

高教深耕計畫教學活動記錄表

1. 授課教師姓名：葉倚任	職稱：副教授	單位：數學系
2. 課程名稱：機器學習	開課年級：三年級	☐ 必修 ☒ 選修
3. 任教學期： ☒ 111 學年度第二學期 ☐ 112 學年度第一學期		
4. 課程類別： ☒ 問題導向的教學 ☐ 探究導向的教學 ☐ 開發新的教學方法及評量工具		
5. 修課人數：15		

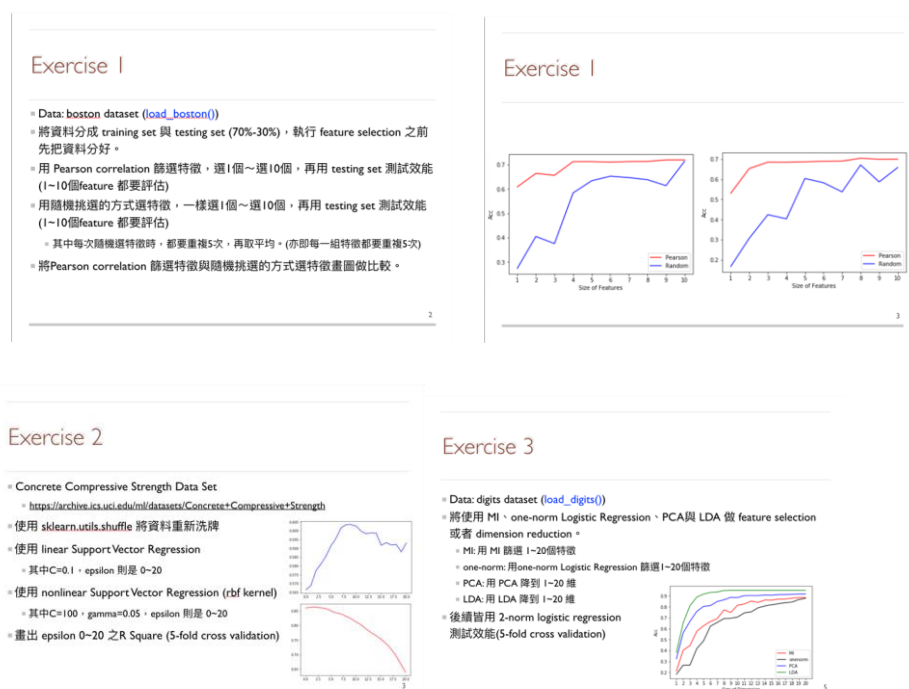
6. 教學概述及成效

(1) 教學目標

此課程除了機器學習的基礎觀念的教授，另一個目標就是以透過資料科學競賽以實戰養成資料分析之技術，然而以實戰養成技術是有前提的，整個教學項目先是 Python 程式設計，接著才是資料分析(機器學習)。主要就是讓學生學習到程式實作能力與資料分析流程掌握度，另外 Python 目前也是很受歡迎的程式語言之一，所以讓學生接觸 Python 除了目前機器學習領域的相關套件支援較多之外，Python 也適用於許多應用，是一個可以廣泛應用的語言，也是讓學生更接近業界所需。

(2) 教學過程

除了講述教學法之外，也搭配問題導向學習法。在學期中，基礎概念會持續利用講述教學法來教授。而在團體合作學習法中，課程會將學生分成四人一組，然後會固定有雙週練習，也就是每兩個禮拜在課程上會有四題題目給學生練習，如圖一。學生需要自己討論由誰實作哪一題，過程中可以互相討論也可以上網查詢，這種團體練習主要是模擬在業界實作的場景：團體合作與資訊尋找能力。這種方式在過去已經實施多年，也發覺學生在學習上也比較認真，也會互相討論。



圖一：資料分析課程雙週練習範例

最後問題導向學習的部分，則是在學期一開始給予學生可能會有興趣且需要課堂所學的知識才能完成的任務，主要的目的是想讓學生課堂之後，還有一個可以繼續練習的目標，以 Python 來說這學期

則是給予實作 1D RPG 之遊戲。



圖四：Python 課程實作任務截圖

而在資料分析的課程，問題導向學習的部分則是直接採用 Kaggle 競賽題目。

(3) 評量方式

在 Python 程式設計方面，除了雙週練習的評量之外，會有一次上機測量。其中分為個人得分題與團體得分題，其中團體得分題一樣是自己做題目，但是得分方式為看整隊答題率。例如某一題為 10，隊員數為四人，答對的隊員有 3 個，則此題四人的分數為 $10 \times 3 / 4 = 7.5$ 分。這樣設計主要的原因為要利用團體組隊的方式來讓小組間的協助更加顯著。另外，課堂一開始的 side project 則是提供學生加分的途徑。而在資料分析課程方面，一樣有雙週評量，但是主要的考核為期末競賽成果或者產學合作資料分析成果。其中期末專題將於最後四周開始上台做進度報告，主要讓學生能夠花更多時間了解資料並互相討論，而不是最後幾天才開始進行資料分析。

(4) 學生學習成效

期末資料分析競賽主要採用下面 4 個資料集，主要都是來自 Kaggle 網站上的資料：

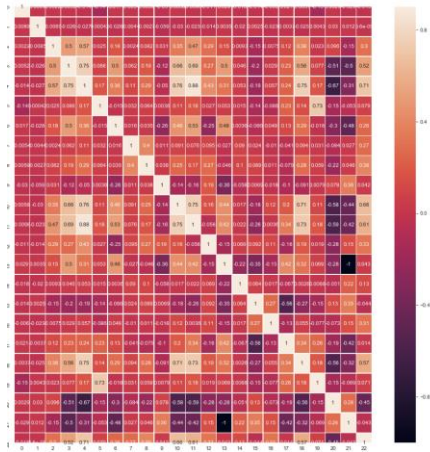
Home Sales Task1: Report the absolute error by using all features Task2: Report the absolute error by using less than 5 features Training set size: 17291 Testing set size: 4322 Regression: predict the price of a house	Gender Recognition by Voice Task1: Report the accuracy by using all features Task2: Report the accuracy by using less than 5 features Training set size: 2536 Testing set size: 632 Class(or label): Male (1) or Female (0)
物流運送類型預測 Training set size: 388,232 Testing set size: 258,822 Feature size: 6 Class (Trip Type): a categorical id representing the type of shopping trip the customer made. This is the ground truth that you are predicting. TripType_999 is an "other" category.	Text Classification: Medline dataset Task: Classification in a high dimensional space Training set size: 1250 Testing set size: 1250 Feature size: 22095

每週競賽成績如下：

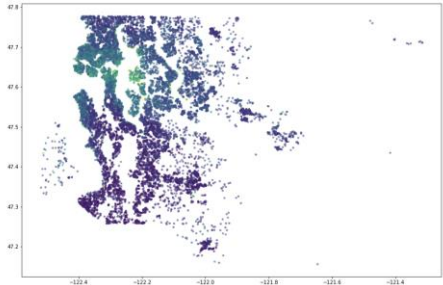
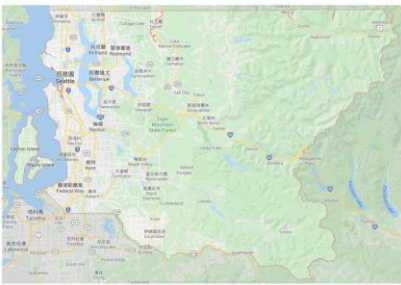
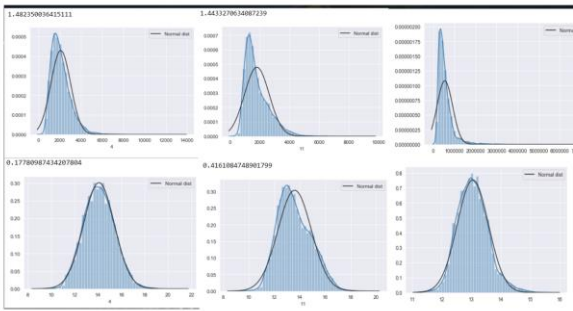
week1	voice_task		House_task		Trip Type		Meddline
	陳韋宏	0.9841	王怡鑫	57821.03	王怡鑫	0.5097	
	王怡鑫	0.9825	林緯捷	60424.11	林緯捷	0.3821	
	林緯捷	0.9810	林亦勳	64716.93	林亦勳	0.3631	
	林亦勳	0.9810	林海民	68303.01	陳韋宏	0.3525	
	紀昆宏	0.9778	紀昆宏	64728.17	紀昆宏	0.3161	
week2	voice_task		House_task		Trip Type		Meddline
	陳韋宏	0.9841	王怡鑫	57821.03	王怡鑫	0.5507	林緯捷 0.9560
	王怡鑫	0.9825	林緯捷	59342.78	陳韋宏	0.5262	陳韋宏 0.9520
	林亦勳	0.9810	林亦勳	60117.09	林亦勳	0.4966	王怡鑫 0.9320
	紀昆宏	0.9810	陳韋宏	63972.91	林海民	0.3427	紀昆宏 0.5176
	林緯捷	0.9794	紀昆宏	64784.61	林緯捷	0.2866	林亦勳 0.5128
week3	voice_task		House_task		Trip Type		Meddline
	陳韋宏	0.9841	王怡鑫	56389.09	王怡鑫	0.5521	陳韋宏 0.9768
	王怡鑫	0.9825	林緯捷	58783.72	林亦勳	0.5395	王怡鑫 0.9696
	林亦勳	0.981012	林亦勳	59844.68	陳韋宏	0.5262	紀昆宏 0.9624
	紀昆宏	0.981012	陳韋宏	61372.74	林緯捷	0.5175	林緯捷 0.9616
	林海民	0.9778	紀昆宏	63748.06	紀昆宏	0.3464	林亦勳 0.9600
	林緯捷	0.9762	林海民	65656.76	林海民	0.3421	李純靜 0.9504

學生 EDA 成果

相關係數

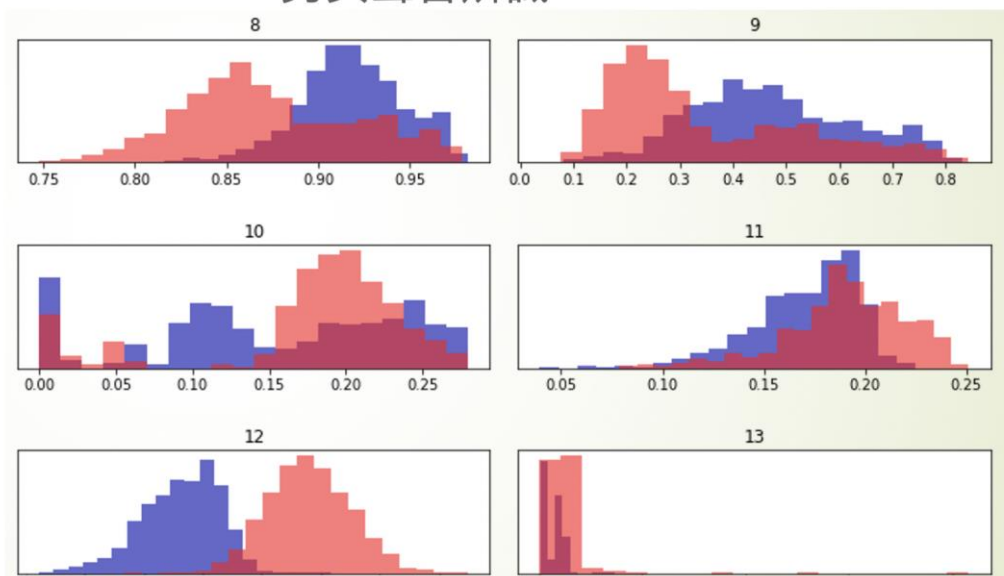


去偏態



- 推測位置為北美洲西邊(華盛頓州)
 - 距離接近
- idea
- ➡ 同一個地區房子價格差距大致上不大

男女聲音辨識EDA



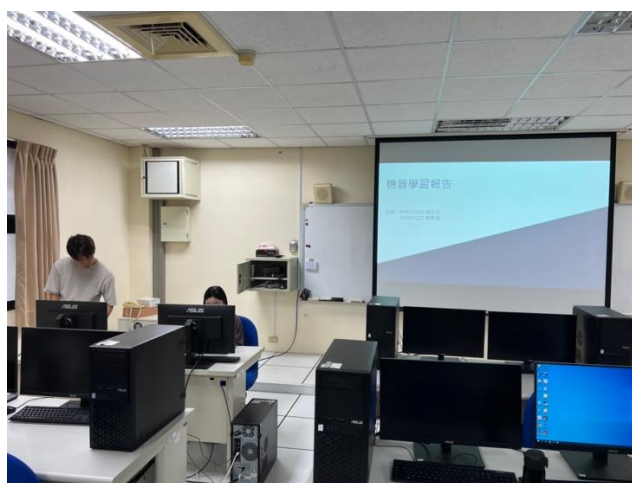
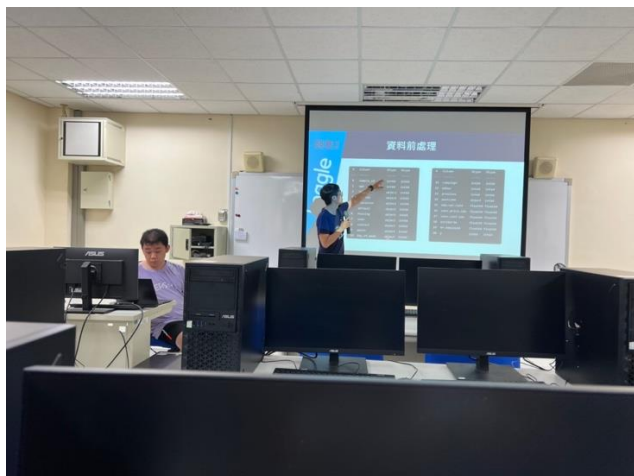
7. 教學反思和評估

[專題形式 vs. 競賽形式] 去年採用專題形式，方式為每組鎖定一個問題，深入分析，但是發覺效果較不好，由於沒有互相比較，所以進度上會比較拖。今年則採用競賽形式，方式為直接對四個資料集做競賽，特地混合容易分析與較難分析之資料集，這樣讓學生成就感較大。另外，透過互相比較，可以讓學生動機會比較強烈一點 每週都會互相分享做法，所以技巧上可以互通有無，讓學生可以進步比較快。

8. 教學照片



團體雙週練習時，是可以一起討論並上網尋找相關的技術與資料處理方法



期末競賽的報告，期末競賽主要的方向是讓學生參與 Kaggle 線上競賽(<https://www.kaggle.com>)，也就是在學完基礎技巧之後，直接讓學生面對資料，透過直接接觸資料，學生才能夠更深刻了解機器學習的流程，進而養成資料科學的技術。