機器語言與深度學習期末報告

主題：Students performance in exams

組員：1044A027 林玳萱、N1046451 陳鏘諼

(原先準備參加 kaggle 比賽,但此資料集沒有在比賽內,因此有與老師商量組員皆交同一份檔案)

摘要

將 Kaggle 題目 Students performance in exams 做分類方法及分析，找出各 feature 與研究結果之間的關聯性。

介紹(研究背景及研究目的)

依照學生性別、父母教育程度、午餐餐費以及考試成績，推斷該位學生是否準備課程。

資料集介紹.資料集來源

Marks secured by the students in college Students from England maybe.

資料集來源：Kaggle

資料預處理

1. 原資料集的種族 feature 以代號表示，因此本組無法判別該種族為何，故此次研究排除該 feature。
2. 在 Excel 中額外新增 pedu 欄位將原先 parental level of education(父母教育程度)中 some high school 取代成 high school， Some college、bachelor's degree、associate's degree 取代成 college，master's degree 取代成 master。取代完成後，再將 parental level of education 這項欄位刪除。
3. 將性別、午餐餐費、是否準備課程欄位之內容，分別轉換成 0 或 1 表示，並將轉換後的結果 merge 至原先的表格裡。
4. 將學歷欄位中的 high school、college、master 分別轉換成代碼 1、2、3。

機器學習或深度學習方法（使用何種方法）

使用整體學習(單純貝式分析、決策樹、邏輯迴歸、KNN、SVC、投票法、裝袋法、隨機森林、AdaBoost、Stacking、XGBoost)

研究結果及討論（含模型評估與改善）

依照各模型結果顯示之準確度

Y 值設為 free/reduced(為清寒家境)：

1. 單純貝式分析：67%
2. 決策樹：69%
3. 邏輯迴歸：69%
4. KNN：59%
5. SVC：70%

6. 投票法：67%
7. 裝袋法：61%
8. 隨機森林：60%
9. AdaBoost：69%
10. Stacking：60%-70%
11. XGBoost：68%

Y 值設為 completed(有準備考試課程)：

1. 單純貝式分析：69%
2. 決策樹：63%
3. 邏輯迴歸：77%
4. KNN：63%
5. SVC：67%
6. 投票法：68%
7. 裝袋法：69%
8. 隨機森林：67%
9. AdaBoost：71%
10. Stacking：60%-70%
11. XGBoost：73%

結論

本組使用兩個不同的 Y 值進行準確度比較,發現以有無準備考試與其他欄位的關聯性較高,所以本組最後 Y 值採用判斷是否準備課程。

參考文獻

1. Kaggle 網站
   https：//www.kaggle.com/spscientist/students-performance-in-exams
2. 許晉龍老師所提供的教材