[NTU ESOE] 109 Data Mining Midterm Project Deadline: 2020.11.17.23:00 (可補交)

Student ID: _r10525069_ Name: __林子傑__ Department: __工程海洋學系碩士班___

ProjectType: __clustering__

*please check your project type in xxxxxxxx.xlsx

報告影片：https://drive.google.com/file/d/14kSlfdxp-CQCHWIFc_K4wfSNXX1h1-hx/view?usp=sharing

1. 使用指定資料集以及 **inanalysis** 進行資料分析，並完成機器學習畫布 **(50%)** (機器學習畫布請使用 MLC.docx 為模板修改)

2. 請詳述你對資料作的前處理，以及調整模型參數的過程邏輯。可參考 **data preview** 以及 **model preview** 的結果來描述。必須超過 **350** 字。 **(35%)**

資料前處理：ID 或是有固定選項的資料不做處理，孩子和寵物數量雖然沒有固定選項，但分布範圍小且資料相對簡單，也不做處理，因此只對其他沒有固定選項且分布範圍廣的資料做 Normalize 和 Flitering Missing Value。

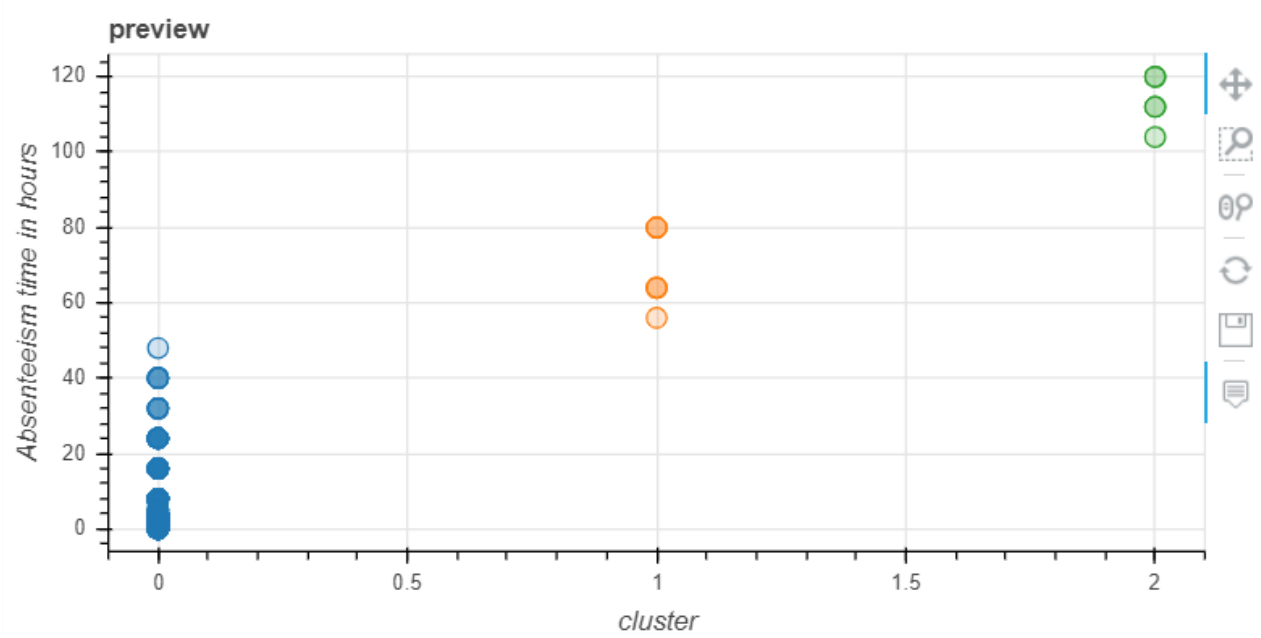
資料選擇：跟缺席時間有關的是缺席原因(不同種類原因所影響缺席的時間程度不同)、通勤狀況(缺席時間包含解決問題後的通勤時間)、健康狀態(如果缺席原因和健康有關，健康狀態會影響缺席時間)、工作表現(會影響到心理健康)，因此把無關的 ID、日期、教育程度、孩子和寵物數量排除。

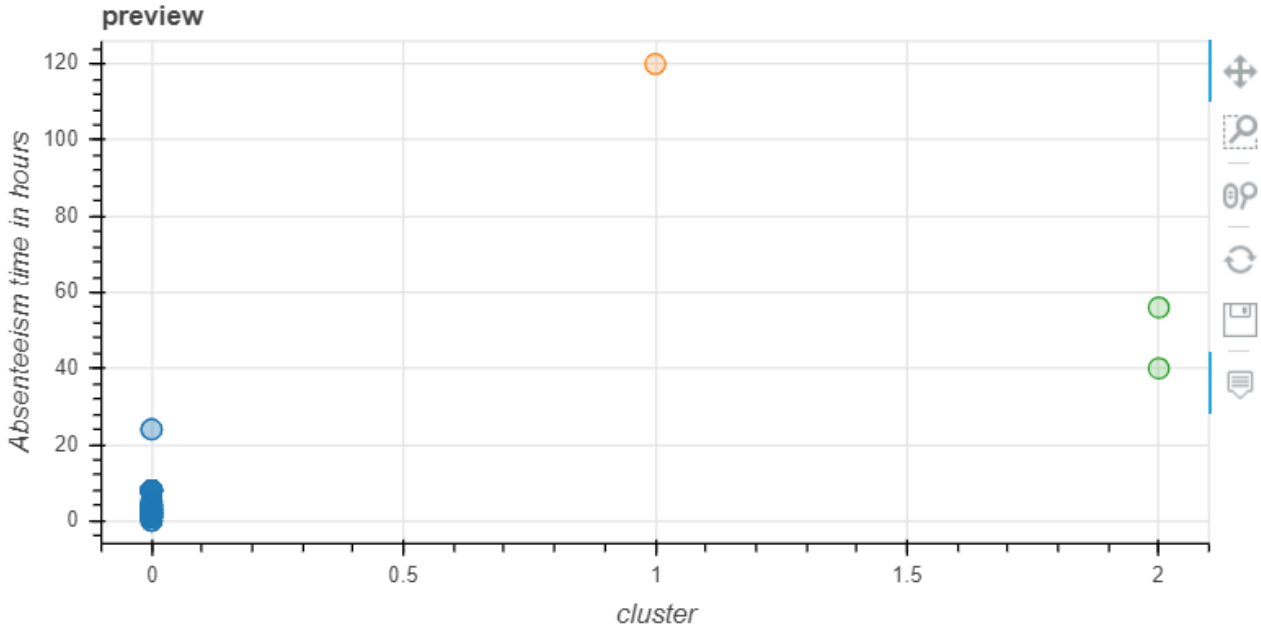
調整模型參數：

首先使用 Kmeans，3 clusters，max_iter=300，Correlation algorithm 為 Pearson Corrleation coefficient，結果三個類別有二個類別的 Absenteeism time in hours 區間是重疊的，因此將 cluster 數量改為 2，結果 2 個分類的 Absenteeism time in hours 完全分開，雖然做到這裡成功地把每個分類成功分開，但是我我目標是把資料分成三個(含)以上分類。

於是我改用 Agglomerative clustering，3 clusters，linkage=ward，affinity=euclidean，這樣分成三類也會有兩個類別的 Absenteeism time in hours 完全分開，後來將 linkage 改成 complete，跑出的結果成果將 3 個分類的 Absenteeism time in hours 完全分開，成功訓練出模型。

3. 你最終的訓練以及測試結果為何**?** 請有邏輯的解釋你的結果 **(15%)**





第一張圖是 Training 的結果，第二張圖是 Testing 的結果，兩種資料都能將資料的 Absenteeism time in hours 分成三種分類，兩張圖的差異在於 Absenteeism time in hours 在 40 左右時，Training 分在最低的一類，Testing 分在中間那一類，推論是 Absenteeism time in hours 在 40 左右的資料較少，因此在判斷上沒辦法明確地分到特定類別。