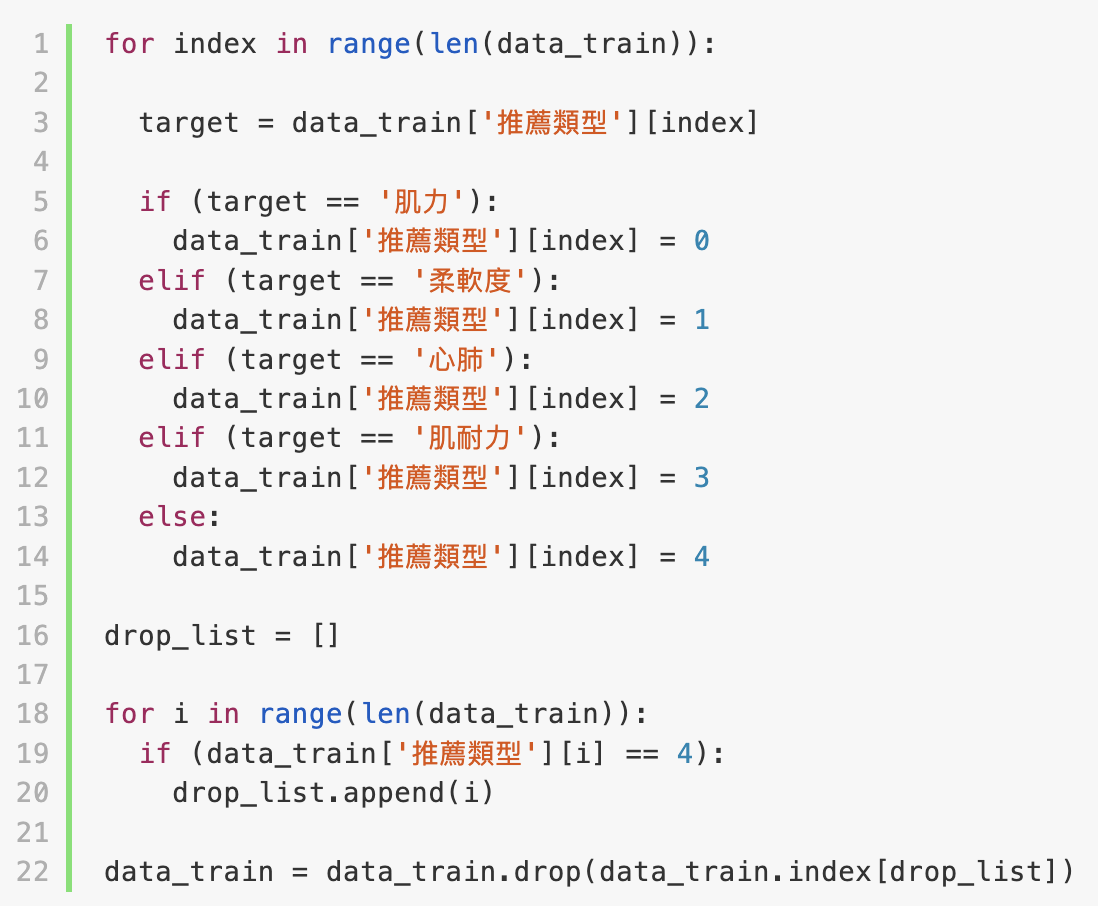
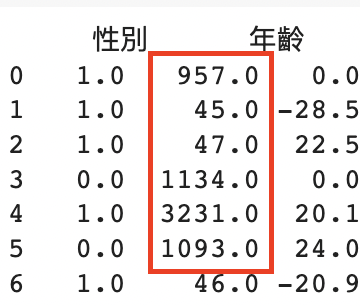
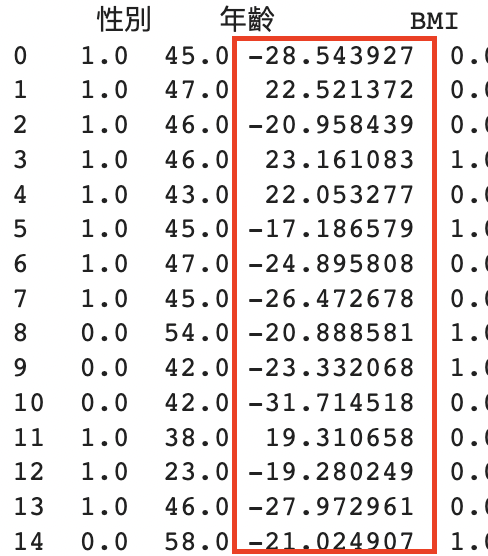
機器學習工程應用 期中競賽報告

|  |  |  |
| --- | --- | --- |
| 姓名:盧柏翰 | 學號:F14081088 | 系所:不分系大四 |

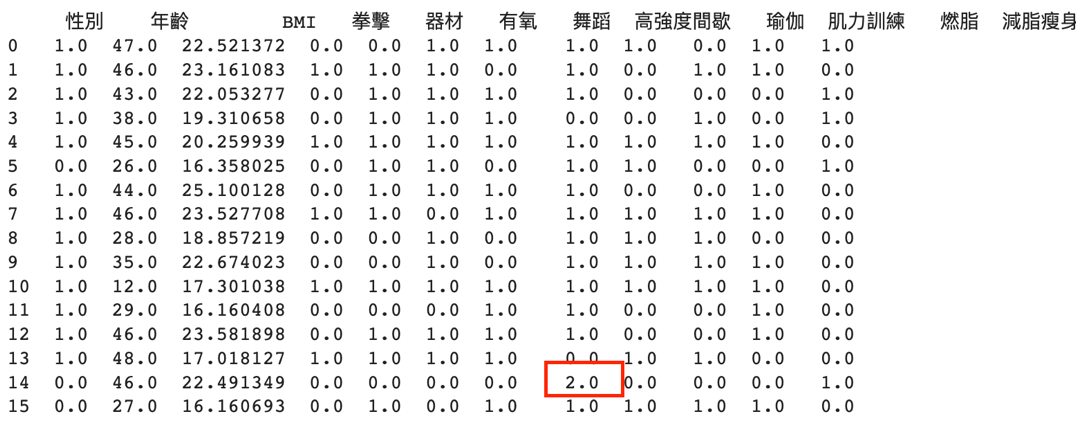
1. 資料前處理
   1. 先導入相關模組，像是 numpy、matplotlib 以及 pandas 後，即使用 pandas 讀取 csv 檔案。
   2. 針對 label 進行處理，按照題目要求，依照「推薦類型」進行分類，若有出現錯誤、無法辨識或是不在題目要求的類別中，則先將其的 label 值設為 4，並將其刪除之。一張含有 桌 的圖片

      自動產生的描述
   3. 下一步即針對年紀進行處理，有些年齡及其不合理，像是：負值、極大或是極小，因此，可判定其為不合理之資料，故刪除之。一張含有 文字 的圖片

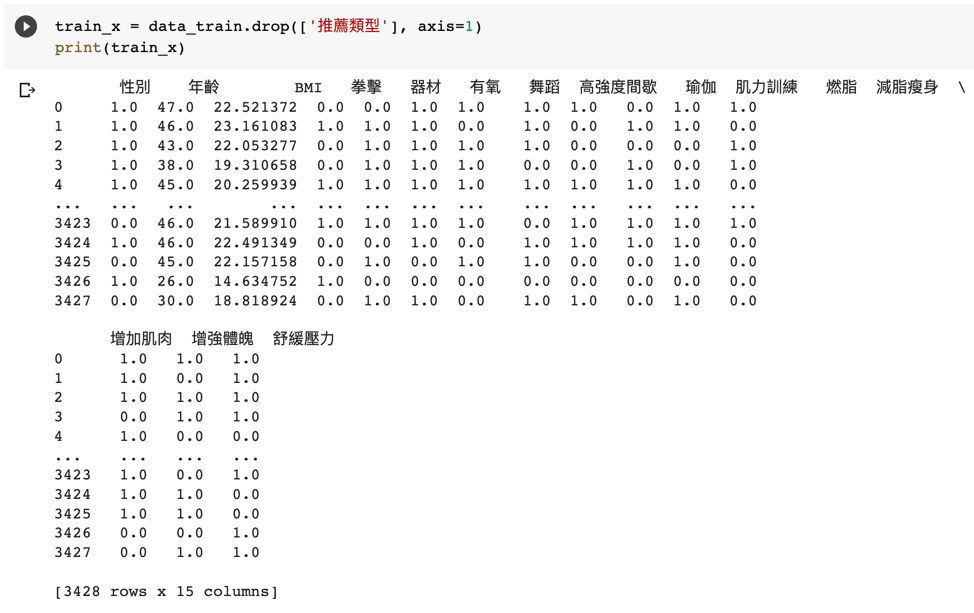
      自動產生的描述
   4. 接下來是處理 BMI 值，因為 BMI 的公式為 體重(公斤) / 身高2(公尺2)，故其應為正值，在 training 的資料中，BMI 為負值之資料應為不合理資料，故刪除之。



* 1. 針對各式興趣，資料的型態應為布林值，或是對應到整數或浮點數應為1（1.0）及0（0.0），非上述所描述之數值，一樣為不合理資料，故刪除之。



* 1. 將不合理的資料刪除後，將「推薦類型」（也就是結果）設為訓練之結果（train\_y），其餘資料去除推薦類型欄位後，設為訓練資料集（train\_x）。
  2. 上述有對資料進行刪除，都續對刪除後的資料進行重新編號，可透過 reset\_index(inplace=True, drop=True) 完成 index 之 reset。
  3. 為使用 sklearn 模組進行訓練，將上述 train\_y 及 train\_x 從 dataframe轉為 numpy 之資料型態（使用 .values）。一張含有 桌 的圖片

     自動產生的描述
  4. 另外，為了測試模型對於未看過資料之準確度，亦有測試將train\_y 及 train\_x 透過 train\_test\_split 分成 X\_train, X\_test, y\_train, y\_test 對模型進行驗證，而非只單純將結果上傳至 Kaggle 去測試模型對於未看過資料的準確度。

1. 訓練模型建立

我使用了四種分類器進行測試，分別為Decision tree、Random forest、SVM 及 GradientBoost。上面也有提到，為了確認模型對於「未看過」之資料的準確度，我也有將訓練集分成 train 和 test，並使用 cross\_val\_score 去追蹤模型的成效。目前而言，最佳的分類器為 Random forest，針對 Kaggle 平台的結果已達到 80% 以上的準確率。一張含有 文字 的圖片

自動產生的描述

1. 參數調整

除了上述所說，我使用四種分類器的模型進行比較外，我也同時使用 GridSearchCV，去測試多組參數，並印出其平均準確率，以評斷模型之效力。透過 GridSearch 即可透過類似暴力解的形式，將所給予的參數排列組合，找出成效最佳的參數組合和模型，CV 則可對資料集處理，將其拆分為 train 及 test（這邊應稱為 validation較佳，因為此題之test為未公開之數據），測試模型對於未看過資料的準確度。

1. 預測結果

我是將結果存入 list，因為先前除了測試分類器外，我亦有使用回歸分析做測試，方式為模型預測結果做四捨五入，以達成分類效果，然而，成效不盡理想。最後，為達成規定的輸出格式，我導入csv和itertools的zip\_longest，將輸出結果寫入csv 中，並加入 index（即為測試資料之資料數量），後再上傳至 Kaggle，目前準確率最高分之模型為透過GridSearchCV 找出的Random forest 分類器，效果達八成。

一張含有 文字 的圖片

自動產生的描述