Data Mining mini project

40947005S 何躍陽

1. 資料集特徵資料說明、屬性特性說明

0 p\_id int64 🡪個人ID

1 no\_times\_pregnant int64 🡪懷孕次數

2 glucose\_concentration int64 🡪葡萄糖耐量試驗後2小時的血漿葡萄糖

濃度

3 blood\_pressure int64 🡪血壓(的舒張壓)()

4 skin\_fold\_thickness int64 🡪三頭肌皮褶的厚度()

5 serum\_insulin int64 🡪2小時血清胰島素()

6 bmi float64 🡪身體的BMI指數 ()

7 diabetes pedigree float64 🡪有糖尿病的譜系

8 age int64 🡪年紀

9 diabetes int64 🡪分類結果：class 1代表測試為陽性

1. 對特徵做甚麼樣的分析? 哪些前處理? 採用哪些特徵? 原因?
   1. 分析資料筆數、dtypes、最小值
   2. 先檢查有沒有missing data，發現missing data的value為0，所以用**SimpleImputer()**填值成mean
   3. 雖然有一兩個的特徵有不少的missing data，但其實訓練只有把p\_id拿掉(僅用來辨識資料用)，其他的特徵都有用到
2. 基於什麼理由選擇哪個分類器?

訓練的時候有測試以下幾種分類器：

MLPClassifier DecisionTreeClassifier

GaussianNB RandomForestClassifier

AdaBoostClassifier

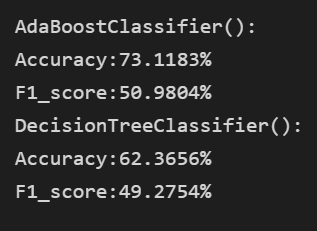
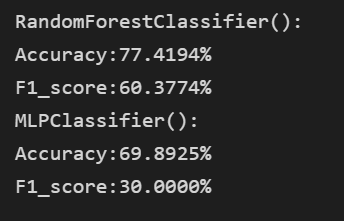
也執行了多次看看不同的不同的train跟test測出來的Accuracy，發現

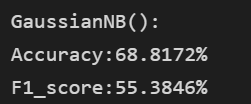
RandomForestClassifier的分數浮動最小，也相對最高分，其他的如：AdaBoostClassifier、GaussianNB等等，雖然也可以訓練到跟RandomForestClassifier接近的成績，但RandomForestClassifier還是相對分數最高、最穩定的那個，所以最後才會選擇用它來進行分類

1. 採用的評估指標結果與觀察

評估指標使用了F1\_score跟Accuracy

用不同分類器測出來的Accuracy都可以有在65%以上， F1\_score則浮動比較大，但會集中在50~60左右





1. 上傳至 kaggle 的預測結果及截圖測試的分數

