數據分析與機器學習期末報告

題目:決策樹預測學生購買平板並載入資料庫

1. 簡介

透過決策樹從訓練資料學習，預測測試集中的資料中，學生是否會買平板，並且將預測會購買平板的學生聯絡資料載入資料庫

1. 原始數據

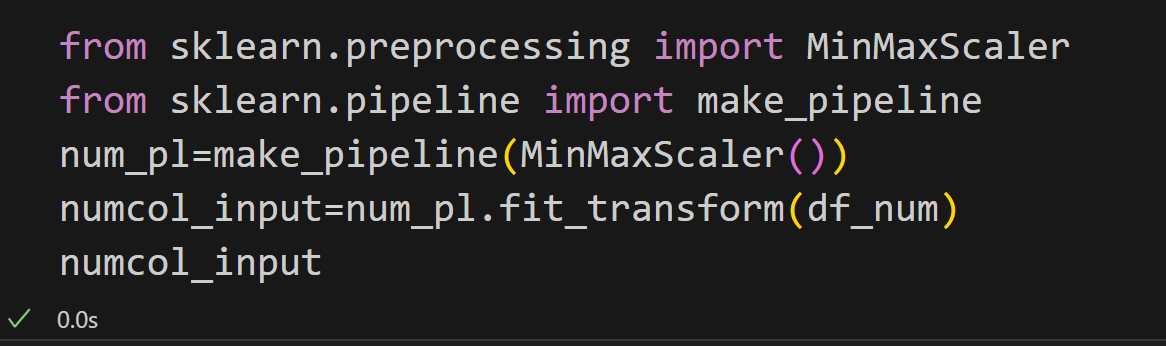
* 欄位
  + 學生名
  + 排名: 1-20名
  + 科系:CS(資訊工程系），EE(electric）(電機工程系)，EE(electronic)(電子工程系)，ME(機械工程系）
  + 生活費(income\_per\_month):範圍為30000元/月至60000元/月
  + 是否購買平板：y(有購買平板），n(未購買平板）
* 原始資料train\_data.xlsx由chatGPT指定範圍後生成，訓練集與測試集數量比例為7:3

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 訓練集(前14筆) | | | | |
| name | rate | subject | income\_per\_month | buy\_tablet\_or\_not |
| Olivia | 12.00 | CS | 45890.00 | y |
| Ethan | 5.00 | CS | 51235.00 | y |
| Ava | 17.00 | EE(Electronic) | 36741.00 | n |
| Mason | 3.00 | ME | 40219.00 | n |
| Isabella | 8.00 | ME | 54380.00 | y |
| Liam | 6.00 | CS | 39874.00 | n |
| Sophia | 14.00 | EE(Electrical ) | 57302.00 | y |
| Jackson | 10.00 | ME | 32987.00 | n |
| Emma | 20.00 | EE(Electronic) | 47250.00 | n |
| Aiden | 2.00 | EE(Electrical ) | 54106.00 | y |
| Mia | 1.00 | CS | 30562.00 | n |
| Lucas | 18.00 | ME | 57643.00 | y |
| Harper | 7.00 | CS | 48795.00 | n |
| Noah | 11.00 | ME | 35321.00 | n |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 測試集(後6筆) | | | | |
| name | rate | subject | income\_per\_month | buy\_tablet\_or\_not |
| Abigail | 13.00 | EE(Electrical ) | 41908.00 | n |
| Ella | 15.00 | EE(Electronic) | 53674.00 | y |
| Alexander | 16.00 | ME | 38899.00 | y |
| Grace | 19.00 | EE(Electrical ) | 45213.00 | y |
| Elijah | 4.00 | EE(Electronic) | 51247.00 | y |
| Scarlett | 9.00 | CS | 34125.00 | y |

1. 資料前處理

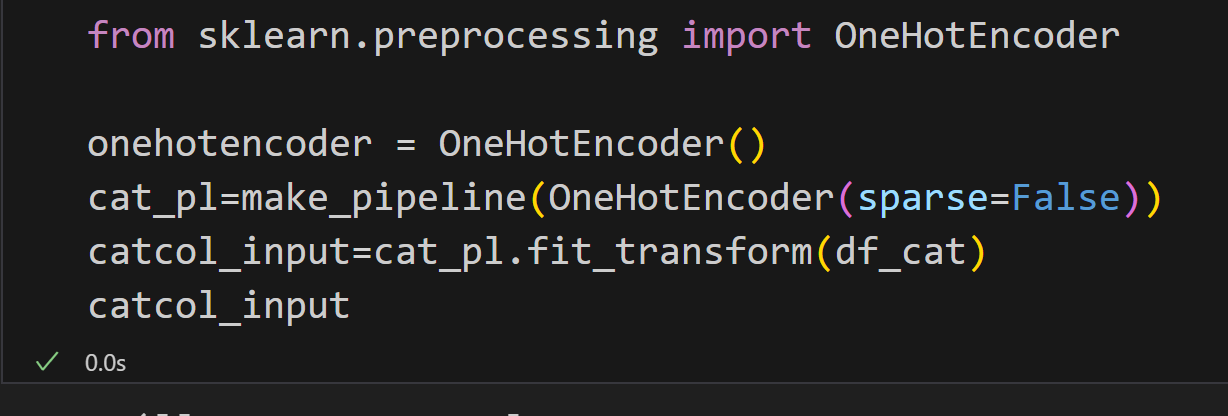
數值型資料:用線性壓縮壓縮



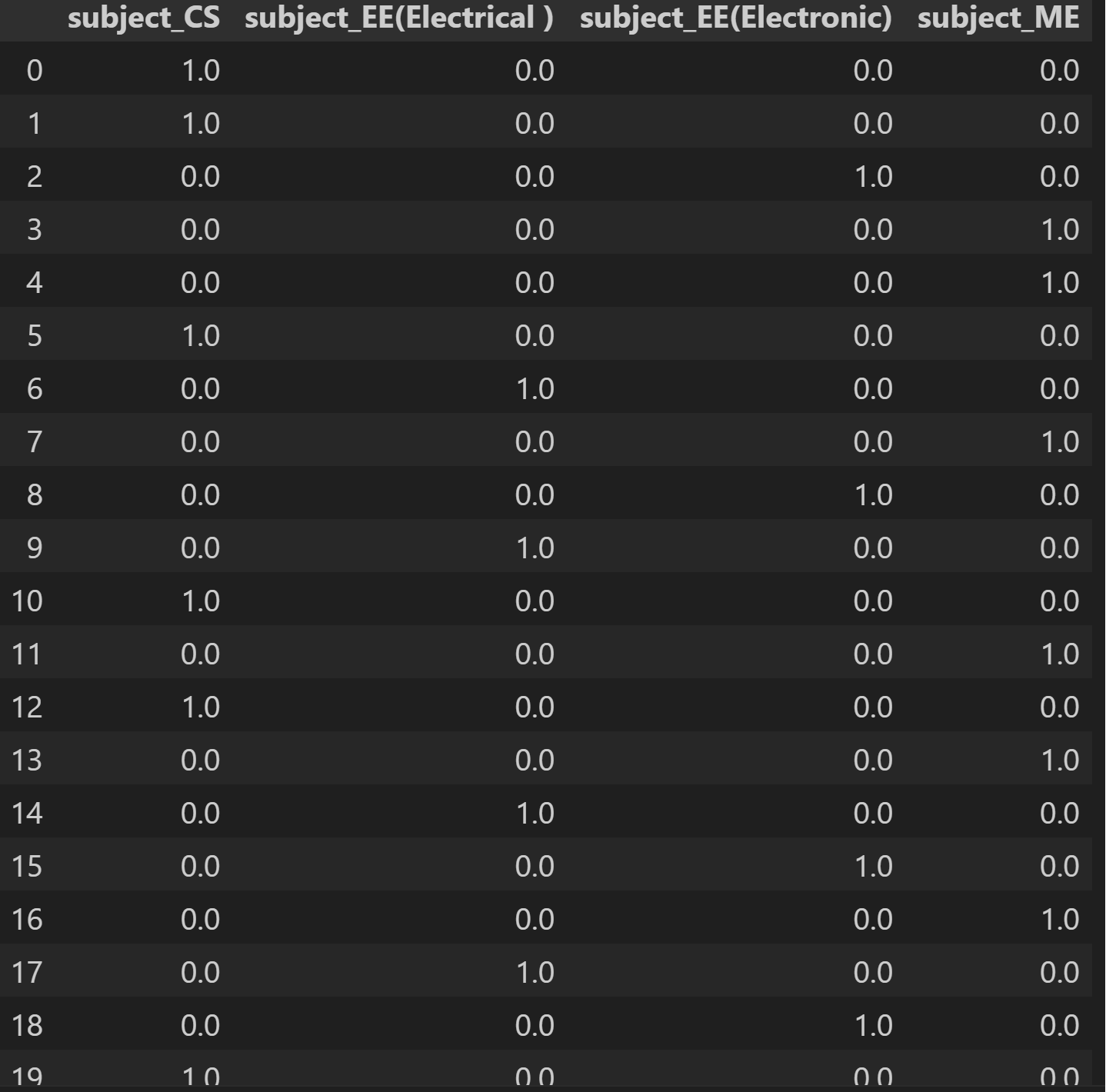
處理結果:



類別型資料:用onehot做讀熱編碼



處理結果:



1. 演算法及參數使用
   1. 演算法:決策樹
   2. 亂度值標準:entropy
   3. random\_state:0
2. 預測結果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 測試結果 | | | | | |
| name | rate | subject | income\_per\_month | predict | buy\_tablet\_or\_not |
| Abigail | 13.00 | EE(Electrical ) | 41908.00 | n | n |
| Ella | 15.00 | EE(Electronic) | 53674.00 | y | y |
| Alexander | 16.00 | ME | 38899.00 | n | y |
| Grace | 19.00 | EE(Electrical ) | 45213.00 | y | y |
| Elijah | 4.00 | EE(Electronic) | 51247.00 | y | y |
| Scarlett | 9.00 | CS | 34125.00 | n | y |

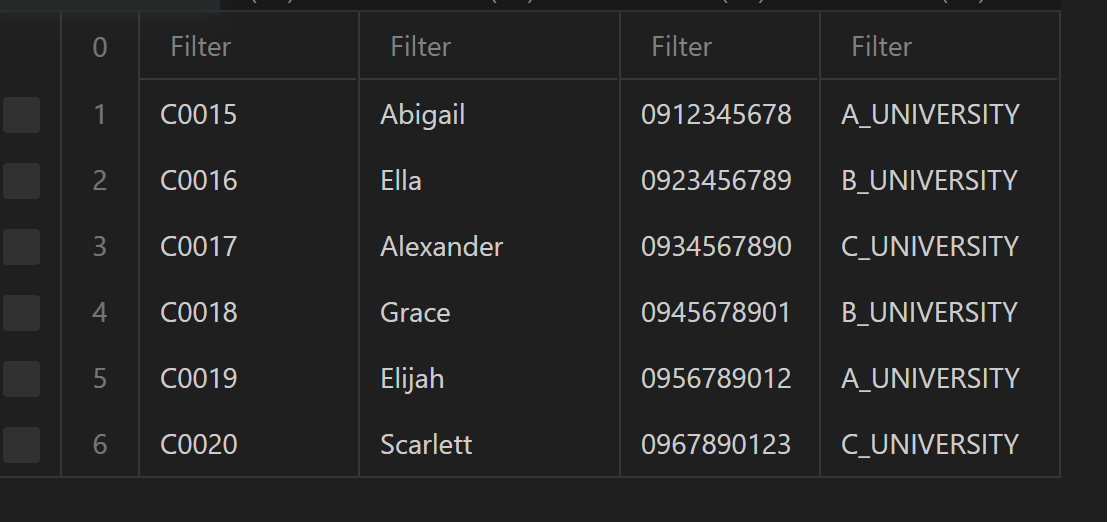
準確率: 0.667

混淆矩陣:

|  |  |  |
| --- | --- | --- |
|  | 預測會買 | 預測不買 |
| 實際會買 | 1 | 0 |
| 實際不買 | 2 | 3 |

因為此處之數據為隨機生成之資料，因此模型預測準確率只有0.667

1. 載入預測會買的客人至資料庫
   1. 原始顧客資料(資料庫db\_for\_tablet的資料表student，亦透過chatGPT生成)，顧客資料中只有測試資料中的學生



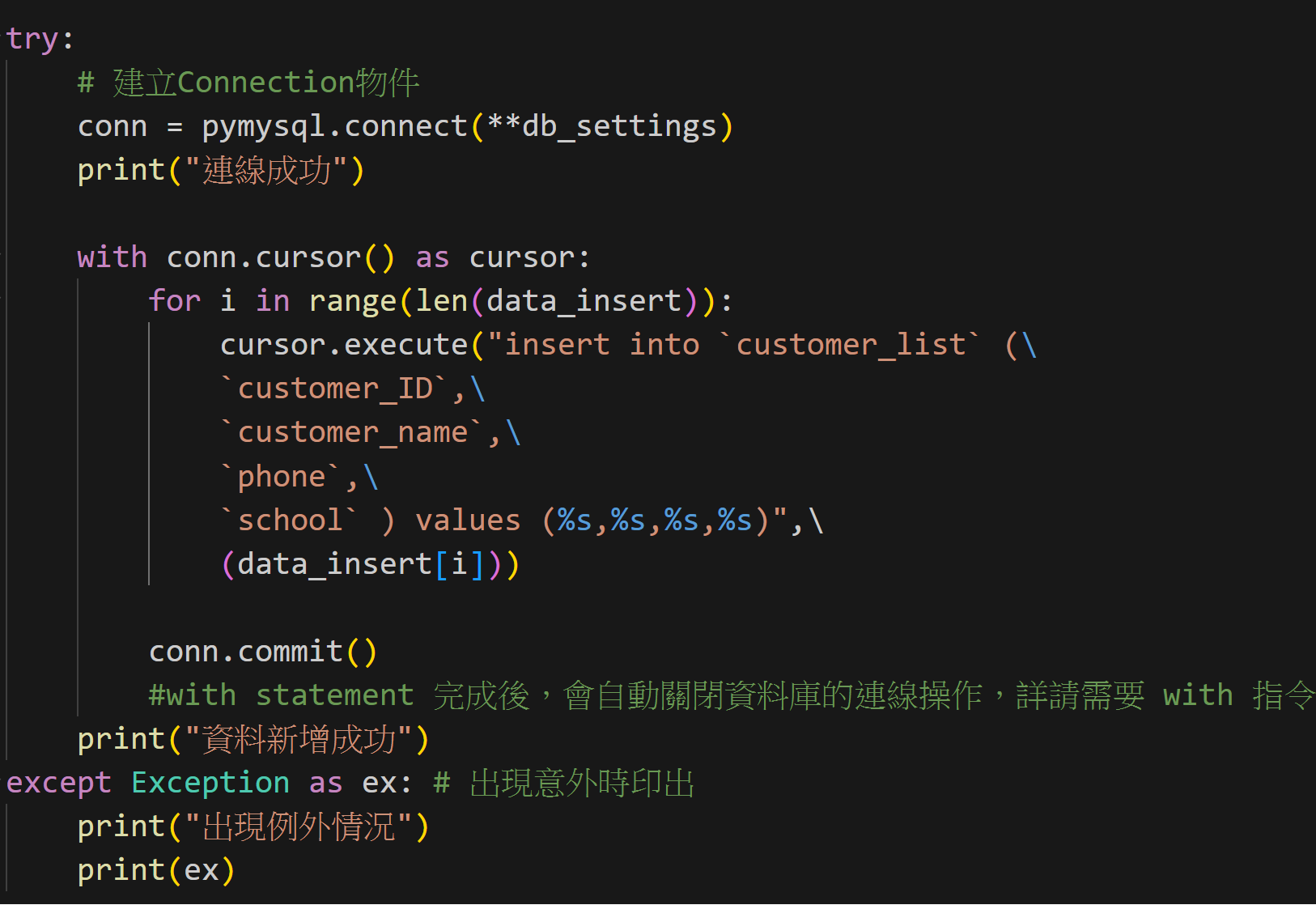
* 1. 找出測試資料中那些學生可能購買

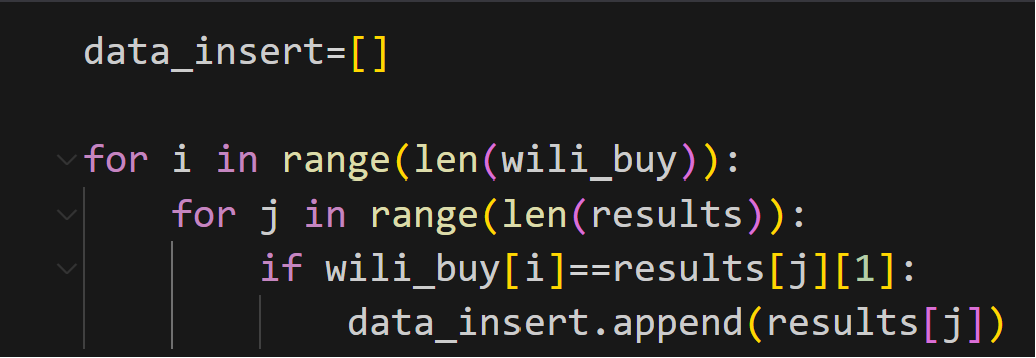


* 1. 設定資料庫連線



* 1. 將資料庫內容放入tuple以供使用



* 1. 篩出可能購買的學生資料
  2. 連線資料庫customer\_list將可能購買平板的學生載入資料庫中的資料表
  3. 至資料庫確認是否順利載入資料