

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)(mm Hg)
4	skin_fold_thickness	int64	→三頭肌皮褶的厚度(mm)
5	serum_insulin	int64	→2 小時血清胰島素(mu U/ml)
6	bmi	float64	→身體的 BMI 指數 (kg/m ²)
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

(2) 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

(3) 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

(4) 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

```
RandomForestClassifier():      AdaBoostClassifier():
Accuracy:77.4194%              Accuracy:73.1183%
F1_score:60.3774%             F1_score:50.9804%
MLPClassifier():              DecisionTreeClassifier():
Accuracy:69.8925%              Accuracy:62.3656%
F1_score:30.0000%             F1_score:49.2754%
```

```
GaussianNB():
Accuracy:68.8172%
F1_score:55.3846%
```

(5) 上傳至 kaggle 的預測結果及截圖測試的分數

✕ Submission Details



my_submission.csv

Complete (after deadline) · 1m ago

Score: 0.75324

Private score: 0.75324

UPLOADED FILES



my_submission.csv (1 KiB)

