3 分類、2 分類，分別進行訓練。以下分別介紹各模型參數設定。

3.5.1 深度神經網路模型架構

為防止過度擬和[31]，設定 Dropout 部分，其值為 0.4。優化函數使用 SGD，損失函數使用 categorical_crossentropy，激勵函數使用 ReLu，輸出層的激勵函數為 softmax，模型架構為 3 層，節點個數為 10，batch_size 為 10，epochs 為 5，評估標準使用 categorical_accuracy。參數設定如下圖 13，以「10 分類」為例，其餘標籤方式僅差在 output 個數。特別說明此處的輸出個數為 9 的原因是資料樣本缺少成績落在 I 區間(10-20]的資料。

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 10)                80
_________________________________________________________________
dropout (Dropout)            (None, 10)                0
_________________________________________________________________
dense_4 (Dense)              (None, 10)                110
_________________________________________________________________
dropout_4 (Dropout)          (None, 10)                0
_________________________________________________________________
dense_8 (Dense)              (None, 9)                 99
=================================================================
Total params: 289
Trainable params: 289
Non-trainable params: 0
```

圖 13、「10 分類」深度神經網路模型架構

藉由測試各種參數設定後依照結果決定使用以上參數設定，這裡將測試過程之準確率結果整理如下表 4。以總平均預測的模型進行測試，選擇意義上參考價值較高「3 分類」及「5 分類」的標籤方式挑選準確率較高的參數設定。

表 4、DNN 決定參數過程

	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：50 batch_size：10 epoch：5	36.36%	27.27%	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：10 epoch：5	36.36%	36.36%	60.61%	84.85%
層數：50 節點數：200 batch_size：10 epoch：5	24.24%	36.36%	60.61%	84.85%
層數：200 節點數：50 batch_size：10 epoch：5	24.24%	15.15%	60.61%	84.85%
層數：200 節點數：200 batch_size：10 epoch：5	36.36%	36.36%	60.61%	84.85%
層數：3 節點數：10 batch_size：50 epoch：50	24.24%	24.24%	60.61%	84.85%
層數：3 節點數：10 batch_size：100 epoch：100	27.27%	36.36%	60.61%	84.85%
層數：50 節點數：50 batch_size：50 epoch：50	24.24%	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%

3.5.2 隨機森林模型架構

隨機森林模型主要需要調整參數有兩個 n_estimators 及 criterion。n_estimators 主要用來限制決策樹個數，避免過度擬和，這裡通過測試(如下表 5)選擇「3 分類」及「5 分類」皆表現較佳的 5。Criterion 決定使用何種不純度，如 2.2 節提到 RF 分類器使用 Gini 不純度。參數決定過程整理如下表 5。

表 5、RF 決定參數過程

	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%

3.5.3 支援向量機模型架構

支援向量機模型需要調整參數有核函式 kernel 及多分類方式 decision_function_shape。Kernel 設為 Poly，參數 degree 僅與 poly 相關，這裡設為 3，代表在 3 維空間做線性分割。在測試選擇參數過程中 OneVsOne(ovo)與 OneVsRest(ovr)表現一樣，故選擇預設值 ovr。Cost(C)參數表示容錯項，數值越高容錯越低，這裡設為 1。參數決定過程整理如下表 6。

表 6、SVM 決定參數過程

	10 分類	5 分類	3 分類	2 分類
kernel：linear shape：ovr C：1	0.00%	9.09%	51.52%	84.85%
kernel：poly shape：ovr degree：1 C：1	0.00%	0.00%	60.61%	84.85%
kernel：poly shape：ovr degree：2 C：1	0.00%	0.00%	45.45%	84.85%
kernel：poly shape：ovr degree：3 C：1	0.00%	6.06%	63.64%	90.91%
kernel：poly shape：ovr degree：4 C：1	3.03%	12.12%	42.42%	87.88%
kernel：poly shape：ovr degree：5 C：1	9.09%	18.18%	39.39%	84.85%
kernel：poly	3.03%	15.15%	39.39%	84.85%

shape：ovr degree：3 C：50				
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%

第四章　資料分析與實驗結果

本研究的各項實驗成果，將於本章詳細的敘述以及討論，實驗內容包括實驗結果的評估方式、實驗結果。本章節分別說明學業成績透過三種機器學習模型於不同資料標籤方式的表現，評估指標為準確率，針對 10 分類及 5 分類增加鄰近正確率(1-Away)[32]：預測結果與測試集答案「前後相符」，意及預測結果與測試集答案相符、加一及減一皆算正確。為了不失去資料標籤分類的意義，5 分類的部分僅針對 60 分以上成績做鄰近正確率。

4.1　實驗結果

實驗結果評估以總平均、程式科平均、數學科平均及基礎科平均四項目標值做說明。每項目標值中對 4 種標籤方式分別呈現。最後展示特徵重要性。

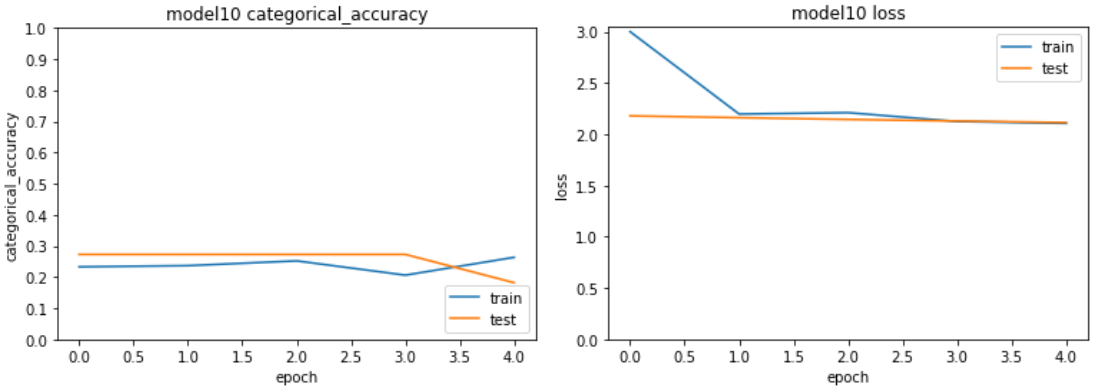
表 7、實驗簡表

模型	深度神經網路(DNN)、隨機森林(RF)、支援向量機(SVM)
特徵欄位	國文、英文、數學、社會、高中 PR 值、高中 GPA
標籤欄位	總平均、程式科、數學科、基礎科
標籤方式	10 分類(A-J)、5 分類(A-D、F)、3 分類(A-B、F)、2 分類(A、F)

4.1.1 學科總平均實驗結果

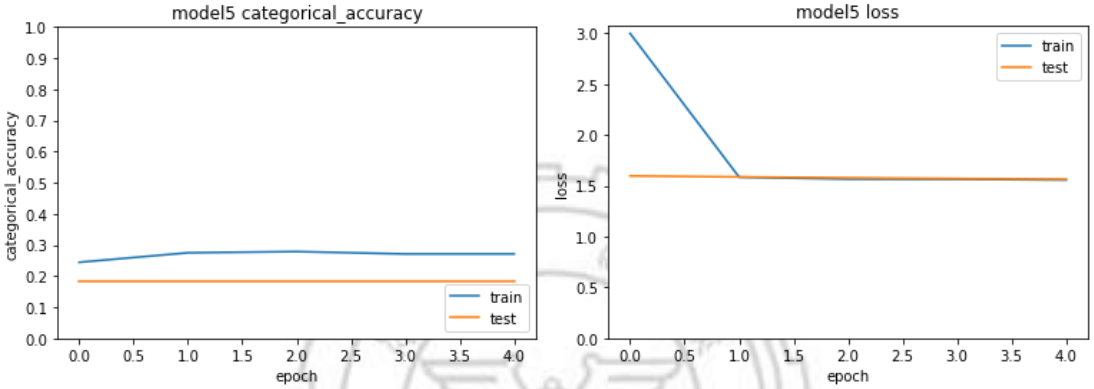
1. 深度神經網路：

訓練過程的準確率與損失值情形如下圖 14 至圖 17。為了更清楚觀察訓練時損失值變化，這裡將 y 軸刻度放大並將大於 3 的值以 3 表示。



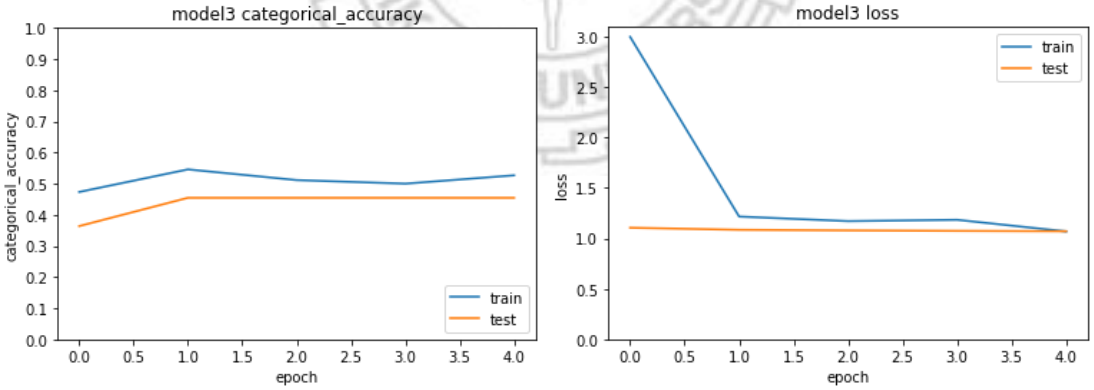
(a)正確率　　(b)損失值

圖 14、總平均 DNN「10 分類」訓練結果



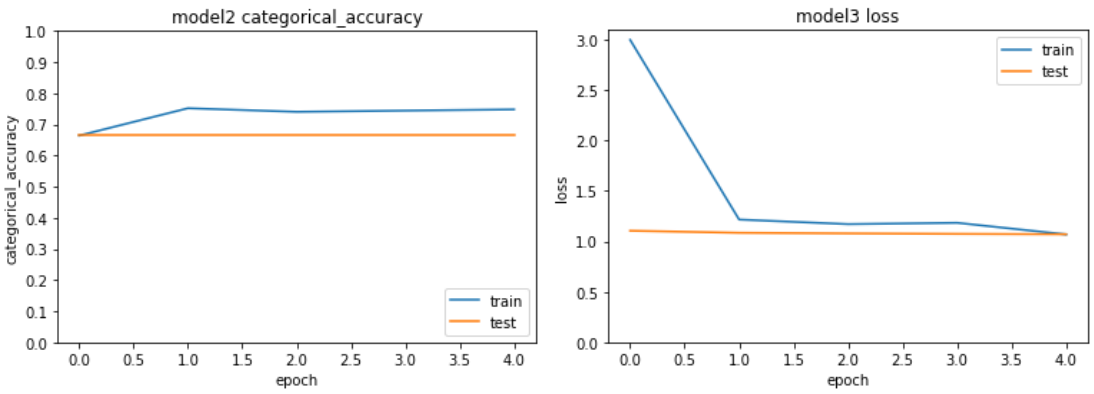
(a)正確率　　(b)損失值

圖 15、總平均 DNN「5 分類」訓練結果



(a)正確率　　(b)損失值

圖 16、總平均 DNN「3 分類」訓練結果



(a)正確率　　(b)損失值

圖 17、總平均 DNN「2 分類」訓練結果

透過 evaluate 函數預測得到準確率及損失值與鄰近正確率如下表 8。

表 8、總平均 DNN 預測結果

	10 分類	5 分類	3 分類	2 分類
準確率	36.36%	36.36%	60.61%	84.85%
鄰近正確	66.67%	63.64%		
損失值	2.05	1.58	1.02	0.54

2. 隨機森林：

透過 accuracy_score 函數得到結果與鄰近正確率如下表 9。

表 9、總平均 RF 預測結果

	10 分類	5 分類	3 分類	2 分類
準確率	27.27%	27.27%	72.73%	87.88%
鄰近正確	66.67%	48.48%		

3. 支援向量機：

透過 accuracy_score 函數得到結果與鄰近正確率如下表 10。

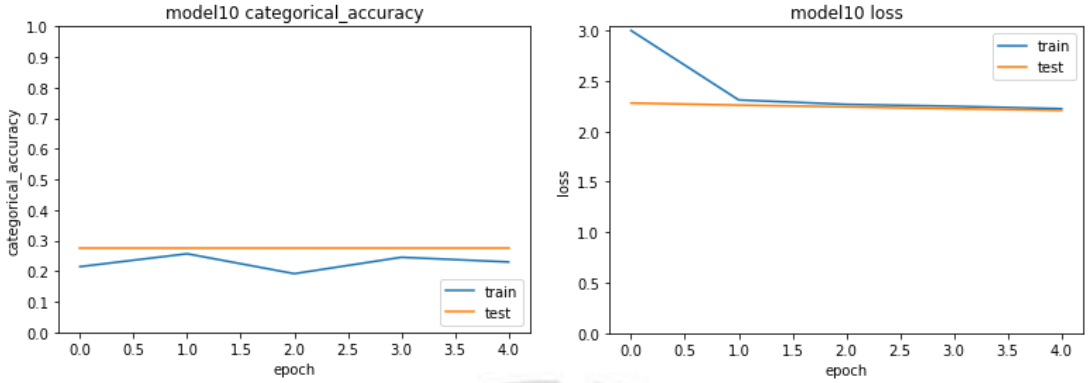
表 10、總平均 SVM 預測結果

	10 分類	5 分類	3 分類	2 分類
準確率	0.00%	6.06%	63.64%	90.91%
鄰近正確	3.03%	9.09%		

4.1.2 程式科平均

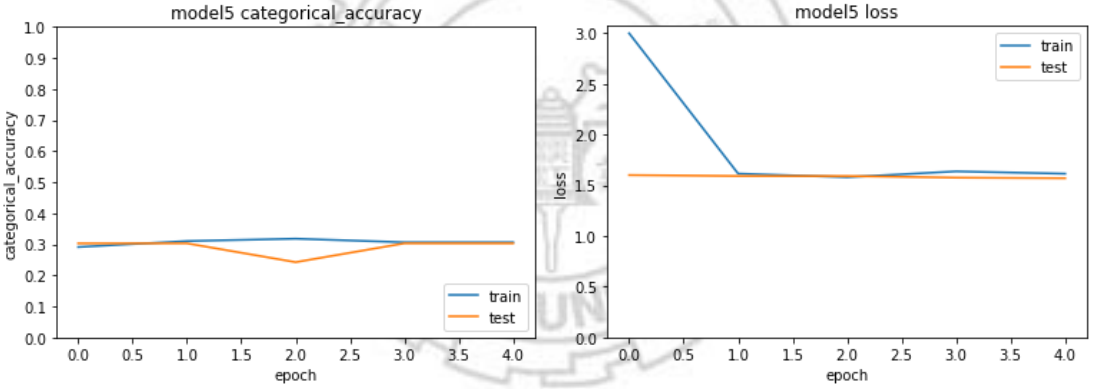
1. 深度神經網路：

訓練過程的準確率與損失情形如下圖 18 至圖 21。



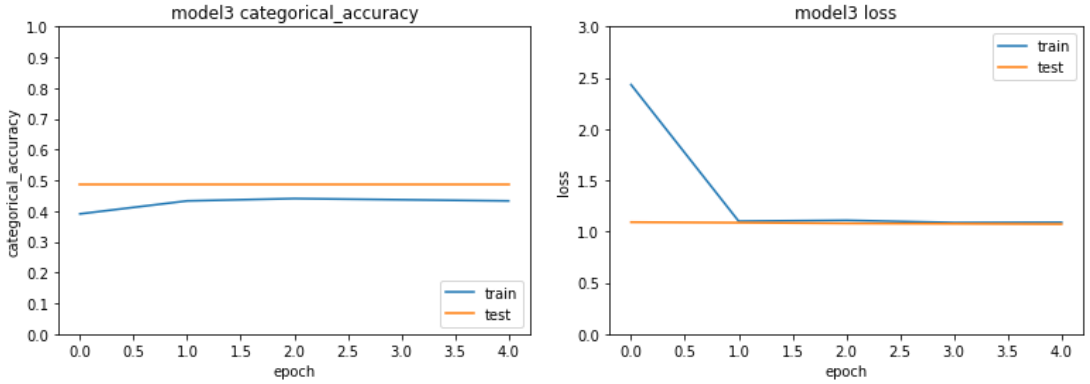
(a)正確率　　(b)損失值

圖 18、程式科 DNN「10 分類」訓練結果



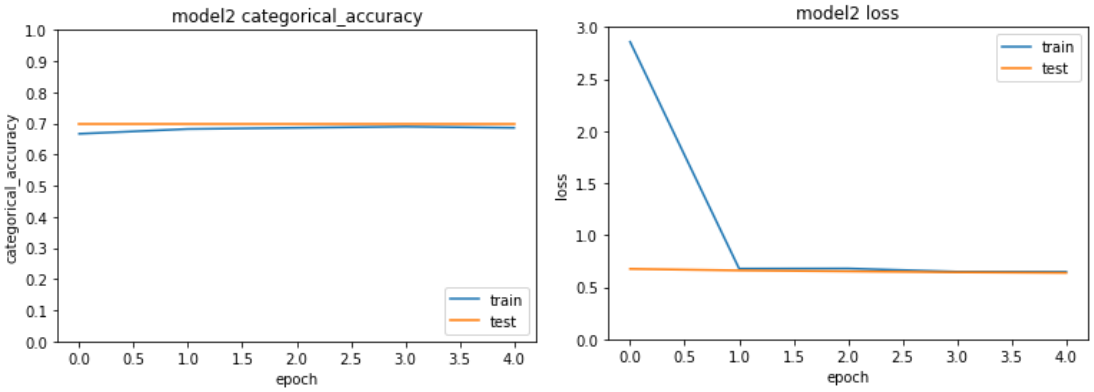
(a)正確率　　(b)損失值

圖 19、程式科 DNN「5 分類」訓練結果



(a)正確率　　(b)損失值

圖 20、程式科 DNN「3 分類」訓練結果



(a)正確率　　(b)損失值

圖 21、程式科 DNN「2 分類」訓練結果

透過 evaluate 函數預測得到準確率及損失值與鄰近正確率如下表 11。

表 11、程式科 DNN 預測結果

	10 分類	5 分類	3 分類	2 分類
準確率	24.24%	24.24%	45.45%	75.76%
鄰近正確	57.58%	27.27%		
損失值	2.24	1.59	1.08	0.61

2. 隨機森林：

透過 accuracy_score 函數得到結果與鄰近正確率如下表 12。

表 12、程式科 RF 預測結果

	10 分類	5 分類	3 分類	2 分類
準確率	18.18 %	30.30%	60.61%	84.85%
鄰近正確	54.55%	57.58%		

3. 支援向量機：

透過 accuracy_score 函數得到結果與鄰近正確率如下表 13。

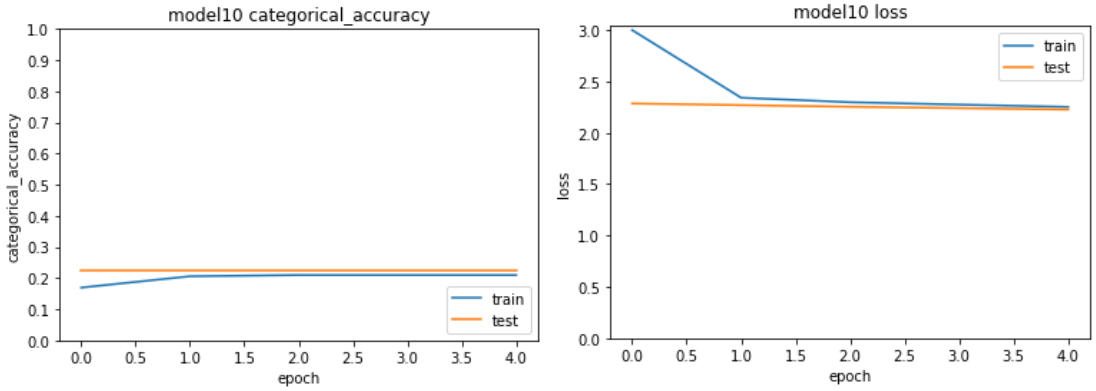
表 13、程式科 SVM 預測結果

	10 分類	5 分類	3 分類	2 分類
準確率	0.00 %	9.09%	36.36%	78.79%
鄰近正確	3.03%	12.12%		

4.1.3 數學科平均

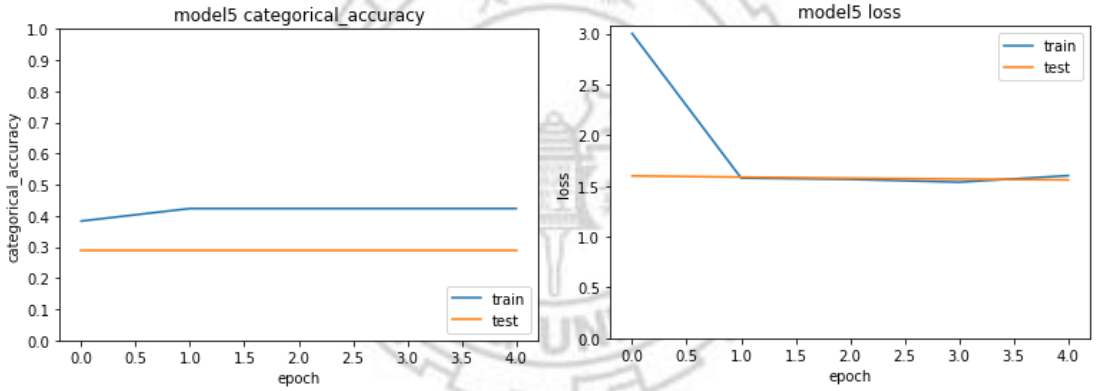
1. 深度神經網路：

訓練過程的準確率與損失情形如下圖 22 至圖 25。



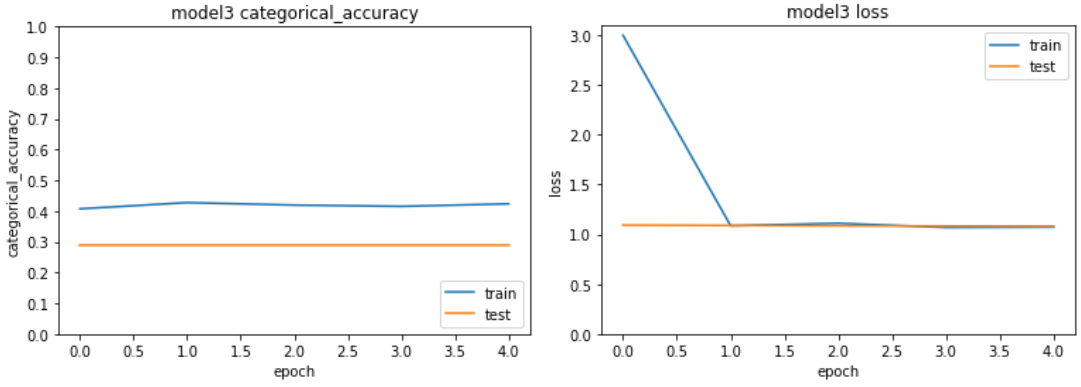
(a)正確率　　(b)損失值

圖 22、數學科 DNN「10 分類」訓練結果



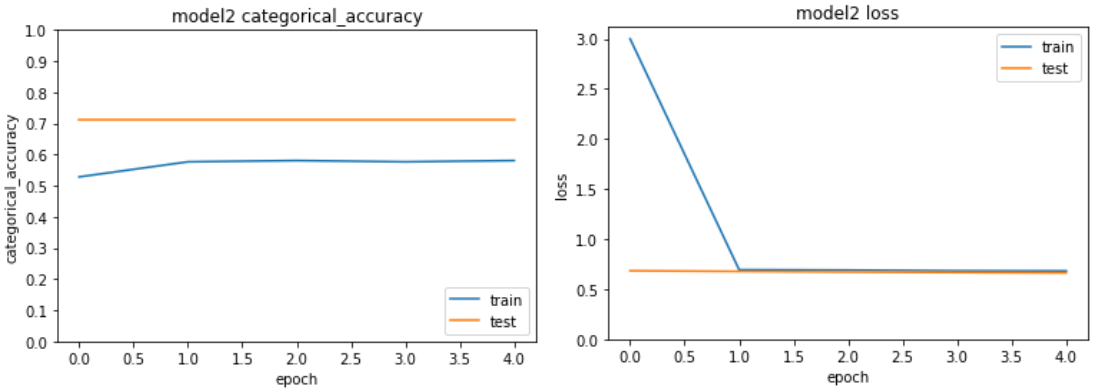
(a)正確率　　(b)損失值

圖 23、數學科 DNN「5 分類」訓練結果



(a)正確率　　(b)損失值

圖 24、數學科 DNN「3 分類」訓練結果



(a)正確率　　(b)損失值

圖 25、數學科 DNN「2 分類」訓練結果

透過 evaluate 函數預測得到準確率及損失值與鄰近正確率如下表 14。

表 14、數學科 DNN 預測結果

	10 分類	5 分類	3 分類	2 分類
準確率	35.48%	41.94%	41.94%	58.06%
鄰近正確	70.97%	45.16%		
損失值	2.23	1.52	1.05	0.68

2. 隨機森林：

透過 accuracy_score 函數得到結果與鄰近正確率如下表 15。

表 15、數學科 RF 預測結果

	10 分類	5 分類	3 分類	2 分類
準確率	6.45%	32.26%	54.84%	70.97%
鄰近正確	29.03%	45.16%		

3. 支援向量機：

透過 accuracy_score 函數得到結果與鄰近正確率如下表 16。

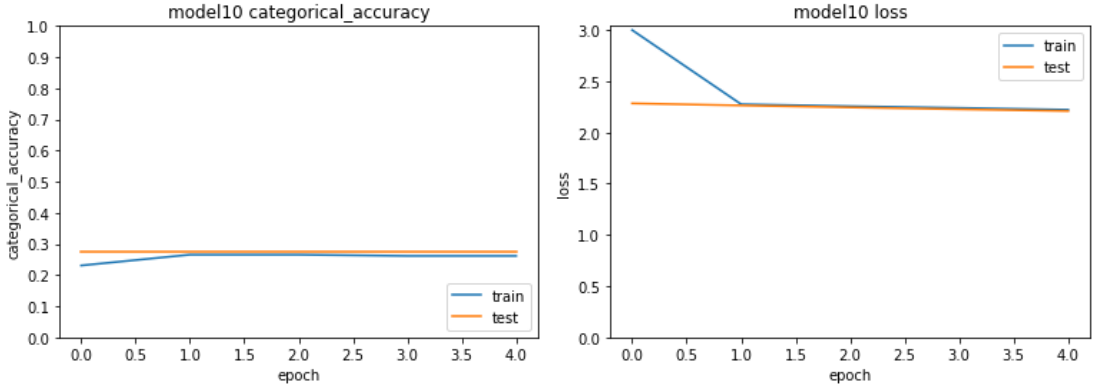
表 16、數學科 SVM 預測結果

	10 分類	5 分類	3 分類	2 分類
準確率	0.00 %	19.35%	22.58%	70.97%
鄰近正確	3.23%	22.58%		

4.1.4 基礎科平均

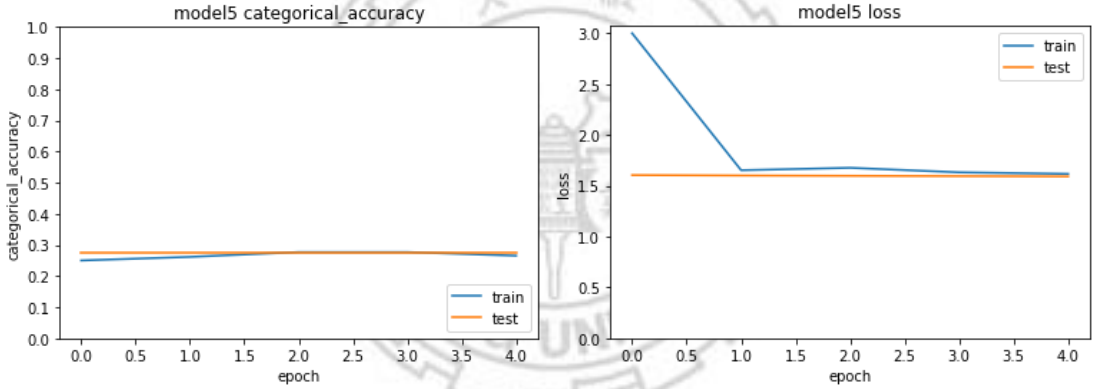
1. 深度神經網路：

訓練過程的準確率與損失情形如下圖 26 至圖 29。



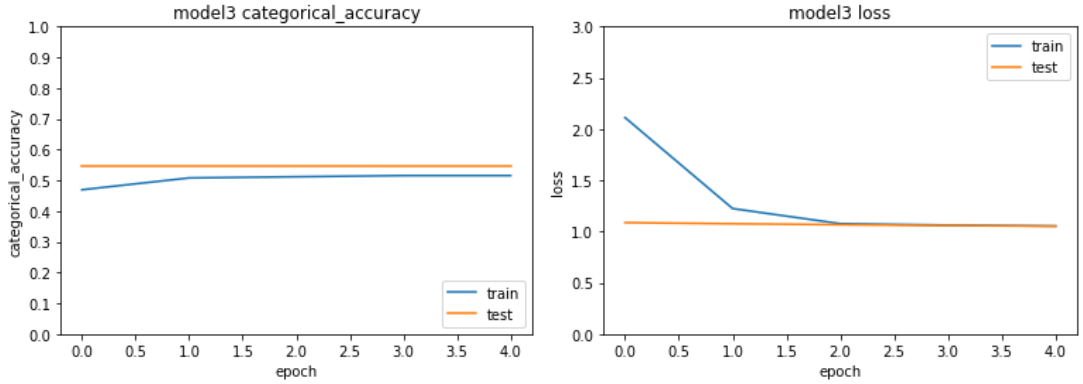
(a)正確率　　(b)損失值

圖 26、基礎科 DNN「10 分類」訓練結果



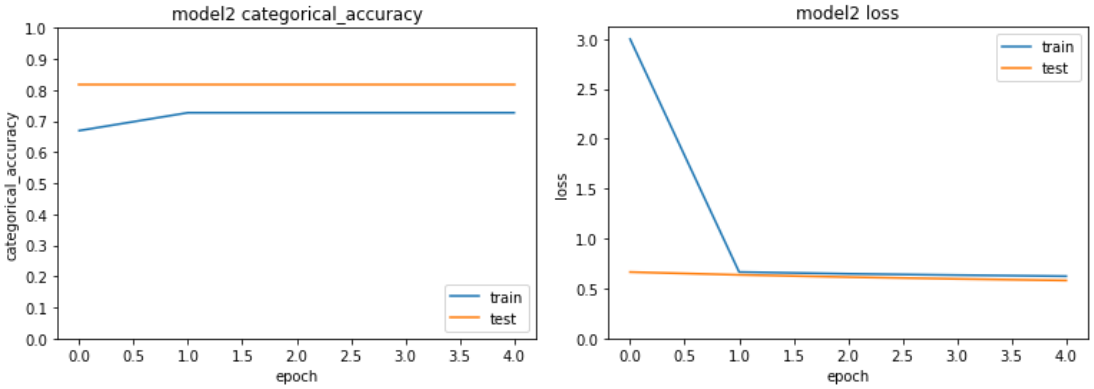
(a)正確率　　(b)損失值

圖 27、基礎科 DNN「5 分類」訓練結果



(a)正確率　　(b)損失值

圖 28、基礎科 DNN「3 分類」訓練結果



(a)正確率　　(b)損失值

圖 29、基礎科 DNN「2 分類」訓練結果

透過 evaluate 函數預測得到準確率及損失值與鄰近正確率如下表 17。

表 17、基礎科 DNN 預測結果

	10 分類	5 分類	3 分類	2 分類
準確率	6.06%	21.21%	30.30%	78.79%
鄰近正確	48.48%	24.24%		
損失值	2.23	1.63	1.27	0.58

2. 隨機森林：

透過 accuracy_score 函數得到結果與鄰近正確率如下表 18。

表 18、基礎科 RF 預測結果

	10 分類	5 分類	3 分類	2 分類
準確率	9.09%	9.09%	30.30%	78.79%
鄰近正確	27.27%	21.21%		

3. 支援向量機：

透過 accuracy_score 函數得到結果與鄰近正確率如下表 19。

表 19、基礎科 SVM 預測結果

	10 分類	5 分類	3 分類	2 分類
準確率	0.00 %	0.00%	24.24%	78.79%
鄰近正確	3.03%	3.03%		

4.1.5 特徵重要性

透過隨機森林模型的 feature_importances_函數得到特徵重要性。依不同目標成績將 4 種標籤方式的特徵重要性平均後整理成下表 20-23。

表 20、總平均特徵重要性

欄位名稱	重要性
國文	0.10
英文	0.14
數學	0.09
社會	0.11
自然	0.10
高中 PR 值	0.12
高中 GPA	0.34

表 21、程式科特徵重要性

欄位名稱	重要性
國文	0.11
英文	0.14
數學	0.12
社會	0.10
自然	0.11
高中 PR 值	0.14
高中 GPA	0.27

表 22、數學科特徵重要性

欄位名稱	重要性
國文	0.11
英文	0.13
數學	0.12
社會	0.14
自然	0.13
高中 PR 值	0.13
高中 GPA	0.25

表 23、基礎科特徵重要性

欄位名稱	重要性
國文	0.14
英文	0.11
數學	0.12
社會	0.14
自然	0.15
高中 PR 值	0.15
高中 GPA	0.18

4.2 實驗之結果比較

實驗結果比較首先展示完全正確率分布圖(圖 30-圖 33)，再來展示特徵重要性將 4 種標籤方式結果平均後的分布圖(圖 34)，接著展示完全正確率與鄰近正確率之比較(圖 35-38)，最後將模型表現依照排名整理成表格(表 24-27)，以更直觀的比較出各個模型的表現。

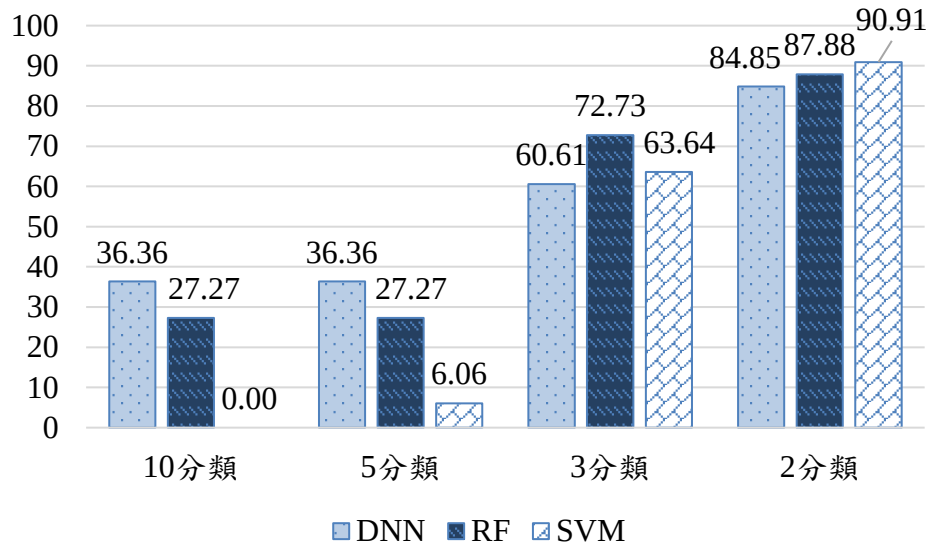


圖 30、總平均預測正確率分布圖

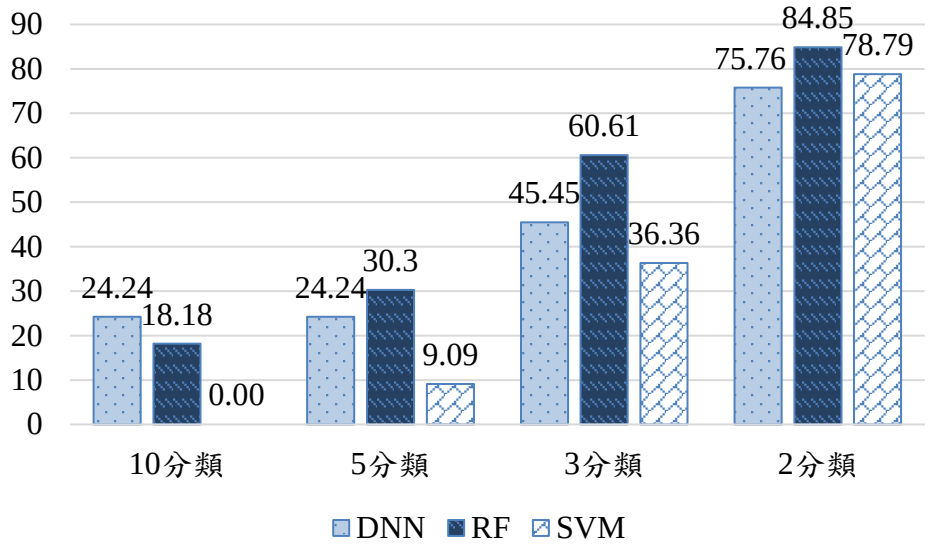


圖 31、程式科預測正確率分布圖

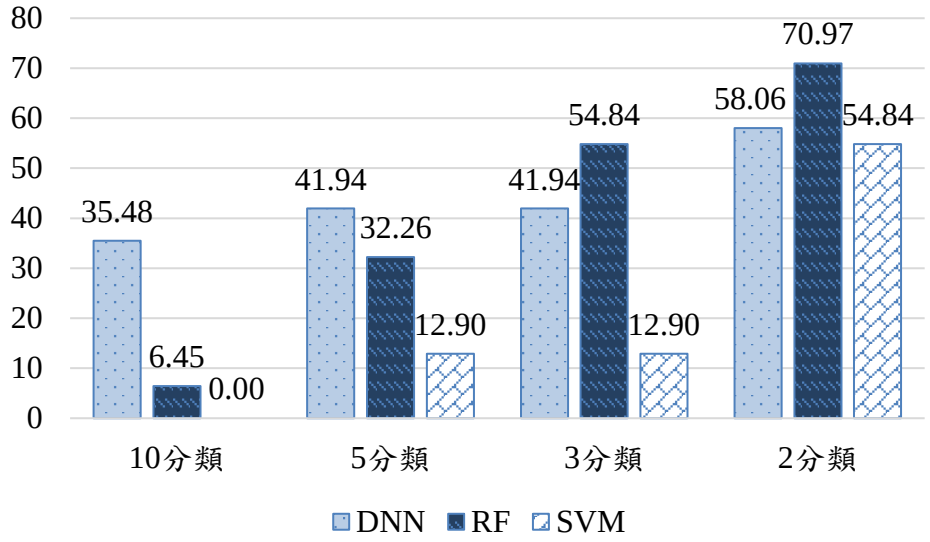


圖 32、數學科預測正確率分布圖

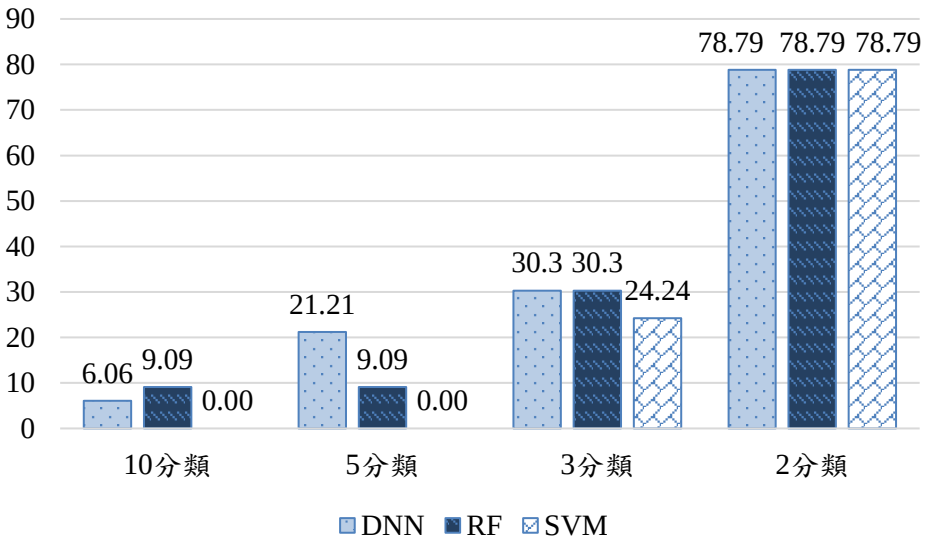


圖 33、基礎科預測正確率分布圖

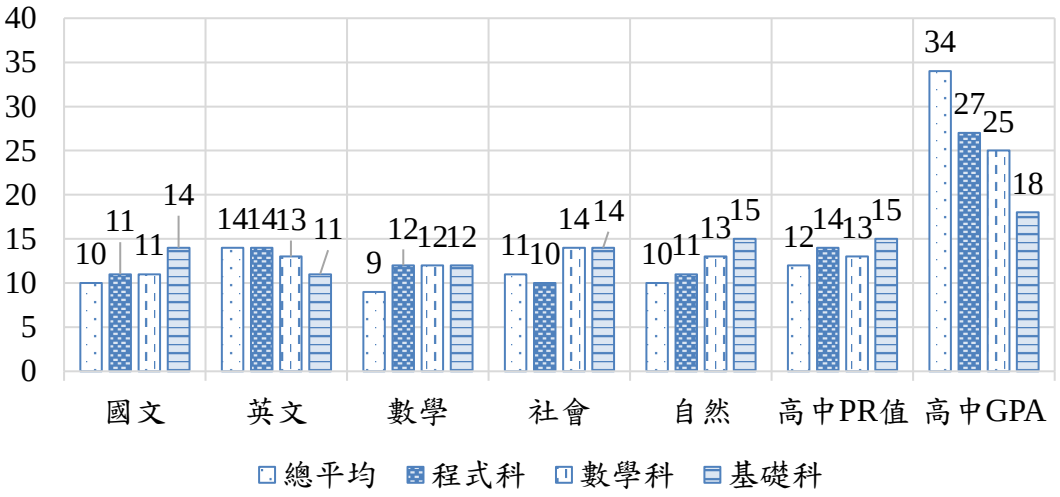


圖 34、特徵欄位重要性分布圖

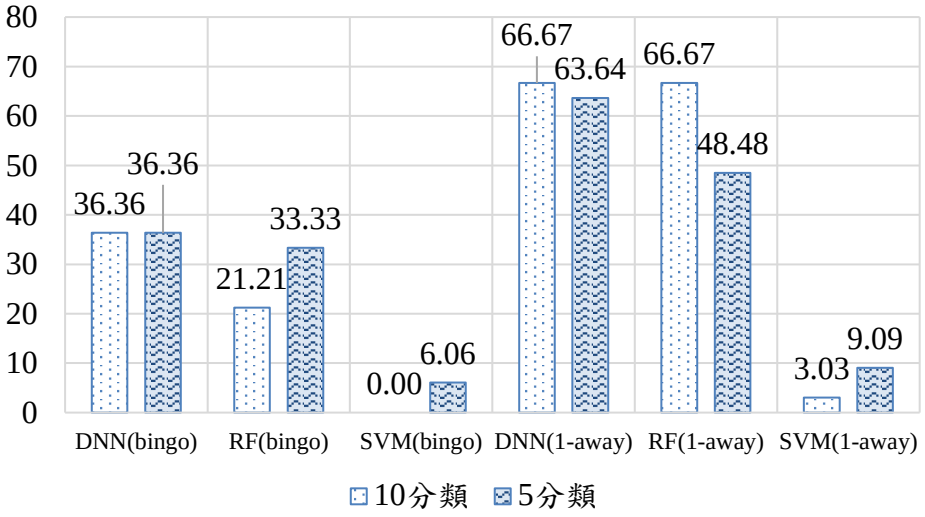


圖 35、總平均原始正確率與鄰近正確率分布圖

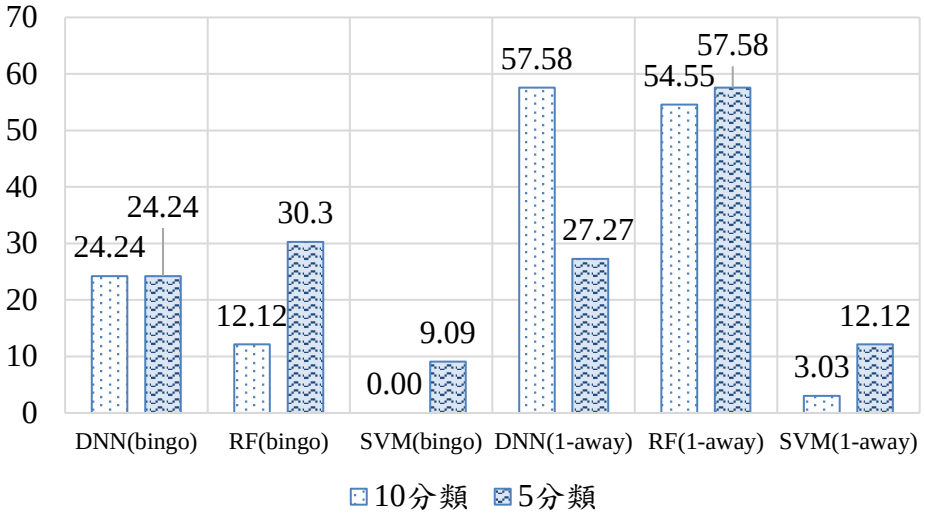


圖 36、程式科原始正確率與鄰近正確率分布圖

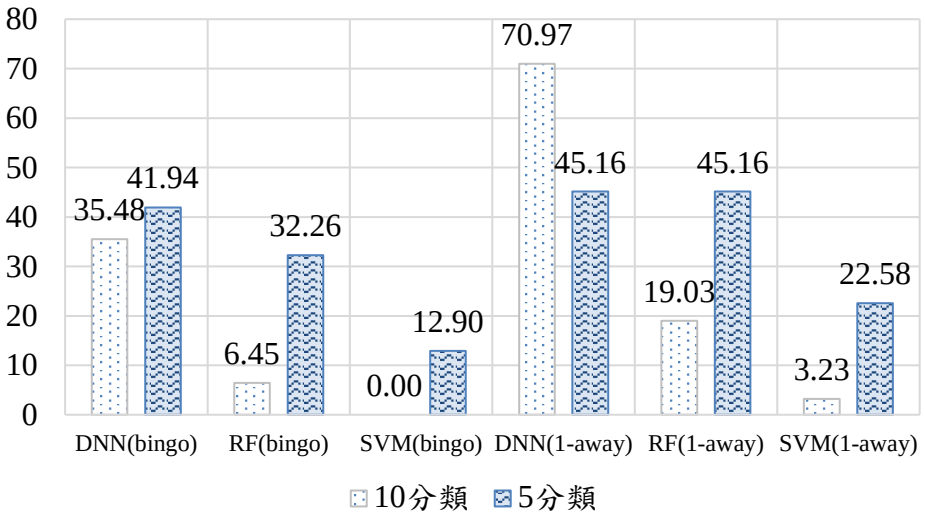


圖 37、數學科原始正確率與鄰近正確率分布圖

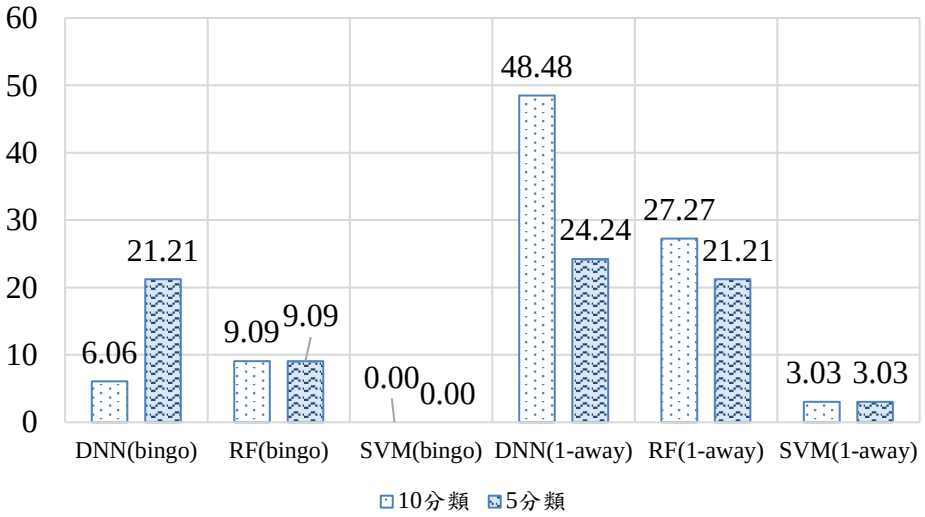


圖 38、基礎科原始正確率與鄰近正確率分布圖

表 24、總平均實驗結果排名

	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	2	3	3	2	1

表 25、程式科實驗結果排名

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

表 26、數學科實驗結果排名

	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	2	2
RF	2	2	2	1	1	1
SVM	3	3	3	3	3	3

表 27、基礎科實驗結果排名

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

綜合上述圖表後觀察出以下結論：

1. 透過將實驗結果整理成排名可以得知由深度神經網路及隨機森林表現較佳。
2. 使用鄰近正確率後可以在不失去分類意義的情況下得到更高的準確率，「10 分類」的準確率提升尤其明顯。
3. 減少資料標籤類別有助於提高準確率。
4. 輸入之特徵欄位與成績表現是有相關性存在的，尤其是高中 GPA 在各成績分類中皆表現出其重要性。

第五章 結論與建議

5.1 結論

本研究以有效的資料集，將總平均與科目類別平均成績作區間劃分，做為學業表現的預測結果分類方式。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。

5.2 研究限制

由於淡江大學資工系入學前學生資料可取得的欄位相較於其他資料欄位少了許多，且少部分的欄位為估算並不是官方釋出的數據，因此在資料的正確性上也有些疑慮。為求資料集的正確性，所能匯整進入資料集的學生資料有限，因此資料集的總數無法增加為本研究之限制，使得各項模型的訓練成果有限。

5.3 未來研究方向

建議後續研究者可從以下方面著手：

1. 入學前資料可以從備審資料下手，取得更多有參考價值的欄位，如高中每學期的學科成績或是人口統計、外部評估、行為模式等特徵。
2. 若特徵充足情況下可嘗試使用不同輸入組合以提升準確率。

參考文獻

[1] 沈慶揚, 楊憲明, 陳志福, 羅瑞玉, 莊勝義, 蘇永明, “臺灣光復四十年來教育發展之回顧,” 中華民國比較教育學會比較教育通訊, 13 卷, p.6-21, 1986.

[2] 秦夢群, “大學多元入學制度實施與改革之研究,” 教育政策論壇, 7 卷 2 期, p.59-84, 200４.

[3] S.K.Mohamad, Z.Tasir, “Educational data mining: A review,” Procedia Social and Behavioral Sciences, Vol.97, p.320-324, 2013.

[4] A.Peña-Ayala, “Educational data mining: A survey and a data miningbased analysis of recent works,” Expert Systems with Applications, Vol.41, No.4, p.1432-1462, 2014.

[5] C.Romero, S.Ventura, “Educational data mining: A review of the state of the art,” IEEE Transactions on Systems Man and Cybernetics, Part C(Applications and Reviews), Vol.40, No.6, p.601-618, 2010.

[6] H.Aldowah, H.Al-Samarraie, W.M.Fauzy, “Educational data mining and learning analytics for 21st century higher education: A review and synthesis,” Telematics and Informatics, Vol.37, p.13–49, 2019.

[7] C.Anuradha and T.Velmurugan, “A comparative analysis on the evaluation of classication algorithms in the prediction of students performance,” Indian Journal of Science and Technology, Vol.8, No.15, p.974-6846,2015.

[8] V.L.Miguéis, A.Freitas, P.J.V.Garcia, A.Silva, “Early segmentation of students according to their academic performance: A predictive modelling approach,” Decision Support Systems, p. 36-51, 2018.

[9] A.M.Shahiri, W.Husain, N.A.Rashid, “A review on predicting Student's performance using data mining techniques,” Procedia Computer Science, Vol.72, p.414-422, 2015.

[10] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, K.Lang, “Phoneme recognition: neural networks v.s. hidden Markov models,”International Conference on Acoustic, Speech,Signal Processing, p.107-110,1988.

[11] M.Mayilvaganan, D.Kalpanadevi, “Comparison of classfication techniques for predicting the performance of students academic environment,” International Conference on Communications, Computation, Networks and Technologies,Sivakasi,India, p.113-118, 2014.

[12] P.M.Arsad, N.Buniyamin, J.L.A.Manan, “A neural network students' performance prediction model(NNSPPM),” IEEE International Conference on Smart Instrumentation, Measurement and Applications(ICSIMA), Kuala Lumpur, Malaysia, p.1-5, 2013.

[13] F.Marbouti, H.A.D.Dux, K.Madhavan, “Models for early prediction of at-risk students in a course using standards-based grading,” Computers & Education, Vol.103, p.1-15, 2016.

[14] G.Gray, C.McGuinness, P.Owende, “An application of classifcation models to predict learner progression in tertiary education,” IEEE International Advance Computing Conference(IACC), p.549-554, 2014.

[15] E.N.Maltz, K.E.Murphy, M.L.Hand, “Decision support for university enrollment management:Implementation and experience,” Decision Support Systems, Vol.44, No.1, p.106-123, 2007.

[16] A.M.Hanan, “Using DataMining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems,” IEEE Access, Vol.8, p.55462–55470, 2020.

[17] https://ithelp.ithome.com.tw/articles/10253192

[18] https://cvfiasd.pixnet.net/blog/post/275774124-%E6%B7%B1%E5%BA%A6%E5%AD%B8%E7%BF%92%E6%BF%80%E5%8B%B5%E5%87%BD%E6%95%B8%E4%BB%8B%E7%B4%B9

[19] T.K.Ho, “Random decision forests,” Proceedings of 3rd International Conference on Document Analysis and Recognition, p.14-16, 1995.

[20] T.K.Ho, “The random subspace method for constructing decision forests,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.20, Issue 8, p.832-844, 1998.

[21] L.Breiman, “Random Forests,” Machine learning, Vol.1, Issue 45, p.5-32, 2001.

[22] H.Guruler, A.Istanbullu, M.Karahasan, “A new student performance analysing system using knowledge discovery in higher educational databases,” Computers & Education, Vol.55, No.1, p.247-254, 2010.

[23] S.Natek, M.Zwilling, “Student data mining solution–knowledge management system related to higher education institutions,” Expert Systems with Applications, Vol.41, p.6400–6407, 2014.

[24] S.Fong, R.Biuk-Aghai, “An automated university admission recommender system for secondary school students,” The 6th International Conference on Information Technology and Applications, 2009.

[25] https://towardsdatascience.com/an-introduction-to-decision-trees-with-python-and-scikit-learn-1a5ba6fc204f

[26] 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

[27] V.N.Vapnik, “The Nature of Statistical Learning Theory. Springer,” New York. for medical diagnosis-application to congenital heart disease. Journal of the American Medical Association.

[28] B.E.Boser, I.M.Guyon, C.Vapnik, “V.N.A training algorithm for optimal margin classifiers.,” Proceedings of the fifth annual workshop on Computational learning theory, COLT.92, p.144, 1992.

[29] C.Cortes, V.Vapnik, “Support-vector networks.,” Machine Learning, Vol.20, p.273-297, 1995.

[30] http://bytesizebio.net/2014/02/05/support-vector-machines-explained-well

[31] https://www.lexico.com/definition/overfitting

[32] A.Ainslie, X.Drèze, F.Zufryden, “Modeling Movie Life Cycles and Market Share,” Marketing Science, Vol.24, Issue 3, p.508-517, 2005.

[33] Q.Zhou, Y.Zheng, C.Mou, “Predicting students' performance of an offline course from their online behaviors,” Fifth International Conference on Digital Information and Communication Technology and its Applications (DICTAP), 2015.

[34] S.Roy, A.Garg, “Predicting academic performance of student using classification techniques,” 4th IEEE Uttar Pradesh Section International Conference on Electrical, Computer and Electronics, UPCON, p.568–572, 2017.

[35] T.Devasia, T.P.Vinushree, V.Hegde, “Prediction of students performance using Educational Data Mining,” International Conference on Data Mining and Advanced Computing (SAPIENCE), p.91–95, 2016.

[36] E.Irfiani, I.Elyana, F.Indriyani, F.E.Schaduw, D.D.Harmoko, “Predicting Grade Promotion Using Decision Tree and Naïve Bayes Classification Algorithms,” Third International Conferenceon Informatics and Computing(ICIC), 2018.

[37] G.E.Hinton, S.Osindero, Y.W.Teh, “A Fast Learning Algorithm for Deep Bwlief Nets,” Neural Computation, Vol.18, No.7, 2006.

[38] N.Srivastava, G.Hinton, A.Krizhevsky, I.Sutskever, R.Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” The Journal of Machine Learning Research, Vol.15, Issue 1, p.1929-1958, 2014.

附錄一　英文論文

Using Machine Learning Techniques to Find the Relationship between University Admission Score and Student Performance－Taking Department of Computer Science and Information Engineering of Tamkang University as an example

Shu-Shiuan Rong
Department of Computer Science and Information Engineering
Tamkang University
Taipei, Taiwan
rss@gms.tku.edu.tw

Abstract: With the impact of the declining birthrate, all higher education institutions are facing enrollment challenges. Collecting useful data can find appropriate students for university admission. In this study, we adopt machine learning techniques like Deep Neural Network (DNN), Random Forest (RF), and Support Vector Machine (SVM) to find the relationship between entrance score and freshman academic performance. Experimental results show that the SVM has the best prediction results on average of total courses in 2-category classification. The RF performs best on mathematical courses and programming courses in 2-category classification. However, on fundamental courses, all methods have the best prediction results in 2-category and perform equally well. Moreover, on other category classifications, like 3-category, 5-category, or 10-category classifications, we cannot find a universal best method. Another result shows that High school GPA has a significant impact on results.

Keywords**:** Machine Learning, Student Performance, Performance Prediction, Deep Neural Networks, Random Forest, Support Vector Machine.

I. INTRODUCTION

With the reform of education, the General Scholastic Ability Test (GSAT) has become the first major test that most students are exposed to. And two of the three existing admission channels use GSAT score as a criterion. With the impact of the Sub-replacement fertility, the school pays special attention to student abilities in addition to performance. How to choose suitable students to study has also become a major issue for the school.

This work aims to use machine learning models to predict students' academic performance in the freshman year. Our goal is to compare the prediction results of different labeling methods through different machine learning models.

II. LITERATURE REVIEW

In the study proposed by Shahiri et al. in 2015 [1] pointed out that the cumulative grade point average (CGPA) is the most

influential feature field, followed by more people using demographic statistics (for example: gender, age, family background, disability) And external evaluation (for example: final exam results for special subjects), and finally behavioral patterns (for example: extracurricular activities, social networking). Shahiri also counted the accuracy of various forecasting methods in the past forecasting research. Among them, the most accurate is the neural network, followed by the decision tree, and then the support vector machine.

1. Deep Neural Network (DNN)

Deep learning is a branch of machine learning. It was proposed by Hinton et al. [2] as early as 2006. The concept is to simulate the neural network of the brain for learning by superimposing multiple hidden layers. In many published studies [3], this method has been used as a method to predict student performance. In the study proposed by Hanan in 2020[8] used three admission scores to predict the cumulative grade point average (CGPA) of students in the first year. In the study, the Artificial Neural Networks (ANN) model was used, and the results showed that the prediction was 79.22% accurate. Accuracy rate. The schematic diagram of the model structure and operation process are as shown in Figures 1 and 2.

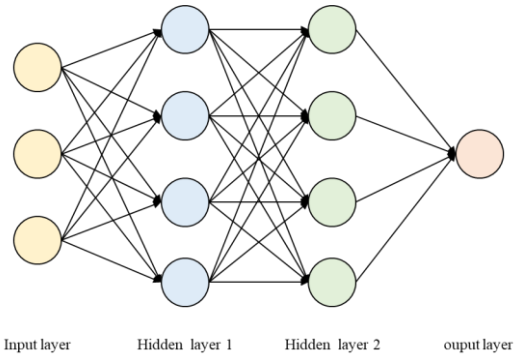


Fig. 1. DNN structure

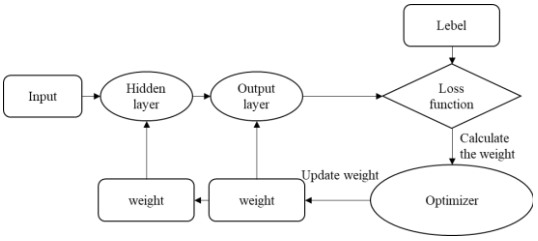


Fig. 2. DNN process

2. Random Forest (RF)

Random forest was first proposed by T.K.Ho of Bell Labs in 1995 [9], and later L. Breiman and Cutler developed and deduced the algorithm of random forest [11]. The basic principle is to combine multiple Classification and Regression Trees (CART), use GINI impurity decision trees, and add randomly allocated training data to greatly improve the final calculation results. It has been widely used in various research related to performance prediction [12]. The study proposed by Anuradha et al. in 2015 used student demographic data and external evaluations to predict the end semester mark (ESM) of students. The decision tree model was used in the research, and the results showed that the accuracy of the prediction was 72.51%. The schematic diagram of the model structure and operation process is as shown in Figure 3.

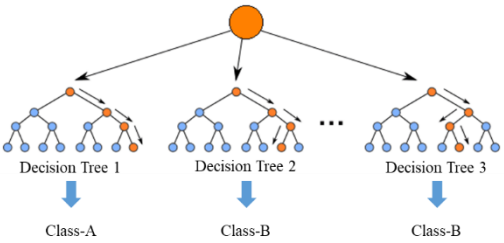


Fig. 3. RF structure

RF summarizes the predictions of all decision trees, and decides the classification results by majority decision. Take Figure 3 as an example, the final classification result is Class-B.

3. Support Vector Machine (SVM)

Vapnik et al. invented the Vapnik–Chervonenkis theory in the 1960s [17]. In 1992, Boser et al. proposed a method of

building a nonlinear classifier by applying the kernel technique to the hyperplane of the maximum interval [18]. The predecessor of the current standard (soft interval) was proposed in 1993 by Cortes and Vapnik, and in 1995 Published [19]. It has been widely used in studies related to learning performance prediction[5][13]. Among them, the research proposed by Miguéis[13] used the results of the first two semesters of university and some demographic data (such as gender, marital status, etc.) to predict academic performance after four years. The prediction result of its SVM model reached 92.6%. The schematic diagram of the model structure and operation process is as shown in Figure 4.

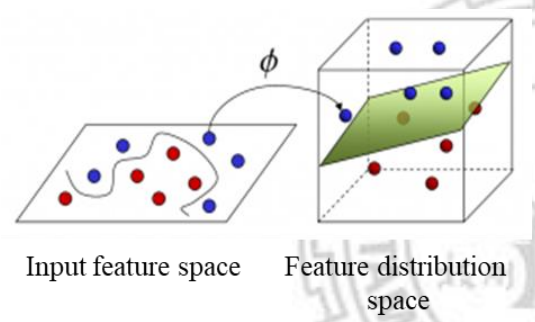


Fig. 4. SVM structure

SVM maps the linearly inseparable samples in the low-dimension space to the high-dimensional space through kernel transformation, and finds a hyperplane to effectively cut these samples (as shown in Figure 4). The distance between the samples on both sides of the hyperplane and the hyperplane itself will determine How good is the model training.

III. DATA PROCESS AND METHODOLOGY

1. Data pre-processing

This work uses data from Center for Institutional Research of Tamkang University and Department of Computer Science and Information Engineering of Tamkang University. Total 328 records. Details in Table 1.

Table 1. Feature and label list

Column name	Type	Pre-processing	Missing
GSAT_Chinese	feature	✓	0
GSAT_English	feature	✓	0
GSAT_Math	feature	✓	0
GSAT_Society	feature	✓	0
GSAT_Science	feature	✓	0
H_PR	feature	✓	0
H_GPA	feature	×	0
Average	label	✓	0
Programming	label	✓	1
Mathematics	label	✓	18
Foundation	label	✓	1

The detail of each feature and label are distributed in the following.

F1. GSAT Scores: GSAT scores are all expressed in grades, which are converted into a unified standard through the percentage of the cumulative number of people.

F2. H_PR: H_PR is converted from the school name. Some private schools can only be collected from the Internet, so that the value is less reliable.

F3. Average: Select the professional subjects according to the list provided by the department office, and then average the professional subjects.

F4. Professional subjects: According to the course classification provided by the department office, the professional subjects are divided into three categories: Programming, Mathematics, and Foundation, and then averaged individually.

2. Labeling methods

The labeling methods are 10-category, 5-category, 3-category, and 2-category, as shown in Figure 5.

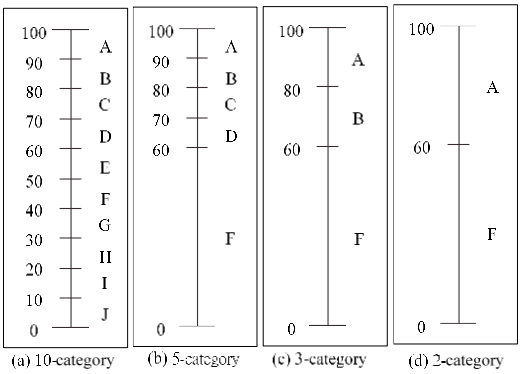


Fig. 5. Labeling methods

3. Model settings

1) Deep Neural Network (DNN)

Figure 6 shows the DNN structure of the experiment. To prevent over-fitting, set the

dropout to 0.4. The optimization function uses SGD, the loss function uses categorical_crossentropy, the activation function for each hidden layer uses ReLU, and the activation function of the output layer is softmax.

```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense (Dense)                (None, 10)                80
_________________________________________________________________
dropout (Dropout)            (None, 10)                0
_________________________________________________________________
dense_4 (Dense)              (None, 10)                110
_________________________________________________________________
dropout_4 (Dropout)          (None, 10)                0
_________________________________________________________________
dense_8 (Dense)              (None, 9)                 99
=================================================================
Total params: 289
Trainable params: 289
Non-trainable params: 0
```

Fig. 6. The summary of DNN structure

2) Random Forest(RF)

n_estimators is mainly used to limit the number of decision trees and avoid over-fitting. In this experiment, set this parameter to 5. Set the Criterion to "gini".

3) Support Vector Machine (SVM)

The kernel function mainly assists in mapping features to high dimensions, In this experiment, set this parameter to "poly" to mean linear segmentation in high dimensions. The parameter degree is only related to poly, set to 3. The parameter decision_function_shape set to "ovr", which means that the n-classification problem is treated as multiple binary classification problems, and each binary problem is divided into two categories: "one" and "rest".

4. Evaluation

In the experiment, we apply two accuracy rates. The first accuracy rate is the ratio of the full answer rate (bingo), the prediction result, and the test set answer, the "completely consistent" result to the total number of samples in the test set. The second accuracy rate is 1-away, which is the prediction result and the test set answer. The "matching before and after" means "the ratio of the test set answer, plus one and minus one" and the ratio of the test sample to the total number of samples in the test set.

IV. RESULT & DISCUSSIONS

The evaluation of the experimental results is illustrated by the standard values of four items: Average, Programming, Mathematics, and Foundation. The four labeling methods are presented separately in the value of each item. Finally, the importance of features is shown.

1. Average:

Obtain the accuracy through the evaluate function then organize accuracy and the 1-away accuracy into Figure 7.

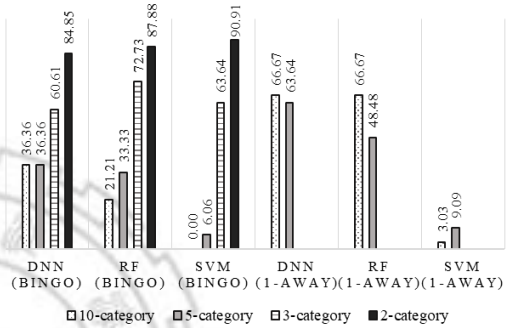


Fig. 7. Result of Average

2. Programming:

Obtain the accuracy through the evaluate function then organize accuracy and the 1-away accuracy into Fig.8.

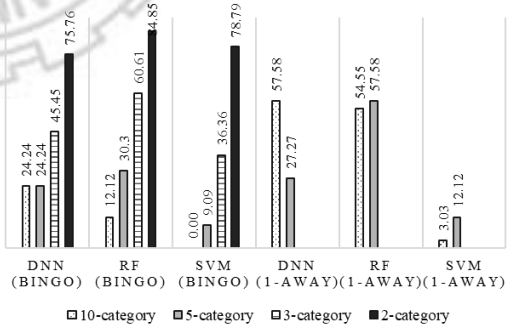


Fig. 8. Result of Programming

3. Mathematics:

Obtain the accuracy through the evaluate function then organize accuracy and the 1-away accuracy into Figure 9.

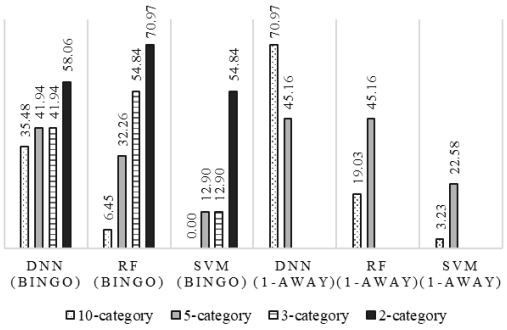


Fig. 9. Result of Mathematics

4. Foundation:

Obtain the accuracy through the evaluate function then organize accuracy and the 1-Away accuracy into Figure 10.

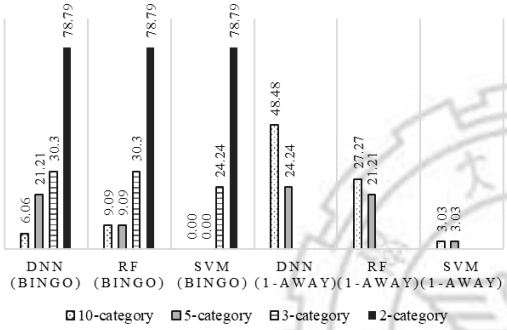


Fig. 10. Result of Foundation

5. Feature Importance in Random Forest:

Obtain the Feature Importance through the feature_importances_ function in RF model. shown in Figure 11.

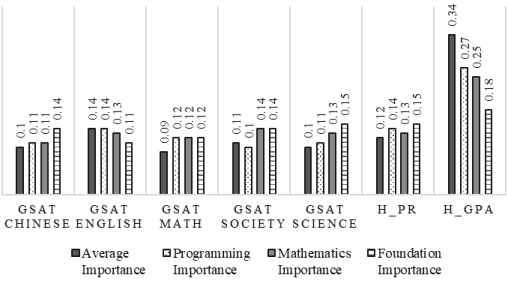


Fig. 11. Result of Foundation

According to the prediction of performance rankings add up organized into Figure 12.

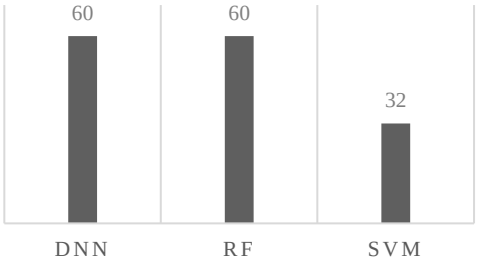


Fig. 12. Prediction of performance rankings add up

Based on the above chart, the following results are obtained:

1) By organizing the experimental results into rankings, we can see that the DNN and RF perform better.
2) After using the 1-Away accuracy, a higher accuracy rate can be obtained without losing the meaning of labeling method. The 1-Away accuracy of "10 cat." is particularly improved.
3) Reducing the label category helps to improve accuracy.
4) There is a correlation between the input feature and performance, especially the high school GPA shows its importance in each performance classification.

V. CONCLUSIONS

We use different labeling methods to predict the results through different models when the features and the number of samples are limited, and obtains a good prediction performance by appropriate use of accuracy evaluation criteria.

Follow-up recommendations from the researchers can begin has aspects:

1) Gather more valuable features, such as student's demographic, external assessments, and personal interest and behavior.
2) If the features are sufficient, also can try different input combinations to improve accuracy.

VI. REFERENCES

[1] A.M.Shahiri, W.Husain, N.A.Rashid, "A review on predicting Student's

performance using data mining techniques," Procedia Computer Science, Vol.72, p.414-422, 2015.

[2] A.Waibel, T.Hanazawa, G.Hinton, K.Shikano, K.Lang, "Phoneme recognition: neural networks v.s. hidden Markov models,"International Conference on Acoustic, Speech,Signal Processing, p.107-110,1988.

[3] M.Mayilvaganan, D.Kalpanadevi, "Comparison of classfication techniques for predicting the performance of students' academic environment," International Conference on Communications, Computation, Networks and Technologies, Sivakasi, India, p.113-118, 2014.

[4] P.M.Arsad, N.Buniyamin, J.L.A.Manan, "A neural network students' performance prediction model(NNSPPM)," IEEE International Conference on Smart Instrumentation, Measurement and Applications(ICSIMA), Kuala Lumpur, Malaysia, p.1-5, 2013.

[5] F.Marbouti, H.A.D.Dux, K.Madhavan, "Models for early prediction of at-risk students in a course using standards-based grading," Computers & Education, Vol.103, p.1-15, 2016.

[6] G.Gray, C.McGuinness, P.Owende, "An application of classifcation models to predict learner progression in tertiary education," IEEE International Advance Computing Conference(IACC), p.549-554, 2014.

[7] E.N.Maltz, K.E.Murphy, M.L.Hand, "Decision support for university enrollment management: Implementation and experience," Decision Support Systems, Vol.44, No.1, p.106-123, 2007.

[8] A.M.Hanan, "Using DataMining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems," IEEE Access, Vol.8, p.55462–55470, 2020.

[9] T.K.Ho, "Random decision forests," Proceedings of 3rd International Conference on Document Analysis and Recognition, p.14-16, 1995.

[10] T.K.Ho, "The random subspace method for constructing decision forests," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.20, Issue 8, p.832-844, 1998.

[11] L.Breiman, "Random Forests," Machine learning, Vol.1, Issue 45, p.5-32, 2001.

[12] C.Anuradha and T.Velmurugan, "A comparative analysis on the evaluation of classication algorithms in the prediction of students performance," Indian Journal of Science and Technology, Vol.8, No.15, p.974-6846,2015.

[13] V.L.Miguéis, A.Freitas, P.J.V.Garcia, A.Silva, "Early segmentation of students according to their academic performance: A predictive modelling approach," Decision Support Systems, p. 36-51, 2018.

[14] H.Guruler, A.Istanbullu, M.Karahasan, "A new student performance analysing system using knowledge discovery in higher educational databases," Computers & Education, Vol.55, No.1, p.247-254, 2010.

[15] S.Natek, M.Zwilling, "Student data mining solution–knowledge management system related to higher education institutions," Expert Systems with Applications, Vol.41, p.6400–6407, 2014.

[16] S.Fong, R.Biuk-Aghai, "An automated university admission recommender system for secondary school students," The 6th International Conference on Information Technology and Applications, 2009.

[17] V.N.Vapnik, "The Nature of Statistical Learning Theory. Springer," New York. for medical diagnosis-application to congenital heart disease. Journal of the American Medical Association.

[18] B.E.Boser, I.M.Guyon, C.Vapnik, "V.N.A training algorithm for optimal margin classifiers.," Proceedings of the fifth annual workshop on Computational learning theory, COLT.92, p.144, 1992.

[19] C.Cortes, V.Vapnik, "Support-vector networks.," Machine Learning, Vol.20, p.273-297, 1995.